UNIVERSITY OF OULU
Faculty of Humanities
Information Studies

Lucie Hradecká

TAILORING ORGANIZATIONAL SUPPORT SERVICES FOR RESEARCH DATA MANAGEMENT IN THE OPEN SCIENCE CONTEXT

Service self-assessment tool pilot

Master's Thesis

Oulu 2023

Lucie Hradecká
Master's thesis, November 2023, 140 pages + 3 Appendices
University of Oulu, Faculty of Humanities, Information Studies

ABSTRACT

TAILORING ORGANIZATIONAL SUPPORT SERVICES FOR RESEARCH DATA MANAGEMENT IN THE OPEN SCIENCE CONTEXT: Service self-assessment tool pilot

This thesis aimed to explore how support services for research data management (RDM) and its open science aspects can be tailored in research organizations to the needs of service users, while also taking into account the external requirements set by policies, legislation, and recognized best practices. The theoretical part outlines the background of open science in sociology and philosophy of science, and how it relates to research data. The various aspects of RDM and potential problems are discussed, followed by a brief review of previous research on the organization of RDM support services. The empirical part is based on the service design approach, gathering information on service users' needs and experiences in the first stage and involving various stakeholders in the formulation of steps towards solutions in the second stage.

The purpose of this thesis was not to collect data for service development at a specific institution, instead, the goal was to evaluate whether the selected methods and tools can provide valuable information for the development of RDM support services. The data was collected at VTT Technical Research Centre of Finland in 2020. A survey questionnaire was used to explore what kind of needs, attitudes and experiences researchers had in topics where RDM support services could be useful. The results were analysed quantitatively with basic descriptive statistics and qualitatively with content analysis in case of open-ended questions. In the second stage, the Research Infrastructure Self Evaluation (RISE) model developed by the British Digital Curation Centre was tested in six semi-structured interviews with stakeholders in RDM services, to explore how these experts would evaluate the usefulness of the RISE model for service self-assessment. The interview data were analysed using theory-driven content analysis. The discussion also explored the overarching research question whether the survey worked together with the service self-assessment interviews as useful tools for RDM support service development.

The results confirmed that the selected tools can offer some useful information for RDM support service development, however, further research would be needed to develop the proposed approach into a standard methodology framework for RDM support service self-assessment, which could be used by organizations to evaluate their own service portfolio. The survey questionnaire could be refined and accompanied by an instrument for follow-up interviews. The suggested modifications to the definitions of capabilities and their maturity levels in the RISE framework to better suit the Finnish research environment would have to be confirmed by at least one more pilot evaluation.

Keywords: research data management, open science, service design

TABLE OF CONTENTS

1 INTRODUCTION

Open science is a paradigm of conducting and disseminating scientific research, with focus on transparency and equal access. It aims at increasing the quality, reliability, and societal impact of research. Open access to research outputs, such as publications and the underlying research data, are important parts of open science. The transition from traditional models of scholarly communication towards open science in practice requires incentives and sufficient infrastructure. The Open Science and Research Initiative (Avoin tiede ja tutkimus -hanke, ATT), funded by the Ministry of Education and Culture and carried out in 2014-2017, laid the foundation of transition to the open science paradigm in Finland. National level policy and infrastructure development continues in the National Coordination for Open Science and Research, also funded by the Ministry of Education and Culture, and organised by the Federation of Finnish Learned Societies (Tieteellisten seurain valtuuskunta, TSV). (Forsstöm & Kutilainen 2019, 4, Avointiede.fi 2020).

The long span of these projects illustrates that transferring the open science principles into practice is a complex task with many challenges. The involvement of academic libraries in this task originates from their basic goal to support the mission and objectives of their institutions - higher education and research. As open science becomes more prominent in higher education and research environment, academic libraries must adapt their services to include support for the practices of open science. Open science values transparency and openness in the scrutiny of reliability, which has brought more attention to the underlying research data in addition to publications (Miedema 2022, 202). Academic libraries have been subsequently attempting to ascertain if and how they should support the process of managing and publishing research data. One of the elements that has become a part of academic library services is the support of research data management planning. (Rice & Southall 2016, 1-2, 54, 157.) Publishing research data requires good research data management, which in turn requires planning ahead to seamlessly introduce good practices from the start. Research data management (RDM) is an innate part of any data-based research, it includes a complex set of practices and decisions, and efficient support requires multidisciplinary skills and knowledge. (Corti, Van den Eynden, Bishop & Woollard 2014, 29.)

Although open science has been adopted in many national and organizational research policies as well as by many funding bodies, there is less clear evidence that it has become an organic part of researchers' workflows across disciplines. Organizational support services are positioned between policymakers and researchers who must implement these policies in practice. In this thesis, academic support services are understood as not only helping implement policies delivered from top down, but they also have an opportunity to hear the feedback and experiences of those who must apply policies in their academic work. The principles of transparency, openness and democracy characteristic for the open science approach also drive us to embrace the diversity of academic disciplines and their various traditions, and to open a dialogue to develop feasible open science practices which really improve the quality and impact of research, with sensitivity to the contexts of various disciplines and the bigger picture of scientific communication, career advancement and evaluation.

This thesis aims to evaluate a concrete tool which can be used to open the dialogue and encourage discussion about research data management and its open science dimension. First, we explore how open science has become a widely accepted paradigm and how research data is understood in the open science context. The theoretical part continues to define the main aspects of research data management and the academic support services required to help with related questions and problems. The empirical part will focus on the development of these support services in a way that is informed by service users' needs as well as relevant policies and the organizational context.

1.1 Research problem and approach

The focus of this thesis is on communities of practice formed by the researchers as well as the support staff in research organizations. Both sides should be included in the development of services to reflect users' needs as well as the affordances and gaps in the existing service infrastructure. The first problem investigated in this thesis is how an organization can gain insight into the practices and support needs in research data management (RDM) specific to their researcher community. RDM is a part of conducting scientific inquiry, which inevitably involves discipline specific variation. The workflows, problems and support needs are therefore not easily generalized, and investigation of research data management is often practice-oriented. (Kruse & Thestrup 2018, 1.) After reviewing common challenges and support needs in RDM and its open science aspects described in literature, a questionnaire was developed striving to map how these topics are perceived by researchers. This questionnaire was piloted at VTT Technical Research Centre of Finland in June 2020 to test whether the resulting information can help increase the understanding of RDM practices, challenges and support needs in the organization.

The second research problem concerns how the organization can review its current RDM support services. Needs for service developments can be recognized based on the data on service user needs and challenges, however, users may not always be aware of already existing services, or support needs driven by external requirements of research funders, legislation, or national and institutional open science policies. There is also the problem of limited resources and what is realistic to provide within each organization’s means. Chapter 3, describing RDM practices and common challenges, shows that RDM has many aspects, and efficient support requires diverse expertise as well as technical infrastructure. Therefore, it is useful to define a baseline set of the various technical and advisory services needed to support RDM. However, in the practice-oriented approach of this thesis, a feasible framework defining recommendations for RDM services should allow flexibility to tailor support services to each organization’s profile, researchers’ needs, community practices and existing service infrastructure.

For this reason, service evaluation was conducted using a tool based on the Capability Maturity Model as described in chapter 5.4. This tool defines RDM services and their capabilities, and each organization can evaluate what maturity level the current services correspond to and whether there is a need to develop the services to a higher maturity level. This evaluation should not be understood as a benchmarking tool between organizations, where each organization should strive for the highest level in every capability. Instead, it can help organizations conceptualize a reasonable basic set of services and prioritize higher level in those services which are most relevant to them. Defining services as capabilities can also allow flexibility of how the services are organized in each institution's structure. The evaluation tool was also piloted at VTT Technical Research Centre of Finland during November-December 2020 as an instrument for semi-structured interviews with experts providing relevant support services. The interview instrument was enriched with selected data from the survey to combine the information about user needs with the experts' knowledge about current services and resources available for development. The pilot aimed to evaluate whether this combination of user needs survey and service self-assessment can help organizations review their RDM support services in a way that uncovers gaps and also helps prioritize development goals based on realistic possibilities and user needs. Figure 1 illustrates how the background information from chapters 2-4 lead to the selection of the RISE tool and to the design of the user needs survey, which was also used to include the users' voice in the service self-assessment pilot.

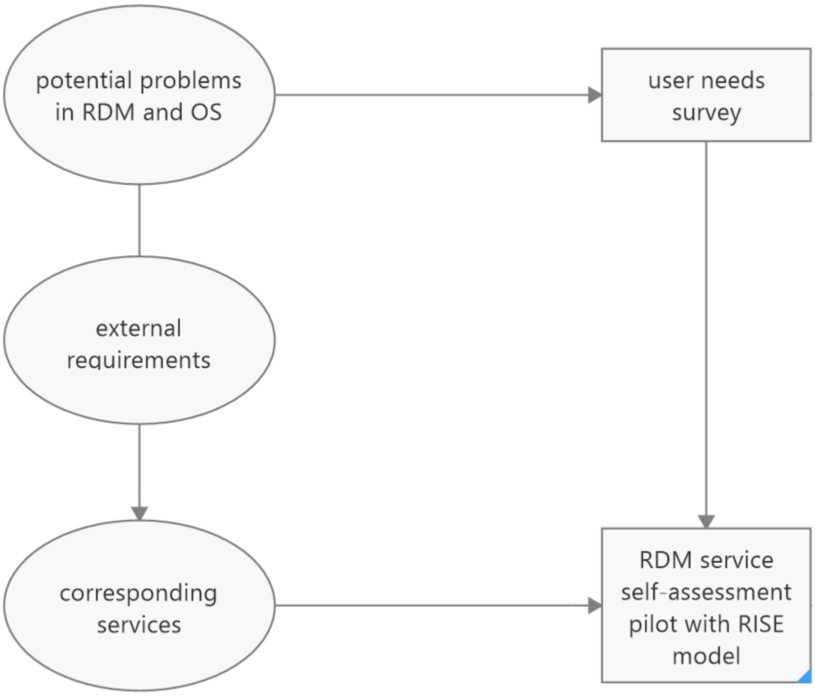

potential problems in RDM and OS
user needs survey
external requirements
corresponding services
RDM service self-assessment pilot with RISE model

Figure 1: Background information and methods.

1.2 Key terms and concepts

Research data in this thesis refer to the factual digital records in various formats (numerical, textual, audiovisual, …) used as primary sources for scientific research, which are commonly accepted in the research community as necessary to validate research findings. Supporting documentation such as laboratory notebooks, preliminary analyses, or communications with colleagues are a part of conducting research and handling research data, but not considered research data per se. Physical objects such as laboratory samples or test animals are not considered research data, as this definition limits data to factual digital records. (Organisation for Economic Co-operation and Development (OECD) 2007, 13-14, White House Office of Science and Technology Policy (OSTP) 2013, 5, 20-21, Rice & Southall 2016, EU 2019/1024.)

Research data management (RDM) refers to the practices applied to the data reused, collected or produced in course of the research. This includes processes and activities conducted during the various stages of the research life cycle: planning, documentation, organisation, storage and sharing during the active phase of the research, publishing and long-term storage. (Corti, Van den Eynden, Bishop & Woollard 2014, 2, Van den Eynden 2018, 43.)

Open Science has been defined as follows*:*

> "In its broadest definition, Open Science covers Open Access to publications, Open Research Data and Methods, Open Source Software, Open Educational Resources, Open Evaluation, and Citizen Science. But openness also means making the scientific process more inclusive and accessible to all relevant actors, within and beyond the scientific community." (Miedema 2022, 244.)

While "science" can be understood as referring only to natural sciences, in this thesis *science* and *research* are used interchangeably, as an umbrella term for scientific inquiry in any discipline including social sciences and humanities.

Open access (OA) means that a research output or other information resource in a digital form is available online free of charge without unnecessary limitations of its reuse and dissemination, within the legal framework (“libre OA”). “Gratis OA” means that the information resource is not behind a paywall, but there are still restrictions regarding its reuse. There is always a moral obligation to attribute credit to the authors regardless of OA status, and the source should be appropriately cited when used. (Suber 2012, 1-5, 21, 65-66, Budapest Open Access Initiative (BOAI) 2023, Max-Planck-Gesellschaft 2023.)

FAIR data principles were coined in 2016 as an approach to good research data management, making data Findable, Accessible, Interoperable and Reusable by humans and machines. These principles emphasise that in order to validate research results or reuse research data in further investigation, the data have to be managed and shared in a way that makes them sustainably accessible and understandable. (Wilkinson et al. 2016.)

2 THEORETICAL BACKGROUND

This chapter introduces the theoretical background of open science and the values of transparency and societal impact, which are connected to a shift in thinking about scientific knowledge building and its relation to society. A brief history of initiatives and policies will illustrate how open science has become an ingrained part of conducting publicly funded research in the Finnish academic environment.

The subsequent chapters will focus on research data in the context of open science, the possibilities and limitations of making research data openly accessible, and the research data management practices necessary to implement the principles of FAIR data and making the data "as open as possible, as closed as necessary". This background information illustrates the wide span of challenges and potential support needs of researchers who need to manage their research data well and follow the open science principles.

2.1 The philosophical and theoretical roots of open science

In his book *Open Science: The Very Idea*, Frank Miedema (2022) traces the roots of the current open science movement to the epistemology of pragmatism. In the pragmatic view, scientific knowledge is created and accepted as valid in communities of practice. The process is based on shared ideas of ethical and reliable methods in the communities, rather than on one normative ideal of objective and neutral "scientific method". If these shared ideas and practices are made transparent and accessible for examination, we can start to understand how scientific inquiry is really performed in practice and how judgements on reliability or significance are made. Pragmatism also supports the idea that science should be responsible to the society it might affect, and which also provides the public funding of research. This understanding of practices of scientific knowledge creation and views on the relationship between science and society also sparked critical examination of values affecting the rewards and incentives favouring certain practices or research outputs in the management of public research organizations and research funding. These three aspects in turn influenced the development of open science as a paradigm of conducting and communicating about research, in which transparency of research data and its management are important.

The dominant tradition in philosophy and sociology of science until the 1960s was a theory of scientific investigation and scientific knowledge creation which Miedema (2022, 18-19) refers to as “the Legend”, or the Standard Model. This view on epistemology in science builds upon Cartesian rationality and was later influenced by positivism, especially the logical positivistic tradition of the Vienna Circle. It is therefore also called the analytical, empirical, or logical-positivistic tradition. The Cartesian dualism of the observer versus the observed, and fact versus value, supported the view that scientific knowledge consists of objective facts free from the bias of cultural and personal value judgements. This knowledge would be generated using a formal mathematical method and the foundation consisting of objective universal principles, making such scientific knowledge inherently reliable.

The positivists later argued that such given universal foundation would not be scientific, because it was not acquired empirically. In their view, scientific theories are accepted or rejected after rigorous experimental (empirical) testing and scientific debate. The empirical, formal, logical method was however still viewed as a guarantee for objectivity and neutrality which separates values from facts. This approach is based on the tradition and methods of natural “hard” sciences where theories and laws can be empirically tested and ideally also expressed formally (mathematically). Under the same criteria, knowledge created in the “soft” social sciences and humanities may not be considered scientific. Miedema also notes that the concept of objective empirical observing of nature can be challenged when we consider that research is often done in laboratory settings or under the condition of *ceteris paribus* (“other things being equal/constant”) which are not common in the real world. (Miedema 2022, 18-19, 54.)

The “Standard model” is defined by Miedema as a combination of “the Legend” with the classical sociological image of science developed by Robert Merton and his followers between 1930–1970. They viewed science as a special category of human activity defined by the scientists' altruistic pursuit of the truth. This is done in an open community of academics engaging in sceptical, but also fair and honest debate about each other's work, with the goal of collectively achieving the best knowledge. The unique internal criteria and norms of best knowledge are all the scientific community as an autonomous social system needs to govern itself. While the existence of incentive and reward system in which researchers get credit for their work, advance their careers and become elite in their field is acknowledged, it is seen as secondary and as a logical consequence of scientific activity, reflecting the natural order. Merton did however point out some unwanted effects of this stratification, such as the Matthew effect ("the rich get richer" – further accumulation of advantages to those who achieve elite status). Merton seemed to believe that the elite scientists would deal with receiving unfair advantage with integrity, and that the stratification would not be problematic, which Miedema considers an idealistic view outdated in the eyes of the current reader. (Miedema 2022, 20-21.)

Pragmatism represents a major shift in thinking about scientific knowledge creation. Its roots go back to the work of C. S. Peirce at the end of the 19th century. Peirce was a proponent of fallibilism, accepting the possibility of error in every claim for truth. Scientific knowledge is not uniquely exempt from revision and correction. What makes scientific knowledge different from other types of claims is its higher confidence in logic of procedure and a "self-corrective" process. In the pragmatist view, knowledge is recognized as scientific when consensus is reached in a community of inquirers in which hypotheses are continuously tested – this intersubjective (social) and iterative character of inquiry and knowledge building is what decreases the bias of limited individual perspectives and enables the self-correction of science. Some authors call this epistemological behaviourism: an approach focused on "the social process by which a community of inquirers come to produce and accept knowledge and beliefs." The normative ideal "unique scientific method" was critiqued for not reflecting this social process and being detached from real-life practices in science. Pragmatism called for the descriptive historical and sociological investigation and understanding of scientific communities, their values, reasoning, and practices in the intersubjective knowledge building process. (Miedema 2022, 39-40, 53, 112, 117-118.)

While in the “standard model”, reliability of scientific knowledge claims is evaluated based on unique internal criteria and norms leading towards the pursuit of best knowledge, researchers in the pragmatist approach aim to understand how these criteria are developed in practice. They point out that scrutiny of evidence (such as research data) and the procedures by which the evidence was obtained are based on established habits, a tradition that is not necessarily under scrutiny itself. The status-based hierarchy within the community can also influence what is considered ‘*the* scientific opinion’. These collectively held beliefs are then transferred to new researchers in their academic education. (Miedema 2022, 11, 44.) Some argue that meaning cannot be derived by formal mathematical or logical method directly from data, models, pictures or other evidence. The conclusions based on evidence are also not judged by the “observed nature” itself, but they are evaluated in the scientific community. Scientific knowledge is therefore not fully “objective”, rather, its reliability is evaluated in intersubjective agreement. This however does not have to mean lesser reliability. Claims validated in a community of inquiry and in the social world can be seen as more reliable and robust as they are scrutinized from more points of view. If we maintained that scientific knowledge building is objective and neutral, we would avoid the responsibility to be transparent about the socially accepted criteria and to reflect whether they may be affected by non-epistemological values such as hierarchy or tradition. (Miedema 2022, 44, 50-51, 125.)

Ravetz (1996, 31-33, 47-48) discussed the dissipation of sense of community with innate codes of behaviour and ideals of best knowledge in the “industrialized science”. This is connected to the increase in scale of the research sector organized in research institutions, increased costs of conducting advanced research and competition for funding, and rapid changes beyond the control of the scientific community itself, more characteristic for the world of industry and trade. The amount of information generated and published has also been increasing, including new ways to share information about research which do not fall under the peer-review process characteristic for traditional scientific publications. Ravetz (1996, 49-50) goes on to point out that such environment can contribute to increased incidence of “shoddy science”. He claims that it is a “dirty secret” usually not acknowledged in philosophy of science that most researchers have encountered examples of bad research when trying to use others’ published results. The quality control system of scientific publications is therefore not perfect. Authors may submit low quality work to increase their number of publications, and publishers in turn can be motivated by profit to accept such work. When the sense of community with shared goals and codes of behaviour is diminished, so is the social control which demands adherence to certain socially agreed quality criteria.

In addition to internal reliability and quality judgements, there is also the question of the value of scientific knowledge in society. Those who defended "the Legend" argued that the acceptance of fallibility and the social (intersubjective) character of scientific knowledge building could invite relativism, and lead to loss of the authority status of scientific knowledge as more certain than personal opinions and experiences. Miedema (2022, 114) however argues that pragmatism with its realistic, open, and democratic view of science allows a way to communicate about science and engage with larger society which is more responsible than maintaining the image of certainty and authority based on unrealistic premises. Because the products of scientific inquiry are not always directly applicable or understandable to larger audiences, it is the quality and reliability of the inquiry process itself that should be able to justify the worth of scientific activities in society (Ravetz 1996, 42). Miedema (2022) concludes that in the various practices of scientific inquiry, there are overarching elements increasing trustworthiness of the process in which scientific knowledge is created: robustness, independence, openness and transparency, continuous rigorous testing, scrutiny and debate which lead to accepting, improving or rejecting claims. Testing of claims in various theoretical contexts is complemented by testing in practice and on real world problems. Reliability judgements are formed in a social process based on sharing ideas and explaining the methods and results, including the empirical research data, and opening them up to debate and scrutiny. (Miedema 2022, 28, 33-34, 56, 62.)

There is a tension between the values of engagement and responsibility towards society versus the autonomy of research and academic freedom, which was largely discussed already in the 1960-1970s. As noted earlier, "The Legend" presumes the neutrality and separation of science from non-scientific values, politics, and society in its pursuit of truth. This approach was important in times when it was necessary to establish the independence of academic research and institutions separate from church, state, and politics. We however need to reconcile the value of independence with the fact that science and society are developing and interacting with each other in a common public sphere. (Miedema 2022, 6, 147.) Miedema (2022, 135) further describes that in the 1960s, the interactions of science and society in common public sphere became apparent in debates on "environmental issues, nuclear energy, radioactive waste and the nuclear arms race, the first signs of the energy crisis and a war in Vietnam for which the motives and logic had long evaporated". The concern and protests made it obvious that the public felt alienated and wanted to be included in issues where politics were interacting with science and technology.

The pragmatist approach suggests that rather than focusing only on the intellectual pursuit of truth, scientific inquiry should also be motivated by situations or problems in the (natural and social) real world which prevent people from "leading the good life". Such investigation is conducted in interaction and cooperation with the environment and with other human beings. Engagement with society is therefore necessary. The cooperation is however effective only when it is based on ethical and democratic grounds and open communication – power play and hierarchy hinder the goals of scientific inquiry, and lack of trust in experts hinders the relationship of science with public. Expert knowledge can often be rejected by members of public when it is not perceived as relevant to the needs and the social situation in question, and when it comes as authoritative knowledge without transparent explanation of the process in which it was scientifically validated and why it is considered reliable. The strive to engage with societal problems in research agenda setting, to encourage participation, and the concept of science furthering the common good are reflected in the values of open science. (Miedema 2022, 112, 118-119, 141-143.)

The traditional approach views all external influence on science as damaging and inherently corrupt. The problems of such harmful external influence are seen as only afflicting applied science, while basic curiosity-based research is seen as pure, value neutral and autonomous. It is however difficult to fully separate basic research from possible practical use. (Miedema 2022, 120-121.) As Ravetz (1996, 35) pointed out, the invention of nuclear weapons was only possible with the contributions of the "pure" basic research in physics, and even in physics, which is usually considered an example of logical, formal, basic “hard” science, experimental work can call for the managerial and political relations to acquire funding for large infrastructure such as particle accelerators. Concerns about abuse of science and threats to free scholarship were largely discussed in the critical theory approach, influenced by the development of critical social science theory and first represented in the works of authors such as Habermas, Foucault and Bourdieu. The critical approach brings attention to the influence of governments’ military interests and the economic interests of large multinational businesses, of which we should be aware, and science should be reclaimed as an “emancipatory force in society”. (Miedema 2022, 130.)

Habermas saw science and technology as “drivers of economic and technologic innovation shaping and dominating our social life”**.** While they can improve everyday life, the influence can also become dominating and repressive when the “needs and problems of the diverse publics” are not included. He suggested a “pragmatistic” democratic model in which critical interaction leads to scientifically informed discussion, allowing legitimation of policies for the public. Further investigation of this model of democratic deliberations discussed the problems of ethics and intentions of parties engaged in the deliberations, achieving consensus and avoiding conflict at the expense of supressing some voices, as well as the issue of language and translating scientific knowledge into a form that will enable the public to be well-informed. (Miedema 2022, 134-135, 138-140.) Miedema (2022, 122, 135-136, 138-139) further conveys that engagement with public striving for inclusion of the less powerful can bring beneficial external influence on problem choice in research and the formation of governmental policies, while science should be protected from unwanted influence of powerful economic or political entities, and the “vulgar democracy” of ill-informed majority vote.

The previously discussed epistemological views and adequacy judgement criteria are not only abstract concepts, but the otherwise tacit assumptions in different traditions of what is considered high quality research have explicit impact in established quality control and evaluation practices such as peer review, grant application committees, academic promotions committees etc. (Miedema 2022, 32). Values affecting problem choice, such as strive for relevance to societal issues or preference for “pure” basic science also play a role. The socially agreed criteria of excellence and significance of research are incentivised in funding distribution and career advancement, which in turn affects what researchers must aim for to succeed. Miedema (2022, 82) quotes that early literature on management of organized science from the 1950s was aligned with the normative ideal of self-governing scientific community with its own internal value system, mentioned in the beginning of this subchapter in connection to Merton’s views on the incentive and reward system as a logical part of scientific activity, reflecting the natural order. When thinking shifted towards investigation of science as a practice, the literature from the 1970s onwards also tried to examine the social system behind governance of science with its rewards and incentives.

Miedema (2022) and Ravetz (1996) both pointed out the need of governments and funding bodies to justify their investments into science and measure its contributions. As societal impact takes longer time to be noticeable and is not easy to measure, short-term quantitative indicators such as number of publications, citations and patents became the staple of performance evaluation metrics on various levels, from national and institutional to the evaluation of individual researchers. For a long time, these metrics were not systematically and openly scrutinized, which could lead to misunderstanding and misuse, for example using journal-level metrics such as the Journal Impact Factor (JIF) to judge the quality of individual articles. As argued above, the rules of evaluation also impact the behaviour of researchers. Metrics such as the JIF favour international journals published in English and attracting wide readership, leaning towards more theoretical basic research. While applied research, multidisciplinary collaborations tackling problems in society or publishing in national languages could have larger societal impact, these activities tend to score fewer points in metrics. The quantitative, formal, analytical type of research that is typical for hard sciences is more valued in the traditional concept of excellence and reliability based on “the Legend”. This preference for quantitative “hard” methods started to influence other fields such as linguistics, sociology, or economy, to the detriment of qualitative research. Replication studies and sharing of negative results also contribute to the robustness of scientific knowledge but are less desirable as novel results are more likely to be published and get into prestigious journals. (Miedema 2022, 74-76, 80, 87-88, 110-111.) To continue, Miedema (2022, 110) remarks that in the past decade the international scientific community has been discussing how to fix this “broken system”.

Ravetz (1996) pointed out the effect of external evaluation criteria on the motivation to publish more and the incidence of “shoddy science”. Miedema (2022, 68) refers to a few high-profile cases of fraud in the Netherlands regarding falsification of research data which brought more attention to the topic. While high-profile fraud scandals are rare and attract attention, it is more likely that less obvious shortcuts affecting reliability of research can be taken not with the intention to mislead, but as a survival strategy in a highly competitive system that incentivizes certain behaviour. Making underlying research data available could improve the quality control of articles presenting results, however, this was not incentivised in the traditional system. In the “credibility cycle” based on traditional ideas of excellence, committees and advisory boards consisting of the elites of the discipline decide on promotions, appointments, and funding. Quality is often measured in quantitative indicators such as number of publications and citations or JIF. Hypercompetition for merit and funding discourages multidisciplinarity, diversity, and working in teams. In the open science way, the credibility cycle is enriched with engagement of societal stakeholders in problem choice, valuing societal impact and the use of research by others in academia and society, open access to publications, data sharing, and improved peer review practices which may include open peer review or post-publication peer review. (Miedema 2022, 68-69, 85-87, 90, 202.)

2.2 Open science initiatives and policies

Before open science developed as the umbrella term uniting a number of practices related to transparency, engagement, inclusivity and pragmatist-leaning notion of excellence and credibility in science, various movements started raising awareness of these issues and values. One of them was the Open Access movement, reacting to the rapidly increasing prices of subscriptions to scientific publications. Ever since the first scholarly journals were established in 1665, researchers have contributed their articles with the goal of impact and addition to knowledge building, but without any financial profit. Publishers on the other hand needed to generate revenue to cover their expenses. (Suber 2012, 10, 18.) Digitalisation has made it easier to disseminate publications to a broader audience, however it has also challenged the business models behind academic publishing and raised questions about the costs. Researchers contribute not only their manuscripts but also peer review and editorial efforts without compensation or profit. The time dedicated to tasks related to publishing is commonly covered by salaries provided by public funding from institutional budgets or grants. Institutional libraries then again spend large sums on purchasing access to read the resulting publications, and for a private person the costs make scholarly publications even more inaccessible. Some have been questioning the logic and fairness of such system. (Corti et al. 2014, 2, Rice & Southall 2016, 147-148.)

The movement pushing for open access to scientific publications started to be more prominent in the early 2000s. In 2001 Open Society Foundation organised a meeting in Budapest that led to publishing of the Budapest Open Access Initiative in 2002 (Budapest Open Access Initiative 2023). In 2003 the Bethesda Statement on Open Access Publishing (Brown et al. 2003) expressed similar goals for the community in the USA. Both declarations were authored by groups consisting of library and publishing professionals, researchers and representatives of scientific societies and funding bodies. The Budapest Open Access Initiative declaration was subsequently signed by thousands of individuals and organizations such as researchers, universities, libraries, journals, publishers or learned societies (Budapest Open Access Initiative 2023).

Also in 2003, the Max Planck Society and the European Cultural Heritage Online (ECHO) project organized a meeting in Berlin that would later be followed up by a series of Berlin Open Access conferences. At this first meeting (now known as “Berlin 1”), delegates from the Max Planck Society presented the principles which later became the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Max-Planck-Gesellschaft 2023). The Berlin Declaration referenced and built upon the previous two statements. These three seminal initiatives agree that it is important to remove barriers and provide equal access to scientific information for the benefit of the scientific community, education, and society. The declarations anticipated that such cultural shift would, however, require development of new operational models and information technology tools.

In the three declarations, Open Access (OA) means a publication is sustainably available online, free of charge and without unnecessary limitations of its reuse and dissemination, retaining the authors' right to be properly acknowledged. (Budapest Open Access Initiative 2023, Max-Planck-Gesellschaft 2023, Brown et al. 2003). This definition was further refined later. A digital publication available online free of charge without restrictions in relation to copyright and licensing, other than strictly necessary within the legal framework, is referred to as libre OA. Gratis OA means the paywall has been removed, but there are still restrictions regarding reuse. The term gold OA was coined for OA to content in scholarly journals, and green OA for content deposited to digital repositories. To recover lost revenue from subscription fees, journals often charge an open access publication fee to be paid by the author instead of the reader. As journal articles have traditionally not generated profit for the authors unlike other types of publications which can generate royalties, the academic journal article was the most logical starting point for advancing OA. There has also been discussion and movement to provide OA to more diverse research outputs, including research data. (Suber 2012, 1-5, 65-66, 111-112.)

The free software movement also influenced thinking about open data and open code in research context. In reaction to the commercialization of software development, the free software movement introduced the idea that digital information 'wants to be free'. This does not necessarily mean 'free of charge'. Rather, it refers to the freedom to handle and use the information to co-create the best possible solutions as a community. This seems to resonate with the values of collective scrutiny, openness and engagement in pragmatism and open science described in the previous subchapter. While freedom to use and build upon information is recognized as valuable, digital information is also a valued resource which can become capital. (Rice & Southall 2016, 147-150.)

In the public sector, access to digital information can be viewed as fundamental right and means to promote democracy and civic engagement. In the EU Directive 2019/1024 on open data and the re-use of public sector information (also called Open Data Directive, formerly known as the PSI Directive) open data is defined as “data in an open format that can be freely used, re-used and shared by anyone for any purpose”. The directive addresses public sector or administrative data, but also takes into account research data generated in publicly funded research. Public sector documents and data are meant to be “open by design and by default”, although certain cases call for protection of public interest objectives, for example when public security or personal data are concerned. Documents held by public or academic libraries, museums, and archives, for which third parties hold the intellectual property rights (IPR) are excluded from the scope of the Directive. The directive requires member states of the EU to adopt national “open access policies” to make the data created during publicly funded research openly available, with the exceptions of concerns related to IPR, personal data protection and confidentiality, security as well as legitimate commercial interests. (EU 2019/1024.)

In Finnish legislation, the Directive (EU 2019/1024) is implemented with the law Laki julkisin varoin tuotettujen tutkimusaineistojen uudelleenkäytöstä (713/2021, published in Finnish and Swedish, can be translated as Law on reuse of data resulting from publicly funded research). The law is applicable to published data resulting from publicly funded research. Data is defined as digital documentation collected or created as part of scientific research which can be used as evidence, or which are necessary to validate research results. Academic publications are not subject to this law, and documents held by libraries, museums and archives are excluded in compliance with the Open Data Directive. The Finnish national law instructs that if research data are published, they must be made available free of charge, under unbiased terms of use. Terms of use must not exclude commercial utilization. Any legal or practical restrictions of use must be made public. (FINLEX 713/2021.)

Open science is also connected to the development of various movements concerning Responsible Research and Innovation (RRI). RRI aims to increase impact of science by opening up to society and diverse stakeholders, while also paying attention to integrity and social responsibility and the ethical and legal aspects. The movement promotes “science for and with society”, encouraging prioritizing societal challenges in research problem selection and engagement of society in research. All of the approaches introduced in this subchapter contributed to the implementation of the values of inclusivity and accessibility of science within and beyond the scientific community in the open science practices. These practices include open access to publications, open research data and methods, open-source software, open educational resources, open evaluation, and citizen science. Transition to open science also requires changes to the system: the infrastructure, culture, incentives, and rewards. (Miedema 2022, 183-185, 191, 244.) Therefore it is important that many major research funding bodies (for example European Commission 2015 and Research Council of Finland 2023a) have included open science practices in their policies. In Finland, the national-level vision, mission and goals are outlined in the Declaration for open science and research 2020-2025, supplemented by policy documents and recommendations for the various open science practices. The declaration can be signed by organisations and individuals who wish to endorse and support the objectives of the Declaration. (Avoimen tieteen koordinaatio 2020.)

Miedema (2022, 39) interprets the concept of paradigm in research as a set of scientific values, rules, techniques, and methods, which are not completely separate from related social, cultural, ethical or practical values, and which affect how results are scrutinized and accepted or rejected in the community of inquiry. We can view open science as a similar paradigm consisting of a set of practices reflecting the value of transparency in the scrutiny of reliability of research, equal accessibility, and engagement with scientific community as well as larger society. History of science documents communities of inquiry within which different schools and paradigms would usually coexist until one became dominant. This paradigm shift is not easy to achieve and change often requires the authority of someone in position of intellectual power. (Miedema 2022, 39, 211.) Although funders and national policies promote open science from the position of power, it is not necessarily the type of intellectual power accepted internally in communities of scientific inquiry. Chapter 2.1 suggested that some epistemological approaches in science can perceive external influence negatively, and information presented as authoritative can be rejected if not perceived as relevant to the experience of the community. Being mindful of diversity in methods and adequacy judgement criteria in different types of research seems to be built into the Open Science way of thinking, and researchers have been involved in the development of the movement, for example by authoring or signing the open access declarations. However, it is important to enable further engagement and make sure we do not approach open science as the new “Legend”, prescribed as the norm without scrutiny.

In addition to diversity of practices and differing epistemological views, geopolitical differences and competitiveness also contribute to the realization that there cannot be one global community with open science practices as a united way of conducting research. Differences in academic cultures, socio-economic politics and national history can influence how academic research is institutionalized in different regions and countries. (Miedema 2022, 195-196, 214.) In his book Miedema (2022, 203-205) also quotes the concerns of experts from outside the wealthy Global North. In their perspective, open science practices may not be effective in the goal of tackling inequalities, instead, they may in some respects trigger Matthew effect for rich and powerful countries, referring to the concept of accumulated advantage in society. Open access publishing has transferred the costs from subscribers to authors, and those who cannot afford expensive open access publication fees are once again excluded. Rich countries could also benefit more from open data due to better access to funding to develop new ideas based on shared data.

To give more space to the concerns of scholars from the Global South, Miedema (2022, 205-207) also reprinted a talk by professor Mamokgethi Phakeng from the University of Cape Town, who emphasized the importance of transparency and openness to scrutiny for building democracy, and the importance of societal impact of research especially for developing countries. She however also noted that researchers cannot “do impact”, their research only enables impact to happen, and it is not always predictable. Impact for the public good is important, but we should also not fall victim to a utilitarian approach, which would exclude curiosity-driven research of larger questions without obvious opportunity for immediate application. Miedema (2022, 140) poses the moral question whether researchers should be made to work only on problems that are estimated to create the most impactful knowledge. He concludes that the democratic deliberations characteristic for open science also include hearing out the researchers. This view is adopted in this thesis as the reason why user needs, experiences and attitudes are important not only for service development, but also for understanding the communities services are supposed to help, and opening up dialogue.

2.3 Research data

The term *research data* usually refers to digital objects in various formats. In the broad sense, any digital information created as an outcome of conducting scientific research can be considered research data. Various research fields and disciplines utilise different methods and instruments and collect or create various types of data, which is why an umbrella definition is usually somewhat vague. Depending on the discipline and methods, research data can mean for example statistics, experiment outcomes, measurements produced by various instruments and tools, survey data, recordings and transcripts of interviews, numerical data, databases, geospatial data, or digital documentation of physical objects, field work or observation. Various file formats and structured information about the data can be collected into a database or form a dataset. (Corti et al. 2014, 57, Kruse & Thestrup 2018, 2, Spichtinger & Siren 2018, 12.)

The diversity of digital objects which can be considered research data in different disciplines makes it challenging to find a common language and terms that would be understood the same way across fields. For example, in the humanities, the term “research data” does not have a long tradition. In fine arts and the related research fields, the qualitative and audiovisual materials which would fit the definition might not be understood as “research data”, resulting in these disciplines being less represented in the discussion and development of research data management policies and guidelines. Although the term *data* in general refers to digital objects, it may be beneficial to broaden the scope when guidelines are developed, also considering the special characteristics of physical materials supporting research in certain disciplines. (Davidson 2013, 84-85, Rice & Southall 2016, 19, Ala-Kyyni, Korhonen & Roinila 2017, 26.) This thesis will however use the definition of research data as digital objects.

Regardless of the format or creation method, research data can be defined as the factual digital records in various formats (numerical, textual, audiovisual) which are collected or produced as part of the research process and used as primary sources for scientific research, and commonly accepted in the scientific community as necessary to validate research findings. A systematic, partial representation of the research subject, for example data validating certain results presented in a publication, forms a dataset. Supporting documentation such as laboratory notebooks, preliminary analyses, drafts of scientific papers, peer review reports or communications with colleagues are a part of conducting research and handling research data, but not considered research data per se. Physical objects such as laboratory samples or test animals are not considered research *data* in most definitions. (Organisation for Economic Co-operation and Development (OECD) 2007, 13-14, White House Office of Science and Technology Policy (OSTP) 2013, 5, Rice & Southall 2016, 20-21, EU 2019/1024.)

The factual records that may be used in research are not always primarily created as research data, meaning they are not produced or collected as an outcome of the research process. Sometimes research can utilise data that was primarily created with other intentions, and which is not considered *research* data by itself outside of the context of reuse for research purposes. Public sector information, also called administrative data or process-produced data, is often reused in research. This data is usually collected or produced as the result of government administration and public authority operation. Public sector information can include for example various maps, statistics, traffic data or business registers. Research can also utilise digital data from the Internet and social media, such as the content or user statistics. (Rice & Southall 2016, 21, Spichtinger & Siren 2018, 13.) Analogue materials and digital data collected by archives such as photographs and documents are important primary sources in some disciplines, for example history research (Rice & Southall 2016, 20-22).

2.4 Open research data: community practices and external requirements

In the previous subchapter, research data was defined as "commonly accepted in the scientific community as necessary to validate research findings". However, subchapters 2.1 and 2.2 discussed that there is no one "scientific community", but there are many methods in research and various practices in credibility judgement of scientific knowledge claims. It may seem obvious that open research data should support transparency, collaboration, or validation of findings. However, in the pragmatist approach, descriptive examination is encouraged to understand real practices. Literature points out some examples of disciplines where sharing research data has been perceived beneficial.

Sharing and reuse of research data have become a practice organically in some disciplines that require collaboration or where data have potential to be used to answer many research questions. For example, in astrophysics and high energy physics, data generated with expensive unique equipment like observatories or particle accelerators is shared in larger collaborations. In climate research, historical weather data or even old diaries describing the weather and environment may be reused to understand long-term developments. Data sharing also played a key role in aggregating enough data to understand the human genome. (Corti et al. 2014, 9-10.)

In social sciences, large-scale surveys and census data have significant reuse potential. Research data collected with contribution from citizens may be similar to public sector data, comparable to a public good which should be open for reuse for the benefit of society. To enable access to reusable data in social sciences, specialized data archives started being established already since the 1960s. The Consortium of European Social Science Data Archives (CESSDA), a major provider of data services in the social sciences, was founded in 1976. (Rice & Southall 2016, 3-6, 10-11, Kruse & Thestrup 2018, 5.) Experiences with reusing data collected by others brought attention to the issues of time and effort necessary to make the data understandable and reusable after the research project: the data organization and documentation necessary for archiving, or the responsibility to answer potential questions were not usually taken into account in project funding (Stouthamer-Loeber & Bok van Kammen 1995, 110).

While in some disciplines, data sharing practices have developed organically, there are also top-down requirements resulting from policies such as those mentioned in chapter 2.2. In life sciences, the National Institutes of Health of the USA adopted a policy on open research data in 2001. The main idea was that data should be shared as openly as possible, while taking into consideration the protection of personal data, confidentiality, and patentable inventions. Applicants were required to submit a plan specifying whether the data can be made openly accessible and how it will be shared. Since then, many public research funders have adopted similar policies, with the goals of enabling open access to research outputs, making the results more transparent and verifiable, and avoiding inefficient funding of duplicate efforts. (Rice & Southall 2016, 69.)

The 2004 Declaration on Access to Research Data from Public Funding by The Organisation for Economic Co-operation and Development (OECD) was followed by the Principles and Guidelines for Access to Research Data from Public Funding (The Organisation for Economic Co-operation and Development (OECD) 2007). These documents emphasize the aim to maximize return on investment of public funding of research, but also acknowledge that there should be balance between the public benefit of open access to scientific information and the concerns regarding for example IPR, personal data protection and confidentiality. (Organisation for Economic Co-operation and Development (OECD) 2004, 2007.)

Another well-known example of funders advocating for open data is the Open Research Data Pilot (ORD) launched within the Horizon 2020 framework of the EU which ran from 2014 to 2020. The pilot was extended in 2017 to be included by default in all new grant agreements, requiring that the underlying data behind research publications generated in the project should be made available. However, exceptions were reserved for valid reasons similar to the previously mentioned policies: IPR, ethical issues such as security and personal data protection, confidentiality and commercial exploitation of the results. There was also an option not to make research data openly available if it would jeopardise the main objective of the project. The policy defined that research data should be made “as open as possible, as closed as necessary”. (Spichtinger & Siren 2018, 14-16.)

Funders’ requirements are creating external pressure to open research data, which may be perceived negatively especially in fields where data sharing and reuse do not have a tradition within the community. As discussed in chapter 2.2, researchers should also be heard in the deliberations on policies and best practices. On the other hand, it is useful that funders participate in the change of the incentive system to include more diverse type of research outputs than the old “credibility system” prioritizing points in publication-based metrics. The funders’ requirements reflect the understanding of funding bodies that the research data management necessary to make data accessible and reusable to others is part of the work conducted in the funded projects, and funders should commit to supporting this work. (Rice & Southall 2016, 69.)

The Finnish national Declaration for Open Science and Research was followed by a policy on open research data and methods (National Coordination of Open Science and Research 2023). The policy component on open access to research data is applicable to research data produced or used by researchers working in or affiliated with Finnish research organizations, or projects which have received funding from a Finnish research funding body. The working group who drafted the contents of the policy included representatives of researchers. The Finnish policy also follows the principle of research data being as open as possible and as closed as necessary, encouraging openness if there are no restrictions resulting from legal, ethical, contractual, or other significant limitations. Responsible management and openness of data applied in the context of the research discipline are seen as supporting the goals of quality and impact, and therefore compatible with the researchers’ right and freedom to choose suitable best practices for conducting scientific inquiry and disseminating the results. The policy recommends including research data related practices in the rewards and incentives systems, emphasizing that responsible evaluation should acknowledge the diversity of methods and the types and role of research data in various disciplines. (National Coordination of Open Science and Research 2023, 6-9, 15.)

2.5 Value of open research data for society, science, and individual researchers

The legislation and policies referred to in previous subchapters show that open access to data is often viewed as ethical in terms of societal impact, supporting democracy, or responsible and transparent research. Research data may play a role in increased access to evidence-based information for decision making, and advancement of research and innovation. Openness and transparency make data available for scrutiny, leading to higher accountability and supporting the quality control mechanisms enabling the self-correction of science, discussed in chapter 2.1. Sharing data for verification and reuse also helps maximize the benefits of research supported by public funding or participation of the public. In disciplines working with human participants, such as social sciences or medical research, the impact on participants must be considered. Reuse of existing datasets lowers the burden on respondents while potentially raising participation rate when the population is not overwhelmed by too many different requests to participate in research. Being able to reuse and combine datasets also helps access a wider sample to investigate topics important to the larger society. (Stouthamer-Loeber & Bok van Kammen 1995, 110, European Commission 2012, 39, EU 2019/1024, Corti et al. 2014, 1, 11-12, Van den Eynden 2018, 47.)

Piwowar, Day and Fridsma (2007) state that while there is a general consensus about the benefits of open data to society and research, it is important to also consider the benefits and burden for the individual researcher carrying the responsibility for the work involved in making data reusable and accessible. Van den Eynden (2018, 49-53) has collected and analysed the results of various interview and survey studies of researchers' attitude towards open research data. The factors affecting researchers' motivation to make research data openly available can be classified as individual or institutional.

Individual factors include the perceived benefits, or on the other hand perceived risks related to the researcher's academic career. Research data management and opening the data are time consuming tasks, and those using their time and effort to open research data may fear falling behind in tasks garnering more merit in career evaluations, such as article publishing. (Van den Eynden 2018, 45-46.) Piwowar and her colleagues (2007) also refer to publications and citations as a "currency of value" in academic careers. They studied whether opening the underlying research data alongside publications leads to more citations, using data from the field of cancer microarray clinical trials. In the studied sample, publications also sharing the underlying research data were cited almost 70 percent more frequently. A positive effect of opening the data on publication citations has also been noted in astrophysics (Drachen, Ellegaard, Larsen and Dorch 2016).

The researcher's aspiration to increase the quality, responsibility and impact of their research can form another individual motivational factor. The studies also suggest motivation is higher when the individual's research data management skills are already good or when they perceive sharing research data as a positive concept linked to academic altruism (viewing the advancement of scientific knowledge as a collective effort). Uncertainty of own research data management skills or necessary steps, as well as high perceived amount of effort necessary to make research data available can however negatively affect this motivation to open research data. (Van den Eynden 2018, 44, 49.)

Institutional factors include the norms within the research discipline, requirements of funding bodies or academic journals, and institutional policies. The funding allocated specifically to research data management, efficient data sharing infrastructure as well as institutional support services tend to have a positive effect on motivation. The same has been observed for common policies, guidelines, and clear instructions on the associated legal issues, all of which advance the appropriate use of the data and receiving credit as the creator. (Van den Eynden 2018, 47-49).

2.6 Limitations and challenges of open access to research data

Alongside the potential benefits and motivating factors, there are also many aspects limiting the possibility or motivation to open research data. As mentioned in the previous subchapter, making data well organized and documented so that it can be understandable and reusable for others requires a significant amount of work. Academia is a highly competitive environment, where chances to win grants and advance in one's career are limited, and the evaluation systems primarily reward for other achievements, mostly writing academic articles. While advocating for the benefits may help some understand why sharing research data might be important, it does not create more time in the researcher's schedule. A researcher who is already very busy with working on their current project while also making time for publishing and applying for new grants must prioritize their tasks. Putting additional effort into research data management and opening their data may not be the highest on their priority list. Especially early career researchers are at risk of being assigned more tasks in research data collection and management for collaborative projects, which will not necessarily be recognized in authorship of publications that create the "currency" in evaluations. (European Commission 2012, 31, Van den Eynden & Bishop 2014, 25, Rice & Southall 2016, 137, 152, Van den Eynden 2018, 44, 47.)

The understanding of research data as a reusable resource may be problematic for some researchers. On one hand, some feel that their data cannot be useful or interesting to other researchers or the public. While there may be discipline-specific differences in how data can be reused, and some studies may not generate data with as many opportunities for reuse as others, perceived value and reuse potential of the data can often be underestimated. On the other hand, researchers may see the reuse potential, and feel that it would be unfair to let others use the results of their hard work for their research and publications. The authors of the data may have their own ideas for further research and fear that these ideas will get scooped by others. However, data does not have to always be openly accessible immediately, allowing their authors some time to finish the current project and develop further research ideas. The development of citable records for published research data also makes it easier to receive credit if anyone does publish new results based on the data before the original authors. The fear that research data could be misunderstood or misused was also mentioned as a limiting factor. While genuine risk of misuse should be considered in decisions whether data should be kept closed, some fear of misunderstanding can be mitigated with support in good research data management and documentation to lower the risk of incorrect interpretation. (Corti et al. 2014, 10-11, Van den Eynden & Bishop 2014, 25, Rice & Southall 2016, 120-123, Van den Eynden 2018, 48.)

There are also institutional factors which can negatively affect the possibilities to make research data openly available. Adverse institutional factors can be for example insufficient funding allocated to the development and maintenance of the necessary IT infrastructure and storage services, insufficient incentives to make the effort worthwhile, such as inability to receive credit for making data available, lack of national or local guidelines, policies and strategies, discrepancies in data ownership and mandate to deposit data to long-term storage after the project, lack of skills or capacity for data management and lack of standards to make the data reusable. (European Commission 2012, 28-30.)

Although proper IT infrastructure, clear guidelines, and legislation as well as support services can help avoid unnecessary limitations to opening research data, there are also necessary restrictions to openness due to legal and ethical requirements. These matters such as personal data protection and confidentiality were found to be a common barrier to open access to data according to the report by European Commission (2012, 28-29). Intellectual property rights (IPR), such as copyright, patents or trade secrets may apply in some projects and legal support services are needed to help researchers define appropriate limitations. It is usually recognized by research funding bodies that the protection of IPR or commercialization plans may grant an exception from open data requirements. (Corti et al. 2014, 143-144, Rice & Southall 2016, 23-24.)

Clearly defined authorship, rights ownership and conditions of reuse are in any case a prerequisite to opening of research data. Academic research projects are often collaborative and international, involving partners from various universities or from industry. Support services may be needed to draft agreements defining the rights ownership, and other rights and responsibilities regarding the research data (for example, how data will be used in resulting publications, who is responsible for data storage during the project, and who will ultimately make the decisions regarding openness after the project). It is not unusual for universities and research institutes to include clauses in employment contracts which grant rights to research data to the employer. While authorship and the right to be cited and receive credit remain with the researcher who produce the data, rights transfer to the employer makes it easier for the institutional services to store data on behalf of researchers and provide continuity of access when project members leave the institution. Support services may also be required to help select a license defining conditions of reuse. If conditions of reuse are unknown, data are not reusable in practice, as the parties reusing it cannot be certain whether they are in breach of IPR. (Corti et al. 2014, 143-144, 147, Kuusniemi 2019.)

Personal data protection was already mentioned as a common barrier to opening research data. In research with human participants, it is likely that the data may be linked to the identity of specific participants. Personal data refer to such information that identifies a person or makes them identifiable. According to the General Data Protection Regulation (EU 2016/679), EU law regulating data protection and privacy in the European Economic Area, personal data include direct and indirect identifiers such as "a name, an identification number, location data, an online identifier or one of several special characteristics, which expresses the physical, physiological, genetic, mental, commercial, cultural or social identity of these natural persons." The regulation also defines special categories of sensitive data, which reveal the person's ethnic origin, political opinions, religious or philosophical beliefs, sexual orientation and behaviour or membership in a trade union. Biometric identifiers and data regarding health also belong to the special categories. The GDPR, as well as Finnish Data Protection Act (Tietosuojalaki, 1050/2018) aim to protect the citizens' right to privacy by defining how and for what purposes personal data and their sensitive categories can be collected and processed.

Both European (EU 2016/697) and Finnish (1050/2018) legislation allow the collection and processing of personal data in scientific research as a task carried out in public interest. The personal data must be minimized, meaning only the personal data necessary for the purpose of the research project can be collected and processed, and they should only be retained for the period necessary to carry out the research. The data must be stored securely so that they cannot be accessed by unauthorized persons. When possible, direct identifiers should be removed or replaced by code (pseudonymized). Research participants must be informed in a privacy notice about the purpose of the data collection and how the data will be processed in the project. Researchers must also confirm the research subjects' consent to participate. If sensitive personal data are processed, researchers must seek out ethical pre-review prior to data collection. (EU 2016/697, FINLEX 1050/2018.)

Personal data collected in research cannot be made openly accessible. The resulting dataset can only be opened if all personal data are removed, and the data are fully anonymized. Privacy protection and participants' consent must be prioritized over any potential benefits of openness. This is recognized in open data policies mentioned in chapter 2.4. In case participants have given consent to reuse of data in further research, the data could potentially be made available with managed access to authorized persons only. However, the GDPR formulates that informed consent can only be given when participants are informed about the specific purpose of data collection and processing (EU 2016/697). This makes it difficult to gain consent for yet unspecified future research. The Finnish act on secondary use of health and social data (Toisiolaki, FINLEX 552/2019) defines certain conditions under which such data can be accessed for research purposes in a secure operating environment.

In addition to personal data, there are other data types that cannot be always made fully open for ethical reasons. Another class of sensitive data is related to the protection of the environment and biodiversity. Data about endangered species or protected natural reservations should be kept confidential if their publishing could jeopardize their protection. (FINLEX 621/1999). Sensitive data can also include information that may pose risk to national security and defence (EU 2019/1024).

3 RESEARCH DATA MANAGEMENT

Regardless of whether data can be opened or not, research data are the cornerstone supporting the research results. As Higman, Bangert and Jones (2019) argue, openness is not a guarantee of data quality, and opening the data may not be useful to anyone if the data have not been properly managed during the research. The appropriate handling of research data is an integral part of any project that uses research data and a matter of research integrity and trustworthiness. Misinterpreting the data in results on purpose, such as data fabrication or data falsification (e.g., selecting only observations supporting the pre-selected result) are considered serious research misconduct. Openness and promoting further use of the data within the limitations of data protection and confidentiality is on the other hand considered best practice. (Finnish National Board on Research Integrity TENK 2023, 11, 17.) Even closed data should be organized and documented in a way that allows transparency in case they need to be reviewed, to support the validation or reproducibility of the research. The increased pressure and incentives to make research data openly available have only drawn more attention and visibility to the already existing practices of research data management.

Research data management (RDM) refers to the practices and steps taken to handle and take care of the data collected or produced during research activities in a way that ensures the accessibility, integrity, and high quality of research data. This includes for example the organisation of data, version control, documentation, backup and digital security and quality control. Various RDM practices are performed at certain stages or throughout multiple stages of the research data life cycle. (Corti et al. 2014, 2, Van den Eynden 2018, 43, Higman, Bangert & Jones 2019, 2-3, 28.) Figures 2 and 3 show two possible ways among others to visualize the life cycle model. While the circular model in figure 2 emphasises the iterative character of the research life cycle, the linear model in figure 3 helps visualise how some RDM practices are relevant across multiple phases.

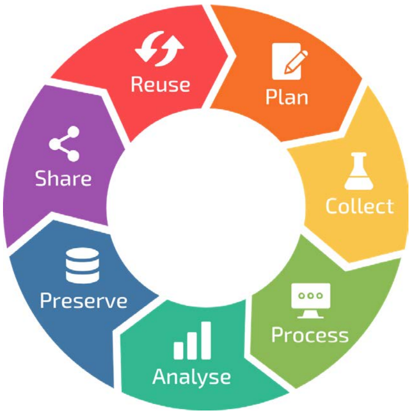


Figure 2: Research data life cycle. Research Data Management Kit (ELIXIR 2021).

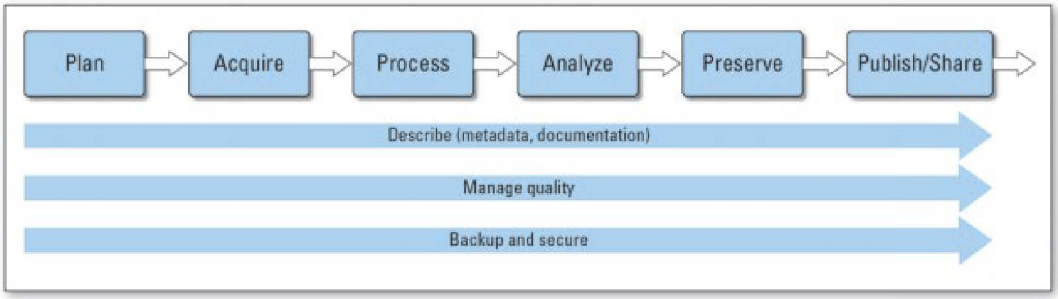


Figure 3: Research data life cycle. U.S. Geological Survey (2014).

Research data management is the overarching concept, which should be distinguished for the purpose of this thesis from the related terms “data curation” and “digital preservation”. Data curation usually refers to the principles and measures taken to ensure the long-span usability and integrity of data beyond the conclusion of the project. Data curation is often performed at the storage service where the data are deposited after the project, for example adding rich descriptive metadata or attaching keywords from controlled vocabularies or ontologies. Some of the curation steps, however, can and should be conducted before the deposition to long-term storage, such as the selection of file formats suitable for long-term accessibility and usability. Digital preservation is offered by specialized services that commit to proactively curating especially valuable data so that they would be still accessible to the next generations, for several decades or even centuries, keeping up to date with changes in technology. (Rice & Southall 2016, 31-32, 114-116.) In Finland, the Ministry of Education and Culture finances the Digital Preservation Service for Research Data “Fairdata PAS”. PAS stands for *pitkäaikaissäilytys*, the Finnish term for digital preservation. (Fairdata.fi 2020.) The literal translation is however “long-term storage” which may cause some confusion between Finnish and English terminology.

3.1 Not just open: FAIR data principles and reproducibility

In chapter 2.1 it was pointed out that social system in which research is performed, with its pressure to publish and academic hierarchies based on the incentive and reward system, can take away resources and attention from those research practices which are less valued. This can negatively impact the robustness of scientific knowledge and invite taking shortcuts or even straight misconduct. As argued at the beginning of chapter 3, making data openly accessible in itself is not a guarantee of reusability or quality of the data. For this reason, the scientific community including research funders and academic publishers has been aware of the need to establish principles guiding research data management that would enable humans and machines to find, access, combine and analyse research data, with the understanding of the associated algorithms and workflows. At a conference in Leiden, Netherlands, in 2014, a group of various academic and private stakeholders suggested a draft of what would be published two years later as the FAIR data principles. The acronym FAIR stands for Findable, Accessible, Interoperable, and Reusable. (Wilkinson et al. 2016, 1-3.)

The perspective of computational agents used by researchers to navigate data discovery and integration in the large scale was an important motivation for the creators of the FAIR data principles. This means machine readability is emphasized in addition to the human approach. Metadata is a crucial element in implementation of the principles for both machines and humans. (Wilkinson et al. 2016, 4.) Metadata refers to ‘data about data’, a type of data documentation providing contextual information presented in a standardised and structured form (Corti et al. 2014, 38-39). Descriptive metadata provide information about the content and characteristics of the resource, as well as the bibliographic information such as title and author. Technical metadata describe how the resource was created, its structure and intended use. Administrative metadata define the conditions of reuse and rights management. (Rice & Southall 2016, 26.)

The principle of findability requires rich metadata that are necessary to index the resource in search engines. Persistent identifiers (PID) improve findability by creating a unique and long-lasting reference to a digital object such as a publication or dataset. (Wilkinson et al. 2016, 4.) PIDs are meant to prevent link rot, where hyperlinks stop working as websites are no longer maintained, or content drift, meaning the link still works but the structure or contents of the website have changed. However, PIDs are not inherently persistent - the organisation responsible for registering the PID must commit to updating the metadata if the URL changes, so that the PID resolves (links) to the correct digital object. (Rice & Southall 2016, 37.)

The data are accessible when they are retrievable by their identifier using an open, free and standardized communications protocol, for example http(s). It is important to note that the FAIR principles are still applicable for data that cannot be made openly available due to legal or ethical reasons. Data do not have to be open to be FAIR. Publicly shared descriptive metadata can still provide findability and accessibility. In case the data can be made available with restricted access, for example only with research permit, a clearly described authentication and authorization procedure should be in place to provide access if conditions are fulfilled. The descriptive metadata should be openly shared, even when the respective data cannot be shared at all or have to be removed. (Wilkinson et al. 2016, 4, Higman, Bangert & Jones 2019.)

Interoperability means the ability of data or tools from various resources to integrate or work together. This requires use of a formal, accessible, shared, and broadly applicable language such as controlled vocabularies or metadata standards. Part of interoperability is also creating qualified references to other data or metadata, meaning that links and relations between them should be defined and explained (e.g. not just ‘X is associated with Y’, but also *how* they relate to each other). (Wilkinson et al. 2016, 4, GO FAIR Initiative 2020.) For example, Bernasconi and co-authors (2020) examined the importance of interoperability in the conceptual models expressing entities and their relationships and controlled vocabularies in integrating viral genome sequence data in the COVID-19 pandemic.

The reusability principle also depends on rich metadata, which is necessary to make the decision on suitability of the data for specific reuse purpose. However, reuse is not possible if the conditions under which the data can be used are not clearly communicated and defined with a license. Standard licenses such as the Creative Commons licenses enable machine readability. The FAIR principles also emphasise the concept of provenance. Provenance refers to the origin and history of how and by whom the data was generated or collected and processed. Following domain-relevant community standards and best practices in research data management and documentation helps make data understandable, and therefore reusable. (Wilkinson et al. 2016, 4, GO FAIR Initiative 2020.)

The concepts of FAIR and open data cover two aspects increasing data reusability. Both can be described as a spectrum, with varying degrees of openness and FAIRness. The third concept is research data management, which covers the whole data life cycle, including processes with internal benefits to the researcher, project or institution, ensuring the quality and integrity of the data regardless of whether it can be made open. Good RDM throughout the project is a prerequisite for openness and FAIRness, since in the planning and active management stages choices are made about crucial aspects such as data ownership and agreements, practicalities of personal data protection, and practices such as data format choices, naming conventions or capturing of documentation. (Higman, Bangert & Jones 2019.)

It is possible to share open data with many elements of FAIRness such as persistent identifier, reusable non-proprietary file format and basic metadata, which is in fact not very well reusable due to lack of more detailed and not necessarily standard types of documentation which should have been captured during RDM. This can mean for example variable descriptions, instrument settings, or information on methodology. On the other hand, the data could have been managed very well, but publishing them as graphs in a supplementary pdf file instead of the underlying numerical data means they cannot be reused. There is an interplay of RDM, FAIR data and open data as closely related but also distinct concepts. They complement each other and focusing on only one of them would overlook important elements of the others. However, they need to be prioritized and implemented differently in different data types and disciplines. (Higman, Bangert & Jones 2019.)

In chapter 2 the quality control mechanisms and criteria for reliability judgements supporting the robustness of intersubjective knowledge building were discussed. These mechanisms included peer review and making the underlying research data available for scrutiny. Depending on the discipline and context, the terms repeatability, reproducibility, and replication are used to describe similar but not identical mechanisms aimed at validating the reliability of one's own or someone else's research results. (Barba 2018.) For such processes it is not enough that data files are openly accessible, but the documentation resulting from following the FAIR data principles and good practices in RDM is also necessary. Even then, the data and the publication explaining the research process and its results are not always enough. To attempt validation of said results, detailed information about the study design, analysis methods and relevant code or software may be required as well. (Rice & Southall 2016, 153-155.) In some fields, such as computer science or some engineering disciplines, the code and software are the cornerstone, while data are only used to test the code. In such cases, the management, documentation and openness of the code are more important to reproducibility than the data. (Higman, Bangert & Jones 2019.)

Ravetz (1996) raised questions about "shoddy science" which should not pass validity judgements, but still often gets published. In his thought-provoking paper, Ioannidis (2005) argues that most published results are unreliable, often due to misunderstanding of statistical significance and bias in the analysis method or representation of results. He suggests confirming significance of single findings by gathering evidence in large studies and meta-analyses as well as improvements in research standards to remedy this issue. In that respect the principles of interoperability and reusability would support such validation studies. According to Ioannidis (2005), registration of studies before they are conducted would also in some fields increase the transparency of study design and help avoid manipulation of data to fit the preferred results.

Although the FAIR principles originate from a group within the scientific community, many researchers learn about them via funders' requirements. The concept has been adopted and included in the research data management plan templates for example in the EU Horizon framework and in Research Council of Finland grants (accessible from DMPTuuli 2023). While reproducibility and transparency can be seen as values internal to the self-corrective practices of scientific inquiry, their realization in FAIR principles or open data is for many introduced as an external requirement rather than a negotiated consensus of the community. Chen and her colleagues (2019) support the view that openness should not be pursued as a goal in itself, since more than openness is needed for reproducibility and reusability. Often the data needs to be accompanied by software, workflows, and explanations, which should be documented and managed throughout the research life cycle. Practices tailored to each discipline's methods and culture should be developed and supported by tools and services to build in the prerequisites of reusability already into the data collection and analysis processes. Support is therefore needed in raising awareness and promoting open discussion as well as practical implementation of the practices enabling reuse and reproducibility of research data.

3.2 Data management planning: policies and practices

The research data life cycle, as demonstrated in the beginning of chapter 3, starts with the planning phase. Funder requirements to provide a strategy of RDM and share research data have led to the formalisation of Data Management Plan (DMP) as a document (European Commission 2012, 34). It is now common for research funders to create a DMP template for their funding programs, such as for example the Horizon 2020 DMP template (Spichtinger & Siren 2018, 17-18; DMP templates relevant in the Finnish academic environment can be found in the data management planning tool, DMPTuuli 2023).

Adding a formal DMP as a required project document increases the researchers' administrative burden of grant applications and project management. As described in chapter 2.5, Van den Eynden (2018) and others demonstrate the amount of time and effort used on tasks less rewarding in career development can be perceived negatively, as taking away time that could be spent on the research and publishing results. On the other hand, chapter 2.4 argued based on Van den Eynden's conclusions that researchers can be driven towards good RDM by the pursuit to increase the quality, responsibility, and impact of their research. When required to write a DMP, this may be the first time the researcher needs to verbalize their RDM practices and harmonize with the language and expectations of the funder. As a best-case scenario, the DMP would not only be a top-down requirement, but it would also serve the researcher as a practical RDM planning tool supporting high quality, high impact and responsible conduct of their research.

A DMP should include information on what kind of data will be collected, produced, or reused, and describe the research data management practices needed to ensure data integrity and security as well as compliance with legal and ethical norms. In the planning stage, it is important to recognize and manage risks, such as loss or destruction of data due to storage solutions not being backed up, or breach of legal or ethical regulations which can rarely be fixed when the damage is done. Such risk assessment also helps understand if data can be made open or whether there are reasons to keep it as closed as necessary. DMP should help identify the necessary tools and services, financial resources, as well as define the RDM roles, responsibilities and rights of the individuals and organisations involved in the research project. A written plan shared with all project members can help keep practices such as documentation or data organization consistent. It can also help find and use data even when research group or project members leave. Naturally not everything can be planned for in detail at the beginning of the project, therefore the DMP is meant to be a living document, updated whenever significant changes occur. (Corti et al. 2014, 22, 24, 27, Briney 2015, 19-22, Kruse & Thestrup 2018, 3-4.)

General Finnish DMP template (Tuuli-project 2021) follows the recommendations outlined by Science Europe (2021), an organization representing a group of European national research funding bodies. The Research Council of Finland, member of Science Europe and an important research funding organization in the Finnish academic environment, has adopted the same DMP template as the general Finnish recommendation. European Commission is a crucial funder of research in Europe on an international level. The DMP template for the EC's current Horizon Europe programme, the Research Council of Finland template and the general Finnish template have been published in the Finnish national data management planning tool DMPTuuli (2023).

While there are some differences in how the DMP is organized into sections in the Horizon Europe and the Finnish DMP templates, both cover the same fundamental themes across the data life cycle. Both start with the general description of the data that will be the subject of the DMP. There are comparable sections covering the ethical and legal issues related to the data, secure storage and backup during the project, and allocation of the responsibilities and resources for RDM. Although both template types mention the FAIR data principles, the approach is slightly different. The Finnish template introduces the FAIR principles in the section on documentation and metadata, which seems more relevant in data management *during* the project (for example, file-naming conventions, version control and folder structure, capturing metadata necessary to understand the data in README files etc.). Elements provided by data repositories such as persistent identifiers or machine-readable licenses are suggested later, in the section covering opening, publishing and archiving the data *after* the research project. The Horizon Europe template has a FAIR data section with subsections covering each of the four principles separately, with guidance focused on the reuse aspect. Other research outputs such as code or software which may be useful for reproducibility and reuse are mentioned in both templates, although more attention to these other outputs is given in the Horizon Europe template. (DMPTuuli 2023.)

As a tool supporting RDM planning for the researcher, the former approach following the life cycle may seem more intuitive and less prescriptive. Even if data cannot be opened for reuse, they can still be well managed and documented to attain a certain degree of FAIRness, useful for potential validation of results or good RDM supporting the quality of research results.

3.3 RDM during the active phase of research

The practicalities of RDM during the active phase of research depend on the research discipline, data type and research methods. Some principles can however be identified as common in striving for digital security, consistent organization, and quality of the data. In collaborative projects it is advisable to agree on some shared practices so that the data are managed consistently and made understandable and reusable within the project, and to others if the data will be made openly available at a later stage. Some of the elements that should be consistent are a clear folder structure and file hierarchy as well as logical naming conventions that will help locate and identify files or other digital objects and their mutual relations. (Corti et al. 2014, 42-43, 68, Briney 2015, 62-74, Aineistonhallinnan käsikirja 2020).

When selecting the software and file formats, priority should be given to solutions that will not put unnecessary restrain on reuse. This means usually non-proprietary file formats that can be accessed without commercial software or special equipment and which will not be affected by changes in different versions of software packages or software becoming obsolete. If the software and formats necessary during the research are proprietary or uncommon, and conversion will be necessary for longer-term storage or data sharing, quality control of the conversion should be taken into consideration. Planning for a versioning system will help keep track of the changes and document how the data is handled, tracking the provenance of the data. As a digital security measure, the master copy of the raw data can be stored separately to be able to compare and locate possible errors, or to redo the analysis. (Corti et al. 2014, 34-35, 71-73, Briney 2015, 80, 132.)

Documentation refers to all the contextual information and description that is necessary to make data findable, accessible, understandable and (re)usable. It provides explanation of how and why the data was created or collected, what is its structure and content, how it has been modified or coded. (Corti et al. 2014, 27, 38.) While in some contexts metadata can be used as a synonym for documentation, it is rather a formal and structured subtype of documentation which should be machine readable to enable searching. Documentation can also include free text such as readme files. Study-level documentation should provide the high-level information about the research context and design and how the data was collected and manipulated, e.g., data collection protocols, sampling design, instruments used, software used, digitization methods etc., the quality assurance processes carried out and modifications between different versions. It should also explain the structure of data files and the relationships between them. Data-level documentation provides information about individual data files and their components such as variables. (Corti et al. 2014, 38-40, Rice & Southall 2016, 24.)

The FAIR principle of interoperability suggests using community-accepted, shared metadata standards or formats that enable understanding and combining of data by humans and machines. For descriptive (discovery) metadata on the study level, some commonly used metadata formats include DDI (Data Documentation Initiative), Dublin Core, DataCite or RDF (Resource Description Framework). On the data level, applicable and suitable metadata formats or other ways of documentation are discipline specific. They depend on the character of the data, the elements which need a standard representation, and what is deemed good practice in the community. (Corti et al. 2014, 45-46, 49-50, Rice & Southall 2016, 24, 26.) The wide range of practices can be observed for example in the disciplinary metadata guide by Digital Curation Centre (2023a) or the FAIRsharing registry of metadata standards (Sansone et al. 2019).

Documentation practices should be consistent and coordinated among the research project members. Documenting already in the active phase of research is advisable, as documenting afterwards from memory can be time consuming and prone to errors. Documentation is necessary and useful to keep track of the research process and to understand the data for individual researchers, research groups or projects. If plans are made to make the data available after the research, it is also useful to consider what should be added or changed so that the documentation facilitates understanding and reuse by someone else. (Rice & Southall 2016, 25-26.)

The DMP prompts researchers to plan for secure storage and sharing of the data during the active phase of the research (live storage) (DMPTuuli 2023, Rice & Southall 2016, 106). For a researcher employed by an academic organization, appropriate IT services and storage capacity are likely provided by the home institution. Back-up method and frequency are important to avoid loss of data. Appropriate levels of physical and digital security measures, such as monitoring access to facilities, prevention of unauthorized access with password protection, firewalls, or not connecting servers to the internet should be considered. Live storage is also more secure if levels of access and user rights are determined, for example, when it is possible to assign specific read-only, read and write or administrator rights to concrete folders or files. This is especially important if the project handles personal or sensitive data, which should be also protected for example with encryption. (Stouthamer-Loeber & Bok van Kammen 1995, 109, Corti et al. 2014, 27, 35, 88.)

In collaboration across institutions, secure transfer and sharing of data in live storage for the active research phase can be challenging. Sharing or collaborating on research data with project members from other institutions requires IT solutions that can be accessed by users affiliated with different organizations. At the same time, these solutions must be secure, with procedures for authorization and authentication of users. Common solutions include providing external collaborators access to the organizational network drives, using a secure file transfer protocol (FTP) server, or cloud services. (Corti et al. 2014, 158, 162-163.) In the European context, the GDPR regulates that personal data shall not be transferred outside the EEA without a special agreement (EU 2016/679). If cloud storage service is used for collaboration, special attention should be paid to where the servers are located and how reliable is the digital security (Rice & Southall 2016, 129-130).

3.4 Long-term storage after the project, publishing, and citations

Research data in form of immutable datasets can be deposited into long-term storage after the project, or during the project, for example if a dataset is ready to validate published results. Long-term storage services are usually called data repositories, but the term data archive is more traditional in certain disciplines, such as social sciences (Corti et al. 2014, 199). A data centre refers to the technical infrastructure behind a data service (see for example European Council for Nuclear Research (CERN) 2023, CSC - IT Center for Science 2023). Rice and Southall (2016, 103) argue that using the term data archive in the long-term storage context may create the impression that data are being actively curated for digital preservation, just like physical archival materials are preserved for future generations. However, the term data archive is often used interchangeably with the term data repository, a long-term storage solution not committed to digital preservation. For clarity, this thesis will use *data repository* to refer to long-term storage solutions for datasets resulting from research projects and activities.

As discussed in the beginning of chapter 3, further confusion about the level of curation can be caused by the Finnish term for digital preservation '*pitkäaikaissäilytys*' literally translating as long-term storage. As defined previously, digital preservation services are a special case of long-term storage solution which commits to curating the data for decades or even centuries. Such curation involves changing the deposited files, for example due to file format obsolescence. While common file formats should not become obsolete in the time frame of regular long-term storage, we cannot foresee technological changes in the time frame of many decades. A digital preservation service has specific requirements for archivable files and collects the depositor's consent to curate the files if necessary. This may mean for example file format migration or using emulation software still capable of opening an obsolete format. Common data repositories for long-term storage will usually not provide such measures. (Rice & Southall 2016, 106, 114-116.)

Although most repositories do not have the resources necessary for active curation over decades or even centuries, measures are in place providing reliable long-term storage of the data as it was deposited. A trustworthy data repository will have methods in place to ensure digital integrity of the deposited data on the bit level. To avoid data corruption ('bit rot'), checksums may be created to act as a 'fingerprint' of the data which can be checked and compared to detect corruption. Data can also be replicated to another storage medium to prevent data loss in case of severe malfunction. There are some certificates of a trustworthy data repository, such as the DSA Data Seal of Approval, later renamed to CoreTrustSeal, or the ISO 16363 standard. However, the uptake of these certifications has been limited so far. Re3data - Registry of Research Data Repositories is considered a reliable database of data repositories suitable for long-term storage. The choice of repository will depend on the discipline-specific best practice, home institute or funder recommendations, and the type of research data. (Corti et al. 2014, 87-88, 96-97, Rice & Southall 2016, 103-104, 115.)

Specialized discipline-specific data repositories can be the best choice especially if they are well-established in the community. Such repositories will often provide discipline-specific metadata fields or vocabularies. General data repositories such as Zenodo, or the Fairdata IDA repository which is a part of the Finnish national research data services, are available for data from various disciplines. The available metadata standards are therefore more generalist. (Corti et al. 2014, 197-199, Rice & Southall 2016, 105, Fairdata.fi 2020.) Institutional data repositories are less common. While publication repositories are well established at universities, developing and maintaining an institution's own data repository would require lots of resources. Current research information systems (CRIS) have been adopted at many universities to gather information about various research outputs and activities beyond publications, however, they are technically more suitable for collection of bibliographic metadata, rather than long-term storage of datasets. (Rice & Southall 2016, 104, Corti et al. 2014, 200-201.)

The benefits of using a data repository, instead of other solutions such as a project website or informal exchange by request, are better digital security and sustainability of access. Repositories are usually better equipped to check data integrity on the bit level and to follow the FAIR principles than a website or storage on an external drive. Repositories will create a landing page for the data with the necessary metadata, which will enhance the findability through a search engine. Most trustworthy repositories will also assign a PID to the datasets. The PID can be linked to the related publication, and it is also important to make the datasets citable, which means researchers can get merit via citations if their data are reused. A repository also usually lets the depositor select a license and define conditions of reuse. Repositories may enable various levels of openness from open access datasets to only providing the descriptive metadata of closed access datasets. Some repositories provide an option to set up authentication and authorisation process for managing access to restricted datasets. Repositories also can be interoperable with other information tools and systems, for example via an API (application programming interface). (Corti et al. 2014, 197, 199, 204, Rice & Southall 2016, 117-118.)

If data cannot be deposited to a repository due to restrictions on data sharing or not being ready to be published as a dataset, there are still RDM procedures which should be considered to keep the data accessible and usable after it is no longer in active use in the original project. One of them is the selection of data which should be kept for validation of published results or potential reuse, or to comply with institutional or data type specific retention policies. For example, there may be special requirements for the retention period of data resulting from clinical trials and medical research, or underlying data related to patented inventions. If data can or should be deleted, proper mechanisms should be applied to fully erase sensitive data, as regular deletion may only delete the reference to the data location, while keeping the files. For data that needs to be retained, it is important to ensure long-term accessibility and reusability accounting for the risks of corruption of data files, device failure, natural disasters or other damages to the hardware. It is also important to prepare the data files to remain reusable by using common file formats and also retaining the documentation necessary for interpretation of the data. (Briney 2015, 127-137.)

As presented in subchapter 2.5, opportunities to receive credit for the work related to RDM are an incentive for opening the data. Citations were reported to be the most common currency of academic merit. Piwowar and her colleagues (2007) have concluded that opening the data may increase citations to the related article and thus offer some kind of reward for the additional work of putting together publishable and FAIR datasets. On the other hand, chapter 2.1 also discussed the problematic dominance of journal and article citation-based metrics in research evaluation and the need to diversify opportunities to receive merit, for example for good research data management and data sharing. Developing practices enabling tracking of citations or reuse of the data itself could be beneficial. However, the system built for measuring journal and article citations is not directly transferable. Data collectors may be given credit in the acknowledgements section of the article, which is not indexed in citation databases. (Corti et al. 2014, 205.)

Data may also be published as supplementary files, which do not have their own PID and other citable bibliographic metadata (Piwowar et al. 2007).

Data repositories support data reuse and citation by providing persistent identifiers and descriptive metadata needed for discovery and citation. The most common descriptive metadata standards used in data repositories (Dublin Core, Schema.org, DataCite, DATS) have metadata fields needed to uniquely identify cited resource: identifier, creator, title, publisher, year of publication and resource type. (Fenner et al. 2019.) Cousijn and colleagues (2018) recommend that publishers should define their policies for citing data used in published articles (both generated as a result of the work described in the article, or previously published data reused in the research). This should include guidelines how the citations should be formatted and where they should appear: in the standard reference list or in other sections of the article. Data Availability Statement section is recommended as a standard part of articles where authors can provide information on the data and explain potential reasons why the data may not be publicly available (such as ethical and legal issues, confidentiality or embargo period needed to allow the authors to complete their research). (Cousijn et al. 2018.) Although these recommendations technically enable data citation, the practices of collecting data citation metrics (data citation databases, harvesting data citations from publications) are not as comprehensive, widespread, and commonly used as services indexing publication citations. It is also questionable whether tracking citations in scientific journals is all we can do to evaluate impact of research data, and what kind of alternative metrics would be needed to capture this impact. (Rice & Southall 2016, 37-38.)

Monitoring of reuse and citation also favours datasets that are available with open or managed access, while the workload that goes into good management of research data which cannot be shared remains invisible. Although the workload is probably lower when the data does not have to be prepared for public sharing and reuse, the FAIR principles and good data management are important for the robustness and reliability of results which the data supports, no matter its openness level (as argued in chapter 3.1). The management of personal and confidential data also comes with additional requirements for secure live storage (see chapter 3.3).

4 RESEARCH DATA MANAGEMENT SUPPORT SERVICES

As the open science paradigm becomes more common in publicly funded research, the pressure from the funding bodies on research institutions to adopt open access policies grows. However, most of the responsibilities to put policies and principles into practice fall on the shoulders of the researchers. Therefore, there is a need for practical guidelines and support services that fit in with the reality of researchers' workflows. To understand what kind of support services would be sufficient, it is important to gain insight into the RDM landscape and researchers' information and support needs. (Davidson 2013, 90-91.)

Documents such as an institutional data policy or open science strategy define the institutional strategic objectives and responsibilities of various stakeholders, including researchers, support services, and management. Ideally, the policy does not only require certain steps from researchers, but also helps secure resources and commitment from the institution management to provide the infrastructure necessary to implement the policy. The need to define what is necessary regarding the responsibilities and roles of support services often motivates institutions to map service user needs first. (Rice & Southall 2016, 46, 69-73.)

As mentioned in chapters 2.5 and 2.6, researchers have in previous studies expressed a need for common policies and guidelines, but also a need for mitigation of the perceived potential risks of opening research data and the disadvantages of using time and effort on work which is not included in evaluation processes. An institutional policy or strategy adopting a highly ambitious form of open science approach without involving the researchers may create friction. In principle, researchers are inherently motivated to create research outputs of high quality and impact, and if they do not feel that open science is facilitating these goals, they often have legitimate concerns. Including them in the policy development and hearing out their concerns can lead to better and more realistic policies as well as to higher level of commitment to the strategy. (Rice & Southall 2016, 71-73.)

Institutional support services are positioned between the researcher and the institutional or wider-scale policies, providing opportunities for communication and mutual learning. Van den Eynden's (2018, 44-50) review of previous studies summarized the individual and institutional motivational factors affecting researchers' motivation to open their data (and pay more attention to RDM), some of which are related to support services. On the individual level, perceived lower level of RDM skills and high level of effort necessary for data opening was shown to affect motivation. Support services can aim to help the researcher build upon their RDM skills or provide support to make time-consuming tasks more efficient (Van den Eynden 2018, 44-50). Support services in the home organization are mentioned directly as a positive factor on the institutional level. Other factors such as efficient IT infrastructure or clear guidelines on the associated ethical and legal issues may require effort on a higher level than individual organizations, however, the support services close to the researcher can still play a major role in helping researchers access the infrastructure and guidelines and incorporate them in their RDM.

The range of RDM topics covered in the previous chapters, such as compliance with funder's requirements, discipline-specific documentation practices, secure storage and sharing with collaborators, legal and ethical issues, implementation of the FAIR data principles or other reproducibility criteria, intellectual property rights or publishing research data is not likely to be answered by one type of expert or service. There can be various stakeholders within the institution and also outside the institution (such as external data centres or repositories) offering support or participating in the various facets of RDM, and it can be confusing and laborious for the researcher to navigate this space and find the right person to help with a specific issue. Ideally the different stakeholders within the institution should be connected in a network with open communication channels and division of roles and responsibilities, presenting to users as a service with an easily accessible contact point. (Davidson 2013, 85, Rice & Southall 2016, 67-68.) The following subchapters will examine how RDM services have been organized and provided in academic libraries and academic institutions in general, and how they can be evaluated.

4.1 RDM related skills and expertise in academic libraries

The Finnish Ministry of Education appointed a committee in 2008 as part of the Structural development of higher education project to outline what was needed to transform university and polytechnic libraries into a digital service network. In the committee memorandum on development of university and polytechnic libraries (Saarti, Poutanen, Kuusinen & Vattulainen 2009), the main task of these libraries is defined as providing services and library collections supporting the tasks of their parent institutions: higher education and research, and the needs of their target groups: researchers, students and other information seekers. Research data was defined as an important part of researchers' workflows which also requires support. (Saarti, Poutanen, Kuusinen & Vattulainen 2009, 10-11.) How exactly academic libraries can or should respond to this support need has been investigated in several studies.

Auckland (2012) explored the information needs and information seeking behaviour of researchers, including needs related to research data, and what kind of skills and knowledge librarians should have to respond to these needs. The focus was on subject librarians, who support researchers in the context of a specific discipline. Based on available literature and survey results, Auckland (2012, 36-38) identified the following areas related to research data management: knowledge of data sources available in the discipline, knowledge to advice on data management and curation (e.g., ingest, discovery, access, dissemination, preservation), knowledge to advise on data manipulation tools used in the discipline (e.g. statistics tools for analysis), knowledge to advise on data mining, being able to support researchers in complying with funders' requirements (including open science requirements), understanding of author rights and intellectual property issues, and knowledge to advocate and advise on the use of metadata, as well as skills to develop metadata schemas or advise on discipline-specific standards and practices. Further survey questions investigating the perceived importance of these areas and existing skills of subject librarians discovered that the respondents felt there was a skill gap and need for professional development in the areas related to metadata, data management and curation, data mining, data manipulation tools, and support in complying with various funders' requirements. (Auckland 2012, 42-43.)

The Association of European Research Libraries (LIBER) established the "E-Science working group" in 2010 to investigate the role of libraries in supporting digital research practices, including research data management. After a series of three workshops held at relevant professional conferences, the working group published the resulting "Ten recommendations for libraries to get started with research data management" (Christensen-Dalsgaard et al. 2012). The recommendations are ordered by priority, starting from the recommendation to offer RDM support, including data management planning, intellectual property rights advice and information materials. This is followed by recommendations to engage in the development of metadata standards and provide metadata services, to create specialized data librarian positions and develop related professional skills of staff, to actively participate in institutional data policy development, to network with researchers, research groups and data repository providers to foster interoperable infrastructure, to support the lifecycle for research data by providing services for sustainable storage, discovery and access, to promote data citation with persistent identifiers, to provide an institutional data catalogue or repository depending on available infrastructure, to engage with discipline specific RDM practices, and finally to collaborate with institutional IT services and cloud service providers to offer secure live storage for research data. (Christensen-Dalsgaard et al. 2012, 1.)

A subsequent recommendation by the Association of European Research Libraries (2017) outlines the possible role of libraries in implementing the FAIR data principles. The recommendation states that libraries can expand their FAIR data expertise starting from existing skills and knowledge concerning certain elements that are already well known from publications, such as PIDs, licenses, bibliographic description, and generic controlled vocabularies. Libraries can start raising awareness of the FAIR data principles and learn how they can be implemented together with institutional IT services and researchers. Bibliographic description and metadata knowledge can be expanded to include discipline specific metadata standards and tools that are compliant with FAIR principles. Libraries should recommend that researchers deposit their data to repositories which support implementation of the FAIR principles. The library can also actively advocate for including the FAIR principles in organisational guidelines and data management plan templates. (Association of European Research Libraries LIBER 2017, 2.)

Tenopir, Pollock, Allard and Hughes (2016) conducted a survey of European academic library directors to investigate what kind of research data related services the libraries offered and planned to develop in the future. The results were compared to a previously conducted survey of academic library directors in the USA and Canada. The libraries represented in the survey provided two major types of research data services: informational/consultative (reference services, consulting with faculty or students, etc.) and technical/hands on services (creating and maintaining a data repository, preparing datasets for deposition, etc.). The first type was more common in both European and American libraries. While hands-on services were less common, the majority planned to offer them in the future. In Europe, the majority of libraries reported at least some level of involvement in RDM support. Over 76 % were involved in discussions of research data services with others on campus, and over 66 % were also involved in the development of research data related policies. Consultative research data services included consulting on data management plans and metadata standards, reference service for finding and citing data, or creating online guides. Direct participation with researchers on a project was the only type of consultative service the majority did not plan to offer. The authors speculate that this may be because such service would require larger time commitment. (Tenopir et al. 2016.) It could also be argued that libraries may not have subject librarians with understanding of discipline-specific practices for every field.

Cox, Kennan, Lyon and Pinfield (2017) studied the current and planned RDM services in academic libraries in Australia, Canada, Germany, Ireland, the Netherlands, New Zealand, and the UK. The survey also investigated how libraries collaborated with other internal services or external organization on the provision of RDM support and explored the knowledge or skill gaps. The majority of respondents stated they had an RDM policy in place or planned to develop one in the next 12 months. Stakeholders involved in RDM policy development included the library, IT services, research office, legal office, and academic contributors, with library or research office most often leading the development. Similarly to Tenopir and colleagues (2016), Cox and others (2017) divided RDM services into advisory (consultative in Tenopir et al. 2016) and technical. Both studies conveyed that technical services (such as developing and running a data repository, advising on curation of active data, selecting and preparing data for deposition) were less developed. The survey questionnaire designed by Cox and his collagues (2017) also asked the respondents to rate the maturity level of their RDM services on a scale of no service, basic, well developed and extensive. Results report mostly basic services or no service in some of the service areas, while extensive services were identified only in the Netherlands (8 %) and the UK (1 %).

According to Cox and his colleagues (2017), most common currently provided type of advisory service across respondents was maintaining an online guide of local advice and useful resources for RDM. The authors observed that this seemed to be a natural starting point for service development, because it is relatively easy to implement and does not require high levels of resources and effort. RDM training and/or data literacy instruction was the second most common type of service. Since academic libraries already offer training and are well known in this capacity to researchers and other stakeholders, this also appeared to be a natural role for libraries to assume in RDM. The lowest level of service provision was reported in advisory service on data analysis, mining or visualization, and, similarly to the findings of Tenopir and collaborators (2016), the direct participation with researchers on a research project. However, Cox and colleagues (2017, 2190) noticed an aspiration of some libraries to be more “embedded within the research space”. In the open-ended comments, some respondents emphasized the need for advocacy and engagement of researchers alongside the service delivery, taking into consideration possible anxiety or negative perception around data sharing and open data. Regarding the provision of technical services, respondents disclosed challenges in understanding the diversity of data types and related metadata. Storage requirements and sufficient storage services were also highlighted as a challenge difficult to tackle by the library alone. (Cox et al. 2017.)

Experience from reference services, guiding the user in dialogue to identify their information need and find suitable sources, can be transferable to finding solutions for the information needs in RDM. The pedagogical skills acquired in various user training and courses, such as information seeking and information literacy training, as well as creating various guidance and training materials can be used in organising RDM and DMP training and creation of educational materials for various audiences. (Rice & Southall 2016, 35-43.) This is supported by the findings of Tenopir and colleagues (2016) and Cox and colleagues (2017) that organizing training and creating guide material seem to be the aspects where libraries most likely get engaged with RDM. Corall (2012, 108-110) comes to a similar conclusion that experience in information literacy education as well as reference and consultation services equip academic libraries with skills useful in providing RDM support services. She also notes that strive for good RDM is aligned with the values of open science, which academic libraries usually advocate for. Libraries also have experience with institutional repositories and information systems which can be transferable to RDM. On the other hand, Corall (2012, 120-121) points out the insufficient amount of RDM and digital curation related courses and opportunities for hands-on practice in curricula of Library and Information Science programs. Cox and his colleagues (2017, 2184, 2191) also emphasize the need for continuing education and training of academic library staff on RDM topics, as well as the importance of practical experience.

DMP guidance is a support need that arises early in research data life cycle, especially if a DMP is required from the funding body already at the proposal stage. The DMPs cover various RDM topics across the life cycle, including ethical and IPR considerations, active and long-term storage. These topics often require input from various experts; however, the libraries are often well connected to these other institutional services and can act as liaisons between the researcher and other relevant support services during the DMP drafting process. (Davidson 2013, 93.) Davis and Cross (2015) published a practice article describing the experiences from the North Carolina State University regarding librarians' hands-on skill development through DMP review service. At their university, a Research Data Committee was established to provide a team based DMP review service, available at the moment of need. The committee would also subsequently be tasked with developing training in RDM for other librarians and researchers. At first, committee members were selected from those who already had some relevant expertise, for example in copyright and licensing, grant funding, open science, working with geospatial and numerical data or digital humanities. Subject librarians were also involved to include their discipline-specific expertise. Team-based way of working enabled knowledge exchange among the members and eventually also the hands-on training of new members with limited or no previous RDM support experience. The various use cases that researchers brought for review also provided an invaluable training ground, helping learn about challenges and disciplinary norms for data acquisition, management and sharing. Recognized gaps in knowledge were filled by expanding the network involved in DMP review, including for example IT services, research proposal development experts, or technology transfer (IPR) experts. (Davis & Cross 2015.)

4.2 Organization and evaluation of research data management support services

The previous subchapter suggests there are gaps in RDM-related skills and knowledge represented in academic libraries. Some of these can be bridged by education and professional development, but others are more likely to be effectively bridged by involving other services and experts. Cox, Pinfield and Smith (2014) explored whether RDM could be categorized as a "wicked problem". A "tame problem" is a type of problem which is familiar and even though it can be challenging, there are well-known approaches which can be used or adapted to solve it. A wicked problem, on the other hand, is unique and highly complex. The approach towards solution is unclear, and there are various stakeholders who have different understanding of the definition of the problem and possible approach to solutions. The character of problems will affect how organizations can respond to them. An attempt to solve a wicked problem will be "clumsy", not satisfying all stakeholders equally, and the different viewpoints should be taken into consideration. One of the approaches suggested for dealing with wicked problems is Design Thinking. In management this refers to a creative process consisting of gathering information and imagining possible solutions and testing out prototypes before implementation. (Cox et al. 2014.)

The authors (Cox et al. 2014) conducted interviews with academic library practitioners to investigate whether the way they described the RDM agenda could fit the criteria for a wicked problem, and whether the ways they approached RDM were suitable for a wicked problem. Results suggested that RDM does indeed fulfil many of the 15 criteria the authors defined based on available literature. The interviewees felt there was no service stakeholder for whom RDM support is an obvious core capability. Rather, many different stakeholders each brought with them part of the answers and expertise. They identified IT services, research support services, legal advisory services, records management services and research leadership as service stakeholders involved in RDM support. These stakeholders view RDM problems differently depending on their roles as well as values and cultures, for example, while IT services focused on active data storage, libraries prioritized long-term preservation and sharing. The interviews suggested there were different approaches towards solutions and uncertainty about whether they would work. The real scope of the problem was also unclear, given that there is not enough information about some of the aspects of RDM such as the discipline specific and local practices or attitudes. Approaches towards solutions mentioned by interviewees could be seen as appropriate for wicked problems. They highlighted flexibility and embracing emerging needs and problems. Networking, collaboration, building and maintaining relationships among stakeholders as well as with students and researchers were considered key capabilities.

Hofelich Mohr, Johnston and Lindsay (2016) describe a specific example how such well-connected, distributed network of services was implemented at the University of Minnesota. They state that support for RDM across the research life cycle “takes a village” – a more effective and comprehensive service can be provided in collaboration than by any service unit alone. As part of the network development, the team reviewed service providers whose input would be beneficial for RDM support. Grant consultants were identified as starting point for research funding related requests for DMP consultations, and grant administration office could help check compliance with funders’ requirements. Research deans at the various colleges were recognized as valuable stakeholders advocating for services and connecting RDM experts with the needs of their colleges' researchers. Commercialization office, legal counsels and copyright librarians helped provide support for licensing, agreements, and intellectual property rights issues. Institutional ethics board typically reviews also RDM practices and their recommendations have impact on data sharing and archiving. Office for research was defined as the unit which creates research policy, provides administrative support and sets the university's research agenda. Data security offices could contribute with understanding of security risks and secure data storage practices. IT services provide the solutions for secure storage, supercomputing centres and other technological solutions researchers may need for their data. Support services for statistical analysis or data collection tools (for example survey tools) can play a role in developing best practices. The role of the library was connecting the researchers to these various stakeholders, providing support for open data repositories, data sharing and preserving, metadata creation and cataloguing.

Pinfield, Cox and Smith (2014) examined the relationships and division of roles between different stakeholders via semi-structured interviews with academic library professionals in the UK. Similarly to what was discussed in the previous chapters, they articulate that RDM includes a complex set of different aspects including "an array of technical challenges as well as a large number of cultural, managerial, legal and policy issues". Support may be needed across the research data life cycle in the form of technical components (software, tools, infrastructure) and organizational components (policies, funding strategies, guidance, and training). The thematic analysis of interview data showed that in addition to libraries, IT services and research support services were the most common stakeholders. Legal advisory services and records management services were also mentioned, as well as senior academic staff. Drivers of development in RDM identified by the interviewees were the need for storage facilities, need for security in accordance with the level of confidentiality of the stored data, need for medium and long-term preservation, compliance with external requirements such as funders' policies and legislation (e.g. data protection), need to adhere to high standards of data quality and robustness of research findings supporting their validation and reproducibility, and the need for tools and systems enabling data sharing with specified persons or making data openly accessible. A driver specific to the staff involved in RDM support concerned the "jurisdiction", meaning the justification of involvement in RDM and division of roles and responsibilities among stakeholders.

Matusiak and Sposito (2017) analysed job announcements for roles in RDM support and interviewed practitioners in order to gain insight into practices and organization of RDM services (both technical and advisory). The participants were from institutions in North America, Australia and Europe. Most commonly RDM services were based in the university library. Lesser developed services provided mostly advisory support and did not focus on the technical aspects of research data management. A more advanced type of RDM services typical especially for European institutions was a collaborative model of "distributed networks" of research data experts, including expertise in IT, copyright, research ethics and scholarly communication. The library usually coordinates the network and distributes support cases to relevant experts. The need for understanding of discipline specific practices is reflected in the model of "embedded services". In this model, a data steward or data manager is assigned to a project, department, lab, or a similar smaller unit to assist with RDM throughout the research data life cycle. "Research data service centres" are a more comprehensive and evolved model of collaboration among network of experts. These centres can answer to a wide range of RMD support needs with advisory services and technical tools and infrastructure. Units involved in the collaboration usually include the library, IT department, legal services, and office for research. Matusiak and Sposito (2017) conclude that there is no right or wrong solution, rather, the various approaches reflect the need to tailor the model to various institutional and national contexts.

There are several models developed specifically to gain insight about user needs and to develop or evaluate RDM services. One of them is the Data Asset Framework (Digital Curation Centre 2023b). The aim of the DAF is to map what kind of research data are produced at the organization, and how are they stored, managed, shared and used, as well as what risks are associated with them, such as data security and protection issues. The DAF evaluation can also open up a communication channel with researchers, eliciting the expression of attitudes and opinions regarding RDM. This in turn helps locate the gaps in RDM services. The framework is designed to be facilitated by an academic librarian, RDM specialist, or another person with experience in academic research environment and research data life cycle. Participants should represent various stakeholders in the RDM landscape. There are four phases of the DAF evaluation: planning, identifying and classification of research data assets, RDM evaluation, and reporting. The methods applied in the second and third phases can be flexibly tailored to the organization's needs, selecting from survey, interviews and searches within the organization's databases or IT interfaces. The published guide for DAF assessment includes some crucial elements which help understand the character of the research data, the steps taken in their management, and potential challenges. There are, however, no standard models for surveys or interview instruments. (Digital Curation Centre 2009, 3, 7, Jones, Ross & Ruusalepp 2009.) A survey and follow-up interviews based on DAF were used for example at the Georgia Institute of Technology in an extensive Research Data Assessment complemented by a content analysis of DMPs (searching for information about, for example, used storage services, repository services or standards) and data archiving case studies (Rolando et al. 2013).

Many of the tools for RDM service assessment are based on the Capability Maturity Model (CMM) first developed by the Software Engineering Institute to enhance software development and maintenance processes. The model defines an immature organization as one that is reactionary, focused on solving immediate crises ad hoc, and whose processes are mostly improvised without an objective way to assess quality. In a mature organization, the processes are defined and consistent with how the work is done in practice. This way the management can communicate the process to staff and new employees. Roles and responsibilities are clear, and the processes are updated when necessary, with evidence collected via methods such as pilot testing or cost-benefit analyses. There is an objective basis for assessment of quality. Schedules and budgets are based on previous experience and therefore realistic. Capability is defined in the original CMM in the context of software process as "a range of expected results that can be achieved by following a software process", and "capability is one way to predict the most likely outcome [of the project]". Maturity is "the extent to which a specific process is explicitly defined, managed, measured, controlled, and effective". (Paulk, Curtis, Chrissis & Weber 1993, 18-20.)

The CMM has been later applied in various contexts outside software development. The definitions of capability and the maturity levels vary, but the main maturity framework with the five levels (1 - Initial, 2 - Repeatable, 3 - Defined, 4 - Managed, 5 - Optimizing) can be used to devise maturity level descriptions (Paulk et al. 1993, 21). Crowston and Qin (2011) presented their capability maturity model for scientific data management. They propose that even outside the software engineering context, the CMM provides a model describing how with increased maturity, the organisation processes become "more refined, institutionalized and standardized, establishing a basis for process management, appraisal and improvement". This seems appropriate to turn an intangible concept of user need into a defined service. Crowston's and Qin's model focuses on the provision of RDM infrastructure such as policies, technology, and guidelines, but does not cover the provision of advisory services and support. (Crowston & Qin 2011). The Research Data Management Framework developed by the Australian National Data Service (ANDS) is another tool based on the capability maturity model. (Rice & Southall 2016, 75). There is, however, little information available on the implementation of the model. A different approach was taken by the Distributed Data Curation Center (D2C2) at Purdue University Library. Their Data Curation Profiles Toolkit provides an instrument for "data interviews" between an RDM service specialist and an individual researcher or research groups. This interview instrument can be applied to gain insight into researchers' RDM practices and support needs. (Carlson 2010.)

The CARDIO model (Collaborative Assessment of Research Data Infrastructure and Objectives) developed by the British Digital Curation Centre (DCC) is also based on evaluating the maturity level of capabilities. This model can be used to conduct a stakeholder meeting for the purpose of coordinating a common RDM service strategy. (Digital Curation Centre 2020.) DCC later developed the Research Infrastructure Self Evaluation (RISE) model based on experiences from the CARDIO tool. In this model, services refer to both the IT infrastructure and the advisory services provided by the organization. The evaluation is conducted by gathering information from various stakeholders: the library, IT services, and other relevant research services, such as legal experts, grant application and project management experts, etc. The RISE model emphasizes the efficiency of coordinating and developing comprehensive services in collaboration. The model identifies 10 areas of RDM, which contain 21 capabilities altogether. Each capability can be provided on one of four levels: level 0 means there is no support in this capability, level 1 means compliance with the basic requirements, level 2 is for services tailored to the organization’s needs, and level 3 is reserved for pioneers in the specific capability. The model describes requirements for each level, in each of the 21 capabilities. The evaluation consists of four phases: 1) understanding the context and planning the scope of evaluation, 2) current state evaluation, 3) setting goals for further development, 4) reporting and recommendations. (Rans & Whyte 2017.)

5 RESEARCH QUESTIONS, DATA COLLECTION AND METHODS

This thesis utilizes the user-centred approach, meaning that service development should be informed by user needs, behaviours, and experiences. The insights should be collected from users themselves rather than from the service staff perceptions of user needs and behaviours (Bury & Jamieson 2013, 43-44). Wilson (2013, 28-31) argues that in the practice-oriented sphere of library and information science research, the problems are fluid with changing requirements. In such context, utility or suitability of a solution is the main concern, rather than the discovery of 'truth' or testing a theory. The solution is, however, also a means to better understand the problem. Design science explores both the problem and the solution space, in an evaluative and iterative manner. This seems to be applicable to the still developing field of open science, RDM practices and infrastructure that are going to continue to evolve in reaction to developments in scientific methods, practices, technology, legislation, and funding policies.

The multifaceted character of the RDM support service landscape does not necessarily disqualify library service development as a possible framework to approach the problem. Even in cases where the library does not play a key role in RDM service coordination and provision, the concepts presented in the library and information science literature cited in this thesis can be useful in information service development. The literature on library service development is often interdisciplinary and applies some concepts and methods that have more tradition in management science (such as service design) and computer science (such as usability studies).

The shift towards service economy has led to increased emphasis on customer service and customer experience, which can be also observed in the non-profit and public service sectors. The customer experience-oriented approach brings into service development the collaborative input from customers (or, more suitably for a non-commercial service, users) and frequent assessment. Service design centres efficiency, seamlessness, customer focus and co-creation. Because a service is an intangible exchange that does not directly result in the possession of a concrete product, it is closely tied to personal experience. User experience with the service is therefore as important as the achieved outcome. Participatory design is a related user-centred approach with a similar toolkit. The service design approach however focuses more on the ecology of services interacting with one another and influencing the user experience, which appears more suitable for the heterogeneous network of services tackling the "wicked problem" of RDM. (de Jong 2014, 138-140, Marquez & Downey 2015.)

The specific methods used in service design can vary depending on the scope, time frame and resources available for the design project; however, gaining insights into user behaviour is crucial in order to create or refine services that meet user needs. On a general level three main phases can be recognized: inspiration (observation), ideation (understanding), and implementation. The inspiration phase includes selection of the design team, recognising and involving the relevant stakeholders. In this phase, preliminary data about user needs and behaviours are gathered and analysed. The methods used can be for example observation of service user behaviours, informal interviewing, or reuse of readily available quantitative data, such as usage statistics. (de Jong 2014, 144, Marquez & Downey 2015.) In the case of research data management services, usage statistics on their own bear minimal value to the purpose of service development. Survey appears to be more suitable to gain insights into the users' behaviour and needs.

The ideation phase involves more advanced data collection and formulation of steps towards solutions. In this phase common methods are ethnography, formal interviewing, focus groups and surveys. It is useful to involve various stakeholders or a group of people to facilitate collaborative and cooperative thinking. Internal stakeholders such as service specialists and management have knowledge of the service ecosystem and available resources. This is important to evaluate the feasibility of the solutions that have emerged from user insight. While users have important insights and often excellent ideas, not everything will be possible to put to practice. (de Jong 2014, 144-145, Marquey & Downey 2015.) For this phase, a focus group or interviews with various RDM key players are both suitable options. The implementation phase leads to demonstration of a solution, which in early stages can be a pilot or a prototype that is tested and assessed based on the feedback. This phase is meant to be iterative to ensure the solutions remain relevant for the users (de Jorg 214, 145). Implementation and production of the services are however beyond the scope of this thesis.

5.1 Research questions

Chapter 1.1 introduced the topic of this thesis - development of tailored services for Research Data Management and its Open Science aspects. This thesis aims to test a specific tool for organizations to gain insight into the practices and support needs in research data management specific to their community, while also staying informed about service needs arising from sources other than users' perceived need. Such sources identified in the previous chapters may be compliance with legislation and ethics guidelines, policies (national, organizational, funders'), FAIR data principles, or technical requirements for data security and long-term storage.

The first research question concerns user insight. Based on the literature review on RDM and open science practices and problems, we can identify potential user needs and aspects in which it would be useful to gain insight into users' experiences and attitudes. Adopting a service design approach, user insight survey questionnaire attempts to organize these potential needs in relation to the points at which users encounter a service. The survey aims to answer the question: *What kind of needs, attitudes and experiences do researchers have in RMD topics where support services could be useful?*

The second research question concerns putting user insight into practice in service development in a way that does not ignore service requirements from other sources listed in the beginning of this chapter. The previous chapter reviewed some tools for RDM service evaluation, and the RISE self-assessment tool shows potential for service development in a way that takes into consideration user needs, organizational research profile and the realities of each organization's resources and operational environment. The RISE self-assessment tool was piloted in a series of interviews with experts from various aspects of RDM services who were be asked to evaluate the tool and its suitability from the point of view of Finnish research environment. The interviews also aimed to discuss how the data from the survey can be taken into account when using RISE to self-evaluate RDM services. The RISE pilot should therefore answer the question: *How did RDM support experts evaluate the usefulness of the RISE framework for service self-assessment?*

The larger aim of this exercise is to test if this approach consisting of user insight survey and a service self-assessment framework could be useful and potentially replicated at different institutions in Finland to tailor RDM service development to user needs and essential external requirements. The thesis will also attempt to discuss these cross-cutting research questions: *Did the survey work together with the RISE model to include users? Based on the pilot, how useful were the survey and RISE as tools for service development?*

5.2 The research environment

The data was collected at VTT Technical Research Centre of Finland during the author's employment there in 2020, and Senior Specialist (now Team Leader) Anssi Neuvonen was the data collection advisor at VTT. VTT was founded in 1942 as a state-owned research centre, and it is steered by the Finnish Ministry of Economic Affairs and Employment. VTT specializes in research, development, and innovation in three main areas: carbon neutral solutions, sustainable products and materials, and digital technologies. These research problems represent societal issues with high relevance, and the mission of VTT is to provide solutions to their customers as well as society. Private sector forms a large customer group. (VTT 2023.) This is bound to bring IPR issues into the way research, development and innovation results can be shared. At the time of the data collection, VTT provided research data management services in a "distributed network" (Matusiak & Sposito 2017). It should be noted that the data collected in 2020 does not reflect the current situation in user needs and support services at VTT.

The data collected for this thesis was not used in the full service design circle to specifically evaluate and develop the services at VTT. Instead, VTT kindly allowed the data collection for the purposes of this thesis to explore whether such approach is usable and applicable to learn more about RDM needs and services in a specific organization. In that sense, this thesis could be characterized as a pilot case study where the “system of interest” forming a case is an academic organization’s RDM environment (the practices and the support system). A case study should use a framework which could enable comparability, and this thesis utilizes an approach which could potentially form such framework (survey questionnaire and the RISE framework for RDM service self-assessment). The kind of data which can arise from this framework in different contexts can however be diverse, because the framework is not standardized, and this pilot aims towards the exploration and expansion of understanding of the RDM service system rather than derivation of generalized rules. (Stake 2009, 3-7.)

5.3 Survey on the RDM landscape and support needs

In the first phase of this study (inspiration/observation), the goal was to gain preliminary understanding about the RDM practices, needs and attitudes of potential service users. Survey and focus group were considered for data collection. While focus groups are more participatory, they are not recommended as basis for any statistical generalizations about behaviours and attitudes. The existing preliminary insight on service user target group was also limited and insufficient for the qualitative purposive sampling recommended to reflect diversity within the investigated group. Purposive or theoretical sampling starts with theorizing about the aspects that are likely to create differences in perceptions or experiences, so that outliers can be included, and comparisons can be made in the generated data. (Barbour 2018, 17-20, 69.) Surveys are more suitable for gaining initial insights because they reach a larger group and allow some level of generalization about the sample group based on the results. The first research question is not directly quantifiable (*What kind of needs, attitudes and experiences do researchers have in RMD topics where support services could be useful?).* While quantitative data is useful to describe the general trends, it will not be used to draw conclusions about researchers in general. The approach is exploratory and involves qualitative elements. (Toepoel 2016, 2-3.)

A survey was planned and carried out to gain insight on researchers' attitudes and support and information needs concerning RDM. The questionnaire was drafted with a service design approach, with user needs and experience in mind. The questions were derived from service moments and service touchpoints. Service touchpoints refer to the physical or communicational points at which a customer or user experiences or perceives the service, such as a library reference desk or an occurrence of communication with service staff. Service moments are the episodes of service use during which the user encounters several service touchpoints. (Koivisto 2007, 66-67, Marquez & Downey 2015.) Based on the RDM practices and problems discussed in previous chapters, the following potential service moment situations can be identified:

1) The researcher is applying for external funding and is faced with the requirement to make data openly available.
2) The researcher has to write a DMP.
3) The researcher needs secure live storage for their data.
4) The researcher has to deposit the data in a repository.
5) The researcher cannot open the data due to legitimate reasons and needs long-term storage internally.
6) The researcher needs RDM advisory service or training.

The first and second draft of the questionnaire were consulted with two specialists from VTT familiar with the researcher workflows and services in the RDM and open science context. Feedback was also provided by thesis supervisor. The data collection advisor suggested involving an IT specialist at VTT to evaluate relevance of questions regarding service moment 3 (live storage). Service moment 5 (internal long-term storage) was added based on recommendation of the IT specialist. This was a fruitful addition as this service moment is important for RDM but can be overlooked if focused narrowly on open data. The questionnaire was piloted with two former researchers who are now in leadership or management positions and had experienced both the practical and administrative side of RDM. After final editing, the survey was launched in June 2020 as an online questionnaire in Microsoft Forms, a survey tool commonly used at VTT at the time. Microsoft Forms allow to collect responses anonymously so that personal data are not collected, and respondents may also feel more comfortable to express negative feedback anonymously. There was an option to add an e-mail address for those who wished to be contacted about a specific concern or question, but for those who did so, the identifiable data was deleted immediately after processing. The cover letter can be found in Appendix 1 and the full questionnaire in Appendix 2.

Before dissemination, it was necessary to define the studied population and find a sampling frame (list or resource used to find and reach the studied population). Since the survey's character was exploratory and there was no need to provide inferential statistical evidence for hypothesis testing regarding the population, the sample selection was not further refined. (Toepoel 2016, 55-57.) The population studied in this part of the survey was defined as researchers at VTT who have been exposed to the external requirements on RDM and open data practices, resulting in higher likeliness of interaction with support services. In practice this would likely be researchers who have been involved in publicly funded research. The data collection advisor at VTT helped identify suitable mailing lists as the sampling frame. One list included 67 principal investigators or project leaders in Research Council of Finland and Horizon 2020 projects. The survey was also sent out to the mailing list for training concerning the Horizon Europe framework (approximately 500 recipients). Out of this list of possible contacts, 59 chose to fill in the questionnaire.

Questionnaire survey can include different kind of questions, mostly intended to be analysed quantitatively, however, some types of questions can also require qualitative analysis. Questions usually aim to explore the respondents' behaviour, knowledge of certain concepts, and attitudes or opinions. Close-ended questions provide a selection of possible answers. Ordered close-ended questions ask the respondent to evaluate a concept on a scale, for example, strongly agree to strongly disagree or excellent to poor. Unordered close-ended questions are answered by selection from a provided list of options. Open-ended questions provide no response options, and respondents can answer freely without being guided into any particular direction. Open-ended questions are useful when the possible answer alternatives are not well known in advance or when it is beneficial to let the respondents elaborate. However, these questions also require more effort from the respondent and are more likely to be skipped. If there are many compulsory open-ended questions, this may lead to respondents abandoning the survey. (Toepoel 2016, 27-32.)

The questionnaire in this thesis contained several unordered close-ended questions exploring the respondents' behaviour related to RDM and open data, as well as preferences for support service delivery. Since the survey aimed to explore rather than gather statistical evidence on existing hypothesis, an additional "other" option was included where relevant, allowing opportunity to add options that could have been overlooked. Attitudinal questions regarding respondents' confidence and support needs in different RDM aspects and opinions about DMPs, RDM and open data were explored using ordered close-ended questions. The response options utilized a five-point Likert scale, a type of scale commonly used in attitudinal survey questions to evaluate level of agreement or disagreement with a statement, with a neutral option in the middle (Vehkalahti 2019, 35). The results of close-ended questions were analysed with basic descriptive statistic, calculating the frequency distributions and relative frequency (percentage of respondents who chose a particular option) (Christopher 2017, 55).

Open-ended questions were included in cases where most likely response alternatives were not previously anticipated, and it was beneficial to allow free response without guiding participants towards assumptions. For example, Q15 explored solutions for FAIR long-term storage of data which cannot be published in a repository, and Q17 investigated where respondents would go for help in DMP/RDM questions. While assumptions about alternatives can be made based on literature and existing institutional services, offering selection from these assumptions could lead the respondents towards what they assume to be the expected "correct" answer. Open-ended option was considered more suitable to learn about real-life practices and awareness of existing services. Optional open-ended questions were offered after attitudinal questions as an opportunity to elaborate or point out overlooked options, in order to encourage dialogue with respondents and sharing of opinions and feedback. The responses to open-ended questions were analysed depending on the character of the collected data. The questions that can be answered with a single concept or list of concepts (for example, Q12 If you have made the data or its metadata available, what service did you use, see Appendix 2) can be categorised with content analysis and quantified. Other answers were analysed with qualitative content analysis, categorised into clusters or conceptual categories that may or may not form larger themes, as some of the questions designed to elicit feedback were voluntary to avoid abandonment of the survey, and response rate could be low. (Given 2012, 121-122.)

5.4 Interviews with RDM service providers: the RISE model

The second research questions focused on providing adequate support services corresponding to the user needs. This is the aim of the ideation (or understanding) phase of service design, where the various stakeholders brainstorm for possible solutions to user needs or problems that emerged in the inspiration phase. There are several existing tools for the assessment of RDM services, some of which are in fact designed to facilitate participatory discussion among the stakeholders.

Out of the evaluation models described in chapter 4.2, the RISE framework was selected because it is comprehensive, covering a wide range of aspects important for RDM in and beyond the context of open science. The framework's background and application are also well documented in an accompanying user guide (Rans & Whyte 2017). The report from application of RISE at the 4TU.ResearchData consortium of Dutch technical universities by Dunning, Verbakel, de Smaele and Böhmer (2017) also provided initial evidence that the framework was perceived as useful and applicable outside its original context. RISE adapts the CMM approach for service development. The set of capabilities was defined based on literature reporting on survey research of institutional RDM services and academic libraries, as well as experiences with the CARDIO tool developed previously at the same institution (British Digital Curation Centre). The model was also tested in a workshop with RDM practitioners prior to publication. (Rans & Whyte 2017)

The authors address that maturity usually refers to the organization's capability to reliably perform and manage the examined process and are often described in quantifiable levels, however, they adapted the maturity concept differently. In the RISE model, capability refers to "the ability to generate an outcome", or provide service value, and the maturity

levels describe the service value available in each area rather than quantifying the maturity of each capability. The maturity levels of service value are described as:
1: Compliance
2: Providing locally-tailored services
3: Sector-leading activity
Level zero is used to describe a complete lack of support activity. The aim of this self-assessment is not to serve as a benchmarking tool between organizations – it is not necessary or perhaps even reasonable to strive for level 3 in all the capabilities. Compliance on level 1 should already cover the basic needs, and the local tailoring on level 2 allows flexibility to serve the needs in a given environment. This way of defining maturity levels can be a useful tool to help research institutions prioritise and consider what is feasible and desirable in their context. Available resources, institutional philosophy and the benefits and risks associated with further service development should be taken into consideration. (Rans & Whyte 2017.) This approach aligns well with the second research problem of this thesis, finding a way for organizations to evaluate and develop services with flexibility and sensitivity to their users and environment.

The user guide for the RISE framework describes four stages of the process. The first stage is setting the scope and identifying context of the self-assessment exercise. In this thesis, the scope is not defined by internal use for service design at VTT, rather, it is an initial pilot of the applicability of the RISE framework in the Finnish academic environment. The next steps in using RISE are ‘Classifying current support provision’ and ‘Designing the future service’. The authors propose RISE can be used as basis for semi-structured interviews or a workshop where these steps are collectively discussed. If future development is beyond the scope of the assessment exercise, the third step can be omitted or limited. The last stage, ‘Reporting and recommendations’, is also optional and the tool can be also used only to initiate discussion and connect the various key players. (Rans & Whyte 2017.)

A working group on Data services for researchers organized within the Openness of research data expert group of the National Coordination for Open Science and Research in Finland has developed a model for research data openness management support

services (Assinen et al. 2020). This model was incorporated in this thesis to add local context. The model describes 3 aspects of organizing RDM services (Management, Human Resources, Support services) on 3 levels similar to RISE:
1: The minimal model: What should be produced in a small organization with minimal resources?
2: The more comprehensive model: What can be done in larger organisations to provide sufficient support to researchers?
3: The vision: What could ideally be produced one day to create an almost automatic RDM workflow?
The levels 1 and 2 are comparable to the RISE framework levels. However, level 3 is described as a vision for the future that is not yet realistic to achieve even in sector-leading institutions, and therefore does not correspond to level 3 in RISE. (Assinen et al. 2020.)

The RISE framework combined with the Finnish model was used in this thesis as a basis for semi-structured interviews. The RISE user guide (Rans & Whyte 2017, 11) recommends that the capability *Advisory services* is evaluated also on the level of specific topics, to enable tailoring to institutional context and priorities. Therefore, the interview instrument was also enriched with visualized results from the user survey regarding the service moment "*the researcher needs RDM advisory service or training*" (see chapter 6.1.5). The same RDM and DMP topics evaluated by service users (researchers) were also used to evaluate advisory service capability in the interviews. The full interview instrument can be found in Appendix 3.

Brinkmann (2014, 286) notes that semi-structured interviews are most common in qualitative research. Compared to structured interviews, which follow a list of pre-set questions and are often used as a survey tool, semi-structured interviews allow more flexibility in following up on issues brought up by the interviewee and give the interviewer a more active role in the dialogue (Brinkmann 2014). In this thesis, structure is provided by the interview instrument. However, since RDM can be characterized as a "wicked problem" (Cox et al. 2014), it should be expected that interviewees with different roles in the organization will perceive the RDM problems and solutions in their own way. For each of them different capabilities of the model were thought to be relevant, and it was expected that the direction of the interview would be guided by their experiences and interpretations.

In the first RISE stage (identifying context), the full interview instrument was piloted with Interviewee 1, who is involved in DMP review and basic RDM support services and was able to help identify other relevant interviewees for specific topics, also reflecting the various areas of expertise usually involved in RDM services as described in chapter 4.2. The expertise and roles of Interviewees 1 and 2 correspond to what is often provided by academic libraries (see chapter 4.1). However, since organizational structures vary, it is more informative to specify the interviewees' expertise area rather than organizational unit. Altogether six experts were interviewed and each interview following the initial pilot focused only on the capabilities corresponding to the area of expertise of the interviewee, pre-selected by the interviewer. However, any other comments that emerged during the interviews were welcome because experts can have insights on topics that may not obviously fit into their usual tasks. Including additional experts in intellectual property rights management, cyber security, electronic laboratory notebooks and laboratory information management systems as well as a representative of researchers involved in data-heavy research was considered useful, but such experts were not available at the time of the data collection to participate in interviews. Table 1 briefly presents the roles and areas of expertise of each interviewee, and Table 2 shows which capabilities were discussed with each interviewee. If a capability was not initially planned to be discussed but relevant comments emerged during the interviews, the interviewee's number is presented in brackets.

Table 1. Interviewees and their expertise relevant to RDM.

Interviewee	**Expertise/role relevant to RDM**
Interviewee 1 "General RDM"	General RDM support and DMP review.
Interviewee 2 "Open Science"	Open Science, general RDM.
Interviewee 3 "IT"	Expert in IT services for research.
Interviewee 4 "Project management"	Project management and funding application expert.
Interviewee 5 "Legal"	Expert in legal services for research.
Interviewee 6 "Records management"	Expert in information classification, records management, and archival policy.

Table 2. Parts of the interview instrument and which interviewees were asked to evaluate them based on their expertise.

Capability/topic	Discussed (emerged) with interviewee:
1) RDM policy and strategy	
1a) Policy development	1, 2, (4), 5, 6
1b) Awareness raising and stakeholder engagement	1, 2, 3, 4, 5, 6
1c) RDM implementation roadmap	1, 2, 6
Finnish model: Management	1, 2, 6
2) Business Plans and Sustainability	
2a) Staff Investment	1, 2, 3, 4
2b) Technology Investment	1, 2, 3, (4)
2c) Cost modelling	1, 2, 4
Finnish model: Human resources	1, 2
3) Advisory Services - general	1, 2, (3), 4, (6)
Advisory Services – relevant individual topics	1, 2, 3, 4, 5, 6
4) Training	
4a) Online training	1, 2, (3), 4
4b) Face to face training	1, 2 (3), 4
Finnish model: Support services	1, 2
5) Data Management Planning	1, 2, (4), 6
6) Active Data Management	
6a) Scaleability and synchronization	1, 3
6b) Collaboration support	1, 3
6c) Security management	1, 3
7) Appraisal and Risk Assessment	
7a) Data collection policy	1, 2, 3, 5, 6
7b) Security, legal and ethical risk assessment	1, 2, (3), 5, 6
7c) Metadata collection to inform decision-making	1, 2, 3, 4, 5, 6
8) Preservation	
8a) Preservation planning and action	1, 2, (3)
8b) Continuity Support	1, 2, (3)
9) Access and Publishing	
9a) Monitoring locally produced datasets	1, 2

9b) Data publishing mandate	1, 2
9c) Level of data curation	1, 2
10) Discovery: Metadata cataloguing scope	1, 2

The interviews were scheduled during November-December 2020. Due to the COVID-19 pandemic, interviews were conducted remotely via Microsoft Teams. The participants were sent the interview instrument beforehand, and the instrument was shared on screen via Microsoft Teams during the discussion, to allow the interviewee to re-read and evaluate the maturity level definitions during the discussion. Interviews were conducted in Finnish, the native language of the interviewees. Participants consented to video recording of the interviews with the option to turn their camera off. The video footage of cursor movements through the shared interview instrument was useful during analysis to follow what was being discussed. The interviewees were informed that video and audio recordings would be deleted after transcription and analysis.

The interviews focused on stage 2 of the RISE self-assessment process, classifying current support, while evaluating the maturity level definitions for relevant capabilities with the interviewees. Stage 3, designing the future service, was also discussed in the interviews, however its purpose was only to facilitate the evaluation of the capabilities and maturity levels. While the discussion could provide the interviewees with new insights and inspiration for future development, evaluation of services at VTT was outside the scope of this thesis. Therefore, the reporting and recommendations stage was not included as part of the data collection and analysis in 2020. The interview data were not analysed quantitatively to place current support or future development goals on corresponding maturity levels, however, comparing the capabilities and maturity level descriptions to what is feasible in the organization's context allowed the interviewees to think about the presented capabilities and maturity levels and reflect on how they relate to their own experiences and understanding of the RDM problems and solutions.

Tuomi and Sarajärvi (2018, 103) state that content analysis is one of the most common methods in qualitative research. It can be used to systematically search qualitative data

such as interview transcripts to find meanings. Qualitative content analysis usually follows these three main steps:
1) Reduction: sorting through the data and selecting what is relevant to the research questions or problems, reducing the original utterances into summarized statements.
2) Clustering: looking for similarities or differences in the reduced statements, grouping them into categories.
3) Abstraction: conceptualization, moving from the categories to larger themes or theoretical concepts they represent.
In theory-driven content analysis, the clustering is based on existing conceptual system, theory, or model. In the data we search for specific observations of the general concepts. The analysis framework is defined as the first step. The data analysis then follows the reducing and clustering stages, and categories are formed not directly from the data, but as subcategories of the existing themes or concepts defined in the analysis framework. Emerging categories which do not fit under the themes can be placed outside the framework. This approach can be used to test or evaluate the prior conceptual system in a new context. (Tuomi & Sarajärvi 2018, 110-111, 117, 122-131.)

The interviews were conducted to answer the research question *How did RDM support experts evaluate the usefulness of the RISE framework for service self-assessment.* Because the question was posed as evaluating a pre-existing model, the theory-driven approach was appropriate. The RISE self-assessment tool served as the analysis framework. While the interview instrument was enriched with the Finnish service model and an excerpt from the survey data, these were mainly used in the interviews to compare with and evaluate the concepts in the RISE framework. The transcribed interviews were reduced to statements which reflect on the usefulness of the tool, describe relevant experiences, and offer opinions on the adequacy of the maturity level definitions in practice. These statements from each interviewee were categorized under the capabilities (themes) they were related to. Second iteration of clustering focused on finding similarities or differences unique to the expertise of a specific interviewee. The results present how thoughts expressed by the interviewees related to the themes – were the capabilities considered adequate, or did new concepts arise from the interviews.

6 RESULTS

In this chapter the results of both the survey and interviews are presented. The survey results are organized based on the service moments identified in chapter 5.3. Interview results are organized into five subchapters combining related capabilities from the RISE model and the Finnish model of RDM services.

6.1 Survey

The survey questionnaire was completed by 59 respondents. Most respondents (95 %) had 5 or more years of experience in research. The rest (5 %) had been working in research for 2-5 years.

6.1.1 Service moment: The researcher needs to write a DMP and comply with funders' requirements

The first potential service moment situation explored was needing to write a Data Management Plan. Only 20 % responded that DMPs were not relevant for them personally, while 63 % had written a DMP before, and 17 % were preparing to write their first DMP for an upcoming grant application. Figure 4 shows how those who had a prior experience evaluated the difficulty of writing a DMP (the question was answered by 68 % of all respondents). The average was leaning towards slightly difficult. Although 14 % found writing a DMP easy and 33 % felt neutral, services should be available for the 45 % and 8 % who found the DMP process quite or very difficult.

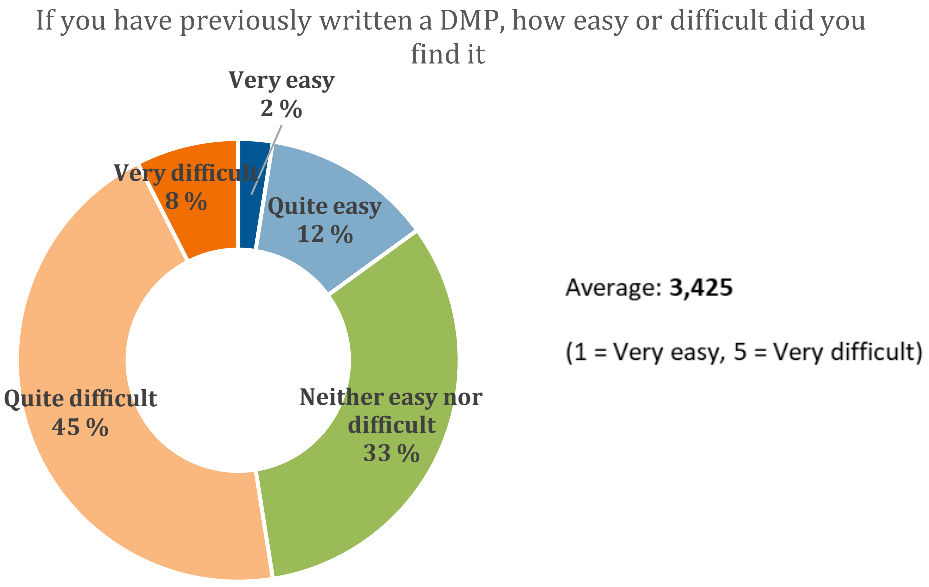


Figure 4. Perceived difficulty of writing a DMP.

The DMP is often required when applying or receiving external funding. Therefore, following funders' open science requirements was explored in the survey as part of the same service moment. Since the focus of the survey is on research data management services, questions about open access to publications and about other aspects of grant application writing were not included. According to the results presented in Figure 5, most researchers appreciated that a good research data management section in grant applications could help them secure funding. While 10 % disagreed with the statement, there were no respondents who strongly disagree.

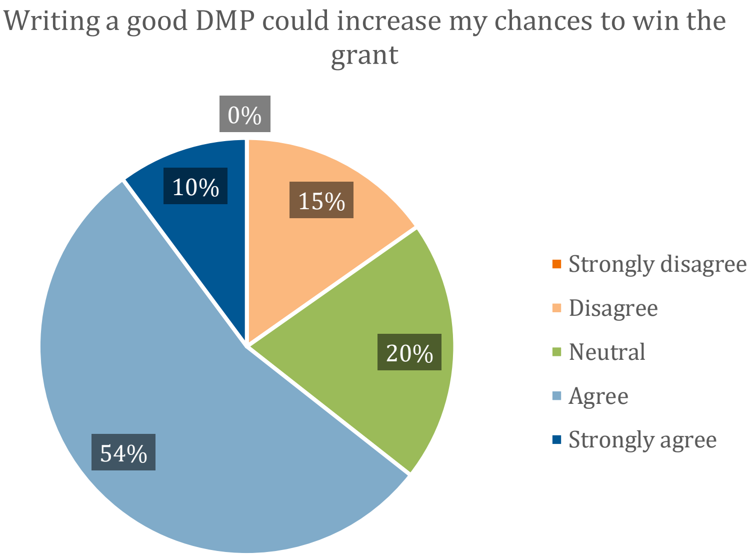


Figure 5. Opinions about effect of DMP quality on success of grant application.

Especially in projects with collaboration of multiple organizations, it is beneficial to reach mutual agreement within the collaboration on what tools or conventions to use during the project. Figure 6 shows how respondents agreed or disagree with the statement that a DMP supports collaboration: 62 % agree or strongly agree that creating a DMP can help establish a smooth collaboration on research data management within the project. Only 16 % disagree or strongly disagree.

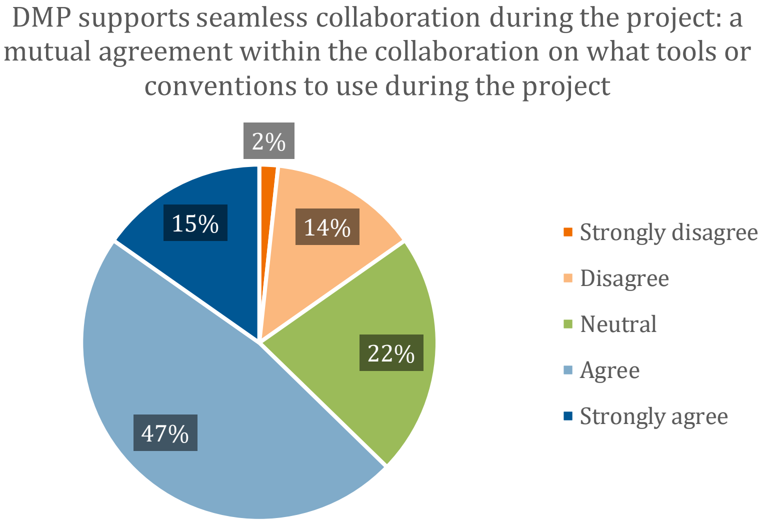


Figure 6. Opinions on whether DMP improves collaboration during the project.

Quality of research and the role of DMP in good and efficient research data management could be a strong motivating factor. As shown in Figure 7, overwhelming majority agreed that planning can make research data management during and after the project more efficient and easier.

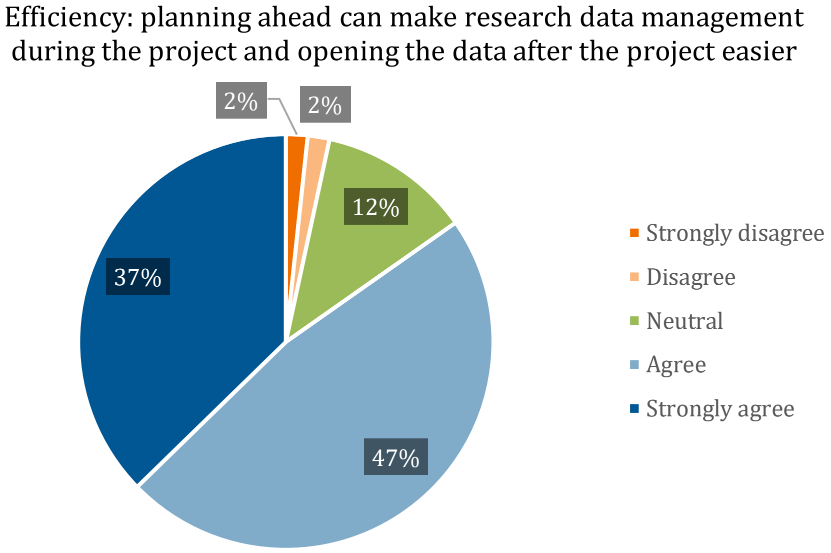


Figure 7. Opinions on whether planning ahead makes research data management during the project and subsequent opening of data easier.

In a following open-ended questions about other pros and cons of creating a DMP, 13 respondents provided additional comments. One respondent mentioned an important advantage that was not included in the survey: the DMP can help recognize possible risk points connected to the data collected or processed in the project, such as risks concerning data protection, privacy or business sensitivity. Risk management could be a potential motivational factor in drawing attention towards research data management. Other open comments helped point out possible problems linked to DMPs. Creating a data management plan is a lot of extra work (n = 4). Especially the initial DMP in the proposal phase can be perceived as “efforts lost” if funding is not granted. Other respondents pointed out that creating the DMP after the project has already started could also be problematic. Some things are already defined by the consortium agreement and researchers might have already started doing things in their own way before a shared plan is created.

Some respondents perceived a gap between the plan or basic principles, and the reality of actual research data management practices (n = 4). The DMP in their opinion does not in fact help establish clearly what needs to be done in practice, and what are the appropriate technical solutions. In connection to this perceived gap between the plan and the reality, some respondents (n = 3) mentioned that a DMP is more a bureaucratic necessity than a useful tool for researchers. They elaborated that the DMP is required by the funders, but not properly reviewed by them with feedback to the researchers.

6.1.2 Service moment: The researcher needs secure live storage for their data

As discussed in chapter 3.3, usual storage solutions include servers or storage drives managed by organisation, allowing centralized back-up solutions and secure data sharing within organization. Cloud storage or other services can be approved by organization as secure solutions for data storage and sharing with partners across organizations and countries. The less desirable solution is local storage on computer hard drives and external hard drives or USB sticks. These physical media are not a part of services that can automatically backed up by the organization, they are susceptible to damage, and may be lost or stolen.

IT experts from VTT provided a list of available solutions for data storage during the research. The specific solutions are internal information which is not necessary for the aims of this thesis and therefore will not be presented here. Respondents were asked to select the solutions from the list that they used in their research (multiple choice question with an option to add others, to be able to gain information about possible alternatives not listed by the organizational IT).

6.1.3 Service moment: The researcher needs to make data openly available via a repository

Respondents were asked whether they agreed with the statement that opening the data could increase the impact and visibility of their research. Figure 8 shows that the majority agreed with the statement, and none of the respondents selected the strongly disagree option. Impact and visibility seemed to be a strong motivating factor for making research data openly accessible.

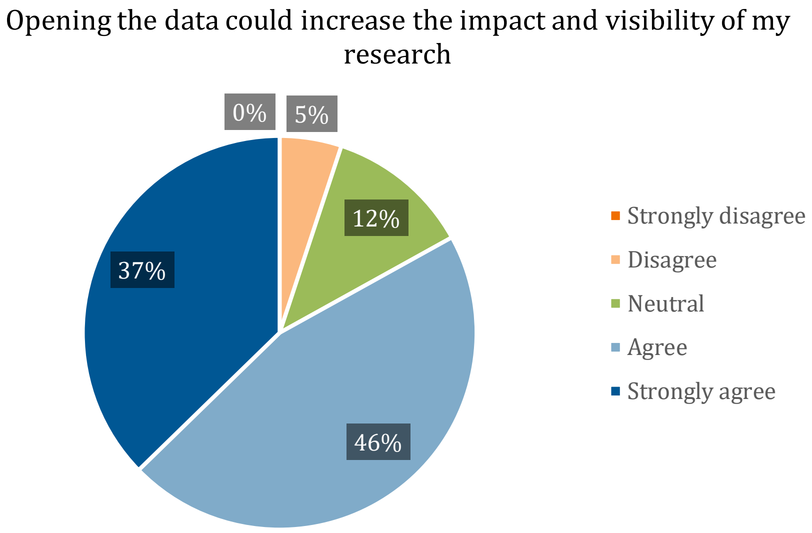


Figure 8. Opinions on whether opening research data increases the impact and visibility of research.

Despite the belief that opening the data can increase the impact and visibility of their research, 71 % said they have not made research data underlying their published results available in a repository or data journal. In a multiple-choice question, 24 % responded that they have made their data available with open access, 17 % with managed access, and 3 % have only published the descriptive metadata of their data. Out of the 59 respondents, 15 answered the optional open-ended question to specify which service they used to share their data (with possibility to enter multiple answers). Zenodo was the most used repository (n = 7) alongside other general-purpose repositories (n = 2), institutional repositories (n = 2), discipline (n = 1) or data type (n = 1) specific repositories. One respondent shared data via a journal.

Use of repositories or data journals is recommended in the FAIR principles because they enable FAIR elements such as persistent identifiers, landing page with descriptive metadata, or ability to define license. Two respondents named a metadata catalogue used to share the descriptive metadata, and two provided managed access to data stored on institutional servers. Providing access to metadata via a metadata catalogue may be considered FAIR-compliant when open access to data is not possible. Managed access restricted based on data protection or confidentiality issues may also be FAIR-compliant (as discussed in chapter 3.1). Two more respondents said they had shared their data via a company or project-maintained website. This solution rarely follows the FAIR principles, because there usually are no persistent identifiers and maintenance may end abruptly when the project ends, or the interests of the company change. Data on websites is also more difficult to find via a common search engine than research data with standard descriptive metadata in repositories.

Question 16 addressed the respondents' experience with reusing research data collected or created by others. Such experience can draw the researcher's attention to how FAIR (findable, accessible, interoperable, reusable) the data is or is not, and what they should pay attention to when they produce, document, and share their own data. Only 42 % reported having experience with reusing data collected and shared by someone outside their own research group or project, and 58 % did not have experience with such data reuse.

Respondents who had made their data available were asked to select the reasons that motivated them to do so. Twenty respondents (34 %) answered this multiple-choice question. The results are shown in Figure 9. The most common reason was improving the visibility and reusability of research results, which corresponds with the belief that opening the data can increase the impact and visibility of research. Second most common reason was funder's requirement to make the data available. These reasons can help support services tailor the information they provide to respond to the needs to raise the visibility and reusability of research, and to comply with funders' requirements.

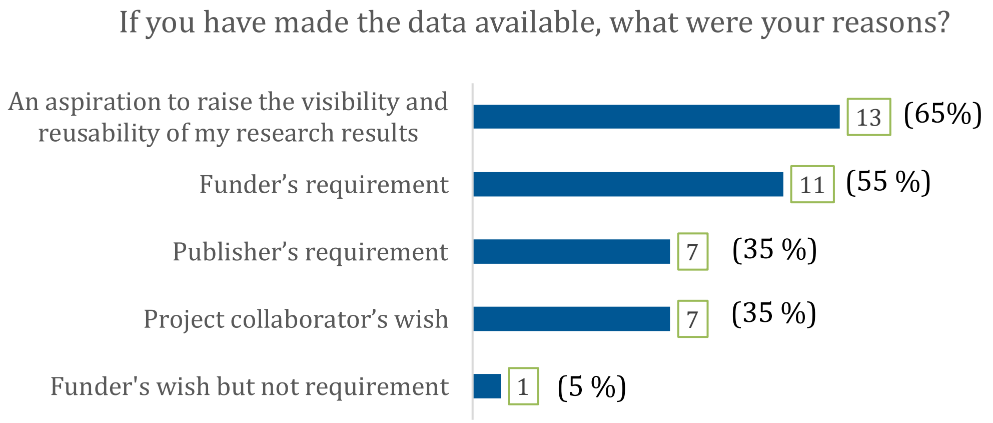


Figure 9. Reasons why respondents made their data available.

6.1.4 Service moment: The researcher needs internal long-term storage

Respondents were also asked if they had encountered any issues that prevented them from opening the research data or a part of it. This was a multi-choice closed-ended question with the option to add other reasons, because respondents could have encountered multiple limitations with different data types or projects. The predominant issues may vary in organizations based on their research profile, and understanding these issues is useful information for support services to further tailor their resources. The results in Figure 10 show that 31 % said they had not encountered any such issues. The most common reason for not opening research data were intellectual property rights and confidentiality, likely due the high proportion of commissioned research and private sector customers (see chapter 5.2). Data protection issues were less common. Lack of resources for the associated work and not enough knowledge about how or where to open data also seem to affect around 25 % of the respondents.

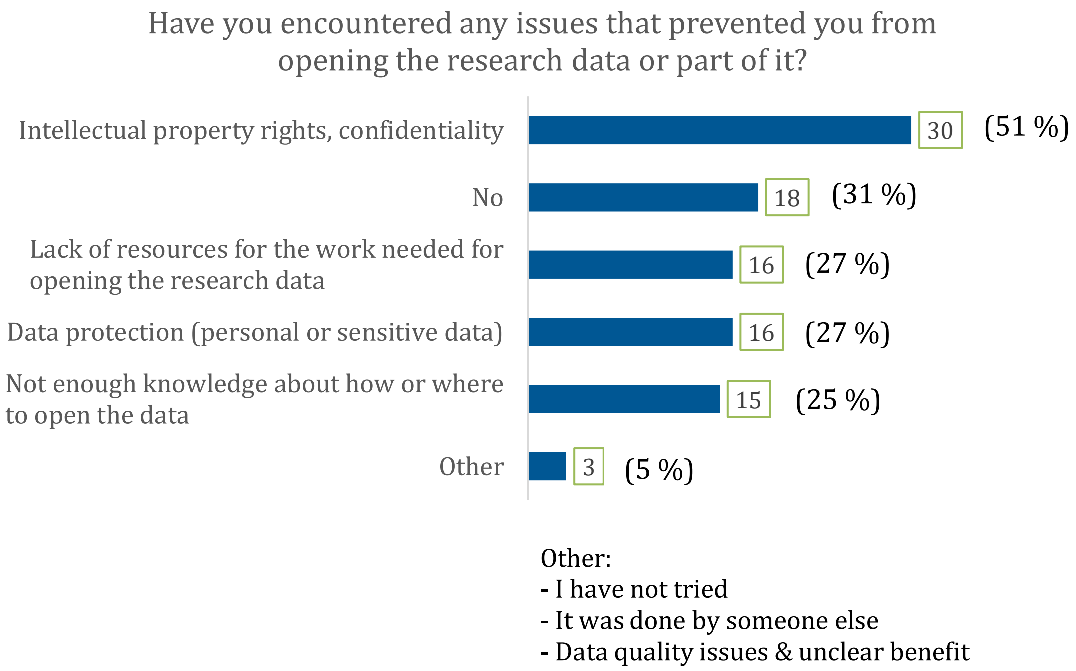


Figure 10. Reasons why respondents did not make their research data openly accessible.

If data cannot be openly accessible, researchers may need other solutions for long-term storage for the final stage of research data life cycle, enabling validation or reuse stage after the original project is concluded. The respondents were asked where they would store data that must be kept internally but should be findable and accessible for further research or sharing within the organisation. This was an open-ended question since the possible alternatives were not straightforward and it was also considered useful to elicit free responses without suggesting the "correct" answers. Many respondents reported using organizational services suitable for the purpose, which are not presented here since they contain VTT internal information. Many also added longer comments to their selected storage solution or lack thereof. They underlined the importance of findability and accessibility of the data stored internally, as well as awareness how this should be done correctly. Also the issues of cost and effort needed for internal findability, accessibility and reusability were recognised.

> That is the million euro question! Data is stored in many places, and old data is very difficult to find or use, as there is no commonly agreed policy on what, how, where and even why. (Survey results, Q15)

6.1.5 Service moment: The researcher needs RDM advisory service or training

This survey also aimed to find out what subjects the respondents need support in, and what form of support they would prefer to receive. The first question was explored by asking the respondents to evaluate their confidence in aspects of RDM defined previously in chapter 3 and framed to correspond to language researchers may know from some of the most common DMP templates, which cover all major RDM issues as well as external requirements (DMPTuuli 2023). The results are visualized in Figure 11. Respondents were also asked to evaluate the benefits of targeted support in these aspects, which can help prioritize or accentuate the most relevant topics in guidance and training materials. The perceived benefit of support is reported in Figure 12.

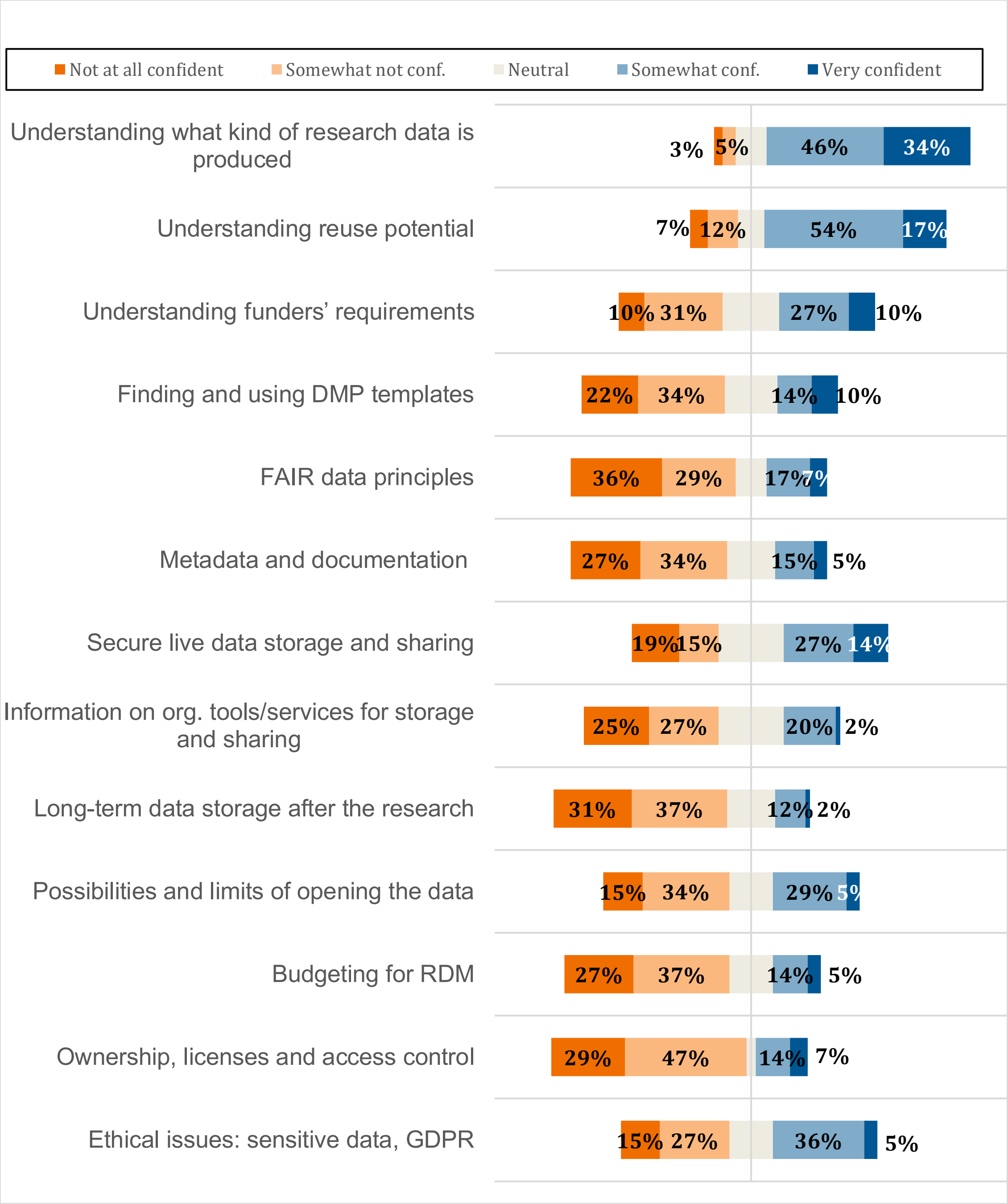


Figure 11. Respondents' confidence in various aspects of data management planning.

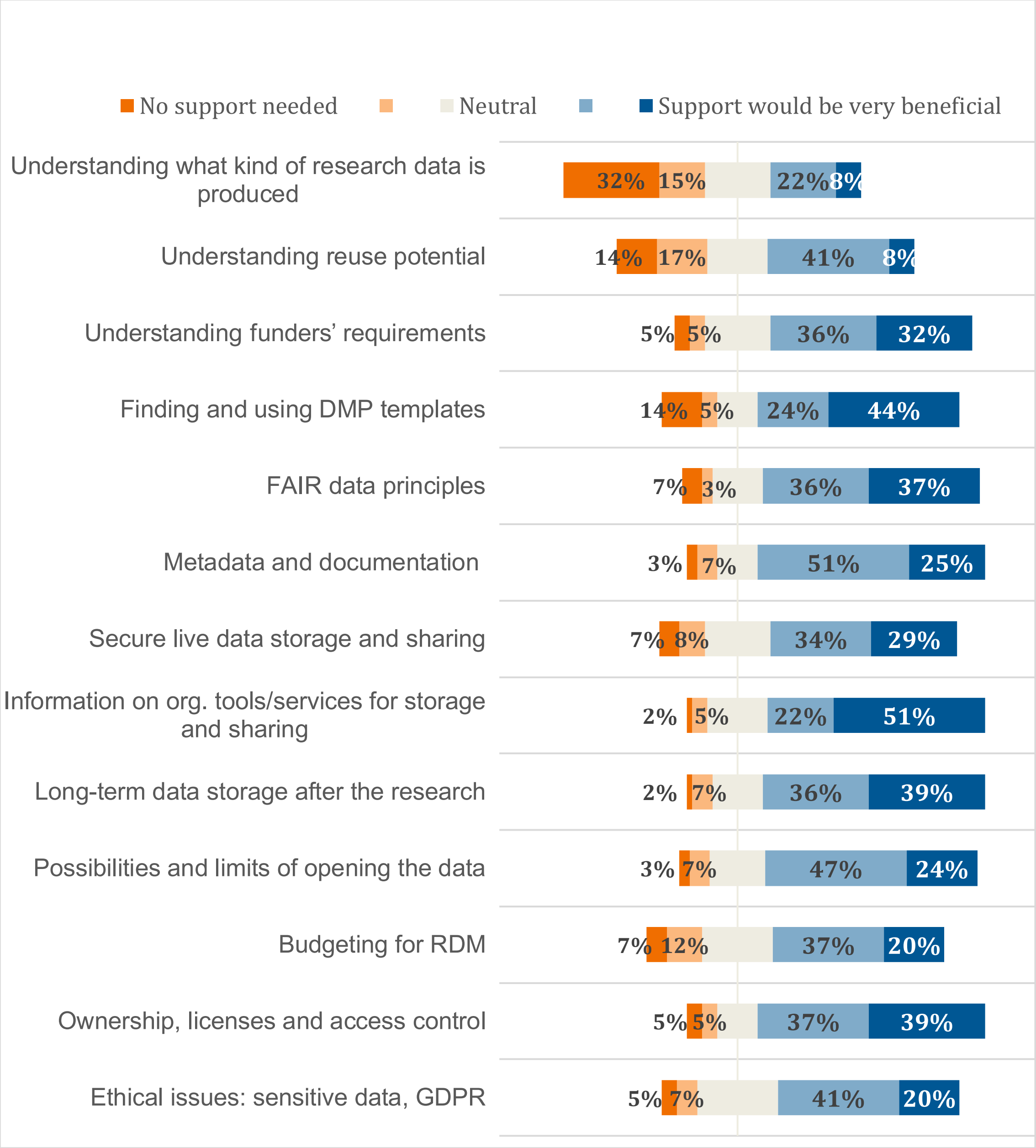


Figure 12. Perceived benefit of support or training in different aspects of DMP.

The following Table 3 sorts the topics in the order of level of confidence (from lowest to highest) and in the second column topics are sorted by the perceived benefit of support (highest benefit to lowest). The highest percentage of uncertainty was expressed in agreements on data ownership, licenses and access control, with a corresponding percentage of respondents who thought that support in this area would be beneficial. This is not very surprising considering that VTT is largely involved in research and development projects for customers, often from the private sector (see chapter 5.2), and in chapter 6.1.4 intellectual property rights and confidentiality were reported as the most common reason why data could not be made openly accessible (51 %). On the other hand, personal data and ethical issues regarding sensitive personal data were much less common (27 %), and that could contribute to the reason why respondents did not express as significant uncertainty (only 42 % not confident). Support was however still perceived as beneficial by 61 % of respondents.

Lowest uncertainty (8 %) was expressed in understanding what kind of research data the project produces, which is usually the starting point of DMPs, the characteristics of the data defining how it should be managed. Despite such high confidence in this topic, 30 % still thought that support would be beneficial. Overall, the reported level of uncertainty was below 50 % in six areas. Only in two of them, perceived benefit of support was also below 50 %: in addition to the previously mentioned understanding what kind of research data the project produces, also the area of understanding how the research data could be beneficial also to other users after the project scored very low uncertainty (19 %), while a much larger percentage (49 %) thought support would be useful. In the vast majority, the perceived benefit of support was higher than expressed lack of confidence, for example, metadata and documentation (61 % not confident, 76 % would appreciate support), information on institutional tools and services for storage and sharing (52 % were not confident, 73 % thought support would be beneficial), or understanding the possibilities and limits of opening data: only 49 % not confident, but 71 % would find support beneficial.

Table 3. Comparison of topics in which respondents were not confident and topics in which support would be beneficial. Low confidence combines percentage of responses "Not at all confident" and "Somewhat not confident", Support would be beneficial combines the responses of "Support would be very beneficial" and "Some support would be beneficial". Where score is equal, the statement with higher proportion of "Support would be very beneficial" is ranked higher.

	Low confidence		Support would be beneficial
Agreements on data ownership, licenses and access control	76 %	Agreements on data ownership, licenses and access control	76 %
Possibilities of long-term data storage after the research	68 %	Metadata and documentation of the research data	76 %
Complying with the FAIR data principles	65 %	Possibilities of long-term data storage after the research	75 %
Budgeting for research data management	64 %	Information on VTT tools and services available for storage and sharing	73 %
Metadata and documentation of the research data	61 %	Complying with the FAIR data principles	73 %
Finding and using DMP templates (incl. the language used in templates)	56 %	Understanding the possibilities and limits of opening the data	71 %
Information on VTT tools and services available for storage and sharing	52 %	Finding and using DMP templates (incl. the language used in templates)	68 %
Understanding the possibilities and limits of opening the data	49 %	Understanding funders' requirements related to data management	68 %
Ethical issues: sensitive data, GDPR	42 %	Secure research data storage and sharing during the active phase of the research	63 %
Understanding funders' requirements related to data management	41 %	Ethical issues: sensitive data, GDPR	61 %
Secure research data storage and sharing during the active phase of the research	34 %	Budgeting for research data management	57 %
Understanding how the research data could be beneficial also to other users after the project	19 %	Understanding how the research data could be beneficial also to other users after the project	49 %
Understanding what kind of research data the project produces	8 %	Understanding what kind of research data the project produces	30 %

In a follow-up open-ended question, the respondents were given opportunity to add any other support needs not covered in the previous question. Seven respondents added their comments. One of them was unsure how much detail needs to be provided in the planning stage, another respondent thought that it would be useful to have access to existing high quality DMPs for reference. Two respondents noted that DMPs need to be maintained during the project, and it would be good to know the principles for updating when new or unforeseen data management issues arise. Two respondents emphasized the importance of the organization thoroughly describing their technical solutions for data storage in terms of their security, with technical and legal vulnerabilities in mind. Two of them also mentioned practical guidance on how to implement personal data protection, including anonymization and secure storage. One respondent highlighted that while data storage during the project is relatively easy, there is usually a need to keep the data available after the project, which requires a plan and funding. Two respondents further emphasized that while writing a DMP in general terms is not difficult, the practical applicability can be more challenging.

> Most if this is relatively clear in the general level and it is easy to find general level support and training. When things get real and you are collecting data that might be sensitive/confidential/GDPR-related, it is much more difficult to find practical knowledge what needs to be done and what is the reasonable level of actions. (Survey results, Q7)

In addition to the topics in which support would be needed, the survey also aimed to identify the current touchpoints where users would go to interact with the support services. Because the various topics were introduced in the context of a DMP, respondents were asked where they would look for help when writing a DMP. The question was open-ended, because the aim was to gauge how aware respondents were of different services and resources. Especially interesting is to note if there are respondents who are not able to name a single service, or to what extent respondents would ask their peers rather than contact a service. Depending on the goals of the organization conducting the survey, it may be preferrable for the service to be known and approached as a service, or for example in smaller organizations it may be preferrable that respondents know specific names and establish personal contact. The responses from the pilot survey at VTT are not relevant for the purposes of this thesis and contain personal identifiers (names) as well as internal information, therefore they are not analysed here.

After exploring the current touchpoints, the survey also investigated how the users want to be approached and interact with the services, i.e.., what would be their preferred service touchpoints or formats of support that would fit in with their usual workflows. The survey offered a multiple-choice list of forms of training and support, including peer support among researchers themselves. As presented in Figure 13, the respondents showed strong preference towards resources that can be used independently on their own schedule. While 78 % would like to have access to online materials such as guides or instructions, only 41 % would like to have courses and training available, and 20 % would like to be able to attend workshops. Over half of the of respondents (56 %) would like to seek support directly from their own contact network within the organization, such as other researchers or someone from the support services they know through their work. An equal percentage selected an organizational data support service with a helpdesk or similar single contact point for various specialists.

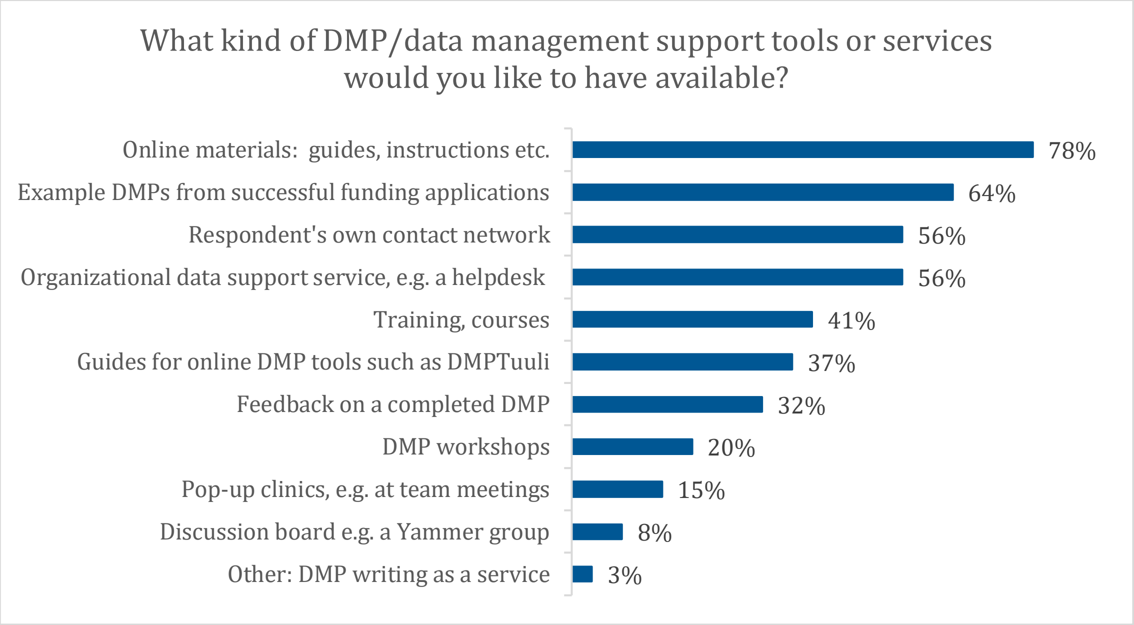


Figure 13. The types of data DMP or data management support tools and services the respondents would like to have available. Multi-choice.

6.1.6 Other comments and feedback from the respondents

Respondents were given opportunities to give feedback or comments throughout the survey. Question 7 “Any other support needs not covered in the previous question?” allowed respondents to point out support needs missed by Q5-6. Question 9 “Any other perceived pros (or cons) of writing a DMP?” gave opportunity to express opinions regarding the benefits of planning the research data management other than the statements provided in Q8. At the end of the survey, more expression of own opinion was encouraged in final Q19 “Other comments or feedback?”. Altogether 16 respondents gave feedback and comments in these open-ended questions. The anonymous respondent IDs generated by the survey tool were checked to identify similar responses to multiple open-ended questions from the same respondent. Such repeated comments from the same respondent were considered one response. The responses to all three open-ended questions covered similar themes.

- Theme 1: Evaluation - what is a high-quality DMP with sufficient level of detail? Need for practical examples, reusable materials. (n = 5)

  Respondents expressed uncertainty about what a good DMP looks like, how much detail is required and how a DMP should be structured when it is a project deliverable. Examples of good practice and learning from others’ experience would be welcome, e.g., DMPs from successful proposals for reference.
- Theme 2: The plan and principles versus the reality: what are the appropriate technical solutions and what needs to be done? (n = 5)

  These examples express the common thread:

> The DMP and the actual data management seem to generally be two different things. Having a plan is needed for the grant application, but this does not mean that the plan will be implemented. […]. The actual data management is probably more important in the long run than making the plan. (Survey results, Q19)
>
> I'd rather take help on the actual work, not just bureaucracy (although the latter helps, too, but I would really appreciate someone helping with the actual work, not teaching how to do it). (Survey results, Q19)

- Theme 3: Highlighting the need for support services (n = 4)

A few respondents have noted their appreciation for the survey as an effort to develop support services or tools that would be more easily accessible and more visible.

6.2 Interviews

As described in chapter 5.4, the RISE framework in combination with the model of RDM services developed in the Finnish National Open Science Coordination was used as a basis for semi-structured interviews with six experts in various aspects of RDM. The interview instrument was also enriched with data from the user insight survey in topics where user preference can be very important (Advisory service and Training). The full interview instrument can be found in Appendix 3.

6.2.1 RDM policy, strategy, and sustainable business plan

Higher-level capabilities 1) RDM policy and strategy and 2) Business plans and sustainability were overall evaluated as useful to reflect on as a research institution. In policy development (1a), funders' requirements are often an agent of change in mindset (Interviewee 4 "Project management"). This has also been noted in chapter 2.4 in regard to funders' policies effect on open science practices. The interviewees also pointed out some challenges in policy and strategy development. These capabilities are perhaps more visible to the support service experts than to users, who are however affected by policy and funding decisions. While legal and regulatory obligations as well as funders' requirements are reflected in institutional policies, the responsibility for knowing their data is on the researchers and oftentimes the policy is defined on a very general level. Policy does not always easily translate into practices, and it is not to be taken for granted that the policies and responsibilities are internalized and implemented.

> … of course, when such policies are written they often stay on the level of as open as possible, as closed as necessary and then no one tells you what next, what is closed as necessary, where does the necessary come from and what issues can cause it. And I feel bad for the researchers, that there are so many things that can trigger

> the requirement to keep data closed such as export control, personal data protection, trade secrets, various things, so do they recognize the "closed as necessary". The researchers sure carry a lot of responsibility for their own actions, projects, knowing and remembering the requirements related to their own data. (Interviewee 5)

The description of Level 2 capability in policy development includes "institutional policy articulates … its rationale for retaining data of long-term value". This was recognized as a useful capability. Internal retention of data of long-term value is often not defined in institutional policies for research data or lacks systematic practices. Interviewee 6 (Records Management) brought up the organizational Records Management Plans (*tiedonohjaussuunnitelma* – TOS) where the policies for retention of documents, their level of openness or protection and related processes are usually defined. While traditionally the TOS would involve administrative records and certain types of outputs such as research reports, this type of document could be extended to apply similar logic to research data produced at the organization.

The capability 1b) Awareness raising and stakeholder engagement somewhat addresses the concerns about application of policies in practice. This capability was discussed with all interviewees, as they were all representing stakeholders in RDM policies. Some interviewees expressed that the wording of capability levels as "policies are promoted" may not be sufficient, because even when policies are promoted, this does not automatically mean the promotion is successful, awareness is increased, and policies are implemented. Promotion campaigns can fail when they come at the wrong time, in the wrong place, or in irrelevant context. Level 2 (Guidance on how to apply policies to the institutional context is provided and promoted) would seem to get closer towards the goal of awareness raising. It is easier for staff, students and researchers to engage with a policy that is linked to relevant practices with concrete guidance what steps to take to fulfil policy requirements. Interviewee 5 (Legal) mentioned practical guidance on "as open as possible, as closed as necessary", and interviewee 6 from their point of view of records management suggested specifying recommended storage solutions and processes suitable for the retention period and openness or protection level of certain types of data.

Level 3 in awareness raising and stakeholder engagement emphasizes the component of engagement. On this level, policy is promoted to staff and students via channels commonly used for engagement with these groups. Interviewee 3 (IT) mentioned communicating about policies and new services to existing customers, making the promotion context and target group specific. Interviewee 4 (Project management) mentioned that researchers often become engaged with RDM when they have to implement institutional and funders' policies in funding application and project management contexts.

> The proposals include a compulsory section where you have to describe these things, how they will be arranged. And also at the proposal stage we have to allocate resources needed to implement these things. Then when we are doing the projects, when we get a positive funding decision, well then of course we have to carry out the project in the way that was presented in the proposed research plan, so nowadays practically in all projects we write the data management plans. And it's just like, when these things are not familiar and you're used to the commercial side then even if you have been to some training then you don't, your thinking doesn't just switch to open data by default. (…) Well, this is not an individual researcher's problem but also the research teams have to think about how they will arrange the data management and the openness questions. (Interviewee 4)

In the evaluation of capability 1c) RDM implementation roadmap, the term "roadmap" was not easily understood, and it could benefit from clearer definition or a specific example in the accompanying guidance. When applying this model to the academic environment of another country, terminology differences and different ways how academic institutions are organized can increase the need to define capabilities in more detail. The interviewer and interviewees agreed on an understanding of "roadmap" as the way RDM policies are implemented and followed up, for example in strategic decisions about what support services and technical infrastructure must be provided and funded. It remained unclear whether the capability requires that such roadmap is formalized as part of institutional policy or strategy. Interviewees considered the implementation roadmap dependent on the development of policy in 1a) and readiness of other basic components. Level 2 description "Roadmap is informed by the institution's strategies and its researchers' priorities" is well aligned with the approach in this thesis that RDM service evaluation tools should take into account the organization's and service users' needs. Researchers' priorities were investigated through the user insight survey. As discussed in chapter 5.2, the RISE evaluation stages 2 "classifying current support and 3 "designing the future service" are not the same for every institution but should be informed by each institution's strategies and operational environment.

The interviewees who work more closely with policy and information management (1 "General RDM", 2 "Open Science", 6 "Records Management") were also asked to evaluate the Management part of the Finnish model, related to the topics of RISE capabilities 1 and 2. As noted in chapter 5.2, the Finnish model focuses on 3 aspects of RDM services in relation to opening research data, a smaller scale than the whole battery of 21 capabilities in RISE. While both models define 3 levels for each aspect or capability, the meaning of these levels is a little different. In RISE, level 1 describes basic compliance and in the Finnish model it describes the minimum level services with which even small organizations can provide the basics. Level 2 in RISE is locally tailored to the needs and priorities of the organization, and in the Finnish model level 2 is a more comprehensive version that larger organizations should provide to fulfil the needs of their researchers. The approach adopted in level 3 differs the most. In RISE level 3 description is a suggestion for existing sector-leading capability maturity level, and in the Finnish model level 3 is the vision or dream for future development.

On level 1, the management understands the importance of open data. Interviewees 2 and 6 agreed that this is indeed the minimum requirement, because any higher education institution and research institute applying for public funding must understand the importance of open science policies and even if there were no other motivations, funders' policies are the basic driver of open data. According to Interviewee 1, the statement that "management understands the importance of open data" is however difficult to define and measure. In the Finnish context, the national open science declaration (Avoimen tieteen koordinaatio 2020) was signed by a number of research organizations including VTT to express their commitment to implementing the policy. Perhaps the management approving signature of the national policy can be interpreted as the management understanding the importance of open science, and therefore also open data. On level 2, the management invests in open data. This is further specified as goals, resources, monitoring/indicators, reward system, and organization of collaboration between various units. The focus on open data in goals, monitoring and rewards system was perceived as somewhat problematic, because in an organization where confidential commissioned research is an important part of the research profile, the focus on open data would exclude a large proportion of research data which cannot be opened.

Level 3 is described as "openness of data is a strategic choice of an organization participating in international networks, as a part of open science". Interviewees 2 and 3 both debated the progression of levels from 1 to 3, and how level 3 seems more easily attainable than level 2. Strategic choice can be construed as aligned with an organization's priorities and strategy. In that sense, an organization where confidential research data play a major role may not prioritize investments in open data necessary for level 2 and may invest more resources and efforts into elements of RDM such as secure storage, legal services for contracts and personal data protection documentation, or internal retention of data for the long term. At the same time an organization which does not prioritize investments in open data can still make strategic choices in how they will practice open science and participate in international networks.

> Although we don't fully reach the goal for level 2, (…) we however have many international projects so that is kind of a conscious choice, and strategic choice, that we want to open that data, so this (…) does not really go quite symmetrically." (Interviewee 2)

The maturity levels of RISE capability 2a) Staff investment distinguish whether RDM services are delivered by dividing responsibilities among existing staff, or whether staff roles have been significantly redesigned, including investment in staff development. This capability and its levels were generally considered adequate. Level 1 can be sufficient if all relevant roles are covered and those involved in RDM services as part of their existing role have enough time to devote to RDM issues. This would require that the RDM tasks are acknowledged as part of the existing role, and do not become an invisible workload. When roles are not defined, the researchers as well as other professionals in the network providing RDM services may not know the right contact point to get help with specific problems, and just contact the person who has helped them before. The workload of initial contacts and distribution of tasks may fall unevenly on these people.

> So if you know (*name 1*) so you send messages to (*name 1*) and if you know (*name 2*) you send messages to (*name 2*), I guess that's how such system goes and of course, for an organization such personification of services is always a risk. (Interviewee 4)

In the level descriptions of capability 2b) Technology investment, the interviewees appreciated that technical infrastructure was considered from the point of view of the whole data life cycle. Interviewee 1 suggested that even when the organization centrally invests into tools, researchers may still opt for other tools (such as free cloud services) if that is what they are used to instead of the service vetted and acquired by the organization. Technology investments could therefore also be linked to capability to raise awareness and ensure uptake. On Level 2, the institution coordinates investment in the central technical services it deems a strategic priority for research data life-cycle support. This was considered a useful capability, given that strategic priorities are defined. It was noted that in the Finnish environment, the Ministry of Education invests in national technical services for long-term storage as well as digital preservation, therefore, Finnish organizations do not need to prioritize such infrastructure. Internal long-term storage after the project, if data cannot be deposited to a repository, can be more challenging for individual organizations. From the project management side, external funding can be only used during the project, and funding for long-term storage after the project cannot be secured from grants. Interviewee 3 (IT) considered strategic planning as a necessity as the allocation of funds to acquire technology infrastructure cannot be short-sighted.

> Strategic planning must be done because the funding is such that one has to be able to and know how to invest into the right objectives. Where the resources are allocated, that always requires this strategic and tactical planning. (Interviewee 3)

In the RISE model, capability 2c) Cost modelling evaluates whether all RDM service costs are covered by overheads on grants (level 1) or whether support exceeding the norm can be charged directly from grants. This raised questions whether this was indeed an organizational capability, because how costs can be covered is defined by what costs each funder considers eligible. It is then more a capability of the funding environment which cannot be flexibly tailored to each organization. Some funders or funding instruments may consider some types of RDM related costs eligible to be charged directly, such as additional working time or storage capacity exceeding the usual quota, however, the needs are difficult to estimate in advance at the stage when proposal budgets are drafted. Centrally acquired servers and storage solutions are commonly not covered from grants, which only cover what was budgeted for the duration of the funded projects. Funding for central services also depends on national funding models for higher education institutions and state research institutes. Overall, this capability could be redefined to better distinguish what the organization can choose or affect, and what are the affordances of the funding environment.

The Finnish model approaches the question of costs and business plans through the lens of human resources. On level 1, there is at least one person who knows the basics of RDM, knows where to find RDM-related information, and can maintain a website. Their working time is allocated to RDM knowledge development and website maintenance. It was not perfectly clear whether “one person” means literally one person who is capable of performing the listed tasks, or if it could also refer to one full-time equivalent (FTE) of working time allocated to RDM. This could potentially be distributed among multiple staff members in organizations where RISE capability 2a) Staff investment is on level 1 and RDM responsibilities are divided among existing staff. Interviewee 1 expressed some doubts about why so much value was given to website maintenance as a specific form, rather than defining the capability as the type of support service provided, for example, creation and maintenance of online guides and instructions. Depending on how the organizational communication channels are designed, this does not necessarily require special website maintenance skills.

On level 2, sufficient resources are assigned to RDM services based on the size of the organization. Competency development is systematic and continuous, roles are defined with division of responsibility between units. National and international level networking is part of RDM work. These were considered useful goals for human resources in RDM, especially recognizing roles and responsibilities. While tailoring resources to the organization's needs was seen as useful, the organization's size is perhaps not the only important criterium. The organization's research profile and orientation can also affect RDM practices and service needs, guiding where the resources should be allocated.

6.2.2 Advisory services and training

Level 1 of advisory services provides generic online guidance that addresses key areas of RDM. Content may be externally sourced, with little relating to the specific institutional context. Pages include a helpdesk email address. This was considered a well-defined basic level of service, with some discussion about the helpdesk email address. The interviewees 1, 2, 3, 4 and 6 all agreed that a single point of contact with RDM services would be beneficial for researchers who wouldn't have to remember specific names for each topic, but also for the division of tasks and sharing knowledge among the various experts. This capability is dependent on capability 2a) Staff investment, since managing a helpdesk relies on clear division of roles and responsibilities and being able to allocate working time to the monitoring of shared helpdesk.

Some interviewees thought that in practice the single point of contact does not necessarily have to be a helpdesk email address. In smaller organization it may be feasible to have a contact person, for example an RDM service coordinator. However, there may be issues when this person is out of office. The IT expert (Interviewee 3) suggested a support channel on an organizational communication platform where issues and their solutions would be visible to others and remain accessible, forming a type of guidance material. This would also enable peer support among researchers and connecting researchers with service staff.

> The generic online guidance, there you could put the basic, or, it would be quite easy to document a lot of those things required in the data management plan, let's say, the principles and techniques for data storage, what's available, the technology and techniques and principles would be possible to document quite well, (…) and indeed when there is a bit trickier question and you need someone from legal for

> example, and an IT specialist, then this type of email address sounds like a really good idea. (…) And one type could of course also be (…) this kind of so-called peer support in some yammer channel or somewhere, this is something we've noticed in information classification for example, and some other things as well, that if someone has already been through some conversation and found a solution, that we would somehow get to share that with everyone, or someone could answer that. So we could have a channel where someone could ask like, hi, has anyone here dealt with this thing and what kind of solution did you get or who has helped you with that. (Interviewee 3)

On level 2, the guidance offers relevant advice on how to use services that comply with institutional policies, and the benefits to researchers of doing so. This focuses on the content of the guidance, while level 1 also defined the mode of service delivery: online guidance, helpdesk email address. Level 2 did not clearly define if it would be provided in the same way, as online guidance improved from generic to more practical with access to helpdesk, or whether the helpdesk staff would also be more capable of providing tailored advisory service relevant to specific use cases. The interviewees agreed that usually a more tailored, advanced advisory service would be needed for some topics and for others level 1 could be enough. The RISE user guide (Rans & Whyte 2017, 11) also recommends evaluating the capability on the level of specific topics, since the needed maturity level of advisory services may differ for each topic based on the organization's priorities and needs. The user guide suggested a few possible topics, but as was explained in chapter 5.2, this thesis utilized the same list of topics that were also evaluated by researchers in the user insight survey questionnaire. The results showing perceived benefit of support in each aspect of DMP/RDM was added to the interview instrument to include the users' voice in the discussion and allow the interviewees to compare their own perception to that of users.

The model does not specify additional maturity level definitions per each topic, instead, the topics can be evaluated using the general maturity levels for the Advisory service capability. To aid the evaluation, the RISE user guide formulates the following "facets to consider" when discussing advisory services:

- "*Which staff deliver support across relevant professional service units, and what scope is there to join this up?*"
- "*On which topics is the advice provision strongest and weakest?*"
- "*Which channels are used to connect researchers to any support already available, and what scope is there for using online more efficiently*?"
  (Rans & Whyte 2017, 11.)

The interviewees were asked to think about the topics relevant to their expertise. They were asked to consider whether these topics could be quite easily covered with level 1 type of advisory service, or whether there is more need for tailoring the guidance and providing individual, case-by-case advice. To draw on the facets listed above, they were also asked to think whether there is staff already providing this capability and if they should join or have already joined the RDM support service network. The interviewees were able to compare their knowledge about existing services and perceptions of where the advice provision is strongest and weakest to how the survey respondents had evaluated perceived benefit of support. The interviewees engaged very actively with this part of the RISE self-assessment. It facilitated brainstorming on how advisory services should be further developed, and helped identify topics in which the demand for individual advisory service could decrease if online guidance was improved with more practical advice. This discussion also brought out some topics in which sufficient level of capability exists, but more communication and engagement might be needed to connect researchers to the services.

The capability 4) Training generated less discussion as the interviewees had less experience with training than with advisory services. The results from user insight survey displayed in the interview instrument also showed that only 41 % of respondents would like to have training and courses available as a form of RMD/DMP support. Interviewee 2 pointed out that in research institutes, which do not have an educational role, training and courses may be less relevant than at universities which also educate students and train future researchers in doctoral programmes. Both interviewees 1 and 2 suggested that researchers may tend to prefer advisory services which are quick to access and available at the time the support need arises. On the other hand, reported preferences are based on current experience and they could change as external requirements which previously were not part of the researchers' workflows become relevant.

> Some need can come up all of a sudden and then you quickly sort it out, but not like training, it's more like advising or some… temporary support, but I wouldn't see it as training. (Interviewee 1)
>
> This made me think that maybe this is because some have not really bumped into this type of FAIR data refinement yet or haven't done it, and then when these maybe more challenging tasks come up, then it's possible that in that situation this kind of hands-on support or embedded services will be needed. (Interviewee 1)

The training capability is divided into capabilities 4a) Online training and 4b) Face-to-face training. In both capabilities the levels were considered well defined, and they provide reasonable options for organizations to tailor their training offer to demand. Level 1 for Online training means externally sourced online courses are linked to from organizational RDM-related pages, and on level 1 of Face-to-face training capability a course in basic RDM principles is available on request. If these are fulfilled, any organization is capable of responding if need for training arises, even if current demand does not suggest the need for more advanced training development.

- It is after all a problem of "just in time". And exactly, you won't have the time and energy to go to such courses if you don't need them right now. That's a big problem. And that's why these things like Opinet (*online learning platform*) are indeed better because when the need comes up, they fulfil the need. But then if you have just some couple hours of training in 6 months then that doesn't really interest you. (Interviewee 2)
- But this kind of face-to-face on request well maybe it would be a good capability to have? (Interviewer)
- Yes (…) We've had that too, (*name*) has sometimes done talks and so on. (Interviewee 2)

The Finnish model puts advisory services and training together into the Support services category. On level 1, the support service is a website with links to resources. The model lists some examples of resources which can be included, such as links to organization's

RDM policy or guidance, information on data storage services, and links to relevant courses or webinars. On level 2, the website is further developed with more comprehensive information on RDM, and it includes a “one-stop-shop” helpdesk address, run in collaboration between units. Personal support service is available on demand. The service is involved in communication or marketing via multiple channels – this was not explicitly mentioned in the RISE model and was considered a good capability. Interviewee 1 noticed that in the Finnish model, helpdesk address and individual advisory service on request were included on level 2 and not on level 1, while in the RISE model a helpdesk address was already a part of level 1. While organizing a helpdesk with defined roles and processes is not trivial, they considered some point of contact enabling access to individual advice a reasonable minimum requirement.

6.2.3 Data management planning, appraisal and risk assessment

Data management planning capability levels seemed well-defined to the interviewees. The compliance-focused level 1 means the institution provides guidance to researchers on completing funder-mandated DMPs. This was considered an essential basic-level capability. On level 2, the institution mandates DMP production at bid stage for all researchers. Guidance and templates are provided. Research Office connects to relevant stakeholders to appraise DMP content and notify them of relevant resource implications. In general, it was considered a good idea and likely useful for organizations to require DMPs for all projects. Interviewee 1 as the expert who usually helps with DMPs emphasized the importance of establishing a process for the DMP appraisal. The purpose is not just to collect additional paperwork, but to help projects plan and implement good practices. The specification of Research Office as the unit responsible for the DMP support service coordination could be removed from the description because the names of units vary in organizations, and the work can be arranged differently. The goal for sector-leading support activity (level 3) would include automated systems which flag researcher requirements to the relevant institutional support services. Interviewee 1 thought this was an interesting capability, although the development of such systems did not seem realistic in the Finnish environment at the time.

The records management expert (Interviewee 6) considered DMPs useful as part of quality assurance and making sure the data is stored with sufficient security according to the organization's information classification. They however thought that the additional workload should be considered, and it could be a useful capability to be able to identify projects which can rely more on a ready template and which projects need to plan in more detail. If organizations formulate their general institutional data management and storage policy well, they could offer a basic DMP template based on this general policy, which could be enough to adapt and follow in some projects. The legal expert (Interviewee 5) suggested that projects working with personal data in their research are one type of such case that would not benefit from ready templates. They agreed that having a DMP review service for all projects would help flag cases where the legal services should be involved.

- (…) does the template reserve space or sections for this, that if you have this type of data then you should consider this. The ones I've seen before, in those the descriptions were on such a high level, basically that we will follow the applicable legislation. Then the actual considerations and planning won't get done, it will become a kind of ceremonial proclamation, which lacks the actual substance of what they're doing and with what data and so on. (…) (Interviewee 5)
- Interviewer: (…) the one who reviews the data management plan could maybe notice that there's something, even if they are just a librarian, they will notice that maybe the legal issues here are not defined well enough and contact the lawyer for the researcher, that's an option here.
- Sounds good! (Interviewee 5)

The capability 7) Appraisal and Risk Assessment similarly includes collection of information by the organization to ensure compliance with legislation and ethical regulation, and to help the organizations keep track of their data resources. Capability 7a) concerns data collection policy. On level 1, service primarily supports data deposit to third-party repositories, and holds datasets in-house when legal/regulatory compliance requires. This was considered good basic capability for compliance. On level 2, service defines criteria for retention of datasets of long-term value to the institution. Interviewee

6, the expert in records management, pointed out that such criteria are commonly in place in the institutional information management plans for documents and textual research outputs (e.g., reports, publications) but they are not as well established for research data. It is however a good capability. Other interviewees (1, 2) also agreed that it is a good idea to define criteria for datasets retained in-house beyond only compliance, but also based on long-term value. The definition for level 3 (service defines criteria for developing datasets as special collections and ensures these meet specialist depositor and user needs) could benefit from some further explanation or examples, since it was not very clear what these special collections and specialist needs could be, and what would be the added value compared to discipline-specific data repositories.

Since the RISE model includes both technical and human infrastructure, it was unclear what is meant by “service” in this section. This was confusing especially in capability 7b) Security, legal and ethical risk assessment. For example, on level 1, service seeks confirmation that data was collected or created in accordance with legal and ethical criteria prevailing in the data producer's geographical location or discipline. It was difficult to understand what is the service that seeks confirmation, is it a technical service such as storage solution or institutional repository that requires checks for legal and ethical criteria before data deposition, or is this a capability of support staff to perform some type of checks? On level 2, service commits to proactively manage legal and ethical risks relevant to its depositors and users, and to relevant professional and technical development for researchers and support staff. The reference to depositors and users would suggest this is a technical infrastructure capability, and the reference to professional development for researchers and support staff suggests some type of support service.

Interviewees 3 (IT) and 2 (Open Science) mentioned that even if the technical infrastructure allowed some confirmation checkbox with guidance, and had the technical capability for various security levels, the data producer is still responsible for knowing their data and making the right decisions. Some human service would be needed to advise and ensure compliance. Interviewees 2 and 5 (Legal) also felt that the capability is defined in too broad terms and covers many things: data security is more of a technical infrastructure capability and legal and ethical issues are a wide range, including also contractual obligations, and the different aspects require different types of advisory service.

> It’s definitely difficult for researchers, and for management to identify what kind of limitations are set by legislation and agreements to what information can be made available, I think a lot nowadays depends on the researcher’s own expertise and awareness so that they know what kind of, first of all, commitments are linked to the project, have they made any non-disclosure agreements, what data have they created themselves and what they got from others. And then on the other hand, where confidentiality obligations come from legislation, such as export control, personal data. Of course, we try to (…) maintain awareness and then some checklists at various stages of the process, for example project kick-off review or such, these checklist type of questions can be used to verify, (…), but always there is the concern that someone maybe out of misunderstanding accidentally publishes something that should have been kept confidential. (…) So I’m not sure I understand this thing here, but this kind of centralized service that would read and know all the agreements and laws is probably not possible to develop, but more dialogue about what these requirements are on a concrete level is definitely important. (Interviewee 5)

Metadata collection to inform decision-making is the last capability in appraisal and risk assessment (7c). Decision-making could perhaps be further defined already in the title, since it isn’t very clear whose and what kind of decision making should be prioritized, researchers and external data users could also benefit from availability of metadata about datasets, but the level descriptions suggests the priority is organization’s decision making about data retention for compliance purposes (level 1). On the locally tailored level (2), metadata is routinely recorded to relate research activity to data and other outputs, and enable better informed decisions on the preservation costs, risks and value to the institution.

Such metadata collection was seen as a useful capability. Some interviewees thought it was useful to have a strategic view of data held in-house as a resource (Interviewees 1,

2) and to be able to estimate the costs of long-term storage (Interviewees 1, 4). Another angle was identifying datasets with possible risks or requirements based on legislation – e.g., personal data that should be deleted, since according to the GDPR, personal data cannot be kept indefinitely, but the retention period should be defined in the privacy notice (Interviewee 5). Organizations would benefit from systematic information gathering on data stored internally, identifying what data there are, where they are stored and for how long, who has access rights to them and how they can be accessed (Interviewees 4, 6). From the technical side it is also important to know where data are stored internally, to develop standards for what metadata needs to be captured and to make this metadata and data findable (Interviewee 3). There are better standards for collecting metadata about publications presenting the data, or published datasets, but practices on how to collect metadata about unpublished research data in a unified format are less developed in general.

6.2.4 Technical services for active data management and preservation

The three capabilities in section 6 described services for active data management, and they were mainly discussed with Interviewee 3 (IT). In general, they found the capabilities and their levels well defined. Capability 6a) Scaleability and synchronization concerns storage capacity. On level 1, the service provides researchers with managed access to networked storage, from multiple devices, of sufficient capacity and performance to satisfy most of the organisation's projects. This was considered a good basic service for research data storage. Levels 2 and 3 describe the mechanisms in which this service could be scaled to provide additional storage capacity, performance, or device networking. On level 2, this is provided on demand, while on level 3 the service would provide automated access to these additional resources. Automation was seen as a good capability, but as is typical for level 3 (sector-leading activity), it wasn't considered realistic in the near future. Flexibility in scaling down was also suggested.

> I would say that often, well often it happens so that more capacity is requested than what is then actually needed. So the understanding of how much data will be

> generated and how much space will be needed is not necessarily always in sync. (Interviewee 3)

The capability 6b) Collaboration support describes services which enable data sharing with collaborators in the active storage phase. The interviewee could easily link the level descriptions to solutions used in practice and considered them reasonable. Only in level 3, they commented on the concept of virtual research environments. While such collaboration environment would be a good capability, they wondered what the official definition was. Technically a combination of tools used for data sharing and collaboration could work together as a virtual research environment, but perhaps what is meant here is a different, more specific type of platform. Regarding capability 6c) Security management, the descriptions of access control requirements and procedures or tools for data de-identification and encryption were also evaluated positively. Level 3 mentions providing a service accredited for analysis of shared sensitive data with standard ISO 27001/2 or equivalent. The standard was known to the interviewee and their colleagues and considered relevant.

Technical capabilities for preservation (capability 8) were difficult to evaluate, because in Finland it is not common for individual universities or research institutes to develop their own data repositories for long-term storage and digital preservation. It was discussed with interviewees 1 and 2 that for long-term storage and preservation, the Ministry of Education and Culture of Finland funds national services (called Fairdata services) produced by CSC – IT Center for Science. The capabilities seemed reasonably defined from a general RDM point of view. Capability 8a) Preservation planning and action described the levels of curation and preservation activities, from ensuring bit-level integrity on level 1 to enabling digital preservation activities such as file migration. Capability 8b) Continuity support described various levels of back-up, with compliance level 1 defined as automatically creating a copy held in another location. These capabilities at least on the compliance level were considered useful also from the point of view of active storage services and data retained internally. Unfortunately, a data security specialist was not available to be interviewed and evaluate the definition of maturity levels in detail.

> If we think about something like this, maybe not a service but this kind of longer-term storage for the data internally, if it cannot be opened, then are the questions of interoperability and integrity and so on considered. (Interviewee 1)

6.2.5 Access, publishing, and discovery

The first capability in section 9) Access and Publishing is 9a) Monitoring locally produced datasets. On level 1, information is gathered from research projects to enable compliance with funders' requirements for research data discoverability. This can be technically enabled by collection of metadata into the institutional current research information system (CRIS) or by getting metadata from repositories, but the reporting for compliance is more likely to be the responsibility of the researchers rather than automated as a service. On level 2 metadata is routinely recorded on locally produced data, and its links to research activity or related outputs, enhancing the quality of the institution's research information, and on level 3 this metadata is sufficiently structured and organised to inform institutional strategy. This seems to overlap with capability 7c) Metadata collection to inform decision-making, however, this capability seems to concern open datasets.

Capability 9b) Data publishing mandate was not easily understood, since the title did not seem to match the level definitions, for example, level 1: service supports minimum external requirements for metadata and publicly accessible data. The descriptions of the higher levels also describe data access, citation, metadata exchange and discoverability. Data publishing *mandate* on the other hand was understood as requirement that data are published, which has more to do with policy than metadata quality. Capability 9c) Level of data curation seems close to 8a) Preservation planning and action, which also dealt with level of curation. The difference is not absolutely clear from the descriptions, but capability 8a) goes into more detail about technical curation for digital preservation, while the descriptions in 9c) are a little vague and seem to refer more to the quality of data and metadata, for instance level 1: service commits to brief oversight of submitted data and metadata e.g. for compliance purposes, and level 2: service commits to maintain or enhance value through routine action across data collections. This wasn't discussed much because it seemed more like a service provided by a data repository, and as mentioned earlier, this is not usually provided by individual organizations.

The last capability was 10) Discovery, specified as metadata cataloguing scope. This capability is related to 9a) Monitoring locally produced datasets, but here the metadata should be also public and make the data discoverable. On level 1, service catalogues metadata for the organisation's publicly funded datasets according to funder expectations that they are discoverable, citable, and linked to related content. On level 2, service catalogues metadata to enhance value of the institutions data assets in accordance with recognised best practice standards. It was discussed that similarly to capability 9a), this could also be technically handled with organizational CRIS systems, although it is not common to be able to collect such metadata from external data repositories routinely as a service.

> In principle there is readiness but in practice of course not yet comprehensive, so that, this kind of doesn't take a stand on if the whole thing should work seamlessly and if it gets through, but there's some kind of readiness and it's been done to an extent. (Interviewee 1)

7 DISCUSSION

This thesis explored research data management and its open science aspects from the point of view of researchers' support needs and the corresponding support services. It adopted the pragmatist views that scientific knowledge is built intersubjectively in communities of practice with their own criteria for adequacy judgements, and that science cannot be separated from the society with which it coexists in a common public sphere. The adequacy judgement criteria often include scrutiny of evidence such as research data, and transparency is important in sharing such evidence as well as in actively reflecting on the judgement criteria. In relation to responsible research and interaction with society, public funding bodies have largely adopted open science policies including open data requirements, enabling the scrutiny of research data underlying published results as well as reuse of data resulting from publicly funded research. The pragmatist approach recognizes the diversity of methods and practices in communities of scientific inquiry which affect how open science practices can be applied. (Miedema 2022.) In the development of support services for research data management and open data, these differences should be considered, and more attention should be given to possible problems faced by researchers in applying open data requirements and RDM recommendations in practice.

From the point of view of service development, it is important to be informed about service users' needs, behaviours and experiences (Bury & Jamieson 2013, de Jong 2014, Marquez & Downey 2015). A user insight survey was conducted at VTT Technical Research Centre of Finland (VTT) to gain insights on the research question: *What kind of needs, attitudes and experiences do researchers have in RMD topics where support services could be useful?* The topics were framed as motivations for six potential service moments (episodes of service use during which the user interacts with the services' communication points) (Koivisto 2007, 66-67). The aim was not to test a hypothesis or gain statistically reliable knowledge about the population of researchers at VTT, but to explore whether the answers to the research question would provide interesting insights for service development.

7.1 User insight survey implications for support service development

The potential service moment in which the researcher needs to write a DMP and comply with funder's requirements was approached through questions about experienced level of difficulty and attitudes towards potential benefits of writing a good DMP. These may affect how likely the respondents are to seek out or accept the offer of support. Davidson (2013) and Davis and Cross (2015) convey that DMP review is a support need that arises early in the research data life cycle, and it is useful for connecting the researcher to various experts or services based on their needs. The findings of Van den Eynden (2018) show that uncertainty about how data should be managed and having to use too much time and effort negatively affect motivation to manage and open research data. The attitudes expressed towards the statements mitigating this uncertainty and amount of time and effort (DMP supports seamless collaboration during the project, or planning ahead makes research data management and opening the data easier) can help understand opinions in the community, which can be utilized in marketing of services and best practice. For example, if most do not agree with such statements, successful use cases can be included in guidance, training, or communication.

The statement that writing a good DMP could increase the chance to win the grant and DMP gives useful insights how the data collection and storage should be described in the rest of the funding application are less relevant than at the time the survey was originally conducted in June 2020, since one of the major funders, Research Council of Finland, used to require a DMP in the proposal phase, but this has since been changed to a much shorter DMP included as a section of the proposal research plan with the full DMP written after positive funding decision (Research Council of Finland 2023b, compared to older instructions in Research Council of Finland 2023c). This can mitigate the fear of lost effort on proposals which don't get funded expressed by some respondents in the open-ended comments. It was useful to offer an open-ended question after these statements where respondents could comment on any other perceived pros and cons of writing a DMP. The statement by one of the respondents that DMPs can help recognize possible risk points connected to the data collected or processed in the project, such as risks concerning data protection, privacy or business sensitivity is an excellent argument for data management planning. It could be added in future surveys exploring opinions and attitudes in organizational communities, or directly applied in the marketing of services and best practices.

It was also useful to give space to freely express the negative experiences and perceptions, because these are more difficult to anticipate and asked about directly in the survey, but addressing weaknesses, gaps or challenges is important. Chapter 3.2 discussed that ideally DMPs would serve the researcher as a practical RDM planning tool. Some respondents perceived data management planning as detached from real data management practices, not helpful in identifying appropriate solutions and applying good practices throughout the research life cycle. Because the funders do not provide feedback, the DMPs were seen by some as pointless bureaucracy, since there was no opportunity to learn or receive confirmation that the plan was good. As formulated in chapter 2.2 based on Miedema (2022), open science practices should support the underlying values of robustness and transparency of research and wider use of scientific knowledge by others in academia and society, and not become top-down requirements. Research funding bodies could perhaps do more to increase the relevance of DMPs. For the organizational support service, it may be useful to tailor or market existing services to highlight the availability of guidance or advisory service beyond DMP.

The questions investigating data storage during the project and internal long-term storage can offer more practical information applicable to the development of technical services. During the planning of the survey, IT experts were consulted to provide a list of storage solutions offered by the organization for active data storage during research. The decision was made to include less secure solutions such as USB sticks on the same list without any expression of value judgement. There is a risk that inclusion of these storage solutions may be perceived as endorsement for some respondents (it is on the list, therefore it is ok), however, it may be useful to gain an honest view of real practices within the organization. In case of low uptake of secure storage and high uptake of less secure storage, more can be done in the organization to promote the recommended solutions. When the respondents choose from the list of possible solutions, it can already raise awareness about organizational storage services previously unknown to respondents.

In case of internal long-term storage, finding out the reasons why data is usually retained internally and not submitted to a data repository can help estimate how common internal long-term storage would be, indicating the associated storage capacity and its costs, and requirements for security level of solutions recommended to archive internal data after the active phase (protection of personal data, confidential data). This insight alone would not be enough to plan investments in potential additional technical infrastructure, but it can give some initial direction. The question of internal long-term storage of data which cannot be shared publicly but could be shared within the organization (or otherwise under specific conditions) and should remain findable and accessible, sparked interesting comments from the respondents about the readiness of internal storage for data findability and accessibility. Since organizational servers, network drives and similar solutions do not commonly share the capabilities of data repositories for capturing metadata which helps make data FAIR (Findable, Accessible, Interoperable, Reusable), this is probably a common gap in services for RDM.

Supporting the findings of Van den Eynden (2018) that the opportunity to increase the quality, responsibility, and impact of their research motivates researchers to manage and share their data, the majority of respondents agreed that opening the data could increase the visibility and impact of their research, however, a much smaller proportion had in fact made their data available with open or managed access. This was likely affected by the reasons specified in the question regarding limitations to opening data, most commonly intellectual property rights issues. Those who were able to open their data reported an aspiration to raise the visibility and reusability of research results was more common motivation to do so than funder’s requirements. However, coming back to the question about reasons why research data was not publicly shared, some respondents also reported lack of resources for the work needed and not enough knowledge about how or where to open the data, common reasons why data are not opened as also observed by Rice and Southall (2016) and Van den Eynden and Bishop (2014).

Combining the results of these questions from two service moments (the researcher needs to make data openly available via a repository, versus the researcher needs internal long-term storage because they could not deposit to a repository) provides the support services with information on bottlenecks of opening research data. In this case the respondents agreed that open data is linked with the values of transparency and impact and these values resonated with them, so open data did not seem like a top-down requirement and maybe advocating for open science and open data would not be as crucial, but more focus could be targeted at improving the required knowledge and skills by additional guidance, advisory services or training. The issue of resources is not as easily mitigated, since allocation of resources for RDM activities in general is a part of a larger structural issue of hypercompetition for merit and funding and what is valued as indicators of excellence in academia, discussed by Miedema (2022) and also mentioned by Piwowar, Day and Prisma (2007) when assessing the burden versus benefits of RDM work for researchers' career.

When respondents evaluated their level of confidence and perceived benefit of support in various areas of RDM as described in common DMP templates, quite often the benefit of support was estimated higher than the lack of confidence in the corresponding topic. This could be because *support* was not specifically defined as advisory service on a topic where researchers do not feel confident. Support could also be construed as hands-on help from experts or peers to save time on time-consuming tasks, perhaps useful in some topics such as metadata and documentation. Maybe respondents felt that while they do not need support themselves, support would be useful for the community. Unfortunately, the survey did not allow sufficient opportunity for respondents to follow up on the reasons why they may appreciate support in topics where they feel confident, or to elaborate what type of support should be available in each topic. Despite this lack of detail, these results offer general direction for support service development.

Preferences for support service delivery were explored on a general level. While respondents largely preferred resources that can be used independently such as online guides and instructions and did not show as much interest in training, courses and workshops, such result does not necessarily mean these service touchpoints are not worth organizing at all. The community of potential customers includes people with different needs and learning styles, who may benefit the most from different forms of support. The preferences are however good for prioritizing efforts in the various support types. Interestingly, respondents valued access to an organizational data support service with a helpdesk or similar single contact point for various specialists equally to relying on their network of contacts, such as other researchers and specialists they have interacted with. This would suggest the need for more informal support among peers and colleagues closer to their usual workflows, perhaps providing discipline-specific information and researcher-driven discussion about good practices. Such community preferences could help an organization decide between the models described by Matusiak and Sposito (2017) to provide a more centralized distributed network or research data service centre, or whether it is worth to develop some type of embedded service close to research units and familiar with the discipline.

7.2 RISE self-assessment tool usefulness evaluation and synergy with survey

The RISE self-assessment tool was applied as an instrument for semi-structured interviews with experts providing various types of RDM-related support for researchers, in order to investigate the second research question, *how did RDM support experts evaluate the usefulness of the RISE framework for service self-assessment?* A similar, albeit much more concise model of recommended services for open research data in Finland was included in the evaluation to see if it would contribute a local perspective on the Finnish environment to the RISE model, originally developed in the UK. The thesis also aimed to collect potential observation regarding the overarching questions: *Did the survey work together with the RISE model to include users? Based on the pilot, how useful were the survey and RISE as tools for service development?*

Overall, an important advantage of the RISE model is that its focus is not limited to open data. It covers a wide range of both technical and advisory services for RDM throughout the lifecycle, taking into consideration the division of roles and responsibilities and sustainable plans for funding of the services. While the value of open research data has been established in terms of transparency and verification, enabling further research, avoiding duplication of efforts, and larger benefit of public funding (see chapters 2.4 and

2.5), it is also clear that there are limitations to open data sharing such as personal data protection (EU 2016/697, FINLEX 1050/2018) or commercialization and protection of intellectual property rights (Corti et al. 2014, Rice & Southall 2016, European Commission 2012). As Higman, Bangert and Jones (2019) argued, favouring one of the interrelated concepts of RDM, FAIR data and open data would overlook important elements of the others.

This view was supported by interviewees who evaluated the Management section of the Finnish model. In this section, level 2 is described "the management invests in open data", including goals, resources, monitoring/indicators, reward system. The interviewees questioned the focus on only open data, referring to the large proportion of confidential commissioned research where data cannot be open, but should still be managed well. Chapter 2.1 discussed the tendency of open science approach towards diversifying incentives and rewards system to include and appreciate various types of research outputs and contributions (Miedema 2022). Although chapter 2.5 showed that receiving merit for RDM work is an important incentive (for example, Piwowar et al. 2007), chapter 3.4 argued that indicators and rewards should be well designed so that RDM efforts are incentivized and not limited to simple metrics such as number of open datasets. Therefore, we must be careful in how goals, indicators and reward systems are constructed to truly appreciate the diverse contributions. This is not a criticism of the Finnish model, which naturally focused on open data since it was developed in a working group on openness of research data in the National Coordination for Open Science and Research in Finland. The model was specifically framed as support services for research data *openness* management. (Assinen 2020.) However, for the purpose of service development, the interviewees saw value in the more holistic RDM approach of the RISE framework.

The capability 7c) concerning collection of metadata kept internally, informing the organizations about the associated risk, value of the data and preservation costs, was considered useful and was commented on in some way by all interviewees. It was also viewed as important by some respondents in the survey, when asked about internal long-term storage supporting findability and accessibility. This aspect of RDM could also be overlooked when emphasizing openness. The RISE user guide (Rans & Whyte 2017, 3) states that another advantage of the framework is that it brings together various stakeholders to discuss RDM services from their perspectives and reach a shared vision. While this advantage might be more prominent when the self-assessment exercise is organized as group workshop, the option of semi-structured interviews also enabled comparing and contrasting the various perspectives on the capabilities. In support of the findings of Cox, Pinfield and Smith (2014) that RDM could be categorizes as a "wicked problem", where the solutions and the problem itself can be understood differently by different stakeholders, some capabilities which sparked the most discussion also showed the various angles from which different experts approached these common topics. These were for example the capability 7c) on collection of metadata about internally stored datasets, but also 1b) Awareness raising and stakeholder engagement, 5) Data management planning, and 3) Advisory services. Including results from the user insight survey about topics where support would be beneficial and about preferred ways to interact with support services helped include the user point of view and helped the interviewed experts reflect on their own perceptions.

There were also some weaknesses of the framework identified in the interviews. The practical implications of some capabilities and their maturity levels were not easy to understand or relate to own experience in cases where the terminology and language were considered unclear. These included for example the term "roadmap" in capability 1c), specifying "research office" as a point of contact in capability 5, vague use of "service" in capabilities 7a)-7c), "decision making" in 7c), and "mandate" in 9b). These terms could be rephrased or clarified in the accompanying user guide. In addition to these examples noted in the interviews, it can also be argued that the capability 8b) Continuity Support in practice focuses on back-up solutions and the preservation activities such as file migration in capability 8a) also support continuity. Back-up to avoid loss of data is also important in live storage solutions, as was conveyed by Corti and colleagues (2014) and Stouthamer-Loeber and Bok van Kammen (1995) (see chapter 3.3). However, in section 6) Active data management the capability 6c) Security management focused only on access control, de-identification and encryption. As a side note, in times post COVID-19 when remote meetings have become common, it could be practical to consider the organization of capability 4) Training in terms of synchronous (webinars, face to face training) and asynchronous (online courses).

The interviewees also suggested modifying the capability 2c) Cost modelling to reflect common funding models in the Finnish environment, and clarifying the capability 7b) Security, legal and ethical risk assessment which seemed to involve many issues with their own processes: data security requirements for storage, ethical concerns in research with human participants and personal data protection, and various legal compliance issues such as export control and confidentiality agreements. Another observation from the interviews showed that in the Finnish environment where national data repository and digital preservation solution exist, capabilities for preservation (8) could be specified in terms of repositories and internal long-term storage.

Metadata collection was addressed in three different capabilities from slightly different angles. In 7c) Metadata collection to inform decision making, the goal was a strategic view of datasets retained in-house and their relation to other research outputs and activity. Capability 9a) Monitoring locally produced datasets also focused on metadata collection and is defined very similarly, but since it is listed in section 9 about Access and Publishing, it seems to concern published datasets. Capability 10 again mentions metadata collection, but this time the metadata is catalogued to enable discoverability. It could be useful to better distinguish the differences and describe the synergies between these three capabilities, which are related and can build on each other.

The user insight survey suggested some areas in which there may be a potential gap in awareness and uptake, even though services are provided. These were for example technical services for storage, guidance and advisory support. A similar view was expressed in the interviews in connection to capability 2b) Technology investment and ensuring uptake of acquired technology and in discussion of the various topics of Advisory services (3). The capability 1b) Awareness raising and stakeholder engagement addressed the issue, but only in terms of policy. Awareness raising and stakeholder engagement can also be important to improve visibility of services and their relevance in practice. Communication and service marketing could be given more attention in the

RISE framework within capabilities or as a separate capability. In the Finnish model, communication (marketing) via multiple channels is mentioned as part of Support services on level 2.

Both the survey and the interviews highlighted the importance of guidance how to implement policies and plans in practice. This may not always be easy, especially given the discipline and data type specific variation in practices (see also Corti et al. 2014, Rice & Southall 2016, Chen et al. 2019). A DMP review service suggested in the RISE model, which would appraise the DMP content and connect researchers to services based on their needs, could help connect the plans and policies with practice. Locally tailored support which reflects the institutional service offer and recommended practices also does not have to be delivered only by advisory service, which is quite demanding on personnel resources. Well written instructions, practical examples, and documentation of technical services defining their suitability for various use cases can also respond to many support needs on a larger scale. Combining user insight from the survey with the assessment of a full range of RDM services can be considered useful in recognizing users' priorities, but also not overlooking requirements set by legislation and policies and not always visible in the users' everyday workflows.

7.3 Limitations and suggestions for further research

While the survey provided some interesting insights for service development, it only reached a small sample of the whole population of those who could benefit from RDM support. This survey was also designed as exploratory and was not concerned with statistical significance or reliability. In further research, it would be advisable to redesign the survey questionnaire with more expertise on service design, research methods and inferential statistics. This similar but improved survey could be conducted in other organizations and developed further into a reusable tool for investigating user needs in RDM.

In some respects, the survey results remained superficial. For example, the questions exploring uncertainty and support need in various topics in RDM were not linked with the service touchpoints, i.e., what kind of support service would be needed to respond to the needs. The question regarding experience with reusing data created by someone else helps understand users' behaviours and experiences but offers little information value for service development without elaboration. Further valuable insights into discipline-specific practices could have been gained from asking the respondents to specify this experience, such as what was the reuse context and what concrete practices improved or hindered the findability, accessibility, interoperability, and reusability of the data. However, in an already lengthy survey this would take several additional open-ended questions which would require a lot of time and effort to answer. An account on past experiences would be better attainable in subsequent use case interviews, where the respondents could be given prompts and follow-up questions to elicit reflection.

Following the user guide (Rans & Whyte 2017, 7), the interview instrument was available to interviewed stakeholders to familiarise themselves with the framework before the discussion. It is recommended that especially the person facilitating the process should study the materials in advance, because they will need to guide the discussion and potentially deal with situations in which the stakeholders interpret the capabilities differently. During the interviews it was indeed sometimes necessary to interpret the specialist language of the descriptions to or with the interviewees, especially if a term was unclear to all. Turning the statements into questions and discussion points and helping interviewees understand the capabilities carries a risk of contaminating the discussion with the interviewer's own views and understanding of the capabilities and RDM problems in general. Especially if time for the interviews is limited, the facilitator might need to ask more questions for smoother flow of the discussion which poses a risk of leading questions and allows less opportunity for the interviewee to evaluate the capability descriptions independently and spontaneously. However, critical thinking and straying from the posed questions were encouraged and all interviewees seemed eager to express their opinions.

The self-assessment interviews were conducted as part of this thesis, and not as a service development exercise initiated by the organization. The participants volunteered their time out of interest in the topic of RDM. The interviews took between 1-1,5 hours each and still perhaps more time could have been used to allow interviewees more time to think and maybe go through some additional sections of the framework, not necessarily as close to their expertise, but still possibly relevant. If conducted as an actual assessment of organizational services, it would be useful to reserve more time, and ensure all relevant stakeholders are included. In the Finnish environment, perhaps the framework could be slightly revised if similar findings regarding the understandability and suitability of the capabilities and their maturity levels are supported by an additional pilot.

This thesis adopted an approach to service evaluation which aims to benchmark services for RDM within an organization, rather than benchmark and rank various organizations (Rans & Whyte 2017). The goal is not to compete for the position of sector-leading organization, but to provide sufficient services responding to the needs of the organization and its community. In the RISE framework, using externally developed guidance or training resources can be sufficient on the basic compliance level, and here organizations can learn from each other and collaborate with those for whom sector-leading activity in some aspect is a priority. The framework utilised in this thesis could be further developed into a standard methodology for organizational RDM service self-assessment, exploring the users' needs and evaluating existing services based on a set of essential service capabilities. This would require further research dedicated to the development of the survey tool and an instrument for optional follow-up use case interviews. The RISE model could be modified as suggested above to better aid the self-assessment exercise with various stakeholders in the Finnish research environment. It would also be beneficial to further research the issue of incentives and rewards for good RDM practices and opening research data, where the field of bibliometrics or scientometrics could contribute to the robustness and quality of indicators, to avoid creating another "broken system" (Miedema 2022, 110).

REFERENCES

Aineistonhallinnan käsikirja (2023). Tampere: Yhteiskuntatieteellinen tietoarkisto. https://www.fsd.tuni.fi/aineistonhallinta urn:nbn:fi:fsd:V-201504200001 (Accessed 13.11.2023)

Ala-Kyyny, J., Korhonen, T., & Roinila, M. (2018). Tutkimusdatan avaamisen esteet: haastattelututkimus Helsingin yliopistossa. *Signum* 49(4), 25–29. DOI: 10.25033/sig.69198

Assinen, P. (8.5.2020). Tutkimusdatan avoimuuden hallinnan tukipalvelut. https://wiki.eduuni.fi/display/csctuha/Tutkimusaineistojen+avoimuuden+hallinnan+tukipalvelut (Accessed 11.11.2023).

Association of European Research Libraries LIBER (2017). Implementing FAIR Data Principles: The Role of Libraries. https://libereurope.eu/wp-content/uploads/2020/09/LIBER-FAIR-Data.pdf (Accessed 29.10.2023)

Auckland, M. (2012). Re-skilling for Research. An investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers. Research Libraries UK (RLUK) Report. https://www.rluk.ac.uk/portfolio-items/re-skilling-for-research/ (Accessed 29.10.2023)

Avoimen tieteen koordinaatio (2020). Declaration for Open Science and Research 2020–2025. *Vastuullisen tieteen julkaisusarja*. DOI: 10.23847/isbn.9789525995251

Avointiede.fi (2020). Koordinaatio. https://avointiede.fi/fi/koordinaatio (Accessed 29.4.2020).

Barbour, R. S. (2018). *Doing focus groups*. London: SAGE Publications Ltd.

Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P. & Ceri, S. (2020). A review on viral data sources and search systems for perspective mitigation of COVID-19. *Briefings in Bioinformatics*, 22(2), 664-675. DOI: 10.1093/bib/bbaa359

Briney, K. (2015). *Data Management for Researchers: Organize, Maintain and Share your Data for Research Success*. Exeter: Pelagic Publishing, UK.

Brinkmann, S. (2014). Unstructured and Semi-Structured Interviewing. In: Leavy, P. (ed.), *The Oxford Handbook of Qualitative Research* (pp. 277-299). New York: Oxford University Press.

Brown, P.O., Cabell, D., Chakravarti, A., Cohen, B., Delamothe, T., Eisen, M., Grivell, L., Guedon, J.-C., Hawley, R.S., Johnson, R.K., Kirschner, M.W., Lipman, D., Lutzker, A.P., Marincola, E., Roberts, R.J., Rubin, G.M., Schloegl, R.,Siegel, V., So, A.D., Suber, P., Varmus, H.E., Velterop, J., Walport, M.J. & Watson, L. (2003). Bethesda Statement on Open Access Publishing. *Digital Access to Scholarship at Harvard (DASH) repository*. URN: http://nrs.harvard.edu/urn-3:HUL.InstRepos:4725199

Budapest Open Access Initiative (BOAI) (2023). Read the declaration: Budapest Open Access Initiative Declaration (2002). https://www.budapestopenaccessinitiative.org/read (Accessed 5.4.2023).

Bury, R. & Jamieson, H. (2013). A service in transition: how digital technology is shaping organizational change. In: Mackenzie, A. & Martin, L. (eds.), *Mastering digital librarianship: Strategy, networking and discovery in academic libraries* (pp. 41-61). London: Facet Publishing.

Carlson, J. (2010). The Data Curation Profiles Toolkit: User Guide. *Data Curation Profiles Toolkit*. Paper 1. Purdue University. DOI: 10.5703/1288284315650

Chen, X., Dallmeier-Tiessen, S., Dasler, R., Feger, S., Fokianos, P., Benito Gonzalez, J., Hirvonsalo, H., Kousidis, D., Lavasa, A., Mele, S., Rodriguez Rodriguez, D., Simko, T., Smith, T., Trisovic, A., Trzcinska, A., Tsanaktsidis, I., Zimmermann, M., Cranmer, K., Heinrich, L., Watts, G., Hildreth, M., Lloret Iglesias, L., Lassila-Perini, K. & Neubert, S. (2019) Open is not enough. *Nature Physics* 15(2)**,** 113–119. DOI: 10.1038/s41567-018-0342-2

Christensen-Dalsgaard, B., van den Berg, M., Grim, R., Horstmann, W., Jansen, D., Pollard, T. & Roos, A. (2012). Then recommendations for libraries to get started with research data management. Final report of the LIBER working group on E-Science/Research Data Management. https://libereurope.eu/wp-content/uploads/2020/11/The-research-data-group-2012-v7-final.pdf (Accessed 29.10.2023).

Christopher, A. N. (2017). *Interpreting and using statistics in psychological research*. Los Angeles, CA: SAGE Publications, Inc.

Corrall, S. (2012). Roles and responsibilities: libraries, librarians and data. In: Pryor, G. (ed.), *Managing Research Data* (pp. 105-134). London: Facet Publishing. DOI:10.29085/9781856048910.007

Corti, L., Van den Eynden, V., Bishop, L. & Woollard, M. (2014). *Managing and Sharing Research Data: A Guide to Good Practice.* Thousand Oaks, CA: Sage.

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M. & Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data* 5(1), 180259. DOI: 10.1038/sdata.2018.259

Cox, A. M., Pinfield, S. & Smith, J. (2014). Moving a brick building: UK libraries coping with research data management as a 'wicked' problem. *Journal of Librarianship and Information Science* 48(1), 3-17. DOI: 10.1177/0961000614533717

Cox, A. M., Kennan, M. A., Lyon, L. & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology* 68(9), 2182-2200. DOI: 10.1002/asi.23781

Crowston, K. & Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology* 48(1), 1-9. DOI: 10.1002/meet.2011.14504801036

CSC – IT Center for Science (2023). The CSC Kajaani data center. https://www.csc.fi/csc-datacenter-in-kajaani (Accessed 28.10.2023).

Davidson, J. (2013). Supporting early-career researchers in data management and curation. In: Mackenzie, A. & Martin, L. (eds.), *Mastering digital librarianship: Strategy, networking and discovery in academic libraries* (pp. 83-102). London: Facet Publishing.

Davis, H. M. & Cross, W. M., (2015). Using a Data Management Plan Review Service as a Training Ground for Librarians. *Journal of Librarianship and Scholarly Communication* 3(2), eP1243. DOI: 10.7710/2162-3309.1243

de Jong, M. E. (2014). Service Design for Libraries: An Introduction. In: Woodsworth, A. & Penniman, W. D. (eds.), *Advances in Librarianship: Vol. 38, Management and Leadership Innovations*. (pp. 137-151). Bingley: Emerald. DOI: 10.1108/S0065-283020140000038003

Digital Curation Centre (2009). Data Asset Framework Implementation Guide. https://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf (Accessed 29.4.2020)

Digital Curation Centre (2020). CARDIO. http://www.dcc.ac.uk/resources/tools/cardio (Accessed 29.4.2020).

Digital Curation Centre (2023a). Data Asset Framework. https://www.dcc.ac.uk/tools/data-asset-framework (Accessed 4.11.1013)

Digital Curation Centre (2023b). Disciplinary Metadata. https://www.dcc.ac.uk/guidance/standards/metadata (Accessed 28.10.2023)

Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). EU-Lex. http://data.europa.eu/eli/dir/2019/1024/oj

DMPTuuli (2023). Public DMP templates. https://www.dmptuuli.fi/public_templates (Accessed 26.3.2023)

Dunning, A., Verbakel, E., de Smaele, M. & Böhmer, J.K. (2017). Research Data Services + 4TU.Centre for Research Data: RISE 2017 REPORT. https://docs.google.com/document/d/1uy7pOnRbSLHJD5C2FkFn1R39K79NUQJOOA9ZBuJ84ok/pub (Accessed 11.11.2020)

ELIXIR (2021). *Research Data Management Kit.* A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075). URL: https://rdmkit.elixir-europe.org (Accessed 13.11.2023)

European Council for Nuclear Research (CERN) (2023). CERN Data Centre. https://home.cern/science/computing/data-centre (Accessed 21.10. 2023).

European Commission (2012). *Online Survey on Scientific Information in the Digital Age*. Luxembourg: Publications office of the European Union.

European Commission (2015). *Open innovation, open science, open to the world*. Luxembourg: Publications office of the European Union.

Fairdata.fi (2020). FAIR-periaatteet. https://www.fairdata.fi/miksi-fairdata/fair-periaatteet/ (Accessed 23.4.2020)

Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M. & Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data* 6(1), 28. DOI: 10.1038/s41597-019-0031-8

Finnish National Board on Research Integrity TENK (2023). *The Finnish code of conduct for research integrity and procedures for handling alleged violations of research integrity in Finland.* Publications of the Finnish National Board on Research Integrity TENK 4/2023. Helsinki: Finnish National Board on Research Integrity TENK. https://tenk.fi/sites/default/files/2023-05/RI_Guidelines_2023.pdf

Forsström, P.-L. & Kutilainen, T. (2019). Kohti yllättäviä löytöjä ja luovia oivalluksia. Avoin tiede ja tutkimus -hankkeen loppuraportti. https://avointiede.fi/sites/default/files/2019-12/ATT-hankkeen%20loppuraportti%20v5_0.pdf (Accessed 29.4.2020)

Given, L. M. (2012). Content Analysis. In: *SAGE Encyclopedia of Qualitative Research Methods* (pp. 121-122). Thousand Oaks, CA: SAGE Publications, Inc.

GO FAIR Initiative (2020). FAIR Principles. https://www.go-fair.org/fair-principles/ (Accessed 30.10.2020).

Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights: The UKSG Journal*, *32*(1), 18.DOI: https://doi.org/10.1629/uksg.468

Hofelich More, A., Johnston, L. R. & Lindsay, T. (2016). The Data Management Village: Collaboration among Research Support Providers in the Large Academic Environment. In: K. Thompson & L. Kellam (eds.), *Databrarianship: The Academic Data Librarian in Theory and Practice*. Chicago: Association of College and Research Libraries (ACRL). Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/181127

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine* 2(8), e124. DOI: 10.1371/journal.pmed.0020124

Jones, S., Ross, S. & Ruusalepp, R. (2009). *Data Audit Framework Methodology, draft for discussion, version 1.8.* Glasgow: HATII. https://www.data-audit.eu/DAF_Methodology.pdf (Accessed 18.11.2020)

Koivisto, M. (2007). *Mitä on palvelumuotoilu? Muotoilun hyödyntäminen palvelujen suunnittelussa*. Taiteen maisterin lopputyö. Helsinki: Taideteollinen korkeakoulu.

Kruse, F. & Thestrup, J. B. (2018). *Research data management: A European perspective*. Berlin: De Gruyter Saur.

Kuusniemi, M.E. (2019). Kuka omistaa tutkimusdatan? *Vastuullinen tiede*. https://vastuullinentiede.fi/fi/tutkimuksen-suunnittelu/kuka-omistaa-tutkimusdatan

Laki julkisin varoin tuotettujen tutkimusaineistojen uudelleenkäytöstä 15.7.2021/713. Finlex. https://finlex.fi/fi/laki/ajantasa/2021/20210713

Laki sosiaali- ja terveystietojen toissijaisesta käytöstä 26.4.2019/552. Finlex. https://www.finlex.fi/fi/laki/ajantasa/2019/20190552

Laki viranomaisten toiminnan julkisuudesta 21.5.1999/621. Finlex. https://www.finlex.fi/fi/laki/ajantasa/1999/19990621

Marquez, J. & Downey, A. (2015). Service Design: An Introduction to a Holistic Assessment Methodology of Library Services. *Weave Journal of Library User Experience* 1(2). DOI: 10.3998/weave.12535642.0001.201

Matusiak, K. & Sposito, F. A. (2017). Types of research data management services: An international perspective. *Proceedings of the Association for Information Science and Technology* 54(1), 754-756. DOI: 10.1002/pra2.2017.14505401144

Max-Planck-Gesellschaft (2023). Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003). https://openaccess.mpg.de/Berlin-Declaration (Accessed 13.11.2023).

Miedema, F. (2022). *Open Science: the Very Idea*. Dordrecht: Springer.

National Coordination of Open Science and Research (2023). Open research data and methods. National policy and executive plan by the higher education and research community for 2021–2025: Policy component 1 (Open access to research data) and 2 (Open access to research methods and infrastructures). *Responsible research series* 4:2023. DOI: https://doi.org/10.23847/tsv.669

Organisation for Economic Co-operation and Development (OECD) (2004). Declaration on Access to Research Data from Public Funding. OECD/LEGAL/0321

Organisation for Economic Co-operation and Development (OECD) (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. DOI:10.1787/9789264034020-en-fr

Paulk, M.C., Curtis, B., Chrissis, M.B. & Weber, C.V. (1993). Capability maturity model, version 1.1. *IEEE Software* 10(4), 18-27. DOI: 10.1109/52.219617

Pinfield, S., Cox, A.M. & Smith, J. (2014). Research Data Management and Libraries: Relationships, Activities, Drivers and Influences. *PLOS ONE* 9(12), e114734. DOI: 10.1371/journal.pone.0114734

Piwowar, H.A., Day, R.S., Fridsma, D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE* 2(3): e308. DOI: 10.1371/journal.pone.0000308

Rans, J. & Whyte, A. (2017). Using RISE: the Research Infrastructure Self-Evaluation Framework. Version 1.1.
http://www.dcc.ac.uk/sites/default/files/documents/publications/UsingRISE_v1_1.pdf
(Accessed 6.11.2023)

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). EU-Lex. http://data.europa.eu/eli/reg/2016/679/2016-05-04

Research Council of Finland (2023a). Research Council policies on open science. https://www.aka.fi/en/research-funding/responsible-science/open-science/academy-policies-on-open-science/ (Accessed 15.10.2023)

Research Council of Finland (2023b). Data management plan. https://www.aka.fi/en/research-funding/apply-for-funding/how-to-apply-for-funding/az-index-of-application-guidelines2/data-management-plan/data-management-plan/
(Accessed 8.11.2023)

Research Council of Finland (2023c). Funding for research post as Academy Research Fellow, all research fields. https://www.aka.fi/en/research-funding/apply-for-funding/calls-for-applications/apply-now2/funding-for-research-post-as-academy-research-fellow-all-research-fields/ (Accessed 8.11.2023)

Rice, R. & Southall, J. (2016). *The Data Librarian's Handbook*. London: Facet Publishing.

Rolando, L., Doty, C., Hagenmaier, W., Valk, A. & Parkham, S. W. (2013). Institutional Readiness for Data Stewardship: Findings and Recommendations from the Research Data Assessment. Atlanta, GA: Georgia Institute of Technology. http://hdl.handle.net/1853/48198

Saarti, J., Hormia-Poutanen, K., Kuusinen, I. & Vattulaien, P. (2009). *Opetuksen ja tutkimuksen toimintaympäristö 2020: Korkeakoulukirjastojen rakenteellinen kehittäminen digitaaliseksi palveluverkoksi*. Opetusministeriön työryhmämuistioita ja selvityksiä 2009:26. Helsinki: Opetusministeriö. http://urn.fi/URN:ISBN:978-952-485-773-4

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M. & the FAIRsharing Community (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology* 37(4), 358–367. DOI: 10.1038/s41587-019-0080-8

Science Europe (2021). Practical Guide to the International Alignment of Research Data Management - Extended Edition. DOI: 10.5281/zenodo.4915862

Spichtinger, D. & Siren, J. (2018) The Development of Research Data Management Policies in Horizon 2020. In: Kruse, F. & Thestrup, J. B. (eds.), *Research data management: A European perspective* (pp. 11-23). Berlin: De Gruyter Saur.

Stake, R. E. (2009). The case study method in social inquiry. In: Gomm, R., Hammersley, M. & Foster, P. (eds.), *Case Study Method.* (pp. 19-26). London: SAGE Publications Ltd. DOI: 10.4135/9780857024367

Stouthamer-Loeber, M. & Van Kammen, W. B. (1995). *Data collection and management: A practical guide*. Thousand Oaks: Sage.

Suber, P. (2012). *Open Access*. Cambridge, Massachusetts: MIT Press.

Tenopir, C., Pollock, D., Allard, S. & Hughes, D. (2016). Research data services in european and north american libraries: Current offerings and plans for the future. *Proceedings of the Association for Information Science and Technology* 53(1), 1-6. DOI: 10.1002/pra2.2016.14505301129

Tietosuojalaki 5.12.2018/1050. Finlex. https://www.finlex.fi/fi/laki/ajantasa/2018/20181050

Toepoel, V. (2016). *Doing Surveys Online*. Los Angeles: SAGE.

Tuomi, J. & Sarajärvi, A. (2018). *Laadullinen tutkimus ja sisällönanalyysi*. Helsinki: Kustannusosakeyhtiö Tammi.

Tuuli-project (2021). General Finnish DMP guidance (Version 2021). Zenodo. DOI: 10.5281/zenodo.5242629

U.S. Geological Survey (2014). The United States Geological Survey Science Data Lifecycle Model. USGS Open-File Report 2013-1265. DOI: /10.3133/ofr20131265

Van den Eynden, V. (2018). What Motivates Researchers to Manage and Share Research Data. In: Kruse, F. & Thestrup, J. B. (eds.), *Research data management: A European perspective* (pp. 43-52). Berlin: De Gruyter Saur.

Van den Eynden, V. & Bishop, L. (2014). Incentives and motivations for sharing research data, a researcher's perspective. A Knowledge Exchange Report. http://repository.jisc.ac.uk/id/eprint/5662

Vehkalahti, K. (2019). *Kyselytutkimuksen mittarit ja menetelmät.* Helsinki: Finn Lectura. DOI: 10.31885/9789515149817

VTT (2023). What is VTT? https://www.vttresearch.com/en/what-vtt (Accessed 16.10.2023)

White House Office of Science and Technology Policy (OSTP) (2013). Memorandum About Increasing Access to the Results of Federally Funded Scientific Research https://www.epa.gov/sites/production/files/2015-01/documents/ostp_memo_increasing_public_access.pdf (Accessed 13.11.2023)

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3**,** 160018. DOI: 10.1038/sdata.2016.18

Wilson, F. (2013). *The quality maturity model: assessing organisational quality culture in academic libraries*. Brunel University, PhD Thesis. http://bura.brunel.ac.uk/handle/2438/8747

Witt, M., Carlson, J., Brandt, D. S. & Cragin, M. H. (2009). Constructing Data Curation Profiles. *The International Journal of Digital Curation*, 4(3): 93-103. DOI: 10.2218/ijdc.v4i3.117

Appendix 1. Survey cover letter

Sent out on the 10th of June 2020

Dear colleagues,

Many of you have written, or will need to write a **research data management plan** as part of a grant application. The aim of this survey is to gain insight about your experiences, potential problems and support needs. The results will be used in the development of research data support at VTT, based on actual needs.

The survey takes approximately 10-15 minutes to complete and your response will remain anonymous. If you want personal follow-up on your feedback or questions, you have the option to leave your details in the questionnaire form. Your personal details will be excluded from the results.

The survey is also a part of my Master's thesis on participatory design of research data management support services, conducted at the University of Oulu and VTT. Anssi Neuvonen and [redacted] in the [redacted] team are supervising this work. Feel free to contact us directly with any questions about the survey.

As the summer holiday season is coming up, we would like to ask you to answer promptly, if possible by the **17th of June**. If you're still working after Midsummer, you can send your answer during the following week. Your participation is highly appreciated!

Open this link to take the survey.

Best regards,

Appendix 2. Survey questionnaire

Research Data Management and Planning Survey

In this survey you will be asked about your research data management practices, planning and support needs. The topic is very complex, but we value your time, so there will be mostly choice and agree-disagree type of questions. There are a few open-ended questions which can be answered briefly with just a few words, but feel free to express your opinions and experiences in as many words as you need. You can write in English or Finnish. There will be space for additional comments and feedback at the end of the survey.
* Compulsory field

Q1 Your VTT Team*: open-ended

Q2 How long have you been working in research?*

- ○ 0-2 years
- ○ 2-5 years
- ○ More than 5 years

Q3 Your current experience with writing a Data Management Plan (DMP)*

- ○ I have previously written a Data Management Plan (DMP)
- ○ I am preparing to write my first DMP for an upcoming grant application
- ○ Not relevant for me personally

Q4 If you have previously written a DMP, how easy or difficult did you find it
1 = Very easy, 2 = Quite easy, 3 = Neither easy nor difficult 4 = Quite difficult 5 = Very difficult

Q5 Please rate how confident you are in the following aspects of data management planning*

1= Not at all confident – 2 = Somewhat not confident – 3 = Neutral - 4 = Somewhat confident - 5 = Very confident

Understanding what kind of research data the project produces
Understanding how the research data could be beneficial also to other users after the project
Understanding funders' requirements related to data management
Finding and using DMP templates (incl. the language used in templates)
Complying with the FAIR data principles
Metadata and documentation of the research data
Secure research data storage and sharing during the active phase of the research
Information on VTT tools and services available for storage and sharing
Possibilities of long-term data storage after the research
Understanding the possibilities and limits of opening the data
Budgeting for research data management
Agreements on data ownership, licenses and access control
Ethical issues: sensitive data, GDPR

Q6 Would you find support or training in these aspects beneficial?*
1 = No support or training needed – 3 = Neutral - 5 = Support would be very beneficial

Understanding what kind of research data the project produces
Understanding how the research data could be beneficial also to other users after the project
Understanding funders' requirements related to data management
Finding and using DMP templates (incl. the language used in templates)
Complying with the FAIR data principles
Metadata and documentation of the research data
Secure research data storage and sharing during the active phase of the research
Information on VTT tools and services available for storage and sharing
Possibilities of long-term data storage after the research
Understanding the possibilities and limits of opening the data

Budgeting for research data management
Agreements on data ownership, licenses and access control
Ethical issues: sensitive data, GDPR

Q7 Any other support needs not covered in the previous question? Open-ended

Q8 Do you agree with the following statements regarding the benefits of planning your research data management:*
1 = Strongly disagree – 2 = Disagree – 3 = Neutral – 4 = Agree – 5 = Strongly agree

Writing a good DMP could increase my chances to win the grant
Writing a DMP gives useful insights into how the research data collection and data storage should be described in the rest of the funding application
DMP supports seamless collaboration during the project: a mutual agreement within the collaboration on what tools or conventions to use during the project (e.g. file naming, sharing, backup…)
Efficiency: planning ahead can make research data management during the project and opening the data after the project easier
Opening the data could increase the impact and visibility of my research

Q9 Any other perceived pros (or cons) of writing a DMP? (Skip if not applicable)

Q10 What tools or services have you used for data storage during the active phase of research?*
(Multi-choice from list of options provided by VTT, internal information)
Other:

Q11 Have you made the underlying data available in a repository or data journal in connection with publishing the results?*

☐No

☐Yes, as open access

☐Yes, with managed access

☐I have published only the descriptive metadata

Q12 If you have made the data or its metadata available, what service did you use? For example Zenodo, Fairdata IDA, any other general, organisational or subject based repository…
Open-ended

Q13 If you have made the data available, what were your reasons? Skip if not applicable.

☐Funder's requirement

☐Publisher's requirement

☐Project collaborator's wish

☐An aspiration to raise the visibility and reusability of my research results

☐Other:

Q14 Have you encountered any issues that prevented you from opening the research data or part of it?*

☐Intellectual property rights, confidentiality

☐Data protection (personal or sensitive data)

☐Lack of resources for the work needed for opening the research data

☐Not enough knowledge about how or where to open the data

☐Other:

Q15 If the data can't be opened publicly, but they can be shared within VTT, where do you store the data after project closure and how do you make sure the data is also findable and accessible for VTTers?*
Answer N/A if not applicable in your case.
Open-ended

Q16 Do you have experience with reusing data collected and shared by someone outside your own research group or research project?*

○ Yes

o No

Q17 Do you know where you would look for help inside VTT if you needed help writing a data management plan? Name any people, teams, websites, guidelines, events to sign up for etc. that come to mind or that you have used before.*
Or you can simply answer No, if nothing comes to mind.
Open-ended

Q18 What kind of DMP/data management support tools or services would you like to have available?*
Pick your preferred option, as many as you want.

☐My own contact network, e.g. colleagues, someone I know in a research support service
☐ Organizational data support service, e.g. a helpdesk that would point me to the specialist I need
☐Online materials: guides, instructions etc.
☐Training, courses
☐Pop-up clinics, e.g. at our team meetings
☐Discussion board such as a dedicated Yammer group
☐DMP workshops
☐Guides for online DMP tools such as DMPTuuli
☐Feedback on a completed DMP
☐Example DMPs from successful funding applications from inside the organization
☐Other:

Q19 Other comments or feedback?
If you would like us to follow up on your questions or feedback and get back to you, please leave your name or e-mail address. Your details will be removed before further processing of the survey.
Open-ended

Appendix 3. Interview instrument

RISE self-assessment model + Avoimen tieteen kansallisen koordinaation Tutkijan datapalvelut -työryhmän Tukipalvelumallit

This model consists of 21 capabilities in 10 areas of research data management (RDM). The model is based on the British experience and capability descriptions may not be fully relevant in the Finnish environment. The Finnish national recommendation for support service levels is added to bring a local approach, however the Finnish model emphasizes Open Science while RISE focuses on RDM.

Note that it is not realistic nor expected to aim for the highest level in each capability. This tool aims

RISE Levels:	The Finnish model levels:
1 = Compliance	1 = Minimal
2 = Locally-tailored services	2 = More comprehensive
3 = Sector-leading activity	3 = Vision for the future (not feasible yet)

1) RDM policy and strategy

1a) Policy development

Level 1	Level 2	Level 3
Institutional policy **articulates roles & responsibilities** for researchers, other staff and students to **comply with legal & regulatory obligations and external funders' RDM policy** expectations.	Institutional **policy articulates the value of good RDM practice to the institution** and its rationale for **retaining data of long-term value**. **Policy is subject to a regular, scheduled review process**.	Institutional **policies** with a bearing on RDM (e.g. FOI, ethics, research conduct, etc.) **are joined up and complementary**. Policies are **externally promoted, aiming to push the sector forward**.

1b) Awareness raising and stakeholder engagement

Level 1	Level 2	Level 3
Research data **policies are promoted to all relevant staff**, students and researchers	Guidance on **how to apply all relevant policies to the institutional context is provided and promoted** to all relevant staff, students and researchers.	Policies are **promoted by the institution through channels designed to engage** with staff, student and researcher groups' specific interests

1c) RDM implementation roadmap

Level 1	Level 2	Level 3
RDM roadmap is **compliance-focussed and defined by funder requirements**	Roadmap is informed by the **institution's strategies** and its **researchers' priorities.**	Roadmap/strategy **seeks to derive competitive advantage from RDM support**. It aims to be sector-leading and innovative.

*Avoimen tieteen kansallinen koordinaatio: Tukipalvelumallit

Johtaminen

Level 1	Level 2	Level 3
Johto ymmärtää datan avoimuuden merkityksen	Johto panostaa datan avoimuuteen • Tavoitteet, resurssit, seuranta/mittarit ja palkitseminen • Yksiköiden välisen yhteistyön organisointi	Datan avoimuus on kansainvälisesti verkostoituneen tutkimusorganisaation strateginen valinta osana avointa tiedettä.

2) Business Plans and Sustainability

2a) Staff Investment

Level 1	Level 2	Level 3
RDM **service is delivered by dividing responsibilities among existing staff.** *(Note: ad hoc solutions, staff not specifically assigned to RDM)*	RDM service is delivered through **significant redesign of staff roles including investment in staff development.**	The RDM service is delivered by **major redesign of staff roles, consistent with the establishment of an RDM service**.

2b) Technology Investment

Level 1	Level 2	Level 3
A **base level of investment** in technical infrastructure, with **commitment to supporting recurring costs**, ensures that researchers can make their data findable and accessible in the long-term.	The institution coordinates **investment in the central technical services it deems a strategic priority** for research data life-cycle support.	The institution **invests in technical infrastructure for all aspects of the research data life cycle**, interoperating with tools and workflows at research group level

2c) Cost modelling

Level 1	Level 2	Level 3
All RDM service costs are covered by overheads on grants.	Standard RDM services are funded through grant overheads. Where support exceeds the norm mechanisms allow for direct charging of grants.	Cost modelling enables specialist, stand-alone RDM services to be offered alongside standard support provision. (e.g. statistical modelling service or data visualisation service).

*Avoimen tieteen kansallinen koordinaatio: Tukipalvelumallit Henkilöresurssit		
Level 1	Level 2	Level 3
Vähintään yksi henkilö • Tuntee RDM:n perusteet ja pyrkii ylläpitämään osaamistaan • Tietää mistä löytyy RDM-tietoa • Osaa ylläpitää verkkosivua tai saa tukea verkkosivun ylläpitoon • Työaikaa on kohdennettu RDM-tiedon hankintaan ja verkkosivun ylläpitoon	• Organisaation koon mukaisesti riittävä resurssointi tutkijan datapalvelujen tarjoamiseen ja kehittämiseen • Roolien tunnistaminen ja vastuunjako yksiköiden välillä • Osaamisen kehittäminen on suunnitelmallista ja jatkuvaa • Kansallinen ja kansainvälinen verkostoituminen on osa RDM-työtä	Hyvin resurssoitu ja monipuolisen osaamisen omaava Datastuerttien tiimi ylläpitää ja kehittää organisaation Datanhallinta-palvelualustaa osana EOSCia ja muita kansainvälisiä federoituja datan hallinnan alustoja.

3) Advisory Services		
(General)		
Level 1	Level 2	Level 3
Generic, online guidance is offered that addresses key areas of RDM. Content may be **externally sourced**, with little relating to the specific institutional context. Pages include a **helpdesk email address**.	Guidance offers **relevant advice on how to use services that comply with institutional policies**, and the benefits to researchers of doing so	Guidance is **significantly tailored to support the specific needs of the institution's researchers** and support staff. Guidance content is externally referenced as sector best practice

RISE Recommendation: *Typically, advisory service provision will vary in capability depending on institutional context and strategic priorities. So it may help to note under the table which topics the service is cable of providing at each level.*

The following topics are from VTT RDM user need survey (see the background data below). Consider if there is staff available to provide advice (at what level?). Note that support or guidance can be sourced externally.

Ethical issues: sensitive data, GDPR

Agreements on data ownership, licenses and access control

Budgeting for research data management

Understanding the possibilities and limits of opening the data

Possibilities of long-term storage after the research

Information on VTT tools and services for storage and sharing

Secure data storage and sharing during the project

Metadata and documentation of research data

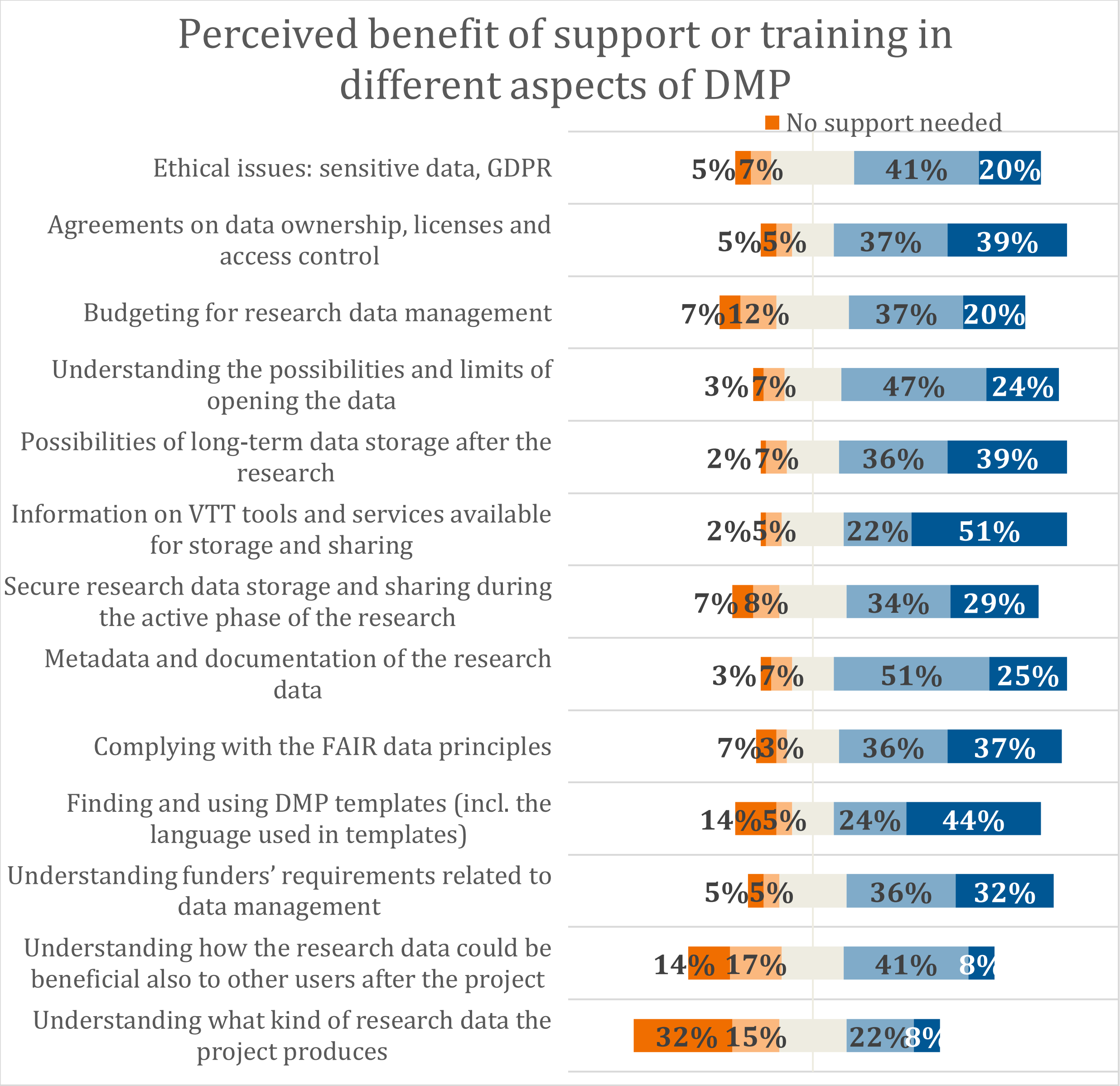
Complying with the FAIR data principles

Finding and using DMP templates

Understanding funders' requirements regarding RDM

Understanding how data can be beneficial, also to other users after the project

Background data: VTT RDM survey 6/2020 (n = 59)



4) Training

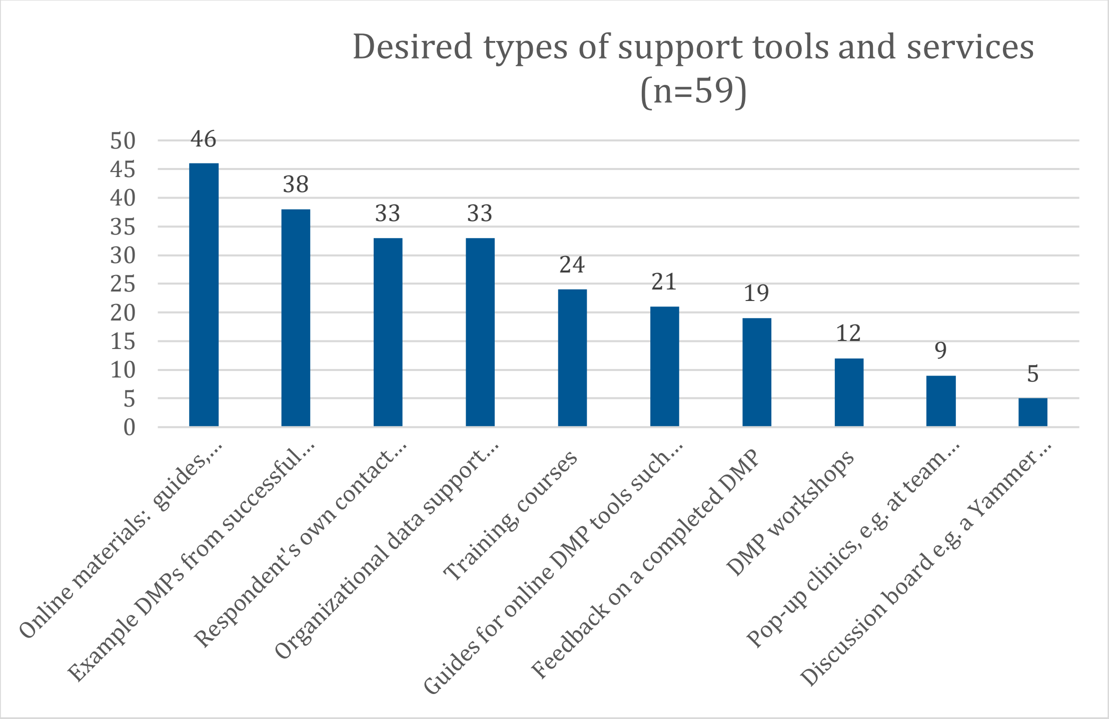
4a) Online training

Level 1	Level 2	Level 3
Externally sourced online courses are **linked to from RDM pages.**	**Externally sourced** online courses are **supplemented with some materials which support local needs** and services.	The institution produces a **significant amount of online training material which meets the needs of its researchers** and staff. Materials are reused by others in the sector.

4b) Face to face training

Level 1	Level 2	Level 3
Face to face training in **basic RDM principles is available on request**. Course content is regularly updated and responsive to feedback.	Regular, **structured face to face RDM courses are available to all**. Training objectives are aligned with the objectives of the institution's RDM strategy.	**Competencies for relevant researchers and professional support staff are defined in standard role descriptions**. Training is provided which facilitates this development.

Background data: results from the RDM survey at VTT from 6/2020



*Avoimen tieteen kansallinen koordinaatio: Tukipalvelumallit Tukipalvelut (mixed Advisory + Training)		
Level 1	Level 2	Level 3
Verkkosivu, jossa on linkkejä, joiden takaa lyötyy esim. • Organisaation RDM ohjeistus • Vastuullinen tiede Aineistot jatkokäyttöön–sivu • Datan tallennuspalveluja ○ Organisaation omat ○ Organisaation tutkijoiden tieteenalojen data-arkistot ○ Yleiset esim.https://zenodo.org/ • Muita datan hallintaan liittyviä palveluja, esim. anonymisointi • Datan avoimuuden kursseja/webinaareja • Avoimen tieteen kursseja/webinaareja	• Verkkosivut, joista löytyy kattavasti RDM-tietoa • Helpdesk yhdessä osoitteessa <> yksiköiden välinen yhteistyö • Henkilökohtainen tukipalvelu tilattavissa • DMP tuki • Koulutukset • Viestintä (Markkinointi) monikanavaisesti	• Verkkosivu, joka esittelee Datanhallinta-palvelualustan käytön • Koulutukset ○ datanhallinnan merkitykseen eri tieteenaloilla eri kohderyhmille räätälöitynä ○ Datanhallinnan-palvelualustan käyttöön • Tekoälyyn ja koneluettaviin (FAIR) metatietoihin ja tutkijan/tukihenkilöiden syöttämiin tietoihin perustuva Datanhallinta-palvelualusta hoitaa datan avaamisen DMP:n tietojen pohjalta ○ DMP:n hallinta ○ Sopimukset ao. osapuolien välillä ○ Datan anonymisointi tarvittaessa ○ Lisenssit ja käyttöehdot ○ Datan kuvailu verkkosivuille ○ Tilastotietojen keruu ja jakelu ○ Rajapinnat datan jakeluun ja hyödyntämiseen/uudelleen käyttöön

5) Data Management Planning		
Level 1	Level 2	Level 3
Institution provides guidance to researchers on completing **funder-mandated DMPs** as part of grant bids.	**Institution mandates DMP production at bid stage for all researchers**. **Guidance and templates are provided**. Research Office connects to relevant stakeholders **to appraise DMP content** and **notify them of relevant resource implications**.	Institution **promotes best practice in data management planning** and facilitates good research design in relation to data generation and preservation**. Automated systems flag researcher requirements to the relevant institutional support services** (e.g. exceptionally large projected data volumes)

6) Active Data Management

6a) Scaleability and synchronization

Level 1	Level 2	Level 3
The service provides researchers with **managed access to networked storage**, from **multiple devices**, of **sufficient capacity and performance to satisfy most** of the organisation's projects.	The service **can provide additional storage on request** to satisfy exceptional storage capacity, device networking, or performance demands.	The service **provides automated access to additional storage** to satisfy exceptional capacity or performance demands.

6b) Collaboration support

Level 1	Level 2	Level 3
The service enables **access to data for external collaborators** by providing them with **local access rights to institutional storage systems.**	The service provides **managed access to tools that enable researchers to share data with external collaborators**.	The service provides **managed access to virtual research environments** that enable researchers to **work on data with external collaborators**.

6c) Security management

Level 1	Level 2	Level 3
The service provides **authenticated access to storage that is protected from unauthorised data access**, and researchers are **made aware of procedures for data protection and de-identification**.	The service provides **tools/environments that enable researchers to de-identify, encrypt or control access** to data as required.	The service provides researchers from across the institution with **access to ISO 27001/2 or equivalently accredited facilities for analysis of shared sensitive data.**

7) Appraisal and Risk Assessment

7a) Data collection policy

Level 1	Level 2	Level 3
Service **primarily supports data deposit to third-party repositories,** and **holds datasets in-house when legal/ regulatory compliance requires**	Service **defines criteria for retention of datasets of long-term value to the institution**	Service defines **criteria for developing datasets as special collections** and **ensures these meet specialist depositor and user needs**

7b) Security, legal and ethical risk assessment

Level 1	Level 2	Level 3
Service **seeks confirmation that data was collected or created in accordance with legal and ethical criteria** prevailing in the data producer's geographical location or discipline	Service **commits to proactively manage legal and ethical risks** relevant to its depositors and users, and **to relevant professional and technical development for researchers and support staff**	Service **offers data producers tailored guidance on risk assessment, and on solutions that offer an appropriate level of risk control** for the data they manage

7c) Metadata collection to inform decision-making

Level 1	Level 2	Level 3
Information is gathered from research projects to enable the **identification of research data that must be kept for compliance purposes**	**Metadata is routinely recorded to relate research activity to data and other outputs**, and enable better informed **decisions on the preservation costs, risks and value to the institution**	Metadata on data and related research outputs is **sufficiently well-structured and interoperable to enable added value to be extracted** for service users' needs.

8) Preservation

8a) Preservation planning and action

Level 1	Level 2	Level 3
Service demonstrates it can ensure **continued bit-level integrity** of the data collections it holds, its **metadata, and its links to any related information submitted with it**	Service enables **preservation plans e.g. file migration** or normalisation to be enacted **at time of ingest or dissemination, and records all actions**, migrations and administrative processes it performs	Service commits to deploy tools and expertise to maintain the significant properties of data, metadata and related information for required retention periods and identified user groups **(full preservation)**

8b) Continuity Support

Level 1	Level 2	Level 3
Service enables retained data to be stored **with a copy automatically held in another location**	Service enables retained data to be stored with **copies automatically held in two separate locations,** at least one off-site	Service enables **data & metadata to be automatically distributed across multiple locations** according to specific policy criteria

9) Access and Publishing

9a) Monitoring locally produced datasets

Level 1	Level 2	Level 3
Information is gathered from research projects to enable compliance with funders' requirements for research data discoverability.	Metadata is routinely recorded on locally produced data, and its links to research activity or related outputs, enhancing the quality of the institution's research information.	Metadata on locally produced research data, and its links to other activities or outputs, is sufficiently structured and organised to inform institutional strategy.

9b) Data publishing mandate		
Level 1	Level 2	Level 3
Service supports minimum external requirements for metadata and publicly accessible data.	Service supports community best practice standards for data access, citation and metadata exchange.	Service supports bespoke content discoverability, access and quality review needs for user groups or organisations.

9c) Level of data curation		
Level 1	Level 2	Level 3
Service commits to brief oversight of submitted data and metadata e.g. for compliance purposes.	Service commits to maintain or enhance value through routine action across data collections.	Service commits to maintain or enhance value through bespoke action on individual collections.

10) Discovery		
Metadata cataloguing scope		
Level 1	Level 2	Level 3
Service catalogues metadata for the organisation’s publicly funded datasets according to funder expectations that they are discoverable, citable, and linked to related content	Service catalogues metadata to enhance value of the institutions data assets in accordance with recognised best practice standards	Service catalogues metadata to enhance potential dataset reuse according to sector-leading standards, or fulfil domain-specific purposes