Research Data Management

University of Oulu
Information Processing Science
Bachelor Thesis
Harri Hirvonsalo
2024

Abstract

Research data management (RDM) is essential for all researchers, as data is central to modern scientific discovery (Tenopir et al., 2020). RDM involves actively and continuously managing data throughout its entire lifecycle (Cox & Pinfield, 2014; Wilms et al., 2020). Good RDM supports knowledge discovery, innovation, and the reuse of data (Wilkinson et al., 2016), while poor data management can lead to unsuccessful and harmful research projects (Kanza & Knight, 2022). RDM specialists, such as librarians, assist researchers by providing services, consulting on the data lifecycle, and offering strategies and guidelines (Bryant et al., 2023a; Redkina, 2019; Sun et al., 2023). However, defined RDM standards can create a high workload for researchers, making it difficult for them to practice RDM effectively (Wilms et al., 2020). To build effective RDM infrastructures and services, it is important to integrate the practices and perspectives of both researchers and support specialists (Sun et al., 2023).

In this literature review, RDM, as presented in literature is reflected against RDM service portfolio of CSC – IT Center for Science (*What CSC?*, n.d.), to find out if the RDM support CSC provides, is how recommended in literature. Results of this study imply that service and support offered by CSC are according to recommendations and objectives of research data management as presented in literature. Results of the study can be used to prepare further academic studies. Results of the study have been used to support work related decision making of author of this thesis.

Keywords
Research Data Management, Research Data Management support, Research Data Management services

Supervisor
PhD, University lecturer, Pertti Seppänen

Foreword

Thanks to all friends and colleagues at University of Oulu, at CSC – IT Center for Science, and at Imperial College London, who have given me instructions for making of this thesis.

Contents

1. Introduction

Research data management (RDM) is essential for all researchers, especially as data becomes central to modern scientific discovery (Tenopir et al., 2020). Data sharing is crucial for scientific progress, making research traceable and reusable for future work (Hackman et al., 2024). Data discovery has evolved into a distinct topic with support from organizations like the Research Data Alliance, and the "Fourth Paradigm" (Hey et al., 2009) of data-intensive science relies on capturing, curating, and analyzing data, i.e. RDM practices (Tenopir et al., 2020).

The volume of digital data is growing rapidly, making effective data management increasingly challenging (Wilms et al., 2020). Good RDM leads to knowledge discovery, innovation, and the integration and reuse of data (by the community) after publication (Wilkinson et al., 2016). Planning of RDM needs to consider not just the current moment, but also long-term use and possible cross-disciplinary sharing in understandable form (Hackman et al., 2024). Researchers invest significant effort in collecting, systematizing, and analyzing data, and managing it before publishing (Redkina, 2019). Poor data management can result in unsuccessful and potentially harmful research projects (Kanza & Knight, 2022). Recommendations and roadmaps from bodies like the European Commission and National Science Foundation along private foundations encourage making publicly-funded research data accessible (Tenopir et al., 2020). Data management plans (DMPs) are a requirement in grant applications, detailing how data will be created, shared, published, and preserved (Tenopir et al., 2020).

Good RDM practices are guided by the FAIR principles (Boeckhout et al., 2018; Jacobsen et al., 2020; Wilkinson et al., 2016), which aim to make data Findable, Accessible, Interoperable, and Reusable. The European Commission's 2018 report emphasizes implementing FAIR principles (European Commission. Directorate General for Research and Innovation., 2018). RDM specialists, such as librarians, help researchers by providing services, consulting on the data lifecycle, and offering strategies and guidelines such as describing semantics of data and selection of data formats. (Bryant et al., 2023a; Redkina, 2019; Sun et al., 2023).

However defined RDM standards can create a high workload for researchers, hindering their ability to practice RDM (Wilms et al., 2020). Building and maintaining digital infrastructure for RDM is typically the responsibility of IT-oriented support staff (Sun et al., 2023). RDM practices are often studied separately from the perspectives of researchers and support specialists, which can lead to infrastructure and services based on perceptions rather than actual practices (Sun et al., 2023). To build effective RDM infrastructures and services, it is crucial to integrate the practices and perspectives of both researchers and support specialists (Sun et al., 2023).

Purpose of this study is to introduce what RDM is and how RDM should be supported according to literature, and find out if service and support offering of national level support organization, CSC – IT Center for Science (*What CSC?*, n.d.), supports RDM as mentioned in literature. Research question of this thesis is *"are research data management services and support offered by CSC, aligned with requirements and recommendations for research data management as presented in literature". I.e. d*oes CSC provide what literature recommends.

The structure of this paper is as follows. Section 2 present research methodology and research problem is defined. Section 3 is a brief overview on RDM as presented in

literature; what is data and RDM, what is data lifecycle, what are FAIR principles as RDM guidelines, what are dataset and metadata, how RDM can be supported. Summary of CSC's role and activities in RDM support are presented in section 4. Section 5 contains a discussion and conclusions of the study.

2. Research methods

This study is conducted as a literature review. Research question of this thesis is *"are research data management services and support offered by CSC, aligned with requirements and recommendations for research data management as presented in literature"*. Research protocol of this literature review was defined by combining guidelines from literature review instructions by Knopf (Knopf, 2006) and from mapping study and systematic literature processes as defined by (Kitchenham & Charters, 2007). Snowballing method defined by Wohlin (Wohlin, 2014) was the main method for searching interesting publication to be included in this literature review.

To prime the information search process for this thesis a very generic *"research data management"* search clause was used at Google Scholar service. Results were limited to publications made after 2015. Few highly cited articles were selected to be included in this literature review.

To discover related scientific publications, *Connected Papers* -service (*Connected Papers | Find and Explore Academic Papers*, n.d.) was used to create similarity graph for each selected Google Scholar result. The graph displays a network of publications as nodes connected by edges, arranged according similarity of the publications (*Connected Papers | About*, n.d.). Similarity metric is based on Co-citation (Donthu et al., 2021) and Bibliographic Coupling (Donthu et al., 2021), where two papers presumed to treat related subject if their citations and references overlap highly (*Connected Papers | About*, n.d.). Graph helps to visualize clusters of similar papers, which help to identify related information and point-of-views of the general topic of *"research data management"*. Example of *Connected Papers* similarity graph is presented in Figure 1.

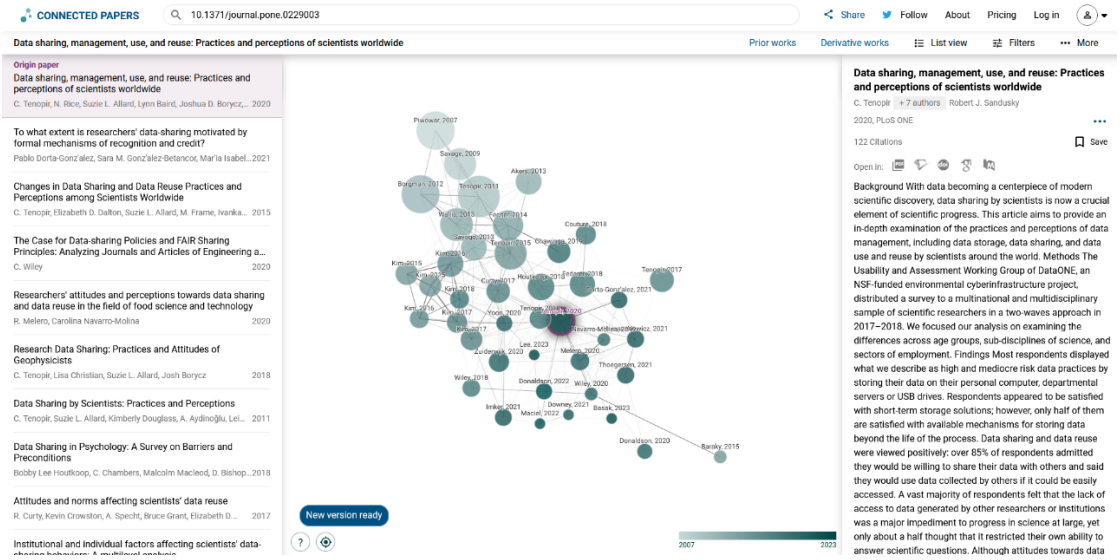

Figure 1. Screenshot of information graph created at Connected Papers -service. In the public domain.

Publications presented in similarity graph were included for further analysis if, based on title and abstract of the publication, the publication could be presumed to contain relevant information in regards to research questions of this thesis. Selected publications were further investigated using *Semactic Scholar* -service (*Semantic Scholar | AI-Powered Research Tool*, n.d.), which is an AI-powered search and discovery tool to support

discovery of scientific literature, by extracting meaning and identifying connections from within scientific publications (*Semantic Scholar | Frequently Asked Questions*, n.d.). Backward and forward snowballing was done utilizing reference lists and citation lists constructed with *Semactic Scholar*. Similarly, as with Connected Papers, interesting and presumably relevant publications selected for further investigation.

At this point, relevance of each selected article assessed by reading the article abstract and section titles of each paper. If needed, relevance was assessed in more detail by skimming through each section of the article. Articles that did not seem relevant were excluded from this literature review.

High impact of articles was seen as mark of relevance. Impact of each paper was roughly evaluated by obtaining citation count from *Crossref* (*Crossref*, n.d.) and *Sematic Scholar* (*Semantic Scholar | AI-Powered Research Tool*, n.d.) -services. Higher citation count was assumed to correlate with higher impact. Articles with higher impact were examined in more detail than articles with low impact.

Finnish Publication Forum (*Publication Forum*, n.d.) was used to assess quality publication channels that scientific source articles were published in. It was seen as a benefit, if publication forum of an article was listed in Finnish Publication forum, but absence of the publication forum did not exclude the article from this study. Articles which had been published in a forum listed by Finnish Publication Forum, were examined in more detail than articles from forums not listed there.

In addition to literature sources, numerous webpages have been used as reference to technologies, tools, services and collaborations, when suitable scientific reference did not exist.

3. Research Data Management

This section describes how research data and research data management (RDM) have been described in literature (section 3.1). After that data lifecycle and data lifecycle models are introduced in section 3.2. Section 3.3 introduce FAIR principles and section 3.4 present definitions that literature offers for dataset and metadata. Conceptual RDM models as well as RDM service models and roles are presented in section 3.5.

3.1 Research data and research data management

To understand research data management (RDM), it's essential to first understand what research data can be.

According to definition by OECD (OECD, 2007), research data are factual recordings, such as numerical scores, text records, images, and sounds, that serve as primary sources for scientific research. (Redkina, 2019) Another comprehensive definition describes research data as any records, files, or other evidence, regardless of their content or form, which are research observations, findings, or outcomes, including both primary materials and analyzed data. (Sheikh et al., 2023)

In general, RDM is the act of managing this research data. RDM involves organizing, describing, storing, and sharing data used in research. It includes various processes like creating, capturing, storing, organizing, documenting, disseminating, reviewing, publishing, discovering, reusing, retaining, archiving, or destroying data according to agreed policies. (Sheikh et al., 2023)

Cox and Pinfield (2014) bind RDM activities to data lifecycle and expand list of RDM activities to include security, preservation, retrieval of research data, while “taking into account technical capabilities, ethical considerations, legal issues and governance frameworks”.

RDM is active and ongoing management of data that covers the entire data lifecycle from entry of the research data to the research cycle, all the way to the dissemination and archiving of valuable results (Wilms et al., 2020). RDM encompasses all the activities, tools, and infrastructure needed to manage research data effectively (Sheikh et al., 2023) including long-term storage and accessibility of research data over time, protection of data, usage of data repositories and facilitating exchange of data (Wilms et al., 2020).

RDM impacts data reuse and reproducibility, from planning the details of data collection to addressing long-term data plans (Sheikh et al., 2023).

Besides collection of primary data or deriving data from existing sources, RDM covers also “any informational data that has been given meaning by way of relational connection, as well as research publications which can be described as a type of stored knowledge” (Wilms et al., 2020)

RDM supports researchers by making it easier to analyze, search, and store data, ensuring reliable verification of results and enabling new research built on existing information. (Terra et al., 2021)

3.2 Data lifecycle

As highlighted in section 3.1, research data management (RDM) is a set of actions, policies, technical means that deal with data used in research. To gain further insight into RDM, it is critical to understand that data is not static or isolated (Surkis & Read, 2015).

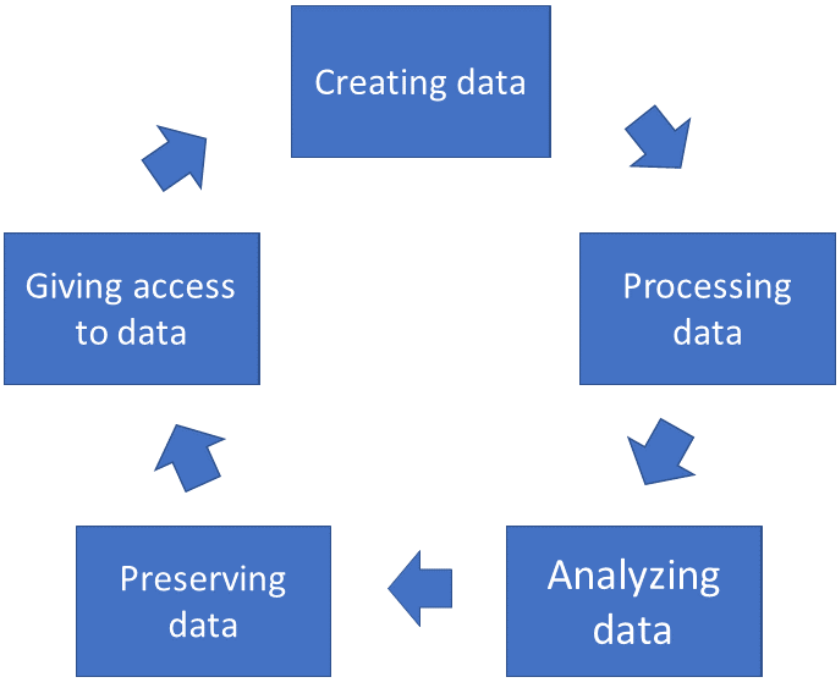

Figure 2. 5-phase data lifecycle model adapted from Surkis & Read (2015). Copyright 2015 by Surkis & Read.

In RDM, the concept of the data lifecycle helps researchers understand the full scope of managing data (Terra et al., 2021). This lifecycle includes stages such as creating or collecting data, processing it for analysis, and analyzing the data to produce academic outputs like journal articles (Surkis & Read, 2015). In addition, data lifecycle also concerns data after publication of journal articles; does the data need to preserved, how the will data be preserved and how it will be published to enable data reuse. Throughout the lifecycle of data, documentation, metadata, copyright, etc. must be properly managed so that data is understandable and usable by others (Sheikh et al., 2023). Stages of data lifecycle also illustrate that also RDM is by nature continuous and often iterative process, not just a single task.

Figure 3. 7-phase research data lifecycle adapted from Sheik et al. (2023). Copyright 2023 by Sheik et al.

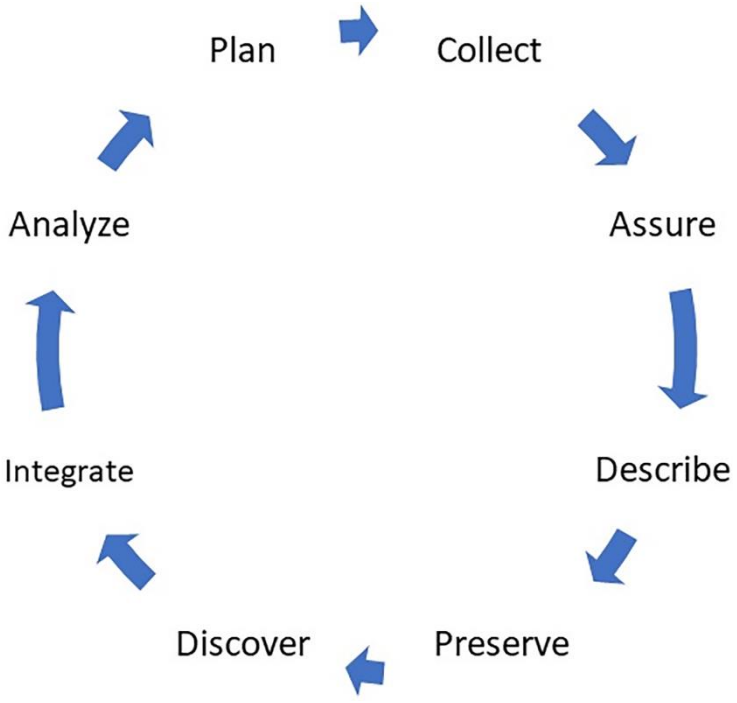

Figure 4. 8-phase data life cycle from Tenopir et al (2020). Creative Commons CC0.

Various data lifecycle models exist. Model presented presented in Figure 2, by Surkis and Read (Surkis & Read, 2015), list creation, processing, analysing, preserving and giving access to data as the 5-phases of data lifecycle. Model by Sheik et al. (Sheikh et al., 2023) list 7-phases; Plan, Collect and capture, Process, Analyse, Publish and share, Preserve

and Reuse (Figure 3). Tenopir et al. (Tenopir et al., 2020) use 8-phase model (Figure 4) for describing phases of data lifecycle. Summary of phases and their names for each model is presented in Table 1.

Table 1. Summary of phases of different data lifecycle models.

5-phase model (Fig. 1)	**7-phase model (Fig. 2)**	**8-phase model (Fig. 3)**
Creating data	Plan	Plan
Processing data	Collect & capture	Collect
Analysing data	Process	Assure
Preserving data	Analyze	Describe
Giving access to data	Publish & Share	Preserve
	Preserve	Discover
	Reuse	Integrate
		Analyze

From these models, it can be seen that data lifecycle is bind to actions; someone or something does something with data, which moves data into a different phase in its lifecycle. Given that data can be reused in another study, data lifecycle is somewhat continuous process. Reason for multiple models is more about emphasis of different point-of-views and granularity of actions, rather than question of right or wrong. For example, Integrate -phase of 8-phase model where "*data from multiple sources are combined into a form that can be readily analyzed"* (*Data Life Cycle: Integrate*, n.d.), is, by its definition, very close to Processing data -phase in 5-phase model, which defined as *"processing data from its rawest form to another form for analysis"* (Surkis & Read, 2015). 7-phase model emphasizes that there is a "Reuse" -phase, which 5- and 8-phase models don't have.

3.3 FAIR principles

Good research data management (RDM) is not an end goal, but means to more and better research. It helps in knowledge discovery, innovation and to integrate data and knowledge during and after research, throughout the data lifecycle (Boeckhout et al., 2018; Wilkinson et al., 2016). Effective RDM and stewardship lead to high-quality digital publications, for both data and results, making it easier for researchers to discover, evaluate, and reuse data in future studies (Tenopir et al., 2020; Wilkinson et al., 2016). However, what defines "good data management" is not clearly defined and is often left to the discretion of the data owner (Wilkinson et al., 2016). Therefore, to foster creation and understanding of best practices, providing clear guidelines and goals for good data management for those who publish and preserve scholarly data are needed. (Wilkinson et al., 2016)

The concept of "FAIR data principles" has been widely adopted by various stakeholders in research, including academia, industry, funding organizations, and scholarly publishers (Sheikh et al., 2023). The principles serve as guidance for facilitating data sharing more

systematically, enabling digital resources to become more Findable, Accessible, Interoperable and Reusable (FAIR) for machines and thus also for humans (Boeckhout et al., 2018; Jacobsen et al., 2020). These four foundational principles are more explicitly and measurably described by 15 FAIR guiding principles (Wilkinson et al., 2016).

Table 2. The FAIR principles.

Findable:	
F1.	Metadata and data are assigned a globally unique and persistent identifier.
F2.	Data are described with rich metadata (defined by R1 below).
F3.	Metadata clearly and explicitly include the identifier of the data it describes.
F4.	Metadata and data are registered or indexed in a searchable resource.
Accessible:	
A1.	Metadata and data are retrievable by their identifier using a standardized communications protocol.
A1.1.	The protocol is open, free, and universally implementable.
A1.2.	The protocol allows for an authentication and authorization procedure, where necessary.
A2.	Metadata are accessible, even when the data are no longer available.
Interoperable:	
I1.	Metadata and data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2.	Metadata and data use vocabularies that follow FAIR principles.
I3.	Metadata and data include qualified references to other metadata and data.
Reusable:	
R1.	Metadata and data are richly described with a plurality of accurate and relevant attributes.
R1.1.	Metadata and data are released with a clear and accessible data usage license.
R1.2.	Metadata and data are associated with detailed provenance.
R1.3.	Metadata and data meet domain-relevant community standards.

Summary of FAIR principles is presented in Table 2. Boeckhout (Boeckhout et al., 2018) describe each principle in more detail.

The principle of Findability requires that data should be clearly identified, described, and registered or indexed. This means assigning a unique and persistent identifier to each dataset, specifying the main characteristics of the data using standard formats, and storing or indexing this information in a public resource like a data archive or institutional repository. (Boeckhout et al., 2018)

The principle of Accessibility states that datasets should be accessible through a clearly defined access procedure, ideally through automated means. When needed, authentication

and authorization procedures must be in place. Even if data is not or no longer available, metadata of the dataset should always be accessible. (Boeckhout et al., 2018)

The principle of Interoperability means that data and metadata should be structured using common, published standards. This involves using standards for technical and semantic data, formats, variables, and ontologies. The standards themselves should also be FAIR, meaning they are published, traceable, and accessible. (Boeckhout et al., 2018)

The principle of Reusability emphasizes that data characteristics, including their origins, should be described in detail according to community standards relevant to the domain. This includes providing accurate data descriptions, access and usage licenses, and documenting the community standards related to creation and use of the data, as well as provenance for each dataset. (Boeckhout et al., 2018)

The FAIR principles apply digital research objects in general, such as algorithms, tools, and workflows. Analytical workflows, for example, are crucial in scholarly research, and their formal publication is necessary to achieve transparency and scientific reproducibility. The FAIR principles can be applied to these assets as well, meaning they should be identified, described, discovered, and reused just like the assets we traditionally see as the data. In best case, all parts of the research process are available to ensure transparency, reproducibility, and reusability. (Terra et al., 2021; Wilkinson et al., 2016)

FAIR principles also push to enable machines to make optimal use of data resources (Boeckhout et al., 2018; Jacobsen et al., 2020; Wilkinson et al., 2016). In a way, FAIR aims to ensure that "the machine knows what we mean.". By Jacobsen et al. (Jacobsen et al., 2020) this has following implications for all foundational principles.

Findability: Digital resources should be easy to find for both humans and computers. Extensive, machine-actionable metadata is required to automatically discover relevant datasets and services. (Jacobsen et al., 2020)

*Accessibility***:** The protocols for fetching digital resources, both data and metadata, should be well and explicitly defined to enable machine-use. This includes mechanisms for obtaining authorization to access protected data. (Jacobsen et al., 2020)

*Interoperability***:** It should be possible for a machine to merge digital resources of same topic or entity into a unified view. Machines should also automatically detect and facilitate interactions between data and services and tools compatible with this data. This requires that the meaning (semantics) of each resource, whether data or service, is clear and comprehensive. (Jacobsen et al., 2020)

*Reusability***:** Machines should be able to determine if a resource is relevant to a task, understand the conditions under which it can be reused, and know whom to credit for its reuse. (Jacobsen et al., 2020)

Overall FAIR is a prerequisite for proper data management and data stewardship. (Wilkinson et al., 2016)

3.4 Dataset and metadata

The term "dataset" is often interwined with "data". In short, dataset is a concept that groups data into a single unit (Borgman, 2012).

The concept of a dataset is common across almost every scientific discipline where data forms the basis for research activities. Although the term "dataset" is frequently used in articles, papers, reports, and informal conversations among scientists, there is no precise, established definition. Datasets are groups of data treated collectively as a unit; a collection of related, specific types data. On the other hand, datasets are considered to group together related data in a way that goes beyond merely being a collection of similar entities. For example, a dataset can be thought of as data related by factors like time, place, instrument, or object of observation. These features emphasize the circumstances around the creation or maintenance of a dataset, rather than any internal characteristics of the data itself. (Renear et al., 2010)

While the term dataset is useful for creating collection of data, it does not explain what data the dataset contains (Borgman, 2012). Metadata of the dataset explains what the data in the dataset is, as well as data about the dataset itself.

Metadata is essentially data about data. Narrowly defined, metadata refers to systematic descriptions and attributes of datasets, similar to bibliographic information about publications. Metadata typically contains structured details about the resource, which help manage access and content; location of the resource, and how resource can be retrieved. More broadly, metadata includes all data about data, such as information about theoretical assumptions, methods and techniques used, and the provenance and context needed for proper interpretation and meaningful reuse. This type of metadata provides further information about a dataset, making it easier to find, access, and use.(Boeckhout et al., 2018; Hackman et al., 2024)

Metadata of dataset is typically grouped together according to specified structure, a *metadata schema* (Hackman et al., 2024). In addition to structure, metadata schema specifies what kind of information about the dataset should be recorded (Hackman et al., 2024). Information specified by the schema could be very generic, discipline-agnostic information, i.e. information that can be recorded for all datasets, such as dataset title, creators, file formats of the data included in the dataset (Sun et al., 2023). Scientific-discipline-specific schema allow to record information that is not relevant, or even possible to record for all datasets (Sun et al., 2023). For example, information about the elevation in atmosphere in meters, where data was gathered, is information that is hardly relevant for all datasets of all scientific disciplines, and could be even impossible to obtain.

Metadata schemas developed and maintained by organizations or institutions are known as metadata standards (Hackman et al., 2024). Schemas such as *Datacite metadata schema* (*DataCite Schema*, n.d.) or *EUDAT Extended metadata schema* (*EUDAT Metadata Schema Documentation*, n.d.), are generic dataset metadata standard, whereas *EML* and *INSPIRE* metadata schemas are domain-specific: *INSPIRE* is for geospatial information in the European scientific communities (*Overview - European Commission*, n.d.) and *EML* is used in the earth and environmental sciences, and increasingly in other research disciplines as well (“Ecological Metadata Language,” 2023).

Metadata are crucial for digital data curation, following the FAIR principles (Findable, Accessible, Interoperable, and Reusable) (Hackman et al., 2024). Review by Chapman et

al. (2019) emphasizes the importance of good metadata for data discovery. The visibility of a dataset depends on the quality of the metadata provided; the better the metadata, the easier it is to find the dataset (Sun et al., 2023). To enable data reuse, the origin and context of production of the dataset has to be recorded (Silva et al., 2016). Therefore a workflow that covers the entire data lifecycle is needed to link data and metadata from the beginning and clearly document the data production process (Silva et al., 2016).

Good quality metadata needs contributions from the researchers involved in creating the data, as their domain knowledge is essential for adequately documenting the context of dataset production so that others can reuse it (Amorim et al., 2017). However, researchers are not experts in data management. They need effective tools that help them produce adequate, standards-compliant metadata records without requiring them to learn about those standards (Amorim et al., 2017). Tools like electronic laboratory notebooks can help motivate researchers to actively describe their data (Silva et al., 2016).

3.5 Research data management models, services and roles

As research data management (RDM) covers a range of complementary yet distinct categories for services, roles and stakeholders. several models have been developed to help in conceptualizing (RDM) (Cox & Pinfield, 2014; Curdt, 2019; Bryant et al., 2023b; Sheikh et al., 2023; Sun et al., 2023).

Model by Jones et al. (2013) presented in Figure 5, outlines the components needed for institutions to build and operate effective RDM services, highlighting the need for a comprehensive strategy and need for sustainability plans; predicted costs and planned expenditures, resource deployment and enhancement, predicted returns on investment, etc. are point-of-views that other models presented in this section do not emphasize.

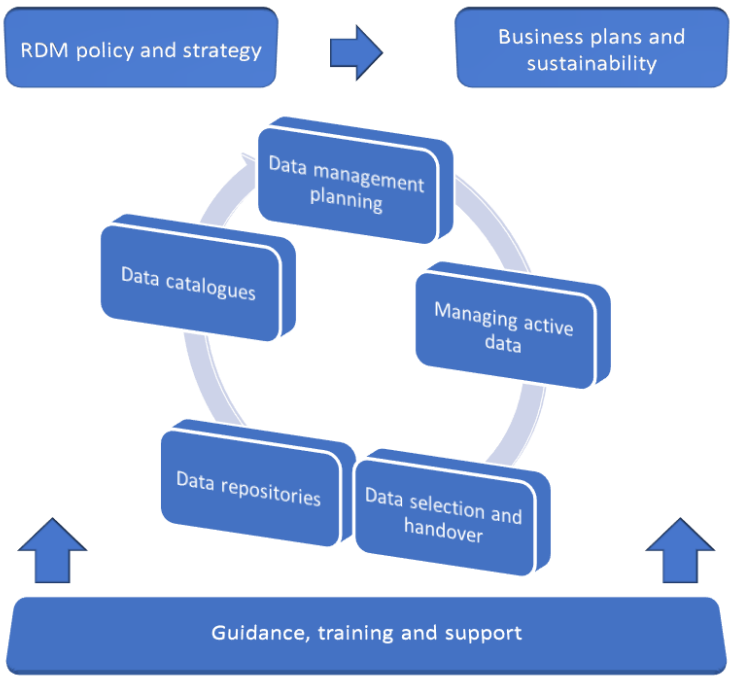

Figure 5. Components of research data management support services. Adapted from Jones et al. (2013). Copyright 2013 by Digital Curation Centre.

Model by Pinfield et al. (2014) presented in Figure 6, emphasizes influences of different stakeholders of RDM through four key factors: what (components), why (drivers), how (influencing factors), and who (stakeholders). Drivers include funders' mandates, security concerns, open-access needs, data storage, preservation, and sharing. Influencing factors are demand, roles, resources, acceptance, and communications. Stakeholders include libraries, IT services, academic departments, university managers, legal offices, research support services, and researchers. (Cox & Pinfield, 2014)

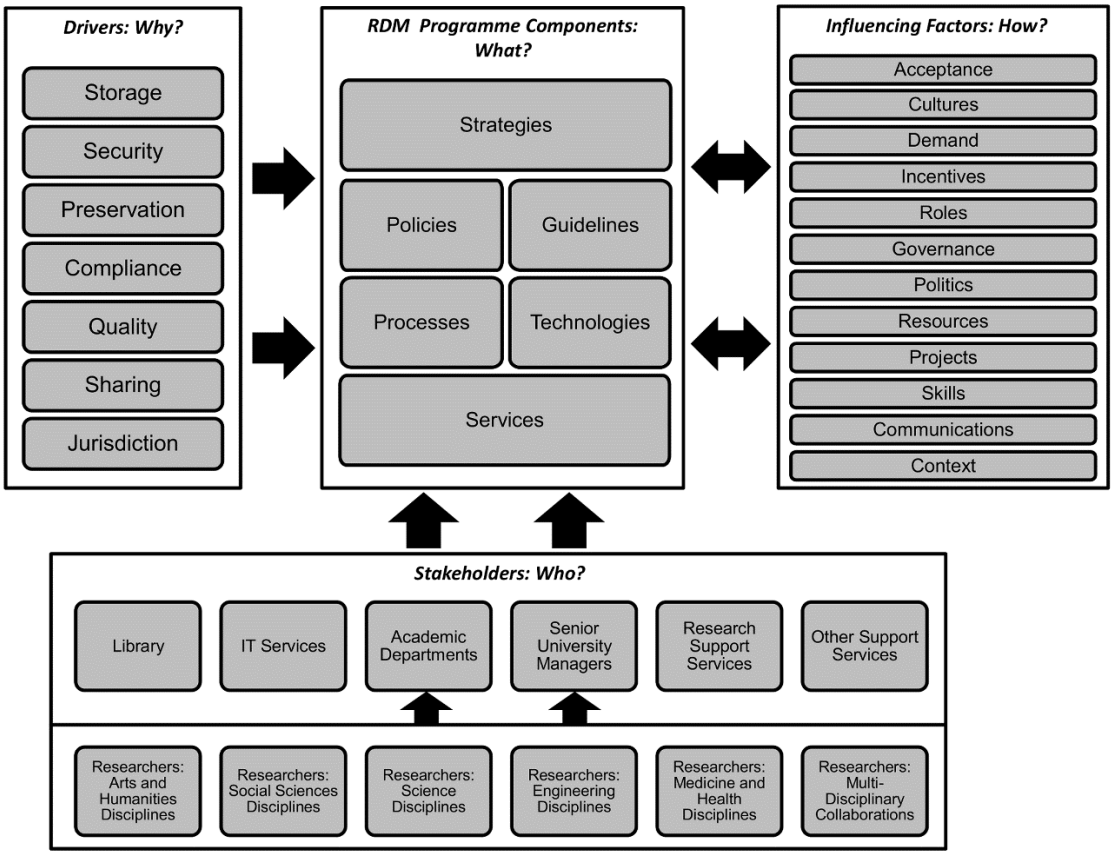

Figure 6. A library-oriented model of institutional RDM from Pinfield et al. (2014). Creative Commons CC0.

RDM service model by Bryant et al. (2023a) presented in Figure 7, divides RDM services into three categories; *Education services, Expertise services, Curation services.*

Education services raise awareness and teach researchers and other stakeholders about RDM. Education services provide information on incentives for practicing RDM; tools to ease RDM; RDM requirements set by funders, agencies, their own institutions, as well as discipline-specific norms and practices. Education services provide general information that can be used independently of any specific research project or organization. (Bryant et al., 2023a)

Expertise services provide human-mediated, specialized knowledge of data librarians, technologists, and other support staff to support RDM activities done by researchers. Rather than relying on unmediated resources like online tutorials, researchers get

customized support at various stages of their research; e.g. support for metadata creation and preparation of data for dataset deposit. Although expertise services include training programs for internal staff who support data management at the institution, key difference to education services training is the customization of the support to specific research project or organization. (Bryant et al., 2023a)

Curation services offer technical means needed to manage datasets throughout the data lifecycle, from near-term (during active research), medium-term (for a few years after research concludes), to long-term (extending indefinitely after data deposit), supported by both local and distributed RDM infrastructure and services. The services encompass persistent storage, unique identifier assignment, access controls, metadata creation and management, versioning, and long-term preservation. National legislative, funder-imposed, scientific community guidelines and local institution policies define important aspects of RDM curation services, such as data retention, metadata requirements, access restrictions, and privacy assurances for sensitive data. (Bryant et al., 2023a)

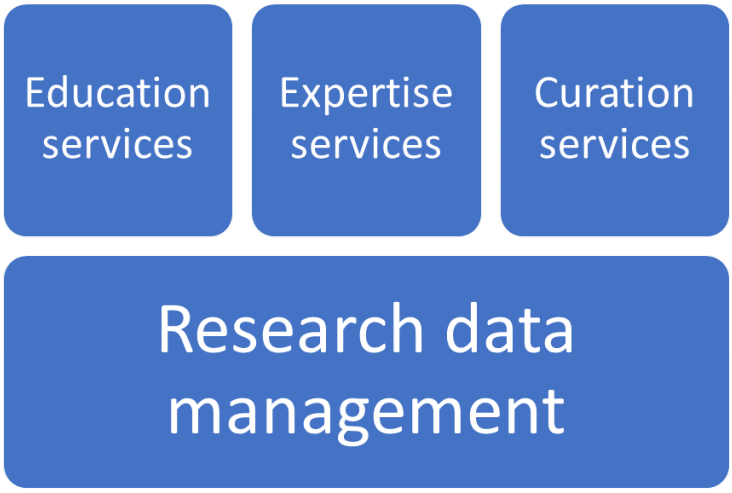

Figure 7. Research data management service categories. Adapted from Bryant et al. (2023a). Copyright 2023 by OCLC, Inc.

RDM service model by Curdt et al. (2019) presented in Figure 8 is based on experiences obtained during developing and operating of an RDM system over 10 years, for long-term (2007-2017) interdisciplinary research-project. Services listed in the model support scientists during their research by providing:

- *training and support service* for data publishing, searching, and using the RDM system efficiently.
- *internal data sharing service* to enable sharing of unorganized raw data which is useful for other researchers of the project.
- *secure data storage and backup service* for storing finalized, processed and prepared data ensuring long-term access and availability of the data.
- *data documentation service* where researchers describe metadata of dataset using existing metadata schemas and controlled vocabularies to ensure the data are findable, accessible and reusable.
- *data publishing service* for the actual publication of datasets with persistent identifier (Digital Object Identifier, DOI) to enable the data to be citable in a publication or referenced in other datasets.

- *data search service* to enable researchers to find data by data types, topics, regions and sites, funding phases or project sections through different search queries.
- *data download and access service* to control access to project datasets, through permissions and license.
- *data statistics service* to offer visibility to availability and reuse of project data at the repository and dataset levels.
- *web mapping services* based on dataset metadata, to inspect geographic coverage of datasets on a map

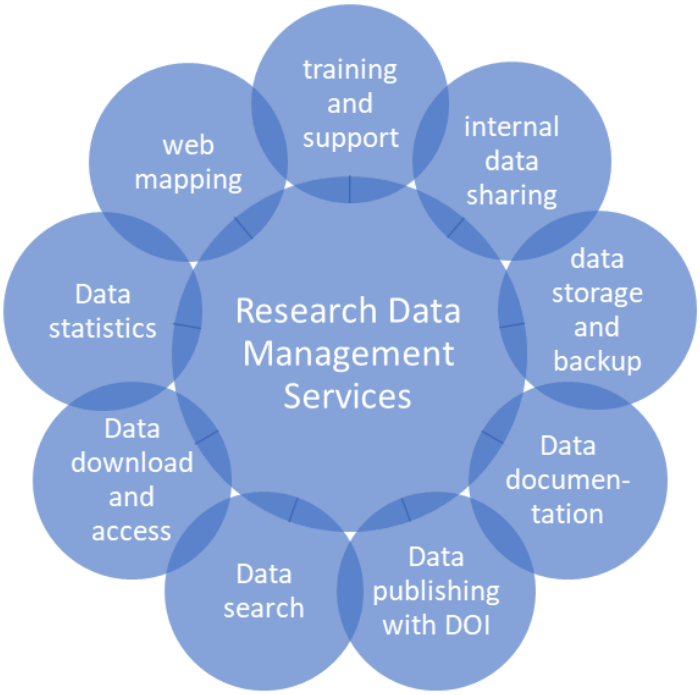

Figure 8. Research data management services model. Adapted from Curdt (2019). CC BY 3.0.

According to Curdt (2019) establishing RDM services and infrastructure, such as data storage, backup, documentation, search, and access, is essential in collaborative, cross-institutional, interdisciplinary, long-term research. Collaborating researchers often rely on data from their colleagues, not just their own data, requires active sharing of research data, documents, and other information in a well-managed, controlled, and structured manner. RDM systems facilitate data exchange within the project and enable data reuse by future project participants. Infrastructures should align with current standards and principles, and they should be set up according to the needs and requests of the scientists, since scientists will only use infrastructures if they are designed to meet their needs. Operating with the model presented in Figure 1. has contributed significantly to the overall success of the research project which the model is based on. (Curdt, 2019)

Sun et al (2023) list "*three primary types of research data management support work: people-oriented roles, e.g., providing consultations; metadata-related roles, e.g., providing data documentation services which are non-technical but focus on enhancing the completeness of research data description and data discoverability; and infrastructure-oriented roles, e.g., providing IT-focused technical tasks"*.

People oriented roles provide information, consulting, training, and active involvement in data management planning. They offer guidance during research on topics like data storage and file security, help with research documentation and metadata, and support data sharing and curation. (Sun et al., 2023)

Managing research data requires building and maintaining digital infrastructures that meets user needs throughout the research data lifecycle. *IT-oriented RDM support staff* typically handle this technical work. The main goal of technical service providers is to maintain research data infrastructure and ensure its long-term availability for data sharing and storage. (Sun et al., 2023)

Sun et al. (2023) emphasize that roles and type of work done in RDM support are not mutually exclusive. Due to complexities of RDM support work RDM specialist benefits blending technical and metadata-oriented skills that allow to work in multiple roles depending on the nature of requested support. (Sun et al., 2023)

In second part of their report series, Bryant et al. (2023b) say that when higher education institutions, research institutes and research communities consider how they should arrange RDM support, it is not enough to just tell researchers about the importance of RDM without providing practical solutions. Researchers need actual, working RDM services, not just promises of future tools Bryant et al. (2023b). Offering comprehensive RDM support and services requires coordination and collaboration among various stakeholders, including libraries, IT services, legal advisors, research support offices, and the research community (Sheikh et al., 2023). Ambition of RDM services is not to gather data itself, but to enable researchers to insert an utilize the data, the information and the knowledge collected during research (Juhás et al., 2017). Therefore, development of the services must happen through requirements set by researchers' practices, not by perception of these practices by specialists (Sun et al., 2023). When services are easy to use and provide both immediate- and long-term added value for researchers, it is more likely that research adopt the services as part of their daily research work (Amorim et al., 2017).

An RDM service bundle can include services built and deployed locally as well as those sourced externally. National centers of RDM capacity are important supplements to the local RDM services as local RDM can rely on resources available on a national or consortial scale in the wider RDM service ecosystem. Even if many services are provided externally, the local institution still plays an important role in RDM support. Local staff need to identify and broker access to appropriate RDM solutions for researchers and sometimes mediate the use of these services to ensure that researchers' needs are met. (Bryant et al., 2023b)

4. RDM support in practice at CSC

In addition to local, institution-arranged research data management (RDM) support, that for example higher education institutes in Finland offer to their researchers, national level support also exists (*Higher Education Institutions and Science Agencies - OKM - Ministry of Education and Culture, Finland*, n.d.).

CSC is a non-profit enterprise entrusted with special state assignment (i.e. the company has an assignment defined, regulated and overseen by the state) to build solutions for digitalization for research, the national education system, culture and public administration together with its customers. CSC is owned by the state of Finland (70% of the shares) and Finnish higher education institutions (30% of the shares). (*What CSC?*, n.d.)

Shareholders form a joint advisory council, through which higher education institutions draw CSC ownership strategy and monitor its implementation (*CSC - Governance*, n.d.). CSC delivers the services needed by its customers within the framework of its mission specified in the ownership strategy and company strategy (*CSC - Values, Vision and Strategy Guide Everything We Do*, n.d.). The customers develop the services they need with CSC, and CSC supplies them at cost price. Customer steering is listed as the most important driving force in the development of a new service and development throughout the service lifecycle (*CSC - Customers are at the center of our work*, n.d.). Higher education institutions, along research institutes, research infrastructures, public administration, archives, libraries and museums are customers of CSC (*CSC - Customers are at the center of our work*, n.d.).

In addition to customer and shareholder steering, goals of CSC service development and offering is affected by goals of research and development projects which CSC participates in. Examples of such projects are HORIZON2020 (*Horizon 2020 - European Commission*, n.d.) funded projects, European Strategy Forum for Research Infrastructures – ESFRI (*About | Www.Esfri.Eu*, n.d.) funded projects, and Finnish Research Infrastructures – FIRI (*Research Council of Finland - Research infrastructures*, n.d.) funded projects by Research Council of Finland (*Research Council of Finland*, n.d.). These funding instruments are research centered and as such, have at least indirect effect on RDM services and support CSC provides. For example, EUDAT B2SHARE -service (*EUDAT B2SHARE*, n.d.), that CSC jointly provides as a member of EUDAT Collaborative Data Infrastructure (EUDAT CDI) (*EUDAT CDI | EUDAT*, n.d.) , has been initially developed under co-funding from projects funded by HORIZON2020.

CSC actively participates in many international collaborations and forums which foster and developed research practices and tools in general, or directed towards particular fields of science (*CSC - Our international collaborations support research*, n.d.). Examples of such collaborations and forums, are Nordic e-Infrastructure Collaboration – NeIC (*Home - NeIC Web*, n.d.), European Genome-phenome Archive – EGA (*About - EGA European Genome-Phenome Archive*, n.d.), Research Data Alliance – RDA (*Research Data Alliance*, n.d.) and European Open Science Cloud – EOSC (*EOSC Association*, n.d.). European Open Science Cloud in particular has been a policy maker on services for research provided by European entities, such as CSC (*CSC - EOSC Association Archives*, n.d.).

Data management is explicitly included in CSC's strategic objectives (*CSC - Values, Vision and Strategy Guide Everything We Do*, n.d.):

> *"We make customized solutions for data management to our customers and develop our competence and services regarding responsible utilization of data. We produce solutions which help in managing open data available for science and research, securely processing sensitive data, and ensuring data preservation long into the future with digital preservation services."*.

CSCs RDM service portfolio is presented in sections 4.1-4.10.

4.1 Fairdata Services

The *Fairdata Services* (*Fairdata | Take Care of Your Research Data*, n.d.; *Fairdata Services - Services for Research - CSC Company Site*, n.d.) are integrated services for storing, sharing and publishing research data. The services are not tied to a specific field of science. The Fairdata service components are *IDA* (*IDA - Store and Organize Your Research Data before Publishing It*, n.d.) for storing research data, *Qvain* (*Qvain – Research Dataset Description Tool*, n.d.) for creating metadata descriptions for datasets, the metadata warehouse *Metax* (*Metax API - Interfaces for Transferring and Reading Metadata*, n.d.), the research data finder *Etsin* (*Etsin - Research Dataset Finder*, n.d.), open data publishing platform *AVAA* (*AVAA – Open Data Publishing Platform*, n.d.) and the *Digital Preservation Service for Research Data* (*Digital Preservation Service for Research Data*, n.d.). The services are offered free of charge to Finnish universities, universities of applied sciences and state research institutes. Digital preservation service is based on agreement. (*Fairdata Services - Services for Research - CSC Company Site*, n.d.)

4.2 EUDAT Services

EUDAT CDI (*EUDAT CDI | EUDAT*, n.d.) is a European research data infrastructure, with a vision *"Data is shared and preserved across borders and disciplines"*. CSC is a member of EUDAT CDI and EUDAT services can be accessed through CSC (*CSC - EUDAT*, n.d.). EUDAT CDI provides researchers and practitioners from any research discipline heterogeneous research data management (RDM) services and storage resources, through a geographically distributed, resilient network distributed across 15 European nations, where data is stored alongside some of Europe's most powerful supercomputers (*EUDAT CDI | EUDAT*, n.d.). *B2FIND* is metadata indexing service of EUDAT and provides a discovery portal which allows users to find data collections within an international and inter-disciplinary scope (*B2FIND | EUDAT*, n.d.). *B2SAFE* is a large-scale data management service, which offers an abstraction layer of large scale, heterogeneous data storages, for community and departmental repositories to implement data management policies on their research data (*B2SAFE | EUDAT*, n.d.). *B2SHARE* is dataset repository for researchers, scientific communities and citizen scientists to store, publish and share research data in a FAIR way (*B2SHARE | EUDAT*, n.d.). *B2DROP* is a low-barrier and user-friendly storage environment which allows users to synchronize their active data across different desktops and to easily share this data with peers (*B2DROP | EUDAT*, n.d.). *B2ACCESS* is a federated cross-infrastructure authorization and authentication proxy for user identification and community-defined access control enforcement (*B2ACCESS | EUDAT*, n.d.). *B2HANDLE* (*B2HANDLE | EUDAT*, n.d.) is the distributed service for storing, managing and accessing persistent identifiers (PIDs) and essential metadata (PID records) as well as managing PID namespaces. *B2HANDLE* is mostly transparent to end-users, shielding them from the complexity of infrastructure details (*B2HANDLE | EUDAT*, n.d.).

4.3 Sensitive Data Services

Sensitive Data services for research (*SD Services for Research - Services for Research - CSC Company Site*, n.d.) are designed to support secure sensitive data management through web-user interfaces accessible from the user's own computer. *SD services* are comprised of five components that together help users manage sensitive data during all the phases of their research. Users can store and share encrypted sensitive data with *SD Connect* (*SD Connect - Services for Research - CSC Company Site*, n.d.) and create a private workspace to compute sensitive data with *SD Desktop* (*SD Desktop - Services for Research - CSC Company Site*, n.d.). At the end of their research, users can publish sensitive data under controlled access with *SD Submit* or biomedical data with *Federated EGA* (*FEGA - Services for Research - CSC Company Site*, n.d.). *SD Apply* (*SD Apply - Services for Research - CSC Company Site*, n.d.) service enables data controllers to manage data access permissions for reuse of published research data. SD services support national requirements for sensitive data and comply with General Data Protection Regulation (GDPR) (*SD Services for Research - Services for Research - CSC Company Site*, n.d.). Secondary SD Desktop service (*SD Desktop - Services for Research - CSC Company Site*, n.d.) exists to be used as a certified computing environment for secondary use of health and social data according to Finnish regulation, specifically according to Findata regulations (*Findata - Regulations*, n.d.) and the Act on Secondary Use of Social and Health Data (*Secondary Use of Health and Social Data*, n.d.).

4.4 Allas storage service

Allas (*Allas - Services for Research - CSC Company Site*, n.d.) is CSC's general-purpose data storage service. It provides an environment for storing and sharing data. Allas support storing static research data that needs to be available for analysis as well as collecting and hosting cumulating or changing data (*Allas - Services for Research - CSC Company Site*, n.d.).

4.5 CSC Noppe -service

CSC Noppe-service (*Notebooks - Services for Research - CSC Company Site*, n.d.) provides easy-to-use electronic notebooks for working with data and programming. Service is accessed and used via web browser and CSC cloud computing environment handles computing activities on the background. Different versions of Jupyter Notebooks (*Project Jupyter*, n.d.), RStudio Server (*Posit*, n.d.) and Apache Spark (*Apache Spark™ - Unified Engine for Large-Scale Data Analytics*, n.d.) tools are provided via Notebooks. (*Notebooks - Services for Research - CSC Company Site*, n.d.)

4.6 Research Information Hub -service

The National Research Information Hub (*Home | Research.Fi*, n.d.) gathers and shares information on scientific research carried out in Finland. All the information on researchers, publications, research data, research projects, and research infrastructures is made available in one place, openly accessible. Service facilitates administrative work and the reporting of researchers as it automates information flow between research organizations, funders, and other research services. (*Research Information Hub - Services for Research - CSC Company Site*, n.d.)

4.7 Cloud- and high-performance -computing environments

CSC offers a variety of cloud computing services: *Pouta* services, *cPouta* (*cPouta - Services for Research - CSC Company Site*, n.d.) and *ePouta* (*ePouta - Services for Research - CSC Company Site*, n.d.) and *Rahti* container cloud service (*Rahti - Services for Research - CSC Company Site*, n.d.). *Pouta* services are Infrastructure-as-a-Service (IaaS) services, where users can launch and manage virtual machines, deploy storage, and create networks for their needs. *cPouta* is targeted towards generic cloud needs, while *ePouta* is mainly used for handling sensitive data. *Rahti* is container orchestration service based on Red Hat's OKD distribution of Kubernetes (Innes, n.d.), offered on top of *cPouta*. Compared to *Pouta*, *Rahti* provides a higher-level service for deploying applications to virtual environments, where *Rahti* automatically takes care of a lot of the lower level management, on behalf of the user. CSC offers two high-performance computing (HPC) environments, *Mahti* and *Puhti*. *Mahti* supercomputer is designed for massively parallel jobs of medium and large scale simulations, that require large floating point performance and a capable interconnect (*Mahti - Services for Research - CSC Company Site*, n.d.). *Puhti* is designed to be flexible enough to run interactive single core data processing to medium scale simulations spanning multiple nodes (*Puhti - Services for Research - CSC Company Site*, n.d.). *Puhti* also offers an artificial intelligence partition (Puhti AI) that is suitable for heavy AI models spanning multiple nodes.

4.8 Persistent Identifier (PID) related -services

CSC offers national services for persistent identifiers, with focus on identifiers for research data used by research data management (RDM) provided by CSC and other national entities. CSC coordinates the DataCite Finland consortium and is a member of the ePIC consortium (*Persistent Identifiers for eResearch*, n.d.). CSC can provide Handle identifiers (*Handle System*, 2024), such as Digital Object Identifier – DOI Digital object identifier, 2024). Several services at CSC also utilize URN identifiers (Uniform Resource Name, 2024) offered by the National Library of Finland (*Home | Kansalliskirjasto*, n.d.). CSC collaborates with other national entities in defining national PID policies. (*CSC - Persistent Identifiers*, n.d.)

4.9 Customer specific services

Kapseli (*Kapseli®*, n.d.) is a Finnish Social and Health Data Permit Authority Findata (*About Findata*, n.d.) provided secure operating environment for the processing of data on individuals. It enables secure processing of sensitive data subject to a Findata permit. (*CSC - A secure service family for sensitive data*, n.d.)

CSC maintains and develops Statistics Finland's (*Tilastokeskus - FIONA*, n.d.) remote access system Fiona (*Tilastokeskus - FIONA*, n.d.), which is the organization's secure processing environment for unit-level data needed in research, including Statistics Finland's micro data. (*CSC - A secure service family for sensitive data*, n.d.)

The Language Bank of Finland (*Language Bank | Kielipankki*, n.d.) is a service for researchers using language resources across digital humanities and social sciences. It has a wide variety of text and speech corpora and tools for studying them. (*Language Research and Other Humanities and Social Sciences - Services for Research - CSC Company Site*, n.d.)

Finnish Meteorological Institute – FMI (*About us—Finnish Meteorological Institute*, n.d.) research data repository METIS (*METIS*, n.d.), is provided by EUDAT and enables the institute data to be preserved, discovered, and accessed. FMI covers a wide range of research on weather, sea, climate and space, and research data produced by FMI is made openly accessible through EUDAT B2SHARE based METIS service (*Open Research Data—Finnish Meteorological Institute*, n.d.).

4.10 National level research support

In addition to services, CSC provides a lot of support channels and materials to Finnish research organizations and higher education institutions (*Accounts and Support - Services for Research - CSC Company Site*, n.d.). *CSC Docs -site* (*Docs CSC*, n.d.) provides documentation about CSC services, but also documentation on best practices in various research related topics, such as research data management (RDM). CSC offers *data management planning support* to researchers on topics such as data types and amounts of data, ethical and legal issues, data documentation and metadata, data storage during a project and information security, making data available for reuse and preserving data, data management responsibilities and resources, and budgeting of data management (*Plan Data Management - Services for Research - CSC Company Site*, n.d.). CSC has participated in development of Data Management Plan tooling and offered some such as *DMPTuuli* (*DMPTuuli*, n.d.), which continues to be offered to all researchers by University of Helsinki. CSC provides *remote and on-site expert support and training* for sciences, methods and tools, especially on biosciences, chemistry, computational engineering, geosciences, social sciences and humanities, mathematics and statistics, physics, data analytics, visualization, and code optimization (*Sciences and Methods - Services for Research - CSC Company Site*, n.d.). CSC produces and provides *training videos* at *Video CSC* (*Video CSC*, n.d.) and *Youtube* -services (*CSC — Tieteen Tietotekniikan Keskus / CSC — IT Center for Science*, n.d.) and *self-study online courses* (*CSC | e-learn*, n.d.) on RDM.

CSC Research Data Management Competence Center promotes and supports the skills and competence development of RDM. Aim of the center is to foster Open Science and FAIR data management with adequate knowledge and best practices for managing, reusing, sharing and analyzing research data. To increase the quality, reproducibility and productivity of research, RDM Competence Center coordinates, develops and provides data management training, provides consultation and support for organizations, data management experts and researchers, works both in international and national projects, and collects feedback on CSC´s data management services, which are used in future service development. (*CSC - Good data management is a requisite for successful research*, n.d.)

CSC – Data Support Network gathers data management experts in research organizations and Finnish higher education institutions to develop their skills and exploit the opportunities for collaboration as a member of the network using train-the-trainer principle. (*CSC - Good data management is a requisite for successful research*, n.d.)

5. Discussion and conclusions

In general, research data management (RDM) is the act of managing research data and all-encompassing definition that would gather all the activities done in context RDM has yet to emerge.

According to literature, RDM should address the whole data lifecycle while considering technical capabilities, ethical considerations, legal issues and governance frameworks (Bryant et al., 2023a; Cox & Pinfield, 2014). RDM should be supported by guidelines and best practices such as FAIR principles, made collaboratively by researchers and experts in forums such as Research Data Alliance (Bryant et al., 2023a; Wilkinson et al., 2016).

RDM is supported by services and tooling as well as education and expert support (Bryant et al., 2023a; Wilkinson et al., 2016). Services and tooling enable and ease RDM practices throughout data lifecycle. Expert support helps researchers in using the services and tools, but more importantly helps researchers to understand why and how RDM is done in the first place (Bryant et al., 2023a). RDM helps in knowledge discovery, innovation and to integrate data and knowledge during and after research, throughout the data lifecycle (Boeckhout et al., 2018; Wilkinson et al., 2016). Effective RDM and stewardship lead to high-quality digital publications, for both data and results, making it easier for researchers to discover, evaluate, and reuse data in future studies (Tenopir et al., 2020; Wilkinson et al., 2016).

Offering comprehensive RDM support and services requires coordination and collaboration among various stakeholders, including libraries, IT services, legal advisors, research support offices, and the research community (Sheikh et al., 2023). Development of the services must happen through requirements set by researchers' practices, not by perception of these practices by specialists (Sun et al., 2023).

National level support for RDM can be addressed in multiple ways, complementary to each other (Bryant et al., 2023b). Higher education institutions and research institutes often offer education and expert support to their employees (*Higher Education Institutions and Science Agencies - OKM - Ministry of Education and Culture, Finland*, n.d.). In Finland, these activities are complemented by discipline-agnostic RDM services and support offered by CSC – IT Center for Science, a non-profit enterprise entrusted with special state assignment to provide support for research and education in Finland (*What CSC?*, n.d.).

The customers develop the services they need with CSC, and CSC supplies them at cost price. Customer steering is listed as the most important driving force in the development of a new service and development throughout the service lifecycle. Higher education institutions, along research institutes, research infrastructures, public administration, archives, libraries and museums are customers of CSC. Since users of CSC's services are mainly associated to these customers institutions, users are main influencer of RDM offerings of CSC. (*CSC - Customers are at the center of our work*, n.d.)

CSC offers wide variety of services for RDM; *Fairdata* services (*Fairdata | Take Care of Your Research Data*, n.d.; *Fairdata Services - Services for Research - CSC Company Site*, n.d.) for national level research, *EUDAT services* (*EUDAT CDI | EUDAT*, n.d.) for international collaboration in research, *Sensitive Data services* (*SD Services for Research - Services for Research - CSC Company Site*, n.d.) made specifically to answer challenges

of sensitive data, *Pouta* (*cPouta - Services for Research - CSC Company Site*, n.d.; *ePouta - Services for Research - CSC Company Site*, n.d.) and *Rahti* cloud computing services (*Rahti - Services for Research - CSC Company Site*, n.d.), high-performance computing environments *Mahti* (*Mahti - Services for Research - CSC Company Site*, n.d.) and *Puhti* (*Puhti - Services for Research - CSC Company Site*, n.d.), petabyte level storage service *Allas* (*Allas - Services for Research - CSC Company Site*, n.d.), research results aggregation service Research.fi (*Home | Research.Fi*, n.d.), persistent identifier services (*CSC - Persistent Identifiers*, n.d.), as well as other services each made to support some part of RDM during data lifecycle.

CSC offers education and expert support and services for RDM through *CSC Research Data Management Competence Center*, which coordinates, develops and provides data management training, provides consultation and support for organizations, data management experts and researchers, works both in international and national projects, and collects feedback on CSC´s data management services, which are used in future service development. *CSC Data Support Network* gathers data management experts in research organizations and Finnish higher education institutions to develop their skills and exploit the opportunities for collaboration as a member of the network using train-the-trainer principle. CSC offers data management planning support to researchers. CSC Docs and Video CSC -websites provide self-learning material. (*CSC - Good data management is a requisite for successful research*, n.d.)

CSC participates in national and international projects, collaborations and forums to exchange and contribute to development of best practices and composing of policies and guidelines that help to advance RDM. (*CSC - Our international collaborations support research*, n.d.).

Conclusions

The terms research data and research data management are briefly introduced in section 3.1. Section 3.2 and 3.3 connect data lifecycle and FAIR data principles to RDM. Dataset and metadata are briefly explained in section 3.4. Examples of services and roles that support RDM are provided in section 3.5. Summary of CSC's role and activities in national and international RDM support are presented in section 4.

Research question of this thesis, *"are research data management services and support offered by CSC, aligned with requirements and expectations for research data management as presented in literature",* was studied by conducting a literature review about RDM. Literature is reflected against RDM service and support portfolio offered by CSC. Results of this study imply that service and support offered by CSC are according to recommendations and objectives of RDM as presented in literature.

Limitations

Organizations (Finnish or non-Finnish, private or governmental connections), other than CSC are not introduced in detail. There are distinct differences between European countries how RDM support has been organized throughout national research communities. Although similarities between RDM services and support provided by CSC and those provided other similar entities in European countries reported in case studies by Bryant et al. (2023b) are easy to identify, this study only implies that CSC's special state assignment works well for RDM support in Finnish research field. It might not be applicable to national RDM support activities done in other countries.

This study does not go into details about important aspects of RDM such as ontologies, vocabularies or provenance information for reproducibility and reuse involved metadata-oriented RDM, or special requirements needed to handle sensitive data in RDM. Being a general overview on RDM and service and support offering done by CSC, large expert topics like these were seen as too complex to be just briefly introduced. They are good candidates for possible future research.

Author of this topic has at least 6 years of RDM support as developer of RDM services, such as computational workflow framework and dataset repositories. This experience might have had effect on research methodology; mapping and snowballing are done through RDM practitioner's mindset, which might not be as objective as mindset of a researcher. This does not affect the implications of the thesis, but might have limited the introduced perspectives on RDM.

Data accessibility statement

The data (bibliographic information and self-contained HTML snapshots of webpages used as references in this study) that support the findings of this study are openly available in EUDAT B2SHARE -repository (https://b2share.eudat.eu) accessible at https://doi.org/10.23728/b2share.5ecf2a081f9245f5beb8ee5926ef6bd5 and at https://hdl.handle.net/11304/209edce7-c1ce-43c3-b61e-d1f66a6e7fa1 .

References

About | www.esfri.eu. (n.d.). Retrieved June 11, 2024, from https://www.esfri.eu/about

About Findata. (n.d.). Findata. Retrieved June 11, 2024, from https://findata.fi/en/about-findata/

About—EGA European Genome-Phenome Archive. (n.d.). Retrieved June 11, 2024, from https://ega-archive.org/about/ega/

About us—Finnish Meteorological Institute. (n.d.). Retrieved June 11, 2024, from https://en.ilmatieteenlaitos.fi/about-us

Accounts and Support—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/support-and-training

Allas—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/allas

Amorim, R. C., Castro, J. A., Rocha da Silva, J., & Ribeiro, C. (2017). A comparison of research data management platforms: Architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, *16*(4), 851–862. https://doi.org/10.1007/s10209-016-0475-y

Apache Spark™—Unified Engine for large-scale data analytics. (n.d.). Retrieved June 11, 2024, from https://spark.apache.org/

AVAA – Open Data Publishing Platform. (n.d.). Fairdata. Retrieved June 11, 2024, from https://www.fairdata.fi/en/avaa/

B2ACCESS | EUDAT. (n.d.). Retrieved June 11, 2024, from https://www.eudat.eu/service-catalogue/b2access

B2DROP | EUDAT. (n.d.). Retrieved June 11, 2024, from https://www.eudat.eu/service-catalogue/b2drop

B2FIND | EUDAT. (n.d.). Retrieved June 11, 2024, from https://www.eudat.eu/service-catalogue/b2find

B2HANDLE | EUDAT. (n.d.). Retrieved June 11, 2024, from https://www.eudat.eu/service-catalogue/b2handle

B2SAFE | EUDAT. (n.d.). Retrieved June 11, 2024, from https://www.eudat.eu/service-catalogue/b2safe

B2SHARE | EUDAT. (n.d.). Retrieved June 11, 2024, from https://www.eudat.eu/service-catalogue/b2share-0

Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: Fair enough? *European Journal of Human Genetics*, *26*(7), 931–936. https://doi.org/10.1038/s41431-018-0160-0

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Bryant, R., Lavoie, B., & Malpas, C. (2023a). *The Realities of Research Data Management: A Tour of the Research Data Management (RDM) Service Space*. OCLC. https://www.oclc.org/research/publications/2017/oclcresearch-rdm-part-one-service-space-tour.html

Bryant, R., Lavoie, B., & Malpas, C. (2023b). *The Realities of Research Data Management Part Two: Scoping the University RDM Service Bundle*. OCLC. https://www.oclc.org/research/publications/2017/oclcresearch-rdm-part-two-scoping-service-bundle.html

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., & Groth, P. (2020). Dataset search: A survey. *The VLDB Journal*, *29*(1), 251–272. https://doi.org/10.1007/s00778-019-00564-x

Connected Papers | Find and explore academic papers. (n.d.). Retrieved June 10, 2024, from https://www.connectedpapers.com/

Connected Papers | About. (n.d.). Retrieved June 10, 2024, from https://www.connectedpapers.com/about

Cox, A., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science*. https://doi.org/10.1177/0961000613492542

cPouta—Services for Research—CSC Company Site. (n.d.). Retrieved June 11, 2024, from https://research.csc.fi/-/cpouta

Crossref. (n.d.). Crossref. Retrieved June 10, 2024, from https://www.crossref.org/

CSC - A secure service family for sensitive data. (n.d.). Retrieved June 11, 2024, from https://csc.fi/en/our-expertise/sensitive-data/

CSC - Customers are at the center of our work. (n.d.). Retrieved June 11, 2024, from https://csc.fi/en/about-us/customers/

CSC - EOSC Association Archives. (n.d.). Retrieved June 11, 2024, from https://csc.fi/researchinfrastructure/eosc-association/

CSC - EUDAT. (n.d.). Retrieved June 11, 2024, from https://csc.fi/researchinfrastructure/eudat/

CSC | e-learn. (n.d.). Retrieved June 11, 2024, from https://e-learn.csc.fi/

CSC - Good data management is a requisite for successful research. (n.d.). Retrieved June 11, 2024, from https://csc.fi/en/our-expertise/research-data-management/

CSC - Governance. (n.d.). Retrieved June 11, 2024, from https://csc.fi/en/about-us/governance/

CSC - Our international collaborations support research. (n.d.). Retrieved June 11, 2024, from https://csc.fi/en/about-us/international-collaboration/

CSC - Persistent Identifiers. (n.d.). Retrieved June 11, 2024, from https://www.csc.fi/palvelukuvaus/-/asset_publisher/Zf77YgC5aqym/content/persistent-identifiers

CSC - Values, vision and strategy guide everything we do. (n.d.). Retrieved June 10, 2024, from https://csc.fi/en/about-us/values-vision-and-strategy/

CSC — Tieteen tietotekniikan keskus / CSC — IT Center for Science. (n.d.). YouTube. Retrieved June 11, 2024, from https://www.youtube.com/channel/UCFv-76jNZlBFp6O9umdnyDA

Curdt, C. (2019). Supporting the Interdisciplinary, Long-Term Research Project 'Patterns in Soil-Vegetation-Atmosphere-Systems' by Data Management Services. *Data Science Journal*, *18*(1). https://doi.org/10.5334/dsj-2019-005

Data Life Cycle: Integrate. (n.d.). Retrieved June 12, 2024, from https://dataoneorg.github.io/Education/bp_step/integrate/

DataCite Schema. (n.d.). [Website]. DataCite Schema. Retrieved June 12, 2024, from https://schema.datacite.org/

Digital object identifier. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Digital_object_identifier&oldid=1226850076

Digital Preservation Service for Research Data. (n.d.). Fairdata. Retrieved June 11, 2024, from https://www.fairdata.fi/en/dps-for-research-data/

DMPTuuli. (n.d.). Retrieved June 11, 2024, from https://www.dmptuuli.fi/about_us

Docs CSC. (n.d.). Retrieved June 11, 2024, from https://docs.csc.fi/

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Ecological Metadata Language. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ecological_Metadata_Language&oldid=1175543063

EOSC Association. (n.d.). EOSC Association. Retrieved June 11, 2024, from https://eosc.eu/

ePouta—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/epouta

Etsin—Research Dataset Finder. (n.d.). Fairdata. Retrieved June 11, 2024, from https://www.fairdata.fi/en/etsin/

EUDAT B2SHARE. (n.d.). Retrieved June 11, 2024, from https://b2share.eudat.eu/

EUDAT CDI | EUDAT. (n.d.). Retrieved June 11, 2024, from https://www.eudat.eu/eudat-cdi

EUDAT Metadata Schema Documentation. (n.d.). Retrieved June 12, 2024, from https://schema.eudat.eu/

European Commission. Directorate General for Research and Innovation. (2018). *Turning FAIR into reality :final report and action plan from the European Commission expert group on FAIR data.* https://doi.org/10.2777/54599

Fairdata | Take care of your research data. (n.d.). Fairdata. Retrieved June 11, 2024, from https://www.fairdata.fi/en/

Fairdata Services—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/fairdata-services

FEGA - Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/fega

Findata - Regulations. (n.d.). Retrieved June 11, 2024, from https://findata.fi/en/services-and-instructions/regulations/

Hackman, L., Mack, P., & Ménard, H. (2024). Behind every good research there are data. What are they and their importance to forensic science. *Forensic Science International: Synergy, 8*, 100456. https://doi.org/10.1016/j.fsisyn.2024.100456

Handle System. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Handle_System&oldid=1227615954

Hey, T., Tansley, S., Tolle, K., & Gray, J. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Higher education institutions and science agencies—OKM - Ministry of Education and Culture, Finland. (n.d.). Opetus- Ja Kulttuuriministeriö. Retrieved June 12, 2024, from https://okm.fi/en/heis-and-science-agencies

Home | Research.fi. (n.d.). Retrieved June 11, 2024, from https://research.fi/en/

Home | Kansalliskirjasto. (n.d.). Retrieved June 11, 2024, from https://www.kansalliskirjasto.fi/en

Home—NeIC web. (n.d.). Retrieved June 11, 2024, from https://neic.no/

Horizon 2020—European Commission. (n.d.). Retrieved June 11, 2024, from https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en

IDA - Store and organize your research data before publishing it. (n.d.). Fairdata. Retrieved June 11, 2024, from https://www.fairdata.fi/en/ida/

Innes, B. (n.d.). *OKD.io*. Retrieved June 11, 2024, from https://okd.io/

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G.,

Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, *2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

Jones, S., Pryor, G., Whyte, A. (2013). *How to Develop Research Data Management Services - a Guide for HEIs*. *DCC How-to Guides*. Edinburgh: Digital Curation Centre.

Juhás, G., Molnár, L., Ondrisová, M., & Juhásová, A. (2017). Data, information and technology services for research and management of science. *2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 1–6. https://doi.org/10.1109/ICETA.2017.8102491

Kanza, S., & Knight, N. J. (2022). Behind every great research project is great data management. *BMC Research Notes*, *15*(1), 20. https://doi.org/10.1186/s13104-022-05908-5

Kapseli®. (n.d.). Findata. Retrieved June 11, 2024, from https://findata.fi/en/kapseli/

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. *2*.

Knopf, J. W. (2006). Doing a Literature Review. *PS: Political Science and Politics*, *39*(1), 127–132.

Language Bank | Kielipankki. (n.d.). Retrieved June 11, 2024, from https://www.kielipankki.fi/language-bank/

Language Research and Other Humanities and Social Sciences—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/language-research

Mahti—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/mahti

Metax API - Interfaces for transferring and reading metadata. (n.d.). Fairdata. Retrieved June 11, 2024, from https://www.fairdata.fi/en/metax/

METIS. (n.d.). Retrieved June 11, 2024, from https://fmi.b2share.csc.fi/

Notebooks—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/notebooks

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

Open Research Data—Finnish Meteorological Institute. (n.d.). Retrieved June 11, 2024, from https://en.ilmatieteenlaitos.fi/open-science-research-data

Overview—European Commission. (n.d.). Retrieved June 12, 2024, from https://knowledge-base.inspire.ec.europa.eu/overview_en

Persistent Identifiers for eResearch. (n.d.). Retrieved June 11, 2024, from https://www.pidconsortium.net/

Plan Data Management—Services for Research—CSC Company Site. (n.d.). Retrieved June 11, 2024, from https://research.csc.fi/data-management-planning

Posit. (n.d.). Retrieved June 11, 2024, from https://www.posit.co/

Project Jupyter. (n.d.). Retrieved June 11, 2024, from https://jupyter.org

Publication Forum. (n.d.). Publication Forum. Retrieved June 10, 2024, from https://julkaisufoorumi.fi/en/publication-forum

Puhti—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/puhti

Qvain – Research Dataset Description Tool. (n.d.). Fairdata. Retrieved June 11, 2024, from https://www.fairdata.fi/en/qvain/

Rahti—Services for Research—CSC Company Site. (n.d.). Retrieved June 11, 2024, from https://research.csc.fi/-/rahti

Redkina, N. S. (2019). Current Trends in Research Data Management. *Scientific and Technical Information Processing*, *46*(2), 53–58. https://doi.org/10.3103/S0147688219020035

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4. https://doi.org/10.1002/meet.14504701240

Research Council of Finland. (n.d.). Research Council of Finland. Retrieved June 11, 2024, from https://www.aka.fi/en/

Research Council of Finland - Research infrastructures. (n.d.). Research Council of Finland. Retrieved June 11, 2024, from https://www.aka.fi/en/research-funding/programmes-and-other-funding-schemes/research-infrastructures/

Research Data Alliance. (n.d.). Retrieved June 11, 2024, from https://www.rd-alliance.org/

Research Information Hub—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/research-information-hub

Sciences and Methods—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/sciences

SD Apply—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/sd-apply

SD connect—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/sd-connect

SD Desktop—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/-/sd-desktop

SD Services for Research—Services for Research—CSC Company Site. (n.d.). Services for Research. Retrieved June 11, 2024, from https://research.csc.fi/sensitive-data-services-for-research

Secondary use of health and social data. (n.d.). Ministry of Social Affairs and Health. Retrieved June 11, 2024, from https://stm.fi/en/secondary-use-of-health-and-social-data

Semantic Scholar | AI-Powered Research Tool. (n.d.). Retrieved June 10, 2024, from https://www.semanticscholar.org/

Semantic Scholar | Frequently Asked Questions. (n.d.). Retrieved June 10, 2024, from https://www.semanticscholar.org/faq#advantages

Sheikh, A., Malik, A., & Adnan, R. (2023). Evolution of research data management in academic libraries: A review of the literature. *Information Development*, 026666692311574. https://doi.org/10.1177/02666669231157405

Silva, F., Amorim, R. C., Castro, J. A., da Silva, J. R., & Ribeiro, C. (2016). End-to-End Research Data Management Workflows. In E. Garoufallou, I. Subirats Coll, A. Stellato, & J. Greenberg (Eds.), *Metadata and Semantics Research* (pp. 369–375). Springer International Publishing. https://doi.org/10.1007/978-3-319-49157-8_32

Sun, G., Friedrich, T., Gregory, K., & Mathiak, B. (2023). *Supporting data discovery: A meta-synthesis comparing perspectives of support specialists and researchers* (arXiv:2209.14655). arXiv. https://doi.org/10.48550/arXiv.2209.14655

Surkis, A., & Read, K. (2015). Research data management. *Journal of the Medical Library Association: JMLA*, *103*(3), 154–156. https://doi.org/10.3163/1536-5050.103.3.011

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE*, *15*(3), e0229003. https://doi.org/10.1371/journal.pone.0229003

Terra, A., Cardoso, J., & Faria, D. (2021). *Development of tools for Data Management*. https://www.semanticscholar.org/paper/Development-of-tools-for-Data-Management-Terra-Cardoso/452edeb8b264b4338bdc83d130ceeebba312de72

Tilastokeskus - FIONA. (n.d.). Statistics Finland. Retrieved June 11, 2024, from https://stat.fi/tup/tutkijapalvelut/fiona-etakayttojarjestelma_en.html

Tilastokeskus - FIONA. (n.d.). Statistics Finland. Retrieved June 11, 2024, from https://stat.fi/org/index_en.html

Uniform Resource Name. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Uniform_Resource_Name&oldid=1220954593

Video CSC. (n.d.). Retrieved June 11, 2024, from https://video.csc.fi/

What CSC?. (n.d.). Retrieved June 11, 2024, from https://csc.fi/en/about-us/what-csc/

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilms, K. L., Stieglitz, S., Ross, B., & Meske, C. (2020). A value-based perspective on supporting and hindering factors for research data management. *International Journal of Information Management*, *54*, 102174. https://doi.org/10.1016/j.ijinfomgt.2020.102174

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 1–10. https://doi.org/10.1145/2601248.2601268