The Processes Behind Research Data Management

Ville Tenhunen[1], James A.J. Wilson[2]

[1]University of Helsinki, PL28, 00014 Helsingin yliopisto, Finland, ville.tenhunen@helsinki.fi
[2]University College London, Gower St, London WC1E 6BT, United Kingdom, j.a.j.wilson@ucl.ac.uk

Keywords
Research data, research data management, RDM, research processes, life cycles, data services

1. Summary

Without understanding the need for compatibility between research data management systems and research processes, it is not possible to design and develop efficient and user-friendly data services for researchers. In this paper we describe the research data processes behind RDM lifecycles and provide an insight into the actual implementations of these services all around the world in different research systems.

2. Research data management lifecycles

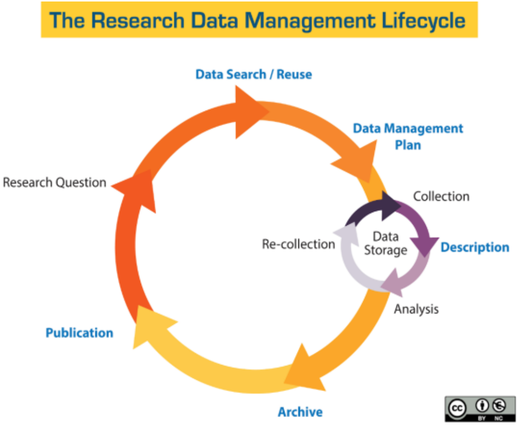


Figure 1. Example of the RDM lifecycle[1]

Research Data Management (RDM) is a combination of processes, systems, and their integration. RDM lifecycles are used to present these processes as a cycle, which starts from the idea or planning phase and ends with data reusability, long term preservation, and publication. Lifecycle models give us one way to visualize the steps involved. When real-world research data processes and architectures are presented, however, we can see that they rarely fit the idealized vision. Lifecycles hide several data use cases of researchers, and more informative and specific models are needed. Process maps help to illustrate several very different paths to move and handle data within the data lifecycle.

3. Process mapping

3.1. RDM processes in general

The multiple paths that data can take during the course of a research project presents a challenge when designing RDM infrastructure and can make life hard for service providers and decision makers. Much research is far more iterative than traditional lifecycle models let on, as data is collected, assessed, cleaned, analyzed, compared, improved, discarded, and so forth.

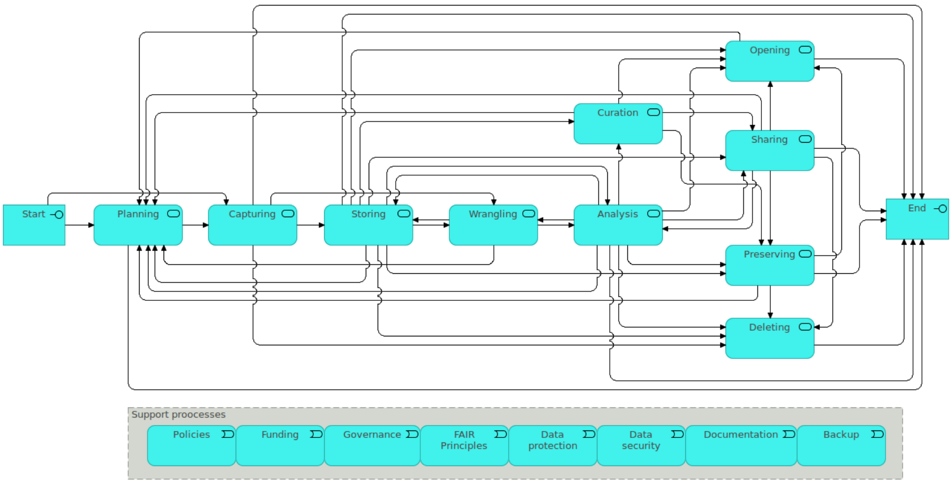


Figure 2. Research data processes and their relationships

In the context of this article a process map is a planning tool that visually describes the flow of data between RDM processes or sub-processes. In general, a process map is also called a flowchart, process flowchart, process chart, functional process chart, functional flowchart, process model, workflow diagram, etc. As a tool, such maps can also be used to define what an organisation or part of it does, who is responsible for processes, illustrate how an organisation's borders affect the processes and so forth.

The purpose of the process mapping is to assist organizations in becoming more effective.

In this article we use simplified process mappings to show the relations of the RDM sub-processes and the research data flows between them. The map tends to show the flow of data from the user's perspective and we call here these routes that data takes as "data paths".

The process map in this article is based on real world observations and experiences, but there are always organization level variations. In deeper analysis it is easy to add elements like process owners, technical services, users etc. to the process maps.

3.2. Processes in short

Figure 2 illustrates common research data processes and the paths data takes.

Planning: In this phase the researcher ideally makes a plan to cover all RDM issues. These plans are generally known as Data Management Plans (DMPs) and consist of a formal document that outlines how data are to be handled both during a research project, and after the project is completed[2]. Research funders sometimes demand these as a part of a project bid.

Capturing: Data capturing can be defined as any method of collecting information and then changing it into a form that can be read and used by a computer[3].

Storing. Data has to be saved or stored somewhere in some phase of the RDM processes. Technically there are large variations between types of storage: e.g. fast parallel storage for computing or slower, cheaper technologies for sharing data between individuals. Storage facilities have different access methods such as filesystems, structured interfaces for clients or web applications, etc. It is possible that above the storage layer of a data architecture there are applications such as databases or object storage middleware.

Wrangling: Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics[4]. Data scientists might spend a significant part of their time with data wrangling before moving to the real analysis of their data[5]. Sometimes this process is combined with the data analysis.

Analysis: This is the phase in which researchers inspect, transform, combine and model data and datasets to discover useful information to inform science, decision making or some other purpose. This is in many respects the core process in the whole RDM lifecycle.

Curation: Data curation is the organization and integration of data collected from various sources. It involves annotation, publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for access and re-use[6]. In this article the curation is connected with the preservation and especially the open publication and sharing of the data.

Sharing: Sharing the data includes work phases where the data is made available to a researcher's collaborators or anyone else. Shared data might be unfinished and not so strictly curated, as the purpose of sharing is to communicate and collaborate with someone who can bring added value to the research process or who can benefit the interpretation of the data before it is published. Researchers may share raw data with their colleagues as well as processed.

Preservation: Data preservation is the act of maintaining both the security and integrity of data. Preservation is done through formal activities that are governed by policies, regulations and strategies directed towards protecting and prolonging the existence and authenticity of data and its metadata[7]. In this article only digital preservation is discussed. Sometimes this phase is called archiving or preservation is considered an aspect of archiving.

Deleting: This phase of the process covers removing the data from storage or other infrastructures permanently and securely. Deleting is connected with data security and data protection principles. Deleting is also important for freeing resource allocations and is a part of effective workflows.

Opening: Open publishing is the most visible part of the RDM process. Idea of opening the data or open publishing makes data freely available to everyone to use and republish as they wish, without restrictions arising from copyright, patents or other mechanisms of control[8]. Sometimes opening the data covers only opening the metadata, especially where the data itself might be personal or otherwise sensitive. Even if the FAIR principles[9] cover all processes of the RDM, making data open is a key aspect of its implementation.

Support processes and principles like policies, funding, governance, the FAIR principles, data protection, data security, documentation and backup are necessary elements of RDM processes. Without these, research falls short of the standards one would expect of a reputable research institution.

In this article we focused on more operational processes of research data management and do not describe these support processes to any great extent. However, a couple of things can be said about documentation and backup.

Documentation in the research data processes is crucial. Without good documentation of the data and the processes behind its creation, confident reusability of the data is not possible to achieve. For example the components of a processed data set[10] are the raw data, tidy data set[11], a code book describing each variable and its values in the tidy data set and an explicit and exact recipe the researcher has used in steps to go from raw data to the tidy data and the code book. Metadata is a necessary part of the process but it is not only documentation needed for full reusability.

Backup is a very important part of research data processes and it has to be performed systematically and in an organized manner. That said, not all research data is worth backing up if it is large and can be produced again in the case of storage problems.

In the next sections we present some examples how to use process mappings for the analysis of the RDM processes.

3.3. Ideal situation of RDM

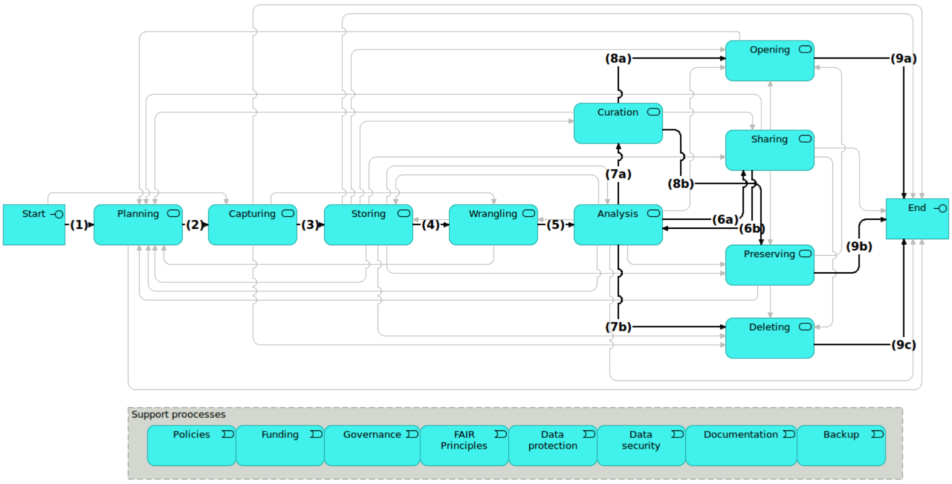


Figure 3. Ideal process map of the FAIR data

In the ideal case (Fig. 3), a researcher first makes a clear data management plan (1) which tracks how the data will be managed in every step of the whole process. Data itself arrives via the capturing process (2). A researcher could receive data from a repository, sensors or measurement equipment, a simulation, or one of various other data capture mechanisms. However it is acquired, data has to be stored (3) in some way, whether on an ordinary storage or disk server, or temporarily on a desktop or laptop.

Alongside the storage process it is possible to take care of data protection, security, and governance issues, and set up a proper back up system for the storage based also on the DMP.

The wrangling process (4) prepares the data for actual analysis. Data wrangling has been considered to be the work process that typically takes the most time[5] in the whole data management life cycle.

Properly wrangled data is then ready for analysis (5). The results of the data analysis might be shared (6a and 6b) with other researchers. Data to be retained after analysis will need to be curated (7a). which is part of the quality assurance process and involves, for example, checking that the documentation of the data is clear and sufficient.

It is common that the data analysis and wrangling processes also generate data of little value – early versions, erroneous readings, duplicates, etc. These should be deleted (7b) in a controlled way, especially if the data contain personal data or sensitive personal data. In any case deleting the useless data saves the storage and possibly researchers' time. Different versions of the same dataset are notorious for causing confusion later down the line.

The next processes facing the curated data are opening (8a) and preserving (8b). Within the opening process data should be assigned a PID[12] or PIDs such as a DOI. The data should be moved to and stored in the publication repository. In the preserving process data will be stored over potentially very long time periods, which necessitates policy-based solutions for continued funding, governance and good documentation.

By the end of the process the research data will have either been opened and preserved or deleted. It's whole journey from beginning to end will have been actively planned and managed.

3.4. Fast track research data

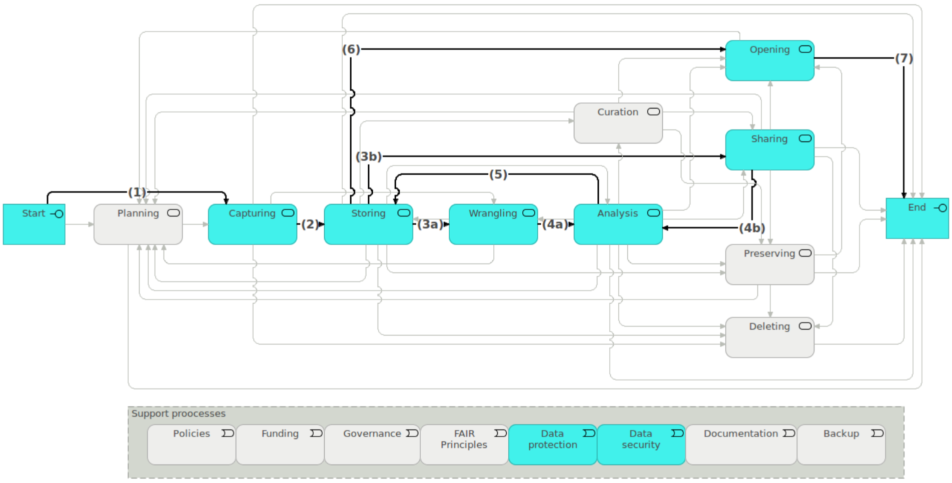


Figure 4. Process map for the simple and fast data handling

The second use case is where a researcher takes some data and moves to analysis as fast as possible, without the planning phases. For example, when a researcher knows the content of the data and methods beforehand and there is a need to use the data for further analysis quickly for an existing project.

The path starts with capturing data (1) from some repository, laboratory, measurement equipment, data stream, etc. The data is then stored in an active data store (2). The data then goes through the wrangling phase (3a), where it is prepared for analysis. Alternatively, the researcher may share the data (3b) with colleagues.

After this, the cleaned data is analysed (4a) one way or another, and the results and processed datasets are stored again (5). It is possible that data will be analysed after sharing (4b) with someone else who makes some contribution or changes to the data. The process may include some iterations between different steps (storing, wrangling, analysis, sharing) depending on the issues encountered.

Eventually the process ends with the open sharing of the data (6). Sometimes this is tied to publication and the reason for opening is to get a PID (such as a DOI) for the dataset because the (textual) publication demands it.

The main issue with this kind of workflow is monitoring and adherence to the FAIR principles. If there is not any plan or the data curation processes are not included then it may be challenging to document the data adequately or provide suitable access or preservation mechanisms.

3.5. Loops without publishing

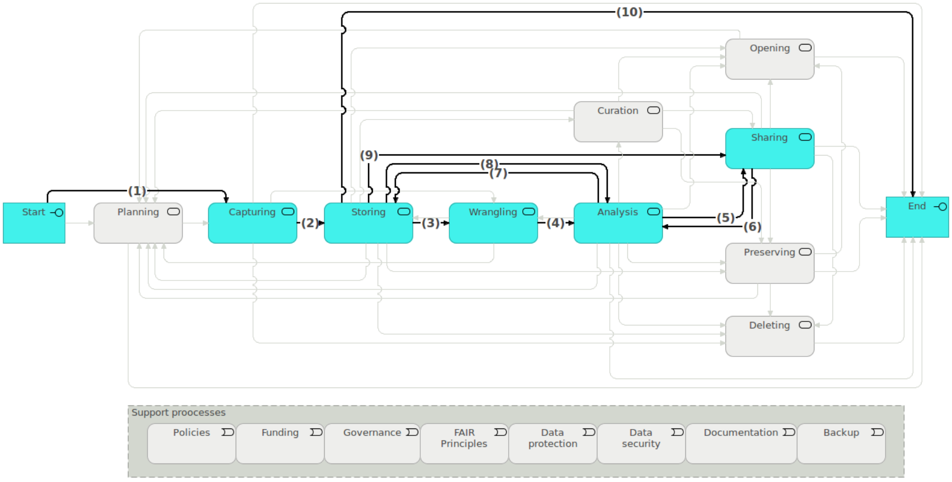


Figure 5. Sometimes data is just discussed

Not all research data processes are intended to lead to the publication or opening of data. Some research processes do not produce such results or outputs. Despite that, this kind of process might nevertheless be valuable and useful for research or science itself, such as testing ideas.

It is possible that research data management includes only a few steps but there are several loops. This kind of process starts from capturing data (1) like in the other examples. Because there is no planned data product or other set of results to be preserved, the process path does not include a coordinated planning phase.

After capturing the data, it is stored in an appropriate storage facility (2). After this the researcher does the customary data wrangling (3), and analysis (4), and possibly shares the data. (5) The data and get some feedback (6). The results of the analysis are stored (7) and possibly taken from there again to inform some new analysis (8). This kind of working data is also shareable (9). Typically this kind of process ends in a situation where different data and datasets are stored somewhere with the working documentation without any quality-controlled metadata. This is a common data path to follow when a researcher wishes to try out some new ideas.

4. Insights into the RDARI survey results

Given that researcher's workflows do not always follow the idealized Research Data Management lifecycle, what can, or should, a research institution do to support their researchers whilst also encouraging good practice?

The Research Data Architectures in Research Institutions (RDARI)[13] Interest Group of the Research Data Alliance (RDA) Survey of Institutional Research Data Services was conducted between July and November 2019.[14] The aim of the survey was to build a picture of the contemporary state of research data management service at universities and other research-conducting institutions across the world, both to assist with benchmarking and also to help put people in touch with each other so they could exchange notes about their approaches (assuming they were happy to share).

The survey covered institutional RDM governance and infrastructure as well as a number of individual services. For each service, the survey asked about the technologies used, costing and pricing models, and uptake. It also covered priorities for new service development and the growth of research data provision, in the context of lifecycles and process maps.

The specific services that were included were selected by the survey authors after consultation with RDARI Interest Group members at successive RDA Plenaries. Whilst the way people referred to the selected services at different institutions varied, they were chosen as they were common across multiple institutions. The services were defined in the survey as follows:

1. Research Data Storage Service (defined as “a centrally-managed research data storage service for use by researchers during their research projects”)
2. Research Data Repository (“a research data archive or repository that provides your researchers with a place where their data can be stored over the long term and which maintains a catalogue of records describing the data holdings”)
3. Research Data Management Advisory Service (“a research data management advisory service, whether [presented] as a dedicated service or as part of a larger researcher support service”)
4. Data Back-up Service (“a research data back-up services, taking a back-up copy of the data [the researchers] generate and store which they could retrieve in the event of data loss”)
5. Data Management Planning Tool (“a Data Management Planning (DMP) support or templating service based on software? (such as, but not limited to, DMP Online or DMP Tool”)
6. Research Database Hosting Service (“a research database hosting service where researchers can host and serve SQL or other types of active database (i.e. not simply flat files”)
7. Electronic Lab Notebook (ELN) Service (“a centrally-supported Electronic Lab Notebook service, or something similar intended as a place where researcher can record their day-to-day research activities”)
8. ‘Data Safe Haven’ (“...in which sensitive data can be securely stored and worked on / analysed during the active phase of a research project”)
9. Research Data Dark Archive (“for the long-term preservation of sensitive data that is not accessible via the Internet”)
10. File synchronization Service (“a local/remote file synchronization service (such as provided by DropBox for example)”)
11. Special Data Collections Showcase (“in which data considered to be of particular importance or interest can be displayed to others as an example, possibly with advanced visualization tools”)
12. Other RDM Services (“any other research data management services that you would like to mention in this survey which might be of interest to people at other institutions”)

That the selected services were indeed common across institutions is supported by the fact that only 16% of the 82 respondents who completed the survey indicated that their institution offered one or more ‘other’ RDM services (and the free text responses indicate that many of these could legitimately be considered to fall within the existing categories).

It is worth illustrating how the services covered map to the process diagrams in section 3 above.

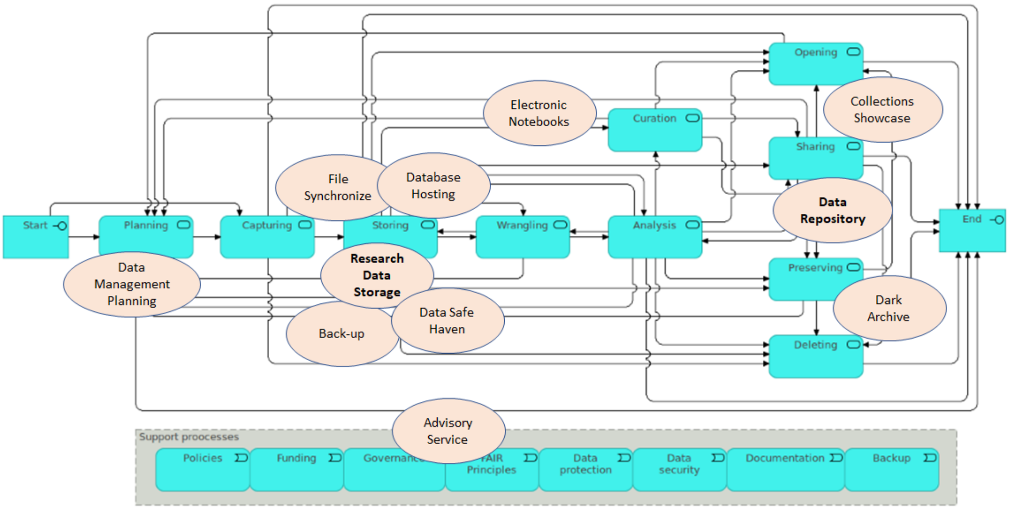


Figure 6: RDM Services loosely coupled to ideal FAIR process map

As can be seen, there are gaps in services around the capture, wrangling, and analysis of data – understandably so given that it is these steps in the research process that are most discipline-specific and therefore hardest to support centrally in a multi-disciplinary institution. They are also where most of the intellectual work of research takes place. As the process diagrams show, whilst researchers can and often do skip the ‘peripheral’ steps of research, such as curation and data publication (opening), the core steps of data capture, storing, wrangling, and analysis are fundamental.

If an institution is going to reach its researchers and intervene in their research processes in a way that can begin to encourage engagement with the wider principles of good data management then the obvious place to start is with storage. This is the only step common to all of the research process diagrams in section 3 that can realistically be addressed by providing infrastructure that is of use across all disciplines. All researchers need to store their data, and if an institutional storage offering is as good as alternative storage options, and cheaper to use, then there is little reason not to use it. 56% of the institutions that responded to the survey already offered a research data storage service, with an additional 24% reporting that such a service was in development. Of the RDM services that research institutions have adopted, research data storage is also one of the ones with the highest take-up.

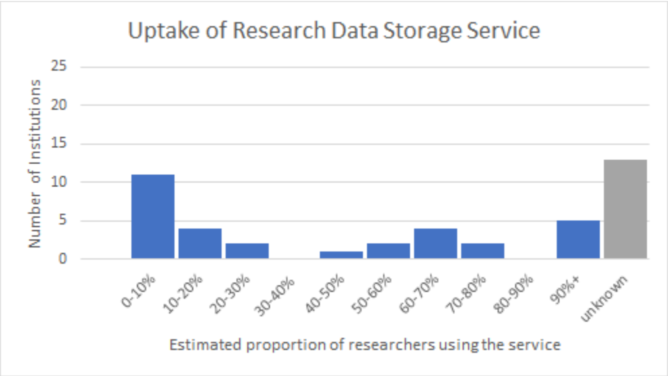


Figure 7. Uptake of institutional research data storage infrastructure

As with all RDM services, take up is influenced by a number of factors, including the time that a service has existed for, and how successfully it has been promoted. Although the data is limited, there is some suggestion that the institutions with the highest levels of uptake are those who do not charge their researchers for using the storage. Whether this is a sustainable model is up to the individual institution, although in some countries the costs of managed storage may be covered by research funders.

The breakdown of other services that institutions already operate, along with those that they are in the process of implementing, are as follows:

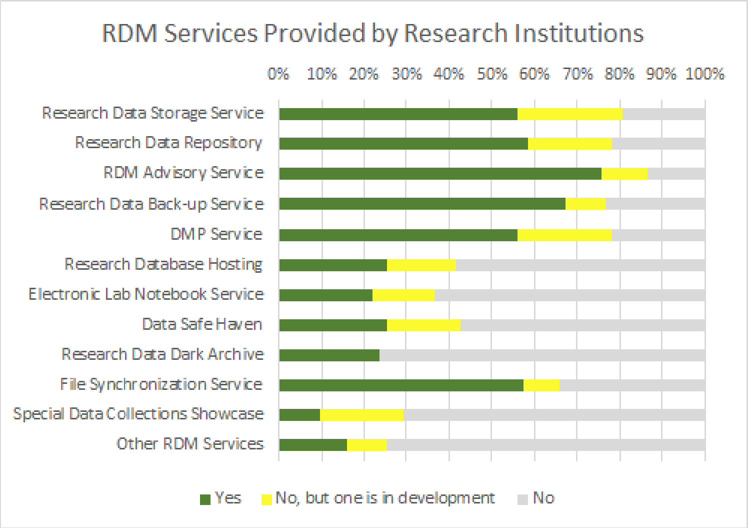
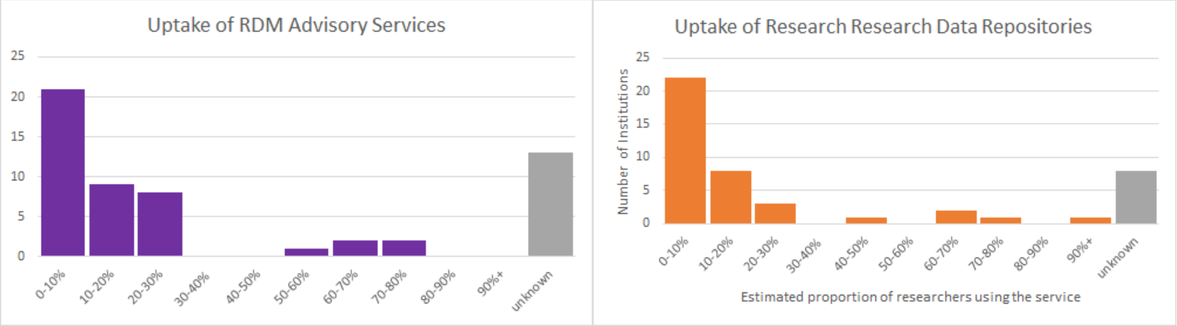


Figure 8. Provisions of common RDM services by research institutions

As is evident from Figure 8, not all institutions have begun their provision of RDM services with technical infrastructure. The most common service at present is an advisory service, probably because such a service can be established with little capital investment and often added to the responsibilities of existing staff, particularly those in the library. The provision of personalized advice and training does not necessarily scale as easily as technical infrastructure, however. Plus it relies on researchers engaging with processes that go beyond the fundamentals of their research and towards the FAIR data model – and if researchers do not feel a need to make their data open, then they are unlikely to seek advice. This is reflected in advisory service uptake levels and emphasizes the need for advocacy as well as merely service provision.



Figures 9 & 10. Uptake of RDM Advisory Services and Research Data Repositories

Other commonly provided services such as research data repositories are also still showing relatively low rates of uptake, as they are also dependent on researchers bring convinced to move from non-data-publishing loops to FAIR Data practices by policies and external pressure rather than by beneficial service interventions that improve their typical current research routines.

Besides advocacy and stringent policy, one way to persuade more researchers to adopt more open workflows might be to better link together the different data management services that institutions can provide, starting from the data storage that all research requires. If completing a data management plan were to trigger the creation of a project space within the storage infrastructure, with correct permissions and appropriate capacity, and it were easy to select datasets from this storage and move them to a repository, with metadata being added and enriched as data is processed rather than all at the end, then it would better map to desired research workflows. When each step of the process is separate and manual, it requires more hand-holding and persuasion to get a researcher to take their data through the full process required to make it FAIR, and it requires more of the researchers' precious time. Other RDM services could be integrated into the lifecycle over time, to further enhance the findability, accessibility, interoperability, and reusability of the data.

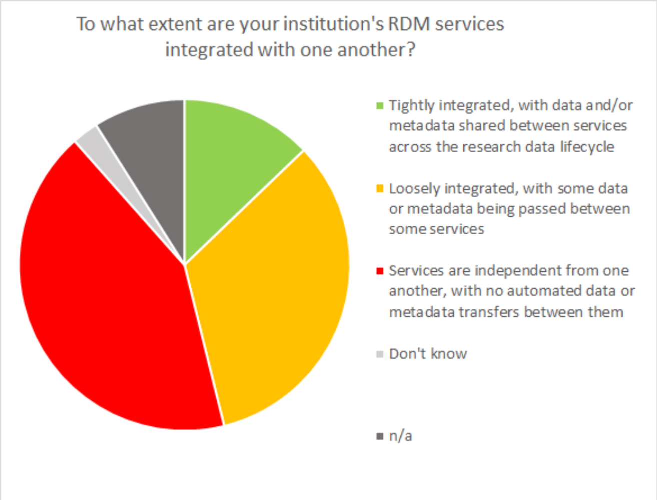


Figure 11: Integration of RDM services

At the time of the survey, fewer than 15% of the institutions who responded to the RDARI survey claimed that they supported a tightly-integrated suite of research data services, although a further third indicated that their services were more loosely integrated. This is too small a sample at present upon which to base any conclusions regarding the effectiveness of a better-integrated suite of services, and there are too many other factors relating to the uptake of services that would need to be accounted for, but it is a principle to consider when architecting institutional RDM infrastructure and services, and one which might fruitfully be further explored over the coming years.

5. Conclusions

Usual research data management models give a general picture of the landscape. For the service development more specific tools for designing user driven data services are needed. Process mapping described in this article and detailed integration description are potential option. Method might help to solve challenges when all research data processes are not suited to the ideal model but they might be useful and valuable for research and science. RDARI survey gives real world data to describe services and their integrations in the research institutions. It also indicates there are gaps in services around the capture, wrangling, and analysis of data. It is understandable but noteworthy because these are important phases for the scientific results.

6. References

1. UC Santa Cruz University Library website. Retrieved 1 March 2020, from https://guides.library.ucsc.edu/datamanagement
2. Wikipedia. Retrieved 15 May 2020, from https://en.wikipedia.org/wiki/Data_management_plan
3. Cambridge Dictionary. Retrieved 15 May 2020, from https://dictionary.cambridge.org/dictionary/english/data-capture
4. Wikipedia. Retrieved 15 May, from https://en.wikipedia.org/wiki/Data_wrangling
5. Armand Ruiz, The 80/20 data science dilemma, Inforworld. Retrieved 15 May 2020, from https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html
6. Wikipedia. Retrieved 15 May 2020, from https://en.wikipedia.org/wiki/Data_curation
7. Wikipedia. Retrieved 15 May 2020, from https://en.wikipedia.org/wiki/Data_preservation
8. Wikipedia. Retrieved 15 May 2020, from https://en.wikipedia.org/wiki/Open_data
9. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. Retrieved 3 May 2020, from https://doi.org/10.1038/sdata.2016.18
10. Jeff Leek, Elements of the Data Analytic Style, p. 10. Retrieved 15 May 2020, from https://leanpub.com/datastyle
11. Wikipedia. Retrieved 15 May 2020, from https://en.wikipedia.org/wiki/Tidy_data
12. Juha Hakala, Persistent identifiers - an overview (2010), Retrieved 15 may 2020, from http://www.persid.org/downloads/PI-intro-2010-09-22.pdf
13. Research Data Architectures in Research Institutions IG. Retrieved 15 May 2020, from https://www.rd-alliance.org/groups/research-data-architectures-research-institutions-ig
14. Wilson, James A J; Tenhunen, Ville; Russell, Keith (2020): RDARI International Survey of Institutional Research Data Services 2019. Figshare. Dataset. https://doi.org/10.5522/04/10283540.v1

7. Authors' biographies

M. Sc. Ville Tenhunen has worked as a team leader and project manager in the University of Helsinki more than 12 years. Last major projects has dealt with research data and its storages. He has been also active in Finnish national open science and research Initiatives. Tenhunen has been also co-chair of the Research Data Architectures in Research Institutions IG of the Research Data Alliance (RDA). He is also member of the Architecture Working Group of the EOSC. Beginning of March 2020 he has worked in the EGI Foundation as Data Solutions Architect.

Dr James A J Wilson has over ten years' experience in research data management and is currently Head of Research Data Services at UCL. He is the service manager of UCL's Research Data Storage Service and the Institutional Research Data Repository. He has formerly developed a database hosting facility and worked with a number of research teams on their research data workflows and documentation. His academic background is in the humanities. Like Ville, he is a co-chair of the Research Data Architectures in Research Institutions IG of the RDA.