Research data repository requirements: A case study from universities in North Macedonia

Fidan Limani*, Arben Hajra**, Mexhid Ferati***, Vladimir Radevski****

ARTICLE INFO

Article history:
Received 14 January 2022
Revised 22 February 2022
Accepted 26 February 2022

Keywords:
Research Data
Institutional Repository
Repository Requirements
Research Data Management

ABSTRACT

With research data generation on the rise, Institutional Repositories (IR) are one of the tools to manage it. However, the variety of data practices across institutions, domains, communities, etc., often requires dedicated studies in order to identify the research data management (RDM) requirements and mapping them to IR features to support them. In this study, we investigated the data practices for a few national universities in North Macedonia, including 110 participants from different departments. The methodology we adopted to this end enabled us to derive some of the key RDM requirements for a variety of data-related activities. Finally, we mapped these requirements to 6 features that our participants asked for in an IR solution: (1) create (meta)data and documentation, (2) distribute, share, and promote data, (3) provide access control, (4) store, (5) backup, and (6) archive. This list of IR features could prove useful for any university that has not yet established an IR solution.

* Research Assistant, ZBW Leibniz Information Center for Economics, Kiel, Germany (f.limani@zbw.eu) (First Author) (Corresponding Author)
** Research Assistant, ZBW Leibniz Information Center for Economics, Kiel, Germany (a.hajra@zbw.eu) (Co-Author)
*** Associate Professor, Department of Informatics, Faculty of Technology at Linnaeus University (mexhid.ferati@lnu.se) (Co-Author)
**** Professor, South East European University, North Macedonia (v.radevski@seeu.edu.mk) (Co-Author)

International Journal of Knowledge Content Development & Technology, 13(1): 75-100, 2023.
http://dx.doi.org/10.5865/IJKCT.2023.13.1.075

1. Introduction

Research undertakings typically generate multiple deliverables, such as research data (RD), executable scripts, source code, workflow models, etc. Their presence in scholarly communication enables a better understanding of a research endeavor, and is important for research reuse, verification and reproducibility. Once available, research deliverables can support many operations: they enable the verification of their quality and principles, can be re-run to reproduce the results reported in the original study, can be reused in new studies, and so on. As a result, focusing the research dissemination on publications alone is often insufficient to capture and communicate all the aspects of a research. Moreover, the ability for an exhaustive research consideration is in line with the emergence of Open Science practice in research (the initiatives of the European Open Science Cloud initiative[1] for a European and the National Open Science Cloud Initiative[2] are two examples of promoting this on a European and national level, correspondingly), which advocates for "transparent and accessible knowledge" in research (Vicente-Saez & Martínez-Fuentes, 2018).

One such deliverable that is gaining importance in scholarly communication are research data (RD). Even though practices around them differ across research communities (for e.g., computationally-driven fields lead in RD generation and reuse; see Hey (2016) for such example), we are witnessing an increase in their role in the research (communication) practice, for both providers (it is being created and managed, and ultimately published more) and consumers (it is cited, reused, etc.). The multiple publication venues and formats that target RD, be it the data paper, RD published alongside publications, or RD "packaged" with other research artifacts (source code, publication, funding details, etc.) testify to an interest for them in practice. This development provides RD with an appropriate recognition − commonly attributed mainly to research publications − which obliges researchers to put adequate effort to manage it.

Different factors shape the evolving RD research practices across disciplines. Publishers, driven from more research transparency and reuse, increasingly require that researchers provide the accompanying RD of the research publications (Lin and Strasser, 2014). On the other hand, initiatives like DataCite[3] ("Welcome to DataCite", n.d.) provide the necessary means (metadata standard and services) to help capitalize the effort that researchers spend in RD generation and curation. RD citation and usage, or impact metrics, present additional incentives to researchers for sharing RD, whereas, (public) funding agencies, another important actor in the research ecosystem, have accommodated new RD requirements for grant holders, which mandate research outputs − and some specifically refer to RD − to be publicly available. In this way, for the Horizon 2020 projects, the European Commission requires a certain level of alignment with a more open access-type form of RD publication. The Research Data Pilot project (H2020 Programme, 2017) is one such example, which, among other things, emphasizes the benefits of open access to publications and data. Namely, under such circumstances, the expectation is that the duplication of the same or similar RD decreases.

1) https://eosc-portal.eu/
2) https://www.nosci.mk/
3) https://datacite.org/

In this context, research communities need the necessary means to handle RD archiving and dissemination. It is clear that an infrastructure – repository at an institutional level – is required in order to support these research expectations, while considering the lifecycle and specific domain practices in the community. Lynch (2003), for example, supports the role of such repositories in scholarly infrastructures, especially in universities.

When managing RD, interested entities can choose from a rich set of repository solutions. In terms of scope, we distinguish several types of such repositories, such as project-specific, discipline-specific, or institutional repositories (IR) (Uzwyshyn, 2018). Being that we focus on a university case study, we choose the latter repository type as the most suitable one for this work. Defining IRs depends on the context, research practices, domain, and so on. For the purpose of this paper, we adopt the definition from Luther (2018), as an entity that "contains digital materials created by the institution and its community members". In the university context, the digital materials targeted for preservation and dissemination include research outcomes that stem from or are used in teaching and research.

It is clear that universities, as part of the research community, regularly deal with RD. Although to different extents, this requires RD management (RDM) support. The idea of an Institutional (data) Repository (IR), for any type of resource - be it publications, research data, scientific workflows, etc. - is enticing and generally accepted as advantageous to individual researchers and institutions alike. Additionally, the number of such repositories by country (Royster, 2019) or those listed in the data repository registry of re3data (Registry of Research Data Repositories) attest to their adoption. The level of adoption of repositories is another relevant indicator in this context. However, having into view the different research practices of these communities, there is a need to specify their RDM requirements and, in turn, map them to the features of an RD repository solution to be adopted.

Considering that, the adoption of RD repositories in university settings is still a novel undertaking, the best approach is yet to be found. This is especially true for young universities. Authors have already published a study (Limani et al., 2020) showing initial requirements for RD repository adoption at university setting. In this paper, we have expanded our data collection to more universities to get a broader understanding on a national level.

In this case study, we survey university faculty and investigate the extent to which RD and RD-related activities are present at the university. Finally, we rely on their feedback to develop a set of RDM requirements and map them to features that an RD repository should support for the targeted institutional context.

2. Related Work

Many aspects of institutional repositories (IR) adoption, in a broader context, have attracted considerable attention in the research community. Namely, cases that focus on the motivations for IR adoption - be it added or perceived value, targeted services, or specific IR functional requirements – are relevant for our research as they ultimately reflect the choice of features in an IR. In this section, we draw upon the literature for these different aspects.

RD repositories of all types are already well-accepted in the scholarly infrastructure, and the

motivation for this varies across domains and research communities. The survey from Asadi et al. (2019) includes 115 publications on the topic of IRs and lists the benefits, challenges, and motivations that universities and individual researchers seek in their deployment. Research dissemination, archiving of research deliverables and reputation increase of the institution and the individual research(ers) are the most cited reasons for adopting an IR in the institutional context. Furthermore, Kipnis et al. (2019) report on IR adoption trends based on the feedback from health sciences libraries, where 70% of participants already use or are in the process of deploying an IR for their needs. The survey also features different aspects of such an undertaking, including research culture (over 57% of participants do not consider an open access policy, for example), technical solutions, required repository features, and so on.

The services an IR supports are another source of requirements determination. Akers and Doty (2013) conducted a survey to show the differences in research (data) management practices among respondents from 4 different research domains. The survey was organized with 6 different categories, including aspects such as data storage and back-up, data sharing, and data preservation, that enabled authors to elicit 9 RD management-related services to support participants' RD management tasks. Similarly, Carnegie Mellon University, during its selection and deployment of an IR, identified five categories to classify the tools and services required to support its staff: discover, organize, create, share, and impact (Scherer & Valen, 2019). These tools and services were provided by different vendors - all supporting different services, which is important to point out another way to organize an IR. Akers & Green (2014) report about another university institution, the Michigan University Library, which, in cooperation with Dryad[4], provides RD management services, such as deposit, access, and share RD. Besides the more common services, this also includes services such as DOI assignment and data-level metrics (page views, downloads, etc.).

When it comes to specific IR features for RD management, the Repository Platform for Research Data interest group is dedicated to providing recommendations for improving repository solutions since 2015 ("Repository Platforms for Research Data IG", n.d.). Bringing together different repository stakeholders, it collects use cases and derives corresponding functional requirements - including their importance - that RD repository solutions should consider. In its latest deliverables, the requirements, 44 in total, are organized based on 13 categories, such as authentication, data access, integration, metadata, etc. In a similar feat, Kim (2018) considered the characteristics of RD, the available RD repository solutions, as well as the feedback from a targeted research community, and identified 75 functional requirements, grouped over 13 categories, which provide a good insight for what such a repository should support. Seeing the importance of Data Management Plans (DMP) and trustworthy data repositories for researchers (based on CoreTrustSeal[5]), Kim (2020) also identified corresponding functional requirements for data repositories. In terms of data discovery, based on 79 use cases collected, Wu et al. (2019) recommended 9 features that users see as beneficial for data repositories to implement. In implementing an institutional data repository for the Institute of Science and Technology in Austria, Petritsch (2017) conducted a survey with 19 participants

4) https://datadryad.org/
5) https://www.rd-alliance.org/groups/rdawds-certification-digital-repositories-ig.html

(each representing a research group in this organization), and derived both mandatory and optional requirements for the IR to be adopted. On the other hand, the University of Bath used user stories to elicit the requirements for their IR (Research360, 2013).

IRs are being adopted across different domains. Franke et al. (2013) synthesized a main use case to derive IR requirements based on the institution's projects involving research imaging from the medical domain, as well as the input from its staff - both researchers and IT specialists. This resulted with 8 categories of requirements, such as data import, export, and search. Gray (2009) tackled the IR requirements for multimedia resources from arts researchers. His approach combined quantitative (a survey) and qualitative (one-to-one interviews) techniques, and the resulting requirements were implemented in a prototype repository, which was used to engage potential users and elicit new requirements. Requirements about the visual aspect of the IR, copyright, and video and audio support were some of the cases that differ from the IR requirements we reviewed so far, and understandably so considering the domain.

On the qualitative side of IR features' spectrum, one of the goals of the FAIRSharing project is to propose a set of criteria − important to publishers and journals − that repositories can adopt in order to enable an easier navigation for researchers that want to publish their artifacts (McQuilton et al., 2020). Additionally, the CoreTrustSeal, groups within the Research Data Alliance[6], or the Confederation for Open Access Repositories (COAR, 2008), represent but few of the initiatives that focus on adding quality-related criteria for repositories. These criteria can be reflected as features in data repositories as they represent requirements that researchers (could) rely on during their research and publication activities.

IR deployment practices are already well represented, including functional, user experience (UX), technical, and other relevant aspects. However, what the IR deployment examples show is that, in such projects, taking an incremental (i.e., start small) approach, driven by the target community, is of great importance. This is precisely the path we take with our target group. Namely, in order to study the RDM practices and ultimately map them to IR features, we target the academic staff members of few national universities in North Macedonia.

3. Methodology

In this section, we provide the design process for the methodology and the rationale behind it, as well as the setup for the survey design and administration. We believe these two aspects would provide enough context to understand the research approach for this work.

3.1 Methodology design

The focus of this work was the RDM user practices and requirements at higher-education institutions in North Macedonia and their matching to IR services or features. Reflecting this in our methodology,

6) https://www.rd-alliance.org/groups/rdawds-certification-digital-repositories-ig.html

we adopted the survey as the approach of choice. To address this focus, we needed to reach a large number of participants, and gather information both in breadth (reach different institutions), and depth (elicit enough details to understand their practices). We designed the questionnaire with this in mind.

As an additional input, we considered existing recommendations about IR services and features from relevant bodies (RDA, for example) and few of the works cited in the "Related work", as a means to frame the survey. This was in terms of both questions and the available options to those questions that participants chose from (for the close-ended questions). In our case, where a preliminary study of the RDM practices in universities in North Macedonia was not available, compiling the questionnaire with this input proved valuable. **Figure 1** shows these different, at times complementary considerations when specifying our methodology, which proved suitable to our target group's specifics.

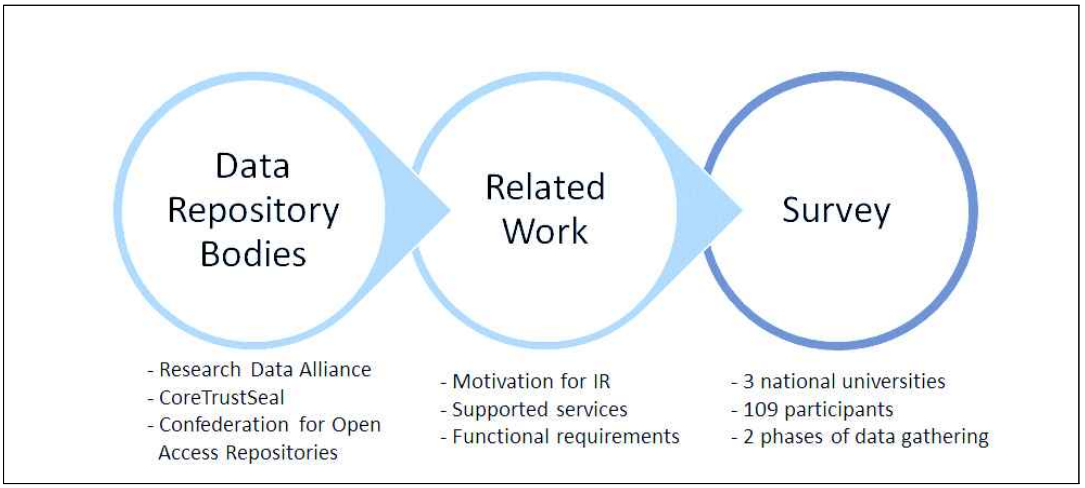

Fig. 1. Methodology components

This methodology clarifies the study context - both based on previous work and past and ongoing initiatives - for IR and RDM practices, and the means to address such requirements for the universities in North Macedonia.

3.2 Survey design and administration

For the survey, we prepared a questionnaire consisting of 20 questions and shared it with the university teaching and research staff, ensuring we covered all the departments. For it, we mainly relied on closed-ended questions, with few exceptions where participants, by choosing the predefined option ("Other"), could provide a free-text answer. These exceptions are especially important for cases where providing a final set of options to choose from is not practical or exhaustive.

The questionnaire consisted of five main sections (see **Table 1**). In the first section the participant demographic was gathered, such as their academic engagement experience, role (teaching, research, or both), department they belong to, etc. The next three sections of the survey aimed at understanding

user practices in relation to different RD lifecycle phases, such as data creation, preservation, and access, correspondingly, whereas the last section focused on participants' understanding of and feedback (and preferences) concerning the features that an IR should provide. In the final section, we also asked participants to (optionally) provide their email, in case they are willing to participate in a qualitative information gathering follow up in the future. In addition to the main survey sections, we also included questions that provided information to introduce the survey topics, gather participants' email addresses, as well as explain the survey aims, along with a request to get the participation consent. The survey was anonymous and confidential, meaning that their identity is not known, unless they opted to be available for a follow up via email. In order to validate our survey, we took into account the work from Guindon & Dennie (2016) and that of Hsu (2016) regarding researcher practice and repository functionalities to finalize our survey.

Table 1. Questionnaire: Categories and questions (Limani et al., 2020)

General Information	
1	Faculty / Department
2	How long have you been working at a higher education or research institution?
3	Your job description requires
Research Data Creation	
4	During your research, do you create/collect research data?
5	How often do you create/collect research data?
6	What is the format of the research data you typically create/collect?*
Research Data Preservation	
7	Who manages your research data?*
8	Do you describe your created/collected research data? This could include providing title, author, publication date, subject discipline, etc.
9	How much description do you provide for your research data?*
10	Where do you store your research data?*
11	How long do you keep your data after the completion of your research project?
12	How do you backup your research data?
Research Data Access	
13	Do you share your research data? (With colleagues, publish it for a broader audience, etc.)
14	When do you typically share the research data?*
15	How do you share it?
16	Is access control important when sharing your research data?

	Research Data Repository Requirements
17	Are you aware of any research data repositories in your area of research?
18	If Yes, please list them
19	Is a research data repository a good support for your research activities?
20	Is access control important when sharing your research data?

Our target group included university staff. In the context of administering the survey, we had to account for a time when they would be relatively free, such as outside of lecture, exam or grading sessions. Our study included two such phases, and for both phases we distributed the survey at the end of the semester, before their summer break, or before their semester started (with lectures), which makes for ideal periods to engage the academic staff in activities such as the survey. In any case, we were aware of the potentially low number of responses typical of surveys (as witnessed during our few distribution attempts throughout the process). Moreover, a lot of part-time research staff are usually harder to reach in such cases, depending on how often they check the official university email account, the period of their engagement at the university (certain semester, for example), and so on.

When collecting practices from one university to that of the national level, we wanted to take an incremental approach. As we experienced with our previous work with a single institution, an incremental approach helps conduct the right message and collect the right feedback, especially since the topic of RDM might be relatively new for them. For this reason, we selected three national universities − potentially with higher chances of being engaged with RD − to survey about RDM practices and their mapping to features of an IR.

The data collection was conducted twice by sending an email invitation to university staff to take part in the survey administered in Google Forms. The first instance was during March - April of 2019 when we collected data from the South East European University. Details of this process are documented in a previous study (Limani et al, 2020). The second instance was during August - November of 2019 when we collected data from two other universities in North Macedonia, namely, Ss. Cyril and Methodius University (UKIM) and Goce Delcev University (GDU). For the former, our responders were exclusively from the Faculty of Computer Science and Engineering (FINKI). For the purposes of completing the study, we relied on fast responsiveness of this department and mostly used this data for their descriptive significance. This university, however, is the biggest one in the country, as well as the largest producer of research data across domains, such as technical, social, medical, and so on. In both instances, participants included teaching and research staff as well as full- and part-time engagements.

Table 2. Characteristics of survey respondents

Teaching / Research Experience				Type of Engagement		
University	Respondents	2 to 5 years	6 to 10 years	More than 10 years	Research only	Teaching only
SEEU	43	1	2	40	36	6
UKIM	8	1	3	4	8	0
UGD	59	4	17	38	56	3
Total	110	6	22	82	100	9

SEEU – South East European University; UKIM – Ss. Cyril and Methodius University; GDU – Goce Delcev University

Reminder emails are a common practice when conducting surveys, as they help participants not lose sight of the survey invitation, thus improving its completion rate. In this perspective, our survey was no different, and we sent a few email reminders to participants about the survey. After the last such reminder, the response rate for each university was as follows: SEEU - 45%, UKIM - 11%, and GDU - 23%. **Table 2** shows the participants' teaching or research experience, and their type of engagement at the university for the three institutions. Considering that our intent with this study was not to compare between different universities, but learn from their practices, we maintained that such discrepancies in the number of responders among universities is irrelevant to the study objectives.

It is worth noting that the surveyed institutions are at relatively different stages between and across them. While some of them have established (public) IR for research publications, others have it restricted to institutional use. In any case, this is a good indication of established standards for research publications management, which gives us confidence in the usefulness of our study focusing on IR features for research data management. Finally, knowing the local circumstances, we are aware that a lot of RD is produced in specific fields of research, such as medicine, biology, chemistry from one side, and (other types of data) social sciences, ethnography, archeology, history, music, etc., from the other side.

4. Analysis and Findings

The analysis of the survey started by downloading the two datasets with raw responses from Google Forms. The dataset from the first phase of the data collection process is that from SEEU, which contained 43 records. The dataset from the second phase involved two other universities, UKIM and UGD, which contained 67 records. These two datasets were combined and the result was a new dataset with 110 records.

The data analysis was mainly of a descriptive nature, aiming to carefully examine each question in order to discover any connection between the participants' answers. Early in the analysis process,

we noticed common aspects between the RD practices among the participants from different universities. We organized these findings in six themes meant to communicate and discuss them better. We provide these themes next in subsections 4.1 to 4.6. Before presenting the themes, please recall that since multiple answers were allowed, thus the totals larger than 110 (the total respondent count) for some of the answers.

4.1 Data storing formats

One question on the survey asked participants to select the kind of format they use to save their research data. Participants could select more than one format. We grouped participants by their field of research in order to see any differences between various fields. We relied on the FRASCATI[7] categorisation provided by the Organization for Economic Co-operation and Development (OECD). FRASCATI covers six categories, namely: Natural Sciences, Engineering and Technology, Medical and Health Sciences, Agricultural Sciences, Social Sciences and Humanities. Our participants belonged to all areas except Engineering and Technology.

In **Figure 2** we can see that Natural Science researchers save their research data mostly in spreadsheet formats (Excel, OpenOffice Calc, and Mac iWork), whereas researchers from Social Science save their data mostly as Text Documents (MS Word, OpenOffice Documents, and PDF). Detailed values for other fields (Medical and Health Sciences; Agricultural Sciences; and Humanities) and formats are shown in the figure. The group at the right end labeled as Unknown are respondents who did not provide their field of research or department in the questionnaire.

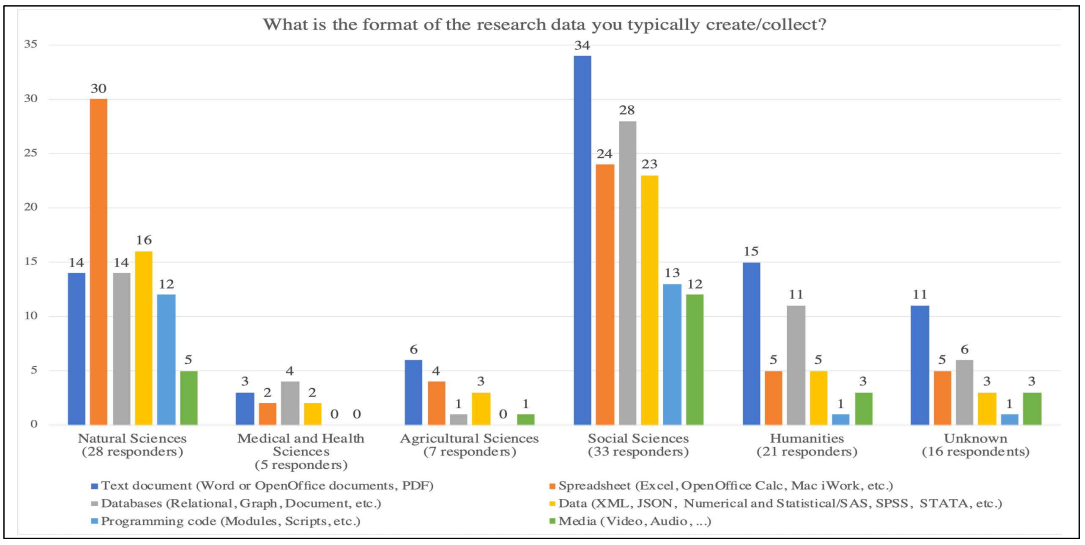

Fig. 2. Research Data storing format by research field.

7) https://www.oecd.org/science/inno/38235147.pdf

A careful observation of the graph shows that text documents and spreadsheets are the two most used formats across all research areas, with 83 and 70 respective counts. The least used are programming code and media with 27 and 24 respective counts. Somewhere in between lie databases and RD with 64 and 52 respective counts.

4.2 Data storage and backup practices

With two of the questions, we were interested to learn where researchers save and back up their research data. As seen in **Figures 3 and 4**, most of the participants prefer to store and backup their research data on their computers including portable storage and flash drive. Essentially, participants exhibit similar patterns - use the same devices - to store and backup their research data. Their second choice is cloud storage, which includes Dropbox, Google Drive, MS FileShare, etc.; the third is Research Repositories, whereas the final choice represents the use of department or laboratory servers.

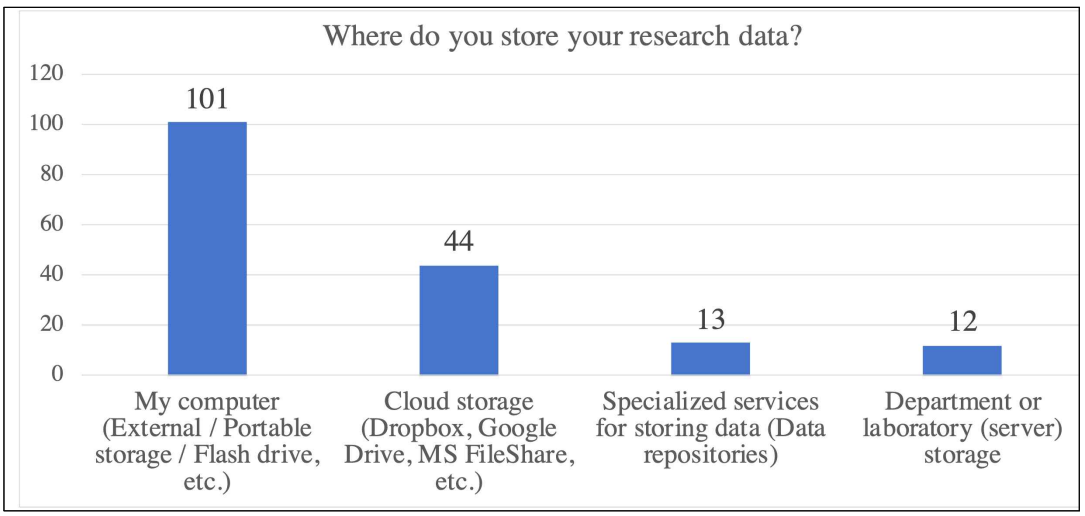

Fig. 3. Location preference when storing research data.

We see the same patterns for both storage and backup, that is, users demonstrate the same choices in terms of the medium used to store and backup RD. Another interesting observation is that of how many of the participants adopt each practice (storage and backup). It may come as a surprise that more of them (179) backup their RD than they store (171), but this could be due to the multiple choices available for these questions. In this case, chances are that a user uses a single (with exceptions) medium to store, and more than one (at least in few of the cases) to back up the RD.

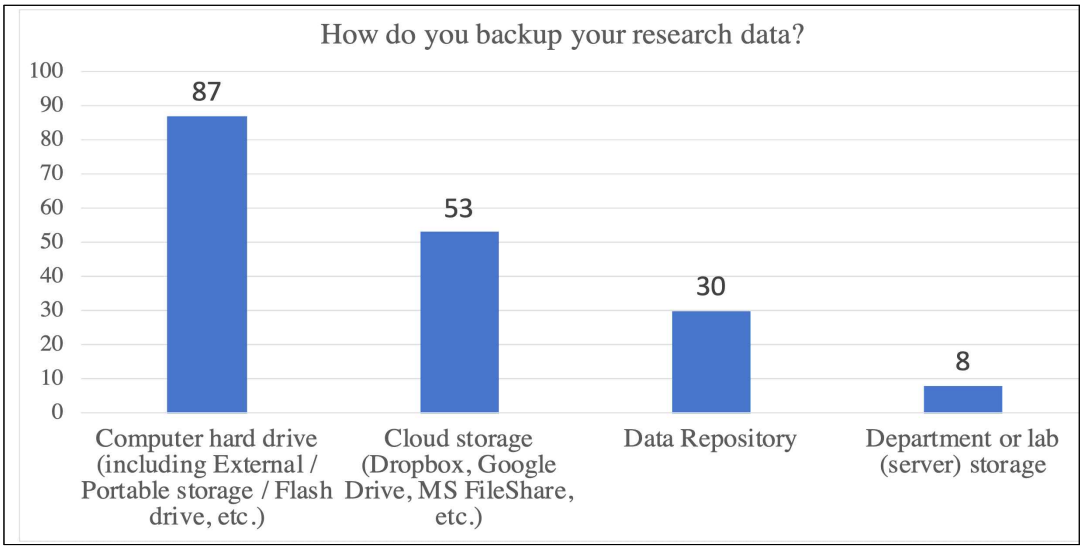

Fig. 4. Location preference when backing up research data.

4.3 Data keeping policies

Most participants indicate that they do not have a regular period of collecting research data, but they do it on a project basis. Some participants collect data monthly, some yearly, whereas a small number does it weekly (**Figure 5**). On the other hand, most participants, as shown in **Figure 6**, do not have a strategy on data storage/archiving after their research activity or project completes. A large number of participants prefer to keep that data three-to-five years after project completion. Similarly, a large number of participants prefers to keep the data indefinitely, whereas a minority keeps their research data up to one year after the project completes.

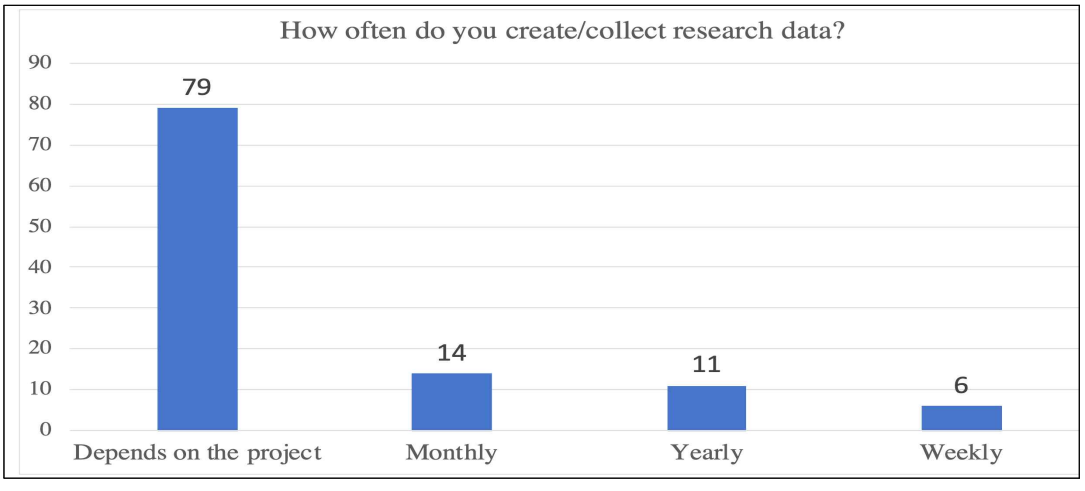

Fig. 5. Frequency of research data creation/collection.

A key element for this theme is that a research project seems to dictate participants' RD creation or collection, and storage timeline patterns. While this is clear from **Figure. 5**, **Figure. 6** seems to show a set timeline for keeping the RD for what could be common project timespan (both for the "3 to 5 years" and "up to 1 year" options).

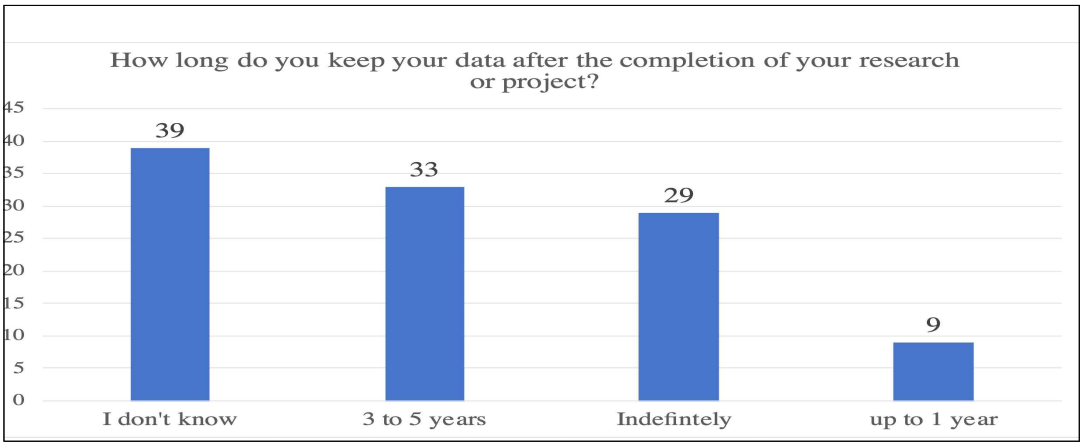

Fig. 6. Period of keeping the data after project completion.

4.4 Data sharing

Researchers who collect research data typically have a need to also share those data with collaborators. The survey results indicate that almost all participants share their data. Participants, however, differ greatly about when they share their data. As shown in **Figure. 7**, most of the participants share their data after project completion, whereas a considerable group does it upon request from their collaborators or people interested in such data. Similarly, a large number of participants share their data immediately after it is collected, whereas a minority shares it during its creation.

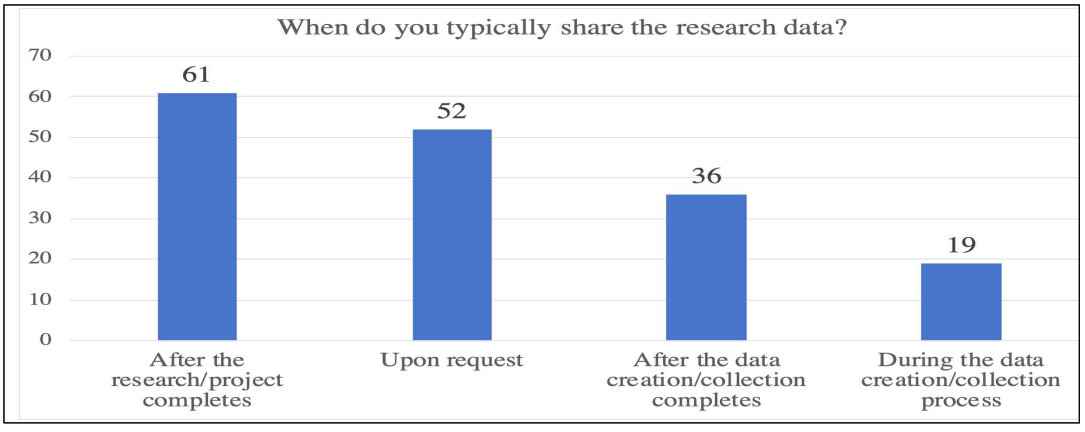

Fig. 7. The period when participants prefer to share their data.

It is interesting to find that most participants share their data as an attachment via emails. Another large proportion of participants prefers to share their data using a portable storage or cloud storage, whereas a minority prefers to do it using the department server or data repositories. **Figure. 8** shows all the data sharing practices from the survey.

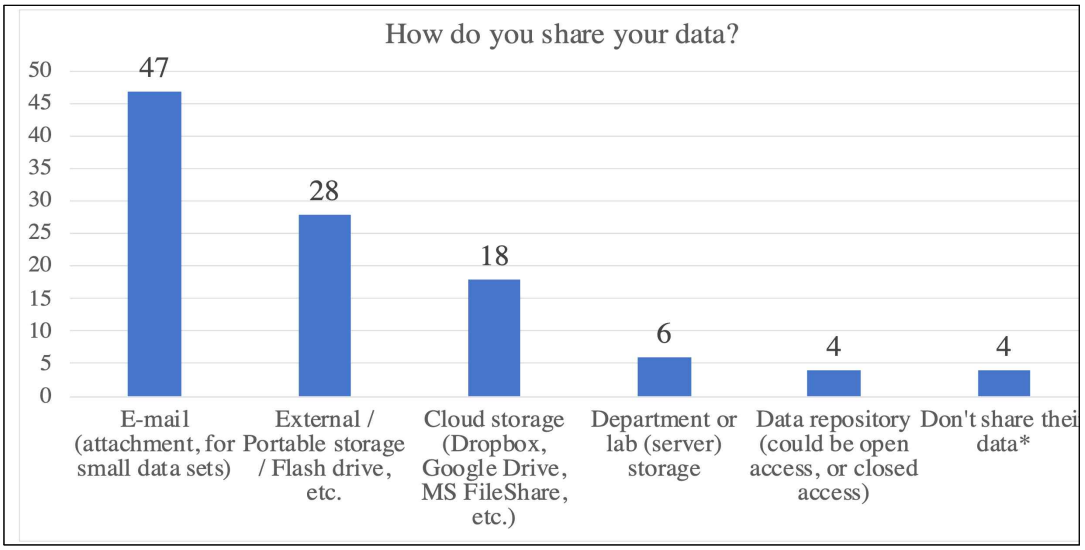

Fig. 8. Preference on how participants share their data.

Similarly to the case of RD creation and preservation discussed in the previous theme, we can infer that RDM policies on when to share RD as well as the mechanism to do it seem to be missing. Not only was this not part of the responses, but no one from the participants mentioned it as a support to their preferences - be it the moment or the means of sharing RD.

4.5 Viewpoint on research repositories

Of the 110 total participants, a majority of them (78) have no awareness of the existence of any research data repositories (**Figure. 9**). Despite this, most participants express interest - said either yes or maybe when answering the survey - in trying such a service, thus demonstrating a need for it. Just a fraction of participants indicated no need for such a service (**Figure. 10** shows the details).

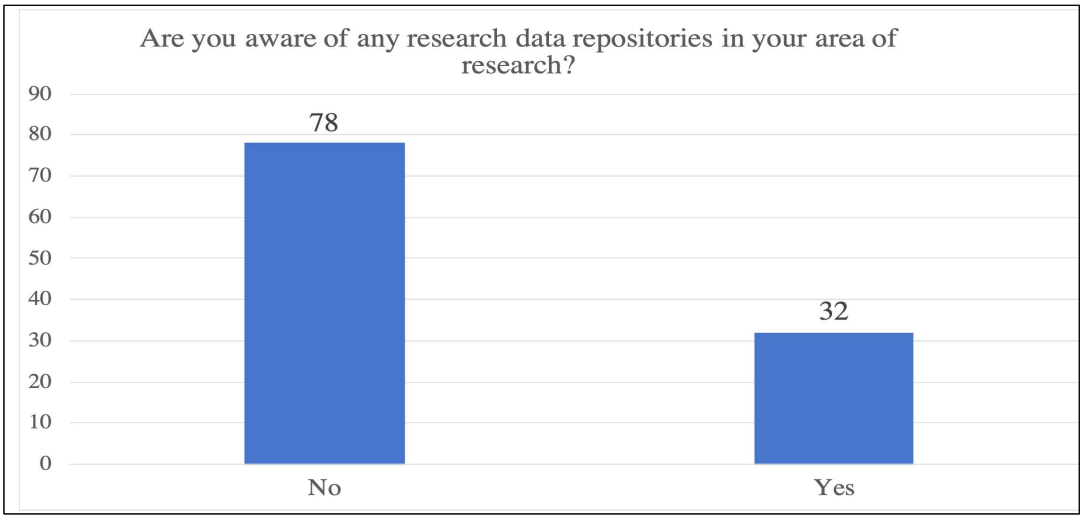

Fig. 9. Participant awareness of existing research data repositories.

A final piece of information that the survey could not capture and thus could not be subject of analysis for this theme relates to explaining the hesitation of those 53, and especially the refusal of those 4 respondents. What could cause these responses is perhaps lack of information about data repositories, a previous experience showing such repositories as unsuitable for their RDM needs, or disciplinary practices without a corresponding match in an IR feature, could have hampered IR adoption.

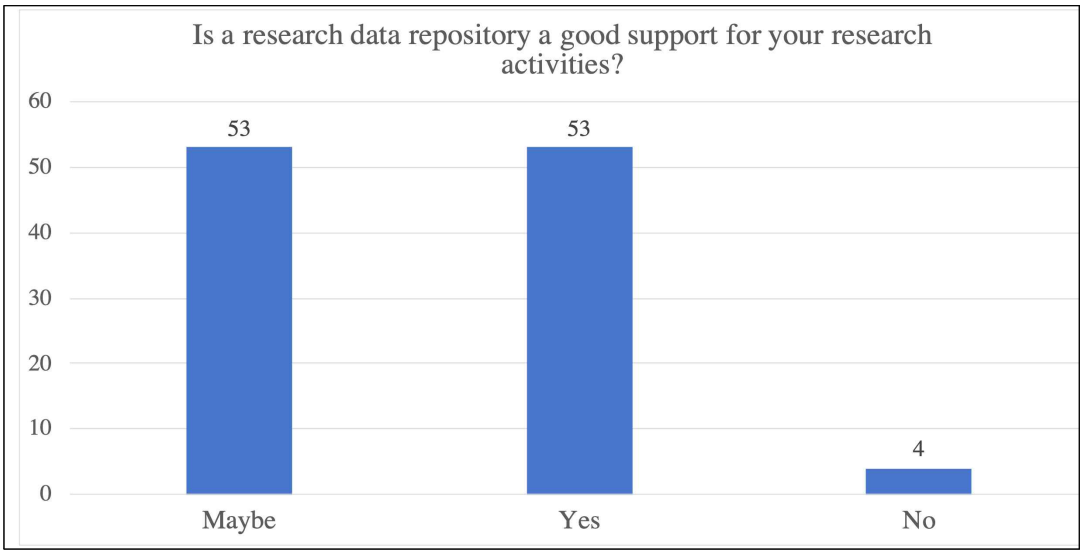

Fig. 10. The need for research data repositories.

4.6 Data repository features to be supported

We were interested to learn what features of research repositories our participants considered important when managing their data. **Figure 11** lists six features with storing data and metadata being the highest and creating metadata and documentation and archiving data being the lowest in terms of features. We were particularly interested in knowing the participants' attitude towards the access control feature.

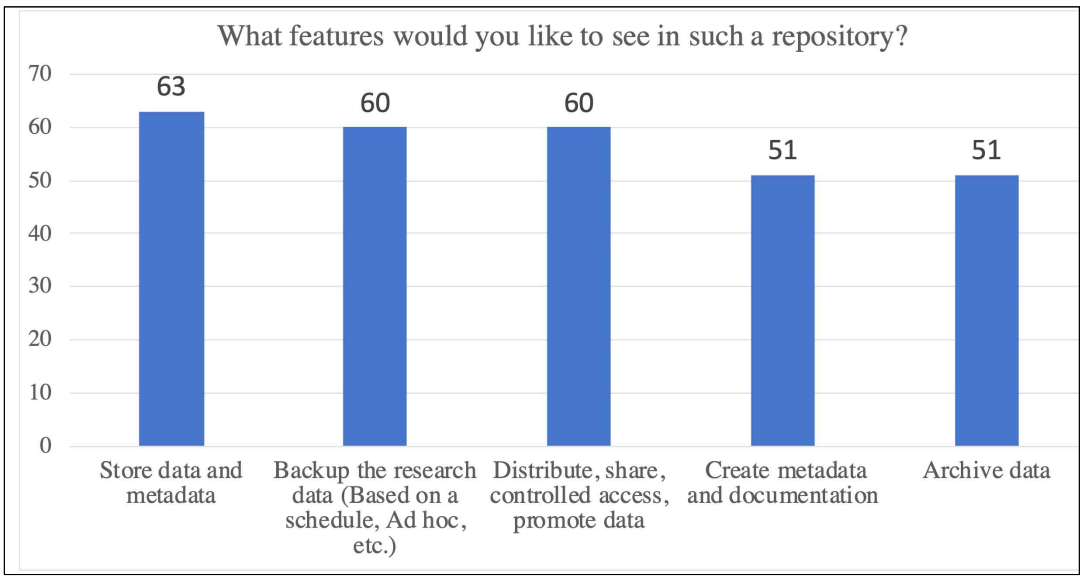

Fig. 11. Research repository features ranked by participant preference.

As indicated in **Figure 12**, the access control feature is preferred by a majority of participants, whereas the rest were either not sure or did not need such a feature. This also reflects the different types of RD that participants create, for which a "fully open" or "fully closed" access control mechanism surely is not ("expressive") enough.

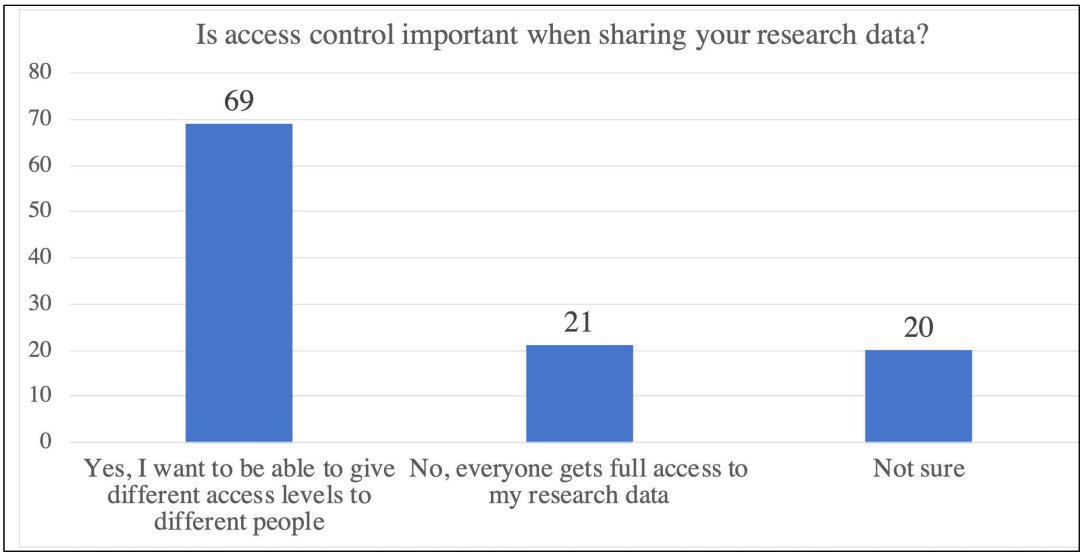

Fig. 12. The importance of access control.

These findings reveal interesting practices by our participants. It is a positive trend to see that RDs are present - generated, published, reused, etc. in a variety of formats and to a different extent (**Figure. 2**). Despite the fact that most participants were not aware of RD and their features (**Figure. 9**), they expressed a very high interest in relying on them, emphasizing the ability to store and distribute data and metadata (**Figure. 11**), or even the need for different access controls during the distribution of the same (**Figure. 12**). As seen in **Figure. 3**, IRs are the lowest used means for data storage and backup. This is in line with the findings shown in **Figure. 6**, which show that participants lack awareness about IRs, although they see them as a potential support tools for RDM activities.

5. Discussions and Recommendations

In this section, we discuss the survey findings and how they compare with the RDM practices and IR requirements from the previous related research. In addition, we see this as an important step to provide more context of RDM practices for the target communities, and to identify relevant recommendations for them. Finally, every community has its own specifics, and imposes certain limitations that reflect in our work, with which we conclude this part.

5.1 RDM practice: Discussions

With the analysis complete, we now discuss the identified RDM practices, including every theme, in relation to recent efforts to identify RDM practices and map them to IR features. The motivation to adopt IRs was one of the aspects we explored in this paper. In this context, similar to Asadi

et al. (2019) and Kipnis et al. (2019), our study participants acknowledged the benefits of IRs. Despite the low awareness of research repositories, survey participants in our study expressed the need for an IR as means to support their RDM requirements at a university setting.

The textual, databases, and spreadsheets remain the most prominent RD types, which is similar to what Guindon & Dennie (2017) report. Additionally, practices such as storage (local, external, or cloud storage), and dissemination (via email, portable or cloud solutions) are based on similar means. On a final note, 42% of respondents of the Guindon & Dennie (2017) study are not aware of data repositories, which does not seem too different from the university staff in North Macedonia. This just goes to show that RD usage is gaining momentum at a global scale, regardless of the minor details of its adoption with different researchers and regions. One aspect that Guindon & Dennie (2017) cite is the lack of incentives around RD, something that Hahnel et al. (2020) also present. The latter study presents some interesting findings about motivations and obstacles of sharing (open) data. Namely, survey participants of their study generally feel that sharing the data is not valued enough. Moreover, data citation, increased impact of a research work, and co-authorship on publications are the top-3 reasons to share, whereas concerns about misuse of data, sharing sensitive data, and lack of credit or acknowledgment are the top-3 reasons not to share the research data. Motivations for RD sharing and, generally, handling remains in our interest - and future work - to study with our participants.

Based on the data analysis in section 4, the need for organized storing, sharing and better management of the data is obvious. This is in line with (a subset of) the categories that Hsu (2016) or the Repository Platforms for Research Data IG have identified. On that note, sharing RD via email is the most popular approach, which, considering its capacity, could be an indication that the amount of data being collected and shared by those participants is relatively small in size. Of course, there remains a portion of respondents (57% of the cases) that we assume work with larger RD sizes, as noted by the other means of sharing it, as discussed in Section 4.4. Since we included staff members across university departments, this was expected as it shows the different practices across domains in terms of RDM.

What do the storage and backup practices show? In our view, we see room for improvement in moving from existing means, especially the local ones (almost 60% of the participants practice this approach; see **Figure 3**), towards more institutional, data repository-based solutions. Thus, although a considerable majority have stated that they currently store and backup their data locally (section 4.2), almost everyone expresses the need to share in different periods of time (see Section 4.4). Consequentl y, the use of an appropriate repository would simplify and facilitate this process, as opposed to current email distribution practices, external portable storage, etc. This need is further reinforced by the need to store and preserve the data for a longer period of time, as presented in **Figure. 6**.

On another note, IRs provide an appropriate path towards data FAIRness. Namely, considering the variety of RD and research practices across disciplines, de León and de Ferrer (2019) provide recommendations that RD repositories could adopt to become more FAIR. The mere adoption of IRs provides a good starting point towards the adoption of FAIR principles. As pointed out by Burns et al. (2013), two aspects that IRs adoption brings can be seen in the “discoverability of work through robust metadata and providing a permanent URI for that work”, which map well

with some of the FAIR criteria. The aspect of FAIR, although not explicitly explored in our work, seems to align well with the participants' practices (as shown in **Figure. 10**) since most of them either want or could try RDM services as supported by an IR.

We would like to complete this discussion with a view on RDM policy, which represents a common thread of the identified RD practices in our survey. While considering the answers from the questionnaire in **Table 1**, we did not come across any reference or even a hint of an RDM policy being the reason for the practices we identified. In this way, how long a researcher keeps their data, or how they share it, or when, to name but a few, are not rooted on an institutional RDM policy, but on, what we can assume, individual practices. We believe that an RD policy, coupled with incentives for complying with it (staff evaluation, promotion, institutional standing, etc.), could impact what RD practices researchers adopt.

5.2 RDM Requirements

The recommendations from the RD repository bodies, the related work on the topic, as well as our survey, provided a holistic context to understand RDM practices in the broader context, requirements that stem from them, as well as the more narrowed focus of RDM requirements for academic institutions. In this part, we present the (functional) requirements we identified as valuable or even explicitly needed for the survey participants.

The survey had a dedicated section on IR features that the participants would want to use for their RDM activity (see "Research data repository requirements" in **Table 1**), meant to both help us understand researchers' needs for IRs, and serve as a guide to universities or other bodies and institutions to potentially close such a gap and invest in providing more RDM support. Please note that, to the best of our knowledge, this is the first such study of RD management practices and their mapping to IR features for universities in North Macedonia.

Survey participants were interested in different IR features, and these features could support activities in many phases of the data lifecycle. In order to discuss their interest through a data lifecycle perspective, we chose the model from the UK Data Archive[8], among many others. As it can be seen from **Table 3**, the presence of data activities in the different phases is different, including a case (Re-using data) for which we did not record any activity in the survey data. This is not to worry, however, since we were not focused on evaluating the adoption of a data lifecycle as means to organize the RD practices at the participating higher education institution staff members. With that said, let's next discuss the IR requirements through the data lifecycle perspective:

- **Collecting data.** In this context, participants were interested in a feature that would allow them to create meta(data) and accompanying documentation (in 51 of the participants).
- **Publishing and Sharing data.** Once meta(data) is there, publishing and sharing it becomes an important next step. We see this with 60 of the participants when they specify distributing, sharing and promoting as a feature they need in an IR. At this point we would like to

8) https://ukdataservice.ac.uk/learning-hub/research-data-management/

point out one requirement in particular that we identified and that is relevant to this feature - that of access control. Two thirds of the participants explicitly were in favor of a feature that enables them to publish and share RD based on a variety of access levels. However, considering that 21 were keen on publishing their RD with full access, while 20 of them were not sure, these two options can also be seen as just additional access control levels. Access control is important as it enables a research team to provide different access levels to a RD collection. For example, one can provide a different (re-use) access within the same research group, department, or institution; the (outer) community, researchers at the national level, and so on.

- **Preserving data.** Most of the requirements we identified are in relation to data preservatio n tasks. In this context, 63 participants selected storage, 60 selected backup, and 51 selected archive as features in an IR. When it comes to backing up RD, users prefer having different considerations to do so, such as based on a schedule, mode (ad hoc vs a more general approach), etc. Also, please note that there is a slight difference between the need to create and store meta(data): generally, more researchers store (download for reuse) rather than create RD themselves.

Table 3. IR requirements categorized based on data lifecycle

Data Lifecycle Phase	Functional Requirement
Collecting data	• Create (meta)data and documentation
Publishing and sharing	• Distribute, share, and promote data • Provide access control
Preserving data	• Store • Back up • Archive
Re-using data	• Not applicable

As discussed in 4.5, the survey participants generally are not familiar with IRs and their capabilities, but those that do, indicate a readiness to use such a solution. Considering a(ny) RD life cycle as a categorization framework, the IR requirements we identified in our survey are part of many life cycle phases (collect, publish, preserve), echoed to a different extent (some were more referred to than others in the replies) among the participants. As we pointed out, while storing meta(data) is the highest sought out feature, there is a relatively small difference with the remaining four IR features selected by the survey participants. Moreover, the requirements do not reflect any specifics of a field or domain the participants adhere to, and rather represent some of the more common IR features, generalizable at an institutional setting, in this case a university.

5.3 RDM practice: Recommendations

The methodology we adopted for this work enabled us to have considerable insights in RDM practices and IRs as means of supporting such practices. We would like to next share some of

this insight, which could be used to explore RDM practices with new or existing academic communities, and perhaps other institutions.

The role and valuing of RD. The research staff would be more keen on focusing on RD as an equally valued deliverable as research publications. Similar to the incentives for researchers to publish, and the accompanying infrastructure - including publication repositories, RD could see an uplift if university staff members, as well as other relevant bodies and institutions, are also valued based on the RD they collect, maintain, publish, share, and so on.

Research groups are ready for RD sharing. We speculate such different strategies are dependent on the level of collaborators involved in the project. If the project has many collaborators, participants have the need to share those data during creation or immediately after the data have been created, perhaps, so that collaborators either complement the data creation process or they use it for analysis. On the other hand, researchers who do not have many collaborators in their project share their data after the project completion or upon request. It seems that most researchers belong to the latter group. One recommendation would be to adopt RDM practices (let's say from "Publication and Sharing") with researchers that are part of a project and for whom sharing RD is more of a need than with individual researchers.

Start with basic activities from the data life cycle phases. In the previous sub-section, we listed the RDM requirements we identified in this study. They represent some of the basic RD-related activities, such as creation, storage, backup, and so on. Seeing that these requirements appear across different university departments implies that they are common enough to be considered as an initial path towards pushing for a broader adoption of RD within research practices at universities.

RDM policy. There are many facets of RDM practices that could impact the culture of RD. Some of the findings in our survey go in the direction of supporting research data management policies. Any such policy initiative - be it at an institutional, broader, or national level - should strive to support the incentives and lessen the eventual barriers as Griffiths (2009) identified in the survey practices. In this context, there seemed to be a lack of such policies at the institutions we surveyed, as there was no rationale given based on institutional RD policies for any of the practices reported by the participants. In a way, this is an umbrella (catch all) for other recommendations that an institution could adopt, and one which enables an institution to pace their RD-related plans according to their goals, available resources, research culture, etc.

5.4 Methodology limitations

After discussing the results and our findings, we would like to briefly discuss the limitations of the methodology we adopted, and provide the rationale for them.

Some of the limitations we see in of our methodology include:

- **The topic scope.** Being that it is one of the first surveys that focuses on RDM practices for universities in North Macedonia, we had to decide on its scope - its breadth and depth. The former would include the spectrum of RD life cycle, and its corresponding phases,

and the latter would include the level of details for each such phase, i.e. activities around RD that go with it. As an initial exploration, we wanted to start "small", i.e. be as close to the targeted participants as possible about their RDM activities. The risk of increasing either the breadth or depth would result in no feedback, letting the participants potentially feel not up to the task or even wrongfully invited to fill out the survey.

- **Target group.** We targeted three universities for the survey. Even though they constitute a considerable part of the academic staff in North Macedonia, we cannot confidently report on RDM practices at the national level. We were aware of this since the very beginning of our exploration of RDM practices in academic institutions, and we wanted to have a smaller scope, and incrementally include more institutions along the way. To maintain this, we chose three of the larger universities in the country, a target group whose RDM practices were manageable for our research scope. Finally, even for these three institutions, not all of the invited staff members participated, which leaves us with limited knowledge of their RDM practices.
- **Data analysis.** Due to the number of participants, the research scope, as well as the topic, which is relatively new for our target group, the data analysis was of a descriptive nature. Moreover, it only includes the practices of this specific target group (the three universities), and only of the participants that responded to the survey. As such, the analysis could not be used to predict RDM practices or potential IR features based on the data we gathered. Thus, it cannot be generalized to broader groups - academic staff - and only depicts the practice of 110 survey participants.
- **Ethical aspects.** When collecting and sharing data, inevitably ethical aspects become highly relevant, such as complying to the General Data Protection Regulation (GDPR), for example. Such aspects should be reflected within IR features and even help researchers easily adopt ethical ways of dealing with the data they collect and share, whether those data belong to private individuals or companies. Considering the complexity of the topic, we do not address these in this study, however, we recognize their relevance.

Lastly, we would like to mention generalizing the results of this study. Namely, since we do not have representative samples from all the departments surveyed, thus their corresponding fields, these results do not reflect the RDM practices for those departments (and corresponding fields) at the national level. While we were able to cover different departments, we could not control a representative response rate from each department. For example, having more feedback from one department or field over the others will necessarily reflect the corresponding practices in the results. This, in turn, would make it seem as these practices are common at the study level, which could not necessarily hold.

A final comment to wrap up this section goes to the lack of an institutional policy for RD. We notice this from the feedback, i.e. the parts where survey participants do not mention any policy as a rationale for their RDM practices. We strongly believe that a policy that promotes and values RD-related tasks - their creation/collection, curation, sharing, publication, etc. - would benefit the research staff at universities, as it would provide guidance and motivation for them to engage with RD similarly to how they engage and treat research publications.

6. Conclusion and Future Work

RDM is gaining in popularity, including support organizations, research focus, and so on, and IRs are seen as possible tools of choice to support its practices. The heterogeneity across domains contribute to a richer, varied RDM picture, and this requires that these domains and communities be studied and their requirements for RDM understood, including mapping to IR features. Thus, despite being based on common patterns, RD life cycle, etc., individual organizations need individual attention to distill, identify, and ultimately map these practices to IR features.

Such is our work in this paper; based on a two-phase approach, we started with a smaller-scoped study of one university, and broadened it to include 3 of the largest universities in North Macedonia. The first phase provided the opportunity to not only study the RDM practices at this university, but also test our survey approach and questionnaire details, before, in the second phase, moving on to a broader audience. This research provides a stepping stone - an initial attempt - at surveying the RDM (requirements) scene of universities in North Macedonia, and mapping the potential IR features to support them.

The feedback from the 110 survey participants showed RD as part of research activities at these universities, in a variety of RD-related tasks, such as creation, publication, dissemination, backup, access control, and more, which we organized as 6 themes in our findings. Moreover, we were able to identify six requirements that participants would like to have in an IR solution. These could prove useful in case any of the universities that participated in the survey decided to adopt such a solution or any other university without established practices.

In the future, we intend to follow up and update some aspects of this work, as well as broaden the scope and introduce new ones to it. For the former, we have a couple of research goals, such as conducting qualitative data gathering in order to increase the understanding of the current RDM practices; including additional universities from North Macedonia in order to partly address the limitations mentioned earlier. For the latter, we plan broadening the RDM topics to cover for an ever more complete RDM context for universities in North Macedonia.

References

Akers, K. G. & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. International Journal of Digital Curation, 8(2), 5-26.

Akers, K.G. & Green, J.A. (2014). Towards a Symbiotic Relationship Between Academic Libraries and Disciplinary Data Repositories: A Dryad and University of Michigan Case Study. Int. J. Digit. Curation, 9, 119-131.

Alcalá Ponce de León, M. & Anglada i de Ferrer, L. M., (2019). FAIR x FAIR. Feasible, Affordable and Implementable Requirements for a FAIR research data repository. Report, p. 44.

Artini, M., Candela, L., Manghi, P. & Giannini, S., 2020, January. RepOSGate: Open Science Gateways for Institutional Repositories. In Italian Research Conference on Digital Libraries (pp. 151-16 2). Springer, Cham.

Asadi, S., Abdullah, R., Yah, Y. & Nazir, S. (2019). Understanding Institutional Repository in Higher Learning Institutions: A Systematic Literature Review and Directions for Future Research, in IEEE Access, vol. 7, pp. 35242-35263. COUNTER: Consistent, Credible, Comparable. (n.d.). Retrieved from https://www.projectcounter.org/.

Burns, C. Sean, Amy Lana, & John M. Budd. "Institutional repositories: Exploration of costs and value." D-Lib Magazine 19, no. 1/2 (2013).

Confederation of Open Access Repositories. (2020, October 8). COAR Community Framework for Best Practices in Repositories. (Version 1). Zenodo. http://doi.org/10.5281/zenodo.4110829

FAIR Principles. (n.d.). Retrieved from https://www.go-fair.org/fair-principles/.

Franke, T., Grutz, R., & Dickmann, F. (2013). Functional Requirements for a Central Research Imaging Data Repository. Studies in health technology and informatics, 192, 298-302 .

Gordon, A. S., Millman, D. S., Steiger, L., Adolph, K. E., & Gilmore, R. O. (2015). Researcher-Library Collaborations: Data Repositories as a Service for Researchers. J. of Librarianship and Scholarly Communication, 3(2), eP1238.

Gray, A. (2009). Institutional Repositories for Creative and Applied Arts Research: The Kultur Project.

Guindon, A., & Dennie, D. (2016). Concordia University Research Data Management Survey 2015-16. Retrieved from https://spectrum.library.concordia.ca/982722/2/RDM-Concordia-Questionnaire-Final.pdf

Guindon, A., & Dennie, D. (2017, April 28). Research Data Management Survey, Concordia University [PowerPoint slides]. Concordia University Library's 15th Annual Research Forum. Retrieved from https://spectrum.library.concordia.ca/982529/1/Guidon-Dennie-library-forum-2017.pdf.

H2020 Programme. (2017). Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 [online] p. 8. Retrieved from https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf [Accessed: 2 Dec. 2019].

Hahnel, M., McIntosh Borrelli, L., Hyndman, A., Baynes, G., Crosas, M., ⋯ Research, N.. (2020). The State of Open Data 2020. doi: 10.6084/m9.figshare.13227875.v2.

Hey, T. (2016). The Fourth paradigm - data-intensive scientific discovery and open science. Book of abstracts, Barcelona: Barcelona Supercomputing Center, p. 34-36.

Hsu, L. (2016). RDA: Functional Requirements for Research Data Repository Platforms. Retrieved from https://my.usgs.gov/confluence/display/cdi/RDA%3A+Functional+Requirements+for+Research+Data+Repository+Platforms.

Kim, S. (2018). Functional requirements for research data repositories. International Journal of Knowledge Content Development & Technology. 8(1), 25.

Kim, S. T. (2020). Functional requirements of data repository for DMP support and CoreTrustSeal authentication. International Journal of Knowledge Content Development & Technology, 10(1), 7-20.

Kipnis, D. G., & Palmer, L. A. (2018). Medical institutional repositories in a changing scholarly communication landscape. Against the Grain, 30(4), 56.

Limani, F., Hajra, A., Ferati, M. and Radevski, V. (2020). Requirements and Recommendations for University Research Data Repository: A Case Study. In 18th International Conference e-Society, 2-4 April, 2020 (pp. 51-58). IADIS Press.

Lin J, & Strasser C. (2014). Recommendations for the Role of Publishers in Access to Data. PLoS Biol 12(10): e1001975.

Luther, J. (2018).The Evolving Institutional Repository Landscape. ACRL/Choice publisher.

Lynch, C. A. (2003). Institutional repositories: essential infrastructure for scholarship in the digital age. portal: Libraries and the Academy, 3(2), 327-336.

MacIntyre, R., & Jones, H. (2016). IRUS-UK: Improving understanding of the value and impact of institutional repositories. The Serials Librarian, 70(1-4), 100-105.

McQuilton, P., Sansone, S. A., Cousijn, H., Cannon, M., Chan, W., Carnevale, I., & Cranston, I. (2020). FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter. Open Science Framework.

Needham, P., & Lambert, J. (2019). Institutional repositories and the item and research data metrics landscape. Insights, 32(1).

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., ... & Dierolf, U. (2013). Making research data repositories visible: the re3data. org registry. PloS one, 8(11), e78080.

Petritsch, B. (2017). Implementing the institutional data repository IST DataRep.

Registry of Research Data Repositories. (n.d.). Retrieved from https://www.re3data.org/.

Repository Platforms for Research Data IG. (n.d.). Retrieved from https://www.rd-alliance.org/groups/repository-platforms-research-data.html.

Research360 Project (2013). Institutional Data Repository User Stories. University of Bath.

Royster, P. (2019). IRs in America: "Land of the Free" or "Free Online Access". Retrieved from https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1164&context=library_talks.

Scherer, D., & Valen, D. (2019). Balancing Multiple Roles of Repositories: Developing a Comprehensive Repository at Carnegie Mellon University. Publications, 7(2), 30.

Uzwyshyn, R. (2018) Research Data Repositories: Developing and Implementing Infrastructures for Institutional and Consortial Environments. Coalition for Networked Information. San Diego, CA. April 12-13 2018. Retrieved from https://www.cni.org/wp-content/uploads/2018/04/cni_researchdata_uzwyshyn.pdf.

Vicente-Saez, R., & Martínez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. Journal of Business Research, 88, 428-436.

Welcome to DataCite. (n.d). Retrieved from https://datacite.org

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

Wu, M., Psomopoulos, F., Khalsa, S. J., & de Waard, A. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. Data Science Journal, 18(1).

Yin, S., Zhang, J., Jia, M., & Hu, J. (2020). How to Evaluate and Select a Data Repository for Humanities and Social Science: A Case Study of Fudan University Data Repository for Humanities and Social Science. Library Trends 69(1), 125-137. doi:10.1353/lib.2020.0024.

[About the authors]

Fidan Limani With a background in computer science and information systems, since 2017 he is associated with the Leibniz Information Center for Economics – ZBW, engaged with (national) research data infrastructure projects. This includes research data management aspects and services implementation for different research communities, with foci on, among other aspects, analysis and implementation of metadata standards, FAIR principles, and FAIR Data Objects, automatic metadata linking to standard bibliographic data, and so on. Another important part of his research includes the Semantic Web and the application of Knowledge Graphs as integration means for different scholarly research deliverables into (digital) library environments. He previously worked as a research and teaching assistant at the Computer Science department of the South East European University in Macedonia for 10 years.

Arben Hajra works as Researcher at Leibniz Information Center for Economics - ZBW, Germany. He is part of the EconBiz development team, with a particular focus on enriching digital resources through machine learning and information retrieval approaches. He also has several years of teaching experience, mainly as a lecturer at the SEE University in North Macedonia. He earned his PhD in Computer Science from Kiel University in Germany. His research interests include machine learning, information retrieval, semantic web technologies, author disambiguation, databases, and NLP.

Mexhid Ferati is an Associate Professor in Informatics at Linnaeus University. He has an international experience working as a researcher and lecturer in the United States, North Macedonia, and Norway. He earned his PhD in Human-Computer Interaction from Indiana University in 2012. His research interests are within Human-Computer Interaction, Interaction Design, Accessibility, Internet of Things, and STEM Education. He worked on projects funded by EU, STINT, EUniWell, Platform eHealth, NSF, and Google and served as a reviewer for several international conferences and journals. He currently leads the Interaction Design Research Group and coordinates the Interaction Design Program at Linnaeus University.

Vladimir Radevski is a Full Professor in Computer Sciences at the Faculty of Contemporary Sciences and Technologies at South East European University in Tetovo, Republic of North Macedonia. He received his PhD in Informatics from University of Paris 13, France (2000) in the field of Artificial Intelligence. Dr. Radevski is with SEE University since 2003 and, besides teaching a wide range of computer science courses, he was also holding many administrative positions, including that of a Vice-rector for Academic Planning and Digitalization. He was member of various national bodies like Expert body of the Macedonian Associations of IT, National Commissions, National body for Strategy in IT development and similar. His research interests are realized through membership in several international research groups in Turkey, France and Germany and are in the area of Applied Artificial Intelligence and Semantic Web.