Research Data Management Commitment Drivers: An Analysis of Practices, Training, Policies, Infrastructure, and Motivation in Global Agricultural Science

SEBASTIAN S. FEGER, LMU Munich, Germany and ICARDA, Lebanon
CININTA PERTIWI, ICARDA, Lebanon
ENRICO BONAIUTI, ICARDA, Lebanon

Scientists largely acknowledge the value of research data management (RDM) to enable reproducibility and reuse. But, RDM practices are not sufficiently rewarded within the traditional academic reputation economy. Recent work showed that emerging RDM tools can offer new incentives and rewards. But, the design of such platforms and scientists' commitment to RDM is contingent on additional factors, including *policies*, *training*, and several types of personal *motivation*. To date, studies focused on investigating single or few of those RDM components within a given environment. In contrast, we conducted three studies within a global agricultural science organization, to provide a more complete account of RDM commitment drivers: one survey study (n = 23) and two qualitative explorations of *regulatory* frameworks (n = 17), as well as *motivation*, *infrastructure*, and *training* components (n = 13). Based on the sum of findings, we contribute to the triangulation of a recent RDM commitment evolution model. In particular, we find that strong support and suitable tools help develop RDM commitment, while policy conflicts, unclear data standards, and multi-platform sharing, lead to unexpected negotiation processes. We expect that these findings will help to better understand RDM commitment drivers, refine the RDM commitment evolution model, and benefit its application in science.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Data-processing science; Reuse; Reproducibility; Human data interventions; Motivation; Research data management; Data management commitment.

ACM Reference Format:
Sebastian S. Feger, Cininta Pertiwi, and Enrico Bonaiuti. 2022. Research Data Management Commitment Drivers: An Analysis of Practices, Training, Policies, Infrastructure, and Motivation in Global Agricultural Science. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 322 (November 2022), 36 pages. https://doi.org/10.1145/3555213

1 INTRODUCTION

The value of comprehensive research data management (RDM) is significant. Through the systematic documentation, preservation, and sharing of scientific resources, all core RDM practices [4, 39, 72], researchers and organizations make their materials reusable [42] and reproducible [56]. RDM is key in demonstrating responsibility in accessing unique experiments, data sources, and populations. Suitable RDM practices and appropriate online RDM tools are prerequisites to validate and advance

Authors' addresses: Sebastian S. Feger, LMU Munich, Munich, Germany, ICARDA, Beirut, Lebanon, S.Feger@cgiar.org; Cininta Pertiwi, ICARDA, Beirut, Lebanon; Enrico Bonaiuti, ICARDA, Beirut, Lebanon.

2573-0142/2022/11-ART322 $15.00
https://doi.org/10.1145/3555213

science [30, 36]. In some fields, RDM even has implications way beyond scientific curiosity and the desire to push the boundaries of our knowledge. In global agricultural and food sciences, for example, the question whether or not scientists follow comprehensive RDM practices impacts how we feed the world [38, 40]. In today's data-intensive world, practices between poor and good RDM can make the difference in supporting farmers in the world's most rural areas [27], providing enough and healthy food to troubled regions, and developing sustainable farming practices that reflect the challenges of the global environmental change [53].

One of the biggest challenges of RDM lies in the effort required to follow comprehensive practices. Selecting, cleaning, describing, preserving, and sharing resources is time consuming [12]. Time that researchers commonly prefer to invest in novel research which promises to advance their careers. This is understandable as the traditional academic reputation economy is heavily focused on novel work, rather than replications [5, 20, 29]. Recent user-centered research focused on those issues around *motivation* and incentives for RDM by exploring and describing new meaningful technology-mediated benefits for scientists who document and share their work [32, 34]. But, *motivation* is not the only barrier. Related work further described opportunities and challenges of RDM *policies*, particularly in the context of implications for infrastructure design [36, 54]. Additional components of RDM frameworks that impact *practice* — and are impacted by *practice* — include level of *training* and suitability of *technical infrastructure*. All those components involved in RDM are often considered in topic-specific scientific explorations. Instead, our work analyzes needs and requirements for effective RDM in global agricultural and food sciences through the lenses of those five key RDM components — i.e., practices, training, policies, infrastructure, and motivation — across three studies within the CGIAR, a partnership of globally distributed international organizations in agricultural and food sciences. In particular, we focus on practices within one of those centers, the International Center for Agricultural Research in the Dry Areas (ICARDA).

In Study I, we mapped current practices related to the five RDM components in a survey study. This study was closely aligned with a theoretical model on RDM commitment evolution, recently described within the Computer Supported Cooperative Work (CSCW) domain [36]. Our work aims at both validating that model through a domain-specific survey study, as well as mapping the current state in agricultural research. For this purpose, we collected responses from 23 participants. Studies II and III were conducted as qualitative explorations focusing on RDM policies (Study II; n = 17) and motivations, infrastructure, and training (Study III; n = 13). Based on that extensive research process and the different lenses we adopted, we provide a detailed account of practices, needs, and requirements across various stages of RDM commitment evolution in agricultural science at ICARDA and the CGIAR. Our findings show a mismatch between the suitability of current infrastructure, in relation to implemented training mechanisms, policy regulations, and motivational drivers. We position our findings within the changing global agricultural research environment, and a major CGIAR reform that is heavily impacted by future RDM strategies and the design of technical infrastructure.

The sum of our findings from the three studies in global agricultural science contribute to the triangulation of the RDM commitment evolution model [36]. In particular, our findings suggest a more positive view on the RDM commitment life cycle in organizations that complement their regulations with strong training support and suitable technical infrastructure. Yet, in contrast, our findings reveal an additional negotiation process around policy conflicts, multi-platform sharing, and unclear data standards, that involves various stakeholders throughout the RDM process. We argue that our research in global agricultural science helps to refine the RDM commitment evolution model and strengthens its application in other domains. In summary, our paper makes three key contributions:

- We present findings from a survey study that is aligned with a recent RDM commitment evolution model [36].
- We mapped five RDM commitment components (i.e., practices, training, policies, infrastructure, and motivation) in global agricultural science through one survey study and two semi-structured interview studies.
- We present implications on how to stimulate and sustain RDM commitment in agricultural science and beyond. Further, we reason about the applicability of the RDM commitment evolution model based on findings from three studies and discuss how our work refines the model through the systematic mapping of RDM components.

This paper is structured as follows. First, we reflect on work related to RDM commitment evolution, with particular regard to the five RDM components practices, training, policies, infrastructure, and motivations. Based on those reflections, we present our research questions. Second, we present our research methodology, including the interconnections between our three studies. Next, we present the procedures, results, and findings of all three studies separately. Finally, we discuss implications for stimulating, improving, and sustaining RDM commitment evolution within ICARDA/CGIAR and across the sciences.

2 RELATED WORK

Our work is framed around and explores ecological validity of the stage-based model of personal RDM commitment evolution, introduced by Feger et al. [36]. Figure 1 shows the four stages: Non-Reproducible Practices, Overcoming Barriers, Sustained Commitment, and Reward. Based on a cross-domain study, the authors described this model that explains how researchers transition from non-reproducible practices to sustained RDM commitment. They described that policies and intrinsic motivation play key roles in the initial *adoption* process, transitioning away from non-reproducible practices. According to the model, they need to overcome barriers related to unsuitable technical infrastructures and lack of formal training, before they can effectively integrate RDM into their work routine. The model further shows that sustained long-term commitment is contingent on various forms of rewards that speak directly to the motivation of researchers, including scientific visibility, citation counts, and promotions. As such, this model relates — through its stages and transitions — directly to the *five key RDM components* we explore in our work (i.e. *Practices*: non-reproducible practices stage; *Policies* and *Motivation*: adoption transition; *Training* and *Infrastructure*: overcoming barriers stage; *Motivation*: rewards stage/transition). While our work is closely related to this model of personal RDM commitment evolution, we are consistently referencing to the five identified key RDM components practices, training, policies, infrastructure, and motivation, rather than to the stages and transitions of the model. We decided to do so because the terms used to reference the model's stages and transitions do not carry an established meaning within the scholarly areas of RDM and digital data practice. Instead, the five key RDM components identified in the model description relate to major areas of interest to researchers and practitioners, as evidenced by this section.

In this section, we first briefly introduce RDM more conceptually. Next, we reflect on related work within each of the five RDM components we target specifically through our research: Practice, Training, Policies, Infrastructure, and Motivation. We conclude this section through a summary of related work and present the research questions that guided our studies.

2.1 Research Data Management: An Overview

Research Data Management (RDM), "the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results" [71] concerns the systematic

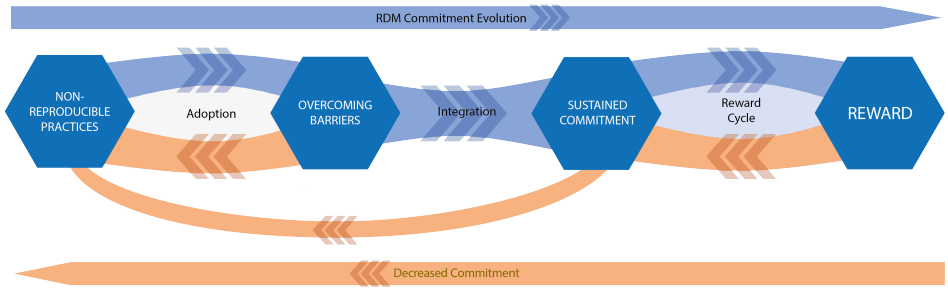

Fig. 1. The Stage-Based Model of Personal RDM Commitment Evolution, described by Feger et al. [36]: "Four stages describe researchers' commitment to comprehensive RDM: Non-Reproducible Practices, Overcoming Barriers, Sustained Commitment, and Reward."

documentation of data and meta-data, as well as their long-term preservation and sharing. Several guidelines and frameworks exist that aim to foster comprehensive RDM for open, reproducible, and reusable science. De Waard et al. [23], for example, contributed a model featuring ten data management aspects classified according to four key factors: *Saved* (Stored and Preserved), *Shared* (Accessible, Discoverable, and Citable), *Trusted* (Comprehensible, Reviewed, Reproducible, and Reusable), and *Successful Data*. The model helps in integrating the ten aspects and concepts between systems, domains, and stakeholders. The authors stressed that "in building systems for data reuse or data citation, the practices of current systems for storing and sharing data need to be taken into account." The well-known FAIR Data Principles [39, 72] set important data standards, as they demand data to be *findable* (F2: "data are described with rich metadata"), *accessible* (A2: "metadata are accessible, even when the data are no longer available"), *interoperable* (I3: "(meta)data include qualified references to other (meta)data"), and *re-usable* (R1.1: "(meta)data are released with a clear and accessible data usage license").

Those models and frameworks share the common understanding that the importance of data processing is ever increasing in science today. While *computational* science was described as sciences' third paradigm, *data-intensive* science was established as the fourth paradigm of science [7], characterized also as "the study of the generalizable extraction of knowledge from data" [26]. With new means for systematic knowledge extraction from increasing data volumes comes the responsibility to make that knowledge accessible and widely usable. Agricultural science represents a very strong example of a field that can turn data and knowledge into action that benefit both individuals as well as the greater good. Effective RDM that makes actionable information accessible and usable enables sustainable farming in rural and low-income regions [27], helps in the creation and maintenance of functioning food chains [38, 40], and allows to introduce change in farming practices that reflect the challenges imposed by the climate change [53].

2.2 Mapping the State of RDM

The systematic mapping of RDM practices is generally important, as it provides tools for stakeholders in the scientific process to plan and implement strategies that counter shortcomings and build on working practices. In addition, this systematic mapping allows us to understand the current state of data management practices and related dependencies. Today, we have data from numerous surveys at our disposal that either map the state of RDM and reproducibility across scientific domains and topics, or focus on individual RDM components in individual fields and organizations. The survey study from Monya Baker [3], published in Nature, provides a very good example of

the former. Baker surveyed more than 1,500 scientists from several fields and found that most perceived a "significant" reproducibility crisis in science, while an additional 38% still referred to a "slight" crisis. Notably, 50% of the participants in the study even described failing to reproduce some of their own work in the past. Additional findings showed that around 80% of the researchers indicated that unavailable code/data, and unavailable raw data always/often/sometimes contributed to irreproducible research. Between 60% and more than 80% of the scientists further indicated that better education, incentives for better practice, incentives for formal reproduction, and journals enforcing standards, could boost reproducibility. These are notable findings, as they point towards several of the RDM components we explored in great detail through our work.

The work of Bishoff and Johnston [8] represents an example of a very focused study, as they reviewed boundaries of RDM through the lenses of NSF data management plans. They described sharing of private and sensitive data as a barrier that can be mapped even in the engineering sciences. This finding resonates also in the work of Akers and Doty [2]. They circulated their survey study among faculty members and concluded that "growing data-intensiveness of scholarly research applies not only to the basic sciences but also to the social science." In addition, they described restricted access to data and resources as one of the most common RDM barriers. In this context, the work of Buys and Shaw [18], surveying faculty, students, and staff showed that formal education in RDM is a big issue: "faculty may have a greater stake in good data management, but the individuals who are managing the data (staff and students) may not have as great an understanding of institutional practices or general practices of good data management." Related to this finding, Hoy [43] advocates for libraries and librarians to support scientists with Big Data challenges. Tang and Hu [66] also investigated institutional support and found that infrastructure limitations represented most often described barriers. Those limitations included storage space, limited support staff, and bandwidth.

This section focused on related work that mapped RDM practices and barriers. The various facets and components of effective RDM have previously either been studied across topics and domains on a high level, or in detail, focusing on individual fields or subject matters. Already in this quantitative exploration of the state of RDM, we identify an opportunity to contribute a combined and detailed analysis of several RDM components and barriers. Our work reflects this understanding and further provides both quantitative and qualitative accounts within agricultural science.

2.3 Providing Suitable Training and Support

Knowledge about RDM and its value for science is key in the process of identifying with and internalizing RDM practices. That way, RDM education and identification plays important roles in the early adoption process of the RDM commitment evolution model. If this training is missing, however, a set of different initial drivers need to stimulate adoption, according to the model. In any case, formal education, training, and support, are key in the *overcoming barriers* stage, where the set of skills and institutional support structures impact whether or not researchers are capable to handle issues and integrate RDM practices into their routine workflow. Reviewing related work, we perceive indications for shortcomings in both education and support. Based on their ethnographic interview study with researchers and faculty, Jahnke and Asher [44] found that scientists had little to no RDM training and were unsatisfied with their knowledge. In a survey study conducted by Bishop and Borden [9], 70% of the participating 81 scientists indicated that they had no prior training in RDM. The authors advocated for more support provided by libraries. Related to formal education, Thielen and Hess [67] stressed that RDM skills are usually not taught in graduate programs and emphasized the value of providing RDM instructions to education graduate students. Read et al. [57] focused on the mismatch between librarians RDM knowledge and researchers'

domain-specific expertise. The authors created online modules that aimed at helping librarians learn the specific constraints of a target domain. Additional resources were deployed, supporting librarians in teaching RDM skills to researchers. Their results showed that the "online curriculum increased librarians' self-reported understanding of and comfort level with RDM. The Teaching Toolkit, when employed by librarians to teach researchers in person, resulted in improved RDM practices." This is a strong example of technology-mediated training and support that we explore more closely in our research.

2.4 RDM Policies and Enforcement

Identified shortcomings in RDM, and consequently irreproducibility, have led to the implementation of various types of regulations and policies. To better understand the role of policies and their constraints, Higman and Pinfield [41] analyzed UK RDM policies and conducted interviews with staff responsible for their development. The authors found that data sharing "is considered an important activity in the policies and services of (Higher Education Institutions) (...) studied, but its prominence can in most cases be attributed to the positions adopted by large research funders." In fact, funding agencies that mandate comprehensive RDM [47, 61] represent one of the key pillars in RDM regulatory frameworks. Other pillars include journals and conferences that demand resource sharing [6, 65], industry partners setting requirements for collaboration [58], and institutional policies [17].

Pasquetto et al. [54] conducted two case studies of large scientific collaborations in astronomy and subseafloor biosphere studies. Their work focused on relationships between policies, open data, and infrastructure development. Their findings also highlighted the role of funding rules and researchers' commitment to prove compliance. Notably, they contributed a description of the interplay between policies and infrastructure: "while policy definitions for open data do shape scientific infrastructure, extant configurations of available infrastructure also shape open data policies in terms of what specific types of data are covered by the policies, and how these data are to be made available., to whom, and under what conditions." The authors further highlighted differences in sharing resources between scientists and the public [14, 50], or only within the science community [55, 74], concluding "that infrastructures are emergent, impact and are impacted by, policy, design, and practice [13, 45]." Our work links to this understanding, as it focuses on an equally weighted exploration of RDM components in agricultural science, including policies, practices, and infrastructure.

2.5 Infrastructure

Infrastructure is an RDM component involved at all stages and transitions of the RDM commitment evolution model. Early adoption can only be successful if the technical requirements are met to preserve and share research adequately, thereby overcoming barriers. At the same time, infrastructure plays an important role in sustaining commitment, as it needs to be maintained and updated, in order to match novelty and creativity in science [32]. Interactive systems further represent a strong base for incentives, relating to the model's Reward stage. In particular, gamification is considered a promising design tool to foster RDM commitment through peer recognition [11, 21, 33, 37, 48, 52, 59]. We will focus on the effects of communication in more detail in the following section, but want to stress, again, that this type of incentive is heavily dependent on the underlying technical infrastructure, as it represents a foundation for any kind of interactive gameful design strategy.

Generally, we distinguish between domain-tailored services and general data repositories [70]. At the intersection of the two, we find institutional repositories [73] that are tailored, yet possibly

spanning several research domains. General repositories include Dryad[1] and Zenodo[2]. Characteristics of those types of platforms are general accessibility across all fields of science and a resulting heterogeneity of preserved data. In contrast, domain-tailored data management services allow for customized preservation and sharing according to the specific needs of the target community. Strong examples include CERN Analysis Preservation (CAP)[3] in particle physics and Monitoring, Evaluation and Learning (MEL)[4] in agricultural science at ICARDA/CGIAR. The latter will be part of our research.

Domain-tailored services can reduce the effort needed for effective RDM, as they map researchers' practices closely [32]. Based on domain knowledge they can, for example, support scientists through auto-suggestion and auto-completion mechanisms and connect to existing databases. However, the design and maintenance of domain-tailored tools is also significantly more difficult and expensive compared to general platforms, especially considering the limited user base [22, 64]. Besides easing RDM, tailored tools can, though, offer unique use cases, as Feger et al. [32] showed in their qualitative study on CAP usage in particle physics. They found that domain-tailored tools can support strong use cases that benefit those who contribute to the system. The authors referred to "secondary usage forms" of technology. In the case of particle physics at CERN, they identified several of those secondary uses, including uncertainty coping, expertise location, and fostering of useful collaboration. These secondary uses have a clear motivational component that shows again the strong interdependency between infrastructure design and incentives that we further explore as part ouf our research in agricultural science at ICARDA/CGIAR.

2.6 Incentives and Motivation for RDM

As reflected in this Related Work section, motivation for RDM plays a central role in the RDM commitment evolution model and ties to most of the RDM components we identified, namely practice, infrastructure, and policies. This can be explained by the variety of motivations that need to be considered. The self-determination theory (SDT) by Ryan and Deci [62] is a psychological framework and macro-theory that provides guidance in this exploration of motivational drivers for RDM. SDT generally distinguishes between intrinsic motivation, different forms of extrinsic motivation, and amotivation. Intrinsic motivation relates to actions and activities that are perceived personally rewarding. In contrast, extrinsic motivation is based on external rewards and incentives, like promotions, salaries, and connected regulations like funding policies and conference submission rules. The organismic integration theory (OIT), one of several SDT mini-theories, describes several regulatory styles that help to distinguish more or less self-determined forms of extrinsic motivation [24]. Those include (from more to less self-determined) integrated regulation, identified regulation, introjected regulation, and external regulation. Examples in the RDM context could include researchers how conduct comprehensive RDM because they identify with the values of RDM and open science and have internalized this attitude (identified regulation), and researchers who follow RDM practices solely because their funding agencies demand it through policies (external regulation). We consider that studying this framework closer, in particular in relation to the various RDM components, is of great value. Thus, we map different types of motivation explicitly through our research.

Gamification, the "use of game design elements in non-game contexts" [25], is an example of how infrastructure design drives motivation. Game design elements like leaderboards can motivate some to demonstrate their engagement while competing with others for recognition and possible

[1]https://datadryad.org/stash
[2]https://zenodo.org/
[3]https://analysispreservation.cern.ch/
[4]https://mel.cgiar.org/user/login

benefits, all good examples of extrinsic forms of motivation. At the same time, game elements like badges can, if designed with researchers' needs in mind, foster positive peer recognition and allow to demonstrate one's identification with perceived valuable practices [33]. Notably, just like RDM services, there are both tailored [33], as well as generic approaches to gamification design that motivates and recognizes RDM practices. Open Science Badges (OSB) [21] represent a strong example for the latter. They are designed to promote and recognize sharing through three badges: Open Data, Open Materials, and Preregistered. The general nature of those game design elements led to their adoption among dozens of journals across numerous fields of science. Kidwell et al. [49] confirmed, through their quantitative analysis, that data sharing significantly increased for submissions to the *Psychological Science* journal after adopting those badges. And Rowhani-Farid et al. [60] concluded, based on their systematic literature review, that OSB were the only evidence-based incentive effectively promoting data sharing in the health and medical domain. The Association for Computing Machinery (ACM) introduced more fine-grained badges that tailor to research within the ACM's scope [1, 11], providing an example of game design elements located between tailored and generic design paradigms. Conceptual similarities between tailored and generic RDM systems design and RDM gamification elements, as well as their interdependencies, are both intriguing and subject to exploration in our research.

2.7 Summary and Research Questions

In this section, we reflected on the five RDM components we target in our work within agricultural science and ICARDA/CGIAR: practice, training, policies, infrastructure, and motivation. Reviewing related work, it becomes apparent that each of those components strongly impacts RDM effectiveness. Thus, stakeholders at various levels and stages of the scientific process need to review, in detail, current limitations and needs in relation to each of those components in order to foster sustainable RDM practices. However, related work, both quantitative and qualitative, also showed that these explorations are usually limited to either individual subject matters or high-level cross-topic explorations. In contrast, the RDM commitment evolution model [36] highlights the value of taking an in-depth joint study approach and exploring the various connections and interdependencies. However, to date, this model, derived from a cross-domain qualitative study, has not been validated in practice. Our research addresses this gap through quantitative (Study I) and qualitative (Studies II and III) studies within a single domain and organization: agricultural science within ICARDA/CGIAR. Our work addresses the following three research questions:

RQ 1: How are policies and regulations aligned with the organization's RDM goals?
Regulatory frameworks within RDM are some of the most often discussed strategies to resolve shortcomings. Institutional enforcement, funding policies, and publication regulations have been implemented in response to ineffective RDM and reuse. They also play a role in the adoption transition of the RDM commitment evolution model. Reflecting this strong role of regulatory frameworks in RDM, we dedicated Study II to the study of the effectiveness and constraints of policies in ICARDA/CGIAR.

RQ 2: What are practices and needs around training, infrastructure, and motivation?
Following our goal to provide an account of the five identified RDM components in ICARDA/CGIAR, we dedicated Study III to the qualitative investigation of training, infrastructure, and motivation. We address the following secondary research questions:

RQ 2a: How do current drivers of motivation impact RDM?
RQ 2b: What incentives can future RDM platforms provide?
RQ 2c: How does formal training and the current support impact RDM practices?

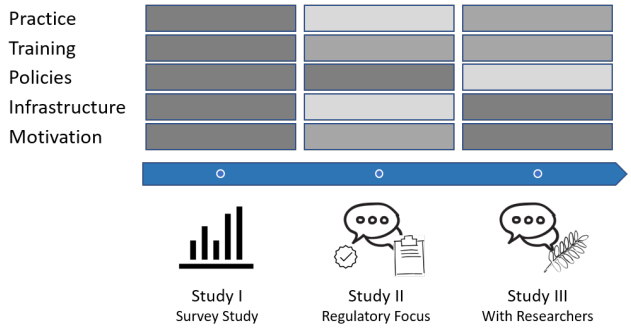

Fig. 2. Conceptual structure of this research revolving around three studies. Darker fields reflect a strong focus on a corresponding RDM component in a given study.

RQ 3: How does the stage-based RDM commitment evolution model apply to global agricultural science?
The RDM commitment evolution model [36] provides a bigger picture view on how RDM practices are established and hindered. To date, this model has not been validated in practice. Our work aims to close this gap through quantitative and qualitative studies (Studies I, II, and III) in agricultural science.

3 METHOD

To provide a detailed account of the various RDM components, we conducted three separate studies at ICARDA/CGIAR: a survey of researchers' existing data practices at ICARDA; semi-structured interviews with data managers about regulatory frameworks; and semi-structured interviews with scientists about infrastructure and motivational components. Figure 2 shows the conceptual structure of those studies in the context of the five RDM components Practice, Training, Policies, Infrastructure, and Motivation. Darker fields reflect a strong focus on a corresponding RDM component in a given study. As depicted, the survey study (Study I) strongly focused on all RDM components. Here, we aimed to initially map the current state of RDM in agricultural science at ICARDA/CGIAR and researchers' ratings of the suitability of the various components. We conducted two additional qualitative studies (Studies II and III) to gain a more in-depth knowledge of the various RDM components. This helped to further reason about the results of the survey study. While Study II focused on regulatory frameworks, in Study III we primarily investigated infrastructure and motivational components with a secondary focus on RDM practice and training.

In this section, we first introduce the chosen research environment: agricultural science at ICARDA/CGIAR. Next, we provide a high-level description of study participants and our recruitment process. We conclude this section by describing our quantitative and qualitative data analysis approaches. In this context, we also describe to which extent we made data and analyses openly available. More detailed accounts of study protocols and participants are provided in the corresponding study sections.

3.1 Research Environment

All three studies were conducted within the same environment: agricultural science at ICARDA and the CGIAR. ICARDA, the International Center for Agriculture Research in the Dry Areas, is an internal organization supporting farmers across Asia, Africa, and the Middle East. In June 2020[5],

[5]https://www.cgiar.org/how-we-work/accountability/gender-diversity-and-inclusion/dashboards/cgiarworkforce/

ICARDA had 421 international employees, representing 46 nationalities. The employees worked across 30 countries. Overall, 35% of the employees were female. And 17% of research staff were female. ICARDA is part of a global partnership of international organizations in the domain of agricultural science, called the CGIAR. The CGIAR network includes a total of 15 international centers across the globe, amounting to a total number of employees of 10,630. CGIAR centers like ICARDA are largely independent organizations with their own boards and regulations. As sustainable farming is a global endeavor that requires collaboration on a global scale, the various centers are collaborating heavily on various projects, thereby providing a valuable and special environment to study collaborative remote RDM practices. Recognizing the global challenges, the CGIAR partnership is currently undergoing a reform process referred to as **One CGIAR**[6]. We note that the One CGIAR reform is part of the mid and long-term 2030 Research and Innovation Strategy[7]. As such, strategic goals have been outlined, but the reform process is in its early stages of planning. Those goals include unified governance and an institutional integration involving the currently independent centers. Notably, the future architecture and use of RDM infrastructure is still subject of ongoing discussions. We consider it a strength of our research to be further involved in this global organizational transformation and reform process. Here, we note that one of the authors of this paper is a long-term employee of ICARDA and an agricultural science expert involved in the reform process. The other two authors consult ICARDA in the context of this research. One author is an agricultural science expert who specialized in qualitative explorations. The other one is a researcher anchored within the CSCW field. We recognize this diversity of perspectives and expertise as a strength of our work.

We consider global agricultural science as a highly valuable setting for our research, as the effectiveness of RDM and resource sharing impacts how we feed the world. The work of Salim et al. [63] represents a strong example. In this recent work, involving ICARDA and CGIAR research programs, the authors reported on the first systematic mapping of genetic diversity in cattle across the African continent. Their work is expected to help design more suitable breeding schemes and to improve resistance to diseases. Yet, this work is also an example of how data sharing is limited to closed repositories. In contrast, MEL DATA[8] provides an overview of openly accessible datasets. The list hints at the diversity of data and formats collected and shared through MEL. Those include quantitative tree planting data[9], yield maps[10], and coded interview results around the adoption of spineless cactus in livestock feed[11]. In this context, we further recognize the global changes caused by climate change. Here, the responsible treatment and sharing of agricultural science data impacts how both global and rural farmers install and maintain sustainable and productive processes. Generally, data plays an increasingly important role in global agricultural science to address these issues. In fact, the CGIAR Platform for Big Data in Agriculture[12] is another example of a data-centered initiative expected to create a lasting impact in agriculture. In addition, we want to stress that ICARDA and CGIAR have previously demonstrated great commitment in designing tools that enable and support RDM, thus providing a good basis to explore the infrastructure components and its interdependence to other components. Case in point is the MEL platform[13]. MEL stands for Monitoring, Evaluation, and Learning. ICARDA is the leading CGIAR center for the development

[6]https://www.cgiar.org/food-security-impact/one-cgiar/
[7]https://cgspace.cgiar.org/bitstream/handle/10568/110918/OneCGIAR-Strategy.pdf
[8]https://data.mel.cgiar.org/
[9]https://data.mel.cgiar.org/dataset.xhtml?persistentId=hdl:20.500.11766.1/FK2/O9LOGI
[10]https://data.mel.cgiar.org/dataset.xhtml?persistentId=hdl:20.500.11766.1/FK2/ZBH02G
[11]https://data.mel.cgiar.org/dataset.xhtml?persistentId=hdl:20.500.11766.1/FK2/UYPZUO
[12]https://bigdata.cgiar.org/
[13]https://mel.cgiar.org/

and maintenance of MEL. The platform is used by several additional CGIAR centers who contribute to the MEL development.

Given our goal to map various RDM components and their interplay in reference to the RDM commitment evolution model, we needed to conduct all studies within the same environment. Naturally, we expect that our findings will be highly relevant to various stakeholders within ICARDA, the CGIAR, and agricultural science more generally. Further, we note that our work is relevant for data managers, information scientists, and librarians involved in the design of infrastructure, training material, and support structures, well beyond agricultural science. In addition, we expect that data managers and administrators in organizations, politics, and funding bodies, will profit from our mapping of policy components and their alignment within the larger RDM commitment evolution. We expect that scholars who research practices around data curation and management will profit from our work that adapts a mixed-method study approach in a domain which mostly focuses on single or individual data management components, explored commonly in quantitative or qualitative studies. Here, our contributions towards a refined and established RDM commitment evolution model will likely benefit diverse scientific domains and the wider scientific discourse.

3.2 Participants

Our research was supported by ICARDA, the driving force behind the MEL platform development. For this reason, we focused on recruiting participants from within this organization. We distributed our various calls for study participation across ICARDA mailing lists and our contact persons. We further invited every person we came in contact with to forward the study descriptions to colleagues at their discretion. For the survey study, Study I, we received a total number of 23 completed responses. We asked participants to voluntarily indicate the center for which they were working. Most responses indicated ICARDA. Several participants were employed at other CGIAR centers like WorldFish. A total number of 14 participants indicated that they were working at a CGIAR center. Since we guaranteed anonymity and did not record any additional information related to the participants, we cannot exclude that the invitation to share the survey link reached scientists outside of the CGIAR. We decided to report on all submitted and completed responses, as we are convinced that doing so contributes to the validation of the RDM commitment evolution model [36].

We conducted semi-structured interviews in Studies II and III, giving us more control over the recruitment process. We recruited 17 data managers for Study II on regulatory frameworks and 13 scientists for Study III. Almost all participants of those two studies were working at ICARDA. Few informants were working at other CGIAR centers. We provide detailed information about individual center employment in the corresponding sections. Since participation in Study I was anonymous, we cannot analyze how many responses were cast from scientists who also participated in Study III.

3.3 Data Analysis and Open Data

In Section 4, we present the full analysis of the responses cast in the survey study. In this context, we want to stress that the entire data set has been made openly available as supplementary data. In addition, we also shared the exported survey sheet. Given the sensitive nature of the qualitative data collected in Studies II and III, we openly share selected resources from these studies. Generally, the qualitative studies were recorded, transcribed, annotated, and coded through thematic analysis [10, Section 5.2]. We used Atlas.ti to organize and code the interview data. We first created an initial set of codes that we discussed and refined within our team. Next, we iteratively constructed, discussed, and refined code groups based on those codes. In the last step, we iteratively created the themes based on those code groups. The themes are represented in the corresponding study

sections of this paper as subsections. Three researchers were involved in the data analysis of all three studies. Two of those researchers coded and discussed the entire qualitative datasets of Studies II and III. The third researcher observed this process through regular meetings and was involved in defining and refining the themes. In agreement with the consent form that we provided to our participants, we made openly available as supplementary materials the Atlas.ti code group reports and the interview protocols.

3.4 Method Reflexivity

In this section, we reflect on biases through three lenses: (1) expertise and employment history of the research team members; (2) biases resulting from the individual motivations of our study participants; and (3) analysis bias resulting from top-down research strategy and bottom-up data analysis.

We recognize that the scientific expertise and employment history of the research team members conducting this study impacts the method, execution, analysis, and reporting of the studies presented in this paper. We aim to be as transparent about this process as possible, to provide insight into our research work, its provenance, and to enable readers to make their own interpretations and conclusions. To this end, we confirm that all three studies were conducted by a research team consisting of three scientists. We note that one team member is a long-term ICARDA employee and agricultural science expert. This team member is involved in the MEL development and in the One CGIAR reform. Further, we note that the other two team members are scientists who consulted ICARDA/CGIAR for the purpose of conducting the reported research. One consultant is an agricultural science expert familiar with qualitative explorations in this domain. The other one is a researcher anchored in the CSCW domain. We note that there was no pressure to steer any part of the research in a direction that would support any political or scientific agenda. In fact, we stress that this mixed research team, mixed in both expertise and ICARDA/CGIAR employment history, allowed for open-minded and independent, yet competent, perspectives. Further, the entire qualitative dataset coding and reporting was handled by the two external advisors. Most of the interviews were also conducted by both external consultants. Yet, we also note that it is difficult or even impossible to completely prevent any form of bias, especially in qualitative studies. For example, one source of bias could be the external researchers' interest to prove their value. Yet, we also note that both researchers had primary employments or employment perspectives completely independent from ICARDA/CGIAR. Also, due to their diverse scientific backgrounds, the two consultants had no competing interests. Finally, their compensation was not contingent on the content or quality of the final report.

Second, we want to briefly reflect on the motivation of our study participants. We note that no remuneration was provided. Further, we note that we ensured that no participant felt pressured in any way to participate in a study. Since we did not provide any remuneration, it is important to consider any form of volunteer bias in our research. Informants likely participated in response to a form of intrinsic motivation, identified regulation, or introjected regulation. In fact, in Study I, researchers showed strong intrinsic and identified regulations for conducting RDM. This might partially indicate a study participation bias, although our interviews hint towards a more general community value system. While we explored sources of motivation for RDM in Studies II and III, we did not map participants' individual motivations. Therefore, we cannot provide an analysis of how statements might have been influenced by the motivation to participate in the study. Yet, we note that the data managers participating in Study II have a professional motivation to report on current practices and to improve data practices. Related to Study III, we note that we focused on sampling concrete experiences from scientists, rather than opinions, to keep information factual and to prevent bias as best as possible.

Role	*Organization*	*Nationality*	*Gender*
PM / Information Scientist	Major Research Funding Agency	German	Female
Postdoc / RDM researcher	Agricultural Science Organization	Indonesian	Female
PM / Former Science Editor	Physics Laboratory	US American	Male
PM RDM tool development	Agricultural Science Organization	Italian	Male
PM / Information Scientist	Physics Laboratory	Greek	Female

Table 1. An overview of the expert judges involved in the survey development. *PM* stands for *Project Manager*.

Finally, we note that while our qualitative data analysis process described by Blandford et al. [10, Section 5.2] is bottom-up, our overall research strategy is aligned with our research questions and impacted by related work. We argue that this does not represent a contradiction. In fact, Blandford et al. [10, p. 54] describe a continuum of overall qualitative research approaches between bottom-up and top-down: "In some cases, the literature review will have guided all the data gathering and analysis. In other cases, you think you have finished your analysis, realise that someone has already written a paper with similar findings to yours [...] Usually, it is somewhere between these extremes." However, we note that a strong existing research framework might impact the data analysis, both intentional or unintentional. Such effects are referred to as *domain summary*, where themes are "organised around a shared topic but not shared *meaning*" [15], and characterized by data topics which are interpreted as themes [16]. In order to prevent an analysis bias as best as possible, two researchers with different sets of expertise and connection to RDM and the CSCW literature analysed the entire dataset, observed by a third researcher on a regular basis.

4 STUDY I: MAPPING KEY RDM DIMENSIONS

In this study, we aimed at systematically mapping the five RDM components (i.e., practices, training, policies, infrastructure, and motivation) through a survey at ICARDA/CGIAR. In this section, we describe the study procedure and present key results.

4.1 Procedure

4.1.1 Survey Development. First, we iteratively defined components out of the RDM commitment evolution model, focusing on practice, training, policies, infrastructure, and motivation. This process was closely aligned with the stages and transitions of the RDM commitment evolution model [36]. Next, we reviewed scales and work related to all dimensions and iteratively generated items for each of the five scale dimensions. Following, we invited five expert judges to review the generated items and dimensions. We recruited personalities heavily involved in the design or support of RDM activities. Table 1 provides an overview of the experts showing that we invited both agricultural science experts, as well as experts from other fields. We decide to include non-agricultural science perspectives, as we aimed for a survey that represent a diverse set of perspectives. The expert judges were instructed to review all items, as well as the survey introduction, to comment and provide suggestions as they see fit. We asked them to review wording, comprehension, fitness of scope, applicability, and invited them to add new items if they believed that certain aspects were not sufficiently covered. Based on the sum of suggestions, we refined the items again.

In the next step, we invited five population judges for 30 minutes formative interviews. All interviewees were active scientists with different roles, backgrounds, and nationalities. Table 2 provides references to the participants and their demographics. We asked the informants to review the instructions and all items and to think out aloud in this process. The interviewer asked for

Background	*Role*	*Nationality*	*Gender*
Mechanical Engineering	PhD Student	Canadian	Male
Psychology	Postdoc	German	Male
Electrical Engineering	Postdoc	Italian	Female
Computer Science	PhD Student	German	Male
Applied Linguistics	Postdoc	Austrian	Female

Table 2. An overview of the population judges involved in the survey development.

additional information about comments and concerns. After completion of this formative study, we discussed the feedback and refined the survey one last time. We made the final version of the Qualtrics survey openly available in the supplementary materials.

4.1.2 Dissemination. We shared a survey description and the survey link with our contacts at ICARDA and through ICARDA mailing lists that predominantly included scientists. In the survey description, we stated that the study targets scientific personnel. We further assured participants that they would remain completely anonymous. We estimated that survey participation would require around 15 minutes. We ran the survey over a period of six weeks.

In total, we received 23 complete responses. Amongst those, 14 specified the center that they worked for (13: ICARDA; 1: WorldFish). We did not make this field mandatory, as we worried that participants might perceive this as a contradiction to the promise of full anonymity. The remaining nine responses might include responses from ICARDA/CGIAR personnel who decided against specifying their respective center. In addition, some scientists might belong to more than one center which could have also caused a decision to not specify a single center in the respective field. Finally, it is possible that our contacts and mailing list recipients forwarded the study invitation to scientists outside of CGIAR. We asked them to share our message with scientists, but did not explicitly mention that this study is exclusive to ICARDA/CIGAR. Analyzing the scientific background specified by some of the participants, we perceive this as a likely cause. However, we cannot exclude participants based on their background, as scientists working at ICARDA/CGIAR come from a range of diverse scientific fields. In addition, we do not consider removing responses as a useful action. While our research focuses closely on agricultural research at ICARDA/CGIAR, we find that a most complete and diverse sample in this first study helps to contribute experiences around RDM commitment evolution. For accuracy, we do, however, report parts of the study results in a manner separating responses that indicate, or miss to indicate, affiliation with ICARDA/CGIAR.

4.2 Results

All items reported in this section are based on a 7-point Likert scale (1: Strongly Disagree; 2: Disagree; 3: Somewhat Disagree; 4: Neutral; 5: Somewhat Agree; 6: Agree; and 7: Strongly Agree). We present results for each of the RDM components practice, motivation, training, and infrastructure next. Each section features one diverging stacked bar chart that is based on *all* collected responses. All those charts, as well as plots based entirely on responses that indicate ICARDA/CGIAR affiliation, are available as supplementary materials.

4.2.1 Practice. As depicted in Figure 3, the participants rated current practices in a neutral to slightly negative manner. For example, responses show a slight tendency towards RDM being "far away from systematic RDM" (mean_all=4.6; median_all=5.0; mean_cgiar=4.8; median_cgiar=5.0). There is also agreement that "experimental resources and data shared by colleagues are usually not

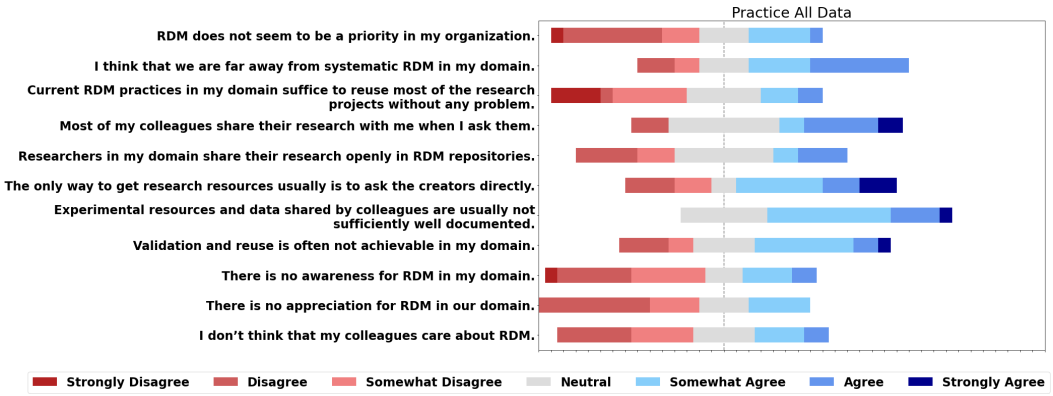

Fig. 3. The plot shows rated agreement of *all* responses to statements related to *practice*.

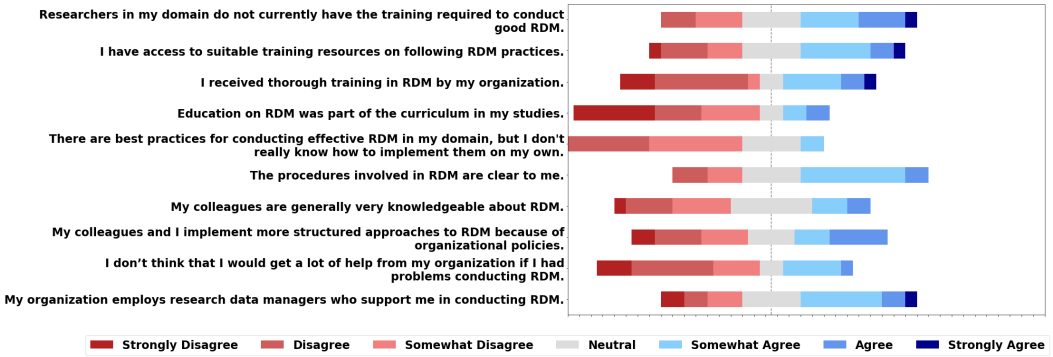

Fig. 4. The plot shows rated agreement of all responses to statements related to *training and support*.

sufficiently well documented": mean_all=5.0; median_all=5.0; mean_cgiar=5.3; median_cgiar=5.0. However, there is slight disagreement with the statement regarding "no appreciation for RDM": mean_all=3.2; median_all=3.0; mean_cgiar=3.5; median_cgiar=4.0.

4.2.2 Training and Support. The results related to overall training in the research domain and institutional support provide first directions related to the RDM state we mapped in the previous section. While, as shown in Figure 4, the survey participants disagreed about not expecting "a lot of help from (their) (...) organization if (they) (...) had problems conducting RDM" (mean_all=3.1; median_all=3.0; mean_cgiar=2.8; median_cgiar=2.0), the scientists also indicated that they lacked formal training in their studies ("Education on RDM was part of the curriculum in my studies": mean_all=2.7; median_all=2.5; mean_cgiar=2.8; median_cgiar=3.0). This echoes a more positive attitude towards provided RDM support, as opposed to the formal training scientists received during their academic studies.

4.2.3 Infrastructure. Results related to infrastructure are notable, as they provide a rather positive account of researchers' assessment of provided resources. As depicted in Figure 5, the respondents leaned towards indicating that that their "organization's RDM infrastructure is suitable" (mean_all=4.7; median_all=5.0; mean_cgiar=4.9; median_cgiar=5.0) and that "the RDM tools (they) (...) have enable (them) (...) to manage data efficiently" (mean_all=4.9; median_all=5.0;

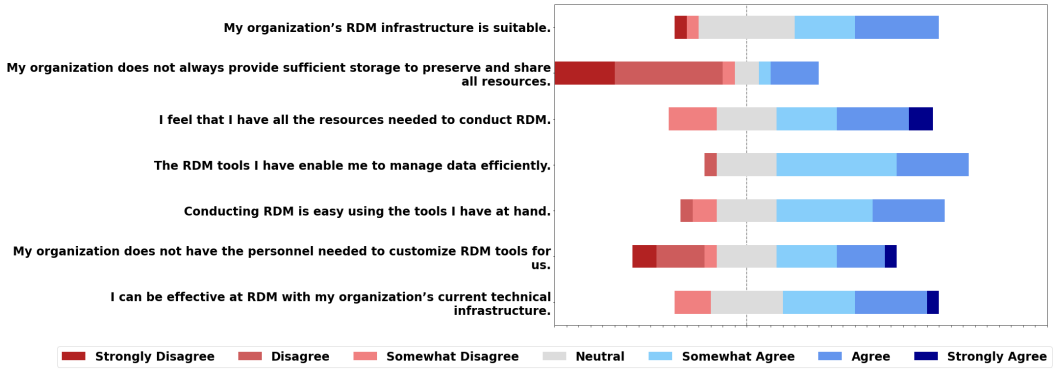

Fig. 5. The plot shows rated agreement of all responses to statements related to *infrastructure*.

mean_cgiar=5.1; median_cgiar=5.0). However, more neutral responses were cast regarding the availability of personnel needed to customize RDM tools for the scientists (mean_all=4.0; median_all=4.0; mean_cgiar=3.9; median_cgiar=4.0).

4.2.4 Motivation. Finally, Figure 6 provides an overview of the distribution of responses regarding different types of motivation for RDM. We notice the strong and steady agreement to intrinsic forms of motivation (green). The responses also show a strong form of identified motivation (gray). Introjected and extrinsic forms of regulation play less dominant roles in the motivational structures of researchers in our study. Yet, policies and funding requirements play a role as the following statements show. "I conduct RDM because of conference and journal publication policies": mean_all=4.3; median_all=5.0; mean_cgiar=4.6; median_cgiar=5.0. "I conduct RDM because my organization demands it through policies": mean_all=4.5; median_all=5.0; mean_cgiar=4.8; median_cgiar=5.0.

4.3 Summary

The results of this study reveal a neutral to slightly negative view on current RDM practices. The participants also did not perceive strong RDM training and support structures, although they rated formal education during their studies even lower. In contrast, the survey respondents indicated consistently stronger agreement towards the suitability of the technical infrastructure in place. Finally, the researchers' responses show strong forms of intrinsic motivation and identified regulation.

5 STUDY II: THE REGULATORY FRAMEWORK

The results of Study I show a positive attitude towards the suitability of the current RDM infrastructure. Yet, the survey respondents also expressed a slight agreement towards being "far away from systematic RDM". They further indicated that research is not usually shared openly in RDM repositories and stressed that "experimental resources and data shared by colleagues are usually not sufficiently well documented". The survey analysis further highlighted the impact of regulations and policies, with particular regard to organizational policies. Conference and journal policies, funding regulations, and supervisor enforcement further impact RDM compliance. To further our understanding of these results and to explore the impact of regulatory frameworks in the commitment evolution model, we conducted a second study that focused on mapping current policies and regulations.

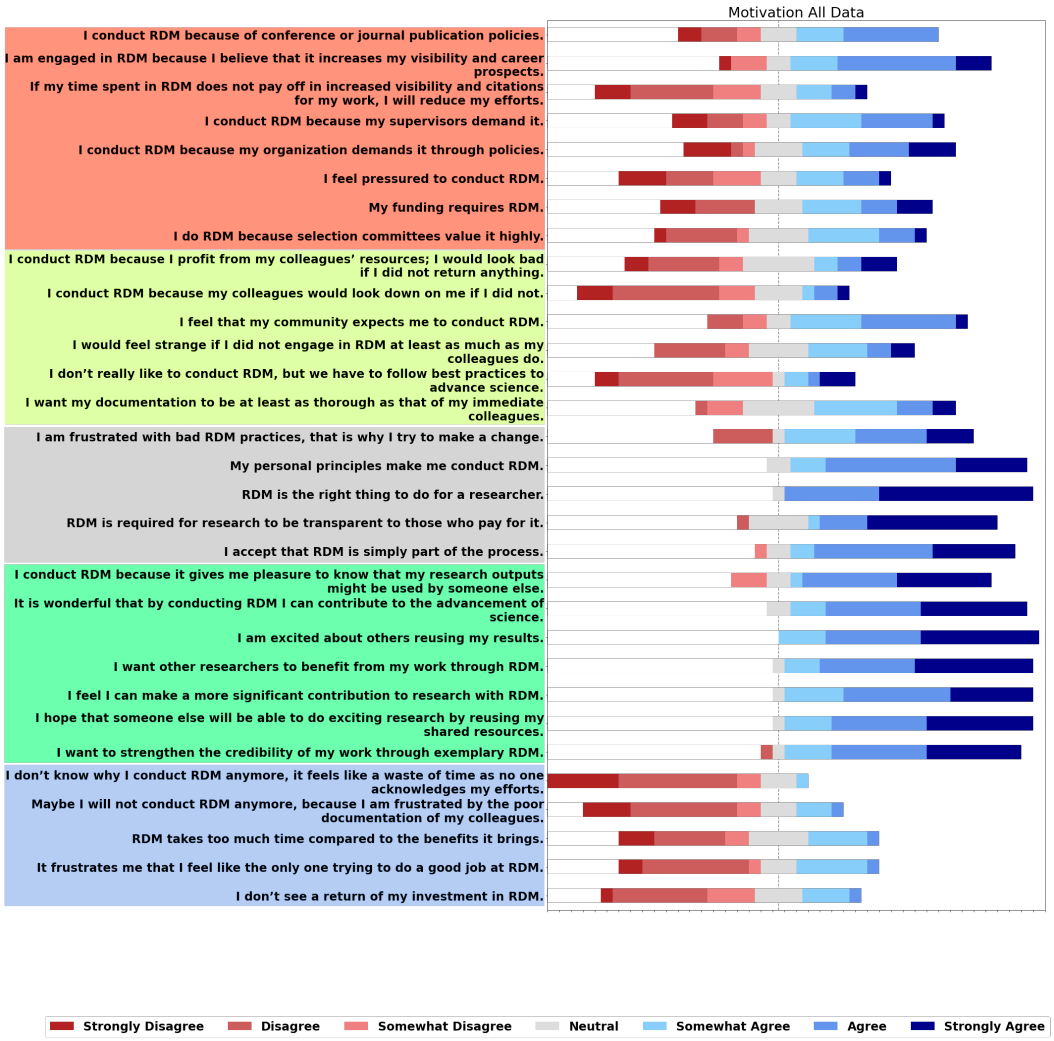

Fig. 6. The plot shows rated agreement of all responses to statements related to *motivation*. The statements are designed to cover all types of motivation and the different forms of extrinsic regulation described by the Self-Determination Theory. The following color coding applies: External regulation: Red; Introjected Regulation: Yellow; Identified Regulation: Gray; Intrinsic Motivation: green; Amotivation: blue.

5.1 Procedure

We conducted a qualitative study with 17 data managers at ICARDA/CGIAR. We decided to recruit data managers for this study, as they are primarily responsible for introducing and designing policies, and for supporting and checking compliance with regulations at all levels. Given the sensitive nature of the topic and the limited number of data managers, we needed to take special considerations into account regarding the protection of participants' anonymity. For this reason, and in agreement with our interviewees, we are not reporting individual characteristics of participants for this study. We refer to the informants as P1-P17 in this section. Based on their requests, P7 and P8, as well as P14 and P15 were interviewed together.

The semi-structured interviews revolved around five key parts. In the first one, we asked the data managers to introduce themselves, to describe their roles and duties, as well as key stakeholders they work with. Next, we explored their perspectives regarding the impact of FAIR data. In the third part, we asked about challenges, especially across four dimensions: technical, cultural, political, and institutional. We continued by asking about concrete examples related to their interaction with the CGIAR Open Access and Data Management Policy and its Implementation Guidelines. Finally, we concluded by asking about participants' needs and required support in their task to stimulate and foster comprehensive RDM.

5.2 Findings

We present findings from this study through the lenses of three themes that resulted from our analysis: Policy Integration, Overlap, and Administration.

5.2.1 Integration. This theme relates to the challenge of integrating existing policies, as well as new regulations, into institutional frameworks and researchers' workflows. The participants highlighted that based on the different mechanisms of policies, this integration process differs heavily. A key challenge for policies that are strictly *mandating* actions, for example, lies in creating awareness. Tools and technical infrastructure can support this process by formalizing regulations through visible platform components, thereby outlining new measures and their concrete requirements. Examples include mandatory fields and reports in submission forms that need to be completed as part of researchers' standard publication workflow. In contrast, another type of policies offers *incentives*, e.g. financial ones for compliance with RDM practices. Other more subtle and implicitly motivating policies might more strictly require resource attribution, thus providing incentives for primary data creators to share their resources more openly. Participants stressed, however, that introducing those institutional policies might not be sufficient on their own to create awareness. Also, incentives might not be strong enough to engage the community at large. In response, the data managers indicated organizing workshops and onboarding sessions and stressed the effect of well-known and respected scientists ("Champion" – P6) advocating the value of new regulations.

The last statements related to one of the roles of workshops: creating awareness and promoting new regulations. Another motivation is to train researchers on how to fulfil the requirements. This is highly important, as the informants stressed that guidelines are often not clear to the research community. Although we did not explicitly intend to study requirements around RDM training in this study, the education component came up repeatedly. The interviewees emphasized the value of policies as a form of implicit training resource. Policies that are very detailed and that relate to steps with which researchers are familiar, can provide a form of check list that can easily be followed, as they provide detailed accounts of what needs to be done to fulfil minimum requirements. Clearly, designing such effective policies is a challenge that needs to be met by including both domain scientists and data managers / policy makers. Installing such mixed teams is also key in developing trust with the science community (P1). Finally, most interviewees highlighted the value of regulations as tools that provide a kind of common protocol between data managers and scientists, a common communication channel that makes exchange easier. Generally, as P2 stressed, policies ultimately need to be designed as a service for research in general, and be perceived accordingly, in order to gain trust.

5.2.2 Overlap. Related to overlap, we present insight into policy hierarchies and conflicts that pose problems for both scientists and data managers. The first type of overlap we identified relates to inter-center regulations. The data managers described challenges posed by conflicting institutional regulations in collaborative projects on different levels: CGIAR centers have individual regulations that can conflict in inter-center collaborative projects; and center policies can conflict with CGIAR

regulations. In particular, questions and conflicts around ownership have been described by the participants. This is a notable finding, given the global scale of agricultural science in general, and ICARDA/CGIAR in particular.

Based on our data analysis, we described two additional types of policy overlaps. P16 described several examples of how national data laws conflicted with data collection and sharing practices. The data manager stressed that this is also caused by a differing understanding of what constitutes FAIR data. Notably, these data laws can further conflict with funding policies, an additional type of policy we identified. Much of the agricultural research conducted is sponsored by donors that introduce there own requirements and rules, conflicting possibly with institutional, inter-institutional, and national regulations.

5.2.3 Administration. The last theme relates to the administration of implemented policies. Here, monitoring of RDM activities is both a requirement to check compliance with current rules, as well as a tool to track the effects of RDM. The latter is a form of external monitoring that aims at finding out *who* is using data and for *what* purpose, in order to further improve processes and tools. This monitoring process, however, is highly resource intensive, as our participants stressed repeatedly. As part of this process, they also uncover additional types of policy issues, including mismatches between center regulations and the practices and needs of local farmers. In this context, P1 also stressed the importance of domain scientists in the design of tools and policies by implying that an earlier focus on software programmers in a data unit caused issues as they were good at developing repositories, but not at understanding the submitted data.

Most study participants reflected on the financial ramifications imposed by RDM policies. Providing staff that support and train researchers in mandatory practices is expensive. The same is true for employing staff members that check compliance with regulations — at least if this check is supposed to be a thorough one. Here, the informants discussed the roles of donors extensively. P1 stated that donors think that throwing "money at it for 6 months (..) will fix it", while in fact a more structured approach to RDM financing is required that takes into account the employment frameworks of support staff. Here, P10 stressed that it would be more important to provide suitable human resources, rather than financial ones. Generally, there was a notion of policies conflicting with or ignoring financial realities.

5.3 Summary

This study focused on regulatory frameworks around RDM at ICARDA/CGIAR. Our interviews with data managers showed that policies need to be designed, maintained, and updated in relation to several related dimensions. They need to both reflect and impact current practice, existing infrastructure, regulation hierarchies, and financial and human support structures. Notably, effective policies do not only act as an initial extrinsic driver for researchers' RDM commitment, but serve as a common protocol between scientists and data managers. Ineffective policies, however, represent barriers if they conflict in collaborative settings and contradict national or funding regulations. In order to ensure that regulations are aligned with practices of researchers and farmers, and technical infrastructure is suitable to implement and follow mandated activities, data managers, software developers, and domain experts need to be involved collaboratively at all stages of the process. This finding is particularly reflective of the current ongoing One CGIAR transformation and reform process that represents an opportunity to resolve policy conflicts and re-design tools that supports future RDM regulations on an inter-center global scale.

Reference	*Center*	*Background/Specialization*	*Role*	*Gender*
P1	ICARDA	Tropical Agriculture	RDM Coordinator	Male
P2	WorldFish	Rural Development	Impact Assessment Researcher	Male
P3	ICARDA	Environmental Science	Scientist	Female
P4	ICARDA	Agricultural Livelihood	Scientist	Male
P5	ICARDA	Environmental Engineering	Team Leader	Male
P6	ICARDA	Agricultural Science	MEL Specialist; Scientist	Male
P7	CRP-RTB	Agricultural Engineering	Gender Specialist	Female
P8	ICARDA	Agronomy	Upper Research Management	Male
P9	CRP-RTB	Agricultural Science	Scientist	Female
P10	WorldFish	Fisheries Science	Scientist	Male
P11	WorldFish	Agricultural Science	Research assistant	Male
P12	WorldFish	Sustainable Aquaculture	Scientist	Male
P13	WorldFish	Agricultural Economics	Scientist	Male

Table 3. An overview of the study participants. Most participants were associated with ICARDA and WorldFish. The gender distribution is reflective of the overall CGIAR research staff statistics.

6 STUDY III: INFRASTRUCTURE AND MOTIVATION

While the previous Study II was limited in scope to regulatory frameworks and data managers as informants, in this final Study III we explored the remaining RDM components practice, training, infrastructure and motivation through a qualitative study with 13 scientists. In this section, we first detail the procedure. Next, we present our findings across four themes: Practice, Platform Design, Motivation, and Organization.

6.1 Procedure

We recruited 13 scientists at ICARDA/CGIAR for semi-structured interviews for this study. Here, we provide details about the participants and the interview protocol.

6.1.1 Participants. All 13 participants were working at CGIAR centers at the time of the interviews. As depicted in Table 3, six of the interviewees were working at ICARDA. Five informants worked at WorldFish, and two worked in the CGIAR Research Program on Roots, Tubers and Bananas (CRP-RTB). The average age of all participants was 42 years (Min: 27, Max: 62). Of all participants, 23% were female. This is reflective of the official employment statistics[14], indicating that 29% of CGIAR research staff were female. We assured participants that we would not detail individual nationalities. However, we can provide an alphabetical list of the diverse set of countries of origin: Austria, Bolivia, Egypt, France, Germany, Italy, Jordan, Kenya, United Kingdom, and Vietnam. The average interview duration was 50 minutes (Min: 40, Max: 59).

During the first stage of the recruitment, we openly shared a study invitation and the study description on an ICARDA email communication channel for scientists. In addition, we asked our contacts to promote the study within their networks. Further, we asked our informants after completion of the interview, if they were willing to refer us to colleagues within their networks. We came in contact with researchers from other CGIAR centers through this type of snowball sampling. We note that we did not provide remuneration to the participants. Potential biases of this approach are discussed in Section 3.4, Method Reflexivity.

[14]https://www.cgiar.org/how-we-work/accountability/gender-diversity-and-inclusion/dashboards/cgiarworkforce/

6.1.2 Interview Protocol. Our semi-structured interviews were closely aligned with our interview protocol featuring 20 questions, classified into three categories and nine sub-categories. The semi-structured approach allowed us to further explore interesting aspects related to any of those questions and to pose new follow-up questions.

In the beginning of the interview, we asked the participants to introduce themselves and to briefly characterize their typical workflows for handling data. After this introduction phase, we followed up with the three interview categories:

- In the first part, we inquired about the communication and information architecture in agricultural research within and across CGIAR centers. In particular, we were interested to learn what role the MEL platform played in this architecture. We further inquired about how CGIAR centers differed with regard to RDM practices and how those differences were anchored in technology use. Finally, we asked about current practices around information needs and communication practices across different user groups and projects.
- In the second part, we wanted to learn about the role Role of MEL and data repositories in agricultural science RDM and reuse. We inquired about challenges for effective RDM, as well as barriers that could be addressed through RDM tool design.
- In the last part, we focused on the alignment of contributions to the knowledge repositories with users' goals and requirements. In particular, we asked about motivations for contributing to MEL and similar platforms, and future use cases to stimulate contributions.

6.2 Findings

We present findings from this study across four themes. In Practice, we summarize our findings regarding a variety of perspectives that show how RDM is currently done. This theme links closely to the Platform Design theme that offers a form of reasoning for several of those observations. The Motivation theme provides further insight into potential incentives that provide opportunities to motivate RDM. This theme is again closely linked to platforms, as participants stressed how use cases supported by tools like MEL could profit RDM contributors. Finally, Organization relates closely to challenges on a global organizational level, including inter-center collaboration and tool usage. In addition, it extensively covers considerations and requirements involved in the ongoing reform and transformation towards the One CGIAR.

6.2.1 Practice. This theme covers a variety of perspectives that show how RDM is currently done in agricultural research at ICARDA/CGIAR and in the context of platforms like MEL. Generally, there is consensus among the participants that services like MEL are valuable as they provide the necessary tools to conduct comprehensive RDM. Also the value of following RDM practices is not contested among our participants. Yet, the challenge remains to transform this acceptance into routine. In this context, P1 stated that *"many scientists have now accepted the use of MEL. But they are not really adapting their work procedure to MEL."* We see evidence that this lack of integration into everyday workflows is partially grounded in lack of fitness of current services. For example, several informants reported that they did not use MEL to link to datasets, even though this is one of the platform's features: *"if internal someone asked this, ah please can I have the last season rainfall data, I would not send you the handle from MEL and say, ah we uploaded it there. I would really send him the raw data as I have it and it's easier for me to communicate more directly during this way."* (P5) This is a form of current communication practice that largely ignores RDM architecture in place. Reasoning behind this observation is documented in the Platform Design theme.

While we did not focus on regulatory frameworks, most participants discussed forms of *enforcement* as a RDM strategy that is closely tied to center regulations and donor rules. Adding to our

findings from Study II, we recognize that regulations can only be effective if the technical infrastructure provides a suitable environment to support those rules. P6 provided a requirements-based account of his experience: *"Open-access is a second stage. First, you need to have a good systematization internally and then you can push easier on open-access."* Closely related to enforcement, the interviewees shared experiences with funding practices. In particular, and adding to Study II, several extensively reported on a double burden to account for both donor and institutional rules. P2 discussed this aspect in the context of MEL, which in the specific case was not required by the donor, but by the center: *"the person in charge for the project is really thinking its priority is to fulfill the donor requirements and not really to basically fulfilling also the organizational requirements in terms of planning and reporting and so on. And so basically people is quite resistant because feels like there is a double process, so one for the donor and one for the MEL platform and they sometimes struggle to understand the value of the platform itself."*

The interviewees shared various concerns and experiences that link to training and educational support. The most common statements related to the fear of judgement regarding the quality of shared data. P9 echoed this aspect through a concrete experience: *"It wasn't I don't want to share my data because I'm afraid that people use it and publish it and take away opportunities that I could have used. No, it was the fear that the data quality wasn't sufficient to be made public."* Additional concerns were raised regarding the selection of data. Several participants stressed that support was required to determine how certain types of data can be shared. Here, a common concern related to privacy regulations. P12 provided an example related to pathogens and farmers: *"So, the information might be, it would be from, like, a pathogen, for example, but that will be linked to an identifier, which will be a farm, a location, the name of the farmers, you know. So, we don't want that to be leaked to the general public. So, data management and privacy is going to be a key and long term, you know, long term storage, medium storage, immediate access."*

6.2.2 Platform Design. Our data analysis showed that the technical infrastructure and tools, in particular the MEL platform, play a central role in all RDM considerations and at all stages of the RDM life cycle. The study participants informed about two major discussion points: (1) platform usability and resulting barriers; and (2) the service design process.

Foremost, scientists described the steep learning curve of MEL, in particular in relation to common and commercial data sharing tools. Here, the informants reflected mostly on general-purpose consumer services like Google Drive and Dropbox, which several participants reported to use for the sharing of scientific data. The participants asked for simpler and more intuitive interaction with MEL. P9 echoed this request in the following way: *"For a broad use, for most people, it would help if it was simpler. Just asking you the absolute minimum number of information in the way that you can handle this. If it's too much, it's just—people just reject it. They get frustrated."* P7 added onto this in the context of infrequent platform interaction: *"Because it wasn't my only or main responsibility. I was a support. I was doing other stuff. So when I had to go back, I would get lost and I couldn't find my way around, had to call and say, so where's that part of the uploading of the innovation? And they say, ah we changed that to change it into this and this other tab."* This is particularly true for scientists who do not regularly engage in RDM activities and do not regularly need to use platforms like MEL. Our informants further discussed this aspect with regard to continuous and frequent service updates, e.g. P1:

> I have some problem with the constant improvement of the platform. Absolutely, I am sure that it's better now because it's more efficient. But since sometimes one of my tasks is also reporting the issue and to check if they were fixed, I find it very frustrating sometimes because I don't know nothing about programming to be clear. Sometimes it feels that the programmers that don't really know what the improvement is for. So

> they implemented the new features, but since they are not really using the platform they are not understanding what problem it will generate to the existing data. [...] But I repeat I don't know anything about programming so I don't know how hard it can be. But I think that sometimes if more strict collaboration with the user maybe can be useful because many user probably can figure that some problem can emerge. – P1

This statement further echoes an interesting aspect that emerged from several interviews with participants who had shared responsibilities, involving research, project management, and data management. They described communication barriers with software programmers and stressed the value of interdisciplinary teams involving scientists, data managers, and software developers. According to the interviewees, such setups are particularly effective in customizing the platform according to specific needs, without causing new issues due to changed meta data or added complexity for the others who are not affected by the customization. The following statement shows how important platform customization is, in particular in collaborative settings that extend beyond an individual center:

> Yeah, I think there are different aspect of the customization. But in my opinion there is one that is absolutely crucial. [...] So, creating an environment for the bilateral projects that is more customized on the reporting system of the specific donor will strongly facilitate the utilization, the use of the platform itself and also facilitate the value of the MEL platform for the project and for the donor, which is I think will rise consistently the profile of the platform also not within our network, but outside. – P2

Overall, the informants described a variety of benefits that MEL offers already at its current stage. Most important, the platform provides means to reason about links and dependencies of projects, with regard to both human and topical considerations. Several participants called for a consolidated system that would allow to further link between stored resources.

The last consideration we reflect upon in this theme is impact assessment. The interviewees stressed how an effective tool can help track impact within the organization. For example, P10 stated how a detailed view on internal processes supports this: *"I find it useful, I suppose repository in itself for finding WorldFish documentation and reports of things that are harder to find elsewhere and also to drill down to specifics in terms of specific people in the organization or specific projects. And then to kind of summarize what impact those projects have had in their own right through kind of publications."* In principal, MEL and RDM platforms can provide similar impact analysis for external use of data as well. However, privacy policies limit both development capabilities and users' abilities to track use of specific resources: *"Only few people I would say between a 5% and 10% that come back to us and say, 'Ah, thank you very much. I've used your data to do this such and such.' The remaining, they just download it. [...] Unfortunately, it's less traceable. If the open access policy would allow us instead to create an interface to request first the information, setting up and maybe storing the personal data of the researcher downloading our data, and then reaching with a survey over time to know, we could have done maybe some interesting feedback or also to our donors that they gave us the funds to generate data."* (P6)

6.2.3 Motivation. The participating researchers described the full spectrum of types of motivation for conducting RDM. To some degree, almost all highlighted the value of open and reusable knowledge to benefit decisions of rural farmers and to create agriculture practices that are fit to respond to today's challenges. This strong expression of intrinsic or identified motivation is likely correlated with the researchers' goals for global sustainable agriculture. Still, the analysis showed that for most scientists it is key to consider additional benefits and incentives to stimulate comprehensive RDM. On the other end of the spectrum are financial incentives. Most participants

discussed them in some form. P7 echoed this in the context of bonuses: *"And in addition to other variables, including quality of the research and publications generated and the teams get the highest scoring teams will get a small bonus for the year to go using some of the activities they're conducting. So that has made some kind of, how do you say this, motivation for teams to put up their information, to show how well they're doing and so that we can also report the system office, but they also get a benefit of the bonus."*

Visibility is another source of motivation. Most informants stressed that having a platform to get their names out and to share resources that others can cite, is clearly a motivation that links to an increase in visibility. However, we know from the Practice theme that researchers are worried about releasing their data and code as they fear judgement for erroneous or low quality data. Several study participants stressed that a mandatory external quality check that is part of the submission to MEL could address this concern. P8 echoed this as follows: *"And the third incentive at least for a scientist is that if through MEL they know their publication or documents will be on one side properly curated and stored, and on the other side for the document they want to share widely. So this is a third motivation. Or I should say four. Because one is the repository of the document and quality control and the fourth one is the capacity to have the information disseminated or accessible to a wider audience."* Notably, this perspective turns obligatory quality checks of data managers into an incentive, if this step is perceived as a quality check service that comes with helpful feedback. Platforms can further support this feeling of impact and visibility through UI elements, as also P13 stated: *"let's say if a published dataset we can see clearly that the number of downloads, for example, which means, let's say, the number of people that are actually interested in my datasets and they are downloading that and they're using it, or how many people are actually reaching out to the data management team in the case, for example, where there is embargo and asking could we utilize that dataset. Okay, you know, that sort of for me as a researcher signals that there's a lot of interest in my research, in my dataset, and that's an incentive in itself."* P13 continued to link this visible and personal impact assessment to a motivating experience around small research grants: *"They provided, for example, it was like a challenge where the person with the most submitted datasets, most archived datasets received like some grant. It's a very small grant, but nonetheless, it's something that sort of appreciate somebody's effort. And I saw that that stimulated a lot people to, you know, to submit datasets."*

Another key motivation for using and contributing to RDM tools was automated reporting and knowledge extraction. On the one hand, the participants described a vision of MEL as a tool that would aggregate information from different sources, including scientific, financial, human resources, and related to the donors. They imagined that such a system could ease reporting through the semi-automated aggregation of resources into a common template. Notably, several agricultural scientists went even further and imagined that future systems could build around Machine Learning (ML) algorithms supporting them even more effectively in time consuming reporting tasks:

> I mean, if there would be a way that through machine learning automatically this curation is done. Let's say you have any form of any kind of excel, CSV, text, whatever and you just put it in the converter and the machine does it and you just quickly manipulate some things if it's not correct. Then I think in such a world, also draft data, that would make much sense to send it in a curated form [...] – P5

6.2.4 Organization. This last theme relates to two major considerations: (1) challenges for RDM in inter-center and inter-institution settings; and (2) requirements, hopes, and needs in the context of the ongoing One CGIAR reform process that allows for valuable live insight into a changing global organization.

Adding to our findings related to regulatory hierarchies in Study II, we found in this study that also the use of multiple diverse tools across centers is a major barrier. The following statement

from P2 shows how these considerations either lead to added workload on the researchers' side or to missing data in a platform:

> But we had a couple of, I think two or three cases maybe, where we projects for instance funded by another research program or by another institute. [...] It was a little bit tricky to manage that project specifically and facilitate the project in MEL platform. Mainly because other centers or other research program using a different platform, so we were in front of a dilemma. And then we start thinking, right, so our researcher, they have any way to report this stuff in the other system, right? So are we going to have them really to report this in our platform too, right? So it's, we felt like it was not really fair, right? So we consider to basically just receiving for, just mapping the final deliverables like final reports, their activities, and not really to go too much in deep in that sense.

Besides the question of choosing common tools, generally, our data analysis showed that informants reported issues related to differing RDM support structures and regulations in collaborative settings. This largely echoes findings from Study II. In addition, we find that a center's general attitude and culture towards RDM and tool usage impacts researchers' awareness of the suitability of practices in collaborative settings. Based on those findings, we see an opportunity to recognize both the value of good practices and shortcomings through inter-center collaboration. P7 echoed this understanding as she talked about an experience she made: *"So we found that the centers using MEL had done that much better because they were using it for center purposes. [...] So you can have a web page for your project and you can upload all the information in that that will reduce the double reporting burden because it reports to extract the report you need for your bilateral project, but it also feeds directly into the CRP (CGIAR Research Program). So that was one of the features that I found very, very useful and one clear difference with the other center not using MEL at center level."*

Observed practices and challenges are of particular interest with regard to the current One CGIAR reform process. In fact, most of the participants discussed expectations for future RDM strategies in light of this ongoing and yet open transformation process. Among those discussions, the participants focused on questions around future system use. To date, decisions regarding the future of the two predominant systems in the CGIAR, MEL and MARLO, have not been taken. It is also not clear yet whether one global system will be promoted or several smaller systems be accepted. Clearly, this is quite problematic for platform adoption and use at the moment, as also P9 stated in the MEL context: *"As long as we don't know if the platform is gonna exist or if it's gonna be migrated into the One CG (One CGIAR) future, I think that people will not engage too much with the platforms beyond reporting and planning."*

There is strong consensus among the participants to choose a single-platform strategy. In fact, some, like P12, even go further and envision a single future CGIAR platform that manages data and communication in a comprehensive manner, from short-term storage to long-term preservation, and from internal chat messages to e-mail correspondence. The following are some statements that highlight the general desire to simplify and unite processes, and to remove duplicate efforts under a central unified platform:

> I think we can expect, you can hope from One CGIAR that they will solve the problem, which has not been solved so far of having several systems for data management and knowledge sharing. You have MEL, MARLO, some CRP also are using different systems, then you have other system managed by the SMO. So this is one thing we can expect it, but there will be One CGIAR approach that would avoid duplication of effort, the need for the same scientist to be involved in different systems, to invest in different systems. So for me, it's one of the prerequisite to expand the use of systems like MEL. – P8

> We are different centers but within the same kind of umbrella. Why are we using different platform? [...] So, you know, you spend a few hours doing the work for the person that is providing the fund and then you have your own PMU colleagues that remind you, you have to do the same on your own platform, which is just, you know, as a scientist, I mean, people really— We're not doing any more science, unfortunately. – P12

7 DISCUSSION

We conducted three studies to systematically map the RDM components practice, training, policies, infrastructure, and motivation in global agricultural science at ICARDA/CGIAR. Our investigation was closely aligned with the RDM commitment evolution model that, to date, was solely grounded in a single qualitative cross-domain exploration [36]. We contribute to the triangulation of this model through the sum of findings from Studies I–III. This triangulation is based on the identification of five RDM commitment drivers in the RDM commitment evolution model, closely described in Section 2, and the design of Studies I–III that focus on specific drivers. This mapping is described in Section 3 and Figure 2. Based on our findings, we confirm that regulations play an important role in RDM adoption and that this adoption is supported by strong training and data management efforts at ICARDA/CGIAR. We note that these efforts play a big role at sustaining commitment and that we did not document commitment decrease as predicted by the model. Further, we note that researchers described negotiation processes around policy conflicts and multi-platform collaborative sharing requirements that introduce an additional perspective to the RDM commitment evolution model. In summary, through our analysis of RDM commitment drivers, we confirm the interplay between those components as predicted by Feger et al. [36] at large. Yet, we also observe differences that motivate further researcher, as discussed in this section.

This section is structured as follows. First, we discuss the findings from our three studies through the lenses of our three key research questions. As part of this discussion, in Section 7.3, we discuss in detail how our findings relate to the RDM commitment evolution model. We conclude by relating research on RDM commitment to the wider data practice literature.

7.1 RQ 1: How are policies and regulations aligned with the organization's RDM goals?

As part of the survey study (Study I), we found that institutional and funding policies play roles as extrinsic forms of motivations for conducting RDM. We explored the current regulatory frameworks at ICARDA/CGIAR as part of our qualitative Study II. In this study with data managers we found that a variety of different rules and regulations are in place that mandate RDM. We described a regulation hierarchy involving funding rules, institutional policies, and national laws. While there is little doubt that the regulations are well-designed with the intention to foster FAIR and open data, this hierarchical framework leads to policy conflicts and duplicate efforts on a regular basis. These issues become most apparent in inter-center and beyond-CGIAR collaborations, as scientists in Study III explained. Here, conflicting regulations might require accessibility through multiple platforms and differing data standards, in turn possibly harming acceptance for RDM policies across the scientific community.

Study II participants stressed that the policies do not only act as enforcement, but represent a common ground for discussion, a protocol, between scientists and data managers. Detailed and specific policies further have a training effect, as they outline requirements for compliance. Still, we found that the organizations need to invest time and effort in the introduction and explanation of regulations. This is noteworthy, as the informants described shortcomings in the administration related to support and verification of compliance with RDM policies.

To answer our original RQ, we find that policies and regulations are aligned with RDM goals, however, effort is needed to harmonize differing policies and to provide clear rules for policy conflict

resolution in collaborative settings. The ongoing reform and transformation towards One CGIAR provides an opportunity to address those issues.

7.2 RQ 2: What are practices and needs around training, infrastructure, and motivation?

Results from our survey Study I showed that researchers had a slightly negative attitude towards the suitability of current RDM practices. While respondents indicated that RDM is important in their domain, they found that shared resources "are usually not sufficiently well documented" and indicated somewhat agreement to the fact that current practices "are far away from systematic RDM" in their domain. The suitability of formal training, educational resources, and training support were rated similarly low to neutral. The respondents indicated that especially formal education on RDM was lacking in their studies. Unfortunately, organizational training and support schemes do not seem to be suitable to compensate for this lack of education. All statements related to training resources received neutral responses. In contrast, all statements related to the suitability of current technical infrastructure received stronger and consistent agreement. We find this positive attitude towards deployed infrastructure reflected in Study III that we conducted with ICARDA/CGIAR scientists. Here, the informants stressed the value of platforms like MEL in their data management workflows. While they offered a variety of possible improvements, with particular regard to customization and usability, the study participants recognized the efforts taken in the MEL development. They further stressed the value of MEL and similar services in reporting and knowledge extraction.

7.2.1 RQ 2a: How do current drivers of motivation impact RDM?. Motivational drivers are key mechanisms of the RDM commitment evolution model involved in the early adoption and the reward cycle. In the adoption phase, they have to be considered in concert with implemented regulations and the researchers' intrinsic motivation to improve scientific processes. In our research, we found that rewarding policies are being implemented that play an increasingly important role. Participants of both Studies II and III referred to financial rewards and bonuses for effective RDM. One participant also referred to a RDM competition that was rewarded with a small research grant. Clearly, these extrinsic motivators do play a role in the early RDM adoption. In particular, the grant competition relates to researchers' goal of increasing visibility and career prospects. A corresponding statement in Study I received the strongest agreement within the extrinsic motivation group ("I am engaged in RDM because I believe that it increases my visibility and career prospects": mean_all=5.2; mean_cgiar=5.7).

Generally, we mapped strong identified motivation for conducting RDM in Study I. Placed in context of Study III, we recognize this as a result of researchers' strong interest to support rural farmers and to make a contribution for sustainable global agriculture. Several participants discussed that they interacted with rural farmers in the past and that they perceived their work as a contribution to global sustainable agriculture. Based on those experiences, the impact of knowledge does not seem to be an abstract goal for the ICARDA/CGIAR researchers, but something rather tangible instead that helps to explain the strong identified motivation for RDM.

7.2.2 RQ 2b: What incentives can future RDM platforms provide? The researchers participating in Study III discussed a set of platform use cases that they could strongly profit from in their work. The most prominent example was that of automated reporting mechanisms. The study participants stressed that administrative procedures take a lot of time and wished for a centralized platform like MEL to aggregate data and to lower the reporting burden through automated mechanisms that would only require manual checks rather than manual creation. In this context, several participants explicitly referred to exploiting new possibilities enabled by advances in machine learning.

An additional strength of platforms that the informants discussed related to general knowledge extraction that helps to identify research gaps, opportunities for collaboration, and provides

arguments for funding requests. Clearly, the sum of all those use cases would primarily benefit researchers who actively contribute to systems like MEL. Based on those contributions, such services can create accurate profiles that are a basis for offering effective solutions around those use cases. This notion relates closely to what Feger et al. [32] referred to as "secondary usage forms" of technology in particle physics RDM that profit mainly scientists who actively contribute detailed resources on dedicated and domain-tailored RDM services.

7.2.3 RQ 2c: How does formal training and the current support impact RDM practices? We know that ICARDA/CGIAR are committed to supporting researchers with their data management and that the organizations go to great lengths to provide assistance. Still, there are indications in Study I that show that scientists do not perceive the current support to be sufficient. The study participants rated both the statements "My organization employs research data managers who support me in conducting RDM" and "My organization does not have the personnel needed to customize RDM tools for us" in a neutral manner. Notably, the latter statement is the only neutral response in the part of the survey related to infrastructure. Our findings related to *Practice* in Study III show that researchers asked for support related to very basic data selection and sharing criteria. They feared that inadequate data would badly reflect on their reputation and were further worried about violating privacy regulations. However, we argue that such low-level selection tasks cannot be covered entirely through existing support schemes. Instead, we see a need to educate researchers about the preparation of data from scratch, to instill trust in their ability to design good quality data. This call for high quality data is also echoed in the work of Trisovic et al. [68]. Further, automated tools should be developed that support researchers in making their data compliant with privacy regulations. One example of such a tool is PII Engine, a tool developed in collaboration with ICARDA[15].

Organizational support becomes especially important in light of lack of formal RDM training. Asked to rate the statement "Education on RDM was part of the curriculum in my studies", Study I participants showed disagreement: mean_all=2.7; mean_cgiar=2.8. However, here we have to note that we did not record the age of the survey participants, which means that we cannot reason about potential differences between early-career scientists and more established researchers. Several of the Study II informants stressed that RDM practice has lately gained importance in academic curricula.

7.3 RQ 3: How does the stage-based RDM commitment evolution model apply to global agricultural science?

Based on the sum of findings from our three studies that we discussed through the lenses of RQ:1 and RQ:2a-c, we present in Figure 7 an overview of RDM commitment evolution at ICARDA/CGIAR. To this end, we mapped our findings to the RDM commitment evolution model proposed by Feger et al. [36]. As depicted in Figure 7, Transition I, researchers *adopt* RDM practices in response to intrinsic and extrinsic regulations. Recognized and accepted champions advocating for the value of data management, as well as researchers' interest to contribute to a sustainable global agriculture, both represent intrinsic forms of motivation that we mapped in our work. In contrast, strong policies and financial incentives were some of the most pronounced extrinsic motivators. While these findings related to the commitment evolution in Transition I correspond to the model, we note that we did not observe evidence for any commitment decrease within this transition cycle. Our findings suggest two explanations. First, ICARDA/CGIAR invest extensively in data curation support and infrastructure development. Thus, the organizations provide strong foundations to *overcome barriers* and to *integrate* RDM practice into researchers' workflows. Second, we note that

[15]https://github.com/SCiO-systems/piiengine

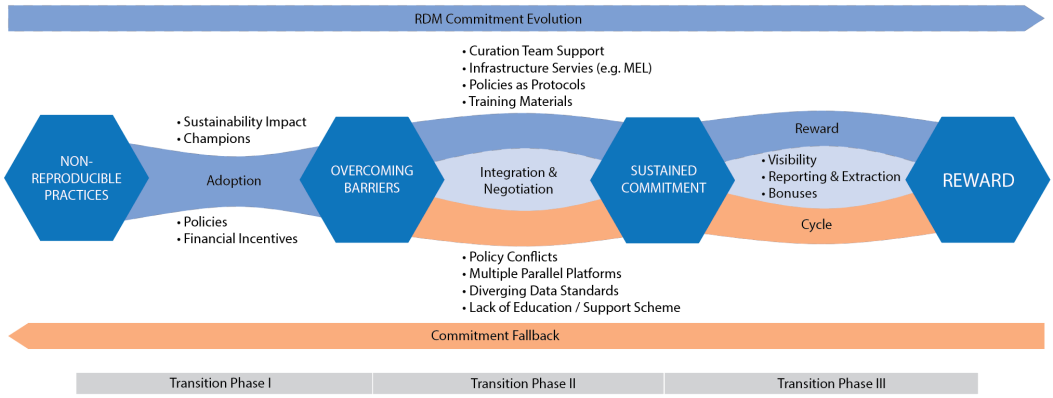

Fig. 7. We mapped our findings to the stage-based RDM commitment evolution model described by Feger et al. [36] and depicted in Figure 1. In this process, we also made adjustments. In particular, we do not see evidence for commitment fallback within Transition I, while we mapped a circular integration and negotiation process in Transition Phase II.

while policies are frequently conflicting, see Study II and RQ:1, they nevertheless continue to be enforced and checked by the data managers. Based on these findings, we argue that RDM practice *adoption* that is at least partially grounded in policies and enforcement, is likely to never fully revert back to the *non-reproducible practices* stage. We suggest future work to consider this implication and to contribute additional experiences to support or reject this model implication.

In the RDM commitment evolution model, Transition II is based on an *integration* of RDM practices into common workflows. Based on our findings, we confirm that ICARDA/CGIAR researchers have a set of effective resources at their disposal to *overcome barriers* and to *integrate* them into their scientific processes. Foremost, we see that the MEL platform is a suitable infrastructure component for most researchers. While our informants described a number of usability barriers, they also stressed the platform's value in storing, managing, and sharing data and reports. Several participants also described that, based on their experience, CGIAR centers who worked with MEL, were better equipped to handle RDM responsibilities in collaborative settings. Our study participants referred to additional resources that helped in the integration: policies that acted both as common protocol and training material; and supportive data managers. Yet, we also find differences between practices at ICARDA/CGIAR and the reference model. While the model does not foresee a RDM commitment decrease between the *sustained commitment* and *overcoming barriers* stages, our findings reveal that practical issues around the RDM implementation lead to *negotiation* processes in which routine RDM tasks and commitment is entangled with attempts to *overcome barriers*. This *negotiation* happens between researchers and data managers, between data managers and management, between collaborating organizations and donors, or simply as an internal process of compromising between different strategies in an effort to complete requirements as thoroughly as *feasible*. Common issues described by our participants that trigger such negotiation processes include policy conflicts, the need to use multiple organizational platforms, and unclear or diverging data standards. We note that this is a process which happens in an attempt to resolve issues and to integrate RDM practices as best as possible, thereby making compromises within this iterative Transition II phase. Again, we find no evidence that any of those issues led to researchers abandoning their RDM responsibilities, as predicted by the reference model. This holds true independent of the rewards provided to researchers within the reward cycle, Transition III. Here, researchers referred

to financial rewards and their expected impact on global sustainable farming as two types of very different rewards. Further, they referred to features that were closely related to desired platform mechanisms, including automatic reporting, knowledge extraction, and increased visibility. In contrast to the reference model, we did not map a substantial commitment decrease resulting from a lack of rewards. Yet, we note that researchers who perceive RDM practices as personally rewarding, are likely to be more inclined to engage in the *negotiation* and *integration* cycle in Transition II, to find effective solutions beyond the bare minimum requirements, in order to deal with various issues and to adapt to novelty and creativity in science [32]. We suggest future work to test if the absence of rewards in the long term does or does not lead to a substantial RDM commitment decrease, and to map such experiences to the effectiveness of policies in place.

In summary, we find that the stage-based RDM commitment evolution model [36] applies well to global agricultural research practices at ICARDA/CGIAR. We note that reviewing concrete practices around different RDM components, and placing them in the context of commitment stages and transitions helps to conceptualise the broader socio-technical framework of RDM practice and to identify weaknesses and open questions. In our setting, we hope that described issues requiring *negotiation* will be addressed as part of the One CGIAR reform. Finally, we highlighted two inconsistencies between our findings and the reference model, related to commitment decrease, and described future research challenges to address and resolve those deviations through additional experiences and records.

7.4 RDM Commitment within the Wider Data Practice Literature

Muller et al. [51] studied digital data practices at IBM. They described five human interventions in data science work practices: discovery, capture, curation, design (e.g. imputing data this is missing), and creation (e.g. inspection of validation data). The authors referred to *Tools & methods* as sensitizing concepts applicable mostly during data design and creation. In contrast, our work, and its application of the RDM commitment evolution model [36], show that tools and technical infrastructure need to be considered earlier, already during interventions related to data curation. Here, platforms can act as common interface between scientists and data managers tasked to support curation activities. Further, Muller et al. describe tasks around *data selection* as part of *capture* interventions. Based on our findings, we note that scientists ask for support from data managers during this stage. Those notions of support and infrastructure requirements are not echoed in the work from Muller et al. We imagine that this is at least partially resonating that data science experts interviewed at IBM are likely to have a stronger background and training in data practices than agricultural science researchers in our study. This suggests that future work exploring data science interventions should carefully map findings to participants' backgrounds.

This call to consider the wider ecosystem of data production and data practice is also echoed in the work from Vertesi and Dourish [69]. The authors reported on their ethnographic research within two robotic space exploration teams and found that the digital data production history directly informs sharing practices. In the context of data curation and data management at ICARDA/CGIAR, we confirm that the collaborative character of agricultural research informs curation and sharing practices. Unlike Vertesi and Dourish, we did not find a direct link between differences in data production and *willingness* to share data. Rather, we found that collaborative frameworks across CGIAR centers and/or third parties led to barriers around data standards, curation practices, and infrastructure requirements. These findings allow to apply a different lense to the data production history and subsequent sharing ability and willingness that perceives (inter-)organizational data production as a potential *barrier* to sharing, rather than a foundation for sharing willingness. This discussion ultimately relates to the question of what constitutes reproducibility. Chen et al. [19] stressed that simply being open is not enough. In this context, Feger and Woźniak [35] proposed a

researcher-centered definition of reproducibility that focuses on data completeness and the trade-off between the complexity of a reproduction attempt and its potential impact.

One key aspect in data management and curation relates to the detailed description of data. In this context, Faniel et al. [28] found that archaeologists require a detailed description of the data collection history when they consider reusing data for their own projects. Here, research in archaeology shows parallels to global agricultural science where researchers need to know about a wide range of background information, including exact geolocation, farming background, and plant and animal characteristics. To support this metadata need, the MEL platform provides a wide range of input fields. However, participants in Study III described them as usability barriers that negatively impact RDM commitment. Faniel et al. describe the development of formal ontologies and the employment of dedicated data editors as solutions to ensure metadata quality and to enable discovery and reuse of data. Our findings related to Study II show that adding more tasks to data managers would put a burden on a system that is already characterized by limited data management resources. This problem could be addressed through political reforms. Another approach relates to the exploration of data collection techniques that automate or ease metadata recording in the field. The concept of Ubiquitous Research Preservation [31] might provide new perspectives for the development and use of data collection devices. Such an exploration might particularly profit domains like agricultural and archaeological science due to the diverse nature of field studies and environments in which scientists operate. Yet, it also poses challenges related to the privacy of data collected and shared. Already today with a mostly manual data description process, researchers expressed that they were uncertain about privacy implications of the data they were supposed to share.

Karasti et al. [46] report on their ethnographic study with scientists and information managers in the Long Term Ecological Research (LTER) network. The authors contrasted short-term and long-term projects and emphasized the need for a consistent and adequate digital archiving of data in long-term studies. Further, they referred to data stewardship across local-global network settings. The LTER network consists of several distributed US sites. Yet, local-global issues in the LTER around data curation and stewardship have been addressed differently than in agricultural science at ICARDA/CGIAR. While data managers and scientists in our study echoed issues around duplicate platform requirements and conflicting policies and standards across CGIAR centers, Karasti et al. note that in "the LTER case, information managers are able to address the divides and boundaries of the local global tension already in their collaborative activities of developing shared technologies and standards." The authors stressed that the LTER network, "[...] sheltered by the exceptionally long and continuous periods of research funding, has been in a privileged position to explore a more science-driven approach to data stewardship [...]". Contrasting these findings, we perceive implications for the One CGIAR reform, as well as general science collaboration, to explore the effects of stable and long-term funding frameworks as a foundation for effect data management practices that involve collaborative data stewardship and not just collaborative data collection and analysis.

7.5 Limitations and Future Work

Our open call to disseminate and share the survey in Study I across the scientific community has possibly resulted in responses coming from outside ICARDA/CGIAR. In turn, we separately reported results from all responses and self-reported ICARDA/CGIAR affiliation. While this represents a limitation to the Study I reporting, we argue that a wider dissemination is also a strength regarding our goal to contribute experiences around RDM commitment evolution. Please note that a detailed reflection on the overall research methodology and potential biases is available in Section 3.4.

We need to stress that most of our study participants were affiliated with ICARDA. We consider this focus on a single CGIAR center as an necessity that allowed us to map perceptions and practices within this domain more closely. Future work will profit from expanding our research strategy across all CGIAR centers and even beyond the CGIAR.

We find that developing a validated RDM scale could profit strategic assessment and decision making in science. We consider our systematic development of a RDM questionnaire, that is closely aligned with the stages and transitions of the RDM commitment evolution model, as a valuable starting point. The research we conducted qualitatively in Studies II and III showed how findings can be placed and discussed in context of the survey results. As described in the following section, we make resources from the survey study freely available in order to foster large-scale validation research.

7.6 Open Data

We openly share a wide set of our resources as supplementary materials, in order to increase the transparency of our research and to enable future work. In the context of Study I, we expect that future research will be able to further validation and creation of a generally valid RDM scale based on our shared resources. We release the exported Qualtrics questionnaire, as well as the entire dataset that we collected. We further share resources related to our qualitative studies. The supplementary materials contain both Atlas.ti code group reports and the Study III interview protocol.

8 CONCLUSION

We reported findings from our empirical research on RDM practices in global agricultural science. Following a systematic research approach aligned with a recent stage-based model of RDM commitment evolution, we conducted three studies focusing on the following RDM components: practices, training, policies, infrastructure, and motivations. For Study I, we created and disseminated (n = 23) a survey designed to map those RDM components. Our results showed that infrastructure developments were perceived as more suitable than current practices and training capabilities. Our results further showed how different types of motivation inform attitudes towards RDM. In Studies II (n = 17) and III (n = 13), we continued to explore individual RDM components through qualitative explorations with data managers and scientists working in the domain of agricultural science at ICARDA/CGIAR. Based on the sum of findings from all three studies, we contribute to the triangulation of the RDM commitment evolution model. We note that strong support and suitable technical infrastructure help develop commitment, while policy conflicts, unclear data standards, and multi-platform sharing, lead to unexpected negotiation processes. Further, we pose questions regarding the lack of observed RDM commitment decrease, as our informants did not report such an effect as a result of current barriers. We discussed this observation in the context of the dominant identified regulation, which stems from researchers' experience and interaction with rural farmers and their goal to contribute to sustainable farming practices. We expect that the sum of findings will help to better understand RDM commitment drivers, to refine the commitment evolution model, and to benefit its application in science.

REFERENCES

[1] ACM. 2018. Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging Retrieved September 10, 2018.

[2] Katherine G Akers and Jennifer Doty. 2013. Disciplinary differences in faculty research data management practices and perspectives. (2013).

[3] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (2016), 452–454. https://doi.org/10.1038/533452a

[4] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29, 2 (2013), 599–611. https://doi.org/10.1016/j.future.2011.08.004

[5] C. Glenn Begley and Lee M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483, 7391 (2012), 531–3. https://doi.org/10.1038/483531a arXiv:9907372v1 [arXiv:cond-mat]

[6] Khalid Belhajjame, Jun Zhao, Daniel Garijo, Kristina Hettne, Raul Palma, Óscar Corcho, José-Manuel Gómez-Pérez, Sean Bechhofer, Graham Klyne, and Carole Goble. 2014. The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web. *arXiv preprint arXiv: 1401.4307* February 2014 (2014), 20. arXiv:1401.4307 http://arxiv.org/abs/1401.4307

[7] Gordon Bell, Tony Hey, and Alex Szalay. 2009. Beyond the Data Deluge. *Science* 323, 5919 (2009), 1297–1298. https://doi.org/10.1126/science.1170411 arXiv:http://science.sciencemag.org/content/323/5919/1297.full.pdf

[8] Carolyn Bishoff and Lisa Johnston. 2015. Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University. *Journal of Librarianship & Scholarly Communication* 3, 2 (2015).

[9] Bradley Wade Bishop and Rose M Borden. 2020. Scientists' Research Data Management Questions: Lessons Learned at a Data Help Desk. *portal: Libraries and the Academy* 20, 4 (2020), 677–692.

[10] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI Research: Going Behind the Scenes*. Morgan & Claypool Publishers, 51–60. https://doi.org/10.2200/S00706ED1V01Y201602HCI034

[11] Ronald F Boisvert. 2016. Incentivizing reproducibility. *Commun. ACM* 59, 10 (2016), 5–5. https://doi.org/10.1145/2994031

[12] Christine L Borgman. 2007. *Scholarship in the digital age: information, infrastructure, and the internet*. MIT Press, Cambridge, MA.

[13] Christine L Borgman, Paul N Edwards, Steven J Jackson, Melissa K Chalmers, Geoffrey C Bowker, David Ribes, Matt Burton, and Scout Calvert. 2013. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. (2013).

[14] Geoffrey Boulton, Michael Rawlins, Patrick Vallance, and Mark Walport. 2011. Science as a public enterprise: the case for open data. *The Lancet* 377, 9778 (2011), 1633–1635.

[15] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597.

[16] Virginia Braun and Victoria Clarke. 2020. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* (2020), 1–25.

[17] Kristin Briney, Abigail Goben, and Lisa Zilinski. 2015. Do you have an institutional data policy? A review of the current landscape of library data services and institutional data policies. (2015).

[18] Cunera M Buys and Pamela L Shaw. 2015. Data Management Practices Across an Institution: Survey and Report. *Journal of Librarianship & Scholarly Communication* 3, 2 (2015).

[19] Xiaoli Chen, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, et al. 2019. Open is not enough. *Nature Physics* 15, 2 (2019), 113–119.

[20] Open Science Collaboration. 2012. An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science* 7, 6 (2012), 657–660. https://doi.org/10.1177/1745691612462588

[21] COS. 2019. Open Science Badges. Website. (2019). https://cos.io/our-services/open-science-badges Retrieved February 5, 2020.

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

[23] A De Waard, H Cousijn, and IjJ Aalbersberg. 2015. 10 aspects of highly effective research data: Good research data management makes data reusable. (11 December 2015). https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data

[24] Edward L Deci and Richard M Ryan. 1985. Toward an organismic integration theory. In *Intrinsic motivation and self-determination in human behavior*. Springer, 113–148.

[25] Sebastian Deterding, Rilla Khaled, Lennart E Nacke, and Dan Dixon. 2011. Gamification: Toward a definition. In *CHI 2011 gamification workshop proceedings*, Vol. 12. Vancouver BC, Canada.

[26] Vasant Dhar. 2012. Data science and prediction. (2012).

[27] Tumsifu Elly and Ephraem Epafra Silayo. 2013. Agricultural information needs and sources of the rural farmers in Tanzania. *Library review* (2013).

[28] Ixchel Faniel, Eric Kansa, Sarah Whitcher Kansa, Julianna Barrera-Gomez, and Elizabeth Yakel. 2013. The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (Indianapolis, Indiana, USA) *(JCDL '13)*. Association for Computing Machinery, New York, NY, USA, 295–304. https://doi.org/10.1145/2467696.2467712

[29] Benedikt Fecher, Sascha Friesike, Marcel Hebing, and Stephanie Linek. 2017. A reputation economy: how individual reward considerations trump systemic arguments for open access to data. *Palgrave Communications* 3 (2017), 17051. https://doi.org/10.1057/palcomms.2017.51

[30] Sebastian Feger. 2020. *Interactive tools for reproducible science.* Ph.D. Dissertation. lmu.

[31] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Pascal Knierim, Passant El.Agroudy, Paweł W. Woźniak, and Albrecht Schmidt. 2020. Ubiquitous Research Preservation: Transforming Knowledge Preservation in Computational Science. *MetaArXiv* (3 March 2020). https://doi.org/10.31222/osf.io/qmkc9

[32] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W. Woźniak. 2019. Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'19*. https://doi.org/10.1145/3290605.3300685

[33] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, and Albrecht Schmidt. 2019. Gamification in Science: A Study of Requirements in the Context of Reproducible Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'19*. https://doi.org/10.1145/3290605.3300690

[34] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, and Albrecht Schmidt. 2019. The Role of HCI in Reproducible Science: Understanding, Supporting and Motivating Core Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312905

[35] Sebastian Stefan Feger and Paweł W Woźniak. 2022. Reproducibility: A Researcher-Centered Definition. *Multimodal Technologies and Interaction* 6, 2 (2022), 17.

[36] Sebastian S Feger, Paweł W Wozniak, Lars Lischke, and Albrecht Schmidt. 2020. 'Yes, I comply!' Motivations and Practices around Research Data Management and Reuse across Scientific Fields. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[37] Sebastian Stefan Feger, Paweł W Woźniak, Jasmin Niess, and Albrecht Schmidt. 2021. Tailored Science Badges: Enabling New Forms of Research Interaction. In *Designing Interactive Systems Conference 2021*. 576–588.

[38] Dimitris Folinas, Ioannis Manikas, and Basil Manos. 2006. Traceability data management for food chains. *British Food Journal* (2006).

[39] FORCE11. 2014. The FAIR data principles. Website. Retrieved August 8, 2017 from https://www.force11.org/group/fairgroup/fairprinciples.

[40] Lisa Harper, Jacqueline Campbell, Ethalinda KS Cannon, Sook Jung, Monica Poelchau, Ramona Walls, Carson Andorf, Elizabeth Arnaud, Tanya Z Berardini, Clayton Birkett, et al. 2018. AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database* 2018 (2018).

[41] Rosie Higman and Stephen Pinfield. 2015. Research data management and openness: the role of data sharing in developing institutional policies and practices. *Program* 49, 4 (2015), 364–381.

[42] HS Hollander, Francesca Morselli, Femmy Admiraal, Frank Uiterwaal, and Marina Noordegraaf. 2018. Guidelines to FAIRify data management and make data reusable: PARTHENOS. (2018).

[43] Matthew B Hoy. 2014. Big data: An introduction for librarians. *Medical reference services quarterly* 33, 3 (2014), 320–326. https://doi.org/10.1080/02763869.2014.925709

[44] Lori M Jahnke and Andrew Asher. 2012. The problem of data: Data management and curation practices among university researchers. *L. Jahnke, A. Asher & SDC Keralis, The problem of data* (2012), 3–31.

[45] Marina Jirotka, Rob Procter, Tom Rodden, and Geoffrey C. Bowker. 2006. Special Issue: Collaboration in e-Research. *Computer Supported Cooperative Work (CSCW)* 15, 4 (01 Aug 2006), 251–255. https://doi.org/10.1007/s10606-006-9028-x

[46] Helena Karasti, Karen S. Baker, and Eija Halkola. 2006. Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. *Comput. Supported Coop. Work* 15, 4 (Aug. 2006), 321–358. https://doi.org/10.1007/s10606-006-9023-2

[47] Karina Kervin and Margaret Hedstrom. 2012. How research funding affects data sharing. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*. ACM, 131–134. https://doi.org/10.1145/2141512.2141560

[48] Mallory C Kidwell, Ljiljana B Lazarević, Erica Baranski, Tom E Hardwicke, Sarah Piechowski, Lina-Sophia Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, Chelsey Hess-Holden, et al. 2016. Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS biology* 14, 5 (2016), e1002456.

[49] Mallory C. Kidwell, Ljiljana B. Lazarević, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Lina Sophia Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, Chelsey Hess-Holden, Timothy M. Errington, Susann Fiedler, and Brian A. Nosek. 2016. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS Biology* (2016). https://doi.org/10.1371/journal.pbio.1002456

[50] Sabina Leonelli. 2013. Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology. *Bulletin of Science, Technology & Society* 33, 1-2 (2013), 6–11. https://doi.org/10.1177/0270467613496768

[51] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 126, 15 pages. https://doi.org/10.1145/3290605.3300356

[52] Daniel Nüst, Lukas Lohoff, Lasse Einfeldt, Nimrod Gavish, Marlena Götza, Shahzeib Tariq Jaswal, Salman Khalid, Laura Meierkort, Matthias Mohr, Clara Rendel, et al. 2019. Guerrilla Badges for Reproducible Geospatial Data Science. *AGILE 2019* (2019). https://doi.org/10.31223/osf.io/xtsqh

[53] Henny Osbahr, Peter Dorward, Roger Stern, and Sarah Cooper. 2011. Supporting agricultural innovation in Uganda to respond to climate risk: linking climate change and variability with farmer perceptions. *Experimental agriculture* 47, 2 (2011), 293–316.

[54] Irene V. Pasquetto, Ashley E. Sands, Peter T. Darch, and Christine L. Borgman. 2016. Open Data in Scientific Settings: From Policy to Practice. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 1585–1596. https://doi.org/10.1145/2858036.2858543

[55] Dirk Pilat and Yukiko Fukasaku. 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding. *Data Science Journal* 6 (2007), OD4–OD11. https://doi.org/10.1787/9789264034020-en-fr

[56] Jian Qin. 2016. Metadata and reproducibility: A case study of gravitational wave data management. *International Journal of Digital Curation* 11, 1 (2016), 218–231. https://doi.org/10.2218/ijdc.v11i1.399

[57] Kevin B Read, Catherine Larson, Colleen Gillespie, So Young Oh, and Alisa Surkis. 2019. A two-tiered curriculum to improve data management practices for researchers. *PloS one* 14, 5 (2019), e0215509.

[58] Michael Rosenblatt. 2016. An incentive-based approach for improving data reproducibility. *Science Translational Medicine* 8, 336 (2016), 336ed5–336ed5. https://doi.org/10.1126/scitranslmed.aaf5003 arXiv:http://stm.sciencemag.org/content/8/336/336ed5.full.pdf

[59] Anisa Rowhani-Farid, Michelle Allen, and Adrian G Barnett. 2017. What incentives increase data sharing in health and medical research? A systematic review. *Research integrity and peer review* 2, 1 (2017), 4.

[60] Anisa Rowhani-Farid, Michelle Allen, and Adrian G. Barnett. 2017. What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review* 2, 1 (2017), 4. https://doi.org/10.1186/s41073-017-0028-9

[61] Jonathan F Russell. 2013. If a job is worth doing, it is worth doing twice: researchers and funding agencies need to put a premium on ensuring that results are reproducible. *Nature* 496, 7443 (2013), 7–8. https://doi.org/10.1038/496007a

[62] Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.

[63] Bashir Salim, Shin-nosuke Takeshima, Ryo Nakao, Mohamed AM Moustafa, Mohamed-Khair A Ahmed, Sumaya Kambal, Joram M Mwacharo, Abeer M Alkhaibari, and Guillermo Giovambattista. 2021. BoLA-DRB3 gene haplotypes show divergence in native Sudanese cattle from taurine and indicine breeds. *Scientific reports* 11, 1 (2021), 1–15.

[64] Alan Schwartz, Cleo Pappas, and Leslie J Sandlow. 2010. Data repositories for medical education research: issues and recommendations. *Academic Medicine* 85, 5 (2010), 837–843. https://doi.org/10.1097/ACM.0b013e3181d74562

[65] Victoria Stodden and Sheila Miguez. 2014. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software* 2, 1 (2014), 21. https://doi.org/10.5334/jors.ay

[66] Rong Tang and Zhan Hu. 2019. Providing Research Data Management (RDM) services in libraries: Preparedness, roles, challenges, and training for RDM practice. *Data and Information Management* 1, ahead-of-print (2019).

[67] Joanna Thielen and Amanda Nichols Hess. 2017. Advancing research data management in the social sciences: implementing instruction for education graduate students into a doctoral curriculum. *Behavioral & Social Sciences Librarian* 36, 1 (2017), 16–30.

[68] Ana Trisovic, Katherine Mika, Ceilyn Boyd, Sebastian Feger, and Mercè Crosas. 2021. Repository approaches to improving the quality of shared data and code. *Data* 6, 2 (2021), 15.

[69] Janet Vertesi and Paul Dourish. 2011. The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 533–542. https://doi.org/10.1145/1958824.1958906

[70] Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE* 8, 7 (2013). https://doi.org/10.1371/journal.pone.0067332

[71] Angus Whyte and Jonathan Tedds. 2011. *Making the case for research data management.* Digital Curation Centre.

[72] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft,

Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris a. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[73] James AJ Wilson, Luis Martinez-Uribe, Michael A Fraser, and Paul Jeffreys. 2011. An institutional approach to developing research data management infrastructure. (2011).

[74] Peter Wittenburg, Herbert Van de Sompel, Jens Vigen, Achim Bachem, Laurent Romary, Monica Marinucci, Thomas Andersson, Françoise Genova, Christoph Best, Wouter Los, et al. 2010. Riding the wave: How Europe can gain from the rising tide of scientific data. (2010). Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission.

Received April 2021; revised November 2021; accepted March 2022