Research Data Management for Open Science

Armel Lefebvre

SIKS Dissertation Series No. 2021-07

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

This work was supported by the Research IT innovation programme of Utrecht University

DOI: https://doi.org/10.33540/447

Research data management for open science

Research Data Management for Open Science

HET BEHEER VAN ONDERZOEKSDATA VOOR OPEN WETENSCHAP

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 15 maart 2021 des middags te 4.15 uur

door

Armel Edmond Jacques Lucien Lefebvre

Promotoren:

Prof. dr. S. Brinkkemper

Prof. dr. B. Snel

Prof. dr. M.R. Spruit

Copromotor:

Dr. ir. B. Van Breukelen

Aan Opa en Oma

À Grand-papa

Acknowledgments

In the last years I got used to be self-absorbed, trying to revise quite a number of my rejected papers to such an extent that it would give me the illusion that my papers and I were the only thing that mattered. Nevertheless, it is clear at the end that there are actually so many better things that happened and so many people that helped me reach the end of this project alive.

Marco, I would like to thank you for supporting my endless IT for research ambitions since my arrival at Utrecht nearly a decade ago! And even more so, to give me an opportunity to explore that domain to its full extend, with some extra research years at Utrecht University! I fuzzily remember one of the techno-conceptual discussions we had during my master thesis, when you asked: “Nou, je wilt gaan promoveren toch?”, and I just answered “Euh, Ja !?”, barely knowing what that would mean. Several years later, I do know what it means slightly better, to “promoveren”, and I must say that it is a cognitive roller-coaster that is worth the entrance ticket. If the two main outcomes of this PhD time can be summarized into some concepts as exotic as “laboratory forensics” and “open science readiness”, it is truly because of your faith in creativity and unbounded exploration that allowed me to pursue new knowledge and go beyond my own limits.

Sjaak, thank you for reviewing and shaping the information systems science behind this dissertation. With your help, I could finally see that much more had been accomplished in these years than I thought, and that there was some strong and original research underlying a unique perspective on research data management!

Bas for literally opening me the doors of a mass-spec lab and confronting me to a bunch of yet to be FAIRified research data. We would need some kind of research infrastructure Horizon Europe lifetime grant to complete all the systems we were envisioning to fix reproducibility in that lab, but it gave such a unique perspective to this research, some “real” life experience of laboratory work. Besides that, you let me join the UBC teaching program with Adrien and Joep where I could see how valuable research data management can be for the future generation of biologists and bioinformaticians who, after many efforts on my teaching skills using (lovely) course evaluations as input, started to appreciate some nice stuff about open science and what it can do for research.

Berend for patiently helping me, an alien dropped in the bioinformatics world for a little while, navigate around my disillusions and delusions, guiding me in a way of thinking and understanding bioinformatics at work as well as the subtleties of the academic system as a whole.

There are so many of these lessons learned that did not make it into papers but that I will carry far beyond this dissertation! Thank you for this.

Jonathan, Rik, Ton, Menno, and Folkert-Jan of ITS for the overambitious digital forensic toolkit, discussions, help, feedback, and support when being a researcher was also being a traveler. I also want to thank Annemiek and Barbara of the library for organizing the UU RDM community events and also work on the workshop “Writing reproducible code”.

I was lucky to have two places where I could go with my observations and questions, one place was close to my computer science colleagues: Siamak, Unal, Anna-Lena, Georg, Matthieu, Frans, Jan-Martijn, Sergio, Fabiano, Sietse, Slinger, Veronica, Başak and the applied data science lab with Lamia, Noha, Ingy, Chaim, Max, and Pablo. Also, thanks to the (then) MBI students Baharak, Elizabeth, Jorien and Jorrit for having conducted great research in RDM during your master studies.

Then, great people in Albert’s Hecklab. There I could spend hours analyzing what is hidden behind scientific publications in Nature with the support of Corine. Scrutinizing unFAIRified data would not have been as enjoyable (really) without all the knowledge of Henk and Jean-Francois and their bio-chem-informatics crash-courses doubled by their huge experience in all these tools and methods that were just keywords in papers to me. It probably did not ease my progress into finding a simple, generic, reproducibility-proof way of managing research data, but I got nearly there at least, and thank you for that.

While I was not doing much besides trying to fix science for a while, there has been the needed support when science broke me. Of course: Bernard, you my friend, never hesitated to bring me buckets of Chouffe (du terroir) – before the maatregelen, I insist - by car, bus or even banana-powered bike. And that since I moved to Utrecht a while ago, and who knows where next (pump up the tires!). Then surely also Guru and Ian, what an adventure the Bunnik Mansion was, completely irreplaceable for so many reasons, we are going to drink beer on that in India and China for the years to come! And Bilge, from office confident to the friend section, that says it all. Hein, bedankt voor alle hulp en moed die je aan Hong en mij hebt gegeven. Also, I cannot end without dedicating this work to Arthur, Pauline, Catherine and Giovanni, I know we are kind of used to not having words to speak that often, but I was so happy to have you besides me at keystones of my life, I’ll see you all soon!

Last but not least, my family, my parents, and my brother. Especially my parents, who probably thought I knew what I was doing and gave all their energy to cross the finish line. We all got greyer hair after all this, but you are still much stronger and wiser than I am. So, now that I am

done with school after a last exam (Coucou, Bonne-Maman!) I'll keep learning from you. Et puis Frérot, toi aussi t'en as fait du chemin, un peu différent de celui narré dans cet ouvrage, mais tout de même!

And last but actually first, Honghong, for being such a courageous and smart person I can trust. We are going to make it Hong' !!

Armel, Zeist, February 19, 2021

Contents

Chapter 1 | Introduction

Chapter 1

In this dissertation, we explore research data management using an information systems research perspective. Research data management (RDM) is a recent phenomenon in academia, at least if we consider RDM as it has been evolving until today, as a service from institutions to researchers. There are many reasons why academic institutions see benefits in professionalizing data management. Research data has steadily gained a more prominent place in recent discoveries involving complicated technical infrastructures as much as redefining the scientific method. Scientific knowledge emerges from a variety of ways, including the meticulous gathering of data and large-scale analyses of thousands of objects and measurements. In the recent history of science, this reliance on infrastructure, data, and computations has come with significant discoveries and, also, a more troublesome side of research.

On the one hand, the last two decades have been rich in significant scientific discoveries where the role of large scientific infrastructures has been essential. In 2003, the human genome project completed the decoding of human DNA after a tough competition where investments in efficient sequencing and computing infrastructure were determinant for the success of the whole genome sequencing project (Collins et al., 2004; Craig Venter et al., 2001). In 2012, we heard about the discovery of the Higgs boson (Aad et al., 2012). Not long after that, in 2016, the LIGO announced the detection of gravitational waves (Abbott et al., 2016), which led to a global media outreach and activities on university campuses to comment on the importance of such advances. More recently, in 2019, the Event Horizon Telescope collaboration showed us the first picture of a black hole (Doeleman, 2019). Those are just three examples from biology, physics, and astrophysics, which were significant in the scale of the scientific communication efforts to make these results known to a broad audience. In other words, those discoveries exemplify science at its best in the eyes of the public.

On the other hand, the last two decades also brought major scientific crises that profoundly impacted several research fields, including in the Netherlands, where studies in psychology contained fabricated data (Levelt, 2012; Schoonen, 2020). Worldwide, there were cases of fraud, sloppy experimental designs, and irreproducible results, which brought a tumultuous side of science to the surface (Mehra et al., 2020; Munafò et al., 2017; Steen et al., 2013). Across the globe, fabricated studies and flawed results led to articles being retracted (Steen et al., 2013) and provoked long-lasting damages as m+uch as public distrust in scientific findings (Jamieson et al., 2019). Furthermore, a tumultuous side of science has been observed by many citizens around the globe. For example, in COVID-19 times, doubts have been cast on studies published too fast with profound policy implications. Besides, the massive production of publications and preprints has been

nurturing an open science agenda for more transparency of (clinical) study results and sharing of scientific data (Homolak et al., 2020).

That being said, not all events of irreproducible results are rooted in similar causes. For instance, irreproducible results may occur due to the wrong application of algorithms during experimentation. In that case, irreproducibility is due to errors and is not committed intentionally, for instance, to deceive the scientific community. Algorithms, software, and datasets are artifacts, i.e., by-products of scientific experimentations, that are created during scientific experimentation. Sharing those artifacts used or produced during experimentation can potentially facilitate the identification of such errors by letting independent researchers rerun the published analyses (Hothorn and Leisch, 2011). To make the detection of such mistakes easier, RDM materializes a solution to this problem by facilitating the preservation and sharing of software and the data from experiments.

Nevertheless, some scientific publications report on a growing list of deficiencies in the academic system such as inefficiency (Burley, 2017), irreproducibility (McNutt, 2014), irrelevance (Miller et al., 2011), untrustworthiness (Yarborough et al., 2019), and sloppiness (C. L. Williams et al., 2019). Hereunder, we explain the rationales behind these, not mutually exclusive, deficiencies:

- Inefficiency in scientific knowledge creation processes is perceived as rooted in the way scholarly communication operates. There are incentives to disseminate poorly designed studies that waste (costly) experimental resources (Ioannidis et al., 2014). At the same time, it is challenging to capture poorly executed research without additional information such as documented protocols, data, and software. Therefore, the lack of research data availability beyond publications is seen by funding agencies as something open science is expected to correct in order to make science more efficient (Burgelman et al., 2019)
- Irreproducibility is due to the unavailability of resources such as code and data; published results cannot be verified easily. This criticism led to a movement known as reproducible research (Peng et al., 2006; Peng and Hicks, 2020; V Stodden et al., 2014) and is now fully integrated as one of the objectives of open science: making science transparent and enabling the verification of previous results (Nosek et al., 2015)
- Irrelevance stems from the idea that the incentives to publish also leads to studies with questionable relevance as they mostly serve career advancement (Miller et al., 2011). Irrelevance has also appeared in research with strong practical incentives such as

translational medicine, where published experiments were hard to implement for drug innovation and clinical trials, pointing to the lack of reproducibility as a reason for irrelevance (Huang and Gottardo, 2013; Rahimzadeh and Bartlett, 2014)

- Untrustworthiness is a critic that occurred in clinical research, where flawed trials raised concerns about how well science (and scientists) can be trusted in the eye of the public (Jamieson et al., 2019; Laine et al., 2007a)
- Sloppiness is a questionable research practice leading to reporting errors that may lead to retractions (Haven et al., 2019; Le Maux et al., 2019; C. L. Williams et al., 2019) and has been reported as being more problematic for (junior) scholars than fraud (Bouter et al., 2016; Haven et al., 2019)

The research community appears to concur with these observations, at least if we consider the opinions from 1576 researchers, collected by Nature Publications in 2016, where more than half of the respondents, representing varying scientific backgrounds, stated that the reproducibility crisis is significant and more than 70 percent failed to reproduce published findings at least once in their careers (Baker, 2016a, 2016b). These (sometimes extreme) cases posed severe challenges to the design and implementation of data management initiatives in academia, as the number of issues to tackle and the divergences between (and within) disciplines is quite large.

Therefore, academia is currently undergoing a close examination of how research is conducted (Grimes et al., 2018). Predominantly, open science tackles communication, reproducibility, incentives, and free access to scientific knowledge (European Commission, 2015). Therefore, a significant part of improving research resides in how scientific data, software, and other relevant material can be better preserved and shared between scientific communities to make research communication more transparent and reproducible (Prager et al., 2019). In consequence, a crucial element of the success of OS is the proper management (also known as stewardship) of scientific artifacts during the whole research lifecycle, from data creation to publication of results (Simms et al., 2016).

Likewise, academic stakeholders, such as public funders and library services, tend to agree that scientific practices such as those listed earlier, need a response from a research governance point of view. For instance, several public funders get more involved at the start of research projects by implementing more stringent rules for managing research data (Akers, 2017; European Commission, 2016b). These new requirements from funders led to universities' efforts to invest in research data management with technology, human resources, and training to support researchers

(van Reisen et al., 2019). The recent uptake of research data management (RDM), as a means for funders to (partly) correct these inefficiencies, is an intriguing attempt to transform scientific practice. Therefore, RDM, as an object of study, has recently attracted more attention from the scholarly community, as it has become the focus of essential stakeholders in academia.

Moreover, RDM has recently become a topic of interest to the Information Systems research community (Wilms, Brenger, et al., 2018; Wilms, Stieglitz, et al., 2018). RDM first emerged in information science and computer science disciplines. Most of RDM research until today embraces a wealth of scopes and perspectives. For instance, computer science focuses on the technological level (e.g., large-scale storage systems, efficient data compression), and search algorithms and information science on the stakeholders and ways to get the value out of the data (e.g., challenges to re-use and curate research data, RDM support services in research institutions). In the literature, developments about RDM are discussed in other scientific disciplines (Pryor, 2012) and even policy initiatives across OECD countries (OECD, 2007). Therefore, RDM is fundamentally a nascent object of study. There is much room to investigate the many aspects of RDM, namely the integration of RDM policy, infrastructure, and RDM for diverse types of research (Paul Ayris et al., 2016).

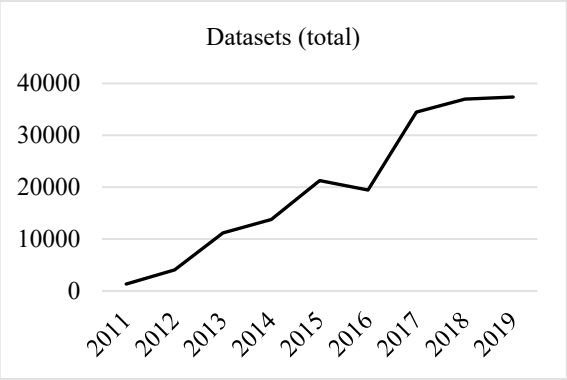
Another reason for diving further into RDM is that the development of RDM also sheds light upon a complex reality of managing the (laboratory) resources needed for the production, preservation, and dissemination of scientific knowledge (Baykoucheva, 2000; Goecks et al., 2010). In an open science context, as described in Section 1.1, RDM is a new frontier of research policymaking with emerging new expectations on the way research results are communicated (Munafò et al., 2018). Remarkably, RDM is evolving in a way that appears unfamiliar to longstanding data management concepts and practices found in business and industry literature (Earley et al., 2017). Hence, many facets of the management of research data have yet to be explored as they are not well represented in traditional data management and governance frameworks. Moreover, data sharing practices are not systematic, and the RDM needs of scientific fields widely differ. As an example, the number of datasets indexed by Dimensions.ai, which is shown in Figure 1.1, with data from Digital Science[1]. A growing number of citable data sets is being deposited in generic repositories (see Figure 1.1.), which then receive a unique identifier that can be cited in scientific publications (Silvello, 2018). Citable data with unique identifiers is one of the foreseen

[1] Data from Digital Science, Dimensions, available from https://app.dimensions.ai and accessed on July 8, 2020, under a license agreement.

outcomes of proper data management from the start of a research project. One side-effect is that datasets can also become the target of quantitative analysis of the publication landscape.

Figure 1.1

The Amount of Citable Research Datasets in Zenodo, Dryad, and Figshare From 2011 to 2019



Overall, this introductory chapter sets the scene and background of RDM and open science. We first elaborate on why RDM is of interest, and hence we motivate our research in Section 1.1. Next, we introduce key concepts such as openness, socio-technical systems, and experimental systems in Section 1.2. In Section 1.3, we introduce the necessary conceptual background. Finally, we present the research design and research questions in Sections 1.5 and 1.6, respectively.

1.1 Why Research Open Science and Research Data Management?

Open science, while recent, has many definitions that can be identified in the literature. One dimension of open science is, in its original form, to be an extension of open access. Open access is the free access to scientific literature at the time of publication and with permissive licenses for reuse in multiple scenarios. As stated by the Budapest open access initiative, open access is "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself." (Chan et al., 2002) Open science is a continuation of open access, promoting open research data and software for reproduction and reuse.

The richness of the open science concept in the literature brought Vicente-Saez and Martinez-Fuentes to define open science comprehensively, encompassing the networked nature of modern scientific practice: "Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks." (2018, p. 434). Fecher and Friesike (2014) opted for a representation of open science as schools of thoughts, showing that open science covers several dimensions that augments Vicente-Saez and Martinez-Fuentes 'definition by giving goals to open science, namely: efficiency, alternative measurements of performance, broader access to knowledge including to citizens and the evolution of infrastructure for science.

Similar to open science, research data management (RDM) is perceived differently by different actors in academia, ranging from the technicalities of software and algorithms to manage large scale scientific data (Shoshani and Rotem, 2009), the management of research data throughout research data lifecycles (Cox and Tam, 2018), and the creation of new roles (e.g., data stewards) to curate and enrich research data (Jones et al., 2020; Wilkinson et al., 2016). Nevertheless, the main ambition of RDM can be summarized as Plomp et al. suggest, where RDM is a means of standardization by "implementing standard practices for accurate data collection and processing, documentation and analysis" (Plomp et al., 2019).

The efficient preservation and dissemination of scientific artifacts require research data management (Corti et al., 2014), as explained earlier. For that reason, academic institutions, and funders, among other stakeholders, introduced research data management (RDM) as a driving force of Open Science (OS). OS is expected to lead to more transparent and reproducible research in academic institutions. Since RDM is seen as a pillar of Open Science, we need to understand RDM's current challenges and the role of information technology in shaping the future of OS. In this dissertation, we consider RDM as a collective means to achieve open science, and therefore opt for the concept of RDM instead of data stewardship. Data stewardship is, in our view, a role in RDM, where RDM is the more generic perspective. This way, we strive to balance challenges experienced by academic stakeholders such as funders, publishers, library services, IT, and researchers; instead of exclusively focusing on a narrower view of research data in laboratories.

According to the (corporate) data management book of knowledge (DAMA-DMBOK, 2009), data management is "the planning, execution, and oversight of policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data and information assets" (Mosley et al., 2010, p. 18). A lighter definition of data management is "managing data to achieve goals" Cervo & Allen (2011). In academia, the goals of managing data are manifold, with open

science emphasizing transparency, reproducibility emphasizing accuracy of reporting and privacy, mixing ethical considerations with the management of research data.

Scientific laboratories are environments operating experimental systems that use or produce research data. The operationalization of experimental systems creates (or re-uses) many resources, some of which fall under the scope of research data management (RDM). Thus, most resources underlying the scientific experimentation lifecycle are relevant for RDM (Corti et al., 2014). Nevertheless, experimental data is a complex object to comprehend, as Bogen and Woodward (1988) note about experimental data representing phenomena instead of observations.

1.1.1 Open Science in the Netherlands

Those changes are especially visible in the Netherlands, the country where our investigations have been conducted. The Netherlands is an excellent illustration of how research institutions seek to transform the way they modernize research resources and infrastructure. In recent years, Dutch organizations such as the VSNU (the association of Dutch universities) and NWO (the national science foundation) started to integrate open science principles in funding and evaluation. While these organizations are affiliated with governmental bodies, much of open science efforts occur in communities of practice; see Hislop et al. (2018) for background about communities of practices. In the Netherlands, the "Landelijk Coordinatiepunt Research Data Management" (LCRDM) pursues the standardization of RDM practices nationally and internationally by collaborating with similar organizations, such as the Research Data Alliance (Treloar, 2014).

The LCRDM and RDA communities gather together academic stakeholders representing governments, researchers, librarians, publishers, and funders to shape the future of research data management and open science. Those community efforts illustrate the vivid activity around open science as well as the early stage in which RDM is developing internationally. Concretely, such collaborative efforts seek to define the governance of research data around seven pillars, namely: support, legal aspects, infrastructure, governance, engagement, and data stewardship (Landelijk Coördinatiepunt Research Data Management, 2019). These pillars have the advantage to serve as an illustration of the context of the studies reported in this dissertation quite conveniently, as this dissertation touches upon four out of the seven pillars of RDM defined by the LCRDM. The pillars (or perspectives) of open science relevant for understanding the context of our study are summarized hereunder.

The first pillar of interest is **funding and governance.** Public research funders drive changes in research policy by disbursing research grants to publicly funded research projects (NWO, 2017). Researchers should also communicate the data sharing conditions to the funders (Mannheimer, 2018; Williams et al., 2017). However, funders' requirements are recent and not yet formally standardized, which raises the question about the efficiency of research data management and governance to positively impact open science (Akers, 2017; Paul Ayris et al., 2016)**.**

Second, research institutions foster **data management and stewardship** across the research data lifecycle. Research institutions offer support in the form of training, data stewards supporting researchers, and IT services. Hence, research institutions shape the field of research data management, together with their academic staff (Corti et al., 2014). For instance, data stewards are recent support roles in research institutions. Data stewards act as intermediates between research individuals (or groups) and funding agencies for creating high-quality data management plans, besides helping researchers with technical or legal advice about research data (The FAIR Guiding Principles for Data Stewardship: Fair Enough? 2018; Victoria Stodden et al., 2019; Thompson et al., 2020). In the business and industry domain, data stewards are also roles in enterprise data management as intermediates between business and data management (Brazas et al., 2017; Plotkin and David, 2013).

Nevertheless, data stewards in academia are hard to compare with data stewards in enterprise data management (EDM). Research data stewards evolve in another context and exercise different responsibilities than EDM stewards. For example, EDM stewards are responsible for providing high-quality reference data in companies by curating enterprise data and forming a role embedded in businesses' data governance strategies. Research data stewards are much more recent, with tasks that may vary across similar job roles.

Third, the development of **research infrastructures** that accommodate data management and stewardship (Houssos et al., 2014; Shoshani and Rotem, 2009; Zondergeld et al., 2020). Research infrastructures proliferate, as a single facility might serve different customers (research groups or businesses). There are also special funds from funding bodies allocated for building or maintaining research infrastructures. The ambition of research infrastructures is to support the increased computational and data-driven nature of several scientific disciplines. Also, exchanging and preserving research data becomes the capabilities of new infrastructure “by design.” In other words, preserving and sharing (operational) data has to be planned and included in the technology used in the labs, such as instruments and software for which principles such as FAIR also apply (Lamprecht et al., 2019; Wilkinson et al., 2016).

Finally, the fourth pillar, communities such as the RDA as well as the LCRDM, also aim at fostering the adoption of proper management and data sharing practices. Thus, **raising awareness** is one of the objectives of networks such as LCRDM. As the responsibility of managing data ultimately lies on the side of researchers (Wilms, Stieglitz, et al., 2018), research institutions feel compelled to educate researchers to use the available services and infrastructure to their maximum potential.

1.1.2 Open Science With a Socio-Technical Lens

The previous section explained how research institutions react to a series of deficiencies found in modern scientific practice. We also depicted open science in the Netherlands, using a local landscape to introduce other local and global organizations involved in RDM. This section positions the RDM landscape previously introduced into an information systems perspective, using a socio-technical systems (STS) lens. STS has a long history in the IS and software engineering domains, as STS broadens the understanding of the context in which (software) systems are designed. To illustrate this, we show how the RDM landscape and technology aiming at supporting the ambitions of RDM align with basic STS constructs found in IS design. We map, in Table 1.1, how academic stakeholders (e.g., funders, services and, researchers) and organizations encountered in the RDM enterprise to Sommerville's socio-technical perspective (Sommerville, 2016; Vicente-Saez and Martinez-Fuentes, 2018).

Table 1.1

Aspects of Research Data Management Classified According to Sommerville's Socio-Technical Items (Sommerville, 2016, p.557)

Scope of STS	Examples of (future) applications	References
National laws & regulations	Amsterdam call for action open science	Amsterdam Call for Action on Open Science (2016)
Organizational strategies & goals	Open science programs	Open Science - Utrecht University (2020)
Organizational culture	Openness and transparency	VSNU (2019)
Organizational policies & rules	Institutional research data management policies	Landelijk Coördinatiepunt Research Data Management (2019)
Operational processes	Data preservation, data exchange, data curation	Corti et al. (2014)

Technical systems	Research information management systems, scientific instrumentation, etc.	Hedges et al. (2007); Hood & Wilson (2001)

In the next section, we further introduce the theoretical elements of technical systems, as exemplified in Table 1.1. In other words, we dive into open science from the perspective of laboratory work and related technologies. Nevertheless, while the legal and institutional efforts about RDM remain intelligible without much background in the topic, laboratory work and scientific experimentation require a more detailed explanation. In Section 1.2, the necessary concepts are introduced. Next, in Section 1.3, the research problem is further specified using the concepts introduced in Section 1.2.

1.2 Concepts of Laboratory Work in an Open Science Context

After briefly introducing the overall context of our study, we now specify the concepts at the core of this dissertation. We first start with "opening laboratories" before explaining why we adopt a socio-technical perspective. Then, we explain why and how a lens of "reproducible experimental systems" is applied during our investigations in laboratories.

1.2.1 Opening Laboratories

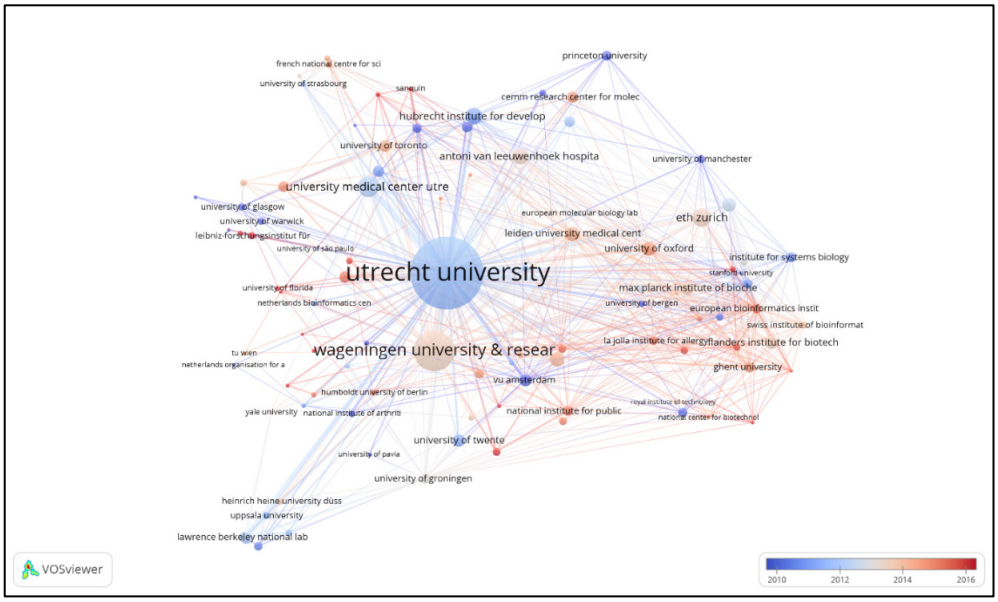
Nowadays, open science initiatives encourage research organizations to adopt stricter procedures for the management of data. These initiatives aim to improve the efficiency and transparency of science (Nosek et al., 2015). At the same time, many difficulties are experienced by working scientists to comply with the open dissemination of research data. In other words, research data is not fully open "yet," and hence the use of the term "opening" instead of open. The use of "opening" emphasizes that opening laboratories is a long-term process with many challenges lying ahead (Borgman, 2020).

The international nature of laboratory work forms an additional issue for understanding laboratory work. To illustrate this, we show a co-authorship network of laboratories of the Mass spectrometry and Proteomics group (see Figure 1.2). The purpose of showing this network is to highlight that research projects are often collaborative, involving groups inside universities and across institutions worldwide. Therefore, studying one laboratory and its data management practices also brings insights into collaborations and data exchanges with other laboratories, which provides evidence on RDM challenges beyond the local level.

Consequently, “opening laboratories” refers to our activities conducted as external observers seeking to obtain more insights about laboratory resources and processes, as well as in the management thereof. In other words, we open a window on a type of organization that often appears to us through publications presenting curated experimental results and methods. By positioning ourselves in a laboratory, we attempt to understand how laboratories are currently preserving and disseminating data as well as the challenges that occur. As we will see in more detail later, challenges in managing data are not purely technological but also emerge from the current situation in research organizations and scholarly communication. Hence, we felt the need for a more comprehensive approach to investigating diverse phenomena that constitute RDM, as further explained in Section 1.2.2.

Figure 1.2

Interactions Between Laboratories in the Last Ten Years Based on Co-Authorship of The Biomolecular Mass Spectrometry Group at Utrecht University.



1.2.2 A Socio-Technical Perspective

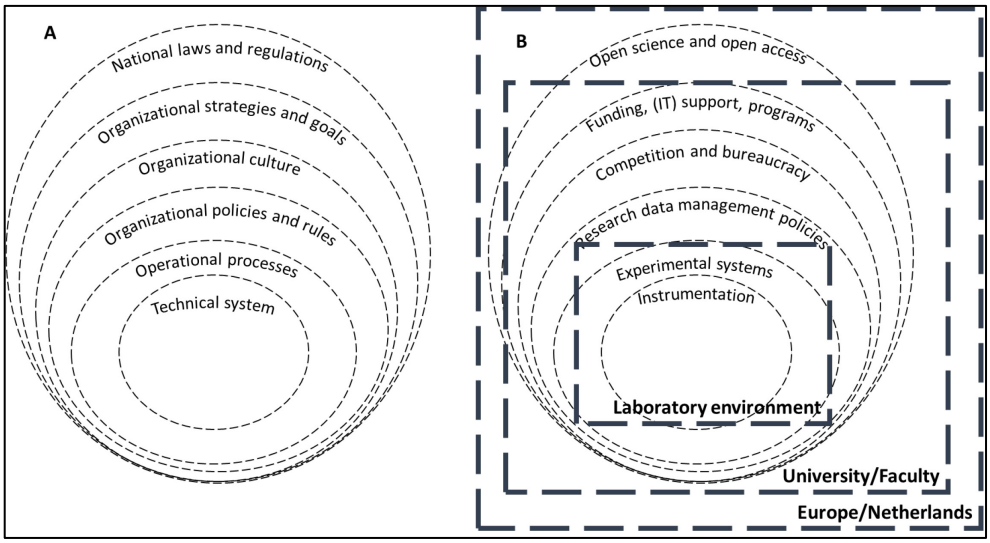
The application of a socio-technical lens on organizational and technological phenomena is not a recent endeavor. Organization theory textbooks place the socio-technical perspective due to the work done by researchers at the Tavistock Institute in the 1950s (Hatch, 2018). Historically,

STS researchers were interested in studying cases where self-managed groups outperformed industrial processes even though they operated less efficient social or technological subsystems.

Later, a socio-technical system (STS) paradigm echoed in information systems (IS) research as an alternative to behavioral methods to study the use of technology in organizations. STS is a lens considering the organizational dynamics in which IS systems are developed (Mumford, 2006). Socio-technical systems design (STSD), discussed in Baxter & Sommerville (2011), seeks to "integrate social and organizational insights from workplace studies into the systems engineering process" (Baxter and Sommerville, 2011).

Figure 1.3

Left (A), a Socio-Technical Framework Introduced by Sommerville (Sommerville, 2016). Right (B), the *Application of Sommerville's Socio-Technical Framework to Laboratory Environments.*



The left side of Figure 1.3, (A) shows Sommerville's STS framework for software systems engineering, which emphasizes the role of extra-technological considerations such as organizational culture and regulations on the success or utility of technical, e.g., novel software systems. Sommerville adds that "generally, large socio-technical systems are used in organizations. When you are designing and developing socio-technical systems, you need to understand, as far as possible, the organizational environment in which they will be used" (Sommerville, 2016). Therefore, a large part of this dissertation has an exploratory nature and aims at understanding RDM as an STS phenomenon. Next, on the right side of Figure 1.3 (B), we show how Sommerville's

STS framework applies to laboratory environments in research organizations. For example, laboratories are situated in research organizations impacted by external factors such as governmental regulations and the competitive allocation of public funds to research projects. In the next sections, we elaborate on experimental systems that describe the core of operations in laboratories, in Section 1.2.3.

As introduced in Table 1.1, there are several layers to consider in socio-technical systems (Sommerville, 2016). Layers go from distant procedures and stakeholders at the national level to the existing system designed and operated by scientists. Briefly summarized, one finds the following layers in Sommerville's framework:

- **National law and regulations** form the legal basis on which (software) systems are designed and are notoriously hard to change, although they influence those systems' behavior (Sommerville, 2016). In open science, an example of changes at a national level is the *Amsterdam Call for Action on Open Science* (2016)
- **Organizational strategies and goals** are the national ambitions translated at the level of research institutions. Universities are well versed in the strategic planning exercise, and open science is lately following a similar path, where bottom-up and top-down initiatives crystallize open science strategy that fit the specific organizational context, such as the open science program at Utrecht University (*Open Science - Utrecht University*, 2020)
- **Organizational culture** in universities is competitive and bureaucratic. Competitive (funding) resources are scarce and primarily financed through competitive means (grant applications) or industry partnerships. The added value of research finds its legitimacy in notions such as impact and excellence. Moreover, universities are bureaucratic cultures where hierarchies are prevalent, specialization (through a person occupying a role), division of labor (e.g., between faculty and support) and, seeking to standardize processes and operations (e.g., tenure tracks, research unit evaluation, among others)
- **In the context of our study, organizational policies and rules are** open science and research data management policies that are developed across faculties and central support services.
- **Operational processes** are the processes related to the functioning of laboratories. There are numerous processes found in laboratories; therefore, we focus exclusively on scientific experimentation and data management.
- **The technical system** refers to the technology employed in laboratories to conduct experiments.

1.2.3 Reproducible Experimental Systems

In this section, we further describe the concept of a reproducible experimental system. Reproducible experimental systems (RES) are a conceptual representation of scientific experiments originating recently from the philosophy of science. The concepts of RES were first introduced by Radder (1992) and later by Rheinberger (1997). The reason why RES was introduced is that, according to Radder, reflections about science and experimentation were limited to methodological, epistemological, or logical discussion, disregarding the actual operationalization of experiments Radder (2012b). To a certain extent, RES show familiarities with STS, for instance, when we look further into the composition of such systems. Rheinberger defined an experimental system (ES) as "a basic unit of experimental activity combining local, technical, instrumental, institutional, social, and epistemic aspects." (Rheinberger, 1997). As a result, Rheinberger's definition of an experimental system includes notions such as instrumentation and social practices distinct from methodological and epistemic considerations of scientific experimentation.

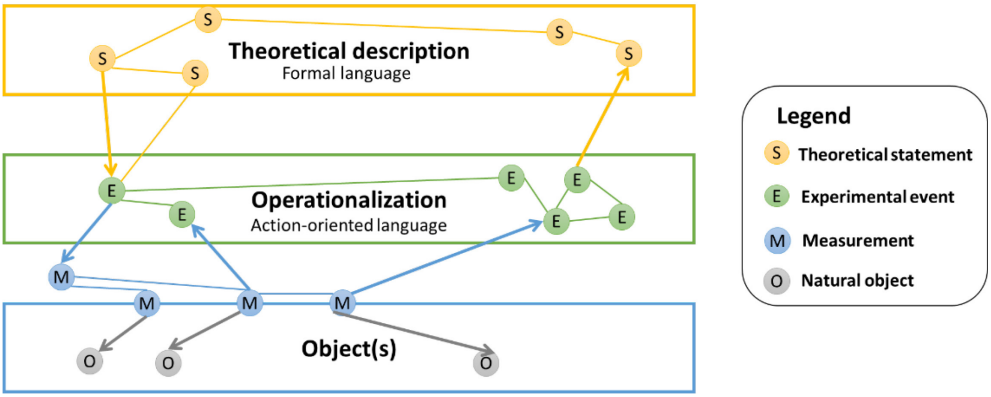
Similarly, Radder uses (successful) experimental systems (ES) to understand how experiments are constructed by experimenters (Radder, 2012b). To that end, Radder defines experimentation as a process that "involves the realization of several manipulations of the object and the equipment, brought into mutual interaction, and the theoretical description (or interpretation) of these manipulations and their results" (Radder, 2012a). In short, as shown in Figure 1.4, measures on the object of interests are mediated by instrumentation and theory. Moreover, experimenters operationalize experiments using concepts that differ from their theoretical equivalent. On the one hand, there is the language of propositions and abstractions (formal language) found in textbooks and scientific publications. On the other hand, operationalization demands an action-oriented language, where theoretical concepts are translated into activities conducted by "theoretically uninformed people," as Radder states. Thus, the action-oriented language is the language of laboratory workers.

People and (technical) resources are used to conduct (operationalize) experiments. Experimental resources cover a wide range of items needed to conduct experiments in laboratories. Among the vast resources existing in laboratories, we introduce only basic kinds of resources related to experimentation. Here, we use Radder's classification (2012), which divides experimental systems into three subcomponents: a theoretical description, a materialization process, and experimental results. As an instance of a general system encountered in general system theory, ESs

are complexes of input, process, and output, with characteristics such as feedback loops and boundaries (Hatch, 2018).

Figure 1.4

The three components of closed experimental systems and their interactions (Radder, 2012a). The arrows indicate the input and output flows. Lines refer to mutually dependent statements, events, and measurements.



In practice, we observed that scientific laboratories distinguish laboratory resources from computational resources and theoretical resources. Laboratory resources correspond to lab instruments, chemicals, samples, physical storage (e.g., fridges and freezers). Computational resources are software, hardware, IT services such as data storage systems, and high-Performance Computing Clusters (HPC). Theoretical resources are used to communicate results and theory through publications, for instance, and serve as inputs and outputs for designing and communicating the results of experiments.

Computational resources might not be directly related to laboratory work, at least not in a single, closed experiment. An example of this is experimental data re-use for secondary analyses, which pinpoints to (simple) computational experiments where no "lab" work is involved. In that case, we deal with an occurrence of experimental systems: computational experiments. To explain this further, we first comment on experimental systems, illustrated in Figure 1.4, and next on computational experiments, which are thus seen here as an extension of experimental systems. Hans Radder defines a *closed* experimental system *S*, as encountered in natural sciences, as a "*complex* of ***object*** and ***equipment*** within a ***specified spatial area*** and **during *a fixed interval of time***" (Radder, 2012a). From Radder's perspective, the instantiation of an ES is guided by a **theoretical description** (i.e., experimental process) and **human intervention** (i.e., operationalization). First, a

theoretical description (TD) delineates the *episodes*[22] of an ES. *Some* episodes have a specific role, which is to determine the *relative closure* of an ES. In short, an experimental system is qualified as being *closed* if non-experimental episodes do not interfere with the operations and results of experimental systems.

Moreover, closure is said to be *relative* to a theoretical system. So, closure is dependent on the *theoretical description* under which the experiment is conducted. During the planning of an experimentation process, experimenters know what kind of interferences might occur based on theoretical knowledge. Nevertheless, theoretical knowledge evolves, also with previously known experimental outcomes. Assuming science is self-correcting (Vuong et al., 2020), published results that are wrong or invalidate, previous findings reflect upon the theory that experimenters will use in future experimental systems. Hence, in Radder's view of experimental systems, it is a possibility that the experimenters do not consider all interferences at time t. The reason is that experimenters might just ignore such interferences as the current knowledge has not yet established their existence.

In addition to a theoretical description (TD), human intervention (HI) is also a determinant of the success of an experiment (Radder, 2012b). Radder sees *HI* as a fundamental activity that initiates the experimental system and *realizes* (i.e., *operationalizes*) the system. Therefore, to initiate an experimental system, human intervention is needed to deviate the sequence of episodes from their "*natural*" behavior, i.e., the behavior episodes would show outside a given ES. Moreover, human action is further required during the experimentation process to translate theoretical knowledge (e.g., a mathematical model) to concrete activities (e.g., manipulation or design of equipment).

Next, computational, experimental systems (CES) are an extension of experimental systems. In this way, we follow the logic of Freire, Bonnet, & Shasha (2012a), who distinguish experiments (i.e., experimental systems) from computational experiments (i.e., computational, experimental systems). (Freire et al., 2012a) define a reproducible experiment as an "**experiment** done by **laboratory L** at **time t** is deemed to be reproducible if it can be repeated at a possibly **different laboratory L'** at **some later time t.'**" Equally, a computational experiment is an "**experiment** that has been **developed at time t** on **hardware/operating system s** on **data d** is

[22] The notion of Episode is used by Radder as an umbrella term covering experimental and environmental aspects of ES such as interactions, processes, results and experimental circumstances (Radder, 2012a).

reproducible if it can be executed at **time t'** on **system s'** on **data d'** that is similar to (or potentially the same as) **d**" (Freire et al., 2012a).

Hence, CESs consist of the same components found in experimental systems except that their environment and the nature of their operations differ. A first difference, as explained by Keller (2003), is that CEs brought computer simulations which "radically extend the range of problems amenable to quantitative analysis" in diverse disciplines such as physics and biology (Keller, 2003). Another difference is that the choice of which software and hardware to combine is (often) left to the researchers' appreciation for running these simulations.

When experimental resources are worth to be shared, such as a data set containing unique measurements of interest to the broader community, those resources can be qualified as research assets. Assets are simply a subset of resources with a value, even after the experiment has been conducted. Thus, computational research assets are digital resources generated by experimental processes in laboratories and which laboratory members find worthy of preserving or sharing. So, not all resources generated by ESs are worthy of preserving, but some that are might still be considered as information waste by experimenters (Ioannidis et al., 2014). An example is the deletion of digital files essential for understanding or repeating experiments or inadequate reporting of results. In that case, RDM needs to effectively guarantee that the experimental resources are preserved so that reproduction and openness are possible.

1.3 Reproducibility of Research

Earlier, we described experimental systems as one system describing the core components and processes of laboratory work. This section discusses how managing data produced by experimental systems can benefit open science even though numerous challenges persist in the current situation. Therefore, openness and reproducibility of science, which often appear linked together in the literature (Braun and Ong, 2014; McCormick et al., 2014), posit conceptual and operational difficulties in the absence of a delineated object of study. First, we present related work about open science (OS) in Section 1.3.1. Then, we further comment on reproducibility in the context of open science in Section 1.3.2.

What emerged from the previous description of RDM in an open science context is that (1) RDM research touches upon different aspects of scientific experimentation, (2) much of the literature discusses experimental systems as closed systems, while open science and reproduction might challenge this assumption. In other words, the control experimenters have on the operationalization of experiments becomes a much more distributed endeavor among academic

stakeholders, where parts of experiments can "easily" be transposed or included in different experiments at any time.

In that context, the answers RDM should offer to the reproducibility issue in an open research landscape are still limited by a misalignment of open science with the (information) technology as used in laboratory work and scholarly communication. The misalignment of technology and open science strategies is where the role (or goal) of reproducibility reaches numerous misconceptions and challenges that we will introduce in the next two sections.

1.3.1 Views on Reproducibility in an Open Science Context

Open science is distributed science, thus distributed operationalization of experiments. Vincente-Saez and Martinz-Fuentes defined open science (OS) as "transparent and accessible knowledge that is shared and developed through collaborative networks" (Vicente-Saez and Martinez-Fuentes, 2018). The authors coined this definition based on an extensive literature review, which revealed two fundamental dimensions of open science. Firstly, the knowledge dimension, which states that knowledge is transparent and accessible. Secondly, a network dimension which states that scientific knowledge is shared and produced collaboratively.

The first dimension, transparent and accessible knowledge, is found in open access and focuses on making scientific results accessible to a broad audience. The network dimension asserts that sharing is an inherent part of the scientific enterprise. For reproducibility, this posits quite some challenges as experimental systems are then distributed efforts and should, at the same time, be communicated transparently. As being STS in nature, experiments, their resources, and experimenters are not as transposable as reproducible resources and methods would require.

Furthermore, reproducible science has some taxonomic and scope issues in the literature that challenge our understanding of what reproducible research entails. In other words, the comprehension and extension of the notion of reproducibility are not reproducible across the literature. The large variety of scientific disciplines, as Barba (2018) demonstrates with an account of the many taxonomies of reproducibility. Therefore, we propose a taxonomy that has the advantage of being aligned with the description of reproducible experimental systems, which we described earlier in Section 1.2.3. Besides, the taxonomy presented hereunder integrates previous classifications, including the Association for Computing Machinery (ACM) artifact badging classification (Ferro and Kelly, 2018).

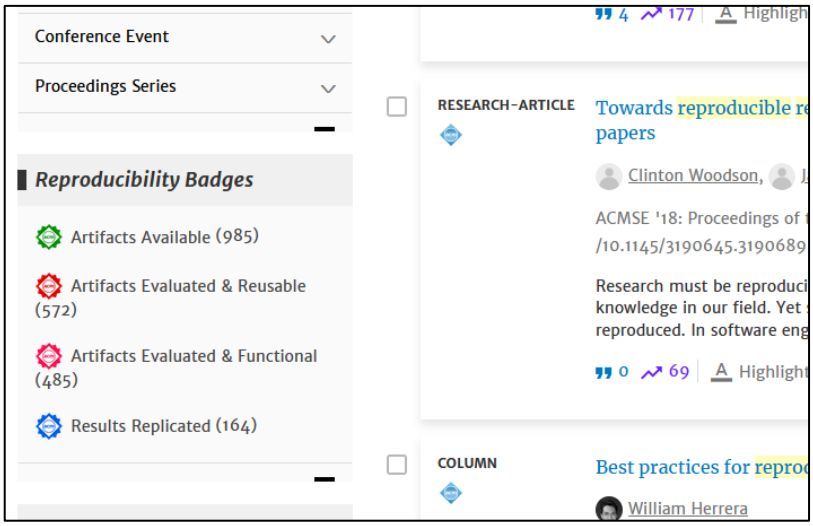
In the ACM classification, which is tailored to computational artifacts (ACM, 2018; Ferro and Kelly, 2018), badges are assigned to publications in the ACM library (as illustrated in Figure

1.5). Figure 1.5 shows search results on the ACM library that displays additional filtering criteria to select publications with a specific artifact quality standard (e.g., artifact available, or one level above, reusable). This classification is an example, among many others (Feger et al., 2019; Wouters et al., 2019), of evaluation systems with enhanced interfaces that implement capabilities that foster reproducible research in (digital) libraries.

Interestingly, the ACM taxonomy classifies research results and artifact quality separately. The results of articles published in the ACM library can be classified as replicated or reproduced. According to the ACM, results are replicated when an independent author (or team) achieve similar results using the original artifacts. Results are said to be reproduced when similar results are achieved without the use of the original artifacts. Regarding the quality of the artifacts, the evaluation uses two levels: functional and reusable. The functional level is achieved when a set of criteria is met, such as the availability of documentation and the fact that the artifact(s) can be operated again. An artifact classified as reusable is a functional artifact that exceeds the requirements of a functional artifact by adding structure to the code and data in a way that facilitates its re-use by others.

Figure 1.5

Reproducibility Badges From the Association of Computing Machinery (ACM) as Shown on the ACM Library Page

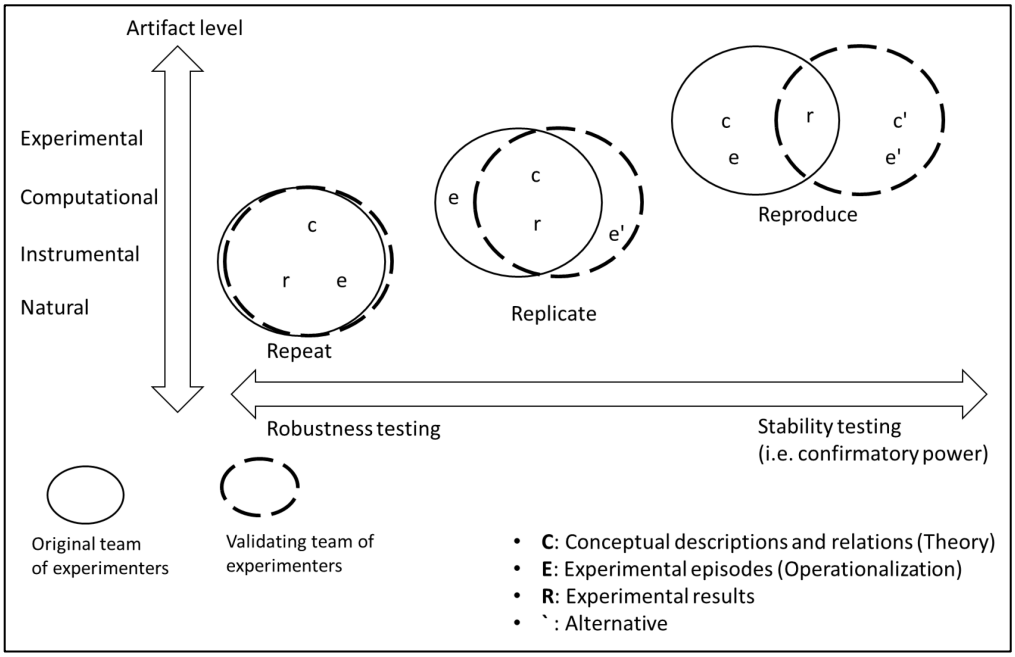


Furthermore, Figure 1.6 shows that experimenters must evaluate at which level and purpose the reproduction of an experiment should fulfill. For instance, the reproduction of

measurements refers to the use of replicated samples in biology so that experimenters can prove that the instrumentation worked as expected. In that case, the purpose of reproducibility is to act as a sort of sieve sorting out experimental results from experimental waste (or noise). Nevertheless, the purpose and level of reproduction are sometimes overlooked in the literature. For instance, when Peng suggests a reproducibility spectrum, it is essentially instrumentally or computationally focused (Peng et al., 2006). When statisticians argue about misreported results, they limit their scope to a methodological level of reproducibility (Simons, 2014), however the methodological level is only a part of what can be reproducible when we depict experiments as experimental systems.

Figure 1.6

Overview of The Notions of Repeat, Replicate, and Reproduce as Different Levels of Artifact Re-Use and Testing Purpose Between two Independent Teams of Experimenters. The 'Symbol (as in C') Means Alternative. Thus C' Stands for Alternative Conceptualization.



The conceptual overview of reproducibility is illustrated in Figure 1.6 and detailed with an example in Table 1.2. In that framework, *repeat* is seen as the most basic form of validation of experimental results. The original team and the validating team of experimenters share identical concepts and episodes and seek comparable results. *Replication leaves* more room than repetition

for the validating team to alternate experimental episodes while still restricted to the original concepts and results. *Reproduce* is the highest level of validation of an experiment, where the validating team seeks to validate the results of the original team by alternating concepts and events. To achieve that, the validating team might adapt the experimental resources used by the original team. The validating team might also opt to adapt the theoretical assumptions that are held by the original team, for instance, due to recent advancements in the equipment or theory since the original team published their results. Nevertheless, all three notions benefit from understanding the precise circumstances in which the original team operated. It implies that the original experimental resources are made available with high-quality meta-data to enable the validating team to conscientiously investigate the original team's assumptions, operationalization, and results.

Also, repeat, replicate, and reproduce different levels of an experiment. *Repeat* is mostly a technical endeavor. Reproduction encompasses far more decisions and knowledge about experimental conditions than necessary to repeat an experiment. Table 1.2 shows each experimental artifact level, namely, Natural, Instrumental, Computational, Methodological, and Experimental. It also maps each level to equivalent artifacts found in the proteomics domain (i.e., chemistry and biology used to study proteins). The objective here is to show how the theory of reproducibility elaborated earlier refers to concrete examples in a scientific discipline.

Table 1.2

The levels of artifacts, each with a corresponding example from the proteomics domain

Artifact level	Scope	Example from the proteomics domain
Natural	Sample	Blood, urine
Technical/Instrumental	Measure	A Mass-spectrometer
Computational	Software and computer hardware	Proteome discoverer, Uniprot
Methodological	Operationalization and Analysis	Liquid Chromatography MS, Ion-mobility MS (which are two separation techniques)
Experimental	Research Design and Theory	Bottom-up mass spectrometry (MS), Top-down MS, Native top-down MS

While representing incomplete levels of reproducibility, reproducibility badges such as ACM reproducibility badges are consistent with a political and societal move towards broadening the impact of academic research. Nowadays, many indicators measure the impact of research, such as publications-focused indicators such as citations, impact factors (IF), and H-index. Recently,

open science has addressed the problems these metrics posit for rewarding scientists and quality research relatively (Fecher and Friesike, 2014). Hence, making experiments reproducible and artifacts more open can be integrated into the performance assessment culture in science. Some argue that the academic system does not favor such moves towards openness, primarily because of the absence of rewards as making high-quality artifacts does not reflect academic performance as much as publishing articles (Grimes et al., 2018).

Now that we have illustrated the inherent complexity of reproducibility and discussed how research data management could help research organizations achieve reproducible research in an open and networked open science environment; we can finally elaborate on the strategic aim of this dissertation, which is to seek to help laboratories design open science with the help of research data management.

1.3.2 Design of Open Science Readiness in Laboratories

Our study's ambition is to develop information technology to shape the future of RDM, more specific approaches and tools that can yield insights into data management practices of laboratories to laboratory managers, principal investigators, data stewards, and other institutional actors involved in RDM programs. As we will see in the next chapters, many challenges lay ahead. Nevertheless, we can refer to Figure 1.7 to show how we design RDM to make laboratories more resilient to the open science future. It starts with inquiring about national developments around open science. Then, we go one layer deeper and investigate data stewardship programs in organizations. We also note that there is a culture of competitiveness in academia doubled by a very bureaucratic structure. Changes at the national level introduced an agenda for open science, data management policies, and support at the university, faculty, or departmental level. That being said, we are not yet at the core of laboratory work as we are still touching upon the structure of academic institutions and their organizational culture. The core of laboratory work starts at the layer of experimental systems and rely upon instrumentation. Here, the instrumentation is seen in a broad sense: lab equipment, software, computational hardware, chemicals, and samples.

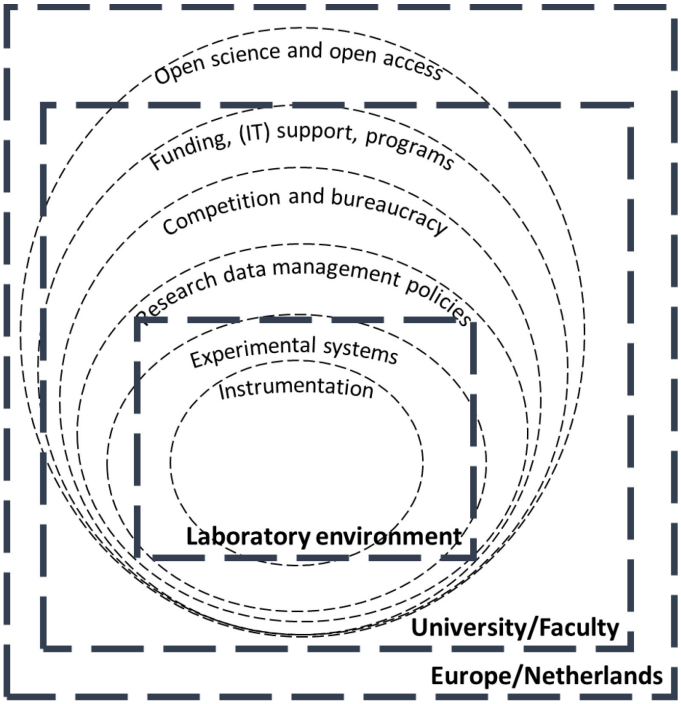
There are trends in policy and research governance that converge towards enforcing the effectiveness, efficiency, and transparency of science. Although seeking efficiency and accountability in research and education institutions is quite an old question (Kelchen, 2018), these more recent policies have gained more visibility in new research funding frameworks, such as the European H2020 program. Therefore, it is vital to keep in mind that the reproduction, management,

and dissemination of research data falls under the bigger picture of high-quality, impactful scientific practice.

Nevertheless, laboratories as organizations producing scientific knowledge are still misaligned with open science, partially to the significant changes in scientific practice required to achieve the ambitions of open and reproducible science successfully. Chapter 7 presents the reconciliation of research data management challenges with open science by allowing scientists to monitor open science readiness according to several(key) dimensions.

Figure 1.7

Research Organizations Depicted as a Layered Socio-Technical System



The concept of readiness, primarily inspired by the digital forensics' domain, refers to the extent to which laboratories can respond to legal, institutional, and experimental requirements for preserving and sharing research data. Readiness is concerned with the capabilities of laboratories to manage their resources in a resilient way, meaning that laboratories comply with recent developments in the academic landscape, namely: openness and reproduction—more about open science readiness in Chapter 7.

1.5 Research Questions

In this section, we detail the research design. We begin with the (main) research questions. Six research sub-questions divide the main research question (MRQ), each answered in a corresponding chapter. First, the main research question (MRQ) is a design science question (Wieringa, 2014). Answering a question of how to design something involves studying the environment, the actual design of artifacts in the studied environments, and finally, the evaluation in real or simulated settings through focus groups or surveys (Wieringa, 2014). Next, the sub research questions (coded SR) are guiding our exploration and intervention across many different paths. We list the research questions below and refer to the next sub-sections for more detailed information about each question. The research methods are described in Section 1.6.

Further, using a socio-technical lens shapes the presentation of the results in this dissertation in a more understandable way. All chapters included in this dissertation tackle one or more STS aspects, as depicted in Figure 1.8. Figure 1.8 describes how this dissertation's research questions relate to a standard STS framework found in the Information Systems literature (Bostrom and Heinen, 1977). The STS framework links the different social and technical components found in organizations and technology. Research data, publications, instruments, laboratory members, and software are elements of a socio-technical system. Another benefit of STS is to emphasize that we are addressing research data management from a holistic perspective, where operational challenges and technology form a subset of our analysis as many other factors are to be considered. Therefore, the choice of a holistic perspective deepens our understanding of research data management, laboratory work, and scholarly communication altogether.

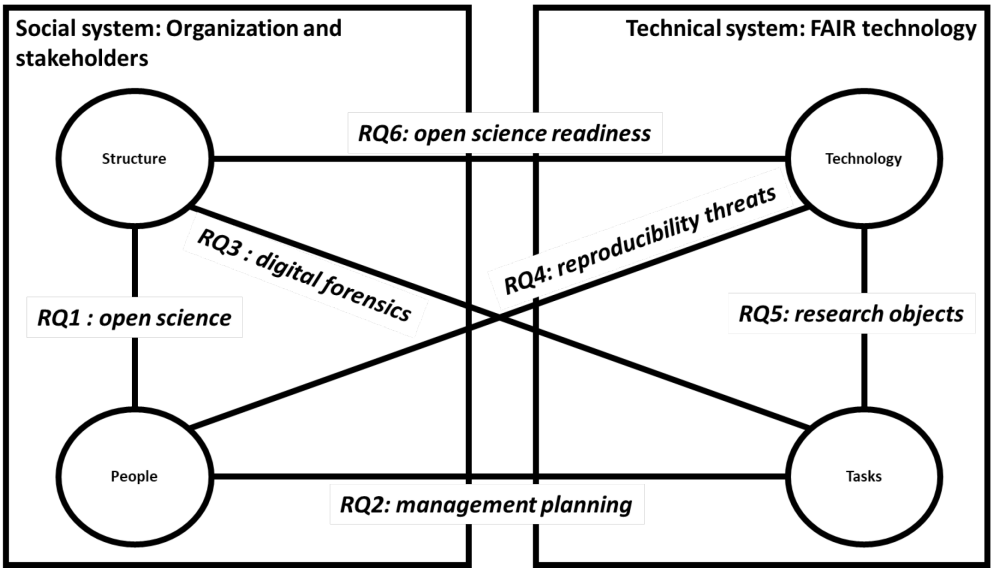
Under a socio-technical view, the laboratory is seen as a workplace where scientists conduct experiments using theory, models, and technology. Laboratories consist of a mixture of closed and open systems integrate into the four STS components, i.e., structure, tasks, people and, technology. The components of STS shown in Figure 1.8 are *structure*: "Informal channels used to communicate and negotiate in the workplace." (Hesketh and Graco, 2015). Then, *people*: "Cultural and other diverse groups and individuals.". Next, *technology*: "The equipment, infrastructure, and technology." (Hesketh and Graco, 2015). And, finally, *tasks*: "The combination of techniques in use." (Bider and Perjons, 2017, p. 99).

For instance, the scholarly infrastructure is open and has a codified way to disseminate research results through peer-reviewing. On the contrary, scientific experiments are closed systems isolated from any uncontrolled interference and require highly knowledgeable workers to succeed.

Although external events might interfere with experimental systems, the experimenters' role is to account for those interferences (Radder, 2012a).

Figure 1.8

The Chapters Included in This Dissertation Glue the Elements Of Socio-Technical Systems. The STS Representation is Adapted From Bostrom & Heinen (1977)



The main research question (MRQ) stated here asks how information systems can support open and reproducible science. The main research question of a design science paradigm (regardless of its chosen flavor) is a design question where artifact building and acquisition of new knowledge work in concert (Thuan et al., 2019; Wieringa, 2014).

MRQ: How can we organize research data management for preserving and disseminating laboratory experiments in a reproducible way?

There are emerging design principles for managing digital resources in the infrastructure landscape, such as findable, accessible, interoperable, and reusable (FAIR) principles for data repositories. We qualify them as emergent as FAIR principles as research communities still discuss and seek to implement FAIR principles (van Reisen et al., 2019). Therefore, the design of information systems for open science is a novel and explorative inquiry. Hence, it is difficult to build upon previous knowledge by observing such information systems in use as they are simply

not available yet. The absence of information management systems tailored to open science invites us to adopt a research paradigm that allows the creation (or design) of a series of artifacts. In information systems research, this paradigm is known as design science research (Gregor, 2013; Hevner et al., 2004; Wieringa, 2014). Design science research (DSR) has several flavors that account for the research domain's scope, maturity, or practitioners' engagement in the DSR project (Peffers et al., 2007). For reasons we further explain in Section 1.6, we have opted for one flavor of DSR named action design research (ADR) (Haj-Bolouri et al., 2018; Sein et al., 2011). Thus, we want to discover and share how information systems that can help mana1ge digital resources can be made by others, under other circumstances, and applicable to a range of (scientific) environments. However, as mentioned earlier in Section 1.2, the issues we are tackling emerge from a very recent phenomenon in academia (hence with a very low maturity). Therefore, the overall study is explorative and requires some innovative approaches to conceptualize the problems and develop the methods and tools to collect evidence from our case study laboratory. It resulted in a rich set of approaches that are guided by the following research questions.

RQ1: How can research data management contribute to efficient and reliable science?

We have seen earlier that open science has several goals, among which reproducibility and transparency. Research data management is a new field of practice in academia that is put forward to achieve the goals of open science. Therefore, open science must be efficient (reducing costs and waste of resources) and useful in developing technology and practices fostering reproducibility. The first question is interrogating the link between RDM and OS using exploratory case studies.

RQ2: What are the current challenges and practices in research data management planning?

Research data management planning is currently the most widely used tool by funding bodies and research institutions to address data management at all stages of a research project. Submitted by researchers as part of a grant application, research data management plans (DMPs) are documents that state how research data will be created, preserved, and shared following laws and regulations that apply to the type of data involved (among other criteria). IT services and libraries offer support for writing and checking data management plans and funders to evaluate those plans. However, despite the significant role of DMPs in guaranteeing proper RDM, there are issues in their content and evaluation, which impede the effectiveness of RDM, which is the object of RQ2.

RQ3: How can digital forensic methods and techniques be applied to the investigation of artifact reproducibility?

Researchers' ability to plan RDM for new research projects depends upon the understanding of (peculiar) needs for RDM related to the technology and field they evolve. At the same time, it is questionable whether the current state of RDM is satisfactory, and new planning could simply extend current practices in laboratories. The question is how we can evaluate the relation between RDM and direct forms of reproducibility, namely that experimental resources are preserved so that (potential) replication of studies is possible. As will be discussed further in Chapter 3, this is a heavily interpretive study, where the structure and names of data resources need to be carefully examined. (Digital) forensics techniques and approaches support the analysis of data resources present on storage systems, although their applicability to experimental resources for evaluating reproducibility is still challenging. Therefore, RQ3 explore the applicability of forensics approaches to the recovery of experimental data and position the outcomes using semiotics to describe RDM shortcomings at an information level.

RQ4: What reproducibility threats occurring in experimental systems stem from vulnerabilities in research data management practices?

In RQ 3, we investigated the resources (the technology) of scientific experimentation and scientific reports. In RQ4, we include people and organizations (e.g., scientific fields) in analyzing a comprehensive set of threats to reproducibility. Some of these threats are in the scope of RDM, and other threats rooted in scientific practice. In the end, we challenge the argument that RDM can make science more reproducible, at least if we account for comprehensive STS analysis of reproducibility. This reflection helps us scope what RDM can achieve in supporting reproducibility and what reproducibility aspects lay beyond RDM.

RQ5: How can we bridge interactive data mining solutions with reproducible research object technology?

Research objects (RO) are an emergent technology leveraging the semantic web to compile structured information about experimental processes and products. As aggregates of experimental products, ROs could potentially drive efforts to curate experimental processes and preserve them in a reusable fashion. Nevertheless, current RDM technology in laboratories (and university institutions) does not widely use RO capabilities. Besides, experimental work driven by interactive exploration of large-scale data shows the limits of making decisions explicit. The

division of labor between bioinformatics and biologists is an additional challenge to consider while developing integrated data management solutions for reproducible research. Biologists emphasize interactive solutions while bioinformatics opts for finer-grained, configurable software packages to analyze their data sets. The question is then, is there a way to bridge these two worlds while keeping results reproducible?

RQ6: How can a laboratory forensics approach help achieve open science readiness?

There is an increasing amount of (structured) data available to understand and evaluate scientific practices. Open science does not escape this trend (Wouters et al., 2019). RQ6 drives our investigation into the design of analytic systems to prepare RDM technology and practices for an open science future. We adopt the concept of readiness from the digital forensics' domain, which was explored mainly for answering RQ3.

1.6 Research Methods

In this section, we introduce the various research methodologies used in this thesis. All the methods introduced here are part of an action design research (ADR) approach. The reason we use ADR is that ADR is particularly suited for developing IS artifacts in a specific organizational context (Sein et al., 2011). First, we describe the environment in which our studies are conducted. In Section 1.5.1, we elaborate on the case study approach. Then, in Section 1.5.2, we explain our choice of an interpretive approach to gathering evidence about the current state of research data management. Finally, in Section 1.5.3, we detail our action design research approach, which is the overarching method of inquiry adopted in this study.

The way we approached laboratories is comparable to previous interpretative studies, as found in Latour and Woolgar (1986). More specifically, we studied laboratory practices as external observers. We have little pre-conceptions about how scientific knowledge is produced in a laboratory, with limited understanding of the object of study and an emphasis on the interpretation of qualitative data found in laboratories. Laboratories (shortened as labs) are controlled environments, where properties of natural objects are transformed into facts from measurements to obtain (novel) insights into the inner workings of nature (Latour and Woolgar, 1986). Scientific knowledge production relies on documents (i.e., inscriptions in Latour's terms) from the very early stages of scientific experimentation and, therefore, are described by Latour as a process transforming measurements into scientific facts. This process is socio-technical, as depicted by

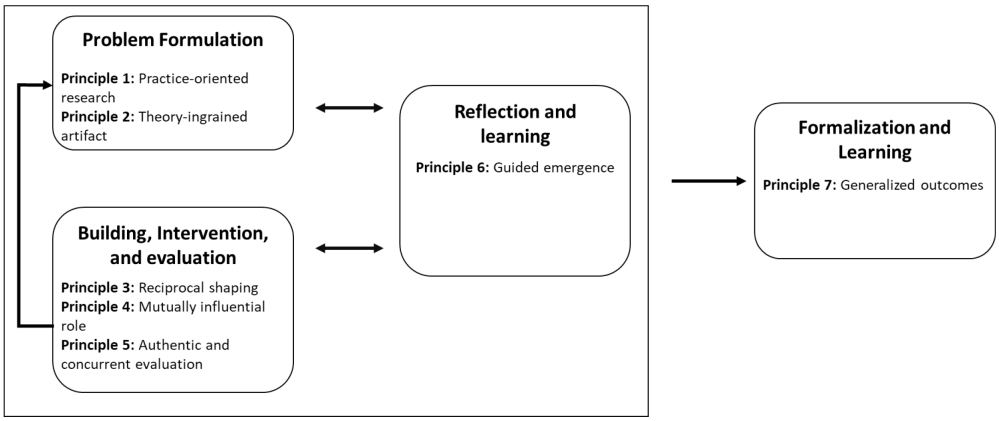
Latour & Woolgar (1986). Experimental resources are a part of the documents, nowadays readily available in digital formats, but not always preserved and exchanged in laboratories.

1.6.1 Action Design Research

Action design research (ADR) is "a method for multi-disciplinary engagement to solve real-world problems" (Haj-Bolouri et al., 2018). ADR provides a comprehensive framework for collecting information about the environment in which IS scholars design and evaluate artifacts. ADR researchers evaluate artifacts according to their utility, as the design of new artifacts has to answer to an intricate organizational or technological issue. ADR is rooted in design science research (DSR), although ADR is more situated in the type of intervention and knowledge that results from applying ADR (Cronholm and Göbel, 2019). Therefore, it shares with DSR the goals of building (novel) artifacts, such as methods and software, to address a real-world problem (A. R. Hevner, 2007).

Figure 1.9

The Action Design Research Phases as Described in M. K. Sein et al. (2011)



The ADR principles shown in Figure 1.9 form the overarching approach that guided our research. The seven ADR principles from Sein et al. (2011) are applied as follows:

Principle 1 of practice inspired research determines that field problems are knowledge creation opportunities (Sein et al., 2011). Open science and RDM start shaping new research practices to which laboratory are confronted. Many questions remain open, as explained earlier. Therefore, our research is primarily an organizational driven attempt to understand and intervene in practice.

Principle 2 of theory ingrained artifacts supposes the use of explanation, prediction, and design theories (D. Jones et al., 2007; Sein et al., 2011). While this is quite challenging to achieve for an exploratory study, we adopted theoretical work such as semiotics (Thellefsen et al., 2018) and experimental systems (Rouse, 2011) as well as previous (technical) approaches that emerged from digital forensics (Rowlingson, 2004) and research objects (Bechhofer et al., 2013), among others.

Principle 3 of reciprocal shaping states that the (IT) artifacts resulting from ADR are interwoven with the domain of study and organizational context, each influencing one another. Therefore, researchers conducting ADR studies develop a better understanding of the environment in which the artifacts operate.

Principle 4 of mutually influential roles is related to principle 3 because learning is also reciprocal among the participants in an ADR project (Sein et al., 2011). This principle is best illustrated by our forensic approach in a laboratory, which required many iterations and discussions among the authors and laboratory workers.

Principle 5 of authentic and concurrent evaluation uses evaluation (e.g., interviews, focus groups) during the ADR project, while artifacts are built, and not once there are complete. This principle has been useful for our study at an exploratory stage. Early feedback had immediate effects on how we apprehended the design of artifacts such as research objects and forensics tooling.

Principle 6 of guided emergence is the extent to which the artifact undergoes refinement emerging from the adoption of the earlier principles (2, 3, 4, and 5). An example of this is the adoption of semiotics in information quality evaluation in a revision of the laboratory forensics (LF) approach (i.e., the ensemble artifact), as it answered a limitation of a previous iteration of LF.

Principle 7 of generalized outcomes is the principle where the ADR solution and its related problem space are generalized through a conceptual stage of formalization and learning (Sein et al., 2011). In Table 1.3, we illustrate the four ADR stages with our study material. We also show how the stages link to the chapters included in this dissertation and the role of the concluding chapter as covering the last two stages of ADR.

Table 1.3

ADR Stages Described In (Sein Et Al., 2011) Applied To This Research

Stage	Explanation
Problem formulation	Laboratory members experience the openness of procedures and artifacts, efficient management, and reproducible preservation of laboratory resources as challenging. Using the lens of experimental systems, we envision an ensemble of artifacts for discovering irreproducible researcher data management practices and supporting open practices in laboratories to extract and monitor evidence from laboratories. Chapter 2 and Chapter 3 help refine the problem of research data management at research institutions.
The building, intervention, and evaluation (BIE)	The artifacts are developed in a laboratory. They serve as support for reflection and learning through discussion and evaluation with practitioners and scenarios/simulations. Chapter 4, Chapter 6, and Chapter 7 describe artifacts created in laboratory settings (and their evaluation).
Reflection and Learning	In Chapter 8, the results of the BIE stage are further linked to a broader context of open science in research organizations and laboratories
Formalization and Learning	In Chapter 8, the knowledge obtained is presented using theoretical lenses such as socio-technical and experimental systems to place our results into a broader context of (open) experimental work.

1.6.2 Exploratory Case Study

As RDM, the object of our study has not precisely been delineated at the start of the project, the application of interpretive methods helps shape, together with stakeholders, the concepts, environment, and (ultimately) technological interventions that fall under the broader topic of RDM for open science. Hence, the application of interpretive data collection techniques to extract the meaning that stakeholders give to reality is recommended (Boland, 1985; Myers, 1997). Among these techniques used to collect data, semi-structured interviews, and document analysis had our preference. We also included information gathered from conversations with laboratory members. While conversations were not as structured and systematic as interviews and focus groups, they were relevant for identifying additional documents, knowledge, or insights into laboratory practices.

Qualitative approaches, such as interpretive case studies, do not exclude the use of quantitative analyses to adopt another perspective on the phenomenon under investigation (Klein and Myers, 1999; Yin, 2009). Mixed methods allow for integrating qualitative and quantitative

methods of inquiry for analyzing an object of study (Creswell, 2013). The mix between quantitative and qualitative data has been used in several chapters of this dissertation.

1.6.3 Interpretive Approaches

The exploratory nature of the work presented in this dissertation involved extensive qualitative data analysis (Ponelis, 2015). As explained earlier, RDM and open science are nascent phenomena, with multiple scopes and definitions that might hold in one (research) organization and be present in altered forms in another.

Therefore, an important data source is expert interviews. New roles, such as data stewards, are emerging. Also, RDM is a novel for funding agencies and research institutions. To identify a broad set of constructs and processes that might affect our understanding of open science in context. Interview data is analyzed following grounded theory (GT) principles of sampling, coding, and constant comparisons (Wiesche and Yetton, 2017). With GT, we seek to obtain a more cohesive view of RDM in the investigated organizations.

Next, focus groups were our approaches, and results were presented to representatives of data stewardship, data management, scholarly communication, and researchers. With focus groups, we present and discuss artifacts and results with an audience of experts. We used exploratory focus groups to improve the design and suggest new developments (Tremblay et al., 2010).

Finally, forensics approaches were used in a case study laboratory. Digital forensics is the systematic retrieval and analysis of digital material from computer systems (Buchholz and Spafford, 2004). Transposed to laboratories, we applied forensics techniques and designed specialized tooling to interpret digital material stored in the lab. We classify our forensics approach as interpretive since the investigated material is mainly textual and necessitates a treatment similar to interview or focus group data.

1.7 Dissertation Outline

We listed six research questions (RQ) earlier. The questions correspond to research articles published as papers in conference venues or articles in scientific journals. Thus, the research questions (i.e., from RQ1 to RQ6) presented in Section 1.5 are investigated in Chapters 2 to 7 of this dissertation. Besides, the six research questions into three sections. Three sections cover the two fundamental aspects of socio-technical systems, as described earlier. The first two sections cover mostly the social part, where we acquire evidence about how people develop RDM in research institutions. We also investigate the current state of RDM in laboratories. This principle is covered

in Section I and II of this dissertation. Section III offers more space to discuss technology and technology governance in laboratories to monitor developments around reproducible research and RDM.

Chapter 1 is the introduction section. In Chapter 1, the background, research questions, and research methods are presented. To summarize, we are exploring research data management in a laboratory in the Netherlands in the context of research organizations evolving towards open science. Open science is a recent phenomenon that required the application of several types of interpretive approaches as well as the design of artifacts to analyze and intervene in the environment.

Section I "Organization and stakeholders" contains Chapters 2 and 3 and deals with organizational and technological issues among stakeholders involved in research data management. First, in **Chapter 2**, we examine the cooperation between researchers and data managers. By doing so, an agenda for open data in academia is proposed based on qualitative research highlighting issues such as lack of proper infrastructure, accountability, legal frameworks, and rewards in research data management. At the same time, new roles such as data stewards and the struggles with data management support are investigated. In **Chapter 3**, a similar exploratory approach is used to discover how funding agencies and data management support develop a research data strategy in the Netherlands (Lefebvre, Bakhtiari, and Spruit, 2020).

Chapter 2 is published as Lefebvre, A., Schermerhorn, E., & Spruit, M. (2018). How Research Data Management Can Contribute to Efficient and Reliable Science. The 25th European Conference of Information Systems, AIS.

Chapter 3 is published as Lefebvre, A., Bakhtiari, B., & Spruit, M. (2020). Exploring Research Data Management Planning Challenges in Practice. It - Information Technology, 62(1), 29–37. https://doi.org/10.1515/itit-2019-0029

Section II "Laboratory work and experimental systems" contains Chapters 4 and 5 and elaborates on the concept of reproducibility in experimental science. In **Chapter 4**, we dive into data management issues from a technological point of view, showing what types of reproducibility issues occur in storage systems with laboratory forensics techniques described in (Lefebvre and Spruit, 2019b). **Chapter 5** investigates reproducibility in research data management by mapping laboratory work and the scholarly infrastructure to a socio-technical model (Lefebvre and Spruit, 2019a). As such, we obtain a more comprehensive view of reproducibility issues and refine organizational and technical aspects of reproducibility challenges in practice.

Chapter 4 is published as Lefebvre, A., & Spruit, M. (2019). Designing laboratory forensics. In 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019 (Vol. 11701, pp. 238–251). Springer. https://doi.org/10.1007/978-3-030-29374-1_20

Chapter 5 is published as Lefebvre, A., & Spruit, M. (2019). A Socio-Technical Perspective on Reproducibility. 13th Mediterranean Conference on Information Systems, MCIS 2019, AIS.

Section III "FAIR technology" consists of Chapters 6 and 7. **Chapter 6** illustrates the need for designing reproducible and reusable research software with reproducible, research-oriented knowledge discovery in databases process (RRO-KDD) (Lefebvre, Omta and Spruit, 2015). **Chapter 7** presents design principles for open science readiness in laboratories. The goal of open science readiness is to pave the way in the form of a research agenda for open science using new RDM capabilities emerging from laboratory forensics investigations. Moreover, we suggest a way to make forensic results accessible to a broader audience at the management level of laboratories and data stewards to identify weak spots in reproducibility and open scholarship in laboratories through the instantiation of a dashboard.

Chapter 6 is published as Lefebvre, A., Spruit, M., & Omta, W. (2015). Towards reusability of computational experiments Capturing and sharing Research Objects from knowledge discovery processes. Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), 1, 456–462. https://doi.org/10.5220/000563160456042

Chapter 7 is currently under review as Lefebvre, A., Spruit, M (submitted). Laboratory forensics for open science readiness: an investigative approach to research data management.

Finally, in **Chapter 8,** we conclude this dissertation by reflecting and formalizing the knowledge gained through our studies. We, therefore, complete the action design research approach followed throughout this dissertation. Besides, we reflect on the limitations of our approaches, our contribution to (open) science, and provide some insights into future work.

We provide a summary of the approaches we followed in each chapter, the related research question, and the primary source of data In Table 1.4. The first chapters focus on interpretive, exploratory case study work, and the final chapters on design and technology.

Table 1.4

A Summary of the Research Publications Included in This Dissertation

Chapter	Goal	RQ	Method	Data
Chapter 2	Explore research data management in research institutions and identify critical challenges	RQ1	Exploratory case study, Interpretive	Twenty Interviews with data stewards and researchers
Chapter 3	Explore research data management planning with research support and funding agencies.	RQ2	Exploratory case study, Interpretive	Ten Interviews with data stewards and policymakers at funding agencies
Chapter 4	Explore research data preservation and dissemination in practice by reconstructing experiments based on storage systems and publications.	RQ3	Exploratory case study, Design science, Digital forensics	Storage systems and publications
Chapter 5	Conceptualize reproducibility threats and provide a framework to understand reproducibility from a socio-technical lens.	RQ4	Mixed methods	Interviews with researchers, Institutional Survey
Chapter 6	Explore the connection between the needs of bioinformaticians and the constraints of reproducible research. This exploration resulted in a process and tool which link knowledge discovery to research objects design	RQ5	Case study/ Design science research (DSR)	Two Focus groups with biologists and bioinformaticians
Chapter 7	Articulate a set of (emergent) design principles for information systems to achieve open science readiness (OSR) with RDM capabilities.	RQ6	Action Design Research (ADR)	Laboratory forensics

Section 1. Open Science in Research Organizations

Chapter 2 | Opening Science: Challenges of Research Data Management

Research data management (RDM) is an emergent discipline that is increasingly receiving attention from universities, funding agencies, and academic publishers. While data management (DM) benefits from a large corpus of data governance and management frameworks adapted to industry, its academic counterpart RDM still struggles at identifying, organizing, and implementing the main functions of RDM. In this study, we explore the status of research data management at two research organizations in the Netherlands. We identify the leading roles and tasks involved in research data governance, services, and research. We show that, while the application of the DAMA-DMBOK functions and RDM structures are overlapping, RDM is coping with fundamentally different organizational structures and roles than the roles and tasks listed in professional DM frameworks. As RDM is developed to make science more efficient and reliable, it is questionable whether its current structure is adequate. We identified several issues based on our interviews with data managers, researchers, and librarians. For instance, at the moment, researchers are responsible for tasks that depend on DM expertise that they, generally, do not possess. At the same time, research data governance as currently implemented fails to capture the complexity of (professional) data management. Similarly, research data support is not well integrated with the vast diversity of research projects. If not addressed, these issues may impede any progress towards open, efficient, and reliable science.

This work was originally published as:

Lefebvre, A., Schermerhorn, E., & Spruit, M. (2018). How Research Data Management Can Contribute to Efficient and Reliable Science. The 25th European Conference of Information Systems. Portsmouth.

2.1 Introduction

Funding agencies are promoting an "as open as possible, as closed as necessary" (Directorate-general for Research and Innovation, 2016, p. 1) principle for research data availability. It is a matter of guaranteeing trustworthy science and more efficient allocation of resources by, for instance, encouraging scientific data reuse by scientists, businesses, and citizens (European Commission, 2016a). This position shared among significant research funders in Europe are driving more substantial investments in (European) research infrastructures, the integration of data management plans with grant proposals and a broader adoption of open access as an option to publish scholarly output (Paul Ayris et al., 2016; European Commission, 2016b; Pryor, 2012; Wallis et al., 2013). Overall, these research governance efforts shape future directions of data management (DM) for publicly funded research projects.

However, while substantial efforts have been made to deploy Open Science at a European level to make science more efficient and reliable (as explained further in Section 2.2), the essential act of publishing research data still encounters significant resistance from academics (Borgman, 2012; Tsai et al., 2016). The problem of "opening" data, which is seen as beneficial for verification and reuse of existing scientific material (Lefebvre et al., 2015; Mannheimer et al., 2016; Peng, 2011), is a striking example of one of the challenges universities or research institutes are facing when offering data management support to researchers. On the one hand, universities implement such programs utilizing data management plans (DMPs) to secure public funding in the coming years (Simms et al., 2016). On the other hand, researchers are coping with numerous ways of producing and analyzing scientific data, making research data management costly, complex, and diverse (Shoshani and Rotem, 2009).

In previous research on DM in academia, several empirical studies report on results of surveys that show the reluctance of researchers to make data open access and the lack of proper means for preserving data (Fecher et al., 2015; Tenopir et al., 2011) although this type of study is informative to obtain more insights about the stakeholders and problematic aspects of RDM (e.g., storage issues, cumbersome data curation, low rewards for publishing data…), they provide limited knowledge about the actual deployment of RDM programs in research organizations.

Therefore, this study investigates the deployment of existing research data governance and RDM programs in two research institutions (a university: *UNI_CASE* and research and healthcare organization: *HEALTH_CASE*) in the Netherlands using an *exploratory, interpretive* case study approach. This approach is chosen to collect experiences from data managers, librarians,

and researchers using qualitative, semi-structured interviews. We use the DAMA-DMBOK (Mosley et al., 2010) as a reference framework that standardizes best data management practices in the industry as a lecture grid for interpreting research data management activities, roles, and infrastructure in academia.

In short, our approach aims at answering the following research question: *How can research data management contribute to efficient and reliable science?* This question implies to define the current state of research data management and the roles which are taking part in RDM programs. More, as we investigate how RDM can (positively) contribute to Open Science, several aspects impeding the deployment of research data management in research institutions are discussed.

By investigating the "interactions" between RDM governance, RDM services, and researchers as they currently occur in research institutions, we aim to provide insights on the challenges of managing research data in a way that is compliant with Open Science requirements. To achieve that, we first depict the current research data management situation in two research organizations in the Netherlands. Next, we focus on a subset of management and governance activities related to the research data lifecycle: the lifecycle of research data from creation to publication and preservation. Finally, we discuss the most frequent issues identified by combining the data policy screening and the two case studies.

2.2 Theoretical Background

This section provides some background knowledge about data governance, data management, and Open Science (OS). As explained earlier, research organizations in the Netherlands seek to implement new research data governance rules to comply with requests from external stakeholders (typically funding agencies and publishers) for making valuable datasets or software available.

2.2.1 Open Science

Open science has, for the European Commission, two goals in addition to openness, these are (European Commission, 2016a):

- ***Reliable science*** relates to the verification of published results. It encourages adequate data quality checks and, in general, better research governance and scientific integrity for more credible and reproducible science.

- ***Efficient science*** focuses on the resources needed to produce scientific knowledge. Efficient science seeks to reuse existing scientific material, thus limiting resource duplication. Additionally, it encourages the use of (web) standards, versioning of research artifacts, and promotes connected tools and platforms. The European Open Science cloud declaration illustrates the commitment to federating research infrastructures in Europe (Ayris et al., 2016).

Data management also has implications for reliable and efficient science with applications outside academic research such as *data science*. For instance, the EDISON project, which formalized knowledge and skills of data scientists (Manieri et al., 2016), and the BDVA reference model (Zillner et al., 2017) have data management as a core priority. The availability of scientific data to the industry and citizens in Europe places research data management and Open Science in direct connection to open innovation (Chesbrough, 2012). To achieve this, two building blocks of RDM must be considered to maximize the quality of available research data:

Research Data Lifecycles describe the steps research data undergoes before, during, and after the project is completed. There is no unified view on research data lifecycles (RDL) discussed in the literature. For instance, RDLs can be oriented towards data management or on data curation. They can also vary from field to field, e.g., geology or social sciences (Higgins, 2008; Pryor, 2012). Although there is no unique reference RDL, several simple steps can be extracted. These are (1) planning, (2) creation/collection, (3) processing, (4) analysis, (5) publication, (6) archival/dismissal, (7) reuse. The point of view of data curation is that research data might be repurposed for other projects. In that case, published data serves as input for another study, which justifies using a cyclic representation of research data.

Lately, ***Data management plans (DMPs)*** are submitted by researchers as part of a funding agreement with a public funding agency. A DMP outlines RDM-oriented activities during the whole lifecycle of the data, i.e., from research design to archival (Corti et al., 2014; European Commission, 2016b; Simms et al., 2016). Funders are not assessing DMPs as part of a grant proposal, but some of them, like NWO, make DMPs a prerequisite for the actual subsidy of granted projects.

2.2.2 DAMA-DMBOK: Data Governance and Management

The DAMA-DMBOK is an industry-standard for data management created by DAMA international (Otto, 2011). It arranges data management into functions that correspond to groups of activities: planning, control, development, and operational activities (Mosley et al., 2010). For instance, Data governance (DG) is "an organizational approach to data and information

management that formalizes a set of data policies and procedures to encompass the full life cycle of data, from acquisition to use and to disposal" (Korhonen et al., 2013, p. 11). The policies and procedures apply to strategic, tactical, and operational decision-making levels in an organization (Korhonen et al., 2013; Mosley et al., 2010).

DG is the core process of Data Management (DM). DG coordinates nine other DM functions by exercising planning and control activities (Mosley et al., 2010). The functions coordinated by DAMA-DMBOK represent vital areas for managing data. These are (1) data architecture management, (2) data development, (3) database operations management, (4) data security management, (5) reference and master data management, (6) data warehousing and business intelligence management, (7) document and content management, (8) meta-data management, (9) data quality management. The *DAMA-DMBOK* also states that IT professionals and businesspeople (named data stewards) are involved in data management programs, each participating in one or more functions.

Next, DM is defined as "the planning, execution, and oversight of policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data and information assets" (Mosley et al., 2010). It is to note that the ambitions of research data management (RDM) are in-line with the definition of DM as given by the DAMA-DMBOK. In the end, RDM aims at enhancing the value of research data, despite the current struggles to identify quality or relevance metrics which could apply to datasets (Belter, 2014). RDM principles known as the **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR) data (Wilkinson et al., 2016) are guiding RDM infrastructure development and practices to generate reliable, quality research data. The FAIR principles are endorsed by major European and Dutch funders (European Commission, 2016b; NWO, 2017).

Accordingly, the initial case study design started with selecting three DM functions from DAMA-DMBOK, which relates to the main topics found in RDM policies. Then, we performed a matching based on RACI charts to identify relevant activities from RDM and link them to functions from the DAMA-DMBOK framework (see Table 2.1). More details about the matching are given in section 2.2.4. Below we present the three RDM functions which resulted from the screening of Dutch RDM policies.

Research data governance: noticeably, RDM governance policies inspected during this case study are not articulated around functions like DAMA-DMBOK. Instead, practical points such as licensing, informed consent, or funder requirements are addressed. At a glance, these activities should be dealt with by researchers (or higher-level Faculty management), who are responsible for

their execution. In contrast to DAMA-DMBOK, no official equivalent to enterprise-council or data stewardship coordination is found, except in one data policy that is attributing (partial) auditing and policy development responsibilities to a "Research data office," a center of expertise.

Table 2.1.

A Preliminary Matching Between RDM Activities Identified From Policy Screening and Formal Definitions of Data Governance, Data Operations, and Data Development Functions Found in the DAMA-DMBOK Framework.

RDM Functions	DAMA-DMBOK Definitions	RDM Activities
Governance DAMA-DMBOK: Data governance	"The exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets" (Mosley et al., 2010)	Audit, Monitoring, Ethics, Legislation, Funder requirements, Establishing Policies and Procedures, Selection of Standards, Licensing, Informed Consent, Authorizations
Services DAMA-DMBOK: Data operations activities	"Planning, control, and support for structured data assets across the data lifecycle, from creation and acquisition to archival and purge" (Mosley et al., 2010)	Raise awareness, Training, Guidelines, IT support, Data Management Plan support, Ownership support, Data stewardship support, Knowledge network
Researchers DAMA-DMBOK: Data development	"Designing, implementing and maintaining solutions to meet the data needs of the enterprise" (Mosley et al., 2010)	Collection, Creation, Processing, Analysis, Publication, Archival, Reuse, Retention, Documentation, Quality, Storage, Back-up, Versioning

Research data management services regroup (1) IT support, which can be close to the researcher or centralized, and (2) centralized library services giving RDM workshops for researchers, support for writing DMPs, and more. These activities correspond to *data operations*, a function of the DAMA-DMBOK, which concentrates on data handling during the entire data lifecycle.

Researchers are matched to the *data development* function as their activities resemble the most to the development activities of the data development function of DAMA-DMBOK. Although this function is the closest to what researchers do with data, as shown by researchers' software and databases in many domains, it is also the less satisfying matching to the DMBOK of all three categories. It indicates that the closest match for this category has no formal function defined in the DMBOK. Existing DMBOK functions do not sufficiently cover the activities of researchers in a single function. For instance, data development is not well suited to researchers only using software

or platforms. Moreover, it emphasizes data modeling, which is not a typical activity executed by researchers.

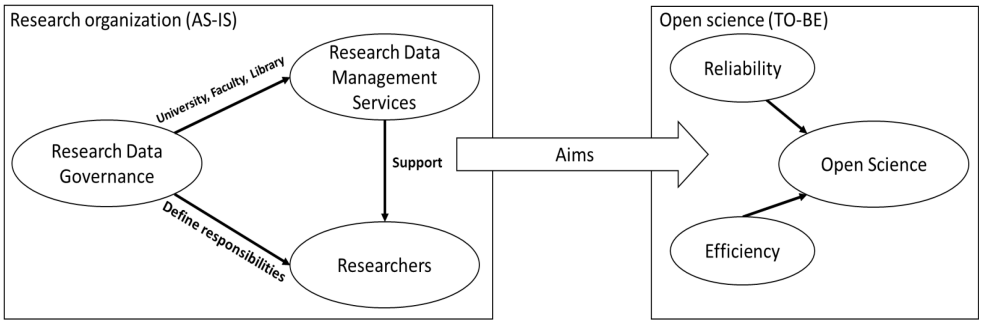
2.2.3 Research Design

We collected evidence from two different data sources, as using multiple sources is recommended by Benbasat et al. (1987) for exploratory case studies where cross-case analysis is performed. Here, cross-case analysis was the option we chose for the reasons stated earlier, i.e., collecting evidence from a diverse organization and a more centralized one. First, we analyzed research data management policies to identify the roles, tasks, and responsibilities that frequently occur in RDM. Next, we conducted 22 interview sessions in one university and its university medical center.

Figure 2.1 shows the relations between the main concepts exposed earlier. Research data governance is developing policies for data management planning and assigns tasks to researchers and data management services. RDM services' tasks are mainly related to supporting. Researchers retain most of the responsibilities of managing research data. RDM is structured as such to help the organization satisfy the goals of Open Science. These are the Reliability and Efficiency of science. These two goals are implemented in the Netherlands in research organizations (such as UNI_CASE and HEALTH_CASE).

Figure 2.1.

Initial RDM Functions Retained for Devising an Interview Protocol.



2.2.4 Governance: Policy Review

To identify several functional areas of the DMBOK, which might be of interest for exploring RDM, Dutch RDM policies were screened before the case studies. The sample of policies consisted of documents from all universities in the Netherlands, including *UNI_CASE* and

HEALTH_CASE. The policy review captures the roles and tasks of data management in research institutions as validated by each organization's administration board.

In total, 13 Dutch RDM policies were scanned using a RACI chart, a type of *Responsibility Assignment Matrix* (Wende and Otto, 2007). As Wende & Otto (2007) explained, RACI stands for **R**esponsible, **A**ccountable, **C**onsulted, and **I**nformed. The matrix's rows are tasks related to data management, and the columns are the roles. The values (R/A/C or I) indicate which type of responsibility a given role received for a task. A summary of the analysis is shown in section 2.4.1.

- Responsible is unique per row in the matrix as it refers to the person who performs the task
- Accountable refers to the role which has ultimate decision power over a task
- Consulted is a role providing input before completion of the task, consult is an optional role
- Informed is simply a role notified of the completion of a task, is optional

The advantage of a RACI chart for policy review is that their interpretation is straightforward. The comparison of policies from different organizations makes inconsistencies in the attribution of responsibility very clear. A disadvantage of this technique, as we experienced, is that the matching between matrices containing unstandardized or equivocal roles and tasks is a cumbersome process for which the correctness is hard to evaluate.

2.3 Exploratory Case Studies

Next, we collected qualitative data employing interviews with practitioners during two exploratory, interpretive, case studies (Benbasat et al., 1987; Klein and Myers, 1999). The first case study (*UNI_CASE)* is a research organization in the Netherlands. The second case is a healthcare and research organization (*HEALTH_CASE*) tightly bound to *UNI_CASE* but having a patient care mission missing from *UNI_CASE*. A subset of the interviewees (total n=23) was recruited during a network meeting, attended by research data managers, held at *UNI_CASE in April 2017*. This networking event attracted participants from *RDM Services* and *RDM Research*. We had no interviews with members of the university's executive board (or faculties) who are formally responsible for developing policies and authorizing their application at an organization-wide level. Nevertheless, people involved in the creation of governance frameworks were interviewed in both organizations.

2.3.2 Sites Selection

Two main strategies are guiding the selection of case study sites. The first one is opportunistic. We had access to a network of data managers working in both organizations, which facilitated interviewees' recruitment. The second strategy was theoretically grounded. UNI_CASE and HEALTH_CASE differ mainly due to the sensitivity of the data analyzed for research purposes. HEALTH_CASE started an RDM program earlier than UNI_CASE for that reason, and it has a stronger centralization of RDM solutions than UNI_CASE, which is a university with very diverse research faculties. They are not strictly speaking of two independent organizations. *HEALTH_CASE* is the *Faculty of medical sciences* of the university *UNI_CASE*. The separation between the faculty of medical sciences and the other faculties is common to all universities in the Netherlands. The two aspects of RDM in context: diversity for UNI_CASE and centralization and sensitivity for HEALTH_CASE.

2.3.3 Interview Protocol

An initial group of interviewees was contacted during a data management network meeting at UNI_CASE. The other interviewees were recommended at the end of the first interviews for their expertise in one or more RDM activities. We designed the interview protocol based on the retained DAMA-DMBOK functions (data governance, development, and operations), the data stewardship lifecycle (RDL), and tasks found in research data governance policies. Also, we retained two aspects of data management to discuss with the participants. The first aspect relates to the three DAMA-DMBOK functions. The interviewees present how data is governed and supported in their organization/division. At the start of each interview, participants were invited to describe their job (or role) to facilitate matching their profile to two of the three categories retained for this study (i.e., service or research). The second aspect is focused on the connection between their tasks and typical stages of the research data lifecycle. Each interviewee had to discuss some of the tasks they perform at each stage of the RDL presented to them as a printed picture.

2.3.4 Data Collection

We conducted the interviews between May 2017 and June 2017. In total, we held 22 interview sessions. The interviews were recorded and transcribed for further analysis in *NVivo 11.4,* a qualitative data analysis software. Twenty-one sessions took place in Dutch, one interview in

English. Table 2.2 shows the role and domain of each participant. The roles refer to the job titles of the participants.

Table 2.2

Background of Interviewees and Mapping to RDM Functions

UN	Role	Domain	RDM	HE	Role	Domain	RDM
U1	1 Consultant	Library services	Services	H1	1 Researcher	Primary Care	Research
U2	1 Information Manager	Faculty Level IT	Services	H2	1 Product Owner 1 IT Architect	Central IT	Services
U3	1 Coordinator	Psychology	Services	H3	2 Data managers	Brain division	Services
U4	1 ICT Manager	Epidemiology	Services	H4	1 Researcher	Neuropsychiatry	Research
U5	1 Librarian	Library services	Services	H5	1 Manager/ Researcher	Infrastructure	Services/ Research
U6	1 Librarian	Library services	Services	H6	1 Researcher	Child division	Research
U7	1 ICT Manager	Epidemiology	Services	H7	2 Data managers	Vital functions	Services
U8	1 ICT Manager	Central IT	Services	H8	1 Data manager	Brain division	Services
U9	1 Information Manager	Geosciences	Services	H9	1 Researcher	Cell screening	Research
U10	1 Researcher	Finance	Research				
U11	1 Consultant	Library services	Services				
U12	1 Researcher	Epidemiology	Research				

2.4 Data Analysis and Interpretation

In this section, we describe the two types of data analyses that were executed. First, the policy screening (see section 2.4.1) explains how policy documents were mapped to RACI charts and how a meta-analysis of 8 Dutch RDM policies was made.

2.4.1 Research Data Policy Screening

The goal of the data policy screening is to collect roles and tasks as they are implemented in several research institutions. The present analysis gives a limited overview of the content of RDM policies. In Table 2.3, a value (R or A, as C and I hardly appear) is added when roles and their

task(s) are mentioned in the documents. Also, a minimum frequency of roles and tasks is used to filter out roles/tasks that differ too much. They had to appear at least two documents to be added to this table. Some of the roles that are not passing the threshold are *data protection officer, data steward,* meaning that only one policy (out of 8) formalized responsibilities about at least one of these two roles.

The small number of Dutch policies we screened has several reasons. First reason, the absence of a data governance policy at 2 Dutch research organizations. Besides, three other documents were guidelines, so not formal policies attributing tasks to defined roles. These three documents could not be analyzed with RACI charts, as only explicit role assignments were considered for screening. Finally, other medical centers in the Netherlands have a *research code of conduct* but no RDM policy could be retrieved online.

2.4.2 Case Studies

The interviews were recorded and transcribed into text files for further analysis in *NVivo*. With *N*Vivo, the interviews were classified per site (UNI_CASE and HEALTH_CASE). A first coding lead to a mapping of statements to predefined categories (e.g., data lifecycle, reuse, general remarks, etc.) using 75 codes for UU_CASE and 67 nodes for HEALTH_CASE. Next, a more open coding process, as defined by Heath & Cowley (2004), was conducted to search for more fine-grained information in the interviews. At a statement level, 355 nodes referred to quotes from interviewees from both organizations. The statement level coding was used to annotate keywords (e.g., agreement, awareness, data quality) and context (e.g., no formal training, data management costs etc.). In the next sections, we substantiate the different tasks with experiences from data managers, researchers, and librarians. All these categories are derived from the research data policy screening. They are presented in the order they appear in Table 2.3. We start with the creation of a data management plan (in section 2.5).

Table 2.3.

The distributions of roles found per task, based on 8 RACI charts. The number between () indicates how many times this relation has been encountered among the eight policies analyzed (status of February 2017).

Task\Role	Function	Researcher	P.I.	Faculty	Executive Board	Library
Create Data Management Plan	Research	R (5)	R (2)	-	-	-
Handle Data Lifecycle (Stewardship)	Research	R (5), A (1)	-	-	-	-
Support IT Infrastructure	Service	-	-	R (2)	R (2)	-
Support Training and Data Handling	Service	-	-	R (1)	R (1)	R (1)
Monitor and Audit	Governance	-	-	R (2)	R (2)	-
Develop and Implement policy	Governance	-	-	R (2)	R (2)	-

2.5 Create a Data Management Plan

Researchers are responsible for creating a data management plan (DMP). Results from the policy screening and interviewees indicate that this rule also applies to UNI_CASE. The separation between RDM services and researchers appears clearly for data management planning. At UNI_CASE, RDM services are available for help but researchers are still fully responsible for writing it. As explained by an RDM consultant: "We offer a review service. We do not write the DMP for the researcher except if it is a big project, then they can hire us. We are hired temporarily to make the DMP, but it is not the goal that everybody does so…" **[U1]**. Another consultant adds: "We provide a DMP check but they [researchers] need to fill it in themselves. We do not have any mandate to validate something. They remain responsible for their DMP and its good quality" **[U11].**

Located closer to researchers, an information manager explained that support for DMP is rerouted to the library: "Now we have to submit a DMP with grant proposals. In that case, they [researchers] know where to find us and we know where to find the library" **[U9]**. Moreover, none of the interviewees from *UNI_CASE* could remember a DMP that would have been written without the request of a funder though *UNI_CASE* has officially enforced the use of DMPs since January 2016 for all new scientific projects.

In HEALTH_CASE, however, the involvement of data managers in RDM planning seems stronger. Two data managers explained: "We support researchers, and we are involved in the creation of the DMP, […]. We help with the data collection tools, the software with which data is collected. This is done in collaboration with the researchers, for surveys and the data from electronic health records" **[H3]**. A manager of a computing facility for bioinformaticians confirms the closer involvement with researchers in their organization: "Fortunately, researchers come more often to us. Nevertheless, the problems arrive, through the principal investigator (PI), to the IT department if it is about storage or something. Fortunately, we have a direct line with IT, so we know where to redirect them to an information point where they know more about it. We called the person who is now busy with audits and data managements plans a data steward, this person is our information point now" **[H5].**

2.6 Handle Data Lifecycle

A DMP describes how data will be handled during and after a research project. Handling research data is the responsibility of researchers (or their principal investigators). Again, some significant differences occur in the implementation of this rule in the two organizations.

First, data collection and data creation are separated activities. **Data collection** refers to data gathered from already existing data sources and **data creation** occurs when researchers use data that did not exist before. **Data collection** is frequent at *HEALTH_CASE,* where operational data from the medical center serves as research data for researchers. The connection between patient data and research data is made by data managers and a centralized research data platform (i.e., a data warehouse). Data managers perform quality checks on the data as it is filled in by multiple people (e.g., nurses, doctors) and are not meant for research but care in the first place. Then, upon request from a researcher, an anonymized subset of the data is transferred. Two data managers state that the fact that operational medical data is used for research purposes have implications on data governance: "Data collection is really important and often different than a university because in a university you start with a study, you select participants and they all consent to the study and they participate. In a medical center, there is already a lot of data that you want to use for your research. This raises other questions regarding governance…" **[H7]**. **[H5]** and **[H6]** add that data can also be created by a *wet lab person* manipulating instruments. After, this data can be integrated with data from health records stored in the *research data platform*.

At *UNI_CASE,* **[U3]** sometimes helps researchers with extracting data from databases and performs some data cleaning. In other cases, data might be collected directly via services from

financial companies that have agreements with the university **[U10]**. Another scenario is that external providers generate data, as it is the case for **[U7]**, a data manager for a longitudinal study with a large cohort: "there are always things that are unclear, and which are not known by the agency [which manages surveys], so we ask them to put memo's or codes if they do not know. And then it starts with the retrieval of memos and codes; we have developed all kinds of rules around that because the study started a long time ago. […] we also have a small data cleaning part". On the sides of librarians **[U5]** and **[U11]** are involved in data collection when it concerns the reuse of published datasets exclusively.

Next, as two data managers from the brain division state: "Data processing, analysis, and publication are done by the researchers themselves" **[H3].** This statement summarizes the roles involved in the **processing and analysis** tasks. Nevertheless, in the case of consortia, tasks are distributed across organizations and processing might be done at another organization, as explained by one researcher in Neuropsychiatry **[H4]**. If third parties do the analysis, researchers might have some difficulties to understand how the data were analyzed: "If we did a part of the processing of data analysis for them [researchers], they come back to us asking what [analyses] you did again?" **[H5].** In **[U12]**'s team (epidemiology), data managers are processing and documenting the data and their role are perceived as really important.

Following the analysis step, **archival** was an issue at UNI_CASE and HEALTH_CASE. There are diverging opinions about what should be done. At HEALTH_CASE, there are rules for preserving raw data for five years and data for verification for 15 years. The difference in retention time is due to the high costs 15-year preservation of raw material would entail **[H5]**. An additional hindrance is the absence of strict rules or guidelines for archival **[H7]** and researchers must decide on what they archive and what not. At UNI_CASE, data managers and researchers are experiencing many similar troubles. The general rule in the Netherlands is that underlying data must be preserved for ten years after publication. For **[U1], [U8]** and **[U11],** nobody else than researchers are solely responsible for deciding about what data to dismiss after a project ended.

Finally, **reuse and publication** suffer from a similar issue in both organizations. The Faculty of Geosciences of UNI_CASE, for instance, shares and reuses data created by others as the analysis methods in their field rely on measurements from different locations on earth **[U9]**. Then, archival techniques impede reusability, not all raw data can be stored, and the archive consists of a processed bulk **[U2]**. According to **[H4]**, even internal reuse is challenging and requires contacting different people in their department to obtain information about where to find relevant datasets. None of the interviewees has put data open at the time of the interviews. In the case of long-term,

longitudinal studies, the cause was informed consents: "In the case of [longitudinal] cohorts, it does not happen. The informed consent signed 20 years ago stated that the data would not be made openly available. So, we are not authorized to do so. Sometimes we must explain that to journals too; they want us to make the data open access until now we were successful. It is not possible. The data is available upon request to individual researchers." **[U12].** At *HEALTH*_CASE, a data manager affirms: "I do not know how it works with open data and patient information […] I do not know what the rules are; therefore we are not doing it" **[H7].** In short, archival issues, privacy, longitudinal studies with no informed consent for open data, limited space in repositories, commercial agreements, unclear regulations and reluctance from researchers are the causes listed by the interviewees not to be able to make data open by default.

2.7 Support IT Infrastructure, Training, and Data Handling

Two initiatives from the *central IT departments* exist to deploy new IT infrastructures for research. In HEALTH_CASE, it is mainly a data warehouse aggregating information from different internal sources. It was developed to provide researchers with integrated, anonymized operational data. Researchers have to fill in a request form where they describe their research questions before receiving the information, they need from data managers. As such, data managers at HEALTH_CASE are there to protect sensitive information hosted in the medical center.

In UNI_CASE, technologies vary per project, disciplines, and faculties. The *Central IT department* is developing a storage environment in collaboration with **[U3]**'s team. At the same time, **[U2]** explained that storage is provided as a paid service for departments inside the faculty but that they are not (yet) competitive against external storage solutions. Meanwhile, **[U11]** stated that library services do not offer any storage too; they only focus on support. As a consultant explained: "RDM support is a combination of a university-wide program for IT and research, the central IT department, the T&T (teaching and research) department, legal affairs, privacy and other departments. It is not a formal organization but a real collaboration inside UNI_CASE" **[U1]**. From the perception of researchers, this scaffold around RDM appears far away from their more urgent issues. One researcher from HEALTH_CASE: "I had no contact with them [RDM support]. Who should be the contact point… right now, everything is really centralized but they are not the people that can help me if I have questions about protocols or data management plans? They are too far away and too generic […]" **[H1].**

Nonetheless, training for managing data is perceived as a benefit for researchers: "I do think that at the start of the career you should receive a workshop or some education on how to

properly do this, this is important. A lot of mistakes have been made at the beginning which I wouldn't do now…" [**U11**]. However, there are no DM training or workshops specialized per discipline of type of analyses in both organizations. The existing workshops are said to be generic and are directed at raising awareness of researchers about RDM topics; they lack some depth as argued by **[H1]** and **[H9]**.

2.8 Monitoring and Developing Policies and Guidelines

Monitoring and audit of research data occurred in one case: when patient data is involved. There is no active monitoring or auditing on how research data management is deployed in the two organizations by the roles that are responsible for these tasks (Faculty board and Executive boards). **[U11]** said it would take time before monitoring is going to happen, the reason given is that the implementation of RDM goes slowly and more time is needed to put these control mechanisms in place.

2.9 Results

In this section, we comment on the results of the policy screening and the interviews. Furthermore, some limitations of our approach are addressed in section 2.10. We also relate our findings to the IS roadmap for open data implementation (Link et al., 2017) to further guide the development of the roadmap.

2.9.1 RDM Policies

What appears from the screening outcomes, summarized in Table 2.3, and are that these documents do not clearly state any accountability, consulted or informed roles (apart from one policy in which researchers are accountable for data handling). Hence, it is not clear how tasks are divided among RDM stakeholders. They differ per project and are not directly linkable to roles present in the DAMA-DMBOK, an industry-standard for data management.

Besides, there are several layers of data policies that are to be developed: deans and executive boards are both responsible for developing data management. They differ in scope, though. While the executive board develops a central (university-wide) policy, refinements are delegated to faculties which are assumed to be able to define more precise and more explicit regulations covering characteristics of their research data. It is to note that there is, to a certain extent, a discrepancy between the tasks related to researchers and those involving executive boards (faculties and university). It can be seen from Table 2.3 that there is an agreement to attribute DMP

and data handling tasks to researchers (5 out of 8) but few data policies explicitly mentioned infrastructure support, training and further policy developments for which libraries or executive boards from the university or faculties are responsible.

Further, Table 2.3 shows that there are no specific RDM roles defined. Unlike the DAMA-DMBOK which provides a set of 32 roles relating to DM responsibilities at enterprise and business unit levels, RDM reshuffles well-known academic positions (e.g., researchers, principal investigator (PI), deans), and appears to not regulate additional roles such as *data manager* or *data steward*, even if these roles actively collaborate with researchers and interact with research data during its lifecycle.

Finally, no monitoring and audit procedures are present, while *data management* (as defined in section 2.2.2) has a significant monitoring component according to industry standards. The monitoring task, as shown in Table 2.3, refers to *responsibilities* to monitor policy development and data handling, not on formal procedures or metrics achieving this. From the case studies, we can also confirm that there is no active monitoring in place, except for strictly legal reasons (i.e., privacy): when patient data is involved.

2.9.2 Research Data at the Medical Centre and the University

We have seen that, for both organizations, a division of RDM in three functions: governance, services, and research, suffices to categorize most of the roles involved. Nevertheless, it is also noticeable that RDM services and RDM research can be further refined. For RDM services, two profiles emerged. The first, the *governance supporter,* profile, support researchers from an *Open Science* perspective. Their activities are constrained to *post*-analysis data curation, data management planning support and raising awareness about open data. Another profile, the *research supporter*, belonging to RDM services, is closer to operational activities that researchers are conducting. The former profile is related to *governance* support more than support for researchers. The latter coincides more with what can be expected from *research* support and encompasses *data managers* and *data stewards*.

The two goals of Open Science, efficiency, and reliability are not the most prominent drivers for *research supporters* at RDM services in *HEALTH_CASE* and *UNI_CASE*. As said earlier, data security is the primary interest for the *research data platform* at HEALTH_CASE, and the federated storage of UNI_CASE serves archival and encryption needs of a longitudinal study. The same can be said about *data management plans*. Data management planning is done when required by funders, and the rule of UNI_CASE to enforce DMP for each new project had no

concrete effect. This might indicate two things: researchers do not see the benefit of data management planning (if no funding depends on it) and central data policies from UNI_CASE have a weak impact on changing the behavior of researchers regarding data planning.

Governance is structured top-down in both organizations. UNI_CASE initiated a central policy that needs to be refined by each faculty, which is still ongoing work at this time. However, there is no evidence that faculties are the most optimal decision layer when it comes to managing research data. For instance, the institute where [U7] is located has no contact with the faculty as their presence in that faculty is purely administrative, the type of research and data differs from the rest of the faculty. Other interviewees agreed that the type of research (e.g., quantitative, qualitative) and type of data (e.g., commercial, medical, experimental, simulated) are significantly more impacting the services needed to plan and handle this data accordingly. Hence, faculty boards might not be a suitable basis for further elaborating on responsibilities and tasks as those tend to share more commonalities with equivalent analysis and data than other departments belonging to the same faculty or institute.

2.10 Discussion and Limitations

There are several limitations to this study. First, the data collection is limited in scope. Indeed, the preliminary RDM division in functions and roles (see Table 2.3) is established based on a limited number of data policies and two case studies in the Netherlands. Identifying to which extent similar RDM implementations exist in research institutions requires further research. Nevertheless, other countries where RDM has gained some maturity appear to use a similar division between functions. In the UK, 79% of the institutional policies “mentioned and specified” a role for RDM support at an institutional level, but only 37% of the 57 policies define explicit control (i.e., review) and responsibilities (Horton and DCC, 2016). Although control is an activity group of data governance according to DAMA-DMBOK, it seems that it is not systematically enforced in policies from research institutions, neither in the UK nor in the Netherlands.

When the scope is broadened to North American research institutions, a study from Tenopir *et al.*, (2014) highlighted that library services are indeed not intervening at a technical level during data analysis and provide (or are planning to provide) support for curation and data management plans. These activities are similar to our findings, where RDM services offered by librarians is mostly restrained to consultancy. We can add that consultancy is insufficiently regulated by RDM Governance, which tends to assign most of the responsibilities to researchers or academic management staff without clear rules for opening data. This absence of guidelines has an

impact on research supporters as well. This makes research supporters (e.g., data stewards, data managers) unsure about how opening data should be done, legally and technically. These findings lead us to contextualize the issues of opening research data into a broader agenda. At a higher-level, Ponte (2015) indicated that technical quality and sustainability of the open data are issues threatening the open data ecosystem. While the research institution we investigated work on increasing data quality (open or not), it is questionable whether the whole enterprise is sustainable if tasks and responsibilities remain vague or if control is not exercised thoroughly.

In addition to data quality and sustainability issues, infrastructure and privacy were two other aspects that are perceived as being complex matters by the interviewees. The motivations for open data implementation introduced by Link et al. (2017) echo the goals of Open Science and RDM functions presented earlier. This allows us to further examine the open data implementation agenda under the light of our findings. The authors classified open data implementation into two dimensions: motivation and implementation. The four motivations we elaborate on here are "mandated sharing," "benefits to the research process," and "extending the life of research data" (Link et al., 2017). We do not discuss "career impact" as the focus of our case study results is more in-line with the three other motivations:

- **Mandated sharing**: Funders, publishers, or universities encourage open data.
- **Benefits to the research process**: reliability of science by facilitating the reproduction of results.
- **Extending the life of research data** refers to the notion of efficiency of science introduced earlier.

Table 2.4 shows several issues identified during our case studies and how they relate to the roadmap of open data implementation (Link et al., 2017). This way, real practical issues that are in-line with the IS roadmap are given to advance the discussions further. These are situations that are to be expected and which negatively impact the to-be situation of Open Science. These findings form discussion points to nurture IS research to become a significant role player in open data implementation and RDM to contribute to Open Science.

Table 2.4.

Classification of our findings according to the IS roadmap for open data (Link et al., 2017).

Motivation		Mandated Sharing	Research Process *(Reliability)*	Life of Research Data (*Efficiency)*
Implementation	Governance	No DMP filled in if not requested by funders	Institutional policies to be overridden by faculties and departments, no standardization of custom policies	Questions arise with data collected from companies (copyrights) or with (old) informed consents (privacy), no clear accountability.
	Socio-technical System	Centralized solutions are developed when data is considered sensitive, open data itself is not sufficient	Institutional storage not competitive with cloud storage, but cloud storage is not used if data is sensitive	No clear rules for archiving research data
	Standards	No standard way of handling data during the lifecycle	Operational data or outsourced data collection might not be standardized	Some “standards” are tailor-made for a project; they are not cross-discipline
	Data Quality	Data might be collected from different sources where researchers have less control over its quality	Freedom when it comes to archive data, researchers are solely responsible	Internal reuse difficult, not immediate insights on how data was generated
	Ethics	Privacy is a reason not to share	Lack of guidelines or infrastructure for opening sensitive data	There are cases were operational medical data is used by researchers which influence the possibility to publish it

2.11 Conclusion and Further Research

This paper investigated how research data management can contribute to Open Science. Open science has two goals: reliability and efficiency of science. Both goals rely on RDM to be attained. At the same time, RDM is lagging industry standards on several aspects, mainly data governance, which stays vague on data management planning and control. For that reason, the open by default strategy is not applied due to regulatory and operational issues that, if adequately

addressed, could ease the data publication process. **The efficiency of science** through better regulated RDM can be achieved by involving *research supporters* and make their responsibilities clearer in data policies, in which they are currently not well represented. **Reliable science** is challenged by other operational issues such as data archival and privacy. *Governance supporters* are more perceived, on the researchers' side, as Open Science champions without the proper infrastructure to make RDM work. Their responsibilities and tasks must be formalized, at least as *consulted* or *informed* roles to foster data publication.

Further research should collect more evidence from other research institutions worldwide following the policy screening and exploratory case study approach. More roles, tasks, and functions can be discovered and refine the three main RDM functions found in the two organizations. Eventually, research data management can benefit from guidelines and assessment instruments grounded in evidence obtained from larger-scale policy screenings and the RDM in diverse research institutions.

Chapter 3 | Research Data Management Planning in Practice

Research data management planning (RDMP) is the process through which researchers first get acquainted with research data management (RDM) matters. In recent years, public funding agencies have implemented governmental policies for removing barriers to access to scientific information. Researchers applying for funding at public funding agencies need to define a strategy for guaranteeing that the acquired funds also yield high-quality and reusable research data. To achieve that, funding bodies ask researchers to elaborate on data management needs in documents called data management plans (DMP). In this study, we explore several organizational and technological challenges occurring during the planning phase of research data management, more precisely during the grant submission process. By doing so, we deepen our understanding of a crucial process within research data management and broaden our understanding of the current stakeholders, practices, and challenges in RDMP.

This work was originally published as:

Lefebvre, A., Bakhtiari, B., & Spruit, M. (2020). Exploring research data management planning challenges in practice. It - Information Technology, 62(1). https://doi.org/10.1515/itit-2019-0029

3.1 Introduction

Public funding agencies and research institutions are facing novel challenges related to the management of research outputs produced in Academia. In recent years, governing bodies across the world have started to promote new open science policies and practices for managing scientific information. Unlike open access policies that exclusively promote the public availability of scientific articles (Chan et al., 2002), open science (OS) policies expand open access to a more extensive set of research outputs into consideration. Accordingly, OS includes scientific publications and their corresponding artifacts, such as software, data, and sample material, into the scope of scientific information (Ardestani et al., 2015; European Commission, 2016a). As a result, research data management (RDM) is introduced by funding agencies as a critical capability of research institutions that benefit from training programs, dedicated IT services, and new roles in research organizations(Lefebvre et al., 2018).

Nevertheless, previous research on research data management planning (RDMP) practices has reported challenges related to a lack of knowledge about the usefulness and best practices of RDMP. So far, other studies have shown that (1) funder policies for RDMP are quite general (Dietrich et al., 2012), (2) researchers are often reluctant to disseminate curated data (Wilms, Stieglitz, et al., 2018), and (3) there is a lack of knowledge and detailed guidelines to support RDMP in research institutions (Lefebvre et al., 2018). Moreover, in the United States (US), (M. Williams et al., 2017) showed that requirements from US funding agencies are inconsistent and emphasize post-publication data management rather than foster an up-stream data strategy, which could guarantee more robust data management from the start of a research project. Likewise, Science Europe, a European association of national public funding agencies, acknowledged the complexity of current policies and recently proposed standardized RDMP guidelines (Science Europe, 2018).

In this work, we investigate research data management planning (RDMP) as a function of research data management (RDM). RDMP is an essential part of the grant application process. In Europe, for instance, the European Commission has (partially) incorporated data management and planning in its grant application procedures for the Horizon 2020 (H2020) funding program (European Commission, 2016a). In Europe, for instance, the European Commission has (partially) incorporated data management and planning in its grant application procedures for the Horizon 2020 (H2020) funding program (European Commission, 2016a). As part of the application procedure of H2020, applicants submit data management plans (DMP). A DMP is an additional document in which grant applicants outline how data is acquired for their research project, which

technology and standards they intend to use (e.g., storage, back-up, software), how data will be preserved and, possibly, shared and, the costs induced by additional resources and services needed to manage data (Michener, 2015).

Therefore, we seek to investigate current RDMP practices in academia from two perspectives: a funder perspective and a research data service perspective. By doing so, we aim at shedding light upon existing practices and challenges in RDMP. Besides, we suggest potential solutions where information technology can play a crucial role in improving the review of RDMP deliverables. The following research question drives this study: ***What are the current challenges and practices in research data management planning?*** To answer this question, we follow a case study approach and collect experiences from representatives of public funding agencies, grant reviewers, and data stewards. Second, as part of the case study, we analyzed 98 data management sections in (draft) research proposals of projects submitted to NWO, the Dutch national science foundation. The goal of the analysis was to investigate whether current data management paragraphs reflect the ambition of producing reusable research data.

3.2 Background and Related Work

From the start of the digital revolution, data management has played a crucial role in organizations. At that time, managers became aware of the potential business value of data stored in companies. However, it appeared that data was also not fully integrated across information systems (IS) (Goodhue et al., 1988). Strategic data planning (SDP) was established as a response to the absence of integration of information systems in firms, which is one of the earlier attempts to plan information architectures integrating data sources in organizations (Goodhue et al., 1988; Shanks, 1997).

The ambition to align data systems with corporate needs is still actively pursued in new data management practices. Accordingly, business and industry environments view data planning as a strategic process. Thus, the business value of data depends upon the capacity to integrate corporate information systems. Research data management planning (RDMP) share similar ambitions with data planning as RDMP aligns theproduction and use of research data with an open science strategy. However, data management in academia has not been as thoroughly investigated as data management in businesses (Corti et al., 2014; Lefebvre et al., 2018).

On the one hand, research data management (RDM) refers to the management of research data during the lifecycle of a research project, from data creation to dissemination. Funders and research institutions' policies state that RDM is the responsibility of researchers. Researchers are

supported by research data management services (RDS). RDS consists of library services specializing in data management matters and metadata curation, IT services for hardware and software and, data stewards as an emerging role in academia, which assist researchers and research groups in managing data.

On the other hand, research data management planning (RDMP) is defined here as the process of planning costs, (storage) technology, formats, documentation, legal matters and openness of data to effectively manage research data during and after (publicly funded) research projects (Dietrich et al., 2012; European Commission, 2016a). Often, the main deliverable of research data planning is a data management plan (DMP) (Michener, 2015). Additionally, data stewards are supporting scientists in establishing a data management plan and curate data (Hartter et al., 2013). The reason why data stewards hold this supporting role is that data management planning is a new practice for scientists who might ignore fundamental data management concepts (Lefebvre et al., 2018). For instance, funding agencies in the Netherlands rely upon a set of guidelines focused on extending the life cycle of research data and producing reusable data. These guidelines are based on FAIR principles, which are a set of guidelines used to create Findable, Accessible, Interoperable, and Reusable research data (Wilkinson et al., 2016).

DMP requirements may vary for each (public) funding agency, as shown in Table 3.1. For instance, depending on the funder, grant applicants need to fill in a data management section in the proposal, and submit an additional DMP before the grant is disbursed (e.g., NWO) or after (e.g., H2020). For the Dutch Research Council (NWO), the first RDMP deliverable is a section called data management, which includes many questions about reusable data. The analysis of data management sections we present in Section 3.4.5 uses those data management sections as input. Then, applicants submit a full DMP. A DMP answers a more comprehensive set of questions about RDM matters for the funded project.

For the European Commission (H2020), the paragraph is labeled “as a data approach.” The data approach statement is a sub-section of the grant proposal. Additionally, both H2020 and NWO make use of DMPs for improving data management.

Table 3.1.

Comparison of RDM Processes of National Funding Agencies

Agency	ST	DMP	Process	Deliverable(s)
NOW	NL	Yes	Data management section in the application then Data management plan four months after the project is granted	Data management section, and the structured data management plan
ZONMW	NL	Yes	DMP after the project has been granted	The structured data management plan
ESRC	UK	Yes	DMP as part of the initial application	The structured data management plan
AHRC	UK	Yes	DMP as part of the initial application	The structured data management plan
NSF	US	Yes	DMP as part of the initial application	The data management plan of a maximum of two pages, directorate-specific

3.2.1 Evaluation of Research Data Planning

Earlier, we touched upon strategic data planning (SDP) to show how business and industry have sought to increase the quality of (strategic) data and extract value from their data by integrating information systems. We showed that a similar planning approach is recently used in academia to produce reusable scientific data. A significant difference between strategic planning (SDP) and RDMP, though, is that SDP programs are deployed inside the boundaries of firms, while RDM is often an inter-organizational effort between funders and research institutions. Hence, it is quite complex to define generic criteria and rules for evaluating the quality of research data management due to the variety of data produced in science.

In RDM, quality criteria are mainly relying on FAIR principles. Recently, Science Europe standardized the areas of data management that data management plans should address. According to Science Europe, DMPs ideally cover the following aspects of data management: Data description and collection or re-use of existing data, documentation and data quality, storage and backup during the research process, legal and ethical requirements, codes of conduct, data sharing and long-term preservation and, data management responsibilities and resources (Science Europe, 2018).

3.3 Method

We conducted an exploratory case study in the Netherlands. We collected data using ten semi-structured interviews with representatives of funding agencies and research data management services in the Netherlands (national level) and one interview with the European Commission, which funds research at the European level through ERC and H2020 funding programs. The list of interviewees is shown in Table 3.2. Interviews lasted approximately 45 minutes and were either conducted face to face or remotely on Skype. All interviews were recorded after obtaining the approval from the interviewees, who signed an informed consent form.

Table 3.2.

Organization, Experience, and Role of the Ten Interviewees. Identifiers (ID) are Used in the Results Sections to Refer to the Interviewees.

ID	Organization	Experience	Role
F1	Government Agency	4 years	Grant support
F2	Dutch Funder	7-8 years	Grant evaluation
F3	Dutch Funder	8 months	Policymaking
F4	Government Agency	7-8 years	Policymaking
F5	European Funder	4 years	Grant evaluation
R1	University/Medical Centre	2 years	Research support
R2	University/Medical Centre	3 months	Research support
R3	University/Medical Centre	10 years	Research support
R4	University/Medical Centre	5 years	Research support
R5	University/Medical Centre	4 years	Research support

The interview protocol contains seven items. The first two items focus on the grant application process and aim at comparing different review processes per agency. The next three items question the criteria for judging the quality of data management planning. These questions help to determine what criteria are used for evaluating data management plans. Finally, the last three items falling under the challenges of data management planning ask the interview participants to describe technical and organizational challenges that occur in practice. The interview protocol was structured, as shown in Table 3.3.

The ten interviews were recorded, transcribed, and anonymized. The transcribed text version was subsequently analyzed with support from NVivo software (version 12.5). Interview

transcripts were classified by type of organization. There are two organizations in the sample: funding agencies (F code) and research data management services (R code). In total, we coded the transcripts with 299 nodes in NVivo to structure the interview data from the (English) transcripts. Examples of such nodes are tools (such as DMPOnline), feedback on DMP, legal issues, and metadata. Next, we categorized those nodes into overarching groups of concepts, namely: checklist/completeness, openness, FAIR items, data archiving requirements, metadata, domain subjectivity, learning process and, institutional support.

Based on the interview results, which are shown in Section 4, we manually screened 98 proposals where a data management paragraph section is to be filled by grant applicants. There, we have seeking to classify proposals according to the reusability criteria communicated by the interviewees. The results of the grant proposal analysis are found in Section 4.2.

Table 3.3.

RDMP Interview Protocol

Grant application process
1. Would you please describe the grant application procedure within your agency and how you proceed with the applications?
2. Why is it important to have a DMP in the proposal phase?
Quality of data management plans
1. Do you believe that grant applicants pay enough attention to the DMP?
2. What feedback do you provide to researchers about DMPs?
3. Would you please explain in detail how DMPs are reviewed before submission (or evaluated after submission)?
Challenges of data management planning in practice
1. What are the challenges in reviewing the DMP?
2. What could help you more to overcome these challenges?
3. Are you aware of any suitable software for RDMP?

3.4 Results

In this section, we present the results of the interviews (see Section 4.1) and the analysis of the data management statements in 98 grant proposals (see Section 4.2).

3.4.1 Interviews

In this section, we present the results of the interviews. This section is divided according to the main categories of the interview protocol, which are: grant application process (Section

4.1.1.), quality of data management plans (DMPs), in Section 4.1.2., and challenges of research data management planning (RDMP) (Section 4.1.3.). We noted that there is a variety of grant application processes. In our study, two types of processes were identified. The first process is when a data section is submitted with the proposal, and more complete DMPs are only required after the project has been granted. The second process only requires a data management plan.

3.4.2 Grant Application Process

As the interviewees explain, the goal of these two deliverables is different. The data management paragraph is used to create awareness, and the data management plan is an elaborated document where successful applicants give more detail about data management matters. "so those are very general questions, and NWO also accepts general answers to those questions they are not going into too much detail because that is something that will come along when the grant is awarded. [...] if the grant is awarded, the researcher has to submit the data management plan within four months, and he or she cannot start the project until the DMP has been approved, so then things will be in more detail in DMP" [R4].

An interviewee F3 employed by a funding agency confirms that data management plans are more elaborated than paragraphs and have to be submitted later on: "But the main aim is to create awareness of the importance of good research data management. After the project is awarded, only the project leaders are requested to write a DMP according to the template the funding agency has, and, in that plan, they have to elaborate on the four questions they have answered in the main application form".

When a paragraph is not needed, funding agencies still expect researchers to be aware of data management matters though this requirement is not formalized in a specific section in the proposal. As explained by [F2], the paragraph has been removed from the proposal as "we did it before, but we did not do anything with it, and then we said if we do not do anything with it then why we do it, so we put it at the responsibility of researcher to take care of this part. I must admit that the researchers are very busy. They will not do that, but we strongly believe, and we strongly encourage that this is the own responsibility of researchers."

3.4.3 Quality of Data Management Plans

Next, we seek to extract a set of quality criteria for research data management plans, the main deliverable of RDMP during the submission of grants. We grouped the quality criteria in three

categories: completeness of the document, openness, and showing that a data management plan is a living document, which means that researchers announce that they will keep the DMP up to date.

The completeness of DMPs is a recurrent criterion appearing in the interviews. For funders and data services, a complete DMP shows that researchers thought seriously about data management matters. As explained by the interviewees: "Our main concern is completeness and whether the impression that the researchers have really thought about their data" [F3]. "if this type of basic things are missing or are explained in a very general way and not detailed, so this kind of check is more like a completeness check" [F5], "When it comes to DM plan I try to check whether they answered the questions and whether their answer is complete because if the questions have several parts, they [= researchers] tend to answer the first part and they tend to forget about the second part" [R4].

Regarding openness and exchange of research data, grant applicants might state in a DMP that their data will be exchanged between partners in a research consortium but not be necessarily made available outside the consortium, as explained by [F1], "there is still the option of opting out from openness of data, and a DMP is also focused on the exchange with consortium partners." The absence of justification on why data is not made openly available might lead to the rejection of the data management plan, as it happened to [R2]: "researchers say they will not open their data and they do not say why. So, the agency will get back and say, please explain why your data is not made openly available". When grant applicants mention closed data without justification, it happens that funders reject the DMP as they believe that researchers should at least share a part of their data "we always advocate this idea of maybe you cannot share all of your data, but certainly, there is some data that you can share. Then you choose a mixed access regime: part of it is open access, part of it is restricted access. But it always needs to be findable, and it is always good description and license or how to access it and, who can access it and what is possible to do with it, who to contact if you cannot get direct access and if they like to use it in follow-up research and so on... So, research funders could be more demanding than they are at this moment it may be a matter of time". Nevertheless, according to [R5], competition for grants play a role against openness: "in my opinion, there is a controversy between open science and the competition within the disciplines as well because in one hand you are competing with others to get the grant then you are asked to put all your data open."

Submitting a data management plan is not the end; funders expect to receive future updates during the project. As explained by [R4], "you can simply start with the project, and within the six months, you were supposed to deliver your first version of DMP where the two standards

agree on that the DMP is a living document. You have strict deadlines for delivering the first version, but you should also deliver an updated version when there is a major change that is relevant, for instance where you planned to collect certain data, which is no longer possible or when for instance a partner that has strong impact on the data quite the consortium and stuff like that". In practice, we did not collect evidence that researchers indeed do DMP updates once the grant is received, on the contrary. Interviewee [R2] states that "No, we usually lose track of the researchers after they submit their DMP. Sometimes I run into them and ask them, and they keep me up to date. But we do not actively ask them about the DMPs after they had come to us for help".

3.4.4 Challenges of Data Management Planning in Practice

As mentioned earlier, the review of DMPs is still in an early phase, and that involves a learning process between funders, data stewards, scientists. More specific challenges identified in the interviews are discussed hereafter.

First, reviewers have no specific guidelines, especially for helping researchers with FAIR principles. "So, these principles need to be operationalized, maybe more concrete, but at the moment they do not. Often, it is not clear what it is, what we need with findable, accessible, interoperable, reusable. We have some ideas, but we still need some guidelines, you talked about guidelines, but with this, we still need a lot more guidance" [R3].

Also, the absence of specific guidelines affect funding agencies as funders cannot provide feedback, "I think in the procedure we have right now it does not happen that often that researchers receive detailed feedback on their DMP as I explained earlier, we have more check on the completeness and adequateness of the answers but the [organization employing F3] has not this capacity as the average employee is not a data manager". Some funders promised feedback but also encounter issues with the required expertise: "Another problem with encounter is that we promised to give feedback on a DMP so that means that my colleagues have to know something about DM because they have to give feedback on DMPs" [F2].

Besides, there are only limited efforts put in planning data due to the low probability of obtaining a research grant. [R1] often sees researchers dealing with data management as a last-minute effort: "my experience at data section in the university was more or less the very last part of the proposal that was written, it was mainly a copy-paste effort because we provided some template so there is not many individuals or thoughtful effort there because that is when the DMP actually needs to be written. So, for the data section, it is more or less a copy-paste effort that the RDM team or other support team provided". [F1] concurs as this phenomenon is quite apparent in projects

depending on the organization of consortia: “And then it is really difficult to also include their opinion in the data and data management plan in time for submission. You see the same thing actually happening also with the consortium agreements which are most of the time also signed after the proposal is granted and it is kind of interesting that also the commission looks for the additional effort before submitting a proposal without even knowing whether a proposal will get granted”.

Another point of concern that funding agencies are aware of is that researchers lack the budget for RDM to make data available for the long term (after the project). At the same time, funding agencies expect the funding of long-term preservation to be a contribution from research institutions, as explained by F3: “a researcher can budget the management of research data within the project but of course not after the project has ended... Then it comes really to the budget of research institutions themselves, and I do not yet have a clear view on the availability of funds on an institutional level yet. I really can imagine that for some research, this is a problem, but also, we think that the university should invest in providing funds to store data for 10 years” [F3].

Lastly, there are academic disciplines where researchers lack knowledge about research data management solutions. Moreover, not only researchers lack insights into RDM matters. Also, when legal issues such as privacy are involved, it appears to be challenging for employees of research data services to provide the right support to researchers: “sometimes researchers are not aware of the technical possibilities of storing data on the network. When they do field research, they store data on the laptop, for instance, with all the possible dangers. Second, they are not aware of the technical possibilities when it comes to storing large data sets. They are very much surprised if they hear they can put up to terabytes without any costs on the network disks of our institution [...] sometimes we notice that they buy their own external hard disk because they just have the assumption that they do not have enough disk space.”. In addition to proper infrastructure, legal knowledge is also not trivial to find at the side of research data services, as [R2] states “and personal data for the most parts, it is not my strong point I mean I am really trying to dive in to it. But that is the most delicate one because there is an actual legal requirement to protect data and these things and not only, I might not be knowledgeable about it. I mean... I know quite a lot about it, but I am not a lawyer. But also, researchers often come to you with I do not have personal data. And you know that clearly, they do, they do interviews, and so then you are telling them hey you have personal data and sometimes trying to explain that there are some issues. There are quite a lot of things that you change because you have personal data. These can possibly be the problem”.

3.4.5 Grant Application Analysis

The study of the data management section in grant proposals aimed at exploring how grant applicants intend to plan research data management, and more specifically, how they intend to satisfy the reusability criterion set by funders. What can be seen from Figure 3.1 is that 42 data management sections in grant proposals met the requirements of reusability (as shown in Table 3.4). Nevertheless, 52 grant proposals (out of 98) did not contain an answer or a very generic answer to the proposal section dedicated to RDM.

Figure 3.2

The Results of the Manual Inspection of 98 Data Management Paragraphs

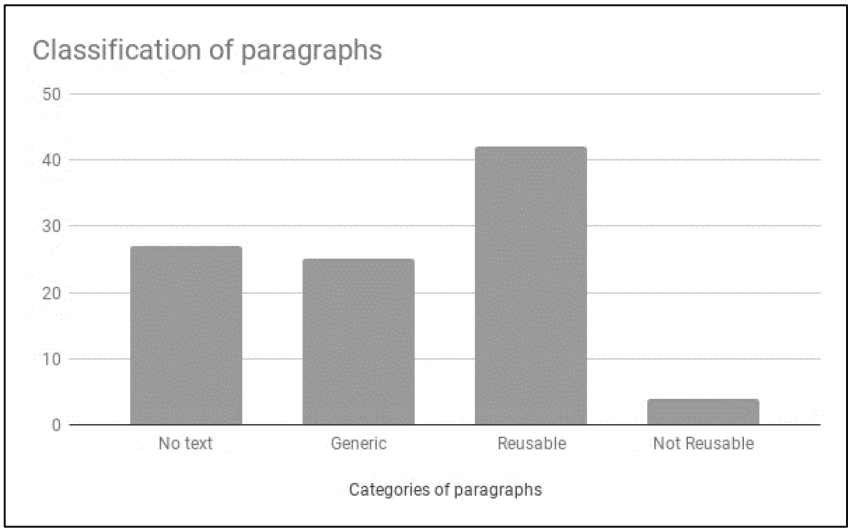


Table 3.4

Criteria for the grant application analysis

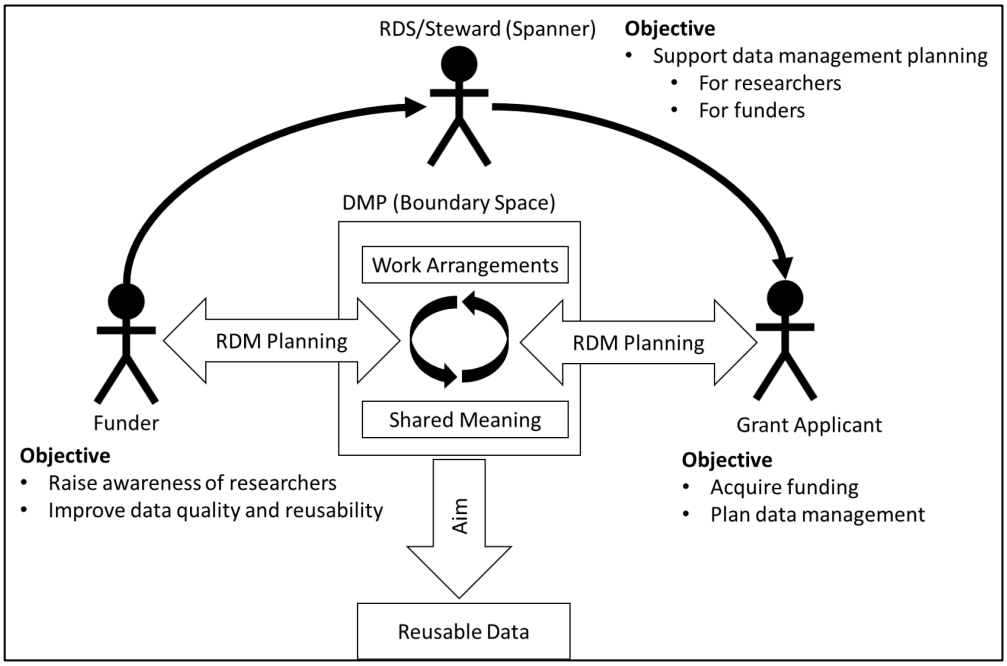
Criteria	Description
No text	The grant applicants have not answered the data management paragraph
Generic	The modalities of data reusability are not explained clearly, or the grant applicants indicate that the data is partially not reusable
Reusable	Grant applicants explain how their data will be reusable
Not reusable	Grant applicants explicitly state the data will not be reusable and open

3.6 Reflection and suggestions

We conceptualized the context of research data management planning in Figure 3.2, where we see that data management plans play a central role in research data management planning (RDMP). The main criteria presented by the interviewees are that DMPs must be complete, describe relevant metadata standards, and address the (absence of) openness of the research data produced during the funded project. Further, the ideal situation is that data management planning remains synchronized with the project, so any updates must be communicated to the funder. Research data services are depicted as a ”boundary-spanning role” (Hislop et al., 2018), where the objective of data management services is to translate the funder’s requirements into acceptable DMPs and, subsequently, proper RDM in practice. Unfortunately, a key lesson from the interviews is that this” boundary-spanning role” is currently limited by the unavailability of appropriate guidelines and the capacity to provide feedback.

Figure 3.3.

RDMP context where funders and grant applicants interact with research data management planning (RDMP) supporters at universities



Furthermore, we provide an overview of the criteria that lead to satisfactory data management sections in grant proposals, i.e., the items that should be addressed by grant applicants. A list of items corresponding to the reusability criteria is shown in Table 3.5. The suggestion is to use the corresponding feedback to grant applicants before the proposal is submitted, with automated feedback technology using natural language processing (NLP) techniques. With automated feedback, the paragraphs in a grant proposal are divided based on a template of the proposal containing the questions to be answered by applicants. Then, sentences are selected, and queries are run on each sentence related to the RDM section. Examples of queries written to extract data from full-text paragraphs are:

["LOWER":"Publicly","LOWER":"available","LOWER":"archive","POS":"NOUN"],["LOWER":"partly","DEP":"advmod","LOWER":"available"].

In case an item, as shown in Table 3.5, is missing, feedback corresponding to the category found in the document can be shown to grant applicants. In the end, the user can see the comments on the DM section as well as the rest of the proposal on his/her screen. So far, this approach has been preliminarily evaluated with "Impacter," an automated feedback tool on grant proposals (https://impacter.eu/). Hence, software solutions such as Impacter could further implement the items and corresponding feedback to also address research data management planning in grant proposals.

From the interviews, we observed that research data management planning is still at an early stage, where the goal of reusable scientific data is not yet fully supported by efficient planning processes and clear quality criteria. A lot of expectations are put on the shoulders of grant applicants, i.e., the researchers, but, at the same time, the technological infrastructure for RDMP is not yet fully functional. The fact that RDM and RDMP are complex for researchers has been investigated by a number of previous studies, for instance, (Akers, 2017; Borgman, 2015). Our study confirms the view of RDM as a network of stakeholders, institutions, and individual grant applicants with competing interests, sometimes leads to the production of non-reusable data.

Moreover, we covered a perspective of RDM where little previous research exists, as most studies focus on funders, researchers or, data services independently (Borgman, 2012). As such, we depicted that RDMP is an ongoing effort where data services also possess a mediating role between researchers and funding agencies, keeping funders informed of concrete issues occurring in practice, which then lead to the revision of funding policies.

Table 3.5

Summary of items expected for reusable data by funders and data management services

Item	Corresponding feedback
Data format	What documentation do you need to make your data format more authentic and transparent [R1]? If your data do not have a common form, please provide a relevant convert tool [F1].
Persistent Identifier	The DOI code and catalog and repository and data format are the critical requirements of Data Management for your project [F2]. "Thinking of a good license is indeed not really a challenge but you need to do it!" [F4]
Metadata	Metadata is important for the findability of data [R4]; to produce reusable data you need to be more specific about metadata. You can follow Dublin Core or Data Cite standards [R4, R2, F3].
Repository standard	"Most repositories they do confirm at least Dublin core standards" [R4]. You can find a list of trusted repositories on the funder's or university's website [R4, F3, R3].
Domain-specific Repository	Data management in different research fields varies a lot and domain experts will be reviewing your data management plan later [F5]; therefore, you have to comply with the demand of your faculty [R5] and get domain-specific guidance [F4].
Certified repository	Open access repositories are able to keep your data even after the project! It is better to calculate the cost of your data maintenance during and after the project [F3, F5, R1, R3].
Licenses	Dataset needs to be findable and have a good description and license or how to access it and who to access it and what is possible to do with it who to contact if you can't get direct access [F4]
Empty Paragraph	At this point in time, you need to talk about the basics of your data management [F1]. The DOI code and catalog and repository and data format are the critical requirements of Data Management for your project [F2].

As a suggestion of future work, one might investigate how technological solutions like automated feedback generation with natural language processing might circumvent some of the knowledge issues experienced by funders, data services and researchers. Data reusability can be operationalized using the quality criteria of funders, as shown in Table *3.5*. For instance, funders might expect that trusted digital repositories support compliant metadata standards, which then leads to an increasing the findability of the data. Such technology could implement entity recognition or more straightforward dictionary-based approaches to detect the presence or absence of these elements in data paragraphs and data management plans. The underlying reason is that we are far from being able to evaluate the effectiveness of RDMP without having insights into the outcome, i.e., the reusability of the data. Therefore, future studies should seek to systematically

study data management plans to obtain a better view of the extent to which DMPs comply with the criteria for reusable data.

Additionally, we covered a perspective of RDM where little previous research exists, as most studies focus on funders, researchers or, data services independently (Akers and Doty, 2013; Corti et al., 2014). In Figure 3.2, we summarize the findings of the case study. As such, we depicted how RDMP is an ongoing effort where data services also possess a mediating role between researchers and funding agencies, keeping funders informed of concrete issues occurring in practice.

However, a limitation of our exploratory study is the limited number of funding agencies and representatives present in our sample. Still, we aimed at offering insights from outside of the Netherlands by including the perspective of the European Commission by interviewing a representative of the H2020 funding program. As we provided the questionnaire, we hope it will foster follow-up studies to conduct additional interviews in other EU countries and elsewhere, to provide a deeper understanding of the variety of grand application processes, and RDMP more specifically. Next, we saw that our findings corroborate studies covering other states than the Netherlands in terms of challenges found in Williams (M. Williams et al., 2017). Williams et al. found a similar trend indicating that funders and services focus on reusability and sharing but less on particular aspects during the project. That being said, none of the previous work attempted to cover the perspectives of funding agencies and data services, highlighting the reciprocal learning process in which both funders and data services co-evolve.

Another limitation is that we focused on the sharing aspect of research data, as it is the aspect that is emphasized in grant proposals. As such, the knowledge gained from the interviews is limited to data sharing, but other elements of RDM could be considered as well. Corti et al. (Corti et al., 2014) described additional aspects than data sharing, which are essential to address during the planning phase: responsibilities, formatting, storing, ethics, copyrighting and, sharing. These aspects could be investigated more in-depth by analyzing data management plans instead of grant proposals, which were limited to data sharing and reusability in our case.

3.7 Conclusion

To conclude, research data management planning (RDMP) has many ongoing challenges, which makes the evaluation of its soundness and effectiveness to generate reusable data a complex task. Furthermore, while there is an agreement of our interviews to expect research data to be reusable, the practicalities and criteria might differ. So, we have identified differing planning

processes, non-standard quality criteria, and a series of complex challenges occurring during the planning phase. At the same time, we have integrated recurring points of improvements from our respondents into actionable criteria to ensure that RDMP is addressing data reusability properly. By doing so, we have strived to contribute to a better understanding of RDMP as a crucial process within research data management.

Section 2. Experiments and Reproduction in Laboratories

Chapter 4 | Opening Laboratories: Evaluating Replicability of Experimental Resources

Recently, the topic of research data management (RDM) has emerged at the forefront of Open Science. Funders and publishers pose new expectations on data management planning and transparent reporting of research. At the same time, laboratories rely upon undocumented files to record data, process results, and submit manuscripts that hinder repeatable and replicable management of practical resources. In this study, we design a forensic process to reconstruct and evaluate data management practices in scientific laboratories. The process we design is named Laboratory Forensics (LF) as it combines digital forensic techniques and the systematic study of experimental data. We evaluate the effectiveness and usefulness of Laboratory Forensics with laboratory members and data managers. Our preliminary evaluation indicates that LF is a useful approach for assessing data management practices. However, LF needs further developments to be integrated into the information systems of scientific laboratories.

This work is an extended version of the paper originally published as:

Lefebvre, A., & Spruit, M. (2019). Designing laboratory forensics. In I. O. Pappas, P. Mikalef, Y. K. Dwivedi, L. Jaccher, J. Krogstie, & M. Mäntymäki (Eds.), 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019. Trondheim, Norway: Springer.

4.1 Introduction

Research data management (RDM) is a pillar of sound data preservation and dissemination practices as encouraged by Open Science (European Commission, 2015). However, RDM has not (yet) reached the maturity of data management in the industry in terms of research, governance, and technology (Lefebvre et al., 2018). These ad-hoc RDM practices result in digital resources being inconsistently preserved in laboratories, thereby increasing the complexity of finding and accessing research data by laboratory members and external parties (e.g., reader, reviewer, another laboratory). Therefore, the consistent documentation of research processes, preservation, and dissemination of the artifacts created is still a complex challenge (Bechhofer et al., 2013).

It can be argued that finding experimental data on storage systems in a laboratory is similar to finding any evidence on any computer. As the original author of the files, a quick scan of the file hierarchy is enough to recover most of the files mostly needed for a given purpose. For instance, finding a file to send with an e-mail does not require an advanced process to locate, verify, and validate the files to send to a correspondent.

In contrast, when laboratory members are responsible for storing research data, it may be difficult for a third party to interpret the file hierarchy and identify relevant files (Federer et al., 2018). In a scientific laboratory, it is not uncommon that files created by a laboratory member need to be retrieved by others, for instance in the case a corresponding author has to respond to a request from another laboratory to access data (Collberg and Proebsting, 2016). At this point, the convenience of a simple file system becomes a significant limitation; the reason is that the understandability of the file structure largely depends on the efforts of the original authors to organize their folders and files.

Once experimental results have been published by a laboratory, the scientific community also benefits from available computational work. As noted by Peng (Peng et al., 2006), any attempt to reproduce published results require the availability of the original artifacts produced by the authors. In an era where computer technology has invaded scientific laboratories, few experimental works can avoid analytics software to study natural phenomena (Stevens, 2013). Still, the resulting publications often refer to a limited number of the original digital resources, if any (Federer et al., 2018). Consequently, the reusability and replicability of published experiments remain challenging due to the lack of available original computational resources (Ince et al., 2012).

In this paper, we present the outcomes of a design science research (DSR) study focused on the design of a forensic approach which evaluates the functional repeatability and replicability of publications based on digital resources preserved on storage systems in a laboratory. The name of the approach is "Laboratory Forensics" as it combines digital forensic techniques on digital evidence. As further explained in Section 4, we aim at providing a set of artifacts that data managers and laboratory members can use to optimize the maximal availability of experimental evidence associated with scientific publications.

The main contribution of this work is a set of forensic techniques applicable to the extraction of experimental data from laboratories. Moreover, the outcomes of several forensic cases are evaluated with laboratory members and data managers in one university. By this, we show the feasibility and utility of laboratory forensics. The research question guiding this work is "How can digital forensics techniques be used to assess the reproducibility of scientific experiments?"

The paper is structured according to Hevner's s DSR model (A. R. Hevner, 2007). More information about DSR is given in Section 4.2. Briefly, the structure of the paper revolves around DSR rigor, relevance, and design cycles. In the literature review section (Section 4.3); we present digital forensics and experimental systems, both of interest for the rigor cycle (A. R. Hevner, 2007). Then, in the Design section, we describe the outcomes of the evaluation of the laboratory forensics approach on four cases (i.e., publications). Finally, we discuss future research and conclude in Section 4.7.

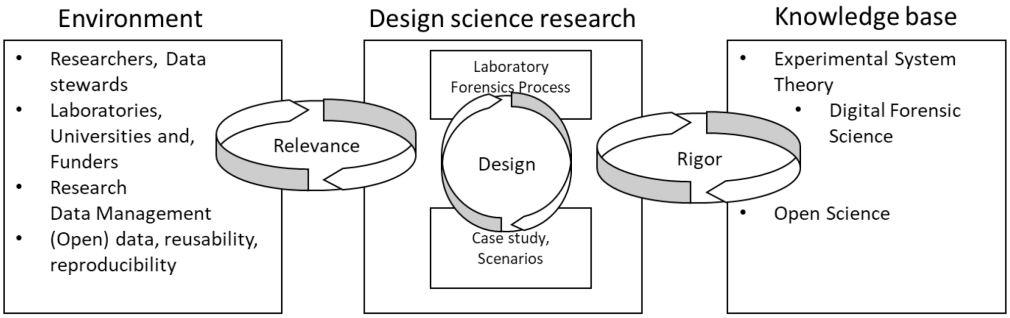
4.2 Design Science Research

Design science research (DSR) addresses organizational problems by designing useful IT artifacts such as software, methods, models, or design theories (Gregor and Hevner, 2013). Hevner (A. R. Hevner, 2007) describes a design process that consists of three cycles named the relevance, design, and rigor cycles. The three-cycle DSR aims to ground artifact design in rigorous construction and evaluation methods. Figure. 4.1 shows a typical three-cycle model adapted to the study presented in this paper.

In DSR, the rigor cycle draws upon a body of knowledge named "knowledge base." There, design scientists describe the theories, processes, and evidence used to justify and evaluate a designed artifact. We explain further the domain and methods which are included in the rigor cycle in the next section. Similarly, we elaborate on the relevance cycle in the domain relevance section, where we give more details about the context of data management in scientific laboratories and the evaluation criteria adopted in this study.

Figure. 4.1

The Three Cycles of a Design Science Project, Based on Hevner (2007)



4.3 Literature Review: The Rigor Cycle

In this section, we elaborate on two key aspects which drive our DSR study. First, we introduce digital forensics. Next, we briefly present a general view on the process and system of scientific experimentation.

4.3.1 Digital Forensics

Digital forensic science (DFS) has been defined as: "the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal or helping to anticipate unauthorized actions shown to be disruptive to planned operations", p.16 (Palmer, 2001). In other words, DFS employs an ensemble of techniques to transform digital resources into evidence usable by third parties in a convincing way. DFS often relates to (cyber)criminal activities. Hence the outcomes of DFS investigations serve judiciary systems by reporting on digital evidence found on computers involved in criminal activities (Casey et al., 2013). As we will explain in the design section, the constraints of reliability and rigorous reporting found in DF investigations form a strong basis of rigor to transfer DF techniques to scientific laboratories. The reason to investigate laboratories with DF techniques is twofold: one is reactive (i.e., something happened) and another proactive. The former refers to the investigation of the preserved evidence underlying scientific publications to reconstruct past experiments. The latter refers to digital forensics readiness, a field of DF which prepare information systems to deal with external threats (Rowlingson, 2004). In the context of Open Science, this translates to evaluating the readiness of data management to comply

with the production, proper preservation, and, dissemination of high-quality scientific data (P Ayris et al., 2018).

Table 4.1

Forensic Processes All Converge Towards a Stable Set of Activities

Step	Årnes	ENSFI	Casey	Candidate techniques (Palmer, 2001)
1	Identification	Identify	Preparation and Preservation	Audit Analysis, Case Management
2	Collection	Acquire	Extraction and Storage	Sampling, Data reduction
3	Examination	-	Examination and reporting	Filtering, Pattern matching
4	Analysis	Analysis	-	Statistical, Timeline
5	Presentation	Report	Sharing, correlating, and distributing	Documentation, Impact statement

What can be seen from Table 4.1 is that the DFS process described by Årnes (Årnes, 2017) is more refined than the two others. The reason is that Årnes makes a distinction between the examination and analysis phases. This distinction is facilitating the decomposition of the forensic process into clearly defined steps. Also, this distinction refines the categorization of candidate techniques that are used at each stage of the process suggested by Palmer. Candidate techniques are methods from other disciplines that belong to an analysis step in digital forensics (Palmer, 2001).

According to Årnes, a DF investigation starts with the identification of the data sources. The next step, i.e., collection, is the actual extraction of the evidence from existing storage systems. Collection requires an image of the disk of interest to the investigators as it would be impractical and even hazardous (e.g., unexpected modifications of files) to investigate the laboratory storage in use. Once the evidence is isolated from a computer device, the examination phase locates potential evidence. After the investigators have recovered potential evidence, the analysis phase takes place. The last step, presentation, is the translation of the findings into a format that can be understandable by the practitioners.

4.3.2 Laboratories and Experimental Artifacts

Scientific laboratories are standard organizational settings encountered in natural sciences such as Physics, Chemistry, and Biology (Franklin and Perovic, 2016; Weber, 2018). At their core, laboratories are organizations producing scientific knowledge by designing and operating experimental systems. Experimental systems are closed systems that enable the observation of

natural phenomena with an ensemble of equipment, theory, and human intervention (Radder, 2012a).

Moreover, experimental systems produce intermediate products from experimental events that are not part of the (communicated) output. These products are, for instance, exports from data analysis software, manuscript's drafts, quality controls, interactions between technicians and researchers (i.e., experimenters) and computer scripts. The association for computing machinery (ACM) has highlighted the need for a better assessment of the quality and availability of digital artifacts underlying publications (ACM, 2018). The ACM classifies artifacts in two categories: functional and reusable (ACM, 2018). Functional are artifacts that are consistent, documented, complete, and exercisable (i.e., runnable on a system).

4.4 Domain Relevance

4.4.1 Application Environment and Case Selection

In laboratories, scientists and technicians transform an object of study into data and data into scientific facts (Latour and Woolgar, 1986). In science, facts are mainly communicated through scientific articles in journals which are presenting a curated version of experimental events (Borgman, 2008). Recently, journals developed new guidelines for more transparent reporting and enriched supplemental information (Federer et al., 2018). Concurrently, public funding agencies encourage proper research data planning and management to foster high-quality data dissemination to mitigate the risks of poor RDM in laboratories. This trend presents several challenges for laboratories.

First, data management is not a priority, as it is often not rewarded by the academic system (Nosek et al., 2015). Second, as explained earlier, laboratories manipulate quite complex experimental processes to obtain results. As experimental processes rely on novel technology and people pushing forward the boundaries of a discipline, it is challenging to keep a record of the experimental evidence and activities produced during several years.

The case laboratory is a proteomics laboratory in the Netherlands that has produced over 400 publications in the last ten years. It makes this laboratory an ideal environment to design and evaluate our approach due to the complexity of the analyses done in the laboratory and a large number of authors (over 100) that worked or is currently working in the laboratory.

4.4.3 Evaluation Criteria

The criteria used to evaluate the outcomes of our LF approach are effectiveness and usefulness. First, we discussed the forensic results with two laboratory members, one experienced post-doc, acting as a data manager in the laboratory, and one a Ph.D. student who is the authors of one of the investigated publications. In addition, the outcomes of one forensic case were presented to 20 participants present at an RDM community event in February 2019. The participants were data managers, senior members of a data governance board, and members of RDM services at the University. Hence, both researchers and data managers could comment on the preliminary outcomes of the laboratory forensics approach presented in this paper.

The forensic cases are all publications originating from the investigated laboratory. The cases, i.e., publications, are selected primarily on the locality of the resources and their year of publication. For this study, we did not include any publication with multiple affiliations to limit the spreading of the files in separate laboratories. The publications are recent: CASE A and CASE C are from 2017, CASE B from 2018 and, CASE D from 2019. The reason is that the storage systems have been adapted recently, making the retrieval of older files an extra challenge due to their relocation, which influenced critical meta-data such as date of creation to a large extent.

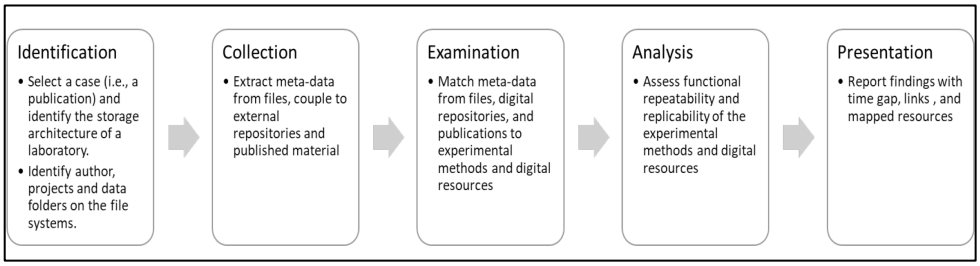
4.5 Design Iterations

The LF process, see Figure 4.2., is designed by integrating digital forensic activities shown in Table 4.1 with evidence typically found in experimental science. The general idea is to merge meta-data collected from a file system in a laboratory (e.g., file names, date of creation, and date of modification) to software, methods, and external sources found in the corresponding scientific article.

To design our laboratory forensics approach, we first started with one publication, hereunder CASE A. The first design iteration was focused on adapting digital forensic techniques to the material available in on storage systems and their specific nature. The fact that (experimental) resources might be domain-specific made it necessary to redefine digital forensic activities, which are increasing the likelihood of locating relevant material. Figure 4.2 shows the current iteration of the laboratory forensic process.

Figure 4.2

An Overview of the Laboratory Forensics Process



4.5.1 Laboratory Forensics in Action

A major challenge encountered during the investigation of CASE A was low confidence that the files retrieved were related to the publication as much more (digital) material was generated by the instruments in the laboratory than necessary for the publication. This fact made the Annotated Experimental Resource Table (AERT) useful to guarantee that the files retrieved remain consistent with the resources reported in the investigated publication. The AERT is a spreadsheet containing all the methods described in the corresponding articles, the software and laboratory instruments used, and external resources such as an online database used by the experimenters. The AERT is helpful for systematic mapping of files and reported resources and excludes irrelevant material. We note that this mapping requires that LF investigators possess in-depth knowledge of storage systems and domain-specific knowledge (here in proteomics).

Table 4.2

The Main Activities of the Laboratory Forensics Process

	Activity	Sub-activity	Example outcomes of CASE A
Identification	Screen experimental report	Identify laboratory employees	There are two authors in the author list and three laboratory members listed in the acknowledgment section.
		Collect information about the editorial process and deposited data	The paper was published in June 2017. It was received by the journal three months earlier, in March 2017. One repository is used to deposit raw data.
Collection	Screen laboratory	Collect administrative data about current and former employees	One author is a principal investigator; two are postdocs. One member is a technician—one member with an unknown status.
		Determine storage systems architecture	There are several shared volumes used by the laboratory members. The data is shared between raw, users, and project volumes.
	Extract resources	Extract experimental methods	There are two types of analyses reported in the paper, each with their software and instrumentation. In CASE A, those are HDX-MS analysis and MS-MS analysis.
		Extract experimental resources	In the publication of CASE A, we extracted: Waters nano-Acquity UPLC system, […], Swiss-Model, Phyre2 server, […] as a list of instruments and software.
		Extract external repositories and supplemental information	In CASE A, files have been deposited on PRIDE, a digital repository, and an access number is present in the additional information.
Examination	Make snapshots	Perform (targeted) snapshots of disk volumes	Select the user and project folders belonging to the laboratory members involved in the publication. Make a snapshot, for instance, in PowerShell (Windows) with the commands Get-ChildItem and Export-CSV.
	Process snapshots	Consolidate snapshots	Merge user and project folders in one (tabular) data sets.
		Reduce noise	Duplicates and system files are deleted from the consolidated snapshot.
	Construct Annotated Experimental Resources	Extract experimental methods	Determining what qualifies as an experimental method depends on the context of the analysis.
		Link experimental resources to methods	The software is extracted for each method mentioning the use of the software.

	Activity	Sub-activity	Example outcomes of CASE A
Analysis	Table (AERT)	Link laboratory instruments to methods	For a method, each instrument used can generate output with specific extensions (such as .RAW files in proteomics).
	Map file paths	Map files to AERT elements	File system meta-data might be inaccurate. Hence, cross-checking with elements in the AERT table is crucial not to include irrelevant files
	Filter file paths	Filter file paths referring to a publication	Not all files in a folder might belong to the analyzed publication; additional filtering is needed to exclude unnecessary files.
	Estimate functional repeatability and replicability	Estimate Method coverage	An estimation of the number of methods reported in the publication that are covered by evidence left on the storage.
		Estimate Software coverage	An indication of the number of software resources that are successfully identified from reading the file paths.
		Estimate Instrument coverage	An indication of the number of laboratory instruments that are identified by the investigator.
Presentation	Report findings	Report file mappings and their confidence	We included the folders of the first and last author in our analysis and reported the folders used during the analysis.
		Report functional repeatability	The extent to which preserved files are consistent, complete, and not fragmented.
		Report functional replicability	The extent to which disseminated files are consistent, complete, and not fragmented.
		Evaluate report with laboratory member(s)	We evaluated the report with one domain expert (laboratory member).

The goal of the LF approach and its activities shown in Table 4.2 is to present to research a report on the state of the storage in terms of reproducibility. To evaluate the functional repeatability and replicability, the following classification indicating the degree to which the preserved evidence corresponds to the requirements of completeness and consistency of the artifacts as described by the Association for Computing Machinery (ACM) (ACM, 2018). The classification shown in Table 4.3, i.e., low, medium, high repeatability/replicability is diverging from the ACM is two ways. First, functional artifacts are divided in terms of locality: repeatable is used for resources preserved in the laboratory, and replicable is used for resources disseminated in supplemental information and digital repositories. Second, in terms of degree (low, medium, high) to account for varying scenarios.

Table 4.3.

Findings From the First Iteration on One Publication (CASE A)

	Outcomes	CASE A	Comment
Identification	Number of laboratory members	5	Each user folder of a laboratory member has to be mapped and investigated.
	Number of external authors	2	Several authors from external research groups are listed. External authors do not have user folders. On the laboratory storage, authors are mentioned by name, identifier, or affiliations.
	Editorial process duration	29 March–28 June, 2017	Filter project folders that were updated (i.e., modified date) at the time of submission.
Collection and Examination	Number of methods*	4	Four method subsections are referring to computational work or laboratory instruments manipulated with software.
	The number of software resources	10	There are ten software resources used (e.g., to preprocess and visualize data).
	Number of Instruments	5	Five laboratory instruments were used to generate raw data.
	Number of files (local/deposited)	3011 / 15	In total, the consolidated mappings contain 3011 files on the storage and 15 files in the external repository.
	Total file size	49.5 GB	The "weight" of digital evidence of the investigated publication is around 50 GB.
	Time delta **	1486 days	The first file included as experimental data was modified more than four years before the last file was included.
Analysis	Corresponding software	5	Files corresponding to 5 software resources are located, which means five other software resources have no (explicit) traces left on the storage or online.
	Corresponding Instruments	4	One instrument could not be mapped to the digital evidence found.
Presentation	Functional Repeatability	MEDIUM	The evidence is complete and entirely consistent with the corresponding experimental report. However, files have not been aggregated in a project folder, which requires to investigate several folders across different folders to obtain the complete (computational) input.
	Functional Replicability	LOW	Only the necessary raw files of one method have been deposited. Direct replicability is, therefore, hindered by the absence of other artifacts that are necessary to replicate the results.

* Computational methods, ** based on file system meta-data, not the exact duration of experiments

4.6 Evaluation of Laboratory Forensics Outcomes

We evaluated the usefulness of LF results with a member of the laboratory responsible for the storage infrastructure. We collected the impressions of our contact person after a short presentation of the forensic process and report of CASE A (see Table 4.3). The impressions formulated by the laboratory member indicate that our approach appears to be rigorous and convincing. The systematic classification of resources into instruments, methods, and software sheds new light on the resources underlying a publication.

Next, the extent of the fragmentation of the files was not expected by our interviewee, which shows that the ability of an LF approach to gathering evidence beyond expected locations by users. Also, the communication of issues with a set of measurable indicators gives an overview of the strengths and weaknesses of data management for specific publications. A critical note was that the indicators and scoring method need further refinements and more transparency. For instance, the indicator of functional replicability should incorporate mandatory openness and non-mandatory openness. This distinction would help to distinguish scientific data dissemination imposed by publishers (mandatory) and self-motivated by researchers in the laboratory (non-mandatory).

Besides, we presented the outcomes of CASE A to the data management community meeting of our University in February 2019, attended by 20 participants. This presentation was meant to collect the impressions of people involved in data management services. There the primary impressions of the participants are that although the approach is time-consuming, it seems worth to conduct such analyses to explain to researchers the importance of acceptable data management practices. Even the fact that LF is challenging to accomplish is, by itself, a powerful example of problematic data management practices that data managers can use to engage with researchers about RDM practices. Further, to collect additional feedback about the effectiveness of the LF approach, we investigated three new cases to obtain better insights into alternative data management practices adopted by other laboratory members. The three additional cases are labeled cases B, C, and D (see Table 4.4).

Table 4.4

Summary of the Outcomes of Additional Cases

Outcome	CASE B	CASE C	CASE D
Size	1.2 GB	2.6 GB	136.9 GB
Number of Preserved / Deposited files	689 / 0	137 / 123	939 / 179
Corresponding software	2 / 8	1 / 5	2 / 6
Corresponding instruments	3 / 4	3 / 4	1 / 2
Functional repeatability	MEDIUM	MEDIUM	MEDIUM
Functional replicability	LOW	HIGH	MEDIUM

The second evaluation was driven by the question of whether a forensic analysis of a storage system in a laboratory retrieves more relevant evidence than laboratory members when they are asked for searching underlying evidence publications. To achieve that, we asked two laboratory members to collect data underlying a publication used in one of the four investigated cases. More, we asked the laboratory members, hereafter participants, to elaborate on their search strategy and judge the extent, according to them, of the repeatability and replicability of the article they received.

The participants reported located files in a word document or during a live demonstration. Their outcomes are consistent with LF assessment, which showed that relevant files are all preserved but fragmented on the storage (hence medium repeatability). Also, the participants expressed their difficulties in locating legacy data or data created by another laboratory member in the past. In that case, we found that the presence of a reference list of files created by the forensic investigation is essential to evaluate whether the participants retrieved the complete list of files or evidence was still not located.

4.7 Discussion and Conclusion

Throughout this study, we answered the following question: “How can digital forensics techniques be used to assess the reproducibility of scientific experiments?” A design science approach has delivered preliminary artifacts and evidence that laboratory forensics (LF) is a useful approach for evaluating storage systems in laboratories. Despite this, LF suffers from significant limitations in its current state. One limitation is that the LF process is yet to be further evaluated on several forensic cases in different environments to increase the rigor and reliability of LF investigations. These limitations are mainly due to the nature of reconstructing events from digital data (Mabey et al., 2018) and the complicated extraction of experimental resources from

publications. Moreover, access to storage systems in laboratories is needed, which might posit some additional challenges related to the privacy of the users. Despite the limitations of the current LF approach, LF has unique strengths compared to approaches for RDM, such as post-publication curation of research data (Bechhofer et al., 2013).

First, LF attempts to locate data despite reporting gaps and unavailable resources, unlike other studies relying on published material exclusively (Federer et al., 2018). Collecting evidence from storage systems allows going beyond the written account of the events that occurred in a laboratory.

Second, an LF investigation actively seeks to reconstruct experiments to accurately report on which experimental resources are used, by whom and locate the underlying materials. This can serve as input for reproducibility studies, where retracing the full life cycle of scientific discoveries is a prerequisite for understanding all steps taken in an experiment to guarantee its reproducibility (Huang and Gottardo, 2013).

Last, the extraction of structured data about experimental methods, resources, and data together with evidence on storage systems might be of high value for designing ontologies representing a particular field of study (Hoehndorf et al., 2013) with a higher ability to manage the artifacts in use in laboratories and guarantee reproducible storage patterns.

To conclude, Laboratory Forensics demands further development, evaluation, automation, and tooling to become readily available for scientists and data managers. Hitherto, we have been able to show that in daily practices (digital) experimental resources are not preserved in a functionally repeatable and replicable way in the investigated laboratory. In short, laboratory forensics support the development of rigorous assessment of data management issues related to laboratory work. In upcoming research, we will further investigate the synergy of laboratory forensics with research data management practices.

Chapter 5 | Identifying reproducibility threats in laboratories

The Open Science paradigm has brought the dissemination of experimental artifacts on the agenda of funding agencies, research institutions, and academic publishers. Managing research data is a crucial part of guaranteeing the reusability and reproducibility of published results. In this research, we suggest a conceptualization of reproducibility based on threats, risks, and vulnerabilities identified in current research data management (RDM) practices. By doing so, we can describe a range of threats to reproducibility and pinpoint areas where current RDM practices and the scholarly communication infrastructure insufficiently address these threats. Further, we elaborate on a socio-technical approach to reproducibility in RDM by collecting evidence from researchers and scientific publications. We show that the STS approach complements current IS research on RDM by offering a holistic view of reproducibility challenges in RDM.

This work was originally published as:

Lefebvre, A., & Spruit, M. (2019). A Socio-Technical Perspective on Reproducibility. MCIS 2019 Proceedings. Napels.

5.1 Introduction

Since the last decade, the reproducibility issue of scientific research becomes apparent and calls for attention. As reported by Laine, Goodman, Griswold, and Sox (2007), the amount of errors or misinterpretations of statistical analyses and reproduction failures of peer-reviewed academic work is surging (Donoho, 2010). As a consequence, a number of academic communities starts to promote a better scrutiny of reported results and encourage replication studies in different fields, which include also information systems (Casadevall and Fang, 2010; Dennis and Valacich, 2014; Laine et al., 2007b; Sandve et al., 2013). Although reproducibility is a notoriously ill-defined term in the literature (Plesser, 2018; Schloss, 2018), reproducibility is defined in the Oxford English Dictionary as "the extent to which consistent results are obtained when an experiment is repeated" (OED Online., 2019).

Moreover, reproduction (or replication) issues have also been discussed in information systems (IS) research in conference panels at the International Conference on Information Systems (ICIS) and the European Conference on Information Systems (ECIS) by Brown et al. (2016) and Olbrich et al. (2017). More, Dennis and Valacich (2014) launched the AIS Transactions on Replication Research, where IS scholars submit replication studies. In their replication, manifesto, Dennis, and Valacich (2014) state that replication falls into three categories: exact replications, methodological replications and, conceptual replications. The distinction made by Dennis and Valacich made is a starting point for our study. We observed that what fundamentally distinguishes exact, methodological, and conceptual replicability mentioned in the manifesto are that these categories are variations of who (i.e., same or other authors), what (i.e., theory, tasks, results) , how (i.e., same or different methods) and, where (i.e., same or different environment) studies are repeated. These categories also apply to scientific experimentation when we opt for a holistic view of the actors, tasks, technology, and structures participating in scientific experiments.

Besides, some communities of researchers took the initiative to underline the necessity of leveraging the scholarly communication infrastructure for managing and making research data findable, accessible, interoperable to be reusable (FAIR) by human and machine consumers (Wilkinson et al. 2016). Although the concept of FAIR data reached research data management policies at international and national levels (European Commission, 2016b), there is no joint agreement on what FAIR data is, nor what reproducible and reusable data entail.

Thus, our work is guided by the following research question: "What reproducibility threats occurring in experimental systems stem from vulnerabilities in research data management

practices?". By answering this question, we seek to contribute to the topics of research data management and reproducible research by (1) characterizing and identifying threats to reproducibility related to challenges encountered in research data management (RDM) (2) articulate reproducibility threats and risks according to a socio-technical perspective on scientific experimentation. By doing so, this paper extends risk management approaches applied previously on digital preservation (Miksa et al., 2014), which is one of the critical tasks of RDM.

The present paper is structured as follows: in the related work section, we make a parallel between experimental systems and socio-technical systems. Next, we introduce research data management activities and concepts. In Section 4, we present the outcomes of a mixed methods (i.e., quantitative, and qualitative) approach to acquire evidence from practitioners, institutions, funders in publishers. Finally, the analysis of the evidence led to the development of an STS reproducibility framework introduced in Section 5.

5.2 Related Work

5.2.1 Socio-Technical Perspective on Experimental Systems

Most of the literature dedicated to scientific experimentation belongs to the area of logic, epistemology, and statistics. Studies dedicated to scientific experimentation from a working scientist's perspective are scarcer than the studies on the logic, validity, and methodology of scientific experimentation. Nonetheless, academic work using a socio-technical view on scientific experimentation emerged in the philosophy of science (Radder, 2012b; H.-J. Rheinberger, 1997) and sociology of science (Latour and Woolgar, 1986; Stevens, 2013).

Therefore, we first need to introduce the experimental system perspective of scientific experimentation developed by Radder (2012). Radder's framework depicts scientific experiments as a system consisting of theory, materialization, and results. Hans Radder defines a closed experimental system S as a "complex of object and equipment within a specified spatial area and during a fixed interval of time" (Radder, 2012b). Radder defines the instantiation of S as a theoretical description (i.e., formal experimental process and theory), results (i.e., the outcomes of experimental events), and human intervention (i.e., operationalization, the translation of theory to experimental procedures). First, a theoretical description (TD) delineates the episodes (i.e., events and activities) occurring inside S. Radder adds that some episodes have a specific role which is to determine the relative closure of S. In short, S is qualified as being a closed system if non-experimental episodes do not interfere with the episodes and results.

Radder's view on experimental systems echoes more generic socio-technical systems (STS). As explained earlier, experimental systems can be decomposed into the production or manipulation of (IT) artifacts by human intervention. According to socio-technical models (Leavitt, 1965; Ahmad, Lyytinen, and Newman, 2011; Silver and Markus, 2013), variables composing ST systems are structure, tasks, technology, and actors. Experimental systems implicitly refer to similar variables. Such a correspondence enables the analysis of scientific experiments as a socio-technical system. The scholarly communication infrastructure (Wallis, Rolando, and Borgman, 2013) corresponds to the structure variable is STS. Further, the variable tasks correspond to the operationalization of experimental designs, as tasks are defined as being the artifacts and rules used by the actors in an STS (Lyytinen and Newman, 2008).

5.2.2 Research Data Management

Academia is facing similar challenges as other sectors such as business and industry to extract valuable knowledge from the increasing amount of data produced worldwide (Borgman, 2012). In experimental science, where sophisticated machines produce large quantities of measures and meta-data about phenomena under investigation in laboratories, the consumption, processing, management, and diffusion of these data are notorious challenges (Baesens et al., 2016; Borgman, 2012). Nevertheless, besides researchers, research data management governance has responsibilities distributed among multiple stakeholders.

Further, external laboratories, stakeholders such as academic funders, are progressively governing the production, preservation, diffusion of scientific data created by (publicly) funded research (OECD, 2007). As a result, researchers are facing new regulations, procedures, and technological challenges for managing data at each step of scientific experimentation.

First, public research funders posit some prerequisites for managing research data generated with public resources. For instance, grant applicants need to describe the (future) data sets, storage systems, and anonymization techniques, among other items, in data management plans (DMPs). Funders pursue societal ambitions of opening data and disseminating scientific knowledge. This ambition can be seen from the evolving National and European regulation, which encourages the dissemination of scientific artifacts in novel ways (European Commission, 2015).

Second, publishers are critical stakeholders of the scholarly communication infrastructure, which is an essential structure of communication in science. In recent years, scholars have investigated the challenges of current scholarship practices to deal with data sharing and reproducible research (Borgman, 2008; Reilly, Schallier, and Schrimpf, 2011). Research data

management seeks to transform the scholarly communication infrastructure to push academic data sharing and preservation forward. RDM is a collective enterprise, for which efforts are shared between research funders, academic publishers, research institutions, and researchers to achieve reusability and reproducibility of scientific output (Lefebvre et al., 2018).

More, academic publishers explicitly integrate research artifacts produced by researchers in their editorial processes. In recent years, publishers introduced data sharing, preservation, and dissemination guidelines and policies (Editorial, 2014). These policies are aimed at grant applicants who have to present data management strategies early in the application process. Experimental artifacts such as datasets, materials, and software must be precisely documented and, possibly, disseminated according to the publisher's and journals' guidelines.

Finally, research institutions reorganize their IT services to support researchers in managing research data at their host institutions. Research institutions deploy institutional repositories and technology for managing research data to secure funding opportunities. This fact led to new managerial and support roles in academia appearing in academia, such as data stewards and research data managers.

5.3 Mixed Methods Approach

We follow a mixed-methods approach (Bergman, 2008); thus, we apply quantitative and qualitative data collection techniques. We conducted semi-structured interviews to gather evidence from laboratory workers and acquired open data to analyze the scholarly communication infrastructure. We divided the data collection and analysis into two periods. During the first period, we gathered information about the management of scientific data by interviewing seven researchers in the bioinformatics community of one University in the Netherlands. There, we collected experiences of researchers in laboratories about data management practices. Throughout the interviews, we identified several challenges related to the preservation, interpretation, and dissemination of scientific artifacts. To increase the contextualization of our interviews, we first obtained a dataset from the administration of the University and analyzed an anonymized version of the data. This survey was submitted to the academic staff of our university in August 2014. For this survey, 829 researchers out of 3197academic staff members (source: annual report of the institution) answered, which is a response rate of 26%. After removing incomplete cases, 489 records were retained for further analysis. As we focus our analysis on reproducibility in experimental systems, we filtered the respondents on faculties that are using scientific

experimentation. Removing non-experimental disciplines further narrowed the sample to 289 respondents.

Next, after the interviews, we screened 323 publications in the domain of Biological Science (i.e., BIOC category on Scopus) as the focus of our study is on experimental work, 252 full-text publications were retained for further analysis (78%). The reason for removing 22% of the articles is that these articles did not report on experimental work (e.g., literature review) or did not produce research data with laboratory work (e.g., computer simulation using open data). Table 5.1 presents the characteristics of the sample of selected publications.

Finally, we sampled author guidelines and policies from seven publishers: Elsevier (ELS), Public Library of Science (PLOS), Cell Press (CP), American Chemical Society (ACS), Nature Publishing Group (NGP), Oxford University Press (OUP), eLife Sciences Publications (ELIFE). The guidelines can vary extensively from publisher to publisher. The rationale is that one single publisher might host several dozens of journals and decide that research data management matters are to be elaborated by each journal. Also, to cover funding agencies, we added documents from public funders: National Institute of Health (NIH), National Science Foundation (NSF), European Commission (EC), and The Netherlands Organization for Scientific Research (NWO, in Dutch).

Table 5.1

Characteristics of the scientific publications screened in this study. SI means Supplemental Information, which are files hosted on the publisher's website. Availability statements are paragraphs where authors describe how to retrieve the underlying data.

Publisher	Average SJR	Average Number of Authors	Average number of files in SI	Average Distinct Formats in SI	Percentage of Availability Statements	N
Cell Press	10.7	24.3	3.5	1.61	69%	42
Elsevier	2.3	10.3	1	0.9	5%	20
Nature Publishing Group	14.2	31.0	4.4	1.88	94%	17
Other	3.3	14.2	2.5	1.26	25%	119
Public Library of Science	1.5	10.2	3.6	1.28	100%	35
eLife Sciences Publications	7.1	11.1	2.31	0.89	31%	19

5.5 Results

5.5.1 People, Technology, and Tasks

The first part of the results section presents the outcomes of the survey and interviews. Next, in Section 5.6, the results of the screening of publications are shown.

5.5.2 Survey

As can be seen from Table 5.2, RDM practices, as reported by researchers to the IT services of their host institution, reveal that reaching the ambitions set by RDM stakeholders is an ongoing effort. Overall, it appears from the survey results that researchers seem reluctant to comply with RDM tasks set by funders, publishers, and research institutions. Also, there seems to be only a minority of respondents who would agree to depend on central IT services of their institutions, except data preservation. Data management planning (around 20%), assistance with lab notebook systems (and assistance with dissemination seem to rank less high than preservation, i.e., assistance and technology to back up research data in the long term.

Table 5.2

Responses From Researchers in the Faculties Of Science, Geosciences, and Veterinary Medicine (N=289). These Faculties Mostly Relied Upon Experimental Systems and Reported to Work With Experimental Data.

Task	Statement	Response	N
Report reproducible results	"You are interested in digital Laboratory notebook systems."	Yes - 60 (20.7%)	289
	"You need to record who accesses and modifies datasets."	No – 150 (46%)	250
Conduct data management planning	"You are interested in expertise in writing data management plans."	Yes – 69 (23.9%)	289
	"You created a data management plan at the start of the project."	Yes – 30 (10.4%)	289
Elaborate preservation strategy	"You want assistance in organizing long term preservation of data."	Yes – 113 (39%)	289
	"You are interested in long-term backup facilities."	Yes – 139 (48%)	289
Elaborate dissemination strategy	"You plan to make data publicly available"	Yes – 79 (27.3%)	289
	"You are interested in expertise for publishing data in a public repository."	Yes – 70 (24.2%)	289

5.5.3 Interviews

This section summarizes information about data management in distinct research laboratories. Seven researchers in the fields of biology and bioinformatics were interviewed. All interviews show different data preservation and dissemination practices as well as technology in place in laboratories. All interviewees are labeled by their laboratory, followed by their position: Principal Investigator (PI) or Postdoc (PD). We purposely discussed with interviewees who had proven experience in their respective domains, see Table 5.3 for an overview of the interviewees.

For analyzing data in **Computational Structural Biology, CSB/PI** has an advanced computation infrastructure (grid computing) and maintain self-developed analysis software utilized internationally. CSB/PI say that assessing the quality of the data needs specific expertise. The files generated have different structures that are specific to the application that generated them. There is also no permanent storage of intermediate processing products. Data sharing is sometimes not done by transferring data but by giving access to where the data is located as its size would be too resource consuming. CSB/PI recommends the use of meta-data to validate the format of the files, but meta-data is more challenging for evaluating the quality of the data itself.

In **Biomedical Genetics**, BG/PI explains that scientific data is reused, but analysis workflows are not as they should be better described. There are intrinsic quality measures in the sequencing files that are used to assess the quality of the sequence reads. BG/PD reuses datasets from different publications and merges them to answer his/her research questions. A lot of this storage is done on a shared network disk and processed on a local desktop. BG/PD says that there are no standards and no description of the data that s/he downloads, which imply to guess the meaning. BG/PD says that for this reason, it is needed to contact the authors who will generally provide the requested information. There is a high turnover of undergraduates and graduates, which means that there is sometimes no follow-up of projects. BG/PD suggests that information should be provided about available code or workflows, indicating that the material works and has been verified by peers (e.g., a stamp).

In a laboratory of **Stem Cell Biology**, the interviewee SCB/PD is a bioinformatician. Although the bioinformatics unit is shared between different groups, there is no appropriate structure or organization of the scientific data. Sequencing data processed by this unit is generated externally. According to SCB/PD, a central repository should be developed to structure this sequence data and identify its location from the start of the data generation process. The available meta-data annotation is considered as weak. Also, bioinformaticians in this group are perceived as

being hesitant to make their code available as its quality might be judged as not being up to standards by peers.

In the Department of **Pediatric Oncology**, the respondent PO/PI took the lead of a newly created research group, which leaves any data management issues open at the time of the interview. A custom-made laboratory information system (LIMS) manages micro-array data, and they seek to develop the same system for sequencing data. The identification of what data can be stored or dismissed is still an open question for which our interviewee believes that better and automated meta-data collection is critical. Regarding data reuse, one scenario is to re-purpose data that was used initially as quality controls.

The role of the interviewee in **Medical Microbiology** (MM), MM/PD, is to establish a bioinformatics pipeline and a private repository to make these datasets findable. Keeping the data consistent is an issue, as illustrated by a legacy issue that occurred when laboratory members in MM moved old files without any identifier assigned to the new repository. According to MM/PD, MM has a conservative attitude regarding data sharing that might evolve with the younger generation.

Table 5.3

Overview of the Interviewees' Domain, Role and A Summary Of Reproducibility Challenges

Domain	Role	Identifier	Reproducibility challenges
Computational Structural Biology	PI	CSB/PI	Expertise required to evaluate data quality
Biomedical Genetics	PI, Post-doc	BG/PI, BG/PD	Absence of standard descriptions of remote data
Stem Cell Biology	Post-doc	SCB/PD	No shared infrastructure between laboratories; Weak meta-data annotations
Pediatric Oncology	PI	PO/PI	Absence of data preservation strategy
Medical Microbiology	Post-doc	MM/PD	Moving data between (legacy) systems. Conservative attitude towards data sharing
Metagenomics	PI	MG/PI	Data is depending upon a range of (online) databases which might not be adequately documented

In **Metagenomics**, the constituents of a biological sample are unknown, and the goal of the analysis is to identify from which organisms the sequenced genomes are originating. A single sample might, therefore, be processed by calling different genomic reference databases to annotate

this material. Still, MG/PI found that the available raw data has poor meta-data description, which evaluates data quality and further processing laborious. MG/PI explained that data sharing is widespread in his field and that data reuse is common for answering new research questions, but not for verification purposes. Intermediate processing products (e.g., files) are not preserved.

5.6 Structure: The Scholarly Communication Infrastructure

The scholarly communication infrastructure has, besides researchers, stakeholders governing and managing the communication of scholarly work. Funders posit requirements to researchers before and after a project, mostly via data management planning. Publishers act as the central governing bodies of science communication. We first introduce tasks as they ought to be conducted (i.e., policy view) and the screening of publications, in Section 4.2.2, to show how things are done (i.e., "real world" view).

5.6.1 Funders and Publishers

The documents we consulted from funders and publishers listed in Section 3 resulted in a classification of several RDM tasks that researchers are expected to complete for getting funding granted on the one side and publishing in journals on the other side. The main tasks which are reported by funders and publishers are shown in Table 5.4. We operated a division between the two main objectives of RDM: facilitating efficient preservation and dissemination. Besides, remaining RDM tasks support data management planning tasks, such as sending data management plans to funders and reporting reproducible results.

Table 5.4

Activities Extracted From Policies of Funding Agencies and Academic Publishers

Task	Origin of Policy	Exemplary Quotes from Policies
Report reproducible results	Publisher	"Authors of research articles in the life sciences, behavioral & social sciences and ecology, evolution & environmental sciences are required to provide details about elements of experimental and analytical design that are frequently poorly reported in a reporting summary" – Nature publishing group
		"Data, methods used in the analysis, and materials used to conduct the research must be clearly and precisely documented and be maximally available to any researcher for purposes of reproducing the results or replicating the procedure." - eLife
Conduct data management planning	Funder	"A data management plan that must be submitted after the proposal has been awarded funding. The approval of this plan is a prerequisite for NWO disbursing the grant." – Netherlands Organisation for Scientific Research (NWO)
Elaborate preservation strategy	Funder	"Which facilities (ICT, (secure) archive, refrigerators, or legal expertise) do you expect will be needed for the storage of data during the research and after the research? Are these available?" – Netherlands Organisation for Scientific Research (NWO)
Elaborate dissemination strategy	Publisher, Funder	"Before manuscript submission, the Authors must deposit the underlying data to an appropriate public repository for public release scheduled no later than the publication date of the article." – Oxford University Press (OUP)
		"Such applicants are expected to contact IC program staff prior to submission and are also expected to include a data-sharing plan in their application stating how they will share the data or, if they cannot share the data, why not" – National Institutes of Health (NIH)

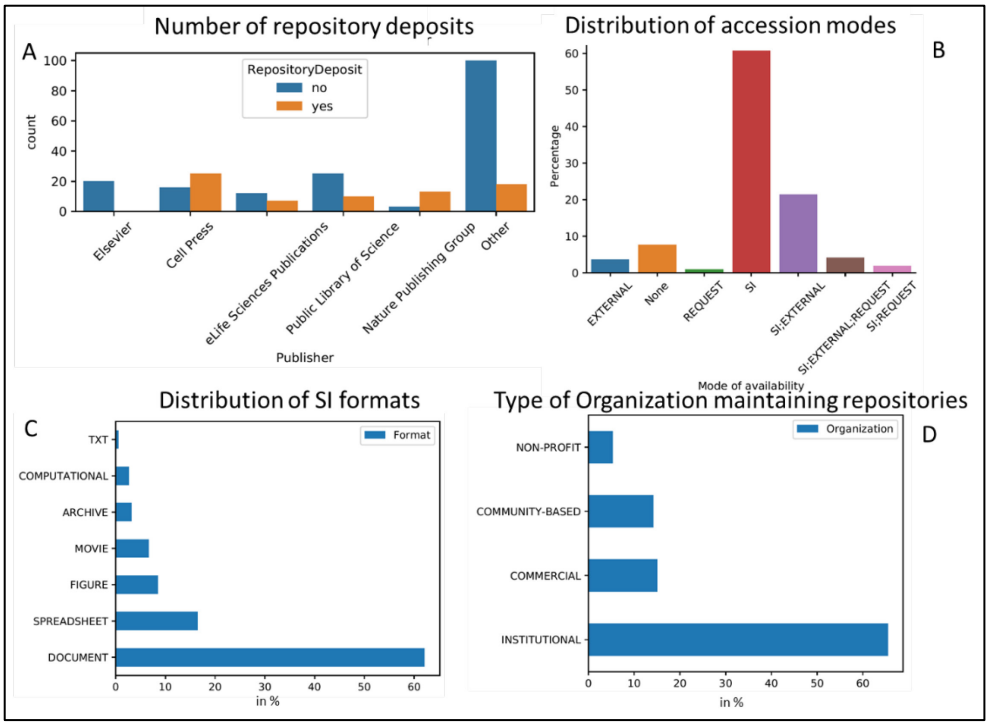
5.6.2 Publications

In addition to publishers and funders' policy, we seek to collect evidence about data sharing practices from screening scientific publications. In other words, we dive specifically into the tasks of reporting reproducible results and observe the consequences of dissemination strategies deployed by publishers.

The results of the screening are shown in Figure 5.1. The analysis illustrates shortcomings in dissemination strategies in terms of the use of digital repositories (Figure 5.1.A), modes of availability (B), available file formats (C) and, type of organizations maintaining repositories.

Figure 5.1

Results of the screening of scientific articles published in 2017 in the category Biochemistry, Genetics, and Molecular Biology (BIOC). Data source exported from Scopus.



In Figure 5.1.A. we observe that most of the screened articles do not refer to any deposited material. The low percentage of deposits is surprising considering that all sampled articles report on experimental work, thus with data acquired or produced. Moreover, Figure 1A shows that few publishers can invert this trend. Cell Press and Nature publishing groups host more prestigious outlets with a more extended history of attempts to improve reporting and data availability, which might explain why these publishers are more successful at convincing authors to deposit data.

In Figure 5.1.B, supplemental information is preferred as an alternative to repository deposits. Supplemental information (SI) files are hosted on the publisher's servers. A limitation of this mode of availability is shown in 1.C, where most of the information available in SI is not in the

original formats. Documents (i.e., PDFs, word documents) and spreadsheets (i.e., excel workbooks) are popular file formats. Original file formats that might prove useful for reproduction purposes, such as computer code, are seldom made available.

Last, Figure 5.1.D shows that authors privilege repositories that are established in their communities and hosted by renowned organizations such as EMBL. This choice might also be guided by the RDM policies of publishers, which mandate authors to deposit material in these types of repositories.

5.7 A Socio-Technical Framework of Reproducibility Threats

In this section, we introduce relevant concepts to decipher the implications on the reproducibility of RDM practices reported in Section 5.4. We also illustrate the treats with the help of Figure 5.2.

5.7.1 Dimensions of Reproducibility

A shared understanding of the concept of reproducibility in scientific research is that reproducibility refers to the capability of re-enacting previous studies. When experimental systems acquire results, the complexity of re-enacting the objects, procedures, digital analysis, and theoretical descriptions calls for a division in terms of "what is reproduced?" Radder (1992), divides types of reproducibility in terms of who is reproducing and what is reproduced. Here, we opt for a slightly different division to consider the many levels at which reproducibility issues might occur in experimental science where laboratory and computer work are combined. In the end, five dimensions are retained:

First, **Phenomenal and technical reproducibility** apply to local laboratory work (Tabb et al., 2010). There are fundamental experimental techniques which ensure that results obtained from the instruments are accurate and biologically sound. Therefore, biological, and technical reproducibility involves several defense mechanisms such as producing data in du/triplicates and comparing measurements on an object of study to positive and negative controls. BTR ensures that the experimental conditions at one location are well set. For instance, those instruments are calibrated and that observations did not occur by chance.

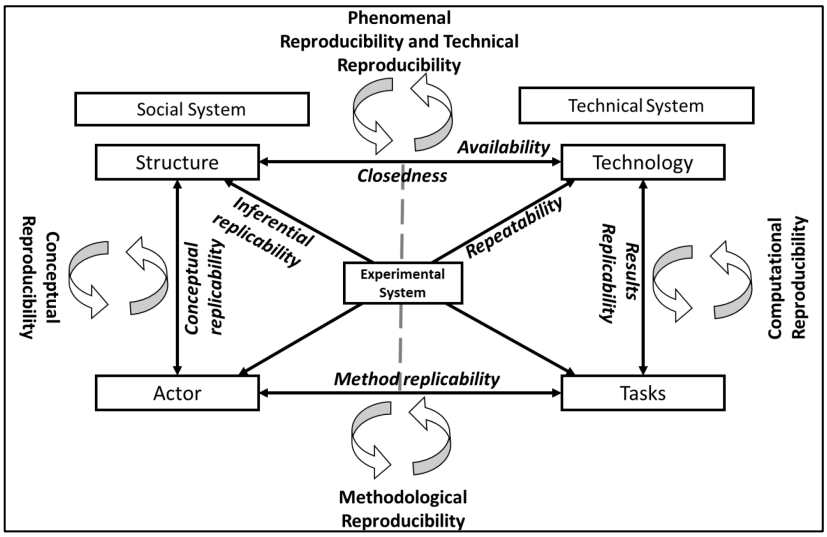
Next, **Computational reproducibility** is a concern when computer software and hardware are used to generate, process, and analyze scientific data (Peng, 2011; Freire, Bonnet, and Shasha, 2012). As we exposed earlier, computing technology is pervasive in modern scientific experimentation. CR is quite diverse in scope. Among many requirements for achieving CR, we

can name a few here: the availability of data and code (Peng, 2011), versioning and logging (Sandve et al., 2013), and the use of literate programming (Knuth, 1984). In short, biological, technical, and computational reproducibility guarantee the robustness of experimental operationalizations. They do not contribute to evaluating if the same results hold under different (experimental) conditions at different locations; this is the role of methodological and conceptual reproducibility (see below). Although CR is crucial for verification and reuse of computational work, it is not enough for evaluating the stability of results as CR focuses on making computations repeatable and reusable.

Finally, **methodological, and conceptual reproducibility** are forms of reproducibility that apply (also) outside the boundaries of a laboratory. At that level, reproducibility is an integral part of the scientific method (Andersen and Hepburn, 2016). Methodological reproducibility (MR) aims at testing the rigor of experimental designs or the stability of experimental outcomes at different points in time and space. Some authors make a clear distinction between methodological and conceptual reproducibility by stating that one can assess the stability of results by applying identical methods on new data or test a similar theoretical framework using new methods (Niederman and March, 2015). Thus, these two types of reproducibility are employed to challenge published theories and results. This is differing from the first three types, which can only say something about the rigor and robustness of data analysis pipelines and laboratory procedures.

Figure 5.2

A Socio-Technical Framework of Reproducibility



5.7.2 Threats to Reproducibility

Previous work by Schloss, (2018) and Goodman et al., (2016) show that reproducibility corresponds to a diversity of threats that occur when an independent team wants to reproduce results. Risks are too often that material is not available nor preserved in optimal conditions. To increase the understanding of the relations between reproducibility and its related threats, risks, and RDM vulnerabilities, we comment hereunder on the relations depicted in Figure 2.

First, we can represent **inferential replicability** as a **Structure ↔ Task** relation. Inferential replicability belongs to the conceptual reproducibility dimension as it relies on the scholarly communication infrastructure on the one hand and the capability to derive the (experimental) tasks that are conducted by the experiments to reach similar conclusions. **Conceptual Replicability** differs from inferential replicability; it is represented as the **Structure ↔ Actor** since it depends on the capacity of the same or alternative (team of) experimenters to reach similar conclusions based on the description of experimental work. Last, the extent to which artifacts are made available by the authors of a study, which we labeled **availability**, is represented the relation **Structure ↔ Technology.** In short, these relations all rely upon the availability of artifacts and detailed reporting of results.

Second, the dimension of computational reproducibility involves the results of replicability and repeatability. The difference between these terms can be explained by referring to the relations depicted in Figure 2. **Results Replicability** is ensured when **Technology ↔ Task** yields consistent results at each run, independently of the fact that the original or another (team of) experimenter(s) use these digital artifacts. Results replicability is facilitated by the publication of re-usable data and software in digital repositories. There is a subtle distinction with **Repeatability**, **Technology ↔ Actor**, as guaranteeing that technology yields repeatable results is the responsibility of the original team of experimenters.

Third, **Method Replicability** is a **Task ↔ Actor** relation, which means that actors (i.e., experimenters) can replicate studies by following the same procedures (or tasks). Methodological reproducibility is only possible in case the experimental system is closed. So, the **closedness** of the system implies that the two sides of the experimental system (social and technical) are consistently communicated and operated (hence closedness is depicted as the bridge between social and technical systems).

Finally, a summary of the threats to reproducibility and their associated risks is shown in Table 5.5. The associated risks and vulnerabilities are compiled from the interviews and the

screening of publications. Also, the terminology retained for classifying reproducibility threats, risks, and vulnerabilities are derived from the ISO standard ISO/IEC 27000:2016, which provides a shared vocabulary for information security. Risk is defined as the "effect of uncertainty on objectives" (2.68). A vulnerability is a "weakness of an asset or control that can be exploited by one or more threats" (2.89). Finally, a threat is the "potential cause of an unwanted incident, which may result in harm to a system or organization" (2.57).

Table 5.4

Overview of threats to reproducibility

Label	Threat	Risk	RDM Vulnerability	Example
Inferential replicability	External experimenters want to use findings from a published study to reach similar conclusions	No relation between reported findings and underlying artifacts which is a risk for Inferential replicability	No (or weak) reproducible reporting	Researchers in biomedical genetics (see interviews) who do not reuse workflows due to poor documentation. (Source: Interviews.)
Method replicability	External experimenters want to produce or acquire new evidence based on published resources and procedures	Underlying artifacts are not disseminated in their original formats which is a risk for method replicability	No (or weak) dissemination strategy	Researchers are reluctant to share code due to their quality perceived as inadequate. (Source: Interviews.)
Result replicability	An independent team wants to conduct an exact or partial evaluation of initial results as reported by the authors using similar (or identical) artifacts (e.g., software).	Custom code, workflows are not available in the laboratory and outside the laboratory which poses a risk for result replicability	No (or weak) dissemination strategy and preservation strategy	Very few computational artifacts are attached to publications. (Source: Screening.)

Label	Threat	Risk	RDM Vulnerability	Example
Closedness	The team of experimenters wants to isolate experimental results from interferences between experimental events and external (uncontrolled) events.	Software versions and computational workflows not preserved which is a risk for closedness as different software versions might give differing results	No (or weak) preservation strategy. No (or weak) reproducible reporting	New laboratories have no systems in place yet to trace experimental processes. (Source: Interviews.)
Repeatability	The team of experimenters wants to obtain similar results by applying the same routines and procedures (i.e., operationalization)	Poor management of software, data and lab notebooks are a risk for repeatability	No (or weak) preservation strategy	Significant turn-over and no follow up on projects. (Source: Interviews.)
Conceptual replicability	External experimenters produce or acquire new evidence to evaluate existing theories.	Experimental conditions not sufficiently described is a risk for conceptual replicability	No (or weak) reproducible reporting	A majority of artifacts are not deposited on curated repositories. (Source: Screening.)
Availability	Make the evidence underlying a preliminary report (i.e., a scientific article) available to readers for further verification and reuse.	Poor planning, versioning of artifacts, sharing habits, unforeseen privacy issues are a risk against the availability of artifacts	No (or weak) research data planning	Few projects consistently plan data management. (Source: Survey.)

5.8 Discussion and Limitations

We presented an approach to reproducibility, articulated in dimensions and threats, which help to categorize the different challenges of reproducibility encountered in experimental sciences. To the best of our knowledge, no such conceptualization of reproducibility and data management has been previously suggested in the literature.

We have introduced an initial framework to answer the question: "What are reproducibility threats occurring in experimental systems stem from vulnerabilities in research data management?". We have seen that preservation, dissemination, planning and reporting practices, standards in experimental science vary per domain (see interviews), and publishers (see publication screening). We worked towards a framework to capture these elements and position them according to reproducibility risks, threats, and RDM vulnerabilities. By doing so, we depict reproducibility threats and RDM vulnerabilities by considering (1) researchers and other stakeholders (2) introduce challenges experienced by the researchers (3) seek to grasp how these challenges translate into the scholarly communication infrastructure. To achieve that, we bridged a gap between experimental systems and socio-technical systems as successful reproduction of experimental work rely upon factors beyond technology.

Besides, reproducibility mechanisms, as depicted in the framework (Figure 5.2) goes beyond scientific experimentation in natural sciences. For instance, in IS research, design science research (DSR) is confronted with similar issues regarding transparent reporting and dissemination of reusable artifacts (Gleasure et al., 2012; Iivari et al., 2018). While some authors cast doubts on the applicability of terms such as reproducibility on a DSR paradigm (Baskerville and Pries-heje, 2016), the challenges experienced during artifact design and experimentation are similar from the perspective of working scientists. A dynamic view on the production of artifacts requires another type of sources than interviews and reports. A suggestion is, therefore, to include laboratory forensics (LF) findings into the current reproducible framework. According to Lefebvre and Spruit (2019), LF adds a perspective from practice by investigating digital files on storage systems and offer insights on RDM vulnerabilities going beyond what can be obtained from interviews and the study of publications alone. Another suggestion is to pursue the evaluation of experimental artifact reusability similarly to evaluation criteria for the reuse of design principles (Iivari et al., 2018), where experimental artifacts and their descriptions in method sections are not only evaluated by their accessibility but with a broader range of criteria such as appropriate guidance and effectiveness of disseminated experimental artifacts.

There are several limitations to our study design and findings, which we elaborate on here. The first limitation is that we conceptualize reproducible for experimental sciences with data covering only a limited sample of experimental scientists in biomedical science. However, we attempted to mediate this narrow view on experimental science in our study by adopting a more general view on scientific experimentation with experimental system theory. Similarly, to the limitations of our interview data, other disciplines might reflect other data sharing and usage patterns that the patterns we found in biomedical disciplines (Gregory et al., 2018).

5.9 Conclusion

A socio-technical approach on experimental systems highlights the dynamics of scientific experimentation from the point of view of working scientists (i.e., the actors) operationalizing experimental design (i.e., the tasks) using laboratory instruments and computers (i.e., technology) to communicate novel findings on the scholarly communication infrastructure (i.e., structure). We believe that understanding RDM practices and reproducibility challenges depends on the capability to frame experimental work in all its dimensions.

However, from the survey, interviews, and screening of publications, we saw that dissemination and preservation strategies are challenging to implement. RDM deals with the fragmentation of policies, ad-hoc data governance in laboratories, and few constraints put on systematic and structured sharing of computational resources in publications. Our results show that reproducibility risks need to be better understood to redesign the research data management and scholarly communication infrastructures effectively.

Section 3. Technology for Open and Reproducible Research

Chapter 6 | Reproducible Experiments: Developing Interactive Reproduction and Re-Use of Experimental Resources with Research Objects

Calls for more reproducible research by sharing code and data are released in a large number of fields, from biomedical science to signal-processing. At the same time, the urge to solve data analysis bottlenecks in the biomedical field generates the need for more interactive data analytics solutions. These interactive solutions are oriented towards wet lab users, whereas bioinformaticians favor custom analysis tools. In this position paper, we elaborate on why Reproducible Research by presenting code and data sharing as a gold standard for reproducibility misses significant challenges in data analytics. We suggest new ways to design interactive tools embedding constraints of reusability with data exploration. Finally, we seek to integrate our solution with Research Objects as they are expected to bring promising advances in reusability and partial reproducibility of computational work.

This work was originally published as:

Lefebvre, A., Spruit, M., & Omta, W. (2015). Towards reusability of computational experiments capturing and sharing research objects from knowledge discovery processes. IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 1, 456–462. https://doi.org/10.5220/0005631604560462

6.1 Introduction

Over the last few years, calls from researchers defending better data and code sharing for computational experiments (CE) are propagated in high-ranked journals (McNutt, 2014; Peng, 2011). Usually grouped under Reproducible Research (RR), these invitations elevate reproducibility or replicability as a central key of the scientific method. One of the interpretations presents reproduction as an application, by independent researchers, of identical methods on identical data to obtain similar results, whereas replication is similar except that different data is selected. According to RR proponents, benefits would be numerous.

First, reproducibility is a prerequisite for verifying the results of a published study (Peng, 2011). Second, for reusing previous work and build new knowledge. While the latter brings a constructive and enriching dimension to reproducible science, the first one is oriented to alleviating scientific misconduct, particularly in Life Sciences (Laine et al., 2007a).

Even though RR proponents are focused on suggesting exchanging code and data as a minimum threshold for "good science," they do not examine the methods used or people participating in CEs. Methods are not of interest to RR as the main focus lays on getting similar results for verification. Hence, the end product of a CE is seen as a script or package that should be made available by the authors of a paper as supplementary material.

The issue investigated in this work emerged from three phenomena: (1) the notorious increase in data generation and resource-intensive analytics. Here in the biomedical domain, (2) ignorance about data generation processes and their impact in terms of modeling. For instance, the sequencing instruments and custom bioinformatics pipelines producing analytical data and how well they represent underlying biological facts and (3) non-specialists, not trained in data analytics, eager to participate in computationally intensive experiments but preferably via convenient end-user interfaces instead of custom scripts or programs (Holzinger et al., 2014).

The phenomena described above were observed during a design science research (DSR) (A. Hevner and Chatterjee, 2010) we conducted in the domain of biomedical genetics. Our research was focused on designing an interactive data mining tool for biologists to identify interesting outliers in RNA-Seq count tables. Ultimately, the goal is to seek how to facilitate access and how to reuse scripts and packages for bioinformaticians and biologists at the same time. After one design cycle of a technical artifact and its evaluation by three focus groups gathering biologists and bioinformaticians (n=15), we collected evidence against some practices proposed by RR and suggest potentially fruitful improvements. Indeed, the reproducibility of CEs should not be reduced

to code and data sharing as it does not cover the fundamental characteristics of modern data analysis in biology. We state that web resources and their support for multiple representations that satisfy the interest of both types of users involved will have a positive impact on reproducibility by facilitating reusability first.

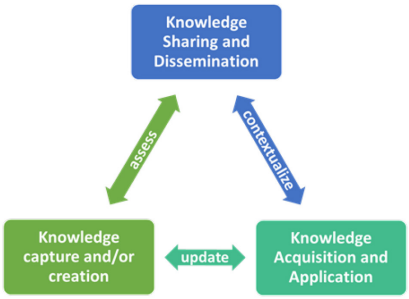
6.2 Background

Two aspects of knowledge creation and sharing are presented. Together, they clarify what issues emerge from code and data sharing when all stakeholders involved in a CE are not considered. We make use of a standard knowledge cycle, the Integrated Knowledge Cycle (IKC) (Dalkir, 2005), to emphasize the issues of codification implied by Reproducible Research. In knowledge management, codification aims at making implicit knowledge (from an individual) available as an object that is separated from the individual (Hislop et al., 2018). This can also be seen as the goal of RR, which distributes experiments as packages.

The IKC is illustrated in Figure 6.1. We focus our discussion on the knowledge capture and creation and knowledge sharing and dissemination phases. The last phase acquisition is not discussed here as we believe it to be the role of academia or industry in general.

Figure 6.1

Integrated Knowledge Cycle With Three Stages (Dalkir, 2005).



6.3 HCI-KDD

We start with Human-computer Interaction (HCI), which is the "study of how computer technology influences human work and activities." (Dix, 2009). Knowledge discovery from databases (KDD) is defined by Fayyad as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al., 1996).

The first aspect is that an end-user should be able to analyze data by using steps from the knowledge discovery process but interactively. This combination of KDD and human-computer interaction was theorized by (Holzinger, 2013). Tailored to the biomedical field, the process emphasizes that an end-user needs powerful visualization tools as much as data management and analytics capabilities. Holzinger also stresses the fact that reproducibility should be investigated further as it represents a significant problem with data-intensive experiments (Holzinger, 2013).

The steps of the HCI-KDD are integration, pre-processing and data mining. Integration is the activity of merging structured or unstructured data sets. Pre-processing applies normalization or transformation techniques to make the data sets suitable for data analysis. Data mining is the design and application of algorithms to identify patterns, associations, or outliers.

6.4 Reproducible Research

The second aspect is the need for better reproducibility of experiments that are conducted with computers. Here we integrate notions belonging to two approaches to reuse context and computational material.

On the one hand, based on literate programming (Knuth, 1984), dynamic documents (Pérez and Granger, 2007) and compendiums (Gentleman and Lang, 2007) constrain design choice to add human and machine-readable context to executable code. Compendiums aggregate dynamic documents. Dynamic documents are executable files that contain code with descriptive information. They are currently available with authoring packages in R (Knitr, Sweave) or Python (IPython notebooks, Jupyter).

On the other hand, an ontology-based approach for the dissemination of reusable components is assured by semantically enriched objects aggregating resources about the context of an experiment and its material. These are called Research Objects (RO) (Bechhofer et al., 2013).

6.5 Discussion

As we noticed, the fact that one end-user deal with each step are, at least, a very optimistic view on data analytics. The HCI-KDD process implemented in our prototype was discussed among participants (see section 3.1). The questions were oriented to the flow of analysis and presence or absence of components (e.g., charts, packages, result tables, context…) in the interface. Additionally, a survey was answered by 11 respondents (n=11) about how they are dealing with data and Reproducible Research.

6.5.1 Focus Groups Result

Inside our three focus groups, we divide participants according to their main interests, i.e., bioinformaticians and biologists. For the first type of participants, bioinformaticians, a friendly user interface, is rejected. Scripts are preferred for analyzing data. Regarding methods applied, a participant indicated that a method is sometimes selected because "it works" and is not a matter of "hidden" assumptions. By assumption, we refer to prior knowledge of the state of the world embedded in packages or statistical models. Not being aware of them makes a package acting as a "black-box" with unknown consequences on the rest of the processing.

For the second type of participant, biologists, they estimated the presence of such methods as appropriate. The indications given on the website (package name, version, reference paper, running environment, and online documentation) are sufficient if kept up to date. The web interface offered the possibility to apply different methods on the same data set. This was judged as beneficial because the influence of a choice could be assessed by the user interactively. In that case, another concern raised by bioinformaticians is about the interpretation of results by users that would not be trained in statistics. Regarding reproducibility, the lab part of an experiment has substantial influences on the rest of the pipeline, and it is perceived as challenging to integrate into the tool. Efforts for improving reproducibility are welcome, but full reproducibility is impossible, as indicated by participants in the third focus group.

6.5.2 Code and Data for Verification

It is the view of Peng (2011) that executable code and data form a gold standard of reproducible research. We argue that these elements are not of interest to each vital type of stakeholder involved in a computational experiment. We may admit, though, that what the author tries to achieve is a minimal level of reproducibility for verification purposes. The idea is that a

reviewer would carefully inspect code-shared with a paper, e.g., as an R package on Bioconductor. With that package, the entire computational workflow is runnable and shows figures that are identical to their online or printed counterpart.

However, as even noticed by Peng (2011), papers validating previous work are rarely acclaimed by publishers who expect "new" knowledge to be submitted. This may be an explanation, while results from our survey showed a reduced interest in full replication on a scale from 1 (never) to 5 (always). The need for full replication has a Mode of 2 (Median=2). Partial replication did slightly better with a Mode of 3 (Median=3).

6.6 Reusability and Interactivity

Regarding Research Objects, they sometimes appear to be developed as external solutions or repositories. We will lose a significant group of researchers if the goal of an application is to manage research objects. Instead, the software application should produce resources that might be automatically aggregated in an RO. This is a transparent manner for users more interested in advanced visualization capabilities.

Therefore, we claim that Research Objects could be a hidden component of any interactive mining tool. By doing this, we encourage RO generation and usage without transforming such tools in a "reproducibility manager" for users interested in getting precious insights from their experiments. Exaggerating any requirement of RO management for these stakeholders will most probably result in a rejection of the entire application. This could be achieved by automatically extracting information from earlier processing stages and intermediate data sets in the analysis flow.

6.7 Resources and Representations

An exciting proposal in compendium design was the notion of a transformer. We present it in this work as the creation of representation (or view) from a single resource. A resource is an object of interest, whereas representation is a usable form of a resource that corresponds to the consumer's interest. We designate by consumers both human and machine readers or interpreters.

In the RO world, it implies to work on ontologies and machine-readable standards. For biologists, it means that a chart resource has to render a dynamic representation. We can imagine that after exchanging an RO, we find a data object resource and a chart resource. A chart shows the content of a data object as, for instance, a scatterplot. We expect an end-user to be willing to select parts of this scatterplot, zoom-in, or display labels. We also expect that this chart resource is identical to what was generated by a team of researchers that created this RO.

As we show in the next section, open-source technologies for visualization “as a resource” exist and are under massive development. They can create JSON or HTML/JS serialization of a chart resource while providing enough interactivity for end-users.

6.9 Solution

The evaluation of our prototype yielded limitations of both HCI-KDD and current practices defended by Reproducible Research. Hence, we suggest an improved knowledge discovery process embedding the HCI-KDD in an extended process named Reproducible Research-Oriented Knowledge Discovery in Databases (RRO-KDD). When conducting a DSR, four stages appear at each design cycle (A. Hevner and Chatterjee, 2010). The *problem specification* resulted from a literature review and meetings with experts in biomedical genetics. The other steps found in design science research are *Intervention*, *Evaluation,* and *Reflection.* Each of them is described in the next subsections.

6.9.1 Specification

The problem addressed in this work encompasses reproducibility and visualization for researchers in biology who are collaborating with bioinformaticians. As explained in the background section, computational experiments are not only conducted on the bioinformatics side of data analysis. Hence, an application enabling *self-service* data analytics for biologists has additional constraints. *Self-service* is understood as letting users perform analytics tasks without advanced knowledge of programming or statistical modeling.

6.9.2 Intervention

As the technical outcome of the DSR we conducted, a prototype was developed and deployed in a research lab for structural genomics at the University Medical Centre Utrecht (UMCU) in the Netherlands. The prototype started from the HCI-KDD process by implementing interactive visualization capabilities together with methods to pre-process and mining data sets. Pre-processing consisted of *normalization* and *transformation* of the table of counts generated by RNA- Seq technologies and tools. A table of counts has samples of patients in columns and a list of genes as rows (60 000 in the files used). This table is the result of a *bioinformatics pipeline*. Hence, analytical data is generated by various levels of data processing from raw DNA sequence quality checks to counting how many RNA fragments found in a patient tissue overlap a gene. Via the web interface, users start with these tables in a *virtual experiment* (gathering data and contextual information). Then a possibility is offered to *normalize* or *transform* data sets by calling packages from Bioconductor. Normalization is an essential pre-processing task to make samples comparable due to the presence of (technical) biases in the raw data.

6.10 Evaluation

Exploratory focus groups with biologists and bioinformaticians provided input for conducting additional iterations, similar to an agile approach. From requirements and discussions with specialists, a set of functionalities for KDD and visualization were implemented. The facet of RR was imposed as it was not a primary requirement from the field experts. Hence, design choices for RR were inspired by previously described literature about compendiums and ROs (Bechhofer et al., 2013; Gentleman and Lang, 2007).

Next, three confirmatory focus groups invited bioinformaticians and biologists to discuss the prototype and judge the applicability of the KDD steps implemented. We addressed the results obtained from the focus groups in section 3. These results are further processed in section 4.4. We present a design proposition, which is an outcome of the evaluation of the prototype. Furthermore, our design proposition covers architectural choices which are mainly grounded in web architecture.

6.11 Reflection

The lessons learned from our DSR are described in the RRO-KDD process. We processed the input of three confirmatory focus groups with 15 participants. We described the results earlier and elaborated on their processing further in the next section.

6.11.1 RRO-KDD Process

The RRO-KDD process depicted in Figure 6.2 is modeled with its related "deliverables" in a so-called process- deliverable diagram (PDD) (Weerd and Brinkkemper, 2008). Here, the elements of the HCI- KDD process are integrated with contextual and technological outputs. These outputs are directed to the reusability of previous experiment code, data, and methods. Below, we shortly describe the steps and deliverables:

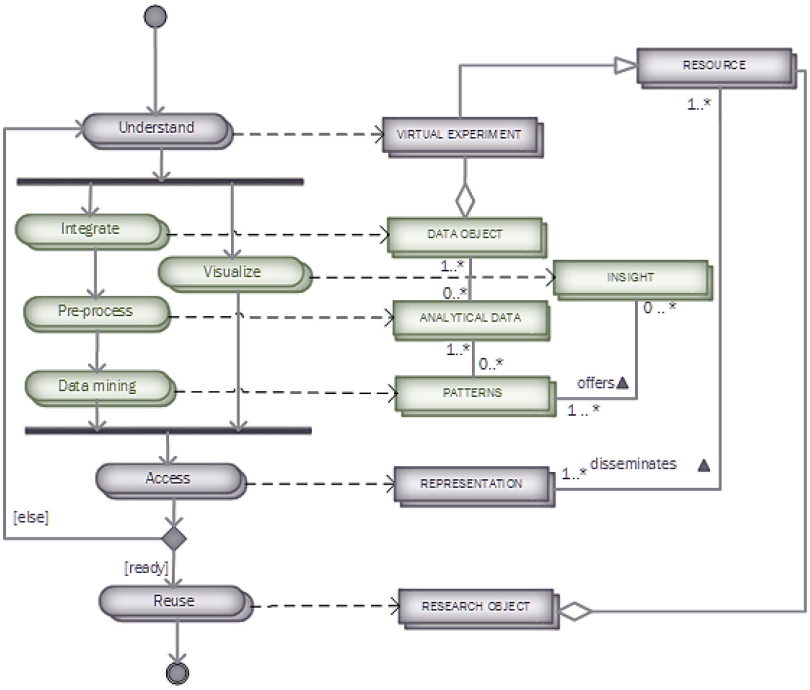
- **Understand** is an activity where a sufficient description of the data sets is provided. For instance, information about instruments, sequencing platforms, sample preparation. It builds a *container* for an experiment which is denoted by *virtual experiment*. Virtual experiments are uniquely identified aggregation of resources and group data sets together with context and methods.
- **Integrate, pre-process, and data mining** are the steps elaborated by the HCI-KDD process. Visualization is an activity that occurs in parallel to KDD and enables us to get an insight into what happens at each step. For instance, it helps the users to judge the

impact of pre-processing methods on the data set. Activity *Integrate* results in data objects, and *Preprocess* will normalize or transform these integrated data sets into analytical data, which are more easily interpretable than raw data, e.g., from sequencing instruments. Finally, data mining results find useful patterns from data, according to Fayyad's definition (Fayyad et al., 1996). Visualization is here a subpart of the whole HCI field of research as it was not extensively investigated in this work.

- **Visualization** has a deliverable called insight, which informs researchers on patterns, scores, or relations in their data in an interactive manner. Interactive plots were rendered with *bokeh*, a python library for creating browser compatible visualizations.

Figure 6.2

The Reproducible Research-Oriented Knowledge Discovery in Databases (RRO-KDD) Process.



The Access feature presents components that are interactively created during an experiment (like charts and new data objects) as REST resources that might be accessed without the user interface via RESTAPIs.

These resources aggregated in a virtual experiment can be semantically enriched for reuse as ROs. This is made possible because each component is uniquely identified and accessible via a

programming interface. As an example, a *mining task* created by a biologist is reusable via an RO with its unique identifier. The code of the prototype is hosted on GitHub under MIT license and is available here: https://github.com/armell/RNASEqTool.

6.12 Conclusion

Our results suggest that reproducibility cannot be reduced to data and code sharing and that the field of biomedical genetics suffers from a lack of software solutions that are both satisfactory for bioinformaticians and biologists who are mutually engaged in CEs. There are overlapping data analytics practices but also serious apprehensions from bioinformaticians to offer such a type of application to biologists if they exceed data visualization.

Despite these concerns, we found that there is a gap to fill both in terms of data analytics and the reuse of previous work. As we have seen, biologists were more inclined to ask more visualization capabilities, whereas bioinformaticians expect a solution where scripting or custom data processing is allowed. A unique identifier of resources and platform-independent information exchange via REST enables this. Nevertheless, HCI alone for biologists is not satisfactory as they want to query data and compare the impact of different methods. These comparisons require pre-processing and mining.

Reusability of data, workflows, or parts of experiments seems to be more enjoyable for the two types of end-users, which evaluated the artifact than reproducibility.

6.13 Future Work

The suggested RRO-KDD is still in a design proposition phase that needs to be evaluated in other settings, and the interest in sharing Research Objects must be assessed. For this assessment, the mining tools have to be upgraded and provide more realistic possibilities to exchange and reuse virtual experiments and their components.

Also, extending the RRO-KDD to distributed systems will have similar problems encountered in previous studies and known as *workflow decay*. This issue still holds in the RRO-KDD context, which is built around web services and URLs that may be inactive after some time. Permanent Identifiers may moderate accessibility issues but not the support of data objects or remote implementations of analysis packages.

Recommendations to face these issues are integration with virtual environments or containers (e.g., Docker), dynamic documents, and proper data management solutions. More research on integrating virtual containers for the reusability of computational experiments for

bioinformaticians and biologists is needed. Dynamic documents generated by the tool could also play a role for bioinformaticians to understand what decisions were taken by biologists processing data via a user-friendly interface.

These investigations should be made by effectively combining HCI and KDD, as suggested by Holzinger. However, the multiplicity of actors, analysis tools, and techniques remains a significant challenge first for reusability then for reproducibility.

Hence, reproducibility arguments in literature should be replaced by better designs for reusability in IT solutions, at least for enhancing collaboration between bioinformatics and biologists. *Reusability* is broader than reproducibility as it enables *repurposing* of previous work and, in essence, *reproducibility*.

Chapter 7 | Towards Open Science Readiness

Recently, the topic of research data management has appeared at the forefront of Open Science as a prerequisite for preserving and disseminating research data efficiently. At the same time, scientific laboratories still rely upon digital files that are processed by experimenters to analyze and communicate laboratory results. In this study, we first apply a forensic process to investigate the information quality of digital evidence underlying published results. Furthermore, we use semiotics to describe the quality of information recovered from storage systems with laboratory forensics techniques. Next, we formulate laboratory analytics capabilities based on the results of the forensics analysis. Laboratory forensics and analytics form the basis of research data management. Finally, we propose a conceptual overview of open science readiness, which combines laboratory forensics techniques and laboratory analytics capabilities to help overcome research data management challenges in the near future.

This chapter is submitted for publication.

7.1 Introduction

Research data management (RDM) is a pillar of future developments in open science, and particularly with regards to the efficiency of data preservation, sharing, and developments of open infrastructure (Higman et al., 2019). Also, in information systems research, the opening of data to the IS community is a current topic of debate (Koester et al., 2020; Link et al., 2017; Wilms, Stieglitz, et al., 2018). One practical reason RDM gains traction is that experimental activities taking place in laboratories increasingly rely upon digital technologies (Huang and Gottardo, 2013). Furthermore, scientific observations themselves are the product of digital technology, as scientific equipment transforms measurements of the physical world into digital entities (November, 2012; Stevens, 2013). This trend is observed in diverse practices encountered in experimental work, e.g., from small science, where all research is conducted in a single laboratory, to more complex projects where scientists employ large-scale, distributed, computational infrastructure (Cragin et al., 2010; D'Ippolito and Rüling, 2019).

Consequently, research software, data files, algorithms, and workflows are widespread (digital) experimental resources. Besides, scientists create, exchange, preserve and share those resources using various channels such as digital files on storage systems, supplemental information sections integrated to publications, online repository deposits, or e-mail attachments, to name a few (Tenopir et al., 2011). To guarantee the re-usability of shared resources, academic publishers implement new guidelines for more transparent reporting and stress research data availability as a prerequisite to publication (Federer et al., 2018). Thus, scientific publishers operate on this matter, along with public funding agencies, to encourage proper research data planning and management to foster (or require) high-quality data dissemination of scientific data (Federer et al., 2018).

Nevertheless, beyond the efforts to manage experimental resources more efficiently lays a wealth of issues stemming from research data management and scientific communication (NAS, 2018). In the biomedical world, for instance, decade-long debates about the trustworthiness of results from lab experimentation and clinical trials pinpointed methodological issues and reporting issues, among others (Huang and Gottardo, 2013; Laine et al., 2007a). Methodological issues were found to vary from misapplications of statistics to poorly designed experiments (Moonesinghe et al., 2007; C. L. Williams et al., 2019). Reporting issues are the result of methodological issues (Ioannidis, 2018), and, more broadly, the lack of fit of the scholarly communication infrastructure to report on the results of activities and resources used in modern experimentation, such as the integration of results generated by computer scripts with scientific articles (Bechhofer et al., 2013).

Studies on data sharing and reproducibility in science are restricted to the analysis of research output, i.e., scientific articles and questionnaires administered to scholars in the forms of surveys and interviews (Adewumi et al., 2021; Federer et al., 2018; Sholler et al., 2019; Tenopir et al., 2011). On the one hand, reproducibility studies focus extensively on information technology development to mediate irreproducibility in defined research fields such as bioinformatics with Galaxy (Goecks et al., 2010) and reproducible software (Napolitano, 2017). On the other hand, studies attempt to give insights into the wicked ecosystem of technology and the practice of data publication (Leonelli, 2013; Sholler et al., 2019; Wilms, Stieglitz, et al., 2018). However, insights on research data in laboratories are incomplete, as scientific publications analyzed in reproducibility studies are curated presentations of experimental processes (Brinckman et al., 2019). Besides, research data has not yet been investigated from an IS perspective, which makes our understanding of the peculiarities of RDM scarce and lagging behind studies addressing data analytics challenges in the corporate world (Mikalef et al., 2018). At the same time, proper RDM practice can lead to improved information quality and, therefore, ease the way to re-use high-quality scientific data at a larger scale.

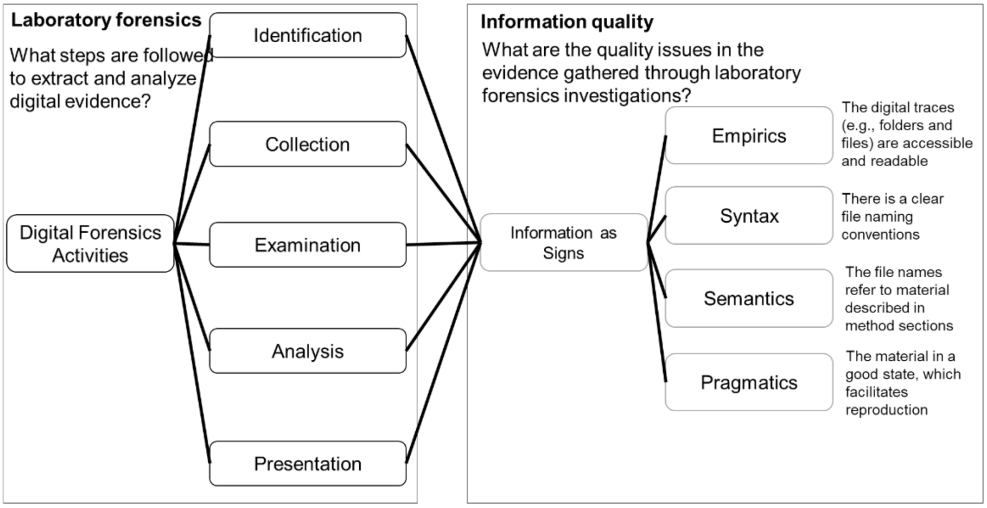
This is the reason why we elaborate here on an approach that enables the systematic extraction and analysis of experimental resources preserved on storage systems in laboratories. The approach we follow combines digital forensics techniques with information quality evaluation in laboratories named Laboratory Forensics, an approach analogous to digital forensics, an already established discipline (Palmer, 2001). By doing so, we aim at uncovering reproducibility issues stemming from data management practices in laboratories. Hence, our main research question is stated as follows: ***"How can a laboratory forensics approach help achieve open science readiness?"*** We propose to answer this question in the first phase of this study by investigating data management in one laboratory to (1) reconstruct the use of experimental data with digital forensics techniques and (2) evaluate the information quality of experimental data through the lens of the descriptive theory of information. Then, the second phase of our study (3) presents a proof-of-concept of an analytic dashboard which introduces and visualizes principles for designing technology that will help laboratories achieve open science readiness (OSR). Briefly, OSR is the laboratory equivalent to digital forensics readiness, a state of IT infrastructure in organizations that speeds up forensic investigations by implementing capabilities to trace (cybercriminal) events and audit information systems (Serketzis et al., 2019).

To further answer the main research question, we first need to gather knowledge about digital forensic methods and techniques that are readily available to extract information from

storage systems in a systematic way. Therefore, our study divides the problem of investigating laboratory storage systems into two parts, (1) the *design* of the laboratory forensics approach and (2) the *application* of the laboratory forensics approach to the evaluation of the quality of experimental artifacts managed by scientists in a laboratory. The former is presented in this article with the results obtained in a case study laboratory, where we systematically conducted forensic investigations in the lab and screened a subset of research data published by the same laboratory. The latter demonstrates how forensics results can translate to insights regarding information quality issues. This division between the development of the laboratory forensics approach and its application is illustrated in Figure 7.1.

Figure 7.1

The First Part of This Work Reports on (1) The Design of the Laboratory Forensics Approach and (2) the Application of Laboratory Forensics Techniques to Report on Information Quality Issues Using a Semiotic Perspective as Found in The Descriptive Theory of Information (DTI)



In the second phase of our study, we define several RDM capabilities that laboratories should consider in order for laboratories to gather evidence about research data management (RDM) practices. These RDM capabilities are devised from the results and lessons learned after our forensic investigations. Then, to illustrate the connection between RDM capabilities and open science readiness, we introduce an analytics dashboard demonstrating the use of RDM capabilities and their corresponding performance indicators.

7.2 Background

The proper management and sharing of research data underlying published studies is still a lively subject of debate in academia (Bajpai et al., 2019; Editorial, 2014; European Commission, 2016b; Freire et al., 2012b). Scientific communities, publishers, and libraries, among others, have concurrently developed numerous solutions to tackle the need for high-quality preservation and dissemination of research data and software (Borgman et al., 2016; Callahan et al., 2006). Furthermore, there are strong methodological incentives to improve data management in academia, as exemplified by the reproducible research movement that emerged more than a decade ago (Peng et al., 2006; V Stodden et al., 2014). More recently, the open science paradigm is perceived as a way of improving information quality in science through citizen science (Lukyanenko et al., 2020). Thus, the increasing number of initiatives to generate high-quality research data leads us here to investigate the difficulties experienced by researchers in documenting the research process using underlying technology such as digital file systems, remote servers, and digital repositories.

Moreover, the analysis of research data preserved in laboratories is an exciting starting point to explore research data further, including data sets and software that are not publicly available. The reason much information is not publicly available is that internal storage systems are meant for exchanging and saving operational data that researchers produce. The operational data created during scientific experimentation is thus not primarily aimed at being exchanged with external parties. Nevertheless, the investigation of operational data with a lens of information quality is at the core of the forensics approach presented here. By conducting forensics, we aim at reporting on the reproducibility of scholarly work uniquely, i.e., through the lens of an information systems theory grounded into semiotics. Previous work in information systems has extensively discussed the usefulness of the semiotic approach to the analysis of information in organizations (Burton-Jones et al., 2005; Stamper et al., 2000). Nevertheless, as noted by Lukyanenko et al. (2020), scientific organizations differ from corporate organizations. Typically, scientific organizations such as laboratories are much more dynamic, and data flows through several actors, processes and, purposes that are not directly relatable to data management in the corporate world (Borgman, 2015; Lefebvre et al., 2018; Lukyanenko et al., 2020).

Thus, in line with semiotics analyses applied to enterprise data integration for investigating data quality (Krogstie, 2015), we apply semiotics analyses to research data management. Experimentation processes produce the research data we analyze in this study. These processes leave a wide variety of (digital) traces from different types of (laboratory) resources. It

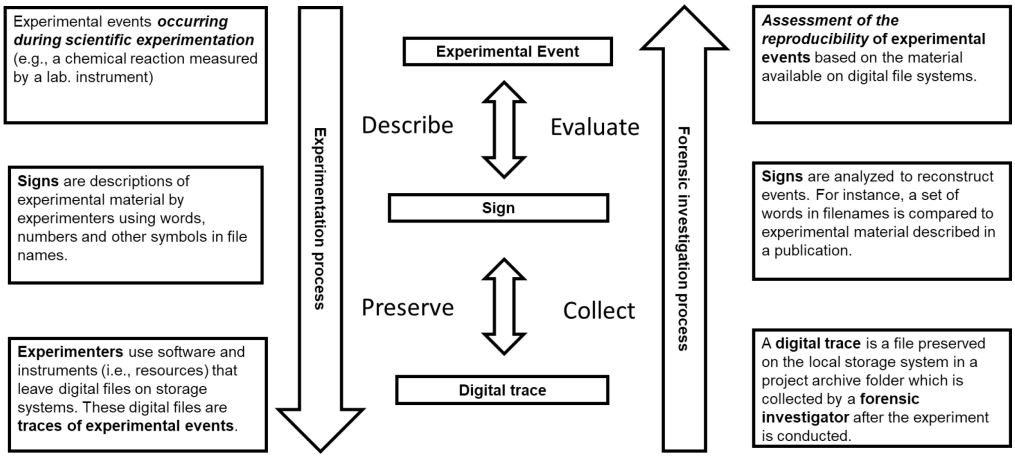
leads to the fact that the interpretation of experimental evidence is not straightforward. For instance, editorial, experimental, and computational processes are of a distributed nature and, therefore, combines the use of a variety of data management systems, software, and laboratory equipment.

7.2.1 Semiotics Perspective on Research Data

From an information point of view, reproducibility is achieved when the experimental materials involved in the experimentation process are located on the storage, systematically named with meaningful concepts that reduce room for interpretation and are adequately documented. In other words, our assumption is here that digital traces that are preserved in such a state that the empirical, syntax, semantic and, pragmatic facets of the information they contain are satisfactory. Forensics techniques are used to extract digital traces with meta-data from laboratory storage systems to judge whether these facets of information are of sufficient quality for reproducing experiments. Therefore, we provide some background about the experimentation process that leads to those digital traces and their interpretation with the DTI (see Figure 7.2).

Figure 7.2

A Comparison of the Use of Events, Signs, and Traces From the Perspective of Experimenters (Experimentation Process) and Forensic Investigators (Forensic Investigation Process).



First, the experimentation process corresponds to the activities, inputs, and outputs of experimental work in a scientific laboratory. Research cycle models are commonly used to represent such processes from the conceptualization of a research problem, the generation of data with instruments, their processing, analysis, and communication to outsiders (Cox and Tam, 2018). The

cyclic representation of experimental processes is emphasizing the re-use of previously generated data for new studies. As computers are involved in many (if not all) of these activities, it is expected to find (digital) traces on storage systems or even in other devices such as USB sticks or cloud storage. From a forensic perspective, experimentation processes are where digital traces originate from, independently of any research field, specific software, or storage architecture involved. The assumption is that experimental activities lead to files that are saved on a storage system. Undeniably, not all activities involved in the experiment are ending as digital files. Nevertheless, there cannot be reproduction without the presence of enough material to verify an experiment.

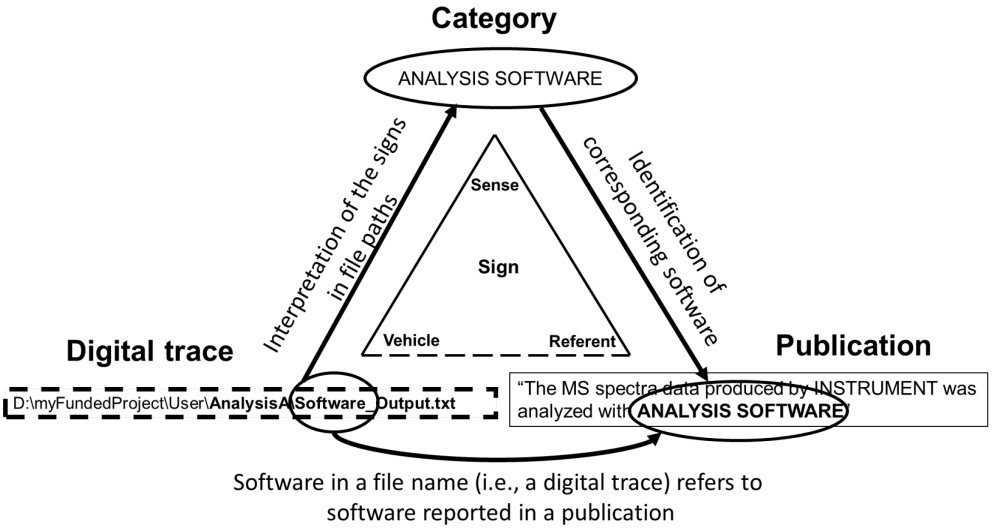
Next, the interaction of **forensic investigation** and the **experimentation process** is understood as follows: software and instruments involved in **experimental events** generate all sorts of **digital traces** found on storage systems during the forensic investigation (Lefebvre and Spruit, 2019b) . Thus, the purpose of a (lab) **forensic investigation** is to report on the information quality of digital traces left by experimenters conducting laboratory experiments. Moreover, the forensic investigation process involves the interpretation of information like **signs**, signs which are used by researchers to describe experimental resources used during scientific experimentation. Signs are elements in filenames such as the identifiers of a lab instrument and an object of study with the date of analysis written in a file name. The **preserved** material is meant for accomplishing the tasks relevant to communicate experimental results. Experimenters describe preserved material to accomplish their tasks. However, a **forensic investigation collects** these **digital traces** to accomplish something different, namely the **evaluation** of the reproducibility of scholarly work originating from the laboratory. In short, experimenters use signs to **describe material** for experimentation, and forensic investigators **interpret** those signs for reproducibility purposes. These two perspectives on the same material tend to provide a rich account of experimental events on the one hand and reproducibility issues, on the other hand. The former perspective is the perspective of an experimenter at work choosing concepts to name the material preserved on storage systems. The latter is the perspective of a third party that attempts to reproduce the experimenter's work.

As will be presented later, the forensic investigation leads to the interpretation of information signs discovered on storage systems in laboratories. There are several models of signs in semiotics, the triadic model of a sign (Klinkenberg, 1996; Nöth, 1990) being the model that conveniently illustrate, see Figure 7.3, the characteristics we investigate in digital files. The first notion is the notion of a vehicle of a **sign**, which corresponds to the digital traces (e.g., a file path). Vehicles are how signs reach their interpreter. **Vehicles** are, for instance, a language with their

written symbols or sounds. Then, the **sense** (or meaning) is an abstraction in one's mind occurring when signs are perceived. In our example, it is a class of objects such as the concept of software. Last, the **referent** is the object itself, for instance, the corresponding software (and version of that software) used to analyze research data reported in a publication.

Figure 7.3

The Interpretation of Digital Traces Depicted as a (Semiotic) Triangle of Ogden-Richards.



For this study, we make use of a descriptive approach rooted named the Descriptive Theory of Information (DTI) to evaluate several aspects of a sign. The DTI was first presented by Boell and Cecez-Kecmanovic (2015). In their work, the authors of the DTI elaborate on a generic approach to the description of the information and provides a critical review of definitions of the concept of information used in IS research (Wang and Strong 1996; Stvilia et al., 2007; Chatterjee et al. 2017). The DTI describes information according to two dimensions. The first dimension of the DTI articulates three different *forms* of information. Therefore, the DTI distinguishes intended information (i.e., stored) from potential information (i.e., potentially relevant to third parties), and information in use (i.e., as interpreted by third parties).

The second dimension of the DTI regroups the four conditions for a sign to be interpreted as information *by* someone. The first branch of semiotics retained in the DTI is named *empirics*, which is at the physical level of information and deals with how information is stored on physical systems. Second, the *syntax* is about how information is structured and obey to rules of a sign

system. Third, *semantics* are conditions of information to provide meaning to information consumers. Last, the *pragmatic* aspect adds dimensions such as interests and socio-cultural context to the previous categories. DTI Facets express each of these branches. Facets are a condition for a sign to become information. Boell and Cecez-Kecmanovic (2015) suggest 15 facets of information (e.g., novelty, physical assets) classified into the four semiotic branches defined earlier.

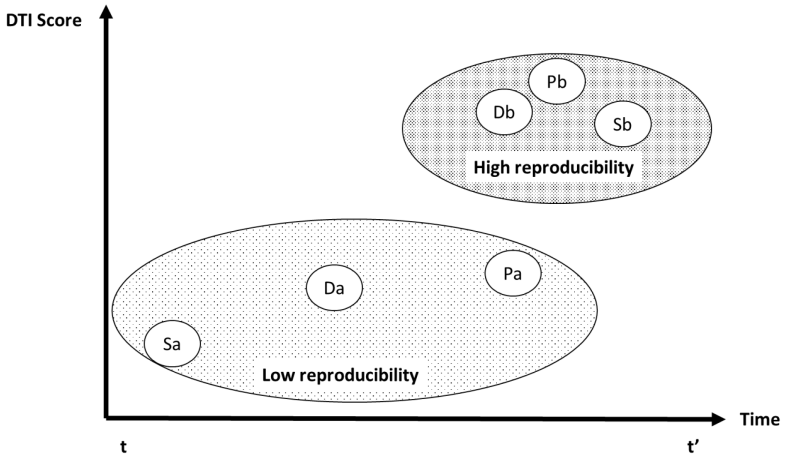
We observed that, in practice, at the stage of preservation, i.e., named intended information in the DTI terminology, a vast amount of research data resides locally and on organization-specific systems (Prost and Schöpfel, 2015; Tenopir et al., 2011). In laboratories, such as in the case study laboratory, the preservation of research data is set up employing shared folders. The digital file system provides basic meta-data structures. The generic architecture of digital file systems defines two types of meta-data: *system-dependent* and *user-defined* metadata (Venugopal et al., 2006). S*ystem-dependent* metadata are analogous to empirics according to the DTI and are focused on physical descriptions of *data objects*. *User-dependent* metadata, on the contrary, might potentially cover syntax, semantic, and pragmatic facets of describing the data in folders and file names. With a mix of both types of meta-data, an investigator can recover experimental resources and obtain knowledge about the time at which they were created as well as other features written in filenames.

The evaluation of experimental material leads to a score of each semiotic level, as can be seen in Figure 7.4. The higher the score on the DTI faces (DTI Score), the higher chance a future experimenter can perform analyses with the material preserved on the local storage or online. For instance, a publication (A) using a software (Sa) to analyze a dataset (Da) and its corresponding manuscript (Pa) are stored in distinct locations that are hard to access. The empirics score of A will be low (e.g., a score of 1 on a scale from 1 to 3). If we add unstructured and ambiguous names (syntax and semantics), as well as the absence of documentation on the workflow (pragmatic), the DTI score will be low, i.e., publication A scores low on the empirics, syntax, semantic, and pragmatic levels.

At a later time (noted t' in Figure 7.4), a laboratory conducted another experiment involving other experimenters who, this time, carefully described the experiment, chose filenames wisely, and kept informative hints about their experimental processes. In that case, we obtain a higher score for B (Pb) than for Pa. As we explained earlier, digital traces that are accessible, well-structured and, holding meaningful information lead to more reproducible experiments. Therefore, from an information point of view, a high DTI means a higher reproducibility potential. The process leading to such an evaluation and scoring with DTI is described later in Section 7.4.

Figure 7.4

The Laboratory Forensics Approach Should Result in The Assessment Of Digital Traces Encompassing Datasets (D), Software (S) and, the Publication (P) Employing a Score Standing for the Quality of Those Traces. For Instance, Here A and B are Two Illustrative Publications, Where A Scores Lower (Harder to Reproduce) Than Publication B as The Score of Its Components (Software, Data, and Publication) Score Lower.



The scoring mechanism forms the basis for developing analytics capabilities further. The scores indicate what type of issues are encountered. Information systems research has developed a wealth of useful capabilities, as introduced in Section 7.2, that can be placed in the context of laboratory work because laboratories are not entirely similar to the corporate context from which analytics capabilities where previously devised. Nevertheless, we can seek to connect analytic capabilities to the information aspects of the DTI that are scored on a level from 1 to 3, low to high, respectively. *Equation 1* shows the criteria for scoring each aspect.

$$DTI\ Score = \sum_{i=1}^{A} \frac{S_i}{A * C} \tag{1}$$

Equation 1 Scoring the information aspects on the laboratory storage and the associated repository. Using this formula, four aspects (A) are scored according to three criteria (C) each.

In Equation 1, the score is divided by 12 to obtain a final DTI score ranging from 0 to 1 after summing up the score of the four DTI aspects, namely empirics, syntax, semantics, and pragmatic. The score (S) is derived from the criteria in Table 7.1.. These criteria help the investigator evaluate the quality of the experimental resources according to their semiotic facets.

Table 7.1

Criteria for evaluating the investigated research data with empirics, syntax, semantics, and pragmatic branches of the DTI.

	Score	Scale	Criterion	Example
Empirics	3	High	All relevant files can be accessed and retrieved	The list of folders that contain documents, raw data, processed material, and other relevant material.
	2	Medium	A part of the files is still accessible on the storage systems; however, some files are not accessible	The raw data might be preserved, but the analysis output has not been preserved.
	1	Low	Some files are located but with low uncertainty and might not belong to the corresponding publication	The scientific data behind the publication is hardly accessible
Syntax	3	High	The structure of file and folder names is consistent in all project folders	The authors follow a strict convention to write file names.
	2	Medium	Files are partially structured	Parts of file names can be delimited by symbols such as – or _, which ease the interpretation of their content
	1	Low	No consistent structure in file names	Date and time in file names can be written in many formats, some of which are confusing, like a date value 02052020, which might refer to February or May
Semantics	3	High	Enough resources mapped with certainty to the corresponding publication	Groups of files are precisely matched to their role in the experiment helped by meaningful names
	2	Medium	Some resources mapped to corresponding publications	A part of the software or data in the method section can be mapped to the preserved resources
	1	Low	No (or a small number of) experimental resources mapped to corresponding publications	A list of figures is found on the storage, but no software output to generate them.

	Score	Scale	Criterion	Example
Pragmatic	3	High	Documentation present and folder structure is logical	A readme file is present, code (scripts), and relevant data sets are described, and the connection between parts of the article and its related resources is unambiguous.
	2	Medium	There is little information about how the resources can be (re)-used	The necessary resources are present but in formats that are not easily modifiable, such data in a spreadsheet with many annotations instead of simpler text files.
	1	Low	Few resources are reusable	A file named such as output.txt does not define which kind of output, when and how it was acquired

7.3 Research Data Management Capabilities for Open Science Readiness

Thus far, we have introduced semiotic concepts used to investigate digital traces found on storage systems in laboratories. In this section, we analyze prior work on RDM capabilities relevant to achieve open science readiness. The concept of readiness is borrowed from the digital forensic domain (Rowlingson, 2004; Serketzis et al., 2019). Forensic readiness is a state of technology that enables organizations to resist (or investigate) external threats, such as cybercriminal events, on their IT infrastructure (Simou et al., 2019). In the context of open science, many events can occur that require information systems in laboratories to be ready to deliver experimental evidence appropriately to (future) laboratory members, reviewers and comply with their research institution's policies.

In information systems research, authors have argued that the targeted use of (big) data analytics can reinforce the organizational capabilities of companies (Mikalef et al., 2018). Nevertheless, data availability is a prerequisite for the success of the Big Data enterprise in business and open science (Austin, 2019; Joubert et al., 2019; Sholler et al., 2019). The extent to which (big) data are in a state that can fulfill the ambitions of reinforcing organizational capabilities, support (national) policies for big data in businesses (Joubert et al., 2019) or the development of governance for open science, reproducible research and, and research evaluation (Austin, 2019). In all cases mentioned above, data quality (or veracity) is a significant factor in the success of big data readiness (Austin, 2019; Joubert et al., 2019).

Transposed to the laboratory domain where experimental work is conducted, we explore how the analysis of research data can reinforce capabilities to manage data in a reproducible way. In the literature, research data management capabilities revolve around the research data lifecycle

and the activities that process data at each step of the lifecycle, from creation to publication (Cox and Tam, 2018). For instance, the SEI CMM is a capability maturity model tailored for research data management that is oriented towards the production, preservation, and dissemination of high-quality research data (Crowston and Qin, 2011). The SEI CMM model lists four focus areas for RDM: (1) data acquisition, processing, and quality insurance; (2) data description and representation; (3) data dissemination; (4) repository service and data preservation (Crowston and Qin, 2011). Open science readiness is the state of a laboratory where RDM is implemented in a way that the transparency and reproducibility of scientific experiments are less challenging to achieve, i.e., do not require extensive forensic investigations to recover experimental evidence.

Furthermore, open science readiness relies on open data value capabilities. Zeleti and Ojo (2017) presented open data value capability areas as data generation, knowledge of data standards, knowledge of data value and, data strategy for generating open data. These open data value capabilities align with research data policies that share the ambition of disseminating high-quality research data using digital repositories, preferably openly or with few access restrictions (Amorim et al., 2015; Jones et al., 2012).

Therefore, we investigate what RDM capabilities can play a role in increasing the availability and quality of research data preserved and shared in laboratories. We first start by applying a digital forensics approach in Section 3. Then, in Section 4, we reflect upon our forensic findings by introducing key RDM capabilities that, once implemented, will decrease the number of challenges occurring when managing research data locally, in laboratories, and online, in publications and digital repositories.

7.4 Information Quality Evaluation With Laboratory Forensics

In this section, we explain the process of extracting experimental evidence from laboratory storage systems. The included publications, listed in *Table 7.2*, are from experiments that were conducted independently; the DTI scores reported in this work are showing the scores of different publications, authors, and years. According to Årnes, a DF investigation starts with the identification of the data sources or interest, which are possibly containing relevant material. The next step is the collection step, where the evidence from existing storage systems is extracted. The collection of evidence requires an image of the data source of interest, as it would be hazardous to investigate storage systems in use. Once the evidence is isolated from a computer device, we proceed with the examination phase to locate potentially relevant evidence. After the investigators have recovered potential evidence, the analysis phase takes place. The last step, presentation, is the

translation of the findings into a format that can be understandable by third parties, who may not grasp the legal and technical details of forensic investigations (Graves, 2013).

Table 7.2

Background Information of the Selected Publications Examined in This Study. The Data in This Study is Reported Anonymously. Only the Year, Journal, and Publisher are Communicated.

Publication Identifier	Year	Journal	Publisher
PUB_1	2019	Chemical science	The Royal Society of Chemistry
PUB_2	2016	Journal of the American Chemical Society	American Chemical Society Publications
PUB_3	2017	Analytical chemistry	American Chemical Society Publications
PUB_4	2017	ACS chemical biology	American Chemical Society Publications
PUB_5	2018	Journal of the American Society for Mass Spectrometry	American Society for Mass Spectrometry
PUB_6	2018	Journal of the American Society for Mass Spectrometry'	Springer
PUB_7	2019	Journal of proteome research	American Chemical Society Publications
PUB_8	2019	Journal of proteome research	American Chemical Society Publications
PUB_9	2019	Molecular & cellular proteomics: MCP	American Society for Biochemistry and Molecular Biology
PUB_10	2019	Analytical and bioanalytical chemistry	Springer

Hence, we followed a number of steps to achieve score the quality of the material underlying each publication, structured around digital forensics approaches:

1) The **collection of digital evidence** is, therefore, a basic set of activities. The output of the forensic investigation depends on the quality of the data sources that are gathered. The investigated digital evidence is produced by e**xperiments** where experimenters combine laboratory work with computational work to produce research results. Once the evidence is gathered and secured with a snapshot of file system meta-data, an **examination** phase follows. During the examination phase, we conduct further quality checks on the data acquired from storage systems.

2) Next, we proceed with the **analysis of experimental evidence**. Once the examination steps confirm the relevancy of the evidence, the selected traces qualify as relevant **experimental evidence** as we are confident at this stage that the traces belong to the experiments reported in the publication of interest. Typical forensic techniques that are applicable at the analysis stage are the production of timelines (where the date of modification of files are plotted together with other information, such as extensions or filenames (see Figure 5 A).
3) Last, we **present findings** as a report mentioning the number of relevant files found during the investigation, the total size of the experimental data, and the duration from the first creation data to the last modification. Besides, we comment on the quality of the material using the DTI to communicate, which issues are prevalent in the storage for each publication.

7.4.1 Identification and Collection of Research Data

In laboratory forensics, publications are used as a starting point for investigating the data disseminated together with the publication. Also, the search space on the storage systems is reduced to folders containing information about authors, methods, and software. The publications are extracted from PubMed. The selection conditions are (1) that a majority writes those publications of authors originating from the case study laboratory and (2) that a full-text version is available in PubMed Central (PMC) in XML format. The reason we opt for publications that can be retrieved in an XML format is to facilitate the extraction of meta-data and paragraphs in the articles.

Next, in the case study laboratory, access to the storage systems was granted by a laboratory member of the case study laboratory. The storage systems in use in the laboratory are remote storage servers, which are logically divided into raw data folders, laboratory computers, users, projects, libraries, groups, and personal folders. Files and folders were first inspected using the file explorer in Windows or PowerShell commands before snapshots were created. We opted for a pre-selection of relevant folders so that the process of copying files does not overwhelm the requests on storage servers, which are used by experimenters. Also, a pre-selection decreases the number of files ending in the snapshot.

The snapshot is preserved as a comma-separated value file (CSV) containing file paths (the location of a file on a file system), file names, dates of creation, modification, and last access. As the snapshot is a text file, it can be analyzed with text and natural language processing techniques during the examination and analysis phases. We used custom analysis software to assist in the

investigation. The path2insights – P2i - software is an analysis toolkit for investigating the content of file systems extracted as text (Lefebvre and Bruin, 2019). It, therefore, combines traditional forensic techniques (timeline creation, matching file extensions to software) with natural language processing techniques such as tokenization, distances (with Levenshtein distances). P2i offers a unified and comprehensive set of tools for analyzing file paths. P2I supports static file systems analysis without requiring access to the original physical storage. A scan of the storage's content exported as a text file suffices to explore the files preserved on the laboratory's storage system. Essentially, P2i brings foundational natural language processing techniques to the analysis of file paths. At this moment, P2i supports the tokenization, similarity, clustering of file paths to compare, and other file paths across different folders. For instance, a file name can be split into subparts so to compare these parts between folders and obtain a comparison of material preserved in different locations. Using a clustering approach, the content of different folders can be compared based on a subset of words (or tokens) extracted from the filenames.

7.4.2 Examination of Research Data

During the examination phase, the collected evidence present in the snapshot is checked and prepared for further analysis. At the end of the examination, unnecessary files are filtered out from the storage snapshot, and their inclusion in the snapshot is made certain. The decision is made based on the information reported in publications. So, experimental resources are identified from the publication and, if applicable, the location where the authors have deposited those resources. Here, we extracted nine concepts that occur in method sections of the publications (see Table 7.3). Besides, these concepts help the investigator detect the origin of resources and specific file formats, such as file formats that belong to laboratory equipment.

When we recovered traces containing signs (e.g., words) referring to software, for instance, we matched those resources to the category "software," as defined in *Table 7.3*. For instance, *proteome discoverer* (Colaert et al., 2011), a software used in proteomics, leaves particular patterns of files on the storage. Therefore, these files can quickly be recovered from their names and extensions, and hence can be mapped with confidence to the publication(s), which refer(s) to them. Nevertheless, in many cases, the evidence collected is not linkable to a publication with high certainty. Depending on the files and folders structure, trial experiments, tests, and other materials used for unpublished activities are confounded with the (relevant) material underlying a publication. In such a case, the DTI score has to be lower to reflect this confusion.

Table 7.3

The Nine Coded Categories Used for Annotating the Ten Articles Published by Our Case Study Laboratory

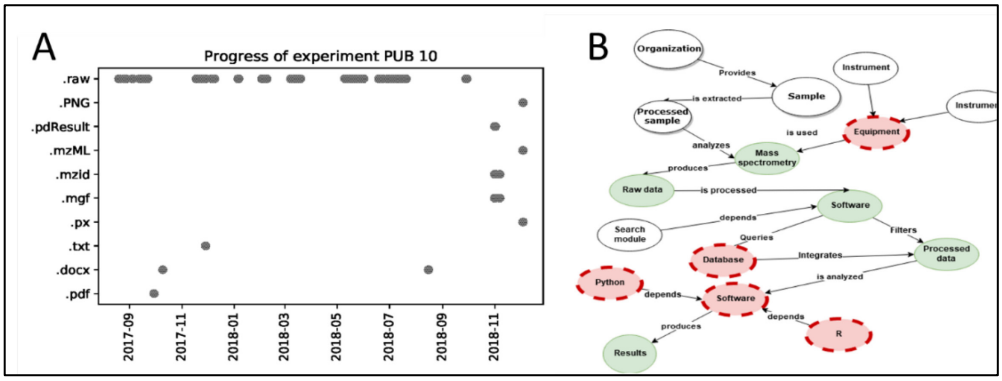
Name	Description	Occurrences in publication
Data	Mentions of the data created by equipment in laboratories or data analysis software reported in a publication	9
Database	A database is a collection of data which is searched/queried to obtain reference material or compare local results with known recorded outcomes	5
Deposit	A dataset or software is deposited in a repository (or website) which is publicly accessible (or with clear guidelines to access the material)	7
Equipment	Equipment groups, instruments, and lab material intervening in the process of experimentation	9
Location	A city or country where material, data, software, and equipment are originating from/manufactured.	6
Method	Laboratory and computational processes used to operationalize experiments.	9
Organization	A company, laboratory, institution, or any other group reported in the publication	7
Software	Similar to equipment but purely computational. Software refers to packages, scripts, analysis software, and so on.	9
Supplemental Information	The authors submit additional files on the editorial system and accessible directly on the journal's website. Supplemental information is referred to from the text.	7
Number of investigated articles (N)	The total number of articles investigated in this study	10

7.4.4 Analysis of Research Data

Once the digital evidence collected from online sources and local storage systems has been examined, as explained in the earlier section, we continue with the analysis of the evidence. The analysis step is where the analysis of information quality issues takes place. From the domain of digital forensics, one can re-use several techniques that help an investigator show when experimental events occurred with timelines and how the identified files fit into the experimental process with link analyses. Timelines are constructed using storage meta-data (which is only applicable to laboratory storage). The timeline in Figure 7.5 A shows the date of modification of files recovered in the laboratory for PUB_10. In the timeline, we can observe that there have been several moments where raw data has been produced for almost a year, with interruptions of a few months between measurements. Then, data processing occurred after the production of raw data, making the total duration of experiments reported in an article an effort longer than a year.

Figure 7.5

In A, an Experimental Process Timeline is Reconstructed by Forensic Investigations. In B, a Link Analysis of Resources as Reported in the Corresponding Publications. Green Circles Refer to Resources Found on the Storage, Red Circles to Missing Resources.



Moreover, to understand the context in which these resources are produced, another useful forensic technique is link analysis (e.g., Figure 7.5 B), which compares the reported experimental data with the traces found on storage. Thus, a network is created using information from a publication. Subsequently, the list of files is consulted, and resources reported in the publication which are not located in the snapshot are labeled as missing. The link analysis of PUB 10 is presented in Figure 7.5 B. The red circle pinpoints the resources that are not recovered (or missing)

on the laboratory's storage. Hence, R and Python scripts mentioned in PUB_10 are not found on the storage server of the laboratory.

7.4.5 Reporting on Research Data

The last step of laboratory forensics is to produce a report summarizing the results of the investigated cases. The results of the scoring are presented in *Figure 7.6*, where the criteria shown in Table 7.2 have been applied on preserved (I.e., locally archived in the lab) and deposited (i.e., accessible online) material. The scatter plot of DTI scores shows that there is a variety of data management situations behind each publication. There are no standard data management practices in the laboratory, as the preservation of data depends on the experimenters and their data management choices. The score of deposited data is lower than the preserved data for half of the publications. The scores of the other half of the investigated publications had no data available with publications that are of sufficient quality to support the reproduction of the published work. Moreover, the material on the local storage is generally of better quality. However, it comes with a significant drawback: it is not available to third parties or teams who wish to reproduce the publication.

Regarding the underlying reasons for the variations in DTI scores, there are several points worth to be noted. All publications investigated in this study shared research data online, one study (PUB_2) had files shared online, but no files were preserved in the laboratory at the time of the investigation. Nevertheless, half of the publications (PUB_5, PUB_6, PUB_1, PUB_10 and, PUB_4) uploaded data to repositories or supplemental information that were only covering a part of the analyses reported in their corresponding publications. Also, the low score on the online deposit (y) axis is caused by the fact that most of the material being available as PDF files in the supplemental information section of publications.

Besides, there are cases where research data is produced outside of the laboratory by external research groups and commercial organizations. The recovery of resources provided by external parties is challenging when equipment and raw data where processed at a different location than the investigated laboratory as they leave no distinguishable traces on internal storage systems. Higher DTI scores are easier to obtain when experiments are entirely produced in the laboratory, while distributed experimental processes and technology led to lower DTI scores.

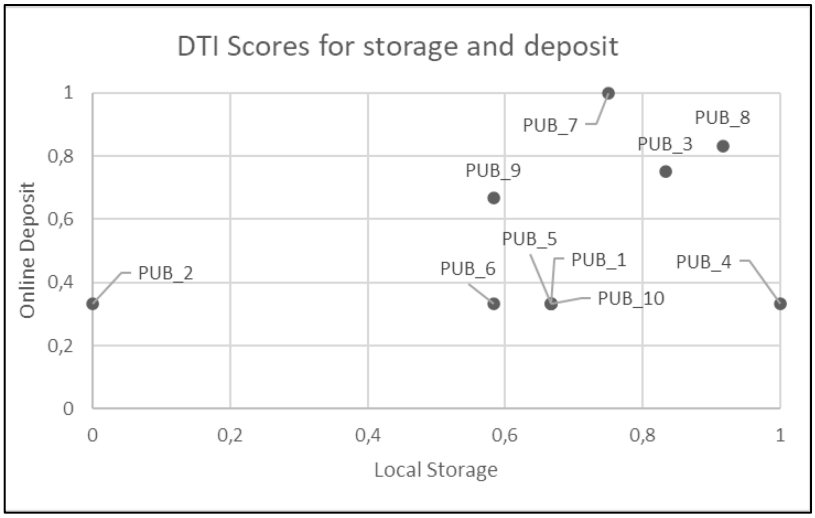
Moreover, most publications are also related to incomplete information on the local storage of the laboratory. While generally, the local storage contained more material underlying publications, the relations of this material to the analyses reported in their corresponding

publications were not clear. One example is PUB_10 that did not differentiate test raw data and raw data from another series of experiments not reported in the investigated publications from the raw data underlying PUB_10. As this influence the recoverability of research data, the DTI score is low (below 0.4) despite the right use of file naming conventions by the authors of PUB_10.

Last, the remainder of this article focuses on transferring the lessons learned from forensic investigations in a laboratory to decision-makers, such as laboratory managers, principal investigators, and support people such as data stewards. In short, how can RDM failures be reduced through the development of RDM capabilities on the one side and analytics on research data on the other side.

Figure 7.6

Overview of the Scores of Information Aspects for Research Data Underlying Each Publication (Local Storage and Deposit). The Closer to 1, the Higher the Information Quality of the Material Extracted From the Storage. In the Case of PUB_2, Our Approach Failed to Recover Files on the Local Storage, Which Explains the DTI Score of Zero.



7.5 RDM Capabilities for Open Science Readiness

In the previous section, we presented the outcomes of the forensics approach. Our findings showed that there is a wide variety of RDM practices that influence the quality of research data. Besides, we show that not all resources were recovered efficiently. The remainder of this article focuses on capabilities that are aimed at reducing the failures of forensics, i.e., the non-

recoverability of essential experimental resources on storage systems in laboratories. To increase the recoverability also means that data availability must be guaranteed. Nevertheless, the results presented in Figure 7.6 show that data availability is not systematic, whether online or locally. Besides, the recovery of relevant research data underlying published experiments is not straightforward, as shown by the efforts and techniques required by a forensics approach to collect digital evidence systematically.

7.5.1 Capabilities

The RDM capabilities for open science readiness cover the four DTI branches that were previously scored: empirics, syntax, semantics, and pragmatic. Each capability could lead to an improvement of the DTI score as they would make the recovery of research data with forensics techniques less error prone. We list the four DTI levels and their corresponding RDM capabilities in Table 7.4. First, to increase the **empirics** part of the DTI score, linking research data on storage systems would enable a smoother retrieval of relevant resources (Bechhofer et al., 2013). Often, research data was retrieved with low certainty during our investigations. Due to a lack of explicit links between folders and files, we retrieved more research data than necessary, files which do not belong to the experiments reported in the investigated publication. A large number of files would then need more intensive processing at the syntax and semantics levels.

Table 7.4

RDM Capabilities for Open Science Readiness

DTI Branches	RDM Capabilities	Description
Empirics	Linked research data	Makes experimental resources discoverable on the file systems by explicitly linking related resources.
Syntax	Traceable resources	Makes use of distinguishable temporal elements, ownership, and sequence in filenames and folder names.
Semantics	Ontology-based data management	Develops consistent naming conventions and lists of materials, people, journal names to be used in filenames with semantically rich aggregates of resources with FAIR objects.
Pragmatic	Open data value strategy	Guarantees the cohesion between laboratory research data and (meta-)data made available on online sources (e.g., articles, repositories) throughout open data value capabilities.

Then, the **syntax** was an issue as crucial elements such as date times, sequences, data creators, experimental conditions where inconsistently written by laboratory workers. It makes those records of experimental operations hard to trace, which is detrimental to reproducibility (M. Williams et al., 2017)**.** For instance, dates and times were alternatively written in US formats and other formats. Data creators were using first names, usernames, and initials to identify themselves and collaborators. Besides, some folders are labeled by journal name, funder, and project name in an inconsistent way. Syntax issues could be circumvented by clear rules that make research data traceable. Traceable research data is, therefore, included here as a capability at the syntax level.

Next, **semantics** is the most challenging branch of the DTI to score based on the forensics approach. A single filename can carry many parts referring to different objects, for instance, objects of study, samples, journals, authors, locations, and domain-specific elements. To remove unambiguous elements, laboratories may use (or develop) ontologies in line with FAIR principles for research data management (Harjes et al., 2020)**.** With an ontology-based (research) data management approach, ambiguity can be reduced by structuring domain-specific knowledge (Lenzerini, 2011). A wealth of ontologies are readily applicable for describing domain-specific knowledge (Mayer et al., 2014), their combination with recent developments in FAIR technology extends semantic capabilities to the whole lifecycle of research data (Harjes et al., 2020).

Last, **pragmatic** relies on empirics, syntax, semantics, and open data value capabilities to provide high-quality research data for reproducibility purposes. Pragmatic is the last level of the DTI score and stands for the (re-)usability of research data. Research data should be preserved and made available following a consistent strategy of documentation and curation to be useful to laboratory members and external parties. Hence, curation is a collaborative effort between many stakeholders to ensure the availability of curated data inside research institutions and on the scholarly communication infrastructure.

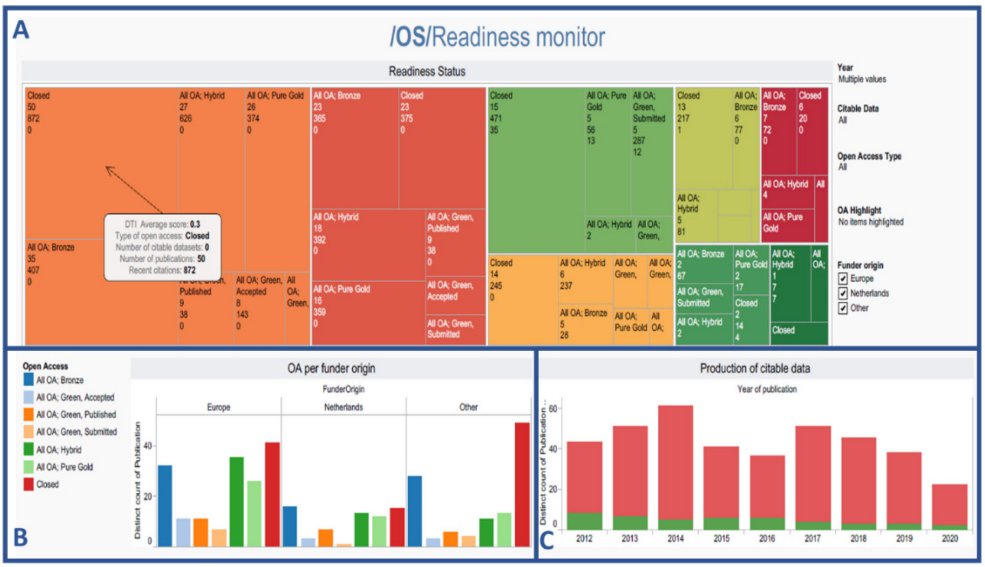
These capabilities, summarized in Table 7.4, are aiming at implementing open science readiness in laboratories. In other words, these are capabilities to achieve the state where a laboratory can responsibly manage research data. However, the dynamic nature of experimentation processes makes the forensics approach hard to scale, and, therefore, automated monitoring of research data quality based on the laboratory's ecosystem is presented here as a future step. Once capabilities that streamline the description of research data are in place, it will allow novel possibilities to interpret data quality in an automated fashion.

7.5.3 Analytics

The analytics dashboard shown in Figure 7.7 is decomposed in three areas representing the "research data strategy" capability of research data management discussed earlier. These three areas are information quality of research data, the openness of research data with citable data, and alignment with stakeholder goals, such as open access programs of funders. We refer to the Tableau public cloud environment for an online interactive version of the dashboard (Lefebvre, 2020). To generate a prototype dashboard for monitoring open science readiness, see Figure 7.7, we extracted data from dimensions.ai about publications and data sets covering the laboratory where we conducted our forensic analyses. Then, we created an additional data set by randomly assigning values to the properties of interest. The reason values are randomly generated to populate the dashboard is that extracting real values from the case study laboratory would necessitate a lengthy forensic investigation process on hundreds of publications. Therefore, the random data simulates DTI scores as those obtained during our forensic investigations, as reported in Figure 7.6.

Figure 7.7

Analytics Dashboard for Monitoring Open Science Readiness



An overview of DTI scores according to the type of openness, citable data, number of publications, and citations are shown in Figure 7.7 A. In Figure 7.7 A, the overview shows how DTI scores could be applied to the scientific output of a lab to flag publications with high, medium,

and low information quality. Publications in a dark-red color indicate that there are serious data quality problems that hinder reproducibility. Compliance with open access to research information per funder is shown in Figure 7.7. B. Figure 7.7. B is an example of how the availability of research data complies with requirements from external stakeholders, in this case, funding agencies. The last part, the state of data openness, is shown in Figure 7.7. C. As mentioned before, the availability of data is crucial for reproducibility. New techniques enabling re-use to rely upon citable data, where research data can be credited in addition to publications (Robinson-García et al., 2016). Thus, the dashboard in its current state simulates a view of data quality using DTI scores to offer a high-level overview to laboratory workers and managers about the state of research data in their organization.

7.6 Discussion

In this study, we have shown the results of a forensics approach conducted in a case study laboratory. The forensics approach, named laboratory forensics, has the purpose of evaluating the quality of information preserved in laboratories as well as the quality of information of research data shared with scientific publications. Next, we described how the outcomes of forensic investigations could nurture a reflection about RDM capabilities and analytics aiming at increasing data quality, and subsequently reproducibility, of published experiments. Here, we present our contribution concerning the existing literature and the practical implications of our findings.

7.6.1 Implications for existing research and future work

We investigated a laboratory that evolves in a chemistry and life sciences, those are scientific domains where one may find a profusion of solutions to preserve, describe and share research data (McQuilton et al., 2016). Still, disparities in the quality of research data exist, showing inconsistencies in the way research data is managed in a research unit. In that sense, our results tend to confirm previous literature emphasizing the responsibility of individual researchers, rather than research units, for managing data (Baykoucheva, 2015; Wilms, Stieglitz, et al., 2018). Nevertheless, we also note that these disparities are rooted in data management practices that are still challenging to align with modern solutions to achieve high quality, reproducible data packages like research objects (Bechhofer et al., 2013).

We found there would be a need for at least four capabilities to make the recovery of research data more robust. Those are linked research data, traceable resources, ontology-based management, and open data value strategy. In the literature, these capabilities are encompassed in findable, accessible, interoperable and, reusable (FAIR) principles and research object principles

and technologies (Ribes and Polk, 2014; Wilkinson et al., 2016). However, a point that current FAIR and research object technology tend to overlook is the multiplicity of actors, equipment, locations, and experimental designs that are currently described by experimenters using standard file management systems in laboratories. A reflection about technology, on the one hand, and research data management capabilities, on the other hand, has to be conducted to make research data management more resilient.

First, we concur that linked research data is a limited part of what makes reproducibility a success (Bechhofer et al., 2013). Nevertheless, many issues arise from the absence of clear links between different outputs generated during experimentation and publication. It further impedes the possibility to automate retrieval techniques and automated assessments of research data preserved on storage systems. At the same time, analytics on linked data posits additional management challenges to integrate a broad diversity of datasets, as shown by the case of big and open linked data analytics (Lnenicka and Komarkova, 2019). As such, it is notable how the application of semiotics, as suggested in our laboratory forensics approach, can account for the enormous diversity of datasets origins and purposes to study research data challenges in greater detail and reconstruct their linkages.

Second, the ambition of making scientific experimentation, at least at the computational level, traceable are the domain of scientific workflows (Cohen-Boulakia et al., 2017; Santana-Perez et al., 2017). In scientific workflows, experimental resources are represented as a graph providing the ability to experiments to repeat experiments by automating the sequence of steps, inputs, and outputs. The difficulty here is that in real laboratory settings, completeness of the archive was a significant issue. As shown in Figure 7.5 B, some of the resources are missing from storage archives. Moreover, the input of the computational experiments is generated by lab equipment. Both types of resources, i.e., laboratory and computational, were (1) not linked correctly in a majority of the investigated cases (2) containing ambiguous information about their usage in an experiment, with the absence of exact version or date-time properties in the file name, for instance.

Third, ontology-based data management refers to a mechanism to access data through a (formal) representation of the domain of interest (Lenzerini, 2011). In the biological domain, and more generally, domains dealing with open data, the use of ontologies for data integration is useful (Mayer et al., 2014; Soylu et al., 2019). The data model on file systems is a hierarchical model that lacks the accuracy of a semantic data model in terms of information that can be preserved. In the investigated laboratory, there was no semantic technology in use to preserve and recover research data. In contrast, much of the information was quite ambiguous as there is no space on the file

system to describe the role of experimental resources in an experimenter. Often, authors, journals, projects are named with abbreviations in filenames, abbreviations that can lead to the uncertain matching of research data to publications. As an example, the authors' initials may be confounded with protein names. The role of ontologies would be to reduce that uncertainty by defining the domain and possible values.

Last, an open data value strategy was missing in the laboratory, despite its utility to forge high-quality data, as shown by Zeleti and Ojo (2017). A missing open data strategy makes the recovery of research data challenging as the material recovered online is not systematically helpful to investigate the data on the laboratory's storage. Except for publications obtaining a high (> 0.7) DTI score (i.e., scoring the maximum on the majority of semiotic branches), online material consisted in supplemental information files with modified names during the editorial process and (extensive) list of files deposited on online archives accessible through a link and identifier mentioned in the corresponding publication. We observed that data deposition is then mostly ad-hoc and dependent on the specific requirements of the outlets in which the investigated articles were published (Wallis et al., 2013). Laboratories should, therefore, work on their internal capabilities to stay in control of data preservation and dissemination technology and mechanisms. Furthermore, a data strategy will foster initiatives to develop a more analytics-driven approach to the evaluation of reproducibility and openness in laboratories that are currently permitted by current RDM practices. In other words, a denser reflection around specific capabilities is necessary for achieving open science readiness. Nevertheless, before reaching a state of readiness where these capabilities can be fully exploited, there are several other practical implications of laboratory forensics that need to be discussed, as explained in the next section.

7.6.2 Practical Implications

Our study aims at contributing to a better understanding of research data management pitfalls as they currently occur in laboratories. Forensic and semiotic techniques help make sense of complex research data and identify shortcomings. For research data professionals, the application of forensic techniques may help shape more specific guidance to laboratories based on their unique RDM strengths and weaknesses, as well as article and data publication practices. Moreover, we recontextualize the scope of FAIR technologies and show their limits when it comes to informing data professionals about the state of RDM in laboratories. That being said, several steps are still necessary before laboratory forensics is fully applicable to professional research data support, as we have learned from a focus group evaluation of laboratory forensics with professionals.

We introduced laboratory forensic techniques to seven participants with expertise in research and scholarly communication. Also, participants had a variety of computer skills, ranging from beginner level to proficient at coding, which is an ideal situation to obtain feedback about the complexity of forensics for a wider audience. The focus group session took place in August 2019 in Los Angeles at the University of California (UCLA) during a six-hour introduction course to laboratory forensics where participants actively applied forensics on a snapshot exported from the case study laboratory and provided feedback on the utility of the forensics approach. Furthermore, limitations and future directions were discussed.

The advantages mentioned by the participants referred to information quality issues, and a lesser extent, governance, and sharing of data. Understanding data to prevent data losses (or finding lost data) served as a basis for discussing conventions or best practices. Several participants even mentioned the benefits of such an approach to develop more robust data organization strategies by discussing conventions in the laboratory. Also, participants considered the practice of forensic investigations as activities that are beneficial for reproducing experiments.

Regarding the challenges of laboratory forensics, the participants discussed the methodological and technical challenges ahead. Regarding the forensic methods, participants experienced difficulties with knowing where the process ends (e.g., when do we obtain the complete set of files, what to write in the report). Also, the fact that, at the empirics level, many files are not coherently aggregated on the storage system. A participant experienced that data in multiple places is challenging. An essential limitation of the forensics approach mentioned by the participants is that, technically, the investigation required participants to be quite comfortable with digital file management systems and python tools such as path2insights (Lefebvre and Bruin, 2019). These technical barriers were still experienced as significant by the participants, so future developments of forensic applications should focus on easier tooling for a wide range of skillsets and audience. Therefore, to apply to a broader audience, laboratory forensics has to be further developed, as explained in the next section.

7.6.3 Limitations

Despite these advantages, the laboratory forensics approach suffers from several limitations in its current state. One limitation is that it is yet to be further applied and evaluated in different laboratories to increase its rigor and reliability. The current results are based on a single site case study, which limits the ability to generalize and compare to other organizational settings. Despite this limitation, the issues encountered also indicate that the investigation of storage systems

in laboratories provides deeper insights into how experiments are conducted, which can serve as a basis for the development of data management systems, scientific workflows and pinpoint specific information issues in laboratories.

The first drawback of the forensic investigation is that thousands of files are created during experiments. On several occasions, their names and folder structure (i.e., signs instead of content) do not always suffice to ensure that the selected digital traces are indeed belonging to the investigated publication. Moreover, a holistic interpretation of such traces is also challenging when filenames do not contain sufficiently informative concepts for third parties. For instance, we found repetitive sequences of filenames that only slightly vary in the experimental conditions. As publications might be based on a fraction of these files, the absence of explicit experimental conditions in a publication has detrimental consequences on the time one investigation might take. In contrast, file names might not be informative enough and require their content to be analyzed (which is out of the scope of this study).

Second, as the case study laboratory has no file naming conventions in every archived folder, the evidence contained in a majority of folders needs to be carefully mapped to publications. At the same time, this issue of mixing experimental data with other types of (non-experimental) data can be mitigated by using discriminative names of folders and files. When no discriminative name, such as the name of a journal, the method of the author is used, the likelihood to include files that are not relevant in the analysis is high. Hence, these limitations are mainly due to the erratic nature of reconstructing events from digital footprints (Mabey et al., 2018) and the error-prone manual extraction of experimental data from storage and publications. Furthermore, the interpretation of signs requires a great deal of knowledge about the experimentation processes and idiosyncrasies of one's field of research.

7.7 Conclusion

In this study, we answered the following question**: "*How can a laboratory forensics approach help achieve open science readiness?*"**. We have developed an approach to investigate experimental evidence in laboratories, including tool support for processing digital files. The purpose of laboratory forensics is to describe reproducibility issues occurring in laboratories in a systematic way, using digital forensic methods and techniques. By investigating the digital files left on storage systems and digital repositories of 10 publications using a variety of tools (e.g., path2insights) and forensic techniques, we have been able to show that in daily practices (digital), experimental data are not systematically preserved or shared online in a reproducible way. We

reached this conclusion by applying the semiotic classification of the descriptive theory of information (DTI) on folders and file names. Besides, we propose that laboratories follow an open science readiness vision on research data management that focuses on increasing information quality for further preservation and dissemination of (open) research data. Subsequently, we demonstrated how our findings from laboratory forensics can assist the digital transformation of laboratories towards open science readiness. In future research, we will further investigate this promising synergy of laboratory forensics with research data management practices. Taking these potential synergies into account, this work contributes to the understanding of scientific data by developing open science readiness to help realize the strategic promise of an open science future.

Chapter 8 | Conclusions

We have examined research data management in an open science context using surveys, interviews, case studies, and laboratory forensics approaches. Our research has resulted in the design of analytics and forensics tools, a framework of replicability threats, and a forensic approach to structuring information quality evaluation using digital forensic techniques. In this final section, we conclude this dissertation by recontextualizing our findings in action design research. Moreover, we seek to achieve a certain level of generalization by discussing and formalizing our findings in Section 8.2. In the introduction section, we posed the following research question:

MRQ - "How can we organize research data management for preserving and disseminating laboratory experiments in a reproducible way?"

To formulate an answer, we first elaborated on the key concepts in Chapter 1. Then, we explored in Chapters 2 and 3 how research data management is paving the way for a more open, reliable, and efficient science despite many organizational difficulties laying ahead for funders and research institutions. In Chapter 2, we conducted two exploratory case studies where we investigated research data management policies and interviewed experts and researchers. We identified three functions encompassing roles and responsibilities in RDM: research, RDM services, and RDM governance. Research covers the wide spectrum of tasks and responsibilities related to the management of research data during its lifecycle. RDM services support research for making the research lifecycle more efficient with tools and practices. RDM governance attributes tasks and responsibilities to roles involved in the research and RDM services functions. One point of attention we found was the overfocus on research for attributing accountability of research data, while the accountability of RDM services was not as clearly defined. At the same time, we observed in Chapter 2 that the quality of RDM deliverables as requested by governing bodies (such as funding agencies) are insufficiently ensuring that RDM will be conducted rigorously.

Second, we also found that research data management and reproducibility are notions that cover many challenging activities and outputs in scientific laboratories in Chapters 4 and 5. First, we developed the laboratory forensics approach and explained some of the technical challenges encountered to evaluate two types of reproducibility: functional repeatability and functional replicability. In Chapter 5, we dive deeper into the details of reproducibility types and provide a more comprehensive view of reproducibility types found in scientific experimentation.

Finally, we introduced tools and approaches to analyze experimental data for biologists and bioinformaticians with research objects in Chapter 6 and laboratory forensics for data stewards and laboratory managers in Chapter 7. The tools developed are a web application for the study of RNA-seq data embedding research object technology on the one hand. On the other hand, a python

module named path2insights combines file path analytics with natural language processing to explore research data preserved in laboratories. Thus, this thesis analyzes data management in practice by identifying the organizational roles and processes of stakeholders, and by intervening with technological tools and approaches specifically designed to answer the MRQ.

8.1 Contributions

To structure the organizational and technological aspects of our study, we opted for an action design research (ADR) that results in (emergent) design principles for research data management through the design and evaluation of an ensemble of artifacts (Sein et al., 2011). Those emergent design principles are discussed in Section 8.1, where we answer the six research questions and summarize each question's main contributions. Next, we present the limitations in Section 8.2 and conclude with future work in Section 8.3.

RQ1: How can research data management contribute to efficient and reliable science?

From a societal point of view, as expressed in funders policies, open science has two objectives: reliability and efficiency of science. Both objectives depend on proper RDM to be achieved. At the same time, we observed that RDM is lagging behind industry standards on several aspects, primarily research data governance (RDG) is currently underdeveloped. Thus far, RDG often merely vaguely implements data management planning and control in its policies. For that reason, **the efficiency of science,** through better-regulated RDM, can be achieved by involving *research supporters* and make their responsibilities clearer in data policies, in which they are currently not well represented. **Reliable science** is challenged by operational issues arising in experimental work such as data archival and privacy to be addressed by RDG policies, supported by governance supporters. However, governance supporters are perceived as open science champions without the proper infrastructure to make RDM work for researchers. The subsequent chapters address the shortcomings in the infrastructure that lead to discrepancies between science at work and how it can be better supported by RDM and RDG.

Insight 1: Implementing research data management for open science requires data governance for reliability. However, we found that, in practice, much of the efforts in RDM is resting on researchers' shoulders and prioritize efficiency (to extend the lifecycle of research data). We followed the categories of Link et al. (2017) to depict an open data implementation strategy. We highlighted that RDM governance supporters have room for improving science reliability (i.e., the traceability of experimental resources). In contrast, RDM supporters and researchers have the

power to make science more efficient if research institutions foster a coherent research data strategy and put control loops in place. Besides, we seek to introduce two support streams of data stewardship: research supporters and governance supporters. Research supporters aim to directly intervene in experimental workflows; governance supporters who are knowledgeable about legal issues can guide policymaking to make research data management more reliable.

RQ2: What are the current challenges and practices in research data management planning?

Research data management planning (RDMP) is the cornerstone of research data management practices and the initial step for researchers to design RDM tailored to their projects. However, RDMP has many ongoing challenges. Funding agencies expect research data to be reusable, however they do not evaluate data management plans and the evaluation criteria differ per funding programme. As an example, we identified differing planning processes, non-standard quality criteria, and a series of complex challenges occurring during the planning phase. Likewise, we have integrated recurring points of improvements from our respondents into actionable criteria to adequately ensure that RDMP addresses data reusability.

In current RDMP settings, data stewards also act as a bridge between funders and research institutions. From that perspective, data stewards possess critical responsibilities in fostering the adoption of RDM by researchers and tailoring RDM services and infrastructure. Research data management plans are documents that serve two purposes: planning research data management for research projects and give directions to governance and support services in identifying weaknesses that can be addressed by better support and infrastructure at institutional or (inter)national level. An issue here is the unstructured nature of answers to data management plans. There are initiatives to collect structured data management plans (Jones et al., 2020). Nevertheless, the profusion of disciplines and the little incentives to elaborate on research data management early make the creation of high-quality data management plans challenging for researchers.

Although data management planning documents are unstructured, we compared data management paragraph content from research grant application documents submitted by researchers to a website providing automated feedback. We showed that natural language processing can play a role in identifying quality issues in data management paragraphs based on a set of criteria used by funding agencies. This resulted in feedback suggestions aimed at completing the automated feedback capabilities of the software. Natural language processing capabilities were used to scan the documents for weaknesses in the data management planning phase based on data management paragraphs' content and preliminary data management plans.

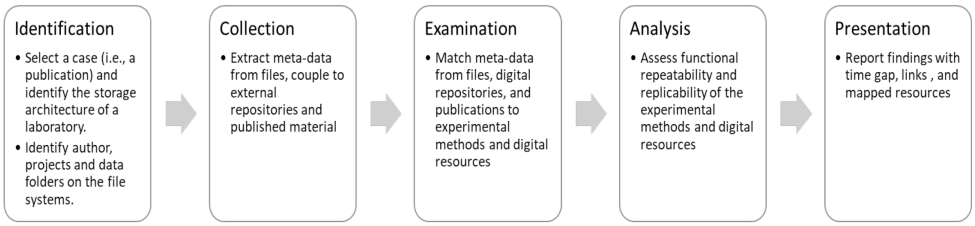
Insight 2: We showed that while research data management planning has the potential of being a tool for improving governance of research data at the institutional level, the current state of research data management planning is misaligned with institutional and funder's policies. Data reusability is perceived as a goal of RDM by funders and research supporters, but the (information) quality criteria necessary for increasing reusability are still not uniformly defined. However, RDMP is also where research supporters and governance supporters can provide feedback for improving research data management and its governance at the early stages of the project. Therefore, means to automate feedback on research proposals and DMP help researchers improve RDM during the writing of their proposals.

RQ3: How can digital forensic methods and techniques be applied to investigate artifact reproducibility?

With Laboratory Forensics (LF), we have shown that in daily practices (digital), experimental resources are not preserved in a functionally repeatable and replicable way. In short, the laboratory forensics approach supports a rigorous assessment of data management issues related to laboratory work. We applied a subset of the criteria of the ACM badging initiative (ACM, 2018) to determine artifact functionality, a quality criteria involved in many other dimensions of reproducibility (see Chapter 1). We also divided artifact reproducibility into functional repeatability and functional replicability. Functional repeatability means that the artifacts (e.g., traces of raw data, software…) found on the laboratory storage are of sufficient quality should anyone *in the laboratory* be willing to reproduce a study using the original author's artifacts. Functional replicability is similar to replicability except that it gives an indication on the quality of the artifacts deposited in online repositories, and therefore aimed at external (teams of) researchers willing to reproduce the results using the original authors' artifacts.

Figure 8.1

Summary of the laboratory forensics process



To investigate reproducibility, we adapted digital forensics techniques to the examination of experimental material in laboratories. There are five main steps that are necessary for an external observer to understand the domain, collect the relevant resources and report on the results. The main steps are depicted in Figure 8. and briefly described below:

- Identification starts with the collection of information about experiments published in a scientific article. Authors, methods, software, organizations, among other elements, are identified from full-text publications. Then, a similar identification is conducted on the storage systems, where potential raw data, user and project folders and files are identified for the next step in the forensics process
- Collection is where the extraction of evidence takes place. More details are given under RQ6 about how collection is operationalized. Briefly, the identified experimental resources are copied from internal and external storage system before the examination step
- Examination is a matching intensive processes where collected resources are related to their corresponding role in an experiment. For instance, a file name containing a protein name and ending with an extension referring to a specific analysis software is matched to a sentence in the corresponding publication
- Analysis is the assessment of functional reproducibility based on the examined evidence. The folder and file structures are compared to the method and result sections of a publication. Missing or ambiguous resources decrease functional reproducibility, whereas complete and documented resources increase functional reproducibility
- Presentation is a step where the findings of a forensic investigations are reported to stakeholders. These reports aim at helping stakeholders improving RDM practices with evidence-based forensic analyses that specifically describe the issues encountered in a given laboratory. In addition, we showed that there are discrepancies between the data preserved in laboratories and data made available online in Chapter 4

Insight 3: We designed a laboratory forensics method to investigate research data in laboratories and create rich insights in reproducibility. In an initial iteration we discovered that data preservation practices insufficiently guarantee the traceability and quality of experimental resources and their corresponding publications. The functional repeatability and replicability of research data varies per research projects and individuals. Nevertheless, well-preserved research data in a laboratory can be significantly degraded when published in repositories and supplemental materials

online. Therefore, we show that reproducibility is challenging to achieve due to data and software being unavailable and low information quality, making the retrieval and interpretation of the operationalizations of experiments inefficient.

RQ4: What reproducibility threats occurring in experimental systems stem from vulnerabilities in research data management practices?

Reproducibility is a complex concept. With the experience gained from the forensic analyses we have sought to integrate functional reproducibility in the broader picture of the reproducibility literature. We constructed a framework that highlights the dynamics of scientific experimentation for working scientists (i.e., the actors) operationalizing experimental design (i.e., the tasks) using laboratory instruments and computers (i.e., technology) to communicate novel findings on the scholarly communication infrastructure (i.e., structure). The framework gives a richer perspective of reproducibility from the perspective of researchers in several disciplines. It circumvents the limitations of our strictly technical view on reproducibility used in laboratory forensics. It is particularly important to underline that experiments scoring high on functional reproducibility after a forensic analysis are not necessarily reproducible from a broader scientific perspective. The framework we introduced in Chapter 4 help to understand why this is the case.

For instance, from the survey, interviews, and screening of publications conducted in Chapter 4, we saw that dissemination and preservation strategies are challenging to implement in laboratories. RDM deals with the fragmentation of policies, ad-hoc data governance in laboratories, and puts few constraints on systematic and structured sharing of computational resources in publications. Our results show that reproducibility risks need to be better understood to redesign the research data management and scholarly communication infrastructures effectively.

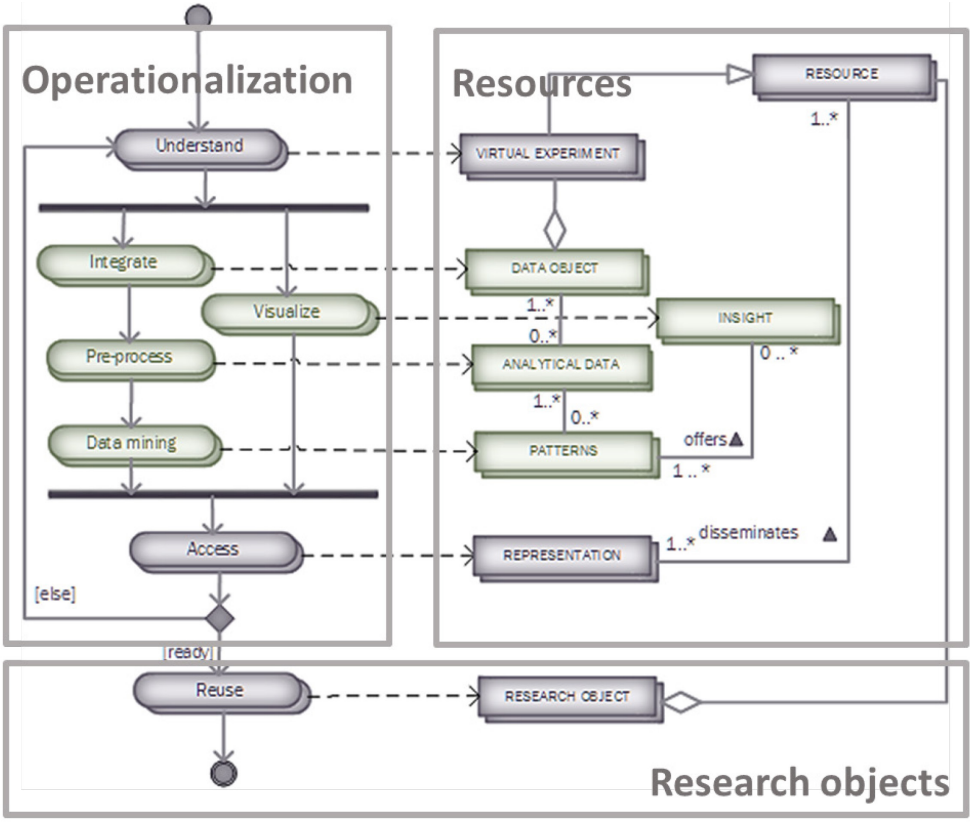
Insight 4: Forensic analyses assess reproducibility from a limited set of criteria. Nevertheless, reproducibility consists of several dimensions that correspond to as many threats to reproducibility, some of which are unrelated to research data management. We can therefore deepen our understanding of reproducibility issues in laboratories. We first decomposed reproducibility into inferential replicability, method replicability, result replicability, closedness, repeatability, conceptual replicability, and availability. By doing so, we describe how RDM activities can support reproducibility in experimental systems. Besides, we are able to set limits to the scope of the reproducibility assessments as presented in our laboratory forensics work.

RQ5: How can we bridge interactive data mining solutions with reproducible research object technology?

To answer RQ5, we conducted a design science project in a biomedical research laboratory. The settings differ somewhat from our main case study organizations, even if similarities are found due to both organizations conducting experimental research. Our results suggest that the software technology that can address reproducibility suffers from a lack of solutions that are both satisfactory to bioinformaticians and biologists who are mutually engaged in computational experiments. There are also serious apprehensions from bioinformaticians to offer advanced analytics to biologists that are going beyond data visualization as many choices and assumptions are contained in bioinformatics software.

Figure 8.2

The translation of experimental operations to research objects as designed for the RNA-Seq miner



To design a solution bridging both experiments in biology and collect annotated traces and resources, we opted for a research object architecture, as shown in Figure 8.2. As explained in Chapter 6, we built a system that enabled biologists to use bioinformatics packages in an interactive way. Next, the system generated a representation of each experiment using a research object model.

For the **operationalization of experiments**, we define two concurrent processes (see left-side of Figure 8.2.). First, the domain of the data needs to be understood by the researchers investigating the phenomenon at hand. Then, bioinformatics pipelines integrate, pre-process and mine multiple sources of data to answer the research questions guiding the experiment. Concurrently, the interest of biologist lays in the visualization capabilities to group, filter and identify relevant values. Therefore, a deeper integration of bioinformatics pipelines and interactive visualization capabilities was at the core of the designed artifact (as explained in Chapter 6).

On the right-side of Figure 8.2, we depict **resources.** Resources are the inputs and output of experiments that are preserved as elements of research objects. The integration of research data by bioinformaticians lead to a (series of) data objects. Next, these data objects are pre-processed into analytical data, suitable for analysis by biologists. On analytical data, patterns can be identified through data mining. The research object architecture is adding a crucial capability labelled **access**. Access is implemented as a REST programming interface where each resource of a virtual experiment can be accessed by members of the research team or by external teams.

Research objects propose a representation of experimental operations and resources in a way that makes them accessible through semantic annotations and interfaces interoperating experiments with third-party software, such as software used by teams that attempt to replicate experiments. Contrary to digital files that required forensic approaches and techniques due to their lack of structure and documentation, research objects aim at aggregating experimental resources in an accessible way to increase **reusability.**

Insight 5: We found that there is a gap to fill both in terms of data analytics and the reuse of previous work. Biologists tend to ask more visualization capabilities, while bioinformaticians opt for a solution where scripting or custom data processing is allowed. However, visualization alone for biologists is not satisfactory to query data and compare the variations in results obtained by different methods. Reusability of data, workflows, or parts of experiments appear to be more pleasant features, for the two types of end-users that assessed the artifact, than reproduction. Nevertheless, a bridge can be made by allowing bioinformatics programs to be included in the visualization pipelines of biologists, as long as data, algorithms and their parameters can be preserved in an annotated way.

RQ6: How can a laboratory forensics approach help achieve open science readiness?

We have developed an approach to investigate experimental evidence in laboratories, including tool support for processing digital files. The purpose of laboratory forensics is to describe reproducibility issues occurring in laboratories in a systematic way, using digital forensic methods and techniques. By investigating the digital files left on storage systems and digital repositories of 10 publications using a variety of tools (e.g., path2insights) and forensic techniques, we have been able to show that in daily practices (digital), experimental data are not systematically preserved or shared online in a reproducible way. Besides, we propose that laboratories follow an open science readiness vision on research data management that focuses on increasing information quality for further preservation and dissemination of (open) research data. Subsequently, we demonstrated how our findings from laboratory forensics can assist the digital transformation of laboratories towards open science readiness. The first step towards open science readiness was to refine the laboratory forensics approach designed for answering RQ3.

Figure 8.3. shows a more complete laboratory forensics approach than the approach used to answer RQ3. The lab forensics approach presented here details the steps resulting in an evaluation of the quality of information stored on file systems. The concepts (on the right side of the process deliverable diagram) are described in table 8.1.

Table 8 .1.

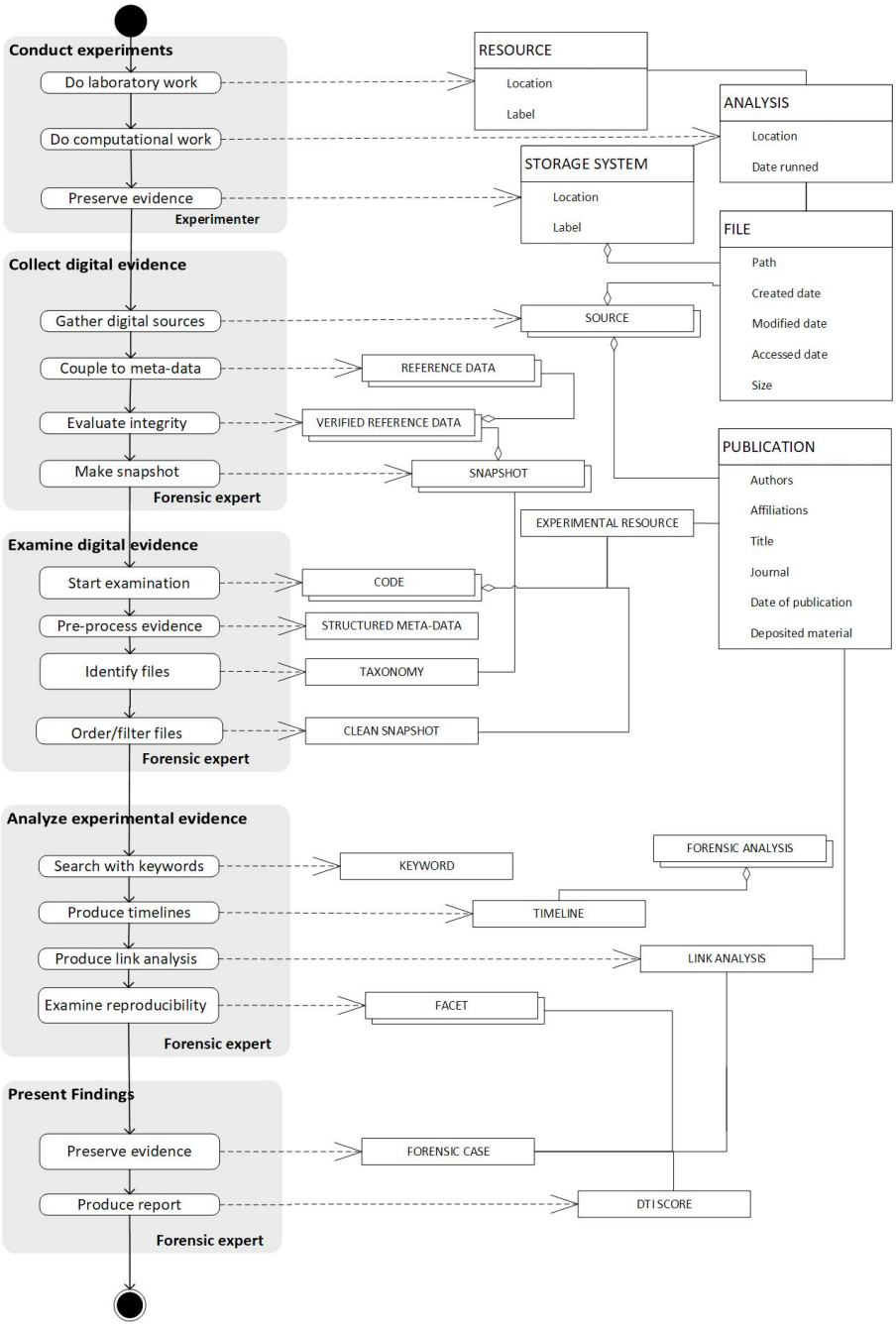
Concepts of Laboratory Forensics

CONCEPT	***Description***
RESOURCE	*Experiments employ materials, software, and equipment to preserve and process samples*
ANALYSIS	*The analysis refers here to the processing of resources with computational means.*
STORAGE SYSTEM	*The storage system is where raw and processed resources are preserved in a persistent way.*
REFERENCE DATA	*Reference data is used in a forensic analysis to verify the likelihood that resources stored on a storage system correspond to resources reported in a scientific publication.*

CONCEPT	***Description***
VERIFIED REFERENCE DATA	*Reference data is verified when the resources, methods, authors, organizations (and any other relevant element) is confirmed to be part of the investigated experiments.*
SNAPSHOT	*At this moment, laboratory forensics is limited to a textual analysis of resources preserved on storage systems. A snapshot generates a text file containing a copy of the (meta)data preserved on storage systems*
CODE	*Coding is a label on resources identified in publications. Labels refer to a taxonomy of methods, equipment, among others.*
TAXONOMY	*Taxonomies classify codes, digital files, and folder structures to generic categories such as person, method, equipment, software, and repository.*
FORENSIC ANALYSIS	*A forensic analysis Is the analysis of the snapshot to generate a reproducibility assessment*
FACET	*Facets are characteristics of information aspects, as defined in the descriptive theory of information (see Chapter 7).*
DTI SCORE	*We devised a DTI Score that approximates the level of information quality of resources stored on storage systems. A low DTI hinders reproducibility as it signifies that resources are missing (or hardly recoverable) from the investigated storage; a high DTI scores means that resources are stored and well documented.*
FORENSIC CASE	*The forensic case comprises all the investigated elements: snapshot, reference data, forensic analyses, and report.*
LINK ANALYSIS	*The resources are linked to represent the relation between the types of resources identified on the storage system. An example of linked analysis is shown in Figure 7.5 (Chapter 7) where it appears that analysis scripts are not included in the archive of research data underlying one of the investigated publications.*
TIMELINE ANALYSIS	*Timelines show when resources are created or modified on the storage and help identify gaps in the preservation of resources*

Figure 8.3

Process-Deliverable Diagram of Laboratory Forensics



The notion of open science readiness extends the outcomes of laboratory forensics to the monitoring of the state of open science and reproducibility in a laboratory. The reasoning behind open science readiness is twofold. First, we observed that current research data management practices in laboratories do not systematically preserve experimental data in a reproducible fashion. Second, research data management governance in research institutions is uninformed about the practicalities of scientific experimentation and the implications of current scholarly practices in research data management. In that context, open science readiness is then the capability of laboratories to implement research data management with the aim of addressing reproducibility threats. Therefore, open science readiness requires capabilities to trace research data underlying scholarly publications and measures bits of research data management practices to point shortcomings to laboratory members and institutional stakeholders.

Insight 6: The second iteration of laboratory forensics sheds light upon information quality issues arising in storage systems. Instead of focusing on reproducibility from a normative perspective (i.e., rules to be followed to make experimental material reproducible), we opted for an assessment of reproducibility from an informational perspective. Open science readiness (OSR) perceives ad-hoc evolutions of open science practices as a threat that laboratories should address. OSR We elaborated on a laboratory forensics approach to counter weaknesses of current data management practices and help identify information quality issues impacting reproducibility. Moreover, we connect the outcomes of our forensic approach to a comprehensive semiotic analysis of information quality.

8.2 Lessons Learned

This section highlights some of the lessons learned and frames our work in the current RDM literature. We discuss several lessons learned from the various studies we have conducted.

First, research data governance is often ad-hoc and research data management policies are not sufficiently implementing control mechanisms. Therefore, the action of research institutions, funders, and support services on collecting evidence about the state of research data management practices is limited and further improvements are challenging to coordinate through organizational units in research institutions. Nevertheless, we observed that the study of research data management plans combined with the acquisition of evidence about RDM practices in laboratories will benefit governance. We showed that current infrastructures lead to information quality issues that make published experiments irreproducible. We then propose a solution to

decrease information quality issues by monitoring the quality of experimental resources, as shown by our work on open science readiness in Chapter 7.

Second, research data management (RDM) in scholarly settings poses challenges that differentiate RDM from enterprise data management. RDM concepts are challenging to understand with concepts from enterprise data management. In Chapter 2, we related RDM to an enterprise data management (EDM) framework from DAMA. However, we found that many activities and roles present in RDM were hardly overlapping with EDM roles. In an academic context, most of the knowledge areas of data management ends in the researchers' hands, whereas enterprise data management presupposes trained professionals for metadata management, information security and data modeling (among others). Also, policies in our case study institutions are written and overruled by faculties and departments while research data knows boundaries that are much more interdisciplinary. It calls for a generalization of the knowledge we have about research data management practices to devise policies in line with experimental work (for instance) instead of organizational divisions that might poorly reflect the actual research settings of a study. This situation leads to difficulties for managing data because of some unique characteristics of data handled in scholarly settings:

- The lack of control in a number of aspects of scientific knowledge production. There are external factors such as funder policies, journal policies, peer-reviewing where external actors interfere with how research data ought to be managed.
- Software used or developed for research purposes may be short-termed. Research data in the lab we investigated was scrupulously managed by projects and derived publications. Nevertheless, the software versions and custom scripts were harder to trace back to elements preserved on the lab ‘storage systems. Experimental software and computer scripts are also hard to inventory, and research organizations miss the global picture of data inputs and outputs, processing software, licenses and more. This is another point where the systematic collection of evidence from lab can help produce better governance of information technology currently “under the radar”

Third, FAIR principles for research data management need empirical grounding. FAIR research data are principles leveraging the linked data architecture for improving the findability, accessibility, interoperability, and reusability of research data. Although these principles are commonly found in the RDM literature, policies, and workshops, we found that the gap with RDM in practice is still significant. Experimenters do not systematically have full control

on the computational resources used for conducting experiments. Dependencies to external software limit the extent to which experimenters can achieve FAIR data. Moreover, experiments may be distributed across several laboratories and facilities, each performing analyses with different methods, laboratory resources and software. It leads to a situation where research data of one laboratory becomes the secondary data in a (collaborating) laboratory while not all resources are to be found in one laboratory or fully online. Then, editorial processes might influence the management of experimental resources by modifying (parts) of the analyses previously conducted, add new analyses, or lead to resubmissions to other journals. We observed that this was a challenge with respect to internal storage systems as it fragments files and folders further and makes it challenging for researchers to preserve the files in a structured way. Also, data availability statements in publications are often too ambiguous to be used to find resources (see Chapter 5 for an analysis of data availability). This section of publications has room to be improved with FAIR and provenance metadata as it is currently limited to full-texts and weblinks in an unstructured way. In that respect, laboratory forensics can help pursue the collection of evidence about research practices in laboratories.

- Identify division of labor in publications and trace back to the research data that has not been (openly) published
- Identify equipment, hardware and software and match their output to storage systems.
- Identify reporting issues in publications (mislabeled software, incomplete descriptions,) and compare disciplines (or research) groups to better specify their RDM needs
- Integrate forensic approaches with (inter)national communities and working groups for open science and data management so that evidence about RDM in practice is acquired by data stewards

Fourth, the application of machine learning and natural language processing techniques offer great potential but also pose great challenges. There is a great potential to integrate information preserved on storage systems with methods and resources available in full-text publications. resources described in methods section can be mapped to resources identified from file names. However, although there is some potential in applying automated techniques to extract structured information from files and publications, digital files and full-text publications remain complex to analyze in-depth algorithmically (Lefebvre et al., 2019). As we discovered during our laboratory forensics study in Chapters 4 and 7, the outcomes of a reproducibility assessment depend on how well external observers can interpret the resources. The successful interpretation of these

resources requires background knowledge and multiple comparisons between the concepts written by researchers in file names, the workflow presented in a publication and the reference data created during the forensic analysis. Therefore, the applicability of automated methods is limited, and our tool Path2Insight is a first step towards the use of natural language processing for research data management.

Fifth, open science is gaining traction, but its broadening scope is weakening efficient intervention for RDM. In recent years, open science has been constituted as a means to address each of the shortcomings of scientific practice. In Chapter 1 we exposed shortcomings from the point of view of working scientists: irreproducible results, lack of data sharing, complexity of the scholarly infrastructures impeding the preservation and sharing of high-quality data to name a few. These shortcomings are at the core of what open science has attempted to address since its inception, as an extension of open access. Nowadays, open science covers a much broader range of shortcomings in scholarly practices, including rewarding individual researchers, performance evaluation of research units, citizen science, and much more. This dissertation only touched upon the core shortcomings and showed that governance and monitoring are lacking at this stage. Therefore, we employ a more stringent definition of open science that reflects upon the concept of open science readiness, which focuses on information quality and accessibility for better reproducibility of experiments exclusively. At the same time, the broadened scope of open science must be addressed, as there is a risk that open science starts to englobe and attempt to address all (perceived) shortcomings of scientific practice, which in our view would dilute the efficiency of intervening on the core shortcomings.

8.4 Bridging the Gaps in Practice

The challenges above result from a number of gaps in the current research data management landscape that we experienced at our case study organization. In this section, we relate the lessons learned to concrete examples encountered in our research. We also reflect further on the choice of an information systems research (ISR) paradigm to study research and the management of research data. This choice influenced many of the assumptions, methods of inquiry, and deliverables included in this dissertation. The goal is to provide further insights into conducting IS research in a scholarly context and develop RDM further than what we have achieved in this work.

Our research project started as an extension of Chapter 6. We would have broadened the research object architecture, a platform harvesting and annotating all sorts of research artifacts produced during scientific experiments. The (longer)-term storage used for the project was a Yoda environment at Utrecht University, a storage solution using iRODS as a backend. At the start of the project, iRODS featured options to store meta-data as a key-value-unit structure. The essence of our study would then be to bridge the meta-data requirements for preserving experimental data in a reproducible way, as well as evaluating such as an environment in one or more laboratories. Nevertheless, the application of ISR methods to study scientific work (mostly bioinformatics) in practice and provide actionable deliverables to research support (more specifically IT services) has proven arduous due to two factors: (1) the conceptual and practical unclarities of research data management for researchers, supporters, and policymakers (2) the design of studies for which the outcomes are usable by several disciplines, such as bioinformatics and ISR.

First, unclarities emerged at the start of our study. In any branch of design science research (DSR), the organizational structures and stakeholders have a prominent place, as design science artifacts are evaluated according to utility criteria. In those circumstances, the vast RDM and open science landscape needed to be framed. Framing is critical for DSR so that the results obtained during evaluation sessions with stakeholders are relatable to the characteristics of our designed artifacts (e.g., a research objects platform) and not contingent factors (e.g., the lab starts using notebooks and it suffices to impact the reproducibility of their work positively). In other words, rigorous applications of design science principles lead to the designed artifact being at the source of (positive) adjustments in organizational processes. Nevertheless, in the context of RDM, the potential changes range from profoundly disruptive to strongly counterproductive.

A profoundly disruptive change is streamlining research artifacts from acquisition (or creation) to scholarly publications effectively. We have observed information quality issues that

posit some severe threats to reproducibility and reusability, as previously described. Nevertheless, those issues are not systematically caused by inadequate data management practices in labs. In Chapter 1, in Figure 1.3., we depicted organizational characteristics that promote laboratories to operationalize new experiments differently. The aim of alternative operationalizations is to create new experiments building upon recent discoveries or addressing shortcomings of previous experiments while also composing with the limits of current technology. Hence, designing research infrastructure for augmenting research data quality is a long-term research project, demanding that research institutions pursue their efforts to invest in RDM infrastructure. The current situation where (many) choices for RDM are left to researchers that make RDM depend on a funding process with few chances of success is detrimental for institutional data management strategies in the longer term. However, reproducibility and data reuse are longer-term achievements.

A counterproductive change is to disrupt research processes while evidence is scarce about how scientific articles unfold from laboratory work and scholarly discourse. Criteria of rigor in data analyses and reporting are better understood than characteristics of "good" data management and researchers who are not well-versed in what might pop up as (yet) another administrative burden that occurs in a crowded competition for funding. Therefore, existing solutions that support better RDM processes are often overlooked, even more so if they require more time and effort to store, exchange, and quickly retrieve research data above other core activities. Besides, journals, funders, and institutions fail to inspect the availability and quality of disseminated research data, leading to little incentives to plan and monitor RDM carefully. In that case, intervening in RDM as IS researchers calls first to understand better what, how, when, how much, and with whom research data is exchanged in laboratories. Input from researchers is, unfortunately, not a strong source of evidence for answering these questions. This is what led us to devise a forensics approach to obtain this evidence without depending on the researchers' input for understanding RDM matters in laboratories. In that sense, improving support for RDM comes from engaging directly with research practices in laboratories and collecting evidence about the state of affairs to nurture governance and institutional infrastructure that fit research work in practice.

Then, the underlying question is then how to intervene in laboratories and collect better evidence for the development of RDM than what we could currently achieve in the scope of this dissertation. Here, direct collaboration with scientific disciplines is key. It is a matter of engaging with knowledge creation processes in laboratories as well as understanding science from a "working scientist" perspective, to use a Latourian concept (see Chapter 1). Even a more basic approach than forensics could foster genuine new insights into data management practices. Spending some time

in laboratories as external observers, starting a series of exercises, together with a sample of scientists, to reconstruct publications, identify where research data is stored, and scan publications for data made openly available will show what the deeper struggles with research data management are, and provide a broader understanding to nurture RDM strategies. There is no need to approach labs from a normative perspective, by, for instance, only pinpointing that data went missing or that it is s not deposited online. These exercises would aim to collect direct experiences from researchers in an as-is situation without the lenses of a "should be" that often occurs in open science.

Second, we have encountered several drawbacks while conducting studies of an interdisciplinary nature. The expectations of an information systems research community and a bioinformatics community regarding artifact development and publication differ. In the information systems community, literature about research data management is scarce (less than 10 results in the AIS library in November 2020 for the keywords "research data management", only counting full articles) and most of the theoretical work underpinning data management in a scholarly context is yet to be developed. It made the scoping of our DSR approaches in RDM hard to tailor for information systems venues and journals. In bioinformatics, software artifacts that are developed aim at answering concrete questions emerging in biology and related fields or increase the efficiency of bioinformatics pipelines for better exploiting data in the boundaries of these fields. The demand for genericity is not as high in bioinformatics as developed artifacts can be proven to adequately deal with the biological data and questions. In information systems research, we were confronted early on with an exigence of abstraction and a well-delineated contribution to the IS body of knowledge. These frictions have had a long-lasting impact on how this dissertation would eventually be shaped. As much effort have been spent to establish RDM as a novel phenomenon of interest for IS researchers, it appeared that the demand for practical utility embraced by bioinformatics would be challenging to achieve all at once.

Nevertheless, information systems research offers the conceptual and methodological toolkit for studying RDM in a field-independent way. Design science research particularly suits the current need for intervening and deploying useful artifacts in RDM. DSR offers a methodological backbone to include many perspectives from stakeholders such as funders, librarians, IT services and researchers in a consistent research framework. Moreover, artifact design is especially welcomed in an evolving scholarly context, where open science calls for a number of changes where information systems could be of valuable help. Collaborating with fields like bioinformatics with offer the opportunity for reality checks and forge a better mutual understanding of what RDM for open science actually means and how it might shape scientific practice in the coming years. At the

same time, the multiplicity of angles to research RDM in its complexity and conduct research on research from an IS perspective in general, still needs more efforts to find common grounds with what is valued by research in practice, as exemplified by our cases in bioinformatics.

8.5 Limitations

The first limitation is the limited scope of our case studies. Most of the evidence we collected originates from one university in the Netherlands. While we have seen in Chapter 1 that the organizational settings observed in our samples also reflect changes in policies at national and international levels, we have not yet satisfactorily included the variation in RDM practices and comparability with laboratories conducting similar experiments. Therefore, we addressed ***external validity threats*** by situating our case study organizations in an international context of open science. That being said, generalization of our results is a limitation, and future research is needed to evaluate and refine our approaches in broader settings. Applying forensics on a single case organization, we showed that information quality issues do exist, both regarding laboratory infrastructure and with respect to the scholarly infrastructure. In the literature, as explained in Chapter 1 and Chapter 5, similar patterns have been found in other domains but are limited by the fact that those findings did not result from in-depth investigations of digital files in laboratories. Thus, while the result of our forensic approach lacks generalizable insights, we can be confident that similar patterns will be observed in other laboratories based on our interviews and focus group discussions.

Second, ***internal validity threats*** to our research are that the artifacts designed and used for collecting and analyzing dealt with a limited scope of experimental activities. We focused on a few scientific domains in relation to life sciences, with few exceptions. Previous studies showed the variability in research data management practices, corroborated by our findings in Chapter 2. Thus, several limitations apply to the relation between the artifacts we designed and the assessment of reproducibility:

1. We used a score (the DTI score, see Chapter 7) that summarizes four aspects of information described in the DTI theory. We summarize a rich set of variating input and experimental designs into a single value after interpreting a limited amount of digital evidence. Therefore, the comparability of DTI scores between experiments (or publications) is problematical, and more evidence and testing are needed to produce an accurate and reliable DTI score.
2. We measure a reproducibility "potential" instead of successful reproducibility. In other words, we can only conclude that there is a chance that a study can be reproduced based

on information quality attributes evaluated during laboratory forensics. There is no possibility to conclude that a study has been reproduced, or that a high-scoring study is more reproducible than a lower-scoring study using the DTI. As shown in Chapter 5, there are many more aspects to consider when establishing the reproducibility of results. Therefore, the DTI scores fits a better purpose for research data management governance and is not to be used to conclude that a low-scoring study is irreproducible.

3. In the first study, presented in Chapter 2, we argue that RDM can result in efficient and reliable (open) science. As explained earlier, we focused on RDM among many other dimensions of science. While RDM demonstrated its positive impact on efficiency and reliability of science, there many confounding variables that are left out of our approach so far. It remains challenging to define a standardized scope of RDM, and hence, to communicate which RDM roles and activities affect efficiency and reliability of science. Next, laboratory forensics is, in its current stage, is too labor intensive to scale. Although the laboratory forensics approach, while labor intensive, was still manageable for an external investigator, we can assume that for a number of laboratories, with a larger number of devices and equipment, or research in clinical settings might vastly complexify the feasibility of forensics, even considering a team of forensic investigators.
4. We mainly collected our data from Utrecht University. We have made several attempts to broaden the scope of the data collection, nevertheless we have also found that collecting data about research practices across universities is challenging to implement. In our experience, research institutions tended to favor local surveys, for instance, for acquiring knowledge about research data management practices. However, laboratory forensic outcomes and research on research data management need to be evaluated against evidence from a larger number of institutions. Hence, research on RDM requires to consider multi-case studies by design, where stakeholders accept to cooperate for acquiring evidence on open science and RDM.

Finally, tool support for forensics and open science readiness are in an early stage of development. We have had limited opportunities to evaluate the laboratory forensics in real settings, partly due to a lack of proper tooling to make the approach suitable for a non-technical audience. Therefore, we discussed the generic approach, outcomes and future directions in focus groups and presentations. The evaluation of the efficiency and accuracy of forensics applied to research data is still in a preliminary stage, and future design iterations may achieve a faster and more in-depth analysis or research data. In addition, open science readiness has not yet been implemented in the

investigated laboratory. The design proposition of open science readiness needs further extensions to be evaluated and used in practice. Then, better evidence will be created for research data governance as the metrics, evidence and, stakeholders will be able to formulate an evidence-based open science strategy.

8.6 Future Work

In this dissertation, we have shown, with a number of case studies, that our approaches led to a rich perspective on social and technological challenges within research data management for reproducible research. We have approached research data in laboratories with digital forensic techniques to obtain a better view on information quality issues that persist in archived data. We have also investigated other types of documents such as research data management plans, scientific publications, and research data policies from funders, research institutions and publishers. Further, we developed tools and approaches to adapt forensic techniques to the investigation of research data. Nevertheless, the current results we obtained are mostly situated into one research organization. Therefore, there are a number of research areas that need further developments in technological support and analytical approaches to further expand our view on research data management in an open science context.

First, the **organizational (i.e., social) dimension of open science in research institutions** deserves more attention. The open science landscape is evolving rapidly and new actors influencing open science strategies in research institutions are now found in communities mixing faculty, junior researchers, and support services. The extent to which these new actors and communities affect core activities such as data management planning discussed in Chapters 2 and 3 is a point of attention. Besides, these communities might play an active role in shaping better governance for open science at institutional levels. Including open science communities in the sample will enable a more accurate picture of more recent open science practices at universities.

Second, we have only shown the utility of laboratory forensics on a small set of cases. These techniques have potential for numerous stakeholders involved in research data management to provide evidence about the state of data management in specific laboratories and disciplines. A major shortcoming that needs to be addressed is the **tool support for streamlining our forensic approaches** and make them useful for a wider audience by automating forensic activities where possible. In Chapter 7, we discussed the shortcomings with a focus group, and future research should focus on the algorithmic and human-computer interaction aspects of forensics. On the one hand, software that supports forensics activities on research data requires more detailed algorithmic,

logging, mining and natural language processing capabilities to produce high-quality reference data, make the results of investigations fully traceable, and help score information quality more interactively. On the other hand, current use of the tooling necessitates programming knowledge that might posit issue for certain stakeholders. Digital forensics tooling offers user-friendly interfaces for exploring storage systems, laboratory forensics can leverage digital forensics software tooling to reach a wider audience of RDM professionals.

Third, as mentioned in Chapter 7, laboratory forensics (as developed in Chapter 3) are a starting point to provide monitoring capabilities of reproducibility and research data management in laboratories. Taking the shortcomings of governance described in Chapters 2, 3 and 5, we **introduced a strategy for open science readiness** that constitutes a horizon for improving the efficiency of RDM infrastructure. With the open science readiness (OSR) dashboard, we make an attempt to give insights into the accessibility and quality of research data. We used a limited number of metrics originating from our experience with laboratory forensics. A future step is to include a broader set of indicators about the domains of informetrics, meta-research, and responsible research to achieve a more robust overview of the many aspects of reproducibility in research labs. Furthermore, the utility in practice of the dashboard should be further evaluated as well as integrated into our overall action design science research approach.

Last, we underline the necessity of working on robust techniques to process scientific information, whether it concerns research data or as full-text publications. **Achieving reproducibility in the studies included in this dissertation** suffers from the same challenges encountered by our interviewees working in other research domains. In other words, we did not escape the common hurdles of sharing evidence for reuse and replication as we worked with licensed, confidential, volatile evidence extracted from paywalled content and file systems that were replaced twice during our investigations in the laboratory, for instance. It is therefore welcomed to investigate further the integration between the scholarly infrastructure and research data management while leveraging any gain obtained by small steps towards open science. Besides, even in design science research, there are recently vivid discussions about the reusability of artifacts published by DSR researchers. Therefore, it is apparent that the need of scholars to access and build upon previous technical artifacts are covering a much larger spectrum of disciplines and domains than bioinformatics and biology. In essence, working on better RDM for reproducibility (and reusability) will foster steps towards transparency and openness of research information by making the governance of open science fairer and more evidence based in all areas of research.

8.7 Personal Reflection

In the final section of this dissertation, I would like to write a few words about why doing research on research is as fulfilling and potentially impactful as surely extremely confusing for a junior scholar.

When I started to work on a topic that involved studying my peers and science at work, I believed the timeliness and relevance of investigating research data management for open science would make it a rather smooth journey. Smooth in the sense that the "normal science" I was observing appeared to be much harder and respectable to conduct, as it involved (and still does) the mastery of an important body of knowledge, the meticulous application of research or experimental techniques and competing in well understood publication and conference arenas. In contrast, five years ago (and still nowadays) research data management is a topic that can become anything, for instance, you could start making the software that makes things easier and make a difference or adapt a data management framework to data management in an academic context, and this would help quite a number of stakeholders.

As a topic of research for a PhD, the multitude of angles and potential contributions are quite overwhelming. At the same time, many of the basics of data management (and even more open science) are not well known. This leads to a situation where my efforts to position this research were not followed by proven approaches that I would have find important to learn, as a (future) researcher. An example is design science research. I learned extensively about its complexity and rigor during a doctoral consortium and conference at DESRIST in 2017. But no matter how hard I attempted to structure this project to leverage design theories and outcomes, I felt that the novelty of RDM made the application of rigorous information systems research basically impossible. My suggestion then would be to give some room to include replication research as part of any PhD trajectory, certainly in information systems research. By doing so, no matter the biases induced by the topic, time and resources are given to become more comfortable with conducting independent research that can be evaluated with proven standards. A perfect opportunity for fostering good science with open science!

Bibliography

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., Abdinov, O., Aben, R., Abi, B., Abolins, M., AbouZeid, O. S., Abramowicz, H., Abreu, H., Acharya, B. S., Adamczyk, L., Adams, D. L., Addy, T. N., Adelman, J., Adomeit, S., … Zwalinski, L. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics*. https://doi.org/10.1016/j.physletb.2012.08.020

Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. X., Adya, V. B., Affeldt, C., Agathos, M., Agatsuma, K., Aggarwal, N., Aguiar, O. D., Aiello, L., Ain, A., Ajith, P., … Zweizig, J. (2016). Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*. https://doi.org/10.1103/PhysRevLett.116.061102

ACM. (2018). *Artifact Review and Badging*. https://www.acm.org/publications/policies/artifact-review-badging

Adewumi, M. T., Vo, N., Tritz, D., Beaman, J., and Vassar, M. (2021). An evaluation of the practice of transparency and reproducibility in addiction medicine literature. *Addictive Behaviors*, *112*, 106560. https://doi.org/10.1016/j.addbeh.2020.106560

Ahmad, A., Lyytinen, K., and Newman, M. (2011). The evolution of process models in is research: from a punctuated social process model to a socio-technical process model. *AIS Electronic Library (AISeL)*.

Akers, K. G. (2017). Going beyond data management planning: Comprehensive research data services. *College & Research Libraries News*. https://doi.org/10.5860/crln.75.8.9176

Akers, K. G., and Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*. https://doi.org/10.2218/ijdc.v8i2.263

Amorim, R. C., Castro, J. A., da Silva, J. R., and Ribeiro, C. (2015). A comparative study of platforms for research data management: Interoperability, metadata capabilities and integration potential. *Advances in Intelligent Systems and Computing*, *353*, 101–111. https://doi.org/10.1007/978-3-319-16486-1_10

Amsterdam Call for Action on Open Science. (2016). https://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science

Andersen, H., and Hepburn, B. (2016). Scientific Method. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University.

Ardestani, S. B., Hakansson, C. J., Laure, E., Livenson, I., Stranak, P., Dima, E., Blommesteijn, D., and Sanden, M. van de. (2015). B2SHARE: An Open eScience Data Sharing Platform. *2015 IEEE 11th International Conference on E-Science*, 448–453. https://doi.org/10.1109/eScience.2015.44

Årnes, A. (2017). *Digital Forensics*. John Wiley & Sons.

Austin, C. C. (2019). A Path to Big Data Readiness. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 4844–4853. https://doi.org/10.1109/BigData.2018.8622229

Ayris, P, Berthou, J.-Y., Bruce, R., Lindstaedt, S., Monreale, A., Mons, B., Murayama, Y., Soedergard, C., Tochtermann, K., Wilkinson, R., and Wilkinson, M. (2018). *Towards a FAIR Internet of Data, Services and Things for Practicing Open Science*. *3*, 0. https://doi.org/10.2777/940154

Ayris, Paul, Berthou, J.-Y., Bruce, R., Lindstaedt, S., Monreale, A., Mons, B., Murayama, Y., Södergård, C., Tochtermann, K., and Wilkinson, R. (2016). *Realising the European Open Science Cloud*. http://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., and Zhao, J. L. (2016). Transformational issues of big data and analytics in networked business. *MIS Quarterly: Management Information Systems*, *40*(4), 807–818. https://doi.org/10.25300/MISQ/2016/40:4.03

Bajpai, V., Brunstrom, A., Feldmann, A., Kellerer, W., Pras, A., Schulzrinne, H., Smaragdakis, G., Wählisch, M., Wehrle, K., Brunstrom, A., Pras, A., Wählisch, M., Feldmann, A., Schulzrinne, H., and Wehrle, K. (2019). The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research. *ACM SIGCOMM Computer Communication Review*, *49*(1), 24–30. https://doi.org/10.1145/3314212.3314217

Baker, M. (2016a). 1,500 scientists lift the lid on reproducibility. *Nature*. https://doi.org/10.1038/533452a

Baker, M. (2016b). Psychology's reproducibility problem is exaggerated – say psychologists. *Nature*. https://doi.org/10.1038/nature.2016.19498

Barba, L. A. (2018). *Terminologies for Reproducible Research*. http://arxiv.org/abs/1802.03311

Baskerville, R., and Pries-heje, J. (2016). *Design Theory Projectability*. 446219–232978. https://doi.org/10.1007/978-3-662

Baxter, G., and Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003

Baykoucheva, S. (2000). Managing research data. In *Managing Scientific Information and Research Data*. Elsevier Ltd. https://doi.org/10.1016/B978-0-08-100195-0.00009-3

Baykoucheva, S. (2015). Managing Scientific Information and Research Data. In *Managing Scientific Information and Research Data*. Elsevier Ltd. https://doi.org/10.1016/B978-0-08-100195-0.00015-9

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., and Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, *29*(2), 599–611. https://doi.org/10.1016/j.future.2011.08.004

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0092590

Benbasat, I., Goldstein, D. K., and Mead, M. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly*, *11*(3), 369. https://doi.org/10.2307/248684

Bergman, M. (2008). *Advances in Mixed Methods Research*. https://doi.org/10.4135/9780857024329

Bider, I., and Perjons, E. (2017). Challenges in assessing parameters of a socio-Technical system. *CEUR Workshop Proceedings*.

The FAIR guiding principles for data stewardship: Fair enough?, European Journal of Human Genetics (2018). https://doi.org/10.1038/s41431-018-0160-0

Boell, S. K., and Cecez-Kecmanovic, D. (2015). What is 'Information' Beyond a Definition? *ICIS 2015 Proceedings*, *Ackoff 1989*, Paper 1363. http://aisel.aisnet.org/icis2015/proceedings/ConferenceTheme/4/

Bogen, J., and Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review*, *97*(3), 303. https://doi.org/10.2307/2185445

Boland, R. J. (1985). Phenomenology: A Preferred Approach to Research on Information Systems. *Research Methods in Information Systems*.

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

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press. https://doi.org/10.1002/asi

Borgman, C. L. (2020). Big Data, Little Data, or No Data? Why Human Interaction with Data is a Hard Problem. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 1–1. https://doi.org/10.1145/3343413.3377979

Borgman, C. L. (2008). Data, disciplines, and scholarly publishing. *Learned Publishing*, *21*(1), 29–38. https://doi.org/10.1087/095315108X254476

Borgman, C. L., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., and Randles, B. M. (2016). Data Management in the Long Tail: Science, Software, and Service. *International Journal of Digital Curation*, *11*(1), 128–149. https://doi.org/10.2218/ijdc.v11i1.428

Bostrom, R. P., and Heinen, J. S. (1977). MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly*, *1*(3), 17. https://doi.org/10.2307/248710

Bouter, L. M., Tijdink, J., Axelsen, N., Martinson, B. C., and ter Riet, G. (2016). Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review*, *1*(1), 1–8. https://doi.org/10.1186/s41073-016-0024-5

Braun, M. L., and Ong, C. S. (2014). Open Science in Machine Learning. In *Implementing Reproducible Research* (p. 343). http://arxiv.org/abs/1402.6013

Brazas, M. D., Brooksbank, C., Jimenez, R. C., Blackford, S., Palagi, P. M., Rivas, J. D. Las, Ouellette, B. F. F., Kumuthini, J., Korpelainen, E., Lewitter, F., Gelder, C. W. G. van, Mulder, N., Corpas, M., Schneider, M. V., Tan, T. W., Clements, D., Davies, A., and Attwood, T. K. (2017). A global perspective on bioinformatics training needs. *BioRxiv*, 098996. https://doi.org/10.1101/098996

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B. D., Nabrzyski, J., Stodden, V., Taylor, I. J., Turk, M. J., and Turner, K. (2019). Computing environments for reproducibility: Capturing the "Whole Tale." *Future Generation Computer Systems*, *94*, 854–867. https://doi.org/10.1016/j.future.2017.12.029

Brown, S., Dennis, A., Samuel, B., Tan, B., Valacich, J., and Whitley, E. (2016). Replication research: Opportunities, experiences and challenges. *ICIS 2016 Proceedings*. https://aisel.aisnet.org/icis2016/Panels/Presentations/3

Buchholz, F., and Spafford, E. (2004). On the role of file system metadata in digital forensics. In *Digital Investigation* (Vol. 1, Issue 4, pp. 298–309). https://doi.org/10.1016/j.diin.2004.10.002

Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., and Schouppe, M. (2019). Open science, open data and open scholarship: European policies to make science fit for the 21st century. *Frontiers in Big Data*, *2*, 43. https://doi.org/10.3389/FDATA.2019.00043

Burley, R. (2017). *How to Fix Peer Review*. https://blogs.scientificamerican.com/observations/how-to-fix-peer-review/

Burton-Jones, A., Storey, V. C., Sugumaran, V., and Ahluwalia, P. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data and Knowledge Engineering*, *55*(1), 84–102. https://doi.org/10.1016/j.datak.2004.11.010

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Vo, T., and Silva, H. T. (2006). VisTrails : Visualization meets Data Management. *2006 ACM SIGMOD IInternational Conference on Management of Data*, 745–747. https://doi.org/10.1145/1142473.1142574

Casadevall, A., and Fang, F. C. (2010). Reproducible science. *Infection and Immunity*, *78*(12), 4972–4975. https://doi.org/10.1128/IAI.00908-10

Casey, E., Katz, G., and Lewthwaite, J. (2013). Honing digital forensic processes. *Digital Investigation*, *10*(2), 138–147. https://doi.org/10.1016/j.diin.2013.07.002

Cervo, D., and Allen, M. (2011). *Master Data Management in Practice: Achieving True Customer MDM*. https://books.google.nl/books?hl=en&lr=&id=PCpWjwq_gFcC&oi=fnd&pg=PA111&dq=master+data+management+allen+cervo&ots=TFUiwh7Ax_&sig=bs-KdwtTY7CRGY3FXsjqq3YbTGA

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.-C., Hagemann, M., Harnad, S., Kupryte, R., Johnson, R., Manna, M. La, Rév, I., Segbert, M., Souza, S. de, Suber, P., and Velterop, J. (2002). Budapest Open Access Initiative (BOAI). *Interlending & Document Supply*, *30*(2), ilds.2002.12230bab.012. https://doi.org/10.1108/ilds.2002.12230bab.012

Chesbrough, H. (2012). Open Innovation: Where We've Been and Where We're Going. *Research-Technology Management*, *55*(4), 20–27. https://doi.org/10.5437/08956308X5504085

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Bras, Y. Le, Lemoine, F., Mareuil, F., Ménager, H., Pradal, C., and Blanchet, C. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, *75*, 284–298. https://doi.org/10.1016/j.future.2017.01.012

Colaert, N., Barsnes, H., Vaudel, M., Helsens, K., Timmerman, E., Sickmann, A., Gevaert, K., and Martens, L. (2011). Thermo-msf-parser: An open source Java library to parse and visualize Thermo Proteome Discoverer msf files. *Journal of Proteome Research*, *10*(8), 3840–3843. https://doi.org/10.1021/pr2005154

Collberg, C., and Proebsting, T. A. (2016). Repeatability in computer systems research. *Communications of the ACM*, *59*(3), 62–69. https://doi.org/10.1145/2812803

Collins, F. S., Lander, E. S., Rogers, J., and Waterson, R. H. (2004). Finishing the euchromatic sequence of the human genome. *Nature*. https://doi.org/10.1038/nature03001

Corti, L., Eynden, V. Van den, Bishop, L., and Woollard, M. (2014). *Managing and sharing research data: A guide to good practice*.

Cox, A. M., and Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, *70*(2), 142–157. https://doi.org/10.1108/AJIM-11-2017-0251

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

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The sequence of the human genome. *Science*. https://doi.org/10.1126/science.1058040

Creswell, J. (2013). Research design: Qualitative, quantitative, and mixed methods approaches. Sage.

Cronholm, S., and Göbel, H. (2019). Evaluation of Action Design Research. *Scandinavian Journal of Information Systems*, *31*(2). https://aisel.aisnet.org/sjis/vol31/iss2/2

Crowston, K., and Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the ASIST Annual Meeting*, *48*(1), 1–9. https://doi.org/10.1002/meet.2011.14504801036

D'Ippolito, B., and Rüling, C.-C. (2019). Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications. *Research Policy*, *48*(5), 1282–1296. https://doi.org/10.1016/j.respol.2019.01.011

Dalkir, K. (2005). Knowledge Management in Theory and Practice. In *Knowledge Management* (Vol. 4).

DAMA-DMBOK. (2009). The DAMA Guide to The Data Management Body of Knowledge. In *Technics Publications, LLC Post*. https://doi.org/10.1161/CIRCULATIONAHA.108.834176

Dennis, A. R., and Valacich, J. S. (2014). A Replication Manifesto. *AIS Transactions on Replication Research*, *1*(1), 1–5. https://doi.org/10.17705/1atrr.00001

Dietrich, D., Adamus, T., Miner, A., and Steinhart, G. (2012). De-mystifying the data management requirements of research funders. *Issues in Science and Technology Librarianship*, *70*. https://doi.org/10.5062/F44M92G2

Directorate-general for Research and Innnovation. (2016). *Open Research Data as the default: Frequently Asked Questions about the exte nsion of the Open Research Data Pilot*. https://ec.europa.eu/research/openscience/pdf/openaccess/ord_extension_faqs.pdf

Doeleman, S. (2019). *Focus on the First Event Horizon Telescope Results*. The Astrophysical Journal Letters. https://iopscience.iop.org/journal/2041-8205/page/Focus_on_EHT

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics (Oxford, England)*, *11*(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028

Earley, S., Henderson, D., and Association., D. M. (2017). *DAMA-DMBOK : data management body of knowledge*. Technics publications.

Editorial. (2014). Journals unite for reproducibility. *Nature*, *515*, 7. https://doi.org/10.1038/nature13193

European Commission. (2015). *Access to and preservation of scientific information in Europe* (Issue May). https://doi.org/10.2777/975917

European Commission. (2016a). EU Open Innovation, Open Science, Open to the World. In *European Comission*. https://doi.org/10.2777/061652

European Commission. (2016b). *Guidelines on Data Management in Horizon 2020*. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. https://doi.org/10.1145/240455.240464

Fecher, B., and Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. In *Opening Science* (pp. 17–47). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2

Fecher, B., Friesike, S., and Hebing, M. (2015). What drives academic data sharing? *PLoS ONE*, *10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., and Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Feger, S. S., Dallmeier-Tiessen, S., Woźniak, P. W., and Schmidt, A. (2019). Gamification in science: A study of requirements in the context of reproducible research. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300690

Ferro, N., and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *ACM SIGIR Forum*, *52*(1), 4–10. https://doi.org/10.1145/3274784.3274786

Franklin, A., and Perovic, S. (2016). Experiment in Physics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Freire, J., Bonnet, P., and Shasha, D. (2012a). Computational reproducibility: state-of-the-art, challenges, and database research opportunities. *Proceedings of the 2012 ACM SIGMOD …*, 593–596. https://doi.org/10.1145/2213836.2213908

Freire, J., Bonnet, P., and Shasha, D. (2012b). Computational reproducibility. *Proceedings of the 2012 International Conference on Management of Data - SIGMOD '12*, 593. https://doi.org/10.1145/2213836.2213908

Gentleman, R., and Lang, D. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, *16*(1), 1–23. https://doi.org/10.1198/106186007X178663

Gleasure, R., Feller, J., and Flaherty, B. O. (2012). PROCEDURALLY TRANSPARENT DESIGN SCIENCE RESEARCH: A DESIGN PROCESS MODEL. *ICIS*, 1–19. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1168&context=icis2012

Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy Team, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), R86. https://doi.org/10.1186/gb-2010-11-8-r86

Goodhue, D. L., Quillard, J. A., and Rockart, J. F. (1988). Managing the Data Resource: A Contingency Perspective. *MIS Quarterly*, *12*(3), 373–392. https://doi.org/10.2307/249204

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Graves, M. W. (2013). *Digital archaeology: the art and science of digital forensics*. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Digital+Archaeology%3A+The+Art+and+Science+of+Digital+Forensics&btnG=

Gregor, S. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*(2), 337–355. http://misq.org/positioning-and-presenting-design-science-research-for-maximum-impact.html

Gregor, S., and Hevner, A. R. (2013). Positioning and Presenting Design Science for Maximum Impact. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.2753/MIS0742-1222240302

Gregory, K., Cousijn, H., Groth, P., Scharnhorst, A., and Wyatt, S. (2018). Understanding Data Search as a Socio-technical Practice. *Journal of Information Science*, 016555151983718. https://doi.org/10.1177/0165551519837182

Grimes, D. R., Bauch, C. T., and Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science*, *5*(1), 171511. https://doi.org/10.1098/rsos.171511

Haj-Bolouri, A., Purao, S., Rossi, M., and Bernhardsson, L. (2018, November 28). Action design research in practice: Lessons and concerns. *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018*. https://aisel.aisnet.org/ecis2018_rp/131

Harjes, J., Link, A., Weibulat, T., Triebel, D., and Rambold, G. (2020). FAIR digital objects in environmental and life sciences should comprise workflow operation design data and method information for repeatability of study setups and reproducibility of results. *Database*, *2020*, 59. https://doi.org/10.1093/database/baaa059

Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., and Strasser, C. A. (2013). Spatially Explicit Data: Stewardship and Ethical Challenges in Science. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1001634

Hatch, M. J. (2018). *Organization theory: modern, symbolic, and postmodern perspectives* (4th ed.). Oxford University Press. https://doi.org/10.5860/choice.35-3404

Haven, T. L., Tijdink, J. K., Pasman, H. R., Widdershoven, G., ter Riet, G., and Bouter, L. M. (2019). Researchers' perceptions of research misbehaviours: a mixed methods study among academic researchers in Amsterdam. *Research Integrity and Peer Review*, *4*(1), 1–12. https://doi.org/10.1186/s41073-019-0081-7

Heath, H., and Cowley, S. (2004). Developing a grounded theory approach: a comparison of Glaser and Strauss. *International Journal of Nursing Studies*, *41*, 141–150. https://doi.org/10.1016/S0020-7489(03)00113-5

Hedges, M., Hasan, A., and Blanke, T. (2007). Management and preservation of research data with iRODS. *Proceedings of the ACM First Workshop on CyberInfrastructure: Information Management in EScience - CIMS '07*, 17. https://doi.org/10.1145/1317353.1317358

Hesketh, B., and Graco, W. (2015). Technological Change and the Sociotechnical System, Applied Psychology of. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 104–108). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.22018-7

Hevner, A., and Chatterjee, S. (2010). Design Science Research in Information Systems. In *Design Research in Information Systems* (Vol. 22). https://doi.org/10.1007/978-1-4419-5653-8

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2), 87–92. https://doi.org/http://aisel.aisnet.org/sjis/vol19/iss2/4

Hevner, March, Park, Ram, Hevner, A. R., March, S. T., Park, J., Ram, S., Hevner, March, Park, Ram, Land, F., and Clemons, E. K. (2004). Design Science in Information Systems Research. *Information Systems*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Higgins, S. (2008). DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *3*(1), 134–140. https://doi.org/10.2218/ijdc.v2i2.30

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

Hislop, D., Bosua, R., and Helms, R. (2018). *Knowledge management in organizations: A critical introduction* (4th ed.). Oxford University Press.

Hoehndorf, R., Dumontier, M., and Gkoutos, G. V. (2013). Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics*, *14*(6), 696–712. https://doi.org/10.1093/bib/bbs053

Holzinger, A. (2013). Human-Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8127 LNCS*, 319–328. https://doi.org/10.1007/978-3-642-40511-2_22

Holzinger, A., Dehmer, M., and Jurisica, I. (2014). Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. *BMC Bioinformatics*, *15 Suppl 6*(Suppl 6), I1. https://doi.org/10.1186/1471-2105-15-S6-I1

Homolak, J., Kodvanj, I., and Virag, D. (2020). Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders. *Scientometrics*, *124*(3), 2687–2701. https://doi.org/10.1007/s11192-020-03587-2

Hood, W. W., and Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291–314. https://doi.org/10.1023/A:1017919924342

Horton, L., and DCC. (2016). *Overview of UK Institution RDM Policies', Version 6 August 2016, Digital Curation Centre*. http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies

Hothorn, T., and Leisch, F. (2011). Case studies in reproducibility. *Briefings in Bioinformatics*, *12*(3), 288–300. https://doi.org/10.1093/bib/bbq084

Houssos, N., Jörg, B., Dvořák, J., Príncipe, P., Rodrigues, E., Manghi, P., and Elbæk, M. K. (2014). OpenAIRE guidelines for CRIS managers: Supporting interoperability of open research information through established standards. *Procedia Computer Science*, *33*, 33–38. https://doi.org/10.1016/j.procs.2014.06.006

Huang, Y., and Gottardo, R. (2013). Comparability and reproducibility of biomedical data. *Briefings in Bioinformatics*, *14*(4), 391–401. https://doi.org/10.1093/bib/bbs078

Iivari, J., Rotvit Perlt Hansen, M., and Haj-Bolouri, A. (2018). A Framework for Light Reusability Evaluation of Design Principles in Design Science Research. *DESRIST*.

Ince, D. C., Hatton, L., and Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485–488. https://doi.org/10.1038/nature10836

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLOS Biology*, *16*(3), e2005468. https://doi.org/10.1371/journal.pbio.2005468

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., and Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. In *The Lancet* (Vol. 383, Issue 9912, pp. 166–175). https://doi.org/10.1016/S0140-6736(13)62227-8

Jamieson, K. H., McNutt, M., Kiermer, V., and Sever, R. (2019). Signaling the trustworthiness of science. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(39), 19231–19236. https://doi.org/10.1073/pnas.1913039116

Jones, D., Gregor, S., and Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information …*, *8*(5), 312–335. http://search.proquest.com/openview/524ec5f1aa1112a1e717d533fe479f34/1?pq-origsite=gscholar&cbl=26427

Jones, S., Pergl, R., Hooft, R., Miksa, T., Samors, R., Ungvari, J., Davis, R. I., and Lee, T. (2020). Data Management Planning: How Requirements and Solutions are Beginning to Converge. *Data Intelligence*, *2*(1–2), 208–219. https://doi.org/10.1162/dint_a_00043

Jones, S., Pryor, G., and Whyte, A. (2012). Developing Research Data Management Capability : the View from a National Support Service. *IPress*, 142–149. https://www.academia.edu/download/47339409/Addressing_data_management_training_need20160718-30010-bgfjp7.pdf#page=152

Joubert, A., Murawski, M., and Bick, M. (2019). Big Data Readiness Index – Africa in the Age of Analytics. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-030-29374-1_9

Kelchen, R. (2018). Higher education accountability. In *Higher Education Accountability*. https://doi.org/10.1353/book.58123

Keller, E. F. (2003). Models, Simulation, and "Computer Experiments." In H. Radder (Ed.), *The Philosophy Of Scientific Experimentation*. University of Pittsburgh Press,. https://doi.org/10.2307/j.ctt5hjsnf.14

Klein, H. K., and Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly*, *23*(1), 67–93. https://doi.org/10.2307/249410

Klinkenberg, J.-M. (1996). Précis de sémiotique générale. In *Points Essais, 411*.

Knuth, D. E. (1984). Literate Programming. *The Computer Journal*, *27*(2), 97–111. https://doi.org/10.1093/comjnl/27.2.97

Koester, A., Baumann, A., Krasnova, H., Avital, M., Lyytinen, K., and Rossi, M. (2020). Panel 1: To Share or Not To Share: Should IS Researchers Share or Hoard their Precious Data? *ECIS 2020 Select Recordings*. https://aisel.aisnet.org/ecis2020_sessionrecordings/6

Korhonen, J. J., Melleri, I., Hiekkanen, K., and Helenius, M. (2013). Designing Data Governance Structure : An Organizational Perspective. *Journal on Computing*, *2*(4), 11–17. https://doi.org/10.5176/2251-3043

Krogstie, J. (2015). Capturing Enterprise Data Integration Challenges Using a Semiotic Data Quality Framework. *Business and Information Systems Engineering*, *57*(1), 27–36. https://doi.org/10.1007/s12599-014-0365-x

Laine, C., Goodman, S. N., Griswold, M. E., and Sox, H. C. (2007a). Reproducible research: Moving toward research the public can really trust. In *Annals of Internal Medicine* (Vol. 146, Issue 6, pp. 450–453). https://doi.org/10.7326/0003-4819-146-6-200703200-00154

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., van de Sandt, S., Ison, J., Martinez, P. A., McQuilton, P., Valencia, A., Harrow, J., Psomopoulos, F., Gelpi, J. L., Chue Hong, N., Goble, C., and Capella-Gutierrez, S. (2019). Towards FAIR principles for research software. *Data Science*, *Preprint*(Preprint), 1–23. https://doi.org/10.3233/DS-190026

Landelijk Coördinatiepunt Research Data Management. (2019). *Landelijk Coördinatie-punt Research Data Management (lcrdm)*. https://doi.org/10.5281/zenodo.3266833

Latour, B., and Woolgar, S. (1986). *Laboratory life the construction of scientific facts*. Princeton University Press.

Le Maux, B., Necker, S., and Rocaboy, Y. (2019). Cheat or perish? A theory of scientific customs. *Research Policy*. https://doi.org/10.1016/J.RESPOL.2019.05.001

Leavitt, H. J. (1965). Applied organisational change in industry: Structural, technological and humanistic approaches. In *Handbook of organizations.*

Lefebvre, A. (2020). *Open science readiness dashboard*. https://doi.org/10.5281/ZENODO.4020379

Lefebvre, A., Berendsen, J., and Spruit, M. (2019). Evaluation of classification models for retrieving experimental sections from full-text publications (Issue UU-CS-2019-002).

Lefebvre, A., and Bruin, J. de. (2019). *Path2Insight : A file path analysis toolkit for laboratory forensics*. https://doi.org/10.5281/ZENODO.3518815

Lefebvre, A., Schermerhorn, E., and Spruit, M. (2018). How Research Data Management Can Contribute to Efficient and Reliable Science. *The 25th European Conference of Information Systems*.

Lefebvre, A., and Spruit, M. (2019a). Designing Laboratory Forensics. In *18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019* (Vol. 11701, pp. 238–251). Springer. https://doi.org/10.1007/978-3-030-29374-1_20

Lefebvre, A., and Spruit, M. (2019b). Designing Laboratory Forensics. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11701 LNCS, 238–251. https://doi.org/10.1007/978-3-030-29374-1_20

Lefebvre, A., Spruit, M., and Omta, W. (2015). Towards Reusability of Computational Experiments - Capturing and Sharing Research Objects from Knowledge Discovery Processes. *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, *1*, 456–462. https://doi.org/10.5220/0005631604560462

Lenzerini, M. (2011). Ontology-based data management. Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11, 5. https://doi.org/10.1145/2063576.2063582

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 503–514. https://doi.org/10.1016/j.shpsc.2013.03.020

Levelt, W. J. (2012). Falende wetenschap: De frauduleuze onderzoekspraktijken van sociaal-psycholoog Diederik Stapel Commissie Levelt Commissie Noort Commissie Drenth.

Link, G. J. P., Lumbard, K., Conboy, K., Feldman, M., Feller, J., George, J., Germonprez, M., Goggins, S., Jeske, D., Kiely, G., Schuster, K., and Willis, M. (2017). Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations. *Communications of the Association for Information Systems*, *41*(1), 587–610. https://doi.org/10.17705/1cais.04125

Lnenicka, M., and Komarkova, J. (2019). Big and open linked data analytics ecosystem: Theoretical background and essential elements. *Government Information Quarterly*, *36*(1), 129–144. https://doi.org/10.1016/j.giq.2018.11.004

Lukyanenko, R., Wiggins, A., and Rosser, H. K. (2020). Citizen Science: An Information Quality Research Frontier. *Information Systems Frontiers*, *22*(4), 961–983. https://doi.org/10.1007/s10796-019-09915-z

Lyytinen, K., and Newman, M. (2008). Explaining information systems change: A punctuated socio-technical change model. *European Journal of Information Systems*. https://doi.org/10.1057/ejis.2008.50

Mabey, M., Doupé, A., Zhao, Z., and Ahn, G. J. (2018). Challenges, opportunities and a framework for web environment forensics. *IFIP Advances in Information and Communication Technology*, *532*, 11–33. https://doi.org/10.1007/978-3-319-99277-8_2

Manieri, A., Brewer, S., Riestra, R., Demchenko, Y., Hemmje, M., Wiktorski, T., Ferrari, T., and Frey, J. (2016). Data science professional uncovered: How the EDISON project will contribute to a widely accepted profile for data scientists. *Proceedings - IEEE 7th International Conference on Cloud Computing Technology and Science, CloudCom 2015*, 588–593. https://doi.org/10.1109/CloudCom.2015.57

Mannheimer, S. (2018). Toward a Better Data Management Plan: The Impact of DMPs on Grant Funded Research Practices. *Journal of EScience Librarianship*, *7*(3), e1155. https://doi.org/10.7191/jeslib.2018.1155

Mannheimer, S., Sterman, L. B., and Borda, S. (2016). *Discovery and Reuse of Open Datasets: An Exploratory Study*. *5*, e1091. https://doi.org/10.7191/jeslib.2016.1091

Mayer, G., Jones, A. R., Binz, P. A., Deutsch, E. W., Orchard, S., Montecchi-Palazzi, L., Vizcaíno, J. A., Hermjakob, H., Oveillero, D., Julian, R., Stephan, C., Meyer, H. E., and Eisenacher, M. (2014). Controlled vocabularies and ontologies in proteomics: Overview, principles and practice. *Biochimica et Biophysica Acta - Proteins and Proteomics*. https://doi.org/10.1016/j.bbapap.2013.02.017

McCormick, M., Liu, X., Jomier, J., Marion, C., and Ibanez, L. (2014). ITK: enabling reproducible research and open science. *Frontiers in Neuroinformatics*, *8*(February), 13. https://doi.org/10.3389/fninf.2014.00013

McNutt, M. (2014). Journals unite for reproducibility. *Science*, *346*(6210), 679–679.

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., and Sansone, S. A. (2016). BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database : The Journal of Biological Databases and Curation*. https://doi.org/10.1093/database/baw075

Mehra, M. R., Ruschitzka, F., and Patel, A. N. (2020). Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis (The Lancet, (S0140673620311806), (10.1016/S0140-6736(20)31180-6)). In *The Lancet* (Vol. 395, Issue 10240, p. 1820). Lancet Publishing Group. https://doi.org/10.1016/S0140-6736(20)31324-6

Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan. *PLOS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004525

Mikalef, P., Pappas, I. O., Krogstie, J., and Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and E-Business Management*, *16*(3), 547–578. https://doi.org/10.1007/s10257-017-0362-y

Miksa, T., Mayer, R., Strodl, S., Rauber, A., Vieira, R., and Antunes, G. (2014). Risk Driven Selection of Preservation Activities for Increasing Sustainability of Open Source Systems and Workflows. *IPres 2014*, 91–100. http://www.ifs.tuwien.ac.at/~mayer/publications/pdf/mik_ipres14-riskDriven.pdf

Miller, A. N., Taylor, S. G., and Bedeian, A. G. (2011). Publish or perish: academic life as management faculty live it. *Career Development International*, *16*(5), 422–445. https://doi.org/10.1108/13620431111167751

Moonesinghe, R., Khoury, M., medicine, A. J.-Pl., and 2007, undefined. (2007). Most published research findings are false—but a little replication goes a long way. *Journals.Plos.Org*, *4*(2), e28. https://doi.org/10.1371/journal.pmed.0040028

Mosley, M., Brackett, M., Earley, S., and Henderson, D. (2010). *DAMA guide to the data management body of knowledge*. Technics Publications. http://agris.fao.org/agris-search/search.do?recordID=US201300002206

Mumford, E. (2006). The story of socio-technical design: Reflections on its successes, failures and potential. In *Information Systems Journal* (Vol. 16, Issue 4, pp. 317–342). John Wiley & Sons, Ltd (10.1111). https://doi.org/10.1111/j.1365-2575.2006.00221.x

Munafò, M. R., Hollands, G. J., and Marteau, T. M. (2018). Open science prevents mindless science. In *BMJ (Online)*. https://doi.org/10.1136/bmj.k4309

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., and Ioannidis, J. P. A. (2017). A manifesto for reproducible science. In *Nature Human Behaviour* (Vol. 1, Issue 1, p. 0021). Nature Publishing Group. https://doi.org/10.1038/s41562-016-0021

Myers, M. D. (1997). Qualitative Research in Information Systems | Serving society in the advancement of knowledge and excellence in the study and profession of information systems. 241–242. https://www.qual.auckland.ac.nz/

Napolitano, F. (2017). repo: an R package for data-centered management of bioinformatic pipelines. *BMC Bioinformatics*, *18*(1), 112. https://doi.org/10.1186/s12859-017-1510-6

NAS. (2018). *Open Science by Design: Realizing a Vision for 21st Century Research*. The National Academies Press. https://doi.org/10.17226/25116

Niederman, F., and March, S. (2015). Reflections on Replications. *AIS Transactions on Replication Research*, *1*(1), 1–16.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. In *Science*. https://doi.org/10.1126/science.aab2374

Nöth, W. (1990). *Handbook of Semiotics*. Indiana University Press.

November, J. (2012). Biomedical computing: Digitizing life in the United States. In *Biomedical Computing: Digitizing Life in the United States*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84905308797&partnerID=40&md5=69be3d57225a27d7af5ea4db4132deaf

NWO. (2017). *Data management protocol*. https://www.nwo.nl/en/policies/open+science/data+management

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

OED Online. (2019). *"reproducibility, n."*. Oxford University Press. https://www.oed.com/view/Entry/163100?redirectedFrom=reproducibility

Olbrich, S., Frank, U., Gregor, S., Louis, S., Rowe, F., Niederman, F., and Rowe, F. (2017). On the Merits and Limits of Replication and Negation for IS Research. *Transactions on Replication Research*, *3*(March), 1–19. https://doi.org/10.17705/1atrr.00016

Open Science - Utrecht University. (2020). https://www.uu.nl/en/research/open-science

Otto, B. (2011). A MORPHOLOGY OF THE ORGANISATION OF DATA GOVERNANCE. In *ECIS 2011 Proceedings*. https://doi.org/10.1007/978-3-8348-9953-8

Palmer, G. (2001). A Road Map for Digital Forensic Research. *Proceedings of the 2001 Digital Forensics Research Workshop Conference*. https://doi.org/10.1111/j.1365-2656.2005.01025.x

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/mis0742-1222240302

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Peng, R. D., Dominici, F., and Zeger, S. L. (2006). Reproducible epidemiologic research. In *American Journal of Epidemiology* (Vol. 163, Issue 9, pp. 783–789). https://doi.org/10.1093/aje/kwj093

Peng, R. D., and Hicks, S. C. (2020). *Reproducible Research: A Retrospective*. http://arxiv.org/abs/2007.12210

Pérez, F., and Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, *9*, 21–29. https://doi.org/10.1109/MCSE.2007.53

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*, 76. https://doi.org/10.3389/fninf.2017.00076

Plomp, E., Dintzner, N., Teperek, M., and Dunning, A. (2019). Cultural obstacles to research data management and sharing at TU Delft. *Insights: The UKSG Journal*, *32*. https://doi.org/10.1629/uksg.484

Plotkin, D., and David. (2013). *Data stewardship : an actionable guide to effective data management and data governance*. Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=2737837

Ponelis, S. R. (2015). Using interpretive qualitative case studies for exploratory research in doctoral studies: A case of information systems research in small and medium enterprises. *International Journal of Doctoral Studies*, *10*, 535–550. https://doi.org/10.28945/2339

Ponte, D. (2015). Enabling an Open Data Ecosystem. *ECIS 2015 Research-in-Progress Papers*. https://aisel.aisnet.org/ecis2015_rip/55

Prager, E. M., Chambers, K. E., Plotkin, J. L., McArthur, D. L., Bandrowski, A. E., Bansal, N., Martone, M. E., Bergstrom, H. C., Bespalov, A., and Graf, C. (2019). Improving transparency and scientific rigor in academic publishing. *Journal of Neuroscience Research*, *97*(4), 377–390. https://doi.org/10.1002/jnr.24340

Prost, H., and Schöpfel, J. (2015). *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3.* http://hal.univ-lille3.fr/hal-01198379

Pryor, G. (2012). *Managing research data*. Facet Publishing.

Radder, H. (2012a). *Experimentation in the Natural Sciences* (pp. 53–71). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4107-2_3

Radder, H. (2012b). *The Material Realization of Science* (Vol. 294). Springer Netherlands. https://doi.org/10.1007/978-94-007-4107-2

Radder, H. (1992). Experimental Reproducibility and the Experimenters' Regress. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1992*, 63–73.

Rahimzadeh, V., and Bartlett, G. (2014). Genetics and primary care: where are we headed? *Journal of Translational Medicine*, *12*(1), 238. https://doi.org/10.1186/s12967-014-0238-6

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., and Wilkinson, M. (2011). Report on Integration of Data and Publications. *Report on Integration of Data and Publications*, *5*, 1–87. https://doi.org/10.5281/zenodo.8307

Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press. https://philpapers.org/rec/RHETAH

Rheinberger, H. (1997). Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube (Writing Science). *Stanford University Press*.

Ribes, D., and Polk, J. B. (2014). Flexibility relative to what? Change to research infrastructure. *Journal of the Association of Information Systems*. https://doi.org/10.17705/1jais.00360

Robinson-García, N., Jiménez-Contreras, E., and Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964–2975. https://doi.org/10.1002/asi.23529

Rouse, J. (2011). Articulating the World: Experimental Systems and Conceptual Understanding. *International Studies in the Philosophy of Science*, *25*(3), 243–254. https://doi.org/10.1080/02698595.2011.605246

Rowlingson, R. (2004). A Ten Step Process for Forensic Readiness. *International Journal of Digital Evidence*, *2*(3). https://doi.org/10.1162/NECO_a_00266

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology*, *9*(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Santana-Perez, I., Ferreira da Silva, R., Rynge, M., Deelman, E., Pérez-Hernández, M. S., and Corcho, O. (2017). Reproducibility of execution environments in computational science using Semantics and Clouds. *Future Generation Computer Systems*, *67*, 354–367. https://doi.org/10.1016/J.FUTURE.2015.12.017

Schloss, P. D. (2018). Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *MBio*, *9*(3), e00525-18. https://doi.org/10.1128/mbio.00525-18

Schoonen, W. (2020, February 15). Fraude en wangedrag: de wetenschap worstelt ermee | Trouw. *Trouw*. https://www.trouw.nl/wetenschap/fraude-en-wangedrag-de-wetenschap-worstelt-ermee~b1e3a98a/

Science Europe. (2018). Practical Guide to the International Alignment of Research Data Management.

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., and Lindgren, R. (2011). Action design research. *MIS Quarterly: Management Information Systems*. http://bada.hb.se/handle/2320/9888

Serketzis, N., Katos, V., Ilioudis, C., Baltatzis, D., and Pangalos, G. J. (2019). Actionable threat intelligence for digital forensics readiness. *Information and Computer Security*, *27*(2), 273–291. https://doi.org/10.1108/ICS-09-2018-0110

Shanks, G. (1997). The challenges of strategic data planning in practice: An interpretive case study. *Journal of Strategic Information Systems*. https://doi.org/10.1016/S0963-8687(96)01053-0

Sholler, D., Ram, K., Boettiger, C., and Katz, D. S. (2019). Enforcing public data archiving policies in academic publishing: A study of ecology journals. *Big Data and Society*, *6*(1), 205395171983625. https://doi.org/10.1177/2053951719836258

Shoshani, A., and Rotem, D. (2009). Scientific Data Management. In A. Shoshani and D. Rotem (Eds.), *Scientific Data Management: Challenges, Technology, and Deployment*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420069815

Silvello, G. (2018). Theory and practice of data citation. In *Journal of the Association for Information Science and Technology* (Vol. 69, Issue 1, pp. 6–20). John Wiley and Sons Inc. https://doi.org/10.1002/asi.23917

Silver, M. S., and Markus, M. L. (2013). Conceptualizing the SocioTechnical (ST) Artifact. In *Signs & Actions An International Journal on Information Technology, Action, Communication and Workpractices* (Vol. 7, Issue 1). http://www.sysiac.org/

Simms, S., Strong, M., Jones, S., and Ribeiro, M. (2016). The Future of Data Management Planning : Tools , Policies , and Players. *International Digital Curation Conference (IDCC16)*, *February 22-25*, 10.

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Simou, S., Kalloniatis, C., Gritzalis, S., and Katos, V. (2019). A framework for designing cloud forensic-enabled services (CFeS). *Requirements Engineering*, *24*(3), 403–430. https://doi.org/10.1007/s00766-018-0289-y

Sommerville, I. (2016). Software engineering (10th edition). In *Pearson Education Limited*.

Soylu, A., Elvesæter, B., Turk, P., Roman, D., Corcho, O., Simperl, E., Konstantinidis, G., and Lech, T. C. (2019). Towards an Ontology for Public Procurement Based on the Open Contracting Data Standard. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11701 LNCS*, 230–237. https://doi.org/10.1007/978-3-030-29374-1_19

Stamper, R., Liu, K., Hafkamp, M., and Ades, Y. (2000). Understanding the roles of signs and norms in organizations – a semiotic approach to information systems design. *Behaviour and Information Technology*, *19*(1), 15–27. https://doi.org/10.1080/014492900118768

Steen, R. G., Casadevall, A., and Fang, F. C. (2013). Why has the number of scientific retractions increased? *PloS One*, *8*(7), e68397.

Stevens, H. (2013). *Life out of sequence: a data-driven history of bioinformatics*. Univeristy of Chicago Press. https://doi.org/10.1080/14636778.2015.1025127

Stodden, V, Leisch, F., Peng, R., Millman, K. J., Pérez, F., Stodden, V., Leisch, F., and Peng, R. (2014). Implementing reproducible research. *Journal of Statistical Software*, *61*(October), 149–184. https://doi.org/10.1201/b16868

Stodden, Victoria, Ferrini, V., Gabanyi, M., Lehnert, K., Morton, J., and Berman, H. (2019). Open access to research artifacts: Implementing the next generation data management plan. *Proceedings of the Association for Information Science and Technology*, *56*(1), 481–485. https://doi.org/10.1002/pra2.51

Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A. J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L., Carr, S. A., Clauser, K. R., Jaffe, J. D., Kowalski, K. A., Neubert, T. A., Regnier, F. E., Schilling, B., Tegeler, T. J., Wang, M., … Spiegelman, C. (2010). Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research*, *9*(2), 761–776. https://doi.org/10.1021/pr9006365

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., and Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS One*, *6*(6), e0118053. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Sandusky, R. J., Allard, S., and Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library and Information Science Research*, *36*(2), 84–90. https://doi.org/10.1016/j.lisr.2013.11.003

Thellefsen, M. M., Thellefsen, T., and Sørensen, B. (2018). Information as signs: A semiotic analysis of the information concept, determining its ontological and epistemological foundations. *Journal of Documentation*. https://doi.org/10.1108/JD-05-2017-0078

Thompson, M., Burger, K., Kaliyaperumal, R., Roos, M., and da Silva Santos, L. O. B. (2020). Making FAIR Easy with FAIR Tools: From Creolization to Convergence. *Data Intelligence*, *2*(1–2), 87–95. https://doi.org/10.1162/dint_a_00031

Thuan, N. H., Drechsler, A., and Antunes, P. (2019). Construction of Design Science Research Questions. *Communications of the Association for Information Systems*, *44*(March), 332–363. https://doi.org/10.17705/1CAIS.04420

Treloar, A. (2014). The research data alliance: Globally co-ordinated action against barriers to data publishing and sharing. *Learned Publishing*, *27*(5), S9–S13. https://doi.org/10.1087/20140503

Tremblay, M. C., Hevner, A. R., and Berndt, D. J. (2010). The Use of Focus Groups in Design Science Research. In *Design Research in Information Systems* (pp. 121–143). https://doi.org/10.1007/978-1-4419-5653-8_10

Tsai, A. C., Kohrt, B. A., Matthews, L. T., Betancourt, T. S., Lee, J. K., Papachristos, A. V., Weiser, S. D., and Dworkin, S. L. (2016). Promises and pitfalls of data sharing in qualitative research. In *Social Science & Medicine*. https://doi.org/10.1016/j.socscimed.2016.08.004

van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A. O., Kirkpatrick, C., and Mons, B. (2019). Towards the Tipping Point for FAIR Implementation. *Data Intelligence*, 264–275. https://doi.org/10.1162/dint_a_00049

Venugopal, S., Buyya, R., and Ramamohanarao, K. (2006). A taxonomy of Data Grids for distributed data sharing, management, and processing. *ACM Computing Surveys*, *38*(1), 3-es. https://doi.org/10.1145/1132952.1132955

Vicente-Saez, R., and Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

VSNU. (2019). *Room for ev eryone's talent towards a new balance in the recognition and rewards of academics*. https://www.nwo.nl/en/news-and-events/news/2019/11/knowledge-sector-sector-takes-major-step-forward-in-new-approach-to-recognising-and-rewarding-academics.html

Vuong, Q.-H., La, V.-P., Ho, M.-T., Vuong, T.-T., and Ho, M.-T. (2020). Characteristics of retracted articles based on retraction data from online sources through February 2019. *Science Editing*, *7*(1), 34–44. https://doi.org/10.6087/kcse.187

Wallis, J. C., Rolando, E., and Borgman, C. L. (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), e67332. https://doi.org/10.1371/journal.pone.0067332

Weber, M. (2018). Experiment in Biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University.

Wende, K., and Otto, B. (2007). A contingency approach to data governance. *Proceedings, 12th International Conference on Information Quality (ICIQ-07), Cambridge, USA*, 14. https://doi.org/10.1002/nml.241

Wieringa, R. J. (2014). Design Science Methodology for Information Systems and Software Engineering. In *Springer Berlin Heidelberg*. Springer Heilderber New York Dordrecht London. https://doi.org/10.1145/1810295.1810446

Wiesche, M., and Yetton, P. W. (2017). GROUNDED THEORY METHODOLOGY IN INFORMATION SYSTEMS RESEARCH. *MIS Quarterly*, *41*(3), 685–701. http://www.misq.org

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

Williams, C. L., Casadevall, A., and Jackson, S. (2019). Figure errors, sloppy science, and fraud: keeping eyes on your data. *Journal of Clinical Investigation*, *129*(5), 1805–1807. https://doi.org/10.1172/JCI128380

Williams, M., Bagwell, J., and Nahm Zozus, M. (2017). Data management plans: the missing perspective. *Journal of Biomedical Informatics*, *71*, 130–142. https://doi.org/10.1016/j.jbi.2017.05.004

Wilms, K., Brenger, B., Lopez, A., and Rehwald, S. (2018). Open Data in Higher Education – What Prevents Researchers from Sharing Research Data? *39th International Conference on Information Systems (ICIS) 2018*, 1–9. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1091&context=icis2018

Wilms, K., Stieglitz, S., Buchholz, A., Vogl, R., and Rudolph, D. (2018). Do Researchers Dream of Research Data Management? *Proceedings of the 51st Hawaii International Conference on System Sciences*, 4411–4420. http://hdl.handle.net/10125/50445

Wouters, P., Ràfols, I., Oancea, A., Caroline, S., Kamerlin, L., Britt, J., and Jacob, M. (2019). *Indicator Frameworks for Fostering Open Knowledge Practices in Science and Scholarship*. https://doi.org/10.2777/445286

Yarborough, M., Nadon, R., and Karlin, D. G. (2019). Point of View: Four erroneous beliefs thwarting more trustworthy research. *ELife*. https://doi.org/10.7554/elife.45261

Yin, R. K. (2009). Case study research : design and methods. In *Applied social research methods series ; Vol. 5.* https://doi.org/10.1097/FCH.0b013e31822dda9e

Zeleti, F. A., and Ojo, A. (2017). Open data value capability architecture. *Information Systems Frontiers*, *19*(2), 337–360. https://doi.org/10.1007/s10796-016-9711-5

Zillner, S., Curry, E., Metzger, A., Seidl, R., Garcia Robles, A., Keneally, J., Perez, M., Hasan, S., Czech, P., Dalle Carbonare, D., Scerri, S., Hahn, T., Södergård, C., Piscitelli, R., Legré, Y., Canet, G., Tonna, C., Walshe, R., Berre, A., … Menasalvas, E. (2017). *European Big Data Value Strategic Research and Innovation Agenda*. http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf

Zondergeld, J. J., Scholten, R. H. H., Vreede, B. M. I., Hessels, R. S., Pijl, A. G., Buizer-Voskamp, J. E., Rasch, M., Lange, O. A., and Veldkamp, C. L. S. (2020). FAIR, safe and high-quality data: the data infrastructure and accessibility of the YOUth cohort study. *Developmental Cognitive Neuroscience*, 100834. https://doi.org/10.1016/j.dcn.2020.100834

Published work

This dissertation includes a series of scientific software, published scientific papers and articles, as listed below.

Lefebvre A., & de Bruin J. (2019, October 25). Path2Insight: A file path analysis toolkit for laboratory forensics (Version zenodo). Zenodo. http://doi.org/10.5281/zenodo.3518815

Lefebvre A., & Spruit M. (submitted). Open science readiness: a research agenda for research data management (*currently under revision for journal publication)*

Lefebvre A., Spruit M., and Omta W., "Towards reusability of computational experiments Capturing and sharing Research Objects from knowledge discovery processes," in *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015)*, 2015, vol. 1, pp. 456–462.

Lefebvre, A., & Spruit, M. (2019a). Designing Laboratory Forensics. In I. O. Pappas, P. Mikalef, Y. K. Dwivedi, L. Jaccheri, & J. Krogstie (Eds.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 11701). https://doi.org/10.1007/978-3-030-29374-1_20.

Lefebvre, A., & Spruit, M. (2019b). A Socio-technical Approach to Reproducibility in Research Data Management. In A. Arenas, L. Mola, & E. D. Zamani (Eds.), the 13th Mediterranean Conference on Information Systems (MCIS 2019). AIS Electronic Library (AISeL).

Lefebvre, A., Bakhtiari, B., & Spruit, M. (2020). Exploring research data management planning challenges in practice. It - Information Technology, 62(1), 29–37. https://doi.org/10.1515/itit-2019-0029

Lefebvre, A., Berendsen, J., & Spruit, M. (2019): Evaluation of classification models for retrieving experimental sections from full-text publications. CS Technical report UUCS2019002

Lefebvre, A., Schermerhorn, E., & Spruit, M. (2018). How Research Data Management Can Contribute to Efficient and Reliable Science. *The 25th European Conference of Information Systems*. Portsmouth.

English Summary

This dissertation investigates research data management practices in laboratories in the context of open science. To achieve that, we first seek to understand what kind of organizational and technological issues are impeding the planning, production, preservation, and dissemination of data in laboratories. Then, we propose a conceptualization of laboratory work using the lens of experimental systems theory, which provides a socio-technical perspective on the building blocks of scientific experimentation. Finally, we apply the lens of reproducible experimental systems further to design a laboratory forensics approach for investigating storage systems in laboratories. The laboratory forensics approach is a starting point of experimental resources discovery and evaluation in labs. Next, we draw upon the results of forensic investigations to shape open science readiness, which is an ensemble of RDM practices and technology that supports reproducible and open practices in laboratories. The goal of pursuing the design of open science readiness for laboratories is to foster evidence-based research data management that effectively achieves the preservation and dissemination of research data in an open and FAIR way. Therefore, the main research question of this dissertation is stated as follows:

How can we organize research data management for preserving and disseminating laboratory experiments in a reproducible way?

First, we start with organizational and technological issues among stakeholders involved in research data management. First, we examine the cooperation between researchers and data managers. By doing so, an agenda for open data in academia is proposed based on qualitative research highlighting issues such as lack of proper infrastructure, accountability, legal frameworks, and rewards in research data management. At the same time, new roles such as data stewards and the struggles with data management support are investigated. To further determine stakeholders' needs and practices, a similar exploratory approach is used to discover how funding agencies and data management support develop a research data strategy in the Netherlands.

Then, we elaborate on the concept of reproducibility in experimental science. To achieve that, we dive into data management issues from a technological point of view, showing what types of reproducibility issues occur in storage systems with laboratory forensics techniques. Moreover, we investigate reproducibility in research data management by mapping laboratory work and the scholarly infrastructure to a socio-technical model. As such, we obtain a more comprehensive view of reproducibility issues and refine organizational and technical aspects of reproducibility challenges in practice.

Finally, we illustrate some applications of “FAIR technology”. First, we show the need for designing reproducible and reusable research software with the reproducible, research-oriented knowledge discovery in databases process (RRO-KDD). Then we present a strategy for open science readiness. The results of this work provide research laboratories and other stakeholders such as libraries, ICT, and funders with insights into reproducibility and open science challenges grounded into an investigation of laboratory work.

Samenvatting in het Nederlands

Dit proefschrift onderzoekt de praktijken van onderzoeksdatabeheer *('research data management')* in laboratoria in de context van open wetenschap *('open science')*. Om dat te bereiken, onderzoeken we eerst wat voor soort organisatorische en technologische problemen de planning, productie, bewaring en ontsluiting van data in laboratoria belemmeren. Vervolgens stellen we een conceptualisering van het laboratoriumwerk voor door gebruik te maken van de lens van de experimentele systeemtheorie, die een socio-technisch perspectief biedt op de bouwstenen van wetenschappelijke experimenten. Ten slotte passen we de lens van reproduceerbare experimentele systemen toe om een forensische laboratoriumbenadering *('laboratory forensics')* te ontwerpen voor het onderzoeken van opslagsystemen in laboratoria. De forensische laboratoriumbenadering is een startpunt voor het ontdekken en evalueren van experimentele bronnen in laboratoria. Vervolgens putten we uit de resultaten van forensisch onderzoek om de bereidheid van open wetenschap vorm te geven, een geheel van onderzoeksdatabeheerpraktijken en -technologie die reproduceerbare en open praktijken in laboratoria ondersteunt. Het doel van het ontwerp van open wetenschap-paraatheid *('open science readiness')* voor laboratoria is het bevorderen van op bewijs gebaseerd beheer van onderzoeksdata waarmee op effectieve wijze de bewaring en verspreiding van onderzoeksdata op een open en FAIRe (*i.e.* vindbare, toegankelijke, interoperabele en herbruikbare) manier wordt bereikt. De hoofdonderzoeksvraag van dit proefschrift is daarom als volgt:

"Hoe kunnen we het beheer van onderzoeksdata organiseren om laboratoriumexperimenten op een reproduceerbare manier te bewaren en te ontsluiten?"

Allereerst beginnen we met organisatorische en technologische vraagstukken bij belanghebbenden die betrokken zijn bij het beheer van onderzoeksdata. We kijken bijvoorbeeld naar de samenwerking tussen onderzoekers en datamanagers. Door dit te doen, wordt een agenda voor open data in de academische wereld voorgesteld op basis van kwalitatief onderzoek waarin kwesties worden belicht zoals het ontbreken van een goede infrastructuur, verantwoording, wettelijke kaders en beloningen in het beheer van onderzoeksdata. Tegelijkertijd worden nieuwe rollen zoals data rentmeesterschap *('data stewardship')* en de worsteling met databeheerondersteuning onderzocht. Om de behoeften en praktijken van belanghebbenden verder te bepalen, wordt een vergelijkbare verkennende benadering gebruikt om te ontdekken hoe financieringsinstanties en databeheerondersteuners een onderzoeksdatastrategie in Nederland ontwikkelen.

Vervolgens gaan we dieper in op het concept van reproduceerbaarheid in de experimentele wetenschap. Om dat te bereiken, onderzoeken we vanuit een technologisch oogpunt verscheidene databeheervraagstukken en laten we zien welke soorten reproduceerbaarheidsproblemen optreden in opslagsystemen met forensische laboratoriumtechnieken. Bovendien onderzoeken we reproduceerbaarheid in onderzoeksdatabeheer door laboratoriumwerk en de wetenschappelijke infrastructuur in kaart te brengen vanuit een socio-technisch model. Als zodanig krijgen we een uitgebreider beeld van reproduceerbaarheidsproblemen en verfijnen we organisatorische en technische aspecten van reproduceerbaarheidsproblemen in de praktijk.

Ten slotte illustreren we enkele toepassingen van FAIR-technologie. Ten eerste laten we de noodzaak zien voor het ontwerpen van reproduceerbare en herbruikbare onderzoekssoftware met het reproduceerbare en onderzoeksgerichte kennisontdekkingsproces in databases (RRO-KDD). Vervolgens presenteren we een strategie voor open wetenschap-paraatheid. De resultaten van dit werk bieden onderzoekslaboratoria en andere belanghebbenden zoals bibliotheken, ICT ontwikkelaars en financiers inzicht in reproduceerbaarheid en open wetenschapsuitdagingen die zijn gebaseerd op onderzoek van laboratoriumwerk uit de dagelijkse wetenschapspraktijk.

Curriculum Vitae

Armel Lefebvre has completed his bachelor's degree in Information technology and Management at the Haute École de la Province de Liège in 2012 followed by his master's degree in Business Informatics at Utrecht University in 2015. He wrote his master thesis on reproducible research and interactive data mining in bioinformatics. After, he pursued a PhD in research data management with the financial support of ITS, Utrecht University's Information and Technology Services department, that launched a Research IT innovation programme in 2016. His PhD project "research data management for open science" has been initiated as a joint effort between the Utrecht Bioinformatics Center (UBC) and the Information and Computing Science department of Utrecht University.

Before joining the department of Information and computing science at Utrecht University, Armel worked as a research software engineer at Sciensano, the Belgian Institute for Public Health until 2013. With his experience in software development in a research context, he has since aimed at reconciliating research and information technology. As a researcher and lecturer, he has presented his work at international conferences and also produced unique teaching material such as a laboratory forensics course for the Force11 community, a research data management for bioinformaticians online course and a Business Intelligence workshop on monitoring research and innovation performance in Europe specifically designed for master students in business informatics and applied data science at Utrecht University.

In March 2020, Armel started his career in research and innovation policy as a Research Information Officer at the Erasmus Research Institute of Management (ERIM) at the Rotterdam School of Management, where he combines research intelligence with novel approaches to research performance evaluation to further support the implementation of open science strategies at the Erasmus University of Rotterdam in the Netherlands.

SIKS Dissertation Series

2011

2011 **01** Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

09 Tim de Jong (OU), Contextualised Mobile Media for Learning

10 Bart Bogaert (UvT), Cloud Content Contention

11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perpective

12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM), Strategic Decision-Making in complex games
19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT), Robust Brain-Computer Interfaces
45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications

05 Marijn Plomp (UU), Maturing Interorganisational Information Systems

06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms

10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems

13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval

30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior

48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data

49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions

50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering

51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013

2013 **01** Viorel Milea (EUR), News Analytics for Financial Decision Support

02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column store Database Technology for Efficient and Scalable Stream Processing

03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics

04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling

05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns

06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience

07 Giel van Lankveld (UvT), Quantifying Individual Player Differences

08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators

09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications

10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.

11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services

12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance

28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience

29 Iwan de Kok (UT), Listening Heads

30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support

31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications

32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere

34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search

35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction

36 Than Lam Hoang (TUe), Pattern Mining in Data Streams

37 Dirk Börner (OUN), Ambient Learning Displays

38 Eelco den Heijer (VU), Autonomous Evolutionary Art

39 Joop de Jong (TUD), A Method for Enterprise Ontology based De- sign of Enterprise Information Systems

40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games

41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning

43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data

02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method

03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions

04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability

06 Damian Tamburri (VU), Supporting Networked Software Development

07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints

09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

10 Ivan Salvador Razo Zapata (VU), Service Value Networks

11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support

12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control

13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

14 Yangyang Shi (TUD), Language Models With Meta-information

15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations

19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games –Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density ModelsIntegrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Soft- ware Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying

38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.

39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital

40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education

41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text

42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models

43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments

44 Paulien Meesters (UvT), Intelligent Blauw.Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.

45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach

46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity

47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015

2015 **01** Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response

02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls

03 Twan van Laarhoven (RUN), Machine learning for network data

04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments

05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding

06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes

07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis

08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions

09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems

10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning

11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins

12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks

13 Giuseppe Procaccianti (VU), Energy-Efficient Software

14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations

15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation

16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork

17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs

18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories

19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners

20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination

21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning

22 Zhemin Zhu (UT), Co-occurrence Rate Networks

23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage

24 Richard Berendsen (UVA), Finding People, Papers, andPosts: Vertical Search Algorithms and Evaluation

25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection

26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure

27 Sándor Héman (CWI), Updating compressed colomn stores

28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO

29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains

30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning

31 Yakup Koç (TUD), On the robustness of Power Grids

32 Jerome Gard (UL), Corporate Venture Management in SMEs

33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Re- sources

34 Victor de Graaf (UT), Gesocial Recommender Systems

35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans an Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Contain- ment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics – Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Inter- active Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meetArt & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UVA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information:from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Explo-ration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TUE), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emo- tion in a group of children
17 Jianpeng Zhang (TUE), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech

30 Wouter Beek, The ”K” in ”semantic web” stands for ”knowledge”: scaling semantics to the web

2019

2019 **01** Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data

06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VU), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VU), Using virtual agents for simulation and train- ing of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learn- ing programming

36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

2020 **01** Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TUE), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Opti- mal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable En- ergy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VU), Where is the robot? Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach:the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst

31 Gongjin Lan (VU), Learning better – From Baby to Better

32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (MU), Intervention or Collaboration ? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021

2021 **01** Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models