Laboratory forensics for open science readiness: an investigative approach to research data management

Lefebvre, A.; Spruit, M.R.

Information Systems Frontiers
https://doi.org/10.1007/s10796-021-10165-1

Laboratory Forensics for Open Science Readiness: an Investigative Approach to Research Data Management

Armel Lefebvre [1] · Marco Spruit [1,2]

Accepted: 26 June 2021

Abstract
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.

Keywords Research data management · Reproducible research · Open science readiness · Digital forensics · Laboratory forensics

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 et al., 2018). One practical reason RDM gains traction is that experimental activities taking place in laboratories increasingly rely upon digital technologies (Huang & 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 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 & 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 & Gottardo, 2013; Laine et al., 2007). Methodological issues were found to vary from misapplications of statistics to poorly

✉ Armel Lefebvre
armelefebvre@gmail.com

[1] Department of Information and Computing Science, Intelligent Software Systems, Organization and Information, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

[2] Department of Public Health and Primary Care, Leiden University Medical Center (LUMC), Leiden University, Albinusdreef 2, 2333 ZA Leiden, The Netherlands

Published online: 03 August 2021

designed experiments (Moonesinghe et al., 2007; 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 practices of data publication (Leonelli, 2013; Sholler et al., 2019; Wilms et al., 2018). However, insights on research data in laboratories are incomplete, as scientific publications analyzed in reproducibility studies are curated representations 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. As such, our study aims at contributing to the evolution of the scholarly ecosystem for.

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 Fig. 1.

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.

2 Background

In this section, we analyze prior work on RDM capabilities relevant to achieve open science readiness and deepens the semiotic concepts underlying our forensic investigations in laboratories. 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. Also, the digitalization of laboratories brings similar organizational challenges as encountered in business, for instance with artificial intelligence readiness (Jöhnk et al., 2020).

The proper management and sharing of research data underlying published studies is a lively subject of debate for a decade (Bajpai et al., 2019; Editorial, 2014; European

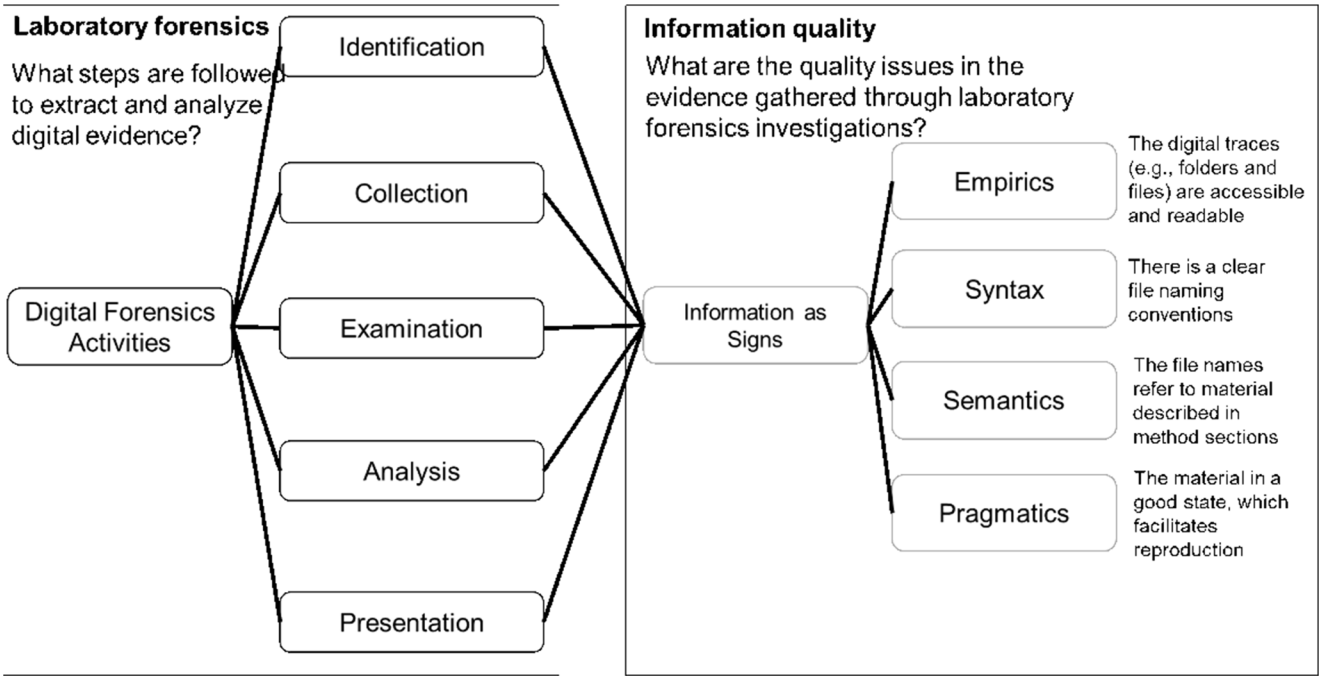

Fig. 1 The first part of this work reports on (1) the design of the laboratory forensics approach and (2) the application of laboratory forensics techniques to report on information quality issues using a semiotic perspective as found in the descriptive theory of information (DTI)

Commission, 2016; Freire et al., 2012). 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; 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 currently experienced by researchers in documenting the research process using underlying technology such as digital file systems, remote servers, and digital repositories.

Similarly, 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, 2021; 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., 2021) or the development of governance for open science, reproducible research and, and research evaluation (Austin, 2019). In all examples above, data quality (or veracity) is a significant factor in the success of big data readiness (Austin, 2019; Joubert et al., 2021).

Transposed to the laboratory domain where experimental work is conducted, we explore how laboratory can better manage research data by streamlining local preservation (inside) the lab and online preservation (i.e., depositing data on the publisher's journal) while keeping experiments reproducible. To achieve that, research data management capabilities need to be developed along the research data lifecycle, i.e., from data creation to publication (Cox and Tam, 2018). For instance, the SEI CMM is a capability maturity model tailored for research data management. The SEI CMM is oriented towards the production, preservation, and dissemination of high-quality research data (Crowston & Qin, 2011), as shown by its 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).

Hence, the analysis of research data preserved in laboratories is a starting point to explore research data capabilities further, including the investigation of 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. Thus, operational data created during scientific experimentation is 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 made openly available and their corresponding data preserved locally, on storage systems in the laboratory. So, specific 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.

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.

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 & 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. Fig. 2 shows how the experimentation process perspective compares to the forensic investigation process.

Next, the interaction of forensic investigation and the experimentation process is understood as follows: software and instruments involved in experimental events generate all sorts

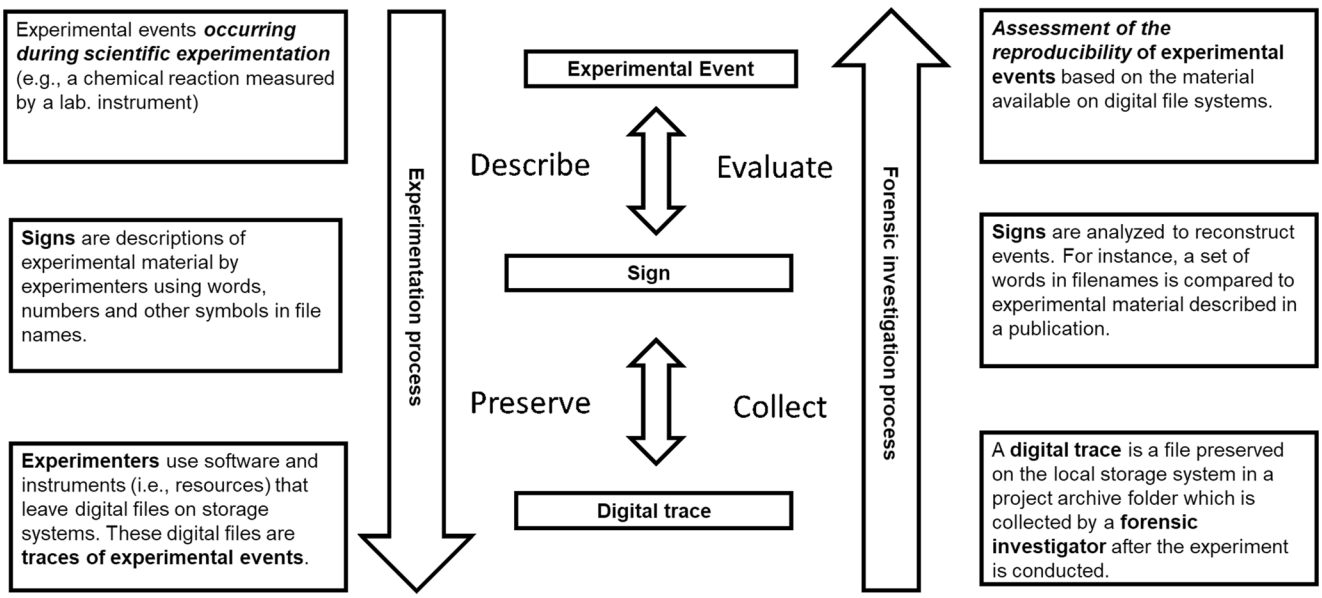

Fig. 2 A comparison of the use of events, signs, and traces from the perspective of experimenters (experimentation process) and forensic investigators (forensic investigation process)

of digital traces found on storage systems during the forensic investigation (Lefebvre & Spruit, 2019). 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 Fig. 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.

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 & 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

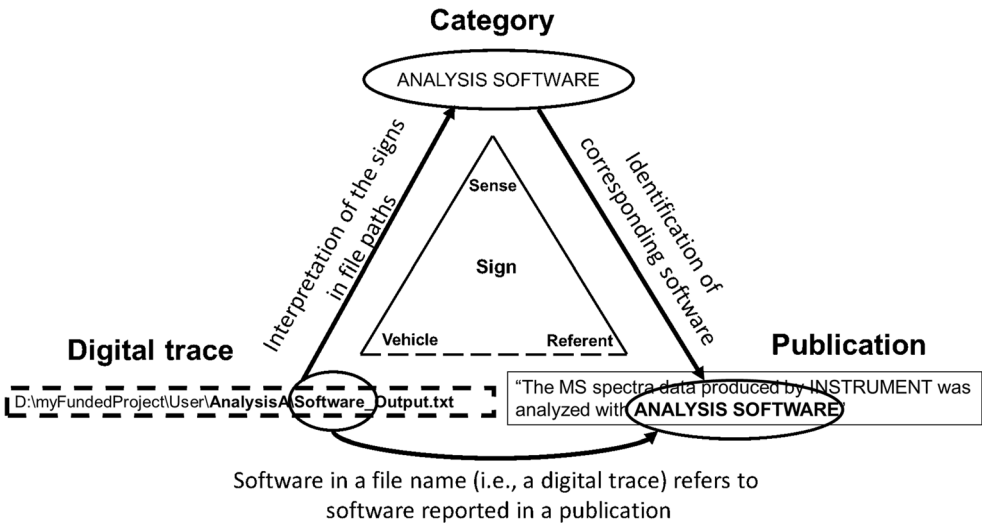

Fig. 3 The interpretation of digital traces depicted as a (semiotic) triangle of Ogden-Richards

vast amount of research data resides locally and on organization-specific systems (Prost & 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 Fig. 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 Fig. 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 3.

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 2.2, 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. In Table 1, the criteria for scoring each aspect are given.

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

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 Eq. 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 1. The criteria for each aspect were derived from our initial attempt to evaluate the repeatability of experiments in Lefebvre and Spruit (2019). We observed that the presence of certain elements on storage systems made the forensic process more efficient, while their absence can decrease the reliability of forensic outcomes. Hence, we have grouped criteria under the four semiotic aspects, each criterion showing what facilitated the recovery of experimental material and at which level (i.e., empiric to pragmatics). These criteria are applied by the forensic investigator for evaluating the quality of the

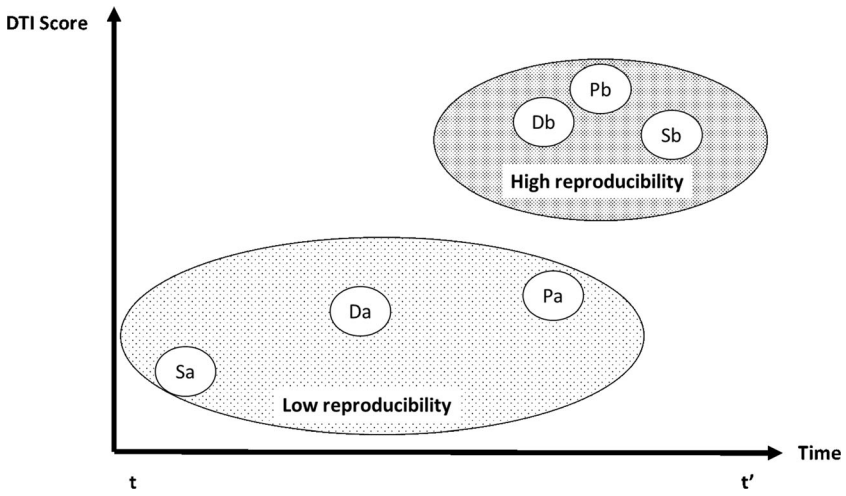

Fig. 4 The laboratory forensics approach should result in the assessment of digital traces encompassing datasets (d), software (s) and, the publication (p) employing a score standing for the quality of those traces. For instance, here A and B are two illustrative publications, where A scores lower (harder to reproduce) than publication b as the score of its components (Software, Data, and Publication) score lower

experimental resources. For instance, at the syntax level, we noticed that date times formats are a recurrent issue to determine if a series of files belong to an experiment. If date times are used consistently in file names, the score obtained for syntax should reflect this. In this study, the results of forensic investigations are used in two ways. First, these results exemplify challenges in research data management based on how research data is preserved in laboratories. Second, research data management challenges identified by forensic approaches form a way to reflect on research data management capabilities in order to enhance reproducibility,

In the two next sections, 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 journals and digital repositories.

3 Information Quality Evaluation with Laboratory Forensics

In this section, we explain the process of extracting experimental evidence from laboratory storage systems. We have conducted the forensic analyses in a chemistry laboratory in the Netherlands. We opted for a chemistry laboratory that combined laboratory work with technology development and bioinformatics so that the research data would cover a wide range of experimental practices. As an example, an experiment investigated in this study will start in a laboratory, with experimenters operating instruments that record measurements with meta-data into files. Next, computational experiments use these files for processing and analyzing laboratory outcomes. From a research data perspective, those activities leave traces on storage systems. This is shown by the included publications, listed in Table 2, which are from experiments that were conducted independently, by different groups of experimenters that combined PhD students with more senior researchers. From a digital forensic point of view, investigating such experiments has the advantage that it offers material of sufficient complexity. Therefore, each publication can be investigated using a broad set of DF activities and techniques. The main forensic activities are grouped in five steps according to (Årnes, 2017). 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

Table 1 Criteria for evaluating the investigated research data with empirics, syntax, semantics, and pragmatic branches of the DTI

Aspect	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.
Empirics	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.
Empirics	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.
Syntax	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
Syntax	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
Semantics	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

Table 1 (continued)

Aspect	Score	Scale	Criterion	Example
Semantics	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.
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.
Pragmatic	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.
Pragmatic	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

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).

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 experiments 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 Fig. 5A).
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.

3.1 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

Table 2 Background information of the selected publications examined in this study

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

The data in this study is reported anonymously. Hence only the year, journal, and publisher are communicated

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 & 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.

3.2 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 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 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.

3.3 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

Table 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

in Fig. 5A 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.

Moreover, to understand the context in which these resources are produced, another useful forensic technique is link analysis (e.g., Fig. 5B), 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 Fig. 5B. 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.

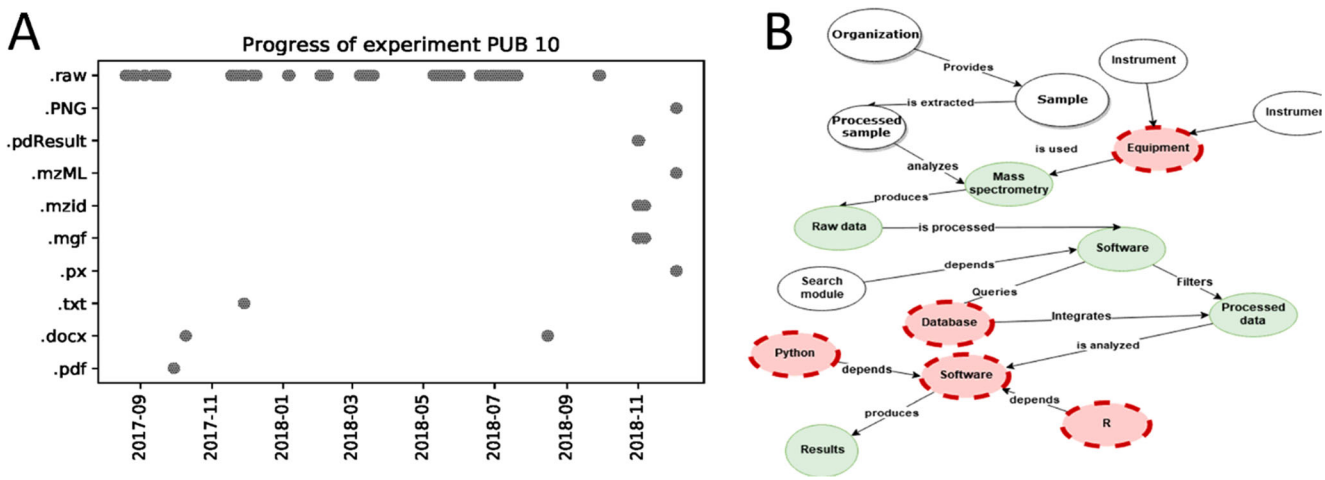

Fig. 5 In **A**, an experimental process timeline is reconstructed by forensic investigations. In **B**, a link analysis of resources as reported in the corresponding publications. Green circles refer to resources found on the storage, red circles to missing resources

3.4 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 Fig. 6, where the criteria shown in Table 1 have been applied on preserved (I.e., locally archived in the lab) and deposited (i.e., accessible online) material. To score the research, the first author of this study examined the files and publications using forensic techniques as shown earlier. Then, a score was given to each publication based on the quality of the data found in the laboratory and online, e.g., data deposited in a repository or on a journal website. Based on a number of forensic investigations, the resulting 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.

4 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

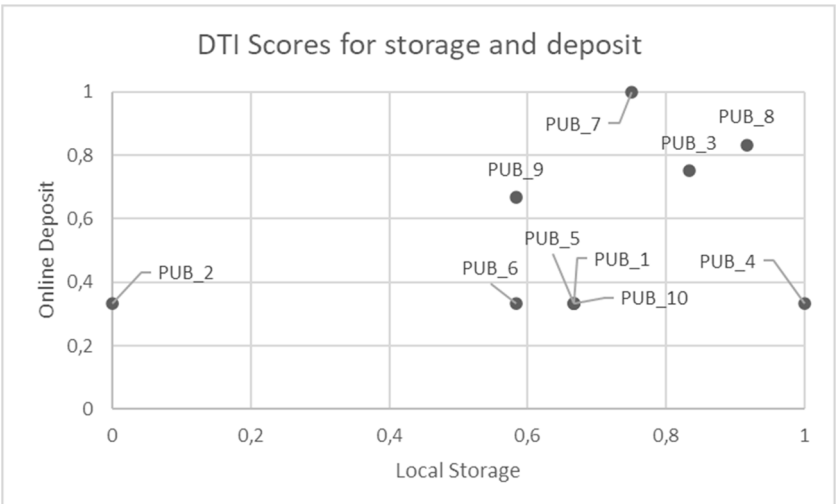

Fig. 6 Overview of the scores of information aspects for research data underlying each publication (local storage and deposit). The closer to 1, the higher the information quality of the material extracted from the storage. In the case of PUB_2, our approach failed to recover files on the local storage, which explains the DTI score of zero

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 Fig. 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.

4.1 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 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.

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 (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 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

Table 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.

streamline the description of research data are in place, it will allow novel possibilities to interpret data quality in an automated fashion.

4.2 Analytics

The analytics dashboard shown in Fig. 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, as shown in Fig. 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 Fig. 6.

An overview of DTI scores according to the type of openness, citable data, number of publications, and citations are shown in Fig. 7A. In Fig. 7A, 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 Fig. 7B. Figure 7B 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 Fig. 7C. 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.

The synergy between the dashboard shown in Fig. 7 and the outcomes of the forensic analyses shown in Fig. 6 appears with the classification of the quality of research data according to DTI aspects. The ambition of such as dashboard is to make reproducibility issues visible to laboratory managers for different experiments, different funded projects and also indicating the progress of open data. The dashboard gives an example of how RDM capabilities can be exploited to provide

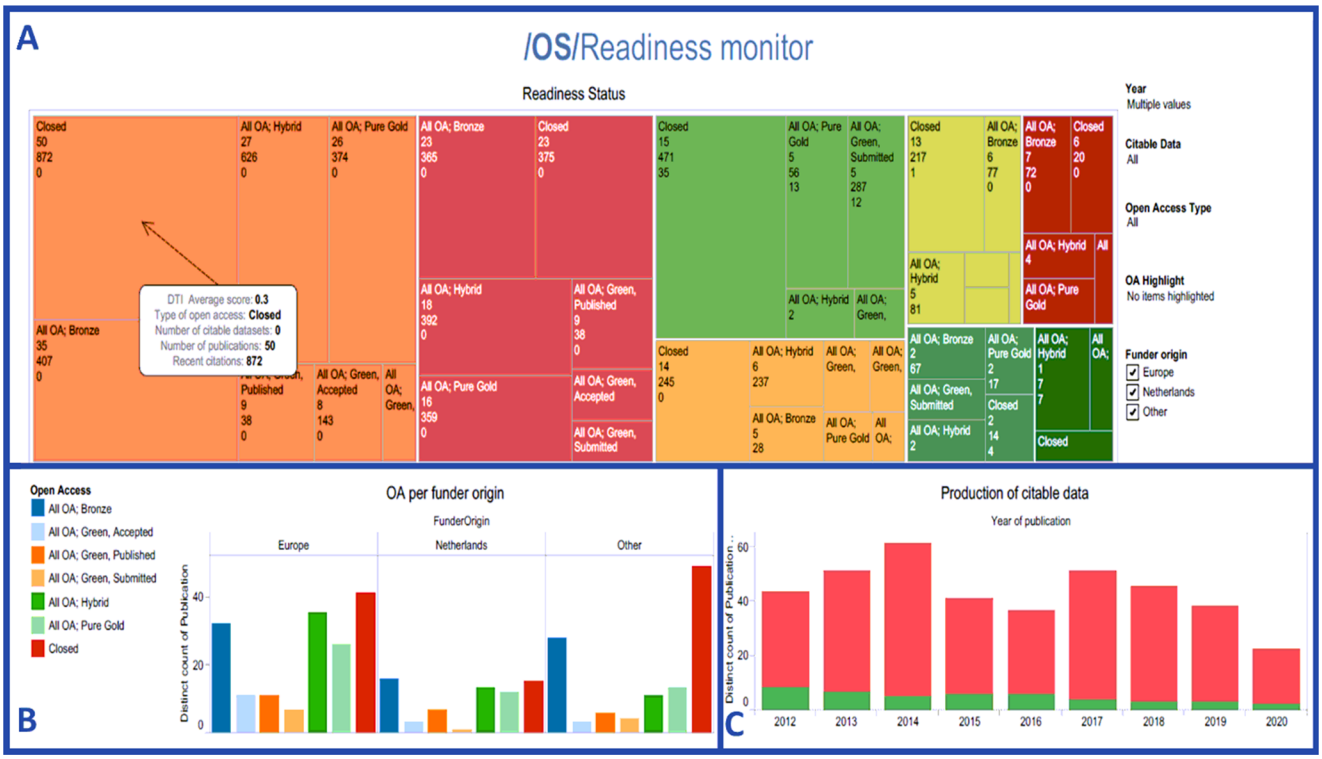

Fig. 7 Analytics dashboard for monitoring open science readiness

insights into reproducibility and open science and prepare the laboratory to deal with reproducibility threats that emerge from low quality research data.

5 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.

5.1 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 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 & 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 & 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 Fig. 5B, 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.

5.2 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. In addition, we expect open science readiness to be fundamental for supporting a robust digitalization of laboratory work. First, we comment on practical implications for research professionals, then we discuss how open science readiness contribute to data analytics ecosystems for generating both business and social value of research outcomes. 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.

Then, open science readiness is aimed at making research data management challenges visible to laboratory workers, laboratory managers as well as helping stakeholders such as research funding agencies/ At the core of open science readiness lays the concept of sustainability of scientific information. More specifically, we introduced capabilities and insights into reproducibility paving the way how research data produced in laboratories can be better preserved and shared with external parties through the scholarly communication infrastructure. OSR seeks to prepare laboratories to be embedded in larger analytics ecosystems like the open science monitor and OpenAire (Manghi et al., 2020). Therefore, the goal of the present article is to offer a path to reflect on the current situation of research data management and shape the digital transformation of laboratories for the coming decade(s). Furthermore, open science readiness reduces the gap with 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). To achieve this, a reflection about new capabilities to manage research data has to be in future research, which leads us to discuss the limitations of this study.

5.3 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. Also, research collecting data through other means than laboratory equipment, such as field experiments need to be included to evaluate forensic on a wider range of scientific practices. 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 for archiving data, 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.

Third, more research is needed about the causes of the tensions between material preserved locally and material shared online. For instance, the level of expertise in software development of an author and the focus on bioinformatic analysis can produce data and software that remain producible, taking advantage of versioned source code, readme files and a logical division of files and folders. On the other hand, analyses that relied upon software that produced a large number of files (i.e., temporary results and configurations) lead to a more challenging investigation. Comparing the forensic outcomes between different laboratories and collecting the comments of experimenters is a next step in the development towards laboratory forensics and open science readiness.

6 Conclusions

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. Open science readiness has the potential to include reproducible laboratory work in the broader reflection about sustainable digital transformation. The peculiarities of the diffusion of scientific information makes the study of scientific experimentation from the perspective of research data management a first step towards robust and complete communication of scientific evidence to society. 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.

References

Adewumi, M. T., Vo, N., Tritz, D., Beaman, J., & 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

Amorim, R. C., Castro, J. A., da Silva, J. R., & 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

Årnes, A. (2017). *Digital Forensics*. Wiley

Austin, C. C. (2019). A path to big data readiness. *Proceedings – 2018 IEEE International Conference on Big Data, Big Data 2018* (pp. 4844–4853). https://doi.org/10.1109/BigData.2018.8622229

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., & 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

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., & 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

Boell, S. K., & Cecez-Kecmanovic, D. (2015). *What is 'Information' Beyond a Definition?* International Conference on Information Systems: Exploring the Information Frontier, ICIS 2015, Paper 1363

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., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., & 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

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., & 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

Burton-Jones, A., Storey, V. C., Sugumaran, V., & 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., & Silva, H. T. (2006). VisTrails: Visualization meets data management. *2006 ACM SIGMOD International Conference on Management of Data* (pp. 745–747). https://doi.org/10.1145/1142473.1142574

Chatterjee, S., Xiao, X., Elbanna, A., & Saker, S. (2017). The information systems artifact: A conceptualization based on general systems theory. In *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)* (pp. 5717–5726). https://doi.org/10.24251/hicss.2017.689

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., & 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., & 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

Cox, A. M., & 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., & 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

Crowston, K., & 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., & 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

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

European Commission. (2016). *Guidelines on Data Management in Horizon 2020*. Retrieved March 1, 2016, from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & 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

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

Goecks, J., Nekrutenko, A., Taylor, J., & 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

Graves, M. (2013). Digital Archaeology: The Art and Science of Digital Forensics (1st ed.). Addison-Wesley Professional

Harjes, J., Link, A., Weibulat, T., Triebel, D., & 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

Higman, R., Bangert, D., & 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

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

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

Jöhnk, J., Weißert, M., & Wyrtki, K. (2020). Ready or Not, AI Comes—An interview study of organizational AI readiness factors. *Business and Information Systems Engineering, 63*(1), 5–20. https://doi.org/10.1007/s12599-020-00676-7

Jones, S., Pryor, G., & Whyte, A. (2012). *Developing research data management capability: the view from a national support service* (pp. 142–149). IPress

Joubert, A., Murawski, M., & 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

Joubert, A., Murawski, M., & Bick, M. (2021). Measuring the big data readiness of developing countries – Index development and its application to Africa. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10109-9

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

Koester, A., Baumann, A., Krasnova, H., Avital, M., Lyytinen, K., & Rossi, M. (2020). Panel 1: To share or not to share: Should IS researchers share or hoard their precious data? *ECIS 2020 Select Recordings*. Retrieved August 22, 2020 from https://aisel.aisnet.org/ecis2020_sessionrecordings/6

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., & Sox, H. C. (2007). 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

Lefebvre, A. (2020). *Open science readiness dashboard*. https://doi.org/10.5281/ZENODO.4020379

Lefebvre, A., & de Bruin, J. (2019). *Path2Insight: A file path analysis toolkit for laboratory forensics*. https://doi.org/10.5281/ZENODO.3518815

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*

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

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

Link, G. J. P., Lumbard, K., Conboy, K., Feldman, M., Feller, J., George, J., Germonprez, M., Goggins, S., Jeske, D., Kiely, G., Schuster, K., & 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., & 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., & 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

Mabey, M., Doupé, A., Zhao, Z., & 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

Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Löhden, A., Bäcker, A., Mannocci, A., Horst, M., Czerniak, A., Kiatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Ottonello, E., Lempesis, A., … Summan, F. (2020). *OpenAIRE Research Graph: Dumps for research communities and initiatives*. https://doi.org/10.5281/ZENODO.3974605

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., & 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

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., & 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

Mikalef, P., Pappas, I. O., Krogstie, J., & 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

Moonesinghe, R., Khoury, M., medicine, A. J.-Pl. (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

NAS. (2018). *Open Science by Design: Realizing a Vision for 21st Century Research*. The National Academies Press. https://doi.org/10.17226/25116

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

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 (1st ed.)*. Johns Hopkins University Press

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

Peng, R. D., Dominici, F., & 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

Prost, H., & Schöpfel, J. (2015). *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3*. Retrieved July 26, 2021 from http://hal.univ-lille3.fr/hal-01198379

Ribes, D., & 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., & 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

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

Santana-Perez, I., Ferreira da Silva, R., Rynge, M., Deelman, E., Pérez-Hernández, M. S., & 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

Serketzis, N., Katos, V., Ilioudis, C., Baltatzis, D., & 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

Sholler, D., Ram, K., Boettiger, C., & 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

Simou, S., Kalloniatis, C., Gritzalis, S., & 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

Soylu, A., Elvesæter, B., Turk, P., Roman, D., Corcho, O., Simperl, E., Konstantinidis, G., & 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., & 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

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., & Peng, R. (2014). Implementing reproducible research. *Journal of Statistical Software, 61*(October), 149–184. https://doi.org/10.1201/b16868

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720–1733. https://doi.org/10.1002/asi.20652

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLoS One, 6*(6), e0118053. https://doi.org/10.1371/journal.pone.0021101

Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of Data Grids for distributed data sharing, management, and processing. *ACM Computing Surveys, 38*(1), 3–3s. https://doi.org/10.1145/1132952.1132955

Wallis, J. C., Rolando, E., & 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

Wang, R. W., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.2307/40398176

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., & 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., & 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., Stieglitz, S., Buchholz, A., Vogl, R., & Rudolph, D. (2018). Do researchers dream of research data management? *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 4411–4420)

Zeleti, F. A., & Ojo, A. (2017). Open data value capability architecture. *Information Systems Frontiers, 19*(2), 337–360. https://doi.org/10.1007/s10796-016-9711-5

Armel Lefebvre is a research information officer at the Erasmus Research Institute of Management (ERIM) in Rotterdam. At ERIM, he contributes to the development of research intelligence for research evaluation and supports the realization of an open science strategy. He published a number of studies on reproducible research and research data management for open science as a former Ph.D. researcher at the Department of Information and Computing Sciences of the Faculty of Science at Utrecht University.

Marco Spruit is professor of Advanced Data Science in Population Health at Leiden University in the department of Public Health & Primary Care (PHEG) of the Leiden University Medical Centre (LUMC) and the Leiden Institute of Advanced Computer Science (LIACS) of the faculty of Science (FWN). His overarching research objective is to establish an authoritative national infrastructure for Dutch Natural Language Processing and Machine Learning to facilitate and popularise Self-Service Data Science. He has notably conducted several European Horizon 2020 studies (OPERAM, SAF21, SMESEC, GEIGER) and nationally funded research projects (e.g. STRIMP).