

7-7-2025

TAPAT Principles for Reliable Research Data Management Enabled by Decentralized Technologies

Arthur Carvalho
Miami University

Chad Anderson
Miami University

Mala Kaul
University of Nevada, Reno

Jeffrey W. Merhout
Miami University

Follow this and additional works at: <https://aisel.aisnet.org/cais>

Recommended Citation

Carvalho, A., Anderson, C., Kaul, M., & Merhout, J. W. (2025). TAPAT Principles for Reliable Research Data Management Enabled by Decentralized Technologies. *Communications of the Association for Information Systems*, 58, 122-145. <https://doi.org/10.17705/1CAIS.05807>

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in Communications of the Association for Information Systems by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

TAPAT Principles for Reliable Research Data Management Enabled by Decentralized Technologies

Cover Page Footnote

This manuscript underwent peer review. It was received 09/01/2024 and was with the authors for five months for three revisions. Andreas Drechsler served as Associate Editor.



TAPAT Principles for Reliable Research Data Management Enabled by Decentralized Technologies

Arthur Carvalho

Miami University
0000-0002-5381-3588

Mala Kaul

University of Nevada, Reno
0000-0002-1327-5666

Chad Anderson

Miami University
0000-0002-7163-5578

Jeffrey W. Merhout

Miami University
0000-0002-9718-2298

Abstract:

Academia has seen increased scrutiny and efforts on data management with the goal of establishing research reliability through openness in science and improving the credibility of data, research methods, and scientific communication. Following a rigorous design science methodology, we contribute to those efforts by proposing a solution that brings transparency and trust to the research data collection and storage process. Through interviews with academic experts, we formulate the TAPAT principles, i.e., five prescriptive design principles for research data collection and storage plans: *Transparency*, *Accessibility*, *Persistence*, *Affordability* (financial), and *Trustworthiness* (data integrity). As a proof of concept, we illustrate the use of such principles by designing a concrete artifact that combines open-source survey applications and decentralized technologies to create a transparent, trustworthy, and affordable data collection process while ensuring highly persistent and accessible data storage. We highlight the practicality and financial affordability of our solution through a cost analysis of a realistic survey, demonstrating its potential for wide adoption in various research communities. Moreover, we evaluate and validate our design principles and artifacts through a focus group with experienced researchers.

Keywords: Blockchain, Decentralized Technologies, Design Science Research, Research Data Management.

This manuscript underwent peer review. It was received 09/01/2024 and was with the authors for five months for three revisions. Andreas Drechsler served as Associate Editor.

1 Introduction

The integrity of scientific inquiry hinges on the veracity of data and the credibility of research findings ("Ensuring the Integrity of Research Data," 2009; Ritchie, 2020). However, as we elaborate throughout the paper, many practices can undermine research quality and erode trust in science. In an effort to address some of those issues, the Center for Open Science developed the Open Science Framework, a research project management framework to manage various stages in the research life cycle through open-source tools and services (Foster & Deardorff, 2017). In a related effort, FAIR principles were established to make research data findable, accessible, interoperable, and reusable (Foster & Deardorff, 2017). Despite these and other developments, there is continued debate on the best practices and design of research data management (RDM) systems due to factors involving data storage processes, sensitive data sharing, and the distribution of costs associated with these efforts (Kaiser & Brainard, 2023). While numerous data storage options exist, their fragmentation, along with limitations in managing the RDM life cycle, make them less effective (Kaiser & Brainard, 2023). Further, there is limited guidance on designing systems that can more effectively support a full cycle of collection, storage, and sharing of research data. Thus, the practical problem of managing research data to promote confidence in its reliability and transparency remains unresolved, leading to our overarching research question: *how to design an information system for managing reliable research data collection, storage, and distribution?*

In information systems (IS) research, it is well recognized that research data quality should be fully transparent to promote trust, provide accountability, and enable replication (Burton-Jones et al., 2021), and this issue has been examined from the communication and reporting perspective (Hui et al., 2019). For example, instead of just a description in the paper, scholars have suggested that authors should make the actual user interfaces that are employed to collect data available (Wagenmakers et al., 2021). Having access to those interfaces, other researchers could replicate the original data collection process. Moreover, the collected data must be easily accessible and persistent, e.g., be widely available to interested parties for a prolonged period of time (American Psychological Association [APA], 2017; GO FAIR, 2016; Vicente-Saez & Martinez-Fuentes, 2018). Additionally, the technologies to collect and store data should be affordable and trustworthy so that ex post manipulation is not feasible (Hahnel et al., 2023; National Science Foundation [NSF], n.d.). Accordingly, we follow the design science methodology by Peffers et al. (2007) to propose the TAPAT¹ — Transparency, Accessibility, Persistence, Affordability [financial], and Trustworthiness [data integrity] — design principles (DPs) that researchers should follow when developing reliable plans for collecting, storing, and sharing research data.

We demonstrate the TAPAT DPs through a concrete artifact, namely, a redesigned online survey using decentralized technologies. From a technical perspective, we discuss how open-source practices, such as hosting the survey's source code on modern platforms for version control and collaboration, can boost transparency. On the backend, our proposed reformulation of surveys relies on decentralized technologies that include the InterPlanetary File System (IPFS) and public blockchains to store respondents' data in a highly distributed and resilient way, showcasing the accessible and persistent features of the artifact. A thorough evaluation of the artifact highlights its affordable nature and how public blockchains, in conjunction with IPFS, can bring trust to research data collection and storage processes. The theoretical contributions of this research include both the TAPAT DPs and the redesigned survey using decentralized technologies as a concrete artifact that demonstrates the potential to make the scientific inquiry process more reliable. This design improvement addresses the general problem of RDM through a new and innovative solution that includes DPs as well as an instantiated artifact (Gregor & Hevner, 2013).

The rest of the paper is organized as follows. Section 2 presents the background information needed to contextualize the study appropriately and reviews the relevant literature. Section 3 elaborates on the design science research methodology (DSRM). Section 4 introduces the TAPAT DPs. In Section 5, we recreate the online survey by Anderson et al. (2023) by using a combination of decentralized technologies to demonstrate a concrete artifact that follows the TAPAT DPs. Moreover, we evaluate the artifact through informed arguments, a cost analysis, and a focus group. Section 6 provides the limitations and proposes avenues for future research, and Section 7 concludes the paper.

¹ TAPAT means *honest* in the Filipino language. It also has connotations embodying *trustworthiness*.

2 Research Background

2.1 RDM

The concept of RDM includes all the activities that scholars engage in to plan, collect, analyze, and share the data created through their research efforts (Kanza & Knight, 2022; Surkis & Read, 2015). For many researchers, those are primarily offline activities with data organized and stored in either non-digital formats (e.g., paper documents arranged in labeled folders in a file cabinet) or in local digital repositories (e.g., Excel, Word, and PDF files arranged in digital folders on a computer hard drive or a portable flash drive). Sharing the data has typically been limited to transferring documents between co-authors or offering the data “upon request” to readers of their publications.

Increasingly, publishers, academic associations, and grant funding institutions are requiring researchers to make their data more accessible. For example, Section 8.14 of the APA (2017) Ethics Code states that “After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis.” Similarly, the NSF has a data-sharing policy that states, “NSF-funded investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF awards” (Walls et al., 1992). These expectations are associated with the concept of “open science,” where research data is made available to others outside the research team (Vicente-Saez & Martinez-Fuentes, 2018) and the methods by which that data was acquired and analyzed to generate published results are sufficiently transparent (Hevner et al., 2024; Marsden & Pingry, 2018).

Interest in open science has grown significantly in recent years (Gentemann, 2023), and the previously mentioned Open Science Framework was created to support researchers’ efforts to engage in open science (Foster & Deardorff, 2017; *Open Science Framework*, 2025). Within the framework of open science, guidelines to improve data management and stewardship have been proposed, including the FAIR principles, which emphasize that data must be findable, accessible, interoperable, and reusable (GO FAIR, 2016).² Efforts to develop principles for more effective data management continued to evolve, with Bellgard (2020) noting that, while FAIR principles are important starting points to effective RDM, a more comprehensive approach is required. For example, the FAIR principles do not explicitly address the need for trust in the platforms through which data is made accessible. In addition to the FAIR principles, other sets of principles have been proposed to address data management and stewardship. For example, the UNESCO (2021) Recommendation on Open Science includes a section on what constitutes open research data in which the FAIR principles are one component. Other sets of principles have been created to address specific types of data or research. For example, the CARE principles were created to emphasize the need to protect the rights and interests of Indigenous people in the use of their data (S. R. Carroll et al., 2023), while the ECSA’s Ten Principles of Citizen Science focus specifically on research activity by the general public and include a principle for the public availability of citizen science data and metadata (Robinson et al., 2018).

Open science efforts have accelerated because of a growing number of RDM issues that include the limited availability of research data for validation and replication (King, 1995), undisclosed conflicts of interest (Friedman & Richter, 2004), selective reporting of studies with significant findings (John et al., 2012), flexible data analysis and data dredging (Gorman et al., 2019), and data fabrication (Scheiber, 2023), to name a few. Addressing this range of issues to support reliable, transparent, and trustworthy research practices requires a sociotechnical approach, as there are social and ethical issues that must be considered when designing technical solutions (Silver & Markus, 2013). Within our sociotechnical approach, we focus specifically on RDM issues that can impact research transparency and reproducibility.

The IS research community has proactively developed guidelines to promote transparent and reproducible research practices (Burton-Jones et al., 2021; Marsden & Pingry, 2018; Pekkola et al., 2019; Wilson et al., 2021). For example, IS scholars have discussed and proposed new standards for null hypothesis significance testing to avoid problems such as p-hacking and HARKing (McShane & Gelman, 2022; Mertens & Recker, 2023; Sen et al., 2022). Such standards can help reduce some questionable

² *Findable* means that data should be easy for humans to locate and that metadata should be machine-readable so that computers can also discover datasets. Once found, it must be clear how the data can be *accessed*, and the access protocol should be open and free. To the extent possible, the data also need to be *interoperable* by being readily usable by applications for analysis, storage, and processing purposes. Finally, the data and its metadata should be adequately described to optimize *reuse*.

behaviors, such as rounding down p -values, a practice admitted to by one in five researchers (John et al., 2012). Moreover, the value in replicability of academic research data has been recognized in IS research (Dennis et al., 2020), with a whole new research venue now dedicated to replication studies in IS (Dennis & Valacich, 2015). The importance of and approaches for improving research transparency have also been debated and documented (Burton-Jones et al., 2021; Hevner et al., 2024; Pratt et al., 2020).

Table 1 summarizes the RDM issues that can arise at each research stage (i.e., planning, data collection, data analysis, and data sharing). For each issue, we note the RDM stage(s) in which the issue is most impactful, some consequences of the issue, and the stakeholders (scientific community, academic institutions, funding agencies, and society) impacted by the issue. Scientific community includes all individuals and teams engaged in research work; academic institutions include colleges and universities that employ researchers; funding agencies include both government agencies and private organizations that provide grants to researchers; and society includes all other entities and individuals not included in the other three stakeholder groups. In particular, (1) *scientific community* was included when issues influenced reproducibility, credibility, or knowledge-building; (2) *academic institutions* were considered where data storage, management costs, retention, or policy compliance were relevant; (3) *funding agencies* were identified when problems involved wasted resources, misallocation of funds, or unethical practices; and (4) *society* was included when issues undermined public trust, ethical standards, or raised privacy and social concerns.

Table 1. Research Data Management Issues

Issue	RDM stage	Consequences	Impacted stakeholders
Limited availability of research data for validation and replication	Planning, sharing	Stagnation of scientific advances	Scientific community, society
High costs to manage research data	Collection, analysis, sharing	Wasted resources	Academic institutions, funding agencies
Limited data archival and retention capabilities	Sharing	Stagnation of scientific advances, unproductive research efforts, reduced confidence in research	Scientific community, academic institutions, society
Lack of confidence in the trustworthiness of research data	Sharing	Wasted resources, reduced confidence in research	Scientific community, academic institutions, funding agencies, society
Flexible data analysis (p-hacking, data dredging)	Analysis	Unproductive efforts to advance knowledge, wasted resources, reduced confidence in research	Scientific community, academic institutions, funding agencies, society
Misrepresentation (selective reporting of significant findings, HARKing)	Sharing	Unproductive efforts to advance knowledge, wasted resources, reduced confidence in research	Scientific community, academic institutions, funding agencies, society
Data censorship	Sharing	Unproductive efforts to advance knowledge	Scientific community, society
Falsification (data fabrication)	Collection, analysis	Unproductive efforts to advance knowledge, wasted resources, and reduced confidence in research	Scientific community, academic institutions, funding agencies, society
Exposure of sensitive data	Sharing	Loss of privacy	Academic institutions, society
Undisclosed conflicts of interest	Sharing	Unethical gains, misappropriation of resources	Scientific community, academic institutions, funding agencies, society
Research data theft	Sharing	Loss of intellectual property, loss of privacy	Academic institutions, society
Note. Issues highlighted in green (dark gray) are those that our proposed solution most directly addresses, and issues highlighted in blue (light gray) are those that our proposed solution partially addresses.			

Focusing on the specific RDM issues our research addresses, we elaborate on the impacts those issues create for stakeholders. When the availability of research data for validation and replication is limited, it impacts the scientific community because researchers are constrained in their ability to build upon and verify prior research, which impacts society through a slowing of scientific advancement. When costs to manage research data are high, academic institutions and funding agencies are impacted by increased risk in supporting researchers who may not produce useful results. Limited data archival and retention capabilities impact the scientific community by constraining researchers' ability to build on and verify prior studies, increasing liability for academic institutions, and limiting scientific advances for society. Lack of

confidence in the trustworthiness of research data impacts all stakeholders because the potential for research to influence societal-level decisions (e.g., public health policy) will be weakened. Flexible data analysis and misrepresentation also impact all stakeholders by weakening the integrity of research, which wastes resources and slows scientific advancement. Data censorship impacts the scientific community by limiting the availability of data for use, and society by slowing scientific advancement. As we elaborate on in subsequent sections, our proposed TAPAT principles address these issues by building on existing principles and emphasizing the need for transparency, accessibility, persistence, affordability (financial), and trustworthiness (data integrity) that are implied but not made explicit in other work.

2.2 IPFS

Decentralized networks are at the heart of our proposed DPs and artifacts to overcome the limitations of existing systems and practices for collecting and storing research data. The IPFS is a decentralized storage system that allows users to store and retrieve files without relying on centralized servers. This decentralization makes IPFS-resilient and censorship-resistant, as files are stored on multiple computational devices, called *nodes*, spread across the globe, and users can retrieve files from nodes on which they are stored. For the purpose of this research, it is essential to highlight that each update or entry related to the research data would create a file. Each file uploaded to IPFS, such as a survey response, receives a unique cryptographic identifier, hereafter referred to as the *hash value*; users can, at no cost, download a file using its hash value from any node storing it. However, although uploaded files get distributed across nodes, there is no guarantee that those nodes will continue storing the file indefinitely. That is, nodes can delete content they are storing as they wish. Users can ask nodes to *pin* important files to ensure they persist and remain available in IPFS. Specifically, a file will remain available in IPFS as long as at least one node has pinned it. Although any entity can join the IPFS network as a node and pin its own files for free, another common practice is to use *pinning services* from different organizations for a fee.

While IPFS offers a transparent, public, open, free, and immutable data storage service, it still suffers from some drawbacks that prevent us from relying solely on it in our work. First, IPFS does not natively keep track of data ownership; i.e., one does not know who uploaded a certain file. That goes against the transparency principle we advocate, as researchers should be able to link data to a certain experiment or academic work. Moreover, IPFS does not provide timestamps regarding when the data was uploaded; instead, it organizes data as a versioned file system. Here, again, the lack of timestamps results in less transparency, as one does not know when the data was produced. We solve the above issues by using IPFS in conjunction with another decentralized technology for data storage, namely, *blockchain*, which we introduce in the following subsection.

2.3 Blockchain

Fundamentally, a blockchain is an append-only and distributed ledger that timestamps and records transactions across multiple independent computers (nodes) in a network. This distributed and decentralized structure ensures that no single entity controls the entire chain of blocks of transactions, making blockchain technology secure and transparent. While Bitcoin introduced blockchain as a means for storing digital currency transactions, Ethereum expanded this technology's potential by incorporating *smart contracts* and *decentralized applications* (DApps) (M. C. Lacity, 2022). For our purposes, smart contracts are algorithms executed in tandem by the nodes within a blockchain network. This feature elevates blockchain technology beyond a distributed database system, transforming it into a consensus mechanism for the correct and consensual execution of algorithms. Smart contracts enable the creation of DApps, i.e., software applications whose logic and/or backend lives on the blockchain, allowing developers to store more than just transactional data on the blockchain.

Another relevant concept for our work regards the nature of blockchain networks. *Public blockchains*, such as Bitcoin and Ethereum, are open to anyone, and everyone can participate as a system user or a node joining the block verification process. In contrast, *permissioned blockchains* may restrict some of the rights to participate in the network, where the underlying permissions are often defined by organizations setting up the blockchain for a specific goal. This distinction is crucial for understanding the varying levels of security, transparency, and speed between different types of blockchain networks. For example, public blockchains offer higher transparency, as all transactions are visible to anyone who accesses the network. In contrast, permissioned blockchains tend to be more efficient, as they have fewer nodes to validate transactions and less stringent consensus mechanisms.

We also note that data storage tends to be costly on public blockchain networks due to their highly distributed and decentralized nature. For blockchains based on the Ethereum Virtual Machine, each transaction has an associated computational cost (*gas*) that can be translated into a cryptocurrency cost, which fluctuates based on, among other factors, blockchain network congestion. At the time of writing, the USD cost of storing a single IPFS hash on Ethereum is \$3.78. Ethereum is a notoriously expensive blockchain network. Other public blockchain networks offer considerably more affordable yet similar services. For example, the costs associated with Polygon and Fantom are, respectively, \$0.0087 and \$0.0009 to store a single IPFS hash. Alleviating such high costs is where IPFS comes into play, as it allows for potentially free storage and access of files across a distributed network, thus reducing the need for large amounts of data to be stored directly on the blockchain. The artifact we develop leverages the synergy between blockchain and IPFS, pairing the affordable storage available through IPFS with the timestamping and provenance capabilities of public blockchains.

2.4 Decentralized Technologies for RDM

We posit that modern decentralized technologies, such as IPFS and blockchain networks, are crucial components to creating concrete artifacts for data collection and storage processes that satisfy the TAPAT principles. On that note, researchers have proposed and evaluated blockchain-based solutions for RDM. For example, Woo and Yoo (2023) applied the technology acceptance model to find the factors influencing the adoption of blockchain-based electronic lab notebooks for managing research data. Zhang (2021) conducted a SWOT analysis focusing on the management of scientific research data at universities. Following that analysis, Zhang (2021) proposed the development of a data management system for university scientific research activities utilizing blockchain technology. Besides rigorously defining the TAPAT principles, our work differs from these previous endeavors in that we create a concrete instance of an artifact that instantiates those principles. This latter point is a major difference between our work and some related work from the computer science community. For example, Tang et al. (2021) suggested storing blocks of massive datasets on IPFS to achieve network efficiency. The hash value representing each block is then stored on a blockchain network to index the blocks. In other words, blockchain works as an indexing system. Our work differs from this and similar solutions in that we are interested primarily in the timestamping and immutability properties of blockchain to complement the storage capabilities of IPFS so that research data can remain available.

To summarize, our research contributes to the RDM literature by prescribing DPs that tackle problems with current research data collection and storage practices. Moreover, we show how the decentralized storage of IPFS, aligned with the properties of blockchain networks, can form the backbone of concrete artifacts that are part of processes that satisfy those principles.

3 Research Methodology

To address our research question, we employed a DSRM due to its recognized value for solving difficult problems through the creation of innovative artifacts and improving our understanding of the problem domain to create new design knowledge (Hevner et al., 2004). We adopted the DSRM proposed by Peffers et al. (2007) based on its proven suitability for developing practically useful artifacts. For example, Silver and Markus (2013) contend that design science must consider both technical and social design features and that the DSRM by Peffers et al. (2007) is structured to support a sociotechnical perspective. That methodology involves six phases: problem identification, solution objective definition, artifact design and development, artifact demonstration, artifact evaluation, and communication. Figure 1, adapted from the work by Peffers et al. (2007), illustrates our activities across those six phases.

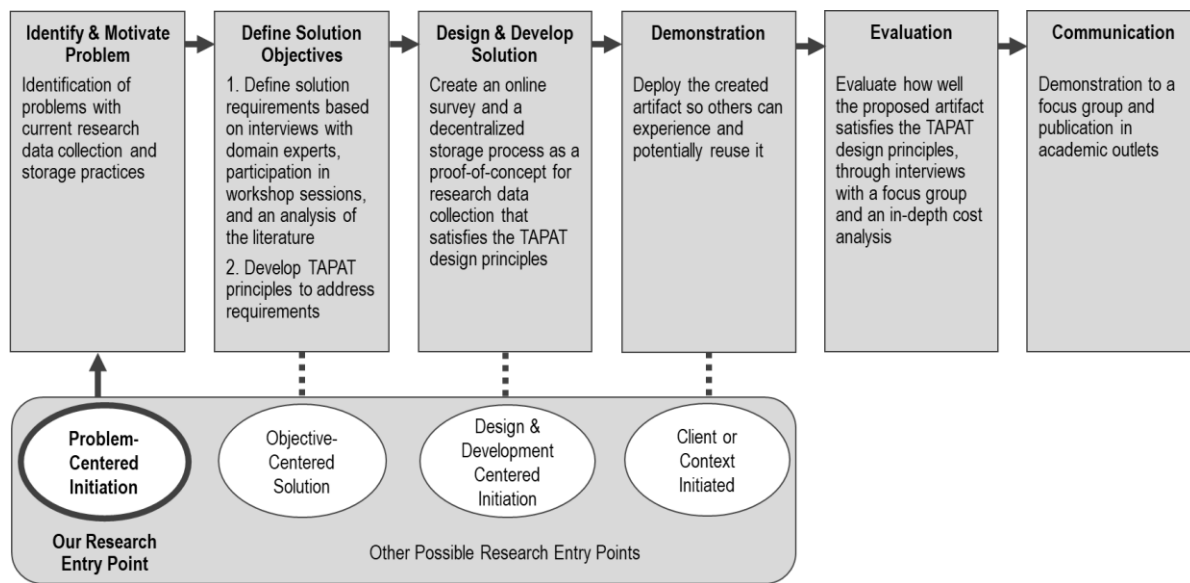


Figure 1. Design Science Research Methodology Applied to this Research (Source: Peffers et al., 2007)

Scholars have discussed how design science research could benefit from the application of open science principles. For example, Doyle et al. (2019) demonstrated how open science principles could be applied to the design of open artifacts, which they define as “artefacts that can be accessed online for free, with an open license that allows use, inspection, modification and reuse” (p. 11). Doing so, they argue, would make such artifacts more accessible to practice. Doyle and Luczak-Roesch (2020) demonstrated how the Open Science Framework and, specifically, research reports (i.e., preregistrations of a research study), could be integrated into design science research. We also note that scholars, such as Brendel et al. (2021), have called for replication studies in design science research, which means design science solutions to RDM problems could directly support the advancement of design science research itself.

Throughout the DSRM phases, we sought to understand different practices regarding RDM and how they could (or could not) solve our motivating research problem, namely, how to design an information system for managing reliable research data collection and storage. In particular, we relied on three major knowledge sources. First, we conducted semi-structured interviews with personnel at the libraries of three major U.S. universities. The interviewees were either directly involved in supporting researchers in managing their research data or had oversight of the personnel who provided those services. Table 2 summarizes the interviewees’ roles and expertise, while Appendix A lists the guiding questions we used when conducting the interviews. Second, we attended local workshop sessions designed to provide faculty with information on the types of research services available through the central library, which included support for managing research data and meeting the data management requirements of funding agencies such as the NSF. Third, we derived information and facts from specialized reports, such as the State of Open Data by Hahnel et al. (2023). We conclude this section by noting that the first DSRM phase — i.e., “define the specific research problem and justify the value of a solution” (Peffers et al., 2007) — has already been addressed in Section 1. Therefore, the following sections elaborate on the next four DSRM phases, with the sixth phase (communication) represented in this paper.

Table 2. Roles and Experience of Interviewees

Interviewee	Current role	Relevant experience
1	Data services librarian	Over 25 years of experience in various information technology roles
2	Director of library technology	Over 25 years of experience in various information technology roles
3	Dean of libraries	Over 25 years of experience in librarian services
4	Repository manager	Over 20 years of experience in information technology roles focused on library services
5	Director of digital archive initiatives	Over 25 years of experience in digital collection and archival services

4 Defining Solution Objectives

The second DSRM phase involves “inferring the goals for a solution based on the defined problem and an understanding of what is achievable and practical” (Peffer et al., 2007). We state such objectives in terms of *design requirements* (DRs) and formulate *DPs* that address the requirements for the design of any system for managing research data collection and storage. The development of the solution objectives and its process, starting from the problem, to defining requirements, to the development of DPs and the features of a concrete artifact, is all summarized in Figure 2.

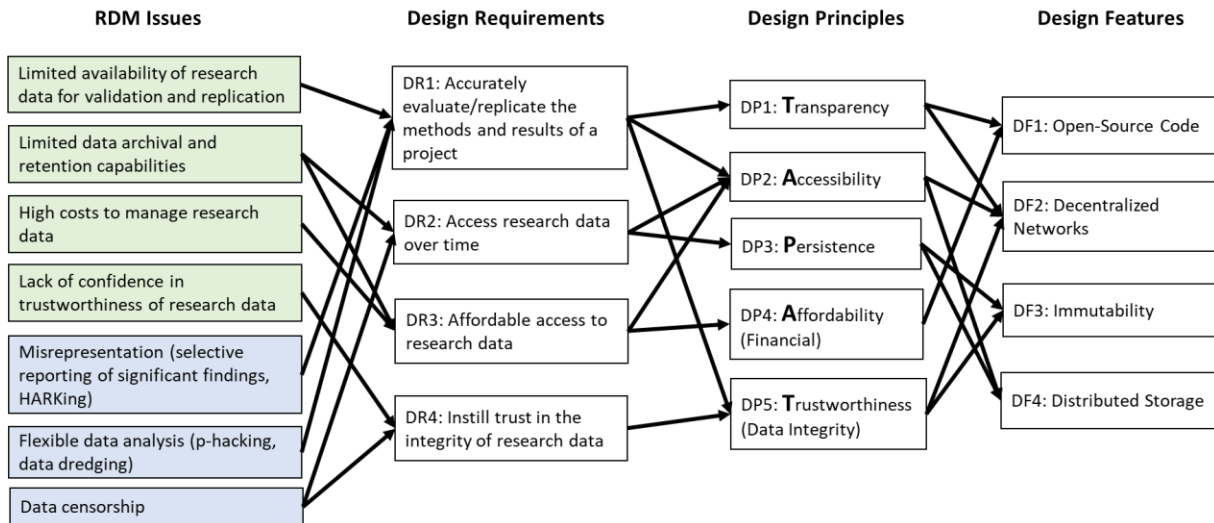


Figure 2. A Sociotechnical Perspective of RDM Issues, Design Requirements, Principles, and Features

4.1 DRs

DRs focus on the “what” of an artifact, i.e., what needs to be achieved or accomplished. In our work, DRs are based on the three sources described in Section 3 (interviews, workshops, and study of documents). Through these sources, we identified four DRs for any solution to effectively manage research data collection and storage. The first requirement is that the solution must enable the accurate evaluation/replication of the methods and results of a research project (DR1). This would address the issue of limited availability of research data for validation and replication. It could also address issues of flexible data analysis and misrepresentation because it would be easier for those manipulations to be discovered. The second requirement is that a solution must provide access to research data over an extended period of time (DR2). This would address the issue of limited archival and retention capabilities. It could also address the issue of data censorship that can result from control over the longevity of research data availability. The third requirement is that a solution must provide access to research data in an affordable manner (DR3). In conjunction with DR2, this would address the issue of limited archival and retention capabilities. It would also address the issue of high costs to manage research data. The fourth requirement is that a solution must engender trust in the integrity of the research data (DR4). This would address the issue of lack of confidence in trustworthiness of research data. It could also address the issues of data censorship.

4.2 DPs

DPs are “prescriptive statements that indicate how to do something to achieve a goal” (Gregor et al., 2020, p. 1622). van Aken (2004) described these as “a chunk of general knowledge, linking an intervention or artifact with a desired outcome or performance,” indicating that a solution to a DR could be design knowledge that was captured in a more general rather than technology-specific format to enable wider applicability across a class of problems (p. 228). Thus, our framing of DPs is done to provide a way of supporting the design of any system that reliably manages research data in a trustworthy and transparent way. We call our DPs the *TAPAT* principles, as they encompass *Transparency*, *Accessibility*, *Persistence*, *Affordability* (financial), and *Trustworthiness* (data integrity). Below, we elaborate on each principle in more detail and present these in the schema of DPs proposed by Gregor et al. (2020).

4.2.1 Transparency

Transparency is defined as “the extent to which one entity discloses relevant information about its own decision processes, procedures, performance, and functioning” (Hosseini et al., 2018, p. 3). Transparency has emerged as a key dimension of the open science and open data movements (Lyon, 2016) since questionable research practices are largely enabled by activities that are not transparent to interested parties outside of a research team (e.g., other members of the scientific community). When not accurately conveyed, choices regarding study design, data collection, data analysis, and result presentation may make it challenging to identify potential scientific errors (Tremblay et al., 2012). A simple yet incomplete solution to tackle transparency issues is to require that research data be made publicly available. Such a solution has some popularity, as the study by Hahnel et al. (2023) shows that 54% of their 6,091 survey respondents agreed that research funders should make sharing research data part of their requirements for awarding grants. However, while sharing research data enables the *verifiability* of research findings, it does not provide *traceability*, which is also essential for full transparency. This leads us to formulate our first DP concerning proper data collection and storage practices in Table 3.

Table 3. Principle of Transparency (DP1)

Components of the design principle schema (Gregor et al., 2020)	
Aim, enactor, implementer, and user	For designers of RDM solutions (implementers) to allow interested parties (users) to accurately evaluate/replicate the methods and results of a research project conducted by a research team (enactors)
Context	Open science research
Mechanism	Ensure the authenticity and veracity of research data through system features that enable both verifiability and traceability
Rationale	Research data collection methods, employed artifacts, and storage procedures that are transparent enhance open science and reduce the potential for scholars to falsify research procedures and findings

The above principle suggests the need for a transparent information flow that unequivocally reveals the entire data journey from gathering to storage, fostering trust, accountability, and clarity throughout the research process, such as through a preregistration process (Nosek et al., 2018). As a consequence, having established both verifiability and traceability, interested stakeholders can not only detect data manipulation issues — as the original data will be available to all — but also investigate the data collection instruments to determine which data was collected, including potential dependent measures, thus helping alleviate questionable research practices such as not disclosing all of a study’s dependent measures (John et al., 2012).

4.2.2 Accessibility

Lack of information accessibility in academic research results from researchers retaining the data created through their research activity in locations and formats not accessible to stakeholders outside the research team (e.g., other members of the scientific community). Historically, researchers have kept their data in non-digital formats, such as paper documents in a file cabinet, or in local digital formats, such as spreadsheets and PDFs saved locally. This was validated by interviewees at all universities, who noted a common practice for faculty at their institutions to store data “under their desks in spare hard drives.” However, they also said that, with the rise in free cloud storage, scholars were moving their research data online to services such as Dropbox and Google Drive. While these online storage options increase the accessibility of data for the research team, the data may remain out of reach of others because access is limited to those with links shared by the researchers. The use of public repositories removes this limitation. Hahnel et al. (2023) found that 92% of their surveyed respondents believe it is important to have research data available in a public repository, which leads us to formulate our second DP in Table 4.

Table 4. Principle of Accessibility (DP2)

Components of the design principle schema (Gregor et al., 2020)	
Aim, enactor, implementer, and user	For designers of RDM solutions (implementers) to allow interested parties (users) the opportunity to assess the results of a research project conducted by a research team (enactors) and to access the data over a period of time
Context	Open science research
Mechanism	Ensure the availability of research data by storing it in public repositories
Rationale	Storing research data in public repositories enhances open science by ensuring that research data is available to interested parties

The above principle shares similarities with the accessibility principle of the FAIR framework (GO FAIR, 2016) in that data should be stored in a way that allows users to access it, even if they are not part of the original research project. We illustrate later how modern decentralized technologies, such as IPFS, can have an essential role in fulfilling the accessibility principle. Moreover, accessibility goes hand in hand with having a stable and long-term data repository, which relates to the DP we discuss next.

4.2.3 Persistence

Often, new research endeavors build on scholarship produced years and often decades in the past. Consequently, accessibility of research data cannot be limited to the near term, which is why mandates from funding agencies often require data to be stored and maintained for extended periods of time. For example, one of the interviewees highlighted how the U.S. National Institutes of Health (NIH) tends to require research data to be publicly available for at least three years and, sometimes, in perpetuity to comply with Health Insurance Portability and Accountability Act (HIPAA) policies. Within this context, cloud storage, an option we discussed in the previous subsection, may not always remain free, making the persistence of such storage options uncertain. For example, one of the interviewees noted how his university had been using Google's suite of apps, which had previously provided unlimited storage space. However, Google's decision to limit the amount of free storage had sent many faculty and students scrambling to find alternative options for storing their larger research datasets. These observations lead to the DP in Table 5.

Table 5. Principle of Persistence (DP3)

Components of the design principle schema (Gregor et al., 2020)	
Aim, enactor, implementer, and user	For designers of RDM solutions (implementers) to allow interested parties (users) the ability to access a research team's (enactors) data over an extended period of time
Context	Open science research
Mechanism	Ensure the persistence of research data by maintaining it in public repositories with redundant data storage capabilities
Rationale	Maintaining research data in public repositories with redundant storage capabilities enhances open science by ensuring that research data is available to interested parties for extended periods of time

Similar to the accessibility principle, we shall later illustrate the role decentralized technologies, such as IPFS and blockchain technology, may play in ensuring persistence. Their redundant nature alleviates problems with individual storage options that may reduce the availability of research data as time passes. For example, investigating how the availability of research data changes with time, Vines et al. (2014) found that the odds of data actually being "available upon request" from data owners fell by 17% each year following the publication of a paper based on the data. This result offers evidence for the assertion that research data cannot be reliably preserved by individual researchers and, as such, decentralized technologies might offer more effective services for storing data for prolonged periods of time.

4.2.4 Affordability (Financial)

Few researchers have the personal resources to conduct the kinds of research they wish to pursue, which is why the academic institutions that employ them provide research support and why researchers seek additional resources from funding agencies. Financial affordability was reflected in our interviews, where the experts we interviewed stated that cost-effectiveness was required for any solution they considered to support RDM, and the cost-effectiveness had to be both in the initial setup and in the ongoing support of the solution. However, financial affordability should not be considered only from the perspective of data gathering and storage but also from the perspective of data querying by interested stakeholders (e.g., other members of the scientific community). On that note, the NSF's (n.d.) data-sharing policy includes the provision that funded researchers are expected to make their data available to others "at no more than incremental cost." This provision is to prevent financial barriers that would prohibit interested parties from engaging in research verification (e.g., other members of the scientific community). Therefore, any solution that intends to promote a more reliable scientific process must include affordability (financial) as a DP. The affordability principle we describe in Table 6 is well aligned with scholars' concerns regarding research costs. For example, Hahnel et al. (2023) found that 33% of the 2,010 principal investigators they surveyed indicated the "cost of data sharing as something that puts them off data sharing."

Table 6. Principle of Affordability (Financial) (DP4)

Components of the design principle schema (Gregor et al., 2020)	
Aim, enactor, implementer, and user	For designers of RDM solutions (implementers) to allow interested parties (users) affordable access to a research team's (enactors) data
Context	Open science research
Mechanism	Ensure that technologies for research data collection and storage, as well as for data querying, are affordable
Rationale	Affordable storage and access enhance open science by ensuring that research data is available to all interested parties, regardless of their ability to pay for those privileges

We shall illustrate the role decentralized technologies, such as IPFS and blockchain technology, can play in making data storage and retrieval financially affordable.

4.2.5 Trustworthiness (Data Integrity)

Our interviewees noted that getting more researchers on a common data-sharing platform would require reducing social barriers to using that platform. Barriers that were noted included a lack of knowledge of what options were available, skill deficiencies in accessing and using available platforms, and limited trust in the platforms. One of the expected outcomes of the open science movement is greater trust in science, with scholars claiming that “if it can achieve its aims of complete clarity and full publicly available content, open science has the potential to become a new trust technology, of benefit to both the scientific community and public groups” (Grand et al., 2012, p. 686). Consequently, trust is an essential DP for technologies that facilitate open science. However, trust can be defined from many different dimensions. In the following DP in Table 7, we define trust in terms of data integrity.

Table 7. Principle of Trustworthiness (Data Integrity) (DP5)

Components of the design principle schema (Gregor et al., 2020)	
Aim, enactor, implementer, and user	For designers of RDM solutions (implementers) to assess the results of the research and to instill trust in the technologies used by interested parties (users) in the evaluation of the methods and results of a research project conducted by a research team (enactors)
Context	Open science research
Mechanism	Ensure that technologies for research data collection and storage do not tamper with or corrupt collected data
Rationale	Technologies that maintain data integrity enhance open science by instilling trust in the collection and storage of research data

The above principle reflects the fact that inaccurate or inconsistent data can lead to incorrect research findings and, thus, poor insights and decisions. Consequently, data integrity failures may help to erode trust in the scientific process. Once again, we shall highlight the role that decentralized technologies, such as IPFS and blockchain networks, might play in ensuring data integrity. Specifically, their immutability prevents research data from being tampered with.

5 Artifact Development, Demonstration, and Evaluation

This section describes the third, fourth, and fifth DSRM phases, which focus on designing an artifact. This includes “determining the artifact’s desired functionality and its architecture” (Peppers et al., 2007), demonstrating the use of the artifact, and comparing the desired and achieved functionality.

5.1 Artifact Development and Demonstration

Our focus in this section is to illustrate how one of the most popular data collection instruments in IS — namely, online surveys — can be redesigned using a combination of decentralized technologies, so that the resulting data collection and storage process satisfies the previously defined TAPAT DPs. We do so by recreating the survey by Anderson et al. (2023), in which the authors collected data from patients to understand their preferences regarding healthcare consent. Throughout the process, we clearly define concrete *design features* that instantiate the TAPAT DPs. Figure 3 shows the architecture of the developed artifact.

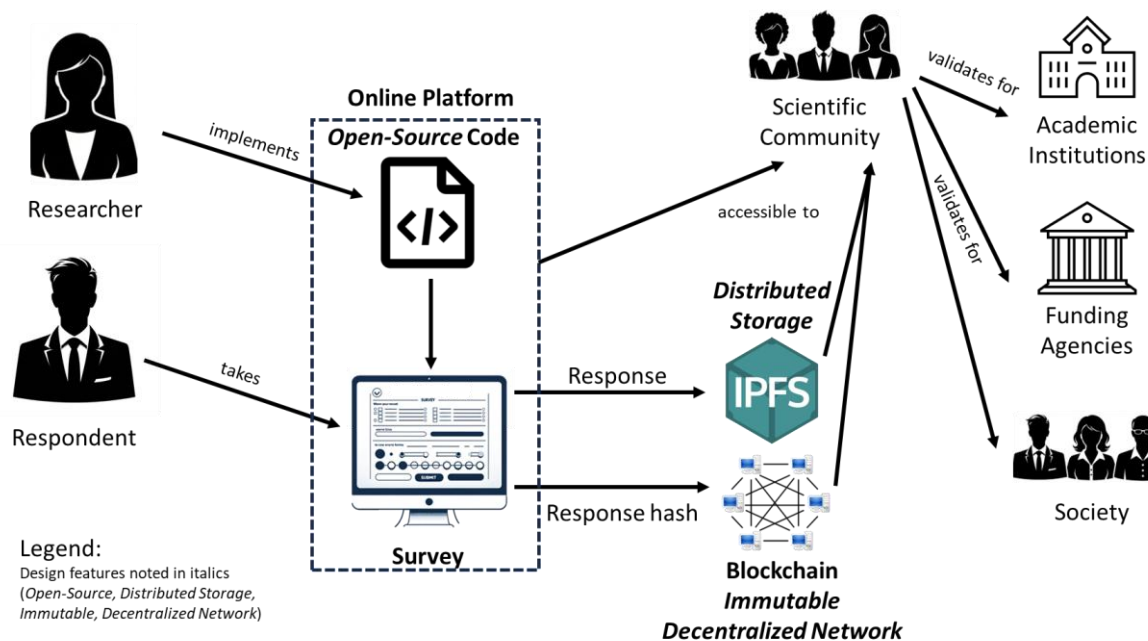


Figure 3. Architecture of the Proposed Artifact

Similar to most modern surveys, a respondent starts by accessing our artifact through a standard web-based application. We developed that application using the Streamlit framework in conjunction with the programming language Python. Importantly, the source code for the application is *open source* and publicly available, features that represent our first design feature (DF1 in Figure 2). By making the code open source, we promote *transparency* (DP1), allowing anyone to inspect how the survey operates, verify its logic, and ensure that no hidden mechanisms bias the results. This openness also supports *affordability* (DP4) by enabling researchers, institutions, or organizations with limited resources to freely reuse, adapt, or build upon the existing code, thereby lowering barriers to entry and fostering wider adoption.

Although making the survey code open source is clearly desirable for transparency and affordability, we argue it is not enough because one must still trust that the code used in the live survey (i.e., deployed application) matches the shared open-source code. Our suggestion to tackle this issue through the second design feature (DF2 in Figure 2) is to host the survey code on a *collaborative development platform* that automatically executes the uploaded source code to produce a live survey. We argue that such a platform improves *transparency* (DP1) in the data collection and storage process, as anyone can inspect the source code — which is timestamped — and access all details, such as what data was collected and where it is being stored. Moreover, the collaborative nature of those platforms enables open science by integrating tools for community engagement, e.g., from creating discussions to pull requests for code changes. In our artifact, we hosted the survey source code alongside the running application (live survey) on the Hugging Face platform.³ Hugging Face offers free hosting tools, which support *affordability* (DP4).

Overall, the survey interface (application's front end) and user experience are no different than other web-based survey applications. The crucial difference between other web-based surveys and our artifact is that our artifact is a DApp where the collected data is stored using decentralized technologies, which represents our third design feature (DF3 in Figure 2). Specifically, the data collected from survey respondents is first saved to IPFS. Although that data is publicly available and can be retrieved by anyone using the generated hash value representing a survey response, a major unsolved issue is that hashes must still be indexed and timestamped. Our artifact fixes this issue by combining IPFS with public blockchains. Specifically, we have developed a smart contract⁴ on the Ethereum network to store IPFS

³ The running survey, source code, and deployment tutorial are available at <https://huggingface.co/spaces/decentralizedsurvey/survey>.

⁴ Information about the deployed smart contract is available at <https://sepolia.etherscan.io/address/0x42b76d8c32f914630627bf924bd1e06055673cf8#code>.

hashes. After storing data on IPFS, the survey application stores the hash value of that data on Ethereum through a smart contract transaction. Given that any Ethereum transaction is timestamped, our artifact can then leverage free and open storage through IPFS while timestamping each survey response using a public blockchain. In this way, we can follow blockchain usage best practices where “the shared ledger [blockchain] itself primarily locks in cryptographic proof of events” (M. Lacity & Van Hoek, 2021). Moreover, the deployed smart contract also serves as an indexing system by storing information needed to retrieve survey responses from the IPFS network. In order to retrieve the survey responses from IPFS, one simply needs the address of the smart contract on the blockchain.⁵ With that address, one can freely query the blockchain to retrieve all the previously stored IPFS hash values and, subsequently, retrieve each survey response from IPFS based on those hash values. By leveraging decentralized storage, our artifact ensures *accessibility* (DP2), *persistence* (DP3), *affordability* (DP4), and *trustworthiness* (DP5) of research data since IPFS provides a distributed, tamper-resistant network where files remain available as long as at least one node retains a copy. Combining this with a public blockchain like Ethereum ensures that each IPFS hash is permanently recorded and timestamped in an immutable ledger. This dual-layered design removes reliance on any single server or authority, making the data resilient to loss or censorship and accessible to anyone with the smart contract address.

5.2 Artifact Evaluation

The fifth DSRM phase involves “comparing the objectives of a solution to actual observed results from use of the artifact in the demonstration” (Peppers et al., 2007). We next evaluate the two main contributions from our work, namely, the TAPAT DPs and the proposed artifact. Venable et al. (2016) developed a framework for evaluation in design science in which evaluations are dimensionalized as formative/summative and naturalistic/artificial. We began our study with formative evaluations that included interviews with RDM experts (naturalistic formative), RDM workshop attendance (naturalistic formative), and a study of relevant RDM documents (artificial formative), all designed to validate whether we were identifying the correct requirements for our design. We also conduct three summative evaluations. First, we self-evaluate how well the proposed artifact aligns with the DPs defined in Section 4 in the context of a broader deployment strategy (artificial summative). Then, we investigate the financial affordability aspect of our proposed artifact by performing a realistic cost analysis (artificial summative). Finally, we report the results of an evaluation conducted through a focus group of IS researchers, discussing the pros and cons of our DPs and artifact (naturalistic summative).

5.2.1 Informed Arguments

Our first evaluation analysis concerns an “informed argument” (Hevner et al., 2004); i.e., we use relevant information to build a convincing argument for the utility of the proposed artifact and its features. In assessing the artifact for *transparency*, we note that having an open-source survey application means interested parties can scrutinize, replicate, or build upon the research process. Interested parties would primarily be other members of the scientific community but could also include academic institutions, funding agencies, or any member of society with an interest in scrutinizing the research process. This directly addresses the RDM issue of limited availability of research data for validation and replication. Hosting that open-source survey code on modern collaborative platforms further boosts transparency so that others can confirm that the live survey used to collect data accurately reflects the shared source code. In conjunction with decentralized data storage, the open-source aspect implies that anyone can inspect how and what data was collected. The result is a digital trail that enhances the credibility and reliability of the research findings by inhibiting selective disclosure of data. Besides satisfying the transparency DP, the above features facilitate compliance with the stringent open-access policies of various research funding institutions, such as from the Gates Foundation (2024), which states that accepted articles “must be accompanied by a Data Availability Statement that describes where any primary data, associated metadata, original software, and any additional relevant materials necessary to understand, assess, and replicate the reported study findings in totality can be found.”

Data *accessibility* and *persistence* are two other TAPAT principles. They denote that research data must be publicly accessible when needed and remain intact over time. Our proposed artifact satisfies the persistence DP by using reliable storage technologies distributed across several independent devices in a peer-to-peer network. Such redundancy implies that data is publicly available and can be accessed from

⁵ A sample application is available at <https://huggingface.co/spaces/decentralizedsurvey/surveydata>.

many different endpoints across the globe, thus satisfying the accessibility principle. The use of IPFS in conjunction with public blockchains also means that we have a unified repository of data, which solves a major concern that is shared by 41% of the respondents in the work by Hahnel et al. (2023), namely, “finding appropriate repositories of data.”

Regarding the *trustworthiness* property of our solution, we note that censorship resistance is a consequence of data storage distribution and decentralization since it becomes significantly harder for any single entity to exert control over the entire network and, thus, take down or corrupt data. Thus, in conjunction with IPFS, blockchains’ append-only, timestamped, and immutable nature can boost trust in the scientific process in that data manipulation is not feasible, leading our artifact to satisfy the trustworthiness (data integrity) DP. Anyone can calculate the hash value of survey responses stored on IPFS and check whether that value matches what is stored and indexed on the blockchain.

5.2.2 Cost Analysis

Any solution aiming to boost the reliability of research data collection and storage processes must be *affordable* if it is to be widely adopted. We address the financial affordability aspect of our solution by discussing the costs behind using the four major components in our proposed design and artifact: (1) the survey application hosting platform, (2) IPFS, (3) smart contract deployment, and (4) blockchain storage. Under the evaluation methods by Hevner et al. (2004), we classify this cost discussion as a mix of static (fixed cost) and dynamic (variable cost) analysis.

Collaborative platforms for hosting source code and applications, such as GitHub and Hugging Face, have different price tiers depending on the required computing resources. At the lower end, the hosting service is free for applications that do not require demanding computations. For example, the 16GB of RAM and two vCPUs currently offered for free by Hugging Face are more than enough to run our survey DApp smoothly. We believe this will also be the case for the vast majority of IS-related surveys.

Similar to application and code hosting, data storage on IPFS is also free as long as one is willing to join the network as a node and pin the collected data. Alternatively, as we discussed in Section 2, one can use API-based services that enable uploading and pinning data on IPFS without joining as a node. For our prototype, we used the services provided by the Infura organization.⁶ At the time of writing, Infura offers a plan that includes the uploading and pinning of up to 5GB of storage for free, which, based on our experience, is considerably more storage capacity than needed by a standard web-based survey.

Next, we estimate the costs of deploying a smart contract and storing IPFS hashes using a public blockchain. At the time of writing, the USD costs of deploying the smart contract and storing a single IPFS hash on Ethereum are, respectively, \$32.45 and \$3.78, whereas the costs associated with Polygon and Fantom are, respectively, \$0.074 and \$0.0078 for smart contract deployment and \$0.0087 and \$0.0009 to store each IPFS hash. Putting the above numbers together, we can now estimate the total cost of running our redesigned survey. Given that Anderson et al. (2023) recruited 200 participants, the total one-time cost of running the survey and storing the data would be between \$0.19 and \$788.45, depending on the choice of a blockchain network. Due to the open nature of IPFS and public blockchains, there is no cost to access any stored data. The above low-end estimation highlights the affordability (financial) aspect of our solution, which leads us to believe that different research communities can indeed widely adopt it.

5.2.3 Focus Group

In our final evaluation, we hosted a focus group of IS researchers who had experience in data collection using survey instruments. Focus groups provide qualitative insights that can reveal user preferences, pain points, and contextual factors (Tremblay et al., 2010) and have been used in the evaluation of design science artifacts (Trieu et al., 2024). The goals for the focus group were to facilitate a comprehensive assessment of the TAPAT DPs, evaluate the effectiveness of the proposed artifact, and identify potential barriers to adoption. In planning for the focus group, we followed established guidelines in the social sciences (Bloor et al., 2000) and design science (Tremblay et al., 2010). For example, the recommended number of participants is at least four but no more than 12 so that a sufficient range of perspectives is captured while also enabling all participants to contribute meaningfully to the discussion. Our focus group consisted of seven IS faculty members selected to ensure a diversity of research experience and survey expertise. Table 8 shows anonymized information concerning the participants.

⁶ <https://www.infura.io/product/ipfs>

Table 8. Focus Group Participant Roles

Participant	Position
P1	Full Professor
P2	Full Professor
P3	Associate Professor
P4	Associate Professor
P5	Assistant Professor
P6	Assistant Professor
P7	Assistant Professor

The focus group was organized into two parts. First, the facilitator explained the research and then provided a demonstration of the prototype in a 30-minute presentation. Thereafter, the participants were invited to discuss and evaluate what was presented. The discussion was structured around three guiding questions:

1. Are the TAPAT DPs valid?
2. For what kind of data collection are they useful/not useful?
3. What are your thoughts on the prototype regarding usability, usefulness, and barriers to adoption?

We note that our goal in the focus group was more on the completeness and overall validity of the principles as a set, rather than examining each one in greater individual depth. That being said, regarding the first question, the participants affirmed the validity of the TAPAT DPs. For example, P5 acknowledged the value of transparency, saying “... *this framework could be very helpful ... [it is] monitoring all the experience you [researcher] are doing, making sure you are just not cherry picking a good one [data/result].*” Noting the trustworthiness (data integrity) of data reusability, P4 said that “*If there's a violation of integrity in the data, someone else can use the data not knowing that has happened and then they become susceptible to the [integrity] problem. But now this [principle] solves that a little bit; it just gives you that ability to be able to say, 'Okay, well, since they [original researchers] have done it the right way, that's the reason why we are using this data.'*” Regarding the affordability (financial) principle, P5 stated, “*We [academics] do face a very critical budget-constrained situation. So, if this open platform can allow us to significantly reuse survey data ... we'll have more data at our fingertips.*” The participants also discussed the range of contexts in which the TAPAT principles would be applicable, mentioning both academic and industry settings. For example, P1 said, “*I think that companies can really start demonstrating the integrity of the decisions they're making based on data they collected, and if they're willing to put that [data] out there, I think we, the consumer base, will embrace that [practice].*” He went on to suggest that, “*It's just like sustainability; everybody is doing sustainability now because we all want companies to be sustainable. I think something like this [TAPAT principles] would catch on in industry as well.*”

Concerns were raised regarding implicit assumptions behind the TAPAT principles, such as the need for Internet access. P3 pointed to research conducted in contexts where “*you can't gather the data electronically.*” To this critique, we acknowledge that the TAPAT principles do require connectivity because, without it, manual intervention is required, which raises new transparency and trust issues. The focus group discussed an example where a person collects data through paper surveys and then manually inputs the data into a digital system. It was agreed that this necessarily breaks the transparent and trustworthy data flow facilitated through the TAPAT principles, as the integrity of the whole process now relies on the person entering the data.

The participants saw value in the TAPAT-compliant survey we had constructed. For example, P4 said, “*The problem with surveys is they are black boxes. I don't know what you [researcher] did. Now [using the proposed artifact, one can see] we did all the right things; [for example] how the questions were arranged.*” P1 agreed with that assessment and highlighted a rather common issue researchers face, which is “*How many times have you used somebody else's measures and they don't hold together, right? Where, for them [original researchers], they held together perfectly.*”

At the same time, concerns were raised about the technical nature of the solution. For example, P7 asked about ease of use stating, “*You mentioned that you use Python code to create such surveys; I'm wondering, from a user side, is it easy to use?*” Although beyond the scope of this research, it is important to note the potential for developing interfaces that allow for the visual creation of surveys, which can then

be automatically translated into code and deployed to appropriate hosting platforms. This approach addresses the concern raised by P1, who pointed out that “*not everybody writes code* [to create a survey interface].” Furthermore, P7 suggested that our solution can be adapted by established survey providers such as Qualtrics and Survey Monkey. Survey providers could maintain their current functionality and user interface while changing the backend systems to store data on IPFS and blockchain, thereby becoming a DApp. A different aspect of the technical nature of the solution was raised by P3, who asked about detecting whether someone is creating fake data with the survey saying, “*Am I good enough to write some code that will take the surveys for me?*” We acknowledge that issue and contend that it is an unfixable problem because the only solution would be to record the people who are taking the survey. We return to this issue in the following section.

The focus group recognized key RDM issues that included limited availability of research data for validation and replication, limited data archival and retention capabilities, high cost to manage research data, misrepresentation, and flexible data analysis. Overall, the feedback from the focus group strongly supports the validity and applicability of the TAPAT DPs. Participants appreciated the framework’s potential to enhance transparency, trustworthiness (data integrity), and affordability (financial) in data collection and reuse, for relevant stakeholders (e.g., scientific community, academic institutions, funding agencies, and society). Additionally, they thought this solution also had potential value for use in industrial contexts, extending the practical significance of our research. The focus group discussion also highlighted important technical considerations, such as the necessity of Internet access for seamless application access and the usability challenges associated with technical implementations. These insights corroborate our belief in the need for TAPAT principles to guide researchers and suggest a promising path for further development and adoption of our proposed artifact, as we further elaborate upon in the discussion section.

6 Discussion

The evaluation presented in the previous section sets up our contributions as well as certain limitations and points of opportunity for future extensions of our work. In this section, we articulate our contributions as well as the limitations of our research and suggest ways in which our research could be extended.

6.1 Contributions

Gregor and Hevner (2013) provide a framework for positioning and presenting design science research in which contributions can be classified as routine, improvement, exaptation, and invention. Our improvement-based research addresses an important practical and theoretical problem of how to design solutions for reliable RDM (Gregor & Hevner, 2013).

Our first contribution to the knowledge bases of design science and RDM domains is the development of DPs for the design and development of RDM systems. Through interviews with RDM experts, participation in sessions of related workshops, and a review of the relevant literature, we developed the TAPAT principles: *transparency*, *availability*, *persistence*, *affordability (financial)*, and *trustworthiness (data integrity)*. These principles offer clear, actionable guidelines for researchers to enhance their data collection strategies. Similarly, they serve as criteria for the scientific community (e.g., reviewers and editors), research institutions, funding agencies, and other members of society to evaluate the robustness and provenance of data collection and management plans in the underlying research of submitted manuscripts.

Our second contribution is a new design and concrete artifact that offers a transparent, trustworthy, and affordable data collection process while providing highly persistent and accessible data storage. Focusing on online survey instruments, one of the most popular data-gathering artifacts in IS research, our proposed artifact integrates open-source survey applications running on collaborative platforms with decentralized technologies, namely, the IPFS and public blockchains, to address critical issues in RDM. By using IPFS to store survey responses and blockchain technologies to store hash values of those responses, our system ensures that survey responses are stored in a decentralized (i.e., highly available and persistent) and immutable (i.e., trustworthy) manner. As such, this architecture not only addresses concerns regarding data manipulation and fabrication but also promotes a more transparent and collaborative research environment. Our cost analysis demonstrates the practical utility and financial affordability of our solution, suggesting its potential for wide adoption across various research communities. Moreover, we see our proposed artifact as a concrete contribution toward the ongoing

development of decentralized science⁷ (DeSci), an emerging paradigm that helps establish provenance by leveraging decentralized technologies to transform how scientific research is conducted, funded, and disseminated.

Our third contribution is the ability of our proposed solution to support preregistration of a research project, which involves “committing to analytic steps without advance knowledge of the research outcomes” (Nosek et al., 2018). While advanced manipulation strategies may be employed to undermine the proposed artifact (e.g., one can create several copies of a survey, each associated with one smart contract that stores hashes of different survey responses, allowing selective reporting of only the smart contract associated with desired responses), a preregistration plan would prevent that type of manipulation. Preregistration would require authors to document their “exact model, their proposed control variables, how they will measure each variable, their hypotheses, and how they will exclude and include any data” (Bogert et al., 2021). While most discussions of preregistration focus primarily on data analysis, our artifact enables researchers to preregister the entire data collection and storage plan. For example, one can preregister the source code of the survey application, a link to the live survey running on a transparent hosting platform, and the address of a smart contract where hash values representing response data will eventually be stored. Such a preregistration plan enables interested stakeholders from the scientific community to understand *ex ante* how and which data will be collected and where the data will be stored *ex post*.

6.2 Limitations and Future Research

A key limitation of our work is that studies in which the capture and storage of confidential data are required would need to be handled through a different process involving authorized access, given that our solution stores the research data in a publicly accessible format (e.g., any data classified as protected health information under the HIPAA could not be stored using our solution). More broadly, it is important to acknowledge that our solution imposes constraints on researchers by requiring careful consideration of what data is collected. For example, researchers should avoid collecting personally identifiable information or any data capable of identifying individuals, adhering to guidelines already mandated by many institutional review boards. A related issue arises due to the use of decentralized technologies in that data removal can be challenging, even for valid reasons, such as complying with right-to-be-forgotten laws (Carvalho et al., 2021). Our artifact still enables data removal when the IPFS nodes agree to “unpin” certain data, such as a survey response. Although the hash value of that response will be forever stored on the blockchain network due to its append-only property, the one-way nature of hash functions means that extracting responses from hashes is not computationally feasible, and thus, removing data from IPFS effectively enables data erasure.

Given the above discussion, a critical challenge to the widespread adoption of TAPAT-based solutions is the need for institutional governance alignment. Many research institutions are bound by stringent internal compliance and data protection policies that limit their ability to outsource or decentralize storage and processing. Implementation at scale would require substantial buy-in from university legal and IT departments, ethics boards, and data stewards. Further governance frameworks must be developed to support how decentralized records would be monitored, validated, or possibly revoked within the boundaries of institutional responsibility.

While our research on data management addresses important issues for research integrity and trust in scientific inquiry, other relevant issues cannot be tackled by our solution alone. For example, data fabrication is an issue that often makes headlines because of the extent to which it undermines trust in the conduct of research (Scheiber, 2023). While the proposed TAPAT principles and artifact enhance transparency in the data collection process and help deter certain data fabrication practices, they cannot currently verify the authenticity or intent of survey takers. As previously noted, ensuring this would require recording participants, an approach likely to violate privacy policies when dealing with human subjects. That said, the TAPAT principles and artifact would increase opportunities for replication that would be more likely to reveal falsification practices. It would also encourage best practices in the proper conduct of research.

Moreover, issues that include undisclosed conflicts of interest and research data theft also fall outside the scope of this work, as such problems are unlikely to be fully resolved through technical solutions alone.

⁷ <https://ethereum.org/en/desci/>

Addressing them effectively will require policies and regulations that shape research practices and guide researchers in adhering to ethical standards. In short, this is a sociotechnical challenge that demands a sociotechnical solution at the macro level, even if some individual components may be primarily technical (Lei & Ngai, 2023).

While our proposed solution marks a significant step forward in addressing some of the prevalent issues in RDM, it also opens avenues for future research. Future studies could experimentally test our solution by creating situations that introduce issues the artifact is expected to address and then asking participants to use the system and assess whether those issues are detected. While our study focused on online surveys in IS research, future work could explore the applicability of our approach to other types of research data and methodologies, especially those with different data collection and storage requirements. For example, audio or video recordings of interviews or focus groups could be stored on decentralized storage platforms like IPFS, and the corresponding metadata (e.g., hash value) could be immutably recorded on a blockchain. Future studies should also contemplate governance dimensions (Beck et al., 2018), such as the ethical and legal implications of decentralized data storage for research data, especially concerning privacy, consent, and data-sharing regulations across different jurisdictions, as the balance between transparency and privacy, particularly in sensitive research areas, requires careful consideration.

7 Conclusion

In this paper, we followed a design science methodology to present a novel approach for enhancing reliability in the process of gathering, storing, and sharing research data. This approach included a set of DPs for designing RDM solutions. We also provided an instantiation, as a proof of concept, of an artifact based on our DPs of transparency, accessibility, persistence, affordability, and trustworthiness. TAPAT is one set of principles available for the governance of research data that is complementary to other similar frameworks that provide guidance for the appropriate collection and storage of data. The best frameworks may be a combination of guidelines and criteria developed by incorporating relevant features from multiple sources. For example, the principle of accessibility significantly overlaps with FAIR, as mentioned previously. There are also other frameworks that scholars can use in conjunction with our proposed solution and perhaps extend TAPAT to cover other unique research domains. For example, the TRUST principles (Lin et al., 2020) of transparency, responsibility, user focus, sustainability, and technology overlap with TAPAT, specifically with regard to transparency and, more generally, by the overarching goal of trustworthiness (data integrity). Moreover, the CARE principles for Indigenous data governance (S. Carroll et al., 2020) might be relevant for data collection from specific, sensitive populations. Together with such frameworks, our work opens new possibilities for addressing long-standing challenges in RDM, laying the groundwork for a more open, collaborative, and trustworthy scientific process.

References

- Anderson, C., Carvalho, A., Kaul, M., & Merhout, J. W. (2023). Blockchain innovation for consent self-management in health information exchanges. *Decision Support Systems*, 174(2), Article 114021.
- APA. (2017). *Ethical principles of psychologists and code of conduct*. American Psychological Association. <https://www.apa.org/ethics/code>
- Beck, R., Müller-Bloch, C., & King, J. L. (2018). Governance in the blockchain economy: A framework and research agenda. *Journal of the Association for Information Systems*, 19(10), 1020–1034.
- Bellgard, M. I. (2020). ERDMAS: An exemplar-driven institutional research data management and analysis strategy. *International Journal of Information Management*, 50(C), 337–340.
- Bloor, M., Thomas, M., & Frankland, J. (2000). *Focus groups in social research*. Sage Publications.
- Bogert, E., Schecter, A., & Watson, R. T. (2021). Preregistration of information systems research. *Communications of the Association for Information Systems*, 49(1), 110–125.
- Brendel, A. B., Lembcke, T.-B., Muntermann, J., & Kolbe, L. M. (2021). Toward replication study types for design science research. *Journal of Information Technology*, 36(3), 198–215.
- Burton-Jones, A., Boh, W. F., Oborn, E., & Padmanabhan, B. (2021). Advancing research transparency at MIS Quarterly: A pluralistic approach. *MIS Quarterly*, 45(2), iii–xviii.
- Carroll, S., Garba, I., Figueroa-Rodríguez, O., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodríguez-Lonebear, D., & Rowe, R. (2020). The CARE principles for Indigenous data governance. *Data Science Journal*, 19, 1–12.
- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodríguez-Lonebear, D., & Rowe, R. (2023). *The CARE principles for Indigenous data governance*. Open Scholarship Press (Curated Volumes: Policy).
- Carvalho, A., Merhout, J. W., Kadiyala, Y., & Bentley, J. (2021). When good blocks go bad: Managing unwanted blockchain data. *International Journal of Information Management*, 57(4), 1–12.
- Dennis, A. R., Brown, S. A., Wells, T. M., & Rai, A. (2020). Replication crisis or replication reassurance: Results of the IS replication project. *MIS Quarterly*, 44(3), iii–x.
- Dennis, A. R., & Valacich, J. S. (2015). A replication manifesto. *AIS Transactions on Replication Research*, 1(1), 1–4.
- Doyle, C., & Luczak-Roesch, M. (2020). *This paper is an artefact: On open science practices in design science research using registered reports*. Hawaii International Conference on System Sciences.
- Doyle, C., Luczak-Roesch, M., & Mittal, A. (2019). *We need the open artefact: Design science as a pathway to open science in information systems research*. International Conference on Design Science Research in Information Systems and Technology, Worcester, MA.
- Ensuring the integrity of research data. (2009). In *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. National Academies Press (US).
- Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, 105(2), 203–206.
- Friedman, L. S., & Richter, E. D. (2004). Relationship between conflicts of interest and research results. *Journal of General Internal Medicine*, 19(1), 51–56.
- Gates Foundation. (2024). *Open access policy*. Bill & Melinda Gates Foundation. <https://openaccess.gatesfoundation.org/open-access-policy/>
- Gentemann, C. (2023). Why NASA and federal agencies are declaring this the Year of Open Science. *Nature*, 613(7943), Article 217.
- GO FAIR. (2016). *FAIR principles*. <https://www.go-fair.org/fair-principles/>
- Gorman, D. M., Elkins, A. D., & Lawley, M. (2019). A systems approach to understanding and improving research integrity. *Science and Engineering Ethics*, 25(1), 211–229.

- Grand, A., Wilkinson, C., Bultitude, K., & Winfield, A. F. T. (2012). Open science: A new “trust technology”? *Science Communication*, 34(5), 679–689.
- Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: The anatomy of a design principle. *Journal of the Association for Information Systems*, 21(6), 1622–1652.
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355.
- Hahnel, M., Smith, G., Scaplehorn, N., Schoenenberger, H., & Day, L. (2023). *The state of open data 2023*.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.
- Hevner, A. R., Parsons, J., Brendel, A. B., Lukyanenko, R., Tiefenbeck, V., Tremblay, M. C., & vom Brocke, J. (2024). Transparency in design science research. *Decision Support Systems*, 182(1), 1–11.
- Hosseini, M., Shahri, A., Phalp, K., & Ali, R. (2018). Engineering transparency requirements: A modelling and analysis framework. *Information Systems*, 74(P1), 3–22.
- Hui, W., Lui, S. M., & Lau, W. K. (2019). A reporting guideline for IS survey research. *Decision Support Systems*, 126(23), 1–10.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532.
- Kaiser, J., & Brainard, J. (2023). Ready, set, share! *Science*, 379(6630), 322–325.
- Kanza, S., & Knight, N. J. (2022). Behind every great research project is great data management. *BMC Research Notes*, 15(1), Article 20.
- King, G. (1995). Replication, replication. *PS: Political Science & Politics*, 28(3), 444–452.
- Lacity, M., & Van Hoek, R. (2021). What we’ve learned so far about blockchain for business. *MIT Sloan Management Review*, 62(3).
- Lacity, M. C. (2022). Blockchain: From Bitcoin to the Internet of value and beyond. *Journal of Information Technology*, 37(4), 326–340.
- Lei, C. F., & Ngai, E. W. T. (2023). Blockchain from the information systems perspective: Literature review, synthesis, and directions for future research. *Information & Management*, 60(7), 1–31.
- Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giarretta, D., De Giusti, M., L’Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST principles for digital repositories. *Scientific Data*, 7(1), Article 144.
- Lyon, L. (2016). Transparency: The emerging third dimension of open science and open data. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 25(4), 153–171.
- Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in IS research and the implications for replication. *Decision Support Systems*, 115, A1–A7.
- McShane, B. B., & Gelman, A. (2022). Selecting on statistical significance and practical importance is wrong. *Journal of Information Technology*, 37(3), 312–315.
- Mertens, W., & Recker, J. (2023). New guidelines for null hypothesis significance testing in hypothetico-deductive IS research. In L. P. Willcocks, N. R. Hassan, & S. Rivard (Eds.), *Advancing information systems theories, volume II: Products and digitalisation* (pp. 385–437). Springer International Publishing.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.
- NSF. (n.d.). *Preparing your data management and sharing plan*. U.S. National Science Foundation. <https://new.nsf.gov/funding/data-management-plan>

- Open Science Framework. (2025). *Open Science Framework*. <https://osf.io/>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77.
- Pekkola, S., Hekkala, R., Rossi, M., & Smolander, K. (2019). The magical “we”: Enhancing collaboration transparency in grounded theory method in information systems research. *Communications of the Association for Information Systems*, 45(1), 251–269.
- Pratt, M. G., Kaplan, S., & Whittington, R. (2020). Editorial essay: The tumult over transparency: Decoupling transparency from replication in establishing trustworthy qualitative research. *Administrative Science Quarterly*, 65(1), 1–19.
- Ritchie, S. (2020). *Science fictions: How fraud, bias, negligence, and hype undermine the search for truth*. Metropolitan Books.
- Robinson, L. D., Cawthray, J. L., West, S. E., Bonn, A., & Ansine, J. (2018). Ten principles of citizen science. In *Citizen science: Innovation in open science, society and policy* (pp. 27–40). UCL Press.
- Scheiber, N. (2023). The Harvard professor and the bloggers. *The New York Times*. <https://www.nytimes.com/2023/09/30/business/the-harvard-professor-and-the-bloggers.html>
- Sen, A., Smith, G., & Van Note, C. (2022). Statistical significance versus practical importance in information systems research. *Journal of Information Technology*, 37(3), 288–300.
- Silver, M. S., & Markus, M. L. (2013). Conceptualizing the sociotechnical (ST) artifact. *Systems. Signs & Actions*, 7(1), 82–89.
- Surkis, A., & Read, K. (2015). Research data management. *Journal of the Medical Library Association*, 103(3), 154–156.
- Tang, J., Jia, T., Chen, H., & Wei, C. (2021). *Research on big data storage method based on IPFS and blockchain*. Proceedings of the 2020 2nd International Conference on Video, Signal and Image Processing, Jakarta, Indonesia.
- Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010). The use of focus groups in design science research. In *Design research in information systems: Theory and practice* (pp. 121–143). Springer US.
- Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2012). Design of an information volatility measure for health care decision making. *Decision Support Systems*, 52(2), 331–341.
- Trieu, V.-H., Vu, H. Q., Indulska, M., & Li, G. (2024). A computer vision-based concept model to recommend domestic overseas-like travel experiences: A design science study. *Decision Support Systems*, 181(5), 1–14.
- UNESCO. (2021). *UNESCO recommendation on open science*. <https://unesdoc.unesco.org/ark:/48223/pf0000379949>
- van Aken, J. E. (2004). Management research based on the paradigm of the design sciences: The quest for field-tested and grounded technological rules. *Journal of Management Studies*, 41(2), 219–246.
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, 25(1), 77–89.
- Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88(4), 428–436.
- Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, 24(1), 94–97.
- Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š, van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, 5(11), 1473–1480.

- Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research*, 3(1), 36–59.
- Wilson, V., Srite, M., & Loiacono, E. (2021). The effects of item ordering on reproducibility in information systems online survey research. *Communications of the Association for Information Systems*, 49(1), 760–799.
- Woo, C., & Yoo, J. (2023). Exploring the determinants of blockchain acceptance for research data management. *Journal of Computer Information Systems*, 63(1), 216–227.
- Zhang, X.-F. (2021). *Application of blockchain technology in data management of university scientific research*. International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Lodz, Poland.

Appendix A: Interview Guide

1. Describe your current role in the university.
2. What kinds of systems does the university have to support faculty for storing and sharing their research data? What are the names of those systems/databases?
3. What types of data can those systems handle?
4. What are the most common types of data that faculty store on those systems?
5. What are some of the most useful features of your existing systems?
6. What are some features or capabilities of the current systems that could be improved upon?
7. How do you promote the use of those systems to faculty?
8. Are there other entities you partner with to extend access to research data outside of the university?
9. What are the different laws, regulations, or policies that govern the collection, storage, and sharing of academic research data?
10. How does the university engage in risk management to protect the academic research data it manages?

About the Authors

Arthur Carvalho is the Dinesh & Ila Paliwal Innovation Chair and an Associate Professor of Information Systems & Analytics at Farmer School of Business, Miami University. Before joining Miami University, Dr. Carvalho worked as an assistant professor at Rotterdam School of Management, Erasmus University. He received his Master's of Mathematics and PhD in Computer Science from the University of Waterloo in 2010 and 2014. Dr. Carvalho's research interests lie at the intersection of artificial intelligence, information systems, and decision analysis. His research has been published in academic journals such as *Decision Support Systems*, *Decision Analysis*, *International Journal of Information Management*, and *Journal of Cleaner Production*. Due to his dedication to students and new educational endeavors, Dr. Carvalho has received prestigious teaching awards, including the Professor of the Year Award at Erasmus University, a Miami University's President Medal, and the Richard K. Smucker Teaching Excellence Award.

Chad Anderson is an Assistant Professor of Information Systems in the Farmer School of Business at Miami University. He received his PhD in Computer Information Systems from Georgia State University. His research interests include the role of technology in the delivery of health care, information security and privacy, and blockchain technology. His work has been published in *Communications of the Association for Information Systems*, *Decision Support Systems*, *Information and Organization*, *Journal of the Association for Information Systems*, *Journal of Management Information Systems*, *MIS Quarterly*, and other journals.

Mala Kaul is an Associate Professor of Information Systems and the Director of Business Analytics in the College of Business at the University of Nevada, Reno. She received her PhD from the Robinson College of Business at Georgia State University. Her research focuses on the design and application of emergent technology, cybersecurity, health information technology, and the foundations of engaged research methods. She has extensive industry experience as an information systems professional in manufacturing, finance, and hospitality industry verticals. Her research has been published in *Management Information Systems Quarterly*, *Journal of Management Information Systems*, *Decision Support Systems*, *European Journal of Information Systems*, *Harvard Business School Case Studies*, and others.

Jeffrey W. Merhout is Associate Professor of IS at Miami University in Oxford, Ohio, where he has taught since 2003. He holds a PhD and an MBA from Virginia Commonwealth University and is a Certified Public Accountant (inactive). His current research interests focus on women in IT, blockchain, information risk management, IT security and information systems auditing, and information technology governance. He has presented and published his research at numerous IS conferences and in journals, including the *Communications of the ACM*, *Communications of the Association for Information Systems*, *Decision Sciences Journal of Innovative Education*, *Decision Support Systems*, *Information Technology & People*, *International Journal of Information Management*, *Journal of the Midwest Association for Information Systems*, *Journal of Information Systems Education*, *Journal of Computer Information Systems*, *International Journal of Accounting Information Systems*, *ISACA Journal*, *Information Systems Control Journal*, *Issues in Information Systems*, *Review of Business Information Systems*, and *The Journal of Effective Teaching*.

Copyright © 2026 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712, Atlanta, GA 30301-2712. Attn: Reprints or via e-mail from publications@aisnet.org.