Research data integrity: A cornerstone of rigorous and reproducible research

Patricia B. Condon[1], Julie F. Simpson[2], and Maria E. Emanuel[3]

Abstract

Research data integrity provides a strong foundation for high quality research outcomes, and it is an essential part of the research data lifecycle due to its critical role in research rigor, reproducibility, replication, and data reuse (the four Rs). Understanding research data integrity is therefore imperative in collaborative interdisciplinary research and collaborative cross-sector research where different norms, procedures, and terminology regarding data exist.

Research data integrity is closely associated with data management, data quality, and data security. Producing data that are reliable, trustworthy, valid, and secure throughout the research process requires purposefully planning for research data integrity and careful consideration of research data lifecycle actions like data acquisition, analysis, and preservation. In addition, purposeful planning enables researchers to conduct rigorous research and generate outcomes that are reproducible, replicable, and reusable. To advance this conversation, we developed two tools: a concept model that visually represents the relationship between data management, data quality, and data security as components of research data integrity, and a schema for implementing these components in practice. We contend that disentangling research data integrity and its components, developing a standardized way of describing their interplay, and intentionally addressing them in the research data lifecycle reduces threats to research data integrity.

In this paper, we break down the complexity of research data integrity to make it more understandable and propose a practical process by which research data integrity can be achieved in a way that is useful for data producers, providers, users, and educators. We position our concept model and schema within the larger dialog around research integrity and data literacy and illuminate the role that research data integrity and its components (data management, data quality, and data security) play in the four Rs. In this paper, we present a concept model and schema for use as tools for instruction/training and practical implementation. Using these tools, we examine the role of research data integrity in rigorous and reproducible research and offer insight into ensuring research data integrity throughout the research process.

Keywords

Research data integrity; data management; data quality; data security; research integrity; data literacy

1. Introduction

The primacy of data in empirical scientific research, trustworthy findings, and the creation of knowledge demands that both institutions and individual researchers allocate adequate resources to ensure research data integrity from the inception of a research project, through the dissemination of findings, and the subsequent sharing of data. In this paper, we focus on the interplay between data management, data quality, and data security to develop a holistic approach to achieving research data integrity in practice and the positive impacts of this approach on research integrity. We aim to transcend disciplinary perspectives to establish a generalized, practical model of research data integrity foundational to the four Rs of research: rigor, reproducibility, replication, and reuse (see Table 1 for definitions).

This paper purposefully discusses the integrity of research data distinct from other types of data. Data integrity can be defined as the ‘state of data (valid or invalid) and/or the process of ensuring and preserving the validity and accuracy of data’ (Ng, 2021). As a concept, however, data integrity is deceptively complex given how it has been used over time and across disciplines and sectors, and therefore not holistically understood by the research community. In this paper, we propose a concept of research data integrity that is comprised of the individual components of data management, data quality, and data security, and realized through documentation and training. Furthermore, we propose that research data integrity is more robust and achievable when approached as a relationship between these components rather than as discrete elements. Understanding these components of research data integrity brings clarity to a complex concept that is difficult to define and implement. It is essential that researchers understand the concept of research data integrity and intentionally implement it in the research lifecycle due to the key role it plays in the achievement of the four Rs. Research data integrity is critical to scientific rigor, and although inability to reproduce or replicate a study’s findings is not always a research data integrity issue, it is an important factor in successful attempts. Moreover, data reuse hinges on research data integrity and the trustworthiness of the original dataset.

Table 1: Definitions of the Four Rs: Rigor, Reproducibility, Replication, and Reuse

Rigor	‘Scientific rigor is the strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation and reporting of results’ (U.S. Department of Health and Human Services, no date).
Reproducibility	‘Reproducibility means computational reproducibility – obtaining *consistent computational results* using the *same input data*, computational steps, methods, code, and conditions of analysis’ (National Academies of Sciences, Engineering, and Medicine, 2019b , p. 6).
Replicability	‘Replicability means obtaining consistent results across studies aimed at answering the *same scientific question*, each of which has *obtained its own data*. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study’ (National Academies of Sciences, Engineering, and Medicine, 2019b , p. 6).
Reuse	‘Data reuse is a concept that involves *using research data for* a research activity or *purpose other than* that for which it was *originally intended*’ (Network of the National Library of Medicine, no date).

* *Italics added for emphasis*

Our interest in exploring the concept of data integrity was raised by a student's question about the difference between data quality and data integrity – a question that the authors were unable at the time to satisfactorily answer for the student. This question was asked in an interdisciplinary graduate seminar on research integrity taught by two of the authors (Emanuel and Simpson) and in which the third author (Condon) guest lectures on the topic of responsible research data management. Not being able to adequately answer the student's question in class revealed the need to further examine and make clear the distinctions among terms related to data integrity. As we tried to refine our own understanding of data integrity, we found that research data integrity was often expressed as being closely associated with data management, data quality, or data security. In general, data management addresses foundational practices to ensure that data are usable and reusable; data quality speaks to the reliability and utility of data; and data security is concerned with safeguarding data from loss or corruption (see Table 2 for working definitions of these terms). Often only vague explanations were provided to distinguish these concepts from research data integrity, or the differentiation was left ambiguous. We concluded that there was a need to further the discussion of research data integrity. We propose to advance the conversation around research data integrity by identifying and disentangling the components of research data integrity, developing a holistic understanding of these components and their interplay, and intentionally addressing their centrality in the research lifecycle.

The aims of this paper are to:

- Establish practical explanations of research data integrity and its components – data management, data quality, and data security – to better articulate the relationships between these terms while clearly conveying their role in the research lifecycle;
- Consider the utility of these components for data producers, providers, users, and educators; and
- Demonstrate the central role of research data integrity in research rigor, reproducibility, replication, and data reuse.

To address these aims, we propose a concept model of the relationships among the components of research data integrity (graphically represented in Figure 1) and a schema for implementing research data integrity in practice, training, and education (see Figure 3). We then discuss using these tools for training and education to foster the four Rs throughout the research lifecycle. Developing this concept model provides a foundation for reducing threats to research data integrity and producing high quality research outcomes.

In the context of this article, we refer to research data as opposed to other types of data, such as organizational, administrative, or product development data. The proposed concept model and schema, however, could be adapted to other types of data. It is additionally not within the scope of this article to provide an in-depth review of the breadth of literature on data integrity and related terms. Rather, we use the literature to provide readers with a foundation for understanding the concept model that we have developed, and we cite materials across disciplines and domains to illustrate the broad reach and complexity of data integrity. Much of the literature addresses data integrity from a discipline- or domain-dependent perspective. Therefore, one of the strengths of the proposed concept model is that we present research data integrity and its components as discipline-independent with the intention that it could be utilized across disciplines.

2. The complexity of data integrity

The National Academies of Sciences, Engineering, and Medicine report, *Fostering Integrity in Research,* explains that '[p]racticing integrity in research means planning, proposing, performing, reporting, and reviewing research in accordance with [the six core values of science: objectivity, honesty, openness, accountability, fairness, and stewardship]' (National Academies of Sciences, Engineering, and Medicine, et al., 2017, p. 38). Integrity of data plays a key role in these core values from objectively analyzing, reporting, and documenting data to responsibly sharing data and code underlying published research. Furthermore, the *Singapore Statement on Research Integrity* includes several professional responsibilities of researchers that align specifically with safeguarding data integrity. These include the responsibilities to 'keep clear, accurate records of all research,' 'share data and findings openly and promptly,' 'create and sustain environments that encourage integrity,' and report 'irresponsible research practices' (World Conference on Research Integrity, 2010, p. 1).

Issues with data integrity are at the center of many cases of research misconduct leading to article retraction (see for instance, Bordewijk, et al., 2021; Carlisle, J. B., 2012; Piller C. and Servick K., 2020; van der Zee et al., 2017). A cursory search of the Retraction Watch Database (Retraction Watch Database, 2018) shows that data integrity issues – such as concerns with or errors in data, unreliable data, or falsification/fabrication of data – account for a concerning number of article retractions. In an editorial about the potential value of data auditing prior to publication to help identify data integrity issues, Shamoo (2020) observes that '[r]esearchers must be rigorous and careful in designing and executing experiments and honest and transparent when reporting data, methods, and results' (p. 325). These responsible data practices are at the core of maintaining data integrity in research and fostering the four Rs.

To promote responsible data practices, data producers, providers, users, and educators in the academic, business, and industry sectors tend to focus on different applications of data integrity. This underscores both the inconsistent use of data integrity as a term as well as the complexity of data integrity as a concept. In academia, data integrity through the management of scientific research data is a key tenet of research integrity education and the practices of scientific and academic professional associations (e.g., U.S. Geological Survey). With more funding agencies now requiring a data management plan as a part of grant applications, increased attention has been given to responsible data management practices and the resources needed to implement them (e.g., National Institutes of Health, 2020). The business sector has embraced the concept of data integrity due to the important role that organizational data play in keeping a company competitive (e.g., sales, marketing, forecasting, revenues versus expenses, and customer privacy) (Caratas, Spatariu and Gheorghui, 2019). Their focus in protecting data integrity, however, is often on the security of data (i.e., keeping data safe from unauthorized access or changes) (Pandey, et al, 2020; Yang, Xiong and Ren, 2020). Certain industries, particularly those in health-related fields (e.g., those whose work falls under the jurisdiction of the U.S. Food and Drug Administration [FDA] regulations), often focus their data integrity efforts on protecting the quality of data due to its importance in the development of their products and interventions (Ahmad, Kumar and Hafeez, 2019; Arroyo-Araujo and Kas, 2022). The FDA, for example, uses the ALCOA (Attributable, Legible, Contemporaneous, Original, and Accurate) standard for data integrity and quality in the current good manufacturing practice (CGMP) for drugs (U.S. Food and Drug Administration, 2018). These multiple applications of data integrity and the lack of a consistent, cross-disciplinary understanding of data integrity among sectors can lead to consequences when communicating, collaborating, conducting research, and sharing information.

The concept of data integrity is often simplified to the discipline-specific applications of data management, data quality, and data security. It is also used simultaneously to describe fundamentally different attributes of data. For example, data integrity may refer to a notion of reliability and trustworthiness of the data; alternatively, it may be considered a characteristic of data throughout the research data lifecycle (National Academies of Sciences, Engineering, and Medicine, 2019a; Ng, 2021; Sandhu, 1993). To add to the confusion, data integrity can be used to describe both the state of data (e.g., whether it is valid or invalid) as well as the process of ensuring and preserving the validity and accuracy of data in a dataset or database (Brook, 2020; Ng, 2021). Finally, data integrity is often categorized as either physical or logical. Physical integrity pertains to the protection of data's wholeness and accuracy as they are stored and retrieved, and logical integrity refers to keeping data unchanged as it is used in different ways in a relational database (Talend, no date). This confusion can be compounded in multi-sector projects or interdisciplinary research where standards, practices, and terminologies may differ.

While useful in certain contexts, the aforementioned applications of the term data integrity lack a holistic approach to identifying and addressing the complexity of data integrity and the relationships among the different components of data integrity. Furthermore, none of the current definitions present either a standardized terminology that can be understood across disciplines or, critically, an overarching, cross-disciplinary framework that addresses all the facets necessary to ensure data integrity throughout the research lifecycle. Subsequently, these applications of data integrity inhibit practitioners (e.g., data producers, providers, and users) and educators from effectively leveraging data integrity and fostering achievement of the four Rs. These varied usages of data integrity do not adequately convey the process required to integrate all of the components into the foundation of a research project. In addition, they do not provide students or early-career researchers a concrete framework for integrating data integrity into the research lifecycle. The lack of consistency among the above examples demonstrates the imprecise use of the term data integrity as both a simplified term and as a multi-dimensional concept depending on the sector or discipline. This imprecision in the understanding and application of data integrity jeopardizes the ability to implement it as a keystone in the research lifecycle.

3. Components of research data integrity and their relationships

To address the definitional and implementation problems raised above, we propose a concept model of research data integrity characterized by the dynamic interplay of its core components of data management, data quality, and data security and as supported by necessary documentation and training. Our first aim is thus to establish practical explanations of these core components to better elucidate the relationships among them. Our objective is to disentangle terminology and look at the interchange between the components to address the limitations when viewing them in isolation from one another rather than as a system. For instance, poor quality data can have integrity. On the other hand, high quality data can be managed according to best practices, but due to issues such as invalid inputs or lack of attribution, the integrity of the data is questionable or compromised. By highlighting the interactions among the components, we better situate their role in practice and illustrate their interconnectedness.

In support of our first aim, we propose a holistic, three-part research data integrity model that integrates a definition of each core component, clarifies relationships among the core components, and emphasizes the fundamental role of documentation and training (see Figure 1 for the CSE Research Data Integrity Concept Model). With this concept model, we attempt to move from a static representation of research data integrity as discrete components to a more dynamic and holistic representation of how research data integrity constitutes the interplay of its core components and in which, ultimately, the whole is greater than the sum of the parts.

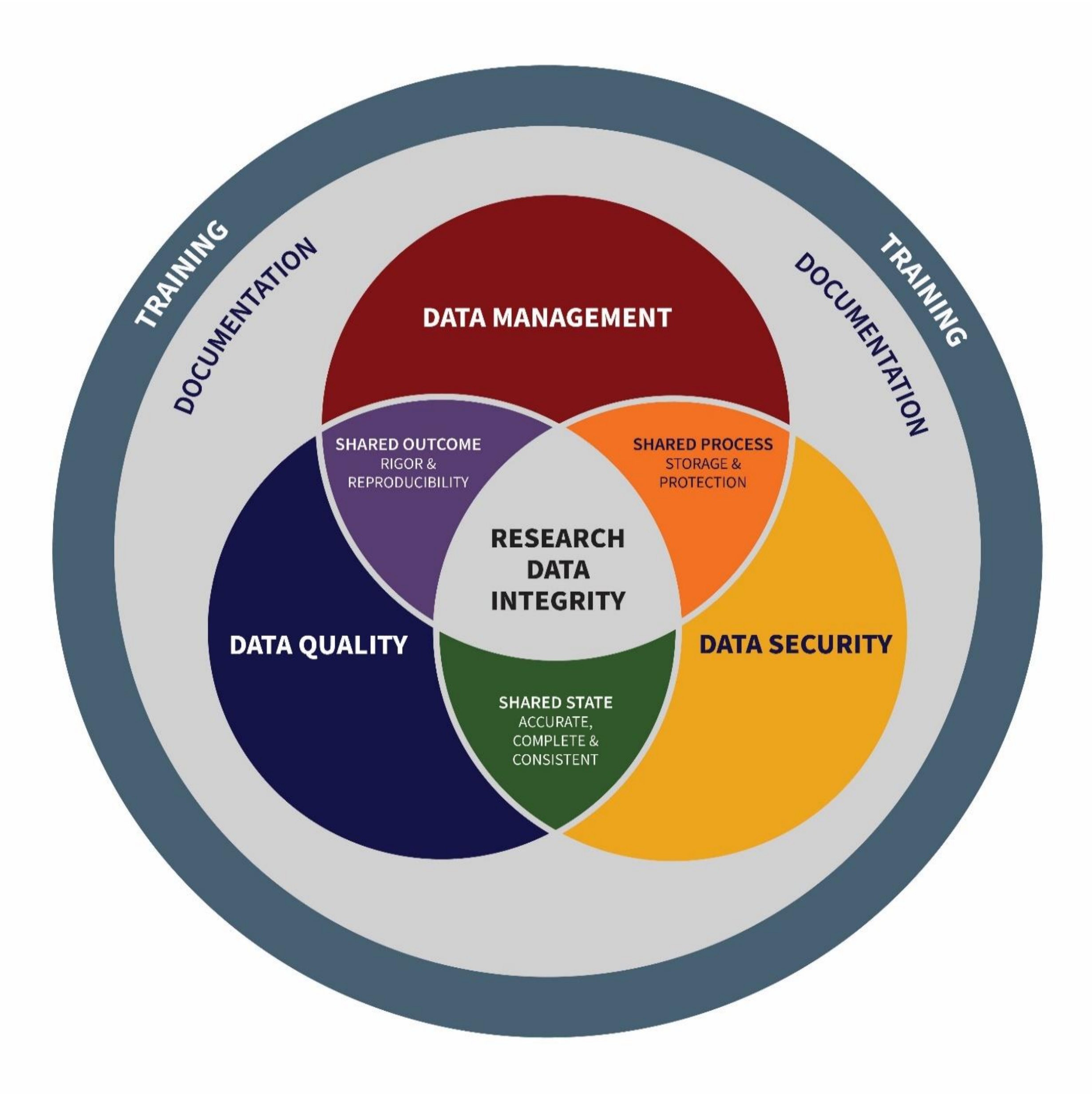


Figure 1: CSE Research Data Integrity Concept Model. A visual representation of the core components of research data integrity, the interplay between these components, and the fundamental role of documentation and training.

3.1. Definitions of core components

Addressing the lack of standardized terminology and definitions was the first challenge we tackled. To develop our concept model, we used the working definitions of research data integrity and the three key components presented in Table 2.

Table 2: Working Definitions of Research data integrity and its Components

Data Integrity	'State of data (valid or invalid) and/or the process of ensuring and preserving the validity and accuracy of data' (Ng, 2021).
Data Management	A set of foundational practices for organizing, documenting, storing, sharing, and preserving data.
Data Quality	'Assurance that data produced is exactly what was intended to be produced and fit for its intended purpose' (Medicines & Healthcare Products Regulatory Agency, 2018, p. 20).
Data Security	Physical security and technological protection of data for safeguarding data from corruption, unauthorized access, or loss.

Data integrity speaks to the trustworthiness of data throughout the research data lifecycle (see Figure 2 for an example of a Research Data Lifecycle). Data integrity as a state characterizes a dataset that is both valid and accurate (trustworthy) now and into the future (McDowall, 2018; Medicines & Healthcare Products Regulatory Agency, 2018; Ng, 2021). Data integrity as a process describes measures used to ensure validity and accuracy (trustworthiness) of a dataset now and into the future (IEEE, 1990; McDowall, 2018; Ng, 2021).

Data management is a set of foundational practices for organizing, documenting, storing, sharing, and preserving data. Data management is concerned with maintaining trustworthiness of data throughout the research data lifecycle and ensuring that data are reusable now and into the future. There are many resources that address best practices for responsible research data management (see Briney, Coates and Goben, 2020; Corti, Van den Eyden, Bishop and Woollard, 2019; Goodman, et al., 2014), as well as several which connect data management to research data integrity and the responsible conduct of research (see Coates, 2014; National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2009). In discussing the importance of reproducible research, Resnik and Shamoo (2017) note that '[r]esearchers need to be able to trust that published data are reliable, and reproducibility problems can undermine that trust' (p. 3), further reinforcing the centrality of responsible data management practices.

Data quality is 'the assurance that data produced is exactly what was intended to be produced and fit for its intended purpose' (Ng, 2021). Definitions of data quality and data integrity are especially difficult to tease apart, with data quality often expressed as an attribute of data integrity or vice versa (Sandhu, 1993; Koltay, 2016). Data quality speaks to the value of the data collected about the phenomenon under investigation and the utility of data collected with respect to the research question(s) or use for which the data were collected. The characteristics of data quality include accuracy, completeness, reliability, traceability, and relevance (National Academies of Sciences, Engineering, and Medicine, 2019a). We position data quality as a component of research data integrity because data quality primarily speaks to fit and utility whereas research data integrity speaks to a much broader context of the trustworthiness of data throughout the research lifecycle.

Data security refers to both the physical security and technological protection of data to safeguard them from corruption, unauthorized access, or loss. Data security emphasizes protection from unintentional changes to data and preventing threats to research data integrity. Data security measures include robust backup systems, controlled access, secure storage, encryption, protection during transmission/transfer among users or devices, and appropriate access controls when sharing (Corti, et al., 2020). Data governance is also an essential data security measure for protecting the integrity of research data and other data assets (Koltay, 2016). Data governance is defined as 'all policies, processes and principles related to the management, use, and security of data' (McCaig and Rezania, 2021, p. 5). In fact, McDowall's four-layer data integrity model designed for the pharmaceutical industry has data governance as its foundation layer (2019). Data governance strategies can help define policy frameworks for maintaining and implementing data security.

3.2 Documentation and training

Documentation and training are integral to the concept of research data integrity and highlighted in our model as supporting elements. Documentation and training serve as communication devices that link the other components, relaying information that is either unique to a component or common among components. Documentation conveys information about data and the research process, and training conveys implementation guidance to producers, providers, users, and educators. Such communication is central to research data integrity, and without documentation or training, each of the other components is weakened.

To achieve effective data management, quality, and security, the systems in place to collect, authenticate, validate, store, secure, backup, access, process, analyze, transmit, and preserve data need to be documented. Documentation and metadata equate to transparency in the research process and facilitates the four Rs. Standards for capturing data-level documentation should be part of the data management plan, built into each stage of the research data lifecycle, and aligned with the FAIR Principles (Partescano, 2021; Wilkinson, et al., 2016). Project-level documentation, including standard operating procedures and record-keeping best practices, is also needed (Murphy, 2019; Schreier, Wilson and Resnik, 2006). This allows any individual working on the project to not only know the correct procedure for any given activity (e.g., recording, validating, backing up, or transmitting data), but also to understand what has happened to the data from the start of the project until the end. Documentation should be kept current, reviewed periodically (on long term projects), easily accessible, broadly disseminated among project team members, and included as part of project and team training.

Research data integrity training (including good documentation practices) is essential for promoting its centrality to the four Rs and the success of a research project. Adequate and effective training of research personnel is critical to ensure everyone understands the project goals, their responsibilities as a researcher as well as part of a team, the governing policies and standard operating procedures for a project, and appropriate conduct. This training thus becomes the foundation for ensuring research and research data integrity. Training may be structured in a classroom environment or occur via mentoring of early-career researchers by seasoned researchers. Although both methods are beneficial, it is important to formalize training as a standard and documented practice. The CSE Research Data Integrity

Concept Model and the following schema were developed for use in training and with education on best practices as a primary outcome.

3.3 Relationships of core components

The above working definitions differentiate the core components from each other. From these definitions, we can further reveal the relationships between the components. In our CSE Research Data Integrity Concept Model (see Figure 1), we highlight how these different components are unique in their function and how they interact with one another as components of research data integrity. No single component can fulfill the requirements of creating or maintaining the integrity of data. While in academia, for example, we often emphasize the critical nature of data management, data management alone is not enough to ensure research data integrity, either as a state or as a process, as represented by our model. To have integrity, data quality and security must be addressed and supported by documentation and training.

It is the shared space between the components that we can define the dynamic and meaningful interactions that occur between the components, which taken together create a holistic approach to research data integrity in practice. One of the strengths of our concept model is that it highlights progress toward an integrated, cross-disciplinary concept of research data integrity that is formed by the interplay between these core components. To illustrate:

- Data management and data security share processes for providing protection and storage of data.
- Data security and data quality share a state of establishing and maintaining accurate, complete, and consistent data.
- Data quality and data management share outcomes of supporting rigorous and reproducible research.

In our proposed model, research data integrity is comprised of the three individual components of data management, data quality, and data security; documentation and training; and the interplay among the three components. This interplay yields a shared process, state, and outcome that creates a dynamic and cohesive concept of research data integrity that while complex, now allows researchers to plan for and realize research data integrity from a practical standpoint.

4. Implementing research data integrity throughout research lifecycle

Our second aim is to consider the utility of the core concept of research data integrity and its components for data producers, providers, users, and educators. To do this, we ask what research data integrity and its components look like in practice and how the CSE Research Data Integrity Concept Model (see Figure 1) may be used for training. We turned to research lifecycle models, visual tools for research activity planning, to answer this question. Here we focus on the Research Data Lifecycle and the Research Project Lifecycle (see Figure 2) to illustrate the embedded nature of research data integrity throughout research activities.

A Research Data Lifecycle provides a high-level overview of working with data, from the planning stage before data are generated to preserving data for future use beyond the life of the original project. Each stage of the lifecycle is implemented through actions and processes (e.g., adding metadata, cleaning data, implementing backup, and selecting a repository) by researchers as they work with data throughout the course of their research project. Research data integrity needs to be maintained throughout each stage and therefore the entire Research Data Lifecycle. The actions and processes taken by researchers outlined in the lifecycle help to ensure well-managed, secure, documented, and quality data.

While the Research Data Lifecycle provides a framework for the creation, use, preservation, and management of research data, the Research Project Lifecycle offers a broad characterization of research project activity and provides an outline for conducting high quality research. While these lifecycles are often portrayed as independent of one another, there is significant interplay between them. Ideally, these lifecycles inform each other in a continuous, dynamic process. The Research Data Lifecycle can be viewed as an aspect of the Research Project Lifecycle. To illustrate their interdependence, Figure 2 presents these two research lifecycles as concentric circles to highlight their interchange, with the Research Project Lifecycle as the outer ring and the Research Data Lifecycle as the inner ring.

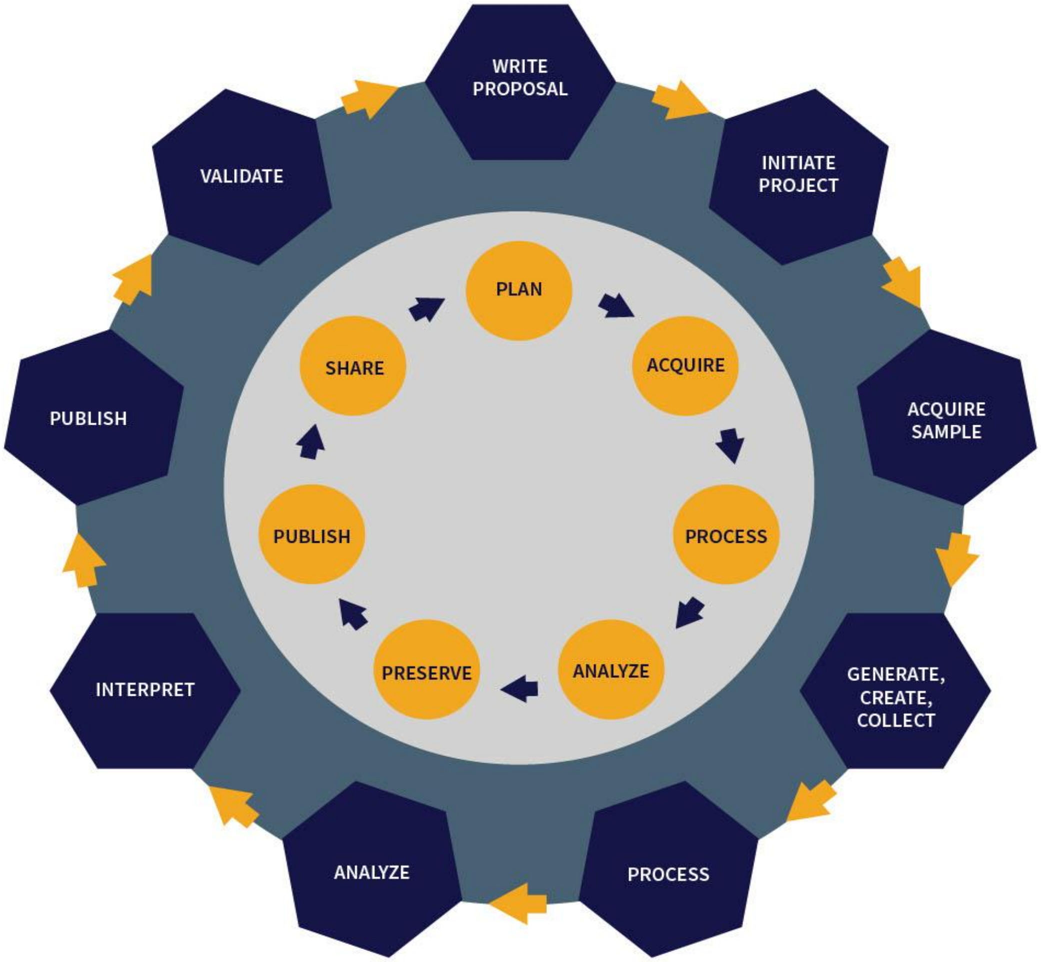


Figure 2: Research Lifecycle. An illustration of the interdependence between the Research Project Lifecycle (outer ring) and the Research Data Lifecycle (inner ring) (adapted from U.S. Geological Survey Science Data Lifecycle, Faundeen, et al., 2014).

4.1 Research data integrity in practice: Research data lifecycle

In practice, researchers may not consciously identify the actions they perform in ensuring research data integrity. Using the Research Data Lifecycle as a guide, we mapped each lifecycle stage to corresponding actions and processes taken by researchers and categorized those activities according to which core components related to research data integrity they address, including documentation and training. By doing this, we achieve the following:

- Identify at which stage in the Research Data Lifecycle the different components of research data integrity are addressed,
- Highlight what activities performed by researchers contribute to each component of research data integrity,
- Merge the holistic research data integrity model (the CSE Research Data Integrity Concept Model in Figure 1) with a practical model (of the Research Data Lifecycle), and
- Create a schema (see Figure 3) for research data integrity that is actionable and may be used in documentation and training.

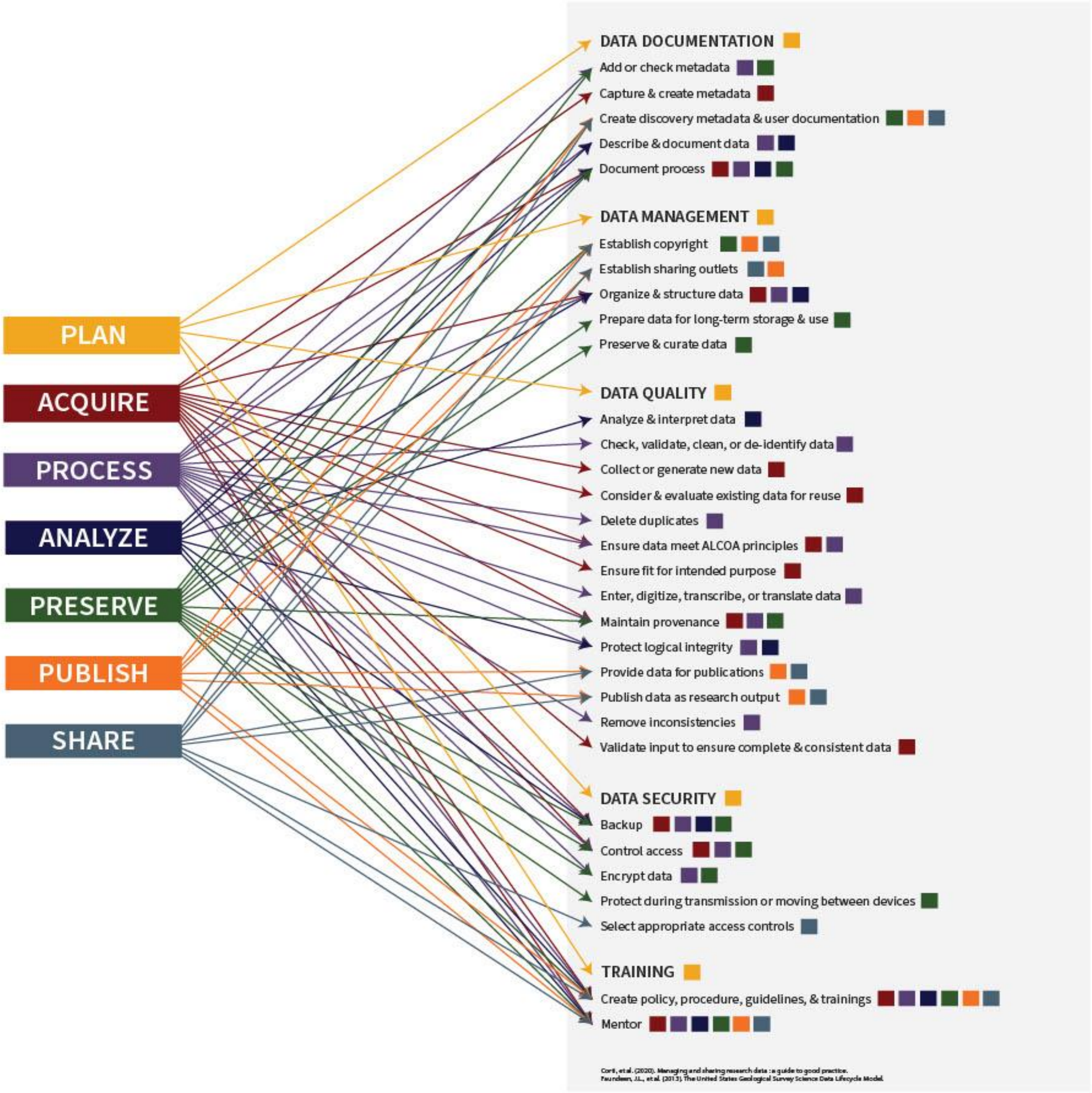
The Research Data Lifecycle that we present in Figure 2 (inner ring) and map to in Figure 3 has seven stages that are summarized in Table 3.

Table 3: Stages of the Research Data Lifecycle

Plan	Establishing and documenting guidelines for a research project, from data acquisition through preservation and sharing
Acquire	Acquiring, generating, creating, or collecting data
Process	Preparing data, new or previously collected, such as cleaning, validating, subsetting, or integrating
Analyze	Exploring and interpreting processed data, such as statistical analysis, visualization, or modeling
Preserve	Ensuring long-term reusability and accessibility of data
Publish	Publishing data; for example, by making data available with published articles
Share	Sharing datasets with other researchers for reuse, often via repository services

Each stage of the Research Data Lifecycle is associated with activities that correspond with multiple, and in most cases all, of the components related to research data integrity. For example, data quality is not achieved at a single point in the Research Data Lifecycle, but rather through multiple activities that take place throughout the lifecycle. The schema (see Figure 3 for the CSE Research Data Integrity

Implementation Schema) offers a mapping diagram that further emphasizes the intersections between the components of research data integrity as illustrated in the CSE Research Data Integrity Concept Model (see Figure 1). Researcher actions and the research data integrity components are not isolated activities, but rather they interact with one another to build a web for achieving research data integrity.


PLAN
ACQUIRE
PROCESS
ANALYZE
PRESERVE
PUBLISH
SHARE
DATA DOCUMENTATION
Add or check metadata
Capture & create metadata
Create discovery metadata & user documentation
Describe & document data
Document process
DATA MANAGEMENT
Establish copyright
Establish sharing outlets
Organize & structure data
Prepare data for long-term storage & use
Preserve & curate data
DATA QUALITY
Analyze & interpret data
Check, validate, clean, or de-identify data
Collect or generate new data
Consider & evaluate existing data for reuse
Delete duplicates
Ensure data meet ALCOA principles
Ensure fit for intended purpose
Enter, digitize, transcribe, or translate data
Maintain provenance
Protect logical integrity
Provide data for publications
Publish data as research output
Remove inconsistencies
Validate input to ensure complete & consistent data
DATA SECURITY
Backup
Control access
Encrypt data
Protect during transmission or moving between devices
Select appropriate access controls
TRAINING
Create policy, procedure, guidelines, & trainings
Mentor

Corti, et al. (2020). Managing and sharing research data : a guide to good practice.
Faundeen, J.L., et al. (2013). The United States Geological Survey Science Data Lifecycle Model.

Figure 3: CSE Research Data Integrity Implementation Schema. A schema for implementing research data integrity throughout the Research Data Lifecycle.

The CSE Research Data Integrity Implementation Schema (see Figure 3) provides a visualization for implementing research data integrity throughout the Research Data Lifecycle. It is significant to note that the Plan stage uniquely maps directly to each of the key components and not to specific actions. While Plan is its own stage in the lifecycle, it is essential for all other actions and processes that take place at each subsequent stage. Researchers should plan for the acquisition, processing, analysis, publication, sharing, and preservation of data. As a result, planning for data management, data quality, data security, as well as documentation and training occurs. Purposefully planning for research data integrity requires careful consideration of its components and their relationships to yield reliable, trustworthy, valid, and secure data throughout each stage of a Research Data Lifecycle. In addition, purposeful planning enables researchers to conduct rigorous research and generate outcomes that not only reflect the desired research data integrity characteristics, but that are also reproducible, replicable, and reusable.

What is also noticeable in the Plan stage, along with Acquire and Process stages, is that most of the activities associated with data quality fall into these three stages of the lifecycle. This is because researchers are actively working with the data (e.g., collecting/generating, cleaning, manipulating, and interpreting data). Consequently, there is ample opportunity within these activities to compromise how data are being handled, their fit for purpose or accuracy, or the rigor of the research being conducted. At the Preserve stage, however, the quality of the data has been established, and addressing integrity focuses on maintaining that quality through security and curation (or long-term data management) of the data to prevent unauthorized changes that would alter the data from their original state. This is emphasized by the role data security plays in the Preserve stage.

Making data available to others allows for the data's quality to be assessed and for others to try to reproduce or replicate the results when made available alongside published papers, protocols, or code. When data providers and producers make data publicly available, the data need to be in a form such that they can be used and evaluated, which requires them to be accompanied by sufficient documentation for their content and value to be ascertained by users (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2009). Thus, data sharing, and data management by extension, supports data quality (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2009). We elected to represent Publish and Share as separate stages in this model to highlight sharing as a mechanism for dissemination of data for transparency and reuse (potentially with restrictions), not only for data associated with a publication. Sharing data in a data repository also supports the long-term preservation of and access to data as responsibilities are transitioned from individual researchers to digital preservation professionals.

Mapping between the Research Data Lifecycle stages and activities associated with research data integrity core components illustrates the complexity of research data integrity in practice. There is not one action that researchers take to achieve research data integrity or, for that matter, one action to achieve quality, security, or the management of data. The web of interactions represented in the CSE Research Data Integrity Implementation Schema (see Figure 3) illustrates the complexity of research data integrity (and its components) as well as its significance when considering the role of data in rigorous and reproducible research, as well as in replication and data reuse.

4.2 Research data integrity: Research project lifecycle

Our CSE Research Data Integrity Implementation Schema (see Figure 3) illustrates research data integrity implementation in the context of the Research Data Lifecycle activities, which is inherently embedded in the Research Project Lifecycle as modeled in Figure 2. Thus, the schema suggests an indirect relationship between research data integrity and the Research Project Lifecycle, via interplay with the Research Data Lifecycle. Research data integrity, however, cannot be left to chance due to its central role in the four Rs. Therefore, the relationship between research data integrity and the Research Project Lifecycle needs to be prioritized. To accomplish this, we suggest three approaches for applying the model and schema in practice: education in the classroom, day-to-day training and support of researchers and scholars, and fostering institutional culture.

Research data integrity can be incorporated into classroom instruction through the exploration of theory as well as hands-on case studies, regardless of discipline. Our CSE Research Data Integrity Concept Model (see Figure 1) and CSE Research Data Integrity Implementation Schema (see Figure 3) can be used as educational and instructional tools for students to explore in context of their discipline; courses about research methods, research integrity, experimental design, or reproducibility/replication are a natural fit. To illustrate the potential, we share the following example, proposed by the student whose question initiated this probe into research data integrity. In this exercise, students use datasets (whether pre-existing or created as part of a class) and multiple checkpoints to evaluate the effects of bias, taxonomy, reference database selections, methodologies, and other elements of the Research Data Lifecycle on research results. The multi-step exercise enables students to develop a working understanding of research data integrity, its complexity, and the dynamic interplay of its components within the broader context of rigor, reproducibility, and replication (Sall, 2020). Incorporating such instruction and exercises in undergraduate and graduate coursework not only helps to address the four Rs but also aids in teaching data literacy to students, an increasingly important and competitive skill not only in contemporary science but also in daily life (Smalheiser, 2017; Teal, et al., 2015; Wolff, et al., 2016).

Outside the classroom, research data integrity training opportunities can be incorporated into existing channels that support students and researchers throughout the Research Project Lifecycle. In labs and research groups, this can be achieved through standardized training for new students or new employees. Research workshop series can be a natural forum for research data integrity training with the benefit of jointly engaging new and experienced researchers in research data integrity discussions. Other opportunities include lab/team/group meetings, journal clubs, online resources or guides, or topic-based events, such as International Love Data Week. Teaching both the complex concept of research data integrity and its components utilizing our CSE Research Data Integrity Concept Model (see Figure 1) and CSE Research Data Integrity Implementation Schema (see Figure 3) to illustrate implementation provides researchers, regardless of their discipline or experience level, a practical model and standardized vocabulary that can be deployed while they address challenges, utilize new analytical tools, or build cross-cutting collaborations.

In additional to education and training, cultivating an institutional culture that includes open dialogue about research data integrity is critical. Policies related to research data ownership, management, use, and sharing are an important tool to establish definitions, procedures, and guidelines that address research data integrity and its components. Institutions can promote research data integrity-related discussions, resource deployment, and policy implementation by hiring dedicated personnel, encouraging collaboration to leverage expertise, or re-envisioning roles of compliance officials, data managers, information technology, librarians, curriculum development specialists, or research development specialists. Institutional research communication channels are a valuable vehicle to inform research and scholarly communities about resources and best practices, share success stories, and encourage ongoing discussions around research data integrity. Collectively, investment in institutional research data integrity education, training, resource development, and personnel support provides a framework that strengthens a researcher's ability to plan and conduct their research activities based on the foundation of research data integrity.

5. Research data integrity and the Four Rs

The third and final aim of this paper is to demonstrate the role of research data integrity in the four Rs and facilitate researchers in achieving the gold standard of research: research that is rigorous, outcomes that are reproducible and replicable, and data that can be reused (the four Rs). Many of the recommendations in the National Academies report, *Reproducibility and Replicability in Science,* speak to efforts that address data integrity and include education and training (National Academies of Sciences, Engineering, and Medicine, 2019b). As illustrated in our CSE Research Data Integrity Concept Model (see Figure 1) and CSE Research Data Integrity Implementation Schema (see Figure 3), researchers need to purposefully plan during each stage of the Research Data and Project Lifecycles to achieve research data integrity. Using best practices to implement each of the research data integrity components throughout the research lifecycle not only can protect research data integrity but can also result in a research effort that is closer to that gold standard – achieving the four Rs.

Planning and managing for research data integrity can serve as a resource allocation tool. Stewardship – the relationship between research, researchers, and society – is a core value of research and being responsible stewards of resources (e.g., people, time, funding, training, etc.) is part of this (National Academies of Sciences, Engineering, and Medicine *et al.*, 2017). The four Rs are characteristics of stewardship and managing for research data integrity is one means for allocating resources to achieve them. Furthermore, stewardship of resources is a responsibility of all researchers and the research enterprise, and it is a key component of research integrity more broadly (National Academies of Sciences, Engineering, and Medicine *et al.*, 2017).

Training for research data integrity can serve as a research management tool. When collaborators – whether they are a data producer, provider, user, or an educator – are operating with the same philosophy and vocabulary about data, a shared understanding of each collaborator's responsibilities, activities, and outputs is more easily achieved. When the four Rs is the goal and research data integrity is viewed as foundational to that goal, researchers can conduct innovative, exemplary, and impactful work. Research data integrity can help an individual or team conduct research efficiently, produce high-quality research, and enable others to make use of research outputs with confidence and fidelity. Further, understanding research data integrity is particularly important in interdisciplinary research and in collaborative cross-sector research (i.e., academia, business, and industry) where different norms, procedures, and terminology regarding data exist.

When planned and intentionally implemented throughout the Research Data and Project Lifecycle, research data integrity can also serve as a risk management tool. The CSE Research Data Integrity Concept Model (see Figure 1) and CSE Research Data Integrity Implementation Schema (see Figure 3) provide a standardized vocabulary, a framework to define and document research activities, and a guide for the roles and interplay of data management, data quality, and data security. Whether implemented by a single researcher or across a team, a methodical and holistic approach to research data integrity can reduce the risk of errors related to both the state of data as well as the process of creating, using, and managing it – a requirement of the four Rs. It can further protect against the use of detrimental research practices, whether intentional or accidental, that erode the rigor of research and can contribute to irreproducibility (National Academies of Sciences, Engineering, and Medicine et al., 2017).

6. Conclusion

Research data integrity provides a strong foundation for high quality research outcomes. Addressing research data integrity with intentionality can significantly impact the way we plan, execute, and advance science. It provides a framework to foster reproducible and replicable science, support resource stewardship, preserve the public trust in science, and inform decision-making about policy. Our initial aim of this undertaking was to disentangle the complexity of data integrity in a way that could be useful for producers, providers, users, and educators. We position our model and schema within the larger dialog around research integrity and illuminate the role that research data integrity and its components play in the four Rs. Our proposed CSE Research Data Integrity Concept Model and CSE Research Data Integrity Implementation Schema enable the complexity of research data integrity to be taught effectively. In addition, they provide a framework for researchers to be intentional in their planning around research data integrity and research integrity more broadly. Demonstrating the central role of research data integrity in the four Rs is particularly pertinent given the growing concern about rigor and reproducibility issues in several disciplines in the past decade (e.g., Fanelli, 2018; Medaglia and Fernandez, 2022, Open Science Collaboration, 2015) and in research that has immediate bearing on human health, the environment, or policy development.

This paper provides a high-level overview of our proposed CSE Research Data Integrity Concept Model and CSE Research Data Integrity Implementation Schema and their value as education and training tools. They are intended for data producers, providers, users, and educators to conceptualize the relationships between research data integrity core components and for intentionally planning and implementing research data integrity during a research project. The model and schema transcend any specific discipline and can be applied to a wide range of contexts, sectors, and disciplines to facilitate the realization of the four Rs. For other types of data (i.e., organizational, administrative, or product development data), the schema can be adapted to map to actions and processes that are more aligned with activities that take place during the lifecycle of those data. In future work, applying the model and schema to discipline-specific case examples will further illustrate their widespread utility in practice and reinforce the role of research data integrity in research rigor, reproducibility, replication, and data reuse across all disciplines. Additional future work using the model and schema involves development of curriculum and training modules for research data integrity and assessing learning outcomes.

When we view the CSE Research Data integrity Concept Model, we notice that the shared spaces between the components consistently support the four Rs. Data management supports the outcome of rigorous and reproducible research, in addition to the process of data protection and storage. Data quality supports the outcome of rigorous and reproducible research, as well as the state of accurate, complete, and consistent data. Data security supports the state of accurate, complete, and consistent data, while also supporting the process of data protection and storage. With the added layers of documentation and training, we demonstrate that the interplay between data management, data quality, and data security comprise a crucial and viable framework to realize research data integrity that enables the achievement of robust science, which ultimately advances our knowledge of the world around us.

Acknowledgements

The authors want to thank the following individuals for their contributions to this manuscript: Izabella Sall for sparking the idea and for allowing them to build on the work that she did on this topic in the GRAD 930 course; Sandra Hickey, the University of New Hampshire Extension's Marketing and Communications Senior Producer for professionalizing the graphics; and Leanna Lostoski-Ho for her thorough review of our manuscript.

References

Ahmad, S., Kumar, A. and Hafeez, A. (2019) 'Importance of data integrity & its regulation in pharmaceutical industry', *The Pharma Innovation Journal*, 8(1), pp. 306–313. Available at: https://www.thepharmajournal.com/archives/2019/vol8issue1/PartF/8-1-44-870.pdf.

Arroyo-Araujo, M. and Kas, M.J.H. (2022) 'The perks of a quality system in academia', *Neuroscience Applied*, 1. Available at: https://doi.org/10.1016/j.nsa.2022.100001.

Bordewijk, E.M., Li, W., Gurrin, L.C., Thornton, J.G., van Wely, M. and Mol, B.W. (2021) 'An investigation of seven other publications by the first author of a retracted paper due to doubts about data integrity', *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 261, pp. 236–241. Available at: https://doi.org/10.1016/j.ejogrb.2021.04.018.

Briney, K., Coates, H. and Goben, A. (2020) 'Foundational Practices of Research Data Management', *Research Ideas and Outcomes*, 6, p. e56508. Available at: https://doi.org/10.3897/rio.6.e56508.

Brook, C. (2020) 'What is Data Integrity? Definition, Best Practices & More', *Digital Guardian*, DataInsider. Available at: https://digitalguardian.com/blog/what-data-integrity-data-protection-101 (Accessed: 17 March 2022).

Caratas, M.A., Spatariu, E.C. and Gheorghiu, G. (2019) 'Privacy and Cybersecurity Insights', *"Ovidius" University Annals*, Economic Sciences Series, XIX(2), pp. 242–246.

Carlisle, J.B. (2012) 'The analysis of 168 randomised controlled trials to test data integrity', *Anaesthesia*, 67(5), pp. 521–537. Available at: https://doi.org/10.1111/j.1365-2044.2012.07128.x.

Coates, H. (2014) 'Ensuring research integrity: The role of data management in current crises'. *College & Research Libraries News*, 75(11). Available at: https://doi.org/10.5860/crln.75.11.9224.

Corti, L., Van den Eyden, V., Bishop, L., and Woollard, M. (2020) Managing and Sharing Research Data: *A Guide to Good Practice*. 2nd edn. London: Sage Publications.

Fanelli, D. (2018) 'Is science really facing a reproducibility crisis, and do we need it to?', *Proceedings of the National Academy of Sciences*, 115(11), pp. 2628–2631. Available at: https://doi.org/10.1073/pnas.1708272114.

Faundeen, J.L., Burley, T.E., Carlino, J., Govoni, D.L., Henkel, H.S., Holl, S., Hutchison, V.B., Martín, E., Montgomery, E.T., Ladino, C.C. and Tessler, S. (2013) The United States geological survey science data lifecycle model. Reston, VA, USA: *US Department of the Interior*, US Geological Survey. Available at: https://pubs.usgs.gov/of/2013/1265/.

Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M. and Hogg, D.W. (2014) 'Ten Simple Rules for the Care and Feeding of Scientific Data', *PLoS Computational Biology*, 10(4), p. e1003542. Available at: https://doi.org/10.1371/journal.pcbi.1003542.

IEEE (1990). Standard Glossary of Software Engineering Terminology. IEEE Standard 610.12-1990. Available at: https://doi.org/10.1109/IEEESTD.1990.101064.

Koltay, T. (2016) 'Data governance, data literacy and the management of data quality', *IFLA Journal*, 42(4), pp. 303–312. Available at: https://doi.org/10.1177/0340035216672238.

McCaig, M. and Rezania, D. (2021) 'A Scoping Review on Data Governance', *Proceedings of the International Conference on IoT Based Control Networks & Intelligent Systems* - ICICNIS 2021. Available at: http://dx.doi.org/10.2139/ssrn.3882450.

McDowall, R.D. (2018) 'How to Use This Book and an Introduction to Data Integrity', in *Data Integrity and Data Governance: Practical Implementation in Regulated Laboratories*, pp. 1-27. Available at: https://doi.org/10.1039/9781788013277-00001

McDowall, R.D. (2019) 'Data Integrity Focus, Part 1: Understanding the Scope of Data Integrity', *LCGC North America*, 37(1), p. 44-51.

Medaglia, J. and Fernandez, K. (2022) 'The "'Crisis' Crisis" in psychology', *Behavioral and Brain Sciences*, 45(E28). Available at: https://doi.org/doi:10.1017/S0140525X21000364 (Accessed: 17 March 2022).

Medicines & Healthcare Products Regulatory Agency (MHRA) (2018) 'GXP' Data Integrity Guidance and Definitions. Available at: https://www.gov.uk/government/publications/guidance-on-gxp-data-integrity (Accessed: 17 March 2022).

Murphy, A.J. (2019) 'Maintaining an Effective Lab Notebook and Data Integrity', in Kennedy G., Gosain A., Kibbe M., LeMaire S. (eds) *Success in Academic Surgery: Basic Science*. Cham: Springer International Publishing, pp. 31–41. Available at: https://doi.org/10.1007/978-3-030-14644-3_4.

National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Committee on Responsible Science (2017)

Fostering Integrity in Research. Washington (DC): National Academies Press (US). Available at: http://www.ncbi.nlm.nih.gov/books/NBK475953/ (Accessed: 21 February 2022).

National Academies of Sciences, Engineering, and Medicine (2019a) *Assuring Data Quality at U.S. Geological Survey Laboratories*. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/25524.

National Academies of Sciences, Engineering, and Medicine (2019b) *Reproducibility and Replicability in Science*. Washington, D.C.: The National Academies Press. Available at: https://doi.org/10.17226/25303.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2009) Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/12615.

Network of the National Library of Medicine (no date) Data Reuse. Data Glossary. Available at https://nnlm.gov/guides/data-thesaurus/data-reuse (Accessed: 18 March 2022).

Ng, C. (2021) 'What is Data Integrity and How Can You Maintain it?', *Inside Out Security*. Available at: https://www.varonis.com/blog/data-integrity (Accessed: 17 March 2022).

Open Science Collaboration (2015) 'Estimating the reproducibility of psychological science', *Science*, 349(6251), p. aac4716. Available at: https://doi.org/10.1126/science.aac4716.

Pandey, A.K., Khan, A.I., Abushark, Y.B., Alam, M.M., Agrawal, A., Kumar, R. and Khan, R.A. (2020) 'Key Issues in Healthcare Data Integrity: Analysis and Recommendations', *IEEE Access*, 8, pp. 40612–40628. Available at: https://doi.org/10.1109/ACCESS.2020.2976687.

Partescano, E., Jack, M.E.M., Vinci, M., Cociancich, A., Altenburger, A., Giorgetti, A. and Galgani, F. (2021) 'Data quality and FAIR principles applied to marine litter data in Europe', *Marine Pollution Bulletin*, 173, p. 112965. Available at: https://doi.org/10.1016/j.marpolbul.2021.112965.

Piller, C. and Servick, K. (2020) Two elite medical journals retract coronavirus papers over data integrity questions, *Science*. Available at: https://doi.org/10.1126/science.abd1697 (Accessed: 17 March 2022).

Resnik, D.B. and Shamoo, A.E. (2011) 'The Singapore Statement on Research Integrity', *Accountability in Research*, 18(2), pp. 71–75. Available at: https://doi.org/10.1080/08989621.2011.557296.

Resnik, D.B. and Shamoo, A.E. (2017) 'Reproducibility and Research Integrity', *Accountability in research*, 24(2), pp. 116–123. Available at: https://doi.org/10.1080/08989621.2016.1257387.

Retraction Watch Database (2018). Available at: http://retractiondatabase.org/RetractionSearch.aspx (Accessed: 17 March 2022).

Sall, I. (2020). How to Bias Undergraduate Students to Critically Care about Data, Unpublished Manuscript, University of New Hampshire, Durham, NH

Sandhu, R.S. (1993) 'On Five Definitions of Data Integrity', in *Proceedings of the IFIP Workshop on Database Security*, pp 257–267.

Schreier, A.A., Wilson, K. and Resnik, D. (2006) 'Academic Research Record-Keeping: Best Practices for Individuals, Group Leaders, and Institutions', *Academic Medicine: Journal of the Association of American Medical Colleges*, 81(1), pp. 42–47. Available at: https://doi.org/10.1097/00001888-200601000-00010.

Shamoo, A.E. (2020) 'Validate the integrity of research data on COVID 19', *Accountability in Research*, 27(6), pp. 325–326. Available at: https://doi.org/10.1080/08989621.2020.1787838.

Smalheiser, N. (2017) *Data Literacy: How to Make Your Experiments Robust and Reproducible*. Cambridge, MA: Academic Press.

Talend (no date) What is Data Integrity and Why Is It Important? Available at: https://www.talend.com/resources/what-is-data-integrity/ (Accessed: 17 March 2022).

Teal, T.K., Cranston, K.A., Lapp, H., White, E., Wilson, G., Ram, K. and Pawlik, A. (2015) 'Data Carpentry: Workshops to Increase Data Literacy for Researchers', *International Journal of Digital Curation*, 10(1), pp. 135–143. Available at: https://doi.org/10.2218/ijdc.v10i1.351.

U.S. Department of Health and Human Services. (no date) 'Guidance: Rigor and Reproducibility in Grant Applications', *NIH Central Resources for Grants and Funding Information.* Available at: https://grants.nih.gov/policy/reproducibility/index.htm (Accessed: 02 March 2022).

U.S. Food and Drug Administration (2018) Data Integrity and Compliance with CGMP Guidance for Industry. Available at: https://www.fda.gov/downloads/drugs/guidances/ucm495891.pdf.

U.S. National Institutes of Health (2020) NOT-OD-21-013: Final NIH Policy for Data Management and Sharing. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html (Accessed: 17 March 2022).

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1), p. 160018. Available at: https://doi.org/10.1038/sdata.2016.18.

World Conference on Research Integrity (2010) Singapore Statement on Research Integrity. Available at: https://wcrif.org/guidance/singapore-statement (Accessed: 17 March 2022).

Wolff, A., Gooch, D., Montaner, J.J.C., Rashid, U. and Kortuem, G. (2016) 'Creating an Understanding of Data Literacy for a Data-driven Society', *The Journal of Community Informatics*, 12(3). Available at: https://doi.org/10.15353/joci.v12i3.3275.

Yang, P., Xiong, N. and Ren, J. (2020) 'Data Security and Privacy Protection for Cloud Storage: A Survey', *IEEE Access*, 8, pp. 131723–131740. Available at: https://doi.org/10.1109/ACCESS.2020.3009876. https://ieeexplore.ieee.org/document/9142202

van der Zee, T., Anaya, J. and Brown, N.J.L. (2017) 'Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab', *BMC Nutrition*, 3(1), p. 54. Available at: https://doi.org/10.1186/s40795-017-0167-x.

Endnotes

[1] Dr. Patricia B. Condon, Assistant Professor, Research Data Services Librarian, University of New Hampshire, Durham, NH, USA. Email: patricia.condon@unh.edu. ORCID: https://orcid.org/0000-0003-3242-6666.

[2] Dr. Julie F. Simpson, Director, Research Integrity Services and Affiliate Assistant Professor of College Teaching & of Education, University of New Hampshire, Durham, NH, USA. Email: julie.simpson@unh.edu. ORCID: https://orcid.org/0000-0003-1067-2823

[3] Dr. Maria E. Emanuel, Partnerships Manager, Research & Large Center Development and Affiliate Assistant Professor of College Teaching, University of New Hampshire, Durham, NH, USA. Email: maria.emanuel@unh.edu. ORCID: https://orcid.org/0000-0002-6469-8378