**Reflecting on Documentation: Impact Statements, Data Management Plans (DMP), and Data Dictionaries**

Cesar Hernandez

School of Computing and Information, University of Pittsburgh

CMPINF 2140: Responsible Data Science

Dr. Matt Burton and Dr. Nora Mattern

4/21/2025

**Reflecting on Documentation: Impact Statements, DMPs, and Data Dictionaries**

**The Importance of Documentation in Data Science**

In any field, proper documentation is the bridge between understanding one's data and helping others navigate the complexities of research they might not be familiar with (Javier, 2020; Prukalpa, 2021). In data science, documents like impact statements, data management plans (DMPs), and data dictionaries are more than just supplementary; they are essential tools that support responsible data science. These documents uphold the core principles of responsible data science — transparency, fairness, accountability, privacy, and nonmaleficence — by outlining how data is managed, clarifying who is involved and their responsibilities, making research more accessible and understandable, and helping to anticipate potential harm and how to prevent it. Good documentation is more than a helpful addition; it makes research ethical, reproducible, and valuable to others.

*Transparency*

In responsible data science, transparency means making the overall project and its finer details visible and understandable. Documents like impact statements and data management plans (DMPs) help achieve this by prompting reflection on ethical risks, privacy, and potential impacts, particularly for vulnerable groups (Henderson, 2021). DMPs also outline how data is collected, managed, and protected, while data dictionaries define variables, structures, and relationships within the dataset (Atlan, 2025; Michener, 2015). The data dictionary is essential because the transparency of the data itself also lends itself to capturing other principles: accountability, privacy, and fairness. These forms of documentation make a project's goals, methods, and data use more open and transparent, which reduces the risk of misuse and supports ethical, trustworthy research.

*Fairness*

The idea of fairness in data science is two-fold: ensuring equitable treatment within the study itself and in how participants or data are recruited. Impact statements encourage researchers to reflect on how their work may affect different groups, particularly those who are marginalized or underrepresented (Henderson, 2021). Identifying potential sources of bias or unequal outcomes early in the developmental process helps guide the creation of fairer models and an overall more inclusive data practice. DMPs support fairness by outlining inclusive collection methods, increasing transparency around existing workflows, and detailing how data will be shared without excluding others. Data dictionaries also provide clear, consistent definitions that reduce the risk of misinterpretation and help maintain high data quality (Atlan, 2025). Transparency across these documents strengthens fairness, as clearly defined processes allow for ongoing evaluation and improvement in data collection, use, and sharing—making them more equitable for everyone involved.

*Accountability*

In data science, accountability means taking ownership of a project's outcomes—like whether things are going well —and being clear about who is responsible for what. The impact statement helps support this by encouraging researchers to think through potential risks early and to show they are serious about addressing issues and reducing harm (Henderson, 2021). A DMP also plays a significant role in laying out how data will be collected and clearly defining each researcher's responsibilities (Michener, 2015). Together, these tools help create a clear structure for accountability by identifying concerns early, assigning roles, and ensuring the project stays transparent from start to finish.

*Privacy*

The importance of privacy is not only to sustain the integrity of one's data but also to protect the data privacy of the individuals. The data science documentation ensures that the data is managed, stored, and shared in ways that will not breach trust. An impact statement helps address these privacy concerns by evaluating how data collection, use, and sharing could affect an individual's privacy (Atlan, 2025). A well-crafted data management plan (DMP) addresses privacy through details on how long the data will be accessible, how it will be stored and protected during the project, and how it will be preserved for future use (Michener, 2015). Overall, privacy is a central consideration in data science documentation, with a concern for protecting an individual's rights, and by clearly documenting these measures, we can mitigate future privacy risks.

*Nonmaleficence*

Data science documentation supports nonmaleficence by encouraging researchers to think critically about potential risks and benefits from the very beginning (Gupta, Lanteigne, & Heath, 2020). Documents like impact statements and data management plans (DMPs) help researchers consider how individuals or groups might be affected at any project stage. DMPs also outline important details like how data is collected, informed consent is handled, and how ethical use is ensured throughout (Henderson, 2021; Michener, 2015). Data dictionaries contribute by clearly defining variables and values, which helps prevent misinterpretation that could lead to harmful or misleading conclusions (Atlan, 2025).

In the end, documentation plays a vital role in upholding the principles of responsible data science. Strong documentation is essential to a project's success, whether it is outlining the process from start to finish, assigning roles and responsibilities, or addressing ethical concerns at every stage—from data collection to sharing and beyond the project's lifecycle. I believe solid documentation can be the difference between research that fades and research that changes lives.

**Personal Integration of Documentation Types**

Of the list of data science documentation provided, the three that are the most valuable and essential in my work are 1) the impact statement, 2) the data management plan (DMP), and 3) data dictionaries. These three documents are important because they capture all five responsible data science principles and combine the reflection of future concerns and the management of current processes.

*Impact Statements*

Impact statements are essential because they prioritize consequences as their central focus. I was initially not taught to think critically about the impact of my work. While I understood that considering social impact was important, I rarely let those concerns influence my process. I'm currently working on a sleep study involving patients with neurodegenerative diseases. I plan to actively consider how the interpretation of results could negatively affect certain groups and whether there are broader implications.

Because I work with several groups to help get their projects off the ground, I will require the researchers I collaborate with to draft an impact statement. I will create one myself, and we will meet to review concerns or brainstorm ways to mitigate potential risks or adverse outcomes. The impact statement will be beneficial because it demonstrates that researchers are committed to driving meaningful change, leaving a positive legacy, and helping others. Those values matter—because research should be about making an impact, not just citations or presentations.

While the impact statement is highly beneficial for a project, utilizing one can be challenging concerning content. It can be difficult to move forward when we identify problems for which we don't yet have solutions. Conceptually, creating an impact statement also has challenges, especially when defining "impact." This definition varies across disciplines, and an agreed meaning would require careful research, discussion, and collaboration. In comparison, measuring impact can also be tough—especially when the short- and long-term effects aren't immediately obvious or quantifiable. And finally, there's the question of causality: if change does happen, how do we know it was because of the research and not something else? Although there are challenges, the benefits prove that a document like this is necessary for the success of any project.

*Data Management Plan (DMP)*

The data management plan (DMP) outlines how data will be handled during and after a project. It typically includes methods for collecting and using data, standards for documentation and metadata, storage needs, ethical and privacy considerations, and plans for long-term access and preservation. In my current work, DMPs are not required but are often recommended. I see them as essential. They push researchers to think through the entire scope of their project from the beginning—something that's especially important when trying to define clear goals and long-term impact. DMPs are also helpful for others outside the original team, especially those who might want to reuse the data or learn from the methods used.

At my institution, most people recognize the value of DMPs, but they're often seen as optional or extra. Because of that, completion is inconsistent and usually left to the discretion of individual teams. When collaborating with researchers, I will also require a draft DMP before we begin working together. Creating a data management plan early in the data science lifecycle allows everyone to visualize workflows, clarify roles, and align expectations. So I will plan to meet with teams to review sections, provide support, and help improve the plan, particularly when responsibly managing, using, and preserving data. The goal is to ensure the plan is clear, complete, and helpful throughout the project.

There are many benefits to having a strong DMP. It helps organize workflow, ensures ethical and transparent data practices, and is a central reference throughout the data lifecycle. A good DMP also reinforces key responsible data science (RDS) principles: it supports transparency through clear documentation, promotes fairness in data collection, outlines privacy protections, defines roles for accountability, and encourages early consideration of risks to support nonmaleficence. In this way, the DMP is not just a plan — it's a tool for ethical, thoughtful research.

Despite these strengths, there are many challenges. DMPs are sometimes created to meet a requirement, with little attention to quality or completeness. Because there's no universal standard, the content and structure of DMPs can vary widely. Some are missing critical details, like how data will be preserved after the project ends. Creating a strong DMP can also be time-consuming and may require a level of data management experience that not all researchers have. These challenges and the perception of DMPs as "extra work" are often why they're left incomplete or underused. To address this, I aim to help collaborators share the workload, fill in gaps, and build DMPs that are practical and impactful. Creating the plan before the project begins helps everyone stay focused, set goals, and move forward more confidently.

*Data Dictionary*

Data dictionaries are essential because they provide comprehensive documentation detailing key elements such as data labels, types, and codes. They are particularly valuable for identifying sensitive fields that require additional privacy measures. I prioritize creating and maintaining data dictionaries for all new datasets, ensuring that the labels are precise, coded values are accurate and up-to-date, and the dictionary is stored in an easily accessible location for future reference.

I plan to implement data dictionaries in all future projects, ensuring that they are created at the start of each dataset's development and updated throughout the project lifecycle. This will ensure that the data remains well-documented, accessible, and easy to understand for both current and future users of the dataset.

The primary benefit of creating a data dictionary is that it ensures a clear and consistent understanding of the data, which helps users interpret the information correctly and make informed decisions. Additionally, it supports transparency and reduces the risk of misinterpretation, especially when datasets are shared across teams or organizations. However, the main challenge lies in the time and effort required to create and maintain data dictionaries, especially when working with evolving large datasets. While many of my data dictionaries are updates of previous versions, ensuring they remain accurate and comprehensive requires ongoing attention and consistency, which can be time-consuming.

**Retrospective Analysis**

A few years ago, I was hired at UCSF as a data analyst and manager to work on a mammography registry. My primary responsibility was building datasets for external clients to use in their models. About three months in, one of those clients returned and asked if we could recreate and update a dataset that had been created the year before. I assumed this would be straightforward—surely, my predecessor had documented everything thoroughly. They hadn't.

There was no documentation besides a general data dictionary describing the registry's variables. I quickly realized that my predecessor hadn't prioritized documentation or knowledge transfer. When they left, they took all of that understanding with them. I had to reverse-engineer everything from scratch with inadequate documentation, missing files, and scripts. After eventually rebuilding the dataset, I realized how much time and effort could've been saved if three key pieces of documentation had existed: a data management plan (DMP), a comprehensive data dictionary that included newly created variables, and README files. These are not just supplementary tools but essential for enabling new team members to step in, understand, and build upon the work that came before them.

A DMP would have shown me the entire workflow—how data was managed, why certain storage decisions were made, and who else had been involved. I spent so much time just trying to piece together what had happened, and while my supervisors helped when they could, they weren't involved in the details. I later discovered that other programmers in different departments had contributed to the project—something I could've known earlier had a proper DMP existed. That document alone would have supported key RDS principles like transparency, privacy, and accountability.

There were also no data dictionaries for the newly created tables or derived variables. The lack of a data dictionary made it incredibly difficult to figure out how specific fields had been calculated or what certain values meant. I had to go back to the client and ask them to walk me through the output they'd received. Eventually, I recreated the same tables and used the comparison to reverse-engineer the process. If variable definitions and derivations had been adequately documented, it would have saved time, prevented confusion, and reduced the risk of misinterpretation. This kind of oversight slowed things down and increased the likelihood of error—impacting fairness, transparency, and even nonmaleficence.

Finally, I discovered that many of the scripts I needed were scattered across various folders—some of which I didn't even know existed until much later. README files would have been a lifesaver. They provide structure, improve access and usability, and help prevent mistakes. A solid README would have shown me where things lived, who had worked on them, and how the pieces fit together. This kind of documentation is powerful—it supports transparency and accountability and fosters collaboration. And just like data dictionaries, README files help prevent harm by reducing the risk of misuse.

**Critical Thinking**

*Challenges*

The potential challenges and limitations in creating extensive documentation for data science projects can be broken down into three categories: resource constraint, insufficient knowledge, and culture. Due to resource constraints, many studies and researchers lack the time and effort to create these extensive documents. Additionally, because many data science projects are incredibly complex or fast-changing, producing documentation that thoroughly explains the processes can take a long time and result in inadequately updated materials. Data documentation uses much technical jargon that can be daunting and lead to incomplete, incorrect, or messy documentation materials. Lastly, data documentation is not required for most people, so shifting this perception requires changing the culture around documentation (Prukalpa, 2021). Many people would rather focus less on documentation until they need to, which is why many projects complete documentation materials after or well into a project. This culture focuses on ad hoc documentation because it is not a priority, and there is no consistency in which documentation should be used.

*Strategies to Balance Through Documentation*

Proper training, realistic goals, and workflow integration are the strategies to balance thorough documentation. One of the key reasons proper documentation is hard to implement is that not many people are thoroughly trained on the best practices. If there is a shift to treating documentation as a necessity and a requirement, with an effort to train people, then data documentation can be thorough. This helps to shift the culture a little, allocate time for documenting, and give people information regarding the technical jargon, which leads to more complete materials (Javier, 2020). Additionally, if training can teach about the technical requirements of data documentation, setting realistic standards and goals can also be taught. By defining which documents need to be completed and teaching how those specific documents are to be completed, we reduce the burden of deciding which is important. Moreover, finally, by combining training and realistic goals, we can integrate data documentation in our project workflows, including DMPs that are updated regularly to include a timeline of changes of both the scripts and overall project changes.

**Conclusion**

This experience has significantly shifted how I approach data documentation. I now push for clear and thorough documentation because I've seen how much time and confusion can result from its absence. When documentation is done well, it's much easier to step into a project, understand the work, and keep things moving without disrupting the workflow. Excellent documentation also plays a critical role in making the research reproducible and reusable by others, leading to more meaningful experiences. While it can be time-consuming and slightly meticulous, documentation is a necessary investment supporting more efficient, ethical, and collaborative research.

**References**

Atlan. (2025, April 20). Understanding the purpose of a data dictionary: 7 essential points for 2025. Atlan. <https://atlan.com/purpose-of-a-data-dictionary/?utm_source=chatgpt.com>

Fleming, G., & Bruce, P. (2021). Responsible data science. John Wiley & Sons.

Gupta, A., Lanteigne, C., & Heath, V. (2020). AI Ethics: Publication norms for responsible AI. Montreal AI Ethics Institute. <https://export.arxiv.org/pdf/2009.07262>

Henderson, C. (2021, November 5). A guide to writing the NeurIPS impact statement. Medium. <https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832>

Javier, P. (2020). The importance of documentation in data science. Medium. <https://princejavier.medium.com/the-importance-of-documentation-in-data-science-f9a39fc95d4b>

Michener, W. K. (2015). Ten simple rules for creating a good data management plan. PLOS Computational Biology, 11(10), e1004525. <https://doi.org/10.1371/journal.pcbi.1004525>

Prukalpa. (2021). Data documentation woes? Here's a framework. Towards Data Science. <https://towardsdatascience.com/data-documentation-woes-heres-a-framework-6aba8f20626c>