Data Management Plan – Part II

Loic Niragire

TIM-8130 v4: Data Curation

Dr. Ulrich Vouama

April 2, 2022

I propose a data management plan that adds provisions for data usage, transformation, and retention of data in our research team to support data analysis activities. Particularly, these activities range from descriptive analysis, predictive modeling, diagnostic, and prescriptive sessions on curated data. The group of researchers that our team supports work in the pulmonology and respiratory medicine department where they study environmental factors and their impact on pulmonary diseases.

Establishing an enterprise data governance is an endeavor that requires continuous commitment and effort from top leaders in an organization. The primary goal of our data governance is to ensure that data is managed as a corporate asset through all phases of its lifecycle, and not as a by-product of processes and applications. Also, I highlight the importance of metadata management in my plan.

To simplify my data management plan, I divided our organization into several business units based on data needs and responsibilities. Thereby adopting a replicated data governance model for the organization. Having the same governance model and standards adopted by each business unit reduces the scope of our policies and stewardship (Weber, Otto, & Österle, 2009) . Governance drivers within our business unit, the research team, will focus on improving data quality and metadata management processes. Improvements in these two areas will help the organization gain more confidence in generated results from our research team. Table 1 below highlights some of the key decisions to be made within our business unit.

I opted to restrict our governance model to two decision domains, namely data quality and metadata management, to reflect our top business needs – simplicity and collaborative. To support a team of scientific researchers conducting data analysis means that our data needs to be accurate, timely, complete, and credible. Additionally, to ease the onboarding of new

scientists on the team or facilitate a better collaborative environment, our data needs to be interpretable.

To assert a sustainable data governance implementation with financial impact, the data analytics business unit will be assigned a dedicated business data steward personnel. This individual will be responsible for representing the interest of all stakeholders by adopting an enterprise perspective. Moreover, this individual is expected to be a subject matter expert reporting directly to the Chief Data Steward. This structure will localize decision making within the organization thus increasing engagement and ownership. Data stewards will also be responsible for outweighing economic impact associated with data acquisition and usage. Meanwhile, the Chief Data Steward is responsible for setting the boundary requirements for the intended uses of data (Khatri & Brown, 2010).

Table 1Framework for data decision domains

	DOMAIN DECISIONS
DATA QUALITY	• What are the standards for data quality with respect to accuracy, timely, completeness and credibility? • What is the program for establishing and communicating data quality? • How will data quality as well as the associated program be evaluated?
METADATA	• What is the program for documenting the semantics of data?

- How will data be consistently defined and modeled so that it is interpretable?
- What is the plan to keep different types of metadata up to date?

Data Transformations

Data transformations are essential for extracting business insights from raw data. Additionally, there is a need to support a variety of data formats and sources in this process. In other words, delivering business insights not only relies on our primary dataset, but also external data sources and internally produced data from our processes. Transformation methods outlined in this section are based on the observation that our ability to remain resourceful in our community directly depends on our data volume, the velocity at which we can process that it, and the variety of data sources considered in our processes.

The process to ingest new dataset into our data lake will follow an Extract-Load-Transform data pipeline model to enable faster loading times while preserving original data format. Although this model requires a powerful data processing engine capable of carrying out requested transformations on demand, it simplifies ingestion of new data sources at a faster rate (Avery & Cheek, 2015). Additionally, preserving raw data format allows our researchers the flexibility to apply ad-hoc or new transformations, which may have been previous unknown.

Policies guiding data ingestion process will be put in place to maintain data integrity, starting with a formal review process requiring approvals from key domain experts for all new

datasets. This process will ensure that all datasets meet our quality threshold and that they are from trusted sources. Additionally, there needs to be an evaluation phase to assess how each new dataset fits into our existing system as well its potential usage. Metadata collected throughout this evaluation process will need to be stored along each new dataset in a dedicated metadata repository. Moreover, visual dashboards need to be updated to reflect the newly added dataset.

Data lifecyle management

This section discusses policies and procedures guiding data retention and archival processes. It is essential that all involved tasks are well documented and require a manager-approved service request. All archives will be stored on a secondary database server through automated database jobs that periodically replicate less frequently accessed data. This will be achieved by first partitioning existing database into archival blocks. Moreover, restoration tests will be performed on a regular basis to ensure that archived data can reliably be restored.

Regulatory compliance will guide purging of archived data. Meanwhile, datasets involved our research publications will need to be permanently made available through public data repository for research reproducibility purpose. Given the slow rate of growth of our dataset, archival storage space is not a current concern prompting an aggressive purging requirement.

References

Avery, A. A., & Cheek, K. (2015). Analytics Governance: Towards a Definition and Framework. *Twenty-first Americas Conference on Information Systems* . Puerto Rico: AIS Electronic Library.

Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM, Volume 53, Issue 1*, 148-152.

Weber, K., Otto, B., & Österle, H. (2009). One Size Does Not Fit All -- A Contingency Approach to Data Governance. *Journal of Data and Information Quality, Volume 1*, 1-27.