Research Data Strategy: framework and motivating factors

Boyan Angelov¹

¹Association for Computing Machinery (ACM), 1601 Broadway New York, USA. Email: boyanangelov@acm.org

Abstract

The need for large amounts of data permeates almost all fields of research. New technologies related to machine learning (deep learning in particular), cloud computing, the Internet of Things (IoT) add to the increased complexity of data-related work and how researchers deal with that. Those trends are accelerating and result in a widespread need for new frameworks to increase the cost-benefit ratio of research work while driving innovation. Modern developments in the private sector, both in technologies and ways of working, can be readily adopted by research organizations. "Data strategy" is an umbrella term for those advancements. This article presents its research variant, Research Data Strategy (RDS), complete with its different elements, the sequence of execution, and supporting activities. This new methodology aims to build on top of existing research data management practices by providing a framework for ensuring innovative science in data-driven research organizations.

Keywords: data strategy, data management, data architecture, innovation, change management

1 Introduction

It has been more than a decade that science has entered its era of big data. Fields such as astronomy, particle physics, bioinformatics have been at the forefront of data-intensive science (Hey, Tansley, Tolle, *et al.* 2009). New advances, such as the "resolution revolution" created new in-roads in the same direction (Kühlbrandt 2014). The next-generation sequencing (NGS) data alone will exceed those from YouTube and Twitter in size (NIH *et al.* 2018). Still, while this field has been ripe with innovation, especially in big data processing, artificial intelligence, and cloud computing, traditional research organizations lag behind their industry counterparts (Larson 2013). Companies such as Google, Microsoft, and Amazon have been developing technologies and algorithms at a breakneck pace. This disparity is not only limited to technology but also ways of working and processes. In recent years a new focus has emerged to help drive the adoption of data-intensive technologies in modern companies. This focus has centered around the new term "data strategy" (DalleMule and Davenport 2017; Kruhse-Lehtonen and Hofmann 2020). It has built on top of the traditional business strategy practice, albeit with quite a few notable differences. Data strategy development is becoming more common in the industry, often conducted by external management consulting firms but is a new concept for research and academia, where adoption is still low.

There have been several notable efforts in implementing a data strategy in a research context - most notably the National Aeronautics and Space Administration (NASA) and the National Institutes of Health (NIH). The former is known as the "NASA Data Strategy Whitepaper", and the latter as the "NIH Strategic Plan for Data Science" (NIH *et al.* 2018). This paper aims to expand on the idea of a business data strategy by establishing a template for its research variant - Research Data Strategy (RDS).

Figure 1. Comparison between data strategy and data management.

2 Discussion

Dealing with large datasets and associated technologies is not a new problem. The research domain has been mostly addressed by the field of Research Data Management (RDM) (Ray 2013). This concept differs in scope from data strategy, as shown on Fig. 1. Data management is defined as ensuring that the organization’s data, associated resources, and processes are utilized to their potential. On the other hand, a data strategy goes further by planning for an innovative future. Data strategy makes sure that data is managed as an asset (Laney 2017), which is then used by the advanced use cases that modern research demands. We can define an RDS as **a framework to deliver research value by applying data and analytics**.

There have been numerous efforts to create a template for a data strategy in the business world (Fleckenstein, Fellows, and Ferrante 2018; Van Rijmenam 2014). Still, consistency is limited when one goes beyond the fundamental components. On top of this challenge, we have to define it in a research context, which can be quite different from business¹.

While there is little agreement on what exactly a data strategy is, and it can depend heavily on the application domain and an organization’s specific situation, there are several fundamental elements available. Those are shown in the overview of an RDS in Fig. 2. There are three main phases of an RDS: 1) Due diligence, 2) Design, and 3) Communication and delivery. While those are almost always executed in sequence, the elements can be rearranged to match the specific research circumstance. For example, research organizations would often have data management practices and infrastructure already established so that RDS design efforts might be better spent on other sections, such as the operating model or solution architecture and technology. One layer of abstraction lower, we can see that some various subtasks and activities support the main elements. The ones presented in Fig. 2. are by no means exhaustive and offer further flexibility to adjust to the research application.

2.1 Preparation

To create an RDS, the participating organization needs to fulfill several criteria: a) alignment with research strategy; b) stakeholder buy-in; c) dedicated strategy ownership. The first of those is often the most time and resource consuming to achieve. An RDS needs to be aligned with the organization’s research goals and not just serve as a support function. It needs to have a core role in its achievement. This in turn depends on the buy-in of the organization’s decision-makers since the

1. For example, Porter’s Five Forces are not suitable for a research organization, since they do not face the same pressures from the market as a business would.

Figure 2. Overview of a Research Data Strategy.

design of an RDS, and its eventual implementation, depends on strong and committed leadership. Furthermore, finally, the completion of the task needs to have dedicated ownership defined. More ambitious organizations often create new roles, such as a Chief Data Officer to support this.

2.2 Current state analysis

The initial phase of the creation of a data strategy is the Current State Analysis (CSA). This phase allows an organization to get its bearings and see where they are in terms of objectives, data assets, processes, and existing workforce. This phase is accomplished mostly by a series of interviews with key members, where more clarity is achieved by answering questions from the two categories below:

Digital maturity assessment

- What are the current initiatives related to data collection, storage, and usage?
- How siloed are those initiatives?
- Who is responsible for the data generating process?

- Who is responsible for manipulating the data?
- Who is consuming the data?
- Who is concerned with the legal and privacy requirements regarding the data?
- What are the skill levels of the people working with data (both producers or consumers)?
- What is the level of technologies used?
- Are they proprietary or open source?

Data due diligence

- What are the current data generating processes?
- Is there a data inventory available?
- Is there good data architecture?
- What are the formats of the current data available?
- What is the data quality?

The documented answers to this questionnaire are the deliverable of the CSA.

2.3 Gap analysis

The gap analysis phase of a data strategy project has a hard requirement to complete the CSA. Taking that information into account, an organization can spend time determining its Ambition Level (AL) and starting to measure the gap to get there.

The AL setting requires knowing what is possible in terms of a data strategy. This process can be completed using the "futures thinking" methodology (Yapp 2005), supported by a research field state of the art (SoA) analysis. To achieve both of those efforts, the RDS owners need to form a steering group together with the research organization's functional leaders. This steering group can then use the CSA deliverable and the field analysis to set the AL. The steering group's decisions will differ from organization to organization due to digital maturity and resource constraints. The difference between the organization's current state and the AL corresponds to the "length" of the eventual roadmap (the final deliverable of the RDS). A gap analysis's deliverables include a hierarchical list of recommendations per strategic objective based on the current state and the AL.

2.4 Data management and governance

Research Data Management (RDM) is a mature research method (Ray 2013). Successful implementation of data management practices is fundamental to an RDS since it enables advanced analytics use cases and other downstream applications. Current datasets are often characterized by the three V's: volume, variety, and veracity (Miloslavskaya and Tolstoy 2016). Data management's primary goal is to improve the data available in the organization to the FAIR standard (Stall *et al.* 2019).

Research organizations are facing unique challenges in terms of data management. Valuable experimental data is often stored on individual contributors (IC) machines or group servers, with little documentation (consisting of a data dictionary, for example) available. The different stages of the data processing (often results of *ad hoc* analysis) are indiscriminately stored. This fragmented data lifecycle is further exacerbated by a frequent lack of backups and clear ownership. Fig. 3 illustrates those issues.

An RDM depends on the completion of the following checklist:

- Unified protocols for the different data steps?
- Correct permission tiers for different users
- Privacy and ethics fulfilled
- Unified systems (removal of siloes)

Figure 3. Centralised data management in a research organization.

- Overview of all assets (i.e., by creating a data lake and catalog)
- Ownership (data stewardship)
- Data provenance (data lineage tracing)
- Adherence to a documented process (such as CRISP-DM)

Benefits of a complete RDM offer increased cross-department collaboration, higher security, lower valuable researcher time spent on mundane and repetitive tasks that can be automated, and lower activation energy for commencing potential citizen science projects.

2.5 Solution architecture and technology

Solution Architecture and Technology (SAT) is often regarded as a subset of data management, but we consider it a separate element due to its importance to data strategy. SAT consists of programming languages, open-source and proprietary software (packages and frameworks), and cloud technologies. RDS addresses this topic by conducting a set of methods to ensure that the SAT forms a coherent, modular, and extensible foundation for downstream advanced analytics use cases. It is common knowledge that data scientists spend north of 80% of their time on tasks that can be automated and have no immediate impact on the results - data cleaning (Lenzerini 2018). An RDS that provides insights on the SAT can address this issue as well.

An example data infrastructure architecture for an advanced analytics use case is provided in Fig. 4. This challenge's complexity is apparent due to the large number of technologies addressing the complete data lifecycle.

During the CSA, the existing SAT of the organization is determined. More often than not, this will affect the recommendations provided by the RDS. A second limiting factor in this area is the resource availability in workforce capacity and skills. The RDS designers' task is to work together with the technical team members to define requirements and prioritize solutions. This can be done with several tools borrowed from traditional business strategies such as the technology radar (Rohrbeck, Heuer, and Arnold 2006), bullseye scale, and impact-effort matrices.

One aspect of RDS that makes it harder to implement than a business data strategy is the low

Figure 4. An example data infrastructure and architecture (modified after [Emerging Architectures for Modern Data Infrastructure](#)).

adoption level of cloud computing in the scientific domain. Many parts of data architecture have been abstracted on a higher level by advances in cloud computing. The popular cloud computing providers, such as Amazon Web Services (AWS), Microsoft Azure Cloud, Google Cloud, and IBM Cloud, make it easier to set up and run scalable data infrastructures. Another benefit of using cloud-based infrastructure is also the cost and security (Bisong, Rahman, *et al.* 2011). There are also other approaches to using the cloud, called hybrid cloud solutions. Still, despite growing pains, there has been an increasing adoption of cloud technologies in different fields (Low, Chen, and Wu 2011).

The SAT part of an RDS needs to be developed in sync with the data management and use case identification elements. They will impose necessary constraints on the scope.

2.6 Operating model

A significant shift in the operating model of an organization is necessary for a successful RDS development and its downstream implementation. This task addresses three topics of equal importance: 1) organizational structure, policies, and culture; 2) workforce; 3) ways of working. A common practice for larger organizations is establishing a so-called "Center of Excellence" (SoC). This is a separate group, solely responsible for the RDS and its wider adoption throughout the organization. Whether this is the approach determined by the RDS or not, the three points mentioned above need to be addressed.

The first two items are addressed by introducing new roles or re-adjusting existing ones. Hiring a Chief Data Officer (CDO) and expanding the existing IT department's capabilities are also viable options. A more formidable task to achieve is a broader spread culture change, where data is seen as an asset. Organizations such as the NIH have achieved this by introducing the concept of "data

Figure 5. Different data team models.

fellows" (NIH *et al.* 2018). Data literacy efforts for domain experts across the organization will be beneficial for adopting the RDS.

The establishment of data roles, which have become the industry norm, can address existing workforce gaps. Those roles include data scientist, data engineer, data visualization expert, or machine learning engineer. The different models explain how those can be integrated within a research organization so that data becomes a core capability and shown in Fig. 5.

The final item in the operating model adjustment recommended by the RDS is the ways of working. Software work is very different from other fields, and data-intensive projects inherit a lot of those aspects. Here research organizations can borrow heavily from industry practices developed during the last decades in software, such as lean and agile methodologies (Hidalgo 2019). There has been skepticism on how many of those can successfully be implemented in the inherently unpredictable and complex research data environment, but there are examples of that in practice. Such methodologies can accelerate scientific research while also improving collaboration and transparency.

2.7 Use case identification

The final component of a complete RDS is the identification of possible use cases. Due to the lag of innovation in academic research compared to industry, many scientific domain experts would be oblivious of the potential applications of data technologies for their research. There is a recent trend of that change, but there are tools in RDS that can accelerate that process.

Ideation (or brainstorming) sessions, facilitated by methods from Design Thinking (DT) (Brown *et al.* 2008), can yield valuable information for researchers as to what to tackle next. This process depends heavily on the participating researchers' data literacy and is related to the operating model changes in the previous section. There is a recent innovation in this space with the creation of Data Thinking (Kronsbein and Mueller 2019). The resulting ideas need to be prioritized, and there are several methods available for that, such as the impact-effort matrix, where the different potential use cases can be ordered relative to each other based on the complexity of implementation and potential research impact.

2.8 Planning and roadmap

With all previous elements in place, the final deliverable of an RDS can be prepared. The input, consisting of recommendations on data management and governance, SAT, operating model, workforce, and use cases, can be prioritized and budgeted. There are methods such as the RACI table to make sure ownership and overview are available. An essential contributing factor to a

thorough RDS is making sure its elements are organized in Objectives and Key Results (OKR)², a technique popularized by Intel in the '70s and adopted to a large extent by Google and many Silicon Valley companies (Doerr 2018).

3 Conclusions

Creating a good RDS is just starting a transformation of a research organization into a data-driven one. This can be seen as a recipe, but the work often begins when this strategy needs to be implemented. The framework presented in this article needs to be adjusted for each case and might require a significant shift in the organization's operating model. With a thorough RDS in place, a research organization will be prepared to tackle the upcoming computational challenges.

References

- Bisong, A., M. Rahman, *et al.* 2011. "An overview of the security concerns in enterprise cloud computing." *arXiv preprint arXiv:1101.5613*.
- Brown, T., *et al.* 2008. "Design thinking." *Harvard business review* 86 (6): 84.
- DalleMule, L., and T. H. Davenport. 2017. "What's your data strategy." *Harvard Business Review* 95 (3): 112–121.
- Doerr, J. 2018. *Measure What Matters: OKRs: The Simple Idea that Drives 10x Growth*. Penguin UK.
- Fleckenstein, M., L. Fellows, and K. Ferrante. 2018. *Modern data strategy*. Springer.
- Hey, T., S. Tansley, K. Tolle, *et al.* 2009. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA.
- Hidalgo, E. S. 2019. "Adapting the scrum framework for agile project management in science: case study of a distributed research initiative." *Heliyon* 5 (3): e01447.
- Kronsbein, T., and R. Mueller. 2019. "Data Thinking: A Canvas for Data-Driven Ideation Workshops." In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- Kruhse-Lehtonen, U., and D. Hofmann. 2020. "How to Define and Execute Your Data and AI Strategy." *Harvard Data Science Review*.
- Kühlbrandt, W. 2014. "The resolution revolution." *Science* 343 (6178): 1443–1444.
- Laney, D. B. 2017. *Infonomics: how to monetize, manage, and measure information as an asset for competitive advantage*. Routledge.
- Larson, E. B. 2013. "Building trust in the power of "big data" research to serve the public good." *Jama* 309 (23): 2443–2444.
- Lenzerini, M. 2018. "Managing data through the lens of an ontology." *AI Magazine* 39 (2): 65–74.
- Low, C., Y. Chen, and M. Wu. 2011. "Understanding the determinants of cloud computing adoption." *Industrial management & data systems*.
- Miloslavskaya, N., and A. Tolstoy. 2016. "Big data, fast data and data lake concepts." *Procedia Computer Science* 88 (300-305): 63.

2. The use case planning can go a level deeper, and each one has to receive a set of Key Performance Indicators (KPI), that would indicate its success.

- NIH *et al.* 2018. "NIH strategic plan for data science." *NIH*, June.
- Ray, J. M. 2013. *Research data management: Practical strategies for information professionals*. Purdue University Press.
- Rohrbeck, R., J. Heuer, and H. Arnold. 2006. "The technology radar-an instrument of technology intelligence and innovation strategy." In *2006 IEEE international conference on management of innovation and technology*, 2:978–983. IEEE.
- Stall, S., L. Yarmey, J. Cutcher-Gershenfeld, B. Hanson, K. Lehnert, B. Nosek, M. Parsons, E. Robinson, and L. Wyborn. 2019. *Make scientific data FAIR*.
- Van Rijmenam, M. 2014. *Think bigger: Developing a successful big data strategy for your business*. Amacom.
- Yapp, C. 2005. "Innovation, futures thinking and leadership." *Public Money and Management* 25 (1): 57–60.