# Data Curation & Management

## Introduction

The benefits of data management have been widely recognised by government, industry, and research sectors as an important practice in their operations (Denison, Introduction – The Benefits of Managing Data, 2015). Particularly in the last decade, where an explosion of data and data products have surfaced, subsequently the amount of data generated and stored have exponentially increased. With the onset of the utilisation of data mining and data science practices, the interest of managing this data in a consistent manner has become a focal point.

Data management can cover a large range of tasks and responsibilities, from obvious tasks such as storage and migration to more complex and abstract responsibilities such as regulatory requirements and privacy (Denison, Introduction – The Benefits of Managing Data, 2015), with the latter being largely the focus of data curation and more or less data governance. It is to be noted, data curation and data governance is often used interchangeably. However, although related, each emphasizes a different practice towards data management (Denison, Data Curation and Management, 2015).

## Data Curation Roles

As defined in Alexandria, the data curation is aimed at enabling or improving the usability and usefulness of data throughout its lifecycle. Consequently, data curation and its scope of works includes, but not limited to, representation, archiving, authentication, management, preservation, retrieval and use (DHCuration, 2015).

Although the data curation process and somewhat the interpretation can vary within industries (Denison, Data Management Associations and Approaches to Data Management, 2015), a general number of core skills have been identified by the Data Curation Centre based in the United Kingdom.

It is highlighted that the skills required reach far beyond just technical expertise and multiple roles play an interrelating and interdependent part in enabling successful data management (Digital Curation Centre, 2004-2017). Figure 1 highlights two important roles that are specific to data curation, namely Data Manager and Data Librarian.

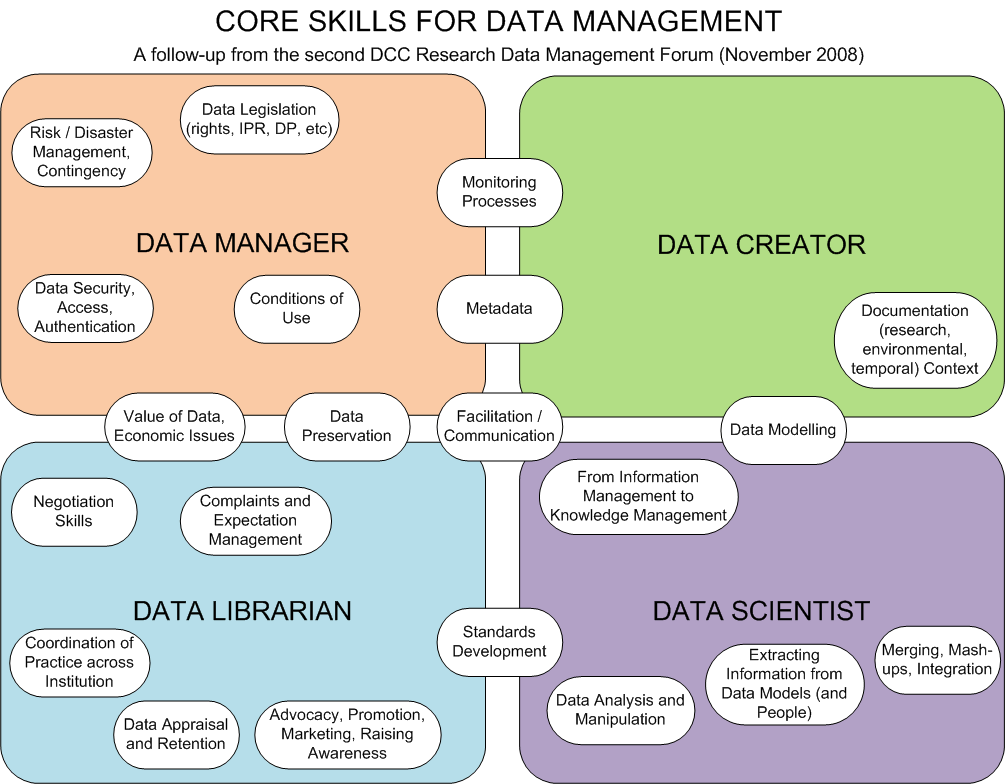


Figure 1: Data Curation Roles (Digital Curation Center)

Data Managers can be identified as the “back-end” of the whole process, whose responsibility is to ensure the development of engineering of the data curation process. Tasks such as data legislation require the ability to implement current data curation and governance principles in the relevant industry into practice. It is also identified that a good technical knowledge is essential to this role, particularly in data security, access and authentication.

Data Librarians on the other hand may be seen as the “front-end” of the process, where the responsibilities lie in the implementation of the data curation process in day to day operations. They have been identified to be responsible for expectation management and negotiation to ensure that data curation processes have been adequately followed. A notable responsibility includes the coordination of practice across institution, which is imperative to data homogeneity. This is essential in reducing inconsistencies not only in the data itself, but the value of the information that is contained within it, such as the semantic meaning of a symbol or code (Amit Rudra, 1999).

Although the number of roles identified for Data Creators is reduced, adequate documentation is key for data re-use across institutions and preservation. A corollary to documentation is the completeness of the data captured, since context is also considered. This is a key in reducing data pollution given that it was identified that the lack of completeness in the data was one of the leading causes of data pollution (Amit Rudra, 1999).

Data Scientists also have an important role in the data curation process. Adequate data curation enables the validity of extracted information and helps in the transformation of information to knowledge. This is particularly highlighted in big data, where Data Scientists mash up data from various different sources to develop data products which make decisions on individuals (Marcus R. Wigan, 2013). This is highlighted in the case of granting loans where a model may be developed using a large number of data that is gathered from various sources. Should the integrity of the data be polluted, then it is the consumer that is impacted from the lack of data curation.

Figure 1 also highlights some critical and overlapping responsibilities in the data curation process. Central to all role types is the facilitation and communication between each other. This is an obvious and crucial task to ensure that the “back-end” engineering is correctly followed, implemented and any changes are communicated to all stakeholders in the data curation process.

The overlapping responsibilities also show a two distinct groups, divisible by a vertical line between the groups. Responsibilities that overlap between Data Manager and Data Creators as well as Data Librarian and Data Scientists seem to be implementation of standards and processes which can broadly be said to be part of data governance. In contrast, the overlap between Data Creators and Data Librarians are predominantly responsibilities of data curation, ensuring data preservation and valued.

## Skills in Data Management

From the roles that have been identified above, a generalised set of skills can already be identified. These include, but are not limited to, stakeholder management, communication, process management and technical skills. Given that the roles differ from industry to industry (Denison, Data Management Associations and Approaches to Data Management, 2015), it is necessary to identify the unique traits of three major sectors that implement data curation.

### Skills in Governmental Data Management

There are a number of challenges in governmental data curation with the obvious being the handling of large number of public data sources. Strategic data governance in this context is very challenging, since standards, abbreviations, context, semantics change widely from individual government organisations. Governmental sector also issues unique challenges in privacy, particularly relating to national security.

Dianna Anderson, chief data officer at Colorado Governor’s Office identified that data governance had a large responsibility to the data creators (Wood, 2014). Furthermore, data governance and curation is identified to be a business issue rather than a technological issue (Wood, 2014), emphasizing the required skills to be more outcome oriented rather than solution oriented.

Specific roles that were present at the time of the development of this article showed that government focused on data quality management. Some of the key skills identified for the roles of Data Quality manger included, data quality policy development, ability to develop data quality standards and processes, stakeholder engagement to implement and evaluate the data quality framework, develop methodologies to monitor and manage data quality, and contribute do addressing data quality issues once identified (Total Resource Solutions, 2017) (Horizon Recruitment Specialists , 2017).

This highlights the interest in government sector in developing good data governance and curation policy. The main interest is shown to be that development of procedure and concepts to enable better data management practice to take place across the wide range of governmental sectors. Less emphasis is shown to be put on technical skills with focus primarily shifted towards stakeholder engagement and training. This approach to data curation and subsequently data quality management can be shown to be a combination of people and process based approaches (Denison, Approaches to achieving data quality, 2015).

### Skills in Corporate Data Management

Data curation in the corporate sector can be identified to be more or less a combination of governmental and research data curation. Business problems in the modern world vary widely, yet the validity and integrity of data have always played a central role in business continuity and information security (Teeling, 2012). Corporate culture and organisational inertia often results in the reluctance of acceptance in data governance arrangements that may be suggested by data curators (TechTarget, 2005-2017). Ownership issues are also produce significant challenges in implementing data curation where many business stakeholders are interested in immediate business concerns, and hold the view that data is a IT issue (TechTarget, 2005-2017).

Specific roles in the enterprise/corporate sector seem to have focuses both on stakeholder management and technical ability. Some common skills required by Master Data Manager and Data Quality Manager roles include proficiency in SQL, T-SQL, SAS, etc. (The Network, 2017) (Morgan McKinley, 2017). In addition to technical skills, data stewardship and accountability is highly regarded in roles (Site Minder, 2017). Data curators are expected to operate closely with data architects, various business departments and establish leadership and direction with senior management (Site Minder, 2017) (Morgan McKinley, 2017).

These position descriptions show that data curators in business need to undertake a number of key roles overarching a wide area of data curation. The expectation is set not only in policy development, but engaging relevant stakeholders and ensuring that policies are correctly implemented in the technical context. Ability to work with data architects to ensure that common ETL tasks adhere by the processes that have been developed by the data curator is also held in high regard. Data curation skills seem to be blended with statistical competencies which in many cases are traditionally being labelled as data science rather than data curation (Palmer, Weber, Muñoz, & Renear, 2017).

### Skills in Research Data Management

In contrast to corporate data management, research data management require a largely an involvement with the research team and subsequently the enabling of the data to the public for use (Palmer, Weber, Muñoz, & Renear, 2017). This requires a permeation on the entire workflow at an unprecedented level, from the initial planning stages of the research project followed by the on-going long term data management in both archiving, securing, and facilitating access to the public.

## Conclusion

It can be identified that although very much overlapping, there are some unique traits to each of the sectors in which data curation is carried out. Some obvious, such as roles presented in the government sector are very much geared towards policy and methodology implementation; aimed at beginning to standardise the data acquisition, merging and storage practices across a large number of different individual organisations.

Business sector seems to be more technical minded with less emphasis on policy development and more emphasis on implementation and data governance planning. As identified, they tend to bundle the data curation responsibilities with traditional technical skills, sometimes under the umbrella of data science.

In the research sector, data curators perhaps have the most expansive and permissive scope being embedded in from the very start of project or series of projects to the management of public access portals.