Assignment 4 Part 2: Data Canonicalization

Felicia Liu (liu318@illinois.edu)
Netid: liu318
CS598: Foundations of Data Curation

Part 2.2 Memo Instructions [20 points]

Write a short convincing memo (500 word max) explaining why data curation services are important. Assume that the memo is written for your new director, who is not familiar with data curation and not convinced whether to keep funding this work. You will want to make sure to introduce data curation within the broader context of data science. You will need to cover the key areas that you think are the most important for data curation at your company. We ask that you incorporate at least two of the following topics into your memo: Provenance, Policy, Metadata, and/or Preservation.

Memo:

To: Taylor Swift, Director of Data Science Department
Subject: The significance and benefits of data curation to our business

Ms Taylor,

Welcome to join our company and I am really thrilled and looking forward to collaborating with you in the future. This memo aims at introducing the current data curation project that my team is working on.
At present, the company makes a determination to decommission the old customer complaint system, and migrate the data to a new system. However, while the data scientists transferred the data for the first time, a great number of issues emerged such as loss of data, inconsistencies between data, duplications of data,and etc. In order to preserve the completeness of data migration, quality of data migration and data

integrity, my team has devoted ourselves to working on a data curation project and would like to bring the importance of data curation for long-term business development to your attention.

Data curation is commonly considered to bridge the gap between various stakeholders in a company or organization, so that the stakeholders are able to coordinate with each other and work seamlessly to present the outcomes and deliver values out of the data. There is no doubt that data curation will be beneficial to our organization in the following aspects:

- **Preservation**: Data curation can assist the company in easily discovering, understanding, and using the data in the future.
- **Provenance**: Data curation can organize, describe, clean and preserve the dataset across different systems, therefore, it ensures the organization to deliver and present high-quality data because it supports identifying what inputs, processes, and calculations are responsible for data values.
- **Policy:** Data curation is well-known for ensuring compliance through data lineage and classification because it always focuses on spotting data quality errors, identifying the root cause of issues, predicting the impact of potential changes and performing the auditing and documenting easier.
- **Metadata:** One of the advantages to conduct data curation in an organization is that it will map the data sets and catalog the metadata connected with the corresponding data sets, allowing the stakeholders to easily understand how the dataset can be used and what it originally contains.

As we may know there are multiple benefits of data curation, I still hope to detailedly introduce how data curation will be profitable to our company in the following two aspects.

Preservation plays an important role in curation activity, ensuring that the data will be understandable and usable in the future and maintaining a documented preservation strategy. It not only includes sequence preservation and syntax documentation, but the documentation of semantics for data elements and the generation and preservation of all metadata as well. To be specific, after transferring the old customer complaint dataset from the old system to the new system, if a customer complains about the same issue, we are able to easily obtain the data, which allows us to spend less time and money on data preparation and more time solving business problems.

Provenance is also a crucial curation activity, concentrating on supporting identifying what inputs, calculations, and actions are responsible for data values. If data curation is applied in our organization, then when the dataset is derived from another, reliable use and understanding requires that the inputs, calculations, and actions responsible for data values can be identified. While migrating the customer complaint dataset from the old system to the new one, we need to ensure the quality of the data sets and ensure that all the transferred values can be accurately identified.

Without successful data curation, it will be difficult to make sure the trust in data, resulting in the difficulty in performing data analysis and providing accurate business decisions and insights. Even if data curation can be a costly and challenging process for the company, it plays an indispensable role in migrating and storing accessible and sustainable data in the long run. Hence, our team highly recommends that we should continue the data curation project and increase the resources to this project.

Thanks and Regards,
Felicia
Data Scientist