Anantharaman Janakiraman,
Data Scientist,
XXXX Corporation.

Date: 11/25/2017

Sir/Madam,

Please treat this memo as a request and recommendation to continue funding data curatorial functions in our area which is an integral and essential precursor to data analysis. The quality of data analysis is dependent on the quality of information analyzed and data curation enhances the quality of data. Data quality issues can have a significant impact on business operations, especially when it comes to the decision-making processes within any organization.

The emergence of new platforms for decentralized data creation such as the digital platforms in our organization, the increasing availability of open data on the web, added to the increase in the number of data sources inside our organization, brings an unprecedented volume of data to be managed. In addition to the data volume, we need to cope with data variety, as a consequence of the decentralized data generation, where data is created under different contexts and requirements. Also consuming third-party data comes with the intrinsic cost of repurposing, adapting, and ensuring data quality for its new context.

Data curation provides the methodological and technological data management support to address data quality issues maximizing the usability of the data. Reusing data that was generated under different requirements comes with the intrinsic price of coping with data quality and data heterogeneity issues. Data can be incomplete or may need to be transformed in order to be rendered useful. Data shifts from a resource that is tailored from the start to a certain purpose, to a raw material that will need to be repurposed in different contexts in order to satisfy a particular requirement. In a scenario where there is a lot of data shift along with the context and to deal with incomplete data that requires transformation, data curation emerges as a key data management activity.

Data curation can be done from data generation perspective too which is also called curation at source and this could help represent the data in a way that maximizes its quality in different contexts. Data curation enables the extraction of value from data, and it is a capability that is required for areas such as ours that are dependent on complex and/or continuous data integration and classification. I would actually recommend improvement of data curation tools and methods that directly provides greater efficiency of the knowledge discovery process, maximizes return of investment per data item through reuse, and improves organizational transparency.

Despite the fact that data heterogeneity and data quality were concerns already present before the big data scale era, they become more prevalent in data management tasks with the growth in the number of data sources such as our shop. This growth as ours brought the need to define principles and scalable approaches for coping with data quality issues. It also brought data

curation from a niche activity, restricted to a small community of scientists and analysts with high data quality standards, to a routine data management activity, which will progressively become more present within the average data management environment.

Data curation processes can be categorized into different activities such as Collection, Organization, Storage, Preservation, Discoverability, Access, Workflow, Identification, Integration, Reformatting, Reproducibility, Sharing, Communication, Provenance, Modification, Compliance, Security and each of the above activities if performed appropriately will improve the data quality significantly that enhances the decision-making processes, transparency and accuracy.

I just want to highlight the importance of few of the activities in the context of data analysis that is performed in our organization on a daily basis. The Metadata that we produce through various processes in our organization can be used to support reuse and long-term preservation. Curating metadata supports discovery, schema changes, understanding and management of other data and information extensively. Similarly, some of the data quality attributes can be made evident by the data itself, while others depend on understanding of the broader context behind the data, i.e. the provenance of the data, the processes, artifacts, and actors behind the data creation. Capturing and representing the context in which the data was generated and transformed and making it available for data consumers is a major aspect of data curation and provenance standards provide the grounding for interoperable representation of data. Curation as we currently employ in our organization supports the ability to retrieve and distribute data – maintain systems, tools and metadata that support the efficient and reliable retrieval and distribution of data.

In essence, with the growth in the number of data sources and of decentralized content generation, ensuring data quality becomes a fundamental issue for data management environments in the big data era. The evolution of data curation methods and tools is a cornerstone element for ensuring data quality at the scale of big data.

I sincerely request you to reconsider your decision to de-prioritize data curation services in our organization and also request you to provide careful consideration when revisiting your decision. I sincerely hope the reasons stated in this memo would emphasize the need for data curation services if not more and help you make an informed decision.

Please let me know if you have further questions or concerns.

Thanks,
Anantharaman Janakiraman.