CS598 – Data Curation Final Project - Memo

By Ved Chugh(vedpc2)

To: Bruce Wayne, Director of Data Science

Subject: Importance and Aspects of benefits of Data Curation to our Division.

Dear Bruce,

Congratulations on your new role, I look forward to working with you.

To give you a detailed picture, I wanted to share details of:

- ⇒ Our team members.
- ⇒ The current scope of work for our Data Science team.
- ⇒ The day to day challenges we face.
- ⇒ And my current learnings and POC that we have been conducting at large scale in our team.

Currently we have 25 team members of which we have Data Analyst/ Engineers/ Scientist, Statisticians, Business Analyst, DBA and Data & Analytics Manager.

Below is visual display of our Day to day work scope. (*Source DS roles and responsibilities*)

Since the advent of this team on day to day basis below are few challenges that we are facing as a team.

- ⇒ Because of the lack of Documentation – Metadata standards in the organization, we have integration problems, we have hours of Analysis delays.
- ⇒ Further there has been a lack of Standard data models, the same schema file existing in different systems within the org have different way of representation. Which leads to inconsistent data sets. And severe integration issues.
- ⇒ Every time we have an updated Tool or Data reformat, we go through painstaking challenges in order to implement those, and severe challenges of backward compatibility
- ⇒ With lack of Metadata there is no clear objective as to why few Data sets exists in first place.
- ⇒ This Leads to often redundant information, and with no Provenance in place it is hard to say or analyze, which dataset is authentic.
- ⇒ Procedures, Practices and formats are not at all well documented, which leads to Identity issues.
- ⇒ Preservation of Data hence becomes a big challenge.
- ⇒ With Lack of Metadata, and proper documentation Data Discoverability is one of the biggest hurdles our Business Units face. And hence we as a team struggle to give deeper insights to Business.
- ⇒ None of the existing work and data is reproducible due to lack of proper Procedures and Practices.

These are a few challenges that we face, and the list goes on.

Well but the good news is that we have identified the solution and approach to solve it.

Solution is Data Curation.

As per Prof. Allen Renear from UIUC - Data science is concerned with all aspects of the **creation, management, analysis,** and **communication** of data focusing particularly on the application of computational methods to digital data.

He further states - Data science **=** Data Curation **+** Data Analytics

Where –

Data curation: Ensuring that data can be efficiently and reliably found and used

Data analytics: Employing specific techniques to extract knowledge from data

Data curation is concerned with all aspects of the management of data in order to efficiently and reliably support the analysis of data and enable reuse over time.

It includes among many other things: acquisition and collection, modeling, workflow, provenance, validity and integrity, metadata, preservation, integration, retrieval, re-use, policy, standards, identifiers, format conversions, processing levels, supporting reproducibility, etc.

He helped us identify the Areas of Curatorial Activities

Objective	Reasoning
Collection	Support the collection and acquisition of data
Organization	Employ an appropriate data model and use appropriate standards
Storage	Support reliable and effective storage
Preservation	Ensure that data will be understandable and useable in the future
Discoverability	Support the ability to search for and locate relevant data
Access	Support the ability to retrieve and distribute data
Workflow	Support the ability to systematize data workflows
Identification	Support the ability to identify, authenticate, and validate data
Integration	Support integration of data from different sources using different data models
Reformatting	Support reformatting for use by different tools or to match new format standards
Reproducibility	Support ability to reproduce results, ensuring scientific validity and reliability
Sharing	Support sharing data between researchers, teams, and institutions
Communication	Support representation, publishing, and visualizations that provide insight
Provenance	Support .identifying what inputs, processes, and calculations are responsible for data values
Modification	Support management of corrections and updates
Compliance	Ensure compliance to legal, regulatory, and local policy requirements
Security	Ensure that data is secure from tampering or inappropriate access and distribution

We recently did a POC on same, where in we had consumer complaints system being migrated to new System. The migration team came to us for solution, wherein they had two xml files and they wanted us to help integrate same.

We were successfully able to integrate the file using DTD and Canonicalization of XML files to achieve the result and below are few results we achieved.

I would be more than glad to setup a demo with you to walk through same.

Objective	Reasoning
Organization	The DTD and appropriate standards help organize the data.
Preservation	Metadata documentation DTD helps ensure that data will be understandable and useable in the future.
Discoverability	With DTD in place and correct readable formatting it supports the ability to search for and locate relevant data.
Access	The xml file also supports the ability to retrieve and distribute data
Identification	The xml file supports the ability to identify, authenticate, and validate data
Integration	Once the future xml file address and adhere to DTD format of Canonicalized file it helps support integration of data from different sources.
Reformatting	Canonicalized file makes reformatting an easy task. As the defaulted attributes and elements have been normalized thus it becomes really easy to reformat the file.
Reproducibility	It definitely supports the ability to reproduce results, ensuring scientific validity and reliability.
Sharing	It helps support sharing data between researchers, teams, and institutions and easy task, due to standardization of file.
Communication	It helps support representation, publishing, and visualizations that provide insight
Modification	The canonicalized file supports management of corrections and updates easy.

Moreover, apart from this we have been able to implement proper Metadata of our existing datasets using ER Models, Ontologies and have extensively used DTD to document our xml datasets.

We have also been able to understand various organization and government standards around Data Curation, Metadata, Provenance, Identity and security.

With onset of new work that we are taking up and further expansion that we see, I strongly believe that our investment in this forte will yield to benefits in Data Science.

I will send you a separate budgeting and funding forecast for same.

Deepest Regards

Ved Chugh

Data & Analytics Manager.