Discussion Board Topic: Metadata – Why is it so important?

What is Metadata?

In the simplest terms metadata is a definition for data that helps the user understand what exactly they're looking at and aids in identifying the differences between variables. A collection of metadata is often referred to as a Metadata Repository or Data Dictionary. Metadata has several use cases such as providing information about other data, summarizing basic information about data and being manually created so that it is up to date with current definitions to be more accurate (1).

Metadata can provide lots of different information, such as the what, when, where, who, why and how descriptive information. It is usually stored some place that is easily accessible to its intended audiences, for example a GitHub repository if it's an open source dataset or a SharePoint site if it's intended for employee access only. It should also have a detailed title so that its purpose is clear, and definitions should be detailed, especially when the values are categorical.

Why is it so Important?

Whenever you don't understand the definition of a word you look it up in a dictionary. Without dictionaries people would often get confused and misinterpret words and sentences. The same is for data; without metadata and data dictionaries we would have a hard time understanding what the data we're working with actually represents and what it is telling us.

Other purposes of metadata include extending data longevity, facilitate data reuse and sharing and maintaining historical records of long-term data sets (2). Without metadata it can become very easy to stop recording data variables because the definition is unknown. When this happens, the data set can become filled with null values and then become useless to work with. Detailed metadata ensures that this does not happen and allows us to share data with other people so that is can be reused for further analysis and decision making.

My Personal Experience with Metadata

I personally believe that the metadata and Data Dictionary are the most important resources when doing any analysis. Often times I find myself confused if the metadata is not easily available. Last semester I worked on a project for visualizing the COVID-19 outbreak in South

Korea. I found a dataset from the CDC in South Korea and it included nicely detailed and easy to understand metadata (3). The metadata broke out how each of the datasets are combined, the variable types and how some variables are created. I found this to be a good experience for an open source dataset.

In one of my previous positions, I worked with a data table that had no clear metadata or data dictionary with any detail. Already familiar with the variable names I needed to gather, I went ahead and queried a table to gather the information I thought to be correct. Little did I know that the data from that table was not complete and was not being recorded for quite some time now. The end result was a deck with incorrect and missing information. I was told "sorry, we don't use that table for so-and-so reporting." When I asked for a data dictionary for our data base, I was told that one wasn't available. From my own personal experience, it can be seen how metadata would've made the difference had it been kept up to date and easily available. I wouldn't have turned in an incorrect report and I would've been able to reference the correct data tables for analysis.

Questions?

- I'm sure metadata is often overlooked, what is/was your opinion on metadata and how has it changed?
- Have you had a bad experience with metadata?
- Do you think your company can do a better job managing their metadata?

Sources

(1) https://www.opendatasoft.com/blog/2016/08/25/what-is-metadata-and-why-is-it-important-data
(2) https://www.villanovau.com/resources/bi/metadata-importance-in-data-driven-world/
(3) https://github.com/jihoo-kim/Data-Science-for-COVID-19/blob/master/dataset-detailed-description.ipynb