Harsiddh Patel

Reflection Prompts

a) Describe your process for canonicalization (i.e., decisions, actions, representation selection, attribute issues, provenance decisions). Report the checksum values after canonicalization.

In the canonicalzation process, I made various changes to get identical checksum for both files. First, I removed extra spacing and whitespaces in both files since spacing adds no value. I removed <submitted/> elements along with via attribute from FileA and added value of via attributes to submissionType attribute to <complaint/> elements like done in FileB. In addition, I added timely attributes to some of the <response/> elements that were missing in FileB. Also, I changed the format of the value in the timely attribute in FileB to match the format in FileA. Finally, I changed the ordering of the attributes (ex. timely, types) in FileB to match with FileA. After making all these changes in both files, I got the same checksum number.

Checksum for canonicalized FileA is *84cab49e05a86074739be1162e3f96f7* and for FileB is *84cab49e05a86074739be1162e3f96f7*

b) How does the way data is represented impact reproducibility?

The way data is represented supports reproducibility since compliant structure is repeated every time with identical child elements and the attributes. This ensures that every compliant element will contain the specific attribute or element. This representation makes sure that the reader can easily find the specific data, or the application can quickly parse and lookup specific data. For example, retrieving list of all complaints who submitted complaints with “web” can be done by looking at submissionType attribute value.

c) How may your canonicalization support the overarching goals of data curation (revisit objectives and activities of Week 1)?

The canonicalization supports Perseveration since the XML can easily be reused in the future and easy to read/flow the different elements and attributes plus after canonicalization, the data was preserved and did not lose integrity plus the object and goal was not lost and was met. Another other goal that was achieved was Sharing, since I canonicalized the old system and the new system, any team can use this file to for their operation by following the standard format and expecting the same result as before. Another goal is Communication, the canonicalized file has clear element names which can be understandable by the reader plus after utilizing the data, teams can easily communicate and represent plus visualized the data.

d) Which additional curation activities would you recommend to enhance the data set for future discovery and use?

One recommendation to enhance the best practices of Cybersecurity which will mitigate risk and accomplish confidentiality, integrity, and availability of data security. Other recommendation is avoiding utilizing attribute lists and rather use nesting element.