REFLECTION PROMPTS

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

Below steps were following for canonicalization of FileA and FileB

I. Convert to a single character encoding and normalize line ends.
II. Remove all comments, tabs, non-significant spaces, etc.
III. Propagate all attribute defaults indicated in the schema to the elements themselves
IV. Put attribute/value pairs on elements in alpha order
V. Expand all character references
VI. Remove any internal schema or declarations.
VII. Now test to see if character sequences are identical.

MD5 Checksum:
File A: ***fed0eda489626463876541c56e93c099***
File B: ***2550c686aafbdf320bd74836a511cd54***

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

FileA and FileB has external reproducability difference. When observing the xml data files, we notice the difference. In terms of DTD as well, we notice some difference internally.

Once canonicalization is done, we don’t see any external and internal differences.

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

- Collection – Since we follow XML standards with a robust DTD schema, collection and acquisition of data is supports.
- Organization – The systematic organization of data in a hierarchical format because of the xml syntax helps in achieving this objective.
- Storage – The xml schema and subsequent data representation allows us to store data in reliably and effectively.
- Preservation – Because of the well defined, hierarchical representation of data, the resultant data/xml can be easily readable, understandable and can be reused depending on the circumstances.
- Discoverability – XML format is a well defined data representation format. Data can be easily searched and retrieved. Since we follow all the standard rules, our data achieves discovery objective.
- Access – We don’t have any access restrictions or permissions assigned for users who would like to retrieved. Hence this objective is not met in this assignment.
- Workflow – Data is systematized because of the xml format standard adherence of our data.
- Identification- We don’t support identity, authenticate and validate data for this assignment. These activities are not targeted in this assignment.
- Integration – Different data models from different data sources cannot be integrated in this assignment results. This activity isn’t targeted and not achieved in this assignment.

- Reformatting – Since we follow standard xml guidelines for data representation and tool that adheres to those rules can be programmed to reformat this data. This assignment doesn’t target at reformatting support internally.
- Sharing – There is no mechanism described or achieved in this assignment to support sharing on the document. But once shared the data can be easily preserved.
- Communication – This assignment doesn’t address this activity.
- Provenance – This activity is partially achieved by adding relevant data place holders in the data format i.e by adding few move XML tags for adding relevant information.
- Modification – There is not process described or supported to achieve this activity in the assignment.
- Compliance – The data representation adheres to XML syntax standards and other relevant standards for describing and defining data. But the assignment in itself doesn’t support this activity of legal, regulatory and local policy requirements although by adding relevant tags, it can be supported.
- Security – There is no mechanism in this assignment to enable security against data tampering in this assignment.

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

- It would be great if we have a standard and automated process for canonicalization