Question 1:
Write a short profile of each file we have given you.
File A (Old System):
The file is an XML document that provides details about the consumer complaints with 8 records. The file has namespaces/indentations propagated down from any parent element. The file does not have DTD defined internally. The elements in are xml for each record are similar except for the last record. Looking at the data, some of the attributes have default values defined. The root element for the XML is cosumercomplaints. This can have multiple complaint records in it with attribute ID as a required value. Complaint element have multiple sub elements as shown in the DTD with nested sub elements.
MD5 Checksum: 8D00F7DE8D007DD2D1641B817E97FE61

File B (New System):
The file is an XML document that provides details about the consumer complaint with 8 records as well. The file does not have proper has namespaces/indentations propagated down from all the parent element. The file has a small DTD defined internally for entity redaction; however, it must be expanded to incorporate the entire structure. Comparing to File A, it seems like some of the default attributes are not defined in File B. The root element for the XML is cosumercomplaints. This can have multiple complaint records in it with attribute ID as a required value and submissionType as an optional value. Complaint element have multiple sub elements as shown in the DTD with nested sub elements.
MD5 Checksum: 47677272E76E1F4332AFE859347C8695

Question 4:
Create and document the DTD of the final, canonicalized data file

Please find attached the document ***"Final canonicalized file and DTD"*** for the DTD.

The DTD for the final canonicalized xml file is stored as an internal DTD. The assumptions for the layout and selection of certain elements/attributes are documented in detail in the Question 6, Section "a" documentation below.

The main element for the final canonicalized file DTD is the consumerComplaints. This element has additional sub element complaint which can have multiple records.

The element complaint has sub elements event, company, response, issue, product and consumer narrative and attributes id which is a required field and the attribute submissionType which has a default value of "Web"

All the sub elements of the element complaints have addition sub elements and attributes nested in the as seen in the DTD.

Since the final canonicalized file has the sub elements in the different order within the different records. The sub element ordering is also specified in the DTD.

Question 6:

a) Describe your process for canonicalization:

- Compared both the file structures to determine if they both had the same data structure. Both the files were structured similarly with few differences listed below. The data structure of the file was updated first to match.
    - "submitted" in File A is a sub element for the element "Complaint" with "via" being passed as an attribute to the element "submitted", however in File B this value is passed as an attribute "submissionType" to the element "Complaint". File A data structure is updated to pass attribute "via" as an attribute to "Complaint" and renamed it as "submissionType"
    - There is only one xml empty line for "submitted" in File B. This has been deleted.
    - File B had submissionType attribute missing in two complaint line. This data is available in File A. Based on the data, a default value of "Web" is assigned the missing entries in File B.
    - File A has the value for attribute "timely" in element response as "Y" while File B has the value for attribute "timely" in element response as "Yes". Also "timely" value is missing for a couple of lines in File B. A default value of "Y' has been assigned to these.
- The XML declaration and document type declaration (DTD) are removed.
- Removed all comments (There was only one comment in File B)
- Converted the encoding of both the files to UTF-8.
- Normalized line breaks to #xA on input, before parsing.
- Attribute "type" in element Event has whitespace before the quotes in file B which has been removed.
- Attribute values are normalized, and value delimiters are set to quotation marks. Whitespace within start and end tags is normalized
- Namespaces were propagated down from any parent element
- Entity references are replaced with values.
- Lexicographic order was imposed on the attributes of each element and default values were propagated to the Elements

MD5 Checksum Before Canonicalizing

File A: 8D00F7DE8D007DD2D1641B817E97FE61

File B: 47677272E76E1F4332AFE859347C8695

MD5 Checksum After Canonicalizing

File A: 7EBB3EBA0945D42CF26C0165CFF99D1F

File B: 7EBB3EBA0945D42CF26C0165CFF99D1F

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

In Data Curation, reproducibility is the ability to reproduce results, ensuring scientific validity and reliability. Since the data is canonicalized using the standard rules, it ensures that the data is accurate and valid. This in turn ensures that the results are reproduceable reliably.

c) How may your canonicalization support the overarching goals of data curation?

The main objectives of data curation are overall management of data for reliable analysis and reuse overtime. Canonicalization is a process for converting data that has more than one possible representation into a "standard", "normal", or canonical form. This supports the following goals of data curation. Since the data in my canonicalized file has a standard structure it ensures that the data

- is organized using standard data model and rules. (Organization)
- data will be understandable and useable in the future. (Preservation)
- supports the ability to search for and locate relevant data. (Discoverability)
- supports the ability to identify, authenticate, and validate data. (Identification)
- supports integration of data from different sources using different data models. (Integration)
- supports reformatting for use by different tools or to match new format standards (Reformatting)
- supports ability to reproduce results, ensuring scientific validity and reliability (Reproducibility)
- supports identifying what inputs, processes, and calculations are responsible for data values (Provenance)

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

Collection: Having some fields as required or having default values for certain attributes ensures that the data is complete, which can help with more accurate analysis of data

Sharing: Data can be easily retrieved for analysis and shared with other organizations and institutions for research.

Compliance: Ensure compliance to legal, regulatory, and local policy requirements. Ensuring that the data is compliant makes sure that we are prepared for Audits.

Security: Access to the data should be restricted to the relevant teams/people. This will ensure that unauthorized entry do not happen and reduce errors in the data.

External References:

https://en.wikipedia.org/wiki/Canonicalization

https://www.xml.com/pub/a/ws/2002/09/18/c14n.html

https://www.w3.org/TR/xml-c14n11/