CS598 – Data Curation Final Project Cont.

By Ved Chugh(vedpc2)

SECTION 6 – Canonicalization steps and reflections

Part a

The Consumer complaints file even though they have same data but are different and that is why they have different MD5 checksum.

- ⇒ There are various differences in files that attribute this.
- ⇒ There are line spacing and unnecessary white spaces.
- ⇒ Attributes in elements don't follow strict order.
- ⇒ There are comments in File B missing in File A.
- ⇒ The data for attribute timely is a text "yes" or "no" rather than a more standard "Y" or "N" flag.
- ⇒ The attribute under submitted element is moved at complaint level in file B.

These are the challenges for which Canonicalization exists.

Canonicalization steps as per Data Curation CS598 coursework is as below -

To determine whether or not two XML files determine the same data structure.

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

Below are the steps followed.
For the sake of Provenance and consistency File A has been considered the legend in terms of Values. Looking at the file, it has formatting in place and thus the starting point for me was to have the DTD of File A as the reference and to clean the file A and File B to have canonicalization in place.

- ⇒ The UTF-8 encoding is used – Both files have UTF-8 encoding thus following same encoding standard.
- ⇒ Line-ends are represented using the newline character 0x0A – To standardize line encoding.
- ⇒ Whitespace in attribute values is normalized – All white spaces in attributes i.e for e.g <event type=> etc. have been normalized to standards in both files A and B

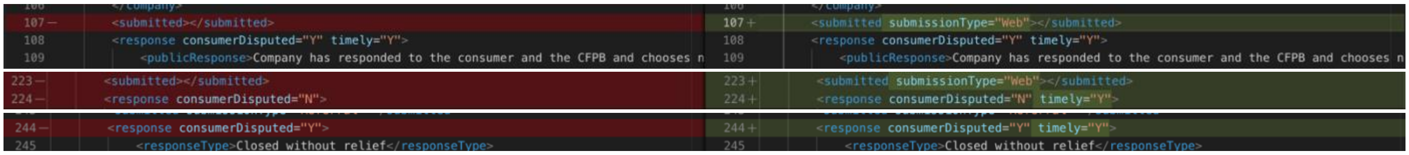
- ⇒ Entity references and non-special character references are expanded - &redaction entity has been deleted from file B and further all reference values have been replaced with defaulted value of entity (XXXX)
- ⇒ Empty elements are encoded as start/end pairs, not using the special empty-element syntax – e.g.- <submitted/> - has been expanded to <submitted></submitted>
- ⇒ In both the files attributes are encoded in a normative order (alphabetical by name) – e.g. - rearranged the attribute “type” and “date” from element “event” in order of “date” and “type”.
- ⇒ Also following above rule – rearranged "consumerDisputed" and "timely" in response element.
- ⇒ Finally the DTD is matched in both files and thus the schema of elements and attributes is matched.
- ⇒ To ensure that XML is valid a final check revealed that the order/sequence of elements last complaint id in file doesn’t match the DTD, thus rearranged the <submitted> element to help Validate the XML.

Now the reality check, to see if the files are identical and if their check sum matches.

Result – NO
Canc_consumer_complaints_fileA.xml - C0CAB6445665D2E8FB3905C7A32B80DE
Canc_consumer_complaints_fileB.xml - 6532E5E29F2A879A0141451524BD8E17

Reason – Difference reveals – that the timely attribute and submissionType attribute has missing values in new system fileB.

- ⇒ There is no attribute “submissionType” and its value for complaint id **2364257**
- ⇒ There is no attribute “submissionType” & “timely” and its value for complaint id **837784**
- ⇒ There is no attribute “timely” and its value for complaint id **14038**



```
-   <submitted></submitted>
    <response consumerDisputed="Y" timely="Y">
        <publicResponse>Company has responded to the consumer and the CFPB and chooses n

-   <submitted></submitted>
-   <response consumerDisputed="N">

-   <response consumerDisputed="Y">
        <responseType>Closed without relief</responseType>
```

```
+   <submitted submissionType="Web"></submitted>
    <response consumerDisputed="Y" timely="Y">
        <publicResponse>Company has responded to the consumer and the CFPB and chooses n

+   <submitted submissionType="Web"></submitted>
+   <response consumerDisputed="N" timely="Y">

+   <response consumerDisputed="Y" timely="Y">
        <responseType>Closed without relief</responseType>
```

Action
Thus, when ensuring that Compatibility with previous version and consistency of data and Provenance – the values are copied over to missing values to File B.

Outcome
Thus, making both files absolutely identical and yielding to matching checksum.
Canc_consumer_complaints_fileA.xml - C0CAB6445665D2E8FB3905C7A32B80DE
Canc_consumer_complaints_fileB.xml - C0CAB6445665D2E8FB3905C7A32B80DE

This results to our final Canonicalized file –

Final_Merged_Canc_consumer_complaints_file.xml - C0CAB6445665D2E8FB3905C7A32B80DE

Part b - How does the way data is represented impact reproducibility?

Reproducibility according to CS598-DC coursework can be defined as the ability to reproduce data/results by ensuring the validity and reliability. It is the key function to ensure future reliability of the chosen format.

With the current representation:

- ⇒ First because the fact file A doesn't have any documentation DTD thus when new system came into picture, there was no metadata standard to adhere to.
- ⇒ Thus, with the missing Documentation in place, missing DTD and missing Metadata info, when fileB was produced, we could see loss of information in some attributes.
- ⇒ Further impacting reproducibility. As the element and attribute sequences between and old and new files didn't match.
- ⇒ This further also impacts tools dependent on the xml file, as the new system file doesn't match or adhere to old format.
- ⇒ For reproducibility one has to spend hours understand the dataset and xml formatting to have first documentation in place and then as a next step reproduce the data.

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

Below are the few objectives which are supported by Canonicalized Xml.

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.

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

Objective	Reasoning
Compliance	To adhere to standard organization and legal standards
Security	To avoid random changes and loss of data the Xml file can be version controlled along with Access control should be implemented.
Provenance	So as to help identify what inputs, processes, and calculations are responsible for data values as in this case the being a consumer complaints file it gives us detail information of same. When one data set (or view) is derived from another, reliable use and understanding requires that the inputs, calculations, and actions responsible for data values can be identified.