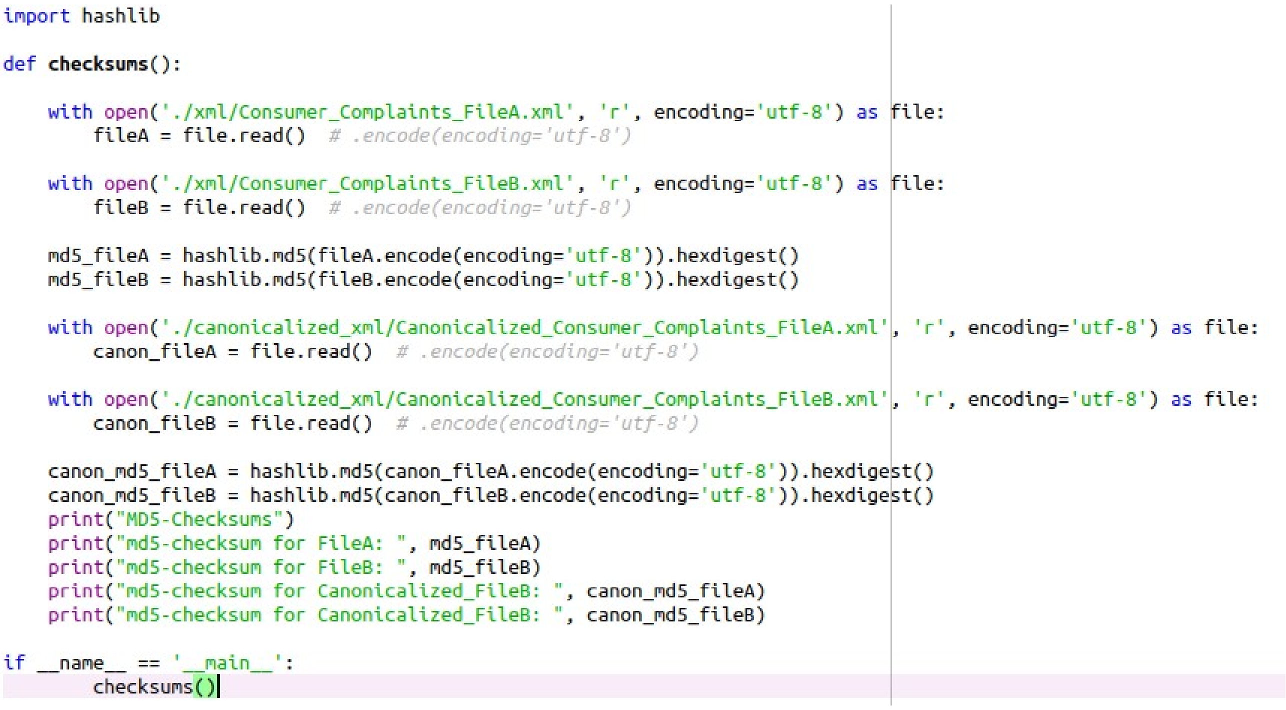
REFLECTION PROMPTS

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

Validator Script:

I created a md5_validator.py script to validate the xml files. Below are the contents of the script.



```
import hashlib

def checksums():

    with open('./xml/Consumer_Complaints_FileA.xml', 'r', encoding='utf-8') as file:
        fileA = file.read()  # .encode(encoding='utf-8')

    with open('./xml/Consumer_Complaints_FileB.xml', 'r', encoding='utf-8') as file:
        fileB = file.read()  # .encode(encoding='utf-8')

    md5_fileA = hashlib.md5(fileA.encode(encoding='utf-8')).hexdigest()
    md5_fileB = hashlib.md5(fileB.encode(encoding='utf-8')).hexdigest()

    with open('./canonicalized_xml/Canonicalized_Consumer_Complaints_FileA.xml', 'r', encoding='utf-8') as file:
        canon_fileA = file.read()  # .encode(encoding='utf-8')

    with open('./canonicalized_xml/Canonicalized_Consumer_Complaints_FileB.xml', 'r', encoding='utf-8') as file:
        canon_fileB = file.read()  # .encode(encoding='utf-8')

    canon_md5_fileA = hashlib.md5(canon_fileA.encode(encoding='utf-8')).hexdigest()
    canon_md5_fileB = hashlib.md5(canon_fileB.encode(encoding='utf-8')).hexdigest()
    print("MD5-Checksums")
    print("md5-checksum for FileA: ", md5_fileA)
    print("md5-checksum for FileB: ", md5_fileB)
    print("md5-checksum for Canonicalized_FileB: ", canon_md5_fileA)
    print("md5-checksum for Canonicalized_FileB: ", canon_md5_fileB)

if __name__ == '__main__':
        checksums()
```

MD5 Checksum (md5_validator.py):

Below are the md5 checksum results after executing the python script:

```
md5-checksum for FileA:  2c4f6af0726ce00d76ea95385c4bb78b
md5-checksum for FileB:  2550c686aafbdf320bd74836a511cd54
md5-checksum for Canonicalized_FileB:  76cb27603b94e4845f7c18f323d4e5a9
md5-checksum for Canonicalized_FileB:  76cb27603b94e4845f7c18f323d4e5a9
```

Steps followed for canonicalization are listed below:

I. Sort all the attributes and sub-attributes in a lexicographical order.

II. Maintain same spacing for the entities and attributes in the xml files.

III. use Y for yes and N for no on timely attribute in fileB. This helps in maintaining consistency.
<response timely="yes" consumerDisputed="Y"> →
<response consumerDisputed="Y" timely="Y">

IV. Replace “submitted via” field with “submissionType” in the attribute
<submitted via="Phone"/> → submissionType="Phone"

V. In fileB, for complaint id 2364257 following attributeType is added from FileA to maintain consistency.
<complaint id=" 2364257" submissionType="Web">

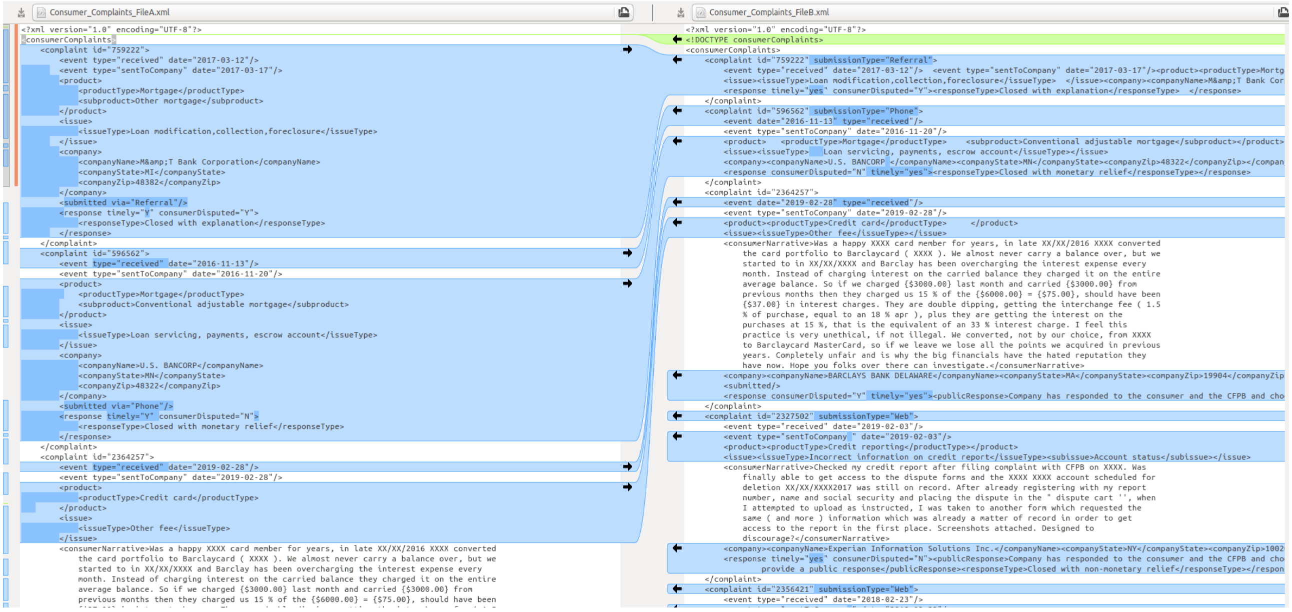
VI. In fileB, for complaint id 14038, following change is added
<response consumerDisputed="Y" timely="Y">

VII. In fileB, for complaint id 837784 following changes are added.
<response consumerDisputed="N" timely="Y">
<complaint id="837784" submissionType="Web">

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

Reproducability comes in many different manners. The choice of XML is significant because the data and the data-carrying structures can be preserved in a number of ways. Directly through remarks in the data files (but not the canonical files which have to be remark free), in the stylesheets or scripts, etc. Additionally, because schema documents like DTD, XSD, etc. can be applied to the documents, the data can be validated and another dimension of preservation can be realized through the schema documents themselves which record the data's structure.

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



```
<?xml version="1.0" encoding="UTF-8"?>
<consumerComplaints>
    <complaint id="759222">
        <event type="received" date="2017-03-12"/>
        <event type="sentToCompany" date="2017-03-17"/>
        <product>
            <productType>Mortgage</productType>
            <subproduct>Other mortgage</subproduct>
        </product>
        <issue>
            <issueType>Loan modification,collection,foreclosure</issueType>
        </issue>
        <company>
            <companyName>M&amp;T Bank Corporation</companyName>
            <companyState>MI</companyState>
            <companyZip>48382</companyZip>
        </company>
        <submitted via="Referral"/>
        <response timely="Y" consumerDisputed="Y">
            <responseType>Closed with explanation</responseType>
        </response>
    </complaint>
    <complaint id="596562">
        <event type="received" date="2016-11-13"/>
        <event type="sentToCompany" date="2016-11-20"/>
        <product>
            <productType>Mortgage</productType>
            <subproduct>Conventional adjustable mortgage</subproduct>
        </product>
        <issue>
            <issueType>Loan servicing, payments, escrow account</issueType>
        </issue>
        <company>
            <companyName>U.S. BANCORP</companyName>
            <companyState>MN</companyState>
            <companyZip>48322</companyZip>
        </company>
        <submitted via="Phone"/>
        <response timely="Y" consumerDisputed="N">
            <responseType>Closed with monetary relief</responseType>
        </response>
    </complaint>
    <complaint id="2364257">
        <event type="received" date="2019-02-28"/>
        <event type="sentToCompany" date="2019-02-28"/>
        <product>
            <productType>Credit card</productType>
        </product>
        <issue>
            <issueType>Other fee</issueType>
        </issue>
        <consumerNarrative>Was a happy XXXX card member for years, in late XX/XX/2016 XXXX converted
            the card portfolio to Barclaycard ( XXXX ). We almost never carry a balance over, but we
            started to in XX/XX/XXXX and Barclay has been overcharging the interest expense every
            month. Instead of charging interest on the carried balance they charged it on the entire
            average balance. So if we charged {$3000.00} last month and carried {$3000.00} from
            previous months then they charged us 15 % of the {$6000.00} = {$75.00}, should have been
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE consumerComplaints>
<consumerComplaints>
    <complaint id="759222" submissionType="Referral">
        <event type="received" date="2017-03-12"/>  <event type="sentToCompany" date="2017-03-17"/><product><productType>Mortg
        <issue><issueType>Loan modification,collection,foreclosure</issueType>  </issue><company><companyName>M&amp;T Bank Cor
        <response timely="yes" consumerDisputed="Y"><responseType>Closed with explanation</responseType>  </response>
    </complaint>
    <complaint id="596562" submissionType="Phone">
        <event date="2016-11-13" type="received"/>
        <event type="sentToCompany" date="2016-11-20"/>
        <product>   <productType>Mortgage</productType>    <subproduct>Conventional adjustable mortgage</subproduct></product>
        <issue><issueType>   Loan servicing, payments, escrow account</issueType></issue>
        <company><companyName>U.S. BANCORP </companyName><companyState>MN</companyState><companyZip>48322</companyZip></compan
        <response consumerDisputed="N" timely="yes"><responseType>Closed with monetary relief</responseType></response>
    </complaint>
    <complaint id="2364257">
        <event date="2019-02-28" type="received"/>
        <event type="sentToCompany" date="2019-02-28"/>
        <product><productType>Credit card</productType>     </product>
        <issue><issueType>Other fee</issueType></issue>
        <consumerNarrative>Was a happy XXXX card member for years, in late XX/XX/2016 XXXX converted
            the card portfolio to Barclaycard ( XXXX ). We almost never carry a balance over, but we
            started to in XX/XX/XXXX and Barclay has been overcharging the interest expense every
            month. Instead of charging interest on the carried balance they charged it on the entire
            average balance. So if we charged {$3000.00} last month and carried {$3000.00} from
            previous months then they charged us 15 % of the {$6000.00} = {$75.00}, should have been
            {$37.00} in interest charges. They are double dipping, getting the interchange fee ( 1.5
            % of purchase, equal to an 18 % apr ), plus they are getting the interest on the
            purchases at 15 %, that is the equivalent of an 33 % interest charge. I feel this
            practice is very unethical, if not illegal. We converted, not by our choice, from XXXX
            to Barclaycard MasterCard, so if we leave we lose all the points we acquired in previous
            years. Completely unfair and is why the big financials have the hated reputation they
            have now. Hope you folks over there can investigate.</consumerNarrative>
        <company><companyName>BARCLAYS BANK DELAWARE</companyName><companyState>MA</companyState><companyZip>19904</companyZip
        <submitted/>
        <response consumerDisputed="Y" timely="yes"><publicResponse>Company has responded to the consumer and the CFPB and cho
    </complaint>
    <complaint id="2327502" submissionType="Web">
        <event type="received" date="2019-02-03"/>
        <event type="sentToCompany " date="2019-02-03"/>
        <product><productType>Credit reporting</productType></product>
        <issue><issueType>Incorrect information on credit report</issueType><subissue>Account status</subissue></issue>
        <consumerNarrative>Checked my credit report after filing complaint with CFPB on XXXX. Was
            finally able to get access to the dispute forms and the XXXX XXXX account scheduled for
            deletion XX/XX/XXXX2017 was still on record. After already registering with my report
            number, name and social security and placing the dispute in the " dispute cart '', when
            I attempted to upload as instructed, I was taken to another form which requested the
            same ( and more ) information which was already a matter of record in order to get
            access to the report in the first place. Screenshots attached. Designed to
            discourage?</consumerNarrative>
        <company><companyName>Experian Information Solutions Inc.</companyName><companyState>NY</companyState><companyZip>1002
        <response timely="yes" consumerDisputed="N"><publicResponse>Company has responded to the consumer and the CFPB and cho
                provide a public response</publicResponse><responseType>Closed with non-monetary relief</responseType></respon
    </complaint>
    <complaint id="2356421" submissionType="Web">
        <event type="received" date="2018-02-23"/>
```

Once canonicalization is successfully performed, we don’t see any external and internal differences in the dtd or the xml files. Please refer to the snapshots in the report.

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. But this isn’t good enough. I believe we need to decided a data model to store all the complaints.
- **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 –** There are no procedure defined to retrieve a complaint information. 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?

- **Organization –** The systematic organization of data in a hierarchical format because of the xml syntax helps in achieving this objective. But this isn’t good enough. I believe we need to decided a data model to store all the complaints.
- **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.
- **Access –** There are no procedure defined to retrieve a complaint information. We need to have multiple procedures defined to access a complaint information.