CS 598 – Data Curation Final Project

Sembian Balasubramanian

Sembian2@illinois.edu

Contents

Overview of CS598 Data Curation Final Project

The government agency manages complaints made by customers against financial institutions like banks and lending companies, the agency switched from old system into the new system. Assuming the role of a data scientist for the agency this project demonstrates the evaluation techniques of the quality of the data transfer ensuring equivalence of the data in the old and new systems, the project also provides an elaborate methodology in the integration and familiarity of concept of identity and canonicalizing data and implementing strategies to ensure equivalence.

Additionally, the document contains a memo to the newly appointed director of the data science department, given the new director has de-prioritized data curation work the memo addressed to the new director the importance of data curation and making the case and to re-prioritize the data curation in the data science department.

File Terminology

In this project report there are two data files containing the information of the customer complaints are referenced for comparison. The report will identify and follows a short naming convention for readability purposes - Consumser_Complaints_FileA.xml will be referred to as **File A** and Consumer_Complaints_FileB.xml will be referred to as **File B**.

Section 1 Data Files Narrative Profile and MD5 Checksum Results

Section 1.1 Narrative description of File A and File B

The File A(Consumer_Complaints_FileA.xml) is a UTF-8 encoded XML file which is denoted in the XML Version 1.0 standards with encoding="UTF-8" and consistent use of line feeds separating each line in XML , as per the instructions the File A contains information about the customer complaints managed by the government agency against various financial institutions. File A is generated from the old system maintained by the government agency. The File A is well formatted and passed validation, File A does not contain any internal DTD specifications. The File A has consumerComplaints as a root element and contains a total of 8 child elements identified with complaint as element name, each of the complaint child element there is an "id" attribute to identify the complaint and two child elements event each with attributes "date" to capture the date of the complaint and "type" as received and sendToCompany. The complaint element has a cosumerNarrative child element with the complain details in a narrative text, but not all complaints contain the consumerNarrative as there are complaints with missing consumerNarrative element. The complaint element also has product element with productType, subproduct, and issue with issueType and company with companyName, companyState, & companyZip. The Complaint element also has submitted with attributes "via" and finally the complaint has a response child element with attributes"consumerDisputed" and "timely" and another child element responseType. The analysis of File A's tree structure in a visual format is shown below. The MD5 checksum value of File A is d30cba6b00308a87fa3a384799c5faf7 also shown in Figure 1.1.a
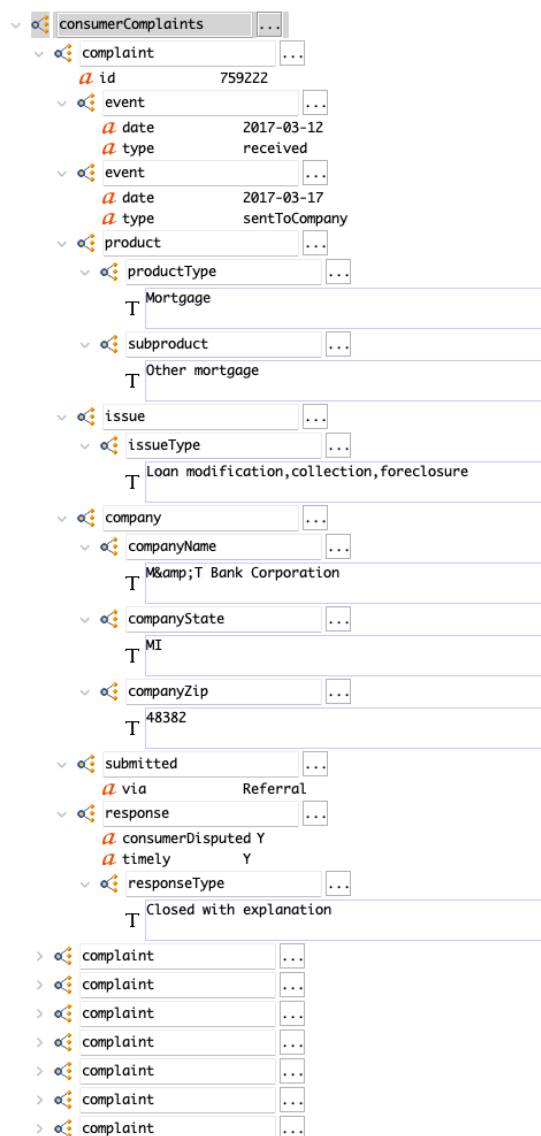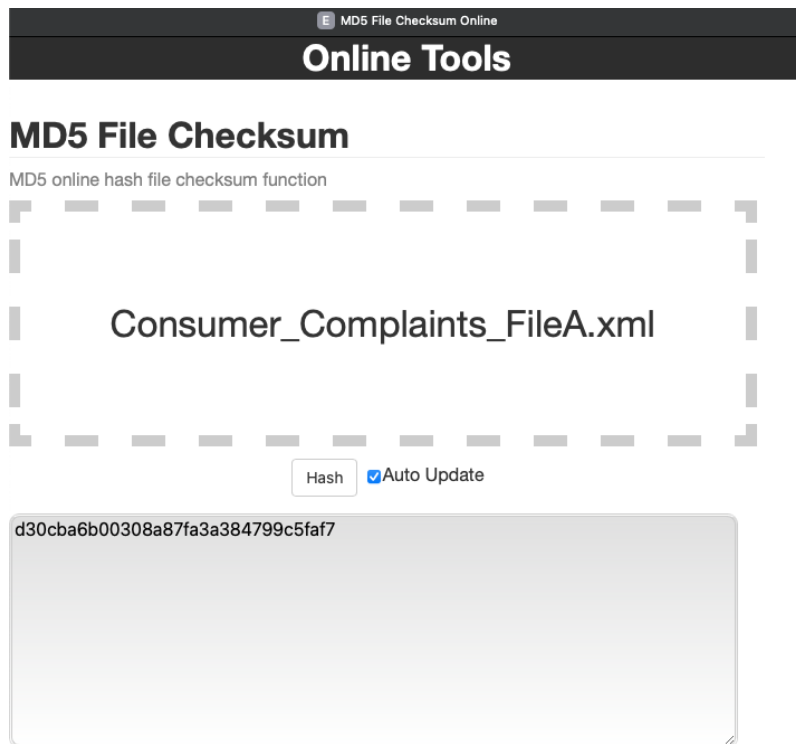
```
∨ ◦⦂ consumerComplaints        ...
  ∨ ◦⦂ complaint              ...
       𝑎 id              759222
     ∨ ◦⦂ event                ...
          𝑎 date          2017-03-12
          𝑎 type          received
     ∨ ◦⦂ event                ...
          𝑎 date          2017-03-17
          𝑎 type          sentToCompany
     ∨ ◦⦂ product              ...
       ∨ ◦⦂ productType          ...
            T  Mortgage
       ∨ ◦⦂ subproduct           ...
            T  Other mortgage
     ∨ ◦⦂ issue                ...
       ∨ ◦⦂ issueType            ...
            T  Loan modification,collection,foreclosure
     ∨ ◦⦂ company              ...
       ∨ ◦⦂ companyName          ...
            T  M&amp;T Bank Corporation
       ∨ ◦⦂ companyState         ...
            T  MI
       ∨ ◦⦂ companyZip           ...
            T  48382
     ∨ ◦⦂ submitted            ...
          𝑎 via            Referral
     ∨ ◦⦂ response             ...
          𝑎 consumerDisputed Y
          𝑎 timely         Y
       ∨ ◦⦂ responseType         ...
            T  Closed with explanation
  > ◦⦂ complaint              ...
  > ◦⦂ complaint              ...
  > ◦⦂ complaint              ...
  > ◦⦂ complaint              ...
  > ◦⦂ complaint              ...
  > ◦⦂ complaint              ...
  > ◦⦂ complaint              ...
```

Figure 1.1.A File A Tree Structure Analysis with elements and attributes

Consumer_Complaints_FileB.xml file is a UTF-8 encoded XML file used by a government agency that tracks customer complaints against financial institutions. Consumer_Complaints_FileB.xml was generated by the newly implemented system. Observed inconsistent linefeeds and XML contains in-line comments. The XML file contains a minimal DTD with an entity definition as "redaction". The file utilizes attributes holding information including identifiers, submission type, dates when the complaint was received and sent, as well as if the complaint was disputed and handled timely. The root node is similar to File A consumerComplaints and <complaints as child elements, the complaint element has two attributes "id" and "submissionType", the other elements are multiple event elements product with productType as child and an optional subProductType, issue with issueType as child

and consumerNarrative, company response and submitted, The submitted element is empty in some cases. The file has extra spaces between tags and its values. The &redaction is used on consumerNarrative element to redact values. MD5 checksum: 47677272e76e1f4332afe859347c8695

Online Tools

E MD5 File Checksum Online

MD5 File Checksum

MD5 online hash file checksum function

Consumer_Complaints_FileA.xml

Hash ☑Auto Update

d30cba6b00308a87fa3a384799c5faf7

Figure 1.1.b – MD5 CheckSum validation of Consumer_Complaints_FileA.xml from
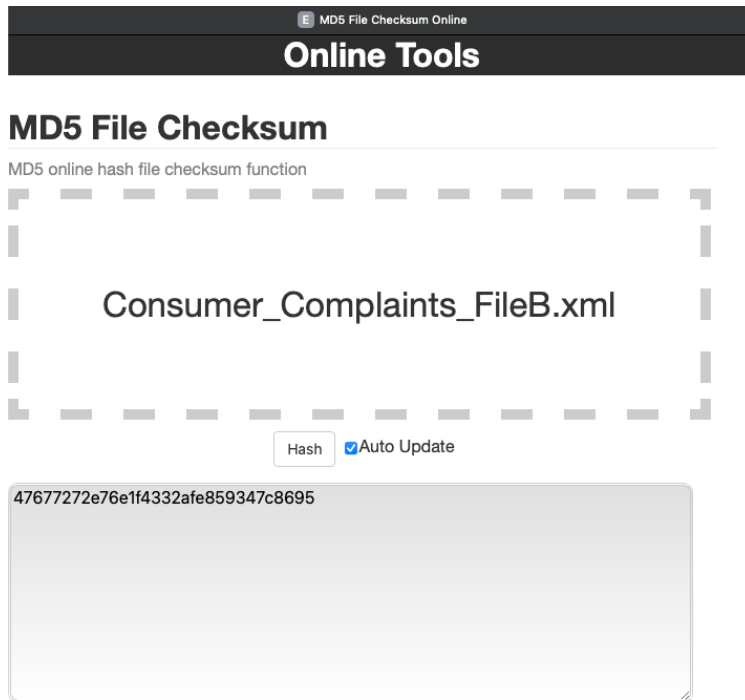
https://emn178.github.io/online-tools/md5_checksum.html

Figure 1.1.b – MD5 CheckSum validation of Consumer_Complaints_FileB.xml from

https://emn178.github.io/online-tools/md5_checksum.html

Section 1.2 Create a DTD for each XML file

1.2.1 File A XML Document DTD Implementation

Below is the embedded DTD implementation for file A. The file has been validated using the website:

http://xmlvalidator.new-studio.org/ with a screen shot of its validity found in Figure Appendix.1.2.1 File A Embedded DTD Validation. This website requires embedded DTD's but an external DTD has also been provided for completeness which was validated using the OxygenXML Editor software package.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE consumerComplaints [
<!ELEMENT consumerComplaints (complaint+)>
```

```
<!ELEMENT complaint ((event, event, product, issue, consumerNarrative?, company, submitted,response) | (company,
event, submitted, issue, product, event, response))>
<!ELEMENT event EMPTY>
<!ELEMENT product (productType,subproduct?)>
<!ELEMENT productType (#PCDATA)>
<!ELEMENT subproduct (#PCDATA)>
<!ELEMENT issue (issueType, subissue?)>
<!ELEMENT issueType (#PCDATA)>
<!ELEMENT subissue (#PCDATA)>
<!ELEMENT consumerNarrative (#PCDATA)>
<!ELEMENT company (companyName,companyState,companyZip)>
<!ELEMENT companyName (#PCDATA)>
<!ELEMENT companyState (#PCDATA)>
<!ELEMENT companyZip (#PCDATA)>
<!ELEMENT submitted EMPTY>
<!ELEMENT response (publicResponse?, responseType)>
<!ELEMENT responseType (#PCDATA)>
<!ELEMENT publicResponse (#PCDATA)>


<!ATTLIST complaint id NMTOKEN #REQUIRED>
<!ATTLIST event type (received | sentToCompany) #REQUIRED >
<!ATTLIST event date CDATA #REQUIRED>
<!ATTLIST submitted via (Phone|Referral|Web) #REQUIRED>
<!ATTLIST response timely (Y|N) #REQUIRED>
<!ATTLIST response consumerDisputed (Y|N) #REQUIRED>
]>

<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>
```

```xml
      </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">
```

```xml
<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
  {$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>
</company>
<submitted via="Web"/>
<response timely="Y" consumerDisputed="Y">
  <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
    provide a public response</publicResponse>
  <responseType>Closed with explanation</responseType>
</response>
</complaint>
<complaint id="2327502">
  <event type="received" date="2019-02-03"/>
```

```xml
    <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>10020</companyZip>
    </company>
    <submitted via="Web"/>
    <response timely="Y" consumerDisputed="N">
      <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
        provide a public response</publicResponse>
      <responseType>Closed with non-monetary relief</responseType>
    </response>
  </complaint>
  <complaint id="2356421">
    <event type="received" date="2018-02-23"/>
    <event type="sentToCompany" date="2018-02-23"/>
    <product>
      <productType>Bank account or service</productType>
      <subproduct>Savings account</subproduct>
    </product>
```

```xml
  <issue>
    <issueType>Deposits and withdrawals</issueType>
  </issue>
  <consumerNarrative>I deposited what turned out to be a fraudulent check drawn on Wells Fargo
    by mobile deposit to my savings account at Wells Fargo on XXXX at XXXX XXXX Time for
    {$2400.00}. They gave me full availability of the {$2400.00} on XXXX at which time I
    withdrew {$2200.00} and the bank then returned the deposited check on XXXX creating an
    overdraft in my account of over {$2000.00}. Wells Fargo rep explained that they do not
    process mobile deposits until late the night one day after the deposit was made. This
    means they honored the withdrawal request before they processed the transaction. That
    gave me the false assurance that the deposited check was good. The cash is gone to the
    perpetrator and now they want me to cover the overdraft. The fact they wait a whole
    business day before processing these deposits is for their convenience and the consumer
    should not be held accountable for the consequences of this delay. Also UCC 4-301 ( b )
    addresses the final payment of on-us checks deposited and states that the payor bank has
    until midnight of the next banking day to decide whether to honor the check. If they do
    n't act by midnight deadline, they lose the right to dishonor the check. 4-214 ( c ),
    4-301 ( b ). The mobile deposit confirmation states " The following mobile deposit was
    made on XXXX at XXXX Time '' and her account statement shows the deposit under the
    posting date of XXXX. Therefore, applying UCC 4-214 ( c ), the deposited check drawn on
    Wells Fargo should have been returned and charged back under the posting date of XXXX.
    It was not. The chargeback is posted under processing date of XXXX.</consumerNarrative>
  <company>
    <companyName>Wells Fargo &amp; Company</companyName>
    <companyState>AZ</companyState>
    <companyZip>85043</companyZip>
  </company>
  <submitted via="Web"/>
  <response timely="N" consumerDisputed="N">
    <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
      provide a public response</publicResponse>
    <responseType>Closed with explanation</responseType>
  </response>
</complaint>
<complaint id="2112558">
```

```xml
    <event type="received" date="2017-09-15"/>
    <event type="sentToCompany" date="2017-09-15"/>
    <product>
      <productType>Debt collection</productType>
      <subproduct>Medical</subproduct>
    </product>
    <issue>
      <issueType>Continued attempts to collect debt not owed</issueType>
      <subissue>Debt is not mine</subissue>
    </issue>
    <consumerNarrative>I am a veteran widow whom is a recipient of Maryland State Medicaid and
      have been for several years. Therefore, the State is responsible for my health bills at
      XXXX cost to me.</consumerNarrative>
    <company>
      <companyName>Round Two Recovery</companyName>
      <companyState>OK</companyState>
      <companyZip>73135</companyZip>
    </company>
    <submitted via="Web"/>
    <response timely="N" consumerDisputed="N">
      <responseType>Untimely response</responseType>
    </response>
  </complaint>
  <complaint id="837784">
    <event type="received" date="2015-05-05"/>
    <event type="sentToCompany" date="2015-05-06"/>
    <product>
      <productType>Student loan</productType>
      <subproduct>non-federal student loan</subproduct>
    </product>
    <issue>
      <issueType>Dealing with my lender or service</issueType>
      <subissue>Need information about my balance/terms</subissue>
    </issue>
    <company>
```

```
        <companyName>Navient Solutions, LLC</companyName>

        <companyState>DE</companyState>

        <companyZip>19802</companyZip>

      </company>

      <submitted via="Web"/>

      <response timely="Y" consumerDisputed="N">

        <responseType>Closed with monetary relief</responseType>

      </response>

    </complaint>

    <complaint id="14038">

      <company>

        <companyName>U.S. BANCORP</companyName>

        <companyState>AZ</companyState>

        <companyZip>85008</companyZip>

      </company>

      <event type="sentToCompany" date="2017-01-22"/>

      <submitted via="Referral"/>

      <issue>

        <issueType>Loan servicing, payments, escrow account</issueType>

      </issue>

      <product>

        <productType>Mortgage</productType>

        <subproduct>Other mortgage</subproduct>

      </product>

      <event type="received" date="2017-01-17"/>

      <response timely="Y" consumerDisputed="Y">

        <responseType>Closed without relief</responseType>

      </response>

    </complaint>

  </consumerComplaints>
```

Figure 1.2.1 File A with embedded DTD

1.2.1 .1 XML Prose for element, attribute, and attribute values for file A

The root element is consumerComplaints. The consumerComplaints element is the parent node for complaint elements.

The complaint provides information pertaining to a complaint received by the legacy system to track from customers complaints. The element complaint contains children elements including either:

- The occurrence of 2 event elements, a product element, an issue element, zero or 1 occurrences of a consumerNarrative, a company, zero or 1 occurrences of submitted, and a response element.
- A company element, an event element, a submitted element, issue element, product element, an event element, and finally a response element.

<!ELEMENT complaint ((event, event, product, issue, consumerNarrative?, company, submitted?,response) | (company, event, submitted, issue, product, event, response))

The complaint element has the fixed required attribute of an id. It was considered that the id could be of type ID, however, that is reserved for unique names which cannot start with an integer value, so a suitable alternative was to use an NMTOKEN which would at minimum ensure no whitespaces are allowed in the id value, however this does not ensure uniqueness of the ID. Placing a character in front of the value (example: id=12345 could be id=M12345) would have ensured uniqueness, this task within the project was creating a DTD for the file as is, so NMTOKEN was an appropriate attribute type.

In order to validate the xml file with a DTD it should be noted, that order does matter, which is why there are two different xml element options to define a complaint element.

event is an empty element with attributes that provide information regarding dates and context to the dates – indicating if a date is in regard to when the information was "received", or when the information was "sentToCompany". The context for the date is stored in a required attribute called type. The date is stored in a required attribute named date. Regardless of order, the event element will be found exactly twice in a complaint element. It will either be the first two items, or the second, and the sixth element found within a complaint.

product is an element providing product information for a consumer complaint. The product element occurs exactly once in the complaint element. The product element will contain exactly one productType and will include 0 or 1 occurrences of a subproduct.

issue is an element providing issue information for a given complaint. An issue element will contain exactly one child element of issueType and will include 0-1 occurrences of a subissue.

consumerNarrative is an element appears to be a narrative of details regarding the complaint. This element will occur 0-1 times in the complaint element provided the element event is the first element found within a complaint. If event is not the first element in a complaint, based on the sample data provided, it will not be included in the complaint element.

company is an element that holds data pertaining to the company that the complaint is filed against. company occurs exactly once within a complaint. The company element includes elements companyName, companyState, and companyZip which each occur exactly once within the company element.

submitted is an element indicating how a complaint was provided. submitted is an empty element which includes an attribute via which can hold the possible values of Phone, Referral or Web. submitted is an element that will occur 1 time in a complaint element.

response is an element that provides information regarding how a response was handled. response occurs exactly once within a complaint element. The response element includes a publicResponse element that will occur 0-1 times within a response as well as a responseType which will occur exactly 1 time within a complaint element. The response element includes a required attribute timely which can be either a Y or N. The response element also includes a required attribute consumerDisputed which can either have a value of Y or N.

No default values were provided to attributes since all attributes were determined to be required.

1.2.1 .2 Difficult decisions in designing DTD for file A XML

Standard DTD convention would be to claim complaint element should be defined as follows below:

```
<!ELEMENT complaint ((event+, product, issue, consumerNarrative?, company, submitted,response) |
(company, event, submitted, issue, product, event, response))
```

Figure 1.2.1.2.1

However, it was noted that there are always 2 events, so in the DTD the choice was made to explicitly call out that there are two events rather than a minimum of 1.

The last XML element, the complaint with an id of 14038 posed interesting choices to be made.

For id 14038 the element follows the convention of:

```
(company, event, submitted, issue, product, event, response)
```

Figure 1.2.1.2.2

Given order is important in the DTD to pass validation, it is concerning that this element does not contain data for a consumerNarrative so we have no indication of where it may or may not be in an element that does follow this

ordering format (referred to as ordering format #2) in other sample data. In addition, since we only have this one data element, it does appear that the submitted is required, when that may be due to the fact there is only one data element that follows this ordering format #2. It was decided to not make the submitted optional, even though it may very well be. It was also determined not to include consumerNarrative since it would be unclear where it would fit in this new ordering. The decision was made to follow the data, and this design decision would require follow-up for either more sample data, or gather requirements from the support team of system A to ensure this DTD could be applied to all data coming from System A. It is very likely that it will require alterations.

1.2.2 File B XML Document DTD Implementation

Below is the embedded DTD implementation for file B. The file has been validated using the website: http://xmlvalidator.new-studio.org/ with a screen shot of its validity found in Figure Appendix.1.2.2 File B Embedded DTD Validation

The external DTD was validated using OxygenXML Editor

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE consumerComplaints [
<!ENTITY redaction "XXXX">
<!ELEMENT consumerComplaints (complaint+)>
<!ELEMENT complaint (
        (event, event, product, issue, consumerNarrative?,  company,  submitted?, response)| (company,
event, issue, product, event, response))>


<!ELEMENT event EMPTY>
<!ELEMENT product (productType, subproduct?)>
<!ELEMENT productType (#PCDATA)>
<!ELEMENT subproduct (#PCDATA)>


<!ELEMENT issue (issueType, subissue?)>
<!ELEMENT issueType (#PCDATA)>
<!ELEMENT subissue (#PCDATA)>


<!ELEMENT company (companyName,companyState,companyZip)>
<!ELEMENT companyName (#PCDATA)>
<!ELEMENT companyState (#PCDATA)>
```

```
<!ELEMENT companyZip (#PCDATA)>

<!ELEMENT submitted EMPTY>


<!ELEMENT response (publicResponse?, responseType)>
<!ELEMENT responseType (#PCDATA)>
<!ELEMENT publicResponse (#PCDATA)>

<!ELEMENT consumerNarrative (#PCDATA)>


<!ATTLIST complaint id NMTOKEN #REQUIRED>
<!ATTLIST complaint submissionType (Phone|Referral|Web) #IMPLIED>
<!ATTLIST event type (received | sentToCompany) #REQUIRED >
<!ATTLIST event date CDATA #REQUIRED>
<!ATTLIST response timely (yes|no) #IMPLIED>
<!ATTLIST response consumerDisputed (Y|N) #REQUIRED>
]>

<consumerComplaints>
  <complaint id="759222" submissionType="Referral">
    <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></comp
any>
    <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"/>
```

```xml
    <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>
    <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 &redaction; card member for years, in late XX/XX/2016 &redaction;
converted
        the card portfolio to Barclaycard ( &redaction; ). 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 &redaction;
        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></comp
any>
    <submitted/>
    <response consumerDisputed="Y" timely="yes" ><publicResponse>Company has responded to the
consumer and the CFPB and chooses not to provide a public response</publicResponse><responseType>Closed
with explanation</responseType></response>
  </complaint>
```

```xml
<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 &redaction;. Was
    finally able to get access to the dispute forms and the &redaction; &redaction; 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>10020</companyZip></company>
  <response timely="yes" consumerDisputed="N"><publicResponse>Company has responded to the
consumer and the CFPB and chooses not to
      provide a public response</publicResponse><responseType>Closed with non-monetary
relief</responseType></response>
</complaint>
<complaint id="2356421" submissionType="Web">
  <event type="received"   date="2018-02-23"/>
  <event type="sentToCompany " date="2018-02-23"/>
  <product><productType>Bank account or service</productType> <subproduct>Savings
account</subproduct></product>
  <issue><issueType>Deposits and withdrawals</issueType></issue>
  <consumerNarrative>I deposited what turned out to be a fraudulent check drawn on Wells Fargo
    by mobile deposit to my savings account at Wells Fargo on &redaction; at &redaction; &redaction; Time
for
    {$2400.00}. They gave me full availability of the {$2400.00} on &redaction; at which time I
    withdrew {$2200.00} and the bank then returned the deposited check on &redaction; creating an
    overdraft in my account of over {$2000.00}. Wells Fargo rep explained that they do not
    process mobile deposits until late the night one day after the deposit was made. This
    means they honored the withdrawal request before they processed the transaction. That
```

gave me the false assurance that the deposited check was good. The cash is gone to the

perpetrator and now they want me to cover the overdraft. The fact they wait a whole

business day before processing these deposits is for their convenience and the consumer

should not be held accountable for the consequences of this delay. Also UCC 4-301 (b)

addresses the final payment of on-us checks deposited and states that the payor bank has

until midnight of the next banking day to decide whether to honor the check. If they do

n't act by midnight deadline, they lose the right to dishonor the check. 4-214 (c),

4-301 (b). The mobile deposit confirmation states " The following mobile deposit was

made on &redaction; at &redaction; Time '' and her account statement shows the deposit under the

posting date of &redaction;. Therefore, applying UCC 4-214 (c), the deposited check drawn on

Wells Fargo should have been returned and charged back under the posting date of &redaction;.

It was not. The chargeback is posted under processing date of &redaction;.</consumerNarrative>

`<company><companyName>Wells Fargo &`

`Company</companyName><companyState>AZ</companyState><companyZip>85043</companyZip></company>`

`<response timely="no" consumerDisputed="N"><publicResponse>Company has responded to the`

consumer and the CFPB and chooses not to

`provide a public response</publicResponse><responseType>Closed with`

`explanation</responseType></response>`

`</complaint>`

`<complaint id="2112558" submissionType="Web">`

`<event  type="received" date="2017-09-15"/>`

`<event type="sentToCompany " date="2017-09-15"/>`

`<product><productType>Debt collection</productType><subproduct>Medical</subproduct></product>`

`<issue>    <issueType>Continued attempts to collect debt not owed</issueType><subissue>Debt is not`

`mine</subissue></issue>`

`<consumerNarrative>I am a veteran widow whom is a recipient of Maryland State Medicaid and`

have been for several years. Therefore, the State is responsible for my health bills at

&redaction; cost to me.</consumerNarrative>

`<company><companyName>Round Two`

`Recovery</companyName><companyState>OK</companyState><companyZip>73135</companyZip></company>`

`<response timely="no"  consumerDisputed="N"><responseType>Untimely response</responseType>`

`</response>`

`</complaint>`

```xml
    <complaint   id="837784">
    <!-- Note: Sally modified this event on 2015-05-06 -->
    <event type="received" date="2015-05-05"/>
    <event date="2015-05-06" type="sentToCompany" />
    <product>      <productType>Student loan</productType><subproduct>non-federal student
loan</subproduct></product>
    <issue><issueType>Dealing with my lender or service</issueType><subissue>Need information about my
balance/terms</subissue></issue>
    <company><companyName>Navient Solutions, LLC</companyName>
<companyState>DE</companyState><companyZip>19802</companyZip></company>
    <response consumerDisputed="N" ><responseType>Closed with monetary relief</responseType>
    </response>
  </complaint>
  <complaint id="14038" submissionType="Referral">
    <company><companyName>U.S.
BANCORP</companyName><companyState>AZ</companyState><companyZip>85008</companyZip></compan
y>
    <event type="sentToCompany" date="2017-01-22"/>
    <issue> <issueType>Loan servicing, payments, escrow account</issueType></issue>
    <product>      <productType>Mortgage</productType><subproduct>Other
mortgage</subproduct></product>
    <event type="received"  date="2017-01-17"/>
    <response consumerDisputed="Y" > <responseType>Closed without relief</responseType></response>
  </complaint>
</consumerComplaints>
```

1.2.2.1 XML Prose for element, attribute, and attribute values for file B

The root element is consumerComplaints. The consumerComplaints element is the parent node for complaint
elements.

The complaint provides information pertaining to a complaint received by the system to process. The element
complaint contains children elements including either:

- Two occurrences of an event, a product element, and issue element, 0-1 occurrences of a consumerNarrative , a company element, 0-1 occurrences of a submitted element , and a response element
- A company element, an event element, an issue element, a product element, and event element and a response element.

<!ELEMENT complaint (

      (event, event, product, issue, consumerNarrative?, company, submitted?, response)| (company, event, issue, product, event, response))>

The complaint element has the fixed required attribute of an id. It was considered that the id could be of type ID, however, that is reserved for unique names which cannot start with an integer value, so a suitable alternative was to use an NMTOKEN which would at minimum ensure no whitespaces could be allowed in the id value, however this does not ensure uniqueness of the ID. Placing a character in front of the value (example: id=12345 could be id=M12345) would have ensured uniqueness, but we are creating a DTD for the file as is, so NMTOKEN was an appropriate attribute type.

The complaint element has an implied attribute of a submissionType which has possible values of Phone, Referral, or Web. The attribute is not in all complaint elements which is why it is an implied value.

To validate the xml file with a DTD it should be noted, that order does matter, which is why there are two different xml element options to define a complaint element.

event is an empty element with attributes that provide information regarding dates and context to the dates – indicating if a date is in regard to when the information was "received", or when the information was "sentToCompany" and "sentToCompany ", note the extra space at the end of the value. Even though this was found, it does pass validation of the DTD. The context for the date is stored in a required attribute called type. The date is stored in a required attribute named date. Regardless of order, the event element will be found exactly twice in a complaint element. It will either be the first two items, or the second, and the fifth element found within a complaint.

product is an element providing product information for a consumer complaint. The product element occurs exactly once in the complaint element. The product element will contain exactly once within the complaint element. A product element will include a productType and will include 0 or 1 occurrences of a subproduct. Product will either be the 3rd or the 4th child element found within a complaint.

issue is an element providing issue information for a given complaint. An issue element will contain exactly one child element of issueType and will include 0-1 occurrences of a subissue. issue will either be the 3rd of 4th child element found within a complaint.

consumerNarrative is an element appears to be a narrative of details regarding the complaint. This element will occur 0-1 times in the complaint element provided a complaint begins with 2 event elements.

company is an element that holds data pertaining to the company that the complaint is filed against. company occurs exactly once within a complaint. The company element includes elements companyName, companyState, and companyZip which each occur exactly once within the company element.

submitted is an empty element that will occur 0-1 times within a complaint. It is unclear what information this element is providing provided the complaint begins with an event. In the case that a complaint begins with company, submitted will not be included in the complaint.

response is an element that provides information regarding how a response was handled. response occurs exactly once within a complaint element. The response element includes a publicResponse element that will occur 0-1 times within a response as well as a responseType. A response will occur exactly 1 time within a complaint element. The response element includes an implied attribute timely which can be either a yes or no. The response element also includes a required attribute consumerDisputed which can either have a value of Y or N.

No default values were provided in the DTD due to the fact that it would be difficult to determine with a single file what the implied value should be and would result in changing the output of the XML file.

It could be noted that for id: 837784 the response attribute of timely in File A was determined to be Y, and given this value is not included in File B, it could be considered that the default value for timely should indicate a value of yes which would be required for the two files to be equivalent.

1.2.2.2 Difficult decisions in designing DTD for file B XML

<!ELEMENT complaint ((event+, product, issue, ((company, response) | (consumerNarrative, company, submitted?, response))) | (company, event, issue, product, event, response))>
<!ELEMENT complaint ((event+, product, issue, consumerNarrative?, ((company, response) | (company, submitted?, response))) | (company, event, issue, product, event, response))>
<!ELEMENT complaint ((event+, product, issue, consumerNarrative?, company, submitted?, response)| (company, event, issue, product, event, response))>

Table 1.2.2.2

Table 1.2.2.2 displays 3 additional DTD definitions for a complaint that pass validation. The first design decision that was made to modify the event+ to be event, event. This ensures that only 2 events can be defined as the DTD definition of an event+ would allow for more than 2 events, and all the complaint data elements had either 2 events, so this will help ensuring data integrity.

Section 1.3 Canonicalize the two data files and run checksums again to check for equivalence

According to Wikipedia (https://en.wikipedia.org/wiki/Canonical_XML, 2016), Canonical XML is "intended to allow relatively simple comparison of pairs of XML documents for equivalence", it further states that:, "in a special context users might care about special semantics beyond the generic logical equivalence with which Canonical XML is associated" which is the case for this project, and further it states that, "Such contexts are beyond the scope of Canonical XML."

The Canonical XML is a great starting point, through analysis of the datasets, it was found that strictly providing a canonical XML file is not be sufficient to determine if the two files can be determined to be identical. Below is the analysis indicating why strictly performing the canonical process on the 2 files will not be sufficient for determining if these 2 files are equivalent.

To evaluate the data found within the two files for logical equivalence, XML canonicalization is not sufficient, as the analysis provided below details.

Violation: According to W3C, it is required to retain all whitespace between consecutive start tags. For complaint ID: **596562**, if we strictly followed the XML canonicalization process, the files would be evaluated as not equivalent. Figure 1.3.2 shows file A has no leading spaces, and as Figure 1.3.3 shows file B has leading spaces.
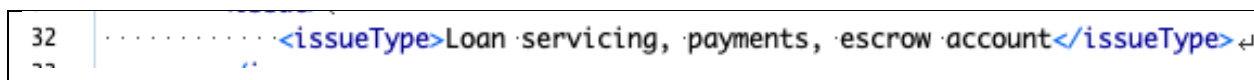

Figure 1.3.2 File A issueType whitespace between tags
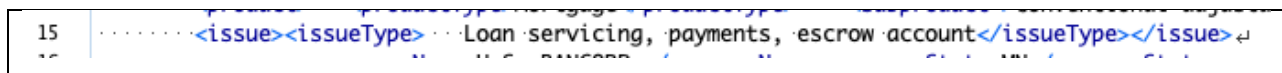
Figure 1.3.3 File B issueType whitespace between tags

Potential Violation: According to W3C, when generating a canionical XML file, a determination can be made if comments are to be saved, or not saved. It should be noted that a comment was found in File B, and not in file A. In order to canonicalize this file, it would be required to not include comments given it is in File B, and not in File A as shown in Figure 1.3.4

Figure 1.3.4 File B

Violation: Retain all whitespace between consecutive start tags. For complaint ID: **596562**, if we strictly followed the XML canonicalization process, the files would be evaluated as not equivalent. It should also be noted that this could result in poor querying of results given the data is not normalized, if a system was returning results based on the company name, this result may not be returned based on the end users query.

Figure 1.3.5 File A

Figure 1.3.6 File B showing extra whitespace.

Recall that for file B, the submissionType is implied and therefore not required (shown in Figure 1.3.9). If the DTD is updated to have default value of Web, the information found within the files can be considered equivalent, however the schema of the files makes it difficult to evaluate. According the W3C, it can be expected that an XPATH implementation with a processor can be used to assist in the canonicalization process. Instead of and generating an XSLT, a Python script using lxml library was generated to perform this task.

```
<!ATTLIST complaint submissionType (Phone|Referral|Web) #IMPLIED>
```

Figure 1.3.9 File B DTD

Violation: For publicResponse the character return is included in File A, but is not included in B.

Figure 1.3.10 File A publicReponse for 2364257

Figure 1.3.11 File B publicResponse does not contain character return

In order to facilitate the canonicalization process I have implemented a Python Script to parse the elements and apply transformation to remove attributes and consitant element naming on both File A and FileB the script is shown below.

```
from lxml import etree, objectify
import hashlib
```

```python
from io import StringIO, BytesIO
import string
# https://lxml.de/xpathxslt.html
#https://python101.pythonlibrary.org/chapter31_lxml.html
from re import sub
import re


def customSemantic(a_string):
    a_string = re.sub("\s+", " ", str(a_string))
    return a_string.strip()


if __name__ == "__main__":
    FileA = 'data/Consumer_Complaints_FileA.xml'
    FileAIntermediate = 'output/FileAIntermediate.xml'
    FileAIntermediateClean = 'output/FileAIntermediateClean.xml'
    FileBIntermediateClean = 'output/FileBIntermediateClean.xml'
    FileB = 'data/Consumer_Complaints_FileA.xml'
    GenericTransform = 'xslt/RemoveAttributesDataTransform.xslt'
    FileBIntermediate = 'output/FileBIntermediate.xml'
    FileBFinal = 'output/FileBFinal.xml'
    #parseXML(FileA)
    #Generic Transform to Remove Attributes with elements
    parser = etree.XMLParser(dtd_validation=False)
    xsl_tree = etree.parse(GenericTransform)
    transform = etree.XSLT(xsl_tree)
    #File A
    FileAtree = etree.parse(FileA)
    resultA = transform(FileAtree)
    resultA.write_output(FileAIntermediate)
    root = objectify.fromstring(etree.tostring(resultA))
    for c in root.complaint:
        submitted = c.submitted.via;
        c.submissionType = submitted;
        c.remove(c.submitted)
        if c.response.timely == "no":
```

```python
            c.response.timely = "N"
        else:
            c.response.timely = "Y"


        for e in c.event:

            if e.type == "received":
                c.receivedDate = e.date
            else :
                c.sentToCompanyDate = e.date
            c.event.remove(c.event.type)
            c.event.remove(c.event.date)
            c.remove(c.event)


objectify.deannotate(root)
etree.cleanup_namespaces(root)
objA_xml = etree.tostring(root,
                pretty_print=True,
                xml_declaration=True, encoding="utf-8")

try:
    with open(FileAIntermediateClean, "wb") as xml_writer:
        xml_writer.write(objA_xml)
except IOError:
    pass
c14n_xmlA = etree.canonicalize(root)


#File B
FileBtree = etree.parse(FileB)
resultB = transform(FileBtree)
resultB.write_output(FileBIntermediate)
rootB = objectify.fromstring(etree.tostring(resultB))
for c in rootB.complaint:
    submitted = c.submitted.via;
```

```python
        c.submissionType = submitted;
        c.remove(c.submitted)
        if c.response.timely == "no":
            c.response.timely = "N"
        else:
            c.response.timely = "Y"
        #print(c.event)


        for e in c.event:


            if e.type == "received":
                c.receivedDate = e.date



            else :
                c.sentToCompanyDate = e.date
            c.event.remove(c.event.type)
            c.event.remove(c.event.date)
            c.remove(c.event)
objectify.deannotate(rootB)
etree.cleanup_namespaces(rootB)
objB_xml = etree.tostring(rootB,
                pretty_print=True,
                xml_declaration=True, encoding="utf-8")

try:
    with open(FileBIntermediateClean, "wb") as xml_writer:
        xml_writer.write(objB_xml)
except IOError:
    pass
c14n_xmlB = etree.canonicalize(rootB)
resultAHash = hashlib.md5(c14n_xmlA.encode())
resultBHash = hashlib.md5(c14n_xmlB.encode())
print(resultAHash.hexdigest())
print(resultBHash.hexdigest())
```

Python Script used for Canonicalization process for File A and File B

Part 1.4 DTD of Final XML File

Below is the embedded DTD implementation for newly created final file (The output was identical for the transformation process). The file has been validated using the website: http://xmlvalidator.new-studio.org/ with a screen shot of its validity found in Figure Appendix.1.4.1 File A Embedded DTD Validation

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE consumerComplaints [
<!ENTITY redaction "XXXX">
<!ELEMENT consumerComplaints (complaint)+>
<!ELEMENT complaint (id, product, issue, company, response, consumerNarrative?,
submissionType,receivedDate, sentToCompanyDate)>
<!ELEMENT id (#PCDATA)>
<!ELEMENT product (productType, subproduct?)>
<!ELEMENT subproduct (#PCDATA)>
<!ELEMENT productType (#PCDATA)>
<!ELEMENT issue (issueType, subissue?)>
<!ELEMENT issueType (#PCDATA)>
<!ELEMENT subissue (#PCDATA)>
<!ELEMENT company (companyName, companyState, companyZip)>
<!ELEMENT companyName (#PCDATA)>
<!ELEMENT companyState (#PCDATA)>
<!ELEMENT companyZip (#PCDATA)>
<!ELEMENT response (timely, consumerDisputed, responseType, publicResponse?)>
<!ELEMENT timely (#PCDATA)>
<!ELEMENT consumerDisputed (#PCDATA)>
<!ELEMENT responseType (#PCDATA)>
<!ELEMENT publicResponse (#PCDATA)>
<!ELEMENT consumerNarrative (#PCDATA)>
<!ELEMENT submissionType (#PCDATA)>
<!ELEMENT sentToCompanyDate (#PCDATA)>
<!ELEMENT receivedDate (#PCDATA)>
]>
```

```xml
<consumerComplaints>
 <complaint>
  <id>759222</id>
  <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>
  <response>
   <timely>Y</timely>
   <consumerDisputed>Y</consumerDisputed>
   <responseType>Closed with explanation</responseType>
  </response>
  <submissionType>Referral</submissionType>
  <receivedDate>2017-03-12</receivedDate>
  <sentToCompanyDate>2017-03-17</sentToCompanyDate>
 </complaint>
 <complaint>
  <id>596562</id>
  <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>
```

```xml
    <companyState>MN</companyState>
    <companyZip>48322</companyZip>
   </company>
   <response>
    <timely>Y</timely>
    <consumerDisputed>N</consumerDisputed>
    <responseType>Closed with monetary relief</responseType>
   </response>
   <submissionType>Phone</submissionType>
   <receivedDate>2016-11-13</receivedDate>
   <sentToCompanyDate>2016-11-20</sentToCompanyDate>
  </complaint>
  <complaint>
   <id>2364257</id>
   <product>
    <productType>Credit card</productType>
   </product>
   <issue>
    <issueType>Other fee</issueType>
   </issue>
   <company>
    <companyName>BARCLAYS BANK DELAWARE</companyName>
    <companyState>MA</companyState>
    <companyZip>19904</companyZip>
   </company>
   <response>
    <timely>Y</timely>
    <consumerDisputed>Y</consumerDisputed>
    <responseType>Closed with explanation</responseType>
    <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
      provide a public response</publicResponse>
   </response>
   <consumerNarrative>Was a happy &redaction; card member for years, in late XX/XX/2016 &redaction;
converted
      the card portfolio to Barclaycard ( &redaction; ). 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 &redaction;

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>

 <submissionType>Web</submissionType>

 <receivedDate>2019-02-28</receivedDate>

 <sentToCompanyDate>2019-02-28</sentToCompanyDate>

</complaint>

<complaint>

 <id>2327502</id>

 <product>

  <productType>Credit reporting</productType>

 </product>

 <issue>

  <issueType>Incorrect information on credit report</issueType>

  <subissue>Account status</subissue>

 </issue>

 <company>

  <companyName>Experian Information Solutions Inc.</companyName>

  <companyState>NY</companyState>

  <companyZip>10020</companyZip>

 </company>

 <response>

  <timely>Y</timely>

  <consumerDisputed>N</consumerDisputed>

  <responseType>Closed with non-monetary relief</responseType>

  <publicResponse>Company has responded to the consumer and the CFPB and chooses not to

   provide a public response</publicResponse>

```xml
    </response>
    <consumerNarrative>Checked my credit report after filing complaint with CFPB on &redaction;. Was
      finally able to get access to the dispute forms and the &redaction; &redaction; 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>
    <submissionType>Web</submissionType>
    <receivedDate>2019-02-03</receivedDate>
    <sentToCompanyDate>2019-02-03</sentToCompanyDate>
  </complaint>
  <complaint>
    <id>2356421</id>
    <product>
      <productType>Bank account or service</productType>
      <subproduct>Savings account</subproduct>
    </product>
    <issue>
      <issueType>Deposits and withdrawals</issueType>
    </issue>


    <company>
      <companyName>Wells Fargo &amp; Company</companyName>
      <companyState>AZ</companyState>
      <companyZip>85043</companyZip>
    </company>
    <response>
      <timely>Y</timely>
      <consumerDisputed>N</consumerDisputed>
      <responseType>Closed with explanation</responseType>
      <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
        provide a public response</publicResponse>
    </response>
```

<consumerNarrative>I deposited what turned out to be a fraudulent check drawn on Wells Fargo by mobile deposit to my savings account at Wells Fargo on &redaction; at &redaction; &redaction; Time for {$2400.00}. They gave me full availability of the {$2400.00} on &redaction; at which time I withdrew {$2200.00} and the bank then returned the deposited check on &redaction; creating an overdraft in my account of over {$2000.00}. Wells Fargo rep explained that they do not process mobile deposits until late the night one day after the deposit was made. This means they honored the withdrawal request before they processed the transaction. That gave me the false assurance that the deposited check was good. The cash is gone to the perpetrator and now they want me to cover the overdraft. The fact they wait a whole business day before processing these deposits is for their convenience and the consumer should not be held accountable for the consequences of this delay. Also UCC 4-301 (b) addresses the final payment of on-us checks deposited and states that the payor bank has until midnight of the next banking day to decide whether to honor the check. If they do n't act by midnight deadline, they lose the right to dishonor the check. 4-214 (c), 4-301 (b). The mobile deposit confirmation states " The following mobile deposit was made on &redaction; at &redaction; Time '' and her account statement shows the deposit under the posting date of &redaction;. Therefore, applying UCC 4-214 (c), the deposited check drawn on Wells Fargo should have been returned and charged back under the posting date of &redaction;. It was not. The chargeback is posted under processing date of &redaction;.</consumerNarrative>
 <submissionType>Web</submissionType>
 <receivedDate>2018-02-23</receivedDate>
 <sentToCompanyDate>2018-02-23</sentToCompanyDate>
</complaint>
<complaint>
 <id>2112558</id>
 <product>
  <productType>Debt collection</productType>
  <subproduct>Medical</subproduct>
 </product>
 <issue>
  <issueType>Continued attempts to collect debt not owed</issueType>
  <subissue>Debt is not mine</subissue>
 </issue>
 <company>
  <companyName>Round Two Recovery</companyName>

```xml
    <companyState>OK</companyState>
    <companyZip>73135</companyZip>
  </company>
  <response>
    <timely>Y</timely>
    <consumerDisputed>N</consumerDisputed>
    <responseType>Untimely response</responseType>
  </response>
  <consumerNarrative>I am a veteran widow whom is a recipient of Maryland State Medicaid and
    have been for several years. Therefore, the State is responsible for my health bills at
    &redaction; cost to me.</consumerNarrative>
  <submissionType>Web</submissionType>
  <receivedDate>2017-09-15</receivedDate>
  <sentToCompanyDate>2017-09-15</sentToCompanyDate>
</complaint>
<complaint>
  <id>837784</id>
  <product>
    <productType>Student loan</productType>
    <subproduct>non-federal student loan</subproduct>
  </product>
  <issue>
    <issueType>Dealing with my lender or service</issueType>
    <subissue>Need information about my balance/terms</subissue>
  </issue>
  <company>
    <companyName>Navient Solutions, LLC</companyName>
    <companyState>DE</companyState>
    <companyZip>19802</companyZip>
  </company>
  <response>
    <timely>Y</timely>
    <consumerDisputed>N</consumerDisputed>
    <responseType>Closed with monetary relief</responseType>
  </response>
```

```
      <submissionType>Web</submissionType>

      <receivedDate>2015-05-05</receivedDate>

      <sentToCompanyDate>2015-05-06</sentToCompanyDate>

    </complaint>

    <complaint>

      <id>14038</id>

      <product>

        <productType>Mortgage</productType>

        <subproduct>Other mortgage</subproduct>

      </product>

      <issue>

        <issueType>Loan servicing, payments, escrow account</issueType>

      </issue>

      <company>

        <companyName>U.S. BANCORP</companyName>

        <companyState>AZ</companyState>

        <companyZip>85008</companyZip>

      </company>

      <response>

        <timely>Y</timely>

        <consumerDisputed>Y</consumerDisputed>

        <responseType>Closed without relief</responseType>

      </response>

      <submissionType>Referral</submissionType>

      <receivedDate>2017-01-17</receivedDate>

      <sentToCompanyDate>2017-01-22</sentToCompanyDate>

    </complaint>

  </consumerComplaints>
```

Figure 1.4.1 Final File A with embedded DTD

In addition strictly for completeness given the two files are identical, the figure below is the Final File B with embedded DTD

```
<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE consumerComplaints [

<!ENTITY redaction "XXXX">
```

```
<!ELEMENT consumerComplaints (complaint)+>

<!ELEMENT complaint (id, product, issue, company, response, consumerNarrative?,

submissionType,receivedDate, sentToCompanyDate)>

<!ELEMENT id (#PCDATA)>

<!ELEMENT product (productType, subproduct?)>

<!ELEMENT subproduct (#PCDATA)>

<!ELEMENT productType (#PCDATA)>

<!ELEMENT issue (issueType, subissue?)>

<!ELEMENT issueType (#PCDATA)>

<!ELEMENT subissue (#PCDATA)>

<!ELEMENT company (companyName, companyState, companyZip)>

<!ELEMENT companyName (#PCDATA)>

<!ELEMENT companyState (#PCDATA)>

<!ELEMENT companyZip (#PCDATA)>

<!ELEMENT response (timely, consumerDisputed, responseType, publicResponse?)>

<!ELEMENT timely (#PCDATA)>

<!ELEMENT consumerDisputed (#PCDATA)>

<!ELEMENT responseType (#PCDATA)>

<!ELEMENT publicResponse (#PCDATA)>

<!ELEMENT consumerNarrative (#PCDATA)>

<!ELEMENT submissionType (#PCDATA)>

<!ELEMENT sentToCompanyDate (#PCDATA)>

<!ELEMENT receivedDate (#PCDATA)>

]>

<consumerComplaints>

 <complaint>

  <id>759222</id>

  <product>

   <productType>Mortgage</productType>

   <subproduct>Other mortgage</subproduct>

  </product>

  <issue>

   <issueType>Loan modification,collection,foreclosure</issueType>

  </issue>

  <company>
```

```xml
    <companyName>M&amp;T Bank Corporation</companyName>
    <companyState>MI</companyState>
    <companyZip>48382</companyZip>
  </company>
  <response>
   <timely>Y</timely>
   <consumerDisputed>Y</consumerDisputed>
   <responseType>Closed with explanation</responseType>
  </response>
  <submissionType>Referral</submissionType>
  <receivedDate>2017-03-12</receivedDate>
  <sentToCompanyDate>2017-03-17</sentToCompanyDate>
 </complaint>
 <complaint>
  <id>596562</id>
  <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>
  <response>
   <timely>Y</timely>
   <consumerDisputed>N</consumerDisputed>
   <responseType>Closed with monetary relief</responseType>
  </response>
  <submissionType>Phone</submissionType>
  <receivedDate>2016-11-13</receivedDate>
  <sentToCompanyDate>2016-11-20</sentToCompanyDate>
```

```xml
  </complaint>
 <complaint>
  <id>2364257</id>
  <product>
   <productType>Credit card</productType>
  </product>
  <issue>
   <issueType>Other fee</issueType>
  </issue>
  <company>
   <companyName>BARCLAYS BANK DELAWARE</companyName>
   <companyState>MA</companyState>
   <companyZip>19904</companyZip>
  </company>
  <response>
   <timely>Y</timely>
   <consumerDisputed>Y</consumerDisputed>
   <responseType>Closed with explanation</responseType>
   <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
    provide a public response</publicResponse>
  </response>
  <consumerNarrative>Was a happy &redaction; card member for years, in late XX/XX/2016 &redaction;
converted
    the card portfolio to Barclaycard ( &redaction; ). 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 &redaction;
    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>
```

```xml
    <submissionType>Web</submissionType>
    <receivedDate>2019-02-28</receivedDate>
    <sentToCompanyDate>2019-02-28</sentToCompanyDate>
 </complaint>
 <complaint>
   <id>2327502</id>
   <product>
     <productType>Credit reporting</productType>
   </product>
   <issue>
     <issueType>Incorrect information on credit report</issueType>
     <subissue>Account status</subissue>
   </issue>
   <company>
     <companyName>Experian Information Solutions Inc.</companyName>
     <companyState>NY</companyState>
     <companyZip>10020</companyZip>
   </company>
   <response>
     <timely>Y</timely>
     <consumerDisputed>N</consumerDisputed>
     <responseType>Closed with non-monetary relief</responseType>
     <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
       provide a public response</publicResponse>
   </response>
   <consumerNarrative>Checked my credit report after filing complaint with CFPB on &redaction;. Was
     finally able to get access to the dispute forms and the &redaction; &redaction; 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>
   <submissionType>Web</submissionType>
   <receivedDate>2019-02-03</receivedDate>
```

```xml
    <sentToCompanyDate>2019-02-03</sentToCompanyDate>
</complaint>
<complaint>
 <id>2356421</id>
 <product>
   <productType>Bank account or service</productType>
   <subproduct>Savings account</subproduct>
 </product>
 <issue>
   <issueType>Deposits and withdrawals</issueType>
 </issue>

 <company>
   <companyName>Wells Fargo &amp; Company</companyName>
   <companyState>AZ</companyState>
   <companyZip>85043</companyZip>
 </company>
 <response>
   <timely>Y</timely>
   <consumerDisputed>N</consumerDisputed>
   <responseType>Closed with explanation</responseType>
   <publicResponse>Company has responded to the consumer and the CFPB and chooses not to
     provide a public response</publicResponse>
 </response>
 <consumerNarrative>I deposited what turned out to be a fraudulent check drawn on Wells Fargo
   by mobile deposit to my savings account at Wells Fargo on &redaction; at &redaction; &redaction; Time for
   {$2400.00}. They gave me full availability of the {$2400.00} on &redaction; at which time I
   withdrew {$2200.00} and the bank then returned the deposited check on &redaction; creating an
   overdraft in my account of over {$2000.00}. Wells Fargo rep explained that they do not
   process mobile deposits until late the night one day after the deposit was made. This
   means they honored the withdrawal request before they processed the transaction. That
   gave me the false assurance that the deposited check was good. The cash is gone to the
   perpetrator and now they want me to cover the overdraft. The fact they wait a whole
   business day before processing these deposits is for their convenience and the consumer
   should not be held accountable for the consequences of this delay. Also UCC 4-301 ( b )
```

addresses the final payment of on-us checks deposited and states that the payor bank has
until midnight of the next banking day to decide whether to honor the check. If they do
n't act by midnight deadline, they lose the right to dishonor the check. 4-214 (c),
4-301 (b). The mobile deposit confirmation states " The following mobile deposit was
made on &redaction; at &redaction; Time '' and her account statement shows the deposit under the
posting date of &redaction;. Therefore, applying UCC 4-214 (c), the deposited check drawn on
Wells Fargo should have been returned and charged back under the posting date of &redaction;.
It was not. The chargeback is posted under processing date of &redaction;.</consumerNarrative>
<submissionType>Web</submissionType>
<receivedDate>2018-02-23</receivedDate>
<sentToCompanyDate>2018-02-23</sentToCompanyDate>
</complaint>
<complaint>
<id>2112558</id>
<product>
<productType>Debt collection</productType>
<subproduct>Medical</subproduct>
</product>
<issue>
<issueType>Continued attempts to collect debt not owed</issueType>
<subissue>Debt is not mine</subissue>
</issue>
<company>
<companyName>Round Two Recovery</companyName>
<companyState>OK</companyState>
<companyZip>73135</companyZip>
</company>
<response>
<timely>Y</timely>
<consumerDisputed>N</consumerDisputed>
<responseType>Untimely response</responseType>
</response>
<consumerNarrative>I am a veteran widow whom is a recipient of Maryland State Medicaid and
have been for several years. Therefore, the State is responsible for my health bills at
&redaction; cost to me.</consumerNarrative>

```xml
  <submissionType>Web</submissionType>
  <receivedDate>2017-09-15</receivedDate>
  <sentToCompanyDate>2017-09-15</sentToCompanyDate>
</complaint>
<complaint>
 <id>837784</id>
 <product>
  <productType>Student loan</productType>
  <subproduct>non-federal student loan</subproduct>
 </product>
 <issue>
  <issueType>Dealing with my lender or service</issueType>
  <subissue>Need information about my balance/terms</subissue>
 </issue>
 <company>
  <companyName>Navient Solutions, LLC</companyName>
  <companyState>DE</companyState>
  <companyZip>19802</companyZip>
 </company>
 <response>
  <timely>Y</timely>
  <consumerDisputed>N</consumerDisputed>
  <responseType>Closed with monetary relief</responseType>
 </response>
 <submissionType>Web</submissionType>
 <receivedDate>2015-05-05</receivedDate>
 <sentToCompanyDate>2015-05-06</sentToCompanyDate>
</complaint>
<complaint>
 <id>14038</id>
 <product>
  <productType>Mortgage</productType>
  <subproduct>Other mortgage</subproduct>
 </product>
 <issue>
```

```
    <issueType>Loan servicing, payments, escrow account</issueType>

   </issue>

   <company>

    <companyName>U.S. BANCORP</companyName>

    <companyState>AZ</companyState>

    <companyZip>85008</companyZip>

   </company>

   <response>

    <timely>Y</timely>

    <consumerDisputed>Y</consumerDisputed>

    <responseType>Closed without relief</responseType>

   </response>

   <submissionType>Referral</submissionType>

   <receivedDate>2017-01-17</receivedDate>

   <sentToCompanyDate>2017-01-22</sentToCompanyDate>

  </complaint>

 </consumerComplaints>
```

Figure 1.4.2 Final File B with embedded DTD

1.4.2.1 XML Prose for element, attribute, and attribute values for final generated file

The root element is consumerComplaints. The consumerComplaints element is the parent node for complaint elements.

The complaint provides information pertaining to a complaint received by either system to process. The element complaint contains children elements including:

An id, product, issue, company, response, 0-1 occurrences of a consumerNarrative, a submissionType, sentToCompanyDate, receivedDate.

An id element is the identifier for a given complaint.

product is an element providing product information for a consumer complaint. The product element occurs exactly once in the complaint element. The product element will contain exactly one productType and will include 0 or 1 occurrences of a subproduct.

issue is an element providing issue information for a given complaint. An issue element will contain exactly one child element of issueType and will include 0-1 occurrences of a subissue.

company is an element that holds data pertaining to the company that the complaint is filed against. company occurs exactly once within a complaint. The company element includes elements companyName, companyState, and companyZip which each occur exactly once within the company element.

response is an element that provides information regarding how a response was handled. response occurs exactly once within a Complaint element. A response includes an element timely, an element comsumerDisputed, an element responseType, and 0-1 occurrences of a publicReponse.

consumerNarrative is an element appears to be a narrative of details regarding the complaint. This element will occur 0-1 times in the Complaint.

submissionType is an element that provides information regarding how the submission was collected. subissionType occurs exactly one time in a Complaint element. With the class implementation this value is given a default value of "Web", and this could be expanded to include an enumeration of the possible types of submission.

sentToCompanyDate is an element that is suspected to be the date the complaint information was sent from the company collecting this information to the consumer of this complaint, which maybe to the company it is against, or perhaps to another organization for analysis. This element will occur exactly one time in a Complaint element.

recievedDate is an element that is suspected to be the date the complaint information was provided to the company collecting this information from a consumer. This element will occur exactly one time in a Complaint element.

1.4.1 .2 Difficult decisions in designing DTD for final XML

In this implementation there really were no difficult decisions that were required to be made. The data was very consistent based on the canonicalization process that was implemented resulting in a consistent data representation which resulted in a simple DTD. In the final XML files the order of the elements was re-verified in order to validate the XML The &redaction entity is retained for values "XXXX" of the consumerNarrative element Value

Appendix

MD5 checksum of the final XML files shows the files are identical after the canonicalization process and both the files are adhered to the same DTD definition

MD5 File Checksum

MD5 online hash file checksum function

AC-FINAL.xml

Hash ☑Auto Update

8685d2fdf995acc82e5711e243bdff1f

MD5 File Checksum

MD5 online hash file checksum function

BC-FINAL.xml

Hash ☑Auto Update

8685d2fdf995acc82e5711e243bdff1f

Figure Appendix.1.1 MD5 Comparision of File A Filal and File B Final

XML Validator Online BETA

XML Validation

○XPath ○Schema ●DTD
Charset: UTF-8 (Unicode, worldwide)
Choose XML file to validate: Choose File AC-FINAL.xml
Due to the technical issue, embedded DTD is allowed ONLY.

Check

XPath Explorer

Run

Validation Result

Validation successful!

Source Code

```
view plain  print  ?
1.   <?xml version="1.0" encoding="UTF-8"?>
2.   <!DOCTYPE consumerComplaints [
3.   <!ELEMENT consumerComplaints (complaint)+>
4.   <!ELEMENT complaint (id, product, issue, company, response, consumerNarrative?, submissionType,receivedDate, sentToCompanyDate)>
5.   <!ELEMENT id (#PCDATA)>
6.   <!ELEMENT product (productType, subproduct?)>
7.   <!ELEMENT subproduct (#PCDATA)>
8.   <!ELEMENT productType (#PCDATA)>
9.   <!ELEMENT issue (issueType, subissue?)>
10.  <!ELEMENT issueType (#PCDATA)>
11.  <!ELEMENT subissue (#PCDATA)>
12.  <!ELEMENT company (companyName, companyState, companyZip)>
13.  <!ELEMENT companyName (#PCDATA)>
14.  <!ELEMENT companyState (#PCDATA)>
15.  <!ELEMENT companyZip (#PCDATA)>
16.  <!ELEMENT response (timely, consumerDisputed, responseType, publicResponse?)>
17.  <!ELEMENT timely (#PCDATA)>
18.  <!ELEMENT consumerDisputed (#PCDATA)>
19.  <!ELEMENT responseType (#PCDATA)>
20.  <!ELEMENT publicResponse (#PCDATA)>
21.  <!ELEMENT consumerNarrative (#PCDATA)>
22.  <!ELEMENT submissionType (#PCDATA)>
23.  <!ELEMENT sentToCompanyDate (#PCDATA)>
24.  <!ELEMENT receivedDate (#PCDATA)>
25.  ]>
26.  <consumerComplaints>
27.    <complaint>
28.      <id>759222</id>
29.      <product>
30.        <productType>Mortgage</productType>
31.        <subproduct>Other mortgage</subproduct>
32.      </product>
33.      <issue>
34.        <issueType>Loan modification,collection,foreclosure</issueType>
35.      </issue>
36.      <company>
37.        <companyName>M&amp;T Bank Corporation</companyName>
38.        <companyState>MI</companyState>
39.        <companyZip>48382</companyZip>
```

Figure Appendix.1.2.1 File A Embedded DTD Validation

XML Validator Online BETA

XML Validation

○ XPath ○ Schema ● DTD

Charset: [UTF-8 (Unicode, worldwide) ▼]

Choose XML file to validate: [Choose File] BC-FINAL.xml

Due to the technical issue, embedded DTD is allowed ONLY.

[Check]

XPath Explorer

[   ] [Run]

Validation Result

✓ Validation successful!

Source Code

view plain print ?

```
 1.  <?xml version="1.0" encoding="UTF-8"?>
 2.  <!DOCTYPE consumerComplaints [
 3.  <!ELEMENT consumerComplaints (complaint)+>
 4.  <!ELEMENT complaint (id, product, issue, company, response, consumerNarrative?, submissionType,receivedDate, sentToCompanyDate)>
 5.  <!ELEMENT id (#PCDATA)>
 6.  <!ELEMENT product (productType, subproduct?)>
 7.  <!ELEMENT subproduct (#PCDATA)>
 8.  <!ELEMENT productType (#PCDATA)>
 9.  <!ELEMENT issue (issueType, subissue?)>
10.  <!ELEMENT issueType (#PCDATA)>
11.  <!ELEMENT subissue (#PCDATA)>
12.  <!ELEMENT company (companyName, companyState, companyZip)>
13.  <!ELEMENT companyName (#PCDATA)>
14.  <!ELEMENT companyState (#PCDATA)>
15.  <!ELEMENT companyZip (#PCDATA)>
16.  <!ELEMENT response (timely, consumerDisputed, responseType, publicResponse?)>
17.  <!ELEMENT timely (#PCDATA)>
18.  <!ELEMENT consumerDisputed (#PCDATA)>
19.  <!ELEMENT responseType (#PCDATA)>
20.  <!ELEMENT publicResponse (#PCDATA)>
21.  <!ELEMENT consumerNarrative (#PCDATA)>
22.  <!ELEMENT submissionType (#PCDATA)>
23.  <!ELEMENT sentToCompanyDate (#PCDATA)>
24.  <!ELEMENT receivedDate (#PCDATA)>
25.  ]>
26.  <consumerComplaints>
27.    <complaint>
28.      <id>759222</id>
29.      <product>
30.        <productType>Mortgage</productType>
31.        <subproduct>Other mortgage</subproduct>
32.      </product>
33.      <issue>
34.        <issueType>Loan modification,collection,foreclosure</issueType>
35.      </issue>
36.      <company>
37.        <companyName>M&amp;T Bank Corporation</companyName>
38.        <companyState>MI</companyState>
39.        <companyZip>48382</companyZip>
40.      </company>
```
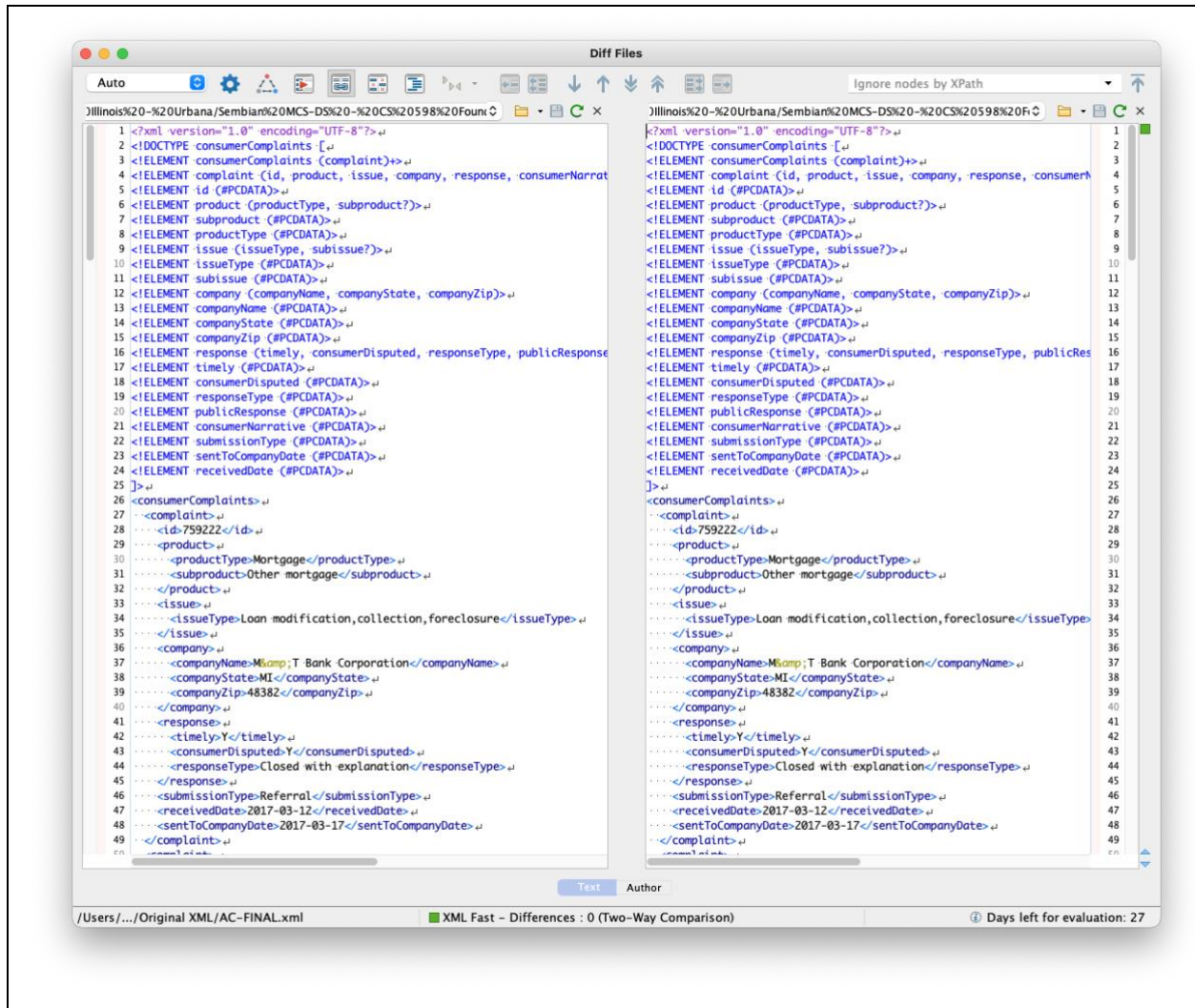
Figure Appendix.1.2.2 File B Embedded DTD Validation

Figure Appendix 1.3.1 OxygenXML File Comparison

References

W3C. (n.d.). *https://www.w3.org/TR/xml-c14n/*

https://oxygenxml.com - A XML editor tool for transformation and DTD validation

https://lxml.de/objectify.html