Sembian Balasubramanian
UIUC CS598 Foundations of Data Curation – Final Project Part 1.6

Contents

1.6. Canonicalization Process

The steps below outline the XML canonicalization process performed to confirm that the datasets for File A and for File B were identical.

Design decision: To perform the task of removing text formatting where required, I created a python script using lxml library etree. To ensure a a balanced transformation solution, an XSLT was generated to transform attributes into elements. This was then fed into a lxml eTree object, and from the etree object and normalize text as required.

In addition to enabling the use of lxml objectify this solution, according to w3schools,

“If you use attributes as containers for data, you end up with documents that are difficult to read and maintain. Try to use elements to describe data. Use attributes only to provide information that is not relevant to the data.” (XML Elements vs. Attributes, n.d.)

This made the design decision easy as arguably only the id would make sense given the information from W3Schools to be an attribute, but given the id cannot be a type of id due to the fact that it cannot start with an integer value, it was determined to enable this process as part of a generic scalable solution, the attributes would be transformed into elements. The steps following were followed as part of the canonicalization process.

Step 1: Making Attributes of the XML files elements. Given between the 2 files, what was chosen to be attributes was not consistent, a standard needed to be established to provide the information for the check sum comparison., the use of attributes provides a challenge, so an XSLT used is provided below:

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
```

```
  <xsl:output method="xml" version="1.0" encoding="UTF-8" indent="yes"/>

  <xsl:template match="@* | node()">
    <xsl:copy>
      <xsl:apply-templates select="@* | node()"/>
    </xsl:copy>
  </xsl:template>
  <xsl:template match="@*">
    <xsl:variable name="namespace">
      <xsl:choose>
        <xsl:when test="namespace-uri()">
          <xsl:value-of select="namespace-uri()"></xsl:value-of>
        </xsl:when>
        <xsl:otherwise>
          <xsl:value-of select="namespace-uri(..)" />
        </xsl:otherwise>
      </xsl:choose>
    </xsl:variable>
    <xsl:element name="{name()}" namespace="{$namespace}">
      <xsl:value-of select="." />
    </xsl:element>
  </xsl:template>
</xsl:stylesheet>
```
Figure 1.6.1.A.1 XSLT

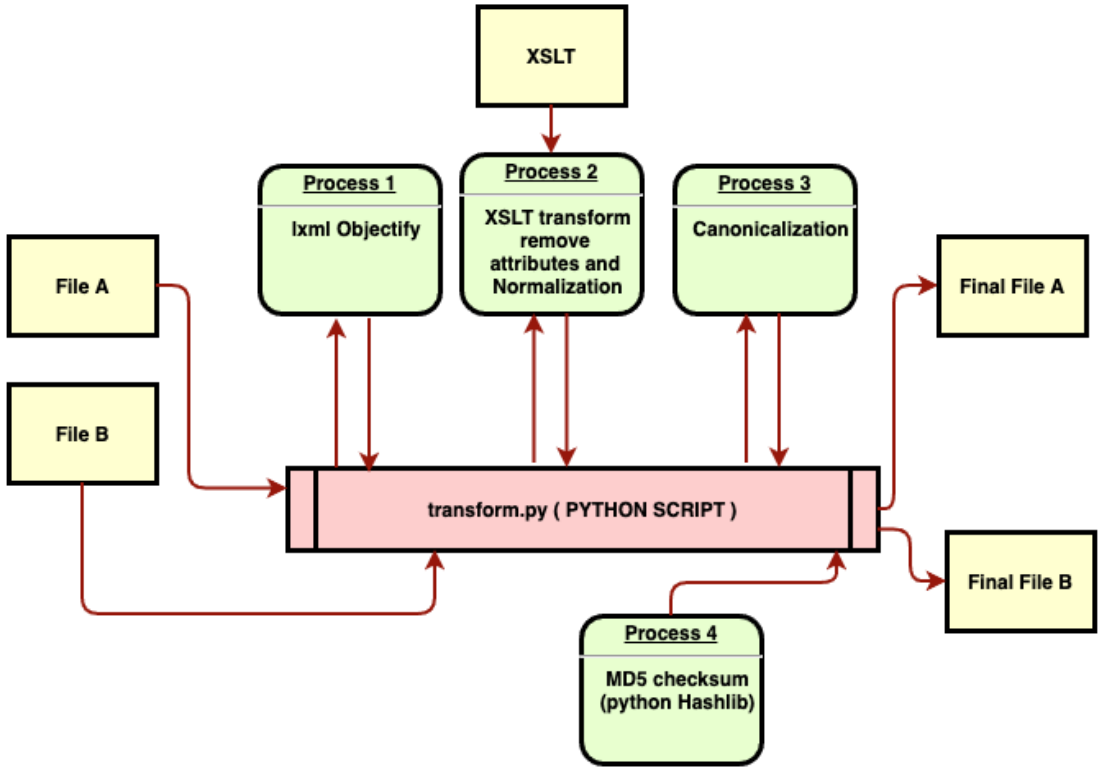


Figure 1.6.1.A.2 Step 1 XML File Transformation

```
19      FileA = 'data/Consumer_Complaints_FileA.xml'
20      FileAIntermediate = 'output/FileAIntermediate.xml'
21      FileAIntermediateClean = 'output/FileAIntermediateClean.xml'
22      FileBIntermediateClean = 'output/FileBIntermediateClean.xml'
23      FileB = 'data/Consumer_Complaints_FileA.xml'
24      GenericTransform = 'xslt/RemoveAttributesDataTransform.xslt'
25      FileBIntermediate = 'output/FileBIntermediate.xml'
26      FileBFinal = 'output/FileBFinal.xml'
27      #parseXML(FileA)
28      #Generic Transform to Remove Attributes with elements
29      parser = etree.XMLParser(dtd_validation=False)
30      xsl_tree = etree.parse(GenericTransform)
31      transform = etree.XSLT(xsl_tree)
32
33      #File A
34      FileAtree = etree.parse(FileA)
35      resultA = transform(FileAtree)
36      resultA.write_output(FileAIntermediate)
```

The output XML of this transform performed on file A can be seen below in Figure 1.6.1.A.3 and is available within this documentation at:

```
<?xml version='1.0' encoding='utf-8'?>
<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;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>
 </complaint>
 <complaint>
  <id>2364257</id>
  <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>
  <response>
   <timely>Y</timely>
   <consumerDisputed>Y</consumerDisputed>
   <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>
  <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>
<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>
<response>
<timely>Y</timely>
<consumerDisputed>N</consumerDisputed>
<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>
<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>
<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 & Company</companyName>
<companyState>AZ</companyState>
<companyZip>85043</companyZip>
</company>
<response>
<timely>Y</timely>
<consumerDisputed>N</consumerDisputed>
<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>
<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>
<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>
<response>
<timely>Y</timely>
<consumerDisputed>N</consumerDisputed>
<responseType>Untimely response</responseType>
</response>
<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>
 <company>
  <companyName>U.S. BANCORP</companyName>
  <companyState>AZ</companyState>
  <companyZip>85008</companyZip>
 </company>
 <issue>
  <issueType>Loan servicing, payments, escrow account</issueType>
 </issue>
 <product>
  <productType>Mortgage</productType>
  <subproduct>Other mortgage</subproduct>
 </product>
 <response>
  <timely>Y</timely>
  <consumerDisputed>Y</consumerDisputed>
  <responseType>Closed without relief</responseType>
 </response>
 <submissionType>Referral</submissionType>
 <sentToCompanyDate>2017-01-22</sentToCompanyDate>
 <receivedDate>2017-01-17</receivedDate>
</complaint>
```

```
</consumerComplaints>
```

Figure 1.5.1.A.3 XML of transformed output of File A

```
73      #File B
74      FileBtree = etree.parse(FileB)
75      resultB = transform(FileBtree)
76      resultB.write_output(FileBIntermediate)
```

Figure 1.6.1.A.3 File B transformation using Python lxml library and applying XSLT transformation

The output XML of this transform performed on file B can be seen below in Figure 1.6.1.A.4

```
<?xml version='1.0' encoding='utf-8'?>
<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;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>
</complaint>
<complaint>
  <id>2364257</id>
  <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>
  <response>
    <timely>Y</timely>
    <consumerDisputed>Y</consumerDisputed>
    <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>
  <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>
  <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>
  <response>
    <timely>Y</timely>
    <consumerDisputed>N</consumerDisputed>
    <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>
  <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>
  <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;amp; Company</companyName>
   <companyState>AZ</companyState>
   <companyZip>85043</companyZip>
  </company>
  <response>
   <timely>Y</timely>
   <consumerDisputed>N</consumerDisputed>
   <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>
  <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>
  <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>
  <response>
   <timely>Y</timely>
   <consumerDisputed>N</consumerDisputed>
   <responseType>Untimely response</responseType>
  </response>
  <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>
  <company>
   <companyName>U.S. BANCORP</companyName>
   <companyState>AZ</companyState>
   <companyZip>85008</companyZip>
  </company>
  <issue>
   <issueType>Loan servicing, payments, escrow account</issueType>
  </issue>
  <product>
   <productType>Mortgage</productType>
   <subproduct>Other mortgage</subproduct>
  </product>
```

```
  <response>
   <timely>Y</timely>
   <consumerDisputed>Y</consumerDisputed>
   <responseType>Closed without relief</responseType>
  </response>
  <submissionType>Referral</submissionType>
  <sentToCompanyDate>2017-01-22</sentToCompanyDate>
  <receivedDate>2017-01-17</receivedDate>
 </complaint>
</consumerComplaints>
```

1.5.1.A.4 XML of transformed output of File B

This data was used to populate the objects. The objects was auto generated using lxml objectify library and manipulated to ensure formatting of text could be handled to ensure the datasets could be properly evaluated and confirmed to be identical as shown in Figure 1.6.1.A.5 below.

```
37      root = objectify.fromstring(etree.tostring(resultA))
```

Figure 1.5.1.A.5 Python lxml objectify

1.6.1.A.1 Data Manipulations Provided on File A After the Transform

For the Transformed File A, the Objectify function was used to serialize the XML and convert to python readable objects for manipulating the elements of File A – the attributes are removed using the XSLT

Event element transform:

The event element which contains the date values indicating the received and sentToCompany normalized using the Python script and converting it to receivedDate and sendToCompanyDate

```
4  ····<event><type>received</type><date>2017-03-12</date></event>↵
5  ····<event><type>sentToCompany</type><date>2017-03-17</date></event>↵

23 ····<receivedDate>2017-03-12</receivedDate>↵
24 ····<sentToCompanyDate>2017-03-17</sentToCompanyDate>↵
```

Submitted:

The submitted element is inconsistent and also had empty and in File A the submissionType is used and on the analysis of the data the submissionType attribute is converted to submissionType element with values this is handled through the Python script

```
39      submitted = c.submitted.via;
40      c.submissionType = submitted;
41      c.remove(c.submitted)
```

1.6.1.A.2 Data Manipulations Provided on File B After the Transform

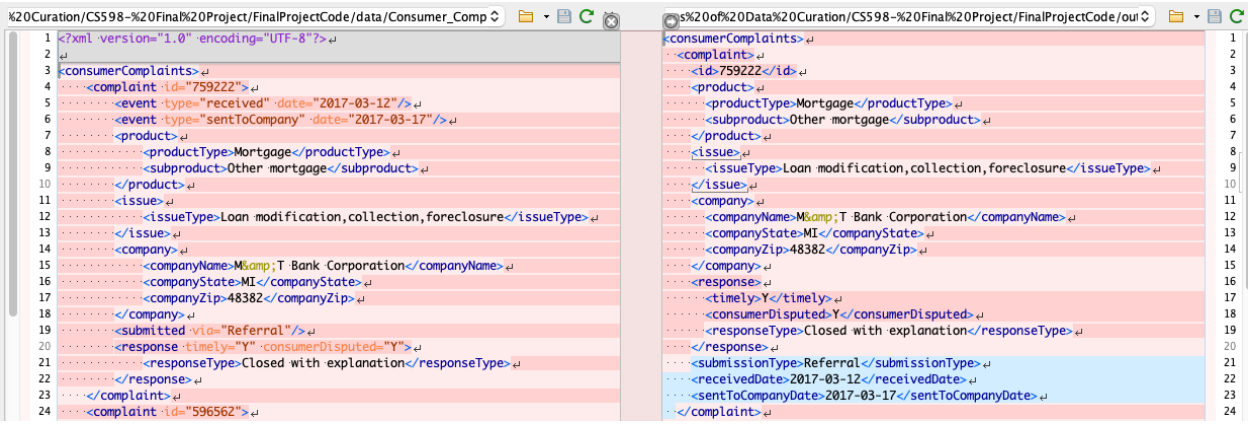
For the Transformed File B, the Objectify function was used to serialize the XML and convert to python readable objects for manipulating the elements of File B

Event element transform:

The event element which contains the date values indicating the received and sentToCompany normalized using the Python script and converting it to receivedDate and sendToCompanyDate

```
4  ····<event><type>received</type><date>2017-03-12</date></event>
5  ····<event><type>sentToCompany</type><date>2017-03-17</date></event>
23 ····<receivedDate>2017-03-12</receivedDate>
24 ····<sentToCompanyDate>2017-03-17</sentToCompanyDate>
```

1.6.1.A.3 Review of XML File Canonicalization



Below are the W3C items detailed on their website defined for establishing a canonical form of an XML document. The changes are summarized in the following list in accordance to the XML File Canonicalization steps provided by W3C (W3C, n.d.):

ID	W3C listed item in XML File Canonicalization	Applicable	Method of fulfillment
1	The document is encoded in UTF-8	Yes	This requirement is fulfilled by the XML Writer in the lxml write and passing the encoding format as UTF-8 objA_xml = etree.tostring(root, pretty_print=True, xml_declaration=True, encoding="utf-8")
2	Line breaks normalized to #xA on input, before parsing	Yes	This step is performed by lxml.canonicalize() by normalizing the line breaks
3	Attribute values are normalized, as if by a validating processor	No	This step was ignored. The attributes were actually made as elements using the transform

4	Character and parsed entity references are replaced	Yes	This was handled during the XSLT transform
5	CDATA sections are replaced with their character content	NA	No CDATA sections are in the file A and File B
6	The XML declaration and document type declaration are removed	Yes	This is requirement is fulfilled by the lxml library lxml.canonicalize() function handles the removal of the declaration
7	Empty elements are converted to start-end tag pairs	NA	While the submitted tag was empty – it was intentionally removed from the solution.
8	Whitespace outside of the document element and within start and end tags is normalized	Yes	This was handled from the lxml.canonicalize()
9	All whitespace in character content is retained (excluding characters removed during line feed normalization)	Violated	This step was intentionally violated for several attributes when a conflict occurred that would prevent the two files from matching specifically the publicResponse element had intentional linefeeds in the values.
10	Attribute value delimiters are set to quotation marks (double quotes)	NA	Attributes were removed as a design decision for easy mappings to other data modeling types and handled through XSLT transform
11	Special characters in attribute values and character content are replaced by character references	NA	Attributes were removed as a design decision for easy mappings to other data modeling types
12	Superfluous namespace declarations are removed from each element	NA	No namespace information was in original file A and File B
13	Default attributes are added to each element	NA	Attributes were removed as a design decision for easy mappings to other data modeling types and through XSLT tranform
14	Fixup of xml:base attributes [C14N-Issues] is performed	NA	Given attributes were not used in this implementation this requirement does not apply.
15	Lexicographic order is imposed on the namespace declarations and attributes of each element	Yes	The information was ordered according to the complaint id as lxml.canonicalize() handles the ordering

1.6.B Data Representation impact on reproducibility

This implementation used an XSLT to provide a generic transformation of datasets removing attributes. This approach enables reproducibility not only for the two datasets but provides an approach for evaluating any two datasets. This is a strategic solution in establishing an organized workflow solution.

This solution addresses syntax heterogeneity in that it ensures that the solution can easily move from XML into Python XML Objects and from Python XML Objects into XML.

With the removal of attributes from the dataset, this makes default integration with the Python lxml language very clean. This design choice enables an organized workflow which ultimately will support the goal of reproducibility by standardizing the parsing of xml in the python code.

If this workflow is chosen as a means of providing a federated integration solution to data scientists of the agency, and to additional systems, it would ensure that the datasets provided would be consistent and would have addressed basic data cleansing issues found between files from the legacy and existing system. This would result in reproducibility of analysis. For example if a data scientist did an analysis of data from System A, and used the same analysis on data from System B, without using a dataset that has gone through a transformation workflow, the analysis would result in varying results due to the variations in the data, thus this solution enables not only reproducibility of datasets, but also reproducibility of the analysis done on the data itself.

1.6.C Canonicalization Support overarching goals of data curation

This approach supports the concept of organization by employing a logical data model that will allow for exchanging data models from XML to Python lxml Objects and from Python lxml Objects to XML. A standard data model has been deployed using Python script and XSLT that removes attributes and creates elements for the given XML file.

The XML file was converted to Python objects using lxml objectify class definition was generated. A class could also have been generated from the XSD, which could have been generated from the initial XML files.

This solution also supports the data curation activity of identification, supporting the activity of validating data regardless of if it originated from Old System A or from New System B.

This transform solution also supports integration. If the legacy system was terminated, the data could be moved from the legacy system either into the current system by deploying a derivation approach where the data from both systems could be preserved in this final data model representation. Alternatively, a federated approach could be used where the data from both systems would be made available using the data model presented in this project.

This data model supports the activity of reformatting, as it enabled using tools that support XML as a data source and the proper formatting is performed.

This solution also supports the activity of reproducibility. Regardless of the source system, the data can be reproduced in this logical data model continuously and at scale using the solution implemented through the Python transform.py code.

This workflow also supports modification, as it modifies the data and normalizes to ensure a consistent representation of the data.

1.6.D Additional Curational Activities to support future discovery and use

The XML file from the new system should questioned. The final dataset that is to be stored with reliable and effective storage should enable data exploration. This final dataset that is generated should be preserved to ensure it can be for analysis in the future.

As datasets are updated, currently in the new system – comments are added to the document to capture its processing, this information could be considered meta data, and for the purposes of data provenance this information should be stored and noted, but not in the dataset itself. As meta data is captured, it's source system should also be preserved for data provenance considerations.

Discoverability should be considered as this information is collected by a government agency, it should be made searchable to ensure its re-use.

Communication of this data should certainly be considered as it is being shared with some entity given the element "sentToCompany".

The &redacted ENTITY supports security and Processing required by the Application generating the data to redact personal information from the data

Considerations should be made regarding security of the XML datasets. The question should be answered, who should have access to these datasets, and how are they shared with companies.

Compliance considerations should also be made. While this sample data did not contain customer information, if a customer included personal information, that would need to be retracted potentially based on who the dataset was being shared with.