Megan Masanz
CS598 – Final Project

Contents

1.5.A Canonicalization Process

The steps below outline the canonicalization process that was under taken to enable in the determination of confirming the datasets for file A and for file B were identical.

An initial solution design decision: To perform the task of removing text formatting where required, the decision was made to use an auto-generated class. C# can generate an XSD from an XML file, and that XSD can be used to generate a class, however the class that is auto-generated is quite complex. An alternative method was used, which is using a class that is generated from json. This required the xml to be transformed into json which is a trivial task. This json is then used to populate the class that is used to remove text formatting as required for elements in the XML file. json2csharp (http://json2csharp.com/, n.d.) will generate a class, but attributes in the data appear as a special character "@" which is not a valid variable name. To ensure a clean solution, an XSLT was generated to transform attributes into elements. This was then fed into a json object, and from the json object a class was generated and used in this solution to normalize text as required.

In addition to enabling the use of auto-generated classes for 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. As I have chose to use a class definition autogenerated from json2csharp (http://json2csharp.com/, n.d.), 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 1.5.1.A.1 XSLT

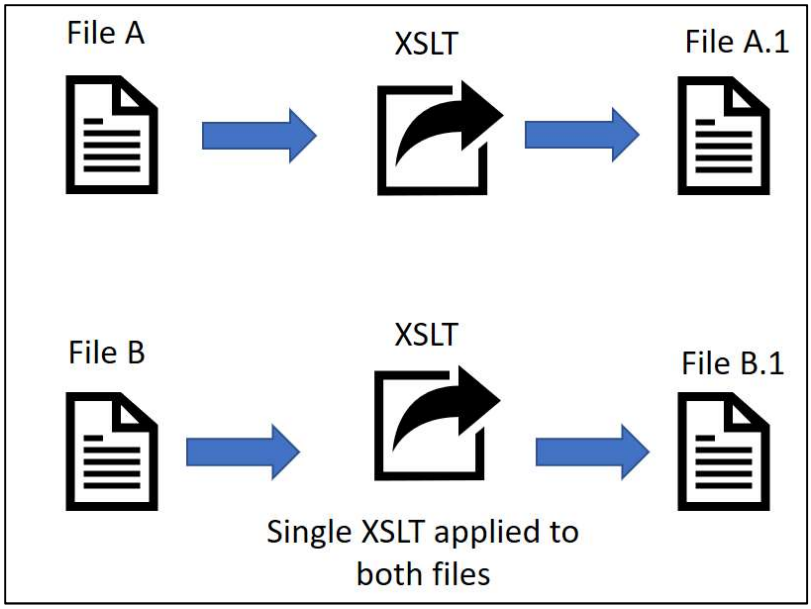

Figure 1.5.1.A.2 Step 1

The XSLT is available within this project solution at:
\Intermittent_TranformedFiles\DataTransform.xslt

The output XML of this transform performed on file A can be seen below in Figure 1.5.1.A.3 and is available within this documentation at:
\Intermittent_TranformedFiles\FileATransformed.xml

```
<?xml version="1.0" encoding="utf-8"?><consumerComplaints>
  <complaint><id>759222</id>
    <event><type>received</type><date>2014-03-12</date></event>
    <event><type>sentToCompany</type><date>2014-03-17</date></event>
    <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>
    <submitted><via>Referral</via></submitted>
    <response><timely>Y</timely><consumerDisputed>Y</consumerDisputed>
      <responseType>Closed with explanation</responseType>
    </response>
```

```
  </complaint>
  <complaint><id>596562</id>
    <event><type>received</type><date>2013-11-13</date></event>
    <event><type>sentToCompany</type><date>2013-11-20</date></event>
    <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</via></submitted>
    <response><timely>Y</timely><consumerDisputed>N</consumerDisputed>
      <responseType>Closed with monetary relief</responseType>
    </response>
  </complaint>
  <complaint><id>2364257</id>
    <event><type>received</type><date>2017-02-28</date></event>
    <event><type>sentToCompany</type><date>2017-02-28</date></event>
    <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</via></submitted>
    <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>
  </complaint>
  <complaint><id>2327502</id>
    <event><type>received</type><date>2017-02-03</date></event>
    <event><type>sentToCompany</type><date>2017-02-03</date></event>
    <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</via></submitted>
    <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>
  </complaint>
  <complaint><id>2356421</id>
    <event><type>received</type><date>2018-02-23</date></event>
    <event><type>sentToCompany</type><date>2018-02-23</date></event>
    <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>
    <submitted><via>Web</via></submitted>
    <response><timely>N</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>
  </complaint>
  <complaint><id>2112558</id>
    <event><type>received</type><date>2016-09-15</date></event>
    <event><type>sentToCompany</type><date>2016-09-15</date></event>
    <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</via></submitted>
    <response><timely>N</timely><consumerDisputed>N</consumerDisputed>
      <responseType>Untimely response</responseType>
    </response>
  </complaint>
  <complaint><id>837784</id>
    <event><type>received</type><date>2014-05-05</date></event>
    <event><type>sentToCompany</type><date>2014-05-06</date></event>
    <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</via></submitted>
    <response><timely>Y</timely><consumerDisputed>N</consumerDisputed>
      <responseType>Closed with monetary relief</responseType>
    </response>
  </complaint>
  <complaint><id>14038</id>
    <company>
      <companyName>U.S. BANCORP</companyName>
      <companyState>AZ</companyState>
      <companyZip>85008</companyZip>
    </company>
    <event><type>sentToCompany</type><date>2012-01-22</date></event>
    <submitted><via>Referral</via></submitted>
    <issue>
      <issueType>Loan servicing, payments, escrow account</issueType>
    </issue>
    <product>
      <productType>Mortgage</productType>
      <subproduct>Other mortgage</subproduct>
    </product>
    <event><type>received</type><date>2012-01-17</date></event>
    <response><timely>Y</timely><consumerDisputed>Y</consumerDisputed>
      <responseType>Closed without relief</responseType>
    </response>
  </complaint>
```

```
</consumerComplaints>
```

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

The output XML of this transform performed on file B can be seen below in Figure 1.5.1.A.4 and is available at:
\Intermittent_TranformedFiles\FileBTranformed.xml

```
<?xml version="1.0" encoding="utf-8"?><consumerComplaints>
   <complaint><id>759222</id><submissionType>Referral</submissionType>
      <event><type>received</type><date>2014-03-12</date></event>
<event><type>sentToCompany</type><date>2014-03-
17</date></event><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</companyZi
p></company>

<response><timely>yes</timely><consumerDisputed>Y</consumerDisputed><responseType>Closed
with explanation</responseType>  </response>
   </complaint>
   <complaint><id>596562</id><submissionType>Phone</submissionType>
      <event><date>2013-11-13</date><type>received</type></event>
      <event><type>sentToCompany</type><date>2013-11-20</date></event>
      <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></comp
any>

<response><consumerDisputed>N</consumerDisputed><timely>yes</timely><responseType>Closed
with monetary relief</responseType></response>
   </complaint>
   <complaint><id>2364257</id>
      <event><date>2017-02-28</date><type>received</type></event>
      <event><type>sentToCompany</type><date>2017-02-28</date></event>
      <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 />

<response><consumerDisputed>Y</consumerDisputed><timely>yes</timely><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</id><submissionType>Web</submissionType>
<event><type>received</type><date>2017-02-03</date></event>
<event><type>sentToCompany </type><date>2017-02-03</date></event>
<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>yes</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>
</complaint>
<complaint><id>2356421</id><submissionType>Web</submissionType>
<event><type>received</type><date>2018-02-23</date></event>
<event><type>sentToCompany </type><date>2018-02-23</date></event>
<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>no</timely><consumerDisputed>N</consumerDisputed><publicResponse>Comp
any 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</id><submissionType>Web</submissionType>
      <event><type>received</type><date>2016-09-15</date></event>
      <event><type>sentToCompany </type><date>2016-09-15</date></event>
      <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>no</timely><consumerDisputed>N</consumerDisputed><responseType>Untime
ly response</responseType>
      </response>
   </complaint>
   <complaint><id>837784</id>
      <!-- Note: Sally modified this event on 2014-05-06 -->
      <event><type>received</type><date>2014-05-05</date></event>
      <event><date>2014-05-06</date><type>sentToCompany</type></event>
      <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</consumerDisputed><responseType>Closed with monetary
relief</responseType>
      </response>
   </complaint>
   <complaint><id>14038</id><submissionType>Referral</submissionType>
      <company><companyName>U.S.
BANCORP</companyName><companyState>AZ</companyState><companyZip>85008</companyZip>
</company>
      <event><type>sentToCompany</type><date>2012-01-22</date></event>
      <issue> <issueType>Loan servicing, payments, escrow account</issueType></issue>
      <product>      <productType>Mortgage</productType><subproduct>Other
mortgage</subproduct></product>
      <event><type>received</type><date>2012-01-17</date></event>
      <response><consumerDisputed>Y</consumerDisputed> <responseType>Closed without
relief</responseType></response>
   </complaint>
</consumerComplaints>
```

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

This data was used to populate the classes. The class was auto generated using json2charp website. (http://json2csharp.com/, n.d.) 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 1.5.1.A.5 below.

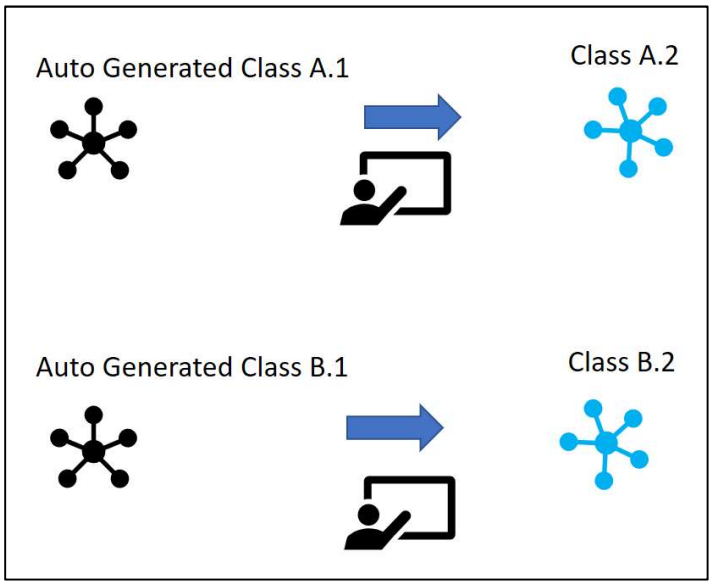

Figure 1.5.1.A.5

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

For the Transformed File A, the only modification that was required was to handle character returns, sort by the id and trim spaces. The figure below captures resolving the trimming of spaces required for file A specifically for the publicResponse. The trimming of white spaces could have been applied to all attributes, but the choice was made to make limited modifications to ensure equality of the files. This will enable a discussion regarding the specific elements that required modifications for analysis tracking. Perhaps there would be an occurrence where an extra space would intentionally be used as a distinguishing component for a given element.

```
public string publicResponse
{
    get { return ppublicResponse; }
    set
    {
        ppublicResponse = Regex.Replace(value, @"\s+", " ");
    }
}
```

Figure 1.5.1.A.1.1 Regex to handle to trim spaces

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

For the Transformed File B, a few modifications were required to the class to handle formatting the data. By modifying the autogenerated class to trim the value of an Event Type, we can handle the spaces which would result in a null value for the sentToCompanyDate element.

```
2 references
public class Event
{
    //public string type { get; set; }

    private string ptype;
    2 references
    public string type
    {
        get { return ptype; }
        set { ptype = value.Trim(); }
    }
    2 references
    public string date { get; set; }
}
```

Figure 1.5.1.A.2.1 Trim

By modifying the issueType to trim the value, we resolve the issue shown in Figure 1.3.2 and Figure 1.3.3 which was dealing with an issue seen for complaint id 596562

```
1 reference
public class Issue
{
    //public string issueType { get; set; }

    private string pissueType;
    0 references
    public string issueType
    {
        get { return pissueType; }
        set { pissueType = value.Trim(); }
    }
    0 references
    public string subIssue { get; set; }
}
```

Figure 1.5.1.A.2.2 Trim

```
1 reference
public class Company
{
    //public string companyName { get; set; }
    private string pcomanyName;
    0 references
    public string companyName
    {
        get { return pcomanyName; }
        set { pcomanyName = value.Trim(); }
    }
    0 references
    public string companyState { get; set; }
    0 references
    public string companyZip { get; set; }
}
```

Figure 1.5.1.A.2.3

By modifying the Company class to trim the input, we resolve the issue shown in Figure 1.3.5 and Figure 1.3.6 in which a given company was stored with and without an extra space.

```
public string publicResponse
{
    get { return ppublicResponse; }
    set
    {
        ppublicResponse = Regex.Replace(Regex.Replace(value, @"\r\n", ""), @"\s+", " ");
    }
}
```

Figure 1.5.1.A.2.4

For File B, we did need to replace both \r\n as well as whitespaces, so the two datasets could be shown to be equal this modification is shown above in Figure Figure 1.5.1.A.2.4.

```
0 references
public Complaint()
{
    submissionType = "Web";
}
```

Figure 1.5.1.A.2.5

For File B, we also needed to provide a default value of “web” to ensure the two datasets can be determined to be identical. The need for this medication can be seen in Figure 1.3.7 and Figure 1.3.8.

Once the class was generated, it needed to be ordered as shown in the figure below

```
//clean JsonA file
var myclassA = Newtonsoft.Json.JsonConvert.DeserializeObject<FinalProjectA.RootObject>(jsonA);
myclassA.consumerComplaints.complaint = myclassA.consumerComplaints.complaint.OrderBy(c => c.id).ToList();

//clean JsonB file
var myclassB = Newtonsoft.Json.JsonConvert.DeserializeObject<FinalProjectB.RootObject>(jsonB);
myclassB.consumerComplaints.complaint = myclassB.consumerComplaints.complaint.OrderBy(c => c.id).ToList();
```

Figure 1.5.1.A.2.6

After the data was sorted, the XML could be generated. In generating the XML file and using the XMLWriterSettings we were able to ensure proper indentation, omitting the XML declaration, using UTF8 encoding and using entities to handle new lines in the text found in a given element. Removing the DTD was already done through the XSLT process.

```
XmlWriterSettings settings = new XmlWriterSettings
{
    Indent = true,
    NewLineHandling = NewLineHandling.Entitize,
    OmitXmlDeclaration = true,
    Encoding = Encoding.UTF8

};
```

Figure 1.5.1.A.2.7

Using the .NET framework, \r\n was replaced with
\n (XmlWriterSettings.NewLineHandling Property, n.d.), which does follow the canonicalization process for

“Encoding of special characters as character references in text (&, <, >,
).” (W3C, n.d.),

outlined in section “3.4 Character Modifications and Character References” of the Canonical XML Version 1.1 provided by w3c (W3C, n.d.).

The final piece in the established workflow for the dataset conversion is depicted in the diagram below which is confirming the MD5 checksum for equivalence.

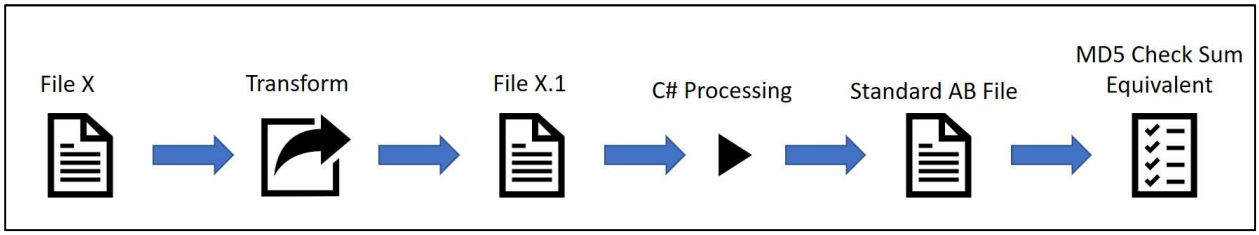

Figure 1.5.1.A.2.8

1.5.1.A.3 Reporting of MD5 checksum

Figure 1.5.1.A.3.1

As can be seen in Figure 1.3.2.10, the MD5 checksum is reported as a4e90dc93f14f2d9eda34c36b0c51f68 for the two files generated, as was confirmed with WinMerge showing the results of the this process resulted in two identical files shown below in Figure Figure 1.5.1.A.3.2.

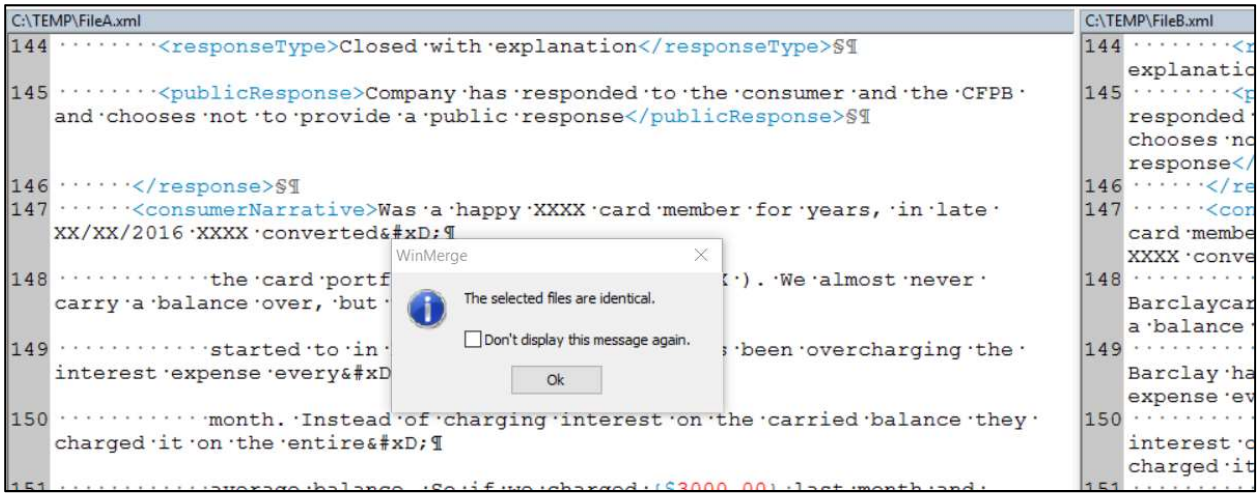

Figure 1.5.1.A.3.2

1.5.1.A.4 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 .NET framework captured in Figure 1.3.1.2.8 (XmlWriterSettings.NewLineHandling Property, n.d.)
2	Line breaks normalized to #xA on input, **before** parsing	No	This step is skipped intentionally. This standard was last updated approximately 10 years. The .NET framework can handle fulfilling the requirements requirement #8.
3	Attribute values are normalized, as if by a validating processor	No	This step was ignored. The attributes were actually made as elements so the class definition would be light.
4	Character and parsed entity references are replaced	Yes	This was handled during the transform
5	CDATA sections are replaced with their character content	NA	No CDATA sections are in the file
6	The XML declaration and document type declaration are removed	Yes	This is requirement is fulfilled by the XML Writter
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 XMLWriter
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 As captured in Figures: 1.3.1.2.1 - 1.3.1.2.5
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
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 files
13	Default attributes are added to each element	NA	Attributes were removed as a design decision for easy mappings to other data modeling types
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 shown in Figure 1.3.1.2.7

1.5.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 implementation included an additional nesting of complaints to support a class generated from a json file. This solution addresses syntax heterogeneity in that it ensures that the solution can easily move from XML into JSON and from JSON into XML. While JSON does support the use of the "@" symbol, it does cause issues as has been noted by w3school, and represented in this implementation that the use of the XML attributes, "attributes are more difficult to manipulate by program code" (XML Elements vs. Attributes, n.d.)

Within a well-designed workflow, the objective of programming language independence is ideal. With the removal of attributes from the dataset, this makes default integration with the c# language very clean. With c# can support attributes without issue, the implementation of classes and variable names makes attribute usage a challenge. This design choice enables an organized workflow which ultimately will support the goal of reproducibility.

If this workflow is chosen as a means of providing a federated integration solution to data scientists, 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.

1.5.C 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 JSON and from JSON to XML. A standard data model has been deployed that removes attributes and creates elements for the given XML file.

The XML file was converted to Json and using http://json2csharp.com/#, a class definition was generated. A class could also have been generated from the XSD, which could have been generated from the initial XML files. It was determined that flexibility to move between data representations would support the goal of data curation in a unique way that would not be supported through class generation from an XSD.

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

This 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 either JSON or XML as a data source.

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 code.

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

1.5.D Additional Activities to support discovery and re-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.

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

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.

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.

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