**FileA.xml description**

**MD5 checksum D30CBA6B00308A87FA3A384799C5FAF7**

This file seems to represent the data for ConsumerComplaints. The parent tag is consumerComplaints which contains multiple complaints pertaining data to multiple complaints.

Each complaint has an id associated with it as an attribute that is unique to every element. Each complaint has events,product,issue, consumerNarrative(which is optional),company,submitted and response. Some of the id don’t follow this exact order of tags.

Event has a type and date associated with it.

Product contains detail about the product and have two child tags productType and subproduct(optional)

Issue contains issueType and may contain subissue

Optionally the complaint may contain consumerNarrative in cases of some complaint.

This is followed by company details containing companyName , companyState and companyZip.

Then we have a submitted element that has attribute via having the details on how the complaint was submitted.

Lastly we have response which indicates timely and consumerDisputed or not as attributes. The response contains a publicResponse(optional) and responseType as well.

**FileB.xml description**

**MD5 checksum D30CBA6B00308A87FA3A384799C5FAF7**

This file also has data related to consumer complaints. The overall structure of this document is very much similar to FileA.xml but there are few differences. Unlike FileA this File has irregular line breaks.

Each complaint has an id as well as submissionType as attributes. Unlike FileA where submitted was an element we have a submissionType attribute. Each complaint has events,product,issue,consumerNarrative(in some cases),company and response. More or less these elements are similar to those present in FileA.xml.

Event has a type and date associated with it.

Product contains detail about the product and have two child tags productType and subproduct(optional)

Issue contains issueType and may contain subissue

Optionally the complaint may contain consumerNarrative in cases of some complaint.

This is followed by company details containing companyName , companyState and companyZip.

Lastly we have response which indicates timely and consumerDisputed or not as attributes. The attributes value are different from that of FileA.xmlThe response contains a publicResponse(optional) and responseType as well.

There is a tag <submitted/> present in the file which seems to be erroneous.

**Steps for canonicalization**

Before we move forward to canonicalization process , we need to make following changes to the documents

1. FileA
   1. We will be removing element submitted that is present inside a complaint and we will instead add an attribute to complaint similar to FileB which is submissionType. The reason for doing is because the element submitted doesn’t have any data and also it makes more sense to have it as an attribute because there will be limited number of values for it.
   2. Replace values Y as yes and N as no. The reason for this is because Yes and No are more readable and don’t introduce any ambiguity.
2. FileB
   1. We will fill the missing values of submissionType attribute for some of the complaint ids by taking these values from FileA
   2. We will remove the tag </submitted> from our file as it has no matching tag and seems to be erroneous.
   3. Replace values Y as yes and N as no. The reason for this is because yes and no are more readable and don’t introduce any ambiguity.
   4. Remove internal DTD and remove &redaction; and replace it with XXXX wherever required.
   5. Add missing values for timely attribute from FileA for two ids

**Above changes can be seen with the github commit** [**https://github.com/elnin09/datacuration/commit/834775f034040fa0e47553ee97a0543397a94c79#diff-985959785319747668373cc6dee294b11db782b03cdd90a2851fbdc0637c6b7b**](https://github.com/elnin09/datacuration/commit/834775f034040fa0e47553ee97a0543397a94c79#diff-985959785319747668373cc6dee294b11db782b03cdd90a2851fbdc0637c6b7b)

1. **Convert to a single character encoding and normalize line ends.**
   1. Both files have UTF-8 encoding.
   2. For normalizing line ends in VSCode I used <https://marketplace.visualstudio.com/items?itemName=sohamkamani.code-eol> extension which helped in normalizing line ends
2. **Remove all comments, tabs, non-significant spaces, etc.**
   1. Removed all the non-significant spaces from both the files and tabs. FileA has
   2. For comments there were only one comment in FileB that was removed

<!-- Note: Sally modified this event on 2015-05-06 -->

1. **Propagate all attribute defaults indicated in the schema to the elements themselves**
   1. There were no attributes present in any of the files with default values so we don’t need to change anything in the files for this issue.
2. **Put attribute/value pairs on elements in alpha order**
   1. For element event in both the files, put date attribute before type attribute