University of Illinois, Urbana-Champaign
Department of Computer Science

CS 598: Data Curation

Project Report

Anantharaman Janakiraman – aj4@illinois.edu

Data Profile:

Consumer_Complaints_FileA.xml

The file is a collection of complaints submitted by consumers for different companies through various modes of submission (Web, Referral or Phone). The XML is organized as a collection of *complaint* elements under a single root element *consumerComplaints.* Each *complaint* element has multiple sub elements or child elements that are associated with the complaint. Each complaint is identified by unique *id* that is an attribute within the *complaint* element. Each *complaint* element has an *event* element with two attributes *type* & *date* that represent the date when a complaint was received and sent to a company. The *company* element has three sub elements (*companyName, companyState, companyZip*) that represents information of a company for which a complaint is submitted. The *product* element and its two sub elements (*productType, subproduct*) hold information about a specific consumer function for which a complaint is submitted. The *issue* element provides information on what type of issue was encountered by the consumer (*issueType* sub-element) and additional information on the issue (*subissue* sub-element). *consumerNarrative* was explanation provided by the consumer for raising the complaint or dispute. The *response* element holds information about response or resolution to complaint from the company and the two sub-elements – *timely, consumerDisputed* are response attributes that indicate (Y or N) if the response was timely and if there was any further dispute on the response from consumer. The *publicResponse* sub-element holds the public response narrative. The mandatory elements & function of each element/attribute is also detailed in the DTD table documentation below. The data file has minimal inconsistencies and needs canonicalization & might require element/attribute reorganization to enable integration with the other file.

MD5: abc4665edea350247bdcacb90dfb9b04
SHA1: 56ea7d9a27b7f555b493bdb73c39018e6d5c6792

Consumer_Complaints_FileB.xml

The file is a collection of complaints submitted by consumers for different companies through various modes of submission (Web, Referral or Phone). The XML is organized as a collection of *complaint* elements under a single root element *consumerComplaints.* Each *complaint* element has multiple sub elements or child elements that are associated with the complaint. Each complaint is identified by unique *id* that is an attribute within the *complaint* element which also has another attribute *submissionType* that identifies one of the three modes of submission (Web, Phone or Referral). Each *complaint* element has an *event* element with two attributes *type* & *date* that represent the date when a complaint was received and sent to a company. The *company* element has three sub elements (*companyName, companyState, companyZip*) that represents information of a company for which a complaint is submitted. The *product* element and its two sub elements (*productType, subproduct*) hold information about a specific consumer function for which a complaint is submitted. The *issue* element provides information on what type of issue was encountered by the consumer (*issueType* sub-element) and additional information on the issue (*subissue* sub-element). *consumerNarrative* was explanation provided by the consumer for raising the complaint or dispute. The *response* element holds information about response or resolution to complaint from the company and the two sub-elements – *timely, consumerDisputed* are response attributes that indicate (Y or N) if the response was timely and if there was any further dispute on the response from consumer. The *publicResponse* sub-element holds the public response narrative. The mandatory elements & function of each element/attribute is also detailed in the DTD table documentation below. The data file has several inconsistencies and needs extensive canonicalization & also requires element/attribute reorganization to enable integration with the other file.

MD5: 1ede428d670ef18e68a3c326efb91bc4
SHA1: 3f6e7154583302094043254 7c12b1592b0663e73

Commonality in Profile between the two datasets :

Based on a high-level review of the datasets (old and new), almost all the elements and attributes are in common between the two files except for a few attributes in one file that is coded as an element in the other file. There are some ordering inconsistencies between the files on certain elements and also within the files like for example – in the event element the order of appearance of its attributes is different both within the file and also between the files. Similarly, not all the response elements appear in the same order within the file and also between the files. Also, there is no consistent ordering in sub-elements within the complaint element both within the file and between the files. There are also some data format inconsistencies between the files, like for example – the Yes or No indicator for one of the attributes in response element is coded as "yes" and "no" while the same attribute for response element is coded as "Y" and "N" in the other file. There are other format issues like excessive spacing between attributes, elements and their values. All these at a very high level indicate that these files require some level of canonicalization and reorganization of elements & attributes including converting some of the attributes to elements and resolving inconsistencies to enable full integration. The following sections detail on how these are accomplished.

DTD Documentation:

DTD table for Consumer_Complaints_FileA.xml

XML DTD ELEMENT DOCUMENTATION			
DTD ELEMENT NAME	**XML CATEGORY**	**MANDATORY**	**SHORT DESCRIPTION**
consumerComplaints	Element	Yes	Root element that constitutes multiple occurrences of complaint.
complaint	Element	Yes	Complaint instance raised by consumer
id	Attribute	Yes	Unique complaint id assigned to each complaint
event	Element	Yes	Identifies the date when the event was received and sent to company
type	Attribute	Yes	Type has two attribute values – received and sentToCompany within the event element
date	Attribute	Yes	Date when the complaint was received and sent to company occurring within the event element
product	Element	Yes	Parent element for product related information
productType	Element	Yes	Product for which complaint is submitted
subproduct	Element	No	Additional product information
issue	Element	Yes	Parent element for issue information
issueType	Element	Yes	Type of issue for which a complaint is being registered
subissue	Element	No	Additional information on the issue
company	Element	Yes	Parent element for Company information
companyName	Element	Yes	Name of the company for which the complaint is submitted
companyState	Element	Yes	State of the company
companyZip	Element	Yes	Zip of the company
submitted	Element	Yes	Method of submission
via	Attribute	Yes	Mode of submission of the complaint. Contains an enumerated list of values
response	Element	Yes	Response to the complaint. This is a group element

timely	Attribute	Yes	This is an indicator attribute with a value of Y or N indicating if a timely response was offered
consumerDisputed	Attribute	Yes	This is an indicator attribute with Y or N value indicating if there was any dispute raised by the consumer on the response
responseType	Element	Yes	Type of response offered for the complaint
publicResponse	Element	No	Narrative on any public response associated with complaint
consumerNarrative	Element	No	Narrative provided by the consumer for registering the complaint

DTD table for Consumer_Complaints_FileB.xml

XML DTD ELEMENT DOCUMENTATION			
DTD ELEMENT NAME	**XML CATEGORY**	**MANDATORY**	**SHORT DESCRIPTION**
consumerComplaints	Element	Yes	Root element that constitutes multiple occurrences of complaint.
complaint	Element	Yes	Complaint instance raised by consumer
id	Attribute	Yes	Unique complaint id assigned to each complaint
submissionType	Attribute	Yes	Mode of submission of the complaint. Contains an enumerated list of values
event	Element	Yes	Identifies the date when the event was received and sent to company
type	Attribute	Yes	Type has two attribute values – received and sentToCompany within the event element
date	Attribute	Yes	Date when the complaint was received and sent to company occurring within the event element
product	Element	Yes	Parent element for product related information
productType	Element	Yes	Product for which a complaint is submitted
subproduct	Element	No	Additional product information
issue	Element	Yes	Parent element for issue information
issueType	Element	Yes	Type of issue for which a complaint is being registered
subissue	Element	No	Additional information on the issue
company	Element	Yes	Parent element for Company information
companyName	Element	Yes	Name of the company for which the complaint is submitted
companyState	Element	Yes	State of the company
companyZip	Element	Yes	Zip of the company
response	Element	Yes	Response to the complaint. This is a group element
timely	Attribute	No	This is an indicator attribute with a value of Y or N indicating if a timely response was offered
consumerDisputed	Attribute	Yes	This is an indicator attribute with Y or N value indicating if there was any dispute raised by the consumer on the response

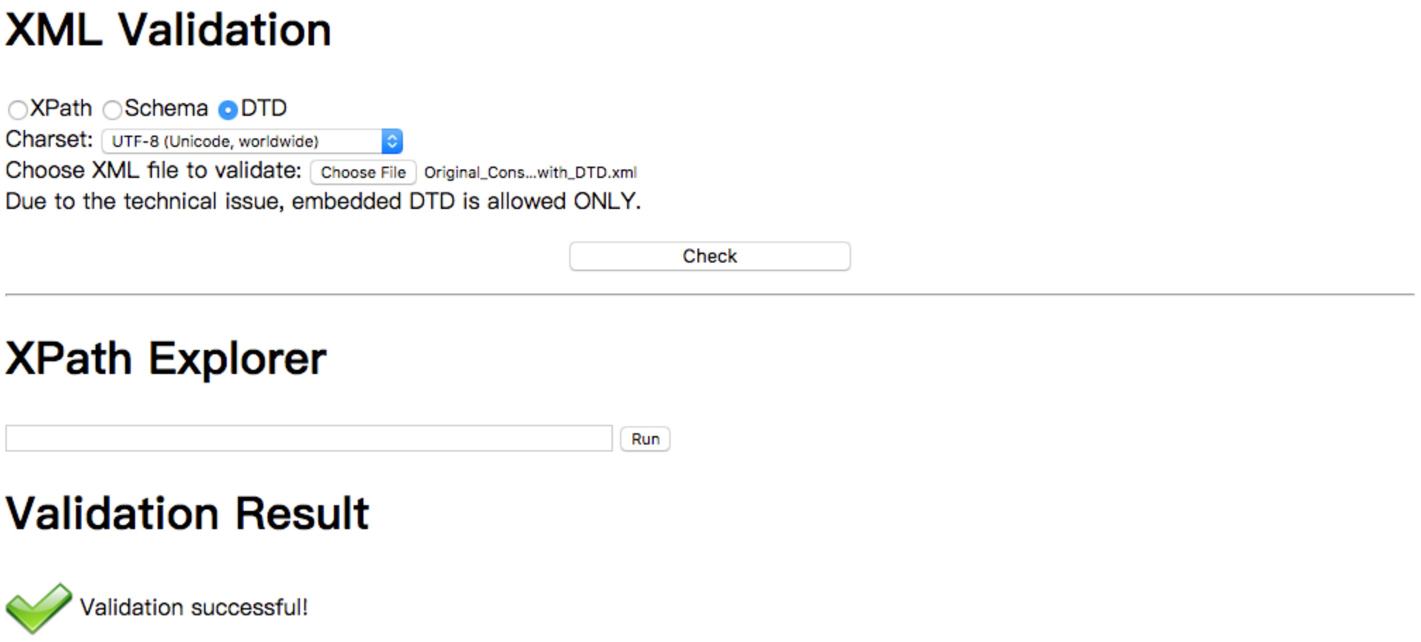
responseType	Element	Yes	Type of response offered for the complaint
publicResponse	Element	No	Narrative on any public response associated with complaint
consumerNarrative	Element	No	Narrative provided by the consumer for registering the complaint

DTD table for Final_Canonicalized_Consumer_Complaints.xml

XML DTD ELEMENT DOCUMENTATION			
DTD ELEMENT NAME	**XML CATEGORY**	**MANDATORY**	**SHORT DESCRIPTION**
consumerComplaints	Element	Yes	Root element that constitutes multiple occurrences of complaint.
complaint	Element	Yes	Complaint instance raised by consumer
id	Element	Yes	Unique complaint id assigned to each complaint
event	Element	Yes	Identifies the date when the event was received and sent to company
Received	Element	Yes	Date when the complaint was received occurring within the event element
sentToCompany	Element	Yes	Date when the complaint was sent to company occurring within the event element
product	Element	Yes	Parent element for product related information
productType	Element	Yes	Product for which a complaint is submitted
subproduct	Element	No	Additional product information
issue	Element	Yes	Parent element for issue information
issueType	Element	Yes	Type of issue for which a complaint is being registered
subissue	Element	No	Additional information on the issue
company	Element	Yes	Parent element for Company information
companyName	Element	Yes	Name of the company for which a complaint is submitted
companyState	Element	Yes	State of the company
companyZip	Element	Yes	Zip of the company
submitted	Element	Yes	Method of submission
via	Attribute	Yes	Mode of submission of the complaint. Contains an enumerated list of values
response	Element	Yes	Response to the complaint. This is a group element
timely	Attribute	Yes	This is an indicator attribute with a value of Y or N indicating if a timely response was offered
consumerDisputed	Attribute	Yes	This is an indicator attribute with Y or N value indicating if there was any dispute raised by the consumer on the response
responseType	Element	Yes	Type of response offered for the complaint
publicResponse	Element	No	Narrative on any public response associated with complaint
consumerNarrative	Element	No	Narrative provided by the consumer for registering the complaint

DTD Validation:

DTD Validation for Original_Consumer_Complaints_FileA_with_DTD.xml



XML Validation

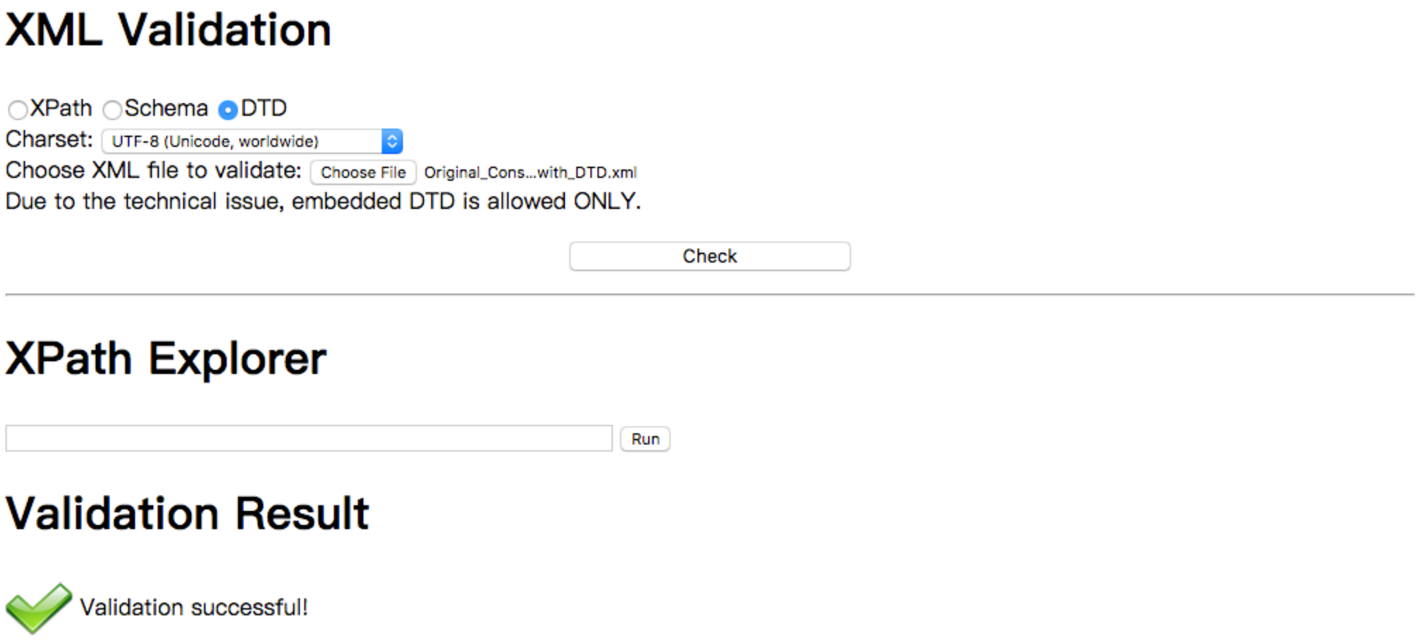
○XPath ○Schema ●DTD
Charset: UTF-8 (Unicode, worldwide)
Choose XML file to validate: Choose File Original_Cons...with_DTD.xml
Due to the technical issue, embedded DTD is allowed ONLY.

XPath Explorer

Validation Result

✓ Validation successful!

DTD Validation for Original_Consumer_Complaints_FileB_with_DTD.xml



XML Validation

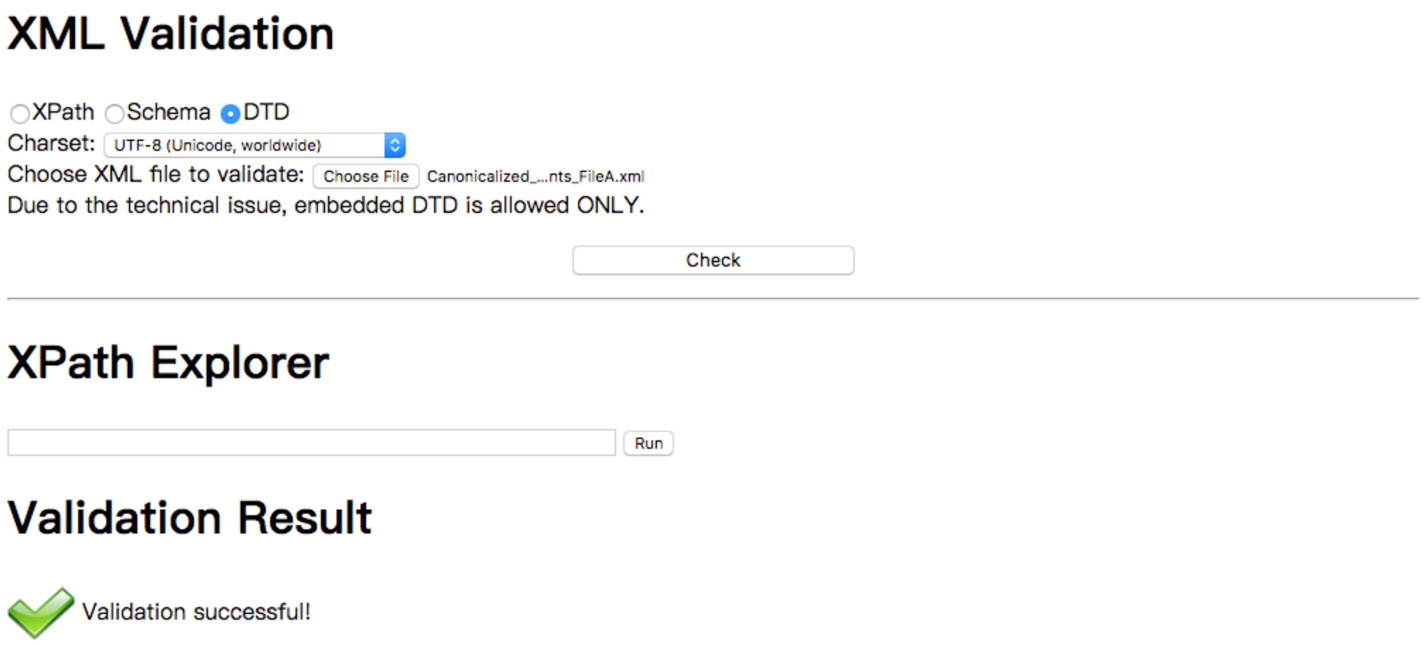
○XPath ○Schema ●DTD
Charset: UTF-8 (Unicode, worldwide)
Choose XML file to validate: Choose File Original_Cons...with_DTD.xml
Due to the technical issue, embedded DTD is allowed ONLY.

XPath Explorer

Validation Result

✓ Validation successful!

DTD Validation for Canonicalized_Consumer_Complaints_FileA_with_DTD.xml



XML Validation

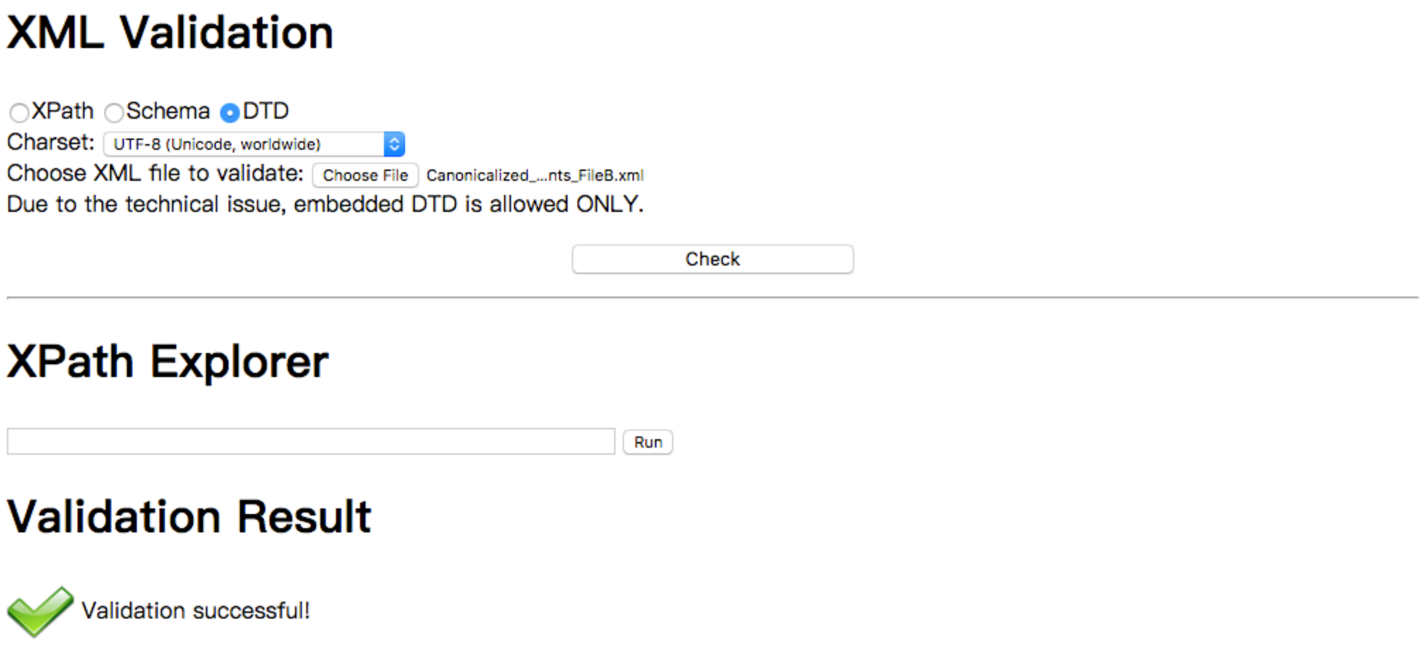
○ XPath ○ Schema ● DTD
Charset: UTF-8 (Unicode, worldwide)
Choose XML file to validate: Choose File Canonicalized_...nts_FileA.xml
Due to the technical issue, embedded DTD is allowed ONLY.

XPath Explorer

Validation Result

✓ Validation successful!

DTD Validation for Canonicalized_Consumer_Complaints_FileB_with_DTD.xml



XML Validation

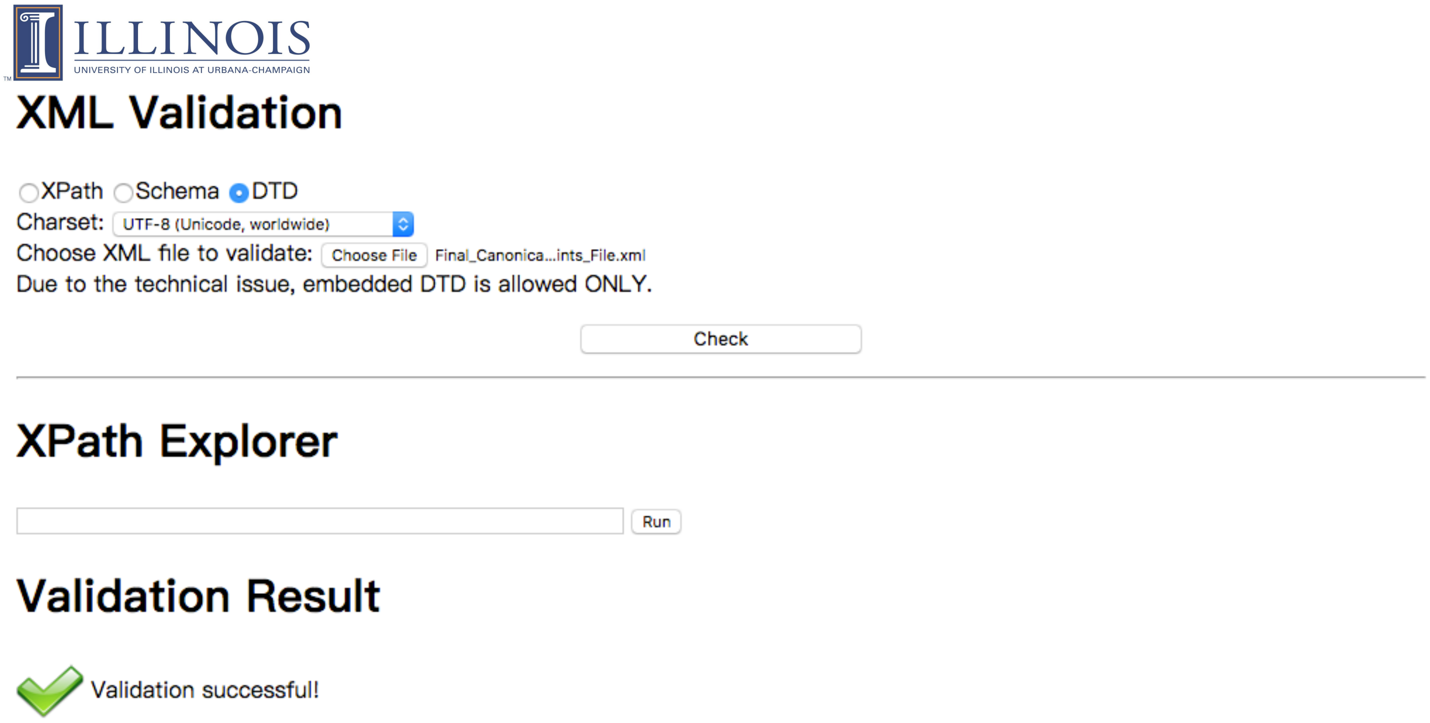
○ XPath ○ Schema ● DTD
Charset: UTF-8 (Unicode, worldwide)
Choose XML file to validate: Choose File Canonicalized_...nts_FileB.xml
Due to the technical issue, embedded DTD is allowed ONLY.

XPath Explorer

Validation Result

✓ Validation successful!

DTD Validation for Final_Canonicalized_Consumer_Complaints_with_DTD.xml



ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

XML Validation

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

Check

XPath Explorer

Run

Validation Result

✓ Validation successful!

Canonicalization:

Canonicalization of Consumer_Complaints_FileB.xml

Below were the actions performed to canonicalize the Consumer_Complaints_FileB.xml after careful review of the provided XML document and the generated DTD.

1. Trailing space in the *subproduct* tag was removed.
2. Leading extra space before the attribute names within the *event* tag was removed.
3. Removed extra space after the attribute names within the *event* element.
4. Removed the trailing extra space from the attribute value of the *event* tag.
5. Removed the single occurrence of empty *<submitted/>* tag.
6. Removed the extra space between attribute name and element name within the *complaint* tag.
7. Removed the trailing extra space within the *response* tag.
8. Placed the elements within the complaint parent element consistently in the following order - *event, product, issue, company, submitted* and *response* which seemed to be the order in most of the *complaint* element occurrences in the original XML.
9. Order of appearance of attributes (*timely* and *consumerDisputed*) within the *response* elements has been made consistent across all occurrences. The attribute *timely* will appear first followed by the *consumerDisputed* attribute.
10. The attribute values for *timely* attribute within the *response* element has been updated to represent the value of Y for Yes and N for No consistently across all occurrences.

MD5: 31df4d082bcd2863115daeb91bf28338
SHA1: 4f965bff80d734eb517a7209d74e9a5ce11cc56d

Canonicalization of Consumer_Complaints_FileA.xml

Below were the actions performed to canonicalize the Consumer_Complaints_FileA.xml after careful review of the provided XML document and the generated DTD.

1. Moved event tag before the *company* tag to keep it consistent.

2. Ensured that all the *event* elements occurred together.
3. Ensured the *company* and *submitted* tags occurred before *response* tag and *response* is the last tag occurring within the *complaint* element.
4. Placed the elements within the *complaint* parent element consistently in the following order - *event, product, issue, company, submitted* and *response* which seemed to be the order in most of the *complaint* element occurrences in the original XML.

MD5: 48a128d50c5b13cc7f90839da2082da3
SHA1: a92f95f0de2be06c7ebc909182a42a541f52cc66

Canonicalized Final Integrated XML

Below were the actions performed to integrate the Consumer_Complaints_FileB.xml and Consumer_Complaints_FileA.xml which seemed to follow similar XML hierarchy along with element & attribute names. The data profile of these two XML files indicate that they exhibit a similar functional purpose and can be integrated into one XML schema with appropriate steps. There were some additional canonicalization steps that were followed during the integration process and are included in the details below. The integration was accomplished after an element to element and attribute to attribute review of each the XML files (old and new) provided along with their respective generated DTD documents and with careful consideration to the functional purpose of each element in the context of the respective XML files.

1. Replaced entities with the actual value where possible. The redacted entity definition in the original DTD has been replaced with the actual value.
2. Converted the attributes within the *event* element to individual elements under the event parent element. The two attributes under the event tag refer to dates and since dates are dynamic with multiple possible values, it made more sense to replace the attributes as elements.
3. Added end tags to the *event* element.
4. Replaced the id attribute from the *complaint* element as a new child element under the *complaint* element. This is because id value is dynamically generated and would be logical to represent them as elements in a XML rather than attributes. Attribute representations are ideal in the case of static values associated with drop downs values or lists.
5. The *timely* and *consumerDisputed* attributes within the *response* element have been updated to represent the values “Yes” and “No” as “Y” and “N” in every occurrence.
6. The *timely* attribute has been made mandatory (#REQUIRED) unlike the Consumer_Complaints_FileB.xml and should either hold a value of “Y” or “N”.
7. The *submissionType* attribute within the *complaint* element from the Consumer_Complaints_FileB.xml has been integrated into the *submitted* element under the *complaint* parent element like in Consumer_Complaints_FileA.xml with the a *via* attribute representing the mode of submission.

MD5: 9ba91cfba61f4ee20749e9bef5bb44d3
SHA1: 68e7262513c1d96e601df612537b819ed56eee8a

Support of Data Curation Objectives :

The canonicalized XML schema supports different curatorial activities and objectives. XML schema in general has several key benefits and some of them include platform independence, flexibility and extensibility. The XML schema above provides data export facilities in a common simple exchange format which both captures the data structure and provides a degree of clarity to the exported data. The canonicalization process itself supports reuse and preservation and ensures the data is understandable and useable in the future. The canonicalization process of ordering the elements, attributes and removing invalid characters supports reproducibility, access and reformatting, the data consistency established enables reproducing results and comparing data for equivalence and in turn ensures validity and reliability. The XML schema supports integration to different disparate data sources and different data types in the distributed data sources. Use of common exchange schemas like XML to define data structures allows to parse data & organize into semantically described data objects conforming to the curatorial objectives. This schema as evident from the structure allows common information types to be represented in a syntactically defined structure thus enabling data documents to programmatically parsed into meaningful information and supports different curatorial objectives like discoverability and provenance too. The semantically consistent data representation in the XML schema above also enables sharing the data within the domains. The above schema also supports the ability to systematize data workflows. There is more work that needs to be done which has not been detailed above in terms of Security, Sharing, Workflow, Compliance and Provenance as well.

DTD for Original and Canonicalized XML:

DTD for Original FileA:

```
<!DOCTYPE consumerComplaints [
<!ELEMENT consumerComplaints ( complaint+ ) >
<!ELEMENT complaint ( company | consumerNarrative | event | issue | product | response |
submitted )* >
<!ATTLIST complaint id NMTOKEN #REQUIRED >
<!ELEMENT event EMPTY >
<!ATTLIST event date NMTOKEN #REQUIRED >
<!ATTLIST event type ( received | sentToCompany ) #REQUIRED >
<!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 >
<!ATTLIST submitted via NMTOKEN #REQUIRED>
<!ELEMENT response ( responseType | publicResponse? )* >
<!ATTLIST response consumerDisputed NMTOKEN #REQUIRED >
<!ATTLIST response timely NMTOKEN #REQUIRED >
<!ELEMENT publicResponse ( #PCDATA ) >
<!ELEMENT responseType ( #PCDATA ) >
]>
```

DTD for Original FileB:

```
<!DOCTYPE consumerComplaints [
<!ELEMENT consumerComplaints ( complaint+ ) >
<!ELEMENT complaint ( company | consumerNarrative | event | issue | product | response |
submitted )* >
<!ATTLIST complaint id NMTOKEN #REQUIRED >
<!ATTLIST complaint submissionType NMTOKEN #IMPLIED>
<!ELEMENT event EMPTY >
<!ATTLIST event date NMTOKEN #REQUIRED >
<!ATTLIST event type CDATA #REQUIRED >
<!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 response ( responseType  | publicResponse? )* >
<!ATTLIST response consumerDisputed NMTOKEN #REQUIRED >
<!ATTLIST response timely NMTOKEN #IMPLIED >
<!ELEMENT publicResponse ( #PCDATA ) >
<!ELEMENT responseType ( #PCDATA ) >
<!ELEMENT submitted EMPTY >
<!ENTITY redaction "XXXX">
]>
```

DTD for Canonicalized FileA:

```
<!DOCTYPE consumerComplaints [
<!ELEMENT consumerComplaints ( complaint+ ) >
<!ELEMENT complaint ( company | consumerNarrative | event | issue | product | response |
submitted )* >
<!ATTLIST complaint id NMTOKEN #REQUIRED >
<!ELEMENT event EMPTY >
<!ATTLIST event date NMTOKEN #REQUIRED >
<!ATTLIST event type ( received | sentToCompany ) #REQUIRED >
<!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 >
<!ATTLIST submitted via ( Web | Referral | Phone ) "Web">
<!ELEMENT response ( responseType | publicResponse? )* >
<!ATTLIST response consumerDisputed NMTOKEN #REQUIRED >
<!ATTLIST response timely NMTOKEN #REQUIRED >
<!ELEMENT publicResponse ( #PCDATA ) >
<!ELEMENT responseType ( #PCDATA ) >
]>
```

DTD for Canonicalized FileB:

```
<!DOCTYPE consumerComplaints [
<!ELEMENT consumerComplaints ( complaint+ ) >
<!ELEMENT complaint ( company | consumerNarrative | event | issue | product | response )*
>
<!ATTLIST complaint id NMTOKEN #REQUIRED >
<!ATTLIST complaint submissionType (Web | Referral | Phone ) "Web" >
<!ELEMENT event EMPTY >
<!ATTLIST event date NMTOKEN #REQUIRED >
<!ATTLIST event type CDATA #REQUIRED >
<!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 response ( responseType  | publicResponse? )* >
<!ATTLIST response consumerDisputed NMTOKEN #REQUIRED >
<!ATTLIST response timely NMTOKEN #IMPLIED >
<!ELEMENT publicResponse ( #PCDATA ) >
<!ELEMENT responseType ( #PCDATA ) >
<!ENTITY redaction "XXXX">
]>
```

DTD for Final Canonicalized File:

```
<!DOCTYPE consumerComplaints [
<!ELEMENT consumerComplaints ( complaint+ ) >
<!ELEMENT complaint (id | company | consumerNarrative | event | issue | product |
response | submitted)* >
<!ELEMENT id ( #PCDATA ) >
<!ELEMENT event (received,sentToCompany)* >
<!ELEMENT received ( #PCDATA ) >
<!ELEMENT sentToCompany ( #PCDATA ) >
<!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 >
<!ATTLIST submitted via ( Web | Referral | Phone ) "Web" >
<!ELEMENT response ( responseType | publicResponse? )* >
<!ATTLIST response consumerDisputed NMTOKEN #REQUIRED >
<!ATTLIST response timely NMTOKEN #REQUIRED >
<!ELEMENT publicResponse ( #PCDATA ) >
<!ELEMENT responseType ( #PCDATA ) >
]>
```

contd..

Memo on importance of Data Curation:

Anantharaman Janakiraman, Date: 11/25/2017
Data Scientist,
XXXX Corporation.

Sir/Madam,

Please treat this memo as a request and recommendation to continue funding data curatorial functions in our area which is an integral and essential precursor to data analysis. The quality of data analysis is dependent on the quality of information analyzed and data curation enhances the quality of data. Data quality issues can have a significant impact on business operations, especially when it comes to the decision-making processes within any organization.

The emergence of new platforms for decentralized data creation such as the digital platforms in our organization, the increasing availability of open data on the web, added to the increase in the number of data sources inside our organization, brings an unprecedented volume of data to be managed. In addition to the data volume, we need to cope with data variety, as a consequence of the decentralized data generation, where data is created under different contexts and requirements. Also consuming third-party data comes with the intrinsic cost of repurposing, adapting, and ensuring data quality for its new context.

Data curation provides the methodological and technological data management support to address data quality issues maximizing the usability of the data. Reusing data that was generated under different requirements comes with the intrinsic price of coping with data quality and data heterogeneity issues. Data can be incomplete or may need to be transformed in order to be rendered useful. Data shifts from a resource that is tailored from the start to a certain purpose, to a raw material that will need to be repurposed in different contexts in order to satisfy a particular requirement. In a scenario where there is a lot of data shift along with the context and to deal with incomplete data that requires transformation, data curation emerges as a key data management activity.

Data curation can be done from data generation perspective too which is also called curation at source and this could help represent the data in a way that maximizes its quality in different contexts. Data curation enables the extraction of value from data, and it is a capability that is required for areas such as ours that are dependent on complex and/or continuous data integration and classification. I would actually recommend improvement of data curation tools and methods that directly provides greater efficiency of the knowledge discovery process, maximizes return of investment per data item through reuse, and improves organizational transparency.

Despite the fact that data heterogeneity and data quality were concerns already present before the big data scale era, they become more prevalent in data management tasks with the growth in the number of data sources such as our shop. This growth as ours brought the need to define principles and scalable approaches for coping with data quality issues. It also brought data curation from a niche activity, restricted to a small community of scientists and analysts with high data quality standards, to a routine data management activity, which will progressively become more present within the average data management environment.

Data curation processes can be categorized into different activities such as Collection, Organization, Storage, Preservation, Discoverability, Access, Workflow, Identification, Integration, Reformatting, Reproducibility, Sharing, Communication, Provenance, Modification, Compliance, Security and each of the above activities if performed appropriately will improve the data quality significantly that enhances the decision-making processes, transparency and accuracy.

I just want to highlight the importance of few of the activities in the context of data analysis that is performed in our organization on a daily basis. The Metadata that we produce through various processes in our organization can be used to support reuse and long-term preservation. Curating metadata supports discovery, schema changes, understanding and management of other data and information extensively. Similarly, some of the data quality attributes can be made evident by the data itself, while others depend on understanding of the broader context behind the data, i.e. the provenance of the data, the processes, artifacts, and actors behind the data creation. Capturing and representing the context in which the data was generated and transformed and making it available for data consumers is a major aspect of data curation and provenance standards provide the grounding for interoperable representation of data. Curation as we currently employ in our organization supports the ability to retrieve and distribute data – maintain systems, tools and metadata that support the efficient and reliable retrieval and distribution of data.

In essence, with the growth in the number of data sources and of decentralized content generation, ensuring data quality becomes a fundamental issue for data management environments in the big data era. The evolution of data curation methods and tools is a cornerstone element for ensuring data quality at the scale of big data.

I sincerely request you to reconsider your decision to de-prioritize data curation services in our organization and also request you to provide careful consideration when revisiting your decision. I sincerely hope the reasons stated in this memo would emphasize the need for data curation services if not more and help you make an informed decision.

Please let me know if you have further questions or concerns.

Thanks,
Anantharaman Janakiraman.

Artifacts :

Link to location of all the Artifacts - https://github.com/ananthajanakiraman/ananjanakiraman.github.io