Assignment 2: XML Schema Design Exercise

Megan Masanz

CS 598: Foundations of Data Curation

Date: 10/14/2018

Contents

Abstract

After reviewing all 7 sections of the DTD tutorial located at: https://www.w3schools.com/xml/xml_dtd_intro.asp students were instructed to select a structured or unstructured document that is of interest from their workplace.

The goal of Assignment 2 is to provide a schema that best represents the document, and its important or essential components and aspects for their purposes.

The assignment requires prose documentation for each element, attribute and attribute value. The document is to be provided in the mark up according to the DTD designed.

The assignment is required to provide a narrative about this process and answer the following questions:

- How did you decide to represent the data in the way that you did? Why did you choose the elements and attributes that you did?
- What were the hardest decisions you had to make in this design process?

- How does your DTD design support data independence?
- How may your DTD design support the overarching goals of data curation (revisit objectives and activities of Week 1)?

The document created for this assignment detailed and provided as part of this assignment did successfully pass the validation provided by the online tool: http://xmlvalidator.new-studio.org/

Document Selection

The JSON document selected for this project is the first revision of a document used by an enterprise corporation to define unique data sets that are to be stored within a data lake. The intention of the data lake is to provide storage for data sets that have not currently been captured and stored for long term use. The initial set of data sets identified as valuable to store will be used by data scientists to provide new insights by leveraging the data for analysis. The data sets will need to be stored and retrieved systematically by both scientists, and by programs written by data scientists for analysis. A data lake is a distributed file system (current implementation uses Azure Data Lake), intended to provide scalable storage.

An ingestion process has been implemented to move on-premise files into the data lake. In addition to storing the files, they are being cataloged for data retrieval.

To assist in this task, the original JSON document is an implementation of a defining a data set based on a limited subset of properties found to define a data set defined by schema.org described in the Appendix. According to the schema.org website, “Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD”. (schema.org, n.d.) The initial JSON document implementation will be used to populate a no-sql database. The information found within this no-sql database will be used to populate a data catalog to enable search, maintenance, and usage of the data sets across an enterprise.

There was an in-depth analysis of business requirements, cataloging strategies, and security concerns used to provide the implementation of the JSON document used within this exercise. The process detailed in this exercise is focused on the translation from the JSON document to XML document detailed in the previous section of this document. Information pertaining to the business requirements which led to the initial implementation can be found within the Appendix.

Original JSON Document

Given the assignment requirement was to provide a single pdf file, file origin JSON document can be downloaded from : https://github.com/megado123/cs-598-DataCuration-Assignment2. Within the repository, this specific file can be downloaded from: https://github.com/megado123/cs-598-DataCuration-Assignment2/blob/master/SampleSchemaOrg.json

```
{
  "@context":"http://schema.org",
  "@type":"Dataset",
  "name":"201186_201186_Samplefile_201809191949_2.SPA",
  "potentialAction":{
    "@type":"Action",
    "name":"refinery",
    "actionStatus":"http://schema.org/CompletedActionStatus"
  },
  "sameAs":"\\\\usfile01\\itdatawhse\\crl_dataIngest\\crl-
dev\\20180919202\\cral\\T1Test\\201186_Samplefile_201809191949_2.SPA",
  "url":"adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl/cral/Spectropscopy-
IRImaging/201186_Samplefile_201808170147_1.SPA",
  "about":{
    "@type":"Thing",
    "description":"N, N - dimethylacrylamide(DMA) is observed throughout both samples(Tables 1 and
2).  The cleaned sample, E18 - 0333 - 02 shows less DMA compared to pre washed sample, E18-0333 -
01.No useful trend was observed in DMA level from “outside” of the roll to the “inside” of the roll.",
    "disambiguatingDescription":"3M Menonmonie supplied a sample from a piece of neoprene
rubber hose.They would like to determine how much N, N Dimethylacrylamide(DMA) has penetrated
through the rubber hose. The side that was exposed to DMA has the largest flat surface.  The side
with the saw marks on it would be towards the inside.",
    "identifier":"260233"
  },
  "creator":{
    "@type":"Person",
    "identifier":"US263070"
  },
  "dateCreated":"2018-09-26T20:17:41.2976598+00:00",
  "fileFormat":"SPA",
  "keywords":"TechniqueCode: IR, TechniqueLongDescription: Infrared Spectroscopy (IR)",
  "sourceOrganization":{
    "@type":"Organization",
    "name":"(US, St Paul) Corporate Research Analytical Laboratory(CRAL)",
    "alternateName":"CRAL"
  },
  "workTranslation":{
    "@type":"CreativeWork",

"url":"adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/refinery/crl/cral/automated/Spectropsco
py-IRImaging/201186_Samplefile_201808170147_1.csv",
    "dateCreated":"2018-09-26T15:17:39.5828728-05:00",
    "dateModified":"2018-09-26T15:17:39.7636674-05:00"
  },
  "distribution":[
      {
        "@type":"DataDownload",
        "url":"https://usw-su1.azuredatacatalog.com",
        "contentUrl":"adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl/cral/Spectropscopy-IRImaging/201186_Samplefile_201808170147_1.SPA",
        "encodingFormat":"CSV"
      },
      {
        "@type":"DataDownload",
        "url":"https://usw-su1.azuredatacatalog.com",
        "contentUrl":"adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl/cral/Spectropscopy-IRImaging/201186_Samplefile_201808170147_1.SPA",
        "encodingFormat":"SPA"
      }
    ],
    "includedInDataCatalog":{
      "@type":"DataCatalog",
      "url":"https://usw-su1.azuredatacatalog.com"
    },
    "measurementTechnique":"IR Spectoscopy",
    "hasDigitalDocumentPermission": [
     {
      "@type": "DigitalDocumentPermissionType",
      "permissionType": "http://schema.org/WritePermission",
      "grantee": {
        "@type": "Audience",
        "name": "cral - data - platform - dev - team - awesome",
        "identifier" : "5c341e13-dfb3-48fc-a776-3dd9b8763293"
      }
     },
     {
      "@type": "DigitalDocumentPermissionType",
      "permissionType": "http://schema.org/ReadPermission",
      "grantee": {
        "@type": "Audience",
        "name": "cral-readers",
        "identifier" : "64932e12-bca9-224l-n237-212de2492124"
      }
     }
   ]
 }
```

XML Document implementation

Given the assignment requirement was to provide a single pdf file, file XML document can be downloaded from : https://github.com/megado123/cs-598-DataCuration-Assignment2 This specific file can be downloaded from: https://github.com/megado123/cs-598-DataCuration-Assignment2/blob/master/Datasets3.xml

```
<?xml version="1.0"?>
<!DOCTYPE datasets [
<!ELEMENT datasets (dataset+)>
<!ELEMENT dataset (name, dateCreated, sameAs, url, fileFormat, measurementTechnique,
potentialAction?, keywords, about, creator, sourceOrganization, workTranslation*, distribution*,
includedInDataCatalog, hasDigitalDocumentPermission+)>
<!ELEMENT name (#PCDATA) >
<!ELEMENT keywords (#PCDATA)>
<!ELEMENT dateCreated (#PCDATA)>
<!ELEMENT sameAs (#PCDATA)>
<!ELEMENT url (#PCDATA)>
<!ELEMENT potentialAction (name, actionStatus)>
<!ELEMENT actionStatus (#PCDATA)>
<!ELEMENT about (description, disambiguatingDescription, identifier)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT disambiguatingDescription (#PCDATA)>
<!ELEMENT identifier (#PCDATA)>
<!ELEMENT creator (identifier)>
<!ELEMENT sourceOrganization (name, alternateName)>
<!ELEMENT alternateName (#PCDATA)>
<!ELEMENT workTranslation (url, dateCreated, dateModified )>
<!ELEMENT dateModified (#PCDATA)>
<!ELEMENT distribution (url, contentUrl, encodingFormat)*>
<!ELEMENT contentUrl (#PCDATA)>
<!ELEMENT encodingFormat (#PCDATA)>
<!ELEMENT fileFormat (#PCDATA) >
<!ELEMENT includedInDataCatalog (url)>
<!ELEMENT measurementTechnique (#PCDATA)>
<!ELEMENT hasDigitalDocumentPermission (permissionType, grantee)+>
<!ELEMENT permissionType (#PCDATA)>
<!ELEMENT grantee (name, identifier)>


<!ATTLIST dataset context CDATA #FIXED "http://schema.org">
<!ATTLIST dataset type CDATA #FIXED "Dataset" >
<!ATTLIST potentialAction type CDATA  #FIXED "Action">
<!ATTLIST about type CDATA #FIXED "Thing">
<!ATTLIST creator type (Organization|Person)  #REQUIRED>
<!ATTLIST sourceOrganization type CDATA #FIXED "Organization">
<!ATTLIST workTranslation type CDATA #FIXED "CreativeWork">
<!ATTLIST distribution type CDATA #FIXED "DataDownload">
<!ATTLIST includedInDataCatalog type CDATA #FIXED "DataCatalog">
<!ATTLIST permissionType type CDATA #FIXED "DigitalDocumentPermissionType">
<!ATTLIST grantee type (Audience|Person)  #REQUIRED>

<!ENTITY SchemaDefinition "http://schema.org/">
<!ENTITY CompletedActionStatus "CompletedActionStatus">
<!ENTITY ActiveActionStatus "ActiveActionStatus">
<!ENTITY FailedActionStatus "FailedActionStatus">
<!ENTITY ReadPermission "ReadPermission">
<!ENTITY WritePermission "WritePermission">
<!ENTITY DataCatalogSolution "https://usw-su1.azuredatacatalog.com">
<!ENTITY AzureDataLakeStore "adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl">

]>
<datasets>
   <dataset context="http://schema.org" type="Dataset">
      <name>201186_201186_Samplefile_201809191949_2.SPA</name>
      <dateCreated>2018-09-26T20:17:41.2976598+00:00</dateCreated>
      <sameAs>\\\\usfile01\\itdatawhse\\crl_dataIngest\\crl-
dev\\20180919202\\cral\\T1Test\\201186_Samplefile_201809191949_2.SPA</sameAs>
      <url>&AzureDataLakeStore;/cral/Spectropscopy-
IRImaging/201186_Samplefile_201808170147_1.SPA</url>
      <fileFormat>SPA</fileFormat>
      <measurementTechnique>IR Spectoscopy</measurementTechnique>
      <potentialAction type="Action">
         <name>refinery</name>
         <actionStatus>&SchemaDefinition;&CompletedActionStatus;</actionStatus>
      </potentialAction>
      <keywords>TechniqueCode: IR, TechniqueLongDescription: Infrared Spectroscopy
(IR)</keywords>
      <about type="Thing">
         <description>N, N - dimethylacrylamide(DMA) is observed throughout both samples(Tables 1
and 2).  The cleaned sample, E18 - 0333 - 02 shows less DMA compared to pre washed sample, E18-
0333 - 01.No useful trend was observed in DMA level from “outside” of the roll to the “inside” of the
roll.</description>
         <disambiguatingDescription>3M Menonmonie supplied a sample from a piece of neoprene
rubber hose.They would like to determine how much N, N Dimethylacrylamide(DMA) has penetrated
through the rubber hose. The side that was exposed to DMA has the largest flat surface.  The side
with the saw marks on it would be towards the inside.</disambiguatingDescription>
         <identifier>260233</identifier>
      </about>
      <creator type="Person">
         <identifier>US263070</identifier>
      </creator>
      <sourceOrganization type="Organization">
         <name>(US, St Paul) Corporate Research Analytical Laboratory(CRAL)</name>
         <alternateName>CRAL</alternateName>
      </sourceOrganization>
      <workTranslation type="CreativeWork">
         <url>&AzureDataLakeStore;/cral/automated/Spectropscopy-
IRImaging/201186_Samplefile_201808170147_1.csv</url>
         <dateCreated>"2018-09-26T15:17:39.5828728-05:00"</dateCreated>
         <dateModified>"2018-09-26T15:17:39.7636674-05:00"</dateModified>
      </workTranslation>
      <distribution type="DataDownload">
      <url>&DataCatalogSolution;</url>
      <contentUrl>&AzureDataLakeStore;/cral/automated/Spectropscopy-
IRImaging/201186_Samplefile_201808170147_1.csv</contentUrl>
      <encodingFormat>CSV</encodingFormat>
    </distribution>
    <distribution type="DataDownload">
      <url>&DataCatalogSolution;</url>
      <contentUrl>&AzureDataLakeStore;/cral/Spectropscopy-
IRImaging/201186_Samplefile_201808170147_1.SPA</contentUrl>
      <encodingFormat>SPA</encodingFormat>
    </distribution>
    <includedInDataCatalog type="DataCatalog">
      <url>&DataCatalogSolution;</url>
    </includedInDataCatalog>
    <hasDigitalDocumentPermission>
      <permissionType
type="DigitalDocumentPermissionType">&SchemaDefinition;&ReadPermission;</permissionType>
        <grantee type="Audience">
          <name>cral - data - platform - dev - team - awesome</name>
          <identifier>5c341e13-dfb3-48fc-a776-3dd9b8763293</identifier>
        </grantee>
    </hasDigitalDocumentPermission>
    <hasDigitalDocumentPermission>
      <permissionType
type="DigitalDocumentPermissionType">&SchemaDefinition;&WritePermission;</permissionType>
        <grantee type="Audience">
          <name>cral-readers</name>
          <identifier>64932e12-bca9-224l-n237-212de2492124</identifier>
        </grantee>
    </hasDigitalDocumentPermission>
  </dataset>
</datasets>
```

XML Prose for element, attribute, and attribute values

The parent node is datasets. The datasets element is a parent node for dataset elements. The datasets element should contain 1-n dataset elements.

A dataset provides information for searching and defining a data set within a data cataloging solution. The element dataset contains elements including: name, dateCreated, sameAs url, fileFormat, measurementTechnique, potentialAction, keywords, about, creator, sourceOrganization, workTranslation, distribution, includedInDataCatalog, and hasDigitalDocumentationPermission. The dataset element has fixed attributes type and context are character data, as this implementation is using the definitions provided by shema.org for a data set (https://schema.org, Dataset, n.d.).

name is an element providing the name of the data set. The element occurs exactly once within a dataset with parsed character data.

dateCreated is an element providing the original date the data set was created. The dataCreated element is intended to be in the format defined by ISO 8601 Chapter 5.4 (https://www.iso.org/iso-8601-date-and-time-format.html, n.d.). The element occurs exactly once within a dataset with parsed character data.

sameAs is an element providing the on-premise location of a data file that was moved into Azure Data lake storage. This is an on-premise file file that is moved into the data lake. The element occurs exactly once within a dataset with parsed character data.
url is an element providing the data lake location of a data file. The element occurs exactly once within a dataset with parsed character data.

fileFormat is an element providing the file extension type of the file located in the data lake. This element occurs exactly once within a dataset with parsed character data.

measurementTechnique is an element providing the scientific measurement technique used to create the data set. This element occurs exactly once within a dataset with parsed character data.

potentialAction is an element providing potential data transformation pipeline actions that may occur on the data set and their status. This element can occur 0-1 times in a dataset with a fixed attribute type of Action (https://schema.org, Action, n.d.), since a potentionAction can only be of type Action as defined by schema.org (https://schema.org, potentialAction, n.d.). The name element found within the potentialAction is parsed data that occurs exactly once. This name specifies the name of the code to run to transform the dataset. The actionStatus element is parsed data that occurs exactly once. It is populated using the concatenation of entities SchemaDefinition and either CompletedActionStatus, ActiveActionStatus or FailedActionStatus. An implementation limitation is that there are no enforcement options for populated with these concatenations for elements.

keywords is an element providing key search terms for the dataset. This element occurs exactly once within a dataset with parsed character data.

about is an element providing description and identification information about a data set. The about element includes the elements description, disambiguatingDescription and identifier that are parsed character data and occur exactly once within the about element. The about element occurs exactly 1 time within the element dataset. about has a fixed attribute type of thing. type is a fixed attribute of character data since its definition is based on the schema.org implementation which requires the about (https://schema.org, about, n.d.) section of the document to follow the class definition of a thing (https://schema.org, thing, n.d.).

creator is an element providing information regarding who created the dataset with a type of either Organization or Person. The creator can be a individual, or a group of individuals defined as either an Organization (https://schema.org, Organization, n.d.) or a Person (https://schema.org, Person, n.d.) following the class definition provided by schema.org. It includes an identifier element that is parsed character data that occurs only once within creator. creator occurs exactly once within the dataset.

sourceOrganization is an element of fixed type Organization (https://schema.org, Organization, n.d.) that includes a name and alternateName element. name and alternateName are parsed character data that occurs only once within the sourceOrganization. sourceOrganization occurs exactly once within the dataset.

workTranslation is an element with an attribute type with a fixed character data with the value of CreativeWork that occurs 0-n times within a dataset. workTranslation is meant to hold information regarding transformations that have occurred on the data set within the data lake. A given data set may or may not require transformations which is why it can occur 0-n times in the dataset. The element workTranslation includes url, dateCreated, and dateModified elements that are parsed character data that only occurs once within the workTranslation. The type is fixed as CreativeWork as it follows the class definition provided by schema.org (https://schema.org, CreativeWork, n.d.). The dateCreated and dateModified are intended to be in the format defined by ISO 8601 Chapter 5.4 (https://www.iso.org/iso-8601-date-and-time-format.html, n.d.). The url element makes use of the entity AzureDataLakeStore to provide an initial root source in the Azure Data lake combined with the detailed location of the physical file within the data lake.

distribution is an element with an attribute type of character data with a fixed value of DataDownload. The element distribution can occur any number of times from 0 to n. The element is intended to capture how the dataset can be accessed within the data lake. distribution includes a url, contenturl and encodingFormat elements that are parsed character data that only occur once within the distribution. The type attribute is fixed as DataDownload because it adheres to the class definitions provided by schema.org for a distribution of type DataDownload (schema.org, DataDownload, n.d.). The distribution element can be found any 0-n times within the dataset. The url is an element of parsed character data that is populated by the entity DataCatalogSolution. The contenturl element makes use of the entity AzureDataLakeStore to provide an initial root source in the Azure Data lake combined with the detailed location of the physical file within the data lake for the file or transformation of the source file. The encodingFormat is is an element providing the file extension type of the file or the transformed in Azure Data Lake. This element occurs exactly once within a dataset with parsed character data.

includedInDataCatalog is an element with an attribute type of character data with a fixed value of DataCatalog that includes a url that is parsed character data that occurs only once within the element includedInDataCatalog. The element includedInDataCatalog occurs exactly once within the dataset. It is intended to provide users with an understanding if the data can be found within a data cataloging solution. The type is fixed as DataCatalog because it adheres to the class definitions provided by schema.org for a distribution of type DataCatalog (https://schema.org, DataCatalog, n.d.).

hasDigitalDocumentPermission is an element that occurs at least once in a dataset as it is the permissions granted to users for the document. It contains element permissionType with a fixed attribute type of character data which is set to DigitalDocumentPermissionType that is parsed character data that occurs once which specifies the permission type being granted. It also contains the element grantee which is of type Audience or Person. The grantee element has 2 elements name and identifier that are parsed character data and occur exactly once within the grantee element. The grantee element must be contained within the hasDigitalDocumentPermission at least once, as it specifies who has the permissions being described.

Process

Logical units of information that could enable search were determined to be elements. Characteristics of the logical units of information were defined in this implementation as attributes. The attributes are metadata associated with a specific element. A few entities were utilized. Entities are used to map elements to processing rules.

To determine what should be considered an attribute, the data set needed to be looked at not only at an individual element perspective, but also from a document level perspective. Schema.org provides class definitions for the information found in the initial JSON document. The class definition (in the JSON document what is called the type) is considered to provide context to a piece of information. A system or a programmer attempting to parse the information can use the type attribute to provide a definition of the nested attributes found within an attribute to assist in processing the data. This type provides the class name definition for the element. This class definition is a characteristic of the information and hence was provided as an attribute. To assist in the understanding that the class definitions for the types are defined by shema.org, the attribute context was also provided as an attribute as the context provides the url for the type definitions.

To determine what should be considered an entity, the processing rules that were well understood were defined. The primary processing of this data that will be implemented by this solution include transformation and user authorization to data sets. Data set transformation activities are defined by the element potentialAction. A potentialAction's status was defined to be an element actionStatus, which would tell a system if transformation needed to be implemented, hence it should be considered as a processing rule, and thus was determined to be an entity. ReadPermission and WritePermission also tell a system how to process the information, if a user or a group of users should be allowed access – authorization of data, and hence was determined to be an entity.

Entities were also used to provide abstraction. The data cataloging solution that will be implemented with this solution is currently influx. An additional processing step that could occur on this data is updating the data set information populated in one data cataloging solution into another. In addition to being related to processing of the information, using an entity to define the cataloging solution provides abstraction from the implemented solution to the data set definitions.

A final entity that was used was for the prefix of the actual data set location within a data lake. Currently in the data lake implementation, a natural hierarchy for storing files has been determined based on business requirements. If these business requirements change, using an entity to separate out the individual file store from the business higher level file store provides abstraction from content (the file location) from file storage (the location within the data lake).

The information found within the data sets xml document will be stored and used to populate a data catalog. Storing the information before using it in a data catalog provides a layer of abstraction between the storage of the documents and the consumption of the documents. Providing the data sets definition in an additional detailed implementation of XML provides an additional method for users to provide this information so the system that places the information into a no-sql database providing flexibility based on the source system of the documents to be cataloged.

The XML document while sharing in content of the information provided in the JSON document, the two documents vary greatly.

The JSON-LD document does not provide a schema, which will present validation issues in the future if one is not included in its own documentation.

The XML document also focuses on the curation of the data, rather than strictly capturing the data for storage. The original intent of the JSON document was to provide an abstraction layer between the data catalog and the data sets and their metadata. This XML implementation shows there are abstractions that should be considered in the original JSON document for additional future proofing of the data to enable management of the data through-out its life-cycle.

Difficult Decisions

XML implementations are surely available for the data set implementation found for the schema.org implementation, but to ensure the integrity of the assignment, any resources available were strictly not used during the assignment to ensure academic integrity. As a follow-up, available implementations will be reviewed and evaluated for use within the company implementing this solution. This was a difficult decision, but ensuring academic integrity was key to this assignment.

The file format element could have been an attribute on the data set and could have come from a list of appropriate file extensions, but given the audience for this data set, it was not considered to be metadata, and the list of file extensions can be exhaustive, requiring constant updates to the DTD, and so it was used as an element.

Another difficult decision was to include `Datasets` as an element or not. For the purposes of this exercise, a data set could have been the highest-level node, however in JSON, the requirement does not exist that there is a single parent node with no parent, so it was included in the XML implementation to indicate that there will be thousands of individual data set definitions included within the `Datasets`.

The element `potentialAction` in this implementation occurs exactly 0-1 times. It could be argued that the DTD should allow for multiple transformations to occur for a given data set. These transformations should be considered as transformation pipelines, so within the single pipeline, multiple transformations can occur, which is why the XML implementation did not stray from the JSON implementation for the dataset. In addition, the DTD does not support limiting the values allowed for this element, so it could be considered that the `actionStatus` should be an element, but based on the user requirements, this is not considered meta data, and was determined not to be an attribute.

The hardest decision was if the `AzureDataLakeStore` entity should be used. It is not actually used for processing, but it's use supports abstraction from the dataset storage, and therefore supports the purposes intended by data curation.

Supporting Data Independence

In the XML implementation, attempts were made to ensure changes made to the definition of the data set would not impact the data stored using entities. This included the acknowledgement of the data cataloging solution may change, the data lake storage location hierarchy may change. A modification to the entity would enable updating all entries of all data sets providing indirection support for the implementation.

This implementation will be used across the corporation, but its initial implementation will be for a single business unit. This initial business unit has identified 83 key data sets. The origin JSON document was designed to define all 83 datasets for storage and retrieval. Each data set requires data independence, but with the use of potential actions, traceability to transformed datasets is enabled. In addition to a consideration for traceability, the data sets captured can be grouped for analysis by their elements and attributes.

DTD Support Data Curation

There are several data curation principals that the DTD implementation will support. While the JSON document schema provided can be validated using a (Google Structured Data Testing Tool, n.d.), the DTD allows for validation of the document without the use of external tool. Providing an XML implementation in addition to the JSON implementation will expand implementation support, as the data sets this solution will capture will be coming from a variety of sources. This implementation will further the support of data collection and acquisition. The use of elements that can be searched against will provide discoverability of data sets enabling search across elements to find and consume data. The potential actions element will support data integration enabling data transformation. This solution also allows for modifications to entities to support quickly updating each data set based on the current understood possible implementation modifications that could be foreseen.

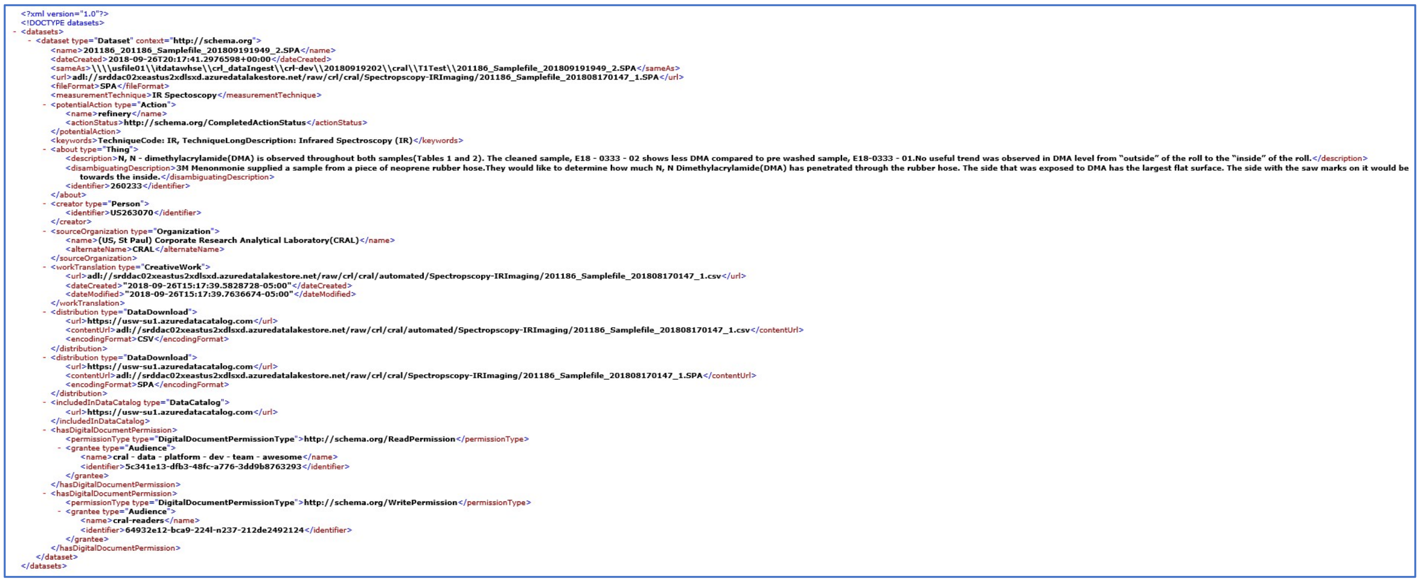
Appendix

Implementation of JSON document: The JSON document was created based on the information found in schema.org to define a data set generically. The motivation behind this limited implementation of a data set defined by schema.org was based on the fact that google is also defining a data set using this implementation and has developed a search engine that can look for these data sets (https://www.zdnet.com, 2018). There was a deviation from the schema given that on schema.org's website, there is also the concept of a digital document. A digital document includes permissions, while a dataset does not. In this implementation, security for a dataset is a concern and thus it was added into this implementation.

For the business in question, for each technical specialty found within their organization generating 83 different data sets, when a data set is created, it needs to be stored and retrievable. Basic file information should be captured as part of the data ingestion process which includes: creator, date and time of creation, modification date and times, technical technique used, file extensions, data lake storage location, project description information which led to the creation of the dataset, file transformations available, how to access the file within a data cataloging solution, and permissions, both user and group based. Each of the 83 data sets has a different format both structured and unstructured. This implementation is an effort to standardize how a data set is stored to enable its retrieval.

The documents can be downloaded from the following github repository:

https://github.com/megado123/cs-598-DataCuration-Assignment2



```
 <?xml version="1.0"?>
 <!DOCTYPE datasets>
- <datasets>
  - <dataset type="Dataset" context="http://schema.org">
      <name>201186_201186_Samplefile_201809191949_2.SPA</name>
      <dateCreated>2018-09-26T20:17:41.2976598+00:00</dateCreated>
      <sameAs>\\\\usfile01\\itdatawhse\\crl_dataIngest\\crl-dev\\20180919202\\cral\\T1Test\\201186_Samplefile_201809191949_2.SPA</sameAs>
      <url>adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl/cral/Spectropscopy-IRImaging/201186_Samplefile_201808170147_1.SPA</url>
      <fileFormat>SPA</fileFormat>
      <measurementTechnique>IR Spectoscopy</measurementTechnique>
    - <potentialAction type="Action">
        <name>refinery</name>
        <actionStatus>http://schema.org/CompletedActionStatus</actionStatus>
      </potentialAction>
      <keywords>TechniqueCode: IR, TechniqueLongDescription: Infrared Spectroscopy (IR)</keywords>
    - <about type="Thing">
        <description>N, N - dimethylacrylamide(DMA) is observed throughout both samples(Tables 1 and 2). The cleaned sample, E18 - 0333 - 02 shows less DMA compared to pre washed sample, E18-0333 - 01.No useful trend was observed in DMA level from “outside” of the roll to the “inside” of the roll.</description>
        <disambiguatingDescription>3M Menonmonie supplied a sample from a piece of neoprene rubber hose.They would like to determine how much N, N Dimethylacrylamide(DMA) has penetrated through the rubber hose. The side that was exposed to DMA has the largest flat surface. The side with the saw marks on it would be
          towards the inside.</disambiguatingDescription>
        <identifier>260233</identifier>
      </about>
    - <creator type="Person">
        <identifier>US263070</identifier>
      </creator>
    - <sourceOrganization type="Organization">
        <name>(US, St Paul) Corporate Research Analytical Laboratory(CRAL)</name>
        <alternateName>CRAL</alternateName>
      </sourceOrganization>
    - <workTranslation type="CreativeWork">
        <url>adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl/cral/automated/Spectropscopy-IRImaging/201186_Samplefile_201808170147_1.csv</url>
        <dateCreated>"2018-09-26T15:17:39.5828728-05:00"</dateCreated>
        <dateModified>"2018-09-26T15:17:39.7636674-05:00"</dateModified>
      </workTranslation>
    - <distribution type="DataDownload">
        <url>https://usw-su1.azuredatacatalog.com</url>
        <contentUrl>adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl/cral/automated/Spectropscopy-IRImaging/201186_Samplefile_201808170147_1.csv</contentUrl>
        <encodingFormat>CSV</encodingFormat>
      </distribution>
    - <distribution type="DataDownload">
        <url>https://usw-su1.azuredatacatalog.com</url>
        <contentUrl>adl://srddac02xeastus2xdlsxd.azuredatalakestore.net/raw/crl/cral/Spectropscopy-IRImaging/201186_Samplefile_201808170147_1.SPA</contentUrl>
        <encodingFormat>SPA</encodingFormat>
      </distribution>
    - <includedInDataCatalog type="DataCatalog">
        <url>https://usw-su1.azuredatacatalog.com</url>
      </includedInDataCatalog>
    - <hasDigitalDocumentPermission>
        <permissionType type="DigitalDocumentPermissionType">http://schema.org/ReadPermission</permissionType>
      - <grantee type="Audience">
          <name>cral - data - platform - dev - team - awesome</name>
          <identifier>5c341e13-dfb3-48fc-a776-3dd9b8763293</identifier>
        </grantee>
      </hasDigitalDocumentPermission>
    - <hasDigitalDocumentPermission>
        <permissionType type="DigitalDocumentPermissionType">http://schema.org/WritePermission</permissionType>
      - <grantee type="Audience">
          <name>cral-readers</name>
          <identifier>64932e12-bca9-224l-n237-212de2492124</identifier>
        </grantee>
      </hasDigitalDocumentPermission>
    </dataset>
 </datasets>
```

References

Google Structured Data Testing Tool. (n.d.). Retrieved from Google Structured Data Testing Tool: https://search.google.com/structured-data/testing-tool

https://schema.org. (n.d.). *about*. Retrieved from https://schema.org: https://schema.org/about

https://schema.org. (n.d.). *Action*. Retrieved from https://schema.org: https://schema.org/Action

https://schema.org. (n.d.). *CreativeWork*. Retrieved from https://schema.org: https://schema.org/CreativeWork

https://schema.org. (n.d.). *DataCatalog*. Retrieved from https://schema.org: https://schema.org/DataCatalog

https://schema.org. (n.d.). *Dataset*. Retrieved from https://schema.org: https://schema.org/Dataset

https://schema.org. (n.d.). *Organization*. Retrieved from https://schema.org: https://schema.org/Organization

https://schema.org. (n.d.). *Person*. Retrieved from https://schema.org: https://schema.org/Person

https://schema.org. (n.d.). *potentialAction*. Retrieved from https://schema.org: https://schema.org/potentialAction

https://schema.org. (n.d.). *thing*. Retrieved from https://schema.org: https://schema.org/Thing

https://www.iso.org/iso-8601-date-and-time-format.html. (n.d.). Retrieved from International Organization for Standardization: https://www.iso.org/iso-8601-date-and-time-format.html

https://www.w3schools.com/xml/xml_dtd_intro.asp. (n.d.). Retrieved from https://www.w3schools.com/xml/xml_dtd_intro.asp: https://www.w3schools.com/xml/xml_dtd_intro.asp https://www.zdnet.com. (2018, 9 5). *Google can now search for datasets first research then the world*. Retrieved from https://www.zdnet.com: https://www.zdnet.com/article/google-can-now-search-for-datasets-first-research-then-the-world/

schema.org. (n.d.). Retrieved from schema.org: https://schema.org/

schema.org. (n.d.). *DataDownload*. Retrieved from https://schema.org/DataDownload: https://schema.org/DataDownload

schema.org. (n.d.). *https://schema.org/Dataset*. Retrieved from Dataset: https://schema.org/Dataset