Assignment 4 Part 1: XML Schema Design

Felicia Liu (liu318@illinois.edu)

Netid: liu318

CS598: Foundations of Data Curation

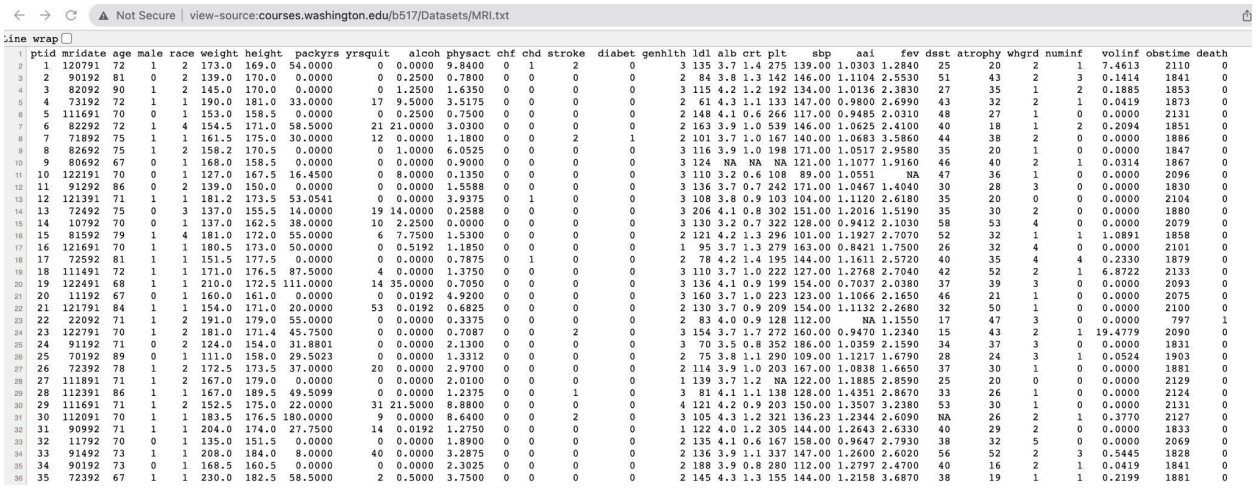
1. [10 points] Choose a document or piece of text that you're interested in from your workplace.

This document can be structured or unstructured. You may choose a text of any sort and any length, as long as it is long enough to meet the following encoding criteria. Be sure to include the text as a separate file in your upload (or if it is online, you may provide a link to it).

The document can be assessed from the below path:

http://courses.washington.edu/b517/Datasets/MRI.txt

The structure of this document is presented as follows:



	ptid	mridate	age	male	race	weight	height	packyrs	yrsquit	alcoh	physact	chf	chd	stroke	diabet	genhlth	ldl	alb	crt	plt	sbp	aai	fev	dsst	atrophy	whgrd	numinf	volinf	obstime	death
2	1	120791	72	1	2	173.0	169.0	54.0000	0	0.0000	9.8400	0	1	2	0	3	135	3.7	1.4	275	139.00	1.0303	1.2840	25	20	2	1	7.4613	2110	0
3	2	90192	81	0	2	139.0	170.0	0.0000	0	0.2500	0.7800	0	0	0	0	2	84	3.8	1.3	142	146.00	1.1104	2.5530	51	43	2	3	0.1414	1841	0
4	3	82092	90	1	2	145.0	170.0	0.0000	0	1.2500	1.6350	0	0	0	0	3	115	4.2	1.2	192	134.00	1.0136	2.3830	27	35	1	2	0.1885	1853	0
5	4	73192	72	1	1	190.0	181.0	33.0000	17	9.5000	3.5175	0	0	0	0	2	61	4.3	1.1	133	147.00	0.9800	2.6990	43	32	2	1	0.0419	1873	0
6	5	111691	70	0	1	153.0	158.5	0.0000	0	0.2500	0.7500	0	0	0	0	2	148	4.1	0.6	266	117.00	0.9485	2.0310	48	27	1	0	0.0000	2131	0
7	6	82292	72	1	4	154.5	171.0	58.5000	21	21.0000	3.0300	0	0	0	0	2	163	3.9	1.0	539	146.00	1.0625	2.4100	40	18	1	2	0.2094	1851	0
8	7	71892	75	1	1	161.5	175.0	30.0000	12	0.0000	1.1800	0	0	2	1	2	101	3.7	1.0	167	140.00	1.0683	3.5860	44	38	2	0	0.0000	1886	0
9	8	82692	75	1	2	158.2	170.5	0.0000	0	1.0000	6.0525	0	0	0	0	3	116	3.9	1.0	198	171.00	1.0517	2.9580	35	20	1	0	0.0000	1847	0
10	9	80692	67	0	1	168.0	158.5	0.0000	0	0.0000	0.9000	0	0	0	0	3	124	NA	NA	NA	121.00	1.1077	1.9160	46	40	2	1	0.0314	1867	0
11	10	122191	70	0	1	127.0	167.5	16.4500	0	8.0000	0.1350	0	0	0	0	3	110	3.2	0.6	108	89.00	1.0551	NA	47	36	1	0	0.0000	2096	0
12	11	91292	86	0	2	139.0	150.0	0.0000	0	0.0000	1.5588	0	0	0	0	3	136	3.7	0.7	242	171.00	1.0467	1.4040	30	28	3	0	0.0000	1830	0
13	12	121391	71	1	1	181.2	173.5	53.0541	0	0.0000	3.9375	0	1	0	0	3	108	3.8	0.9	103	104.00	1.1120	2.6180	35	20	0	0	0.0000	2104	0
14	13	72492	75	0	3	137.0	155.5	14.0000	19	14.0000	0.2588	0	0	0	0	3	206	4.1	0.8	302	151.00	1.2016	1.5190	35	30	2	0	0.0000	1880	0
15	14	10792	70	0	1	137.0	162.5	38.0000	10	2.2500	0.0000	0	0	0	0	3	130	3.2	0.7	322	128.00	0.9412	2.1030	58	53	4	0	0.0000	2079	0
16	15	81592	79	1	4	181.0	172.0	55.0000	6	7.7500	1.5300	0	0	0	0	2	121	4.2	1.3	296	101.00	1.1927	2.7070	52	32	1	1	1.0891	1858	0
17	16	121691	70	1	1	180.5	173.0	50.0000	0	0.5192	1.1850	0	0	0	0	1	95	3.7	1.3	279	163.00	0.8421	1.7500	26	32	4	0	0.0000	2101	0
18	17	72592	81	1	1	151.5	177.5	0.0000	0	0.0000	0.7875	0	1	0	0	2	78	4.2	1.4	195	144.00	1.1611	2.5720	40	35	4	4	0.2330	1879	0
19	18	111491	72	1	1	171.0	176.5	87.5000	4	0.0000	1.3750	0	0	0	0	3	110	3.7	1.0	222	127.00	1.2768	2.7040	42	52	2	1	6.8722	2133	0
20	19	122491	68	1	1	210.0	172.5	111.0000	14	35.0000	0.7050	0	0	0	0	3	136	4.1	0.9	199	154.00	0.7037	2.0380	37	39	3	0	0.0000	2093	0
21	20	11192	67	0	1	160.0	161.0	0.0000	0	0.0192	4.9200	0	0	0	0	3	160	3.7	1.0	223	123.00	1.1066	2.1650	46	21	1	0	0.0000	2075	0
22	21	121791	84	1	1	154.0	171.0	20.0000	53	0.0192	0.6825	0	0	0	0	2	130	3.7	0.9	209	154.00	1.1132	2.2680	32	50	1	0	0.0000	2100	0
23	22	22092	71	1	2	191.0	179.0	55.0000	0	0.0000	0.3375	0	0	0	0	2	83	4.0	0.9	128	112.00	NA	1.1550	17	47	3	0	0.0000	797	1
24	23	122791	70	1	2	181.0	171.4	45.7500	0	0.0000	0.7087	0	0	2	0	3	154	3.7	1.7	272	160.00	0.9470	1.2340	15	43	2	1	19.4779	2090	0
25	24	91192	71	0	2	124.0	154.0	31.8801	0	0.0000	2.1300	0	0	0	0	3	70	3.5	0.8	352	186.00	1.0359	2.1590	34	37	3	0	0.0000	1831	0
26	25	70192	89	0	1	111.0	158.0	29.5023	0	0.0000	1.3312	0	0	0	0	2	75	3.8	1.1	290	109.00	1.1217	1.6790	28	24	3	1	0.0524	1903	0
27	26	72392	78	1	2	172.5	173.5	37.0000	20	0.0000	2.9700	0	0	0	0	2	114	3.9	1.0	203	167.00	1.0838	1.6650	37	30	1	0	0.0000	1881	0
28	27	111891	71	1	2	167.0	179.0	0.0000	0	0.0000	2.0100	0	0	0	0	1	139	3.7	1.2	NA	122.00	1.1885	2.8590	25	20	0	0	0.0000	2129	0
29	28	112391	86	1	1	167.0	189.5	49.5099	0	0.0000	1.2375	0	0	1	0	3	81	4.1	1.1	138	128.00	1.4351	2.8670	33	26	1	0	0.0000	2124	0
30	29	111691	71	1	2	152.5	175.0	22.0000	31	21.5000	8.8800	0	0	0	0	4	121	4.2	0.9	203	150.00	1.3507	3.2380	53	30	1	0	0.0000	2131	0
31	30	112091	70	1	1	183.5	176.5	180.0000	9	0.0000	8.6400	0	0	2	0	3	105	4.3	1.2	321	136.23	1.2344	2.6090	NA	26	2	1	0.3770	2127	0
32	31	90992	71	1	1	204.0	174.0	27.7500	14	0.0192	1.2750	0	0	0	0	1	122	4.0	1.2	305	144.00	1.2643	2.6330	40	29	2	0	0.0000	1833	0
33	32	11792	70	0	1	135.0	151.5	0.0000	0	0.0000	1.8900	0	0	0	0	2	135	4.1	0.6	167	158.00	0.9647	2.7930	38	32	5	0	0.0000	2069	0
34	33	91492	73	1	1	208.0	184.0	8.0000	40	0.0000	3.2875	0	0	0	0	2	136	3.9	1.1	337	147.00	1.2600	2.6020	56	52	2	3	0.5445	1828	0
35	34	90192	73	0	1	168.5	160.5	0.0000	0	0.0000	2.3025	0	0	0	0	2	188	3.9	0.8	280	112.00	1.2797	2.4700	40	16	2	1	0.0419	1841	0
36	35	72392	67	1	1	230.0	182.5	58.5000	2	0.5000	3.7500	0	0	0	0	2	145	4.3	1.3	155	144.00	1.2158	3.6870	38	19	1	1	0.2199	1881	0

2. Evidence of in-depth examination of text document (steps 2 & 4)

Step 2: [20 points] Make an XML DTD for that text. Your DTD should specify the following

- At least 10 different elements
- At least 5 different attributes (you may have multiple attributes per element; not every element requires attributes). For at least 2 attributes, a controlled list of values the attribute may take
- Indicate, of course, whether each element and attribute is required, optional, an either/or, repeated, etc
- Indicate how elements may nest.

Step 4: [10 points] Mark up the text you have chosen according to the DTD you designed. Make sure you include either a <!DOCTYPE> reference to an external DTD file, or an internal DTD within the <!DOCTYPE> element in the XML document.

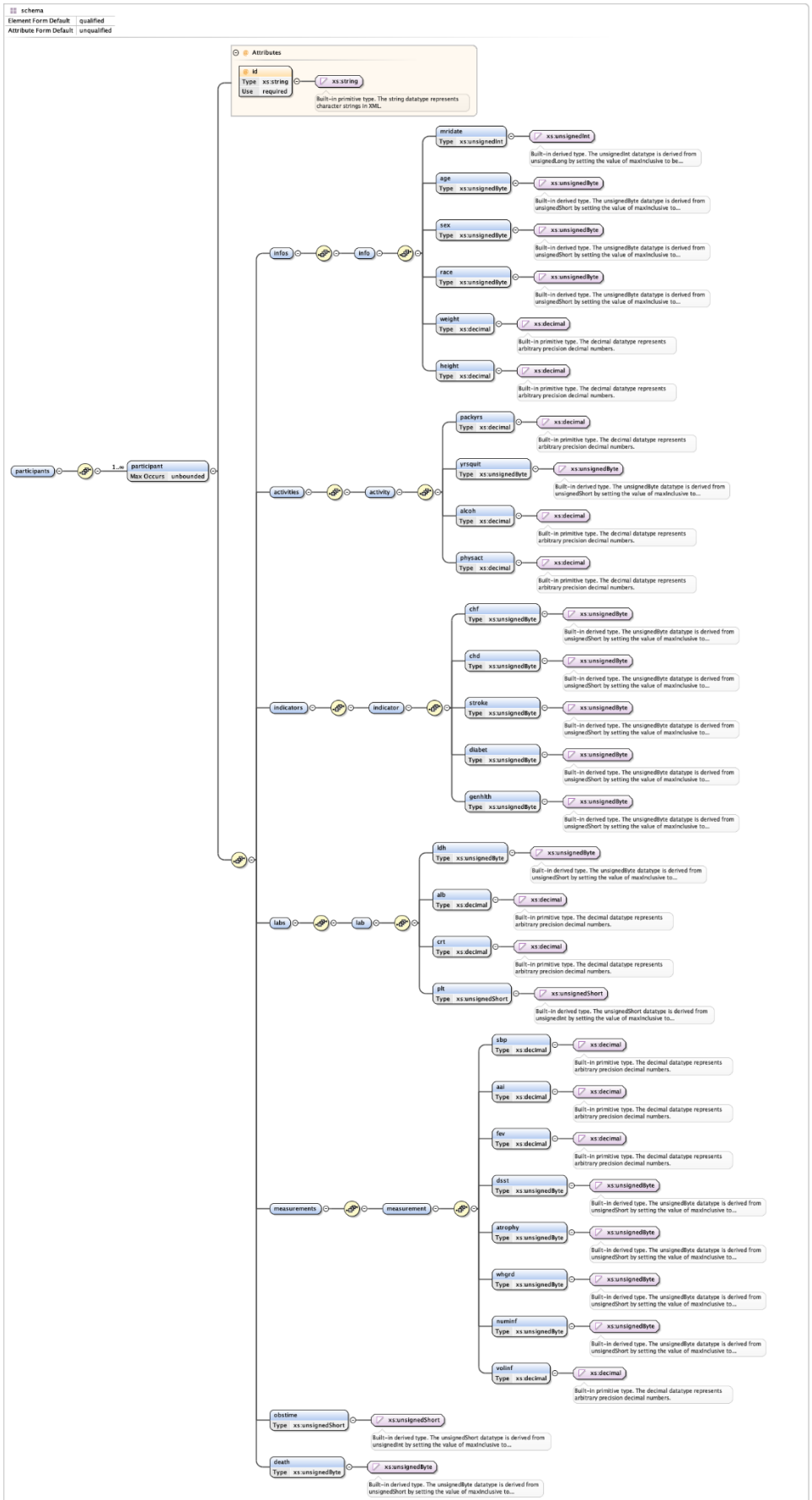
The DTD document is designed manually and presented as below.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT participants (participant+)>
<!ELEMENT participant (infos+,activities+,indicators+,labs+,measurements+,obstime,death)>
<!ATTLIST participant id ID #REQUIRED>
<!ELEMENT infos (info+)>
<!ELEMENT info (mridate,age,sex,race,weight,height)>
<!ELEMENT mridate (#PCDATA)>
<!ELEMENT age (#PCDATA)>
<!ELEMENT sex (#PCDATA)>
<!ELEMENT race (#PCDATA)>
<!ELEMENT weight (#PCDATA)>
<!ELEMENT height (#PCDATA)>
<!ELEMENT activities (activity+)>
<!ELEMENT activity (packyrs,yrsquit,alcoh,physact)>
<!ELEMENT packyrs (#PCDATA)>
<!ELEMENT yrsquit (#PCDATA)>
<!ELEMENT alcoh (#PCDATA)>
<!ELEMENT physact (#PCDATA)>
<!ELEMENT indicators (indicator+)>
<!ELEMENT indicator (chf,chd,stroke,diabet,genhlth)>
<!ELEMENT chf (#PCDATA)>
<!ELEMENT chd (#PCDATA)>
<!ELEMENT stroke (#PCDATA)>
<!ELEMENT diabet (#PCDATA)>
<!ELEMENT genhlth (#PCDATA)>
<!ELEMENT labs (lab+)>
<!ELEMENT lab (ldh,alb,crt,plt)>
<!ELEMENT ldh (#PCDATA)>
<!ELEMENT alb (#PCDATA)>
<!ELEMENT crt (#PCDATA)>
<!ELEMENT plt (#PCDATA)>
<!ELEMENT measurements (measurement+)>
<!ELEMENT measurement (sbp,aai,fev,dsst,atrophy,whgrd,numinf,volinf)>
<!ELEMENT sbp (#PCDATA)>
<!ELEMENT aai (#PCDATA)>
<!ELEMENT fev (#PCDATA)>
<!ELEMENT dsst (#PCDATA)>
<!ELEMENT atrophy (#PCDATA)>
<!ELEMENT whgrd (#PCDATA)>
<!ELEMENT numinf (#PCDATA)>
<!ELEMENT volinf (#PCDATA)>
<!ELEMENT obstime (#PCDATA)>
<!ELEMENT death (#PCDATA)>
```

The XSD design shown below is designed to represent the relationship between various elements and attributes in the XML document. From this XSD design, it can be clearly identified that element “participants” contains the sub-element “participant”, and the sub-element “participant” is mainly divided into 6 parts: element “infos”, “activities”, “indicators”, “labs”, “measurements”, “obstime” and “death”.

In addition, some sub-elements also include other sub-elements. For instance, the element “mridate”, “age”, “sex”, “race”, “weight”, “height” are nested in the element “infos”.

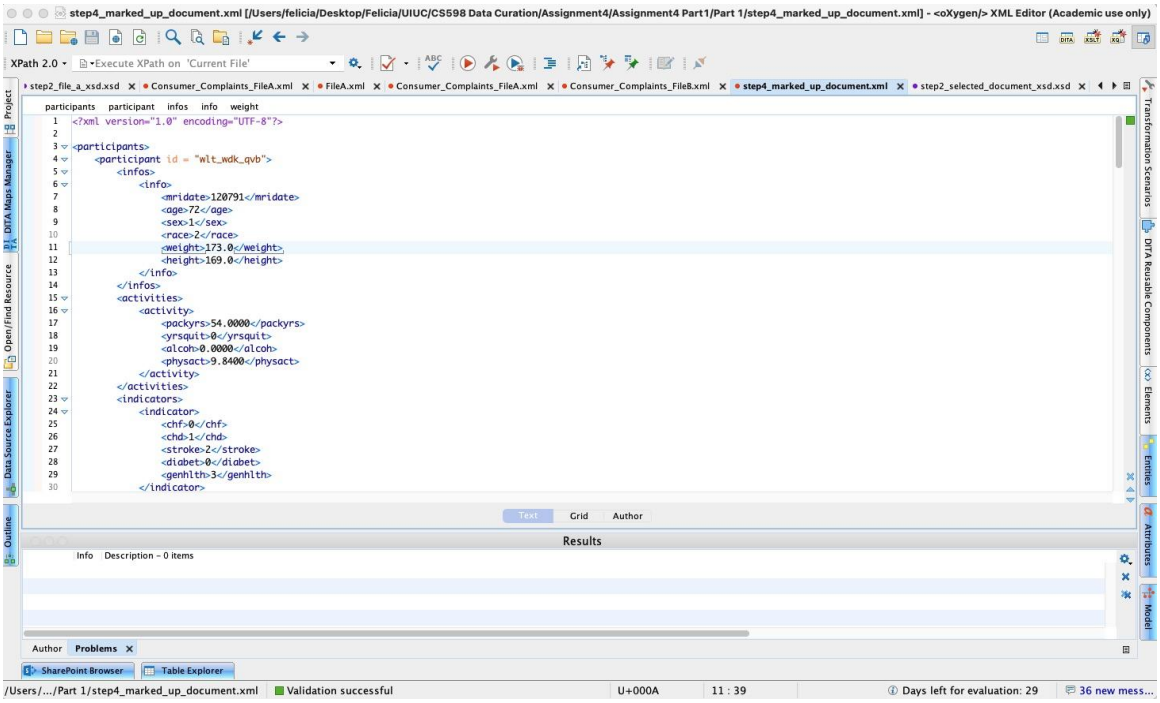
```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="participants">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" name="participant">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="infos">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element name="info">
                      <xs:complexType>
                        <xs:sequence>
                          <xs:element name="mridate" type="xs:unsignedInt" />
                          <xs:element name="age" type="xs:unsignedByte" />
                          <xs:element name="sex" type="xs:unsignedByte" />
                          <xs:element name="race" type="xs:unsignedByte" />
                          <xs:element name="weight" type="xs:decimal" />
                          <xs:element name="height" type="xs:decimal" />
                        </xs:sequence>
                      </xs:complexType>
                    </xs:element>
                  </xs:sequence>
                </xs:complexType>
              </xs:element>
              <xs:element name="activities">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element name="activity">
                      <xs:complexType>
                        <xs:sequence>
                          <xs:element name="packyrs" type="xs:decimal" />
                          <xs:element name="yrsquit" type="xs:unsignedByte" />
                          <xs:element name="alcoh" type="xs:decimal" />
                          <xs:element name="physact" type="xs:decimal" />
                        </xs:sequence>
                      </xs:complexType>
                    </xs:element>
                  </xs:sequence>
                </xs:complexType>
              </xs:element>
              <xs:element name="indicators">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element name="indicator">
                      <xs:complexType>
                        <xs:sequence>
                          <xs:element name="chf" type="xs:unsignedByte" />
                          <xs:element name="chd" type="xs:unsignedByte" />
                          <xs:element name="stroke" type="xs:unsignedByte" />
                          <xs:element name="diabet" type="xs:unsignedByte" />
                          <xs:element name="genhlth" type="xs:unsignedByte" />
                        </xs:sequence>
                      </xs:complexType>
                    </xs:element>
                  </xs:sequence>
                </xs:complexType>
              </xs:element>
              <xs:element name="labs">
                <xs:complexType>
                  <xs:sequence>
                    <xs:element name="lab">
```


schema
Element Form Default qualified
Attribute Form Default unqualified
Attributes
id
Type xs:string
Use required
xs:string
Built-in primitive type. The string datatype represents character strings in XML.
participants
1..∞
participant
Max Occurs unbounded
infos
info
mridate
Type xs:unsignedInt
xs:unsignedInt
Built-in derived type. The unsignedInt datatype is derived from unsignedLong by setting the value of maxInclusive to be...
age
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
sex
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
race
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
weight
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
height
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
activities
activity
packyrs
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
yrsquit
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
alcoh
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
physact
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
indicators
indicator
chf
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
chd
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
stroke
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
diabet
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
genhlth
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
labs
lab
ldh
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
alb
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
crt
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
plt
Type xs:unsignedShort
xs:unsignedShort
Built-in derived type. The unsignedShort datatype is derived from unsignedInt by setting the value of maxInclusive to...
measurements
measurement
sbp
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
aai
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
fev
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
dsst
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
atrophy
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
whgrd
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
numinf
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...
volinf
Type xs:decimal
xs:decimal
Built-in primitive type. The decimal datatype represents arbitrary precision decimal numbers.
obstime
Type xs:unsignedShort
xs:unsignedShort
Built-in derived type. The unsignedShort datatype is derived from unsignedInt by setting the value of maxInclusive to...
death
Type xs:unsignedByte
xs:unsignedByte
Built-in derived type. The unsignedByte datatype is derived from unsignedShort by setting the value of maxInclusive to...

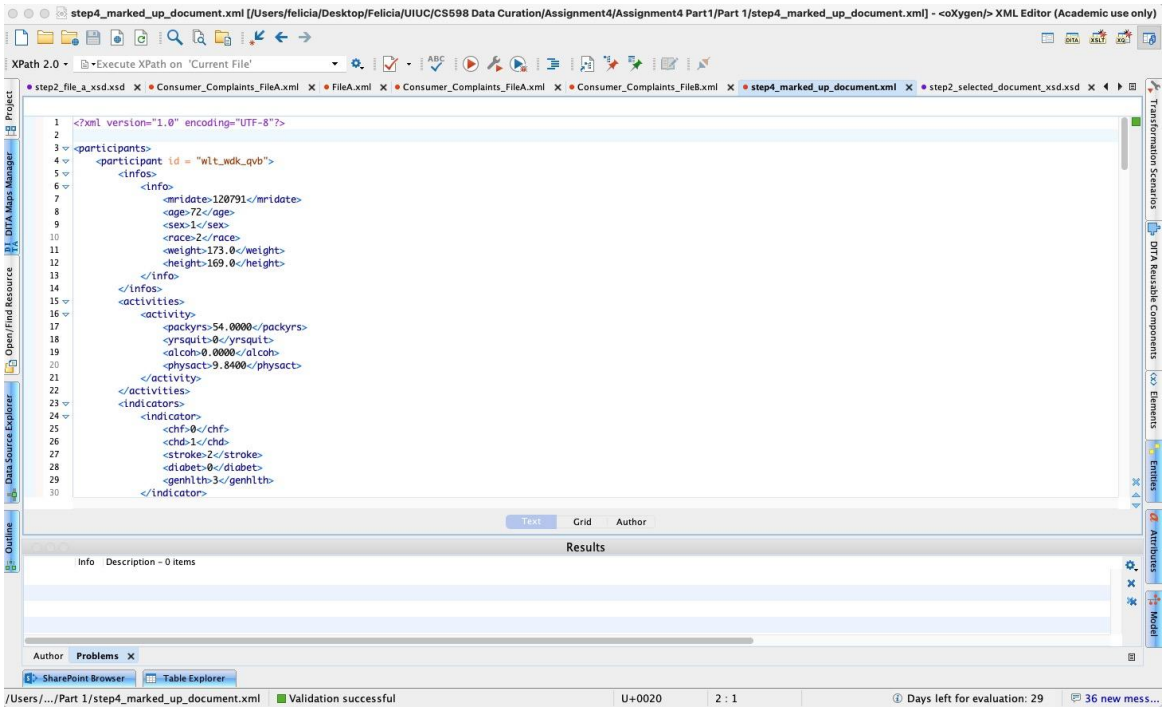
The XML design for the DTD design and XSD schema design is shown as below:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE participants SYSTEM "MRI2data.dtd">
<participants>
  <participant id = "wlt_wdk_qyb">
    <infos>
      <info>
        <mridate>120791</mridate>
        <age>72</age>
        <sex>1</sex>
        <race>2</race>
        <weight>173.0</weight>
        <height>169.0</height>
      </info>
    </infos>
    <activities>
      <activity>
        <packyrs>54.0000</packyrs>
        <yrsquit>0</yrsquit>
        <alcoh>0.0000</alcoh>
        <physact>9.8400</physact>
      </activity>
    </activities>
    <indicators>
      <indicator>
        <chf>0</chf>
        <chd>1</chd>
        <stroke>2</stroke>
        <diabet>0</diabet>
        <genhlth>3</genhlth>
      </indicator>
    </indicators>
    <labs>
      <lab>
        <ldh>135</ldh>
        <alb>3.7</alb>
        <crt>1.4</crt>
        <plt>275</plt>
      </lab>
    </labs>
    <measurements>
      <measurement>
        <sbp>139.00</sbp>
        <aai>1.0303</aai>
        <fev>1.2840</fev>
        <dsst>25</dsst>
        <atrophy>20</atrophy>
        <whgrd>2</whgrd>
        <numinf>1</numinf>
        <volinf>7.4613</volinf>
      </measurement>
    </measurements>
    <obstime>2110</obstime>
    <death>0</death>
  </participant>
```

I used the Oxygen XML editor to validate the DTD design and XML document. The result is shown as “Validation successfully”.



I used the Oxygen XML editor to validate the XSD design and XML document. The result is shown as “Validation successfully”.



Overall, the XML and external DTD documents contain these elements:

- Element “participants” contains sub-element “participant”, more sub-elements are nested into element “participant” to display the data
- Required attlist “participant id” is generated to record each attribute
- Element “participant” contains sub-elements “infos”,”activities”,”indicators”,”labs”,”measurements”,”obstime”,”death”
- Element “infos” contains sub-element “info”, and element “info” can be divided as sub-element “mridate”, “age”, “sex”, “race”, “weight”, “height”
- Element “activities” contains sub-element “activity”, and element “activity” can be divided as sub-element “packyrs”, “yrsquit”, “alcoh”, “physact”
- Element “indicators” contains sub-element “indicator”, and elements “chf”, “chd”, “stroke”, “diabet”, “genhlth” are nested into element “indicator”
- Element “labs” contains sub-element “lab”, and elements “ldh”, “alb”, “crt”, “plt” are nested into element “lab”
- Element “measurements” contains sub-element “measurement”, and elements “sbp”, “aai”, “fev”, “dsst”, “atrophy”, “whgrd”, “numinf”, “volinf” are nested into element “measurement”
- Elements “obstime” and “death” are nested into element “participant”

3. Evidence of understanding data representation, schemas and independence (steps 3)

Step 3: [10 points] Write prose documentation for each element, attribute, and attribute value. For e.g. explain what type of data they have, if there is a default value then how did you decide to choose one etc.

In order to better understand the elements and attributes, and better design the XML DTD file, the text document was further analyzed and the following table was presented to better comprehend the structure of this document:

Element	Description	Data Type	Nullable
ptid	Participant identification number	Integer	Not null
mridate	The date on which the participant underwent MRI scan in MMDDYY format	Integer	Not null
age	Participant age at time of MRI (years)	Integer	Not null
sex	Indicator of participant's gender	Boolean	Not null
race	Indicator of participant's race	Integer	Not null
weight	Participant weight at time of MRI	Float	Not null
height	Participant height at time of MRI	Float	Not null
packyrs	Participant smoke history in pack years	Float	Not null
yrsquit	Number of years since quitting smking	Integer	Not null

alcoh	Average alcohol intake for the participant for the two weeks prior to MRI	Float	Not null
physact	Physical activity of the participant for the week prior to MRI	Float	Not null
chf	Indicator of whether the participant had been diagnosed with congestive heart failure prior to MRI	Integer	Not null
chd	Indicator of whether the participant had been diagnosed with coronary heart failure prior to MRI	Integer	Not null
stroke	Indicator of whether the participant had been diagnosed with cerebrovascular event prior to MRI	Integer	Not null
diabet	Indicator of whether the participant had been diagnosed with diabetes prior to MRI	Integer	Not null
genhlth	An indicator of the participant's view of his/her own health	Integer	Not null
ldh	A laboratory measure of a certain kind of cholesterol in the participant's blood at the time of MRI	Integer	Null
alb	A laboratory measure of a certain kind of protein in the participant's blood at the time of MRI	Float	Null
crt	A laboratory measure of creatinine in the participant's blood at the time of MRI	Float	Null

plt	A laboratory measure of the number of platelets circulating in the participant's blood at the time of MRI	Integer	Null
sbp	A measurement of the participant's systolic blood pressure in his/her arm at the time of MRI	Float	Null
aai	The ratio of systolic blood pressure measured in the participant's ankle at the time of MRI	Float	Null
fev	A measure of forced expiratory volume in the participant at the time of MRI	Float	Null
dsst	A measure of cognitive function for the participant at the time of MRI	Integer	Null
atrophy	A measure of global brain atrophy detected on MRI	Integer	Null
whgrd	A measure of white matter changes detected on MRI	Integer	Null
numinf	A count of number of distinct religious identified on MRI scan which were suggestive of infarcts	Integer	Not null
volinf	A measure of the total volume of the infarct-like lesions found on MRI scan	Float	Not null
obstime	The total time that the participant was observed on study between the date of MRI and death or 9/16/1997, whichever came first	Integer	Not null
death	An indicator that the participant was observed to die while on study	Integer	Not null

The XML document is designed manually and presented as below. To simplify the design, the XML document only includes the top 5 attributes in the dataset (links: view-source:http://courses.washington.edu/b517/Datasets/MRI.txt) instead of presenting the whole 735 attributes.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE participants SYSTEM "MRI2data.dtd">
<participants>
  <participant id = "wlt_wdk_qyb">
    <infos>
      <info>
        <mridate>120791</mridate>
        <age>72</age>
        <sex>1</sex>
        <race>2</race>
        <weight>173.0</weight>
        <height>169.0</height>
      </info>
    </infos>
    <activities>
      <activity>
        <packyrs>54.0000</packyrs>
        <yrsquit>0</yrsquit>
        <alcoh>0.0000</alcoh>
        <physact>9.8400</physact>
      </activity>
    </activities>
    <indicators>
      <indicator>
        <chf>0</chf>
        <chd>1</chd>
        <stroke>2</stroke>
        <diabet>0</diabet>
        <genhlth>3</genhlth>
      </indicator>
    </indicators>
    <labs>
      <lab>
        <ldh>135</ldh>
        <alb>3.7</alb>
        <crt>1.4</crt>
        <plt>275</plt>
      </lab>
    </labs>
    <measurements>
      <measurement>
        <sbp>139.00</sbp>
        <aai>1.0303</aai>
        <fev>1.2840</fev>
        <dsst>25</dsst>
        <atrophy>20</atrophy>
        <whgrd>2</whgrd>
        <numinf>1</numinf>
        <volinf>7.4613</volinf>
      </measurement>
    </measurements>
    <obstime>2110</obstime>
    <death>0</death>
  </participant>
```

4. Discussion of curation activities, needs, and decisions (step 5)

Step 5. [25 points] Write a narrative about this process, answering the following reflection questions:

Q:

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?

A:

The text document I selected is a structured dataset, which has already contained the dataset title like “mridate”, “age”, “race”, “packyrs”, “yrsquit”, and also has contained the attribute values like “120791”, “1”, “173.0”, “169.0” etc.,

To better represent the data, I defined an element “participants”, and all the attribute values are included in the element “participants”. What is more, I have grouped several elements and made these elements nested into another element. To be specific, some information like “mridate”, “age”, “sex”, “race”, “weight”, “height” are pointed to the participant’s personal information, as a consequence, I came up with an element “infos” and the elements like “mridate”, “age”, “sex”, “race”, “weight”, “height” are all nested into the element “infos”.

For another example, elements like “chf”, “chd”, “stroke”, “diabet”, “genhlth” are all indicators that evaluate the participant’s health situation. The dataset uses different numbers like 0,1,2 to indicate the participant’s health. Hence, I concluded them as “indicators” and leveraged the element “indicator” to summarize this information.

Q:

What were the hardest decisions you had to make in this design process

A:

The hardest part is to decide the “parent” elements and “child” elements. To begin with, I determined not to group the information and put all the elements under the one single element “participants”. However, the XML design will be messy and not clear if all the elements are considered as “parent” elements and no “child” element is presented. Since the dataset presents all varieties of information in different columns, and as a matter of fact, some information is correlated with each other. Hence, I have decided to group some elements and make them nested into another element. Based on my initial design, all the elements like “mridate”, “age”, “sex”, “race”, “weight”, “height”, “packyrs”, “yrsquit”, “alcoh”, “physact” are all under element “participants”, and comparing with this design, I divided the 30 elements into 8 groups.

Q:

How does your DTD design support data independence

A:

The DTD design shows what information is stored in the database logically and how the data is managed inside the database.

The DTD is liberated from the actual attribute values, therefore, if the user needs to understand the structure of the database, or needs to make some modifications of the elements, he/she can leverage this DTD design to understand the dataset without making actual changes.

Q:

How may your DTD design support the overarching goals of data curation (revisit objectives and activities of Week 1). Discuss at least **10** objectives, how many of you were able to achieve and how many of them were not able to.

A:

The DTD design clearly defines the XML schemas and it has successfully supported the overarching goals of data curation. After revisiting the objectives and activities of data curation and comparing the concepts with the DTD design, the following curatorial activities have been satisfied:

- **Collection**: Support the collection and acquisition of data.
- **Organization**: Employ an appropriate data model and use appropriate standards
- **Storage**: Support reliable and effective storage
- **Preservation**: Ensure that the data can be understandable and usable in the future.
- **Discoverability**: Support the ability to search for and locate relevant data
- **Access**: Support the ability to retrieve and distribute data
- **Workflow**: Support the ability to systematize data workflows
- **Identification**: Support the ability to identify, authenticate, and validate data
- **Integration**: Support integration of data from different sources using different data model

The DTD design provides information like how the data is stored in the document, what kind of data is stored in the document, the workflow to generate the XML file etc. Therefore, from the DTD design, the user can clearly understand the data model and workflow of the XML document, leading to the achievements of these curatorial activities.

However, some curatorial activities still cannot be achieved:

- **Reformatting:** Support reformatting for use by different tools or to match new format standards
- **Reproducibility:** Support ability to reproduce results, ensuring scientific validity and reliability
- **Sharing:** Support sharing data between researchers, teams, and institutions
- **Communication:** Support representation, publishing, and visualizations that provide insight
- **Provenance:** Support identifying what inputs, processes, and calculations are responsible for data values
- **Modification:** Support management of corrections and updates
- **Compliance:** Ensure compliance to legal, regulatory, and local policy requirements
- **Security:** Ensure that data is secure from tampering or inappropriate access and distribution

From the DTD design, we are unable to support identifying the accuracy of each attribute value and making corrections or updates if the attribute values are incorrect. Additionally, merely by understanding the DTD design, it is unlikely to judge if the document satisfies the compliance requirements. Hence, some curatorial activities are still not satisfied.

Q:

What are the design's pros and cons? Write at least 1 pro and 1 con.

A:

Pros:

- A XML document and an external DTD document are generated, what is more, a XSD XML schema design is generated. It allows the user to easily validate the XML file using the DTD and schema.
- The elements are clearly defined and presented. Several elements are nested into specific elements, reducing the redundancy of the XML syntax and making the XML file easier to read.

Cons:

- The text document I selected does not contain an array, however, if it contains an array, the XML DTD design cannot support storing this data type.
- To simplify the XML design, I only included 5 attributes. However, the original text document contains more than 700 attributes, yet when it comes to large volumes of data, there exists redundancy in syntax and the XML programming language will be difficult to read.

4. Overall quality analysis and completeness

Overall, the XML design, DTD design and XSD design clearly indicate how the data is represented and the designs ensure data independence. The design is readable and understandable, since by simply reading the DTD file and XSD design, we can comprehend what elements are included and the data type of the elements.
However, there are still several improvements that can be conducted to implement the design and analysis:

- Due to time limitation, the XML and DTD design are manually generated, hence, only top 5 attributes of the text document are included in the XML file. Based on relevant papers and materials, we can also design C++, Java and Python scripts to generate the XML file, so that we are able to include all the attributes and validate the attributes.
- The relationship between “parent” elements and sub-elements are manually designed. In total, the 30 elements are divided into 8 groups and we have 3 layers(e.g., element “participants” - element “infos” - element “age”) in the design. We can design more layers, or divide into less groups to make the design clearer.