XML Schema Design

SHRASHTI SINGHAL
Data Curation- IS 531
Department of Information Science
The University of Illinois at Urbana Champaign

Abstract

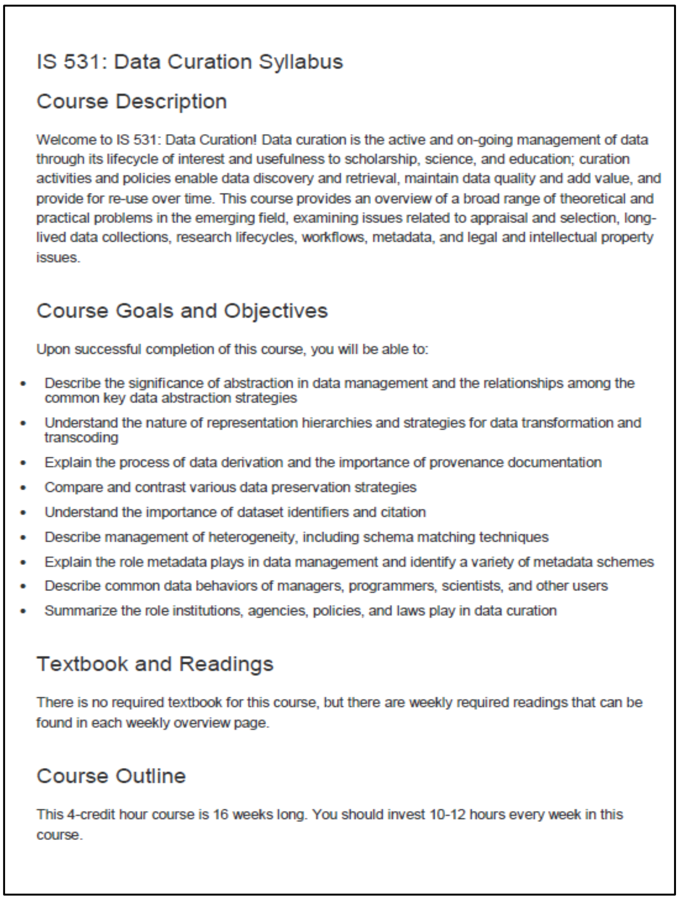
The aim is to write a XML Schema for a regular text document. The digital format of this document could be anything. It can be a docx, pdf, txt etc. The document could be structured or unstructured. The goal is to write a schema that best represents the document, in particular its important components, highlights and purpose. These will include data curation activities. To develop a schema of such a document would need the analysis of this document. The analysis includes categorization of the document into elements, child elements and attributes, and also defining constraints and default values for them. Before designing any logical schema, it is suggested to design the conceptual schema for the same. We would need to execute few data cleaning activities, if required to make the document more readable and usable. We would finally then document the process of data cleaning, analysis, conceptual schema, XML and DTD Schema, Errors occurred during the process and decisions made.

I. Document

The document selected for this purpose is a pdf file of Syllabus of Course –Foundation of Data Curation from University- University of Illinois at Urbana Champaign. The document is six pages long. To pose a challenge, the documents contain tables, headings, sub-headings, paragraphs etc. This document is mostly a structured document, arranged in a hierarchical order of headings and sub headings. Though, at few places the document doesn't follow this order. Examples, there are random notes at places, which doesn't fall under any headings. One such example is at the bottom of the document which says, "Syllabus are subjected to change". Keeping in mind the structure of this document, we have to wisely make the decision of categorizing the documents in elements and attributes and similarly take decisions on their constraints, types and default values.

The document is available at https://cs.illinois.edu/sites/default/files/docs/syllabi/CS598_IS531_DataCuration.pdf .

Below are the screenshots of the document. There are 6 screenshots of the six pages of the document.



IS 531: Data Curation Syllabus

Course Description

Welcome to IS 531: Data Curation! Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities and policies enable data discovery and retrieval, maintain data quality and add value, and provide for re-use over time. This course provides an overview of a broad range of theoretical and practical problems in the emerging field, examining issues related to appraisal and selection, long-lived data collections, research lifecycles, workflows, metadata, and legal and intellectual property issues.

Course Goals and Objectives

Upon successful completion of this course, you will be able to:

- Describe the significance of abstraction in data management and the relationships among the common key data abstraction strategies
- Understand the nature of representation hierarchies and strategies for data transformation and transcoding
- Explain the process of data derivation and the importance of provenance documentation
- Compare and contrast various data preservation strategies
- Understand the importance of dataset identifiers and citation
- Describe management of heterogeneity, including schema matching techniques
- Explain the role metadata plays in data management and identify a variety of metadata schemes
- Describe common data behaviors of managers, programmers, scientists, and other users
- Summarize the role institutions, agencies, policies, and laws play in data curation

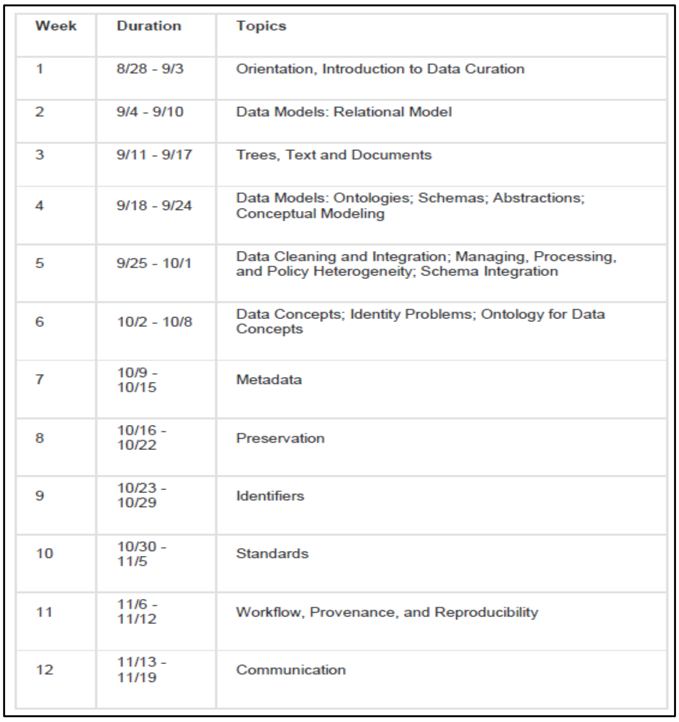
Textbook and Readings

There is no required textbook for this course, but there are weekly required readings that can be found in each weekly overview page.

Course Outline

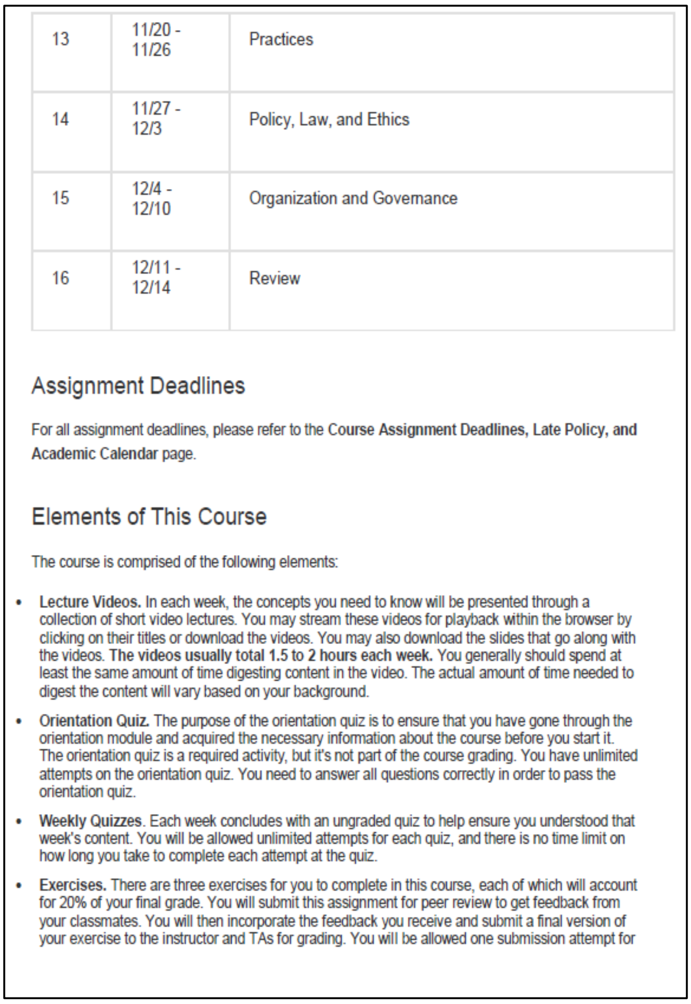
This 4-credit hour course is 16 weeks long. You should invest 10-12 hours every week in this course.

Image 1: Document Page 1



Week	Duration	Topics
1	8/28 - 9/3	Orientation, Introduction to Data Curation
2	9/4 - 9/10	Data Models: Relational Model
3	9/11 - 9/17	Trees, Text and Documents
4	9/18 - 9/24	Data Models: Ontologies; Schemas; Abstractions; Conceptual Modeling
5	9/25 - 10/1	Data Cleaning and Integration; Managing, Processing, and Policy Heterogeneity; Schema Integration
6	10/2 - 10/8	Data Concepts; Identity Problems; Ontology for Data Concepts
7	10/9 - 10/15	Metadata
8	10/16 - 10/22	Preservation
9	10/23 - 10/29	Identifiers
10	10/30 - 11/5	Standards
11	11/6 - 11/12	Workflow, Provenance, and Reproducibility
12	11/13 - 11/19	Communication

Image 2: Document Page 2



13	11/20 - 11/26	Practices
14	11/27 - 12/3	Policy, Law, and Ethics
15	12/4 - 12/10	Organization and Governance
16	12/11 - 12/14	Review

Assignment Deadlines

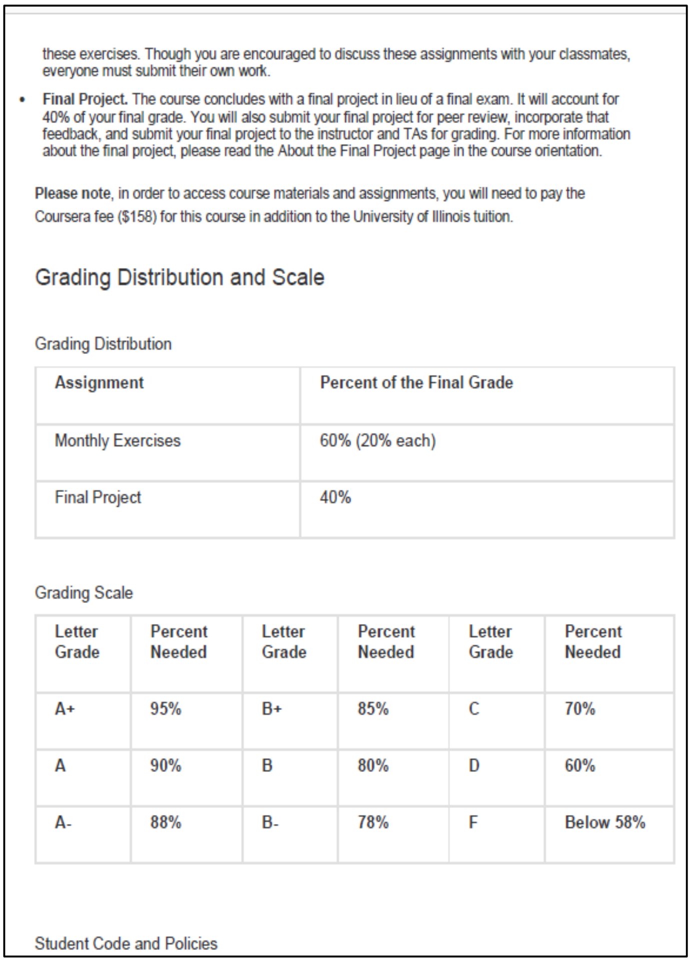
For all assignment deadlines, please refer to the **Course Assignment Deadlines, Late Policy, and Academic Calendar** page.

Elements of This Course

The course is comprised of the following elements:

- **Lecture Videos.** In each week, the concepts you need to know will be presented through a collection of short video lectures. You may stream these videos for playback within the browser by clicking on their titles or download the videos. You may also download the slides that go along with the videos. **The videos usually total 1.5 to 2 hours each week.** You generally should spend at least the same amount of time digesting content in the video. The actual amount of time needed to digest the content will vary based on your background.
- **Orientation Quiz.** The purpose of the orientation quiz is to ensure that you have gone through the orientation module and acquired the necessary information about the course before you start it. The orientation quiz is a required activity, but it's not part of the course grading. You have unlimited attempts on the orientation quiz. You need to answer all questions correctly in order to pass the orientation quiz.
- **Weekly Quizzes.** Each week concludes with an ungraded quiz to help ensure you understood that week's content. You will be allowed unlimited attempts for each quiz, and there is no time limit on how long you take to complete each attempt at the quiz.
- **Exercises.** There are three exercises for you to complete in this course, each of which will account for 20% of your final grade. You will submit this assignment for peer review to get feedback from your classmates. You will then incorporate the feedback you receive and submit a final version of your exercise to the instructor and TAs for grading. You will be allowed one submission attempt for

Image 3: Document Page 3



these exercises. Though you are encouraged to discuss these assignments with your classmates, everyone must submit their own work.

- **Final Project.** The course concludes with a final project in lieu of a final exam. It will account for 40% of your final grade. You will also submit your final project for peer review, incorporate that feedback, and submit your final project to the instructor and TAs for grading. For more information about the final project, please read the About the Final Project page in the course orientation.

Please note, in order to access course materials and assignments, you will need to pay the Coursera fee ($158) for this course in addition to the University of Illinois tuition.

Grading Distribution and Scale

Grading Distribution

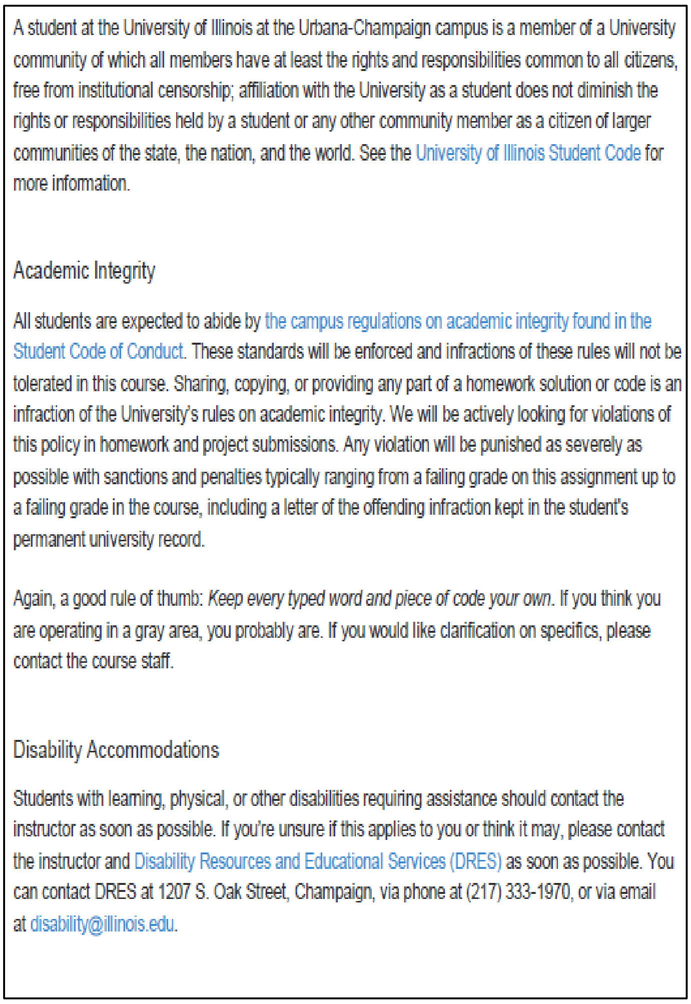
Assignment	**Percent of the Final Grade**
Monthly Exercises	60% (20% each)
Final Project	40%

Grading Scale

Letter Grade	**Percent Needed**	**Letter Grade**	**Percent Needed**	**Letter Grade**	**Percent Needed**
A+	95%	B+	85%	C	70%
A	90%	B	80%	D	60%
A-	88%	B-	78%	F	Below 58%

Student Code and Policies

Image 4: Document Page 4



A student at the University of Illinois at the Urbana-Champaign campus is a member of a University community of which all members have at least the rights and responsibilities common to all citizens, free from institutional censorship; affiliation with the University as a student does not diminish the rights or responsibilities held by a student or any other community member as a citizen of larger communities of the state, the nation, and the world. See the University of Illinois Student Code for more information.

Academic Integrity

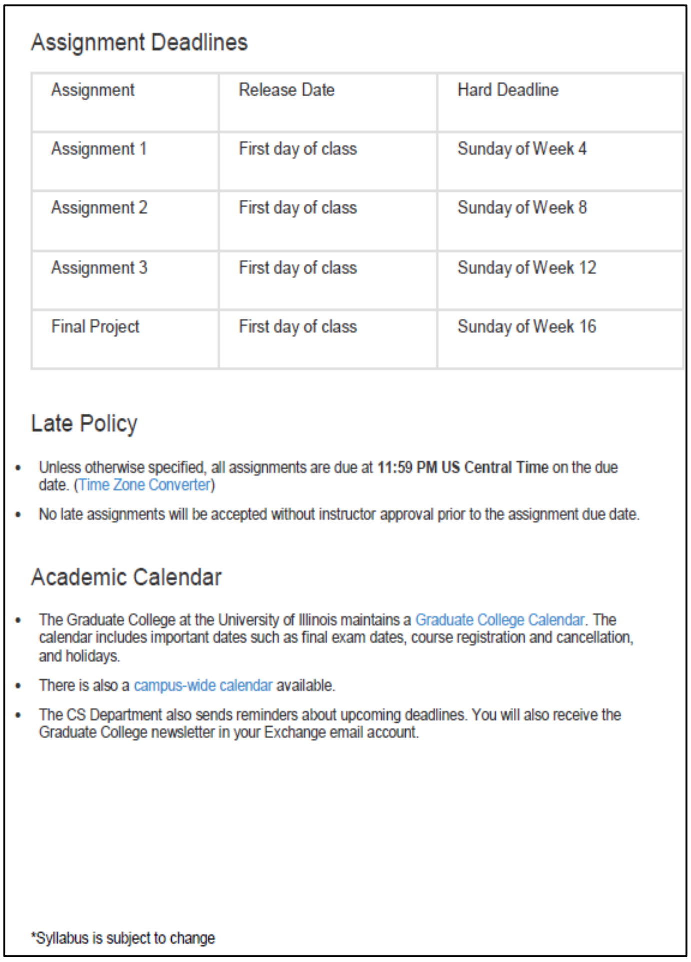
All students are expected to abide by the campus regulations on academic integrity found in the Student Code of Conduct. These standards will be enforced and infractions of these rules will not be tolerated in this course. Sharing, copying, or providing any part of a homework solution or code is an infraction of the University's rules on academic integrity. We will be actively looking for violations of this policy in homework and project submissions. Any violation will be punished as severely as possible with sanctions and penalties typically ranging from a failing grade on this assignment up to a failing grade in the course, including a letter of the offending infraction kept in the student's permanent university record.

Again, a good rule of thumb: *Keep every typed word and piece of code your own*. If you think you are operating in a gray area, you probably are. If you would like clarification on specifics, please contact the course staff.

Disability Accommodations

Students with learning, physical, or other disabilities requiring assistance should contact the instructor as soon as possible. If you're unsure if this applies to you or think it may, please contact the instructor and Disability Resources and Educational Services (DRES) as soon as possible. You can contact DRES at 1207 S. Oak Street, Champaign, via phone at (217) 333-1970, or via email at disability@illinois.edu.

Image 5: Document Page 5



Assignment Deadlines

Assignment	Release Date	Hard Deadline
Assignment 1	First day of class	Sunday of Week 4
Assignment 2	First day of class	Sunday of Week 8
Assignment 3	First day of class	Sunday of Week 12
Final Project	First day of class	Sunday of Week 16

Late Policy

- Unless otherwise specified, all assignments are due at **11:59 PM US Central Time** on the due date. (Time Zone Converter)
- No late assignments will be accepted without instructor approval prior to the assignment due date.

Academic Calendar

- The Graduate College at the University of Illinois maintains a Graduate College Calendar. The calendar includes important dates such as final exam dates, course registration and cancellation, and holidays.
- There is also a campus-wide calendar available.
- The CS Department also sends reminders about upcoming deadlines. You will also receive the Graduate College newsletter in your Exchange email account.

*Syllabus is subject to change

Image 6: Document Page 6

II. Analysis:

II.I ELEMENTS: The document can be divided into 15 main categories. The 11 main headings are evident by their higher font size, moreover these headings do the justice to be the elements as these are the broad categories of this document. These elements are:-

1. TITLE
2. COURSE_DESCRIPTION
3. COURSE_GOALS_AND_OBJECTIVES
4. TEXTBOOKS_AND_READINGS
5. COURSE_OUTLINE
6. ASSIGNMENTS_DEADLINES
7. ELEMENTS_OF_THIS_COURSE
8. GRADING_DISTRIBUTION_AND_SCALE
9. DEADLINES
10. LATE_POLICY
11. CALENDAR

The below headings are not the main headings of the course syllabus but doesn't share any main heading, therefore, we will add these to the elements list as well.

12. CODE_POLICIES
13. INTEGRITY
14. DISABILITY
15. END_NOTE

II.II. CHILD ELEMENTS: There is always a confusion in addressing a data as child elements or attributes. The child elements are same as elements and have all the properties of the main elements, the only difference is that they follow a hierarchical structure or has a parent element.

The below could classify as child elements, as these contain further sub divisions among them. If we classify them as attributes, their values would not be available for expansion.

The table under COURSE_OUTLINE, the 5 sub-headings under ELEMENTS_OF_THIS_COURSE, a note under ELEMENTS_OF_THIS_COURSE, the 2 tables under GRADING_DISTRIBUTION_AND_SCALE, along with their titles marked as child elements and the table under ASSIGNMENTS_DEADLINES

1. TABLE
2. VIDEOS
3. ORIENTATION_QUIZ
4. WEEKLY_QUIZZES
5. EXCERCISES
6. FINAL_PROJECT
7. NOTE
8. GRADING_DISTRIBUTION
9. GRADING_SCALE
10. DEADLINES_TABLE

There is one table each under the sub headings (child elements) GRADING_SCALE and DEADLINES_TABLE.

11. TABLE2
12. TABLE3

We need to make tables as elements because, these tables contain further sub divisions. There are more values in every table in its rows and columns. If whole table id defined as an attribute, we would not be able to classify it any further. Moreover, the table would just be a plain text not a computational value. Therefore, we made a point to categorize all the tables as elements.

We have classified sub-headings as child elements so that a distinction can be made between sub-headings and the value of the sub-heading and paragraphs, as we intend to classify paragraph values and notes under sub headings as attributes. Moreover, the order of the sub-headings might be important, and attributes cannot be ordered. Therefore, we classify subheadings as elements (child elements) as well.

II.III. ATTRIBUTES: The values under the tables columns are selected as the attributes. These values are single values, which do not need expansion. Categorizing data with multiple values and more child elements as elements become necessary. But here as these values are single values, we can safely treat these values as the attributes.

1. WEEK
2. DURATION
3. TOPICS
4. ASSIGNMENT1
5. PERCENT_OF_THE_FINAL_GRADE
6. LETTER_GRADE
7. PERCENT_NEEDED
8. ASSIGNMENT2
9. RELEASE_DATE
10. HARD_DEADLINE

III. TREE STRUCTURE/ CONCEPTUAL SCHEMA

Before developing a XML, based upon our analysis of elements and attributes, we have designed a conceptual Schema.

All the elements (main or child) are shown in Ellipse while attributes are shown in Rectangle boxes.

Elements: 15 (headings) elements are marked in Blue Ellipses.

Child Elements: Sub-Headings of the main elements and direct tables of elements are marked in Green Ellipses.

Child Elements of Child Elements: 2 Child elements have tables as their Childs. These are also categorised as elements as these are tables, tables have multiple values and they fit to be an element. These are marked in Green Ellipses.

Attributes: Attributes of Elements and Child elements are marked in orange rectangles. These are single valued data often with constraints.

Root Element: To develop a XML, we must have root element at the top of the schema. This seems to be a necessary requirement for an XML Document. As we can see, out document doesn't have a root element. I considered to categorize TITLE as the root element of the document, but COURSE_DESCRIPTION, TEXTBOOKS, DISABLITY etc. doesn't fit to be a sub-heading/Child element of the element TITLE. Therefore we developed a root element, which can be a best fit as a parent of all the elements. We named this element as SYLLABUS. This element is marked in yellow rhombus.

NOTE: Child elements, Child elements of Child elements are not a standard terminology. We are naming them so for easy recognition and better understanding of differences. This nomenclature is followed throughout this paper.

A tree structure, conceptual schema, is developed for understanding purposes. From the below image, the relationship between different elements and attributes can be comprehended properly. A legend at the bottom left of the image is provided for reference.

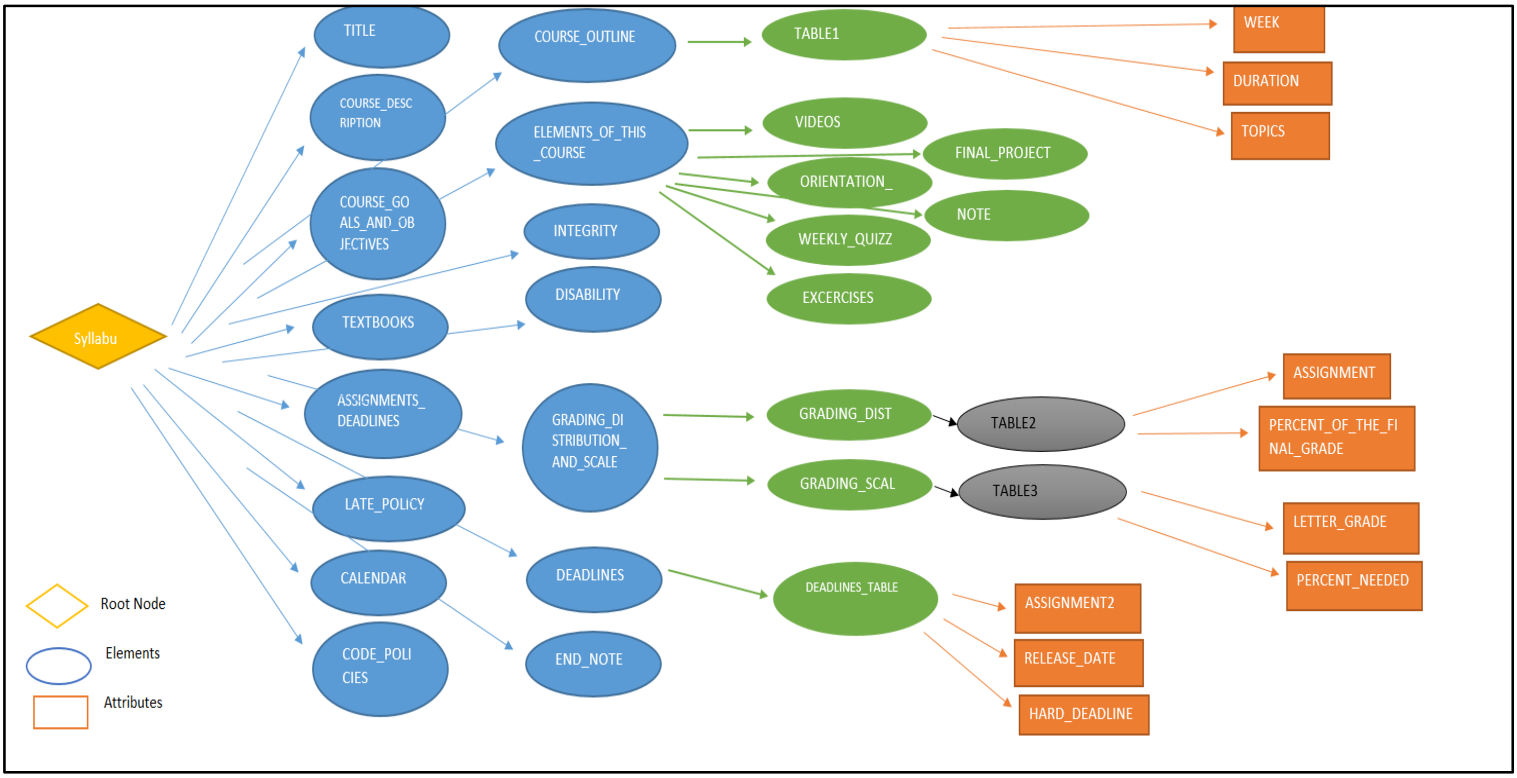


Image 7: Tree Diagram- Conceptual Schema

IV. Designing XML

XML document is designed by hand. The entire PDF file is divided into a hierarchical structure,

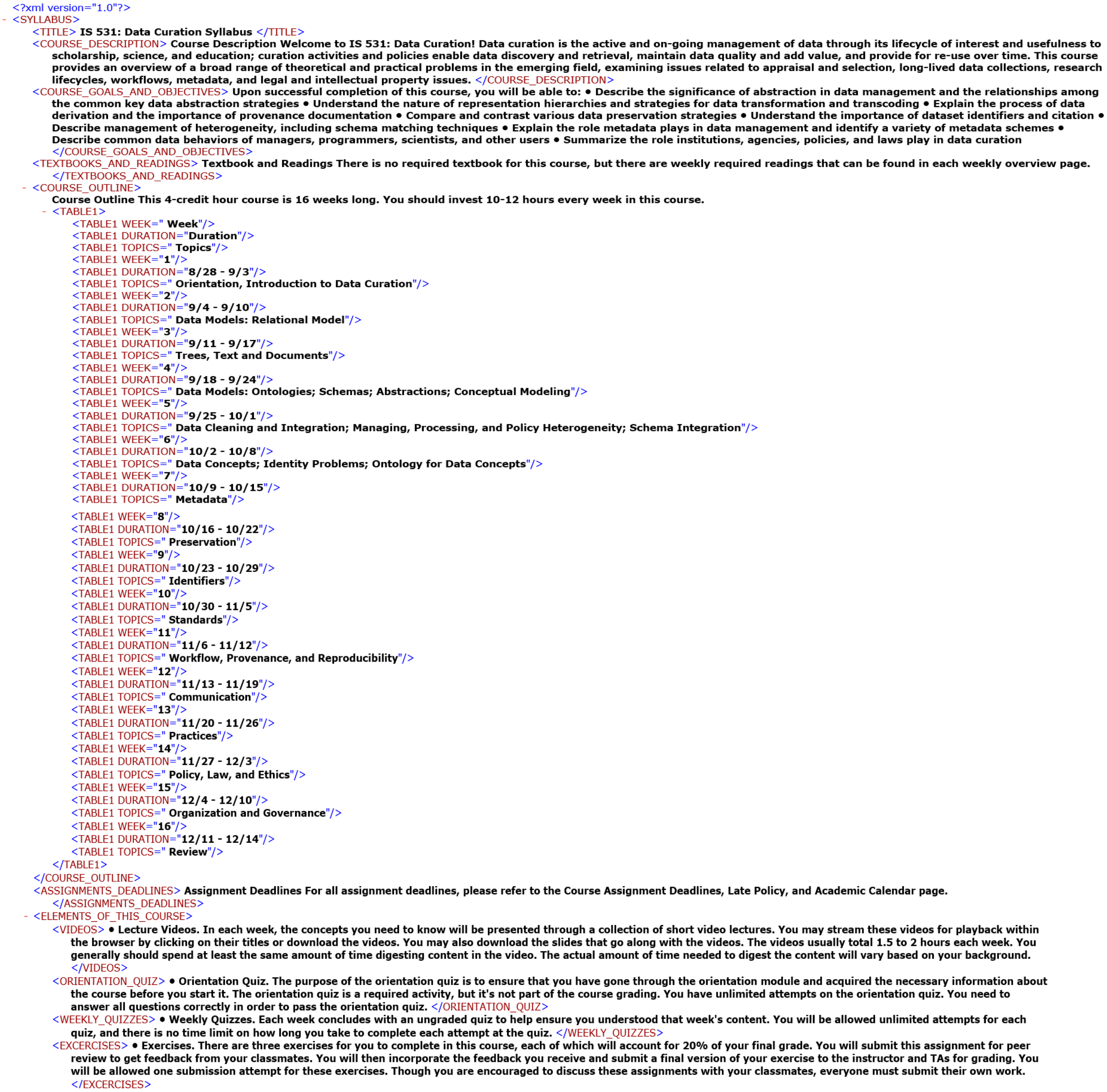
The name of the document is the root element. Under the root element, there are the headings of the PDF Document. Under each heading, there are subheadings, or tables or values of the headings.

- The document is started with the XML version definition.
- The first definition is of the root element-SYLLABUS

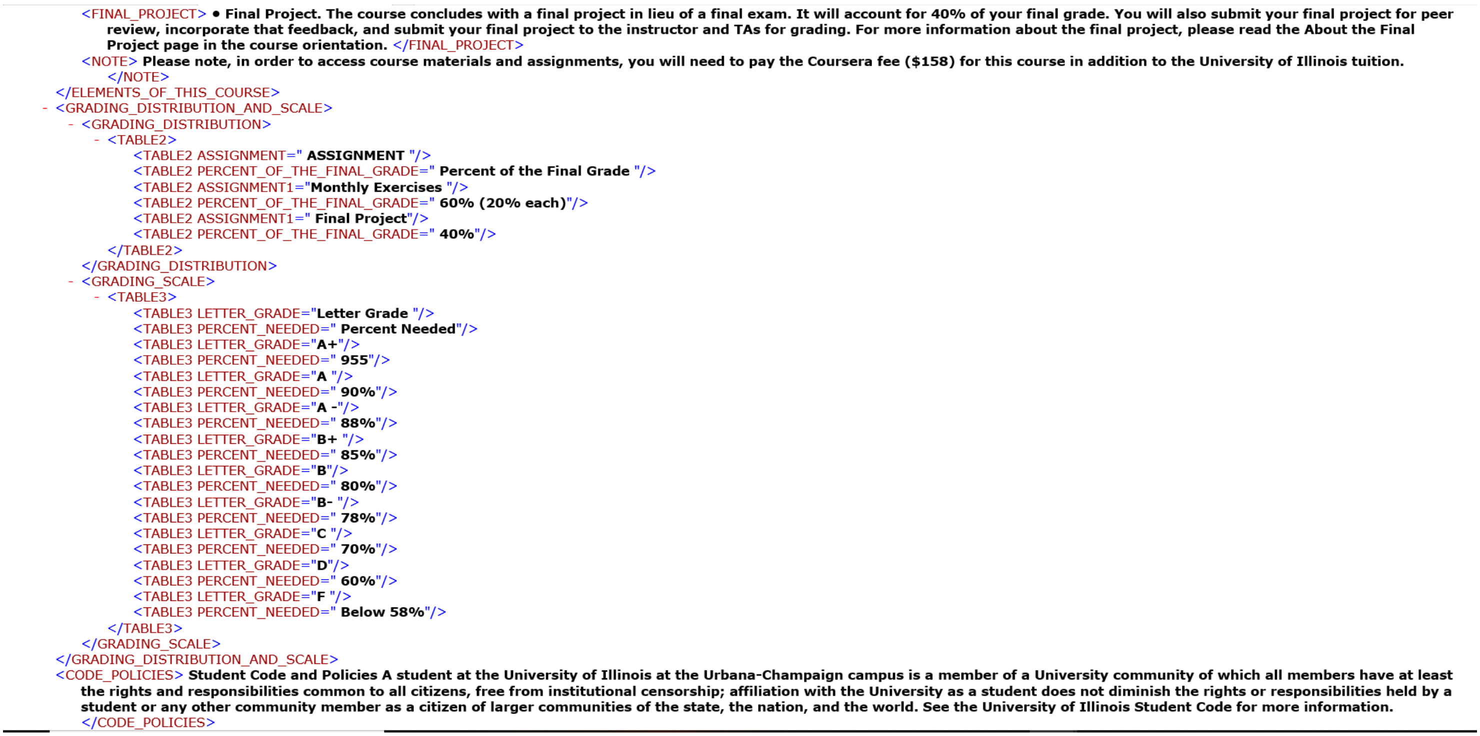
- Tags for the (main) elements are defined. Main elements values are inserted inside the tags.
- Sub- elements are defined inside the tags of the main element.
- Attribute tags are defined for the values of subheadings. There values are inserted in the sub-headings tag.

The designed XML is validated through a validator. The below XML is generated from the validator, which clearly shows the hierarchical structure. The tags marked in red by the validator are the elements or attributes. Remaining text marked in black is the value of the elements or attributes.

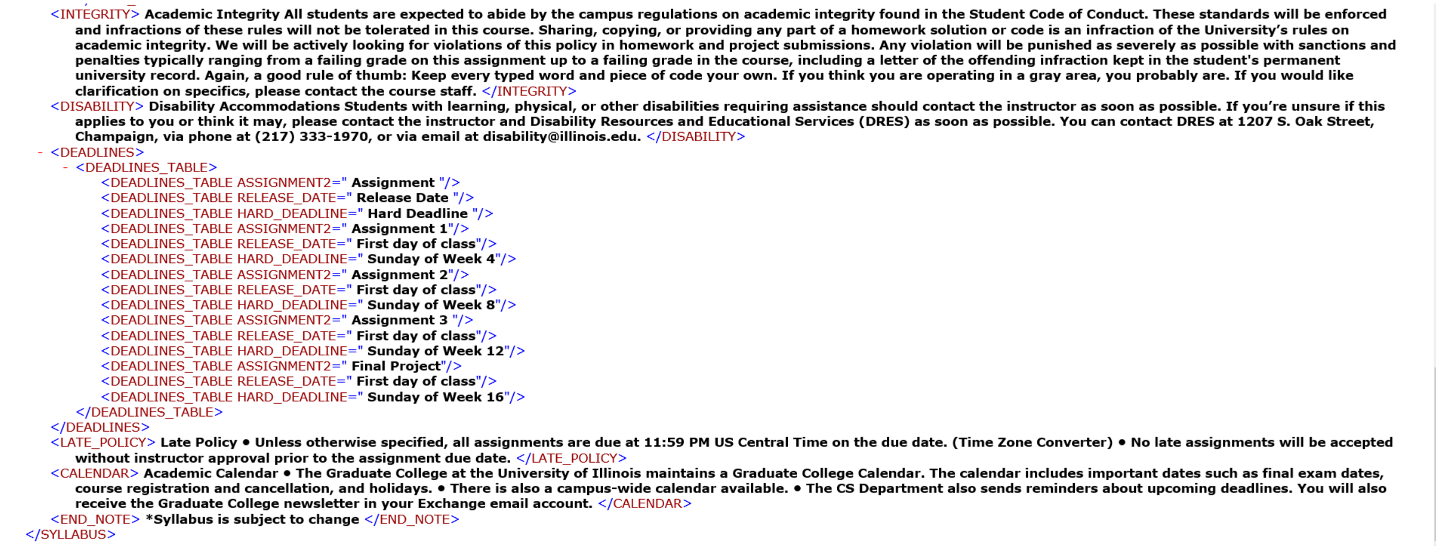
Note: There were several errors encountered while validating the XML. The error record is available in the next component of this paper.



```
  <?xml version="1.0"?>
- <SYLLABUS>
    <TITLE> IS 531: Data Curation Syllabus </TITLE>
    <COURSE_DESCRIPTION> Course Description Welcome to IS 531: Data Curation! Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to
      scholarship, science, and education; curation activities and policies enable data discovery and retrieval, maintain data quality and add value, and provide for re-use over time. This course
      provides an overview of a broad range of theoretical and practical problems in the emerging field, examining issues related to appraisal and selection, long-lived data collections, research
      lifecycles, workflows, metadata, and legal and intellectual property issues. </COURSE_DESCRIPTION>
    <COURSE_GOALS_AND_OBJECTIVES> Upon successful completion of this course, you will be able to: • Describe the significance of abstraction in data management and the relationships among
      the common key data abstraction strategies • Understand the nature of representation hierarchies and strategies for data transformation and transcoding • Explain the process of data
      derivation and the importance of provenance documentation • Compare and contrast various data preservation strategies • Understand the importance of dataset identifiers and citation •
      Describe management of heterogeneity, including schema matching techniques • Explain the role metadata plays in data management and identify a variety of metadata schemes •
      Describe common data behaviors of managers, programmers, scientists, and other users • Summarize the role institutions, agencies, policies, and laws play in data curation
      </COURSE_GOALS_AND_OBJECTIVES>
    <TEXTBOOKS_AND_READINGS> Textbook and Readings There is no required textbook for this course, but there are weekly required readings that can be found in each weekly overview page.
      </TEXTBOOKS_AND_READINGS>
  - <COURSE_OUTLINE>
      Course Outline This 4-credit hour course is 16 weeks long. You should invest 10-12 hours every week in this course.
    - <TABLE1>
        <TABLE1 WEEK=" Week"/>
        <TABLE1 DURATION="Duration"/>
        <TABLE1 TOPICS=" Topics"/>
        <TABLE1 WEEK="1"/>
        <TABLE1 DURATION="8/28 - 9/3"/>
        <TABLE1 TOPICS=" Orientation, Introduction to Data Curation"/>
        <TABLE1 WEEK="2"/>
        <TABLE1 DURATION="9/4 - 9/10"/>
        <TABLE1 TOPICS=" Data Models: Relational Model"/>
        <TABLE1 WEEK="3"/>
        <TABLE1 DURATION="9/11 - 9/17"/>
        <TABLE1 TOPICS=" Trees, Text and Documents"/>
        <TABLE1 WEEK="4"/>
        <TABLE1 DURATION="9/18 - 9/24"/>
        <TABLE1 TOPICS=" Data Models: Ontologies; Schemas; Abstractions; Conceptual Modeling"/>
        <TABLE1 WEEK="5"/>
        <TABLE1 DURATION="9/25 - 10/1"/>
        <TABLE1 TOPICS=" Data Cleaning and Integration; Managing, Processing, and Policy Heterogeneity; Schema Integration"/>
        <TABLE1 WEEK="6"/>
        <TABLE1 DURATION="10/2 - 10/8"/>
        <TABLE1 TOPICS=" Data Concepts; Identity Problems; Ontology for Data Concepts"/>
        <TABLE1 WEEK="7"/>
        <TABLE1 DURATION="10/9 - 10/15"/>
        <TABLE1 TOPICS=" Metadata"/>
        <TABLE1 WEEK="8"/>
        <TABLE1 DURATION="10/16 - 10/22"/>
        <TABLE1 TOPICS=" Preservation"/>
        <TABLE1 WEEK="9"/>
        <TABLE1 DURATION="10/23 - 10/29"/>
        <TABLE1 TOPICS=" Identifiers"/>
        <TABLE1 WEEK="10"/>
        <TABLE1 DURATION="10/30 - 11/5"/>
        <TABLE1 TOPICS=" Standards"/>
        <TABLE1 WEEK="11"/>
        <TABLE1 DURATION="11/6 - 11/12"/>
        <TABLE1 TOPICS=" Workflow, Provenance, and Reproducibility"/>
        <TABLE1 WEEK="12"/>
        <TABLE1 DURATION="11/13 - 11/19"/>
        <TABLE1 TOPICS=" Communication"/>
        <TABLE1 WEEK="13"/>
        <TABLE1 DURATION="11/20 - 11/26"/>
        <TABLE1 TOPICS=" Practices"/>
        <TABLE1 WEEK="14"/>
        <TABLE1 DURATION="11/27 - 12/3"/>
        <TABLE1 TOPICS=" Policy, Law, and Ethics"/>
        <TABLE1 WEEK="15"/>
        <TABLE1 DURATION="12/4 - 12/10"/>
        <TABLE1 TOPICS=" Organization and Governance"/>
        <TABLE1 WEEK="16"/>
        <TABLE1 DURATION="12/11 - 12/14"/>
        <TABLE1 TOPICS=" Review"/>
      </TABLE1>
    </COURSE_OUTLINE>
    <ASSIGNMENTS_DEADLINES> Assignment Deadlines For all assignment deadlines, please refer to the Course Assignment Deadlines, Late Policy, and Academic Calendar page.
      </ASSIGNMENTS_DEADLINES>
  - <ELEMENTS_OF_THIS_COURSE>
      <VIDEOS> • Lecture Videos. In each week, the concepts you need to know will be presented through a collection of short video lectures. You may stream these videos for playback within
        the browser by clicking on their titles or download the videos. You may also download the slides that go along with the videos. The videos usually total 1.5 to 2 hours each week. You
        generally should spend at least the same amount of time digesting content in the video. The actual amount of time needed to digest the content will vary based on your background.
        </VIDEOS>
      <ORIENTATION_QUIZ> • Orientation Quiz. The purpose of the orientation quiz is to ensure that you have gone through the orientation module and acquired the necessary information about
        the course before you start it. The orientation quiz is a required activity, but it's not part of the course grading. You have unlimited attempts on the orientation quiz. You need to
        answer all questions correctly in order to pass the orientation quiz. </ORIENTATION_QUIZ>
      <WEEKLY_QUIZZES> • Weekly Quizzes. Each week concludes with an ungraded quiz to help ensure you understood that week's content. You will be allowed unlimited attempts for each
        quiz, and there is no time limit on how long you take to complete each attempt at the quiz. </WEEKLY_QUIZZES>
      <EXCERCISES> • Exercises. There are three exercises for you to complete in this course, each of which will account for 20% of your final grade. You will submit this assignment for peer
        review to get feedback from your classmates. You will then incorporate the feedback you receive and submit a final version of your exercise to the instructor and TAs for grading. You
        will be allowed one submission attempt for these exercises. Though you are encouraged to discuss these assignments with your classmates, everyone must submit their own work.
        </EXCERCISES>
```



```
      <FINAL_PROJECT> • Final Project. The course concludes with a final project in lieu of a final exam. It will account for 40% of your final grade. You will also submit your final project for peer
        review, incorporate that feedback, and submit your final project to the instructor and TAs for grading. For more information about the final project, please read the About the Final
        Project page in the course orientation. </FINAL_PROJECT>
      <NOTE> Please note, in order to access course materials and assignments, you will need to pay the Coursera fee ($158) for this course in addition to the University of Illinois tuition.
        </NOTE>
    </ELEMENTS_OF_THIS_COURSE>
  - <GRADING_DISTRIBUTION_AND_SCALE>
    - <GRADING_DISTRIBUTION>
      - <TABLE2>
          <TABLE2 ASSIGNMENT=" ASSIGNMENT "/>
          <TABLE2 PERCENT_OF_THE_FINAL_GRADE=" Percent of the Final Grade "/>
          <TABLE2 ASSIGNMENT1="Monthly Exercises "/>
          <TABLE2 PERCENT_OF_THE_FINAL_GRADE=" 60% (20% each)"/>
          <TABLE2 ASSIGNMENT1=" Final Project"/>
          <TABLE2 PERCENT_OF_THE_FINAL_GRADE=" 40%"/>
        </TABLE2>
      </GRADING_DISTRIBUTION>
    - <GRADING_SCALE>
      - <TABLE3>
          <TABLE3 LETTER_GRADE="Letter Grade "/>
          <TABLE3 PERCENT_NEEDED=" Percent Needed"/>
          <TABLE3 LETTER_GRADE="A+"/>
          <TABLE3 PERCENT_NEEDED=" 955"/>
          <TABLE3 LETTER_GRADE="A "/>
          <TABLE3 PERCENT_NEEDED=" 90%"/>
          <TABLE3 LETTER_GRADE="A -"/>
          <TABLE3 PERCENT_NEEDED=" 88%"/>
          <TABLE3 LETTER_GRADE="B+ "/>
          <TABLE3 PERCENT_NEEDED=" 85%"/>
          <TABLE3 LETTER_GRADE="B"/>
          <TABLE3 PERCENT_NEEDED=" 80%"/>
          <TABLE3 LETTER_GRADE="B- "/>
          <TABLE3 PERCENT_NEEDED=" 78%"/>
          <TABLE3 LETTER_GRADE="C "/>
          <TABLE3 PERCENT_NEEDED=" 70%"/>
          <TABLE3 LETTER_GRADE="D"/>
          <TABLE3 PERCENT_NEEDED=" 60%"/>
          <TABLE3 LETTER_GRADE="F "/>
          <TABLE3 PERCENT_NEEDED=" Below 58%"/>
        </TABLE3>
      </GRADING_SCALE>
    </GRADING_DISTRIBUTION_AND_SCALE>
    <CODE_POLICIES> Student Code and Policies A student at the University of Illinois at the Urbana-Champaign campus is a member of a University community of which all members have at least
      the rights and responsibilities common to all citizens, free from institutional censorship; affiliation with the University as a student does not diminish the rights or responsibilities held by a
      student or any other community member as a citizen of larger communities of the state, the nation, and the world. See the University of Illinois Student Code for more information.
      </CODE_POLICIES>
```



```
    <INTEGRITY> Academic Integrity All students are expected to abide by the campus regulations on academic integrity found in the Student Code of Conduct. These standards will be enforced
      and infractions of these rules will not be tolerated in this course. Sharing, copying, or providing any part of a homework solution or code is an infraction of the University's rules on
      academic integrity. We will be actively looking for violations of this policy in homework and project submissions. Any violation will be punished as severely as possible with sanctions and
      penalties typically ranging from a failing grade on this assignment up to a failing grade in the course, including a letter of the offending infraction kept in the student's permanent
      university record. Again, a good rule of thumb: Keep every typed word and piece of code your own. If you think you are operating in a gray area, you probably are. If you would like
      clarification on specifics, please contact the course staff. </INTEGRITY>
    <DISABILITY> Disability Accommodations Students with learning, physical, or other disabilities requiring assistance should contact the instructor as soon as possible. If you're unsure if this
      applies to you or think it may, please contact the instructor and Disability Resources and Educational Services (DRES) as soon as possible. You can contact DRES at 1207 S. Oak Street,
      Champaign, via phone at (217) 333-1970, or via email at disability@illinois.edu. </DISABILITY>
  - <DEADLINES>
    - <DEADLINES_TABLE>
        <DEADLINES_TABLE ASSIGNMENT2=" Assignment "/>
        <DEADLINES_TABLE RELEASE_DATE=" Release Date "/>
        <DEADLINES_TABLE HARD_DEADLINE=" Hard Deadline "/>
        <DEADLINES_TABLE ASSIGNMENT2=" Assignment 1"/>
        <DEADLINES_TABLE RELEASE_DATE=" First day of class"/>
        <DEADLINES_TABLE HARD_DEADLINE=" Sunday of Week 4"/>
        <DEADLINES_TABLE ASSIGNMENT2=" Assignment 2"/>
        <DEADLINES_TABLE RELEASE_DATE=" First day of class"/>
        <DEADLINES_TABLE HARD_DEADLINE=" Sunday of Week 8"/>
        <DEADLINES_TABLE ASSIGNMENT2=" Assignment 3 "/>
        <DEADLINES_TABLE RELEASE_DATE=" First day of class"/>
        <DEADLINES_TABLE HARD_DEADLINE=" Sunday of Week 12"/>
        <DEADLINES_TABLE ASSIGNMENT2=" Final Project"/>
        <DEADLINES_TABLE RELEASE_DATE=" First day of class"/>
        <DEADLINES_TABLE HARD_DEADLINE=" Sunday of Week 16"/>
      </DEADLINES_TABLE>
    </DEADLINES>
    <LATE_POLICY> Late Policy • Unless otherwise specified, all assignments are due at 11:59 PM US Central Time on the due date. (Time Zone Converter) • No late assignments will be accepted
      without instructor approval prior to the assignment due date. </LATE_POLICY>
    <CALENDAR> Academic Calendar • The Graduate College at the University of Illinois maintains a Graduate College Calendar. The calendar includes important dates such as final exam dates,
      course registration and cancellation, and holidays. • There is also a campus-wide calendar available. • The CS Department also sends reminders about upcoming deadlines. You will also
      receive the Graduate College newsletter in your Exchange email account. </CALENDAR>
    <END_NOTE> *Syllabus is subject to change </END_NOTE>
  </SYLLABUS>
```

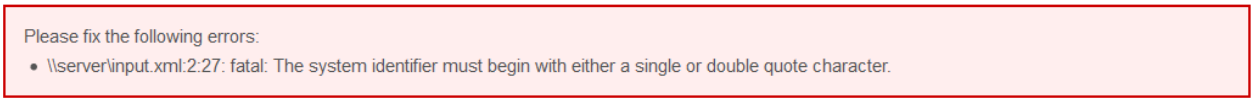
V. Error Record

There were several errors encountered while validating the hand written XML. Errors are documented as a part of Data Curation activities.

Error 1 : <!DOCTYPE SYLLABUS SYSTEM "a.dtd">
The above line in the XML document threw the error of system identifier must begin with single or double quote character.

The error was removed after removing the above statement.
This statement was added to as a part of external DTD document and generate a user defined name of the DTD document.

The format of the above statement was incorrect and is corrected in the final schema. For validating XML purposes, the above statement is removed.

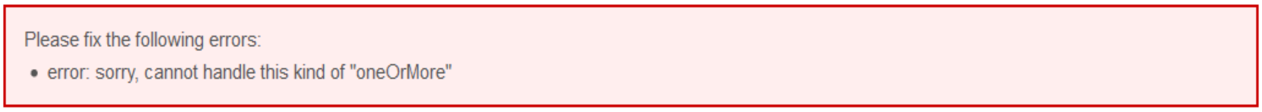


Please fix the following errors:

- \\server\input.xml:2:27: fatal: The system identifier must begin with either a single or double quote character.

Image 8: System Identifier Error

Error 2: Cannot handle this kind of oneorMore error. This error occurs when under one heading, it has subheadings. When subheadings have their values and heading also have its value. This error can be resolved by assigning a subheading to the value of the heading, or completely eliminating the value for the heading.

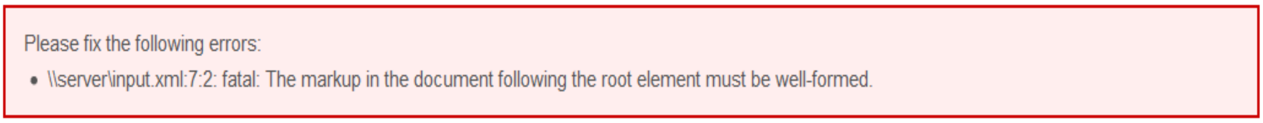


Please fix the following errors:

- error: sorry, cannot handle this kind of "oneOrMore"

Image 9: Cannot handle this kind of oneOrMore

Error 3 : The mark-up in the document following the root element must be well formed. This error occurred when no root element was provided for the XML. After assigning a root element which is syllabus, this error was resolved



Please fix the following errors:

- \\server\input.xml:7:2: fatal: The markup in the document following the root element must be well-formed.

Image 10: The mark-up in the document following the root element must be well formed

VI. Designing DTD

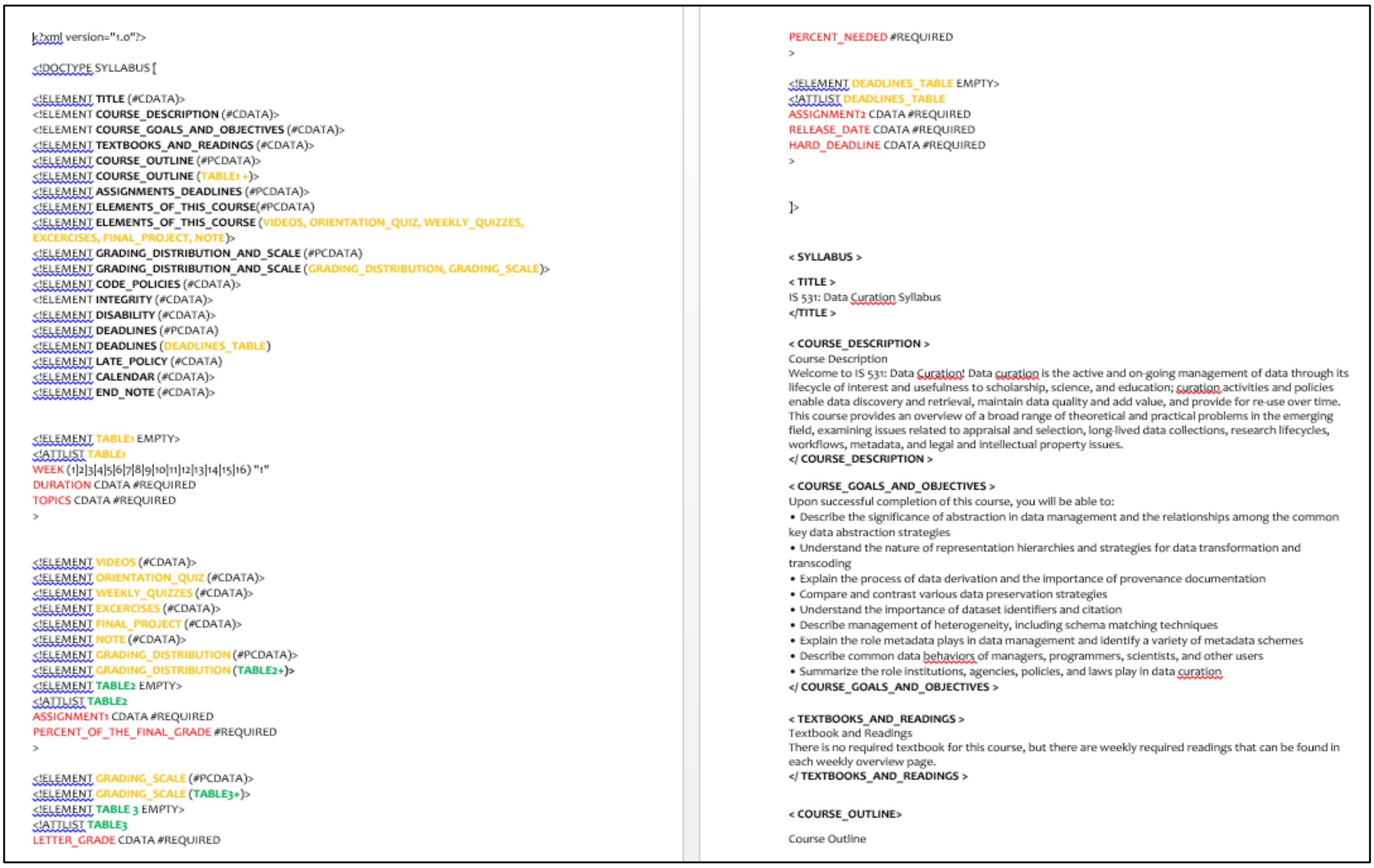
A Document Type Definition (DTD) defines the legal building blocks of an XML document. It defines the document structure with a list of legal elements and attributes. We could have designed a DTD first before the XML, but I prefer to develop it after XML, as I get all the possible values for elements and attributes after developing XML document. It becomes easier to define type and constraints for elements and attributes. We have to write the declaration of all the elements & attributes, with their tag type and PCDATA or CDATA type. For Elements, We also declare child elements here, along with the declaration of number of occurrence of element. Attributes are declared with an ATTLIST declaration. The attribute-value can be one of the following: Default value, REQUIRED, IMPLIED or FIXED value.

To differentiate between various elements, child elements, attributes and values, the text of the DTD file has been colour with different inks. <!ELEMENT > declares an element. All main elements are declared in Bold Black. First Child elements are declared in Yellow and second child elements are declared in Green. All attributes are declared in Yellow. Attributes are declared with <!ATTLIST > tag.

```
<?xml version="1.0"?>
<!DOCTYPE SYLLABUS [
<!ELEMENT TITLE (#CDATA)>
<!ELEMENT COURSE_DESCRIPTION (#CDATA)>
<!ELEMENT COURSE_GOALS_AND_OBJECTIVES (#CDATA)>
<!ELEMENT TEXTBOOKS_AND_READINGS (#CDATA)>
<!ELEMENT COURSE_OUTLINE (#PCDATA)>
<!ELEMENT COURSE_OUTLINE (TABLE1 +)>
<!ELEMENT ASSIGNMENTS_DEADLINES (#PCDATA)>
<!ELEMENT ELEMENTS_OF_THIS_COURSE(#PCDATA)
<!ELEMENT ELEMENTS_OF_THIS_COURSE (VIDEOS,
ORIENTATION_QUIZ, WEEKLY_QUIZZES, EXCERCISES,
FINAL_PROJECT, NOTE)>
<!ELEMENT GRADING_DISTRIBUTION_AND_SCALE (#PCDATA)
<!ELEMENT GRADING_DISTRIBUTION_AND_SCALE
(GRADING_DISTRIBUTION, GRADING_SCALE)>
<!ELEMENT CODE_POLICIES (#CDATA)>
<!ELEMENT INTEGRITY (#CDATA)>
<!ELEMENT DISABILITY (#CDATA)>
<!ELEMENT DEADLINES (#PCDATA)
<!ELEMENT DEADLINES (DEADLINES_TABLE)
<!ELEMENT LATE_POLICY (#CDATA)
<!ELEMENT CALENDAR (#CDATA)>
<!ELEMENT END_NOTE (#CDATA)>


<!ELEMENT TABLE1 EMPTY>
<!ATTLIST TABLE1
WEEK (1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16) "1"
DURATION CDATA #REQUIRED
TOPICS CDATA #REQUIRED
>
<!ELEMENT VIDEOS (#CDATA)>
<!ELEMENT ORIENTATION_QUIZ (#CDATA)>
<!ELEMENT WEEKLY_QUIZZES (#CDATA)>
<!ELEMENT EXCERCISES (#CDATA)>
<!ELEMENT FINAL_PROJECT (#CDATA)>
<!ELEMENT NOTE (#CDATA)>
<!ELEMENT GRADING_DISTRIBUTION (#PCDATA)>
<!ELEMENT GRADING_DISTRIBUTION (TABLE2+)>
<!ELEMENT TABLE2 EMPTY>
<!ATTLIST TABLE2
ASSIGNMENT1 CDATA #REQUIRED
PERCENT_OF_THE_FINAL_GRADE #REQUIRED
>
<!ELEMENT GRADING_SCALE (#PCDATA)>
<!ELEMENT GRADING_SCALE (TABLE3+)>
<!ELEMENT TABLE 3 EMPTY>
<!ATTLIST TABLE3
LETTER_GRADE CDATA #REQUIRED
PERCENT_NEEDED #REQUIRED
>
<!ELEMENT DEADLINES_TABLE EMPTY>
<!ATTLIST DEADLINES_TABLE
ASSIGNMENT2 CDATA #REQUIRED
RELEASE_DATE CDATA #REQUIRED
HARD_DEADLINE CDATA #REQUIRED
>
]>
```

VII. Complete XML Schema with Internal DTD

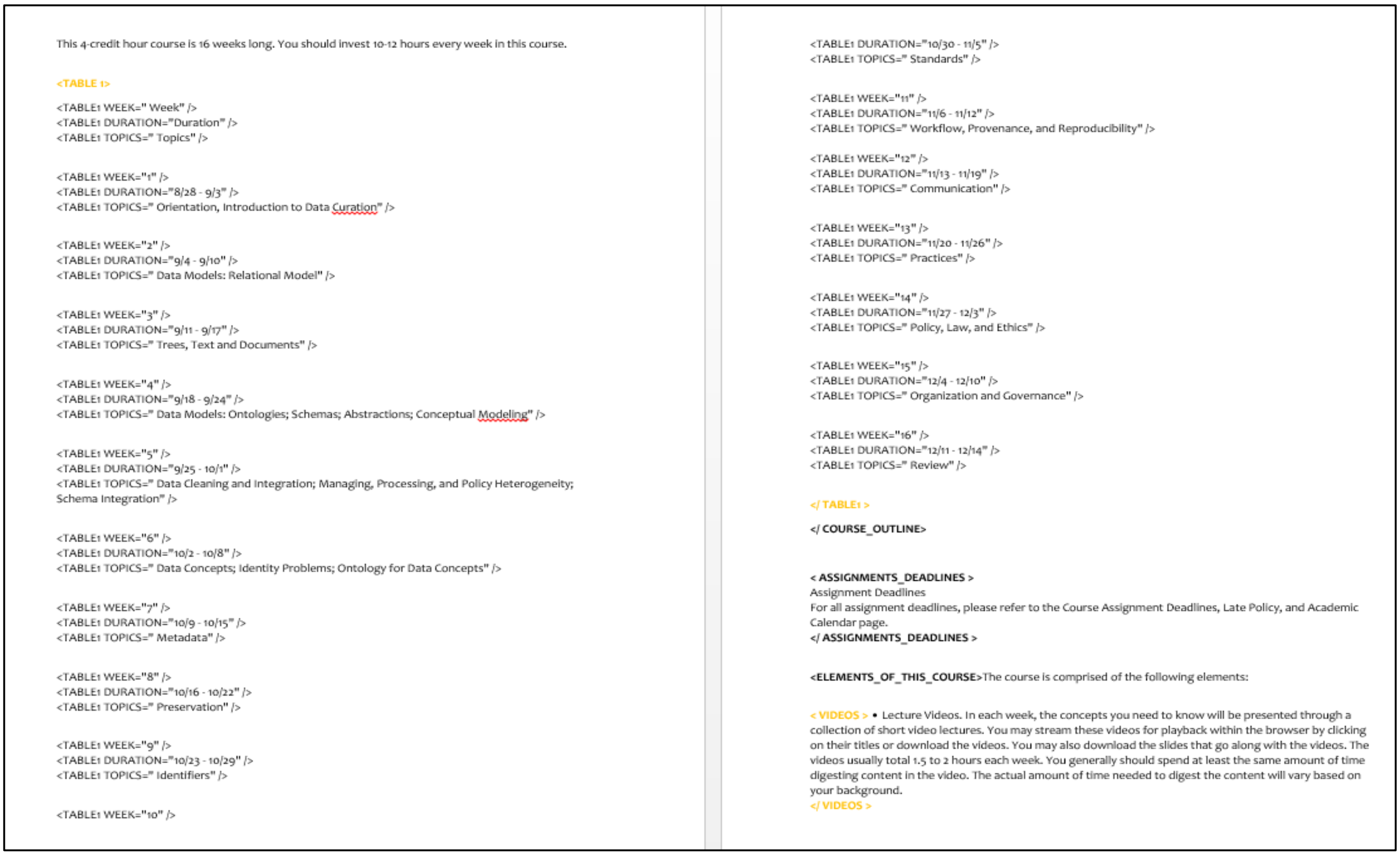
Below are the screenshots of complete XML schema with internal DTD. The validated schema has been provided above. Below is the text version of the XML Schema. There are 8 pages of this XML Schema. Both the DTD & XML files are merged here with addition of the root element SYLLABUS and DOCTYPE reference in way to indicate the internal DTD with XML.



```
<?xml version="1.0"?>

<!DOCTYPE SYLLABUS [

<!ELEMENT TITLE (#CDATA)>
<!ELEMENT COURSE_DESCRIPTION (#CDATA)>
<!ELEMENT COURSE_GOALS_AND_OBJECTIVES (#CDATA)>
<!ELEMENT TEXTBOOKS_AND_READINGS (#CDATA)>
<!ELEMENT COURSE_OUTLINE (#PCDATA)>
<!ELEMENT COURSE_OUTLINE (TABLE1 +)>
<!ELEMENT ASSIGNMENTS_DEADLINES (#PCDATA)>
<!ELEMENT ELEMENTS_OF_THIS_COURSE(#PCDATA)
<!ELEMENT ELEMENTS_OF_THIS_COURSE (VIDEOS, ORIENTATION_QUIZ, WEEKLY_QUIZZES,
EXCERCISES, FINAL_PROJECT, NOTE)>
<!ELEMENT GRADING_DISTRIBUTION_AND_SCALE (#PCDATA)
<!ELEMENT GRADING_DISTRIBUTION_AND_SCALE (GRADING_DISTRIBUTION, GRADING_SCALE)>
<!ELEMENT CODE_POLICIES (#CDATA)>
<!ELEMENT INTEGRITY (#CDATA)>
<!ELEMENT DISABILITY (#CDATA)>
<!ELEMENT DEADLINES (#PCDATA)
<!ELEMENT DEADLINES (DEADLINES_TABLE)
<!ELEMENT LATE_POLICY (#CDATA)
<!ELEMENT CALENDAR (#CDATA)>
<!ELEMENT END_NOTE (#CDATA)>


<!ELEMENT TABLE1 EMPTY>
<!ATTLIST TABLE1
WEEK (1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16) "1"
DURATION CDATA #REQUIRED
TOPICS CDATA #REQUIRED
>


<!ELEMENT VIDEOS (#CDATA)>
<!ELEMENT ORIENTATION_QUIZ (#CDATA)>
<!ELEMENT WEEKLY_QUIZZES (#CDATA)>
<!ELEMENT EXCERCISES (#CDATA)>
<!ELEMENT FINAL_PROJECT (#CDATA)>
<!ELEMENT NOTE (#CDATA)>
<!ELEMENT GRADING_DISTRIBUTION (#PCDATA)>
<!ELEMENT GRADING_DISTRIBUTION (TABLE2+)>
<!ELEMENT TABLE2 EMPTY>
<!ATTLIST TABLE2
ASSIGNMENT1 CDATA #REQUIRED
PERCENT_OF_THE_FINAL_GRADE #REQUIRED
>

<!ELEMENT GRADING_SCALE (#PCDATA)>
<!ELEMENT GRADING_SCALE (TABLE3+)>
<!ELEMENT TABLE 3 EMPTY>
<!ATTLIST TABLE3
LETTER_GRADE CDATA #REQUIRED
```

```
PERCENT_NEEDED #REQUIRED
>

<!ELEMENT DEADLINES_TABLE EMPTY>
<!ATTLIST DEADLINES_TABLE
ASSIGNMENT2 CDATA #REQUIRED
RELEASE_DATE CDATA #REQUIRED
HARD_DEADLINE CDATA #REQUIRED
>


]>


< SYLLABUS >

< TITLE >
IS 531: Data Curation Syllabus
</TITLE >

< COURSE_DESCRIPTION >
Course Description
Welcome to IS 531: Data Curation! Data curation is the active and on-going management of data through its
lifecycle of interest and usefulness to scholarship, science, and education; curation activities and policies
enable data discovery and retrieval, maintain data quality and add value, and provide for re-use over time.
This course provides an overview of a broad range of theoretical and practical problems in the emerging
field, examining issues related to appraisal and selection, long-lived data collections, research lifecycles,
workflows, metadata, and legal and intellectual property issues.
</ COURSE_DESCRIPTION >

< COURSE_GOALS_AND_OBJECTIVES >
Upon successful completion of this course, you will be able to:
• Describe the significance of abstraction in data management and the relationships among the common
key data abstraction strategies
• Understand the nature of representation hierarchies and strategies for data transformation and
transcoding
• Explain the process of data derivation and the importance of provenance documentation
• Compare and contrast various data preservation strategies
• Understand the importance of dataset identifiers and citation
• Describe management of heterogeneity, including schema matching techniques
• Explain the role metadata plays in data management and identify a variety of metadata schemes
• Describe common data behaviors of managers, programmers, scientists, and other users
• Summarize the role institutions, agencies, policies, and laws play in data curation
</ COURSE_GOALS_AND_OBJECTIVES >

< TEXTBOOKS_AND_READINGS >
Textbook and Readings
There is no required textbook for this course, but there are weekly required readings that can be found in
each weekly overview page.
</ TEXTBOOKS_AND_READINGS >


< COURSE_OUTLINE>

Course Outline
```

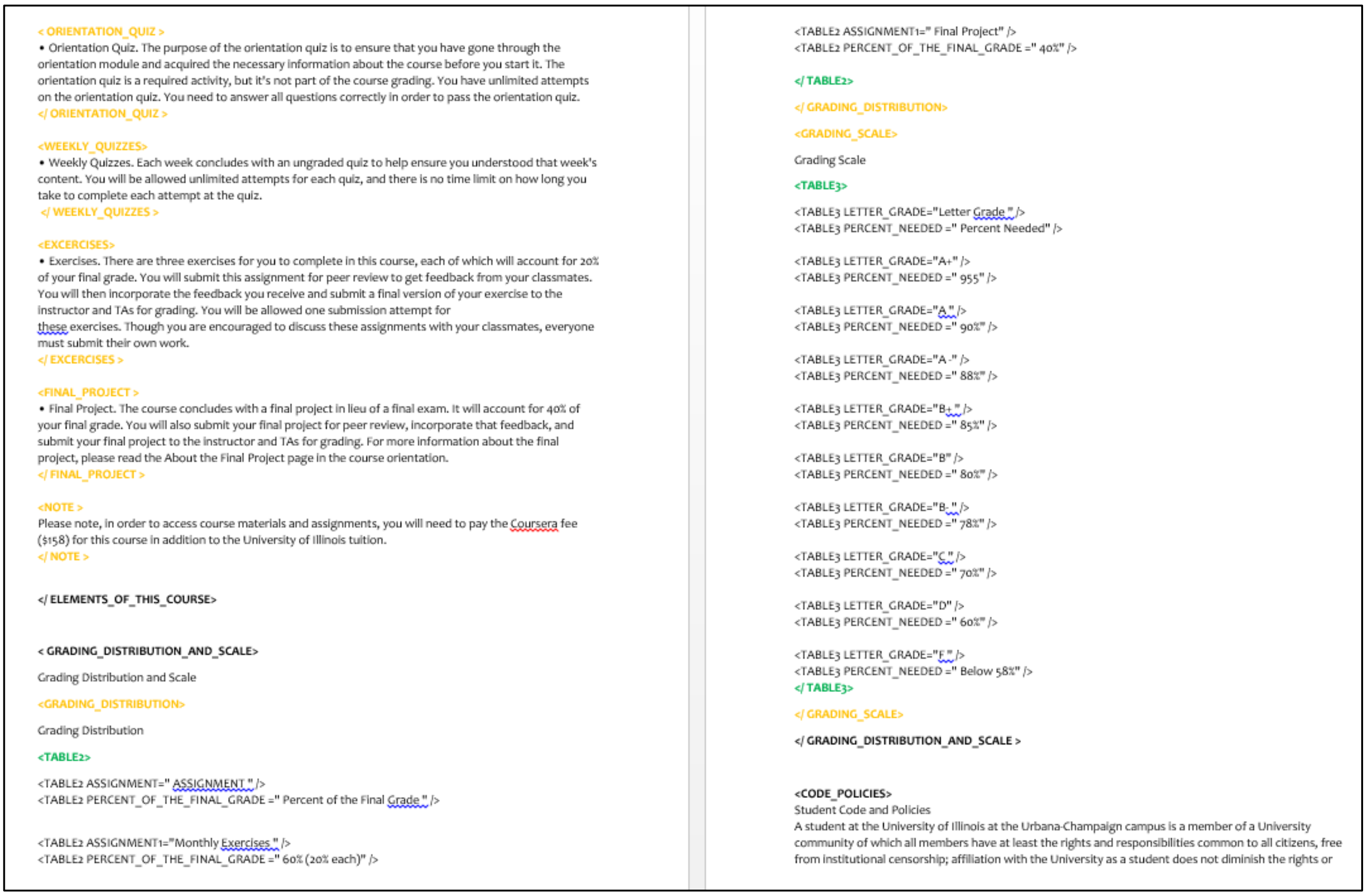
Image 11: XML Schema page 1 & 2



```
This 4-credit hour course is 16 weeks long. You should invest 10-12 hours every week in this course.

<TABLE 1>

<TABLE1 WEEK=" Week" />
<TABLE1 DURATION="Duration" />
<TABLE1 TOPICS=" Topics" />


<TABLE1 WEEK="1" />
<TABLE1 DURATION="8/28 - 9/3" />
<TABLE1 TOPICS=" Orientation, Introduction to Data Curation" />


<TABLE1 WEEK="2" />
<TABLE1 DURATION="9/4 - 9/10" />
<TABLE1 TOPICS=" Data Models: Relational Model" />


<TABLE1 WEEK="3" />
<TABLE1 DURATION="9/11 - 9/17" />
<TABLE1 TOPICS=" Trees, Text and Documents" />


<TABLE1 WEEK="4" />
<TABLE1 DURATION="9/18 - 9/24" />
<TABLE1 TOPICS=" Data Models: Ontologies; Schemas; Abstractions; Conceptual Modeling" />


<TABLE1 WEEK="5" />
<TABLE1 DURATION="9/25 - 10/1" />
<TABLE1 TOPICS=" Data Cleaning and Integration; Managing, Processing, and Policy Heterogeneity;
Schema Integration" />


<TABLE1 WEEK="6" />
<TABLE1 DURATION="10/2 - 10/8" />
<TABLE1 TOPICS=" Data Concepts; Identity Problems; Ontology for Data Concepts" />


<TABLE1 WEEK="7" />
<TABLE1 DURATION="10/9 - 10/15" />
<TABLE1 TOPICS=" Metadata" />


<TABLE1 WEEK="8" />
<TABLE1 DURATION="10/16 - 10/22" />
<TABLE1 TOPICS=" Preservation" />


<TABLE1 WEEK="9" />
<TABLE1 DURATION="10/23 - 10/29" />
<TABLE1 TOPICS=" Identifiers" />


<TABLE1 WEEK="10" />
```

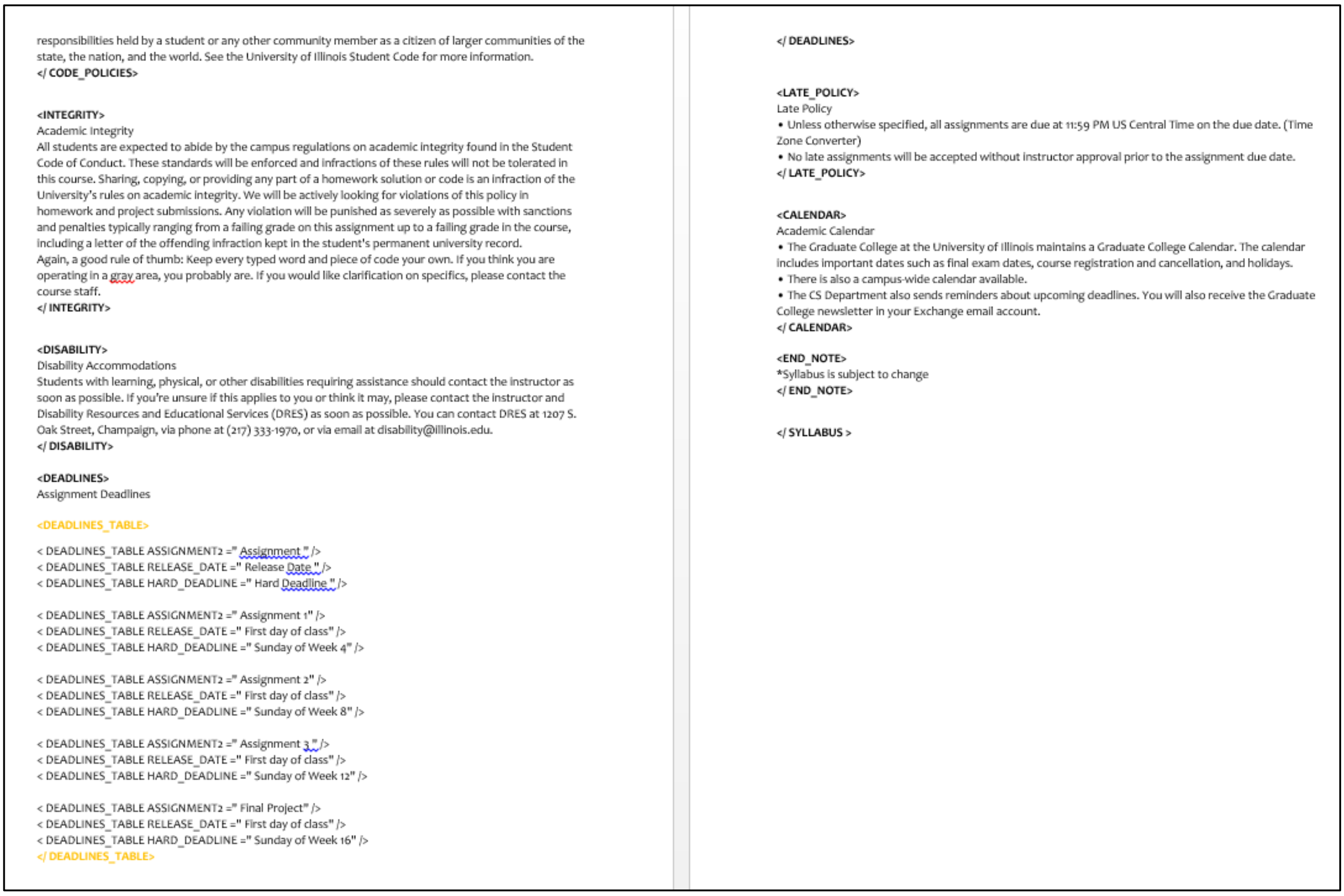
```
<TABLE1 DURATION="10/30 - 11/5" />
<TABLE1 TOPICS=" Standards" />


<TABLE1 WEEK="11" />
<TABLE1 DURATION="11/6 - 11/12" />
<TABLE1 TOPICS=" Workflow, Provenance, and Reproducibility" />

<TABLE1 WEEK="12" />
<TABLE1 DURATION="11/13 - 11/19" />
<TABLE1 TOPICS=" Communication" />


<TABLE1 WEEK="13" />
<TABLE1 DURATION="11/20 - 11/26" />
<TABLE1 TOPICS=" Practices" />


<TABLE1 WEEK="14" />
<TABLE1 DURATION="11/27 - 12/3" />
<TABLE1 TOPICS=" Policy, Law, and Ethics" />


<TABLE1 WEEK="15" />
<TABLE1 DURATION="12/4 - 12/10" />
<TABLE1 TOPICS=" Organization and Governance" />


<TABLE1 WEEK="16" />
<TABLE1 DURATION="12/11 - 12/14" />
<TABLE1 TOPICS=" Review" />


</ TABLE1 >

</ COURSE_OUTLINE>


< ASSIGNMENTS_DEADLINES >
Assignment Deadlines
For all assignment deadlines, please refer to the Course Assignment Deadlines, Late Policy, and Academic
Calendar page.
</ ASSIGNMENTS_DEADLINES >


<ELEMENTS_OF_THIS_COURSE>The course is comprised of the following elements:


< VIDEOS > • Lecture Videos. In each week, the concepts you need to know will be presented through a
collection of short video lectures. You may stream these videos for playback within the browser by clicking
on their titles or download the videos. You may also download the slides that go along with the videos. The
videos usually total 1.5 to 2 hours each week. You generally should spend at least the same amount of time
digesting content in the video. The actual amount of time needed to digest the content will vary based on
your background.
</ VIDEOS >
```

Image 12: XML Schema page 3 & 4



```
< ORIENTATION_QUIZ >
• Orientation Quiz. The purpose of the orientation quiz is to ensure that you have gone through the
orientation module and acquired the necessary information about the course before you start it. The
orientation quiz is a required activity, but it's not part of the course grading. You have unlimited attempts
on the orientation quiz. You need to answer all questions correctly in order to pass the orientation quiz.
</ ORIENTATION_QUIZ >

<WEEKLY_QUIZZES>
• Weekly Quizzes. Each week concludes with an ungraded quiz to help ensure you understood that week's
content. You will be allowed unlimited attempts for each quiz, and there is no time limit on how long you
take to complete each attempt at the quiz.
 </ WEEKLY_QUIZZES >

<EXCERCISES>
• Exercises. There are three exercises for you to complete in this course, each of which will account for 20%
of your final grade. You will submit this assignment for peer review to get feedback from your classmates.
You will then incorporate the feedback you receive and submit a final version of your exercise to the
instructor and TAs for grading. You will be allowed one submission attempt for
these exercises. Though you are encouraged to discuss these assignments with your classmates, everyone
must submit their own work.
</ EXCERCISES >

<FINAL_PROJECT >
• Final Project. The course concludes with a final project in lieu of a final exam. It will account for 40% of
your final grade. You will also submit your final project for peer review, incorporate that feedback, and
submit your final project to the instructor and TAs for grading. For more information about the final
project, please read the About the Final Project page in the course orientation.
</ FINAL_PROJECT >

<NOTE >
Please note, in order to access course materials and assignments, you will need to pay the Coursera fee
($158) for this course in addition to the University of Illinois tuition.
</ NOTE >

</ ELEMENTS_OF_THIS_COURSE>


< GRADING_DISTRIBUTION_AND_SCALE>

Grading Distribution and Scale

<GRADING_DISTRIBUTION>

Grading Distribution

<TABLE2>

<TABLE2 ASSIGNMENT=" ASSIGNMENT " />
<TABLE2 PERCENT_OF_THE_FINAL_GRADE =" Percent of the Final Grade " />


<TABLE2 ASSIGNMENT1="Monthly Exercises " />
<TABLE2 PERCENT_OF_THE_FINAL_GRADE =" 60% (20% each)" />
<TABLE2 ASSIGNMENT1=" Final Project" />
<TABLE2 PERCENT_OF_THE_FINAL_GRADE =" 40%" />

</ TABLE2>

</ GRADING_DISTRIBUTION>

<GRADING_SCALE>

Grading Scale

<TABLE3>

<TABLE3 LETTER_GRADE="Letter Grade " />
<TABLE3 PERCENT_NEEDED =" Percent Needed" />

<TABLE3 LETTER_GRADE="A+" />
<TABLE3 PERCENT_NEEDED =" 955" />

<TABLE3 LETTER_GRADE="A " />
<TABLE3 PERCENT_NEEDED =" 90%" />

<TABLE3 LETTER_GRADE="A -" />
<TABLE3 PERCENT_NEEDED =" 88%" />

<TABLE3 LETTER_GRADE="B+ " />
<TABLE3 PERCENT_NEEDED =" 85%" />

<TABLE3 LETTER_GRADE="B" />
<TABLE3 PERCENT_NEEDED =" 80%" />

<TABLE3 LETTER_GRADE="B- " />
<TABLE3 PERCENT_NEEDED =" 78%" />

<TABLE3 LETTER_GRADE="C " />
<TABLE3 PERCENT_NEEDED =" 70%" />

<TABLE3 LETTER_GRADE="D" />
<TABLE3 PERCENT_NEEDED =" 60%" />

<TABLE3 LETTER_GRADE="F " />
<TABLE3 PERCENT_NEEDED =" Below 58%" />
</ TABLE3>

</ GRADING_SCALE>

</ GRADING_DISTRIBUTION_AND_SCALE >


<CODE_POLICIES>
Student Code and Policies
A student at the University of Illinois at the Urbana-Champaign campus is a member of a University
community of which all members have at least the rights and responsibilities common to all citizens, free
from institutional censorship; affiliation with the University as a student does not diminish the rights or
```

Image 13: XML Schema Page 5 & 6



```
responsibilities held by a student or any other community member as a citizen of larger communities of the
state, the nation, and the world. See the University of Illinois Student Code for more information.
</ CODE_POLICIES>


<INTEGRITY>
Academic Integrity
All students are expected to abide by the campus regulations on academic integrity found in the Student
Code of Conduct. These standards will be enforced and infractions of these rules will not be tolerated in
this course. Sharing, copying, or providing any part of a homework solution or code is an infraction of the
University's rules on academic integrity. We will be actively looking for violations of this policy in
homework and project submissions. Any violation will be punished as severely as possible with sanctions
and penalties typically ranging from a failing grade on this assignment up to a failing grade in the course,
including a letter of the offending infraction kept in the student's permanent university record.
Again, a good rule of thumb: Keep every typed word and piece of code your own. If you think you are
operating in a gray area, you probably are. If you would like clarification on specifics, please contact the
course staff.
</ INTEGRITY>


<DISABILITY>
Disability Accommodations
Students with learning, physical, or other disabilities requiring assistance should contact the instructor as
soon as possible. If you're unsure if this applies to you or think it may, please contact the instructor and
Disability Resources and Educational Services (DRES) as soon as possible. You can contact DRES at 1207 S.
Oak Street, Champaign, via phone at (217) 333-1970, or via email at disability@illinois.edu.
</ DISABILITY>

<DEADLINES>
Assignment Deadlines

<DEADLINES_TABLE>

< DEADLINES_TABLE ASSIGNMENT2 =" Assignment " />
< DEADLINES_TABLE RELEASE_DATE =" Release Date " />
< DEADLINES_TABLE HARD_DEADLINE =" Hard Deadline " />

< DEADLINES_TABLE ASSIGNMENT2 =" Assignment 1" />
< DEADLINES_TABLE RELEASE_DATE =" First day of class" />
< DEADLINES_TABLE HARD_DEADLINE =" Sunday of Week 4" />

< DEADLINES_TABLE ASSIGNMENT2 =" Assignment 2" />
< DEADLINES_TABLE RELEASE_DATE =" First day of class" />
< DEADLINES_TABLE HARD_DEADLINE =" Sunday of Week 8" />

< DEADLINES_TABLE ASSIGNMENT2 =" Assignment 3 " />
< DEADLINES_TABLE RELEASE_DATE =" First day of class" />
< DEADLINES_TABLE HARD_DEADLINE =" Sunday of Week 12" />

< DEADLINES_TABLE ASSIGNMENT2 =" Final Project" />
< DEADLINES_TABLE RELEASE_DATE =" First day of class" />
< DEADLINES_TABLE HARD_DEADLINE =" Sunday of Week 16" />
</ DEADLINES_TABLE>
</ DEADLINES>


<LATE_POLICY>
Late Policy
• Unless otherwise specified, all assignments are due at 11:59 PM US Central Time on the due date. (Time
Zone Converter)
• No late assignments will be accepted without instructor approval prior to the assignment due date.
</ LATE_POLICY>


<CALENDAR>
Academic Calendar
• The Graduate College at the University of Illinois maintains a Graduate College Calendar. The calendar
includes important dates such as final exam dates, course registration and cancellation, and holidays.
• There is also a campus-wide calendar available.
• The CS Department also sends reminders about upcoming deadlines. You will also receive the Graduate
College newsletter in your Exchange email account.
</ CALENDAR>

<END_NOTE>
*Syllabus is subject to change
</ END_NOTE>


</ SYLLABUS >
```

Image 14: XML Schema Page 7 & 8

VIII. Questions:

1. 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?

As some of the problems with attributes are that they contain multiple values, are not easily expandable for future changes, cannot describe structures, are more difficult to manipulate by program code, and attribute values are not easy to test against a DTD. The data which didn't not need all these properties were kept as attributes. The remaining data was made as elements (child elements) to provide flexibility and independence to the data model.

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

There were one liner notes at few places under headings. They do not qualify to be a child element. But there presence at the end of the entire text below the table, with no connection with the table or at the end of the document made us leave with no choice than to address them as the child elements.

Another concern was there was no single root element given in the document, as all the main headings were independent of each other. We had to define our own root element to prepare an XML document as XML document does not work without one root element.

3. How does your DTD design support data independence?

The design support data independence because our XML DTD Schema majorly contains Elements which can be easily modified to add multiple values, or a structure, they are expandable if needed in future. Elements also can be tested against a DTD. The DTD follows a tree structure, where any node can also be added in future if required. Therefore it supports Data Independence.

Moreover, attributes are declared and defined only at places where their modification would not impact the conceptual schema, in turn logical schema would also remain intact.

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

The overarching goal of Data Curation is concerned with all aspects of the data management, therefore, I have implemented most of the data Curation techniques/activities such as, Collection, designing Schema, Formatting, Organization, Modification, Integration, Reformatting, Workflow, Communication, and Discoverability including Data modelling in Conceptual and Logical Layer. This will allow efficient and reliable support to the analysis of data, and will also enable reuse over the time. How these activities have been incorporated and have enhanced the database design is mentioned above in this paper.

This schema has made the documents digital. There are various advantages of Digital documents, such as: Digital documents are

- Easier to create
- Easier to maintain
- Easier to convert (new formats, new delivery software)
- better integrated with workflow in organization
- better integrated with other applications and tools (databases, word processing templates, indexes,) etc.
- more accessible to varied audiences
- easier to accommodate different technological circumstances (varying hardware, operating systems, browser software (brands and versions both), connectivity (bandwidth), etc.
- Easier to accommodate different perceptual abilities (blindness, other sight disabilities, dyslexia, etc.)

IX. Conclusion

An XML Schema is designed for a PDF document. This descriptive markup documents have advantages over the pdf document in

Authoring, Editing, and Transcribing:

- Composition is simplified
- Writing tools are supported
- Alternative views and links facilitated

Publishing:

- Formatting generically specified and modified
- Apparatus automated
- Output device support enhanced
- Portability maximized

Retrieval and Analysis:

- Information retrieval supported
- Analytical procedures supported