CS-598 – Data Curation Assignment 2 - XML Schema Design Exercise

This is an attempt to create XML DTD Schema design. And in the process, one has been able to gain understanding of XML Schema design.

Write a narrative about this process, answering the following reflection 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)?

Towards the fulfillment of the assignment, I was trying to find a dataset that would help answer the questions. Unable to make up my mind towards same, I came up with a "Catalogue" Structured Dataset.

Catalogue dataset is the movie catalogue that helps identify different attributes of it.

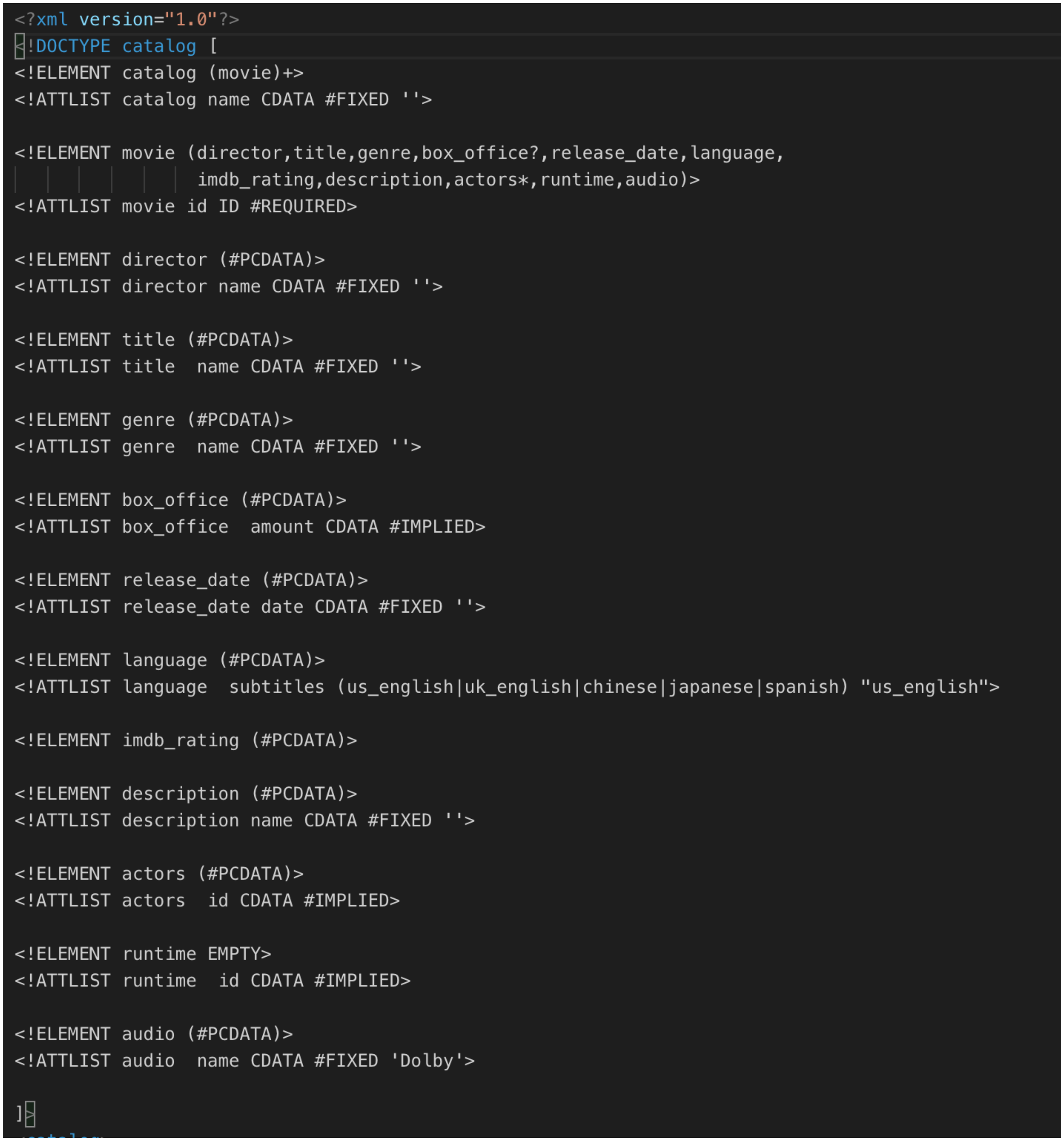
Attribute	**Description**	**Data Type**	**Nullable**
id	Unique Id of the Movie	Alphanumeric	Not null
director	Name of Director	Character - String	Not null
title	Name of the Movie	Character - String	Not null
genre	Genre of the Movie	Character - String	Not null
release_date	Date of Release in US	Date	Not null
language	Language of Movie	Character - String	Not null
imdb_rating	Movie Rating	Alphanumeric	Not null
description	Movie Information	Character - String	Not null
actors/1	Cast Details	Character - String	Null
actors/2	Cast Details	Character - String	Null
box_office	Box Office Collections	Alphanumeric	Null
runtime	Duration of Movie	Numeric	Null
audio	Audio Quality	Character - String	Null

The above Dataset shows a real-world example, wherein information is useful, but the same time can at times have attributes that have missing information. Such free and widely available datasets are quite frequently used in public sites that help end users make decision on various factors.

- 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?

This dataset was chosen as it has the versatility to represent all scenarios of DTD XML schema. It further also helps in ease of fit a formal grammar and a few other constraints. And thus, can be reliably parsed and has an ease of read for end user.

It was further represented with following DTD details:



```
<?xml version="1.0"?>
<!DOCTYPE catalog [
<!ELEMENT catalog (movie)+>
<!ATTLIST catalog name CDATA #FIXED ''>

<!ELEMENT movie (director,title,genre,box_office?,release_date,language,
                 imdb_rating,description,actors*,runtime,audio)>
<!ATTLIST movie id ID #REQUIRED>

<!ELEMENT director (#PCDATA)>
<!ATTLIST director name CDATA #FIXED ''>

<!ELEMENT title (#PCDATA)>
<!ATTLIST title  name CDATA #FIXED ''>

<!ELEMENT genre (#PCDATA)>
<!ATTLIST genre  name CDATA #FIXED ''>

<!ELEMENT box_office (#PCDATA)>
<!ATTLIST box_office  amount CDATA #IMPLIED>

<!ELEMENT release_date (#PCDATA)>
<!ATTLIST release_date date CDATA #FIXED ''>

<!ELEMENT language (#PCDATA)>
<!ATTLIST language  subtitles (us_english|uk_english|chinese|japanese|spanish) "us_english">

<!ELEMENT imdb_rating (#PCDATA)>

<!ELEMENT description (#PCDATA)>
<!ATTLIST description name CDATA #FIXED ''>

<!ELEMENT actors (#PCDATA)>
<!ATTLIST actors  id CDATA #IMPLIED>

<!ELEMENT runtime EMPTY>
<!ATTLIST runtime  id CDATA #IMPLIED>

<!ELEMENT audio (#PCDATA)>
<!ATTLIST audio  name CDATA #FIXED 'Dolby'>

]>
```

1. Catalog is the root/head node of the tree having element movie defined with movie+ (+) sign. Indication that child element can occur one or more times inside parent element. It further has a Character Data Attribute name which is fixed with default value of ‘‘.

2. Further Element movie has following child elements in an order and they appear in same sequence in Document:

   a. director – Sub element has a Character Data Attribute name which is fixed with default value of ‘‘.

   b. title - Element has a Character Data Attribute name which is fixed with default value of ‘‘.

   c. Genre - Element has a Character Data Attribute name which is fixed with default value of ‘‘.

   d. box_office? - Element has a Character Data Attribute amount which is Optional thus has a tag of #IMPLIED – also the element is defined with (?) token which indicated it can have 0 or one definition in parent element.

   e. release_date - Element has a Character Data Attribute date which is fixed with default value of ‘‘.

   f. Language - Element has a Character Data Attribute subtitles which has an enumerated list of values - (us_english|uk_english|chinese|japanese|spanish) with default of "us_english"

   g. imdb_rating – Doesn’t have any attribute list.

   h. Description - - Element has a Character Data Attribute date which is fixed with default value of ‘‘.

   i. actors* - Element has a Character Data Attribute id which is Optional thus has a tag of #IMPLIED – also the element is defined with (*) token which indicated it can have 0 or more definitions in parent element.

   j. runtime? - Element has a Character Data Attribute id which is Optional thus has a tag of #IMPLIED – also the element is defined with (?) token which indicated it can have 0 or one definition in parent element.

   k. Audio - Element has a Character Data Attribute date which is fixed with default value of ‘Dolby‘.

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

In the process of creating the DTD the hardest decision was finding the data, unable to find one that fit the criteria, I had to create one dataset with help of IMDB, this sample dataset was created to fulfill the need of the hour here to cover all aspects of the assignment.

- How does your DTD design support data independence?

The creation of DTD Document, helps abstracting away from storage, abstracting away from processing. Thus change in physical storage won't affect the document, thus supporting Data Independence.

And when the logical schema changes, it would be easier to update the DTD document easily.

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

1. It helps with **Organization** of Data – by clearly helping define the model.
2. It helps with **Preservation** of Data – by ensuring understandability and future usability. As the DTD gives a well-defined details of XML.
3. It helps with **Discoverability** of Data - by ensuring and supporting the ability to search and locate relevant data.
4. It helps with **Access** of Data – by helping in retrieving and distributing data.
5. It helps with **Identification** by clearly helping identify, authenticate and validate the data as per the DTD Schema.
6. Supports easy **Sharing**.
7. Supports **Communication**.
8. Supports **Modification** with ease.