CSCI 585 - Database Systems
Spring 2011
Homework Assignment 3
Due: 04/29/2011

The goal of this assignment is to design an XML Schema and an XML Stylesheet, and to use XQuery to query the XML data.

Part 1: Design an XML Schema to represent the following data specification (28 points).

You need to store the information about the different research conferences that are being held at a specific hotel for computer science graduate students.

For each Student:

SSN (1 occurrence, required)
Name (1 occurrence, required)
Institution name (1 occurrence, required)
Age (1 occurrence, optional)
Email (1 or more occurrences, optional)

For each Conference:

Conference Name (1 occurrence, required)
Conference Year (1 occurrence, required)
Chair Name (1 or more occurrences, required)
Topic (1 occurrence, required)
Original City (1 occurrence, required)
Original Country (1 occurrence, required)
Hotel Chain Name (1 occurrence, required)
Hotel Branch Name (1 occurrence, required)

For each Hotel:

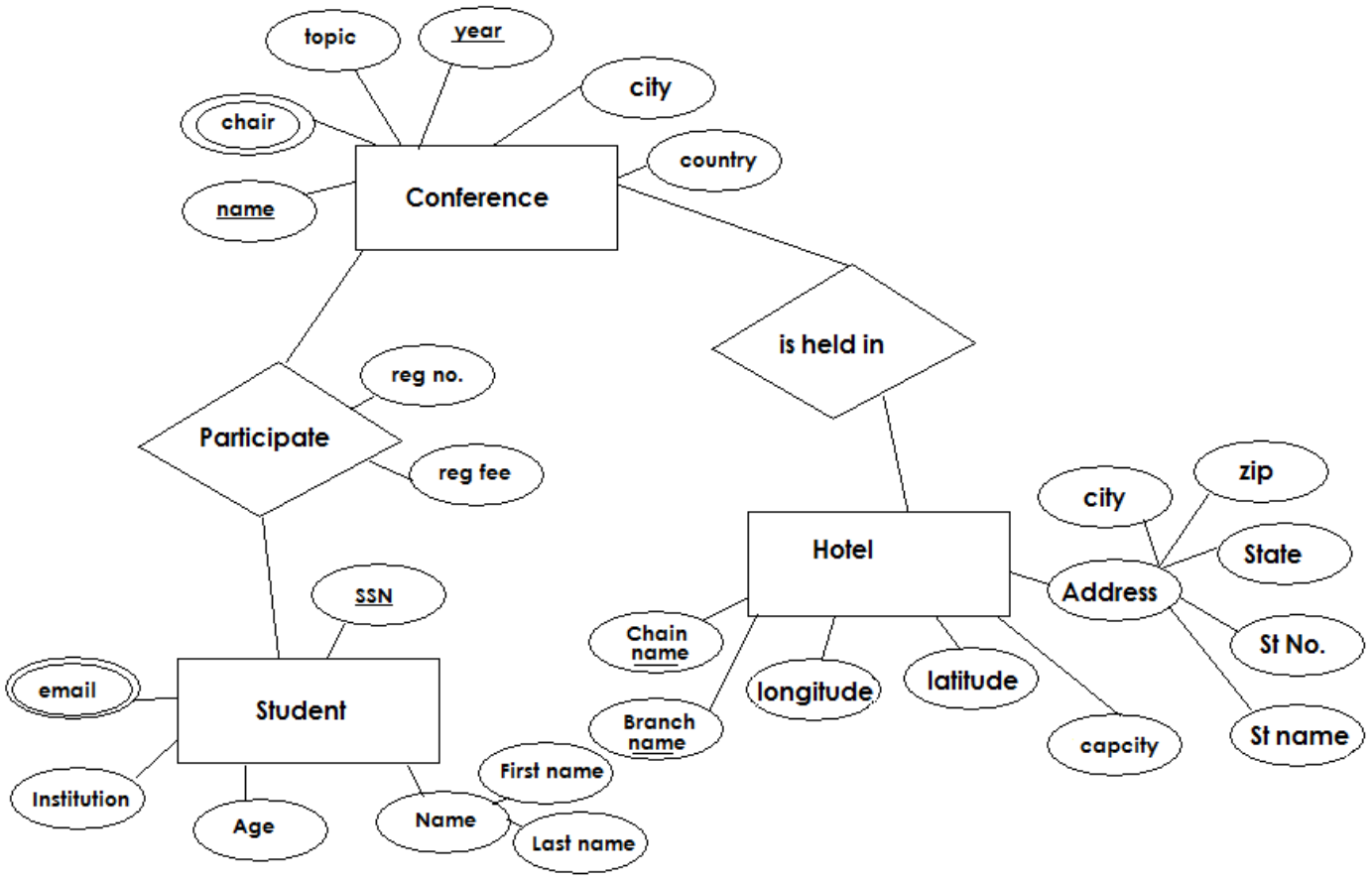
Hotel Chain Name (1 occurrence, required)
Hotel Branch Name (1 occurrence, required)
Address (1 occurrence, required)
Capacity (1 occurrence, optional)
Latitude (1 occurrence, required)
Longitude (1 occurrence, required)

For each Student-Participation-in-Conference:

Student SSN (1 occurrence, required)

Conference Name (1 occurrence, required)
Conference Year (1 occurrence, required)
Registration Fee (1 occurrence, optional)
Registration Number (1 occurrence, required)

(For simplicity, assume each conference is held at only one specific hotel each year.
Also note that the conferences original city does not have to be the same as the city of the hotel at which the conference is being held.)
The details of these relationships and primary keys are shown in the ER diagram below.



Grading Policy for part 1:
Four XML Schema files student.xsd, hotel.xsd, conference.xsd, participation.xsd should be turned in for this part. (28 points– 7 points each)
Missing each attribute: -1 point
Not capturing MultiValue attribute: -1 point
Not capturing primary key attribute: -1 point

Part 2: Take the following data and the XML Schema you designed in Part 1 to create a valid XML data file (16 points).

Student Information:

SSN	First Name	Last Name	Age	Institution Name	Email
622-12-5748	Alexander	Mart	26	UCSF	Alexander@yahoo.com
630-44-3532	Ali	Daria	42	UC Berkeley	Ali@yahoo.com, Ali@hotmail.com
618-58-4662	Amir	Tark	18	Stanford	Amir@yahoo.com
623-62-7824	Antonio	Milsh	34	MIT	Antonio@yahoo.com
613-84-8679	Ashish	Moore	20	Caltech	Ashish@yahoo.com
625-88-5088	Benjamin	Timber	23	Iowa	Benjamin@yahoo.com, Benjamin@gmail.com
625-42-8940	Bob	Shruder	46	USC	Bob@hotmail.com
628-58-8156	Carlos	Gomez	53	USC	Carlos@yahoo.com
627-48-8192	David	Deber	39	UCLA	David@yahoo.com, David@hotmail.com
624-40-5028	Dimitris	Harel	46	Google	Dimitris@hotmail.com
616-55-3688	Fernando	Alexander	48	Google	Fernando@yahoo.com
620-16-0198	Goran	Alav	42	Texas A&M	Goran@yahoo.com
622-93-3200	Hans	Amiri	36	AT&T labs	Hans@yahoo.com
612-58-6604	Harry	Antonio	23	Bell labs	Harry@hotmail.com
613-45-9153	Henrik	Ashish	27	UCSF	Henrik@yahoo.com
630-09-3217	Ivan	Benjamin	51	USC	Ivan@yahoo.com
616-18-5901	Jack	Neuman	28	MIT	Jack@yahoo.com, Jack@hotmail.com, Jack@gmail.com, Jack@aol.com
628-67-3227	Jason	Koch	40	MIT	Jason@yahoo.com, Jason@hotmail.com
622-66-5110	John	Shin	51	Harvard	John@yahoo.com, John@hotmail.com
615-07-7886	Jose	Barnes	28	Harvard	Jose@yahoo.com, Jose@hotmail.com
628-42-4276	Juan	Lopez	50	Delaware	Juan@yahoo.com
615-38-0956	Kun	Gassr	35	Minessota	Kun@yahoo.com, Kun@hotmail.com
617-27-3364	Lee	Liew	48	UMD	Lee@yahoo.com
612-05-8786	Luis	Smith	22	UCSB	Luis@yahoo.com, Luis@hotmail.com, Luis@gmail.com, Luis@aol.com
620-06-4581	Marcus	Stacker	39	ICT	Marcus@yahoo.com
617-71-1016	Maria	Ghosh	49	ISI	Maria@yahoo.com
623-53-0255	Mark	Marr	36	CMU	Mark@yahoo.com, Mark@hotmail.com
618-54-6905	Micheal	Davidson	32	Delft	Micheal@hotmail.com

Hotel Information:

Hotel Chain	Hotel Branch Name	Latitude	Longitude	Capacity	Street No	Street Name	City	State	Zip
Hilton	B1	47.6062095	-122.3078	3 00	406	Washington	Seattle	WA	98104

Doubletree	B2	14.020137	-118.29	9 00	132	Main	Los Angeles	CA	90007
Hilton	B3	37.3297849	-121.8996	1 00	300	Almaden Boulevard	San Jose	CA	95110
Hyatt	B4	14.026535	-118.265	5 00	148	Exposition	Los Angeles	CA	90007
Hilton	B5	30.120635	-70.27	15 0	678	East Monroe Street	Chicago	IL	60603
Doubletree	B6	34.024547	-118.288	14 0	910	Hoover	Los Angeles	CA	90007
Doubletree	B7	40.7142691	-74.239	6 00	160	Main	New York	NY	11209
Doubletree	B8	35.055635	-120.30	13 00	2050	Gateway Place	San Jose	CA	95110
Caesars	B9	44.026535	-121.265	9 00	797	California	Portland	OR	97405
Caesars	B10	46.120635	-122.6008	5 00	314	3rd	Seattle	WA	98104
Ario	B11	40.7142691	-74.8288	13 0	649	3rd	New York	NY	11209
Ario	B12	44.020137	-76.29	2 00	863	Main	Seattle	WA	98104
Carlton	B13	34.055635	-110.30	12 0	991	Vermont	Los Angeles	CA	90007
Carlton	B14	37.026535	-120.265	1 00	282	Magnolia	San Jose	CA	94086
Marriott	B15	17.120635	-116.27	3 00	504	Adams	Los Angeles	CA	90089
Marriott	B16	40.542691	-76.28	7 00	660	Main	New York	NY	11209
Hilton	B17	40.7142691	-73.29	5 00	755	Rose	New York	NY	11209
Hyatt	B18	41.7142691	-73.30	9 00	998	Magnolia	New York	NY	11209
Westin	B19	40.142691	-74.265	9 00	412	Main	New York	NY	11209
Best Western	B20	45.120635	-122.27	4 00	819	Washington	Portland	OR	97405
Doubletree	B21	30.024547	-122.288	8 0	595	Magnolia	Portland	OR	97405
Hilton	B22	64.020137	-108.29	11 0	349	Magnolia	San Francisco	CA	94086
Hilton	B23	45.5234515	-121.30	8 00	252	Rose	Portland	OR	97405
Holidays Inn	B24	34.024547	-117.288	15 0	980	Washington	Los Angeles	CA	90007
Holidays Inn	B25	40. 691	-73.29	4 0	503	Magnolia	New York	NY	11209
Westbound	B26	40. 2691	-74.09729	10 0	532	Ocean	New York	NY	11209
Westin	B27	47.362095	-122.5708	8 0	134	Main	Seattle	WA	98104
Best Western	B28	34.120635	-118.27	13 0	852	Main	Los Angeles	CA	90007

(Please note that the latitude longitude data value might not correspond to the zip code or exact address provided in the table)

Conference Information:

Name	Year	Topic	Chair	Original city	Original Country	Hotel Chain Name	Hotel Branch Name
ACM Siggraph	2010	Computer Graphics	Debevec, Baylis	Boston	USA	Hilton	B1
ACM Sigir	2010	Information Retrieval	Jordan	Moscow	Russia	Westbound	B26
ACL	2010	Computational Linguistics	Traum	Los Angeles	USA	Marriott	B16
ACM Sigdial	2010	Dialogue	Leuski, Arkinson	Madrid	Spain	Westin	B19
CVPR	2010	Computer Vision	Medioni, Feifei	Portland	USA	Hilton	B3
ECCV	2010	Computer Vision	Belongie	San Diego	USA	Holidays Inn	B24
VLDB	2010	Databases	Shamsian, Shahabi	Tehran	Iran	Hyatt	B18
ACM Siggraph	2011	Computer Graphics	Neumann, Baylis, Ghosh	London	UK	Carlton	B13
ACM Sigir	2011	Information Retrieval	Greyson	New Orleans	USA	Hilton	B3
ACL	2011	Computational	Kinght	New York	USA	Doubletree	B8

		Linguistics					
ACM Sigdial	2011	Dialogue	Traum, Morency	Chicago	USA	Hilton	B17
CVPR	2011	Computer Vision	Morency, Medioni, Navatia	Atlanta	USA	Carlton	B14
ECCV	2011	Computer Vision	Koller	Los Angeles	USA	Hyatt	B18
VLDB	2011	Databases	Grey	Seattle	USA	Hilton	B3

Participation Information:

Student SSN	Conference Name	Conference Year	Registration Number	Registration Fee
622-12-5748	ACL	2010	2010100	100 $
622-12-5748	ACL	2011	2011890	120 $
622-12-5748	ACM Sigdial	2010	S9769	170 $
630-44-3532	CVPR	2011	P0000001	400 $
618-58-4662	CVPR	2011	P0000005	400 $
623-62-7824	VLDB	2010	VL5742	300 $
613-84-8679	ACM Siggraph	2010	SIG009	600 $
625-88-5088	ACM Sigdial	2011	S7976	270 $
625-42-8940	ACL	2010	2010	100 $
625-42-8940	ACM Sigir	2011	124	200 $
628-58-8156	ACM Siggraph	2010	SIG0089	600 $
627-48-8192	VLDB	2010	VL1841	300 $
624-40-5028	ACL	2010	2010	100 $
616-55-3688	CVPR	2010	P0000008	400 $
616-55-3688	ACM Sigdial	2010	S9080	190 $
620-16-0198	ACM Sigir	2011	194	200 $
622-93-3200	ACL	2011	20118734	120 $
612-58-6604	ECCV	2011	1	200 $
613-45-9153	ACM Sigir	2010	198	200 $
613-45-9153	ACM Siggraph	2011	SIG006	700 $
630-09-3217	ACM Siggraph	2011	SIG004	700 $
616-18-5901	ECCV	2011	2	200 $
628-67-3227	VLDB	2010	VL8770	300 $
622-66-5110	VLDB	2011	VL9070	330 $
615-07-7886	ACM Sigdial	2011	S9070	250 $
615-07-7886	ACL	2010	2010231	100 $
615-07-7886	ACL	2011	2011342	120 $
615-07-7886	ACM Sigdial	2010	1000001	200 $
628-42-4276	ACL	2010	20105234	100 $
615-38-0956	ECCV	2010	3	200 $
617-27-3364	VLDB	2010	VL1740	300 $
612-05-8786	ACM Sigir	2011	187	200 $
612-05-8786	CVPR	2011	P000098	400 $
620-06-4581	ACL	2010	2010324	100 $
617-71-1016	ACM Siggraph	2011	SIG002	700 $
617-71-1016	VLDB	2011	VL1740	330 $
617-71-1016	VLDB	2010	VL1000	300 $
623-53-0255	CVPR	2010	P000089	400 $
618-54-6905	ACM Siggraph	2011	SIG021	700 $

Note: Make sure that the XML files are validated against the XML Schema you designed using "Validate file" in *XML Spy*. For this, you can download "XMLSpyEnt2010_x32(64).exe" file from:
http://www.altova.com/download/xmlspy/xml_editor_enterprise.html

Grading Policy for part 2:
Four XML data files student.xml, hotel.xml, conference.xml, participation.xml (16 points– 4 points each)
Not matching/referring to the corresponding XML schema definition: -2 points

Not having all the data rows entered to XML data file: -2 points

Part 3: Generate a Simple Style file (XSL) so that the XML files you created in Part 2 can be displayed with MS Internet Explorer 6+ (16 points).

The display must be in the form of tables and the following tables must be displayed:

• Student information (Font=Verdana, Border=1, Background=yellow, font-weight= '400')
• Hotel information (Font=Times, Border=2, Background=blue, font-weight= '600')
• Conference information (Font=Arial, Border=1, Background=orange, font-weight= '400')
• Participation information (Font=Nina, Border=3, Background=green, Font-style=italic, font-weight= '500')

The entire data in the XML file must be included in the corresponding table.
The style file should also accommodate the display of the table's title (e.g.: student information, hotel information, conference information, participation information) as a header above the display of the data table and the title of the columns of the table (e.g.: "SSN" column for "student information" table) in bold font.

Grading Policy for part 3:
Four XSL style files student.xsl, hotel.xsl, conference.xsl, participation.xsl (16 points– 4 points each)

Missing any attribute of the tables in the definition of the style file: -1 point
Missing the display of the table information (the table title and columns information): -2 point
Not showing the tables in the specified format: -3 (You get partial credit of 1 point if you show a table with different formatting rather than the one specified here.)

Part 4: Use the XQuery to answer 8 queries (40 points – 5 points each).

Use XMLSpy for Xquery implementation. Required queries should be implemented in separate files each: q1.xquery, q2.xquery, q3.xquery, q4.xquery, q5.xquery, q6.xquery, q7.xquery, q8.xquery.
The output of the queries in xml format should accommodate the <row> elements as required for each query. You can also define XSL files for the queries and show the resulting xml data in tables but this is not required.

The required queries are as follows:

1. Display all the hotels on Google Earth.

".kml" files, which are similar to XML files with specific tags, could be displayed by Google Earth. You can download "GoogleEarth.exe" from: http://earth.google.com/download-earth.html
The result of your query should be as follows. Save the result in a kml file. Make sure that each "Placemark" tag is repeated for each hotel:

```
<kml xmlns="http://earth.google.com/kml/2.0">
<Folder>
<name>Hotel</name>
<open>1</open>
<Placemark>
<name> Hilton, B1</name>
<Point>
```

<coordinates -122.3078, 47.6062095, 0</coordinates>
</Point>
</Placemark>
</Folder>
</kml>

(TA Note: You can find reference for working with KML files here:
http://code.google.com/apis/kml/documentation/)

2. Display the First Name and Last Name and the email address(es) of each Student. (If a student has more than one email address, each email address should be separated by ","). Sort the result by the student Last Name in ascending order.

Example row in result set:

First-Name	Last-Name	Email
Alexander	Mart	Alexander@yahoo.com

Required XML output:

<row>
<first-name> Alexander </first-name>
<last-name> Mart </last-name>
<email> Alexander@yahoo.com, Alexander@gmail.com </email>
</row>

3. Display different zip codes, and the number of Hotels and their total capacity of all the hotels for each Zip Code(find the sum of capacities of all the hotels in the given zip code). Sort the result by the capacity of hotels in descending order.

Example row in result set:

Zip Code	Number of Hotels	Total Capacity
60603	1	150

Required XML output:

<row>
<zip-code> 60603 </zip-code>
<number-of-hotels> 1 </number-of-hotels >
<capacity> 150 </capacity>
</row>

4. Display the conference name, year and the average age of students participating in each conference. Sort the results by the average age in descending order. (If no student is participating in the conference the average age is 0 or the corresponding row should not be shown)
(TA Note: ACM Siggraph 2010 and ACM Siggraph 2011 are considered to be two different conferences.)

Example rows in result set:

Name	Year	Average age
ACM Siggraph	2011	34

ACM Siggraph	2010	27

Required XML output:

```
<row>
        <name> ACM Siggraph </name>
        <year> 2011 </ year>
        <average-age> 34 </average-age>
</row>
```

5. Display the SSN, the name of the institution and the total amount of money each student has spent on the registration fee of different conferences in each year, and the year. Sort the result by the total fee in descending order. Sort rows by students' SSN number in ascending order.

Example rows in result set:

SSN	Institution	Year	Total fee
617-71-1016	ISI	2010	1030
617-71-1016	ISI	2011	300

Required XML output:

```
<row>
        <SSN> 617-71-1016 </SSN>
        <institution > ISI </institution >
        <year> 2010 </ year>
        <total-fee> 1030 </total-fee>
</row>
<row>
        <SSN> 617-71-1016 </SSN>
        <institution > ISI </institution >
        <year> 2011 </ year>
        <total-fee> 300 </total-fee>
</row>
```

6. Display the name, year, the topic and the number of student attendants of the conference(s).
Example row in result set:

Name	Year	Topic	Number of attendees
ACM Siggraph	2011	Computer Graphics	3

Required XML output:

```
<row>
        <name> ACM Siggraph </name>
        <year> 2011 </ year>
        < topic > Computer Graphics </topic >
        <number-of-attendees> 3 </ number-of-attendees >
</row>
```

7. Display the name of the chair(s) of the conference, the conference name and year, and the total amount of income from student registration fees for the conference(s) .

Example row in result set:

Chair Names	Name	Year	Total income
Neumann, Baylis, Ghosh	ACM Siggraph	2011	2100

(TA Note: If a field has more than one entry, each entry should be separated by ",")

Required XML output:

```
<row>
        <chair-names> Neumann, Baylis, Ghosh </chair-names>
        <name> ACM Siggraph </name>
        <year> 2011 </ year>
        <total-income> 3 </ total-income >
</row>
```

8. Display the name of the hotel chains and the number of conferences they have held over the past years (all years) sorted by the number of conferences in descending order.

Example row in result set:

Name	Number of conferences
Doubletree	3

Required XML output:

```
<row>
        <name> Doubletree </name>
        <number-of-conferences> 3 </number-of-conferences>
</row>
```

(TA Note: Please note that the examples rows provided here might not be the results of the actual running of the query on the provided data tables.)

Grading Policy for part 4:
Eight XQueriy files q1.xquery, q2.xquery, q3.xquery, q4.xquery, q5.xquery, q6.xquery, q7.xquery, q8.xquery (40 points – 5 points each)

You will be deducted 1 point if you have not shown the multi value attributes, separated by commas.

You will be deducted 2 points if you do not represent the required xml elements.

You will be deducted 3 points if your result includes the row we are looking for but also includes other unrelated rows.

After executing the queries, the output should be displayed properly by selecting the "Browser" tab in *XMLSpy* otherwise you will be deducted 1 point.

Submission Instructions:

• You need to submit the following files (the filenames should be EXACTLY the same as mentioned below):

1. The Readme.txt file that contains your full name, USC ID and the list of the submitted files. There is a 10 points penalty if this file or some of the required information is missing from your submission.

2. The XML Schema files student.xsd, hotel.xsd, conference.xsd, participation.xsd (28 points– 7 points each)
3. The XML data files student.xml, hotel.xml, conference.xml, participation.xml (16 points– 4 points each)
4. The XSL style files student.xsl, hotel.xsl, conference.xsl, participation.xsl (16 points– 4 points each)
5. Eight XQueries files q1.xquery, q2.xquery, q3.xquery, q4.xquery, q5.xquery, q6.xquery, q7.xquery, q8.xquery (40 points – 5 points each)

• You must make one .tar file that includes all these files (e.g., hw3.tar) using the following command:

> **tar cvf hw3.tar *.xsd *.xml *.xsl *.xquery readme.txt**
Please note that you can do this in your aludra.usc.edu account.

• You need to submit your homework electronically using your DEN account and through the digital DropBox. Do NOT try to submit any file other than the (hw3.tar) tar file.
The submit button will immediately respond with a SUCCESS message if your submission is successful.
That will be the way you can make sure that your homework has been received properly. Your submissions will be time stamped, so we will know the exact time when you made the submission. Make sure you submit before the deadline.

• Start working on your assignment early.

Late Submission Policy:

This is the late submission grading policy: 10% penalty for submission within one day, 20% penalty for submission within two days, 50% penalty for submission within one week.
No grades would be given for submissions later than one week.