**XML Problem**

Consider the following specification for the ‘Publication’ entity and its corresponding ER model, which is designed to represent every aspect of this specification as much as possible.

**Specification of the ‘Publication’ Entity:**

“A publication must have a title, one or more authors, a published year, and a publication type (PType). The publication type should be one of the “workshop”, “conference”, or “journal”, with “conference” as the default value. Only the publications published between 2000 and 2005 (including 2000 and 2005) are considered, and the format of the year value is ‘YYYY’ (e.g., 2002).

**ER Schema:**

![ER Diagram]

- Title
- PType
- Publication
- Author
- Year
a) Fill out the missing parts of following XML Schema inline element/attribute declarations to represent the given ER model, including the additional constraints described in the specification as much as possible. Note that both ‘PType’ and ‘Year’ are treated as the attributes of the ‘Publication’ element (4pts).

```xml
<xsd:__name="publication">
  <xsd:complexType>
    <xsd:__________>
      <xsd:element name="title" ________________________________>
      <xsd:element name="authors" ________________________________>
    </xsd:__________>
    <xsd:attribute name="ptype" ________________________________>
      <xsd:________ base="xsd:string">
        <xsd:________ value="workshop" __________>
        <xsd:________ value="conference" __________>
        <xsd:________ value="journal" __________>
      </xsd:________>
    </xsd:__________>
    <xsd:attribute name="year" ________________________________>
      <xsd:__________>
        <xsd:restriction base="________">
          <xsd:________ value=______________>
          <xsd:________ value=______________>
        </xsd:restriction>
      </xsd:__________>
    </xsd:attribute>
  </xsd:complexType>
</xsd:____________>
```
b) Write the corresponding DTD declaration of the above schema. **Note that both ‘PType’ and ‘Year’ should be treated as the attributes of the ‘Publication’ element** (3pts).
XML Problem

Consider the following DTD declarations and write the corresponding Xqueries for the queries that follow.

**ACTORS.DTD:**

```xml
<!ELEMENT actors (actor+)>
<!ELEMENT actor (name, contract*)>
<!ELEMENT name #PCDATA>
<!ELEMENT contract (agent, earning)>
<!ELEMENT agent (#PCDATA)>
<!ELEMENT earning (#PCDATA)>
<!ATTLIST actor ssn CDATA #REQUIRED
gender (male|female) #REQUIRED>
<!ATTLIST contract year CDATA #REQUIRED>
<!ATTLIST earning currency CDATA "USD">
```

**MOVIES.DTD:**

```xml
<!ELEMENT movies (movie+)>
<!ELEMENT movie (title, director+, starring+, rate)? >
<!ELEMENT title (#PCDATA)>
<!ELEMENT director #PCDATA>
<!ELEMENT starring #PCDATA>
<!ELEMENT rate EMPTY>
<!ATTLIST title year CDATA #REQUIRED
oscar (true|false) "false">
<!ATTLIST starring ssn CDATA #REQUIRED>
<!ATTLIST rate authority (NYT|LAT) "LAT">
```
a) Display the name of all actors who have worked with both directors “Ron Howard” and “Steven Spielberg”. The output should be in the following format (6pts).

```
<result>
    <actor>
        <name>Tome Hanks</name>
        <title>The DaVinci Code</title>
        <title>Saving Private Ryan</title>
    </actor>
    ...
</actor>
</result>
```

b) Find all directors who have had at least two movies starring “Angelina Jolie”. Display them in the descending order of movie counts. The output should be in the following format (6pts).

```
<result>
    <director>
        <name>Doug Liman</name>
        <moviecount>3</moviecount>
    </director>
    ...
    <director>
        ...
</director>
</result>
```
c) Find the title of all movies starring female actors. DO NOT display the movies without female actors. The output should be in the following format (6pts).

```xml
<result>
  <movie>
    <title>Chicago</title>
    <actress>
      <name>Catherin Zeta-Jones</name>
      <name>Renee Zellweger</name>
    </actress>
  </movie>
  ...  
  <movie>
    ...  
</result>
```
**OLAP Problem**

Consider the following one dimension cube:

\[ A[1:16]= \begin{array}{cccccccccccccc} 
8 & 4 & 4 & 8 & 4 & 2 & 6 & 4 & 7 & 3 & 5 & 7 & 9 & 5 & 4 & 8 \\
\end{array} \]

a) Draw the corresponding SRPS cube where \( k_1=2 \). (2 points)

b) Draw the corresponding SRPS cube where \( k=4 \). (2 points)

c) Draw the corresponding SRPS cube where \( k=8 \). (2 points)

d) Draw the corresponding SRPS cube where \( k=16 \). (2 points)

e) Answer the query \( \text{Sum}(A[i]) \) where \( 3<=i<=12 \), using SRPS cubes in each of the above four cases (a,b,c,d). (Specifically show the SRPS cells you are using) (8 points)

f) Considering the update \( A[3]=6 \), rewrite all the updated cells for each of the four cases (a,b,c,d). (4 points)

g) What is the best value of \( k \) for this problem based on the query and update costs. (2 points)

h) Generalize the intuition you gained (if any) to discuss what is the best value of \( k \) for SRPS in general case. (3 points)

---

1 The length of a box
DDBMS Problem

Consider the following three relations with their keys underlined:

ARTIST (ANo, AName, ArtGenre)
GALLERY (GNo, GName, Loc)
SHOW (ANo, GNo, StartDate, Duration)

Furthermore, assume the following two queries:

- Query 1 (q1):
  
  ```
  select ARTIST.ANo, AName, SHOW.GNo, StartDate 
  from ARTIST, SHOW 
  where ARTIST.ANo = SHOW.ANo and Duration = 5
  ```

  Suppose q1 is executed by an application that is located at sites S1 and S2, with frequencies 1 and 2, respectively.

- Query 2 (q2):
  
  ```
  select ANo, Duration 
  from SHOW
  ```

  Suppose q2 is executed by an application that is located at sites S2 and S3, with frequencies 2 and 1, respectively.

a) Construct the usage matrix UA for the attributes of the relation SHOW. (4 points)

(Reminder: element eij of the UA matrix is use(qi, Aj), the usage value for the attribute Aj by the query qi).
b) Construct the affinity matrix $AA$ containing all attributes of the relation $SHOW$. (6 points)

c) Transform the affinity matrix $AA$ using the Bond Energy Algorithm into a clustered affinity matrix $CA$. (6 points)

d) In the $CA$ matrix, you should be able to recognize at least one cluster (you do NOT need to run the cluster splitting algorithm to distinguish the clusters). What is the corresponding vertical fragment for this cluster? What is the remaining vertical fragment? (4 points)
DDBMS Problem

Consider the relation Office (ID, City, State, Sales), where \{ID\} and \{City, State\} are two candidate keys.

(a) List all valid vertical fragmentations of the relation “Office” into exactly two fragments (4pts).
b) Assume that 50% of the queries access the attributes City and Sales, and the other 50% of the queries access ID, City, and State.

Which of the valid vertical fragmentations is (are) optimal in terms of the number of accesses? Justify your answer (8pts).