

Semi-structured Data Model



SDSC SAN DIEGO
SUPERCOMPUTER CENTER

After this video you will be able to

- Distinguish between structured and “semi-structured model”
- Recognize that most semistructured data are tree-structured
- Explain why tree navigation operations are important for XML, JSON

Let's talk Web Pages

A Simple HTML Example

This is the first paragraph.

- List item 1
- List item 2

This is a bolded text.

```
<!DOCTYPE html>
<html>
<body>

<h1>A Simple HTML Example</h1>

<p title="undecided so far">
This is the first paragraph.
<li> List item 1 </li>
<li> List item 2 </li>
</p>

<p><b>
This is a bolded text.
</b></p>

</body>
</html>
```

The diagram illustrates the mapping between HTML code and its rendered output. Colored arrows point from specific code elements to their corresponding visual representations on the left. A light blue arrow points from `<html>` to the top of the content area. A red arrow points from `<body>` to the start of the content area. A dark green arrow points from `<h1>` to the title 'A Simple HTML Example'. A blue arrow points from `<p title="undecided so far">` to the first paragraph. A yellow arrow points from `` to the list items. Another blue arrow points from `<p>` to the end of the first paragraph. A final blue arrow points from `` to the bolded text.

XML

- Allows the querying of both schema and data

```
- <experiments version="1.2" revision="100915" total="58" total-samples="4011" total-assays="3847">
- <experiment>
  <releasedate>2007-11-22</releasedate>
  <species>Mus musculus</species>
- <miamescores>
  <reportersequencescore>1</reportersequencescore>
  <factorvaluescore>1</factorvaluescore>
  <measuredbioassaydatascore>0</measuredbioassaydatascore>
  <protocolscore>0</protocolscore>
  <derivedbioassaydatascore>1</derivedbioassaydatascore>
  <overallscore>3</overallscore>
</miamescores>
<assays>18</assays>
<samples>18</samples>
<rawdatafiles>0</rawdatafiles>
<fgemdatafiles>18</fgemdatafiles>
- <sampleattribute>
  <category>CellType</category>
  <value>primary chondrocyte</value>
  <value>primary dermal fibroblast</value>
  <value>primary osteoblast</value>
</sampleattribute>
- <sampleattribute>
  <category>Organism</category>
  <value>Mus musculus</value>
</sampleattribute>
- <experimentalfactor>
  <name>CellType</name>
  <value>primary chondrocyte</value>
  <value>primary dermal fibroblast</value>
```

JSON

```
{
  "status": 200,
  "photos": [
    {
      "typeName": "Facebook",
      "type": "facebook",
      "typeId": "facebook",
      "url": "http://graph.facebook.com/amoghnatu/picture?type=large",
      "isPrimary": true
    }
  ],
  "contactInfo": {
    "familyName": "Natu",
    "fullName": "Amogh Natu",
    "givenName": "Amogh"
  },
  "demographics": {
    "gender": "male"
  },
  "socialProfiles": [
    {
      "id": "1839143973",
      "typeName": "Facebook",
      "username": "amoghnatu",
      "type": "facebook",
      "typeId": "facebook",
      "url": "http://www.facebook.com/amoghnatu"
    }
  ]
}
```

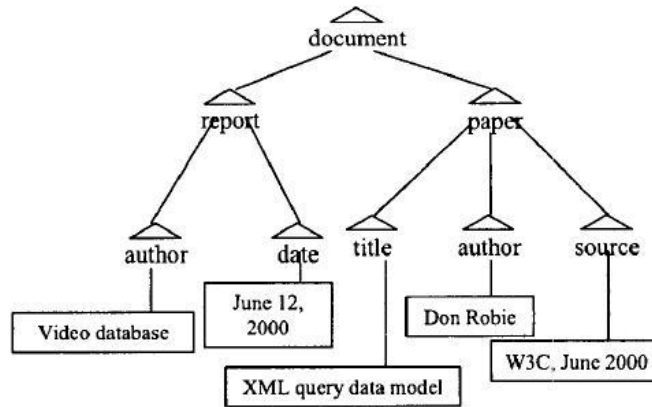
Key-value pair

Tuple

*Square
brackets
indicate arrays*

Tree Data Structure

```
<document>
  <report>
    <author>Video database</author>
    <date>June 12, 2000</date>
  </report>
  <paper>
    <title>XML query data model</title>
    <author>Don Robie</author>
    <source>W3C, June 2000</source>
  </paper>
</document>
```



Tree Operations

- **Paper**

- getParent → document
- getChildren → title, author, source
- GetSibling → report

- **“Video database”**

- Root-to-Node path → document/report/author/“video database”

- **Queries need tree navigation**

- Author of “XML query data model”

