Setting Up and Representing Data with XML

INFO20002: Foundations of Informatics
Tutor: Yang Lu
08/03/2016

XML data representation and validation

- Today’s outcomes:
  - –Use python library “lxml” to handle xml data as a tree model
  - –Understand the mechanism of DTD as well as XML validation

Launch your python environment

1. Setting Up Python

- **Ananconda** (https://www.continuum.io/downloads)
  - –Spyder - IDE
  - –Jupyter http://jupyter.org/ (iPython notebook)
  - – C:\Anaconda2
  - –Setting the path: http://dreamind.github.io/ix/pages/page.html?src=/ix/workshops/infrastructure/python.md

- **IVLE** (http://ivle.informatics.unimelb.edu.au/+login)
  - –University providing; web-based platform

- **Pythonanywhere** (https://www.pythonanywhere.com/)
  - –Cloud; self-learning

2. Representing XML Data

Reviews on XML

eXtensible Markup Language (XML)

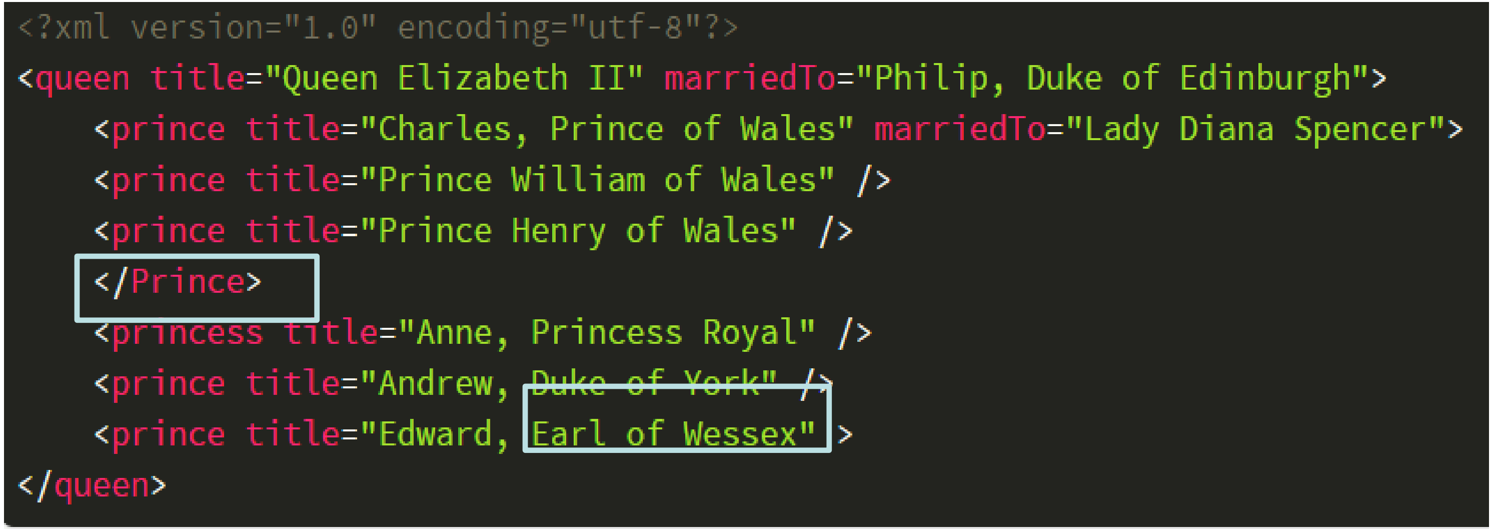
-Extensible; User define

-With Meaning

-Structured

-Hierarchy

Exercise 1 – Fix the XML file

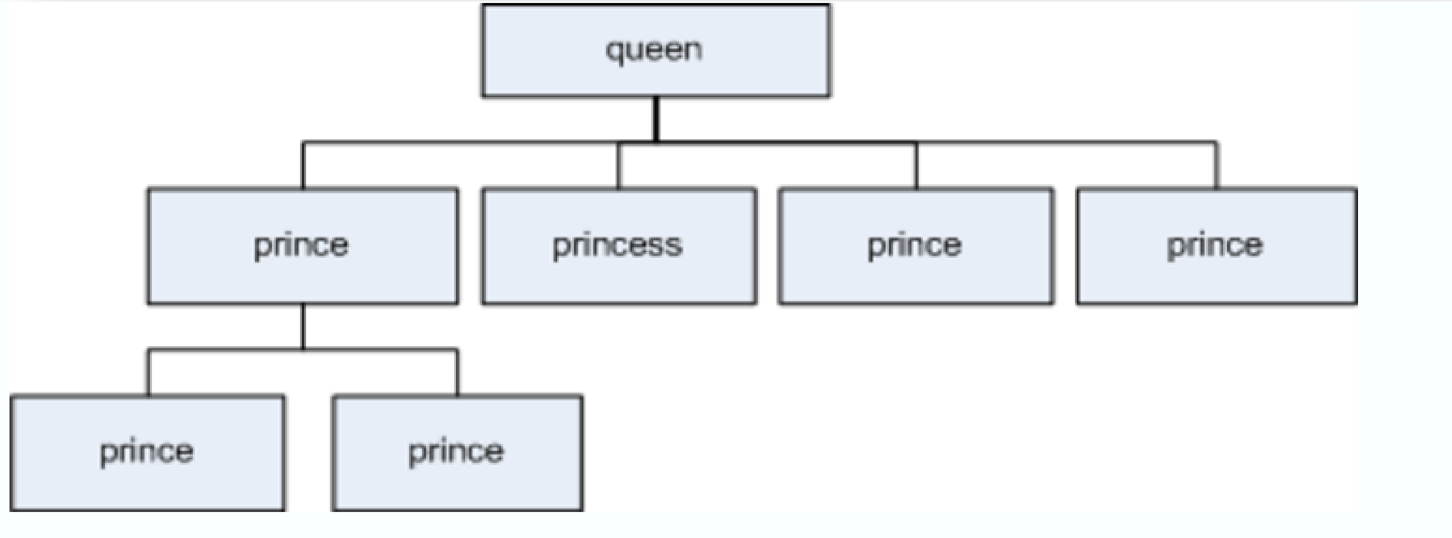


```
<?xml version="1.0" encoding="utf-8"?>
<queen title="Queen Elizabeth II" marriedTo="Philip, Duke of Edinburgh">
    <prince title="Charles, Prince of Wales" marriedTo="Lady Diana Spencer">
    <prince title="Prince William of Wales" />
    <prince title="Prince Henry of Wales" />
    </Prince>
    <princess title="Anne, Princess Royal" />
    <prince title="Andrew, Duke of York" />
    <prince title="Edward, Earl of Wessex" >
</queen>
```

Case sensitive *<patient>and </Patient>*

Opening and closing (*<patient>…</patient>*) (empty element-*<patient/>*)

Exercise 2: recognize the XML structure

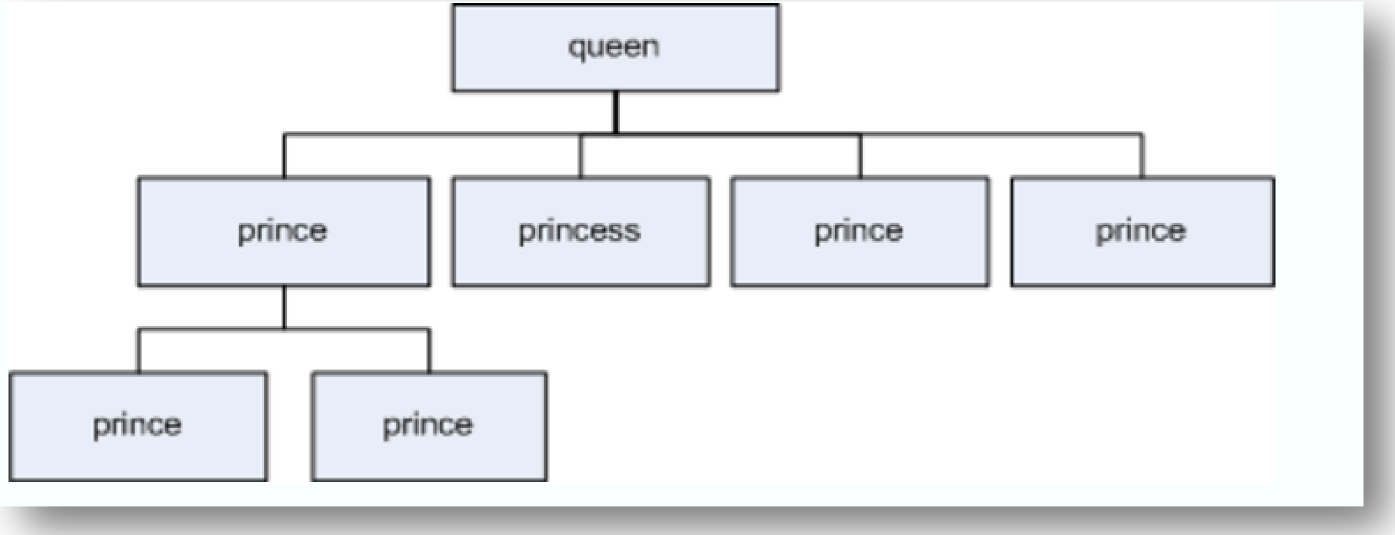


- 1. How many **XML elements** in the whole XML tree?

- 2. What attributes belong to the **first child** of the root element? What are their **values**?
  - –Title (Charles, Prince of Wales), marriedTo (Lady Diana Spencer)

“lxml”

- Python package, specific to XML/HTML processing
- "etree” – data process in a tree-like structure
  - `etree.parse()`



Exercise 3 – access XML data

- Using the royal.xml:
  - 1. Write a Python code to **get the title** of queen's **grandsons**.
  - 2. Write a Python code to **get the title** of the only **princess** in the family tree.

Edit XML data – create and specify

Two procedures:

1. Create new elements- `.Element()`
2. Attach to the tree – `.append()`

Alternatively, create the new element as a child:

– `.SubElement(element, <tag>)`

The new element is attached as the last element (root [-1])

Insert it at a particular position: `.insert([position],<tag>)`

- Attribute – `set(“name”, “value”)`

- Serialising – `.tostring()`

Exercise 4 – edit the xml file

- 1. Assuming you have completed the tasks above, **replace** the text and the attribute of the **price element** to set the book **price** to **25 AUD**.
  - –Hints:
    - Create another <price>
    - **Replace** the old <price> by the new one
- 2. Create a new element called **pages**, set its content to 277, and **append** it to the root. Confirm that the new element is created, you can issue the following command:

```
>>> print etree.tostring(root[-1]).
<pages>277</pages>
```