Information Infrastructure II

INFO I211 – Spring 2014 – Sections 18530 & 22719

Lecture 25 – 2014.04.23 & 2014.04.24

Instructor:
Mitja Hmeljak,
http://mypage.iu.edu/~mitja
mitja@indiana.edu

XML: *ID* ←→ *Name Match* (Group Work)

Using the **students.xml** file available on *Oncourse* ,

and Python 2.6 on your laptop computer (or Python 2.6 and an editor such as Pico on silo.soic.indiana.edu),

write a function called **id_find** that takes an *id number* as a string and prints out the name of the student who matches that id.

Then test your fuction from the main code, for example:

id_find("0019846768") -> Jack Sparrow

Name Match (Solution 1)

```
import xml.etree.ElementTree as ET

def id_find(num):

    root = ET.parse(source="students.xml")

    elements = root.getiterator()

    first = ""
    last = ""

    for elem in elements:
        if elem.tag == "first":
            first = elem.text
        elif elem.tag == "last":
            last = elem.text
        elif elem.tag == "id" and elem.text == num:
            print first, last, "found."

id_find("0019846768")    #Jack Sparrow
id_find("0019846789")    #Jason Bourne
```

Name Match (Solution 2)

```
import xml.etree.ElementTree as ET

def id_find(num):

    root = ET.parse(source="students.xml")

    students = root.findall("Student")

    for student in students:
        if student.find("id").text == num:
            print student.find("name/first").text, \
                  student.find("name/last").text, "found. "

id_find("0019846768")   #Jack Sparrow
id_find("0019846789")   #Jason Bourne
```

Namespaces in XML

The name of an element in XML is just a string

So our list of Students could include different types:

- Informatics Students
- Computer Science Students
- Philosophy Students

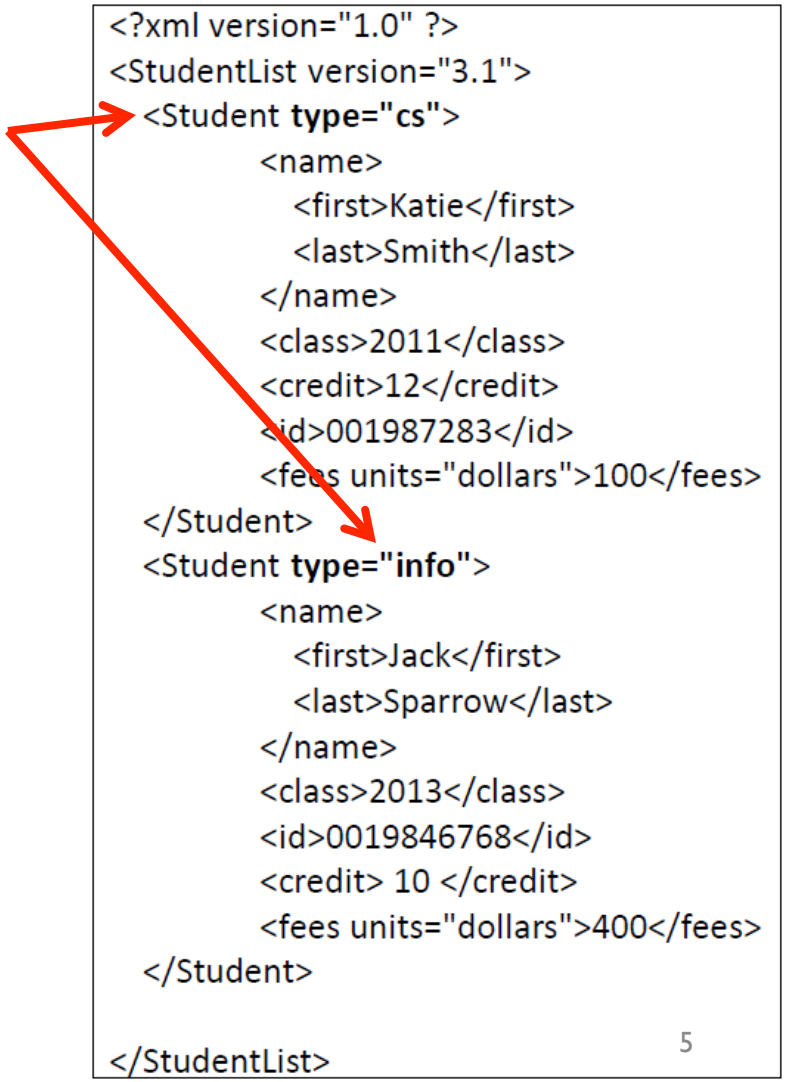
How do we tell them apart?

Namespaces in XML

We could use an attribute

But this is less searchable

Instead, we use
namespaces



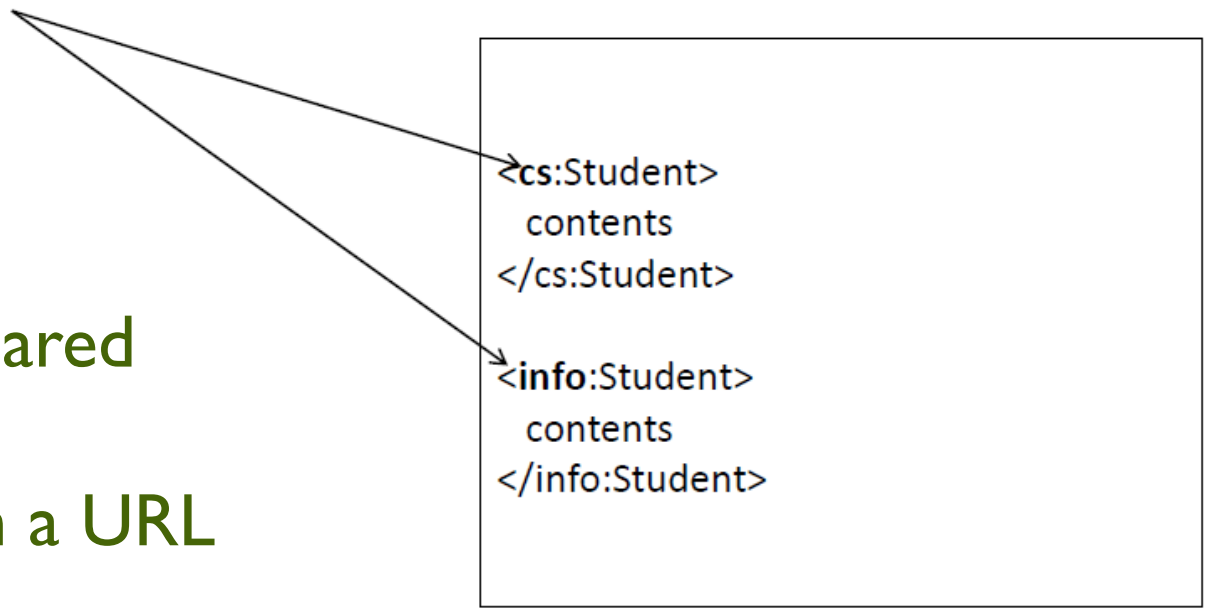
```
<?xml version="1.0" ?>
<StudentList version="3.1">
  <Student type="cs">
        <name>
          <first>Katie</first>
          <last>Smith</last>
        </name>
        <class>2011</class>
        <credit>12</credit>
        <id>001987283</id>
        <fees units="dollars">100</fees>
  </Student>
  <Student type="info">
        <name>
          <first>Jack</first>
          <last>Sparrow</last>
        </name>
        <class>2013</class>
        <id>0019846768</id>
        <credit> 10 </credit>
        <fees units="dollars">400</fees>
  </Student>

</StudentList>
```

Namespaces in XML

The use of a *namespace* is indicated by the presence of a "**:**" in the element's name

- Namespace

Namespaces
have to be declared
and are usually
associated with a URL



Namespaces in XML

```
<?xml version="1.0" ?>

<StudentList version="2.1" xmlns:info="http://http://www.soic.indiana.edu/" xmlns:phil="http://www.indiana.edu/
~phil/">
    <info:Student>
        <name>
            <first>Katie</first>
            <last>Smith</last>
        </name>
        <class>2011</class>
        <credit>12</credit>
        <id>001987283</id>
        <fees units = "dollars" c = "usa">100</fees>
    </info:Student>
    <phil:Student>
        <name>
            <first>Jack</first>
            <last>Sparrow</last>
        </name>
        <class>2013</class>
        <id>0019846768</id>
        <credit>10</credit>
        <fees units = "dollars" c = "usa">200</fees>
    </phil:Student>
</StudentList>
```

Namespace declarations at the root node

Namespace use

Closing tags need namespace as well!

Namespaces in XML

Download **students_ns.xml**

Before, we did this:

```
import xml.etree.ElementTree as ET
root = ET.parse(source="students.xml")
students = root.findall("Student")
print students
```

So, this will work, right?

```
import xml.etree.ElementTree as ET
root = ET.parse(source="students_ns.xml")
info_students = root.findall("info:Student")
print students
```

Namespaces in XML

That did not work.
Instead, we need to **prepend** the namespace to the element name:

```
import xml.etree.ElementTree as ET

root = ET.parse(source="students_ns.xml")

info_students = root.findall("{http://http://
www.soic.indiana.edu/}Student")
phil_students = root.findall("{http://www.indiana.edu/~phil/}
Student")

print info_students
print phil_students
```

Namespaces in XML

The use of namespaces is very common in XML obtained from the web, such as RSS feeds.

For example, view the IU Technology news feed at
http://www.iu.edu/~iunews/services/newsrooms/feeds/?format=rss20&id=1232a17b814f4e1c77a8fb801f237996&sort=date

Let's look at this RSS feed!

Namespaces in XML

https://www.iu.edu/~iunews/services/newsrooms/feeds/?format=rss20&id=1232a17b814f4e1c77a8fb801f237996&sort=date

To obtain the web page (as a string containing XML) and parse it into an XML structure:

```
import urllib
import xml.etree.ElementTree as ET

# all in one line:
conn = urllib.urlopen("https://www.iu.edu/~iunews/services/newsrooms/feeds/?format=rss20&id=1232a17b814f4e1c77a8fb801f237996&sort=date")

lines = conn.read()
conn.close()

root = ET.XML(lines)

print root.tag              #this just prints out 'rss'
```

Namespaces in XML

We're interested in getting the source/credit of the new feed

View the page source and look at the top to see the namespace declaration:

```
<rss xmlns:media="http://search.yahoo.com/mrss/" version="2.0">
```

So let's try adding this:

```
news_items = root.findall("{http://search.yahoo.com/mrss/}credit")
print news_items
```

Does it work?

Namespaces in XML

Look at the file again… what the path to the text element?

channel/item/…

So we change it to:

```
news_items = root.findall("channel/item/{http://
search.yahoo.com/mrss/}credit")
print news_items
```

It's all about finding the right path!!

News Titles (Group Work)

Write a Python program that prints the title and credit of *each news item* in the feed, like this:

```
IDLE 2.6.6
>>> ================================ RESTART ================================
>>>
Current News Items:
--------------------------------------------------------------------------------
Title: MIT professor's speech at IU will focus on 'Big Data, Big Brother and Financial Regulation'
Credit: Indiana University
--------------------------------------------------------------------------------
Title: IU nuclear physicist receives NSF's top award for junior faculty
Credit: Indiana University
--------------------------------------------------------------------------------
Title: Department of Homeland Security officials to keynote the CACR Cybersecurity Summit
Credit: Indiana University
--------------------------------------------------------------------------------
Title: Complex networks researcher at IU fighting crime with mobile phone data
Credit: Indiana University
--------------------------------------------------------------------------------
Title: IURTC Spin Up company YC Bioelectric receives $307,787 award from the National Institutes of Health
Credit: Indiana University
--------------------------------------------------------------------------------
Title: IU cybersecurity expert: Heartbleed Bug puts computer users at risk
Credit: Indiana University
--------------------------------------------------------------------------------
Title: Star Trak: April 2014
Credit: Indiana University
--------------------------------------------------------------------------------
Title: $1.4 million DARPA grant to IU Informatics professor for streamlining programming
Credit: Indiana University
--------------------------------------------------------------------------------
```

News Titles (Solution)

```
import urllib
import xml.etree.ElementTree as ET

# the following is all in one single line of code:
conn = urllib.urlopen("https://www.iu.edu/~iunews/services/newsrooms/feeds/?
format=rss20&id=1232a17b814f4e1c77a8fb801f237996&sort=date")

lines = conn.read()
conn.close()

root = ET.XML(lines)

print "Current News Items:\n", "-"*80
news_items = root.findall("channel/item")

for news in news_items:
    print "Title:", news.find("title").text
    print "Credit:", news.find("{http://search.yahoo.com/mrss/}credit").text
    print "-"*80
```