

XML and HTML

APIs and Web Scraping with Python

vanderbi.it/py

Steve Baskauf



Structure of XML

```
<bookstore>
  <book category="fiction">
    <title lang="en">Harry Potter and the Philosopher's Stone</title>
    <author>J. K. Rowling</author>
    <year>1997</year>
    <price>15.95</price>
  </book>
  <book category="nonfiction">
    <title lang="de">Kritik der reinen Vernunft</title>
    <author>Immanuel Kant</author>
    <year>1781</year>
    <price>9.97</price>
  </book>
</bookstore>
```

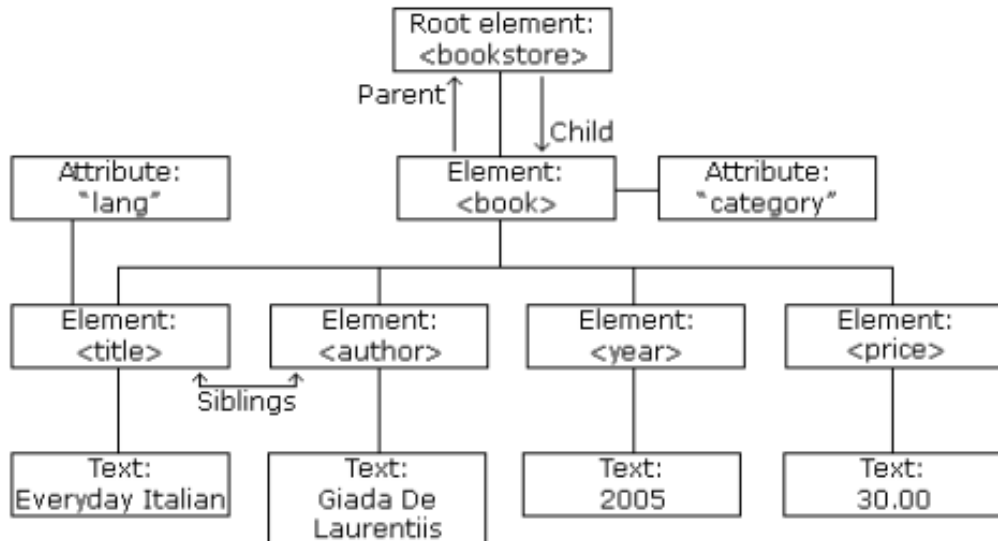
XML Tree Structure

Elements:

- have names
- contain text
- can have attributes
- can contain other elements

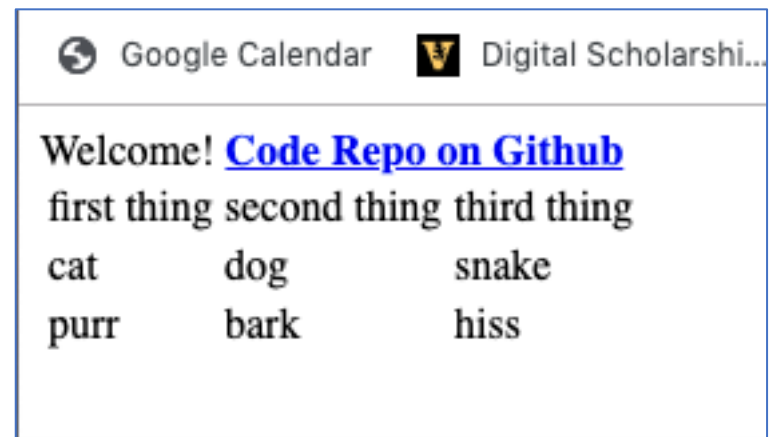
Relationships among elements:

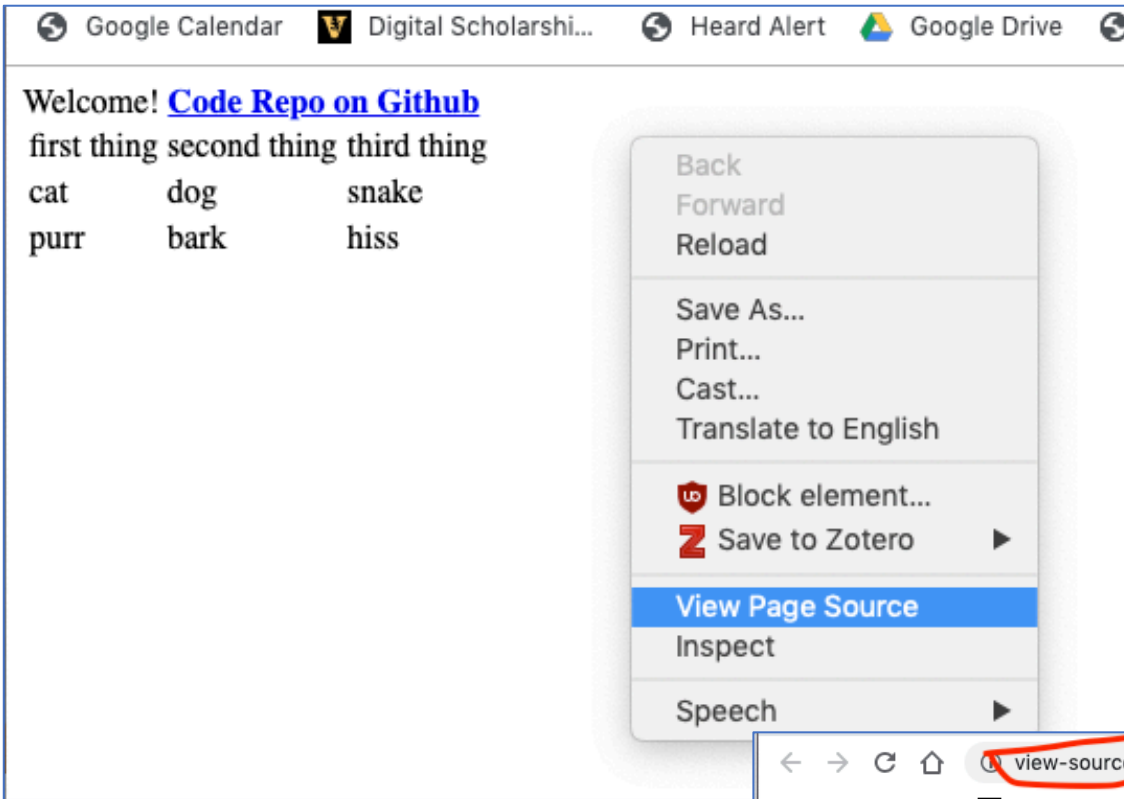
- parent (indent left)
- child (indent right)
- sibling (same indent)



HTML is XML (mostly)

```
<html>
  <head>
    <title>Digital Scholarship Resources</title>
    <meta charset="utf-8" />
  </head>
  <body>
    <div id="banner">
      <span id="logo">Welcome!</span>
      <a href="https://github.com/hey"><strong>Code Repo on Github</strong></a>
    </div>
    <table>
      <tbody>
        <tr>
          <td>first thing</td><td>second thing</td><td>third thing</td>
        </tr>
        <tr>
          <td>cat</td><td>dog</td><td>snake</td>
        </tr>
        <tr>
          <td>purr</td><td>bark</td><td>hiss</td>
        </tr>
      </tbody>
    </table>
  </body>
</html>
```





Viewing the page source

Right click on page
and select View Page Source

```
view-source:/Users/baskausj/Vanderbilt/Digital%20Scholarship%20and%20Communicati
Google Calendar Digital Scholarshi... Heard Alert Google Drive VU people finder Staff | Jean ar
1 <html>
2   <head>
3     <title>Digital Scholarship Resources</title>
4     <meta charset="utf-8" />
5   </head>
6   <body>
7     <div id="banner">
8       <span id="logo">Welcome!</span>
9       <a href="https://github.com/hey"><strong>Code Repo on Github</strong></a>
10    </div>
11    <table>
12      <tbody>
13        <tr>
14          <td>first thing</td><td>second thing</td><td>third thing</td>
15        </tr>
16        <tr>
17          <td>cat</td><td>dog</td><td>snake</td>
18        </tr>
19        <tr>
20          <td>purr</td><td>bark</td><td>hiss</td>
21        </tr>
22      </tbody>
23    </table>
24  </body>
25 </html>
26
```

Inspecting the HTML of a page

The screenshot shows a web browser with a page titled "Welcome! [Code Repo on Github](#)". The page content includes the text "first thing second thing third thing" and a table with animal sounds:

cat	dog	snake
purr	bark	hiss

A right-click context menu is open over the "snake" cell, with the "Inspect" option highlighted. The menu items are: "Look Up 'snake'", "Copy", "Search Google for 'snake'", "Print...", "Block element...", "Save to Zotero", "Inspect", "Speech", and "Services".

The developer tools panel is open, showing the "Elements" tab. The HTML structure is displayed as follows:

```
<html>
  <head>...</head>
  <body>
    <div id="banner">...</div>
    <table>
      <tbody>
        <tr>...</tr>
        <tr>
          <td>cat</td>
          <td>dog</td>
          <td>snake</td> == $0
        </tr>
        <tr>...</tr>
      </tbody>
    </table>
  </body>
</html>
```

The breadcrumb trail at the bottom of the developer tools shows: `html > body > table > tbody > tr > td`. The "Styles" panel is also visible at the bottom.

Right click on a page element and select Inspect

Drilling down on elements

Welcome! [Code Repo on Github](#)

first thing	second thing	third thing
cat	dog	snake
purr	bark	hiss

tbody 226 x 68

mouse over to highlight

click to expand or collapse

```
<html>
  <head>...</head>
  <body>
    <div id="banner">...</div>
    <table>
      <tbody> == $0
        <tr>...</tr>
        <tr>...</tr>
        <tr>
          <td>purr</td>
          <td>bark</td>
          <td>hiss</td>
        </tr>
      </tbody>
    </table>
  </body>
</html>
```

html body table tbody tr td

Styles Computed Event Listeners >>

Filter :hov .cls +

HTML of more complex pages

Many tags (div, span, tr) are used at different places in the document, but can be distinguished by their attributes

The screenshot shows the SEC Edgar website interface. The top navigation bar includes "Home | Latest Filings | Previous Page" and the "U.S. Securities and Exchange Commission" logo. The main heading is "Filing Detail". Below this, there are navigation links: "SEC Home » Search the Next-Generation EDGAR System » Company Search » Current Page".

The main content area displays filing information for Form 10-K. A yellow banner at the top of this section contains the text "attribute is an ID" with an arrow pointing to the "SEC Accession No. 0000034088-16-000065". Below this, a table lists filing details with columns for "Filing Date", "Period of Report", "Accepted", and "Documents". A red arrow points to the "Accepted" date with the text "attribute is an inline style".

Below the filing details is a table titled "Document Format Files". The first row is highlighted in blue. A red arrow points to the "Size" column of this row with the text "highlighted element". Another red arrow points to the "class" attribute of the first row in the table with the text "attribute is a CSS class".

The browser's developer tools are open on the right side, showing the HTML structure. Red arrows connect the text annotations to the corresponding HTML code: "attribute is an ID" points to the "id" attribute of a "div" tag; "attribute is an inline style" points to the "style" attribute of a "p" tag; "highlighted element" points to the "tr" tag in the table; and "attribute is a CSS class" points to the "class" attribute of the "tr" tag.

Document	Document	Type	Size
1	FORM 10-K	10-K	7694659
2	RESTATED CERTIFICATE OF INCORPORATION	EX-3.(I)	260323
3	BY-LAWS	EX-3.(II)	69809
4	EXTENDED PROVISIONS FOR RESTRICTED STOCK UNIT AGREEMENTS-SETTLEMENT IN SHARES	EX-10.(III) (A3)	35677
5	EARNINGS BONUS UNIT INSTRUMENT	EX-10.(III) (B2)	18923
6	STANDING RESOLUTION FOR NON-EMPLOYEE DIRECTOR CASH FEES	EX-10.(III) (F4)	4537
7	COMPUTATION OF RATIO OF EARNINGS TO FIXED CHARGES	EX-12	72539
8	SUBSIDIARIES OF THE REGISTRANT	EX-21	215233
9	CONSENT OF PRICEWATERHOUSECOOPERS LLP, INDEPENDENT REGISTERED PUBLIC ACCOUNTING	EX-23	17133
10	CERTIFICATION (PURSUANT TO SECURITIES EXCHANGE ACT RULE 13A-14(A)) BY CHIEF EXEC	EX-31.1	16086
11	CERTIFICATION (PURSUANT TO SECURITIES EXCHANGE ACT RULE 13A-14(A)) BY PRINCIPAL	EX-31.2	16570
12	CERTIFICATION (PURSUANT TO SECURITIES EXCHANGE ACT RULE 13A-14(A)) BY PRINCIPAL	EX-31.3	16764