XML and HTML

APIs and Web Scraping with Python

vanderbi.lt/py

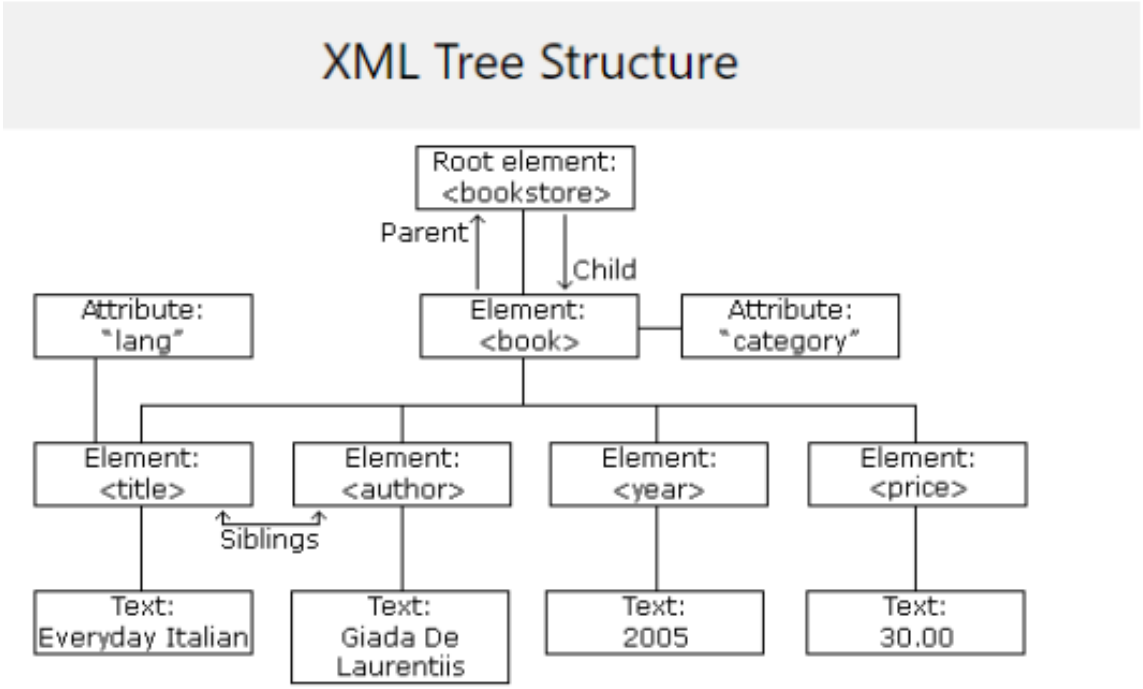
Steve Baskauf

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>
```

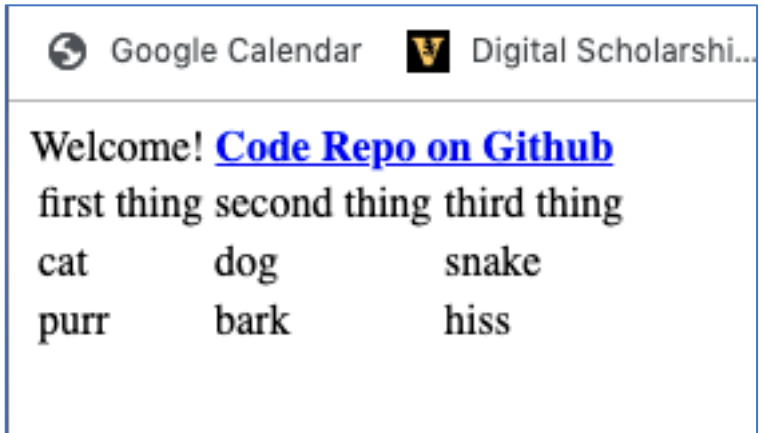
Elements:
- have names
- contain text
- can have attributes
- can contain other elements

Relationships among elements:
- parent (indent left)
- child (indent right)
- sibling (same indent)



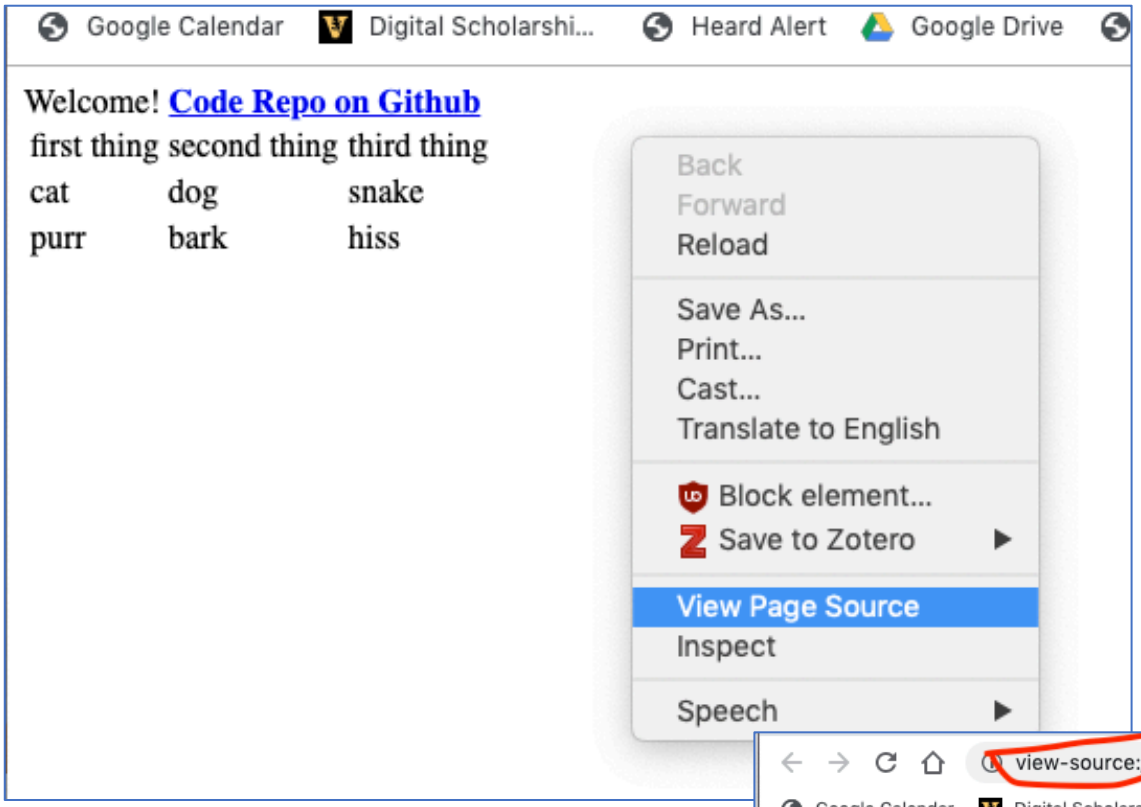
HTML is XML (mostly)

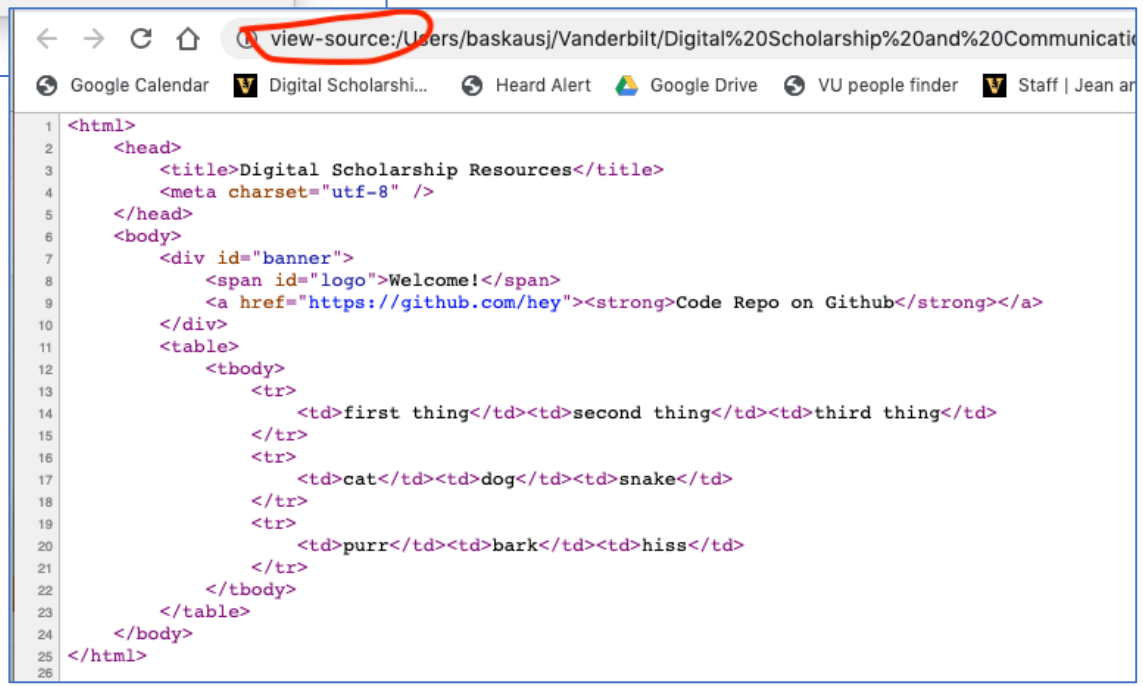
```
<html>
     <head>
          <title>Digital Scholarship Resources</title>
          <meta charset="utf-8" />
     </head>
     <body>
          <div id="banner">
               <span id="logo">Welcome!</span>
               <a href="https://github.com/hey"><strong>Code Repo on Github</strong></a>
          </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>
```



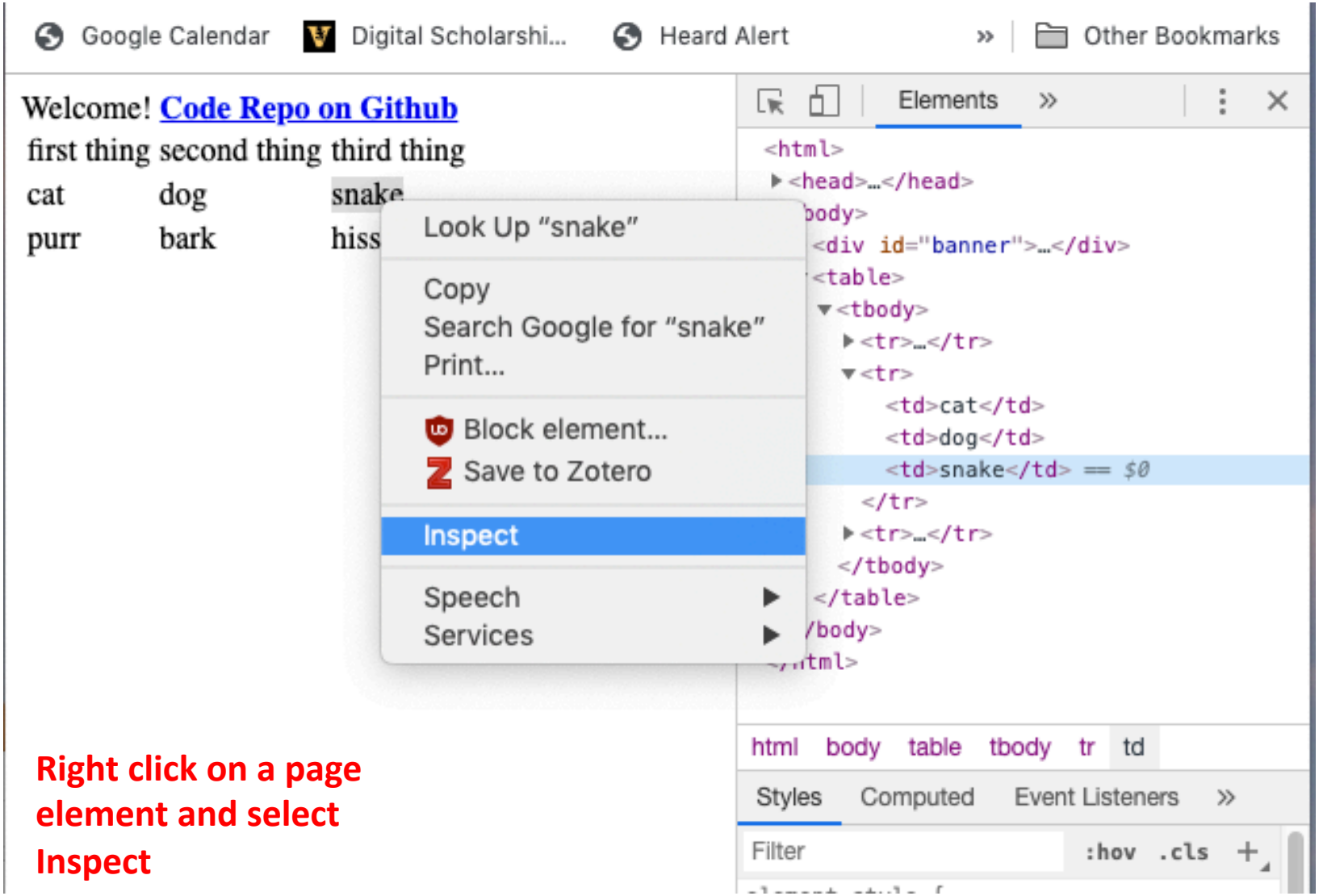
Viewing the page source

Right click on page and select View Page Source



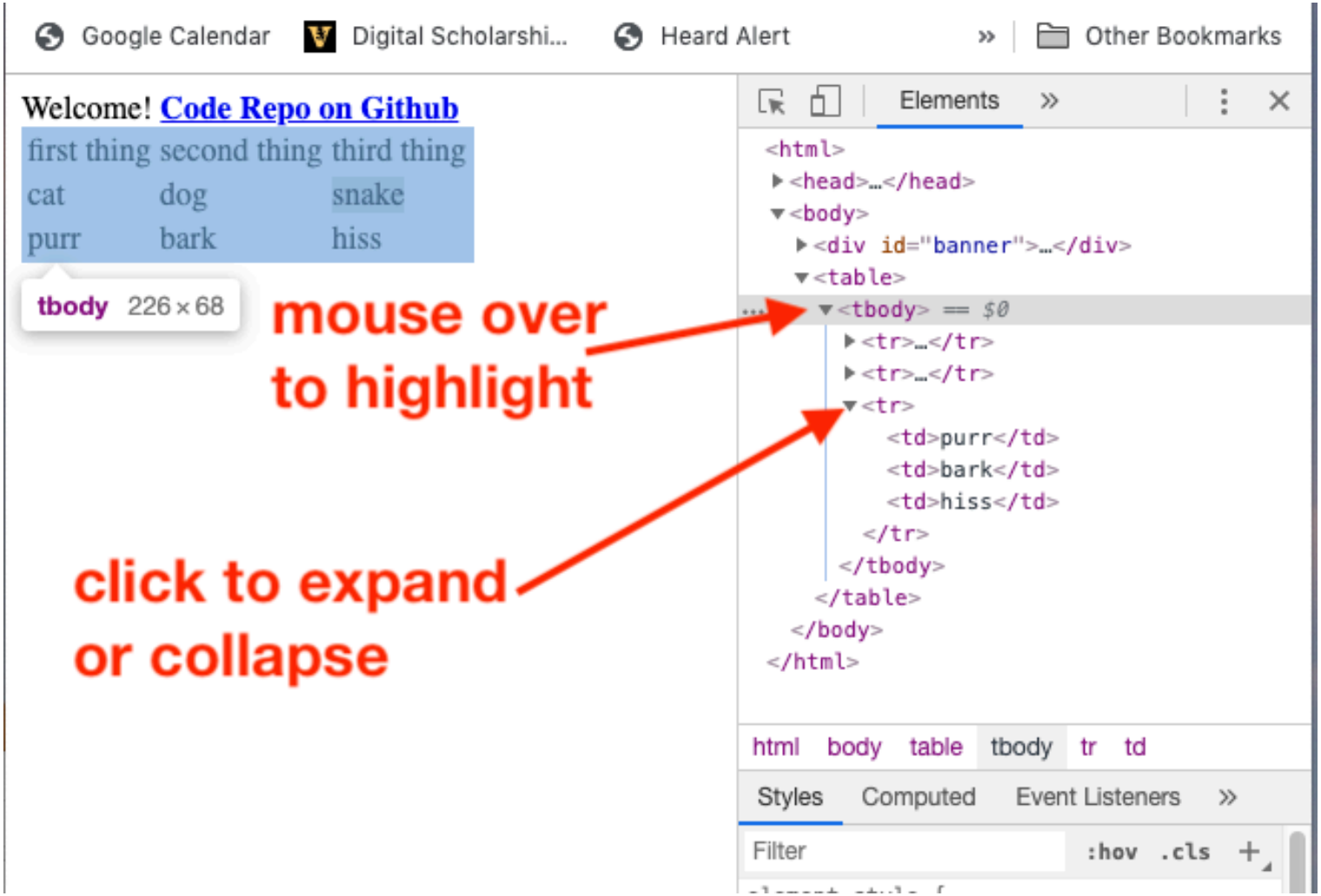


Inspecting the HTML of a page



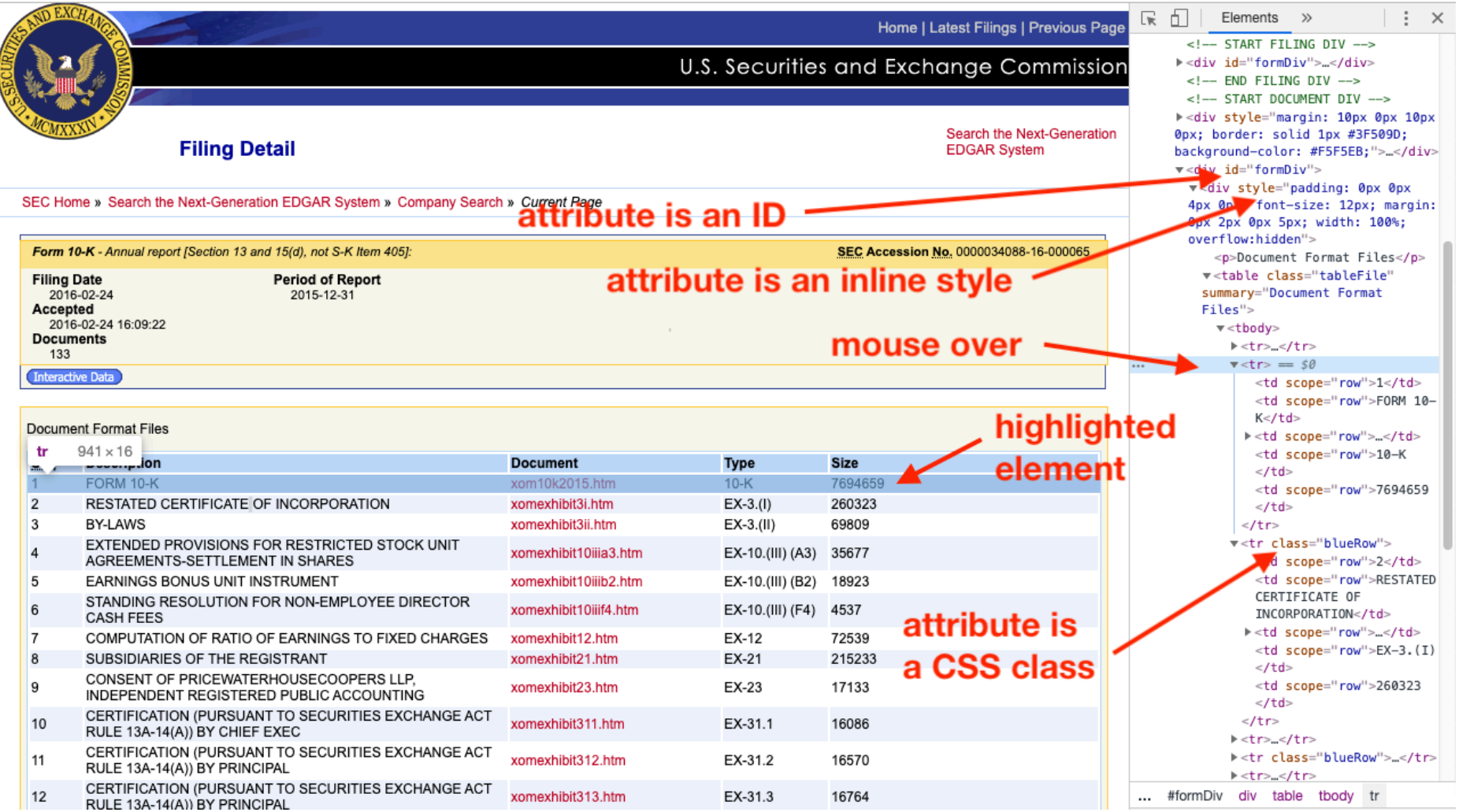
Right click on a page element and select Inspect

Drilling down on elements



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



Seq	Description	Document	Type	Size
1	FORM 10-K	xom10k2015.htm	10-K	7694659
2	RESTATED CERTIFICATE OF INCORPORATION	xomexhibit3i.htm	EX-3.(I)	260323
3	BY-LAWS	xomexhibit3ii.htm	EX-3.(II)	69809
4	EXTENDED PROVISIONS FOR RESTRICTED STOCK UNIT AGREEMENTS-SETTLEMENT IN SHARES	xomexhibit10iiia3.htm	EX-10.(III) (A3)	35677
5	EARNINGS BONUS UNIT INSTRUMENT	xomexhibit10iiib2.htm	EX-10.(III) (B2)	18923
6	STANDING RESOLUTION FOR NON-EMPLOYEE DIRECTOR CASH FEES	xomexhibit10iiif4.htm	EX-10.(III) (F4)	4537
7	COMPUTATION OF RATIO OF EARNINGS TO FIXED CHARGES	xomexhibit12.htm	EX-12	72539
8	SUBSIDIARIES OF THE REGISTRANT	xomexhibit21.htm	EX-21	215233
9	CONSENT OF PRICEWATERHOUSECOOPERS LLP, INDEPENDENT REGISTERED PUBLIC ACCOUNTING	xomexhibit23.htm	EX-23	17133
10	CERTIFICATION (PURSUANT TO SECURITIES EXCHANGE ACT RULE 13A-14(A)) BY CHIEF EXEC	xomexhibit311.htm	EX-31.1	16086
11	CERTIFICATION (PURSUANT TO SECURITIES EXCHANGE ACT RULE 13A-14(A)) BY PRINCIPAL	xomexhibit312.htm	EX-31.2	16570
12	CERTIFICATION (PURSUANT TO SECURITIES EXCHANGE ACT RULE 13A-14(A)) BY PRINCIPAL	xomexhibit313.htm	EX-31.3	16764