

XML AND BASEX

XQuery Working Group
Text Mining at Scale

```
<?xml version="1.0"?>
<quiz>
  <qanda seq="1">
    <question>
      Who was the forty-second
      president of the U.S.A.?
    </question>
    <answer>
      William Jefferson Clinton
    </answer>
  </qanda>
  <!-- Note: We need to add
       more questions later.-->
</quiz>
```



XML

```
<?xml version="1.0" encoding="UTF-8"?>
<meta xmlns="http://xml.house.gov/schemas/uslm/1.0"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:dcterms="http://purl.org/dc/terms/">
  <!-- sample metadata from the US Code -->
  <dc:title source="United States Code">Title 51</dc:title>
  <dc:type>USCTitle</dc:type>
  <docNumber>51</docNumber>
  <docPublicationName>Online@116-56</docPublicationName>
  <dc:publisher>OLRC</dc:publisher>
  <dcterms:created date="2019-06-07T11:10:31">June 7, 2019</dcterms:created>
  <dc:creator>USCConverter 1.5.3</dc:creator>
</meta>
```

Diagram illustrating the structure of the XML document:

- Root Element:** Points to the opening tag of the root element, <meta>.
- XML Declaration:** Points to the XML declaration, <?xml version="1.0" encoding="UTF-8"?>.
- Namespace Declarations:** Points to the xmlns declarations within the root element.
- Comment Node:** Points to the XML comment, <!-- sample metadata from the US Code -->.
- Attribute Node:** Points to the source attribute of the dc:title element.
- Element Node:** Points to the dc:creator element.
- Start Tag:** Points to the opening tag of the root element, <meta>.
- Text Node:** Points to the text node "Online@116-56" within the docPublicationName element.
- End Tag:** Points to the closing tag of the root element, </meta>.

Well-Formed XML

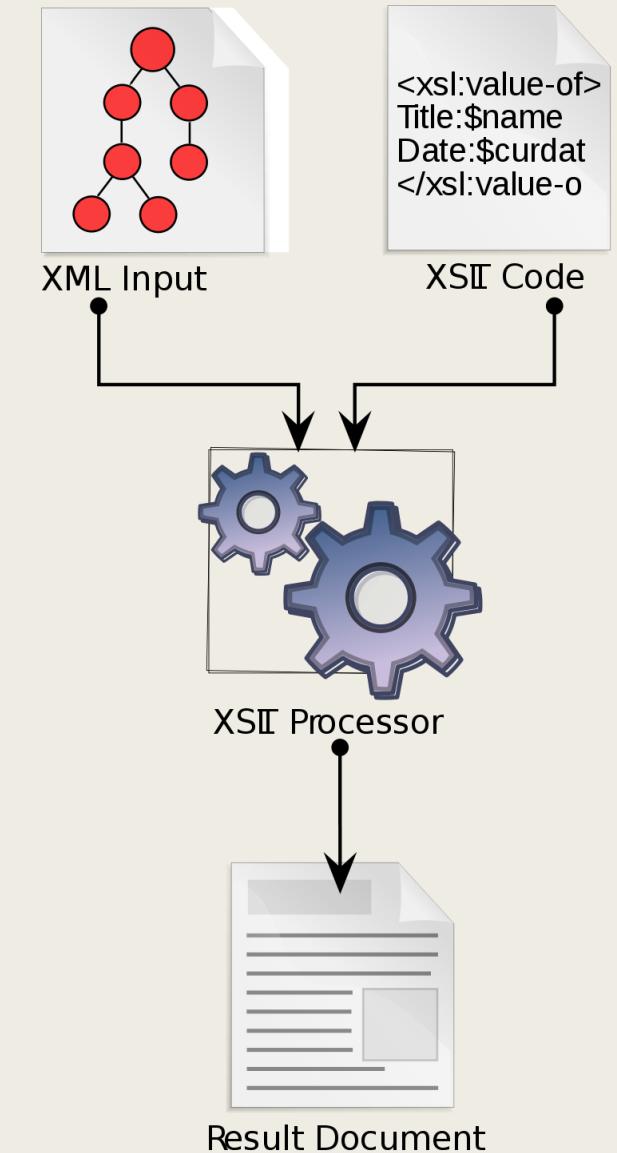
- Single “root element”
- Other elements are nested hierarchically without overlapping tags
- Contains only unicode characters
- Element names cannot begin with numbers
- Element names cannot contain spaces
- Certain characters are reserved and must be escaped
 - & → &
 - < → <

Validating XML

- Grammar-based validation
 - *Document Type Definitions (DTDs)*
 - *XML Schema Definitions (XSD)*
 - *RelaxNG*
 - XML Syntax
 - Compact Syntax
- Rule-based validation
 - *Schematron*
- Text Encoding Initiative
 - *One Document Does it All (ODD)*

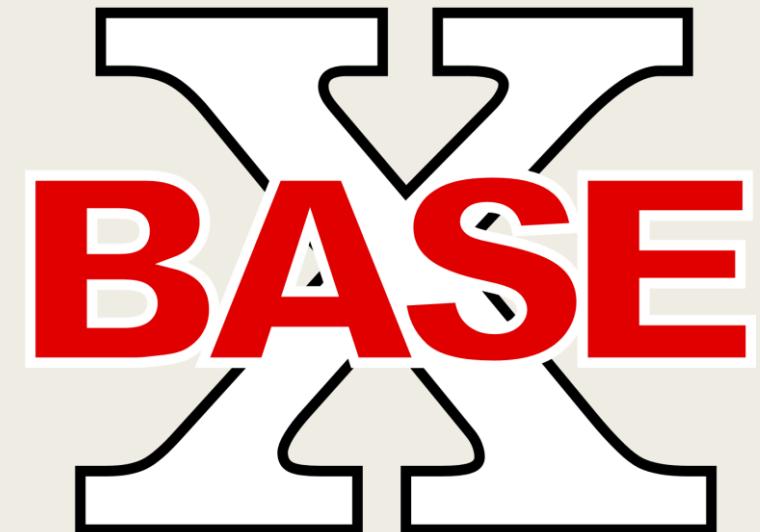
XML Programming Languages

- DOM and SAX
- XPATH
- XSLT
- XQuery



XML Processors and Databases

SAXONICA.COM
XSLT AND XQUERY PROCESSING



Installing BaseX

- Install Java Runtime Environment (JRE)
 - <https://www.java.com/en/download/>
- Install BaseX
 - *Windows*
 - Windows Installer: <http://basex.org/download/>
 - *Mac OSX*
 - Open Terminal
 - Install Homebrew: <https://brew.sh/>
 - Install BaseX with Homebrew: ‘brew install basex’