Introduction XML Technologies

Mark Graves

Agenda: XML Technologies

- Background
- Advantages and Disadvantages
- W3C Standards
- Document Processing (XPath, XSL)
- Databases (XML Query, XML Schema)
- Parsing (DOM, SAX)

What is XML?

- eXtensible Markup Language
- Syntax for data exchange
- Separates data from presentation on the web
- Hierarchical representation language
- Family of related Web standards

XML Document Example

```
<?xml version="1.0"?>
<genes>
  <gene id="14680">
    <name>BRCA1</name>
    <organism>Homo sapiens</organism>
    <chromosome_loc chr="17">17q21</chromosome_loc>
    <protein id="U37574"/>
    <DNA_sequence>atggattta</DNA_sequence>
    <db_xref gi="555931"/>
  </gene>
</genes>
```

When to use XML?

- Need to present data to both user and applications
- Desire a simple user interface
- Need to exchange data between applications
- Need to store complex relationships
- Need to merge data and documents

WWW Consortium (W3C)

- Develops specifications, guidelines, software, and tools for the WWW
- Develops common protocols to ensure interoperability of WWW
- Standards include:
  - HTML, URL, PNG, DOM
  - XML, XPath, XSL
  - XML Schema, XML Query

www.w3c.org

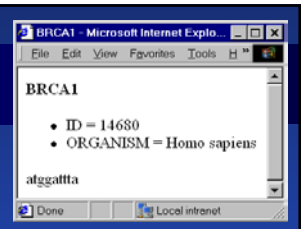
XPath

Purpose is to address parts of XML document

- "/" -- root element
- "." -- current element
- ".." -- parent element
- "gene" -- any gene
- "gene[@id=14680]"
- "gene/name"
- "gene/name/text()"
- "gene[@id=14680]/chr_loc"
- "gene[@id=14680]/protein/@id"

```
<?xml version="1.0"?>
<genes>
 <gene id="14680">
  <name>BRCA1</name>
  <organism>Homo sapiens</organism>
  <chr_loc chr="17">17q21</chr_loc>
  <protein id="U37574"/>
  <DNA_sequence>atggattta</DNA_sequence>
  <db_xref gi="555931"/>
 </gene>
</genes>
```

XSL Stylesheet



XSLT

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"
      xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="html"/>
<xsl:template match="gene">
  <B><xsl:value-of select="name"/></B>
  <ul>
    <li>ID = <xsl:value-of select="@id"/></li>
    <li>ORGANISM =
        <xsl:value-of
    select="organism"/></li>
    </ul>
  <xsl:value-of select="DNA_sequence"/>
</xsl:template>
</xsl:stylesheet>
```

XML

```
<?xml version="1.0"?>
<?xml:stylesheet type="text/xsl" href="gene .xsl"?>
<genes>
  <gene id="14680">
    <name>BRCA1</name>
    <organism>Homo sapiens</organism>
    <chr_loc chr="17">17q21</chr_loc>
    <protein id="U37574"/>
    <DNA_sequence>atggattta</DNA_sequence>
    <db_xref gi="555931"/>
  </gene>
</genes>
```

HTML

```
<B>BRCA1</B>
<ul>
<li>ID = 14680</li>
<li>ORGANISM = Homo sapiens</li>
</ul>
atggattta
```

XSL Stylesheet - FASTA

XSLT

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"
      xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="text" encoding="UTF-8"/>
<xsl:template match="gene">
&gt;<xsl:value-of select="@id"/>
<xsl:text> </xsl:text>
<xsl:value-of select="name"/>
<xsl:text> </xsl:text>
<xsl:value-of select="organism"/>
<xsl:text>
</xsl:text>
<xsl:value-of select="DNA_sequence"/>
</xsl:template>
</xsl:stylesheet>
```

XML

```
<?xml version="1.0"?>
<?xml:stylesheet type="text/xsl" href="fasta .xsl"?>
<genes>
  <gene id="14680">
    <name>BRCA1</name>
    <organism>Homo sapiens</organism>
    <chr_loc chr="17">17q21</chr_loc>
    <protein id="U37574"/>
    <DNA_sequence>atggattta</DNA_sequence>
    <db_xref gi="555931"/>
  </gene>
</genes>
```

FASTA

```
>14680 BRAC1 Homo sapiens
atggattta
```

XML Query

XML Query

```
<genes> {
 FOR $g IN
  document("genes.xml")/genes/gene
 WHERE $g/chr_loc/@chr = "17"
 RETURN
   <gene id={ $g/@id }>
      { $g/name }
   </gene>
} </genes>
```

XML

```
<?xml version="1.0"?>
<genes>
 <gene id="14680">
  <name>BRCA1</name>
  <organism>Homo sapiens</organism>
  <chr_loc chr="17">17q21</chr_loc>
  <protein id="U37574"/>
  <DNA_sequence>atggattta
  </DNA_sequence>
  <db_xref gi="555931"/>
 </gene>
</genes>
```

Result

```
<genes>
 <gene id="14680">
   BRCA1
 </gene>
</genes>
```

XML Schema

- **DTD -- Document Type Definition**
  - **Compatible with SGML definitions**
  - **Simple**
  - **Structure of Elements & Attributes**
- **XML Schema**
  - **Datatypes**
  - **Structure**
  - **XML Syntax**

XML Schema

XML

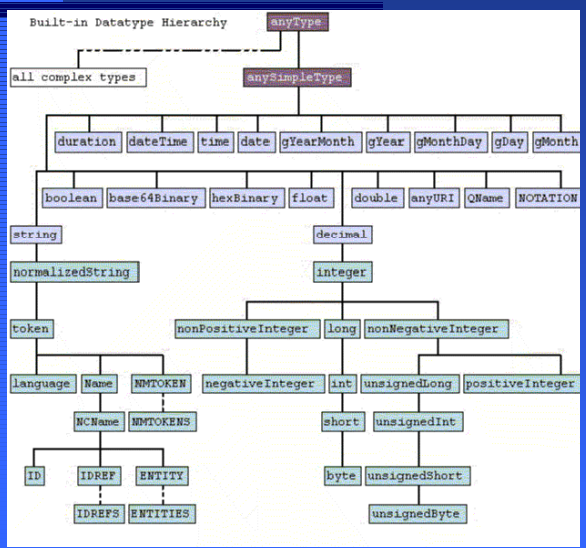
```
<?xml version="1.0"?>
<genes>
 <gene id="14680">
  <name>BRCA1</name>
  <organism>Homo sapiens</organism>
  <chr_loc chr="17">17q21</chr_loc>
  <protein id="U37574"/>
  <DNA_sequence>atggattta</DNA_sequence>
  <db_xref gi="555931"/>
 </gene>
</genes>
```

DTD

```
<!ELEMENT genes (gene*)>
<!ELEMENT gene (gene, name, organism, chr_loc,
                protein, DNA_sequence, db_xref?)>
<!ATTLIST gene
  id     CDATA #REQUIRED>
<!ELEMENT name CDATA>
<!ELEMENT organism CDATA>
<!ATTLIST chr_loc
  chr    CDATA #REQUIRED>
<!ELEMENT DNA_sequence CDATA>
<!ELEMENT db_xref EMPTY>
<!ATTLIST db_xref
  id     CDATA
  gi     CDATA>
```

```
<?xml version="1.0"?>
<schema
 xmlns="http://www.w3.org/2000/10/XMLSchema">
<element name="genes" type="genesType"/>
<complexType name="genesType">
 <element name="gene" type="geneType"/>
</complexType>
<complexType name="geneType">
 <attribute name="id"/>
 <element name="name"/>
 <element name="organism"/>
 <element name="chromosome_loc">
  <simpleType>
   <attribute name="chr"/>
  </simpleType>
 </element>
 <element name="protein">
  <simpleType>
   <attribute name="id"/>
  </simpleType>
 </element>
 <element name="DNA_sequence"/>
 <element name="protein">
  <simpleType>
   <attribute name="gi"/>
  </simpleType>
 </element>
</complexType>
</schema>
```

XML Schema Datatypes

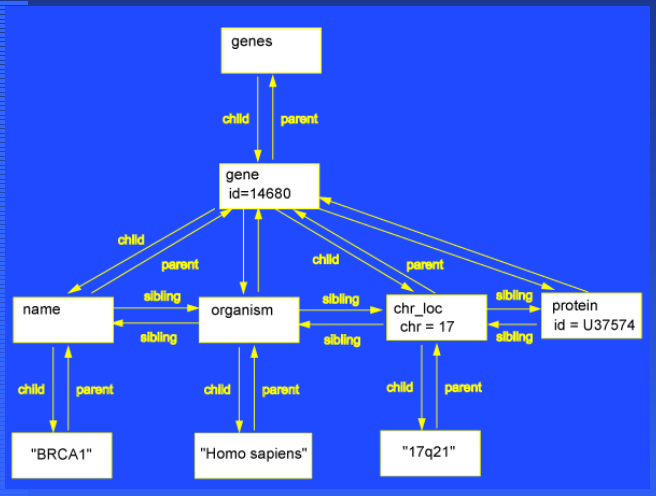


http://www.w3.org/TR/xmlschema-2/

XML Parsers

- **DOM**
  - Document Object Model
  - Defines tree-like data structure
  - In-memory access to data
- **SAX**
  - Simple API XML
  - API used to process or create custom data structure
  - Event-driven parser

Document Object Model (DOM)



Simple API XML (SAX)

- Event-driven parsing --- Method is called for each parsing event.
- Events:
  - start document
  - start element: name, AttributeList
  - character: char[], start, length
  - end element: name
  - end document
- SAX/SAX2

SAX Trace

XML

```
<?xml version="1.0"?>
<genes>
  <gene id="14680">
    <name>
      BRCA1
    </name>
    <organism>
      Homo sapiens
    </organism>
    <chr_loc chr="17">
      17q21
    </chr_loc>
    <protein id="U37574"/>
    <DNA_sequence>
      atggattta
    </DNA_sequence>
    <db_xref gi="555931"/>
  </gene>
</genes>
```

```
document start
start: genes
  start: gene {id="14680"}
    start: name
      chars: BRCA1
    end: name
    start: organism
      chars: Homo sapiens
    end: organism
    start: chr_loc {chr="17"}
      chars: 17q21
    end: chr_loc
    start: protein {id="U37574"}
    end: protein
    start: DNA_sequence
      chars: atggattta
    end: DNA_sequence
    start: db_xref {gi="555931"}
    end: db_xref
  end: gene
end: genes
document end
```