Data Formats and XML

Charles Severance
www.dr-chuck.com

Data on the Web

- With the HTTP Request/Response well understood and well supported there was a natural move toward exchanging data between programs using these protocols

Outline

- XML - Marking up data
- XHTML - HTML + XML
- XML Schema - Contracts for XML
- RSS - Data Feeds
**XML**

Marking up data to send across the network...


---

**eXtensible Markup Language**

- Primary purpose is to help information systems share structured data
- It started as a simplified subset of the Standard Generalized Markup Language (SGML), and is designed to be relatively human-legible


---

**XML Basics**

- Start Tag
- End Tag
- Text Content
- Attribute
- Self Closing Tag

```
<person>
  <name>Chuck</name>
  <phone type="intl">+1 734 303 4456</phone>
  <email hide="yes"/>
</person>
```

---

**White Space**

Line ends do not matter. White space is generally discarded on text elements. We indent only to be readable.

```
<person>
  <name>Chuck</name>
  <phone type="intl">+1 734 303 4456</phone>
  <email hide="yes"/>
</person>
```
Some XML...

```
<recipe name="bread" prep_time="5 mins" cook_time="3 hours">
  <title>Basic bread</title>
  <ingredient amount="8" unit="dl">Flour</ingredient>
  <ingredient amount="10" unit="grams">Yeast</ingredient>
  <ingredient amount="4" unit="dl" state="warm">Water</ingredient>
  <ingredient amount="1" unit="teaspoon">Salt</ingredient>
  <instructions>
    <step>Mix all ingredients together.</step>
    <step>Knead thoroughly.</step>
    <step>Cover with a cloth, and leave for one hour in warm room.</step>
    <step>Knead again.</step>
    <step>Place in a bread baking tin.</step>
    <step>Cover with a cloth, and leave for one hour in warm room.</step>
    <step>Bake in the oven at 180(\textdegree)C for 30 minutes.</step>
  </instructions>
</recipe>
```


XML Terminology

- Tags indicate the beginning and ending of elements
- Attributes - Keyword/value pairs on the opening tag of XML
- Serialize / De-Serialize - Convert data in one program into a common format that can be stored and/or transmitted between systems in a programming language independent manner


Sending Data across the “Net”

```
Python
Dictionary
```

```
Java
HashMap
```

a.k.a. “Wire Protocol” - What we send on the “wire”

Agreeing on a “Wire Format”

```
Python
Dictionary
Serialize

<person>
  <name>Chuck</name>
  <phone>303 4456</phone>
</person>
```

```
Java
HashMap
De-Serialize

<person>
  <name>Chuck</name>
</person>
```
**XML “Elements” (or Nodes)**

- Simple Element
- Complex Element

```xml
<people>
  <person>
    <name>Chuck</name>
    <phone>303 4456</phone>
  </person>
  <person>
    <name>Chuck</name>
    <phone>303 4456</phone>
  </person>
</people>
```

**XML as a Tree**

```
<root>
  <a>
    <b>B</b>
    <c>
      <d>D</d>
      <e>E</e>
    </c>
  </a>
</root>
```

**XML Text and Attributes**

```xml
<a>
  <b x="5">B</b>
  <c>
    <d>D</d>
    <e>E</e>
  </c>
</a>
```

**XML as Paths**

```
/a/b B
/a/c/d D
/a/c/e E
```
XHTML - HTML as XML

http://en.wikipedia.org/wiki/XHTML

XHTML Rules (versus HTML)

- Makes HTML also be well-formed XML
- Allows more flexible parsing of HTML
- All tags must have end tags
- All tag names are lower case (in XML case matters)
- Attributes must have quotes

XML Schema

Describing a “contract” as to what is acceptable XML.


Old HTML

<HTML>
<body>
  <p>Some text.</p>
  <p>Some more text</p>
  <a href="http://www.cnn.com">link</a>
</body>
</HTML>

New XHTML

<html>
<body>
  <p>Some text.</p>
  <p>Some more text</p>
  <a href="http://www.cnn.com">link</a>
</body>
</html>
XML Schema

- Description of the legal format of an XML document
- Expressed in terms of constraints on the structure and content of documents
- Often used to specify a “contract” between systems - “My system will only accept XML that conforms to this particular Schema.”
- If a particular piece of XML meets the specification of the Schema - it is said to “validate”

http://en.wikipedia.org/wiki/Xml_schema

XML Validation

XML Document

XML Schema Contract

Validator

Many XML Schema Languages

- Document Type Definition (DTD)
- Standard Generalized Markup Language (ISO 8879:1986 SGML)
- XML Schema from W3C - (XSD)
XSD XML Schema (W3C spec)

- We will focus on the World Wide Web Consortium (W3C) version
- It is often called “W3C Schema” because “Schema” is considered generic
- More commonly it is called XSD because the file names end in .xsd

http://www.w3.org/XML/Schema

XSD Structure

- xs:element
- xs:sequence
- xs:complexType

```
<xs:complexType name="person">
  <xs:sequence>
    <xs:element name="lastname" type="xs:string"/>
    <xs:element name="age" type="xs:integer"/>
    <xs:element name="dateborn" type="xs:date"/>
  </xs:sequence>
</xs:complexType>
```

```
<person>
  <lastname>Severance</lastname>
  <age>17</age>
  <dateborn>2001-04-17</dateborn>
</person>
```

XSD Constraints

- xs:element
- xs:sequence
- xs:complexType

```
<xs:complexType name="person">
  <xs:sequence>
    <xs:element name="full_name" type="xs:string" maxOccurs="1" minOccurs="1"/>
    <xs:element name="child_name" type="xs:string" maxOccurs="10" minOccurs="0"/>
  </xs:sequence>
</xs:complexType>
```

```
<person>
  <full_name>Tove Refsnes</full_name>
  <child_name>Hege</child_name>
  <child_name>Stale</child_name>
  <child_name>Jim</child_name>
  <child_name>Borge</child_name>
</person>
```

XSD Data Types

- xs:element
- xs:sequence
- xs:complexType

```
<xs:element name="customer" type="xs:string"/>
<xs:element name="start" type="xs:date"/>
<xs:element name="startdate" type="xs:dateTime"/>
<xs:element name="prize" type="xs:decimal"/>
<xs:element name="weeks" type="xs:integer"/>
```

```
<customer>John Smith</customer>
<start>2002-09-24</start>
<startdate>2002-05-30T09:30:10Z</startdate>
<prize>999.50</prize>
<weeks>30</weeks>
```

http://www.w3schools.com/Schema/schema_dtypes_numerical.asp
RSS - Rich Site Summary

- A way to list “what’s new” on a site
- Also called a “feed”
- Designed to be “scraped”
- Special Programs called “RSS Reader”, “News Reader”

http://en.wikipedia.org/wiki/Aggregator
Advertising an RSS Feed in HTML

```html
<head>
<title>Dr. Chuck's Interactive Personal Portfolio</title>
<link rel="alternate" type="application/rss+xml" title="RSS" href="http://www.dr-chuck.com/csev-blog/index.rdf" />
<link rel="alternate" type="application/atom+xml" title="Atom" href="http://www.dr-chuck.com/csev-blog/atom.xml" />
</head>
```

Several Versions of RSS

- There are two threads of RSS specification - the numbers overlap a bit - there must have been a conflict of approaches
- RDF RSS - RDF Site Summary
- RSS 2.0 - Really Simple Syndication

http://en.wikipedia.org/wiki/RSS_(file_format)
Summary

- We have protocols that allow cooperating applications to exchange data
- One common data format is XML
- XML Schema can be used to establish a contract about the format of a set of XML documents
- RSS is a format used for “feeds” that are published by sites and consumed by readers