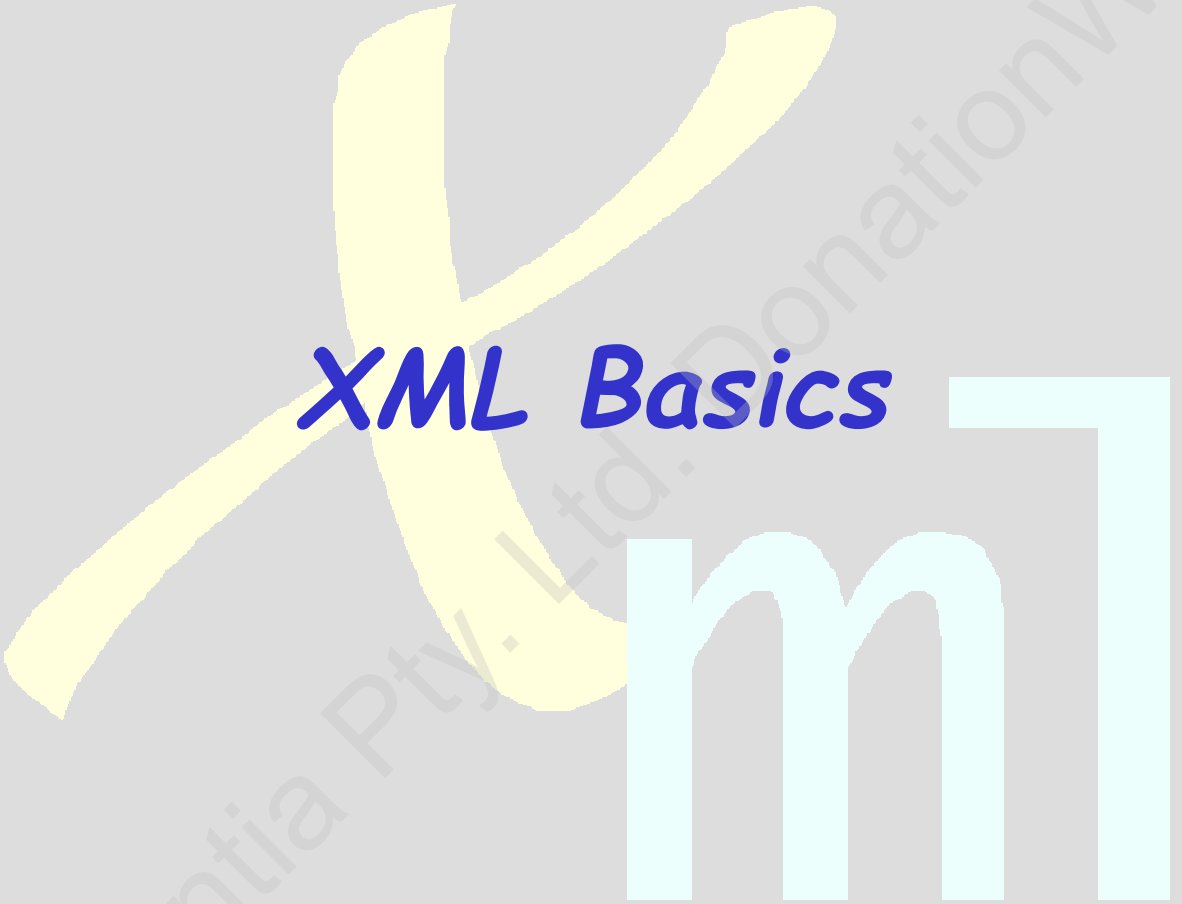
XML Basics



Oh Dear!

"Developers have a hard time agreeing on anything. They are fiercely loyal to the technologies they use and they don't like to compromise by adding support for the other guy's technology. Long and bitter technical wars have ensued, leaving interoperability the victim."

We can only hope…!

- **XML:**
  - is a method of structuring data
    - *to facilitate open interchange formats*
  - looks like HTML but isn't HTML
    - *permits developers to "get up to speed" quickly*
  - is text, but isn't meant to be read by humans
    - *it is mainly for software to parse and manipulate*
  - is actually a family of technologies
    - *XSL, XLink/XPointer, DOM, Schemas, etc.…*
  - is verbose, but not inefficient
    - *documents perhaps larger to store but may be more efficient to process than other formats*
  - is new, but not that new
    - *traces its roots to the 'sixties'*
  - is licence-free, platform and vendor neutral
    - *as all good technologies should be ☺*

- **Markup languages**
  - describe a document for processing purposes. Typically this has simply meant "make the data 'pretty.'"
    - *difficult/impossible to integrate with other business processes*
  - origins:
    - *'prehistory':RUNOFF, [ntg]roff, rtf, PostScript, proprietary formats*
    - *'70: Charles F. Golfarb (attorney): DCF GML*
    - *'86: SGML/HyTime*
    - *~ '94: HTML*

July 3, 2009

XML Basics...

MusicXMLDataIsland.html - Notepad

File Edit Search Help

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<HTML>
  <HEAD>
    <TITLE>Music Inventory</TITLE>
  </HEAD>
  <BODY>
    <XML ID=music>
      <INVENTORY>
        <ITEM>
          <NAME>Brandenburg Concerto No. 2</NAME>
          <AUTHOR>Bach, J. S.</AUTHOR>
          <PRICE>17.95</PRICE>
```

XML Basics.ppt - Notepad

File Edit Search Help

```
ÐÏ■à¡±■á                    >
■ þÿ      ■              ■   ■
■  þÿÿÿ     þÿÿÿ     ■    ■
ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
ÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿÿ
```

```
" CELLPADDING="2">
        <TD><DIV DATAFLD="AUTHOR"></TD>
        <TD><DIV DATAFLD="PRICE"></TD>
      </TR>
    </TABLE>
  </BODY>
</HTML>
```

nroff.1 - Notepad

File Edit Search Help

```
.TH NROFF 1 "26 June 1995" "Groff Version 1.11"
.SH NAME
nroff \- emulate nroff command with groff
.SH SYNOPSIS
.B nroff
[
.B \-h
]
[
.B \-i
]
```

XML Basics Outline.rtf - Notepad

File Edit Search Help

```
{\rtf1\ansi\ansicpg1252\uc1
\deff0\deflang1033\deflangfe1033
{\fonttbl{\f0\froman\fcharset0\f
prq2{\*\panose
02020603050405020304}Times New
Roman;}{\f1\fswiss\fcharset0\fpr
q2{\*\panose
020b0604020202020204}Arial;}
{\f3\froman\fcharset2\fprq2{\*\p
```

JDBC.pdf - Notepad

File Edit Search Help

```
%PDF-1.1■%âãÏÓ
1 0 obj■[/CalRGB■<<■/WhitePoint
[0.9505 1 1.089]■/Gamma [1.8 1.8
1.8]■/Matrix [0.4497 0.2446
0.02518 0.3163 0.672 0.1412
0.1845 0.08334 0.9227]■>>■]■e
ndobj■2 0 obj■<<■/Type /XObject■
Subtype /Image■/Name /Im1■/Width
320■/Height 240■/
```

ad

h Help

```
JL SET
=OFF■@PJL SET
PJL SET
IMAGEADAPT=AUTO■@PJL SET
RET=MEDIUM■@PJL SET
ECONOMODE=OFF■@PJL ENTER
LANGUAGE=PCL
■E■*r0F■&l10■&l1H■&l26a8c1E■*p0x0Y
```

XML Basics...

- XML: "The ASCII of the Future"
  - according to Microsoft, that is…
- XML follows SGML and allows markup according to content, not formatting

```
<?xml version="1.0"?>
<EMAIL>
  <TO>king@transentia.com.au</TO>
  <FROM>wizard@transentia.com.au</FROM>
  <CC>queen@transentia.com.au</CC>
  <SUBJECT>Marking up an email with XML</SUBJECT>
  <BODY>How does this seem to you?</BODY>
</EMAIL>
```

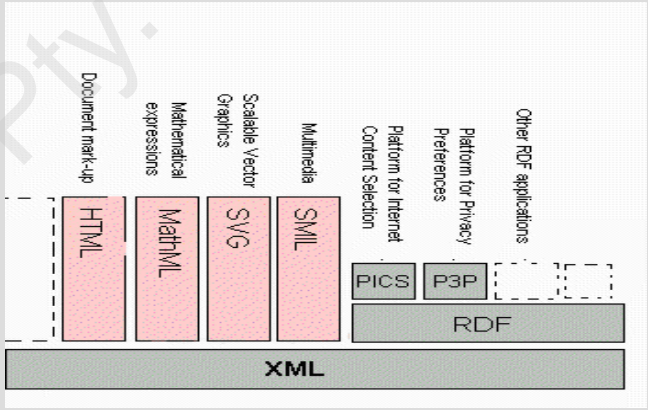
```
<HTML>
  <HEAD><TITLE>Memo</TITLE></HEAD>
  <BODY>
    <P>
      <STRONG>To:</STRONG> king@transentia.com.au
    </P>
    <P>
      <STRONG>From:</STRONG> wizard@transentia.com.au
    </P>
    <P>
      <STRONG>CC:</STRONG> queen@transentia.com.au
    </P>
    <P>
      <STRONG>Subject: </STRONG>Marking up an email with HTML
    </P>
    <P>How does this look to you?</P>
  </BODY>
</HTML>
```

HTML has been dismissed as "Macbeth Multimedia… full of sound and fury, signifying nothing."

- **An Application Profile of SGML**
  - a subset designed more with an eye to efficient processing than extreme expressiveness
  - SGML regarded as being too 'big' for the Web
    - *an appropriate tool for a 4 million page 747 manual, not for a 2K web page…*
    - *thus HTML…a small application of SGML*
  - SGML too processing-intensive
    - *don't want to take 5 mins. to see a 2K page...*

XML...

- like SGML, XML is a metalanguage: a language for defining other languages
  - *these languages can be tailored for specific purposes: describing equations; defining data channels; describing the data format for health records or financial data, etc.*
  - *XML Vocabulary*
    - a set of the actual elements and the structure for a particular document type
      - *developed to support the operation of various vertical industries*
    - many are in development



- XML provides 80% of the power of SGML for 20% of the cost

Goals of XML

- **To be *easily* usable over the Internet**
  - or any situation requiring data interchange; the Internet is simply the most 'trendy' example…
    - *but we all know how the Internet is going to become the center of the universe Real Soon Now...*
  - must be simpler than SGML
- **To support a wide variety of applications**
  - authoring tools, search engines, databases, publishing engines, etc.
- **To be compatible with SGML**
  - allows easy adoption; compatible with government requirements; existing SGML tools can process XML easily
- **It must be easy to process XML**
  - so writing applications is easy; designers originally had the idea of a "two week" benchmark

More Goals

- **Should have very few optional features**
  - ideally zero; SGML suffered by trying to be everything to everybody
- **To be human-readable**
  - XML's textual basis makes life a lot easier for a developer
  - SGML allows for many strange abbreviations and shortcuts; terseness is of minimal importance for XML
  - *may* be **marginally** less efficient
- **XML's design shall be formal and concise and also prepared as quickly as possible**
  - formality should make life easier for everybody
  - needs to be defined according to "Internet speed"
- **XML documents shall be easy to create**
  - implies that it should also be easy to make good tools

Semantic Markup

- **HTML is oriented towards *appearances***
  - *"this should look like…"*
  - the *meaning* of the data is practically ignored
    - *makes subsequent processing difficult*
  - reflects HTML's simple goals
- **XML is oriented towards highlighting data *structure***
  - *"this is a…"*
  - the *appearance* of the data is ignored
  - many people ***wrongly*** assume that if the structure of data is explicit, then the meaning (semantics) of that data is also known; hence they say that XML is useful for semantic markup
    - *but semantics of a piece of data is encapsulated in the* user *of that data, not in its representation…*

More Semantics

- **Capturing the meaning of the data (through its structure) has a number of potential benefits**
  - *'Using XML, the word "bill" can be tagged as a name, a charge, a paper currency, a proposed law, or the mouth of a bird. Tagging the data appropriately (e.g. "<person-name>Bill</person-name>") allows for efficient machine processing (by search engines, for example). Without this ability to distinguish the different semantics, a lawyer would have to weed out a lot of irrelevant data mixed in with the appropriate search results, for example.'*

Semantics...

- **Semantics are in the eyes of the beholder…**
  - the structure of this data is quite explicit:

```
<female>
  <shape>
    hour-glass
  </shape>
  <age>
    22
  </age>
  <eye-color>
    green
  </eye-color>
  <hair>
    <length>waist</length>
    <color>redhead</color>
  </hair>
  <legs>long</lips>
  <lips>pouting</lips>
  <image>http://www.women-galore.com/xt77zz93490t.html</image>
</female>
```

  - is this beauty?
    - *a semantic attribute dependent on the processor's (your!) operation...*

Related Standards

- **Synchronized Multimedia Integration Language (SMIL ["SMILe"])**
  - covers audio, video, and animations. Also addresses the issue of synchronizing elements.
- **Mathematical Markup Language (MathML)**
  - deals with the representation of mathematical formulas
- **Scalable Vector Graphics (SVG)**
  - covers the representation of vector graphic images
- **Drawing Meta Language (DrawML)**
  - a W3C note that covers 2D images for technical illustrations
- **Commerce XML (cXML)**
  - a RosettaNet (www.rosettanet.org) standard for setting up interactive online catalogs for different buyers

Related Standards...

- **Common Business Library (CBL)**
  - a library of element and attribute definitions maintained by CommerceNet (www.commerce.net)
- **Desktop Management Task Force (DMTF)**
  - XML-based standards to remotely administer desktop equipment
- **Web Distributed Authoring and Versioning (WebDAV)**
  - an effort from the IETF that uses XML to maintain/synchronise web servers
- **Resource Description Framework (RDF)**
  - a data modelling framework that serves as a foundation for processing metadata (data about data)

Allied Technologies

- **XSL**
  - essentially two items
    - *transformations*
    - *styling/formatting*
- **Xlink/XPointer**
  - hyperlinking and addressing of data
- **DOM**
  - Document Object Model
- **Schemas**
  - a sophisticated mechanism for making an XML document "self describing" and ensuring validity of the described data
- **Namespaces**
  - differentiating sets of tags
- **XML Patterns/XPath**
  - querying XML-based data repositories

- **XML development carried out under auspices of W3C**
  - an industry consortium, not a formal standards body
- **W3C issues recommendations**
  - XML 1.0 issued Feb. '98
  - XML Namespaces in Jan. '99
  - many other associated bodies of work-in-progress
    - *XSL (styling); Xlink/XPointer (linking); Schemas; Patterns; XML data fragments; RDF (metadata), etc.*

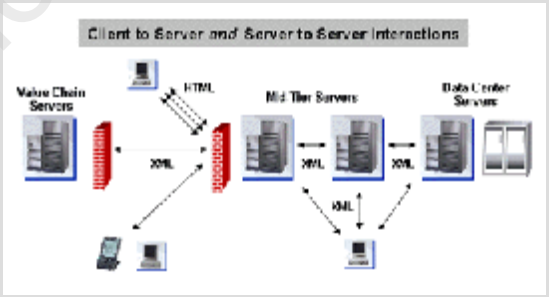
> *"Since the World Wide Web Consortium (W3C) ratified XML 1.0 as an internet standard two years ago, vendors and working groups have created a plethora of XML standards. While this bumper crop clearly demonstrates XML's wide level of adoption and support, it also creates confusion in the marketplace as developers struggle to tell their OASIS from their OAGIS."*

- **W3C recommendations may become true standard eventually**
  - *"The process…for creating a Recommendation is an alternative to, and not a replacement for,…the standards process…"*
- **XML is in a "state of flux"**
  - lots of input from many sources
    - *a few false starts taken*
  - conflicting vendor 'interests'
  - implementing, testing and evaluating ideas takes time and effort

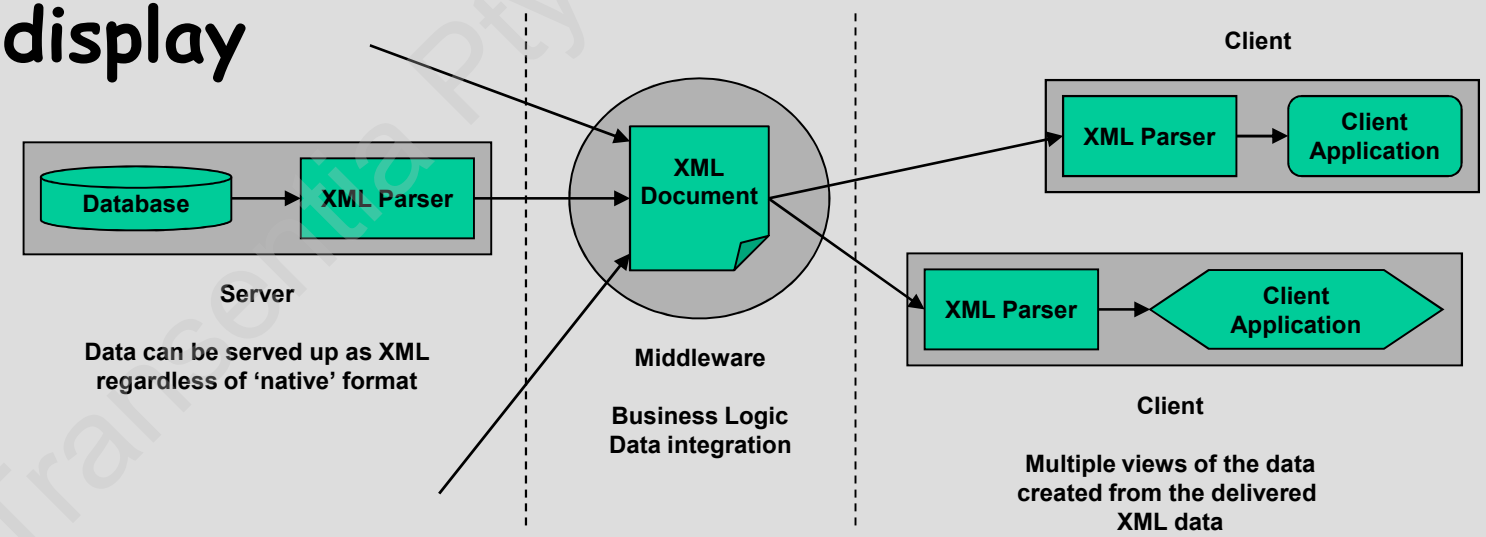
XML and HTML

- **XML is not backwards-compatible with HTML**
  - some 'retooling' needed
  - HTML 3.2/4.0 docs. can easily be converted
- **DHTML/XHTML**
  - DHTML aims to provide greater interactivity in an HTML document
  - XHTML is a reformulation of HTML 4.0 in XML 1.0

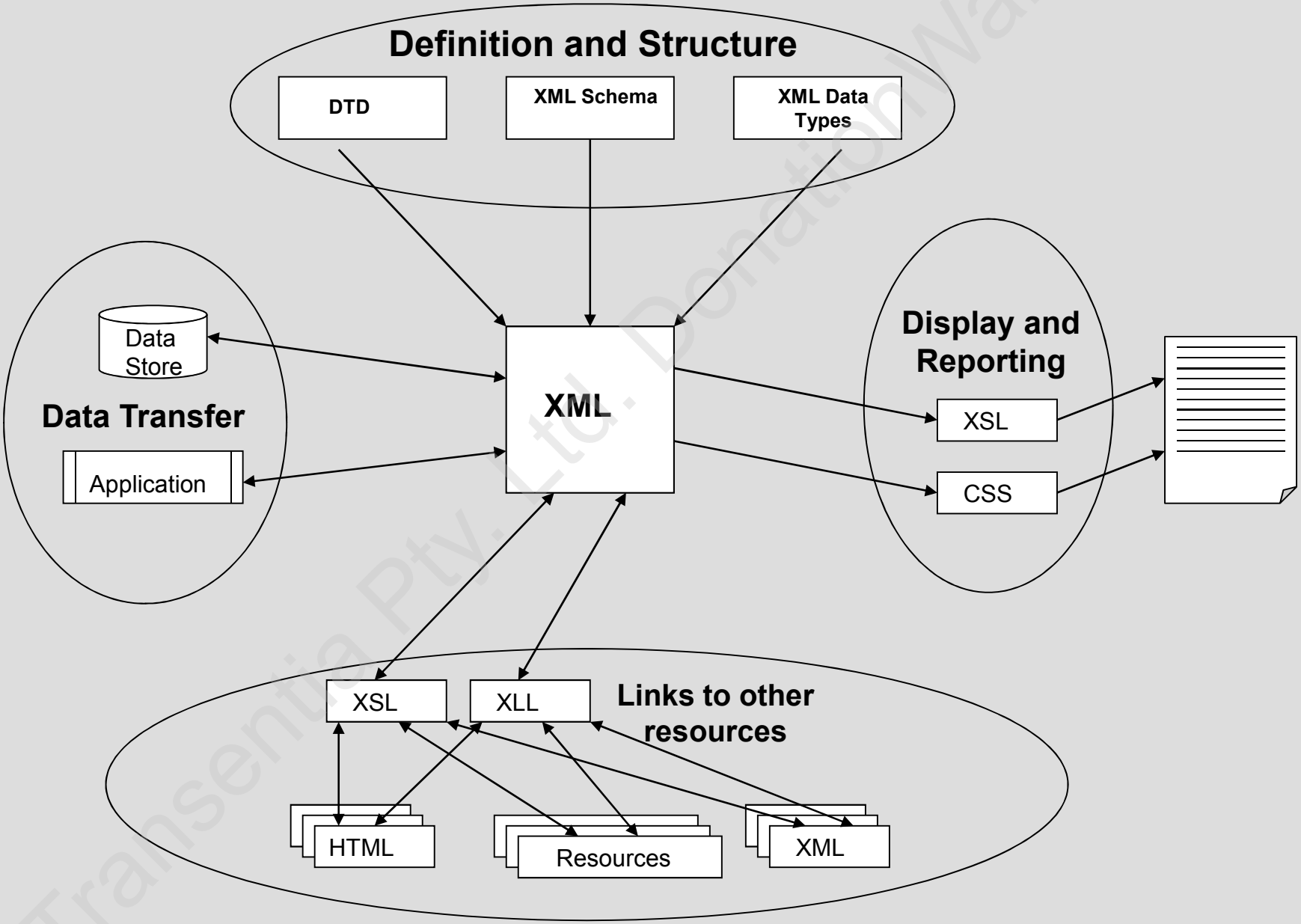
- XML is helping to facilitate the move to "a global information sharing society."
- XML has relevance throughout a tiered enterprise architecture
  - XML offers a robust solution as the underlying architecture for data in *n-tier* architectures



- With XML, structured data is maintained separately from the business rules and the display



All the Pieces



- **XML & Java enjoy a special partnership**
  - portable processing/portable data
  - much XML software is written in Java
  - XML and Java are two tools operating in the same "problem space":
    - *the correspondence between Document Type and Class Type makes Java and XML natural partners*
    - *heterogeneity of data/platforms*
    - *long lifetimes of data/services*
    - *need to take account of internationalization*
    - *etc.*

"Let's face it; XML by itself is just another data format that is annoying to access from your Java programs."

- **XML is rapidly becoming the lingua franca of e-Commerce**
  - according to IBM:
    - *"In the business domain there will soon be specific XML languages to describe orders, transaction, inventory and billing. These open XML languages will allow manufacturers, retailers, and consumers, even banking and accounting systems to share the same data."*
  - much activity…
    - *Consortia*
      - BizTalk, CommerceNet, FinXML, OASIS, RosettaNet
    - *Commerce*
      - Small to Medium Business eXtensible Markup Language
      - Common Markup for micropayment per-fee-links
      - Digital Receipt DTD
      - Digital Signatures for XML
    - *Etc.*

- **BizTalk**
  - an industry initiative started by Microsoft and supported by a wide range of organizations, from technology vendors like SAP and CommerceOne to technology users like Boeing and BP/Amoco
  - has the goal of driving the rapid, consistent adoption of XML to enable electronic commerce and application integration
  - the BizTalk Framework™
    - *a set of guidelines for how to publish schemas in XML and how to use XML messages to easily integrate software programs together in order to build rich new solutions*
  - the BizTalk Server
    - *a server that will facilitate the interchange of information that is encoded according to the BizTalk Framework*
  - www.biztalk.org
    - *a repository for locating, managing, learning about and publishing XML, XSL and the associated information models*

- **A well-understood set of elements and structure for a specific document type**
  - used to develop data interchange systems for specific vertical industries
  - examples:
    - *Channel Definition Format, Open Financial Exchange, Open Software Description, HL7, Wireless ML, First Retail ML, FinXML*
    - *many others under development…*

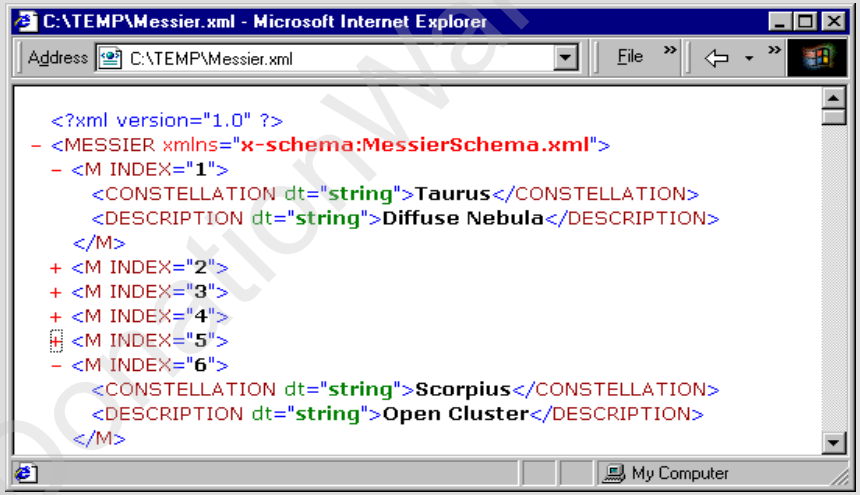
- **Microsoft Office**
  - HTML & XML as a native format
  - enable “round-tripping” allowing customers to save an Office document as HTML and reopen it in Office...XML data will be used to annotate HTML files, preserving Office-specific functionality.

```
<html xmlns:o="urn:schemas-microsoft-com:office:office"
xmlns:w="urn:schemas-microsoft-com:office:word"
xmlns="http://www.w3.org/TR/REC-html40">

<head>
<meta http-equiv=Content-Type content="text/html; charset=windows-1252">
<meta name=ProgId content=Word.Document>
<meta name=Generator content="Microsoft Word 9">
<meta name=Originator content="Microsoft Word 9">
<link rel=File-List href="./Hello_files/filelist.xml">
<title>Hello, Word</title>
<!--[if gte mso 9]><xml>
 <o:DocumentProperties>
  <o:Author>Bob Brown</o:Author>
  <o:Template>Normal</o:Template>
  <o:LastAuthor>Bob Brown</o:LastAuthor>
  <o:Revision>1</o:Revision>
  <o:TotalTime>1</o:TotalTime>
  <o:Created>2000-02-23T23:35:00Z</o:Created>
  <o:LastSaved>2000-02-23T23:36:00Z</o:LastSaved>
  <o:Pages>1</o:Pages>
  <o:Company>Transentia Pty. Ltd.</o:Company>
  <o:Lines>1</o:Lines>
  <o:Paragraphs>1</o:Paragraphs>
  <o:Version>9.2720</o:Version>
 </o:DocumentProperties>
</xml><![endif]-->
```


```
<style>
<!--
 /* Style Definitions */
p.MsoNormal, li.MsoNormal, div.MsoNormal
                {mso-style-parent:"";
                margin:0in;
                margin-bottom:.0001pt;
                mso-pagination:widow-orphan;
                font-size:12.0pt;
                font-family:"Times New Roman";
                mso-fareast-font-family:"Times New Roman";}
h1
                {mso-style-next:Normal;
                margin-top:12.0pt;
                margin-right:0in;
                margin-bottom:3.0pt;
                margin-left:0in;
                mso-pagination:widow-orphan;
                page-break-after:avoid;
                mso-outline-level:1;
                font-size:16.0pt;
                font-family:Arial;
                mso-font-kerning:16.0pt;}
@page Section1
                {size:8.5in 11.0in;
                margin:1.0in 1.25in 1.0in 1.25in;
                mso-header-margin:.5in;
                mso-footer-margin:.5in;
                mso-paper-source:0;}
div.Section1
                {page:Section1;}
-->
</style>
</head>
```

```
<body lang=EN-AU style='tab-interval:.5in'>

<div class=Section1>

<h1>Hello, Word</h1>

<p class=MsoNormal>What does this look like?</p>

</div>

</body>

</html>
```

- **IE5; Mozilla**
  - but in different ways…
- **Others**
  - various specific applications such as Amaya (MathML) and Jumbo (ChemML); Microsoft's Chrome (2/3D animations)
  - also supported by: Oracle 8i, ArborText, Inso, Chrystal Software, POET, and Microstar, DataChannel Inc., UserLand Software, Vignette Corp.



To be Successful...

- **People and organisations must see the value of XML**
  - as they did with HTML
    - *but now we need more than "pretty pictures"...*
- **Better tools must become available**
  - *"...the web browser must become a stable building block for site designers, just as standardisation on Windows has encouraged innovation in the PC space."*
- **Standardization needs to continue**
  - and then be adopted properly

"...despite all the hype, XML is really just a new format for data stored in text files. However, its simplicity, combined with its platform and application independence, means that it is being used in an increasing number of areas where the exchange of data is required—especially between disparate systems."