2. XML & Data Management Usage

Data Reduction Savings

- Reduction in the number of physical servers (to as few as 10% of the original number)
- Reduction in the technical complexity of the operating environment (increasingly Linux)
- Reduction in the number of TCP/IP connections (75% seems to be the median number quoted)
- Reduction in the per-connection cost of integration.

Similar to server consolidation

- $3 million in annual savings;
- $500,000 per month savings in disk and tape;
- 100% payback within 14 months;
- 72% reduction in support resource costs;
- 58% increase in disk utilization;
- 57% reduction in tape;
- 700-square-foot reduction in floor space;
- 10-25% performance enhancement keeping availability the same or better

Smaller volume of data

- Smaller volume of data
- Common metadata model implementation
- Understanding data structures as XML
- Pareto Analysis
  - 20% of your data implements 80% of your functionality

Three Laws of Data Management

1. Upon identification, it is a data manager's role to ensure that a data's metadata is managed in a manner that provides metadata repository functionality
2. It is a data manager's role to report regularly the amount of money that you as a metadata manager have saved the organization (always delineating savings from increases in data quality)
3. Data managers work at the pleasure of business users

To get there …

- In order to be able to manage less physical data …
- We must create and manage more metadata

XML Hype

- *Did you know XML can cure cancer?*
- *Make you lose 20 pounds?*
- *Provide renewable, non-polluting energy?*
- *Sorry, but it almost seems like this when XML continues be touted as the new be-all and end-all of IT. To be fair, Aiken presented many wonderful scenarios and situations where, of course, it would be wonderful if Everything talked to Everything Else.*
  - Comment by a delegate after one of my seminars

Remember Wordstar?

- To change text attributes you embedded tags into the text. For example if you wanted the next word to be bolded, you would add the tag before the next word. To turn the bolding off, you used a to remove future bolding. Similarly, <u> began underlining and </u> ended underlining formatting. Italics was <i>, etc.

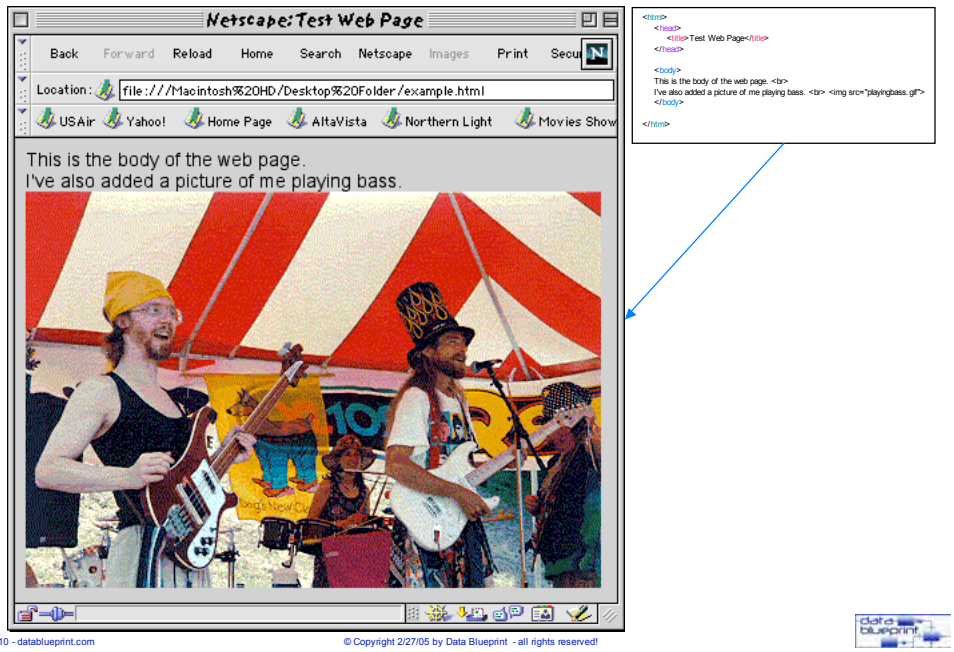
Interpreted Text

- To change text attributes you embedded tags into the text. For example if you wanted the next word to be bolded, you would add the tag **before the next word. To turn the bolding off, you used a** to remove future bolding. Similarly, began underlining and ended underlining formatting. Italics was*, etc.*

HTML Example

```
<html>
  <head>
     <title>Test Web Page</title>
  </head>

  <body>
  This is the body of the web page. <br>
  I've also added a picture of me playing bass.
  <br> <img src="playingbass.gif">
  </body>

</html>
```


HTML versus **XML**



Markup in HTML

- Tag-based language to define document display properties
- Standards set by W3C, so all browsers know how to handle
  - paragraph
  - bold
  - italic
  -
 line break
  - <hr> horizontal rule

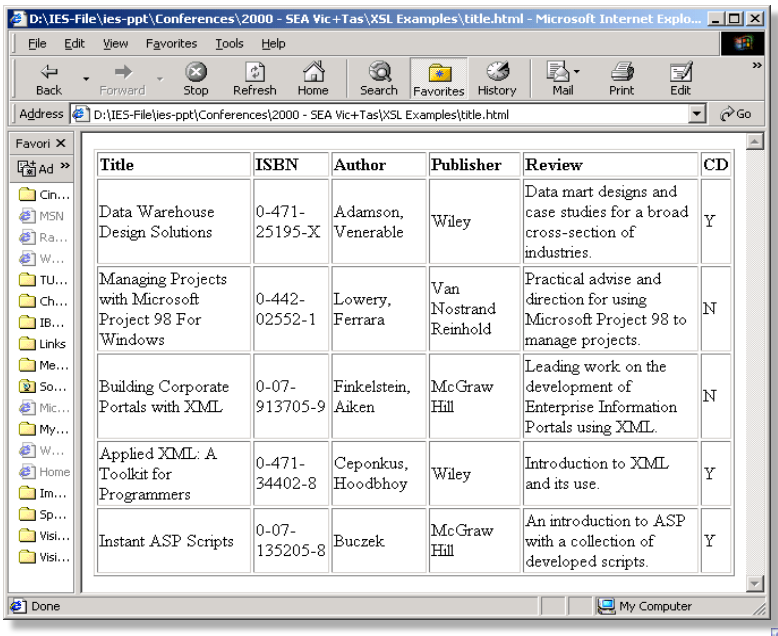
Markup in HTML = Content Loss

- HTML tags specify display format, but not context

```
<TR>
        <TD><b>Title</b></TD>
        <TD><b>ISBN</b></TD>
        <TD><b>Author</b></TD>
        <TD><b>Publisher</b></TD>
        <TD><b>Review</b></TD>
        <TD><b>CD</b></TD>
</TR>
<TR>
        <TD>Building Corporate Portals Using XML</TD>
        <TD>0-07-913705-9</TD>
        <TD>Finkelstein, C. and P.H. Aiken</TD>
        <TD>McGraw-Hill</TD>
        <TD>This book is about portal design, and uses an example that is
              simple in concept, yet meaty in production.</TD>
        <TD>N</TD>
</TR>
```

Markup in XML

Representation in Browser



Title	ISBN	Author	Publisher	Review	CD
Data Warehouse Design Solutions	0-471-25195-X	Adamson, Venerable	Wiley	Data mart designs and case studies for a broad cross-section of industries.	Y
Managing Projects with Microsoft Project 98 For Windows	0-442-02552-1	Lowery, Ferrara	Van Nostrand Reinhold	Practical advise and direction for using Microsoft Project 98 to manage projects.	N
Building Corporate Portals with XML	0-07-913705-9	Finkelstein, Aiken	McGraw Hill	Leading work on the development of Enterprise Information Portals using XML.	N
Applied XML: A Toolkit for Programmers	0-471-34402-8	Ceponkus, Hoodbhoy	Wiley	Introduction to XML and its use.	Y
Instant ASP Scripts	0-07-135205-8	Buczek	McGraw Hill	An introduction to ASP with a collection of developed scripts.	Y

HTML Drawbacks

- *What's odd about HTML is that while it does a perfectly good job of describing how a web page will look and function and what words it will contain, HTML has no idea at all what any of those words actually mean. The page could be a baby food ad or plans to build an atomic bomb. While HTML knows a great deal about words, it knows nothing at all about information.* **- Robert X. Cringely**

Data, Know Thyself by Robert X. Cringely
http://www.pbs.org/cringely/pulpit/pulpit20010412.html

HTML Drawbacks

1. **No effective way to identify content of page**: HTML tags describe the layout of the page. Web browsers use the tags for presentation purposes, but the actual text content has no specific meaning associated with it. To a browser, text is only a series of words to be presented on a Web page for display purposes.
2. **Problems locating content with search engines**: Because of a lack of meaning associated with the text in a Web page, there is no automatic way that search engines can determine meaning – except by indexing relevant words, or by relying on manual definition of keywords.
3. **Problems accessing databases**: We discussed earlier that Web pages are static. But when a Web form provides access to on-line databases, that data needs to be displayed dynamically on the Web page. Called "Dynamic HTML" (DHTML), this capability enables dynamic content from a database to be incorporated "on the fly" into an appropriate area on the Web page.
4. **Complexity of dynamic programming**: DHTML requires complex programming to incorporate dynamic content into a Web page. This may be written as CGI, Perl, ActiveX, JavaScript, or Java logic, executed in the client, the Web server, the database server, or all three.
5. **Problems interfacing with back-end systems**: This is a common problem that has been with us since the beginning of the Information Age. Systems written in one programming language for a specific hardware platform, operating system, and DBMS may not be able to be migrated to a different environment without significant change or a complete rewrite. Even though it is an open architecture, HTML also is affected by our inability to move these legacy systems to new environments.

What XML Developers Want You To Know about XML

- For centuries, humans have successfully done business by exchanging standardized documents: purchase orders, invoices, manifests, receipts and so on.
- Documents work for commerce because they do not require the parties involved to know about one another's internal procedures.
- Each record exposes exactly what its recipient needs to know and no more.
- The exchange of documents is probably the right way to do business on-line, too.
- But this was not the job for which HTML was built.

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

What XML Developers Want You To Know about XML

- Commerce has fueled Web growth
- Business-to-business commerce is moving as quickly as retail sales
- Flow of goods through the manufacturing process, begs for automation
- "Schemes that rely on complex, direct program-to-program interaction have not worked well in practice, because they depend on a uniformity of processing that does not exist!"

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

What XML Developers Want You To Know about XML

"Librarians figured out a long time ago that the way to find information in a hurry is to look not at the information itself but rather at much smaller, more focused sets of data that guide you to the useful sources: hence the library card catalogue. Such information about information is called metadata."

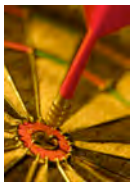
http://dublincore.org/

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

What XML Developers Want You To Know about XML

- XML tags offer no inherent clues about how the information should look on screen or on paper.
- Advantage for publishers
- Write once and publish everywhere
- Distill the substance then pour it into myriad forms, both printed and electronic.
- Tagging content to describe its meaning, independent of the display medium.

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999



XML Aim

- XML rules prescribe a simple architecture
- Clearing away programming details
  - *"People with similar interests can concentrate on the hard part -- **agreeing on how they want to represent the information they commonly exchange**."*
  - *"This is not an easy problem to solve, but it is not a new one, either."*

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

What XML Developers Want You To Know about XML

Just a few XML rules

- These ensure the development of efficient parsers and avoid browser feature bloat.
- Tags almost always come in pairs
  - Like parentheses, they surround the text to which they apply.
- Tag pairs can be nested inside one another to multiple levels
  - Nesting rule ensures computer science tree parsing techniques can be applied
- Unicode-based
  - Allows exchange across national and cultural boundaries

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

What XML Developers Want You To Know about XML

Just a few XML rules

Designers of a new **XML vocabulary**, must agree on three things:

1. which tags will be allowed
2. how tagged elements may nest within one another
3. how they should be processed

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

Java, XML = Similar Excitement

Why the excitement over Java?

- Write it once
- Play it back anywhere

Why the excitement over XML?

- Wrap it once
- Utilize it anywhere
- XML is electronic data interchange (EDI) for the rest of us
- XML is to data what Java is to programming languages

EDI & XML

Much complimentaryness!

EDI has:

- Maturing product market and supportive processes
- Been limited to data exchanges between large organizations
- Required
  - Customized, proprietary software
  - Expensive, private communications networks

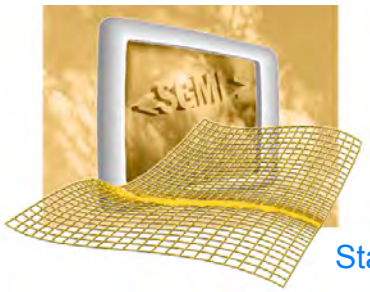
XML

- Requires significantly less customization
- Piggy-backs onto Internet for data communications

EDI isn't quite as readable …

```
ISA*00*          *00*          *ZZ*GATEC       *ZZ*PUBLIC      *960508*...
GS*RQ*GATEC*PUBLIC*960508*1237*000721330*X*003010
ST*840*000721331
BQT*00*F3360196T7174001*960508*106*960509
REF*IL*FM230061280242
PER*IC**EM*F33601@EC099.LLNL.GOV
DTM*002*960517
PO1*1*54*BX***FT*8940*SI*5499*FS*8940011728888*MF*SANDOZ ...
1*MF*SANDOZ NUTRITION*MG*NDE 00212-4580-01
PID*F****SUPPLEMENT, TOLEREX, DIETARY,
CTT*1
SE*16*000721331
GE*1*000721330
IEA*1*000721332
```

Source: DoD



What XML Developers Want You To Know about SGML Foundations

Standardized General Markup Language

- Developed in 1974 by Charles F.Goldfarb and a host of others as a means to create a single basis for any type of markup language
- ISO Standard as of 1986
- This is the foundation on which HTML and XML are based.
- XML is a subset of SGML designed to facilitate the exchange of structure documents over the internet

http://www.oasisopen.org/cover/sgml-xml.html
www.lexum.umontreal.ca/ conf/sgmlottawa/fr/

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

XML versus HTML

- Formats can be specified for any application
- XML compliant browsers can correctly interpret any data sent to it
- XML compliant applications can be used to interpret any type of data

What XML Developers Want You To Know about XML

XML Developers wanted XML to address two fundamental web-based technology problems

1. Internet congestion
   - With HTML updating a quantity field requires transfer to server and reload of page
   - With XML more work can be accomplished on client side
   - "As XML spreads, the Web should become noticeably more responsive

2. Information location difficulties
   - With HTML you cannot search for anything marked as "price"
   - XML-based information is "self describing"
   - Devices will be more capable of local analysis of information

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

What XML Developers Want You To Know about XML

Solution

- Use tags that say what the information is and not what it looks like
- For example, parts of an order for shirt:
  - HTML
    - boldface, paragraph, row, column
  - XML
    - price, size, quality, color

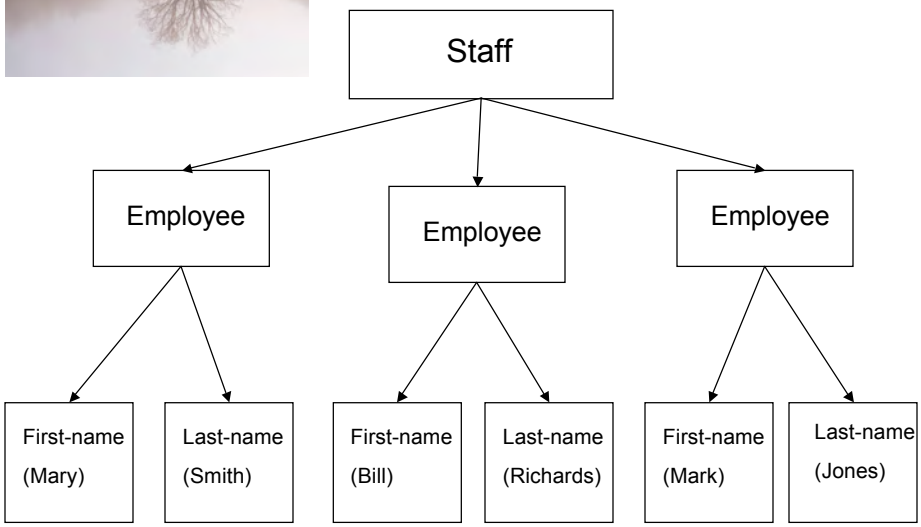
Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

A Sample XML Document

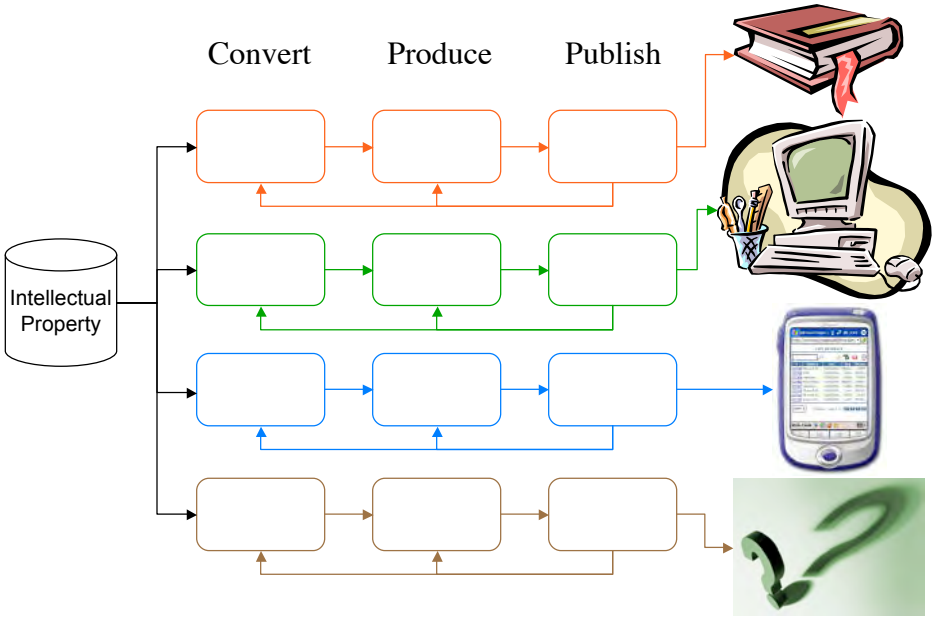
```
<?xml version="1.0"?>
<staff>
  <employee>
    <first-name>Mary</first-name>
    <last-name>Smith</last-name>
  </employee>

  <employee>
    <first-name>Bill</first-name>
    <last-name>Richards</last-name>
  </employee>

  <employee>
    <first-name>Mark</first-name>
    <last-name>Jones</last-name>
  </employee>
</staff>
```

- Looks like HTML
- Easy to Read
- Create your own tags
- Tags give their data meaning
- Composed of elements and attributes

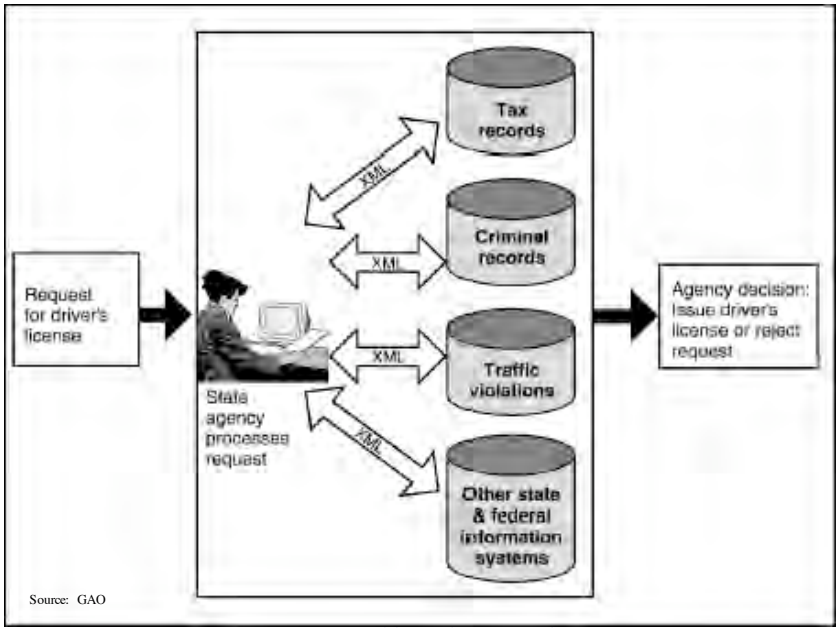
XML Documents as Trees



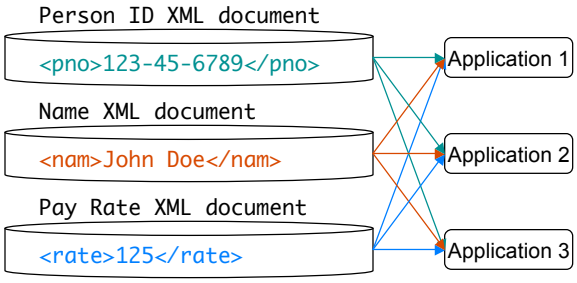
One Ready Made XML Business Case



Hypothetical XML-based Driver's Licensing

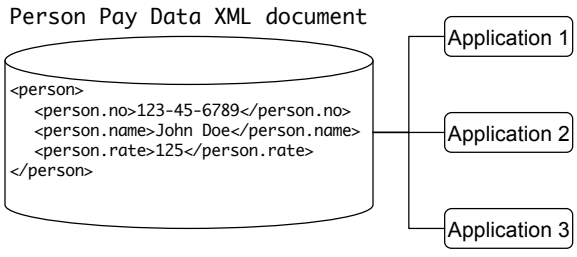


Economies available when wrapping Data Items instead of Data Structures



When wrapping **items** you must:

- Manage three different tags
- Track three different sources and uses of tags
- Create individual structural metadata for each processing application
- Manage tag set availability for each processing application

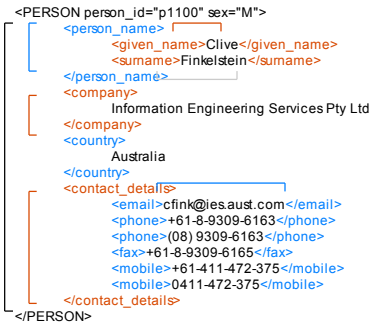


When wrapping **structures** you can:

- Manage data more efficiently by managing just the <person> structure
- Manage just one source and use of metadata
- Maintain one structural definition for all processing applications
- Manage just one set of tags (more like object processing with theme/variations processing options)

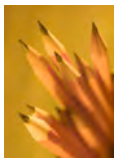
An example of an XML document with metadata tags (surrounded by < … >) identifying the meaning of following data

```
<PERSON person_id="p1100" sex="M">
    <person_name>
        <given_name>Clive</given_name>
        <surname>Finkelstein</surname>
    </person_name>
    <company>
        Information Engineering Services Pty Ltd
    </company>
    <country>Australia</country>
    <contact_details>
        <email>cfink@ies.aust.com</email>
        <phone>+61-8-9309-6163</phone>
        <phone>(08) 9309-6163</phone>
        <fax>+61-8-9309-6165</fax>
        <mobile>+61-411-472-375</mobile>
        <mobile>0411-472-375</mobile>
    </contact_details>
</PERSON>
```



Markup in XML

Markup in HTML



- XML tags are not standardized
- Browsers understand XML tags
- The author defines structure and context of tags/document

```
<TITLES>
   <TITLE CD="N">
      <NAME>Building Corporate Portals Using XML </NAME>
      <ISBN>0-07-913705-9</ISBN>
      <AUTHOR>Finkelstein, C. and P.H. Aiken </AUTHOR>
      <PUBLISHER>McGraw-Hill </PUBLISHER>
      <REVIEW>This book is about portal design, and uses an example
              that is simple in concept, yet meaty in
   production.</REVIEW>
    </TITLE>
</TITLES>
```

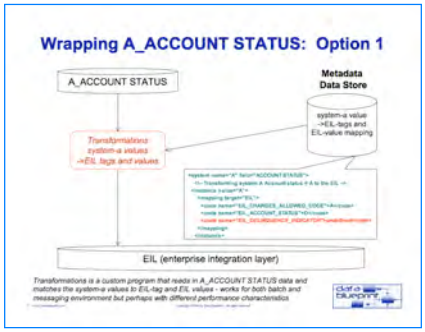
Information Exchange Example

- HTML-based example
  - Search for "stockbroker jobs" will yield lots of ads but few jobs
  - Most are hidden behind classified ads for newspaper Web sites
- XML-based example
  - Newspaper Association of America is building its XML-based markup language for classified ads
  - Ads will be more broadly searchable

Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

XML Processors (aka Parsers)

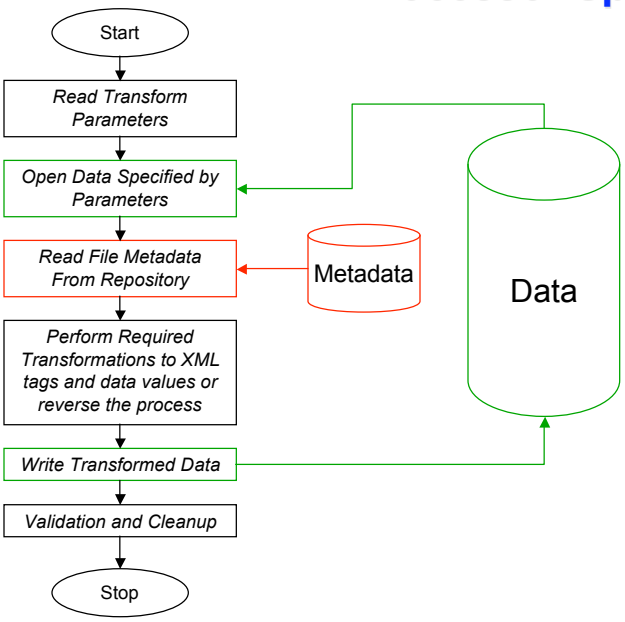
- Terms
  - XML processor
  - XML applications
  - XML vocabularies
  - XML encapsulated
  - XML wrapped



- XML Parser/Processor
  - A software module called an XML processor is used to read XML documents and provide access to their content and structure. It is assumed that an XML processor is doing its work on behalf of another module, that needs access to the data parsed by the parser.

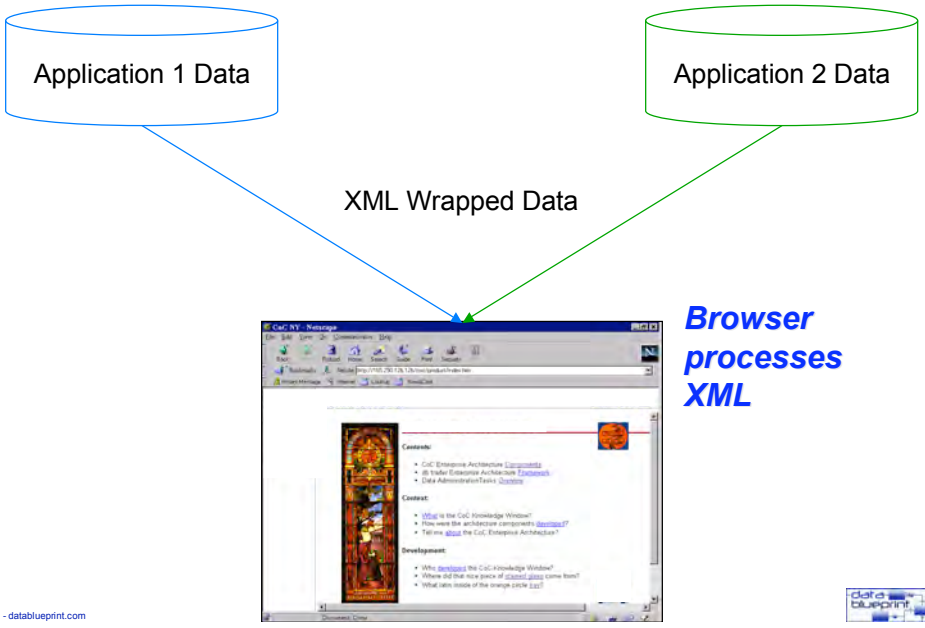
Extracted from: XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job by Jon Bosak and Tim Bray Scientific American May 1999

XML Processor Specifications



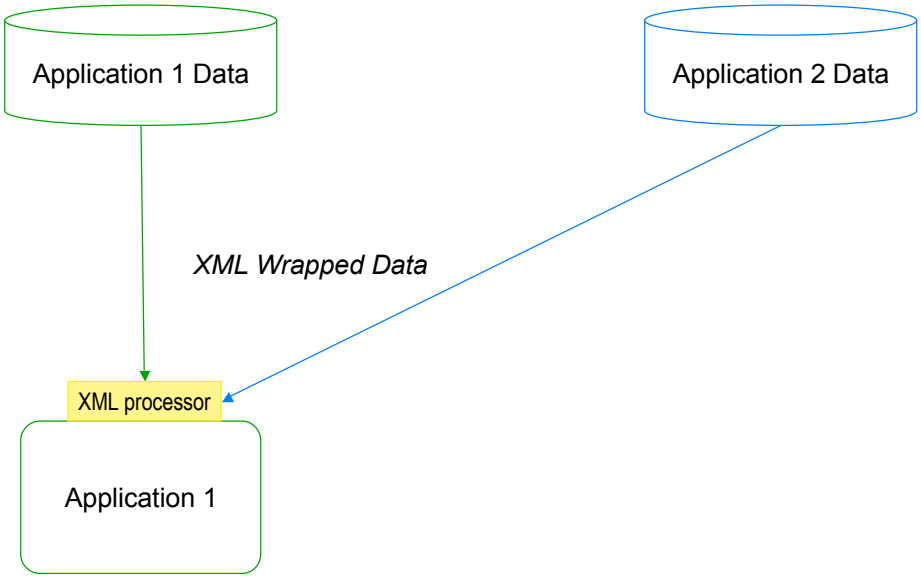
Note: Batch versus message processing results in the different ways of achieving throughput. Batch processing is optimized for large amounts of data throughput. Messaging is optimized for rapid program execution.

XML Integration at the Browser

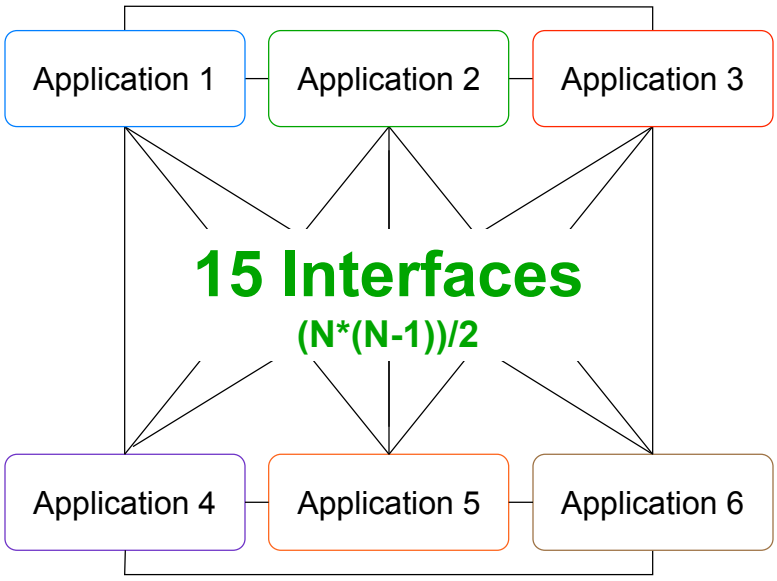


Browser processes XML

XML-based Application Integration

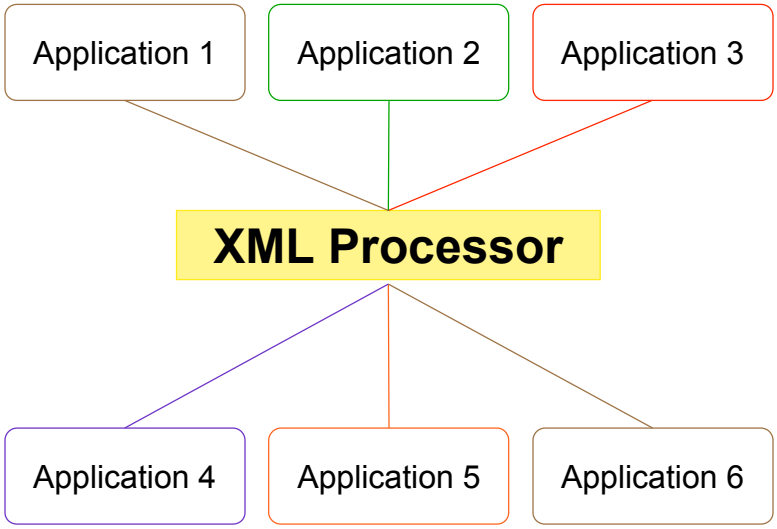


How many interfaces are required to solve this integration problem?

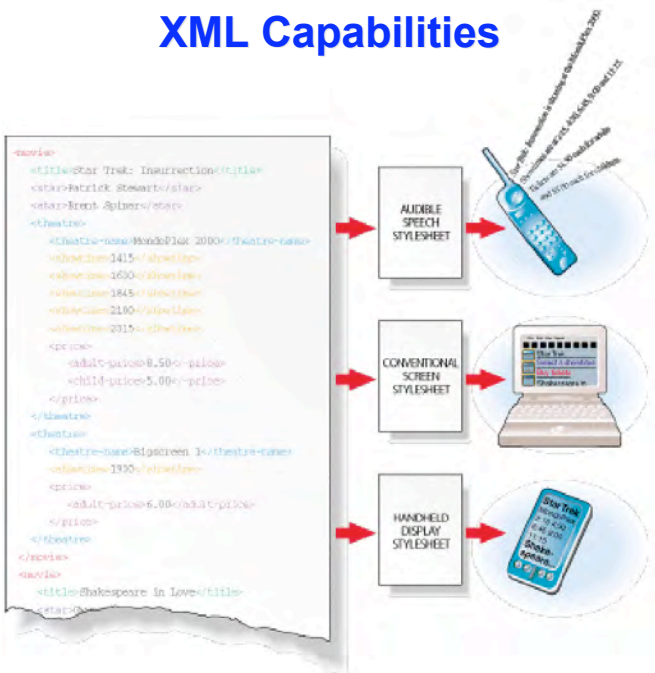


RBC: 200 applications - 4900 batch interfaces

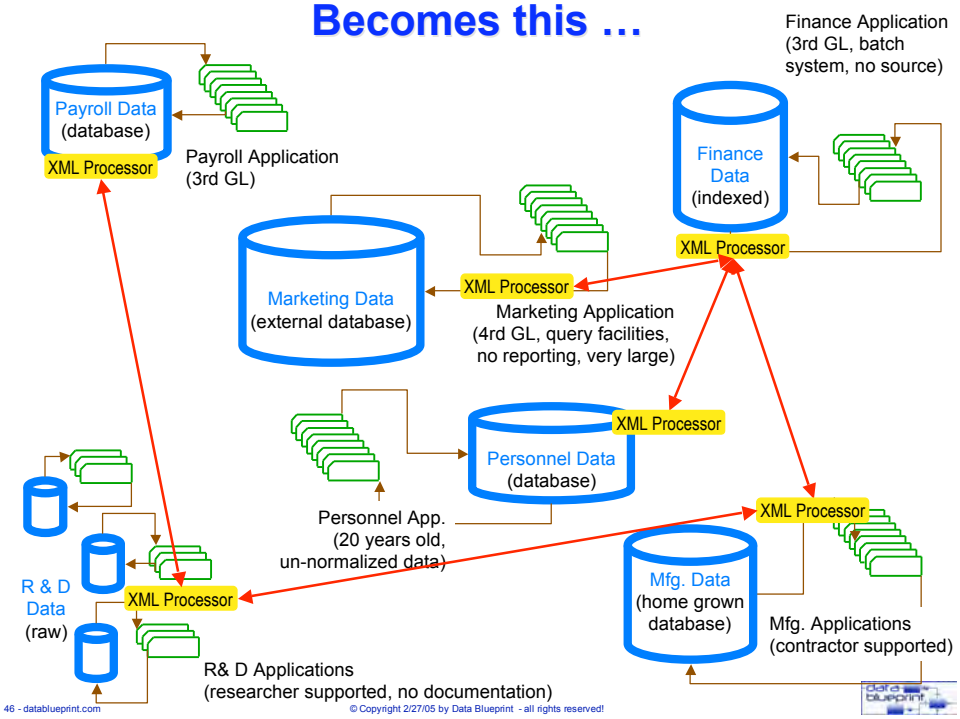
XML-based Integration Solution



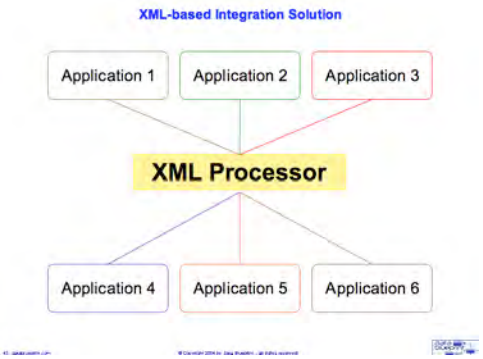
XML Capabilities



Becomes this …



3-Way Scalability



Expand the:

1. Number of data items from each system
   - How many individual data items are tagged?
2. Number of interconnections between the systems and the hub
   - How many systems are connected to the hub?
3. Amount of interconnectability among hub-connected systems
   - How many inter-system data item transformations exist in the rule collection?