XML in Print Production

Deborah Aleyne Lapeyre
Mulberry Technologies, Inc.
17 West Jefferson St.
Suite 207
Rockville, MD 20850
Phone: 301/315-9631
Fax: 301/315-8285
info@mulberrytech.com
http://www.mulberrytech.com

Version 1.0 (January 2006)

XML in Print Production

Administrivia.....1

Remember How Different XML is from Formatted Pages.....1

What XML Looks Like.....2

XML Without Formatting.....4

What We Would Like to See (Print or Screen).....5

XML Versus Print Pages.....6

XML in the Print Production Cycle.....6

Workflow: XML after Page Production.....7

Post-production XML 7

Post-production XML Has Its Advantages.....8

Downsides of Post-production XML 8

Post-production XML Decision Points.....9

How to Make XML after Print.....10

Workflow: XML Introduced in Composition.....11

Workflow: XML Before Composition.....13

Making XML from Word Processors.....14

Advantages of XML Before Composition.....14

XML Validation as an Editorial and Production Tool.....15

Potential Downsides of Pre-Composition XML.....17

Getting from XML to Print.....17

How to Get from XML Tags to Pages.....18

Transform XML Directly into HTML.....18

HTML Tags Imply Formatting.....19

Flow XML into a Proprietary Publishing System.....19

XML-to-Pages: XML-Aware Formatting Software.....20

A Few Representative XML Composition Systems

XML-Aware Composition Engines.....20

XML-Aware Desktop Engines 21

XSL-FO: Format Produced Directly from XML.....22
The *Idea* of XSL-FO.....22
How XSL-FO Formatting Works.....22
Architecture of a Full XSL System.....23
Formatting Objects are XML Elements.....23
Is XSL-FO Ready for Prime Time?.....24
XSL-FO is a Great Report Writer.....25
Flowing XSL-FO into a Composition Engine.....25
In Conclusion: XML Belongs in Print Production.....26
Colophon.....26

XML in Print Production

slide 1

Administrivia

- How this will work
- Questions are always in order
- Why this talk
- Anything else?

slide 2

Remember How

Different

XML is from Formatted Pages

What XML Looks Like

```
<?xml version="1.0"?>
<!DOCTYPE article SYSTEM "../../publishing.dtd">
<article article-type="research-art">
<front>

<journal-meta>
<journal-id>HR-23632987</journal-id>
<abbrev-journal-title>History Revealed
</abbrev-journal-title>
<issn>9704546-3436753</issn>
<publisher-name>Mulberry Press, Ltd.</publisher-name>
<publisher-loc>Rockville, MD</publisher-loc>
</journal-meta>

<article-meta>
<article-id>HR-23632987-8973</article-id>
<subj-group>
<subject>New World discoveries</subject>
<subject>English 16th century history</subject>
<subject>American colonial history</subject>
</subj-group>
<article-title>Raleigh&apos;s Discoveries in
the New World</article-title>
<contrib><name><surname>Gaillard</surname>
<given-names>Tonia Renae</given-names></name>
<degrees>Ph.D.</degrees>
<role>Department Chair</role>
<aff><institut>University of Kentucky</institut>
<addr-line>Lexington, KY</addr-line>
<email>tgaillard&commat;uky.hist.edu</email></aff>
<bio><p>A native of Lexington, Dr. Gaillard completed
her post-graduate studies in Colonial American
History in 1982. Following these studies, she became
an Associate Professor at Murray State University
before joining the University of Kentucky&apos;s
faculty. She has chaired the university&apos;s
Department of History since 1997.</p>
<p>Dr. Gaillard&apos;s extensive writings
include &ldquo;The ...</p>
</bio>
...
```

Print Pages Look More Like This

Raleigh's Discoveries in the New World

Tonia Renae Gaillard
University of Kentucky*

In 1584 Queen Elizabeth I charged Sir Walter Raleigh to discover and colonize lands in the New World on behalf of England This article discusses the efforts and expeditions organized under Raleighís patronage to found a colony in the current day North Carolina Going beyond the chronological history of the doomed Roanoke colony the author provides insights into relations between the colonists and their native counterparts discusses the botanical riches of the region and argues that Raleighís efforts did not end in failure as commonly might be thought

Introduction

On March 25, 1584, Queen Elizabeth I of England charged Sir Walter Raleigh to idiscover, search, find out, and view such remote, heathen and barbarous lands, countries, and territories, not actually possessed of any Christian Prince, nor inhabited by Christian People . . .î [Thorpe, 1997] That same year Raleigh sent two captains, Philip Amades and Arthur Barlowe, from England to Hispaniola and the Canary Islands; from there, the captains were instructed to scout the lands northeast of those already claimed by Spain, to wit, Florida. This land ó now encompassing the Carolinas and Virginia ó was claimed on behalf of England and named iVirginia,î in honor of the iVirgin Queen.î

Securing a Permanent Colony in the Claimed Lands

With land claimed in the New World, an expedition was mounted to establish a settlement. The first expedition failed. Led by Sir Richard Grenville in April 1585, it encompassed 600 men of which 105 remained in the colony while Grenville returned to England for additional provisions. (*See* Appendix 1.) However, when almost a year passed without Grenvilleís return, the remainder of the expeditionary force took advantage of Sir Francis Drakeís arrival to seek return passage to England.[1]

The second expedition, organized by John White in 1587, fared better. It sailed with seven ships filled with Devon families intent upon establishing a colony in that part of Virginia called Roanoke, a word deriving from the speech of native peoples. (*See* Appendix 2.) Two years after founding the iCittie of Raleigh,î houses had been built for almost all families residing in the colony, and the colony had celebrated the birth of its first children born in the New World. The first child, grandchild of John White and child of Ananias and Eleanor Dare, was been named Virginia in honor of the sovereign.

Government of the New Colony

Notwithstanding the initial military nature of the early expeditions, civil government was instituted in the colony; to wit, a subpatent was been granted by Raleigh to several settlers. Under the authority granted by Her Majestyís charter, a council, comprised of a Governor and several assistants, was formed. Those persons named to this council were themselves iplanters,î that is, men who each made investment in the colony and received 500 acres of land in Roanoke upon their arrival there. Each was charged to make laws and ordinances necessary to the success of the colony, except where those laws would circumvent the power of the queen and

History Revealed Volume 42 no 1

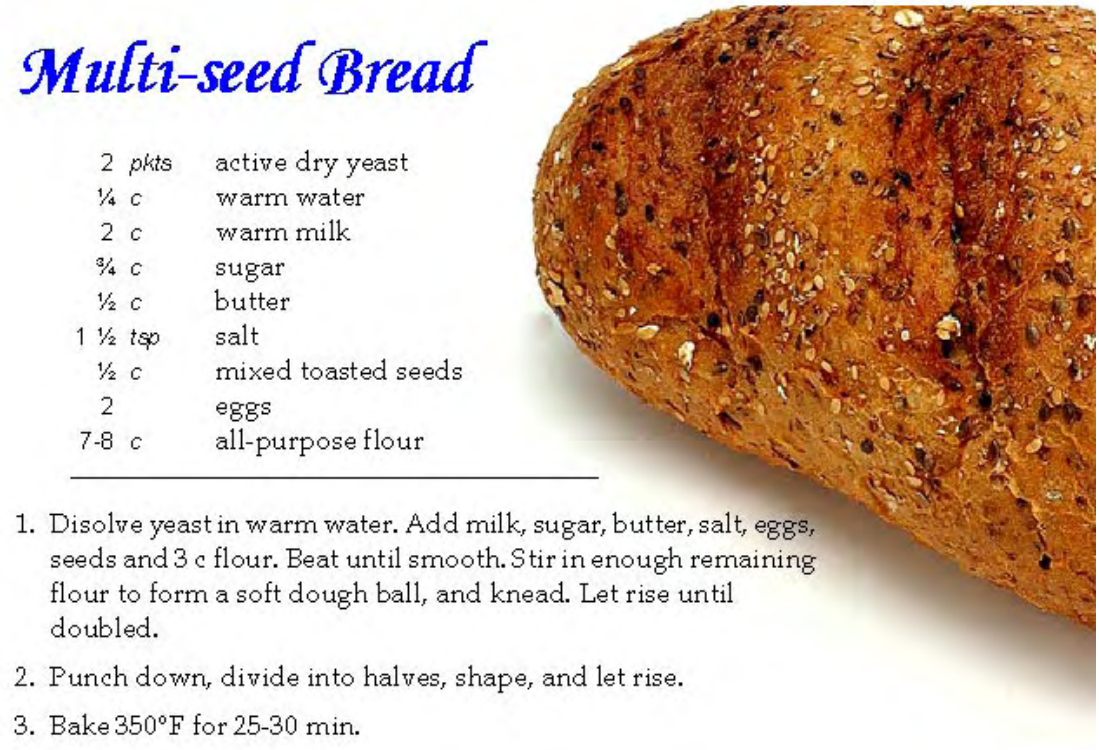
XML Without Formatting

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE collection SYSTEM "CollectionDTD/collection.dtd">
<collection>
<title>Recipe Collection: Breads and Soups</title>

<recipe>

<title>Multi-seed Bread</title>

<class name="type of dish">yeast bread</class>

<component>
<ingredients>
<ingredient>
  <quantity>2</quantity>
  <measure>pkts</measure>
  <foodstuff>active dry yeast</foodstuff></ingredient>
<ingredient>
  <quantity>&frac14;</quantity>
  <measure>c</measure>
  <foodstuff>warm water</foodstuff></ingredient>
<ingredient>
  <quantity>2</quantity>
  <measure>c</measure>
  <foodstuff>warm milk</foodstuff></ingredient>
<ingredient>
  <quantity>&frac34;</quantity>
  <measure>c</measure>
  <foodstuff>sugar</foodstuff></ingredient>
<ingredient>
  <quantity>&frac12;</quantity>
  <measure>c</measure>
  <foodstuff>butter</foodstuff></ingredient>
<ingredient>
  <quantity>1 &frac12;</quantity>
  <measure>tsp</measure>
  <foodstuff>salt</foodstuff></ingredient>
<ingredient>
  <quantity>&frac12;</quantity>
  <measure>c</measure>
  <foodstuff>mixed toasted seeds</foodstuff></ingredient>
<ingredient>
  <quantity>2</quantity>
  <foodstuff>eggs</foodstuff></ingredient>
<ingredient>
  <quantity>7-8</quantity>
  <measure>c</measure>
  <foodstuff>all-purpose flour</foodstuff></ingredient>
</ingredients>

<directions>
<step><p>Disolve yeast in warm water. Add milk, sugar, butter,
salt, eggs, seeds and 3 c flour. Beat until smooth. Stir in
enough remaining flour to form a soft dough ball, and knead.
Let rise until doubled.</p></step>
```

```
<step><p>Punch down, divide into halves, shape, and
let rise.</p></step>

<step><p>Bake 350&deg;F for 25-30 min.</p></step>
</directions>
</component>

<source>Becky</source>

<yield>2 loaves</yield>

<illustration filename="Graphics/long-loaf.jpg"/>

</recipe>
```

slide 6

What We Would Like to See (Print or Screen)



slide 7

XML Versus Print Pages

- XML separates content from format
- XML does not say
  - how it looks (16 point Helvetica Bold)
  - what it does (starts a javascript)
- Page production specifies
  - sizes, fonts, leading
  - color, style
  - all the physical aspects

So they aren't addressing the same issues!

slide 8

XML in the Print Production Cycle

Most production using XML is a variation on one of these

- XML after page production (maybe long after)
- XML as part of composition
- XML before composition

slide 9

Workflow: XML after Page Production

- Manuscript from authors
- All editorial correction and proofing cycles
- Pages made (correction cycles)

  -
- After pages delivered, XML is made
  - by compositor
  - by conversion house
  - by publisher

slide 10

Post-production XML

- Is the way
  - the 10-year backlog becomes XML
  - libraries and repositories get XML
  - companies without XML capabilities get XML
- Is the “safest” XML
  - the content is frozen and will not change

slide 11

Post-production XML Has Its Advantages

- Current processes need not change
- It's in somebody else's ballpark
- Production deadlines have already been met
- Immediate costs go down
- Can add rich subject tagging/indexing not possible on tight production schedule

slide 12

Downsides of Post-production XML

- Total costs go up
  - print and XML are separate, sequential flows
  - sequential flows takes more time/effort
- Web waits for print
- XML and print must *both* be checked
- There are no XML editorial or production advantages
- Without subject expertise, tagging quality may be low
- May be problems: tables, special characters, equations

Repeat: XML and print must both be checked

slide 13

Post-production XML Decision Points

- Choose tagging style
  - obvious structure and format tags or
  - "rich" semantic tagging
- Preserve or discard content that could be generated?
- Preserve or unify duplicate content?
- Fix errors or preserve original?
- *Whose job* is it to check the XML?

(A file that makes perfect pages *may not* be a clean electronic file.)

slide 14

When XML after Print Makes Good Sense

- Pages are already made (e.g., 3 years ago)
- Design does not flow from content
  - each page individually designed and laid out (1st grade math text book)
  - design is art
- The current production process *cannot be disturbed*
- XML is for the long-term only

slide 15

How to Make XML after Print

(Getting XML out of Publishing Packages such as Quark and InDesign)

Typical scenario:

- Require consistent use of styles
  (then clean up to enforce it)
- Map paragraph and character styles to XML tags
- Make a second pass (programs) to add structure
  (formatter styles are flat, not hierarchical)
- Make a third pass (people) to add rich semantic tagging

slide 16

The Areas of Greatest Difficulty

- Tables (especially for older QuarkXPress)
- Special characters
- Equations
- Multiple small stories in one document
- Mapping same tag in different contexts to different styles

slide 17

Aside: Getting XML out of QuarkXPress is a Software Industry

- QuarkXPress internal
  - XML Plus in 6.5 and up
  - `avenue.quark`
  - the ability to import XTags
- PCI's ICPS claims to be the oldest and the best
- Easy Press sells ATOMIC (good at tables)
- North Atlantic Publishing sells a GUI version
- Apropos' Roustabout does it in batch using XML control files
- Gluon (holds Quark together) has one a good one, too
- There are many more!

(Most of these tools also support XML *import* into QuarkXPress)

slide 18

Workflow: XML Introduced in Composition

(Let the compositor do it)

- Manuscript or word-processor from authors
- First editorial correction cycle and proofing
- Manuscript or word-processor to compositor

- Compositor converts to XML (maybe their XML, maybe yours)
- Pages made from XML
- Compositor returns XML and page proofs

slide 19

Advantages of XML Introduced in Composition

- Simultaneous web and print
- Current editorial processes not changed (much)
- Tagging and composition overlap (savings and knowledge)

(Many compositors have chosen XML for their own benefit!)

slide 20

Potential Downsides of XML Introduced in Composition

- Some XML literacy required for staff
- XML and print must still *both* be checked
- Production rhythm changes
  - tagging takes time
  - production is then faster

slide 21

Workflow: XML Before Composition

(An XML publishing application)

- Manuscript or XML from authors
- Converted to clean XML (ideal)
- Editorial/Production done in XML (or XML is added here)
  - content editing
  - peer review
  - QA and validity
  - WYSIROTS proofs (return of the galley!)
  - Live links for references
- Final page adjustments postponed until all else is done
- XML transformations used to produce
  - pages (or rough starter pages)
  - HTML for the web
  - RSS feeds
  - content for aggregators
  - new products

slide 22

Implications of XML before Composition

In most XML production

- Print pages are one of *many* output products
- Concentration is on content (the XML)
  - not appearance
  - not particular output product
- Print, web, other products each *separately designed*
- XML is used to make the output products

slide 23

Making XML from Word Processors

Few authors work in XML

- Give authors an HTML form and make XML behind the scenes
- Give them word-processing "templates" and make XML behind the scenes
- Translate *clean* word-processing styles to XML tags
  (or clean first, then transform)
- Third-party products where authors *think* they are in a word-processor,
  but behind the scenes it's XML
- Microsoft Word 2003 and beyond
  (edit awkwardly in XML, export XML)

slide 24

Advantages of XML Before Composition

- One source for all: web, print, CD, database, etc.
  - simultaneous production
    (no product X waiting for product Y)
  - no parallel maintenance
- Almost-instant galleys or editors' proofs
- Fewer communication steps
- Generated text
  - the numbers in a numbered list (1., 2., 3.)
  - the enumerator on a footnote
  - Chapter 1.
  - Figure 3.6
  - (See Figure 3.4: All Cars Eat Gas)

slide 25

XML Validation as an Editorial and Production Tool

- No structural surprises or “creative” styling
- Bibliographic references checked and made live before publication
- Checklists (terms, graphics, surnames)
- New proofing and QA techniques (false-color)

Quality may increase *a lot*

Downstream turnaround may get much much faster

slide 26

New Proofing/QA Possibilities

More intelligent proofing than “Does it look right?”

- List any element/combination
- Print content of reference next to place where reference is made
- Some automated link checking
- False color proofs
  - add color to make things stand out
  - add numbering
  - add links to what needs to be checked

False Color Proof

Water is blue (italic), land is yellow (bold), and “features” are purpley (display font in the print)

The rivers of the state flow in general from NW. to S.E., across the **Blue Ridge**, the **Piedmont** and the **Coastal Plain**, following courses which were established before erosion had produced much of the present topography. But in the **Newer Appalachians** the streams more often follow the trend of the structure until they empty into one of the larger, transverse streams. Thus the *Shenandoah* flows N.E. to the *Potomac*, the *Hoiston* S.W. toward the *Tennessee*. A part of this same province, in the S.W. part of the state, is drained by the *New* river, which flows N.W. across the ridges to the *Kanawha* and *Ohio* rivers in the **Appalachian Plateau**. In the limestone regions caverns and natural bridges occur, among which LURAY CAVERN and the NATURAL BRIDGE are well known. The drowned lower courses of the SE. flowing streams are navigable, and afford many excellent harbours. *Chesapeake Bay* covers much land that might otherwise be agriculturally valuable, but repays this loss, in part at least, by its excellent fisheries, including those for oysters. In the S.E., where the low, flat **Coastal Plain** is poorly drained, is the *Great Dismal Swamp*, a fresh-water marsh covering 700 sq. m., in the midst of which is *Lake Drummond*, 2 m. or more in diameter. Along the shores of *Chesapeake Bay* and the *Atlantic Ocean* are low, sandy beaches, often enclosing lagoons or salt marshes. ¶

slide 28

More Pre-composition Advantages

- Total costs go down
  (although immediate costs may go up)
- The biggest benefits may be long term
  - increased opportunity
  - building corporate resources
  - ability to create new products more easily

(The expense and hard work are immediate)

slide 29

Potential Downsides of Pre-Composition XML

- XML literacy absolutely required for staff
- Accurate word-processor “styles” may become critical
- Author and editor resistance to WYSIROTS
- If everything flows from the XML,
  all changes need to be in the XML
- Additional author-proofing required for coded math and chemistry

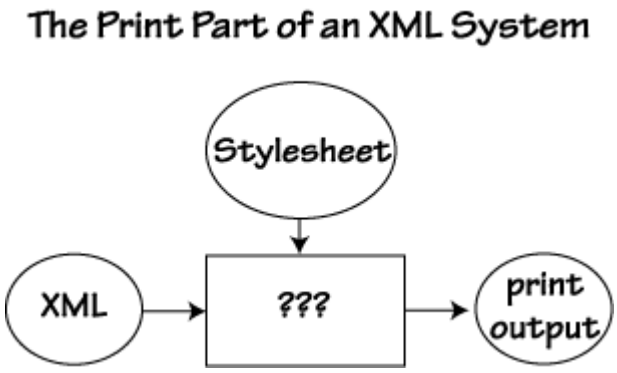
slide 30

Getting from XML to Print

- Biggest challenge/opportunity of pre-composition XML
- Now you need to **make print from the XML**

slide 31

Getting from XML to Print



slide 32

How to Get from XML Tags to Pages

Make XML into HTML, then print the HTML

Flow XML into a non-XML composition system

Use an “XML-aware” desktop publishing system

Use an “XML-aware” composition system

Use XSL-FO stylesheet with an XSL-FO engine

slide 33

XML-to-Pages: Transform XML Directly into HTML

- Use HTML tags as is or add CSS styling
- Transform XML elements to HTML elements (XSLT)

```
<report> into <html>
<report><title> into <h1>
<section><title> into <h2>
```

- Lot of XML is converted this way now!

slide 34

HTML Tags Imply Formatting

- Browser makes a tag look like something
  - `<h1>` is a first level heading
  - `<h2>` is a second level heading
  - `<bold>` and `<strong>` are bigger and darker
  - CSS may be added
- You only have as much control as your browser gives you
- Print from the browser
- Import to a word-processor and print

slide 35

XML-to-Pages:
Flow XML into a Proprietary Publishing System

Common ways to accomplish:

- Text and graphics from XML file
  - poured into publishing system
  - then styles added
- XML tags transformed into proprietary codes (styles) that accomplish layout/typography
- Publishing package imports XML directly
  - maps tags to styles

slide 36

XML-to-Pages: XML-Aware Formatting Software

Two Main Ways XML-Aware Typography Can Work

- Work natively in XML
  Proprietary codes hidden (in XML processing instructions)
- Alternatively, import/export XML
  - take in XML files
  - translate them to internal format
  - proprietary codes accomplish typography
  - then export XML files

A Few Representative XML Composition Systems

slide 37

Representative XML-Aware Composition Engines (alpha order)

- 3B2 (Arbortext)
- DL Pager/DL Composer (DataLogics)
- Genera and Oasys (Miles33)
- Publisher (Penta — Version 3.0 and above do XML also*)
- TopLeaf XML Composition (Metaformix)
- XPP (XML Professional Publisher from XyEnterprise)

(*Penta is not dead; it has gone to its geeks)

slide 38

Representative XML-Aware Desktop Engines (alpha order)

- DeskTopPro (Penta)
- Documorph (Xenos Group)
- Epic (Arbortext — Epic 4.2 and above support XSL-FO)
- FrameMaker 7.0 (Adobe)
- Life*TYPE (Corena UK Ltd)
- PowerPublisher using UltraXML (WebX Systems — supports XSL-FO)

slide 39

These Composition Systems Use XML

But XML Does Not Control

- But the composition system (not the XML) controls
  - Page layout and text flow into layout objects
  - Hyphenation and Justification
  - Column balancing
  - Selective spreading of vertical whitespace at justification points
- The composition system does this too (but XML metadata may advise)
  - Floats and keeps
  - Windows and orphans (but metadata may advise)
  - Real color (in all its meanings)
    although tagging may suggest
    (such as `color="BK#0000"` in HTML)
  - Illuminated initial caps

slide 40

XML-to-Pages:
XSL-FO: Format Produced Directly from XML

Transform XML into XSL Formatting Objects

- Called XSL-FO (Extensible Stylesheet Language Formatting Objects)
- Goes directly from XML to pages (postscript, PDF, etc.)
- XSL-FO supports both online display and printing

slide 41

The *Idea* of XSL-FO

- Definition of layout and styles
  - platform-independent
  - vendor-independent
- Automated formatting from structured data

The Dream: high quality, high volume, content-driven publishing
(*Not* for layout-driven publishing!)

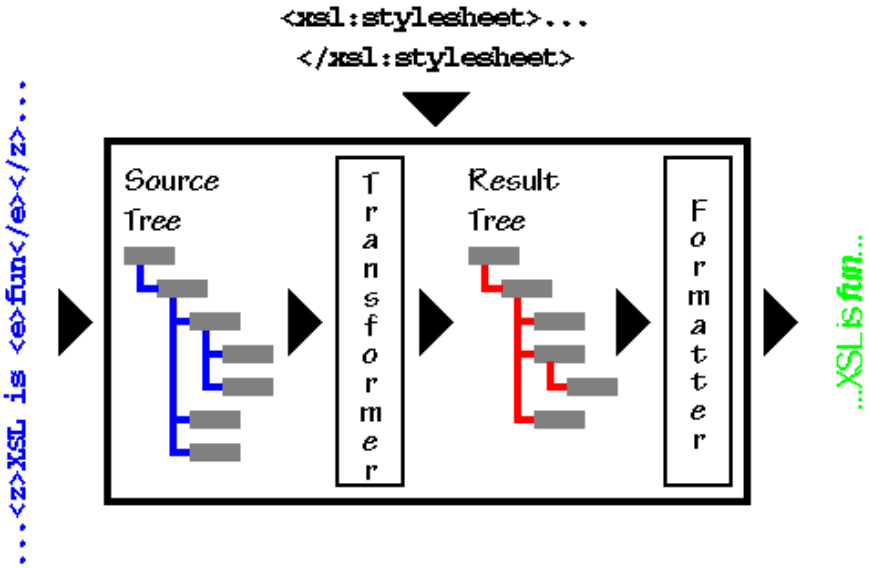
slide 42

How XSL-FO Formatting Works

- XSL-FO provides a tag set into which XML documents may be transformed (using XSLT)
- The XSL-FO tags describe
  - the layout geometry of the page (into which you pour content)
  - a set of *formatting objects*
    - that say how to put content on the page
    - that describe how the document should be rendered
- An XSL-FO *rendering engine* makes pages/display from these tags
- An XSL-FO document is
  - an XML document
  - with text and graphic content wrapped in formatting object tags

slide 43

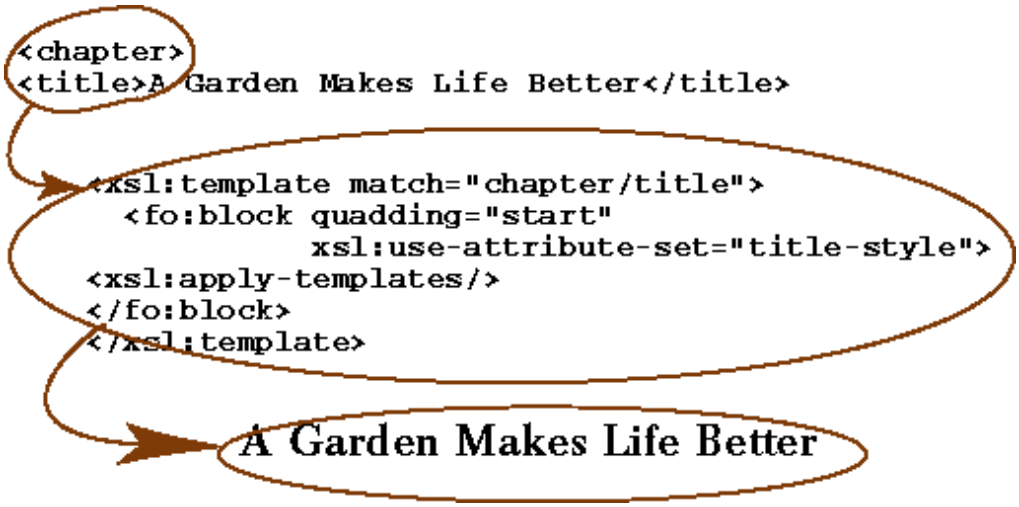
Architecture of a Full XSL System



slide 44

Formatting Objects are XML Elements

that take "properties"



slide 45

Is XSL-FO Ready for Prime Time?

Depends on what "prime time" is

By design, XSL-FO is desk-top publishing, *not fine typography*

- Page layout, column balancing, vertical justification "basic" only
- Keeps, floats, widows, and orphans a little weak
- No CMYK and only modest pantones

slide 46

What the Commercial XSL-FO Vendors Say

- XSL-FO is cost-effective in situations where typography is too expensive
- XSL-FO vendors claim
  - 2002 — XSL-FO could probably meet 30-40% of publishing needs
  - 2006 — 70-85% (they might claim higher)
  - "We can do *most* of traditional high-end composition"
  - Their clients have "*modified their [formatting] requirements* to enable them to live with the [FO] limitations"

slide 47

XSL-FO is a Great Report Writer

(where pagination is not a problem)

- Credit card and bank statements
- Investment portfolios
- Hospital records and patient medical records
- Insurance policies and claims
- State legislatures for bills, resolutions, and reports
- Directory and catalog products

(Anybody where lights out works is a good candidate)

slide 48

Flowing XSL-FO into a Composition Engine

(Maybe best of both worlds)

- Run XSL-FO stylesheet to make formatting objects
- Open results in a composition engine
- If content change, change XML and reflow
- Make non-content tweaks inside the composition system
  - graphics placement
  - column balancing
  - floats, keeps, widow, orphans
- Several engines allow this (XyEnterprise, Arbortext, etc.)

slide 49

In Conclusion: XML Belongs in Print Production

- Real advantages to content creators
  - cost saving (in some situations)
  - time saving (almost always)
  - quality improvement (content gets checked)
- Content creators want XML *and* print
  - repurpose and multi-use
  - customization and internationalization
  - long-term archive
- You can make
  - Good pages from XML
  - XML from pages

slide 50

Colophon

- Slides and handouts created from single XML source
- Slides projected from HTML which was created from XML using XSLT
- Handouts created from XML
  - source XML transformed to Open Office XML
  - Open Office XML opened in Open Office
  - pagination normally adjusted
  - Saved as PDF
- Slideshow materials available at `http://www.mulberrytech.com/slideshow`