

An XML Document Standards Interoperability Framework

DITA East 2007

Raleigh, NC
October 2007

Jim Earley
Scott Hudson

Jabberwocky

When I use a word... it means just what I
choose it to mean – neither more nor less

- Humpty Dumpty, "Through The Looking Glass", Lewis Carrol

You Talk Funny



If you want to talk like us, just open your mouth and say "ah," as if you're at the *doctah*. As in: "Nomah hit a homah!" We save the Rs for words ending in A, like *Chiner*. It sounds *b'zah*, but remember, we're not the ones with the accent. We've been here since 1630. John Winthrop dropped the R into the Hahbah one day on his way to the State House and we didn't find it until the Big Dig.

- John Powers, Boston Globe 07/25/2004

Bubbla = Water Fountain
Tonic = Soda / Pop
Grinder = (Hot) Sub
Side-by-each = Next to each other
Book it = Hurry
Packie = Liquor Store
Directional = Turn Signal
Soft = Gutsy/Brave
So Don't I = Me Too
Frappe = Milkshake
Rotary = Roundabout
Jimmies = Chocolate Sprinkles
Route 128 = I-95 (or I-93)

XML Document Variants

- As XML has become more pervasive, the number of XML documents standards have grown
- Early on
 - DocBook
 - Home-grown content models
- The Last 5 Years
 - DITA
 - ODF (OpenOffice Document Format)
 - Microsoft Office Open XML

A Brief History

- Desktop Publishing (ca. 1980s)
 - Incompatible formats
 - Conversion Nightmares
 - One-Offs
- SGML (ca. 1990)
 - Promises to mitigate conversion problems
 - “The best-laid schemes o' mice an' men ...”
 - Lack of tools
 - Too expensive but for a few large corporations

A Brief History, Continued

- HTML (ca. mid-1990s)
 - The World Wide Web and HTML Enable new ways of sharing content
 - Core publishing components (Headings, Tables, Lists, Images, etc.)
 - Problem: HTML is formatting, no semantics; browser incompatibilities

- XML (Today)
 - “SGML for the Web”
 - Living up to the promise
 - Rich, diverse toolsets
 - Inexpensive, even free

Looking Into the Horizon

- Stronger Emphasis On Collaboration
 - OEMs and Partnerships
 - Intra-Organizational
 - Mergers/Acquisitions
 - Syndication
 - Distributed Authoring
- Work Smarter
 - Leverage Content From Multiple Sources
 - “Take and Bake”
 - Mitigate Conversion Costs
 - Expectation of “conformant content”

In Search of the Holy Grail

- How do I reconcile/leverage/reuse content to and from different XML document standards?

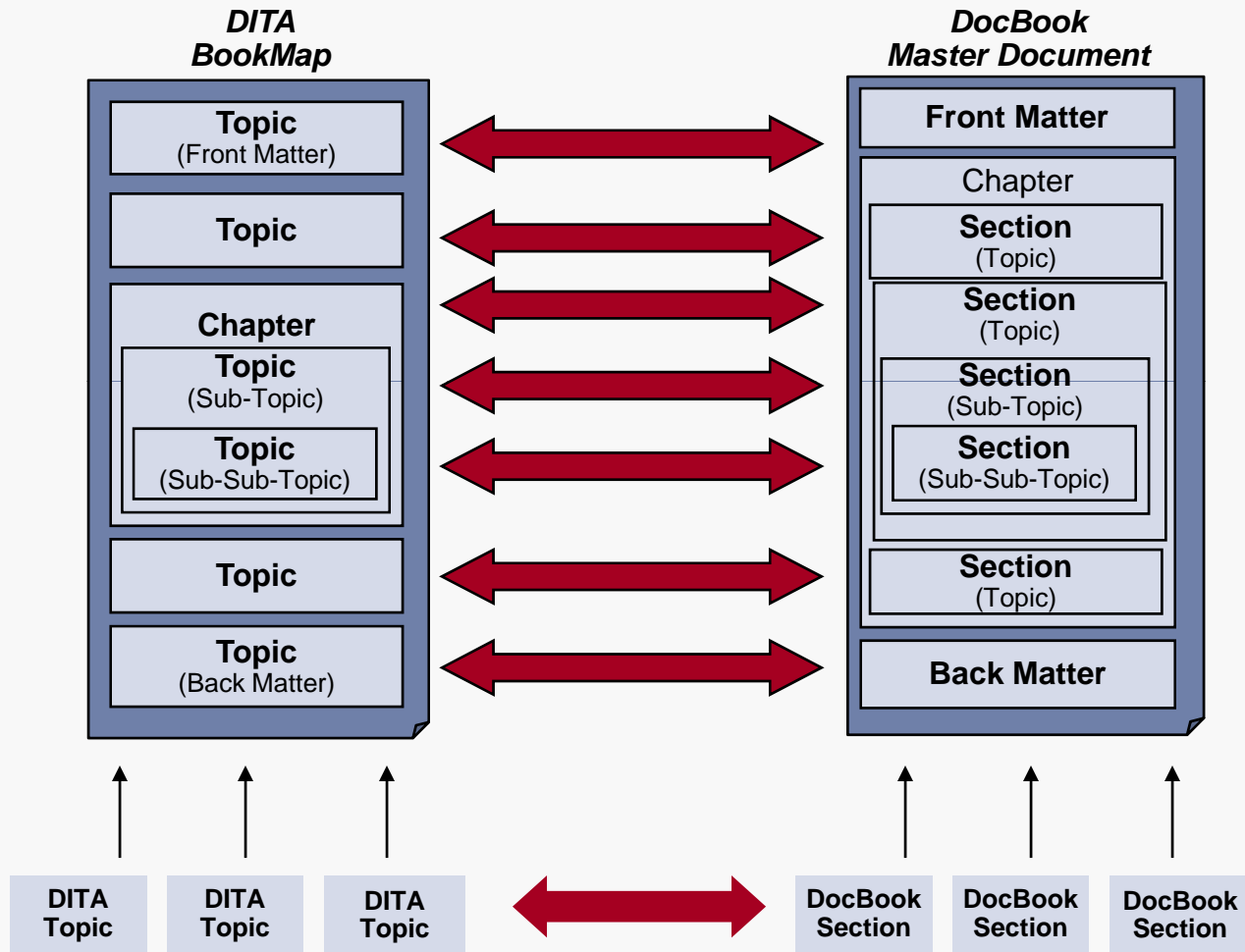


Interoperability: Current State

- Because DocBook, DITA and other standards will co-exist, these standards need to be interoperable, but they're not... Yet.
- A common set of element definitions and models as a base for each standard will require much more collaboration between standards.
- Standards implement different methods to extend their respective standard:
 - Specializations in DITA
 - Customization layers in DocBook

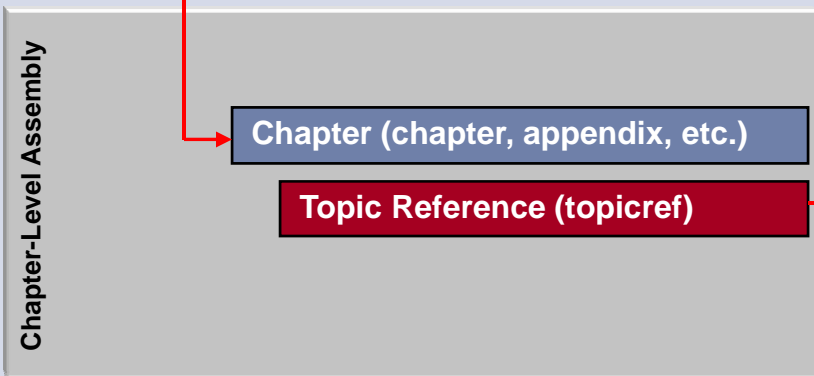
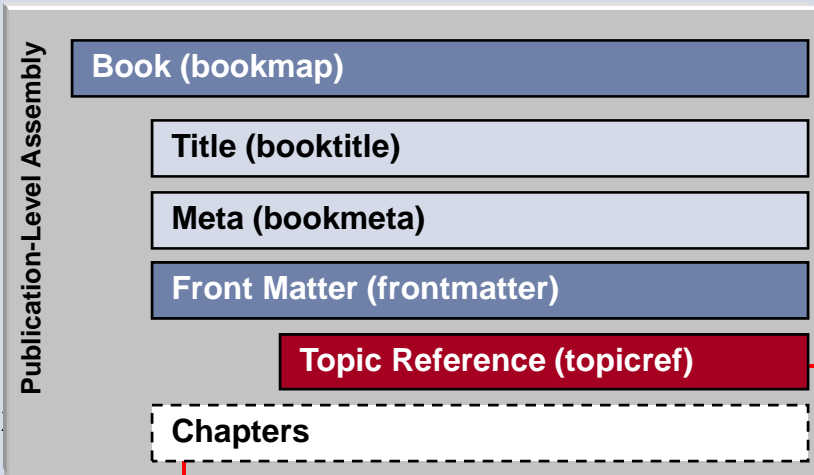
Finding Common Ground

Insight: Parallel View of DITA and DocBook

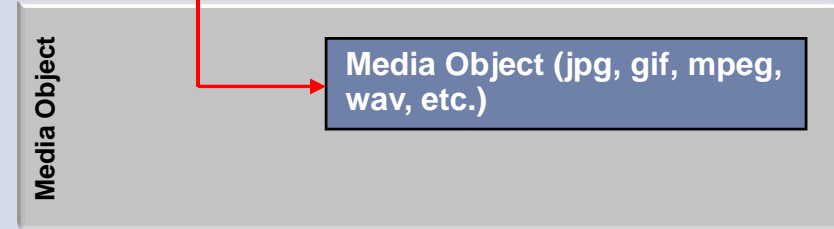
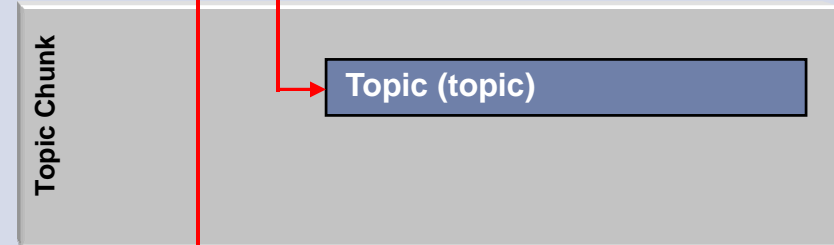
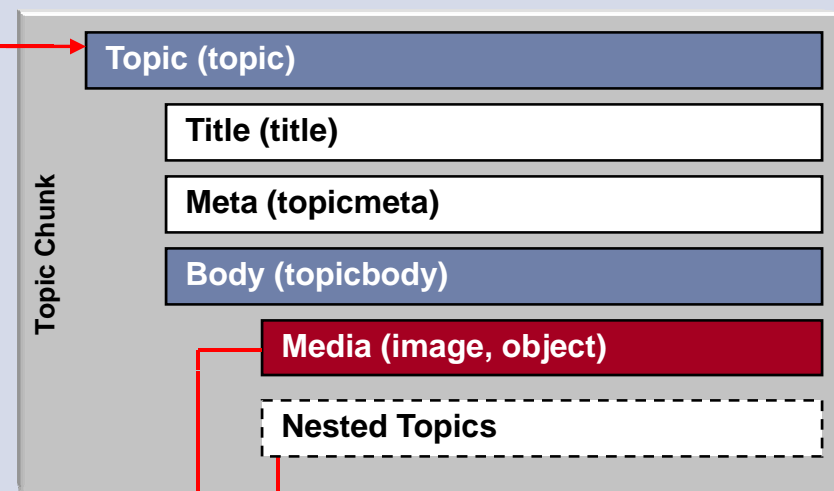


A Closer Look At DITA Structures

Publications



Topics



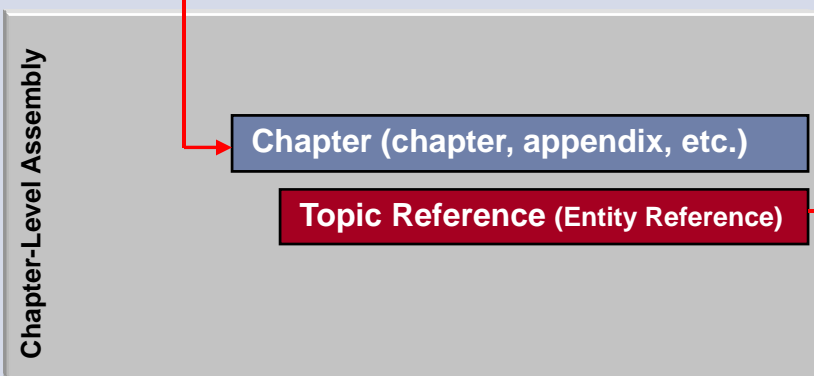
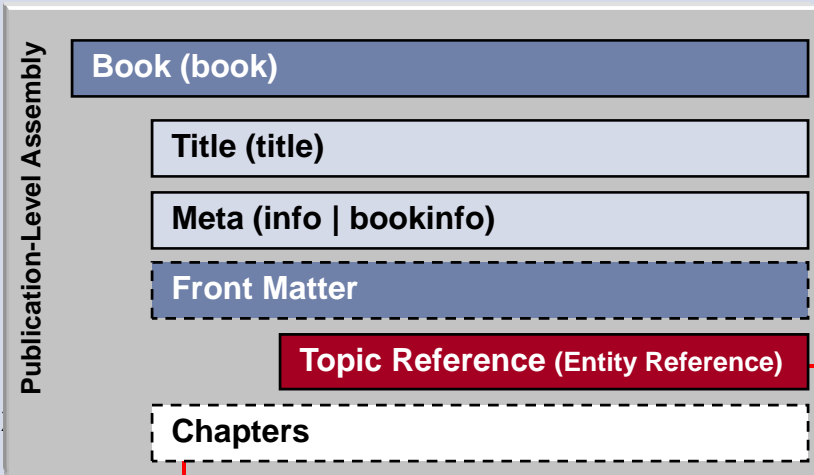
topicref (href)

conref

href

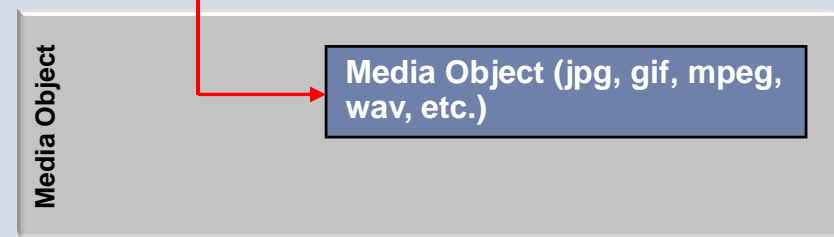
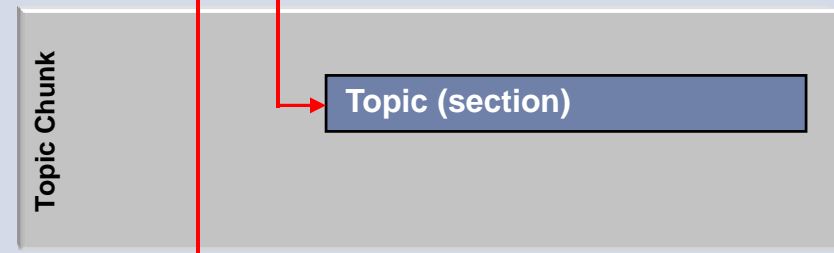
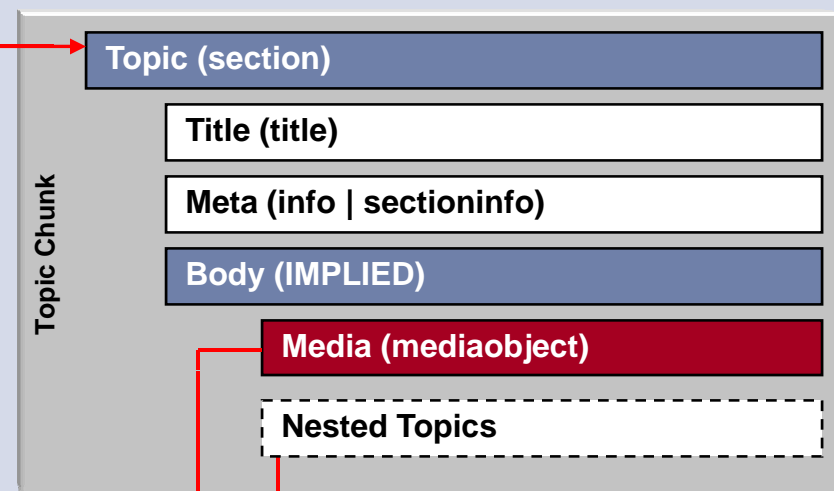
A Parallel Look At DocBook Structures

Publications



Entity Reference or XInclude

Topics



Entity Reference or XInclude

href

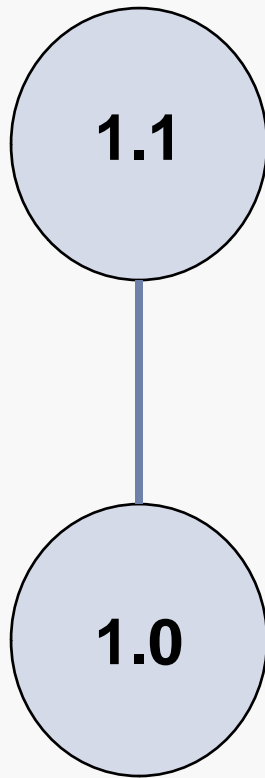
Interoperability Strategies

Strategies for Interoperability

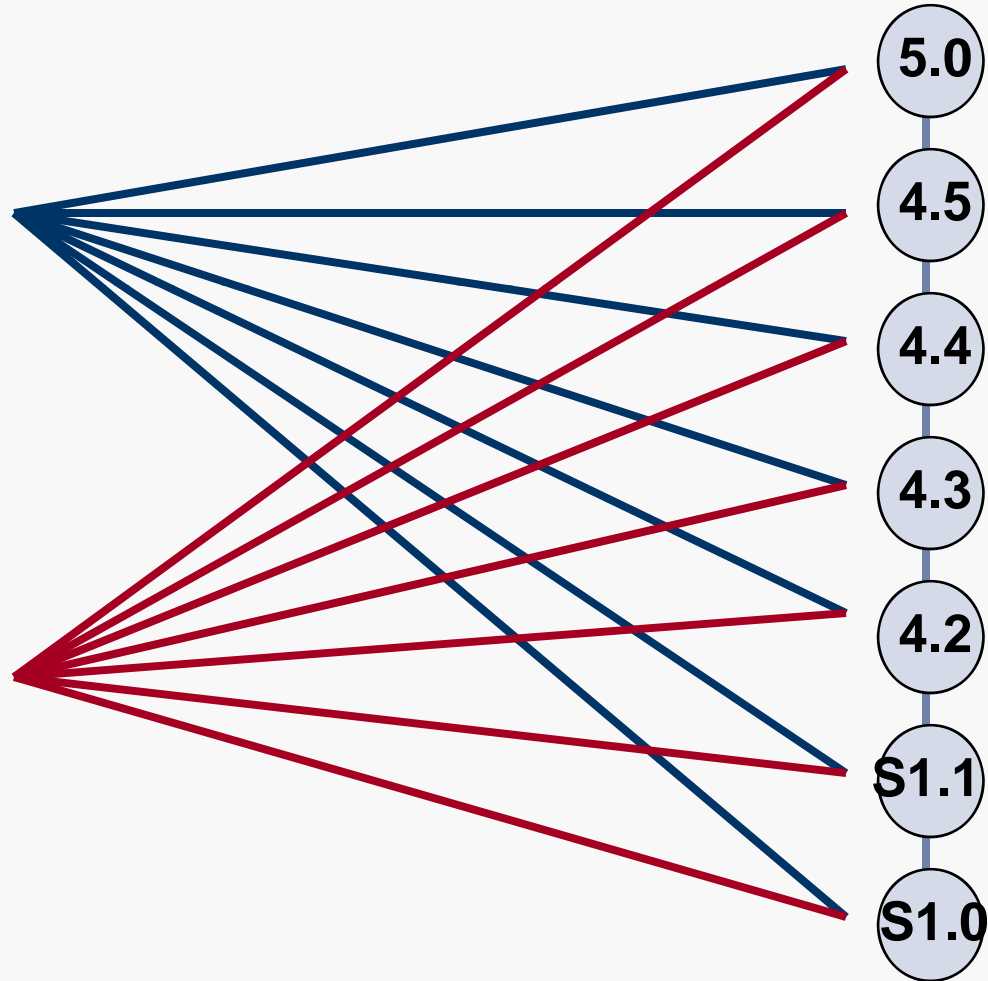
- Content Interoperability – “Shoehorning”
- Processing Interoperability – “Uni-directional”
- Roundtrip Interoperability – “Bi-Directional”

Version Chaos

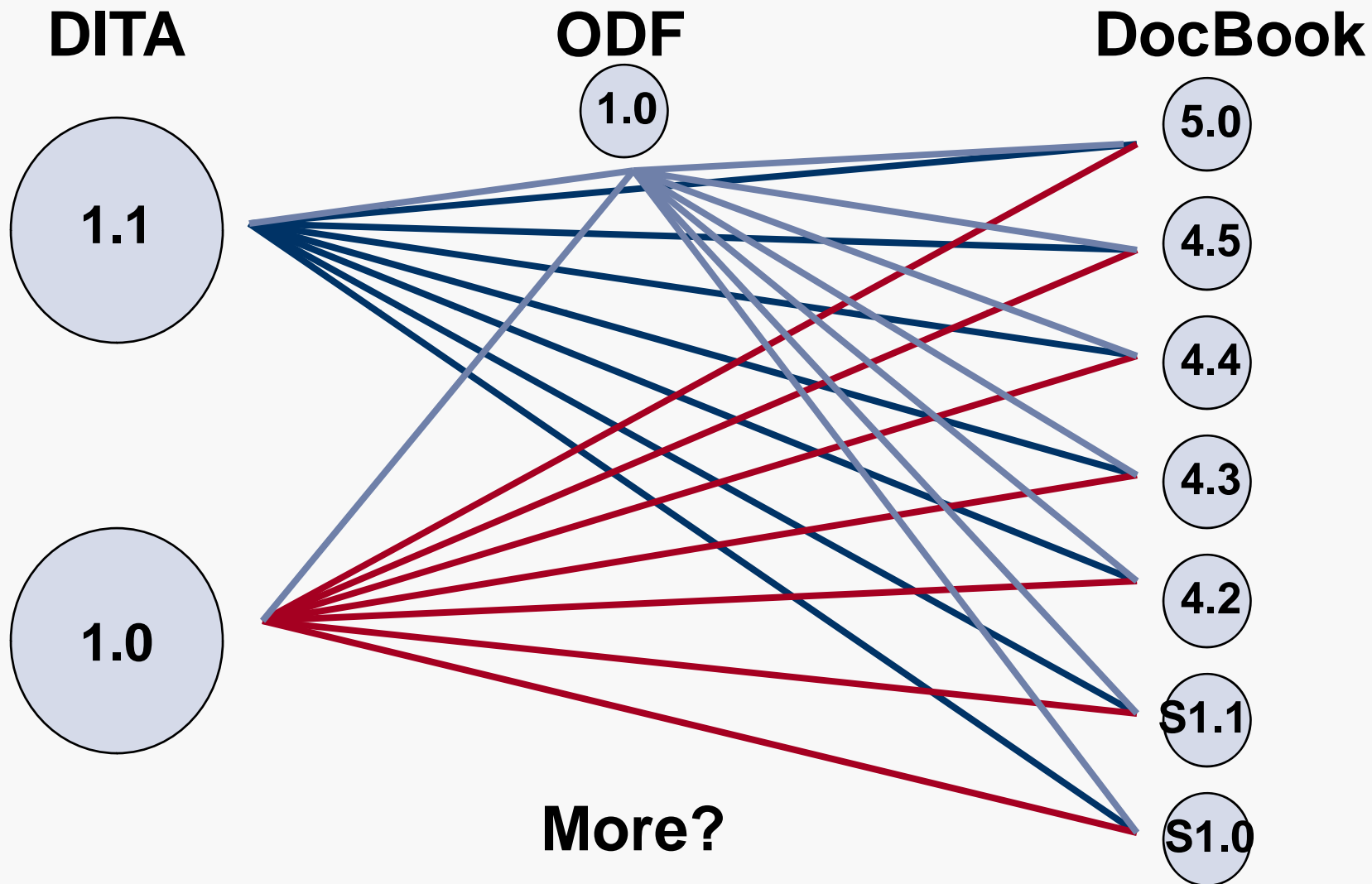
DITA



DocBook



“What a Tangled Web we Weave”



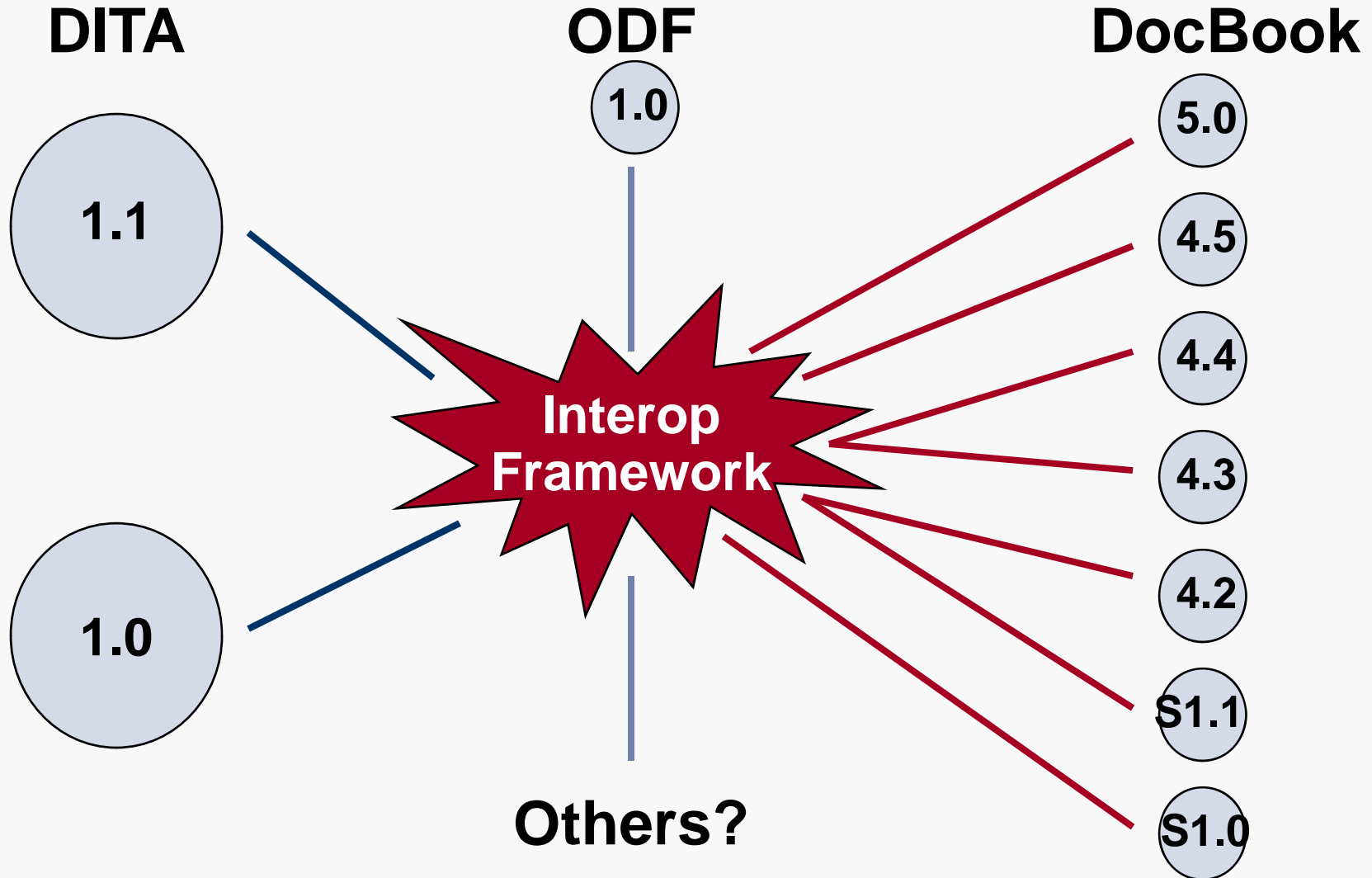
Why Do These Strategies Fail?

- Short-term remedies
 - Can work well if source and destination content stays in the same version of the same standard
- Tight coupling between “source” and “destination” transforms
 - Different versions introduce additional obstacles
 - Interoperability between more than two standards is at best cumbersome

“Lingua Franca”



Version Control



The Interoperability Framework

Choosing an Interoperability Framework Standard

- Not another new grammar! Please?!?
 - Hard to keep up with existing standards
 - Re-Use/Re-purpose an existing standard
 - Leverage existing tools, technology
- What DITA, DocBook, ODF (and others) have in common?
 - Designed for producing Content!
 - Common Structural Components
 - Headings (Sections)
 - Paragraphs
 - Lists
 - Tables
 - Images
 - etc.

Behind the Interoperability Framework

- XHTML
 - XML version of HTML
 - Widely adopted and understood
 - Reduces the learning curve for adoption
 - All XML Document Standards share common structures:
 - paragraphs, lists, tables, images, etc.,
 - Generalized elements to accommodate more specialized inline and block structures
 -
 - <div>
 - Proven to be quite versatile and extensible
 - <http://microformats.org/about/>
 - All of these standards have HTML renditions anyway

Implementing XHTML

- Use XHTML elements for “common” structures
 - `<table>`, `<p>`, ``, ``, ``, `<code>`, `<pre>`, `<abbr>`, `<acronym>`
- Use “generic” structural elements to reflect the structural intent of the source element
 - `<div>`, ``
- `<head>` element used to store metadata about the source XML content
- `<body>` contains the source XML content in the Interop Framework markup

Mapping the Standards

- To develop the interoperability framework, a mapping of content elements between the standards will be needed:

DITA Element

- title
- steps
- step
- substeps
- p
- ul
- ol
- note type="note"
- note type="caution"
- note type="warning"
- b
- u
- i

DocStandards
Interoperability
Framework
Elements

Other Doc
Standards
(ODF, etc.)

DocBook Element

- title
- procedure
- step
- substeps
- para
- itemizedlist
- orderedlist
- note
- caution
- warning
- emphasis role="bold"
- emphasis role="underline"
- emphasis role="italic"

Mapping Rules

- Not all elements have a 1:1 mapping
- Some markup will be *implied* to and from framework:

Interoperability Framework (XHTML)

```
<i mg src="foo.png" />
```

DocBook

```
<medi aobject>  
  <i mageobject>  
    <i magedata href="foo.png" />  
  </i magedata>  
</medi aobject>
```

DITA

```
<i mage href="foo.png" />
```

Preserving Semantic Intent

- XHTML provides the foundation, but we need additional information for standards-specific structures
- Many of these structures will map to <div> and
- The Solution:
 - Define “Semantic Classifications”
 - “contentblock” – chapters, appendices, etc.
 - “topicblock” – DITA topics, DocBook sections, etc.
 - “parablock” – block level elements
 - etc.
 - Store these in the ubiquitous *class* attribute:

```
<div class="contentblock">
```

```
...
```

```
</div>
```

Preserving Source Mappings

- Additional information may be necessary related to the source content
- Three Types of Source Metadata:
 - **Content Level Metadata** – DTD/Schema of the original XML content (e.g., DITA Topic 1.1)
 - **Element Level Metadata** – Name and any attributes of the original XML element
 - **Retroactive Metadata** – Information stored in the source XML content that allows us to determine the original semantics in a round-trip scenario

Content Level Metadata

- Information about the source XML
 - DTD / Schema Name
 - DTD / Schema Version
 - Root Element
 - Namespaces (if applicable)
 - Interop Format Name
 - Interop Format Version
- Stored as <meta> elements in the <head>

```
<head>  
  <meta name="source-schema-name" content="DocBook" />  
  <meta name="source-schema-version" content="version" />  
  <meta name="source-root-element" content="article" />  
  <meta name="interchange-format" content="interop-framework" />  
  <meta name="interchange-version" content="0.1" />  
</head>
```

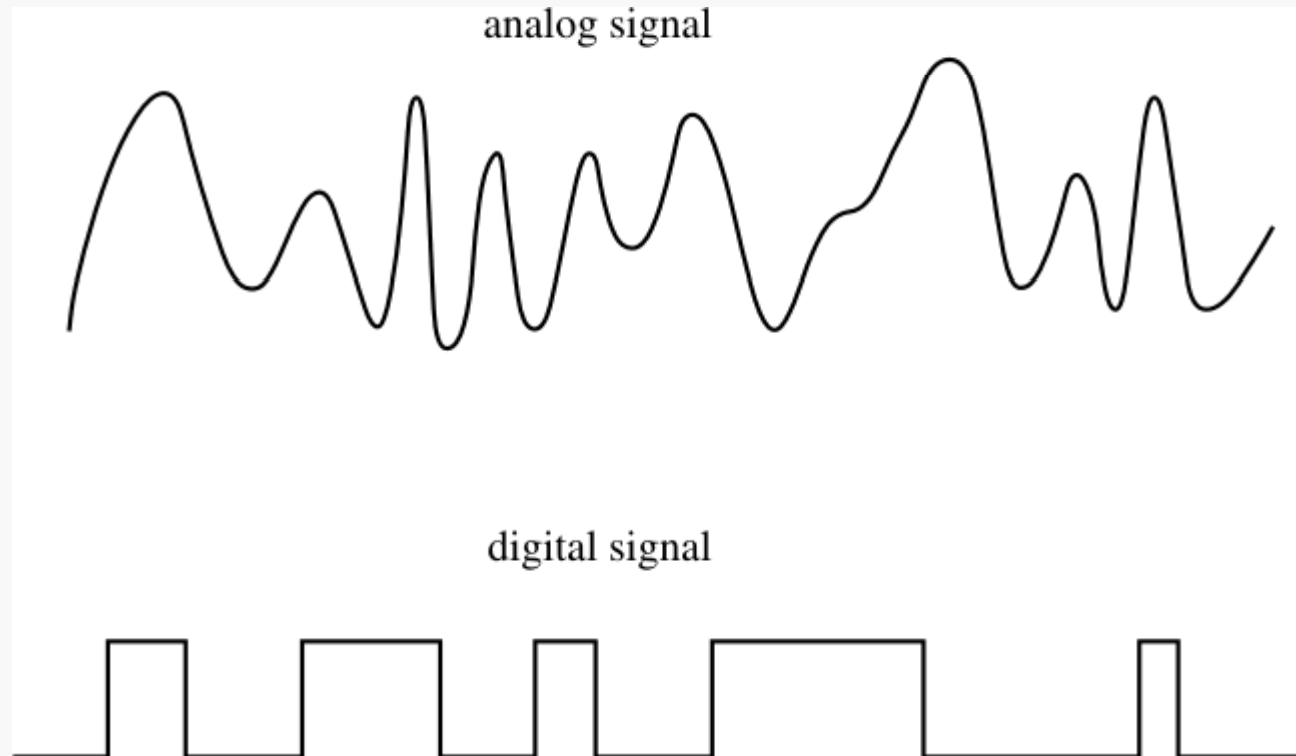
Element Level Metadata

- Source element's name and attributes
 - Enabled for introspection and potential refined translations
- Element Name stored in ubiquitous *title* attribute
<div class="contentblock" **title**="topic">
- Attributes:
 - Stored at *namespaced* attributes, bound to Interop Framework namespace URI. For example:
 - DITA element with *class* attribute: <p class="- topic/p />
 - Stored in Interop Framework: <p **interop:class**="- topic/p />
 - Namespace ensures consistent access to source attributes
 - Avoids potential local XHTML attribute name collisions
 - Plays nicely with XHTML

Retroactive Metadata

- Embed metadata about the source XML into the destination XML
- Provides a mechanism for round-tripping content to and from the source XML
- Only needed if collaborative content interchange is a real issue
- Implicit introspection of *current* XML standard
- Dependent on built-in metadata tags/attributes in the standards:
 - DocBook: *remap* attribute, possibly the *info* element
 - DITA: *outputclass* or *otherprops* attribute, possibly the *data* element
 - ODF: *meta* element
 - Others?

How High is the Fidelity?



Roundtrip Fidelity

```
<emphasi s rol e="bol d" >
```



```
<b ti tle="emphasi s" i nterop: rol e="bol d" >
```



```
<b outputcl ass="emphasi s" otherprops=" rol e=bol d" />
```



```
<b ti tle="b" i nterop: outputcl ass="emphasi s"  
i nterop: otherprops=" rol e=bol d" />
```



```
<emphasi s rol e="bol d" remap="b" >
```

DITA Roundtrip Example

```
<concept>
```



Interoperability Framework

```
<div title="concept" class="topicblock">
```



DocBook

```
<section remap="concept">
```



Interoperability Framework

```
<div title="section" class="topicblock"
  interop:remap="concept">
```



DITA

```
<topic outputclass="section"> OR
```

```
<concept outputclass="section">
```

Purity vs. Practicality

- “It’s not pure [Insert XML Standard Here]”



What Can You Do With the Framework?

- Enable interoperability between two or more standards
- Enable interoperability between different *versions* of each applied standard
- “Unlock” content in proprietary formats for initial migration to XML Document Standards
- Apply the 80/20 rule to semantic accuracy

Next Steps

- Flatirons will be posting a whitepaper explaining the Interoperability Framework
- Flatirons has proposed an OASIS Technical Committee to continue evolving the Interoperability Framework
- We're in the process of creating a charter for TC Formation with OASIS
 - Standards members include:
 - Michael Priestley (DITA)
 - Scott Hudson (DocBook, DITA)
 - Don Harbison (ODF)
 - Jim Earley (DITA)

A Call to Arms

- We need your:
 - Awareness
 - Support
 - Participation!
- For more information:
 - Email thoughtleader@flatironssolutions.com to subscribe to our whitepaper mailing list
 - Download our whitepaper at flatironssolutions.com
 - docstandards-interop-tech@lists.oasis-open.org