



Managing Semantics with Content Using DITA XML

Eric Hennum, IBM


IBM User Technology
<h2>Managing semantics with content using DITA XML</h2>
presented by Erik Hennum IBM STG User Technology
March 2006 Managing semantics within content
© 2006 IBM Corporation

IBM User Technology	
<h3>The talk at a glance</h3> <ul style="list-style-type: none">▪ The need for content semantics▪ Background about SKOS and DITA▪ The XML implementation▪ Lessons learned, limitations, and future directions	
Managing semantics within content	© 2006 IBM Corporation

2006 Semantic Technology Conference

San Jose, California ● March 6-9, 2006

IBM User Technology

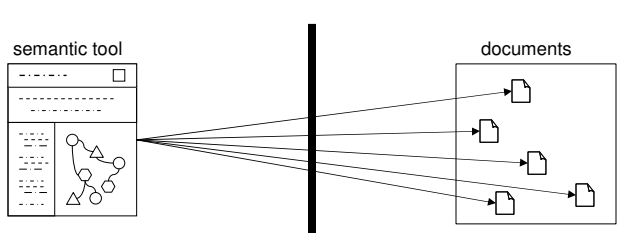
Content – the other semantics

- **The human-readable documents (discourse)**
 - Not the values that are processed only by software (data)
 - Classification identifies what the content is about
- **Important for the vision of the Semantic Web**
 - The worldwide distributed database should support text blobs
 - The issues of semantic interoperability and integration also apply
- **Why the semantics of content are useful**
 - Discover the relevant content
 - Filter the irrelevant content
 - Compose views of content based on relevance
- **Content provides a human interface for data semantics**
 - Where people have to understand a thing or an activity
 - Isolated fragments of black box text can't handle all cases

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Problems with divorcing the semantics from the content

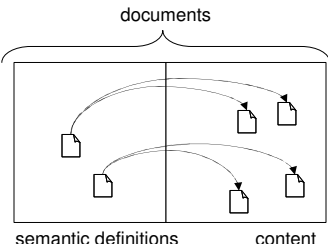


- **Typically semantics and content are maintained separately**
 - By different people and with different tools
- **Classifiers have to read and understand the content**
 - Expensive if they do and inaccurate if they don't
- **Classification and content have to be maintained in parallel**
 - Hard to coordinate and inaccurate when it isn't coordinated
- **Focus on the tool instead of the content**

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Benefits of keeping the semantics close to the content



The diagram illustrates the relationship between documents, semantic definitions, and content. A bracket labeled 'documents' spans across a box. Inside the box, a vertical line separates the left side, labeled 'semantic definitions', from the right side, labeled 'content'. On the left side, there are two document icons. On the right side, there are two document icons. Arrows point from the top-left icon to the top-right icon, and from the bottom-left icon to the bottom-right icon. Additionally, there are curved arrows pointing from the top-right icon back to the top-left icon, and from the bottom-right icon back to the bottom-left icon, indicating bidirectional relationships between the semantic definitions and the content.

- **Use document tools to define the semantics**
- **Writers maintain the content and its classification**
 - Two different ways of expressing the same subject matter
 - Use the classification to improve the content
- **Low barrier to entry**

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

Theoretical background


- **Hypertext theory**
 - “The relationship between hypertext and semantic networks has long been realized.” – Horrocks, McGuinness, and Welty
 - “Hypertext can be seen as a logic representation, where semantics are encoded in both the textual nodes and the graph of links.” – Millard, Gibbins, Michaelides, and Weal
 - Coarse-grained, evolving, tacit, or contextual knowledge benefit from a less formal representation – Shipman and Marshall
- **TopicMaps standard**
 - Formalized content relationships – table-of-contents, index, ...

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

Ingredients of the solution



- **SKOS**
 - W3C RDF vocabulary (Simple Knowledge Organization System)
 - Formal concepts and their relationships
- **DITA**
 - OASIS XML standard (Darwin Information Typing Architecture)
 - Human-readable, semantic content objects and their relationships
- **DITA taxonomy specialization**
 - Extends DITA to provide an authorable XML format for the SKOS model
 - Uses hypertext relationships to specify semantic relationships
 - Defines a taxonomy
 - Classifies the content

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

SKOS (Simple Knowledge Organization System)

- **W3C Public Working Draft**
 - “... expresses the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, 'folksonomies', other types of controlled vocabulary, and also concept schemes embedded in glossaries and terminologies.”
- **Fills a hole in the RDF stack**
 - Between ad hoc RDF properties and fullblown OWL ontologies
- **Enriched by a cross-section of perspectives**
 - Library Science experts
 - Terminology experts
 - RDF and TopicMaps standards leaders
 - Open Source project leads (content management)

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

DITA (Darwin Information Typing Architecture)

An OASIS XML standard
Goal of usability for writers and vocabulary designers (not just processing)

High-level features:

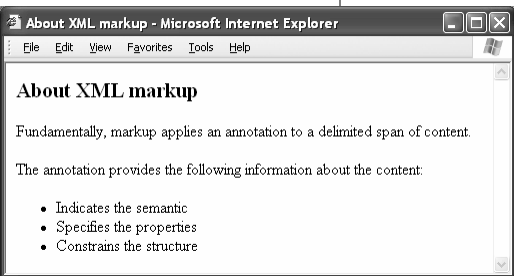
- 1. Topics (documents)**
Human-readable content objects
Emphasis on semantic focus
- 2. Maps**
Hierarchical or associative relationships between topics
- 3. Specialization**
Extensibility to add modular XML vocabularies
Increase semantic precision and constrain content structures

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

DITA 1: A topic is a content object

```
<topic id="xmlmarkup" xml:lang="en-us">
  <title>About XML markup</title>
  <shortdesc>Fundamentally, markup applies an
  annotation to a delimited span of content.</shortdesc>
  <body>
    <p>The annotation provides t
    <ul>
      <li>Indicates the semantic</li>
      <li>Specifies the properties</li>
      <li>Constrains the structure</li>
    </ul>
    <example>...</example>
  </body>
</topic>
```



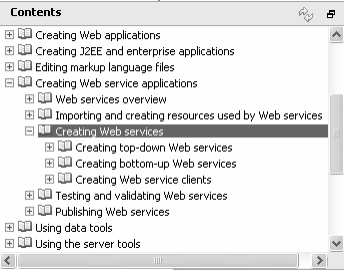
- **Semantic focus, granularity, and independence**
Rich text elements from HTML such as <p>, , <dl>, ...
Emphasis on structure and semantics of content instead of presentation

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

DITA 2: A map defines relationships between topics

```
<map>
  ...
  <topicref navtitle="Creating Web service ap
  ...
  <topicref navtitle="Creating Web services
  <topicref navtitle="Creating top-down W
  ...
  <topicref navtitle="Creating bottom-up
  ...
  <topicref navtitle="Creating Web servic
  ...
</map>
```



- **Relationships and properties defined outside of the topics**
 - Hierarchical relationships – for instance, a navigation such as a sitemap
 - Matrix or group associative relationships – for instance, related links
- **A topic can have different relationships in each context**
 - Organize subsets of the same content in many different ways

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

DITA 3: New markup can be specialized

```
General topic
<topic id="installstorage"
<title>Installing a hard
<body>
<ol>
<li><ph>Unscrew the co
<itemgroup>The drive
</li>
<li><ph>Insert the dri
<itemgroup>If you fe
</step>
</ol>
</body>
</topic>


Specialized task
<task id="installstorage">
<title>Installing a hard drive</title>
<taskbody>
<steps>
<step><cmd>Unscrew the cover.</cmd>
<stepresult>The drive...</stepresult>
</step>
<step><cmd>Insert the drive...</cmd>
<info>If you feel resistance...</info>
</step>
</steps>
</taskbody>
</task>
```

- **Derive new XML markup from existing markup**
 - Extension by substitution to increase semantic precision
 - Modules pluggable into the base DITA vocabulary
- **Adapts to support new requirements**

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Step 2: Define a taxonomy with a hypertext map



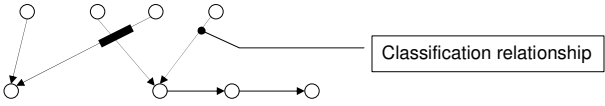
```
<subjectScheme ...>  
  <hasKind>  
    <subjectdef navtitle="Application server technology" ...  
    <subjectdef navtitle="Web Services" href="WebServices.dita">  
    ...  
  <relatedSubjects>  
    <subjectdef navtitle="Service Oriented Architecture" ...  
    <subjectdef navtitle="Web Services" href="WebServices.dita"/>  
    ...  
</subjectScheme>
```

- **An extension of the familiar markup for hypertext relationships**
Applied to the documents that define formal subjects
- **Hierarchy of subject documents**
Like a sitemap but for hasKind, hasPart, or hasInstance relationships
- **Associations between subject documents**
Like related links

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Step 3: Classify the content with a hypertext map



```
<map>  
  ...  
  <topicref navtitle="Creating Web service applications"  
    href="creatingwsapp.dita">  
    <topicsubject>  
      <subjectref navtitle="Web Services" href="WebServices.dita"/>  
      <subjectref navtitle="Application Development" ...  
    ...  
</map>
```

- **Insert the classification into the standard navigation**
Use the <subjectref> element to refer to subject documents
Classify individual content documents or entire collections of documents
- **Most writers work with this view**
Information architect maintains the subject documents and taxonomy map
- **Change the taxonomy without changing the classification**

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Generate the runtime representation in SKOS RDF

```
<skos:Concept rdf:about="&SubjectBase;WebServices">  
  <skos:prefLabel xml:lang="en-us">Web Services</skos:prefLabel>  
  <skos:definition xml:lang="en-us">A method for interaction...  
  <skos:scopeNote xml:lang="en-us">Covers WSDL, SOAP, BPEL, ...  
  ...  
  <skos:inScheme rdf:resource="..."/>  
  <skos:broader rdf:resource="&SubjectBase;ApplicationServer"/>  
  <skos:isSubjectOf rdf:resource="&ContentBase;creatingwsapp.html"/>  
</skos:Concept>  
...  
<foaf:Document rdf:about="&ContentBase;creatingwsapp.html">  
  <rdfs:label xml:lang="en-us">Creating Web service applications...  
  <skos:subject rdf:resource="&SubjectBase;WebServices"/>  
</foaf:Document>  
...
```

Subject definition

Content classification

- **XSLT transforms convert DITA source files to runtime SKOS RDF**
Also transform content to HTML pages (where appropriate)
- **RDF APIs can query or traverse the SKOS model**

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

Runtime 1: the Open Source SWED facet browser

The screenshot shows a web browser window with the URL <http://www.swed.org.uk/swed/servlet/Entry?action=v>. The page title is "Search results: Topicofinterest=Uswed_toi:climate_and_meteorology - Microsoft Internet Ex...". The page content includes a search bar, a "Current Search Results" section with a "Topic of interest" filter set to "Climate and Meteorology", and a list of results. The first result is "Campaign for the Protection of Rural Wales (CPRW)", and the second is "Climatic Research Unit (CRU)".

Classified content

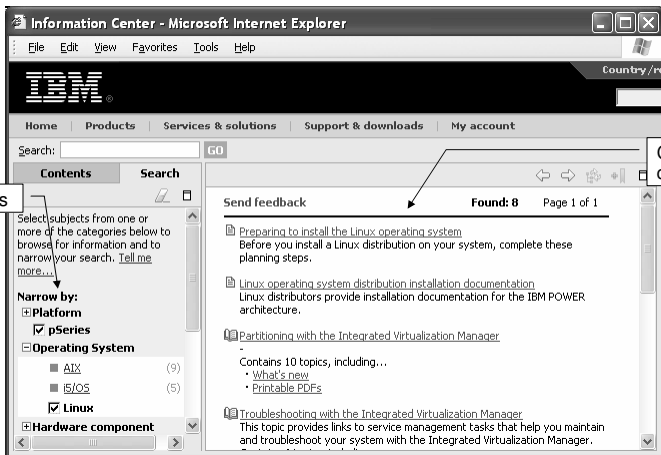
<http://www.swed.org.uk/swed/servlet/Entry?action=v>

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

Runtime 2: IBM STG product information browser



The screenshot shows a Microsoft Internet Explorer browser window displaying the IBM Information Center search results. The search query is "Linux". The left sidebar shows a navigation tree with "Subjects" and "Narrow by:" categories. Under "Operating System", "Linux" is selected. The main content area shows search results for "Found: 8" items, including "Preparing to install the Linux operating system" and "Linux operating system distribution installation documentation". A "Classified content" label points to a search bar area, and a "Subjects" label points to the left sidebar.

Subjects

Classified content

Snapshot of work in progress
<http://publib.boulder.ibm.com/infocenter-beta/eserver/>

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Demonstration

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Variation 1: A document as a subject and content

```
<map>
...
<topicref navtitle="Creating Web service applications"
  href="creatingwsapp.dita">
  <topicsubject>
    <subjectref navtitle="Web Services" href="WebServices.dita"/>
  ...
<topichead navtitle="Glossary">
  ...
  <topicref navtitle="Web Services" href="WebServices.dita"/>
  ...
</map>
```

Document as subject

Document as content

- **Useful when readers might need the subject definition**
Typically unfamiliar glossary terms or conceptual background
- **Refer to the same topic with different elements**
Use <subjectref> to classify content with the subject
Use <topicref> to include the subject document in the navigation
- **Standard formatting can process the specialized subject topic**

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Variation 2: Classify HTML or PDF documents

```
<map>
...
<topicref navtitle="Web Services Activity"
  format="html" href="http://www.w3.org/2002/ws/">
  <topicsubject>
    <subjectref navtitle="Web Services" href="WebServices.dita"/>
  ...
<topicref navtitle="Business Process Execution Language 1.1"
  format="pdf" href="ws-bpel.pdf">
  <topicsubject>
    <subjectref navtitle="Web Services" href="WebServices.dita"/>
  ...
</map>
```

Public web content

- **Local or remote resources in other formats**
Use the format attribute to distinguish from DITA content documents
Supply a title for readable source and simple processing
- **Classify the resource with DITA subject documents as usual**

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Variation 3: Define subjects in SKOS or OWL

```
<subjectScheme ...>
  <hasKind>
    <subjectdef navtitle="Application server technology"
      format="skos" href="http://some.org/subjects.rdf#appserver">
      <subjectdef navtitle="Web Services" href="WebServices.dita">
      ...
    </subjectScheme>
```

Public definition

- **Integrating with public semantics or local ontologies**
 - Use the format attribute to distinguish from DITA subject documents
 - Best to supply a title for readable source and simple processing
- **Use external subjects in taxonomy or classification**
 - RDF-based or TopicMaps formats
- **Extend shared general semantics for local specific semantics**
 - External subjects provide the trunk and branches of the taxonomy
 - DITA subjects provide the twigs and leaves for content semantics

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Variation 4: Tagging without a taxonomy

Steps:

1. **Define the subject documents as usual**
 - Don't organize the subjects in a taxonomy
2. **Classify the content with the subjects as usual**


Benefits of defining tags as subjects:

- **Minimize single tags with many meanings**
- **Minimize multiple tags with the same meaning**
- **Increase the semantic precision of tagging**

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Challenges for writers




- **Focusing on the subject meaning instead of the label**
The titles are only reminders of the subject definition and content meaning
- **Defining the taxonomy hierarchy**
 - Creating hasKind relationships instead of hasPart relationships
 - Applying the intersection of existing subjects instead of defining a new compound subject
 - Balance – consistent depth of coverage
 - Pragmatic discipline – avoid obsession with subdivision of meaning
- **Modifying the content during classifying**
Avoid treating the current content or navigation as a carved in stone

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Benefits of the approach

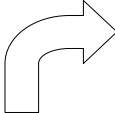


- **Leverage document familiarity and infrastructure**
 - Demystification – you already create semantics when you create content
 - Apply hypertext understanding to taxonomy definition and classification
 - Use content tools to edit, format, and archive the semantic definitions
 - A low cost and open solution for basic semantics of content
- **Improve the content**
 - Identify holes in your coverage of the subject matter
 - Identify content with a blurred focus
 - Identify duplicate content and avoid contradiction nightmares

Managing semantics within content © 2006 IBM Corporation

IBM User Technology

Limitations of the approach



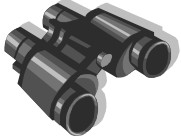
- **Not a format for rigorous ontologies**
 - Useful where an ontology would be hard to create or maintain
 - Integrate content semantics with ontologies as part of a continuum
- **Not a format for common linguistics**
 - Best in a domain with precise concepts
- **Not a method for classification of static inventories**
 - Best where content is maintained with semantic focus
 - Integrate with text mining (such as UIMA) for large content archives
- **Not a power tool for semantic management**
 - Can provide a document interface for semi-formal knowledge

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

Future directions



- **Add a display-oriented map for subjects**
 - Separate presentation from taxonomy definition concerns (SKOS idea)
- **Represent deeper knowledge about the subjects**
 - Associative relationships and properties for subjects
 - Specialize the subject definition elements – aligns with SKOS
 - speculations about subclassing in parallel with conceptual hierarchy
- **Leverage more of the content semantics**
 - Specialized DITA markup provides semantic annotation for content
- **Public definitions of technical subjects**
 - ACM taxonomy not detailed enough
 - CIM model isn't user oriented

Managing semantics within content

© 2006 IBM Corporation

IBM User Technology

Future: specialized subject relationships

- **Source for a specialized associative relationship**

```
<containerFor>
<subjectdef href="toolbox.xml"/>
<subjectdef href="tools.xml"/>
</containerFor>
```
- **Runtime RDF**

```
<rdf:Property rdf:about="&garage;containerFor">
<rdfs:subPropertyOf rdf:resource="&skos;
  related"/>
...
<skos:Concept rdf:about="&garage;toolbox">
<garage:containerFor
  rdf:resource="&contentSubject#tools"/>
```

Managing semantics within content

© 2006 IBM Corporation


IBM User Technology

Summary

- **Importance of semantics for content**
 - Not data vs documents but both
 - Not formal vs informal semantics but a continuum
- **Maintain semantic declarations with the content**
 - Improve the content instead of trying to bolt on semantic precision
- **Leverage SKOS and DITA standards and tooling**
 - Use the familiarity of writers with hypertext and documents:
 1. Define subject documents
 2. Organize subjects in a taxonomy map
 3. Classify content in a navigation map

Managing semantics within content

© 2006 IBM Corporation

	IBM User Technology	
<h2>Resources</h2> <ul style="list-style-type: none">▪ SKOS W3C - http://www.w3.org/2004/02/skos/▪ DITA OASIS – http://www.oasis-open.org/committees/dita Cover page – http://xml.coverpages.org/dita.html Forum – http://groups.yahoo.com/group/dita-users/ DITA Open Toolkit – http://dita-ot.sourceforge.net/▪ DITA Taxonomy specialization Article – http://www-128.ibm.com/developerworks/xml/library/x-dita10/ Plugin available at the SourceForge site for the DITA Open Toolkit <p>Feedback welcome – Erik Hennum – ehenum@us.ibm.com</p>		
	Managing semantics within content	© 2006 IBM Corporation