

Ontologies and User Needs in Publishing

NISO/BISG 4th Annual Forum
The Changing Standards Landscape
Focus on the Item:
Understanding the end-user
perspective

Jabin White
Director of Strategic Content
Wolters Kluwer Health – P&E
June 25, 2010

Agenda

- Introductions
- Some basic definitions
 - Vocabularies, Taxonomies, and Ontologies, Oh My!
- Semantic Markup in Publishing
- Tying All of This to User Needs
 - AKA, “It’s all about the use case!”
- The Semantic Web

Introductions: My Company

- Director of Strategic Content for Wolters Kluwer Health – Professional & Education
- Wolters Kluwer Health includes:
 - Lippincott Williams & Wilkins titles
 - Ovid
 - UpToDate
 - Provation Order Sets
 - Drug Facts & Comparisons
 - Medi-Span
 - Clin-eGuide

Introductions: Me

- Started as Editorial Assistant
- Dove into SGML in the mid-90s working on drug reference
- Six years at Elsevier in Electronic Production
- Joined WK Health in May 2009
 - Responsible for making sure content flows through company more efficiently (DTDs, Content Management, Authoring Tools, Semantic Enrichment, Product Information Management, etc.)

The Web - Stop the Insanity!

- A few humble web stats:
 - There are 2 billion (billion!) Google searches daily
 - There are 1 trillion (**1,000,000,000,000**) unique URLs in Google’s index
 - There are **2,695,205** articles in English on Wikipedia
 - It would take 412.3 years to view all the content on YouTube (3/08), but don’t try, because there are 13 hours of video uploaded every minute

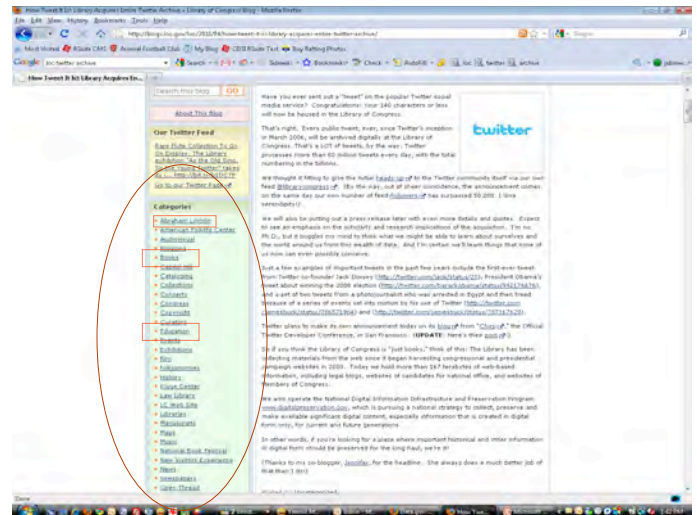
** Source: Adam Singer’s “Social Media, Web 2.0 and Internet Stats site:
<http://thefuturebuzz.com/2009/01/12/social-media-web-20-internet-numbers-stats/>

Wow!

- There is even an “Information Overload Research Group”
- <http://iorgforum.org/>

So What?

- Clay Shirky's concept of "Filter Failure"
- When the capacity of people to "keep up with" information is exceeded, curation becomes the value differentiator



So How Do Publishers' Solve this and Provide Value?

Ontologies

- The use of ontologies allows publishers to add *meaning* to content from within the tag set
- The process of using ontologies to describe content is called Semantic Tagging, Markup, or Indexing (all synonymous)
 - Publishers are getting better at this, but need to get much, much better

Semantic Basics

- Semantics is tagging that describes what content *is* and not how it should *look* on the page or screen
- Contrast to structural tagging, which is made of elements such as <para>, <list>, and <title>
- Both are XML, but semantics is like XML on steroids!
- Doing semantic tagging without a controlled vocabulary is madness for scholarly publishing
 - Think "folksonomies"

* FUD Around Semantic Search

- When people say "we do semantic search," they could potentially mean two very different things:
- Semantic Search engines
 - TEMIS, Collexis, NetBase, Vivisimo, OpenCalais
 - Finding semantic concepts based on entities and search algorithms
 - Finding a needle in a haystack
- Semantic Tagging
 - People (SMEs) identify concepts and tag accordingly
 - Drives precision in search and other things
- Finding the right needle in a stack of 10 needles

Implications for Search

- Machines don't know the difference between hypertension and high blood pressure
 - More accurately, machines don't know they are the SAME
- How this is handled is a matter of User Experience (did you mean? ... give them the result ... etc.), but the content must be tagged first
- A "matter of User Experience" means you need a Use Case

But First, Some Definitions

- Controlled vocabulary: a bunch of words, no relationships
 - But there is advantage if all users use the same terms to describe things
- Taxonomy: is a controlled vocabulary with hierarchy
- Thesaurus: is interchangeable with controlled vocabulary, also sometimes referred to as an ontology
- Ontology: all of the above; think neural network with a bunch of relationships
- MetaData: data about data (I really hate this definition, but it's accurate)
- MetaThesaurus – a collection of all of these things
 - EXAMPLE: UMLS

Some Level-Setting

- Unfortunately, these definitions have been diluted to the point of uselessness by their misuse
 - Think "Content Management" around the year 2000

Taxonomies in STM

Taxonomies in STM Publishing

- UMLS
- MeSH
- SNOMED-CT
- ICD-9 and ICD-10
- RxNORM
- LOINC, ICPC-93, and VA/KP Subset of SNOMED

UMLS – Unified Medical Language System

- More than 5 million terms or named entities
- Divided into concepts, and each term has unique identifier
- Not a vocabulary, but a mapping BETWEEN vocabularies
- Vocabularies included in the UMLS:
 - MeSH Headings in 8 languages (MeSH = Medical Subject Headings)
 - ICPC-93 in 14 languages (ICPC = International Classification of Primary Care)
 - WHO Adverse Drug Reaction Terminology in 5 languages
 - SNOMED-2, SNOMED-3, and UK Clinical Terms (former Read Codes) (SNOMED = Systemized Nomenclature of Medicine)
 - ICD-10 in English and German (ICD = International Classification of Diseases)
 - ICD-10-AM (Australian Modification)

One Concept, Many Names

| TERM | SOURCE VOCABULARY |
|--------------------------------|-----------------------|
| Atrial fibrillation | ICD-9-CM |
| AF | NCI Thesaurus |
| Afib | MedDRA |
| Atrial fibrillation (disorder) | SNOMED Clinical Terms |
| Atrium; fibrillation | ICPC2-ICD10 Thesaurus |

Semantic Markup is One Type of Metadata

- Reading most definitions of metadata and related standards is like trying to resolve disputes with my kids
- Metadata is “data about data”
 - But what does that mean?
- Any information that describes or adds further value to content is metadata
- Its use may be increasing, but metadata is NOT new

Types of Metadata

- **Classifying Metadata**
 - ISBN (I told you this wasn't new)
 - Dewey Decimal System
 - Books in Print/CIP/Library of Congress data
 - MARC records
 - DOI (Digital Object Identifier)
- **Descriptive Metadata** (sorry, my examples are from STM)
 - ICD-9 and ICD-10 Codes
 - MeSH
 - SNOMED-CT
 - NANDA, NIC, NOC for Nursing
 - NDC, HCPCS for drugs

OLD

NEW

Types of Metadata

- **Classifying Metadata**
 - ISBN (I told you this wasn't new)
 - Dewey Decimal System
 - Books in Print/CIP/Library of Congress data
 - MARC records
 - DOI (Digital Object Identifier)
- **Descriptive Metadata** (sorry, my examples are from STM)
 - ICD-9 and ICD-10 Codes
 - MeSH
 - SNOMED-CT
 - NANDA, NIC, NOC for Nursing
 - NDC, HCPCS for drugs
- DOI (Digital Object Identifier)

OLD

NEW

Types of Metadata

- **HTML Metadata**
 - `<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">`
 - `<meta name="verify-v1" content="kBoFGUuwppiWVWGx4Ypzkw1Cs1GgMYEMMbfNr7FY65w=" />`
 - `<meta name="description" content="International publisher of professional health information for physicians, nurses, specialized clinicians & students. Medical & nursing charts, journals, and pda software.">`
 - For people
 - `<meta name="keywords" content="springhouse, medical book, nursing journal, medical pda software, lippincott medical reference, lww, lippincott, lww com, medical publisher">`
 - `<link rel="stylesheet" href="/css/style.css" type="text/css">`

Semantic Metadata

- Using controlled vocabularies, extra power can be added to content via semantic tagging to drive:
 - More precise searching
 - Contextually-based connections
 - Lowering of “two terms meaning the same thing” syndrome (hypertension vs. high blood pressure; heart attack vs. myocardial infarction)
 - Filling in of content gaps
 - Asking questions of data (aka, querying): “How many chapters do we publish that are tagged with the term “pediatric oncology” or “leukemia” that also contain the treatment “interferon therapy”

Tying All of this to User Needs

Importance of Use Cases

- Use Cases should drive strategy and justifications for ontologies
- One ontology size/coverage does not fit all
- One method of tagging/indexing does not fit all
 - There is a fundamental difference, tension, and ultimately tradeoff between large concept coverage over a massive amount of data, and precise conceptual expressiveness
- Approach should be tailored to content set and goals for that content set

Use-Case Driven Development

- Publishers would *like* to make semantic tagging a “normal” part of the production cycle
- However, tying WHAT USERS WANT out of a content set, and how semantic tagging can assist with that/add value, is critical
- This is done with a Use Case
- Doing use cases or digital products has become a necessary evil for publishers
 - This is new, because print books had 1 use case

The Semantic Web

Semantic Web

- Current web (mostly HTML) is “undefined” information, and the growth is making this even worse
- Semantic web concept would ensure that content providers classify their information, so the web would become more of a smart database of information


Jabin’s Shopping List

HTML

```
<H1>Jabin’s Shopping List
</H1>
<ul>
<li>Bread</li>
<li>Milk</li>
<li>Bananas</li>
<li>Beans</li>
</ul>
```

XML

```
<list type=“grocery” date=“6-25-2010”>
<title>Jabin’s Shopping List</title>
<grain>Bread</grain>
<dairy>Milk</dairy>
<fruit>Bananas</fruit>
<veggie>Beans</veggie>
</list>
```



The semantic web both requires and acts on this kind of tagging

A new idea? ... Not so much

- May 2001 issue, "Scientific American"
- ***The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities***

By Tim Berners-Lee, James Hendler and Ora Lassila

- The entertainment system was belting out the Beatles' "We Can Work It Out" when the phone rang. When Pete answered, his phone turned the sound down by sending a message to all the other *local* devices that had a *volume control*. His sister, Lucy, was on the line from the doctor's office: "Mom needs to see a specialist and then has to have a series of physical therapy sessions. Biweekly or something. I'm going to have my agent set up the appointments." Pete immediately agreed to share the chauffeuring.

Semantic Web vs. semantic Web

- Grand vision of Semantic Web is a great goal, but will take time
- Meanwhile, each industry has its own vocabulary(ies), which can drive their own semantic webs
- Resource Description Framework (RDF) can and will "bind" these webs together, but each industry vertical can make progress in the interim

Implications

- If every industry has its own language, how is that language *expressed*?
- Answer: Ontologies
- How are those ontologies applied?
- Answer: Semantic Tagging

THANK YOU

Jabin White
Director of Strategic Content
Wolters Kluwer Health
Jabin.White@wolterskluwer.com
215.521.8911
Twitter: @jabinwhite
Blog: [Technically Speaking](#) at
<http://www.bookbusinessmag.com/channel/technically-speaking>