Embedding Metadata and Other Semantics In Word-Processing Documents

Peter Sefton (University Southern Queensland)
Ian Barnes (Australian National University)
Ron Ward (University Southern Queensland)
Jim Downing (University of Cambridge) (presenting)

[breathe]

The paper supporting this presentation provides important detail and can be obtained from http://www.dspace.cam.ac.uk/handle/1810/206423
Agenda

* Motivations
* Axioms of choice
* Interoperability is Hard
* The approach
* Examples (+ chemistry)

http://www.flickr.com/photos/forezt/524108228
Why is this interesting?

* We want to move towards semantically-rich documents for e-Research. In some disciplines 100% of documents start life in a word processor.

* Introduction of real world constraints yields interesting result
Semantically Rich Documents

- Enable automation
- Prevent information loss
- Better discovery
- Improved presentation

Automation – zero click upload, not filling in redundant forms etc
Information loss – rich data reduced to tables, images.
Semantic information leads to richer alternatives for discovery and communication of research.
Fully Supported Research – all the supporting data delivered with the text
Constraints

Work in the real world, today

Solution had to work in ICE – the Integrated Content Environment, a distributed authoring system in production at USQ.

Therefore the approach is PRAGMATIC!
Real World

- Metadata, semantics and data not easily distinguished
- Document creation == Metadata creation
- Not separable activities
- Metadata is in the document
- Documents have multiple, distributed authors
Tools and Formats

- Microsoft Word [Adoption]
- OpenOffice.org writer [Access]
- ICE - Integrated Content Environment

- .doc, .docx, OOXML, ODF
- HTML, PDF
This is the test that semantic solutions must work inside to be useful in production – once semantics are created, they must survive when the document is edited in the wild.
This is the simple subset of document interop we’re talking about, including only word and OO Writer.

In the wild you can’t control what formats people use to save, or the software they use.

If any of these routes destroys semantics, then we’ve lost interoperability.

There are a lot of standards already involved in this space, but none of them on their own deliver semantic data interoperability.
Interoperability in Publishing

PDF – scholarly publishing now
HTML – the medium term future of scholarly publishing.

Converter needed since HTML and PDF creation in OOo Writer and MS Word produce pretty poor results.
When you apply these interoperability constraints, the solution space gets very small. <metaphor>Like walking along a ridge, keep it simple and take small steps. The paths off to the side lead quickly to peril.</metaphor>
Approaches Ruled Out

* MS Word “Smart Tags”
  * No interop with OOo, but not necessarily a bad idea
* MS Word foreign namespace XML encoding
  * Expensive, no interop with OOo, lock-in issues
* ODF 1.2 embedded semantic
  * No Word equivalent in sight
* Things that would destroy WYSIWYG such as using wiki markup in the word processor.
Codifying a standard wouldn’t work unless vanilla wp software can be shown not to destroy the information.

For delivering interoperability in this area, standards are not sufficient.
Microformats!

http://www.flickr.com/photos/onion/2046003604
Encoding Microformats

- Tables: for, like, tabulating things
- Styles: The original extensible inline semantic mechanism for word processing and still working!
- Links
- Frames: fragile
- Bookmarks and fields: require lots of field testing, not all that reliable in an interop situation

The paper contains much more detail about the mechanism.
The style approach is:

* Simple
* Metadata schema agnostic
* User extensible

It doesn’t need any plugin / customized software to work.
Embedding metadata and other semantics in word processing documents

Peter Sefton,
Australian Digital Futures Institute, University of Southern Queensland
Ian Barnes,
Digital Resource Services Program, Division of Information,
The Australian National University
Ron Ward,
Australian Digital Futures Institute, University of Southern Queensland
Jim Downing,
The Unilever Centre for Molecular Science Informatics, University of Cambridge
July 2008

Abstract

This paper describes a technique for embedding document metadata, and potentially other semantic references inline in word processing documents, which the authors have implemented with the help of a software development team. Several assumptions underly the approach: It
Tables are also useful since the layout implies semantics
The toolbars are implemented for Word and Writer. They provide easy access to the common microformat encoding styles and structures. They also contain macros for communicating with the ICE system, and uploading the document to the Institutional Repo / publisher system etc.
To make it even easier, templates can be used that include sample text in the relevant places – all the user has to do is replace the sample text.
Dublin Core metadata can be extracted directly from the document.

Embedding metadata and other semantics in word processing documents

This paper describes a technique for embedding document metadata, and potentially other semantic references inline in word processing documents, which the authors have implemented with the help of a software development team. Several assumptions underly the approach: It must be available across computing platforms and work with both Microsoft Word (because of its user base) and OpenOffice.org (because of its free availability). Further the application needs to be acceptable to and usable by users, so the initial implementation covers only small number of features, which will only be extended after user-testing. Within these constraints the system provides a mechanism for encoding not only simple metadata, but for inferring hierarchical relationships between metadata elements from a flat word processing file. The paper includes links to open source code implementing the techniques as part of a broader suite of tools for academic writing. This addresses tools and software, semantic web and data curation, integrating curation into research workflows and will provide a platform for integrating work on ontologies, vocabularies and folksonomies into word processing tools. [The work here can/will be demonstrated in a presentation if the paper is accepted]

Peter Sefton

Australian Digital Futures Institute, University of Southern Queensland

Jan Barnes
As can RDF metadata using the ORE vocabulary.
ICE-TheOREM

- Semantics in chemistry thesis documents
- Structural elements, Chapters, Appendices etc
- Data (molecules, spectral data etc)
- Chemical entities in text

http://wwmm.ch.cam.ac.uk/trac/theorem/
This text from a synthetic chemistry thesis.

Highlights the grey area between data and metadata – the compound name is data, but also the subject of the document.
These screenshots taken from CML in ICE demo at http://ice.usq.edu.au/presentations/demos/index.htm
These screenshots taken from CML in ICE demo at http://ice.usq.edu.au/presentations/demos/index.htm
CML Example

As an example of ICE's integration with data-driven research, when viewed online the chemical molecule below is shown as an interactive 3d applet, while for print there is a two dimensional graphic.

C<sub>9</sub>H<sub>11</sub>N<sub>2</sub>O<sub>4</sub> (SOURCE: Crystaley)
These screenshots taken from CML in ICE demo at http://ice.usq.edu.au/presentations/demos/index.htm
CML Example

As an example of ICE's integration with data-driven research, when viewed online the chemical molecule below is shown as an interactive 3d applet, while for print there is a two dimensional graphic.

C_{16}H_{12}N_{2}O_{4} (SOURCE: Crystaleyce)
Thank you.

http://www.flickr.com/photos/jaysun/367670007
ICE - Integrated Content Environment  http://ice.usq.edu.au/
Demos at http://ice.usq.edu.au/presentations/demos/

ICE-TheOREM. Tag: jisctheorem
https://wwmm.ch.cam.ac.uk/trac/theorem
http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/
  theoremice.aspx

  Peter Sefton
  sefton@usq.edu.au
  http://ptsefton.com/

  Jim Downing
  ojd20@cam.ac.uk
  http://wwmm.ch.cam.ac.uk/blogs/downing/