

# 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)

[breath]

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



<http://www.flickr.com/photos/amirjina/2281612876>

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

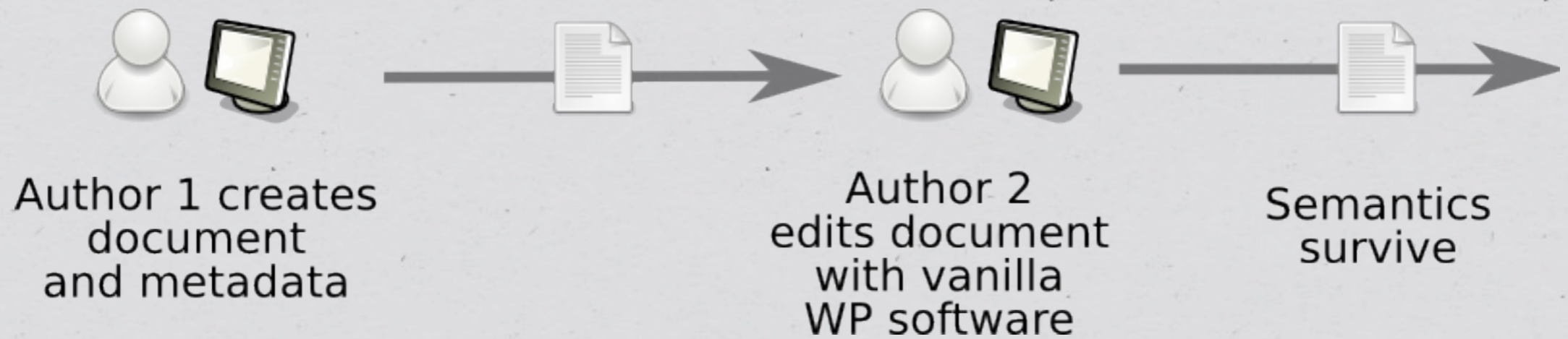
# Tools and Formats

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



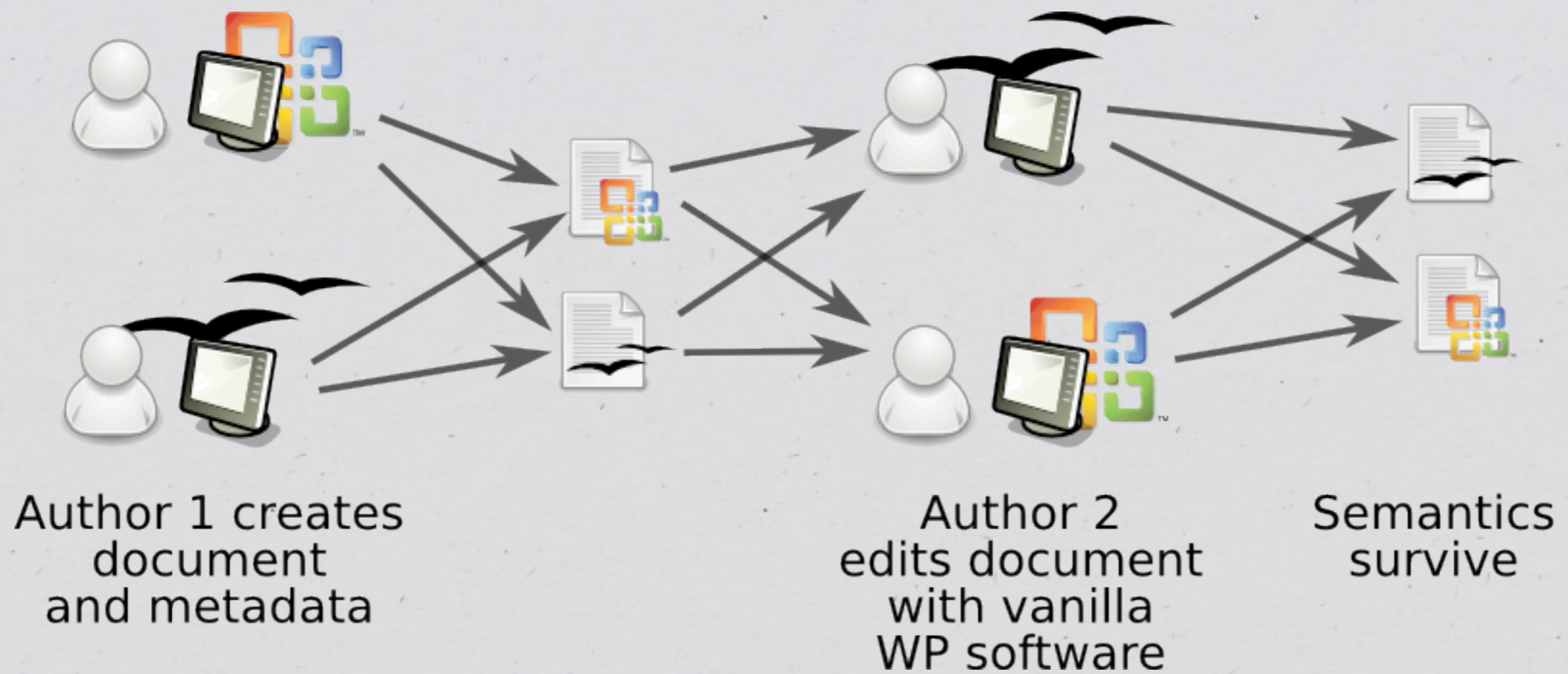
- \* .doc, .docx, OOXML, ODF
- \* HTML, PDF

# The Difference Between Standards and Interoperability



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.





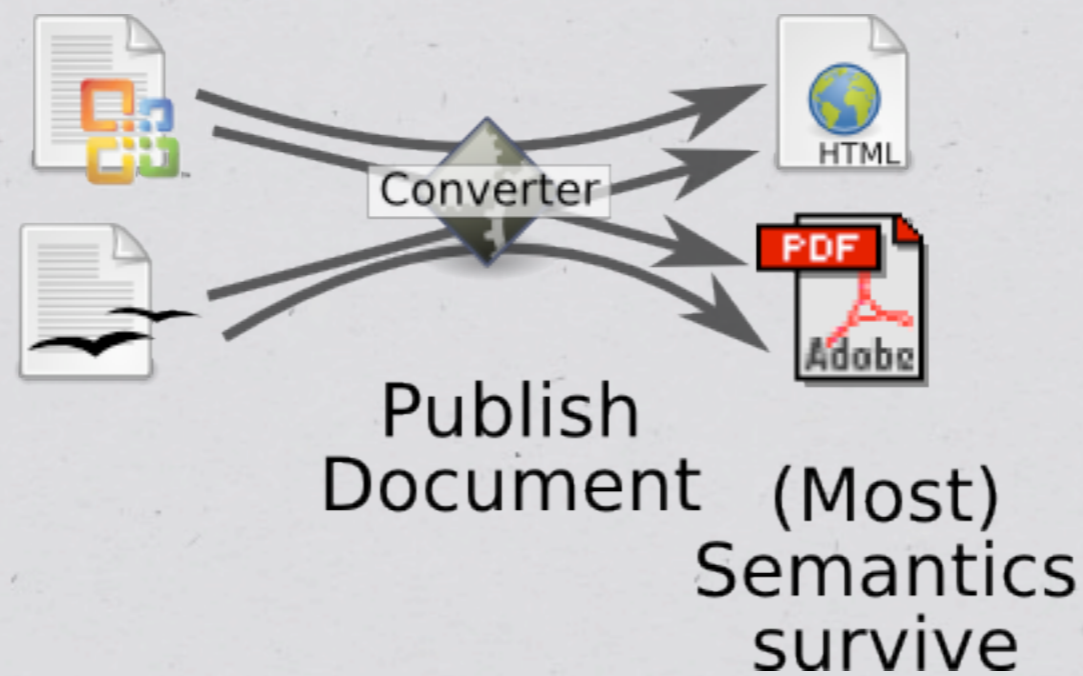
This is the simple subset of document interop we're talking about, including only word and OO Writer.

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.



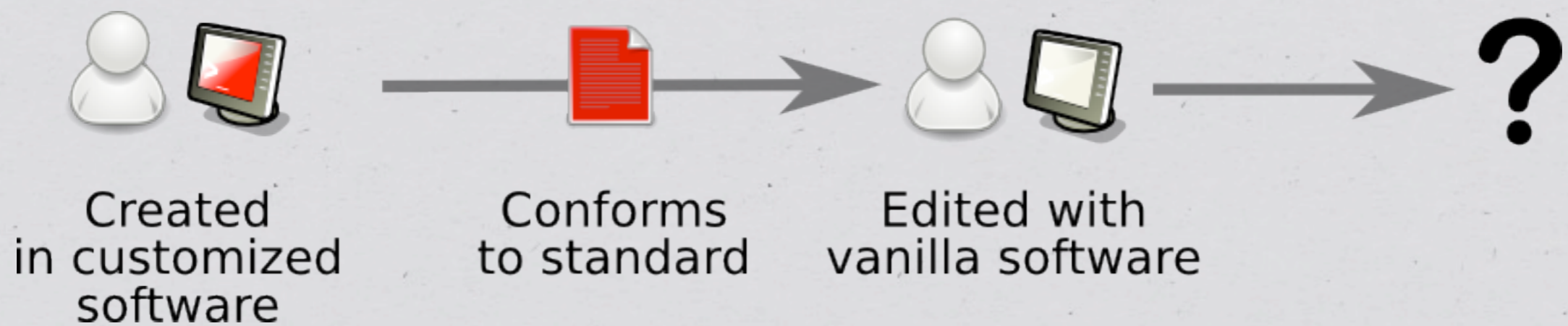
<http://www.flickr.com/photos/druclimb/289636172>

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.

# Define A New Encoding Standard?



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!



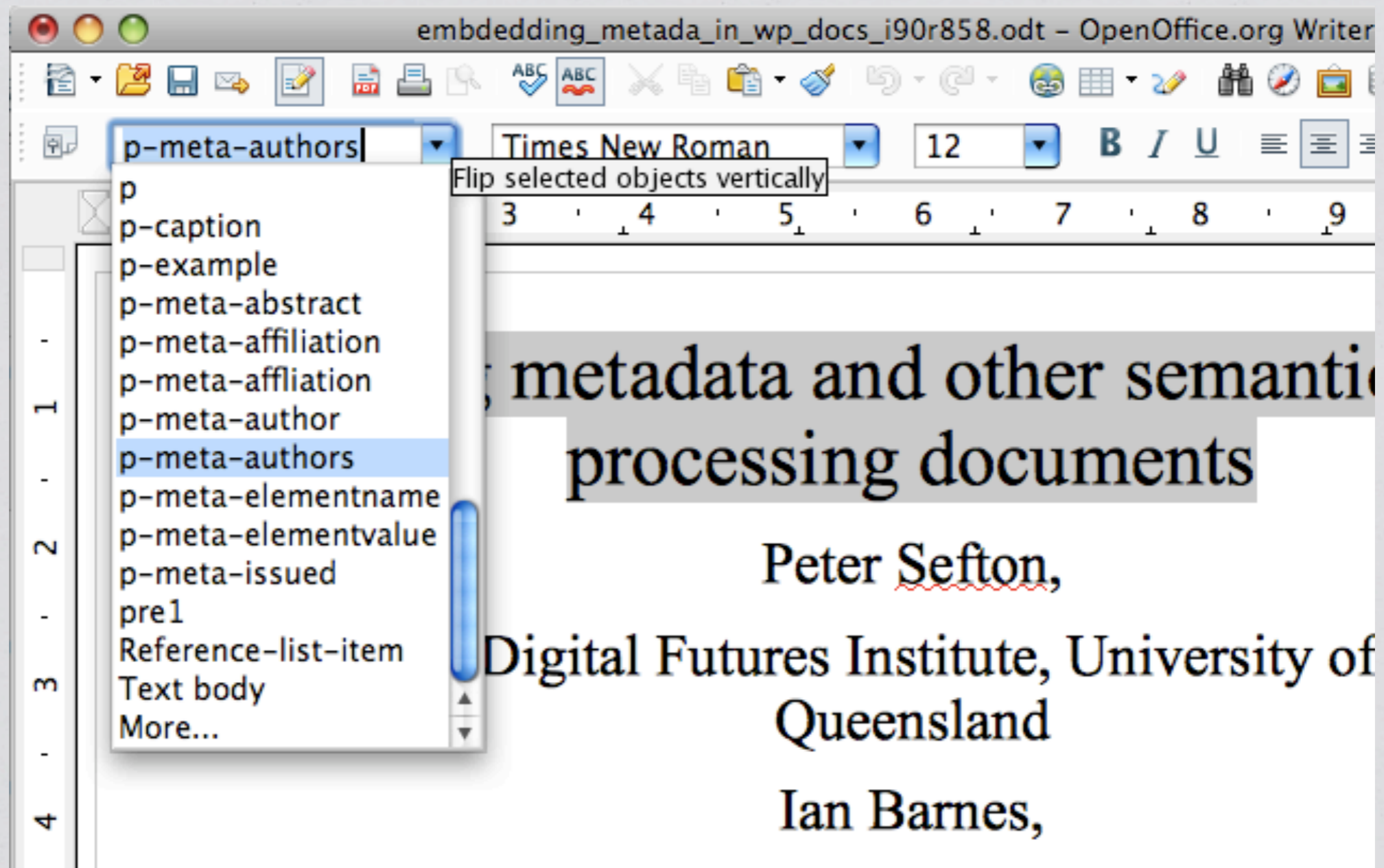
<http://www.flickr.com/photos/onion/2046003604>

# 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.

# Styles



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,

Style: p-meta-author

Australian Digital Futures Institute, University of Southern Queensland

Ian Barnes,

Style: p-meta-affiliation

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

Style: p-meta-issued

July 2008

Style: p-meta-abstract

## 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

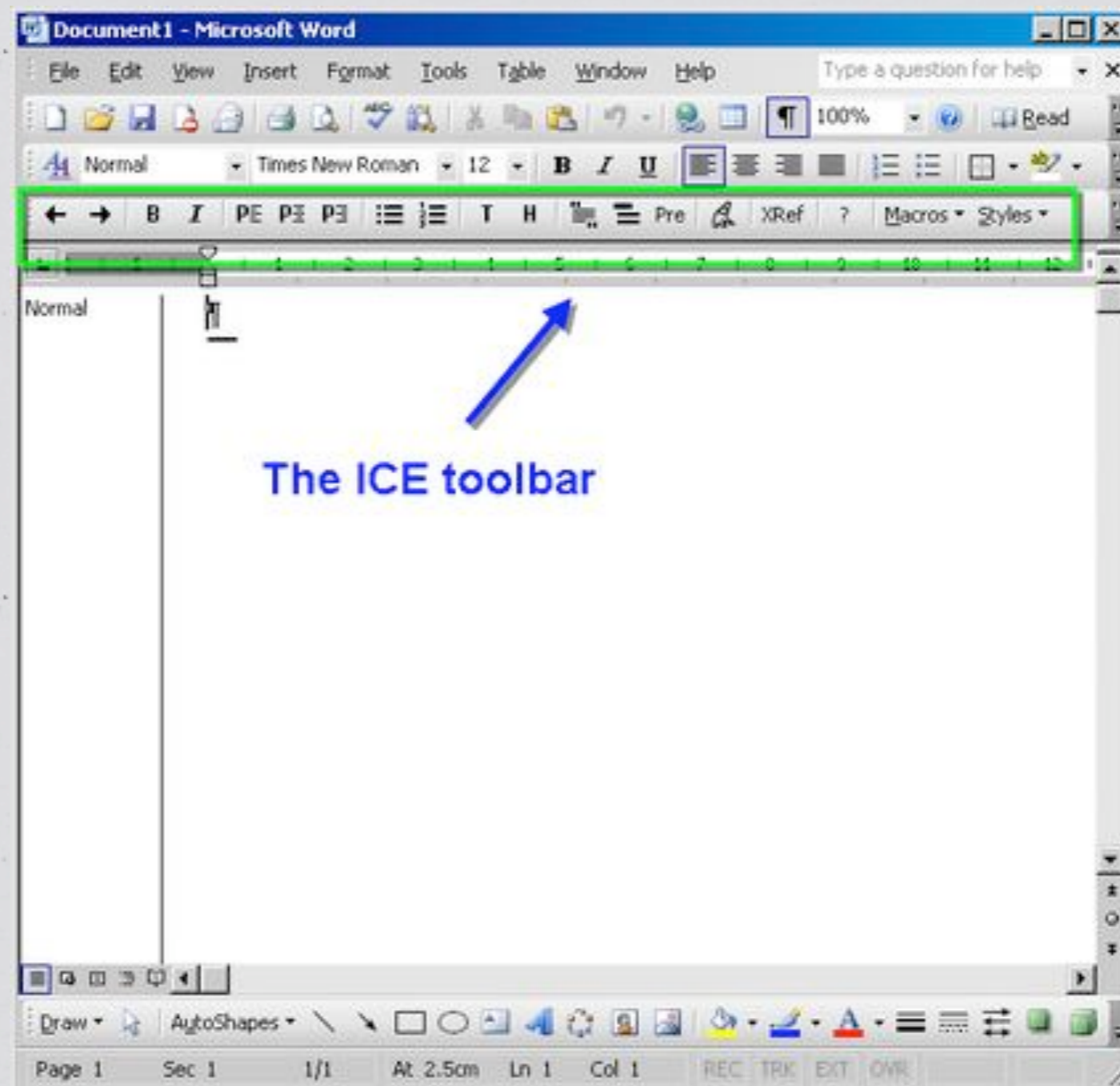
Styles can be nested by placing inline styled text within styled paragraphs.

<i><u>Metadata</u> {meta-document-information}</i>	
Title	Metadata in ICE documents
Author Name	Ian <u>Barnes</u>
Author Affiliation	ANU
Author Email	<a href="mailto:Ian.Barnes@anu.edu.au"><u>Ian.Barnes@anu.edu.au</u></a>

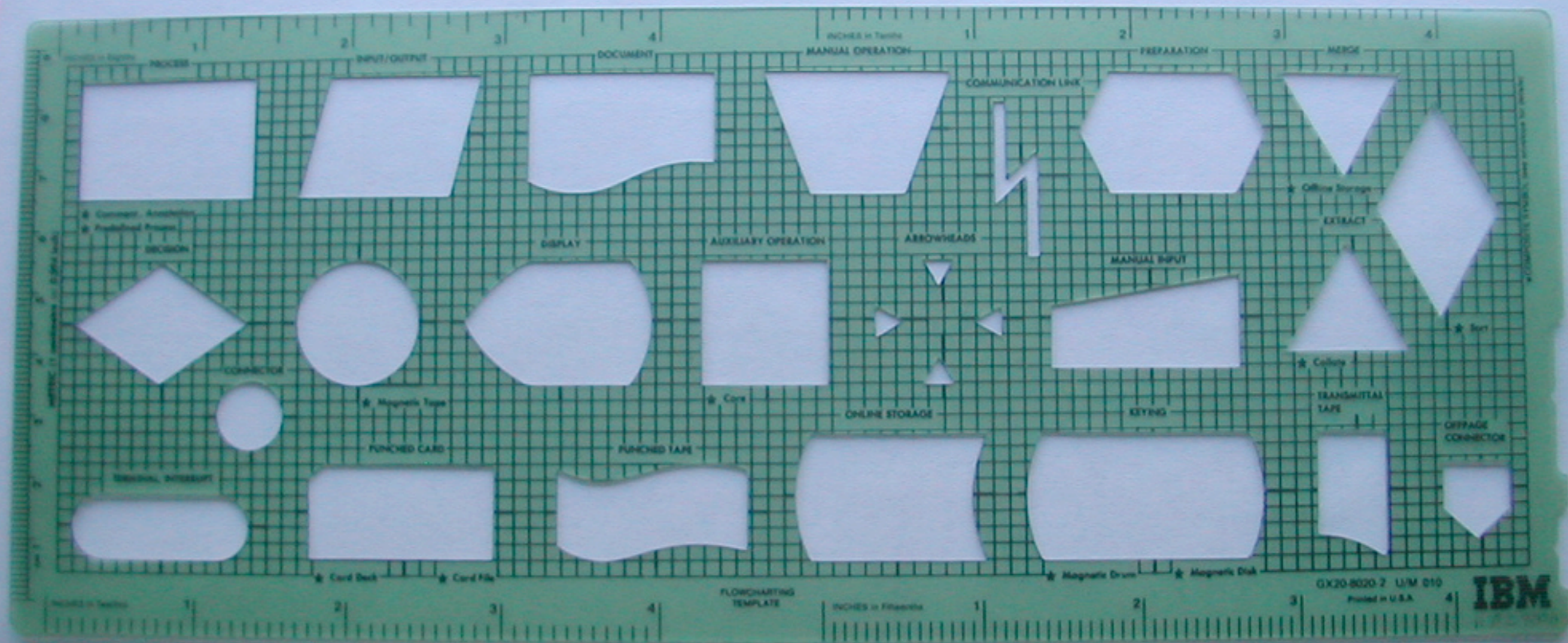
```
{ 'title': ['Metadata in ICE documents'],  
  'author': [{ 'name': 'Ian Barnes',  
                'affiliation': 'ANU' },  
              { 'name': 'Peter Sefton',  
                'affiliation': 'USQ' }  
            ]  
}
```

Tables are also useful since the layout implies semantics

# Toolbars



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.



<http://www.flickr.com/photos/jima/460348206>

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.

The screenshot shows a Mozilla Firefox browser window with the address bar displaying 'adding\_metada\_in\_wp\_docs.dc'. The browser's bookmark bar is visible, showing several bookmarks including 'Amazon...', 'Google ...', 'EAHIL-...', 'htt...dc', and 'ICE file ...'. The main content area displays the following Dublin Core metadata:

```
- <dc:title>
  Embedding metadata and other semantics in word processing documents
</dc:title>
<dcterms:issued>July 2008</dcterms:issued>
- <dc:description>
  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]
</dc:description>
<dc:creator>Peter Sefton</dc:creator>
- <dc:contributor>
  Australian Digital Futures Institute, University of Southern Queensland
</dc:contributor>
<dc:creator>Ian Barnes</dc:creator>
```

The status bar at the bottom of the browser window shows 'Done' and the Zotero extension icon.

Dublin Core metadata can be extracted directly from the document.

The screenshot shows a Mozilla Firefox browser window with the address bar containing the URL `a/embedding_metada_in_wp_docs.rdf`. The browser's bookmark bar is visible with several open tabs. The main content area displays the following RDF metadata:

```
</rdf:Description>
- <rdf:Description rdf:about="http://www.openarchives.org/ore/terms/ResourceMap">
  <rdfs1:label>ResourceMap</rdfs1:label>
  <rdfs1:isDefinedBy rdf:resource="http://www.openarchives.org/ore/terms/" />
</rdf:Description>
- <rdf:Description rdf:about="http://139.86.13.177:80/rep.ICE-Research/papers/embedding_metdata
/embedding_metada_in_wp_docs.rdf#aggregation">
  - <dc:title>
    Embedding metadata and other semantics in word processing documents
  </dc:title>
  <dc:format>application/rdf+xml</dc:format>
  <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/Aggregation" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/default.css" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/inactive_previous.gif" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/slideicon.gif" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/up.gif" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/inactive_up.gif" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/pdf.gif" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/previous.gif" />
  <ore:aggregates rdf:resource="http://139.86.13.177:80/rep.ICE-Research/skin/next.gif" />
```

The status bar at the bottom of the browser window shows the word "Done" on the left and the Zotero logo on the right.

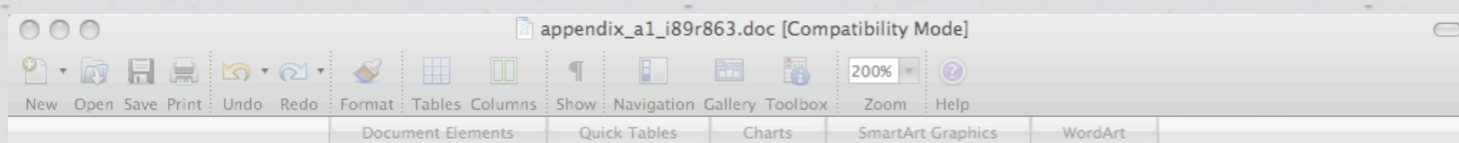
As can RDF metadata using the ORE vocabulary.

# ICE-TheOREM

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

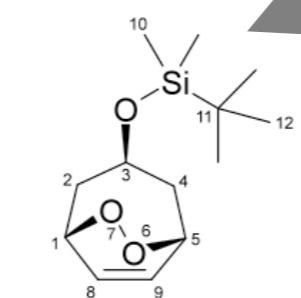
The logo for JISC (Joint Information Systems Committee) is displayed in orange text within a white rounded rectangular box.

# Chemistry



( $\delta_C = 11$  ppm, triplet) or  $C_6H_6$  ( $\delta_C = 128.0$  ppm) was used as the internal reference. The spectra were assigned as fully as possible using a variety of 2-D and nOe techniques. High resolution mass spectra were obtained on a Waters LCT Premier spectrometer with Micromass MS software using electrospray ionisation (+ESI).

## A1.2 Experimental Data for Section 2.1



54

(1,1-dimethylethyl)[(1R,3r,5S)-6,7-dioxabicyclo[3.2.2]non-8-en-3-yloxy]dimethylsilane

5,10,15,20-tetraphenylporphin (8 mg, 0.013 mmol) was added to a solution of diene **61** (300 mg, 1.33 mmol) in dichloromethane (21 mL). Oxygen was bubbled through the deep purple solution whilst it was irradiated with a 500 Watt tungsten-halogen lamp and partially cooled by

Style: p-exptl-compound

Link to data.

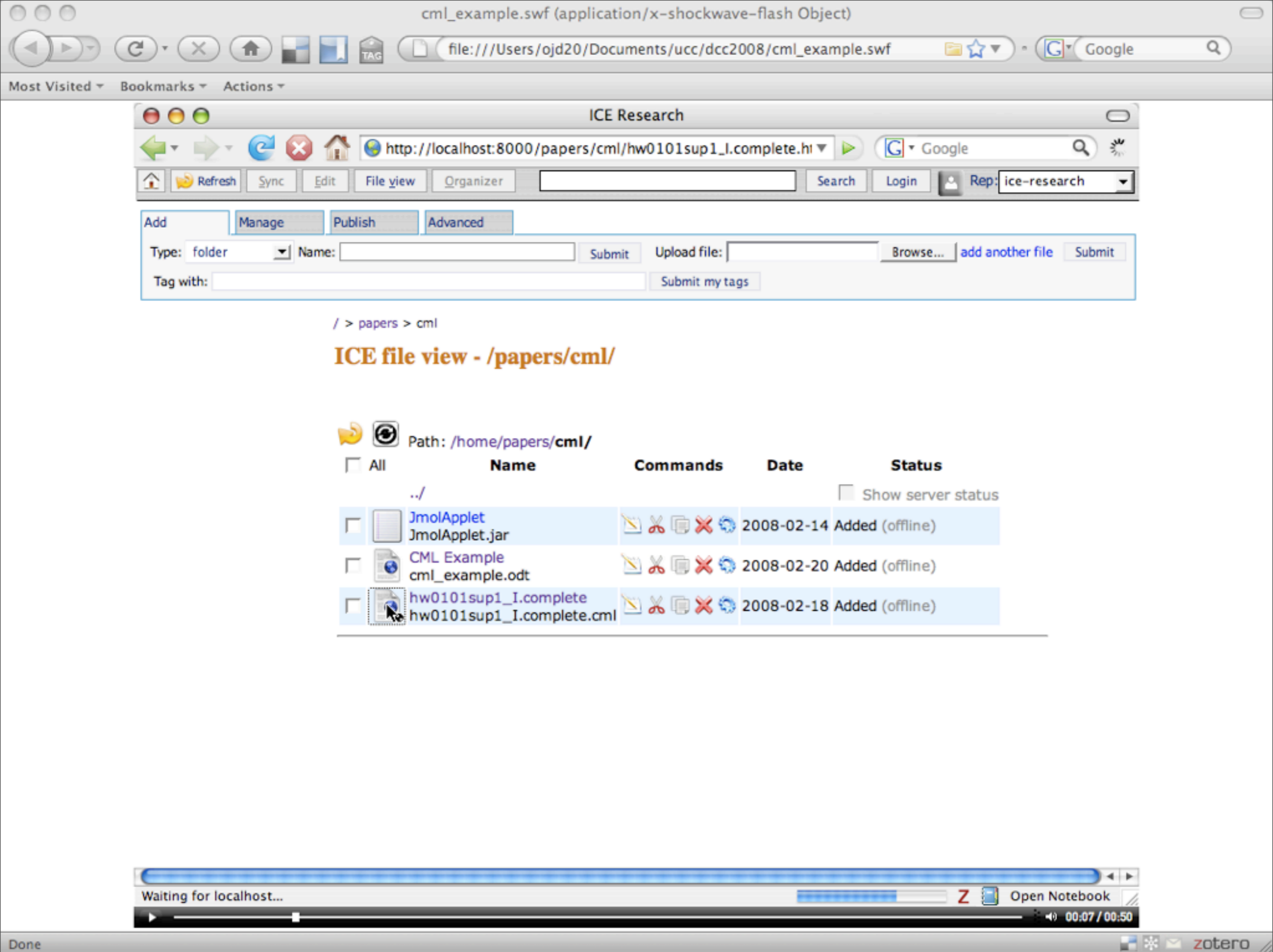
Style: p-exptl-compnum

Style: p-compound-name

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>

cml\_example.swf (application/x-shockwave-flash Object)

file:///Users/ojd20/Documents/ucc/dcc2008/cml\_example.swf

Most Visited ▾ Bookmarks ▾ Actions ▾

ICE Research

http://localhost:8000/papers/cml/hw0101sup1\_l.complete.htm

Search Login Rep: ice-research

Publish Advanced

Submit Upload file: Browse... add another file Submit

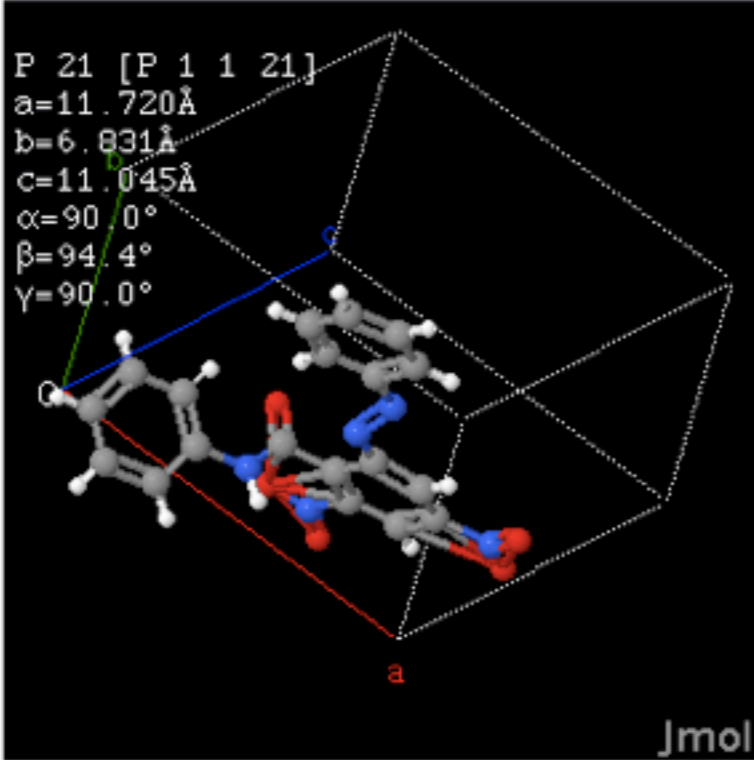
Submit my tags

/ > papers > cml > hw0101sup1\_l.complete.htm

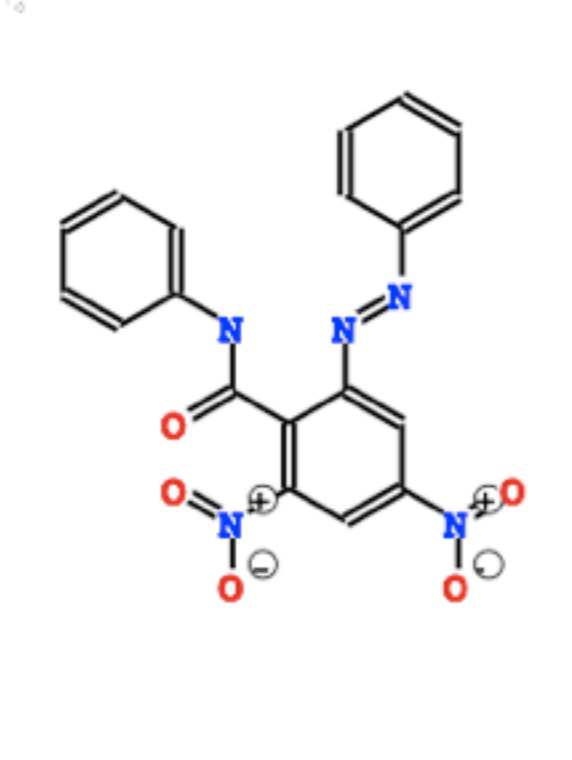
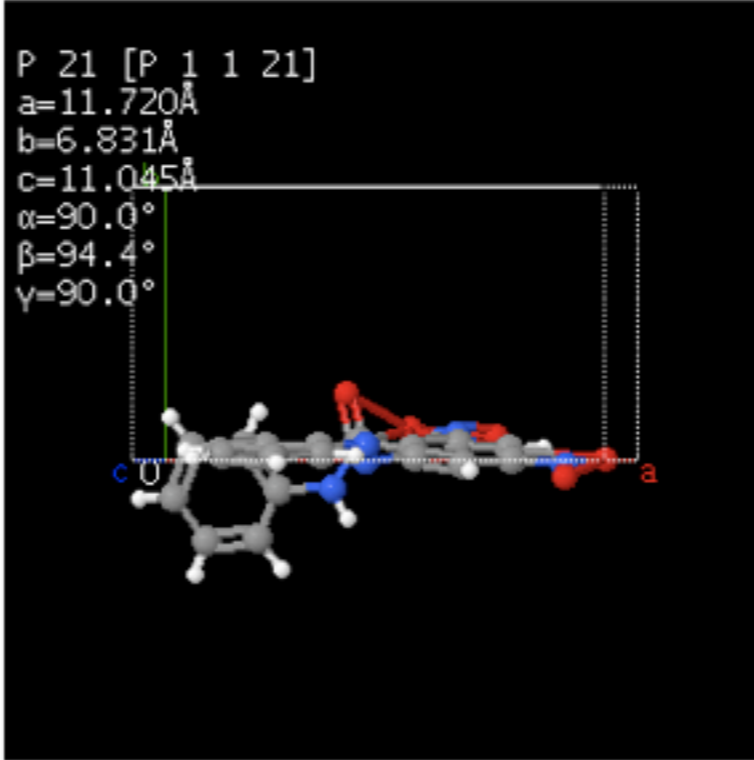
### ICE Research

Interactive Preview (PNG) Preview (SVG)

P 21 [P 1 1 21]  
a=11.720Å  
b=6.831Å  
c=11.045Å  
α=90.0°  
β=94.4°  
γ=90.0°



Jmol



Jmol script terminated

Open Notebook

00:16 / 00:50

Done

zotero

These screenshots taken from CML in ICE demo at <http://ice.usq.edu.au/presentations/demos/index.htm>

cml\_example.swf (application/x-shockwave-flash Object)

file:///Users/ojd20/Documents/ucc/dcc2008/cml\_example.swf

Most Visited ▾ Bookmarks ▾ Actions ▾

cml\_example.odt - NeoOffice Writer

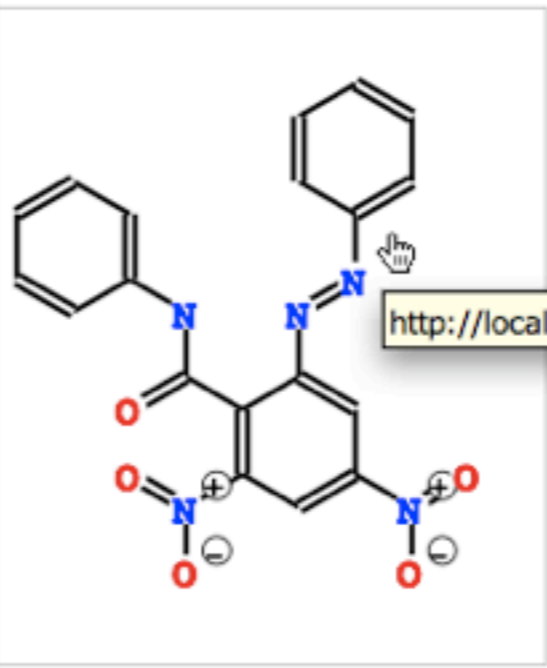
Default Times New Roman 11

CRYSTALEYE URL:

3 2 1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

### 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.



[http://localhost:8000/papers/cml/hw0101sup1\\_I.complete.htm?embed&width=300&he](http://localhost:8000/papers/cml/hw0101sup1_I.complete.htm?embed&width=300&he)

$C_{19}H_{13}N_5O_5$  (SOURCE: [Crystaleye](#))

Page 1 / 1 Default 100% INSRT STD HYP

Done Open Notebook 00:29 / 00:50

Done zotero

These screenshots taken from CML in ICE demo at <http://ice.usq.edu.au/presentations/demos/index.htm>

cml\_example.swf (application/x-shockwave-flash Object)

file:///Users/ojd20/Documents/ucc/dcc2008/cml\_example.swf

Most Visited Bookmarks Actions

CML Example

http://localhost:8000/papers/cml/cml\_example.htm

Refresh Sync Edit File view Organizer Search Login Rep: ice-research

Add Manage Publish Advanced

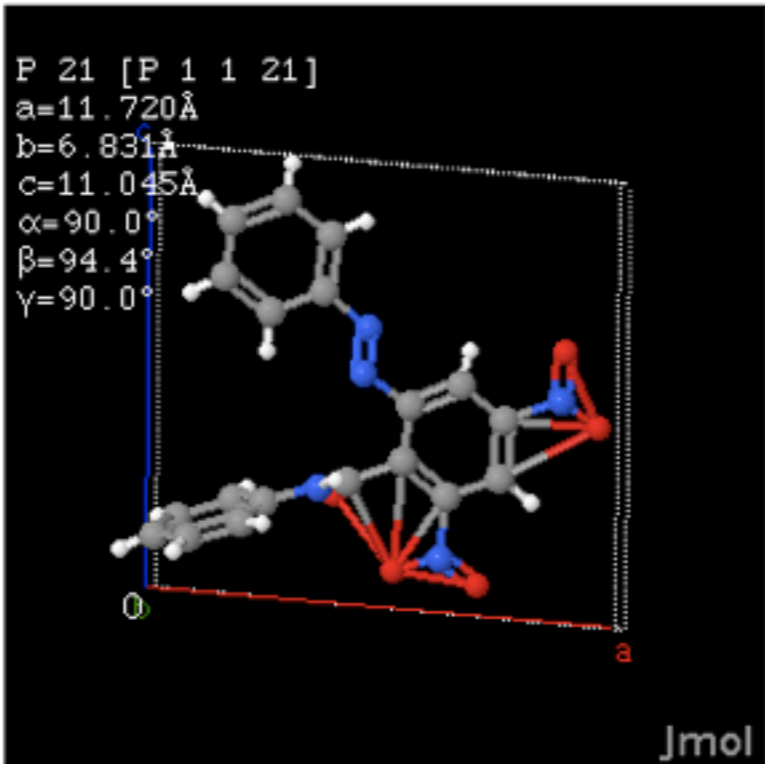
Type: folder Name: Submit Upload file: Browse... add another file Submit

Tag with: Submit my tags

/ > papers > cml > cml\_example.htm > CML Example

## 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.



```
P 21 [P 1 1 21]
a=11.720Å
b=6.831Å
c=11.045Å
α=90.0°
β=94.4°
γ=90.0°
```

Jmol

C19H13N5O5 (SOURCE: Crystaleye)

Jmol script terminated

Open Notebook 00:40 / 00:50

Done zotero

These screenshots taken from CML in ICE demo at <http://ice.usq.edu.au/presentations/demos/index.htm>

cml\_example.swf (application/x-shockwave-flash Object)

file:///Users/ojd20/Documents/ucc/dcc2008/cml\_example.swf

Most Visited | Bookmarks | Actions

CML Example

cml\_example-2.pdf (1 page)

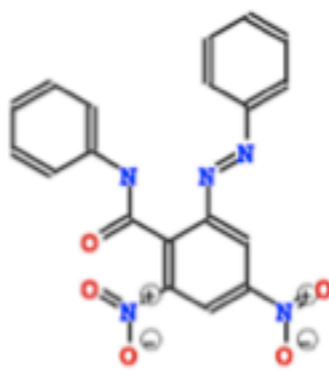
Drawer | Previous | Next | Page | Back/Forward | Zoom In | Zoom Out | Tool Mode

Add  
Type: f  
Tag with

CML Example 1

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.



C19H17N3O5 (SOURCE: Crystaleye)

the chemical dimensional

add another file | Submit

http://lo

Open Notebook 00:47 / 00:50

Done | zotero

These screenshots taken from CML in ICE demo at <http://ice.usq.edu.au/presentations/demos/index.htm>



**FIN**

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](http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/theoremice.aspx)

Peter Sefton  
[sefton@usq.edu.au](mailto:sefton@usq.edu.au)  
<http://ptsefton.com/>

Jim Downing  
[ojd20@cam.ac.uk](mailto:ojd20@cam.ac.uk)  
<http://wwmm.ch.cam.ac.uk/blogs/downing/>