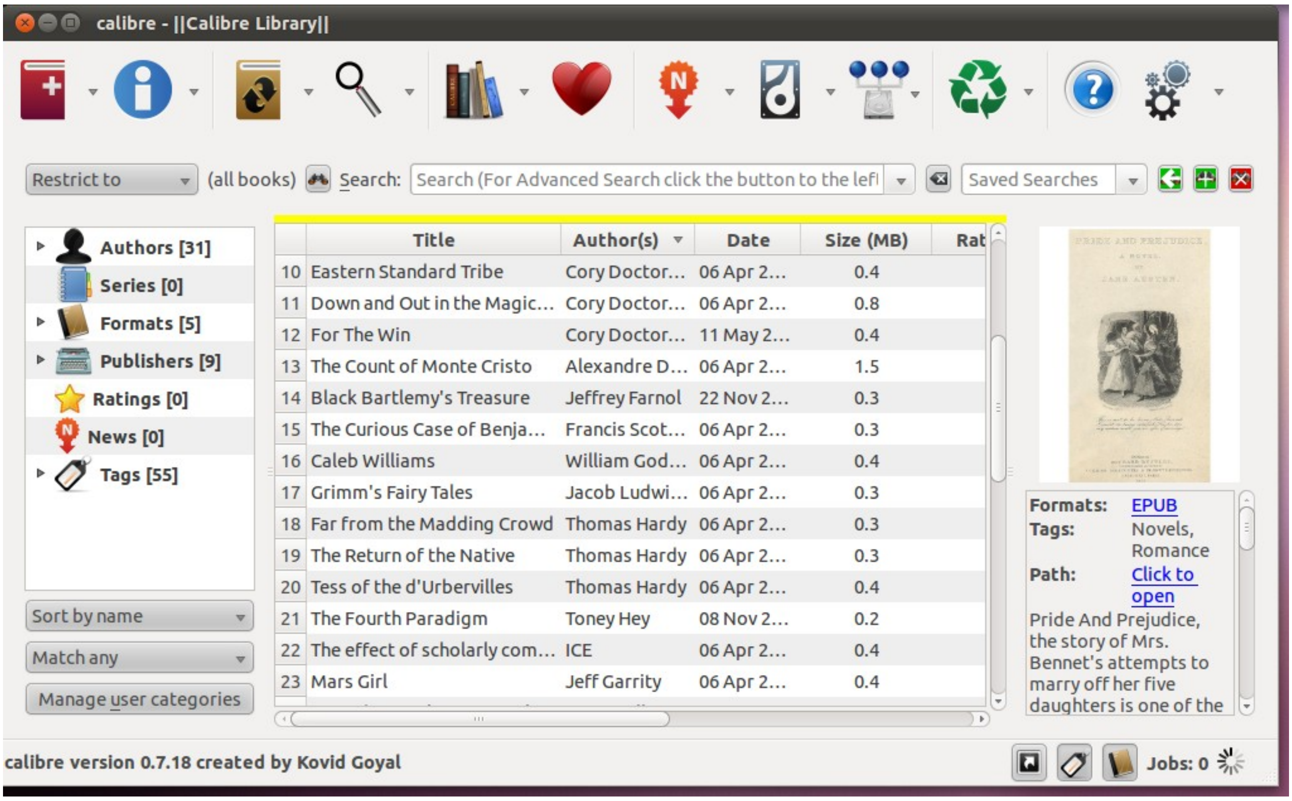
Metadata in word processing monographs

[This is a repost of a document I posted to the jiscPub blog – posting here as well to reach more people but please use the comments over there.]

Introduction – why worry about metadata?

I have been working on a simple service to take word processing documents – Word and OpenOffice.org mainly – and create mobile-readable EPUBs from them. One of the issues in this process is metadata: how do we get quality metadata into the EPUB format?

EPUB readers, like music applications use metadata to provide browse and search access to content.



	Title	Author(s)	Date	Size (MB)	Rat
10	Eastern Standard Tribe	Cory Doctor...	06 Apr 2...	0.4	
11	Down and Out in the Magic...	Cory Doctor...	06 Apr 2...	0.8	
12	For The Win	Cory Doctor...	11 May 2...	0.4	
13	The Count of Monte Cristo	Alexandre D...	06 Apr 2...	1.5	
14	Black Bartlemy's Treasure	Jeffrey Farnol	22 Nov 2...	0.3	
15	The Curious Case of Benja...	Francis Scot...	06 Apr 2...	0.3	
16	Caleb Williams	William God...	06 Apr 2...	0.4	
17	Grimm's Fairy Tales	Jacob Ludwi...	06 Apr 2...	0.3	
18	Far from the Madding Crowd	Thomas Hardy	06 Apr 2...	0.3	
19	The Return of the Native	Thomas Hardy	06 Apr 2...	0.3	
20	Tess of the d'Urbervilles	Thomas Hardy	06 Apr 2...	0.4	
21	The Fourth Paradigm	Toney Hey	08 Nov 2...	0.2	
22	The effect of scholarly com...	ICE	06 Apr 2...	0.4	
23	Mars Girl	Jeff Garrity	06 Apr 2...	0.4	

Formats: EPUB
Tags: Novels, Romance
Path: Click to open
Pride And Prejudice, the story of Mrs. Bennet's attempts to marry off her five daughters is one of the

Illustration 1: Calibre's metadata-driven management interface

Obviously, for books to be useful to readers, and to store-owners, publishers and repositories, metadata is an issue.

But it's not just for ebook delivery that this is an issue. A thesis has to be submitted for examination, and sent to an institutional repository, and maybe to a discipline repository or a publisher. And papers are often submitted to multiple sites over their lives – conference management systems, journal management systems, repositories and so on. The current state of scholarship is that every time you make such a submission you have to re-enter metadata. Upload a paper to a conference site, and chances are you will have to enter the author names into a form, even if they are already on the paper. Not to mention that every time you type in a name, you are generating low-quality string-based non-linked data. Some of us think there is a slow revolution happening in metadata, using URIs and making links.

So one of the things I would like to consider for this project is how to embed metadata within documents so that the various applications that process them can do all the hard work. And I want to think not just about strings but high-quality linked-data metadata. To discuss this I will work through one of the use cases for the jiscPUB project and look at the life-cycle of a thesis.

Thesis workflow

The aspect of workflow we're interested in here is that:

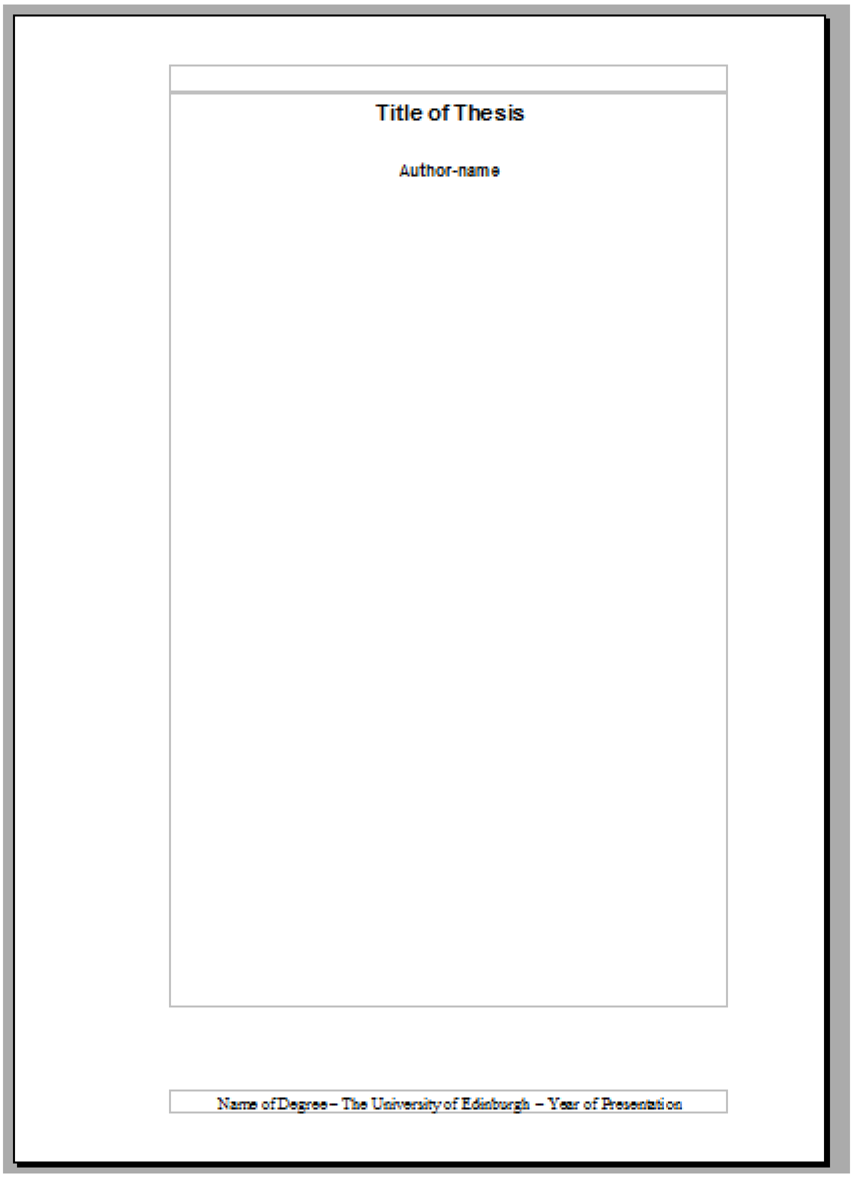
1. If the candidate is lucky the university or supervisor provides them with a template for writing up their thesis.
2. The candidate writes up the thesis and sends it to their supervisor and possibly other reviewers during this process.
3. Depending on the quality of the template there is work to do for submission, generating tables of contents, making PDF files – maybe, probably, in future, making web and mobile-ready versions.
4. Someone deposits the thesis file into (at least) the repository at the university, maybe also other databases, entering metadata about it who knows how many times.
5. Also in the future making sure all the provenance for all claims is available via data that is linked to or bundled with the thesis. (Out of scope for this post, but I will come back to it).

In this post I am going to look at 1-4 above, looking at how template design might aid in preparing a thesis for mobile delivery. I've been thinking for a few years now that the university should not just provide a template but pre-fill as much of it as possible with machine-readable metadata. And note that there's probably a much more compelling case for machine readable metadata in articles, which tend to be submitted to more places.

Thesis metadata

The university of Edinburgh, host of this jiscPUB project via EDINA, has a word template for PhD theses on its wiki. I showed in the last post that if you feed that template, sans any content, through the experimental Word to EPUB converter I've been working on, then it more or less worked, but without very much metadata (it was also dropping heading numbering, which I have now, sort-of, kinda, fixed).

To add the metadata that *should* be in the EPUB you would have to type it in somewhere. Either I could add fields to the conversion service, or you could use something like Calibre, but the thing is, most of the metadata you need is in the document – it's just not marked as such. The title page has the Title (in `AUTHOR` style) the author's name, and the name of the institution, degree and date in the footer.



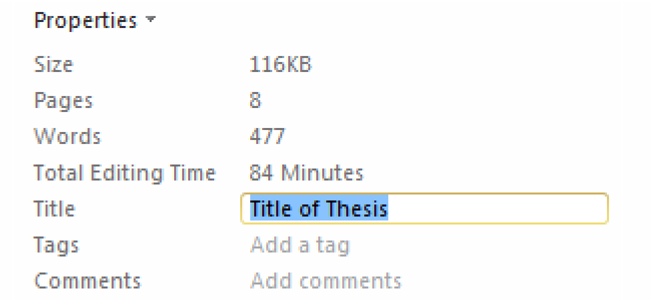
Title of Thesis

Author-name

Name of Degree – The University of Edinburgh – Year of Presentation

Illustration 2: Thesis metadata is there - in the text, just not marked as such

So – it should be possible, given that this metadata is all there to mark it in such a way that downstream processing systems can recognise it. One of the best places to start is with the document metadata fields. The Edinburgh template does use document metadata for the title.

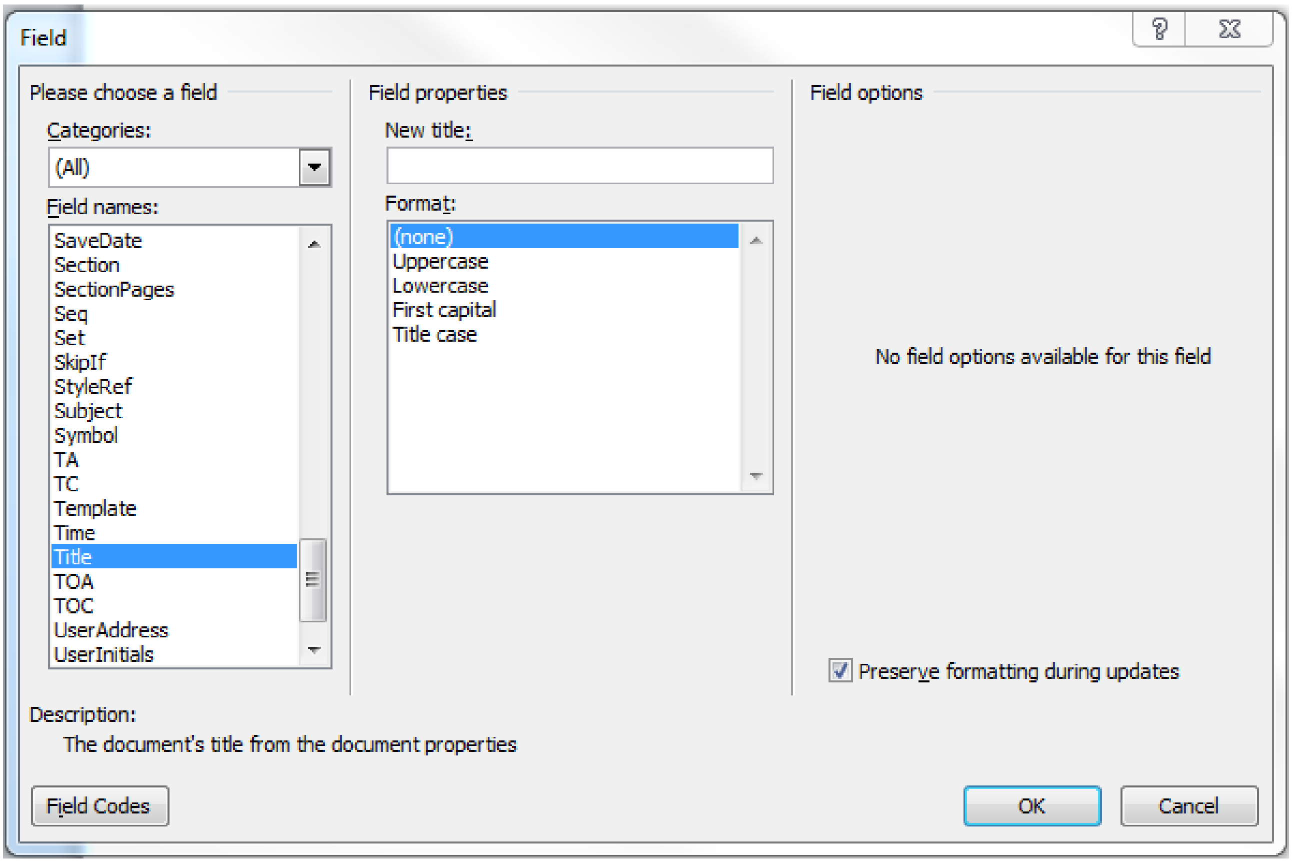


Properties ▾

Size	116KB
Pages	8
Words	477
Total Editing Time	84 Minutes
Title	Title of Thesis
Tags	Add a tag
Comments	Add comments

Illustration 3: Document metadata in Word 2010

But it could go one step further, and instead of requiring the author to enter the same thing in two places, use a field to show the title on the title page. In Word 2010 the field function is hiding in the Ribbon under Insert, Quick Parts.



Categories: (All)

Field names: Title

New title:

Format: (none)

No field options available for this field

☑ Preserve formatting during updates

Description:
The document's title from the document properties

Illustration 4: Adding a field so the title entered in the document metadata can be placed on the title page without re-typing.

Now the title is linked to the document properties, and any application, such as search engine can extract that metadata. But there is a cost – you have to be able to explain to your authors that they need to set the title in the properties, and how to do it, for the different word processing applications they're using.

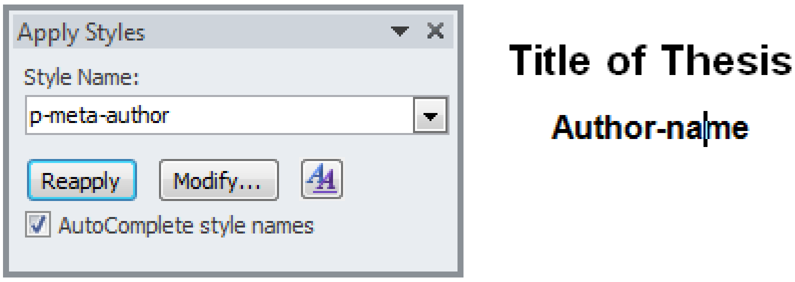
The same thing works for the author field as well. That's OK for theses but it is less useful for other kinds of scholarly content where there are often many authors. Word 2010 supports multiple authors in its metadata– but the fields don't – all you can get using a field is a semicolon separated list of authors, which is not useful for laying out the content. An approach I think is useful for scholarly templates in general is to embed the metadata in-line.

Some colleagues and I wrote up some of the approaches for embedding inline metadata for the Journal of Digital Curation[1]. The short version of that is that the most reliable cross-platform way of adding semantics like metadata in-line to documents is to use styles, or a new technique I have been developing since that work, using links. Both styles and links are supported by major word processors, so they tend to survive being loaded into different word processors or different versions of the same word processor. I will give examples of both approaches here.

Styles are fiddly to apply if you are expecting people to manage the process for themselves, but in the case of a template like this one for theses they should be robust enough – thesis candidates are not going to be changing the title page except to fill in their details. Even better – why doesn't the university do it for the candidate? I'll come back to this idea. Using tables for metadata like the one on the top of this document is also a reliable approach the metadata can be identified using style, or just text in a cell adjacent to each metadata item.

So – to demonstrate the use of styles for metadata in the Edinburgh thesis template, I:

1. Used style p-meta-author instead of AUTHOR so the ICE conversion system would recognise it.



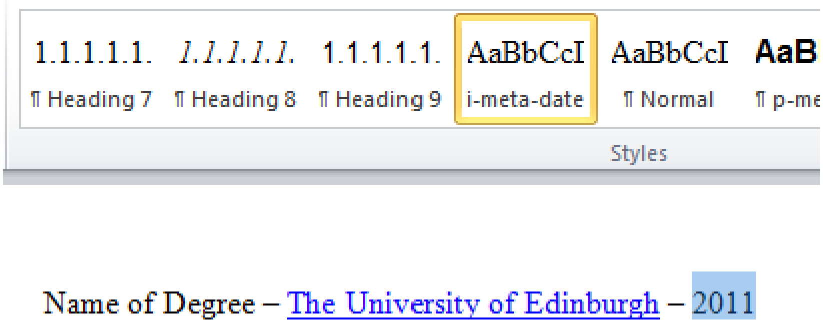
Style Name: p-meta-author

Title of Thesis

Author-name

Illustration 5: Applying the style p-meta-author the author name in the template. This dialogue box is a bit hard to find, good luck.

2. Added an inline/character style for the date i-date. [TODO: get this working or remove from post]



Name of Degree – The University of Edinburgh – 2011

Illustration 6: The inline style for the date, i-meta-date. It has no special formatting.

Getting both of these to work required a bit of hacking on ICE itself, as this metadata handling was only partially implemented.

The result is that both author and date are now included in the metadata for the EPUB file.

There is a problem with this approach, though, in that it is not giving us very high quality metadata in a linked-data sense. The author name is just a string, which as we know is not a good way to uniquely identify an author. More than one person might be identified by a string, and more than one string often identifies an author[2]. It would be much better if we could give the Author an HTTP URI. That is to name them using a URL that will be stable and unambiguous whether they are called “Name of Author” or “Author, N” or they change their name to “Nom de Plume”, which might occur as a string like “de Plume, N” or many other variants.

There's a big project coming, ORCID, which will aim to give researchers URIs, but an university could easily give each candidate a URI now, and match up with ORCID later.

I have included a demonstration of how to identify a party, the Publisher, using a URI. Here's a walk-through of a possible technique for including URIs for metadata in a template. Remember, only the template designer has to do this, not the poor candidate. And if we wanted to use this technique for personal names we could automate it and use a university-assigned URI for each candidate:

1. I chose a URI for the university: http://www.ed.ac.uk/ . Just using that as a link does not amount to metadata though. Instead I,
2. Visited http://www.ed.ac.uk/ - which redirects to http://www.ed.ac.uk/home
3. Clicked my Publisherize.me bookmarklet.
4. Copied the resulting link – which is encodes an RDF statement/triple, and wrapped it around the text in the template.

```
http://ontologize.me/?
```

```
tl_p=http://purl.org/dc/terms/publisher&triplink=http://purl.org/tr
iplink/v/0.1&tl_o=http://www.ed.ac.uk/home
```

5. Now, when documents using that template are fed through ICE, including the word-processing-to-EPUB service I have been prototyping, ICE recognises the metadata, extracts it into a data structure so it can be passed-on to Calibre, which makes the EPUB.

   ```
   ebook-convert ... --title "Title of Thesis" --authors "Author-name" --publisher "The University of
   Edinburgh (http://www.ed.ac.uk/home)" --pubdate "2011-05-01"
   ```

   But wait, there's more! ICE also embeds the metadata in the HTML it produces, like so (I did edit out some cruft that it should not be producing):

   ```
   <span rel="http://purl.org/dc/elements/1.1/publisher"
   resource="http://www.ed.ac.uk/home">

   <span property="http://xmlns.com/foaf/0.1/name"
   resource="http://www.ed.ac.uk/home">

   <a href="http://ontologize.me/?
   tl_p=http://purl.org/dc/terms/publisher&amp;triplink=http://purl.or
   g/triplink/v/0.1&amp;tl_o=http://www.ed.ac.uk/home">The University
   of Edinburgh

   </a>

   </span></span>
   ```

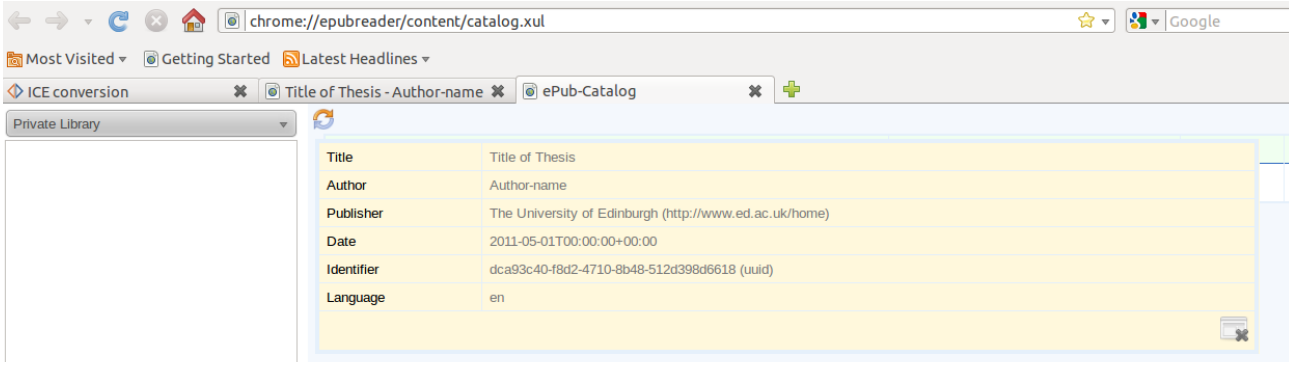
   This is intended to be compatible with RDFa 1.1, and this approach for embedding metadata in scholarly documents is on of the approaches we're promoting in the nascent Scholarly HTML movement.

Summary

In this post I have looked at three ways to embed metadata in a word processing document, so that when people *use* the template the metadata they or the template designer, enter can be machine-processed from then on.

1. Using the metadata fields in the document: good for very basic metadata like titles, but limited and not particularly interoperable for other kinds of metadata.
2. Using styles: flexible but fragile, and requires that each processing system knows about the styles you are using.
3. Using my proposed way of making linked data metadata statements encoded in links; triplinks, as seen on my demo site: http://ontologize.me. This is potentially quite robust, and could be supported by tool-chains that are much easier to use than the current half-baked infrastructure provided by yours truly.

Here's a final screenshot showing how the embedded metadata has made its way from the sample template using those three methods to the EPUB metadata, as seen in the Firefox EPUB plugin.



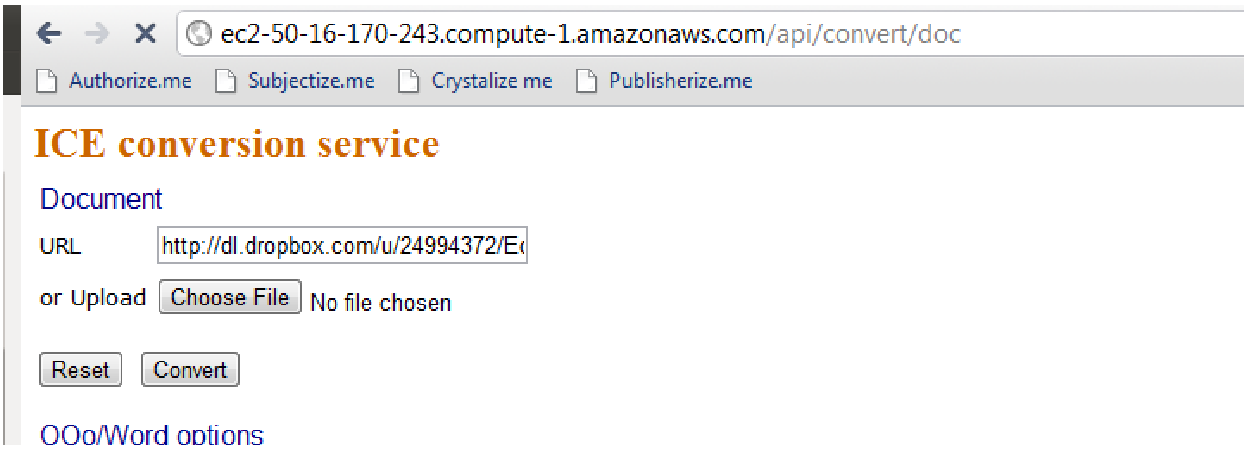
Title	Title of Thesis
Author	Author-name
Publisher	The University of Edinburgh (http://www.ed.ac.uk/home)
Date	2011-05-01T00:00:00+00:00
Identifier	dca93c40-f8d2-4710-8b48-512d398d6618 (uuid)
Language	en

Illustration 7: Metadata from the thesis template demo, in the Firefox EPUB plugin.

All three of these require that software systems know how to find and process metadata – what we're trying to achieve over at the Scholarly HTML site (when I get time to add pages on conventions for encoding metadata) is to document common ways of doing this so that tool-builders can create interoperable systems.

To try this out for yourself:

1. Go here in your browser: http://ec2-50-16-170-243.compute-1.amazonaws.com/api/convert/doc
2. Either:
   - Download this document and upload, or
   - paste in this URL: http://dl.dropbox.com/u/24994372/Edinburgh-ThesisSingleSided-plus-inline-metadata.doc and click Convert.



URL http://dl.dropbox.com/u/24994372/Ed

or Upload No file chosen

Illustration 8: Converting the test thesis doc to EPUB via a URL.

The default check-boxes at that service will make you an EPUB – if you don't have an EPUB reader you can change the file extension to .zip, open it up and have a look. If you do, you'll see something like this:

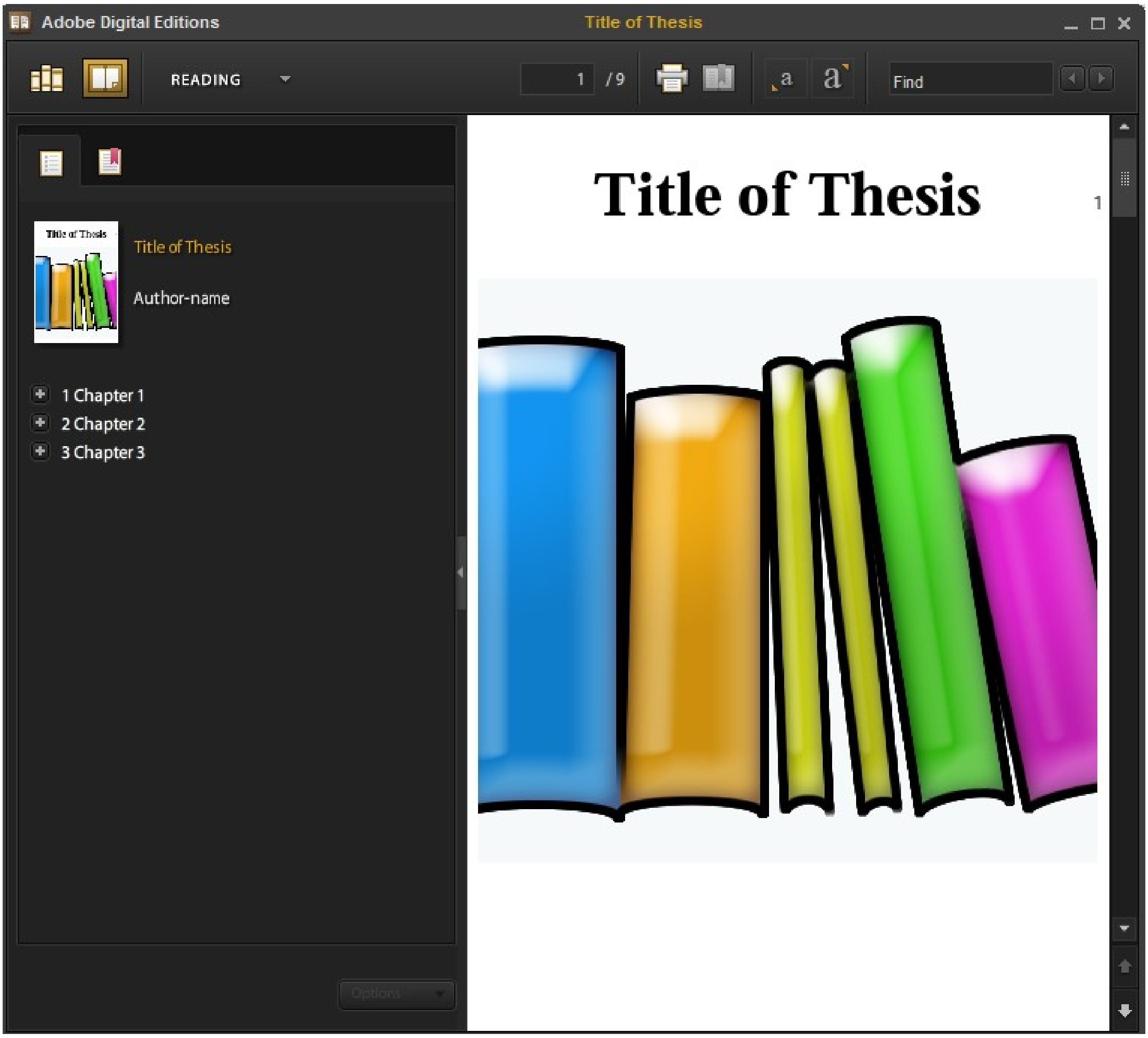


Illustration 9: Test thesis template in Adobe Digital Editions - note the title and author have been automatically extracted from the Word document.

Where now?

There's potential here to test some of this stuff out with the folks who support thesis candidates and their supervisors, or in journal templates.

- I will keep working on the Edinburgh template – to show how we might add to it in ways that increase the utility of the documents it produces, by making it easier to build ebooks. My thinking is to provide demos of what can be done for Word, OpenOffice.org/LibreOffice both using generic styles, or for people prepared to invest a little more time using the ICE styles.
- I'd love to do something with a repository – I'm thinking that it would be great to deposit theses in EPUB format – and the repository could provided a web-based reader, along the lines of IbisReader, which Liza Daly and company created. I'm looking at you, Eprints! Eprints already almost supports this, if you upload a zip file it will stash all the parts for you in a single record. All we would need would be something like this little reader my colleagues at USQ made. It would just be a matter of transforming the EPUB TOC into JSON, and loading the JavaScript into an Eprints page.

- There are improvements to be made to ICE – currently the style-based metadata does not produce Scholarly HTML / RDFa output, and is in a separate part of the code from the link-based metadata; these could be brought together.
- Is it worth adding Scholarly HTML / RDFa metadata support to Calibre so it can auto-detect metadata in HTML input?

Longer term I would like to see:

- A properly resourced end-to-end thesis project looking at how an institution could provide technical resources to candidates and supervisors, from templates, and a content, data and annotation management system . I will be showing demo service of some of this later in the project, but at the moment the demos are just toys – we need some real users and some institutional commitment to trying this stuff out.
- A journal and conference paper service where authors can write once and then submit to multiple journals. This idea comes from Timo Hannay who I met when I was in the UK – he's worked with Nature – where there is a 95%-ish rejection rate, so a service that could automatically re-work your document for you and submit would be really useful. Also sounds a bit like the Repository Junction project that Theo Andrew is involved in.

1. Sefton P, Barnes I, Ward R, Downing J. Embedding Metadata and Other Semantics in Word Processing Documents. *International Journal of Digital Curation*. 2009;4(2). Available at: http://www.ijdc.net/index.php/ijdc/article/view/121. Accessed October 22, 2009.

2. Salo D. Name Authority Control in Institutional Repositories - Cataloging & Classification Quarterly. *Cataloging & Classification QuarterlyWhere*. 2009;47(3 & 4):249 - 261. Available at: Accessed September 9, 2009.

[This is a repost of a document I posted to the jiscPub blog – posting here as well to reach more people but please use the comments over there.]

This post was written in OpenOffice.org, using templates and tools provided by the Integrated Content Environment project.