Compound documents in ICE and beyond: referencing parts of things

Ben O'Steen has put up some thoughts[1] on what he refers to as 'compound' documents and how to store them in repositories and allow for referencing of parts of a document, such as a table, a graph or even a paragraph.

Why did I add the scare quotes to *compound*?

While to a computer scientist a research paper with its graphs and tables and paragraphs might be compound, I suspect most authors tend to think of a research article as a single entity. Until we start giving them access to services that make it clear that it's not monolithic, that is.

As background, Ben gives four rules:

> Note that the four rules of the web (well, of Linked Data technically) are in essence:
>
> - give everything a name,
> - make that name a URL ...
> - which results in data about that thing,
> - and have it link to other related things.
>
> I strongly believe that applying this to the individual components of a document is a very good and useful thing.
>
> http://oxfordrepo.blogspot.com/2008/08/four-rules-of-web-and-compound.html[2]

Agreed.

He goes on to talk about repository services will have to have an explicit contract with authors that lets them know that their document is not just going to be presented in one monolithic format, by default the dreaded PDF.

> One thing first, we have to get over the legal issue of just storing and presenting a bitwise perfect copy of what an author gives us. We need to let author's know that we may present alternate versions, based on a user's demands. This actually needs to be the case for preservation and the repository needs to make it part of their submission policy to allow for format migrations, accessibility requirements and so on.

As we get authors using a system like ICE[3] then this will be:

a. Easier for them to understand because they can see multiple formats generated automatically.
b. Easy to implement, by hooking up ICE (or similar) directly to repositories. Just this week Oliver Lucido has ICE putting content straight in to ePrints via OAI-ORE – that's automatically adding an HTML and PDF view.

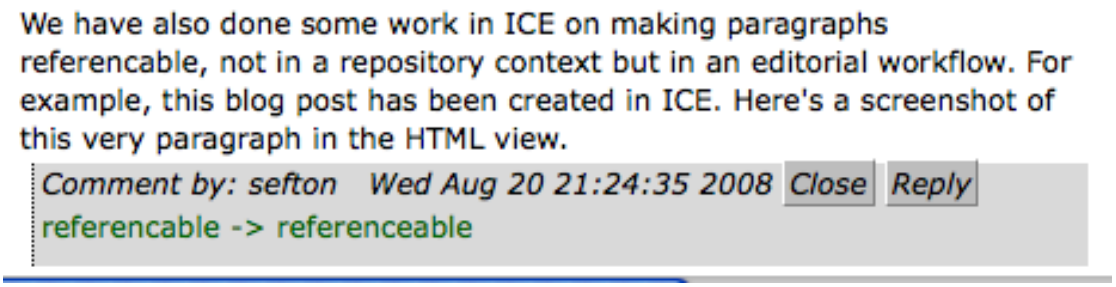
So far with ICE we have done a number of demo hook-ups to repository software. It's now time to turn this on for real – we will get ICE hooked up to USQ ePrints ASAP. This will mean that all the images in a document will automatically become referenceable. That is, in Ben's terms each image will have a name which is a URL.

Going beyond images, we have already done some work in ICE on making paragraphs referenceable, not in a repository context but in an editorial workflow. For example, this blog post has been created in ICE. Here's a screenshot of an earlier version of this very paragraph in the HTML view.

> We have also done some work in ICE on making paragraphs referencable, not in a repository context but in an editorial workflow. For example, this blog post has been created in ICE. Here's a screenshot of this very paragraph in the HTML view. ¶

See the blue pilcrow? That's the symbol that Tim Bray uses on his blog[4] to make each paragraph referenceable. Go and have a look, you can link to or refer to any part of any post on his site. In ICE, however, the plicrow is not for referencing elsewhere, it's for commenting.

See the spelling error? I can annotate the document:



Now, if I fix the paragraph, the comment will disappear from the main body of the text but the old, broken version of the paragraph is kept – it shows at the bottom of the page until I delete it.

So, ICE already knows how to identify any paragraph and has some rudimentary version control for document parts*, but the context matters.In an authoring context we needed something that was not too sensitive to document order, and it had to work with documents created by word processors, so we can't just assign unique IDs to paragraphs the way Tim Bray can in his bespoke workflow. But when it comes to pushing (or pulling) a document into a repository, where there is some expectation that it will not change, there is no reason that we can't mint IDs for parts of a document, and figure out a way to make them obviously citable along the lines of Tim's purple pilcrows.

Coming back to Ben's post. Why not make the HTML view the 'normal' way to look at an article where possible? This would mean that you don't have to store a document in fragments, merely label the parts of the HTML. I guess I'm agreeing with Ben's tentative suggestion that HTML might be a good format to hang this on:

> I have yet to settle on basing it on the content XML format inside the OpenDocument format, or on something very lightweight, using HTML elements, which would have a double benefit of being able to be sent directly to a browser to 'recreate' the document roughly.

Forget 'roughly', at least for documents created with an HTML-ready workflow like ICE. It would even less rough if authors choose something like the Article Authoring Add-in for Microsoft Office Word 2007[5]. But Ben's right; for documents that are deposited in PDF or in unstructured word processing formats then HTML is going to be rough.

Just how we might handle the user interface issues for exposing names (URLs) of the parts of a document is unresolved, but we'll give it a go here at USQ with our ICE and ePrints systems.

* There's the current version and then there are obsolete versions. ICE of course has rich version control at the document level courtesy of subversion

1 http://oxfordrepo.blogspot.com/2008/08/four-rules-of-web-and-compound.html

2 http://oxfordrepo.blogspot.com/2008/08/four-rules-of-web-and-compound.html

3 http://ice.usq.edu.au/

4 http://www.tbray.org/ongoing/When/200x/2004/05/31/PurpleAgain

5 http://ptsefton.com/2008/08/05/another-look-at-the-article-authoring-add-in-for-microsoft-office-word-2007.htm