Desktop Repositories: Smashing up PowerPoint

Les Carr has been experimenting with desktop repository services. He started by wondering how he might manage the thousands of PowerPoint slides and presentations he has, moved on to converting them into images, with embedded textual metadata, then put them in ePrints on the desktop and started speculating on how slides might be reassembled into new presentations and exported.

These workflows are exactly what we have been looking at with The Fascinator Desktop, our nascent eResearch repository platform. Our goal is to index and understand everything on an academic's desktop, including presentations, documents, video, images, audio, data of all kinds, everything; via a plugin architecture which will be easily scriptable. We're in the middle of a two week development sprint getting some of the pieces in place for this, so I thought that picking up on Les Carr's PowerPoint work would make for a good target for the end of next week.

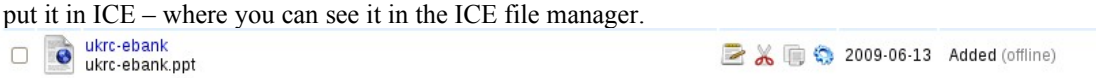
The goal is that by next week we have an automated system that can:

1. Watch your home directory.
2. Extract metadata and index it so we can construct a faceted browse interface.
3. And in particular, break all your Microsoft PowerPoint or OpenOffice.org Presenter files into a set of searchable images, just as Les has done.

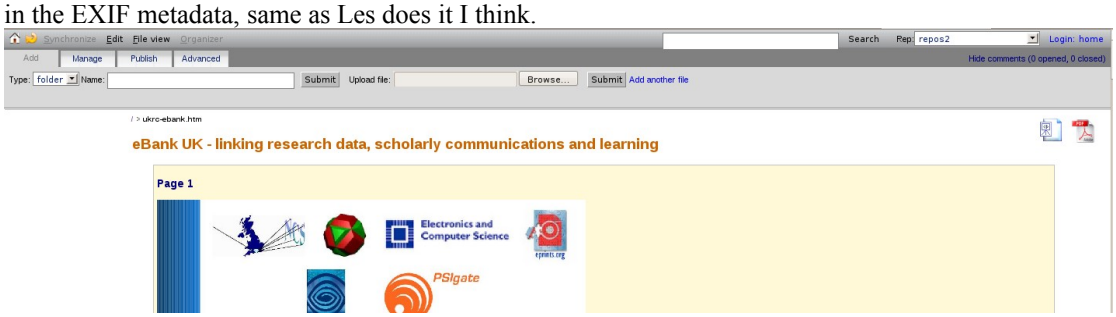
   We think we will be able to build a pretty cool interface for this, so that you can text-search for individual slides but also get a sense of their context. So if you search for `dog cat`, it will find the presentation with 'dog' on slide one and 'cat' on slide seventeen with a neat interface to show how they are related.

We have made one step towards our goal by doing something we should have done ages ago, adding PowerPoint support to ICE. It's still a bit rough, but ICE can now turn ODP and PPT into an HTML slide presentation with images of each slide embedded in the page using a method inspired directly by Les Carr's work. Here it is with a presentation by Les himself (Carr, Leslie, Coles, Simon and Lyon, Liz (2004) Archiving research data and research publications. At, *Research Councils UK Workshop on Publication of Research Results, London, UK, 18 Oct 2004*.)

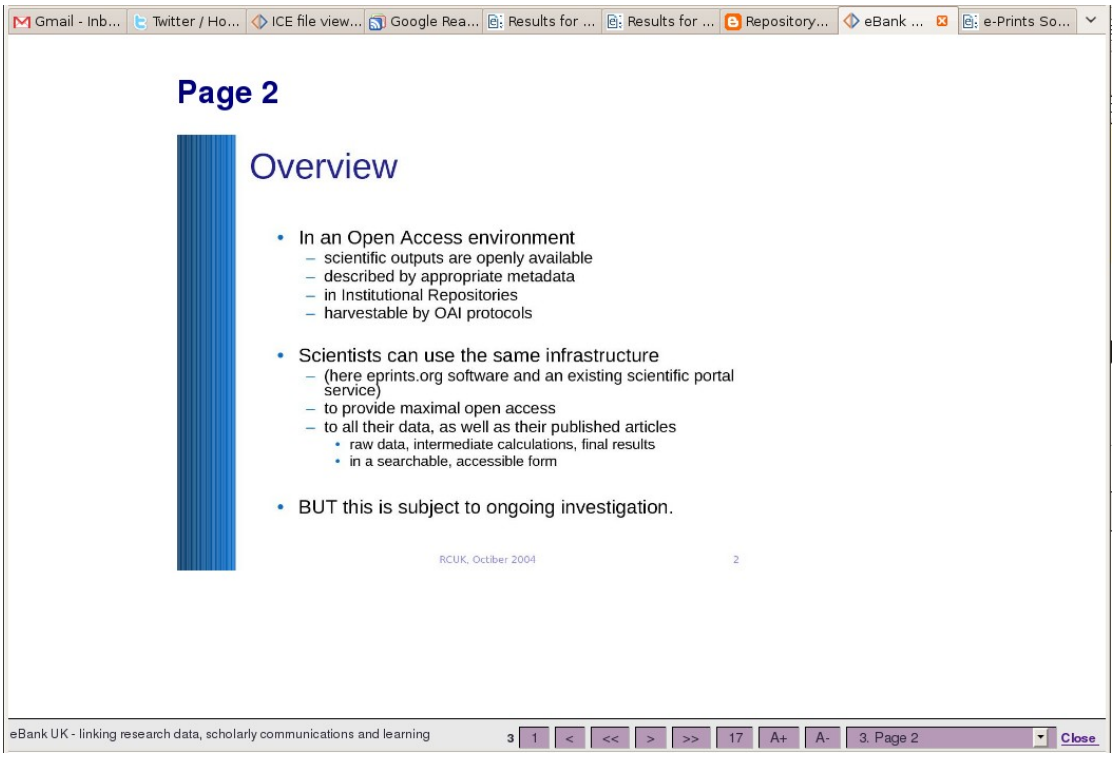
I put it in ICE – where you can see it in the ICE file manager.



I click the link and ICE renders the document into an HTML file, with an image for each slide. OK, so at the moment we're not getting the title of each slide quite right – but the text for each slide *is* sitting in the EXIF metadata, same as Les does it I think.



Now, using ICE's inbuilt presentation mode, I can press the button and get this – that's the slide with presentation controls:



This will be useful in ICE – where it will give people more options, flip through the slides in web mode, grab a PDF of the whole lot, or get the original format. But this is not the end of the story – we're going to use ICE as a conversion service, and view the slides using The Fascinator Desktop instead. That's next week.

Once we have this done we could look at services such as:

- Fill a shopping-cart with slides and or other slide-like resources and build a new presentation, via the magic of OAI-ORE aggregations. As Les seems to be saying, you could reassemble the slides from their source documents into new presentations by grabbing individual slides. We have lots of experience with this kind of work on ICE – and while it's quick to get started we know it is the sort of area where lots of maintenance is required.
- Post a presentation to a repository, in fully-exploded form[1].
- Post the presentation to a blog as a series of images with PPT and PDF download.
- Set up an automated zero-click deposit system so that if I tag the presentation in a certain way it automatically goes off to the archive. As demonstrated by the Australian team at OR08.

(If we had all that in place we could finally help Peter Murray-Rust with his presentations, which are made up of web pages selected from a huge library of un-slides many of which included embedded data visualizations. By indexing all his individual pages we could let him 'shop' for the ones he wants, order them and then create a presentation-by-reference which could be de-referenced and blogged or reposited. Peter, can you make your slide library available to us for experimentation?)

So far we have made some good progress on this goal of having a continuously updated repository view of all your files.

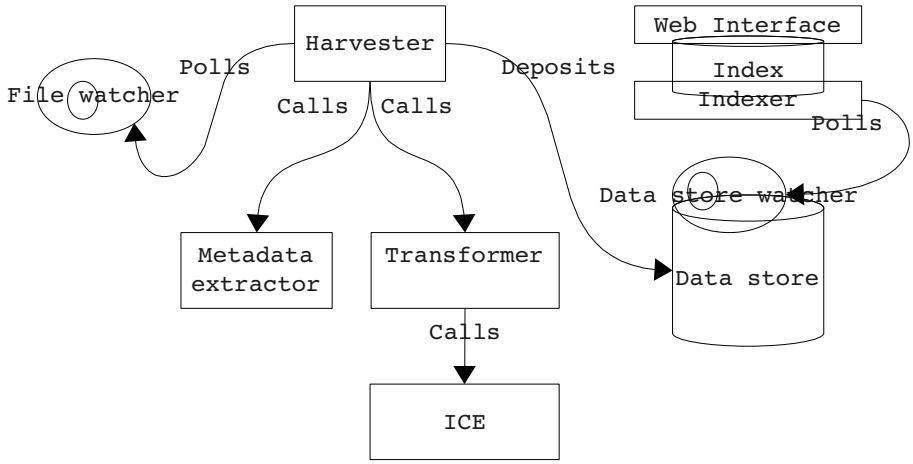
1. Oliver Lucido has built a new abstract interface on Fedora so that we can swap in other data stores. Ron Ward will be trying out Couch DB, and we'll probably build a simple repository layer using

1 If I cycle in to work this weekend and pick up the 7GB virtual machine we made I will be able to demo depositing the PowerPoint into ePrints from ICE.

Pairtree and Dflat. I guess we could use ePrints, or Zentity if we were so inclined.

2. Cynthia Wong and Linda Octalina have built slide handling into ICE as seen above. The approach taken is to load the presentation into OpenOffice.org, save as PDF, then use Imagemagick to break it into images, one per page, then look inside the .odp (ODF presentation) XML to find slide content and use ExifTool to add the slide content as metadat to each image. Lots more to do here, but even as-is it's useful.
3. Linda built a file watcher application, which is going to be cross platform but which at the moment is for Linux only. It, you know, watches your files and other services can ask it via HTTP for a list of recent changes, which it gives in a simple JSON format with RDF inside.
4. Duncan Dickinson and Bron Chandler have been building a harvester; the bit that sits in between the data store where we will be storing all these slide images and so on, and the file watcher which will notice when you add a new PowerPoint, or any other file to the system:
    i. A metadata extractor: This will be based around an bit of Software called Aperture which can extract text and RDF metadata from all sorts of files, such as images and PDFs. It puts the metadata in the data store to be indexed.
    ii. A transformer application which can render PPT to HTML/images, render word processing documents to HTML, generate low-res video from master files and so on. We already have lots of file transformers in ICE so the transformer will be able to call ICE if it needs to. (Some of this is slow, so we have plans to move this to another queue so it doesn't slow down the main harvester). The renditions go in the data store.

A lot of this is similar to Jim Downing's Lensfield project – we have talked about harmonizing our projects.



This may look like a lot of stuff, but we think that it will provide a very flexible platform for doing useful work for academia, discovering and routing files of all kinds. While parts of this are connected by HTTP calls we are prepared to optimize if that proves too slow, but we think on a desktop scale this will probably work alright. What's missing from that diagram are all the other things you will be able

to do via the web interface – tag stuff, label with