Desktop eResearch revolution

It seems to me that there is a bit of buzz at the moment around the need for a desktop eResearch tool that can organize your stuff locally and push it up to a managed store.

In no particular order.

There was a conversation between a couple of Sydneysiders on Twitter on the #eresearch channel:

> usyd_dpa: @jimrhiz The research groups I work with have schemas and taxonomies. either domain- or project specific (e.g. HISPID, VRA). **#eresearch**
>
> 1 day ago from TweetDeck · Reply · View Tweet ·
>
> jimrhiz: @usyd_dpa For taxonomy tool, is NSDL Registry http://metadataregistry.org/ of any relevance? Need to consider sustainability! #eresearch
>
> 1 day ago from twitterrific · Reply · View Tweet
>
> jimrhiz: @usyd_dpa Could also look at @iand 's Open Vocab http://is.gd/likA (expand)
>
> 2 minutes later from web · Reply · View Tweet
>
> usyd_dpa: @jimrhiz The research groups I work with have schemas and taxonomies. either domain- or project specific (e.g. HISPID, VRA). #eresearch
>
> 5 minutes later from TweetDeck · Reply · View Tweet
>
> usyd_dpa: @jimrhiz #eresearch Different disciplines, common need for tools to enable discipline - specific metada & taxonomy creation / maintenence
>
> 2 minutes later from TweetDeck · Reply · View Tweet
>
> usyd_dpa: @jimrhiz #eresearch ... maybe like joomla or plone with a taxon plugin? - so long as its robust, usable, adaptable, export struct. packages
>
> 5 minutes later from TweetDeck · Reply · View Tweet
>
> usyd_dpa: @jimrhiz #eresearch This sort of tool wouldn't even need a flash front end. Index, search, presentation could be handled elsewhere.
>
> 11 minutes later from TweetDeck · Reply · View Tweet
>
> jimrhiz: @usyd_dpa To spare #eresearch , maybe this discussion needs its own tag, say #taxontools
>
> 5 minutes later from twitterrific · Reply · View Tweet
>
> jimrhiz: @usyd_dpa Are these links from @maheshcr any use: http://is.gd/lir2 (expand) ? What about Protégé from Stanford? #taxontools
>
> 3 minutes later from web · Reply · View Tweet

Subsequently, Rowan Brownlee (usyd_dpa) has started a conversation on the ANDS group asking about tools for researchers to organize their stuff, label it using taxonomies and share it. So far no response to that. I would have expected someone to mention Field Helper, which is from Sydney.

> Field Helper is a desktop application that enables you to quickly view and categorise groups of related digital files and then submit the resulting package to a repository for long term preservation and access. Digital repositories require a submission to be formated in a specific way and be described according to a standard meta data encoding schema. Working with Field Helper results in a ZIP file containing compressed versions of your files along with a METS (Metadata Encoding

> and Transmission Standard) file which contains a detailed description of each file and its relationship to other files in the submission. METS is a standard that works with most repositories and - where required - can be easily translated into a form that non METS compliant repositories can work with.

I have looked at Field Helper in the past – I don't think that the metadata tagging is going to scale very well and the system for mapping tags onto formal metadata seems a bit clumsy but it does some of what I think Rowan is asking for.

Also in the last few days, Les Carr mourns another lost disk drive and other lost week of work, and tells us why he needs trusted storage. He says:

> So an intelligent store should help me understand what I have - a bit like the way that user tools like iPhoto help you understand and organise thousands of images. It should be possible to get a highly distilled overview/representation/summary/visualisation of all my intellectual content/property/achievements as well as a detailed and comprehensive store of all my individual documents and files.
> I guess you can see where I'm going with this. I've gone and got the ideal desktop storage and the dream repository all mixed up. Well perhaps I have - but why not?

Yes – I can see where Les is going – at least I think he's hinting at ePrints on the desktop.

I made the same sort of mixup in my mind back in December:

> Thinking about this led me to the idea of putting something like The Fascinator on the desktop, letting it find all your stuff, giving you a simple way to organize it into projects, embargo bits of it and so on, and then automate the process of disseminating it to the institutional and other places you'd like it go. I'm thinking of something like Picasa (which finds all your pictures on your hard drive no matter how embarrassing or not safe for work they are) and iTunes which although in my opinion potentially evil has some nice ways of browsing and organizing content, but with a connection to the world wide repository grid. More on this idea soon.

And finally, we have Dorothea Salo with this one liner. There are a lot of other lines you can read as well in her response to this piece at Library Journal dot com.

> Data curation and IR population need to be reframed as *collection-development challenges*.

Now, maybe Rowan's plea to the ANDS group will turn up an application that does what we want, but I think we might have to take up the *collection-development-challenge*.

Dorothea says:

> Bluntly, DSpace and EPrints are completely inadequate to meet the data-curation challenges you [Clifford Lynch] outline; and Fedora can mostly do the job, but only with major hacking. This is unacceptable. How can we offer data services when we don't have basic building-blocks to work with?

Here at the Australian Digital Futures Institute (ADFI) we're up for a bit of 'major hacking', although we find it soothes our management team to call it 'software development'. That's why we pretend that I'm the manager of the Software Development Research and Development team, not just the one with the biggest mouth in a feral mob of hackers.

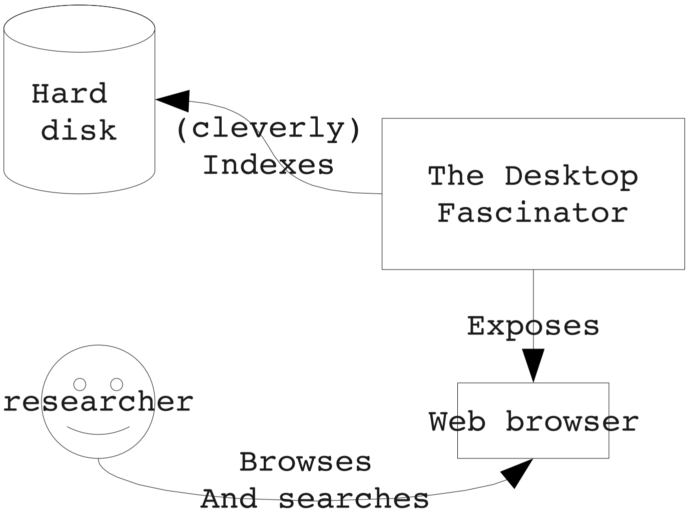
At the moment we are working with Chris Lee's Public Memory Research Centre on a repository for the humanities. Ultimately it will have creative arts content and research materials – we're starting with a military history project run by Leonie Jones.

Leonie has been doing exactly what Rowan describes: "They each manage binary and text content on their desktops or departmental fileservers using spreadsheets and/or sql databases". Leonie has spreadsheets.

We are *considering* trying the following based on our Fedora-based lightweight repository solution known as The Fascinator.

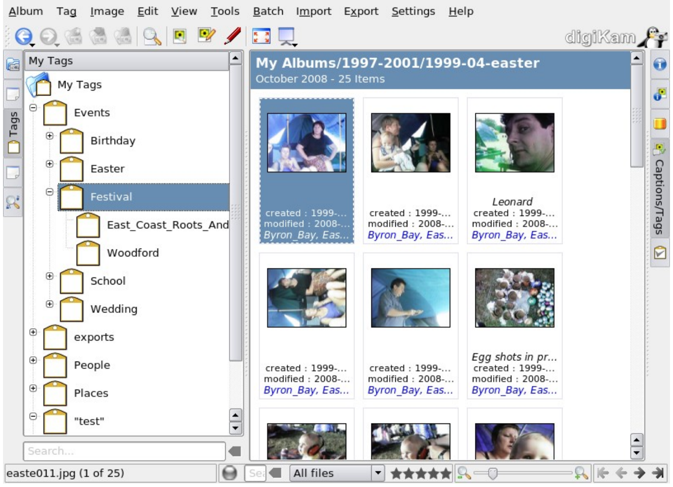
We'll start with a local installation of The Fascinator – that puts Fedora 3 and Apache Solr on your desktop. Don't worry, we have a simple installer. It's all Java, so it might be painful for the programmers at times but it should install pretty much anywhere.

Then we will add a file-system indexer for The Fascinator – pretty much like what Picasa does, it will index *all* of your stuff. It will grab whatever metadata it can, including properties from office documents, EXIF metadata and tags from images . We will also treat the file system as a source of metadata so you will be able to explore using metadata facets and file system facets using the same interface. This should be a very straightforward addition to the existing software, it's just a matter of bolting together some standard software libraries.

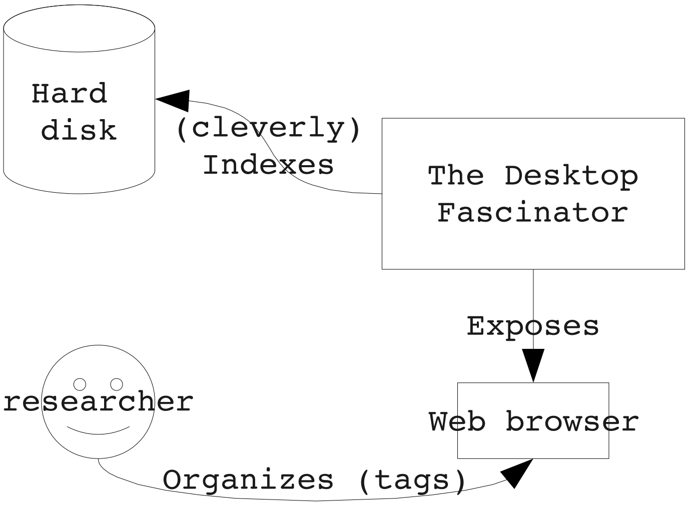


Next comes the taxonomy/tagging bit: we need a way to import tag-sets and taxonomies that you might want to apply to your content and then let you tag it. I think it will be important to support both formal metadata and informal tagging. For example, you might want to set up your own tag hierarchy with home/work at the root, and with work broken down into teaching/research and research broken up by project.

I think the tag hierarchy in digiKam is a good start. Here's a screenshot showing my home-grown tag set applied to my own photos. I think a new tool should allow both ad hoc DIY sets and more formal ontologies.



So we can add a new action for the user:



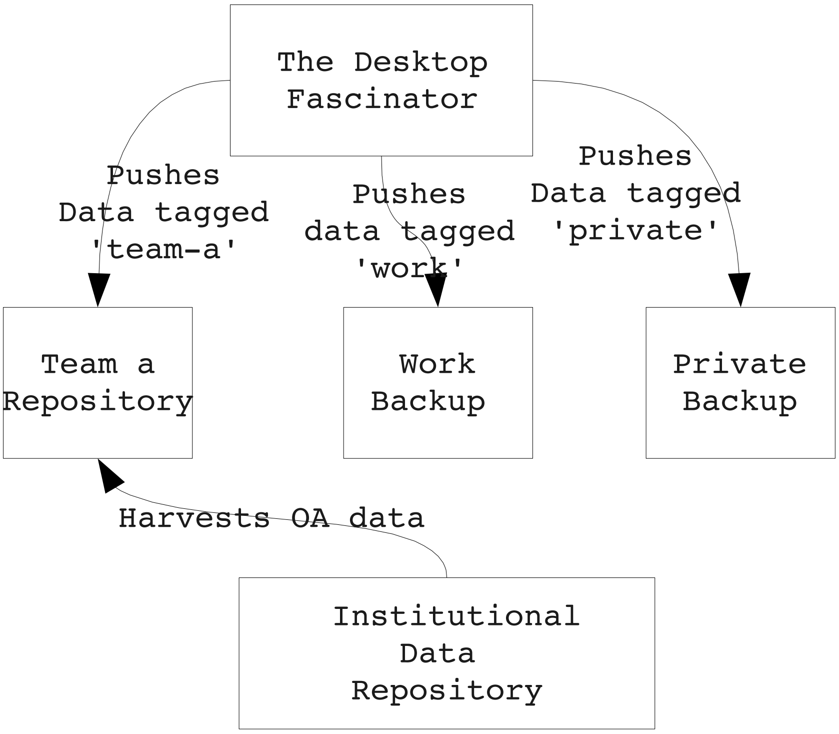
There are a couple of things we'd like to look at here:

1. There is often metadata inherent in the file system. I can point to my Music folder and say 'that's all owned by me'.
2. There are relationships between files; Leonie has video transcripts with time-codes. If we plug in a smart indexer then we should be able to get our text index to let you find words and jump to the right part of the video. So, we need plugins – which may often have to be one-offs. Ben O'Steen has a great blog post where he talks us through one such curation exercise.
3. Which bits of metadata should be written back into the files? For my own images I have been adamant that I want metadata written back into the files, but what if different members of the family wanted to classify things in different ways? Stand-offish metadata may be better if we can build trusted systems that know how to keep things linked up.
4. Add 'playlists' to group content. This is a actually just like what you do in content packaging, for example the organizer in an IMS content package.

At this point in the story we can index everything, label it, and explore it. Next step would be replicating it up into a cloud of repository services. I can imagine a couple of use cases here:

- Everything with the tag `work` is to be backed up to the university system. This would be a mirror of what's on my desk and by default only accessible to me. There's going to be a lot of data and we can't leave it all in our houses offices and labs.
- Everything under `/Music` is to be backed up to a private data store – maybe via another copy of The Fascinator running at home, or even a copy in the cloud somewhere.
- Everything tagged with ePrints goes to you-know-where (Hi Les).
- Everything tagged with `thesis` is to be replicated to the departmental thesis repository where my supervisor will be able to see it as well.
- Everything tagged with PMRC goes to the centre's repository where the repository's curator can, you know, curate it. This could be as simple as adding a tag `public`, that means that it will then be disseminated to the public institutional repository.

To get this kind of data federation going I'd look at Atom Archive for the feed mechanism (that's going to kill off OAI-PMH, right?) with OAI-ORE to package the metadata we've added with the source-objects.


The Desktop Fascinator
Pushes Data tagged 'team-a'
Pushes data tagged 'work'
Pushes Data tagged 'private'
Team a Repository
Work Backup
Private Backup
Harvests OA data
Institutional Data Repository

One feature of this setup which might not be immediately obvious is having the same kind of repository interface on the desk as you would have in a web based repository. The idea is to encourage people to see their research data as 'in the repository' from the moment of creation and to be able to take control over how their stuff is disseminated. This is a like the ICE approach of previewing early and often so that people get used to seeing their documents both in paper format and web format.

There are some free desktop applications that do *some* of what I'm talking about here. Duncan Dickinson is collecting a list of them. They include the aforementioned Field Helper and things like

Mendeley which manages research papers, but not, it seems research data.

And lets be clear here. We have never, ever been under the illusion that if we build it they will come, not with our ePrints system, not with ICE, not with any system. We know that if we decide to build then we will have to build something that either the users need, or didn't know they needed. In this case I am betting the first selling point is the backup feature but funder requirements to place data in repositories may be a motivator in the distant future.

Should we try building this thing? It would only take a few weeks to prototype.