

Creating an eResearch Desktop for the Humanities

Duncan Dickinson¹, Peter Sefton²

¹University of Southern Queensland, Toowoomba, Australia, duncan.dickinson@usq.edu.au

²University of Southern Queensland, Toowoomba, Australia, peter.sefton@usq.edu.au

INTRODUCTION

The Australian Digital Futures Institute (ADFI) at the University of Southern Queensland (USQ) has been working with USQ's Public Memory Research Centre (PMRC) to discover how public memory research can be enhanced through desktop eResearch software that assists the researcher in connecting the components within their "Knowledge Network" of files, online resources, metadata, research notes and publications. Leonie Jones from the PMRC has worked with Australian and Vietnamese servicemen to research the events of the Battle of Fire Support Base Coral to provide a point of reference in the ongoing narrative of Australian military history. This battle constituted the largest fought by Australian troops in the Vietnam War. "They'll come looking for you"¹ is a 60-minute documentary that summarizes Jones' research and is the tip of the iceberg in terms of the artifacts brought together through this work which constitutes a wide array of public memory artifacts, including: video interviews, official texts, private diaries, photos and maps.

This partnership seeks to provide solutions for a range of questions encountered within the digital humanities and the broader eResearch effort. (a) How to manage large amounts of information, including the very fundamental "Is it backed-up" (b) How to coordinate tools used on the data; (c) How to leverage concepts such as Linked Data without obstructing the researcher with technical intricacies; and (d) How to move these research artifacts to the data commons. The key to all of this is to treat everything as a web-resource from the moment it is created by developing a system that gives researchers a web-view of their own desktop or lab environment. The presentation will introduce the desktop eResearch system, The Fascinator Desktop, being developed by ADFI and will demonstrate solutions to the questions posed above. Core to this demonstration will be: data and metadata storage, sharing and conversion technologies; taxonomy-based tagging; and automated object grouping.

SYSTEM OVERVIEW

The Fascinator Desktop has been developed to bring the web to the researcher's desktop. The central idea is to provide a web view of data from disparate sources from the moment data are created, so that the web becomes a natural way to work with data from the very start, rather than a deposit challenge at the end of a project. The aim is to enhance, rather than replace, existing software. To this end, the project seeks **not** to provide a monolithic solution to everything from data analysis to video editing. Rather, the software assists the researcher to combine and describe various sources via a rich browser-based web interface that allows researchers to locate, combine and organize their information in ways that free it from the hierarchical file system, and particular applications allowing them to follow Linked Data² practices without having to know they are doing so. The system enables researchers to share collections of digital objects to repositories via the growing set of services available from the Australian National Data Service (ANDS)³, including the Register My Data⁴ and Identify My Data⁵.

RELATED WORK

The Fascinator has some similarities to various efforts to produce Personal Information Management (PIM) and Semantic desktop software⁶ as well as to desktop search tools. The closest in scope is the Nepomuk project⁷ (closed in December 2008) which provided a semantic desktop capable of collecting data from a variety of sources and utilising them within an RDF framework. From a usability standpoint, however, the system is quite technical and does not hide the inner workings of RDF from the user. The Fascinator is using components created or contributed to by the Nepomuk project – ontologies, RDF extractors and RDF2Go but aims to provide a more intuitive interface. The Haystack⁸ and Similie⁹ projects at MIT have produced a range of software tools for managing information. The projects do not presently form a single desktop application and many require browser extensions. In contrast, The Fascinator provides a pure-web interface that allows remote, cross-browser access to the researcher's data.

The Fascinator is also tightly integrated with The Integrated Content Environment¹⁰ (ICE), a word processor based academic publishing system and is being developed in parallel with the Lensfield¹¹ system for desktop chemical informatics, developed by our collaborators at the University of Cambridge.

PROCESS

STEP 1. HARVEST DATA FROM A RANGE OF SOURCES

For an eResearcher, the "Knowledge Network" is no more concentrated on a PC than it was contained within a bookshelf in the print-centric world. Research materials are spread across a range of sources: files, email, blogs, web sites, social networks and bibliographic software. The Fascinator provides a flexible system for bringing these information sources together into a unified index, with human and machine readable web-views using open source components, loosely joined. The Fascinator watches data sources, including the file system and OAI-PMH. A plugin architecture allows feeds from systems such as email, Zotero, RSS & ATOM feeds (e.g. Twitter and Delicious) and websites. [There will be a range of these ready at conference time]

STEP 2. EXTRACT THE METADATA, APPLY CONVERSIONS AND STORE

Once changes are detected in the data sources, the system extracts metadata and converts the original object into alternative renditions (such as Microsoft Word, Excel and PowerPoint into HTML and video formats into easily accessible formats such as Flash Video). So as to ensure extensibility within the framework, digital object metadata is represented by RDF. To avoid redundancy in the system, file-system objects aren't copied into the store, merely referenced – this reduces the footprint of large files. A similar approach could be utilized for web-based content/data being managed as part of the ANDS Data Commons.

STEP 3. USER INTERFACES FOR ERESEARCHERS

At its most basic level, The Fascinator holds an Apache Lucene index of the data and can be searched via the Apache Solr faceted search engine. At this level of functionality, there is an easy comparison with popular desktop search engines. However, this search interface provides a broader user interface than traditional filesystem directories and can be supplemented by other user-interaction. Several examples from Jones' data are useful when considering such interaction. Firstly, a user tagging interface has been developed that provides three methods for tagging data: open tagging for traditional, uncontrolled tags; hierarchical tagging to allow for the grouping of objects in an informal manner; and taxonomical tagging to allow for classification based on a known taxonomy. Further work is being undertaken to automatically tag items based on elements such as keyword, file paths and metadata. Secondly, each interview consists of a video, transcript, list of questions and timecodes. With some general definitions, The Fascinator will group these objects together so as to provide improved search results and user interfaces that combine these objects. Using Jones' data as an example, The Fascinator can provide the user with the ability to jump to sections of an interview or find common responses across multiple interviews. Lastly, the method for examining Jones' data is based on an intersection of personal and official recounts, time and location. Allowing the user to “mash-up” this data is the basis for explorations into further work that provides concept mapping, timeline and annotation interfaces.

STEP 4. SHARE THE DATA

Data sharing and backup in The Fascinator is designed to be as invisible as possible to the researcher. The act of tagging resources in order to organize them will have the side-effect that they are routed to appropriate backups and sharing services using standards such as RIF CS (Describe my Data), Atom and OAI-ORE. Integral to this work is the desire not just to make the data available but to make it part of the web. To this end, the ANDS Register My Data and Identify My Data services will be integrated into the Fascinator in the form of buttons such as “Publish this collection” (this will be done by conference time). The Fascinator does have an authorization model which can be used to restrict access to materials in a user configurable way – as described in our work on an eThesis system. More broadly we are using the tools developed with the PMRC to manage the ERA data collection and evidence gathering process.

ACKNOWLEDGMENTS

The authors acknowledge the collaborative nature of this work. Leonie Jones and Chris Lee from the Public Memory Research Centre have been involved in describing the research paradigm and members of ADFI have been involved with the technical design and implementation: Bronwyn Chandler, Oliver Lucido, Linda Octalina and Ron Ward.

REFERENCES

1. Jones, L. *They'll come looking for you*. (Maenjin Entertainment: 2008).at <http://12fieldregiment.com/add_dvd.htm>
2. Berners-Lee, T. *Linked Data - Design Issues*. (2009).at <<http://www.w3.org/DesignIssues/LinkedData.html>>
3. Australian National Data Service. *Australian National Data Service* (2009).at <<http://ands.org.au/index.html>>
4. Grant, H. Register My Data. *Australian National Data Service* (2009).at <<http://ands.org.au/services/register-my-data.html>>
5. Grant, H. Identify My Data - Overview. *Australian National Data Service* (2009).at <<http://ands.org.au/services/identify-my-data.html>>
6. Sauermann, L., Grimnes, G. & Roth-Berghofer The Semantic Desktop as a foundation for PIM research. *Personal Information Management Workshop* (2008).
7. NEPOMUK - The Social Semantic Desktop. at <<http://nepomuk.semanticdesktop.org/xwiki/bin/view/Main1/>>
8. Haystack Group. at <<http://groups.csail.mit.edu/haystack/>>
9. SIMILE Project. at <<http://simile.mit.edu/>>
10. Sefton, P. The integrated content environment. *AUSWEB 2006* (2006).at <http://eprints.usq.edu.au/archive/00000697/01/Sefton_ICE-ausweb06-paper-revised-3.pdf>
11. Downing, J. lensfield - Google Code. *Project website* at <<http://code.google.com/p/lensfield/>>