

# Rich Tags: Cross-Repository Browsing

Daniel Alexander Smith, Joe Lambert and mc schraefel  
{das05r,jl2,mc} at ecs.soton.ac.uk

IAM Group, School of Electronics and Computer Science, University of Southampton, UK

## **Abstract**

We present RichTags, a system for cross-site browsing and exploration of digital repositories. Categorical and faceted search across repositories is poorly supported, especially compared to the support of keyword search through internet search engines. We combine a variety of information retrieval techniques to determine categories of papers, to enable cross-repository browsing by category. The browsing and exploration of this metadata is achieved through a multi-faceted dynamic exploration interface. Social interaction features have also been added to enable cross-repository tagging, commenting and sharing of papers into groups. These social features are available via an API to enable future work to add plugins to pull comments back to the repositories.

## **The Problem**

Category search within digital repositories is poorly supported. This means that people wishing to access the assets of digital repositories are largely limited to keyword search, which means they must know what they want in order to look for it. Our participant studies of digital repositories use have shown that, when restricted to keyword search, it is perceived as often easier to use a search engine like Google rather than keyword search on a local repository, even if this is to find a local artefact. An advantage that local repositories currently have over massive search services, however, which is not being leveraged, is local or community-based knowledge. This knowledge of context, such as who works with whom; how one project "Over Here" relates to another project "Over There."

## **Cross-Repository Browsing**

We present RichTags, a system that enables cross-repository browsing of digital repositories. RichTags aggregates metadata from multiple repositories, and derives additional metadata based on document contents, so that it is possible to browse through categories of multiple repository contents in a single interface. RichTags is comprised of two key elements: Back-end processing that aggregates repository metadata and derives additional metadata based on document content, and a rich user interface, mSpace [1], that provides rapid triage of the faceted metadata.

## **Back-end Processing**

All EPrints repositories make available the data they contain via an OAI-PMH feed [2]. These feeds can be used to extract each e-print stored within a repository and also metadata associated with these e-prints. Richtags makes use of these feeds to store a local cache of each known repository's metadata. In addition to the metadata available through each repository's OAI feed, the Richtags back-end server produces additional metadata based on the information it has collected. These additional metadata fields are termed 'Richtags' as they are tags computationally connected with an e-print which have some semantic meaning, such as 'category' or 'Institution'.

An explanation of each Richtag follows.

## **Institution - whois Lookup**

When interrogating a repository's OAI feed, no information is returned about the Institution to which the repository belongs. However using the URL for each repository, Richtags is able to make use of the WHOIS lookup service to identify the Institution.

This is of particular importance when Institutions do not rename their EPrints installation, resulting in multiple repositories having the same default title.

## **Year & Decade**

Date is one of the available pieces of metadata already available from the OAI feed. From this information, a year and decade can be extracted for each e-print. Using range based identifiers such as year and decade allow for temporal browsing of the information which is not so easily done with just a date field.

## **TF-IDF Keywords**

Although EPrints offers authors the ability to apply their own keywords at the time of submission, their availability and worth differ from repository to repository and e-print to e-print. Richtags generates its own keywords from an e-print's title and abstract using the TF-IDF algorithm.

To further improve the quality of the keywords, stopwords and word stemming are used. Stopword removal is the process of removing all words that appear commonly in the English language that offer no significance to a body of text other than to aid its fluency. Word stemming aims to reduce all words to their most basic form such that, for example, 'cancer' and 'cancerous' would appear as the same word. These two additional techniques aid the removal of unimportant words and give a higher word count to repeated topics even if not used in the same context, before applying the TF-IDF ratio.

## **Category Inference**

By default, EPrints uses the Library Of Congress Subject Headings (LCSH) as Subjects. Not all repositories use this system, some opting for their own internal categorisation scheme. Richtags attempts to match all the varying subject hierarchies used by various Institutions by examining the Journals or Conferences that an e-print has been published in, or submitted to.

Depending on the repository's version of EPrints, screen scraping may be required to extract the Journal or Conference title for a particular e-print submission. Once known, a Journal or Conference is searched for within the DMOZ [3] directory. DMOZ is the "largest, most comprehensive human-edited directory of the Web, constructed and maintained by a vast, global community of volunteer editors", and as such a large proportion of the Journals and Conferences can be found, and a two-tier category can be extracted.

These inferred categories provide another new way to browse the information space, producing links between articles that have not been possible before, with the information from the repository alone.

Future work is involved with examining other possible Journal/Conference sources such as EBSCO to improve the match percentage.

## **Author Ambiguity**

Author ambiguity is caused by different text formatting conventions being used from one repository to another. Although not currently implemented, Richtags aims to alleviate this problem by using the UK Je-S lookup service to get additional information about an author than just their name. With this extra information it would be possible to differentiate 'A Jones' or 'Jones, A.' in one repository from 'A Jones' in another.

# User Interface

The Richtags front end is built around mSpace. mSpace is a faceted browser that allows for exploration of a space using a series of columns [Figure 2]. Columns can be re-ordered at will to change the focus of exploration and additional columns can be added and removed from a list below the main window [Figure 2].



Figure 1: Richtags Full Screenshot [4]

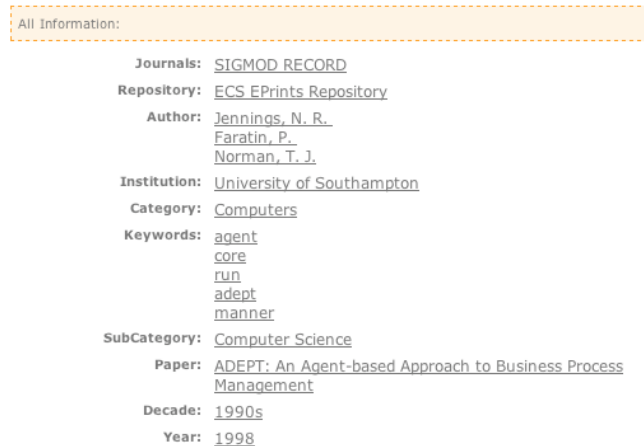
Back highlighting (shown in green) is used to represent the items in a column that are possible matches to items the user has selected themselves. In Figure 2 the user has made selections in the "SubCategory", "Institution" and "Paper" columns, the back highlights make it immediately obvious that the selection the user has made is in the "Decade": 1990's and in the broader "Category": Computers.



Figure 2: Column Browser

When a user makes a selection to view an e-print within the Richtags system, all known information about that e-print is displayed. The title and abstract, along with any known authors are displayed [Figure 1] at the top of the page. At the bottom of the page is a summary of all the metadata that is known about the current e-print [Figure 3]. Some of this metadata is the same as that pulled from the original repository and some are additional Richtags that have been added.

Selecting any of the metadata in this list will add that item to the Column Browser [Figure 2] as an additional filter, producing a new list of matching e-print results.



All Information:

Journals: [SIGMOD RECORD](#)

Repository: [ECS EPrints Repository](#)

Author: [Jennings, N. R.](#)  
[Faratin, P.](#)  
[Norman, T. J.](#)

Institution: [University of Southampton](#)

Category: [Computers](#)

Keywords: [agent](#)  
[core](#)  
[run](#)  
[adept](#)  
[manner](#)

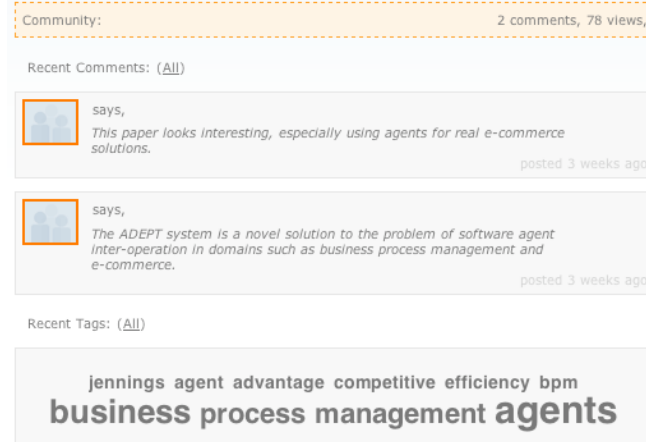
SubCategory: [Computer Science](#)

Paper: [ADEPT: An Agent-based Approach to Business Process Management](#)

Decade: [1990s](#)


Year: [1998](#)


Figure 3: All metadata known about an e-print



Community: 2 comments, 78 views

Recent Comments: (All)

 says,  
This paper looks interesting, especially using agents for real e-commerce solutions.  
posted 3 weeks ago

 says,  
The ADEPT system is a novel solution to the problem of software agent inter-operation in domains such as business process management and e-commerce.  
posted 3 weeks ago

Recent Tags: (All)

jennings agent advantage competitive efficiency bpm  
**business process management agents**

Figure 4: Social tagging & commenting feature of Richtags

## Social Interaction (Web 2.0)

Richtags provides a social, community based tagging system [Figure 4]. Users can register for an account and add their own tags to any article in the system. It is then possible to search for other items with the same tag and this can be scoped to any user or just a particular user. This allows the user to apply tags for their own categorisation or to allow other users find related works. Users can also create and join groups, this allows for a collection of articles to be found and shared with a number of users. Commenting is another feature made possible on e-print submissions from all repositories, encouraging input from the community in discussion of new articles.

An API is available to export the user tags and comments for a given e-print. The API can return a number of formats, including XML and JSON, to accommodate the widest number of uses of the data.

## References

1. m. c. schraefel, Daniel A. Smith, Alisdair Owens, Alistair Russell, Craig Harris, and Max Wilson. The evolving mspace platform: leveraging the semantic web on the trail of the memex. In HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, pages 174–183, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-168-6.
2. H. Van de Sompel, M.L. Nelson, C. Lagoze, and S. Warner. Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine, 10(12):1082–9873, 2004.
3. DMOZ Open Directory Project <http://www.dmoz.org/>
4. Richtags beta, <http://beta.richtags.net/>