TRANSFORMING CURRENT AWARENESS THROUGH RSS: How two projects (ticTOCs and Gold Dust) are using RSS to improve the information landscape for the 21st century researcher

by Lisa J Rogers†, Simon Hodson‡ and Roddy MacLeod†

† Heriot-Watt University, ‡ University of Hull

ABSTRACT

This paper looks at the current situation with respect to RSS and then reports upon the findings of the ticTOCs and Gold Dust projects. We will look at the lessons learnt from developing the ticTOCs service, and also report on two iterations of the Gold Dust development and use cycles. We will deliver an appraisal of the effectiveness of the raft of techniques being employed by Gold Dust. How effective are current data mining and pattern matching techniques for such an application? How useful is RSS metadata in this context? These findings will be of considerable pertinence both for future services which may use RSS Feeds, and for future research and development in the area of adaptive personalisation using RSS.

FEEDS, INFORMATION OVERLOAD AND RSS ADOPTION

Really Simple Syndication (RSS) has emerged as an extremely useful mechanism through which researchers, practitioners and educators can obtain current awareness information and be alerted to the existence of new content. RSS is a XML format that allows content providers to easily share information on the internet[1]. Users can subscribe to RSS Feeds using a feed reader and are alerted to new items in the feeds to which they have subscribed each time they view their feed reader. Typically, an RSS Feed contains a title, together with either a short summary of the content or the full text, and a link back to the original site[2]. However RSS can also be extended[3] through the use of modules to provide more descriptive metadata such as Dublin Core[4] and PRISM[5]. Many publishers of Journal Tables of Contents are already taking advantage of this, allowing them to include metadata in their RSS feeds such as authors, ISSN, volume, issue and page numbers along with article titles and abstracts.

There are many types of RSS feeds. Typically, news and blog updates are the most common use of RSS. Hammond *et al*[3] describe some of the feeds that science publishers are starting to provide, including other content aside from Journal Tables of Contents, such as citation alerts, news, jobs, product releases, press releases, reviews and events. Through the work on the Gold Dust project[6] a number of feeds were identified that may be of interest to engineering academics and researchers, and these can be placed in the following categories: journal tables of contents (TOCs), calls for papers, forthcoming conference and events announcements, funding opportunities, new theses and dissertations, new book announcements, subject related news, new items in institutional repositories,

---

[1] Moffat M (2003) "RSS - A Primer for Publishers & Content Providers" http://www.techxtra.ac.uk/rss_primer/

[2] Cooke CA (2006) "Current Awareness in the New Millennium", *Medical Reference Services Quarterly*, 25:1, pp 59 —69

[3] Hammond T, Hannay T, Lund B (2004) "The Role of RSS in Science Publishing: Syndication and Annotation on the Web' *D-Lib Magazine* 10:12

[4] RSS 1.0 Modules : Dublin Core http://web.resource.org/rss/1.0/modules/dc/

[5] RSS 1.0 Modules : PRISM http://www.prismstandard.org/resources/mod_prism.html

[6] Gold Dust Project http://www.hull.ac.uk/golddust/

professional society news, patents, component announcements, suppliers, standards and new teaching and learning resources.

Yet a recent White Paper[7] from Forrester Research noted that RSS adoption by users within the Internet population was only 11%. Not only that, but the report showed that relatively few non-users had any interest in using feeds. In addition, for those who do use and subscribe to RSS feeds, it is only too easy to become swamped by the resulting information overload. Although RSS use has grown over the last few years (An OCLC Report[8] in 2005 suggested only 5% of information users used RSS based upon a survey of over 3000 respondents) there is still some lack of willingness on the part of researchers in its take up. In 2005 a White Paper[9] from Yahoo suggested that although only a small percentage of users know they are using RSS, 27% use it unknowingly via personalised home page services such as MyYahoo or iGoogle through the syndication of news headlines, etc.

In 2007 Hardesty and Sugarman conducted a survey[10] of Academic Librarians about the methods they used to keep up to date with professional literature. Out of the 707 respondents 15% used RSS Feeds. Only 23% of those surveyed said that information overload did not limit their ability to keep up to date, with the remainder commenting that information overload did limit their ability to keep up to date to a lesser or greater extent. Chen and Tai[11] describe the problem with information overload on the internet in general as a major problem, with the continual growth of the internet they estimated that the content available doubled every 18 months. They suggest new approaches are needed to combat this problem so that users are presented with relevant information.

What can information professionals do to make sure that the opportunities afforded by RSS are not wasted? Should they concentrate on teaching researchers about the benefits of RSS in the hope that uptake will subsequently increase, or should they work with, and regard, RSS as a component of an invisible infrastructure through which to develop simple user-facing services? The two projects described below are focussed on the latter.

PROJECTS

ticTOCs[12] and Gold Dust are two large-scale consortium JISC-funded projects which are based around RSS and which have implications for the international information community. ticTOCs has developed a freely available journal current awareness service which aggregates tables of contents RSS feeds for over 12,000 journals from more than 430 publishers. Gold Dust is exploring ways in which the development of pervasively generated Personal Interest Profiles (PIPs) from ticTOCs usage data may be used to pan the current information flood of RSS feeds for items of significance to individual researchers - the eponymous 'gold dust'.

Whilst ticTOCs responds to a very clearly defined use-case: the need of researchers to keep up to date with current journal tables of contents, Gold Dust is in contrast

---

[7] Katz JM, Spivey Overby C, Owyang JK, Cummings T, Murphy E (2008) "What's Holding RSS Back? Consumers Still Don't Understand This Really Simple Technology *Forrester Research*. http://www.forrester.com/Research/Document/Excerpt/0,7211,47150,00.html

[8] De Rosa C et al. (2005). "Perceptions of libraries and information resources: A report to the OCLC membership". Dublin, OH: OCLC Online Computer Library Center Inc.

[9] Grossnickle, J. et al. 2005. RSS–Crossing into the mainstream. Yahoo!Inc. http://publisher.yahoo.com/rss/RSS_whitePaper1004.pdf#search=%22%22RSS%20%E2%80%93%20Crossing%20into%20the%20mainstream%22%22

[10] Hardesty S, Sugarman T (2007). "Academic Librarians, Professional Literature, and New Technologies: A Survey" *The Journal of Academic Librarianship* Vol 33 No 2, pp. 196-205.

[11] Chen CT, Tai WS (2003). "An Information Push-Delivery System Design for Personal Information Service on The Internet" *Information Processing and Management* 39, pp. 873-888.

[12] ticTOCS Journal Tables of Contents Service http://www.tictocs.ac.uk

very much a research project, yet the projects are connected, as Gold Dust is using ticTOCs usage logs in order to produce PIPs.

TICTOCS PROJECT

ticTOCs has worked with multiple publishers to aggregate, normalise and present journal TOCs within an easy to use interface which facilitates the discovery, display and reuse of TOC content in a personalisable web based environment. ticTOCs has also developed advocacy materials on journal TOC RSS standardisation, and has created APIs which enable journal TOC RSS feeds to be embedded in other information services.

ticTOCs was developed in mind of the low take up of RSS feeds, and allows users to easily consume journal TOC RSS feeds from multiple publishers, without having to have knowledge of RSS or have access to a feed reader. However, ticTOCs also caters to the needs of the more technically minded audience of users who would want to use a feed reader and manage their journal TOC RSS feeds alongside other feeds. These users are able to use ticTOCs to find feeds for journals from many different publishers and export them, individually or via OPML, into their own feed readers. This saves the effort of visiting each publisher's website and individually subscribing to feeds. Wusteman[13] provides guidance on OPML[14], commenting that "OPML is used to exchange subscription lists between programs that read RSS files, such as feed readers and aggregators". Figure 1 shows the features incorporated into ticTOCs.

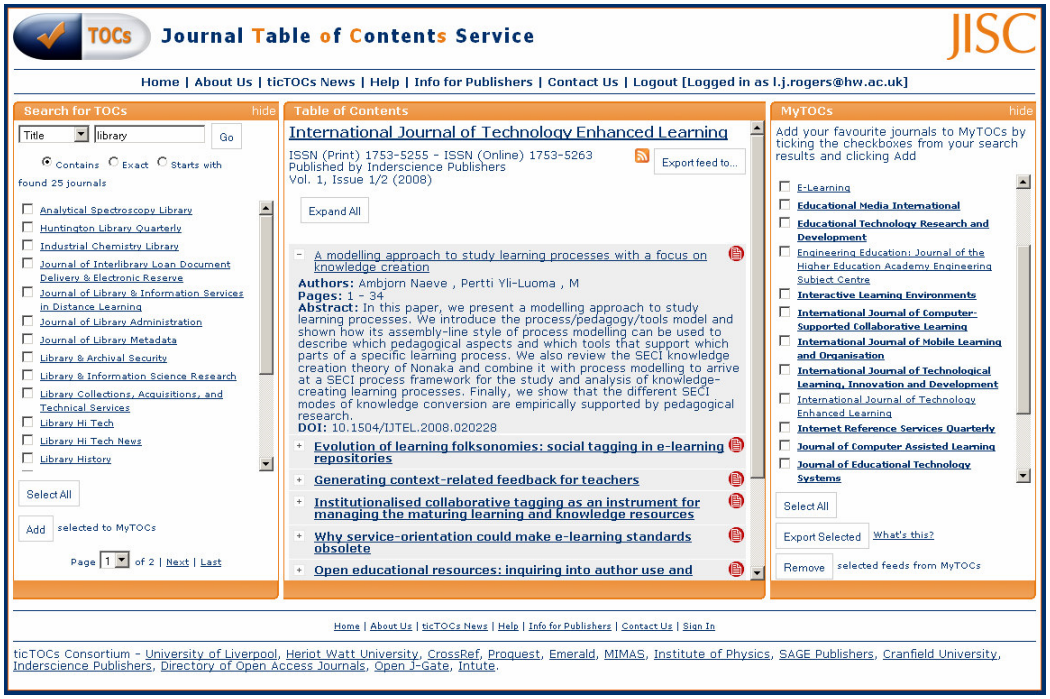

Figure 1 Screenshot of ticTOCs Website

Recommendations for Publishers of TOC RSS Feeds

An outcome of the ticTOCs project was "Recommendations on RSS Feeds for Scholarly Publishers" to address the huge variation in the provision of RSS Feeds. Hammond *et al*[3] summarises the use of the various formats of RSS together with

---

[13] Wusteman J (2004). "RSS: The Latest Feed" *Library Hi Tech* Vol 22 No 4, pp 304-413

[14] OPML Outline Processor Markup Language http://www.opml.org/

the use (or non use) of additional metadata. ticTOCs can confirm this by providing a snapshot of publishers' use of the various formats (see Table 1). The Royal Society of Chemistry[15] is an example of a publisher which includes images in their RSS feed using the RSS 1.0 Content[16] Module to enhance the quality of their feeds for end users.

Table 1 Summary of Feed Formats in ticTOCs

Format	Additional Metadata	No of Publishers	No of Feeds
Atom	-	2	6
RSS 1.0	-	14	196
RSS 1.0	DC	27	1075
RSS 1.0	DC + Content	1	16
RSS 1.0	DC + PRISM	241	3978
RSS 1.0	DC, PRISM + CONTENT	4	396
RSS 2.0	-	148	4096
RSS 2.0	DC	5	1930

ticTOCs initially analysed the content of RSS TOC feeds in order to provide information to guide the development of recommendations for publishers. At that stage, ten publishers' feeds included the following variations in the title element alone:

- Nature i.e. <Journal title>
- BMJ Current Issue i.e. <Journal abbreviation> + <Content description>
- Journal of Geophysics and Engineering latest papers i.e. <Journal title> + <Content description>
- Journal of managerial psychology: table of contents i.e. <Journal title> + <Content description>
- Science Direct Publication: Energy i.e. <Journal Publisher> + <Journal title>
- SpringerLink – Journal i.e. <Journal Publisher> + < Content description >
- Blackwell Synergy: International Journal of Cosmetic Science: Table of Contents i.e. <Journal Publisher> + <Journal title> + < Content description >
- NATURE –LONDON i.e. <Journal title> + <Location of Publisher>

Variations in practice amongst publishers' feeds can be irritating for end users, but can be insurmountable for service providers attempting to aggregate the feeds. A particular problem is invalid feeds as checked with RSS validators such as the W3 feed validation service[17]. Many feeds are quite messy and would not pass validation using RSS validators (e.g. some feeds contain a number of undefined <channel> and <item> elements including: <year>, <month> <day> <volume> <issue> etc). In general it appears that the RSS 2.0[18] feeds are probably worse than the RSS 1.0[19] feeds. Occasionally RSS 1.0 feeds appear to have been created with some RSS 2.0 elements (e.g. some RSS 1.0 feeds have <guid> and <pubDate> which are RSS 2.0 elements. This is probably a simple error on the part of the publisher. Other frequently occurring issues include HTML tags in the description element, undefined characters, invalid XML such as missing a closing tag or empty elements that are required to contain data such as <link>.

Representatives from CrossRef[20], journal publishers and ticTOCs are participating in a working group to produce recommendations for publishers of TOC RSS Feeds. Some of the proposed guidelines are:

---

[15] Royal Society of Chemistry http://www.rsc.org

[16] RSS 1.0 Modules : Content http://web.resource.org/rss/1.0/modules/content/

[17] W3C Feed Validation Service http://validator.w3.org/feed/

[18] RSS 2.0 Specification http://cyber.law.harvard.edu/rss/rss.html

[19] RDF Site Summary RSS 1.0 http://web.resource.org/rss/1.0/

[20] CrossRef http://www.crossref.org/

- **Use the RSS 1.0 specification** because of its greater flexibility. RSS 1.0 extended with RSS 1.0 modules is ideally suited to the provision of RSS table of contents type materials. A number of publishers are already utilising this approach. Some publishers offer a choice of formats for their TOCs (e.g. RSS 1.0 and RSS 2.0).
- **Use RSS 1.0 Modules** (e.g. Dublin Core Module, CONTENT Module and PRISM Module) to extend TOC RSS feed functionality
- **Validate TOC RSS** feeds using an RSS validation tool. (e.g.W3C feed Validator or Redland RSS 1.0 validator[21]).
- **Do not include HTML markup in standard RSS Feed elements** e.g. Avoid using HTML tags (such as <b> <p> <a href> etc) in the <item><description> element. The <item><description> should only include plain text as it is not possible to know how the feed will be presented and including markup can prevent your feed from being correctly displayed.
- **Use the RSS 1.0 Content Module to present HTML marked up content.** For example the <content:encoded> element can be used to provide an alternative marked up version of the <item><description>.
- **Include abstracts/summaries in your feeds.** There is increasing evidence that providing users with more information in a feed will drive more users to your site.
- **Do not restrict access to TOC RSS feeds**. RSS is an excellent and cost-effective way of driving traffic to, and increasing brand awareness of publishers' content. Restricting access to the feeds themselves (e.g. to subscribers only) negates many of the potential benefits that RSS can bring.
- **Understand the purpose of each RSS TOC feed you provide**. Provide multiple feeds, rather than diluting the message of one with information irrelevant to its audience. For example it may appropriate to provide separate feeds for the current TOC issue only and a combined TOC for a number of recent issues.
- **Provide OPML files** to enable aggregators or end users to utilise a number of your feeds. In some instances it may be appropriate to provide a range of OPML files (e.g. OPML files for each subject category) or to enable end users to create custom OPML files on the fly.

More discussion on these guidelines is given in the article - RSS and Scholarly Journal Tables of Contents: the ticTOCs Project, and Good Practice Guidelines for Publishers.[22]

ticTOCs Users Awareness of RSS

The ticTOCs project conducted an evaluation of the initial ticTOCs website in order to understand the requirements of its users and influence future design. 17 users were interviewed by members of the project team. The users were asked about their awareness and usage of RSS. 6 users were not aware of RSS and 4 were aware of RSS but chose not to use it. Upon further investigation 2 users were using RSS pervasively by subscribing to news headlines in iGoogle without realising that this functionality was powered through RSS. It was therefore important that the ticTOCs web site catered for these two types of users, the ones that simply wanted to quickly find and export journal RSS feeds into their own feed reader, and those who wished to find and display the latest issue of these journals within the ticTOCs environment.

---

[21] Redland Feed Validator http://librdf.org/rss/

[22] Rogers, L. (2008) RSS and scholarly journal tables of contents: the ticTOCs project, and good practice guidelines for publishers. *FUMSI Magazine,* October 2008 [online] Perma Link: http://web.fumsi.com/go/as/3356

ticTOCs APIs

The ticTOCs project produced some simple APIs; one of which is a simple text file[23] listing all the ticTOCs journal titles, feed URLs together with ISSNs and eISSNs. This data has allowed libraries to embed TOC information into their library catalogues. An example is the Wageningen University and Research Centre Digital Library[24] whose catalogue now includes RSS feed URLs and embedded RSS content. This feature lets users find feeds and view the latest content from within the OPAC, thus allowing them either to export the feeds or simply browse the latest issue within the catalogue. Provision of these APIs has had benefits not only for external users and developers, but also for ticTOCs itself. When the Wageningen University compared the ticTOCs data to its own catalogue it was able to identify a number of journals that ticTOCs did not have RSS feeds for. This enabled ticTOCs to enhance its own directory of feeds, by collecting feeds for these additional titles.

GOLD DUST PROJECT

A Personal Interest Profile (PIP), also known simply as a "profile" can be thought of as a representation of information needs[25]. The most basic means of generating these profiles is to have users specifying their preferences. However the premise of the Gold Dust project was to create these profiles pervasively, by logging the actions of volunteers' use of ticTOCs and then extracting significant terms.

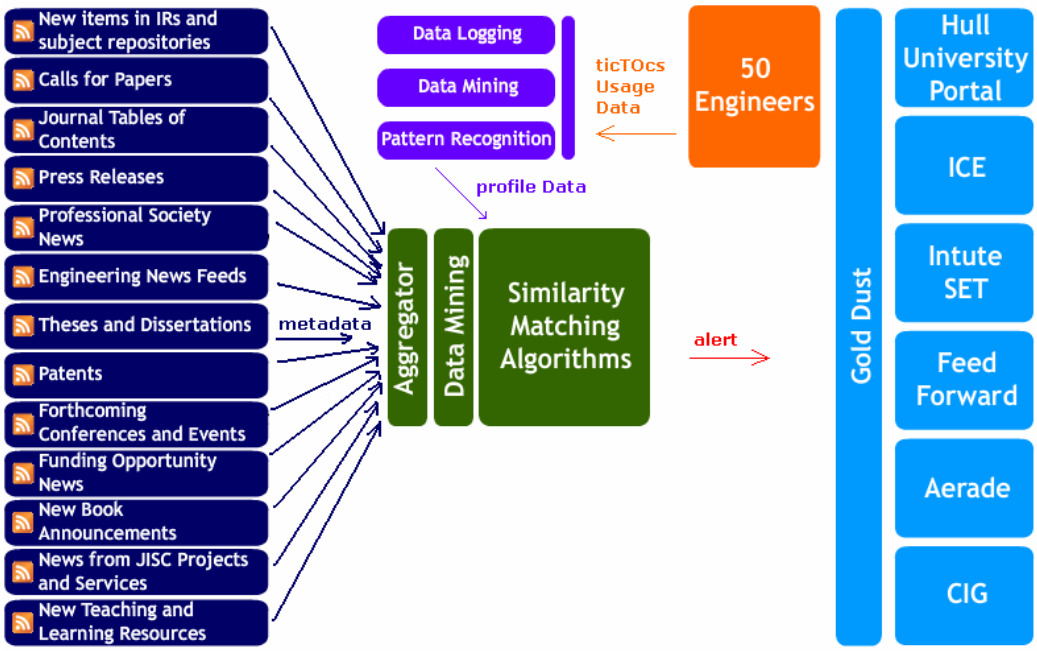

Figure 2 Diagram of Gold Dust Concept[26]

The project worked with a user group of 50 engineering academics and researchers and investigated data mining, pattern matching, latent semantic analysis and the use of automated, hierarchical classification systems for the purposes of producing

---

[23] ticTOCs Data Set http://www.tictocs.ac.uk/text.php

[24] Wageningen University and Research Centre Digital Library eJournals Catalogue http://library.wur.nl/desktop/vrr/

[25] Campbell DR, Culley SJ, McMahon CA (2005) "Pushed-Based Strategies for Improving the Efficiency of Information Management in Design" *Proceedings of International Conference on Engineering Design* (ICED 05).

[26] Gold Dust Concept Diagram Adapted from http://www.macs.hw.ac.uk/~mthljr/golddust/gd.html

PIPs and matching them with content from thousands of hand-selected RSS feeds in order to retrieve relevant items. The flow of data in the Gold Dust project is illustrated in Figure 2.

Although an initial user group of 50 engineers was recruited, some of them did not participate fully in the project, with only 37 registering with the ticTOCs service. Of these 37 users, 6 registered but did not perform any actions, and a further 11 users did not have significant enough usage levels to produce results. Hence the first iteration was limited to only 20 users (19 of whom provided feedback on the results). Three of these users did not participate fully in the second iteration leaving 16 users completing both stages, various reasons were given for non participation including already having established mechanisms for literature searching, ticTOCs not having an article cross search facility and time constraints.

Iteration 1

Phase one of the project set out to compare two text mining approaches. Approach A involved feeding article titles and abstracts that the users had viewed within the ticTOCs environment into NaCTeM's TerMine web service[27] to return a list of key-phrases with corresponding c-Values[28]. Key-phrases with higher c-Values were deemed more likely to represent users' interests than those with lower c-Values. The ranked lists of key-phrases constituted the users' PIPs. These key-phrases were then matched using MySQL queries to two databases of RSS feed items: the ticTOCs database of journal articles, gathered from the engineering related journals in ticTOCs; and what we shall call the Gold Dust database – a collection of items from over 800 engineering related RSS Feeds, aggregated into 16 different categories, including 'Calls for Papers', 'New Items in Institutional and Subject Repositories', 'Theses and Dissertations', 'Teaching and Learning Resources' etc. This produced a lot of results, therefore results were ranked by accumulative c-Values, meaning that items would be ranked highly if they contained multiple key-phrases with high c-Values. For evaluation purposes, users were shown up to 5 items per category from the Gold Dust database and up to 15 journal articles.

Approach B used ExtMiner[29], an open-source, density-based (DBSCAN) clustering and text mining software. It can perform document clustering and text mining on Lucene-indexed data – for single words as well as two-word terms. It was chosen precisely to allow exploration of the potential of combining single word and two-word term extraction. Results obtained by each of these operations are combined and then matched using semantic vectors against Lucene-indexed Gold Dust and ticTOCs data. The top 30 items from the Gold Dust database (split into respective categories) and the top 30 from the ticTOCs database would be presented to users.

Upon analysis of the test results it was noticed that many key-phrases in the users' PIPs were deemed not appropriate to return matched results. These key-phrases were grouped into 4 categories: HTML elements (such as img src=), data relating to locations or dates (such as Tue, Mar, New York etc), other metadata included in the description field of the RSS feeds, which was in addition to the abstract (such as authors, DOIs, publisher etc) and generic terms relating to the investigation process (such as 'case study', 'literature review', 'numerical solution'). A stop list was manually created to remove these key-phrases from the PIPs and the matching was repeated, however this is obviously not a sustainable approach. A method using the Google search API[30] to return the number of hits per key-phrase and eliminate terms above a certain threshold was trialled, but was not

[27] NaCTeM's TerMine Web service http://www.nactem.ac.uk/software/termine/

[28] Frantzi K, Ananiadou S, Mima H (2000): Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, Vol 3 No 2 pp. 115-130

[29] ExtMiner http://extminer.sourceforge.net/

[30] Google Search API http://code.google.com/apis/ajaxsearch/

implemented in time to give results to the users. Future projects should strongly consider the Google API mechanism for creating automatic stop lists.

Approach A presented the items through a web interface, in order of category then relevance, with the users asked to rate the items from 1 to 10; 1 was deemed not relevant whilst an item rated 10 was deemed highly relevant. In total 1175 items were presented between 19 users, with 10.8% of items rated as gold dust (an item with rating of 8 or above was considered 'gold dust'). The average rating given to the items was 3.3. The results were also analysed per category, with 'Journal Articles' returning the highest percentage of gold dust (22.7%) with a mean rating of 4.7. Other categories with high ratings were; 'New Items in Institutional and Subject Repositories' (9.5% gold dust; mean rating 3.3), 'Engineering News Feeds' (13.7 % gold dust; mean rating 3.2), 'Component Announcements' (9.9% of gold dust; meaning rating 2.5), 'Teaching and Learning Resources' (13.9% gold dust; mean rating 3.4). Results for each category are shown in Table 2.

Items matched through Approach B were presented through a similar interface to Approach A and were again rated 1-10 by the users. This method presented 1126 items to the 19 users with 9.86% rated as gold dust (mean rating 3.03). The highest rated categories using this approach were 'Journal Articles', 'Theses and Dissertations' and 'Engineering News Feeds'. Results per category are also listed in Table 2.

Table 2 First Iteration Results

	Approach A			Approach B		
Category	**Total Items**	**Gold Dust**	**Mean Rating**	**Total Items**	**Gold Dust**	**Mean Rating**
Calls for Papers	**60**	**1.67%**	**1.78**	**4**	**0%**	**1.25**
Items in Institutional/ Subject Repositories	**95**	**9.47%**	**3.26**	**110**	**3.64%**	**2.23**
Funding Opportunity News	**18**	**0%**	**2.67**	**1**	**0%**	**1**
Patents	**87**	**3.45%**	**2.66**	**64**	**0%**	**3.28**
Press Releases	**78**	**6.41%**	**2.47**	**20**	**5%**	**2.7**
Professional Society News	**46**	**8.7%**	**3.33**	**4**	**50%**	**7**
Engineering News Feeds	**95**	**13.7%**	**3.19**	**234**	**9.34%**	**2.57**
Component Announcements	**81**	**9.88%**	**2.54**	**34**	**0%**	**1.91**
Teaching and Learning Resources	**36**	**13.9%**	**3.39**	**7**	**0%**	**1**
Forthcoming Conferences and Events	**54**	**3.7%**	**3.17**	**3**	**0%**	**2.33**
Theses and Dissertations	**87**	**8.05%**	**3.69**	**67**	**11.9%**	**3.25**
News from JISC Services and Projects	**4**	**0%**	**1.5**	**0**	**0%**	**0**
Suppliers	**21**	**0%**	**1.95**	**2**	**0%**	**1**
New Book Announcements	**51**	**7.84%**	**3.24**	**3**	**0%**	**2**
Standards	**56**	**3.57%**	**2.18**	**12**	**0%**	**1.75**
Others	**28**	**3.57%**	**2.54**	**6**	**0%**	**1.33**
Journal Articles	**278**	**22.7%**	**4.7**	**555**	**13.2%**	**3.59**
All Categories	**1175**	**10.8%**	**3.3**	**1126**	**9.86%**	**3.03**

On analysis of the two approaches it was noted that the category which returned the most gold dust was 'Journal Articles', with items from 'Institutional and Subject Repositories' also rating highly. This could be a positive indication that the methods

employed have the potential to work well for information recommendation on the basis of Personal Interest Profiles, particularly when journal articles and scholarly items are the targets for matching. Given the source of the user data (journal abstracts) it was always likely that matching would perform best against data with the same type of profile. Indeed, it seems fair to observe that the overall mean rating and the proportion of gold dust rated items produced by Approach B were actually boosted by the higher representation of journal articles in its results.

What the results for Approach B demonstrate very well is which categories of feed, because of the nature of their linguistic content, are most susceptible to matching. A relatively high number of results were drawn from the categories 'Engineering News Feeds', 'Theses and Dissertations' and 'Items from IRs and Subject Repositories', possibly indicating a greater level of semantic similarity between that content and the content found in journal abstracts.

Some conclusions can be drawn from the first iteration.

- Too many non-specific key-phrases were extracted by both ExtMiner and TerMine. Although this was eased by including a stop list, a more sustainable approach would be preferred.
- The ticTOCs usage data was limited. Many users did not interact with ticTOCs enough in the period up to the first iteration to produce an adequate profile. In addition some of the TOC items did not hold sufficient information to build a profile (i.e. they had no abstracts, or contained other metadata where the abstract should be).

User Feedback

A formative evaluation was conducted after the first iteration, including a face-to-face focus group session and a feedback questionnaire. The results of the feedback drew up several conclusions.

- Although some users appreciated the possibility of being sent items from all of the identified categories, some users declared there were certain categories they wished to exclude from their results.
- Many users commented that Approach A found items that were of interest to them from sources that they would not generally search.
- Many users reported that items found in Approach B often fell into their broad subject area without corresponding to their specific research interests.
- Users would like the option of influencing or teaching the system about their interests, whether it be marking items as relevant, removing items from their usage data or topping up their PIPs with data from other sources.
- Some users felt that ticTOCs did not fit well into their current mechanisms for searching for literature and were reluctant to use it. At the time of this project, ticTOCs did not implement keyword searching across articles, with some users unwilling to use it without this feature.

Iteration 2

Based upon the user feedback and the conclusions drawn in the first iteration, two approaches for the second iteration were outlined. Approach C would be a refinement of Approach A and would replicate several features that a full scale service would offer. One such feature would be to remove certain key-phrases from their Personal Interest Profiles. We replicated this by asking each user to rate the key-phrases in their PIP: low rated terms were then removed from that user's PIP for the second iteration. Some users requested that they did not wish to receive items from specific categories, and this too was taken into account. PIPs created in Approach C would be derived from *all* the ticTOCs usage data, including any usage of ticTOCs between iterations, as well as items rated highly in the first iteration. No items that the user had previously seen, either within the ticTOCs environment or

from the first Gold Dust iteration, would be returned to users in the second iteration. Users were again asked to rate each item between 1-10 in terms of relevance and were presented with the top 20 Journal Articles and the top 20 items from other categories in an amalgamated list.

Approach D also used the same underlying technologies as Approach A, using TerMine to produce PIPs and matching to items in the ticTOCs and Gold Dust databases using MySQL queries. However this approach did not depend on ticTOCs usage data, instead users were invited to submit complete documents that reflected their research interests. These documents could include papers they had written, and journal articles that they had identified as being of particular importance to their research. These results were presented to the user as specified for Approach C.

Table 3 Second Iteration Results

	Approach C			Approach D		
Category	**Total Items**	**Gold Dust**	**Mean Rating**	**Total Items**	**Gold Dust**	**Mean Rating**
Calls for Papers	**6**	**0%**	**1.83**	**4**	**0%**	**1.25**
Items in Institutional/ Subject Repositories	**96**	**11.5%**	**3.57**	**83**	**16.9**	**4.01**
Funding Opportunity News	**0**	**0%**	**0**	**0**	**0%**	**0**
Patents	**59**	**0%**	**2.02**	**111**	**3.6%**	**1.95**
Press Releases	**7**	**14.3%**	**2.86**	**1**	**0%**	**1**
Professional Society News	**3**	**66.7**	**6.33**	**3**	**100%**	**9**
Engineering News Feeds	**98**	**20.4%**	**3.69**	**68**	**14.7%**	**2.76**
Component Announcements	**8**	**0%**	**1.5**	**6**	**0%**	**1.5**
Teaching and Learning Resources	**1**	**0%**	**1**	**3**	**33.3%**	**4.33**
Forthcoming Conferences and Events	**8**	**0%**	**2.5**	**1**	**0%**	**6**
Theses and Dissertations	**15**	**0%**	**2.67**	**19**	**10.5%**	**4.53**
News from JISC Services and Projects	**0**	**0%**	**0**	**0**	**0%**	**0**
Suppliers	**0**	**0%**	**0**	**0**	**0%**	**0**
New Book Announcements	**4**	**0%**	**2.5**	**5**	**20%**	**3.4**
Standards	**13**	**0%**	**2.85**	**16**	**0%**	**3.38**
Others	**2**	**50%**	**4.5**	**0**	**0%**	**0**
Journal Articles	**272**	**20.2%**	**4.38**	**308**	**16.6%**	**3.97**
All Categories	**592**	**15.2%**	**2.7**	**628**	**13.7%**	**3.47**

Table 3 show the results for Approaches C and D, indicating some improvement on the approaches taken in the first iteration. However there was no significant difference between Approaches C and D, suggesting more investigation is required to show whether user submitted documents or ticTOCs articles provide better input data to profile the users.

Results Categorised by User

Table 4 displays the percentage of items returned for the 4 different approaches that were rated as gold dust by each user. This table shows that Approach A provided a higher percentage of gold dust than Approach B for 15 out of the 19

users. This justifies the decision to build on the techniques used in Approach A in an attempt to develop more sophisticated techniques for the 2nd iteration. In general the results for the 2nd iteration were better than those in the 1st iteration. Out of the 16 users who completed both iterations 9 had better results in the 2nd than the first, 6 had worse results with 1 user staying the same. Approach C produced better results for 10 out of the 16 users from the 2nd iteration, worse results for 5 users and the same for 1 user, when compared to Approach A. Approach D produced better results than A for 7 out of the 16 users.

Table 4 Gold Dust Percentage split by user

	1st Iteration		2nd Iteration	
User ID	**Approach A**	**Approach B**	**Approach C**	**Approach D**
38	**1.79%**	**0%**	**7.5%**	**2.5%**
49	**6.67%**	**3.33%**	**12%**	**2.5%**
50	**12.5%**	**1.67%**	**0%**	**3.57%**
62	**4.11%**	**8.33%**	**5%**	**7.5%**
162	**11.32%**	**6.9%**	**32.5%**	**7.5%**
169	**38.46%**	**32.2%**	**62.5%**	**32.5%**
175	**3.51%**	**35%**	**12.5%**	**22.5%**
201	**0%**	**0%**	**0%**	**0%**
206	**8.62%**	**0%**	**7.5%**	**2.5%**
208	**26.15%**	**16.67%**	**32.5%**	**30%**
211	**30.26%**	**13.79%**	**10%**	**15%**
212	**3.8%**	**0%**	**0%**	**0%**
217	**4%**	**0%**	**13.04%**	**12.5%**
222	**4.94%**	**1.69%**	**0%**	**10%**
223	**7.32%**	**0%**	**12.5%**	**5%**
278*	**3.45%**	**3.33%**	**---**	**---**
297*	**10.14%**	**3.33%**	**---**	**---**
302*	**47.37%**	**13.36%**	**---**	**---**
316	**18.75%**	**45%**	**27.5%**	**62.5%**

*These users only completed the first iteration of the project.

Some users had worse results in the 2nd iteration than the 1st iteration. This result was unexpected. One factor was revealed by the discovery that some of the key-phrases which had been rated low by these users between the iterations, and thus added to the user-generated stop lists, were actually key-phrases that had returned highly rated items in the first iteration. Such a finding would seem to confirm the need for an automatic stop list as opposed to manually-generated, user-created stop lists, which may be excessively vulnerable to conflicting instructions or changes in interest.

CONCLUSIONS

Data

When using Approaches A, B and C ticTOCs usage data was fed into TerMine. This data had originated from RSS feeds and was stored in a database. There were some problems with the input data, such as HTML tags, data not encoded in UTF-8 and non abstract metadata in the description filed. A certain amount of pre-processing using PHP was required to remove or correct this data, including removing any non ASCII characters, as TerMine does not accept them. The majority of these issues would be resolved if publishers were to adopt the RSS guidelines proposed by the ticTOCs project.

When the project decided to use user-submitted documents instead of RSS feeds as input data, we did not realise the additional problems that would occur when trying to convert the data (generally in PDF format) into XML or put into a MySQL

database. Some PDF files were protected, encrypted or simply scanned as an image, meaning the data was not accessible. Some PDF files contained other typeface associated problems such as the use of ligatures to replace the character sequences ffi, ff, fi and fl. Care was taken to replace these ligatures where possible, but this process was very time consuming.

The format of the data in some categories of the Gold Dust database (such as 'Calls for Papers', 'Component Announcements' and 'Suppliers') was also of poor quality, with many items not having any description or only providing minimal data. This made matching to these items very difficult and this is reflected in the results for these particular categories.

One aim of the Gold Dust project was to determine whether or not RSS TOC data could be sufficiently mined to give a profile of the user in order to return items from other RSS feeds that may be of interest. The project was somewhat successful in this aim, as almost every user indicated that some of the returned items were of high interest to them. For Journal TOC data, adoption of the guidelines for provision of RSS feeds created by the ticTOCs project would improve the amount of information available to mine, as abstracts would be more readily available in the feeds. Data would also be presented in a cleaner format, allowing machine to machine interfaces, such as TerMine, better input data.

Specificity

It was generally considered by members of the project team that the results for users whose interest areas were the narrowest were better than those who had wider interest areas. In order to test this theory, following the second iteration, the key-phrases that had produced matches for Approaches A, C and D, were analysed using the Google API. Users whose matched key-phrases were ranked lower in Google tended to have better overall results. Due to the small number of users involved in this trial, the results were not statistically significant, therefore further analysis should follow.

Future Approach

The aim of the Gold Dust project was to investigate methods of pervasively profiling users in order to serve them with other relevant content. Future projects should build on the outcomes of the Gold Dust project by further investigating the use of TerMine together with the Google Search API to create PIPs. This method has had some success in matching highly relevant items to users' profiles, especially where the area of research is narrow and highly specific. The items returned to users were filtered down from over 250,000 unique items from over 3,000 different feeds. Users commented that they would not have found the items of interest from these feeds if it had not been for the Gold Dust methodology. A future system would allow users to train it in order to improve results.

The Role of Information Professionals

Johnson *et al*[31] suggest that the majority users of the Ebling Library[32] did not readily recognise or were not easily able to access RSS feeds for Journal Tables of Contents. Services such as ticTOCs and Ebling Library's E-Journal Feed Service are examples that enhance scholarly literature current awareness and utilise the functionality of RSS but which do not rely on users knowing about, or wanting to use, RSS. This, together with the reported low take up of RSS by end users, reinforces the value of developing these simple user-facing services utilising RSS technology.

---

[31] Johnson SM, Osmond A, Holz RJ (2009) "Developing a current awareness service using really simple syndication (RSS) " *Journal of the Medical Library Association* Vol 97 Issue 1 pp 52–54.

[32] Ebling Library http://ebling.library.wisc.edu/

However, the benefit of RSS to researchers is the time saved in managing and finding new information. Lee *et al*[33] describe some of the additional benefits of RSS in higher education including the low cost, ease of use, potential for productivity gains and the ability to combine, remix and share information from a variety of sources. Information professionals should therefore continue to encourage users to use RSS and where necessary teach them how to use the tools associated with RSS[34].

ABOUT THE AUTHORS:

Lisa J Rogers graduated with MSc with distinction in IT (Multimedia Design) from Heriot-Watt University in Edinburgh and is now a Research Associate at the Institute for Computer Based Learning at Heriot-Watt University. She provided project support for the ticTOCs and Gold Dust projects and is a member of the ticTOCs RSS Working Group which is producing guidelines for publishers on the provision of RSS Feeds. Home Page http://www.macs.hw.ac.uk/~mthljr/

Simon Hodson, D.Phil (Oxon), was Project Manager of the Gold Dust Project based at the University of Hull where he had also been Project Manager and Research Associate on a number of projects including the 'Virtual Research Environment for the History of Political Discourse'. As well as his interest in the use of technology to assist research, he has published in *French History*, *Women's History Review* and is co-editor of *European Political Thought 1450-1700* (Yale, 2008). He is now the JISC Programme Manager with responsibility for the Research Data Programme.
Contact Details: http://www.jisc.ac.uk/contactus/staff/simonhodson.aspx

Roddy MacLeod, MA, DipLib, MCILIP is Senior Subject Librarian at Heriot-Watt University, UK. He edits the Internet Resources Newsletter, manages the TechXtra service (a free service for technology information), and is a past manager of the Pilot Engineering Repository Xsearch (PerX) project. He provided management support to the ticTOCS project, and subject advice to the Gold Dust project. He was Information World Review Information Professional of the Year in 2000. He co-edited the 4th edition of 'Information sources in engineering', published by KG Saur. Home page: http://www.hw.ac.uk/libwww/libram/roddy.html

[33] Lee MJW, Miller C and Newnham L (2008) "RSS and Content Syndication in Higher Education: Subscribing to a New Model of Teaching and Learning" *Educational Media International* Vol 45 No 4 pp 311-322

[34] Mu (2008) "Using RSS feeds and social bookmarking tools to keep current" *Library Hi Tech News* Vol 25 Issue 9 pp 10-11