Project Document Cover Sheet

Project Information			
Project Acronym	MERLIN		
Project Title	Metadata Enrichment for Repositories in a London Institutional Network		
Start Date	01 May 2009	**End Date**	28 February 2011
Lead Institution	UCL (University College London)		
Project Director	Dr Paul Ayris		
Project Manager	Martin Moyle, UCL Library Services		
Partner Institutions	University of London Computing Centre (ULCC)		
Project Web URL	http://www.ucl.ac.uk/ls/merlin		
Programme	Resource discovery strand, Information Environment 09-11		
Programme Manager	Balviar Notay		

Document Name			
Document Title	Final report		
Reporting Period	Final		
Author(s) & project role	Martin Moyle, Project Manager		
Date	August 2011	**Filename**	
URL	http://discovery.ucl.ac.uk/1321559		
Access	General dissemination		

MERLIN
Metadata Enrichment for Repositories in a London Institutional Network

Final project report

Martin Moyle, UCL Library Services

June 2011

Table of Contents

Acknowledgements

MERLIN was funded by the JISC (Joint Information Systems Committee) under the JISC Information Environment Programme 2009-2011.

The project was a collaboration between UCL Library Services and the University of London Computing Centre. The partners worked in association with NaCTeM, the National Centre for Text Mining).

Thanks to Sero Consulting Ltd, particularly Ian Chowchat, Helen Harrop and David Kay, for their contribution to project evaluation. This report makes extensive use of Sero's thoughtful feedback on the MERLIN initiative. Thanks also to Bill Hubbard, Head of the Centre for Research Communications, University of Nottingham, for contributing to the evaluation.

Project Team

- Josh Brown, SHERPA-LEAP Project Officer, UCL Library Services
- Richard Davis, University of London Computing Centre
- Robert Drinkall, UCL Library Services
- Rory McNicholl, University of London Computing Centre
- Martin Moyle, UCL Library Services, Project Manager

Steering Group

- Dr Sophia Ananiadou, Director, National Centre for Text Mining
- Dr Paul Ayris (Chair), Director of UCL Library Services and UCL Copyright Officer
- Dr Jacqueline Cooke, Goldsmiths, University of London
- Richard Davis, University of London Computing Centre
- Michael Day, Research and Development, UKOLN
- Martin Moyle, UCL Library Services
- Christopher Pressler, Director, University of London Research Library Services

Executive summary

The MERLIN project began with some observations which, although untested, rang true: that traditional subject cataloguing is not generally employed in institutional repositories; that, where subject cataloguing does feature, it is employed at such a high level as to offer little to the search experience; and that it is, in any case, usually overshadowed by full-text search. If it were carried out to its full potential, it was felt, detailed subject cataloguing could in principle have value in the repository context; but it is a highly specialised skill, costly to implement and maintain. There is, nonetheless, a gap between the precision offered by traditional library cataloguing services and the blunt instrument of full text repository search, whether carried out locally or through search engines. However, it would be unrealistic to expect humans to bridge this gap with accurate, structured metadata in the 'traditional' manner at repository scale.

MERLIN, therefore, grew out of an interest in exploring the cost-effective integration of automatic subject description in repository services, enriching search without creating significant new resource implications for repositories. Primarily, the project wished to demonstrate opportunities for using the tools produced by NaCTeM, the National Centre for Text Mining, over repository full text. The project additionally sought to explore opportunities for improving repository discovery through interactions between text-mined keywords and on-line thesauri, specifically the HILT tools.

MERLIN used the London Universities aggregation service, LASSO, to demonstrate term extraction and weighting and thesaurus integration with full text repository content. A user interface was developed to support the extended functionality. Formative evaluation, user testing and final evaluation all preceded the release of an open source, re-usable web application, to allow the MERLIN metadata enrichment technology to be incorporated into any repository on any platform.

The project successfully demonstrated an approach to the integration of off-the-shelf text-mining tools into repository search. Text mining was incorporated into the demonstrator, various issues being surmounted along the way, and a fresh interface was delivered to accommodate the enrichments to the search experience. The feasibility of interactions between mined terms in search result sets and external thesauri was also shown. The 'MERLIN tool' is highly adaptable: it supports the integration of any suitably-formatted external thesaurus; and the interface will interact with any RSS-formatted search results. Indeed, the user interface may be implemented without the text mining extensions, if desirable, as a simple enhancement to a repository front-end.

The summative evaluation led to some useful recommendations, among them that the potential for the approach developed by MERLIN to improve the accuracy of truly large-scale search should be explored further.

MERLIN offers the potential for repository content to be enriched with few of the resource overheads traditionally associated with subject cataloguing. It provides a simple means of enhancing repository discovery, capable of bringing both precision and serendipity. The MERLIN tool offers a means by which repository owners can maximise the value of their own content for the benefit of their users.

1. Background

Introduction

MERLIN aimed to test the utility of off-the-shelf text mining tools to enhance resource discovery in institutional repositories. The test environment for MERLIN was LASSO, the London repository consortium SHERPA-LEAP's pilot repository aggregation service. The project also developed and demonstrated a stand-alone version of the MERLIN tool for integration in institutional repositories.

SHERPA-LEAP

The project originated in discussion within SHERPA-LEAP (London E-prints Access Project, a partner in SHERPA). SHERPA-LEAP is a consortium of London-based Higher Education Institutions, founded in 2004 and led by UCL, which helps London's universities to develop and maintain their institutional repositories. Within the LEAP partnership there is substantial diversity of institutional size and mission, ranging from the large, multi-disciplinary and research-led, to the smaller and highly-specialised, and a substantial range of research interests. These differences are reflected in the content of the consortium's repository cross-searching service, LASSO (LEAP Aggregated Search Service On-line), making it an ideal testbed in which to expose and examine issues relating to the application of text mining techniques across institutions and disciplines. LASSO is a simple OAI-PMH-based aggregation service which was developed in 2008 as a demonstrator by UCL Library Services; it offers cross-searching of the institutional repositories of several SHERPA-LEAP member institutions.

Subject cataloguing in the LEAP repositories

The LEAP repositories, and, by extension, the LASSO aggregation service, do not offer a great deal of subject description in their repository metadata. There are several reasons for this. SHERPA-LEAP rejected the idea of a shared subject taxonomy at an early stage, the partners recognising that a shared taxonomy for subject description would have to be so large and unwieldy - supporting research into specialist subjects ranging from clinical biomedicine to ancient South-East Asian cultures - as to be entirely off-putting to depositors, and unworkable by administrators. Subject classification is a specialist and resource-intensive skill, while the highest priority for repository managers, in many cases working to establish their services with piecemeal funding, is often simply the rapid acquisition of content. Many repositories are founded on self-archiving by authors, who cannot be expected to possess library-standard subject classification skills, or the time or inclination to attempt to apply them. The outcome is that even in those SHERPA-LEAP repositories which do employ subject description, resource constraints often mean that structured keywording is only implemented at the highest level, offering little extra benefit to researchers. Meanwhile, full text search, whether carried out locally or via search engines, tends to be used by researchers in preference to more subtle, subject-based approaches to repository discovery.

The MERLIN approach to subject description

MERLIN aimed to investigate and demonstrate the cost-effective integration of automatic subject description in repository services, using off-the-shelf tools, and without creating significant new resource implications for the participating repositories. At the planning stages, several benefits of the MERLIN approach were foreseen:

- improving the discoverability of repository content.
- allowing cost-effective subject description.
- using researchers' own vocabularies.
- catering for interdisciplinarity.
- going beyond simple full-text indexing by bringing selectivity and weight to index terms.
- creating the opportunity to use weighted keyords as the basis for structured navigation.

To investigate and demonstrate this potential, the project aimed to deploy tools produced by NaCTeM, the National Centre for Text Mining, in a repository context. The main demonstration environment for MERLIN was the SHERPA-LEAP aggregation service, LASSO. MERLIN used TerMine term extraction technology to derive terms from full text digital objects held at LASSO's source repositories and, after a weighting process, enriched the LASSO database with these derived keywords to support discovery. In a supplementary strand of the project, MERLIN used the multi-subject terminological cross-searching aids developed by the HILT project to pilot a thesaurus-driven approach to discovery based on the weighted keywords. Appropriate user interface enhancements were designed and tested. Formative evaluation, involving end-users, of the accuracy, usability and efficiency of the automated enhancements to the LASSO aggregator was conducted, and an independent final evaluation was commissioned. Finally, an open source, re-usable web application was developed, to allow the MERLIN metadata enrichment technology to be incorporated into any repository on any platform, and a demonstration of MERLIN in a single, stand-alone repository was constructed.

2. Aims and objectives

The original project aims were:

- To use the TerMine text mining tool to enrich the LASSO repository cross-searching service with weighted keywords automatically derived from source repositories.
- To design and implement modifications to the LASSO interface to surface relevant derived terms at collection, sub-collection and item levels.
- To incorporate automatically-derived terms as a target within the LASSO Advanced Search interface.
- To engage end-users in developmental evaluation of the enhancements to the LASSO service.
- To use HILT resources to construct a pilot navigable subject tree from text-mined keywords, and to present it through the LASSO interface.
- To carry out a full evaluation of the MERLIN enhancements to repository discoverability.
- To make the MERLIN enrichment technology available as a reusable, platform-neutral, open source web application.

These aims, in essence, remained unchanged in the course of the project.

3. Methodology

In the spirit of the JISC Call, which emphasised the use of off-the-shelf tools, the MERLIN technical methodology was relatively straightforward. The main methodological stages were:

1) Identification of relevant NaCTeM products and their integration in the LASSO service.
2) Retrieval of full text into the LASSO environment (LASSO being a metadata harvesting service).
3) Experimental integration of HILT thesaurus with the text-mined terms.
4) Design and testing of revised user interface, to accommodate search enrichments.
5) Appointment of external evaluator and facilitation of their work.
6) Development of stand-alone version of MERLIN, for plug-in to individual repositories.
7) Public release of MERLIN code.

The project used an iterative, agile development methodology, prototyping each deliverable and then refining it in response to feedback and/or evaluation. The project aimed to utilise the many language-processing tools provided by or recommended by NaCTeM, and consulted closely with NaCTeM on best practice and ongoing developments in the field. A Google Code project environment, including a version control system, was created to support the open development of the MERLIN code, under an open licence.

4. Implementation

Figure 1 gives a simple overview of the architecture that the team set out to implement. MERLIN's technical 'back-end' is bounded in red.

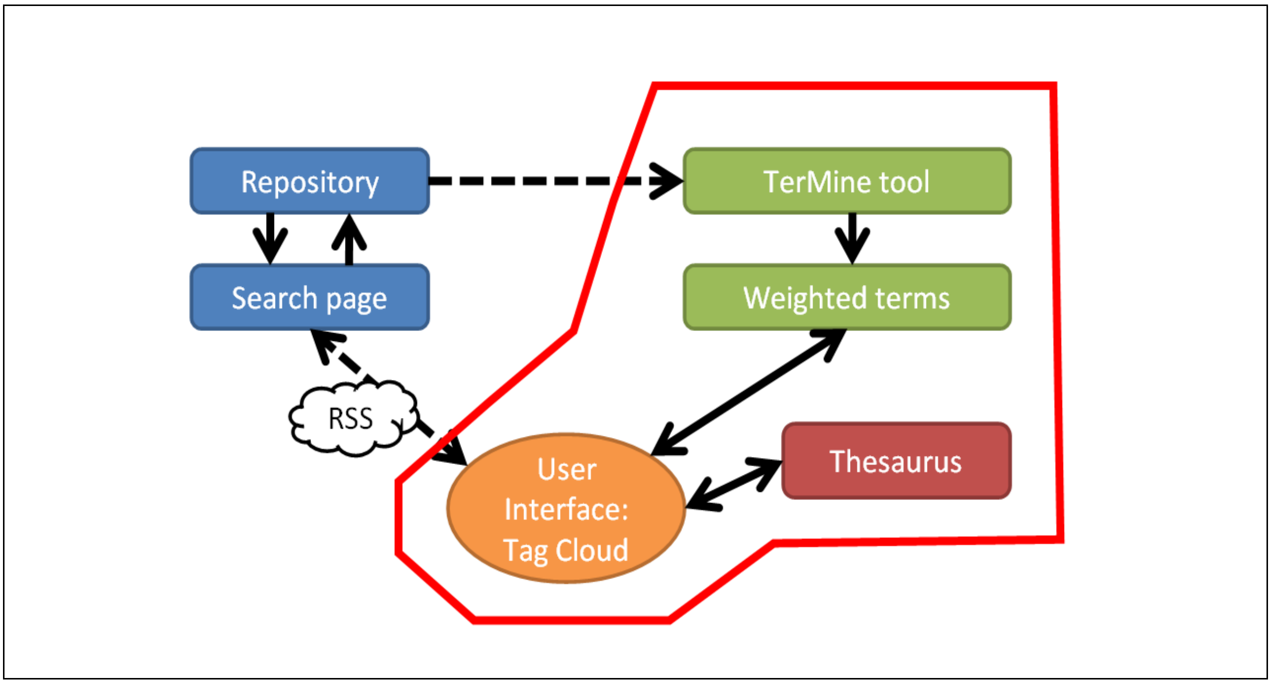


Fig.1. Basic MERLIN architecture (Josh Brown, 2010)

This report chapter describes the team's work in term acquisition; tuning the mined terms; thesaurus integration; interface design; and stand-alone and mobile code release.

4.1. Term acquisition

The acquisition of mined terms consisted of the following steps:

- Collecting full text files from the source repositories
- Extracting plain text from formatted documents
- Processing the text using the Termine text mining service, returning XML
- Storing mined terms in a MERLIN terms database

Digital object retrieval

To collect full text files for mining, the daily OAI-PMH metadata harvesting schedule by which data is acquired for LASSO was extended. As part of the regular LASSO metadata ingest, a test for the presence of <dc:format> is carried out as a simple guide to whether or not the repository object is a full-text resource, and a flag is assigned. The MERLIN extension checks for this flag and, where appropriate, visits the source repository to retrieve the full text.

An initial problem here was the inconsistent use of OAI Dublin Core across the source repositories: some contributors use the <dc:identifier> element for a full text URL, whereas some use <dc:relation>; some do not include a direct URL to the full text in their OAI export metadata; and some contributors include repository items that are under embargo, with some URLs resolving to HTML login pages for staff-only items. With the exception of the latter, these issues were largely surmounted by the use of METS as the metadata format for the MERLIN harvesting stage, which generally provided an unambiguous URL for each resource and enabled file retrieval.

Conversion to text

Three programmes were installed to facilitate the extraction of full text from formatted documents:

1. antiword (http://www.winfield.demon.nl/) for MSWord docs

2. pdftotext (part of xpdf) for PDF documents - http://www.foolabs.com/xpdf/

3. Java OpenDocument Converter for other formats - http://www.artofsolving.com/opensource/jodconverter

(In theory, JODConverter precluded the need for antiword. However, as antiword is a one-step solution, it was felt that defaulting to it for MSWord documents would limit the opportunities for corruption to occur.)

Few difficulties were experienced here, the most common being the failure of pdf conversion when non-text binaries - often scanned documents presented as PDFs, rather than born-digital PDFs - were encountered.

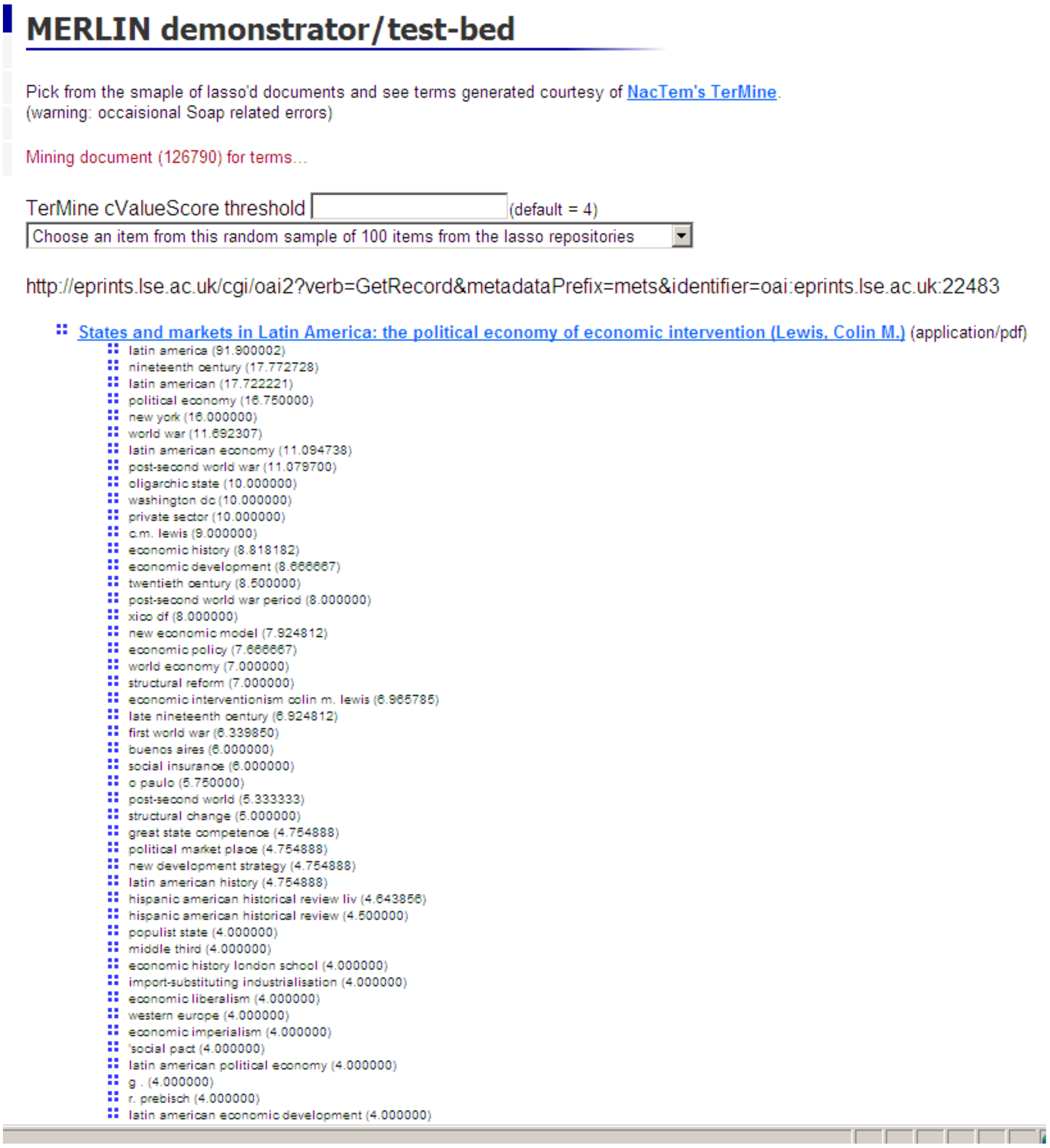
TerMine processing

The text mining stage of MERLIN content acquisition used tools developed and made publicly available by NaCTeM. The plain text files created in the MERLIN environment by the two preceding stages are passed to the NacTeM sentence splitter, which employs heuristic rules for identifying the boundaries of sentences and paragraphs. The resulting chunks of text are passed to the TerMine service, which recognises terms and applies statistical analysis to derive ‘weights’, showing the relative importance of each term in its parent document.

(For interest, the TerMine output from a text-only version of this final report is included at Appendix B.)

MERLIN uses the TerMine SOAP service. Although the sentence splitter is also available as a web service, the team found that in this case a local installation delivered more consistent results

A rough-and-ready tool was created to give a browser view of the output from the NaCTeM tools in the MERLIN context, to help the project team to understand the technical processes and challenges. The tool (Figure 2) displays a random list of full-text records from the LASSO database, from which the user may select a publication to be mined for terms and their relative weights. Figure 2 shows a typical response. The TerMine score threshold - the minimum weight - may be changed.



MERLIN demonstrator/test-bed

Pick from the smaple of lasso'd documents and see terms generated courtesy of NacTem's TerMine.
(warning: occaisional Soap related errors)

Mining document (126790) for terms...

TerMine cValueScore threshold (default = 4)

Choose an item from this random sample of 100 items from the lasso repositories

http://eprints.lse.ac.uk/cgi/oai2?verb=GetRecord&metadataPrefix=mets&identifier=oai:eprints.lse.ac.uk:22483

- States and markets in Latin America: the political economy of economic intervention (Lewis, Colin M.) (application/pdf)
  - latin america (91.900002)
  - nineteenth century (17.772728)
  - latin american (17.722221)
  - political economy (16.750000)
  - new york (16.000000)
  - world war (11.692307)
  - latin american economy (11.094738)
  - post-second world war (11.079700)
  - oligarchic state (10.000000)
  - washington dc (10.000000)
  - private sector (10.000000)
  - c.m. lewis (9.000000)
  - economic history (8.818182)
  - economic development (8.666667)
  - twentieth century (8.500000)
  - post-second world war period (8.000000)
  - xico df (8.000000)
  - new economic model (7.924812)
  - economic policy (7.666667)
  - world economy (7.000000)
  - structural reform (7.000000)
  - economic interventionism colin m. lewis (6.965785)
  - late nineteenth century (6.924812)
  - first world war (6.339850)
  - buenos aires (6.000000)
  - social insurance (6.000000)
  - o paulo (5.750000)
  - post-second world (5.333333)
  - structural change (5.000000)
  - great state competence (4.754888)
  - political market place (4.754888)
  - new development strategy (4.754888)
  - latin american history (4.754888)
  - hispanic american historical review liv (4.643856)
  - hispanic american historical review (4.500000)
  - populist state (4.000000)
  - middle third (4.000000)
  - economic history london school (4.000000)
  - import-substituting industrialisation (4.000000)
  - economic liberalism (4.000000)
  - western europe (4.000000)
  - economic imperialism (4.000000)
  - 'social pact (4.000000)
  - latin american political economy (4.000000)
  - g . (4.000000)
  - r. prebisch (4.000000)
  - latin american economic development (4.000000)

Fig.2. Browser view of TerMine output from sample document. See http://lasso.ucl.ac.uk/merlin/index.php

Term storage

Mined terms are stored in MySQL tables in a simple extension to the LASSO schema. To help performance, two tables are used. One table holds terms, and a second stores details of links between the terms and the full text document records in LASSO. The second table also stores the TerMine c_value (weighting) for every term in each record.

4.2. Noise reduction and tuning

TerMine, naturally, mines documents for all the text it can find. In a repository context, this can lead to supererogatory and/or unhelpful results. One obvious example is TerMine's processing of references lists at the end of research publications, with journal abbreviations in particular causing the return of large amounts of unhelpful results. Efforts to improve the quality of the text-mining integration were made in three areas.

1. Pre-processing

A certain amount of pre-processing was implemented for PDF documents, prior to their exposure to TerMine. A combination of pdf2html and XML DOM manipulation was used to implement the following:

- identification and removal of boilerplate text

- identification and removal of any tabular data
- identification of any “References” headings and the removal of any following text

2. Post-processing

Some post-processing of TerMine data was undertaken, namely the identification and truncation of ‘overlong’ (according to local criteria) terms with repeated words. Regex was used for this cleanup.

Consideration was given to boosting the TerMine scores with ‘secondary weighting’, specifically to give more prominence to terms found in titles or abstracts. Such terms are flagged as part of the data acquisition process, but that information is not currently used to apply any further uplift to recorded values in the demonstrator.

3. Final results from text mining

A snapshot of the enriched LASSO database, after the introduction of PDF cleanup, gives the following figures (rounded):

- 15,000 records are flagged as having full text at source
- 10,000 of those texts have been successfully mined by TerMine (the remainder being embargoed documents, non-text binaries, or incorrectly identified as ‘full text’ by LASSO)
- TerMine has derived 650,000 unique terms from those 10,000 records
- There are 1,000,000 associations between TerMine terms and LASSO records

4. TerMine across multiple documents

TerMine is designed for single documents. The scores that it produces are relative to the document in which a term is found. A frequently-repeated term in a long document may be assigned a high weight; by contrast, a term which is highly important in a different document may acquire a lower weight. A comparison of scores across a set of search results is, in this sense, meaningless.

This presented a normalisation challenge. The solution implemented was a simple one: all the documents that have contributed terms to a set of search results are sampled, and the *n* top-scoring terms from each document are pooled into the cloud. This helps to engineer a more useful and representative cloud for a multiple document set.

Naturally, when the user drills down to view the terms for any single document, TerMine is serving its original purpose, and the scores shown have true statistical meaning.

4.3. Thesaurus integration

The team undertook some experimentation with the HILT tools to investigate the addition of structure to the text-mined search experience.

The system was originally configured to make CURL requests to the HILT SRU/W server, one for each set of thesaurus terms (broader, narrower, etc). The subsequent long-term unavailability of the SRU/W server prompted a redesign, substituting a CURL request to the HILT SOAP client. This also brought a simplified final model, as follows:

1. User enters a search term
2a. LASSO is searched, including the mined terms
2b. The search terms are incorporated in a CURL request to the HILT SOAP client
3. Associated concepts, broader terms, narrower terms and related terms are retrieved from HILT. The MERLIN demonstrator uses the UNESCO thesaurus.
4. When a user selects ‘Thesaurus’ mode, broader, narrower and related terms are made available in the cloud.

The data from the HILT response is stored as a Javascript object.

In principle, any XML-represented thesaurus could be incorporated in MERLIN in this way.

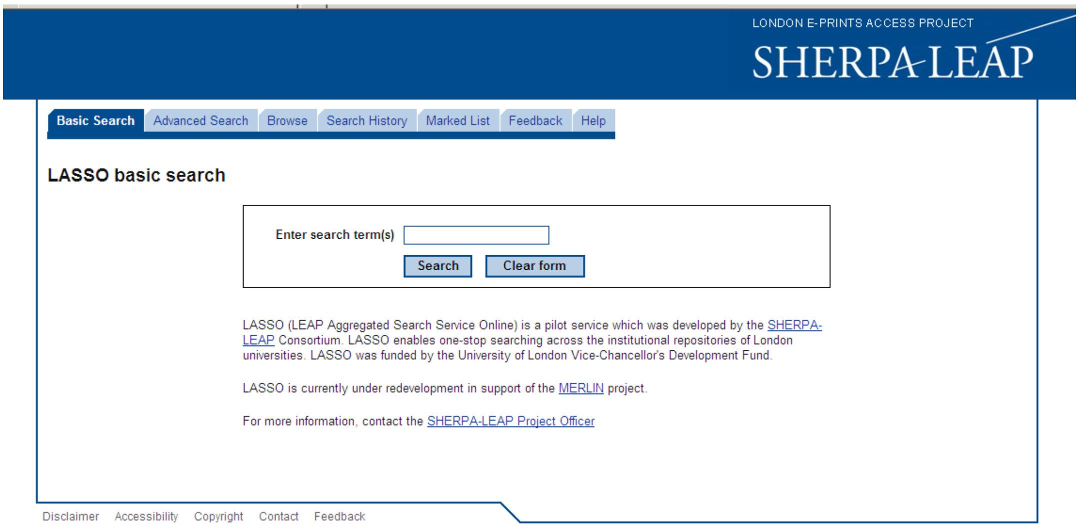
4.4. Interface design, testing and refinement

The team considered ways of incorporating the extracted and weighted keywords, and the thesaurus integration work, into the LASSO discovery service. The team was conscious of the fact that LASSO is an aggregation service, suitable for demonstrating the MERLIN technology but with certain points of difference from the stand-alone Institutional Repository that ultimately the MERLIN work might benefit, and so aimed to focus on the generic issues around the most useful and user-friendly exposure of the derived terms in repository search.

An interface was designed, tested, re-developed and finalised. This report section describes that process in more detail.

The LASSO interface

The LASSO demonstrator used to pilot MERLIN originally offered a simple, clean interface (fig.3), augmented by a more detailed 'advanced' search (fig.4) and various browse indexes. Conventional, sortable results lists were supplied in response to searches (fig.5).



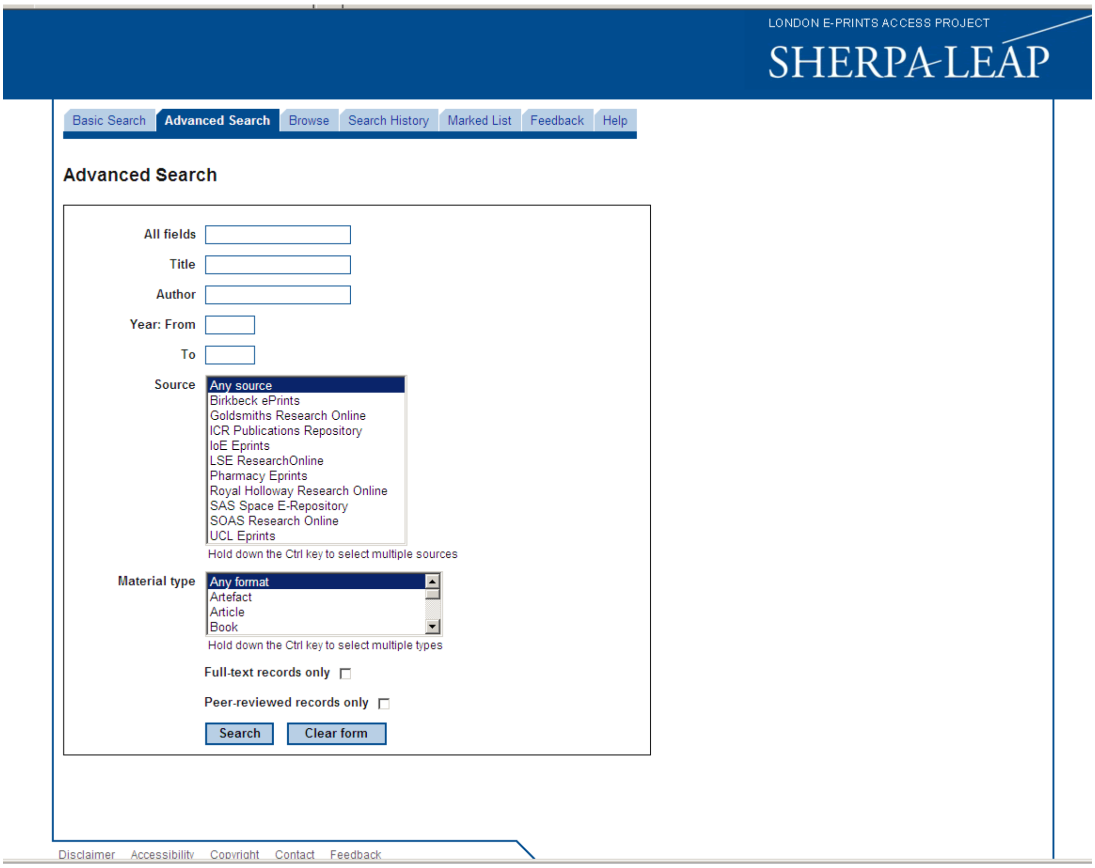
LASSO basic search

LASSO (LEAP Aggregated Search Service Online) is a pilot service which was developed by the SHERPA-LEAP Consortium. LASSO enables one-stop searching across the institutional repositories of London universities. LASSO was funded by the University of London Vice-Chancellor's Development Fund.

LASSO is currently under redevelopment in support of the MERLIN project.

For more information, contact the SHERPA-LEAP Project Officer

Fig.3. LASSO basic search, pre-MERLIN



Advanced Search

All fields

Title

Author

Year: From

To

Source
- Any source
- Birkbeck ePrints
- Goldsmiths Research Online
- ICR Publications Repository
- IoE Eprints
- LSE ResearchOnline
- Pharmacy Eprints
- Royal Holloway Research Online
- SAS Space E-Repository
- SOAS Research Online
- UCL Eprints

Hold down the Ctrl key to select multiple sources

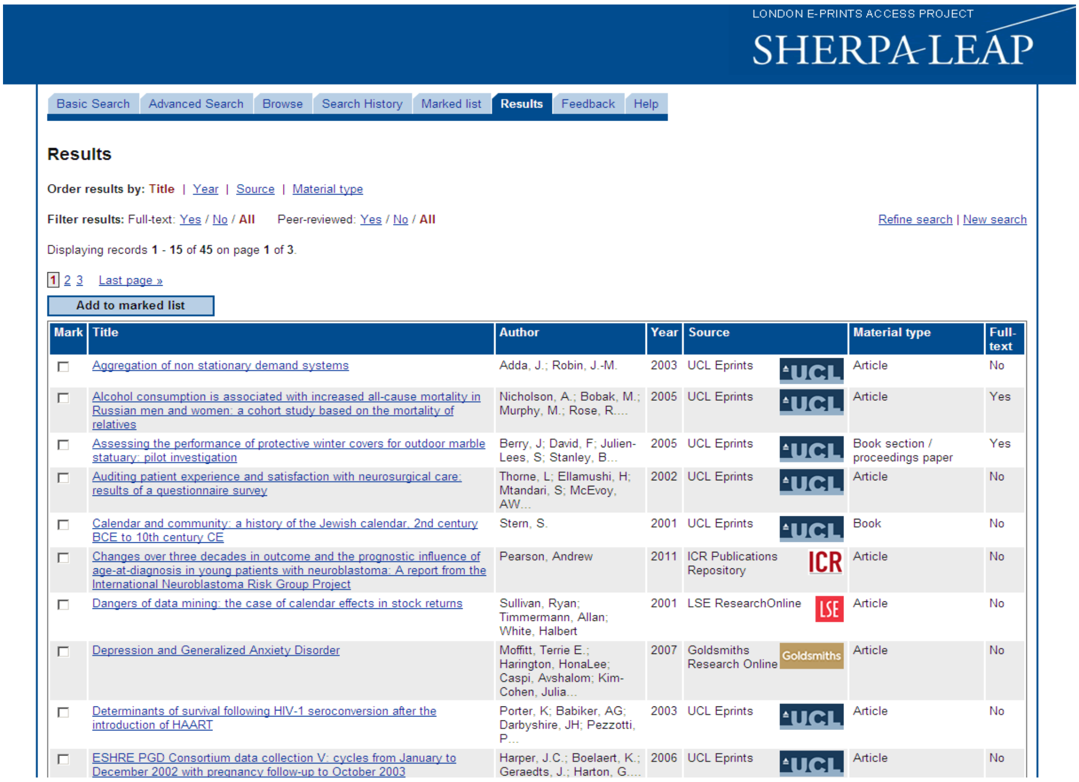
Material type
- Any format
- Artefact
- Article
- Book

Hold down the Ctrl key to select multiple types

Full-text records only □

Peer-reviewed records only □

Fig.4. LASSO 'advanced' search, pre-MERLIN



Results

Order results by: Title | Year | Source | Material type

Filter results: Full-text: Yes / No / **All** Peer-reviewed: Yes / No / **All** Refine search | New search

Displaying records **1** - **15** of **45** on page **1** of 3.

1 2 3 Last page »

Mark	Title	Author	Year	Source	Material type	Full-text
□	Aggregation of non stationary demand systems	Adda, J.; Robin, J.-M.	2003	UCL Eprints	Article	No
□	Alcohol consumption is associated with increased all-cause mortality in Russian men and women: a cohort study based on the mortality of relatives	Nicholson, A.; Bobak, M.; Murphy, M.; Rose, R....	2005	UCL Eprints	Article	Yes
□	Assessing the performance of protective winter covers for outdoor marble statuary: pilot investigation	Berry, J; David, F; Julien-Lees, S; Stanley, B...	2005	UCL Eprints	Book section / proceedings paper	Yes
□	Auditing patient experience and satisfaction with neurosurgical care: results of a questionnaire survey	Thorne, L; Ellamushi, H; Mtandari, S; McEvoy, AW...	2002	UCL Eprints	Article	No
□	Calendar and community: a history of the Jewish calendar, 2nd century BCE to 10th century CE	Stern, S.	2001	UCL Eprints	Book	No
□	Changes over three decades in outcome and the prognostic influence of age-at-diagnosis in young patients with neuroblastoma: A report from the International Neuroblastoma Risk Group Project	Pearson, Andrew	2011	ICR Publications Repository	Article	No
□	Dangers of data mining: the case of calendar effects in stock returns	Sullivan, Ryan; Timmermann, Allan; White, Halbert	2001	LSE ResearchOnline	Article	No
□	Depression and Generalized Anxiety Disorder	Moffitt, Terrie E.; Harington, HonaLee; Caspi, Avshalom; Kim-Cohen, Julia...	2007	Goldsmiths Research Online	Article	No
□	Determinants of survival following HIV-1 seroconversion after the introduction of HAART	Porter, K; Babiker, AG; Darbyshire, JH; Pezzotti, P...	2003	UCL Eprints	Article	No
□	ESHRE PGD Consortium data collection V: cycles from January to December 2002 with pregnancy follow-up to October 2003	Harper, J.C.; Boelaert, K.; Geraedts, J.; Harton, G....	2006	UCL Eprints	Article	No

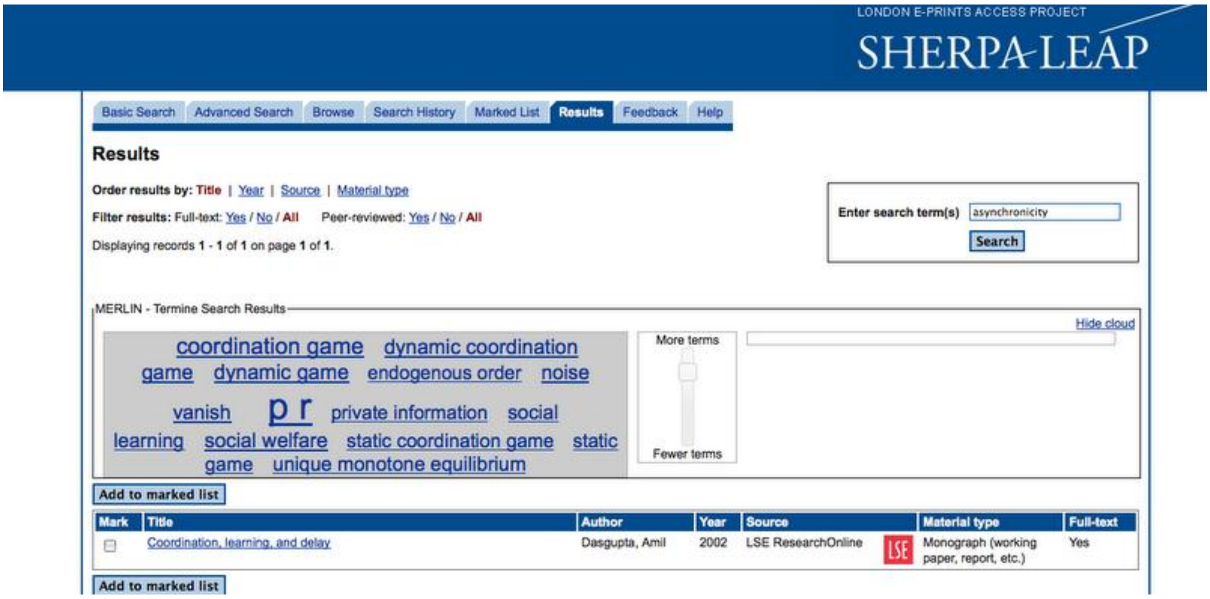
Fig.5. LASSO results, pre-MERLIN

Various issues were considered by the team, including how to offer search against mined terms; where to expose the terms in search results; the display of weighting thresholds and the extent to which they should be manipulable by the searcher; the merits of clouds and other visualisations against text-based results; the merits of aggregated and item-level term presentation; and how best to connect the user, the text mining results, and the integrated thesaurus; all the while providing the researcher with a useful and intuitive experience.

Interim interface

An interim interface was developed by the team, as a first response to some of these questions (it predated the thesaurus work). This is shown in figures 6-9.

Figure 6 shows a search of metadata within LASSO for the word ‘asynchronicity’ returning one record (foot of screen). In the left-hand middle window, a number of associated terms, based on stored Termine algorithmic analysis of the full text, are displayed. ‘Hide cloud’ removes the text-mining add-on.



Results

Order results by: Title | Year | Source | Material type

Filter results: Full-text: Yes / No / All Peer-reviewed: Yes / No / All

Displaying records 1 - 1 of 1 on page 1 of 1.

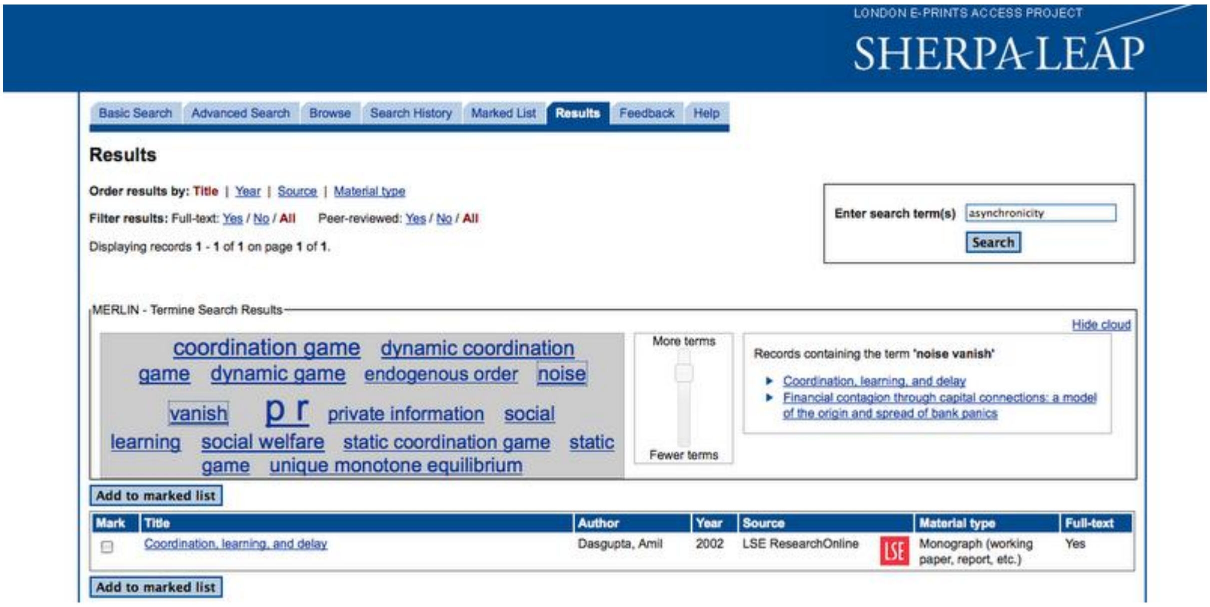
Enter search term(s) asynchronicity

MERLIN - Termine Search Results

Mark	Title	Author	Year	Source	Material type	Full-text
□	Coordination, learning, and delay	Dasgupta, Amil	2002	LSE ResearchOnline LSE	Monograph (working paper, report, etc.)	Yes

Fig.6. Interim interface, default search results.

Clicking on any cloud term returns a list of records whose full text also contains that term. These are displayed to the right of the cloud and slider (Fig.7).



Results

Order results by: Title | Year | Source | Material type

Filter results: Full-text: Yes / No / **All**   Peer-reviewed: Yes / No / **All**

Displaying records **1** - **1** of **1** on page **1** of **1**.

Enter search term(s) asynchronicity

MERLIN - Termine Search Results

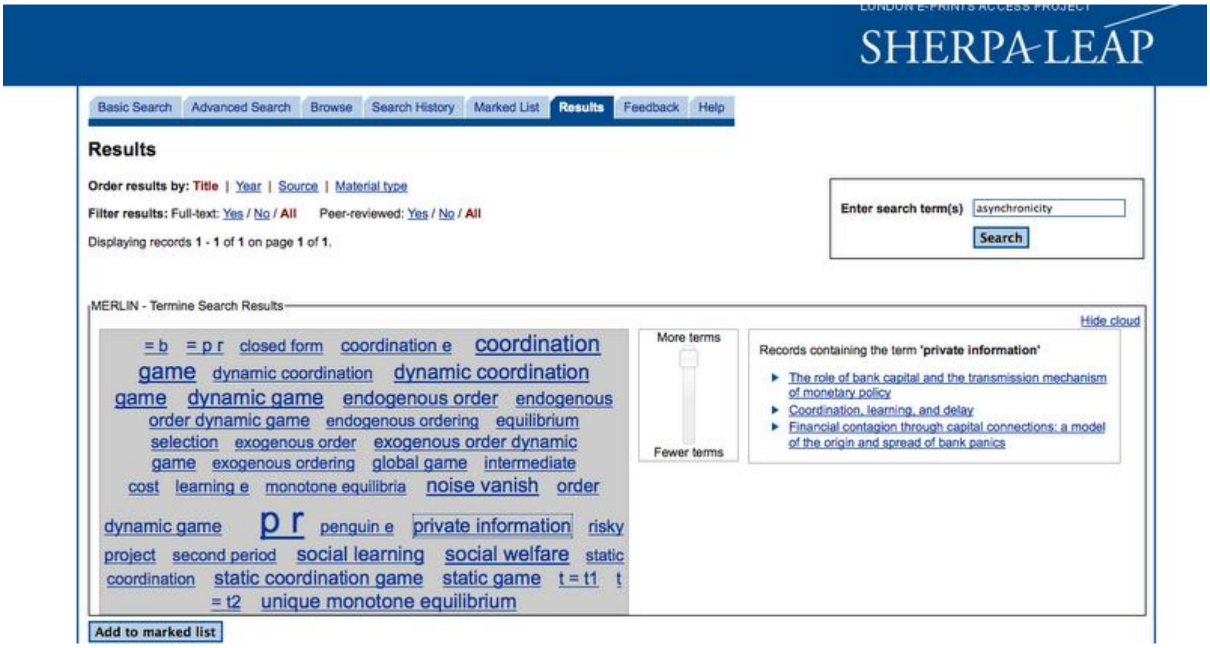
Records containing the term **'noise vanish'**

- Coordination, learning, and delay
- Financial contagion through capital connections: a model of the origin and spread of bank panics

Mark	Title	Author	Year	Source	Material type	Full-text
□	Coordination, learning, and delay	Dasgupta, Amil	2002	LSE ResearchOnline LSE	Monograph (working paper, report, etc.)	Yes

Fig.7. Interim interface, ‘lateral’ search from text-mined terms.

The central slider allows the user to filter terms by raising or lowering the TerMine score threshold for the terms associated with the result set (one document, in this example). Figure 8 shows the same records with ‘More terms’ requested, lowering the threshold to allow terms with lighter weights into the interface.



Results

Order results by: Title | Year | Source | Material type

Filter results: Full-text: Yes / No / **All**   Peer-reviewed: Yes / No / **All**

Displaying records **1** - **1** of **1** on page **1** of **1**.

Enter search term(s) asynchronicity

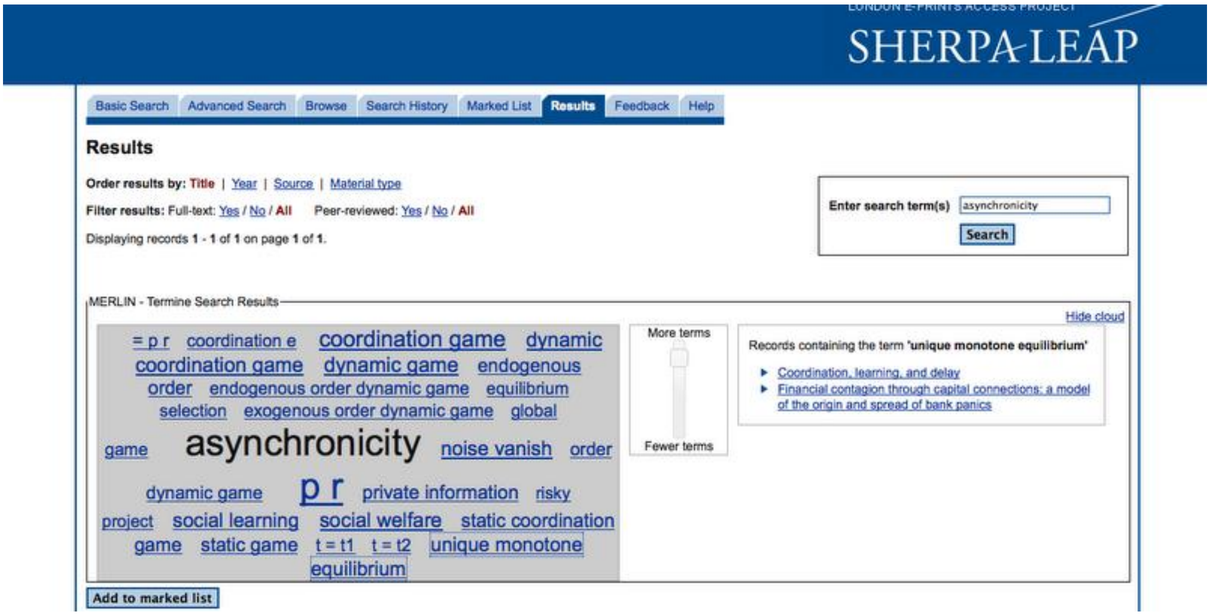
MERLIN - Termine Search Results

Records containing the term **'private information'**

- The role of bank capital and the transmission mechanism of monetary policy
- Coordination, learning, and delay
- Financial contagion through capital connections: a model of the origin and spread of bank panics

Fig.8. Interim interface: effect of requesting ‘More terms’

In a final refinement of the interim service, it was arranged for the original search string to be displayed, emboldened, in a fixed position in the cloud (Figure 9) throughout the search session.



Results

Order results by: Title | Year | Source | Material type

Filter results: Full-text: Yes / No / All Peer-reviewed: Yes / No / All

Displaying records 1 - 1 of 1 on page 1 of 1.

Enter search term(s) asynchronicity

MERLIN - Termine Search Results

Records containing the term **'unique monotone equilibrium'**

- Coordination, learning, and delay
- Financial contagion through capital connections: a model of the origin and spread of bank panics

Fig.9. Interim interface with original search term in cloud centre

User testing

To help to validate the interim demonstration interface, a user testing session was arranged in conjunction with the UCL Department of Information Studies.

The questions and tasks from the user testing session are documented at Appendix A. A summary of the results follows.

The session had 12 participants, 10 of whom had never heard of MERLIN or LASSO, and none of whom had significant prior knowledge of text mining.

The tasks undertaken by the users produced consistent results, so verifying basic search and retrieval.

A few disliked the cloud *prima facie* - it was described as 'annoying' and 'confusing' - but the majority took it in their stride. The slider was readily understood. The click-through options were easily grasped, with almost all criticism levelled at the appearance of the text box - better, bolder display required.

Half found it easy to navigate the site at once, and three found it easy once they had become accustomed to the cloud. The remainder continued to find the cloud an impediment. By the end of the process, eight participants admitted to finding the tag cloud helpful, albeit with reservations, not least a need for more explanation of how to use it. Nine liked the slider. The users liked the results display, but there was a lack of clarity about the interaction between the cloud and the results list. All remained unclear about the text mining process underpinning the service, with most participants assuming that the results came from tagging, metadata or a thesaurus.

Generally, the feedback from the user testing was positive and supportive, although most of the participants felt, correctly, that the interface needed more work.

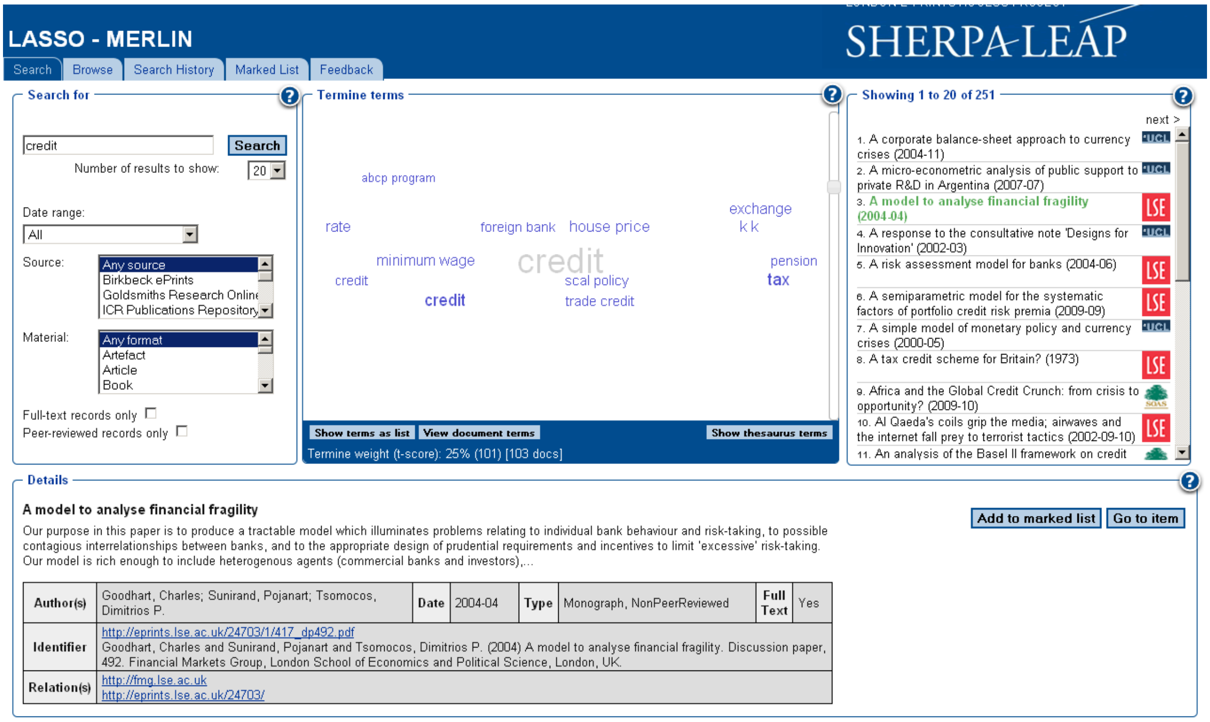
At this stage in the development, an informal evaluation involving members of the SHERPA-LEAP repository network was undertaken. This also provided valuable feedback, to a large extent echoing the sentiments of the cohort of test users.

Final interface

The MERLIN demonstrator interface was finalised and overlaid onto LASSO in response to the developmental evaluation described in the preceding section. The final project UI is available through the project web site. An overview is given here.

Searching, results and TerMine interaction

Figure 10 shows the results of a simple search in the redesigned interface.



1. A corporate balance-sheet approach to currency crises (2004-11)
2. A micro-econometric analysis of public support to private R&D in Argentina (2007-07)
3. **A model to analyse financial fragility (2004-04)**
4. A response to the consultative note 'Designs for Innovation' (2002-03)
5. A risk assessment model for banks (2004-06)
6. A semiparametric model for the systematic factors of portfolio credit risk premia (2009-09)
7. A simple model of monetary policy and currency crises (2000-05)
8. A tax credit scheme for Britain? (1973)
9. Africa and the Global Credit Crunch: from crisis to opportunity? (2009-10)
10. Al Qaeda's coils grip the media; airwaves and the internet fall prey to terrorist tactics (2002-09-10)
11. An analysis of the Basel II framework on credit

A model to analyse financial fragility

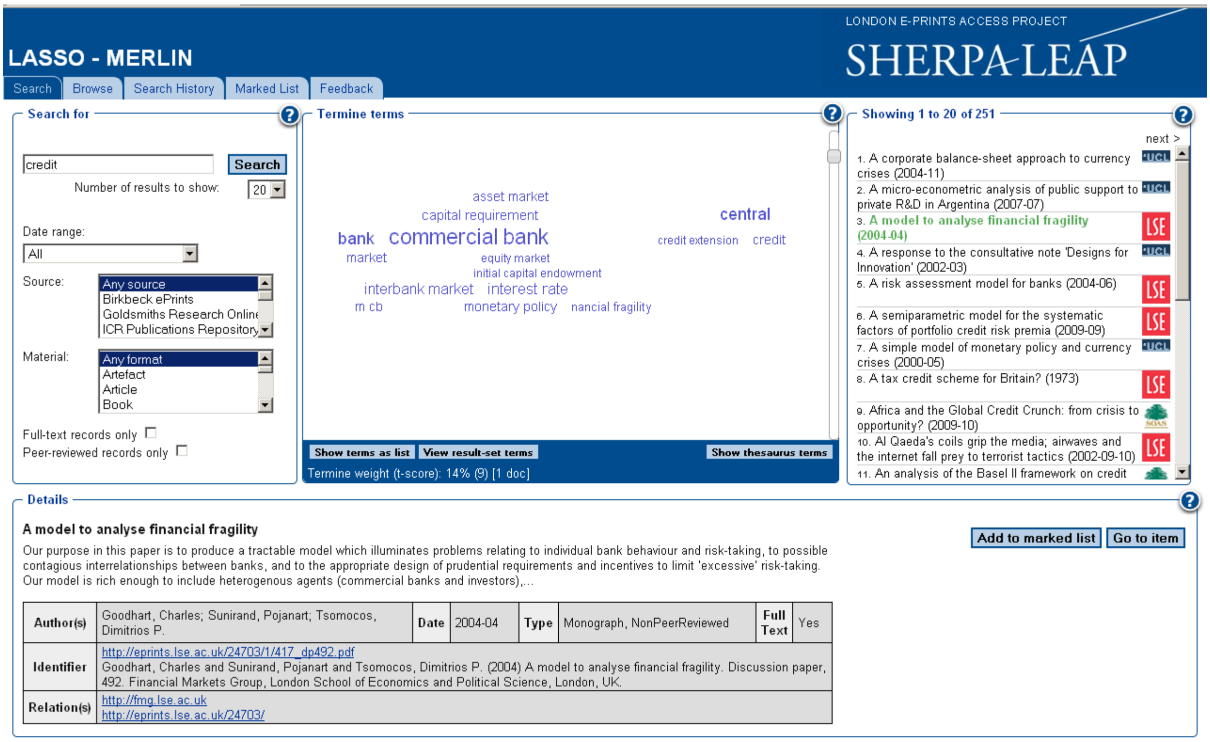
Our purpose in this paper is to produce a tractable model which illuminates problems relating to individual bank behaviour and risk-taking, to possible contagious interrelationships between banks, and to the appropriate design of prudential requirements and incentives to limit 'excessive' risk-taking. Our model is rich enough to include heterogenous agents (commercial banks and investors),...

Author(s)	Goodhart, Charles; Sunirand, Pojanart; Tsomocos, Dimitrios P.	**Date**	2004-04	**Type**	Monograph, NonPeerReviewed	**Full Text**	Yes
Identifier	http://eprints.lse.ac.uk/24703/1/417_dp492.pdf Goodhart, Charles and Sunirand, Pojanart and Tsomocos, Dimitrios P. (2004) A model to analyse financial fragility. Discussion paper, 492. Financial Markets Group, London School of Economics and Political Science, London, UK.						
Relation(s)	http://fmg.lse.ac.uk http://eprints.lse.ac.uk/24703/						

Fig.10. MERLIN - search results.

The left-hand panel conflates the Simple and Advanced search pages of the original LASSO service. The results set is shown on the right. When documents from the results set are highlighted, further details are shown below. The central panel is reserved for the display of terms mined from the full text documents in the results set. In Figure10, TerMine found 103 ‘mineable’ documents in the results set.

It is possible to toggle between terms mined from the whole results set, and terms mined from a single document. It is also possible to use the slider to raise or lower the minimum TerMine strength of the terms displayed. Figure 11 shows only the terms from the selected document, with the TerMined threshold lowered to include more terms.



credit

Number of results to show: 20

Date range: All

Source: Any source, Birkbeck ePrints, Goldsmiths Research Online, ICR Publications Repository

Material: Any format, Artefact, Article, Book

Full-text records only ☐

Peer-reviewed records only ☐

Termine weight (t-score): 14% (9) [1 doc]

Showing 1 to 20 of 251

next >

1. A corporate balance-sheet approach to currency crises (2004-11)
2. A micro-econometric analysis of public support to private R&D in Argentina (2007-07)
3. A model to analyse financial fragility (2004-04)
4. A response to the consultative note 'Designs for Innovation' (2002-03)
5. A risk assessment model for banks (2004-06)
6. A semiparametric model for the systematic factors of portfolio credit risk premia (2009-09)
7. A simple model of monetary policy and currency crises (2000-05)
8. A tax credit scheme for Britain? (1973)
9. Africa and the Global Credit Crunch: from crisis to opportunity? (2009-10)
10. Al Qaeda's coils grip the media; airwaves and the internet fall prey to terrorist tactics (2002-09-10)
11. An analysis of the Basel II framework on credit

Details

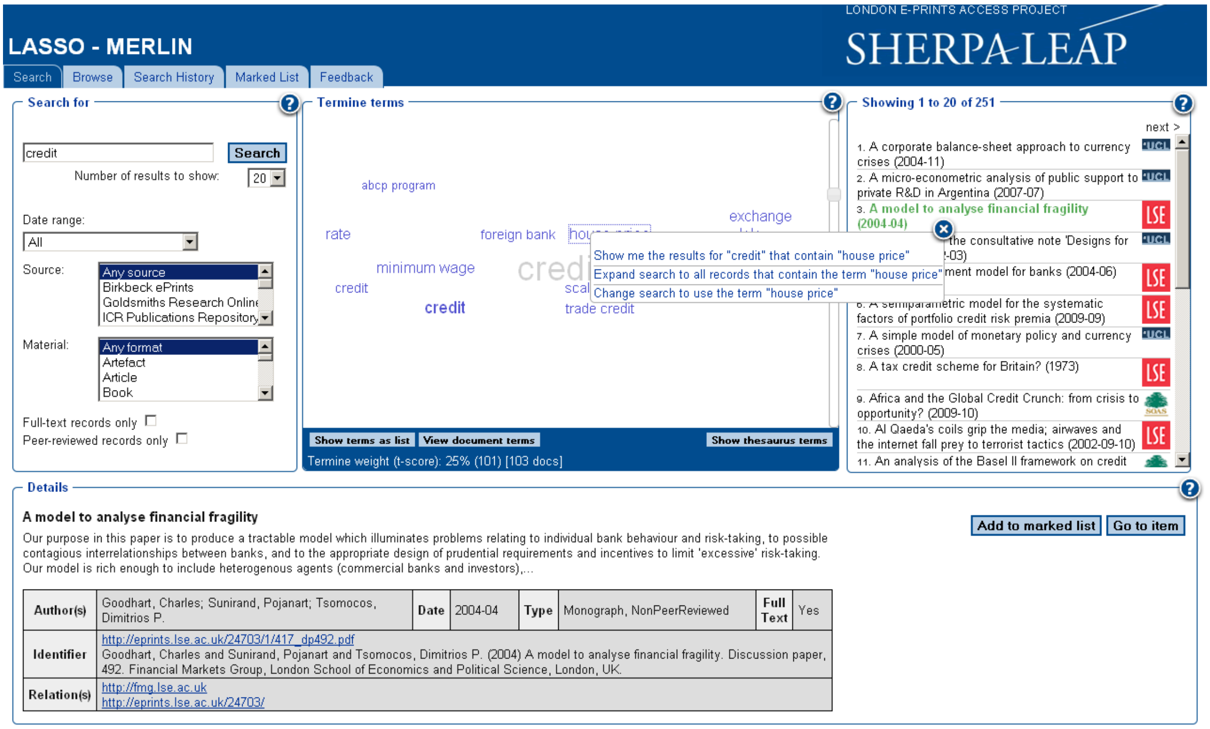
A model to analyse financial fragility

Our purpose in this paper is to produce a tractable model which illuminates problems relating to individual bank behaviour and risk-taking, to possible contagious interrelationships between banks, and to the appropriate design of prudential requirements and incentives to limit 'excessive' risk-taking. Our model is rich enough to include heterogenous agents (commercial banks and investors),...

Author(s)	Goodhart, Charles; Sunirand, Pojanart; Tsomocos, Dimitrios P.	**Date**	2004-04	**Type**	Monograph, NonPeerReviewed	**Full Text**	Yes
Identifier	http://eprints.lse.ac.uk/24703/1/417_dp492.pdf Goodhart, Charles and Sunirand, Pojanart and Tsomocos, Dimitrios P. (2004) A model to analyse financial fragility. Discussion paper, 492. Financial Markets Group, London School of Economics and Political Science, London, UK.						
Relation(s)	http://fmg.lse.ac.uk http://eprints.lse.ac.uk/24703/						

Fig.11. Single document terms, threshold lowered

Any term in the cloud may be clicked, to offer options for narrowing, expanding or exchanging the original search (Figure 12).



credit

Number of results to show: 20

Date range: All

Source: Any source, Birkbeck ePrints, Goldsmiths Research Online, ICR Publications Repository

Material: Any format, Artefact, Article, Book

Full-text records only ☐

Peer-reviewed records only ☐

Show me the results for "credit" that contain "house price"

Expand search to all records that contain the term "house price"

Change search to use the term "house price"

Termine weight (t-score): 25% (101) [103 docs]

Showing 1 to 20 of 251

next >

1. A corporate balance-sheet approach to currency crises (2004-11)
2. A micro-econometric analysis of public support to private R&D in Argentina (2007-07)
3. A model to analyse financial fragility (2004-04)
the consultative note 'Designs for
-03)
ment model for banks (2004-06)
6. A semiparametric model for the systematic factors of portfolio credit risk premia (2009-09)
7. A simple model of monetary policy and currency crises (2000-05)
8. A tax credit scheme for Britain? (1973)
9. Africa and the Global Credit Crunch: from crisis to opportunity? (2009-10)
10. Al Qaeda's coils grip the media; airwaves and the internet fall prey to terrorist tactics (2002-09-10)
11. An analysis of the Basel II framework on credit

Details

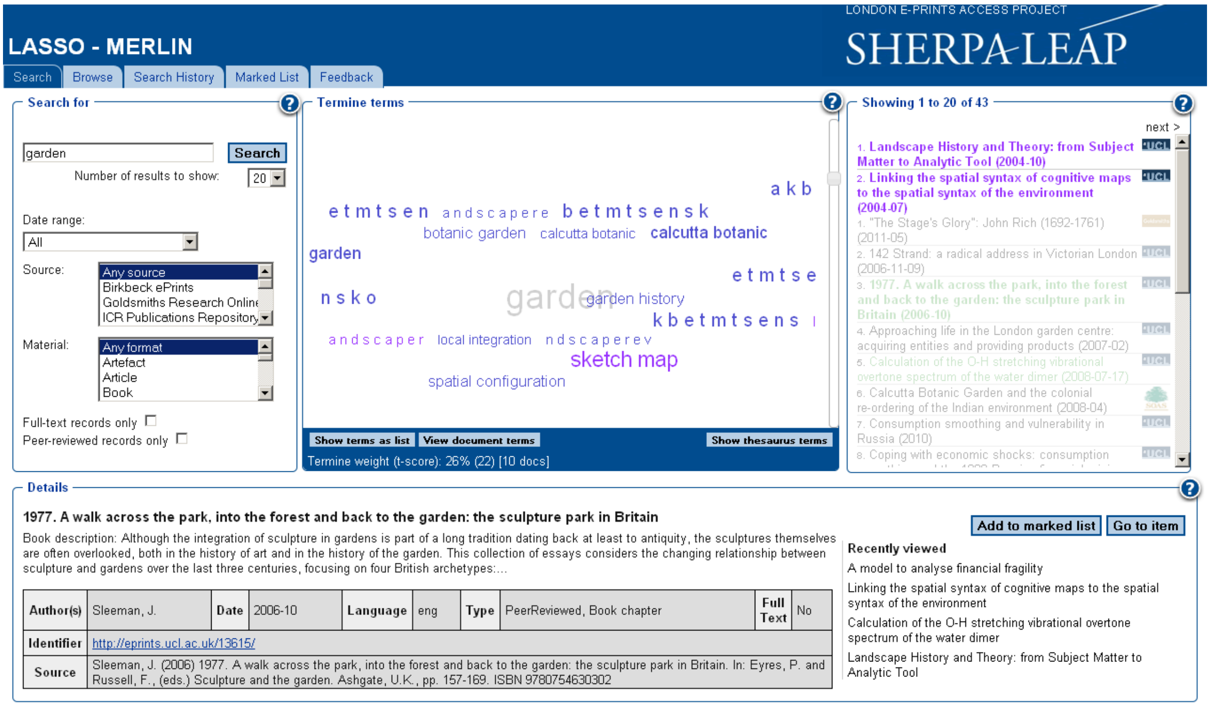
A model to analyse financial fragility

Our purpose in this paper is to produce a tractable model which illuminates problems relating to individual bank behaviour and risk-taking, to possible contagious interrelationships between banks, and to the appropriate design of prudential requirements and incentives to limit 'excessive' risk-taking. Our model is rich enough to include heterogenous agents (commercial banks and investors),...

Author(s)	Goodhart, Charles; Sunirand, Pojanart; Tsomocos, Dimitrios P.	**Date**	2004-04	**Type**	Monograph, NonPeerReviewed	**Full Text**	Yes
Identifier	http://eprints.lse.ac.uk/24703/1/417_dp492.pdf Goodhart, Charles and Sunirand, Pojanart and Tsomocos, Dimitrios P. (2004) A model to analyse financial fragility. Discussion paper, 492. Financial Markets Group, London School of Economics and Political Science, London, UK.						
Relation(s)	http://fmg.lse.ac.uk http://eprints.lse.ac.uk/24703/						

Fig.12. Working with the terms cloud

The results list is colour-coded, to help to distinguish between sets, with the most recent set prepended to the list (Figure 13).



Termine weight (t-score): 26% (22) [10 docs]

1. Landscape History and Theory: from Subject Matter to Analytic Tool (2004-10)
2. Linking the spatial syntax of cognitive maps to the spatial syntax of the environment (2004-07)
1. "The Stage's Glory": John Rich (1692-1761) (2011-05)
2. 142 Strand: a radical address in Victorian London (2006-11-09)
3. 1977. A walk across the park, into the forest and back to the garden: the sculpture park in Britain (2006-10)
4. Approaching life in the London garden centre: acquiring entities and providing products (2007-02)
5. Calculation of the O-H stretching vibrational overtone spectrum of the water dimer (2008-07-17)
6. Calcutta Botanic Garden and the colonial re-ordering of the Indian environment (2008-04)
7. Consumption smoothing and vulnerability in Russia (2010)
8. Coping with economic shocks: consumption

1977. A walk across the park, into the forest and back to the garden: the sculpture park in Britain

Book description: Although the integration of sculpture in gardens is part of a long tradition dating back at least to antiquity, the sculptures themselves are often overlooked, both in the history of art and in the history of the garden. This collection of essays considers the changing relationship between sculpture and gardens over the last three centuries, focusing on four British archetypes:...

Author(s)	Sleeman, J.	Date	2006-10	Language	eng	Type	PeerReviewed, Book chapter	Full Text	No
Identifier	http://eprints.ucl.ac.uk/13615/								
Source	Sleeman, J. (2006) 1977. A walk across the park, into the forest and back to the garden: the sculpture park in Britain. In: Eyres, P. and Russell, F., (eds.) Sculpture and the garden. Ashgate, U.K., pp. 157-169. ISBN 9780754630302								

Recently viewed

A model to analyse financial fragility

Linking the spatial syntax of cognitive maps to the spatial syntax of the environment

Calculation of the O-H stretching vibrational overtone spectrum of the water dimer

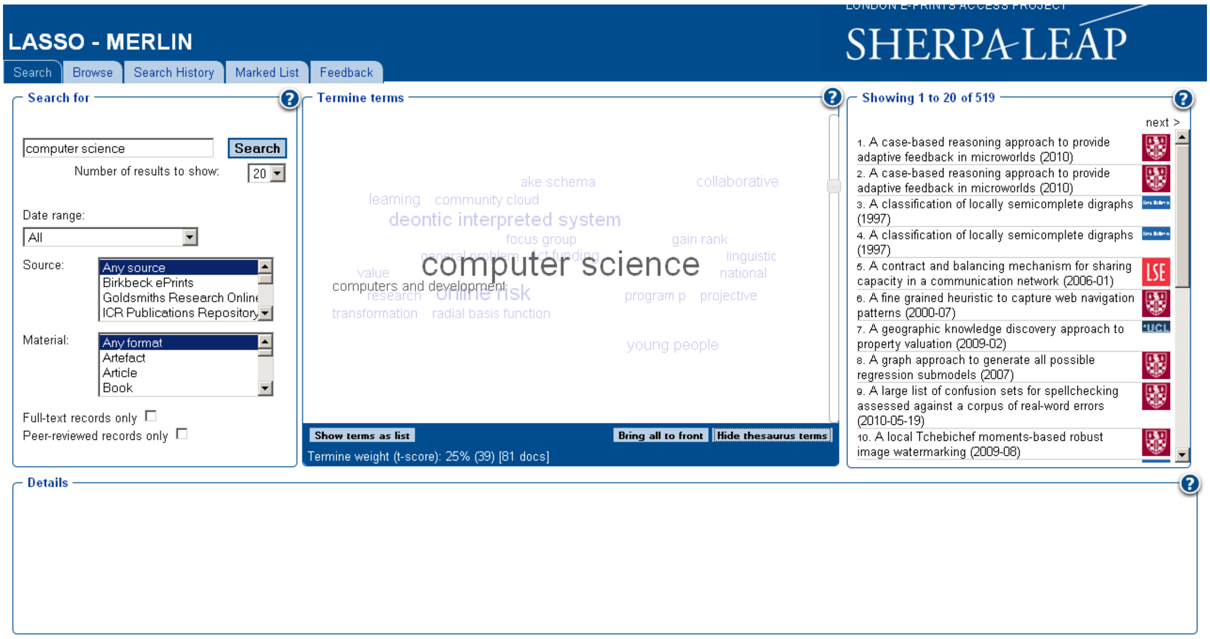
Landscape History and Theory: from Subject Matter to Analytic Tool

Fig.13. Colour-coding applied to results list to aid search

'Show terms as list' caters for the cloud-averse by displaying the terms in columnar format, ranked according to TerMine weight.

Thesaurus interaction

'Show thesaurus terms' is available from cloud mode. This invokes interaction with the external thesaurus (see 4.3 for a technical overview), importing a list of concepts associated with a search term into the cloud. The imported thesaurus terms are displayed one at a time. The user may scroll through them with the mousewheel or up/down cursor keys. Figure 14 shows the initial thesaurus display for the search 'Computer Science'.

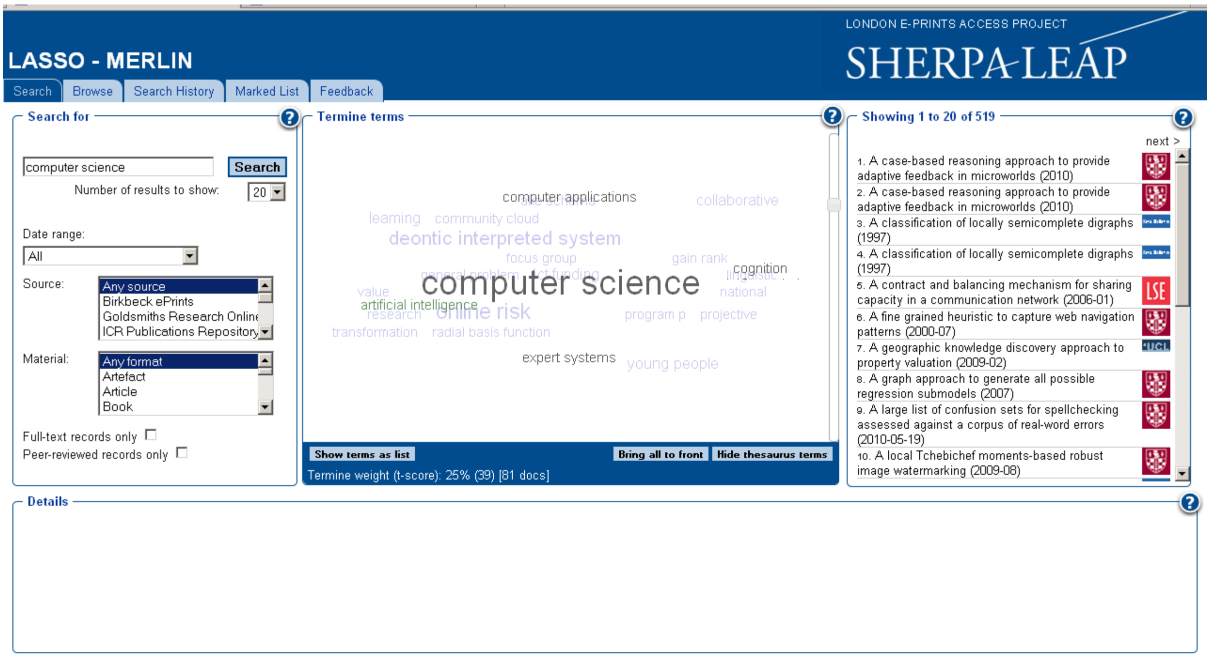


Termine weight (t-score): 25% (39) [81 docs]

1. A case-based reasoning approach to provide adaptive feedback in microworlds (2010)
2. A case-based reasoning approach to provide adaptive feedback in microworlds (2010)
3. A classification of locally semicomplete digraphs (1997)
4. A classification of locally semicomplete digraphs (1997)
5. A contract and balancing mechanism for sharing capacity in a communication network (2006-01)
6. A fine grained heuristic to capture web navigation patterns (2000-07)
7. A geographic knowledge discovery approach to property valuation (2009-02)
8. A graph approach to generate all possible regression submodels (2007)
9. A large list of confusion sets for spellchecking assessed against a corpus of real-word errors (2010-05-19)
10. A local Tchebichef moments-based robust image watermarking (2009-08)

Fig.14. List of related concepts for 'Computer Science' is available for scrolling.

Selecting an associated concept - here, ‘artificial intelligence’ - imports, clockwise, lists of broader, related and narrower terms into the display (Figure 15). Again, the user may scroll through the lists of broader, related and narrower terms with the mousewheel.



Search for: computer science

Number of results to show: 20

Date range: All

Source: Any source, Birkbeck ePrints, Goldsmiths Research Online, ICR Publications Repository

Material: Any format, Artefact, Article, Book

Full-text records only □

Peer-reviewed records only □

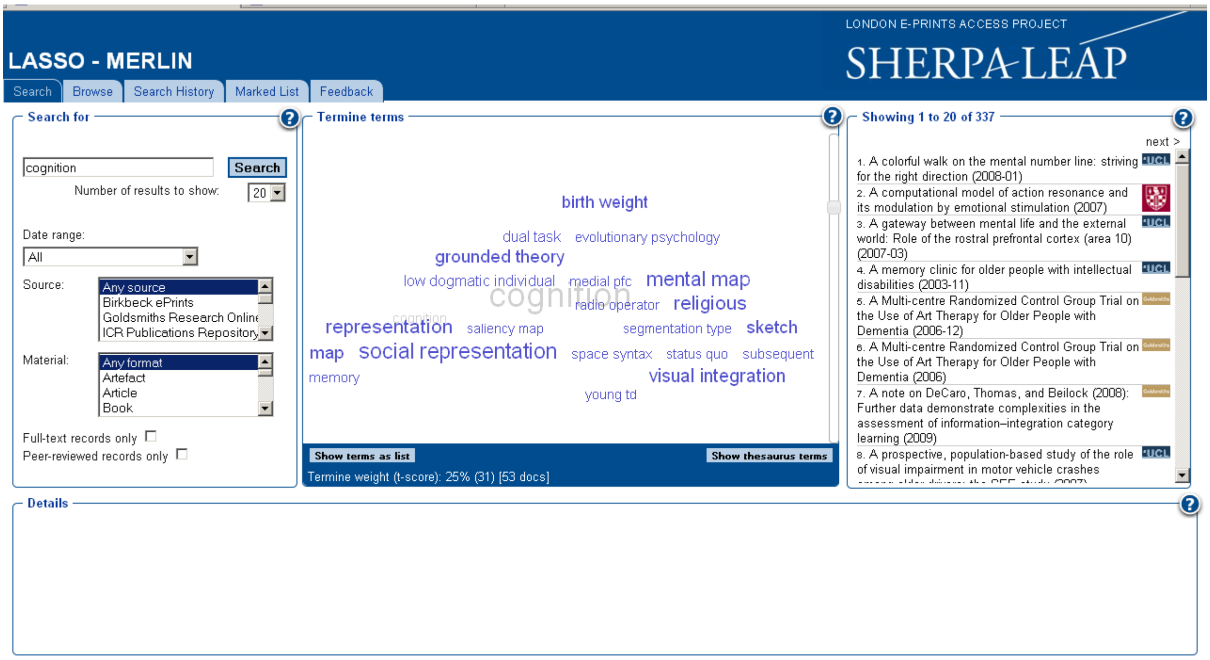
Termine weight (t-score): 25% (39) [81 docs]

Showing 1 to 20 of 519

1. A case-based reasoning approach to provide adaptive feedback in microworlds (2010)
2. A case-based reasoning approach to provide adaptive feedback in microworlds (2010)
3. A classification of locally semicomplete digraphs (1997)
4. A classification of locally semicomplete digraphs (1997)
5. A contract and balancing mechanism for sharing capacity in a communication network (2006-01)
6. A fine grained heuristic to capture web navigation patterns (2000-07)
7. A geographic knowledge discovery approach to property valuation (2009-02)
8. A graph approach to generate all possible regression submodels (2007)
9. A large list of confusion sets for spellchecking assessed against a corpus of real-word errors (2010-05-19)
10. A local Tchebichef moments-based robust image watermarking (2009-08)

Fig.15. Related, broader and narrower terms for ‘Artificial Intelligence’ are now accessible for browsing

Any of the thesaurus terms may be selected as new search terms for the database (Figure 16).



Search for: cognition

Number of results to show: 20

Date range: All

Source: Any source, Birkbeck ePrints, Goldsmiths Research Online, ICR Publications Repository

Material: Any format, Artefact, Article, Book

Full-text records only □

Peer-reviewed records only □

Termine weight (t-score): 25% (31) [53 docs]

Showing 1 to 20 of 337

1. A colorful walk on the mental number line: striving for the right direction (2008-01)
2. A computational model of action resonance and its modulation by emotional stimulation (2007)
3. A gateway between mental life and the external world: Role of the rostral prefrontal cortex (area 10) (2007-03)
4. A memory clinic for older people with intellectual disabilities (2003-11)
5. A Multi-centre Randomized Control Group Trial on the Use of Art Therapy for Older People with Dementia (2006-12)
6. A Multi-centre Randomized Control Group Trial on the Use of Art Therapy for Older People with Dementia (2006)
7. A note on DeCaro, Thomas, and Beilock (2008): Further data demonstrate complexities in the assessment of information–integration category learning (2009)
8. A prospective, population-based study of the role of visual impairment in motor vehicle crashes

Fig.16. Lateral search from ‘Artificial Intelligence’ to ‘Cognition’ via thesaurus.

In the examples shown in Figures 14-16, therefore, the user has searched for ‘Computer Science’, browsed a list of associated concepts, selected the associated concept ‘Artificial Intelligence’, browsed terms related to ‘Artificial Intelligence’, and selected one, ‘Cognition’, as the basis for a new search.

The ‘Bring all to front’ button appears in Thesaurus mode. This will, at any time, reactivate the TerMine terms and display them alongside the thesaurus terms.

Help

A screencast demonstrating the features of LASSO's MERLIN interface was produced and made available from the home page. Originally the screencast was displayed by default in the central panel at the start of each session, but on the advice of the evaluators it was relegated to 'on-demand' availability, since users would most likely only benefit from the video during their first one or two visits. The screencast was made using Screentoaster, a free, cloud-based screen recording service which is now defunct. Its closure came without formal warning and with no facility to migrate recordings. The screencast is unavailable, but in principle was a good means of showcasing the range of search options on offer.

The thesaurus features are supported by hover text. This is intended to ensure that the interface remains relatively compact, despite the range of functionality that it offers.

4.5. Stand-alone code

The final part of the funded phase of MERLIN was to release the demonstration work as open source stand-alone code, for incorporation into other repositories. The source code is packaged and freely available through the project web site. The stand-alone version of MERLIN is provisionally named MER (Metadata Enrichment for Repositories ~~in a London Institutional Network~~). The code incorporates the MERLIN search interface, which will use both the native repository indexes and output and complementary index of TerMine-derived terms. The stand-alone version of MERLIN can interact with any SKOS-based thesaurus to further enhance the discovery experience, as demonstrated in the project.

At the time of writing, the stand-alone MER code has been tested successfully with GNU EPrints 3.

Figure 17 shows the MERLIN architecture; Figure 18 is an overview of MER.

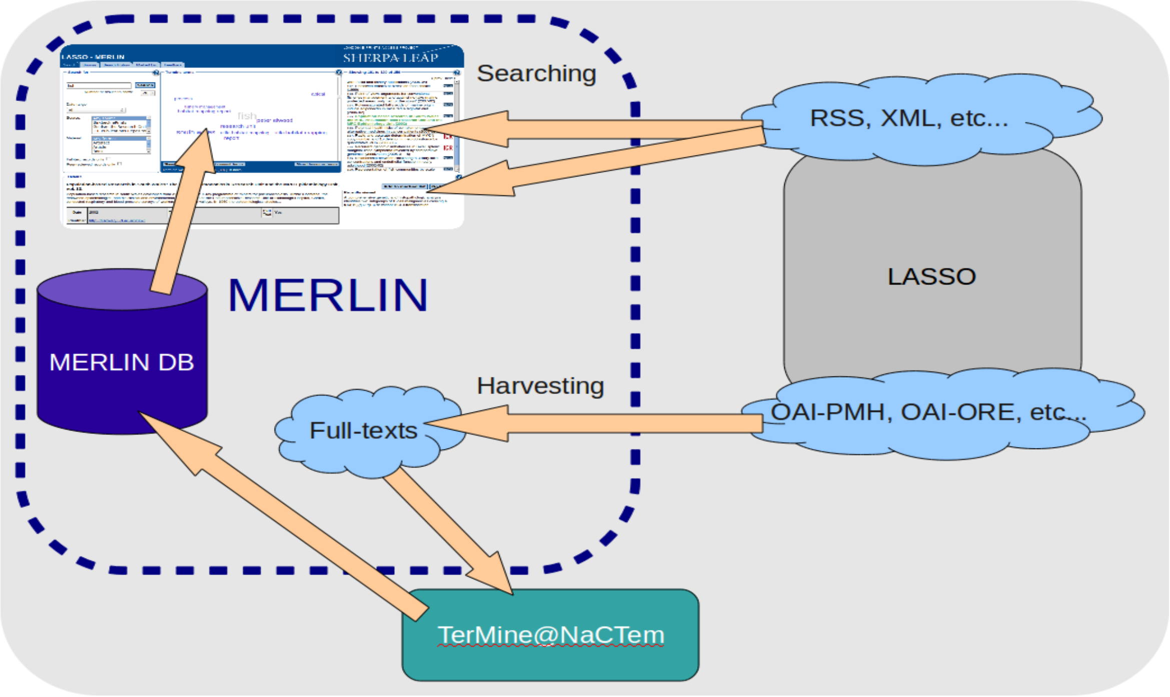


Fig. 17. MERLIN architecture

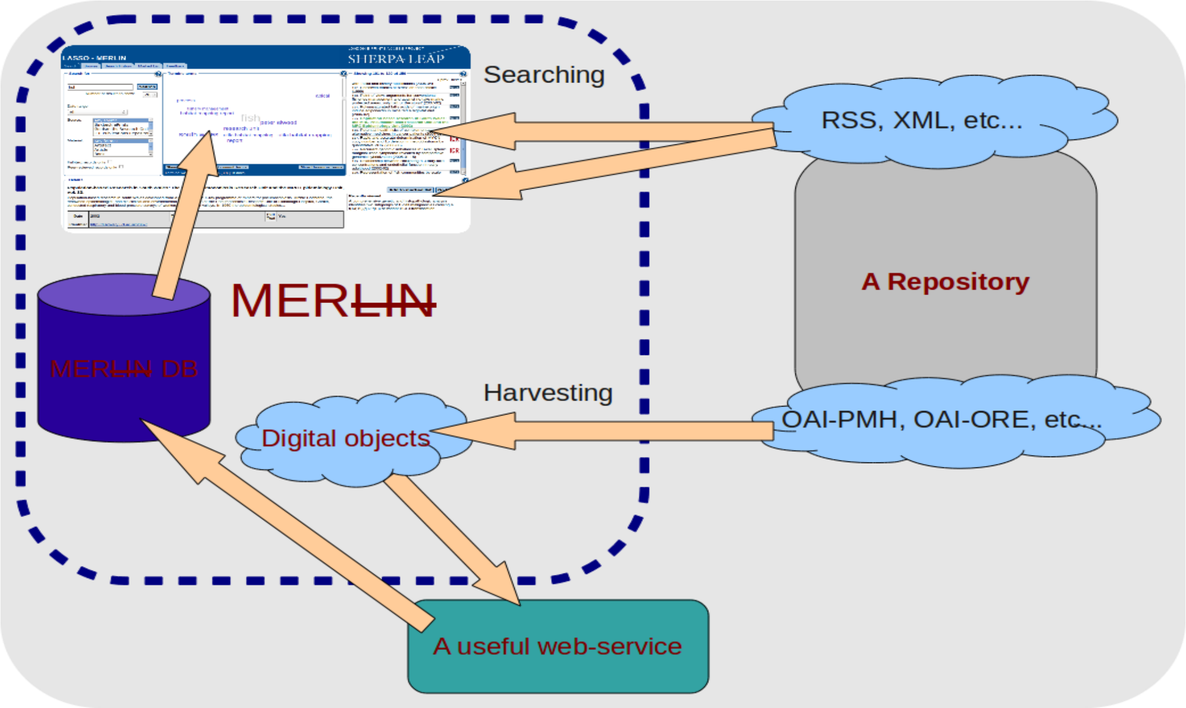


Fig.18. MERLIN stand-alone architecture

4.6. Mobile MERLIN

Post-project, the ULCC team experimented with a touchscreen MERLIN app for smartphones and tablets, resulting in the entry of a rapidly-developed demonstrator in the OR11 Developer Challenge. The app, in honour of OR11's hosts, was christened TEXAS - Touchscreen Enhanced Cross-Search with Augmented Serendipity. Figures 19 and 20 show the TEXAS interface. The app gives search access to the LASSO database, with MERLIN functionality incorporated.

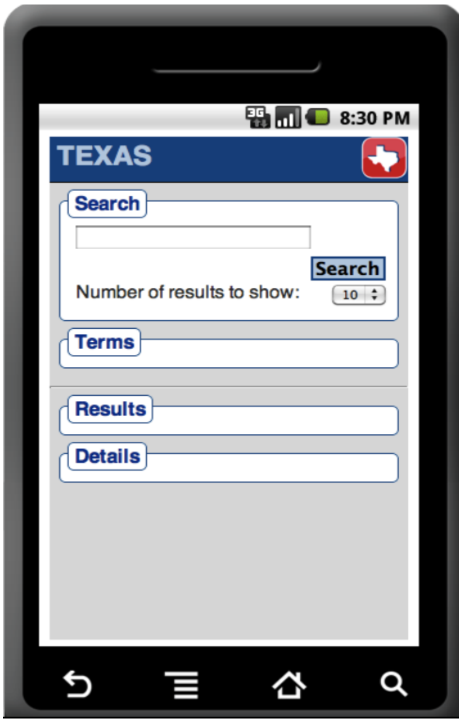
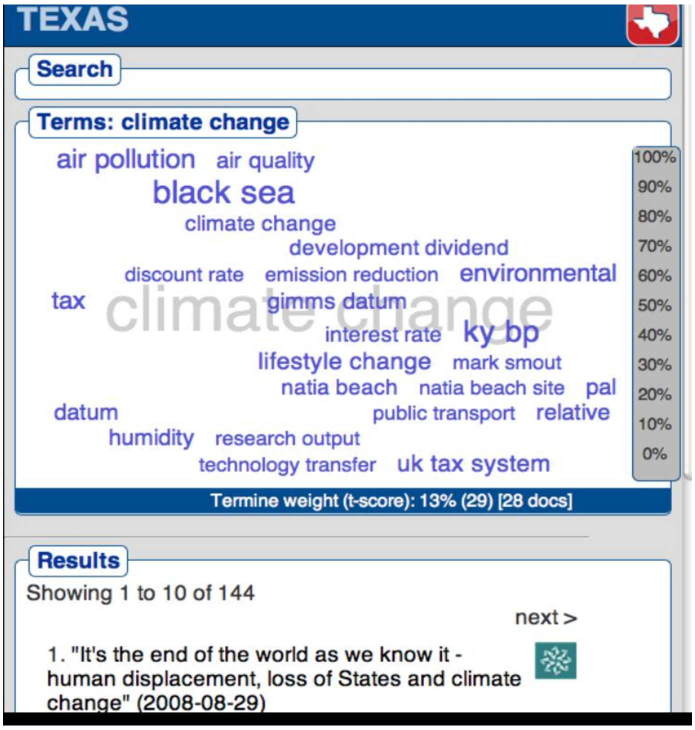


Fig.19. TEXAS - mobile MERLIN.



Termine weight (t-score): 13% (29) [28 docs]

Results

Showing 1 to 10 of 144

1. "It's the end of the world as we know it - human displacement, loss of States and climate change" (2008-08-29)

Fig.20. TEXAS - search results

5. Outputs and Results

5.1. Final evaluation

An external final evaluation of the project was undertaken by Sero Consulting Ltd. In a short timeframe, the evaluators were assigned three objectives:

- To evaluate the user perspective on MERLIN
- To assess the suitability of the technical solution adopted
- To make recommendations about future work that could be undertaken either by JISC or by the project team

Method

Three main methods of investigation were employed:

1. A selection of users were consulted to collect evidence on:

- The user experience of the solution
- Whether MERLIN's functionality matches user needs
- Whether MERLIN is perceived to produce accurate and relevant recall of records
- Views on future development and implementation.

For practical reasons of time and availability this was a survey of those interested in the issue of resource discovery in repositories rather than a more general cohort of student or researcher end users. The method used was to send an email with the survey attached as a Word document. Respondents were directed to the MERLIN website and pointed to the introductory video on the site by way of orientation, and invited to explore the site using search terms of their choice before and while completing the survey.

2. To supplement the technical background of the project two interviews were conducted, with Rory McNicholl the lead technical developer for MERLIN at the University of London Computer Centre, and Bill Hubbard, Head of the Centre for Research Communications (CRC) at the University of Nottingham. An interview was also sought with a representative of the National Centre for Text Mining who had been consulted earlier in the project's development, but he was unable to make himself available on this occasion.

3. To complete a triangulation of views Sero conducted its own usability analysis of the MERLIN site, using its in-house web usability expert Helen Harrop. This focused on issues of navigation and ease of use, applying expertise in website usability testing. Findings from this analysis were compared in particular to those emerging from the user survey.

Summary of results
The overall conclusion drawn by the evaluators was that MERLIN had successfully demonstrated the possibilities of repository search using text mining and thesaurus tools. MERLIN had shown that such an approach can be implemented, and that is has sufficient potential value to merit further investigation and development.

The evaluators also noted that MERLIN had usefully opened up issues which further development work would need to address, among them:

- Technical issues around the harvesting of full-text documents, which knowledge gained through the MERLIN project can now probably resolve;
- User interface issues (some of which are highlighted in the following section);
- Comparison with related approaches, such as the IRS tool referred to above and also OCLC's OAIster service;
- The issues and benefits of implementing MERLIN at scale.

Specific recommendations for improvement
The evaluators made a number of recommendations for improvement, particularly relating to the user interface. It is noted that in some cases there were conflicting views among those consulted, and if time and resources allowed it would be desirable to redesign the interface on the basis of more extensive and systematic user consultation. Nevertheless, several points were highlighted as worth addressing. Among those which relate to MERLIN generally and not to the deficiencies or otherwise of the LASSO service underpinning the demonstrator are the following:

- A need to map out key user journeys (including entry and exit points) and develop the interface to support a smooth, intuitive transit through those journeys.
- Particular issues relating to the user journey include helping users to keep track of their search trail, and making it easier to return to earlier points in their journey. Some help is given but it is not intuitively obvious – there is no 'back' button, and the search history is formidably technical to ordinary users.
- Having a more 'faceted' approach to the search would help users understand how fruitful their search is, e.g. item counts should be added to search options and termine search results
- The prominence of technical terminology should be reduced
- The operation of the results list and how it changes in colour coding and ordering to reflect searches being refined and expanded is far from transparent
- There is no way to manually reorder the results list
- It was also noted that the Help video on the MERLIN home page is crucial to orientating the first-time user and specific recommendations to do with this were made:
    - the prominence of the video on the homepage may give users the initial impression that the search is going to be difficult to use.
    - users only need to watch the video the first time they visit the website, therefore it should not stay as the main item on the homepage [the video has subsequently been moved].
    - ideally the user should have some control over video playback and ability to resize the video player.
    - it would be helpful to introduce the screencast with a sentence about the overall benefits of using MERLIN search.

Recommendations for further development
The evaluators' recommendations for further development are incorporated in Section 8, below.

6. Outcomes

The project successfully demonstrated one approach to the integration of off-the-shelf text-mining tools into repository search. The stated aims of the project were realised. Text mining was integrated into the demonstrator, various issues being surmounted along the way, and the interface was redesigned to accommodate the data enrichments in the search experience. A full and independent evaluation was carried out, leading to some useful recommendations, and the MERLIN source code was made available for download. The aims pertaining to thesaurus integration were realised in spirit, rather than to the letter - a '...navigable subject tree from text-mined keywords...' was not created, but it is felt that the eventual implementation, offering simple and serendipitous enriched discovery opportunities, improved on the original vision.

The approach of MERLIN in creating clear divisions between the harvesting, the indexing technologies and the user interface was welcomed by the evaluators. The technology is scalable: the search interface is compatible with any search that returns an RSS feed; the database can in principle engage with any useful web service.

The MERLIN user interface is novel and, although it could be improved by further user testing and refinement, it may have some merit in repository search even without the text mining and thesaurus overlays.

There are other possible approaches to automated metadata/search enrichment, for instance that piloted by Intute Repository Search (IRS), which was released shortly after the MERLIN project was initiated. IRS demonstrates the incorporation of text-mining and conceptual search in a repository aggregation service in a technically more sophisticated and tightly-coupled way than MERLIN's LASSO overlay, the notable difference being the employment of clustering techniques in IRS. Detailed comparison of IRS and MERLIN is out of scope of this report, but IRS is well documented (see References)

The outcomes of MERLIN may be beneficial both to researchers and to Institutional Repository managers. As the technical interview strand of the evaluation discussed, despite the large scale work of BASE (the Bielefeld service), OCLC OAIster and Google itself, search remains a key challenge. The reach of search services is growing, making the need to address accuracy of search and ability to support filtering/faceting based on subject categorisations and subject-specific vocabularies all the more pressing. Resourcing the construction and maintenance of adequate human-generated metadata seems impossible, and there is a need to find better ways by which improvements in search precision might be derived from automated approaches. This need relates not only to large-scale search, but to individual Institutional Repositories, for which the MERLIN software is also available.

7. Conclusions

The approach investigated by MERLIN is clearly relevant to real-world issues, and the work in text-mining and thesaurus integration carried out by the project does begin to show opportunities, both for low-cost search enhancement and for improving the repository search experience.

The method employed by MERLIN was basically sound, but the decision to pilot MERLIN in an aggregation service had pros and cons. On the positive side, the LASSO aggregator offers diversity, and the specialist content knowledge of those members of the LEAP community who contribute to the aggregator was very helpful in informing the development work. Moreover, LASSO is not a full production service, helping the agility of the project by allowing rapid development. On the other hand, the vagaries of OAI-PMH metadata harvesting threw up some quality issues that on from time to time impeded the fundamental work on text mining and thesaurus integration, and the evaluative and user testing work of the project was occasionally hampered by unexpected or inconsistent results. It might also perhaps, with hindsight, have been easier to design and develop an interface from scratch, rather than to shoehorn new functionality onto a working model.

In terms of outcomes, MERLIN originated in thoughts about the costs and efficiencies of subject cataloguing in repository search services, and clearly the final MERLIN outputs do not amount to subject cataloguing by another route. MERLIN does, however, enrich the discovery experience. It makes resources more accessible by adding statistical precision to full text search, drawing out terminology in researchers' own vocabularies and highlighting it where appropriate, with an optional thesaurus link-up to provide both standardisation of terminology and a jumping-off point for new searches. The project’s external evaluators concluded that the intention behind the MERLIN project to investigate alternatives to manual metadata generation and cataloguing for resource discovery in repositories meets a clear need, and that the approach taken, using text mining software and a thesaurus, was innovative and of considerable wider interest in the field. MERLIN succeeded in helping to open up the potential for this approach and demonstrating how these tools can be used for effective resource discovery, while highlighting several issues and areas for further development.

8. Implications

MERLIN offers the potential for any repository service, whether an aggregation or an institutional service, to enrich its search, at little cost. In so doing, MERLIN helps repository owners to maximise the value of their own content. MERLIN adopters may implement text-mining techniques, as demonstrated by the project; they may implement additional interactions with external web services, notably thesauri; and they may take advantage of a user-tested interface, whether to assist with the management of additional text mining-based search functionality or simply to replace their default repository search.

The interface would benefit from more development and testing, away from the LASSO aggregator. Some achievable recommendations for further improvement, emerging from the final evaluation, are listed in Section 5, above.

To evaluate further the text mining and thesaurus extensions, it would be of interest to try out the entire MERLIN package on ‘real’ researchers from a range of different disciplines. This would also enable a comparison between native full text search and the MERLIN-enriched search that incorporates derived and weighted keywords. Linking stand-alone MERLIN implementations with specialist thesauri - for instance, Getty’s Arts and Architecture Thesaurus (AAT) - for discipline-specific IRs could also be explored, perhaps with a view to providing explicit ‘plug-in’ support for a range of thesaurus add-ons within the MERLIN code release.

Questions of performance and scale for local repository implementations should also be explored. The demonstrator produced over 1,000,000 TerMine terms for only 10,000 parsed documents. The impact of these term-generation and storage overheads on a production service has not yet been assessed.

Interest emerged from the evaluation in seeing whether MERLIN's harvesting and indexing processes would work at scale, for example using the directory of global repositories offered through API by OpenDOAR. Testing both the technical feasibility and end-user benefits at such scale would be desirable.

Although some work was undertaken, post-project, on delivering MERLIN to a range of non-desktop platforms, further development in this area would help to support future MERLIN adopters.

9. Recommendations

The following recommendations were made by the project evaluators.

While the MERLIN project so far has demonstrated the potential for this approach to resource discovery, there are several areas where further development is needed or is desirable to explore its full potential and to realise all the potential benefits. The main lines of development we would recommend in this respect would be:

- JISC or another body to make a small investment in further work to identify the strengths and weaknesses of the Merlin text mining approach for wider large (global) scale application

- If that indicated potential, further experimentation to understand the best ways of linking the resulting terms with specialist thesauri
- As part of that work, or subsequent to it, the MERLIN approach to be compared to similar approaches such as that undertaken in the IRS project
- The MERLIN user interface to be fully redesigned based on systematic user consultation, so that it better matches the specific features of the discovery approach being taken
- Development of the interface to match the range of delivery platforms now in use including tablet and mobile devices.

10. References

MERLIN: http://www.ucl.ac.uk/ls/merlin
MERLIN downloadable code: https://code.google.com/p/jisc-merlin/source/browse/#svn%2Ftrunk
MERLIN-LASSO demonstrator: http://lasso.ucl.ac.uk/merlin-ui
Mobile MERLIN (TEXAS): http://dablog.ulcc.ac.uk/2011/06/14/open-repositories-2011-part-2-the-developer-challenge/

HILT: http://hilt.cdlr.strath.ac.uk/

Intute Repository Search (IRS): http://irs.mimas.ac.uk/demonstrator/
IRS Ariadne article, October 2009: http://www.ariadne.ac.uk/issue61/lyte-et-al/

NaCTeM: http://www.nactem.ac.uk/
TerMine: http://www.nactem.ac.uk/software/termine/webservice/
Termine article: Frantzi, K., Ananiadou, S. and Mima, H. (2000) Automatic recognition of multi-word terms. International Journal of Digital Libraries 3(2), pp.117-132.
Sentence splitter: http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector

SHERPA-LEAP: http://www.sherpa-leap.ac.uk

Appendix A

User testing of interface, mid-project: questionnaire and tasks.

Before visiting the website:

1. Have you heard of any of the following projects: LEAP, LASSO, MERLIN? If so, from where?

2. How do you normally find scholarly information?

3. What do you expect an ‘institutional repository’ to provide?

4. Have you heard of text mining? If so, what do you understand by the term?

Visit <superseded MERLIN UI>.html:

5. What strikes you first about the front page?

6. Is the explanation of what the service is for clear?

7. Professor Smith would like to refer to a colleague’s research into shellfish deposition. How many items will be returned by a search for this term?

8. Please take a moment to explore the results page and describe what you see.

9. Use the mouse to push the ‘slider’ in the centre of the screen upwards. What effect does this have on the display?

10. Please click on the term ‘landscape change’. What do you see?

11. Are the options, and the differences between them, clear?

12. Please click on the second option. How many additional results does this provide?

13. Why do you think that these papers have been included in the results?

14. Please select the last paper in the results list. What type of publication is it?

15. Go back to the ‘basic search’ page. Professor Jones is an Economist who would like to find papers about endogenous growth. How many items will be returned by a search for this term?

16. The results contain a number of non-economic terms. Why do you think that this is the case?

17. Click on the term ‘central bank’ in the tag cloud. Select the option to search within the results for this term.What is the name of the item that contains both terms?

18. Click on the item title to view more information. Which of the search terms (‘endogenous growth’ and ‘central bank’) appear on this page?

19. If the terms do not appear in this page, why do you think that the search service has returned this item in response to your search?

20. How easy did you find it to navigate the site?

21. Does the labelling on the different features in the results make it clear what they are?

22. Did you find it easy or difficult to choose between the different options?

23. Did you like the way that the choices were presented?

24. Did you find the ‘tag cloud’ of search terms helpful in searching within your results?

25. How do you think the tag cloud could be improved?

26. Did you find the ‘slider’ to add and remove terms from the cloud useful?

27. How do you think that the slider device could be improved?

28. What did you think about the general appearance of the search results? What changes, if any, would you suggest?

29. Have you used similar search services before? If so, which one(s)?

30. Do you have any other comments or feedback for the project team?

Appendix B

TerMine C-value analysis of a text version of this report.

556 terms found.

Rank	Term	Score
1	text mining	23.857143
2	full text	17.6
3	ucl library service	12.6797
4	repository search	9.833333
5	user testing	8.5
6	user interface	7.8
7	interim interface	7
7	institutional repository	7
7	search term	7
10	aggregation service	6.714286
11	london computing centre	6.33985
12	source repository	5
12	national centre	5
12	merlin project	5
12	artificial intelligence	5
16	lasso database	4
16	mobile merlin	4
16	lasso interface	4
16	thesaurus term	4
16	merlin metadata enrichment technology	4
16	tag cloud	4
16	search service	4
16	final evaluation	4
16	termine output	4
16	project team	4
16	narrow term	4
27	full text search	3.754888
27	repository aggregation service	3.754888
29	user testing session	3.169925
29	re-usable web application	3.169925
29	project web site	3.169925
29	digital object retrieval	3.169925
29	sero consulting ltd	3.169925
29	systematic user consultation	3.169925
29	termine score threshold	3.169925
29	hilt soap client	3.169925
29	london institutional network	3.169925
29	intute repository search	3.169925
29	full text file	3.169925
29	merlin user interface	3.169925

29	off-the-shelf text-mining tool	3.169925
42	termine term	3
42	lasso aggregator	3
42	appendix b	3
42	final interface	3
42	open source	3
42	appendix a	3
42	interface design	3
42	term acquisition	3
42	thesaurus integration	3
42	single document	3
42	repository service	3
42	repository context	3
42	repository discovery	3
42	curl request	3
42	martin moyle	3
42	lasso service	3
42	resource discovery	3
42	search experience	3
42	merlin approach	3
42	computer science	3
62	in-house web usability expert helen harrop	2.584963
63	leap aggregated search service on-line	2.321928
63	additional text mining-based search functionality	2.321928
65	navigable subject tree	2.169925
65	repository cross-searching service	2.169925
65	full text document	2.169925
68	london repository aggregation service	2
68	simple oai-pmh-based aggregation service	2
68	web service	2
68	whether merlin	2
68	full text digital object	2
68	josh brown	2
68	cost-effective integration	2
68	large-scale search	2
68	termine text mining tool	2
68	discovery experience	2
68	off-the-shelf text mining tool	2
68	subject cataloguing	2
68	pilot navigable subject tree	2
68	central bank	2
68	leap repository	2
68	full text document record	2
68	appropriate user interface enhancement	2
68	jisc information environment programme	2
68	sentence splitter	2

68	please click	2
68	open source stand-alone code	2
68	merlin code	2
68	resource implication	2
68	text-mined term	2
68	user journey	2
68	hilt tool	2
68	merlin tool	2
68	joint information systems committee	2
68	off-the-shelf tool	2
68	original search	2
68	research communications	2
68	bill hubbard	2
68	library-standard subject classification skill	2
68	london-based higher education institutions	2
68	term storage	2
68	google code project environment	2
68	pilot repository aggregation service	2
68	endogenous growth	2
68	native full text search	2
68	full text repository search	2
68	rory mcnicholl	2
68	london research library service	2
68	ancient south-east asian culture	2
68	lasso advanced search interface	2
68	repository content	2
68	multi-subject terminological cross-searching aid	2
68	basic search	2
68	noise reduction	2
68	merlin text mining approach	2
68	more term	2
68	termine processing	2
68	stand-alone code	2
68	weighted keyword	2
68	browser view	2
68	automatic subject	2
68	repository owner	2
68	termine text mining service	2
68	lasso repository cross-searching service	2
68	search interface	2
68	termine term extraction technology	2
68	merlin architecture	2
68	non-text binary	2
68	text-mined keyword	2
68	serendipitous enriched discovery opportunity	2
68	specific recommendation	2

68	termine interaction	2
68	external evaluator	2
68	central panel	2
68	pdf document	2
68	london e-prints access project	2
68	open source web application	2
68	formative evaluation	2
68	developmental evaluation	2
68	metadata enrichment	2
68	richard davis	2
68	search engine	2
68	executive summary	2
68	london repository consortium sherpa-leap	2
146	repository search service	1.584962
146	considerable wider interest	1.584962
146	? merlin tool	1.584962
146	repository front end	1.584962
146	nactem sentence splitter	1.584962
146	repository full text	1.584962
146	datum acquisition process	1.584962
146	cloud prima facie	1.584962
146	original project aim	1.584962
146	institutional repository manager	1.584962
146	full text url	1.584962
146	website usability testing	1.584962
146	text-mined search experience	1.584962
146	or11 developer challenge	1.584962
146	sherpa-leap aggregation service	1.584962
146	merlin code release	1.584962
146	stand-alone merlin implementation	1.584962
146	minimum termine strength	1.584962
146	simple full-text indexing	1.584962
146	final project ui	1.584962
146	merlin source code	1.584962
146	hilt sru/w server	1.584962
146	verifying basic search	1.584962
146	default repository search	1.584962
146	oai dublin core	1.584962
146	item-level term presentation	1.584962
146	technical interview strand	1.584962
146	merlin term database	1.584962
146	regular lasso metadata	1.584962
146	original lasso service	1.584962
146	external web service	1.584962
146	xml dom manipulation	1.584962
146	sherpa-leap repository network	1.584962

146	researcher end user	1.584962
146	relevant nactem product	1.584962
146	native repository index	1.584962
146	initial thesaurus display	1.584962
146	merlin enrichment technology	1.584962
146	merlin stand-alone architecture	1.584962
146	merlin downloadable code	1.584962
146	sherpa-leap project officer	1.584962
146	merlin search interface	1.584962
146	local repository interface	1.584962
146	merlin content acquisition	1.584962
146	adequate human-generated metadata	1.584962
146	original search string	1.584962
146	touchscreen merlin app	1.584962
146	sherpa-leap member institution	1.584962
146	entire merlin package	1.584962
146	specialist content knowledge	1.584962
146	text mining process	1.584962
146	termine soap service	1.584962
146	optional thesaurus link-up	1.584962
146	future merlin adopter	1.584962
146	text mining software	1.584962
146	lasso aggregation service	1.584962
146	irs ariadne article	1.584962
146	daily oai-pmh metadata	1.584962
146	dr jacqueline cooke	1.584962
146	termine algorithmic analysis	1.584962
146	java opendocument converter	1.584962
146	mobile code release	1.584962
146	original search term	1.584962
146	local repository implementation	1.584962
146	traditional subject cataloguing	1.584962
146	effective resource discovery	1.584962
146	london computer centre	1.584962
146	oai export metadata	1.584962
146	external final evaluation	1.584962
146	lasso basic search	1.584962
146	manual metadata generation	1.584962
146	plain text file	1.584962
146	html login page	1.584962
146	lasso discovery service	1.584962
146	final merlin output	1.584962
146	full production service	1.584962
146	merlin demonstrator interface	1.584962
146	user interface issue	1.584962
146	dr sophia ananiadou	1.584962

146	single document term	1.584962
146	merlin technical methodology	1.584962
146	text mining technique	1.584962
146	basic merlin architecture	1.584962
146	merlin home page	1.584962
146	stand-alone mer code	1.584962
146	oai-pmh metadata harvesting	1.584962
146	agile development methodology	1.584962
146	dr paul ayris	1.584962
146	merlin extension check	1.584962
146	independent final evaluation	1.584962
146	ucl copyright officer	1.584962
146	suitably-formatted external thesaurus	1.584962
146	text mining extension	1.584962
146	touchscreen enhanced cross-search	1.584962
146	low-cost search enhancement	1.584962
146	user testing questionnaire	1.584962
146	repository search experience	1.584962
146	stand-alone institutional repository	1.584962
146	lead technical developer	1.584962
146	left-hand middle window	1.584962
246	local criterion	1
246	oaister service	1
246	rss-formatted search	1
246	low weight	1
246	relevant recall	1
246	irs tool	1
246	ongoing development	1
246	thesaurus add-on	1
246	on-line thesauri	1
246	lateral search	1
246	long document	1
246	search trail	1
246	broad term	1
246	staff-only item	1
246	weighting threshold	1
246	minimum weight	1
246	repository manager	1
246	prior knowledge	1
246	external thesauri	1
246	traditional library	1
246	browse index	1
246	merlin technology	1
246	indexing process	1
246	subject-based approach	1
246	final part	1

246	default search	1
246	boilerplate text	1
246	digital libraries	1
246	simple search	1
246	merlin website	1
246	termine c_value	1
246	shellfish deposition	1
246	hilt response	1
246	discovery approach	1
246	positive side	1
246	simple extension	1
246	institutional size	1
246	hilt project	1
246	open licence	1
246	repository scale	1
246	search option	1
246	subject-specific vocabulary	1
246	thesaurus feature	1
246	simple means	1
246	thesaurus tool	1
246	representative cloud	1
246	working model	1
246	xml-represented thesaurus	1
246	columnar format	1
246	such term	1
246	technical background	1
246	cloud mode	1
246	search benefit	1
246	help video	1
246	mine document	1
246	item count	1
246	video player	1
246	hide cloud	1
246	msword document	1
246	david kay	1
246	principle engage	1
246	indexing technology	1
246	thesaurus-driven approach	1
246	metadata/search enrichment	1
246	merlin functionality	1
246	text-mining integration	1
246	formal warning	1
246	search enrichment	1
246	repository discoverability	1
246	main method	1
246	specialist thesaurus	1

246	institutional repositories	1
246	thesaurus mode	1
246	clear division	1
246	termined threshold	1
246	discipline-specific ir	1
246	simple guide	1
246	full evaluation	1
246	gnu eprint	1
246	merlin software	1
246	valuable feedback	1
246	view sero	1
246	merlin context	1
246	augmented serendipity	1
246	automatically-derived term	1
246	specialist subject	1
246	leap partnership	1
246	initial problem	1
246	few difficulty	1
246	research interest	1
246	left-hand panel	1
246	hilt resource	1
246	cloud-based screen	1
246	cloud centre	1
246	mobile device	1
246	merlin environment	1
246	heuristic rule	1
246	initial impression	1
246	file retrieval	1
246	robert drinkall	1
246	steering group	1
246	lasso overlay	1
246	user drill	1
246	project manager	1
246	random list	1
246	implement modification	1
246	msword doc	1
246	substantial range	1
246	ideal testbed	1
246	lasso record	1
246	substantial diversity	1
246	plain text	1
246	video playback	1
246	merlin interface	1
246	full-text record	1
246	complementary index	1
246	original vision	1

246	external thesaurus	1
246	rough-and-ready tool	1
246	many repository	1
246	ordinary user	1
246	interim service	1
246	subject classification	1
246	top-scoring term	1
246	merlin search	1
246	parent document	1
246	nactem tool	1
246	termine weight	1
246	hover text	1
246	sru/w server	1
246	usability analysis	1
246	thoughtful feedback	1
246	resource constraint	1
246	search session	1
246	index term	1
246	up/down cursor	1
246	cloud term	1
246	user-friendly exposure	1
246	technical process	1
246	architecture thesaurus	1
246	merlin initiative	1
246	original purpose	1
246	independent evaluation	1
246	embargoed document	1
246	ian chowchat	1
246	colour coding	1
246	user experience	1
246	first-time user	1
246	delivery platform	1
246	additional interaction	1
246	non-desktop platform	1
246	lasso environment	1
246	statistical meaning	1
246	subject taxonomy	1
246	achievable recommendation	1
246	detailed comparison	1
246	cost-effective subject	1
246	automatic recognition	1
246	ulcc team	1
246	ucl department	1
246	simple enhancement	1
246	front page	1
246	clinical biomedicine	1

246	global repository	1
246	home page	1
246	professor smith	1
246	storage overhead	1
246	michael day	1
246	item title	1
246	technical terminology	1
246	user perspective	1
246	normalisation challenge	1
246	final project	1
246	funded phase	1
246	termine score	1
246	resource-intensive skill	1
246	termine-derived term	1
246	merlin adopter	1
246	landscape change	1
246	search precision	1
246	final refinement	1
246	search access	1
246	termine datum	1
246	summative evaluation	1
246	merlin enhancement	1
246	? bring	1
246	direct url	1
246	supplementary strand	1
246	eventual implementation	1
246	term extraction	1
246	introductory video	1
246	control system	1
246	show term	1
246	clean interface	1
246	short timeframe	1
246	test user	1
246	future development	1
246	redesigned interface	1
246	non-economic term	1
246	helen harrop	1
246	final model	1
246	test environment	1
246	datum enrichment	1
246	text-mining technique	1
246	typical response	1
246	simple overview	1
246	statistical precision	1
246	real-world issue	1
246	lasso schema	1

246	repository item	1
246	experimental integration	1
246	examine issue	1
246	statistical analysis	1
246	central slider	1
246	technical issue	1
246	text box	1
246	subject categorisation	1
246	generic issue	1
246	production service	1
246	pdf cleanup	1
246	termine search	1
246	scale application	1
246	full-text resource	1
246	merlin demonstrator	1
246	termine service	1
246	metadata format	1
246	open development	1
246	particular issue	1
246	slid device	1
246	rss feed	1
246	weighting process	1
246	language-processing tool	1
246	skos-based thesaurus	1
246	merlin ui	1
246	merlin site	1
246	journal abbreviation	1
246	professor jones	1
246	stand-alone repository	1
246	secondary weighting	1
246	thesaurus interaction	1
246	local installation	1
246	full-text search	1
246	leap community	1
246	thesaurus extension	1
246	unambiguous url	1
246	word document	1
246	light weight	1
246	practical reason	1
246	intuitive transit	1
246	rapid acquisition	1
246	merlin-lasso demonstrator	1
246	intuitive experience	1
246	merlin-enriched search	1
246	main line	1
246	thesaurus overlay	1

246	piecemeal funding	1
246	blunt instrument	1
246	unesco thesaurus	1
246	structured navigation	1
246	international journal	1
246	bolder display	1
246	store detail	1
246	javascript object	1
246	sherpa-leap repository	1
246	search history	1
246	jumping-off point	1
246	user-tested interface	1
246	lasso demonstrator	1
246	research publication	1
246	christopher pressler	1
246	jisc call	1
246	exit point	1
246	public release	1
246	click-through option	1
246	repository object	1
246	conceptual search	1
246	specific feature	1
246	full-text document	1
246	rapid development	1
246	reference list	1
246	hilt thesaurus	1
246	oclc oaister	1
246	multi-word term	1
246	quality issue	1
246	informal evaluation	1
246	mysql table	1
246	rapidly-developed demonstrator	1
246	main item	1
246	source code	1
246	small investment	1
246	termine article	1
246	tabular datum	1
246	institutional service	1
246	technical overview	1
246	repository metadata	1
246	long-term unavailability	1
246	born-digital pdf	1
246	texas interface	1
246	bielefeld service	1
246	project evaluator	1
246	search page	1

246	user survey	1
246	pdf conversion	1
246	frequently-repeated term	1
246	information study	1
246	irs project	1
246	various issue	1
246	weighted keyord	1