Intute Repository Search service (www.intute.ac.uk/irs): A collaborative project to showcase UK research output through advanced discovery and retrieval facilities

Institutional repositories are a major element of the Open Access movement and more specifically in research and education. Their main purpose is to make available as much research output of an institution as possible. Technological changes and developments have an impact on search and discovery functionalities. This impact in turn inspires ideas and projects about useful and efficient ways of searching for academic research output.

Intute Repository Search[1] is a project that was set up to develop a UK repository search service to support academic activity. It is funded by the JISC[2] and led by MIMAS[3] in partnership with SHERPA[4], UKOLN[5] and NaCTEM[6]. Intute Repository Search is designed to serve as a showcase for UK research and education. The technological developments in Intute Repository Search operate to reach the project's main targets. This paper will discuss these targets and the project's achievements.

The project has a series of high level aims:

- To identify, develop and support high-value research knowledge communities within search and discovery process by delivering free targeted search and discovery facilities derived from UK Higher Education institutional repositories.
- To encourage the embedding of repository search in familiar and day-to-day research desktop environments.
- To provide improved services to individuals including the ability to personalise information based on user profile, directed browse and dynamic navigation.
- To provide richer, more meaningful conceptual and semantic search facilities using text mining technology, including full-text document searching.
- To embed IRS to Web 2.0 technologies aiming at richer personalisation and contribute to developments within the Semantic Web.

[1] www.intute.ac.uk/irs
[2] Joint Information Systems Committee: www.jisc.ac.uk
[3] National data centre, University of Manchester, UK: www.mimas.ac.uk
[4] Securing a Hybrid Environment for Research Preservation and Access:www.sherpa.ac.uk
[5] UKOLN, University of Bath: www.ukoln.ac.uk
[6] National Centre for Text Mining:www.nactem.ac.uk

The project has identified and successfully carried out specific development paths – namely, simple metadata search, conceptual search and clustering, full-text indexing of documents, text-mining of full-text documents, automatic subject classification, clustering of results and browsing/visualisation of the search results. Together with our work with NaCTeM, this extends to term based document classification and query expansion. Intute Repository Search currently searches over 95 UK institutional repositories that are listed in the Directory of Open Access Repositories, OpenDOAR[7], and harvested using an aggregation system developed by UKOLN.

The evolution of the project involved growing from simple search at the beginning of the project to the more advanced conceptual search, clustering and text mining based search facilities, to be integrated and fully operational by the end of the project. The advanced discovery and retrieval features that Intute Repository Search offer personalisation of searching; and concept visualisation from automated clustering.

The project has combined two complementary technologies. One is conceptual, parametric search and automated clustering, using the IDOL engine 7 provided by Autonomy[8], and the second is based on text mining technology provided by NaCTeM.

Conceptual search, more specifically involves:

- Benefits in the use of unstructured information search algorithms supported by metadata and full text, thus allowing automated taxonomy generation, and concept matching across related documents or artefacts.
- The ability to search for a document, based on words that are related to a concept rather than a document that contains the actual search word or phrase.
- The use of unstructured retrieval algorithms (Bayesian Inference and Shannon's Information Theory[9]) provided by the Autonomy IDOL 7 engine.

Conceptual search then allows for a richer contextual search facility for users who want to view documents that are ranked according to their relation to the query.

[7] Open Directory of Open Access Repositories:www.opendoar.org
[8] Autonomy IDOL: www.autonomy.com/content/Products/products-idol-server/index.en.html
[9] http://www.autonomy.com/content/Technology/autonomys-technology-unique-combination-technologies/index.en.html

Text mining includes term extraction for improved browsing options, supporting a more dynamic, intuitive and semantically enhanced discovery in both a specific domain and cross disciplinary domains.

Text mining allows automated term-based document classification and clustering. This is useful for identifying conceptually similar documents and discovering documents otherwise hidden. Document classification and clustering is based on automatically created metadata from text and associated search by grouping semantically similar terms to retrieve pertinent documents (according to the user's interest).

Positive feedback data has been obtained from formal evaluation with academic end users and researchers. User group requirements have been integrated into the project's development iterations to ensure that the project adequately reflects what researchers want from a service such as Intute Repository Search.

The benefits that this search service provides are:

a) For the research community in providing a more effective and personalised search and discovery facility, addressing the problem of information oversight.

b) For the institutions themselves in providing a useful tool for their research output to attract a global audience.

c) For society as a whole, in ensuring that publicly funded research is not only made easily reached through Open Access but that it is also more clearly identifiable for the organisation or person who searches for a particular study.

Intute Repository Search is a valuable tool that does not work in isolation. It is a collaborative project that shares expertise from established organisations within the UK. At the same time, it shares links and experiences with other projects such as UKPubMed and the EC funded European project DRIVER. By linking with other international projects and initiatives, the project's intention is joint information gathering and the setting up of strategic alliances, thereby broadening the research and teaching knowledge domain and making sense of the repository landscape together in an effort to prioritize search targets and share best practice.