Aiding Research and Electronic Content Discovery through Meta-search based on Open Source Initiatives

Collin D'mello Amit Jha

Abstract

The goal of electronic content discovery with Metasearching is to investigate issues related to locating, retrieving, and promulgating information in large networked environments. The Internet and WWW provide a challenging environment for deployment of these services. The needs of information publishers - to maximize audience reach - and the user - to minimize information overload - require advanced technical solutions and investigative research. As the deep web of information grows at an exponential rate, efforts to make the technology more manageable and affordable are highly in demand. Applying advanced information retrieval techniques i.e. meta-search and open link resolver are some approach to such efforts. In this paper we present our attempt to apply open source initiatives for retrieval model based on NISO Standards, relevance mechanisms and a meta search technique as an integrated electronic resource discovery system for the Deep Web. This paper discusses challenges and issues involved in Meta-Searching, as well as practical issues such as retrieval effectiveness, usability and scalability.

Keywords: Open Source Initiatives, Meta-Search, Resource Discovery, Deep Web, Information Retrieval.

1. Introduction

As education becomes more competitive and the number of institutions rapidly increases, libraries can be a distinguishing factor in establishing the identity of institutions. The knowledge infrastructure of libraries determines the quality of research output of an institution's academic community. With advent of Internet and WWW content delivery has taken a new dimension and more and more content is now delivered online. With so many online sources of research library is faced with new challenges of aiding quality content discovery and also ensure that all online resources well used within the library. In addition there is a continuous challenge to train its user community on various e-resource platforms.

From the interaction with the library community it was felt that best way to address the content delivery challenges would be through single window search environment which is easy to use and is affordable. However while there were solution available they did not seem to be either affordable or providing complete environment required for aiding content discovery. Library community felt that it would be best address by implementing open source technologies and hence many libraries are now exploring various open source technologies to aid content discovery.

GIST took up the challenge of delivering this much required solution and explored various open source technologies and finally started developing a metasearch platform on various open source initiatives. This paper brings presents GIST finding and the solution that finally emerged after extensive

7th International CALIBER-2009,
Pondicherry University, Puducherry, February 25-27, 2009
© INFLIBNET Centre, Ahmedabad

research, library community interaction and researcher feedback.

2. Features and functionality desired for aiding and researching through Meta searching in the dynamic environment of Innovation:

Based on the extensive research after interaction with the researchers, library community and technology experts, it was desired that Metasearch platform goes beyond just enabling searches but provides a complete environment for research. The researches desired some critical functionality some of which are reproduced below

- ♦ A Google like interface for easy Basic and Advance Search across quality research sources through single window.
- ♦ Results Clustering for better results management by Topic, Author, Year.
- ♦ Article Relevancy across multiple resources searched.
- ♦ Subscription Identifier to identify articles to which researchers have full text access.
- ♦ Limiters to navigate across results for various resources searched.
- ♦ Search within Search to narrow results.
- ♦ Counter compliant usage statistics.
- ♦ Select, Export and save results.

3. Challenges for Implementing the Features Desired by Researches above using Open Source Initiatives

- ♦ Understanding the guidelines provided by NISO Standard for Metasearching Initiatives and the recommendations given by NISO for implementing Metasearching.
- ♦ To find open source meta search solution that shall address the above functionality of the users while complying NISO Metasearch Initiatives and functional recommendation.
- ♦ After extensive research on finding various open source solution it was concluded that no one software can addresses all the above functionality and most of them were concentrated on enabling part metasearching only and most of them were not following the complete recommendations of NISO Standards being Open Source. Hence the challenge was to identify various open source software that would be required to implement the complete solutions and comply to NISO Standards of Metasearching.
- ♦ To identify various open source technologies / components.

It was decided that to implement the complete solution open source software are found in three important areas and are further evaluated

- o Metasearching
- o Clustering
- o Relevancy

And few customized programs and algorithms are developed in areas of Customizable User Interface.

- o Subscription Identifiers
- o Usage Statistics.

- ♦ For metasearching we evaluated various software's like Dbwiz from SFU, library find, Metakey and finally concluded that DBWIZ shall power the Metasearching because SFU offered the complete suite which could be customized to meet the need of the researchers as well as libraries.
- ♦ For Clustering we were only able to find one open source Document Clustering Server from Carrot and hence have to use the same without evaluation. However we found that this

clustering engine is very widely used by leading commercial vendors and search engines which gave us the confidence.

- For Relevancy we did not find any open source utilities however found lingo algorithm which could be used for developing relevancy.
- Integrating the various open source components found

  After finalizing the various components the next challenge was to integrate them to work seamlessly. However being open source software the components were developed using various technologies. For e.g. DBWIZ was based on Perl, Clustering was based on Java which presented various technical challenges.
- Developing the parsers for enabling search across various e-resources.
- As per NISO the parsers need to be developed using XML gateway, z.39.50 and screen scrapping.
- Browser Compatible user interface.
- As most the open source software is based on LINUX environment they were mostly compatible with Mozilla FIREFOX or OPERA however it was important the platform is browser independent.
- Develop customized programs for Subscription Identifier, counter compliant usage statistics, Export/save feature, and email.

Solution – Integrated Meta search platform – GIST*Find*

4. The Technical Aspect of GIST*Find*

4.1 Open Source Technology

GIST Find has been developed by integrating many open source initiatives to bring to libraries the integrated functionality in a single window. GIST FIND is powered by open source applications like the DBWIZ from the Researcher Suite developed at Simon Fraser University Library (researcher.sfu.ca). Researcher is an award winning integrated suite of open source products for locating and managing electronic information resources, designed for use by students and researchers in academic, government and corporate libraries.

4.2 Technology:

The clustering technology is powered by software developed by The Carrot2 Project (www.carrot2.org).

4.3 W3C Standards:

GIST Find is built on W3C standards i.e. XHTML, DOM, XML which makes the applications very robust, Interoperable and customizable.

4.4 NISO Metasearching Initiatives:

GIST Find follows the meta-searching standards set by NISO and uses Z39.50 protocol, XML Gateway, HTML Parsing which ensures results integrity between GIST Find and Original Source.

4.5 Harvests Usage Statistics using NISO Protocol:

GIST Find uses NISO Protocols to harvest usage statistics providing the ability to extract data in acceptable formats like Excel, CSV etc.

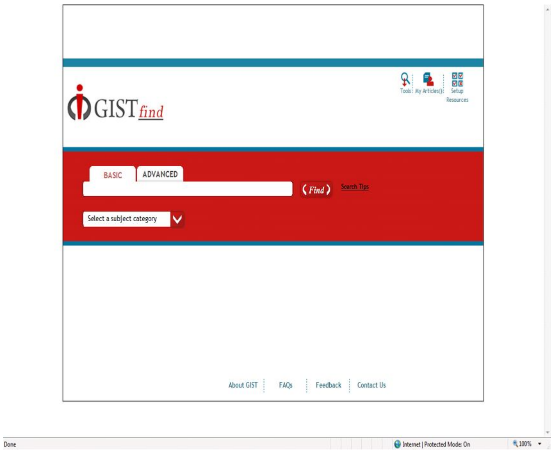
4.6 Generates Counter Compliant Reports and Customized Reports:

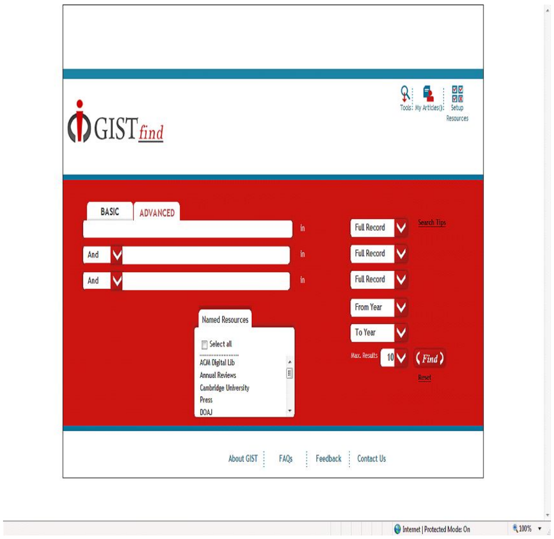
GIST Find has built usage statistics platform which provides libraries with counter compliant reports and also ability to develop customized reports for consortium and institution.

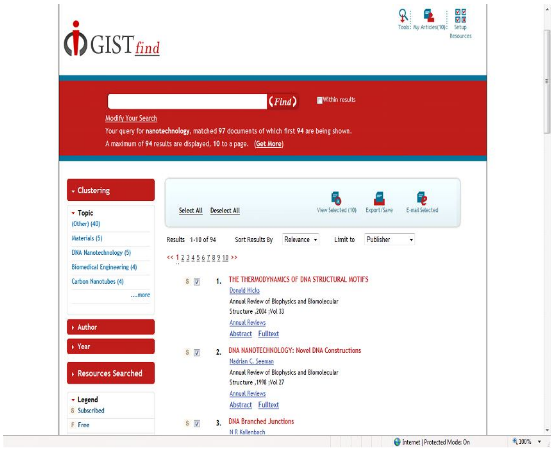
5. Conclusion

Despite the potential benefit of open source technology in reducing users information overload and improving the effectiveness of access to on line information and applying advanced information retrieval techniques with open source solution is one approach to manage and developing an affordable and reliable solution for the libraries. We believe that there is still room for improvements and many strategies yet to be explored. The work presented in this paper is but an early stage of such study in India in library community.

GIST FIND – Integrated Metasearch Platform : Screen Shots



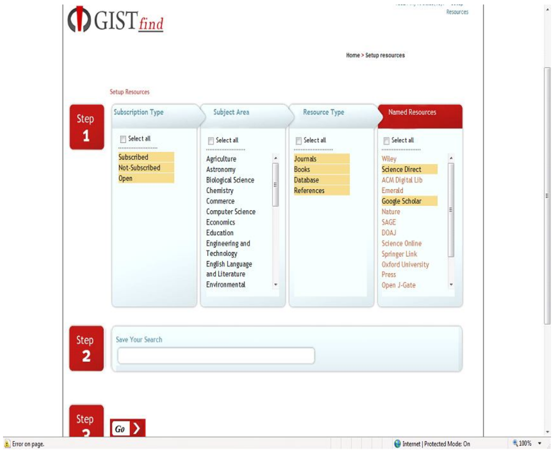




Your query for **nanotechnology**, matched **97** documents of which first **94** are being shown.
A maximum of **94** results are displayed, **10** to a page. (Get More)

Results 1-10 of 94

1. THE THERMODYNAMICS OF DNA STRUCTURAL MOTIFS
   Donald Hicks
   Annual Review of Biophysics and Biomolecular Structure ,2004 ;Vol 33
   Annual Reviews
   Abstract Fulltext
2. DNA NANOTECHNOLOGY: Novel DNA Constructions
   Nadrian C. Seeman
   Annual Review of Biophysics and Biomolecular Structure ,1998 ;Vol 27
   Annual Reviews
   Abstract Fulltext
3. DNA Branched Junctions
   N R Kallenbach



References

1. **Enterprise Search Summit;** November 6-7 2007; San Jose McEnery Convention Center, San Jose, CA

2. **Contemplating Federated Search, Melissa Rethlefsen,** Library Journal, 7/15/2008

3. **Federated Search** 101, Alexis Linoski and Tine Walczyk, 7/15/2008

4. **Free your Search with Open Source, Karen Koombs,** 7/15/2008

5. **Building Bearcat, Lisa A. Ellis, Joseph Hartnett and Michael Waldman,** 7/15/2008

6. Web Links and Resources:-

http://www.dwheeler.com/oss_fs_why.html
http://www.dwheeler.com/oss_fs_eval.html
http://open-source.gbdirect.co.uk/migration/benefit.html
http://www.osv.org.au/index.cgi?tid=5
http://sandro.groganz.com/wiki/Category:Open_Source_Marketing
http://en.wikipedia.org/wiki/Open_source#Terms_based_on_open_source
http://infomotions.com/musings/ossnlibraries/
http://en.wikipedia.org/wiki/Library_2.0
http://www.libraryjournal.com/article/CA6365200.html
http://connect.educause.edu/Library/EDUCAUSEQuarterly/OpenSourceSoftwareinEduca/46592?time=1230267273
http://www.degreetutor.com/library/managing-expenses/open-source-library
http://www.acm.org/ubiquity/views/v4i47_dorman.html

About Authors

Mr. Collin D'mello, Vice President, Sales & Publisher Relations, Global Information Systems Technology Pvt. Ltd. (GIST)
E-mail: collin@gist.in

Mr. Amit Jha, Manager,
Research & Development Global Information Systems Technology Pvt. Ltd. (GIST)
E-mail: ajha@gist.in