Bielefeld Academic Search Engine (BASE): an end-user oriented institutional repository search service

Dirk Pieper and Friedrich Summann
Bielefeld University Library, Bielefeld, Germany

Abstract

Purpose: This paper describes the activities of Bielefeld University Library in establishing OAI based repository servers and in using OAI resources for end-user-oriented search services like BASE (Bielefeld Academic Search Engine).

Design/methodology/approach: BASE uses the search engine technology Fast Data Search.

Findings: BASE is able to integrate external functions of Google Scholar. The search engine technology can replace or amend the search functions of a given repository software. BASE can also be embedded in external repository environments.

Originality/value: The paper provides an overview over the functionalities of BASE and gives insight into the challenges that have to be faced when harvesting and integrating resources from multiple OAI servers.

Keywords: Search engines, University libraries

Paper type: Technical paper

In a position paper of the Scholarly Publishing and Academic Resources Coalition (SPARC) published in 2002, Raym Crow defined an institutional repository as a "digital collection capturing and preserving the intellectual output of a single or multi-university community" (p. 4). Repository servers can help institutions to increase their visibility and, in addition, they are changing the system of scholarly communication.

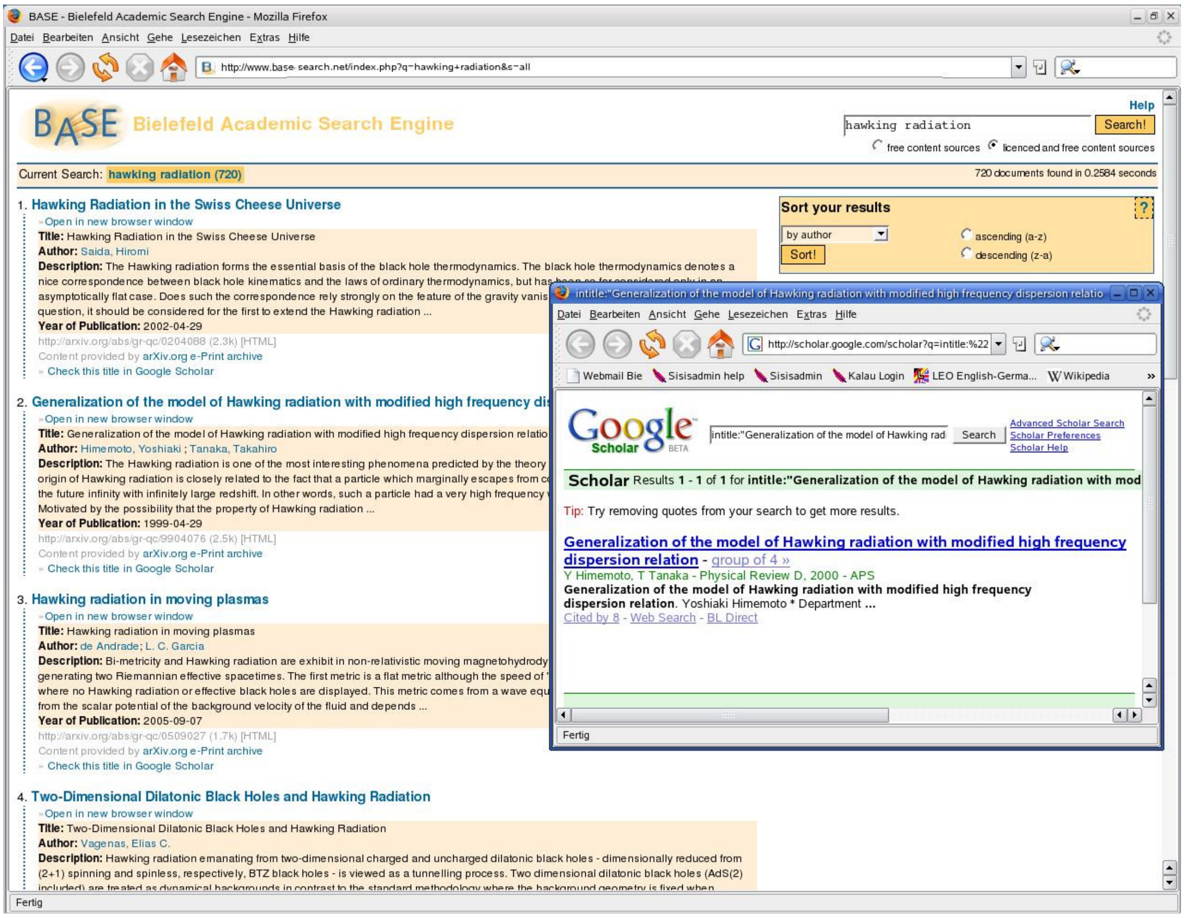
For several years libraries have been facing price increases for scientific journals, which has led to the fact that the proportion of published journal articles libraries can provide access to has decreased. Aside from the necessity of developing new subscription models for journal articles, libraries now can do a lot to increase the availability of journal articles by providing access to open access journals and documents and building repositories for their home university.

In addition to several self-designed repository based services, Bielefeld University Library has developed, with support from the Norwegian company Fast Search & Transfer [1], an end-user-oriented search service for multiple scholarly full text archives, digital repositories and preprint servers on the World Wide Web called Bielefeld Academic Search Engine (BASE) [2] (see Lossau, 2004; Lossau and Summann, 2004). At the time of the 8th International Bielefeld Conference in February 2006 BASE contained about 2.7 million documents in 189 collections. An up-to-date overview, including the content providers, is available online in a comprehensive list [3]. Characteristics of BASE include:

- intellectual selection of resources;
- indexes contain only quality-assured academic online resources from all academic disciplines;
- transparency about the data resources included in BASE;
- searches metadata and full text (depending on the data source);

- discloses Internet resources of the "deep web" (such as 500,000 digitised pages of historical journals and review organs of the German Enlightenment);
- displays search results as bibliographic data and full text hits;
- various options to sort result sets;
- search refinement for authors, keywords, document type, or language.

The newest feature of BASE, which was first presented to the public at the 8th International Bielefeld Conference, is the ability to check BASE results in Google Scholar by a title search, so that users can directly see if, and how many times, an article is cited in Google Scholar. Figure 1 shows the result of a simple search for “hawking radiation”. After clicking on the link “Check this title in Google Scholar” for the second hit, a window with the result in Google Scholar pops up:


BASE - Bielefeld Academic Search Engine - Mozilla Firefox
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://www.base-search.net/index.php?q=hawking+radiation&s=all
BASE Bielefeld Academic Search Engine
Help
hawking radiation
Search!
free content sources
licenced and free content sources
Sort your results
by author
Sort!
ascending (a-z)
descending (z-a)
Fertig

Current Search: hawking radiation (720)

720 documents found in 0.2584 seconds

1. Hawking Radiation in the Swiss Cheese Universe
» Open in new browser window
Title: Hawking Radiation in the Swiss Cheese Universe
Author: Saida, Hiromi
Description: The Hawking radiation forms the essential basis of the black hole thermodynamics. The black hole thermodynamics denotes a nice correspondence between black hole kinematics and the laws of ordinary thermodynamics, but has
asymptotically flat case. Does such the correspondence rely strongly on the feature of the gravity vanis
question, it should be considered for the first to extend the Hawking radiation ...
Year of Publication: 2002-04-29
http://arxiv.org/abs/gr-qc/0204088 (2.3k) [HTML]
Content provided by arXiv.org e-Print archive
» Check this title in Google Scholar

2. Generalization of the model of Hawking radiation with modified high frequency dis
» Open in new browser window
Title: Generalization of the model of Hawking radiation with modified high frequency dispersion relatio
Author: Himemoto, Yoshiaki ; Tanaka, Takahiro
Description: The Hawking radiation is one of the most interesting phenomena predicted by the theory
origin of Hawking radiation is closely related to the fact that a particle which marginally escapes from c
the future infinity with infinitely large redshift. In other words, such a particle had a very high frequency
Motivated by the possibility that the property of Hawking radiation ...
Year of Publication: 1999-04-29
http://arxiv.org/abs/gr-qc/9904076 (2.5k) [HTML]
Content provided by arXiv.org e-Print archive
» Check this title in Google Scholar

3. Hawking radiation in moving plasmas
» Open in new browser window
Title: Hawking radiation in moving plasmas
Author: de Andrade; L. C. Garcia
Description: Bi-metricity and Hawking radiation are exhibit in non-relativistic moving magnetohydrody
generating two Riemannian effective spacetimes. The first metric is a flat metric although the speed of '
where no Hawking radiation or effective black holes are displayed. This metric comes from a wave equ
from the scalar potential of the background velocity of the fluid and depends ...
Year of Publication: 2005-09-07
http://arxiv.org/abs/gr-qc/0509027 (1.7k) [HTML]
Content provided by arXiv.org e-Print archive
» Check this title in Google Scholar

4. Two-Dimensional Dilatonic Black Holes and Hawking Radiation
» Open in new browser window
Title: Two-Dimensional Dilatonic Black Holes and Hawking Radiation
Author: Vagenas, Elias C.
Description: Hawking radiation emanating from two-dimensional charged and uncharged dilatonic black holes - dimensionally reduced from (2+1) spinning and spinless, respectively, BTZ black holes - is viewed as a tunnelling process. Two dimensional dilatonic black holes (AdS(2) included) are treated as dynamical backgrounds in contrast to the standard methodology where the background geometry is fixed when

intitle:"Generalization of the model of Hawking radiation with modified high frequency dispersion relatio
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://scholar.google.com/scholar?q=intitle:%22
Webmail Bie
Sisisadmin help
Sisisadmin
Kalau Login
LEO English-Germa...
Wikipedia
Google Scholar BETA
intitle:"Generalization of the model of Hawking rad
Search
Advanced Scholar Search
Scholar Preferences
Scholar Help
Fertig

Scholar Results **1** - **1** of **1** for **intitle:"Generalization of the model of Hawking radiation with mod**

Tip: Try removing quotes from your search to get more results.

Generalization of the model of Hawking radiation with modified high frequency dispersion relation - group of 4 »
Y Himemoto, T Tanaka - Physical Review D, 2000 - APS
Generalization of the model of Hawking radiation with modified high frequency dispersion relation. Yoshiaki Himemoto * Department **...**
Cited by 8 - Web Search - BL Direct

Figure 1: Checking citations of BASE search results in Google Scholar

The integration of a citation counting functionality in institutional repository servers is something for which there is a high demand from academics. So the basic idea behind this feature is that while our search engine software, like most repository software, does not provide this yet, there is no reason why an external system like Google Scholar should not be used for this functionality. BASE is also flexible enough to combine data collections in special views, e.g. for all institutional repository collections. It is also possible to replace or to amend the search functions of a given repository software, which we will demonstrate in the near future when establishing an e-scholarship repository server for Bielefeld University.

Regarding the technical background of repository based services, Bielefeld University Library is working on both sides – establishing OAI services and using those of other institutions. It has been running the publications server BieSOn (Bielefeld Server for Online Publications) since 2004 and BieTAS (Bielefeld Text Archive Server), a platform for the comfortable dissemination of distributed contents under different systems, since 2005. Both services are registered OAI servers. The e-scholarship repository server mentioned above will soon follow. On the other hand, collecting metadata via OAI harvesting plays an important role in different search environments. We feed more than 550,000 documents into the local library catalogue as online-accessible material. This includes documents such as theses and dissertations, digitized books and journals. 2.4 million journal articles, harvested for example from Citeseer, PubMed, ArXiv, the Directory of Open Access Journals (DOAJ), and Biomed Central, are added to the local article database as electronic references. The most relevant dataflow is feeding all type of material (images, maps, videos, multimedia components and web pages) into BASE. In addition to the crawling of web pages, the data processing of OAI metadata has become the main focus of the BASE data workflow. To process this data we have established a pre-processing stage to transform OAI metadata Dublin Core XML files into an internal XML format. As the next step these files are transformed by a series of different internal and external processing stages into a file which can be indexed directly afterwards. This includes, in particular, format transforming (e.g. PDF, PS, ZIP or Office files), language detection, normalizing and lemmatizing. After indexing, all this information is ready for retrieval - in our case for accessing it via the BASE search interface based on PHP scripts.

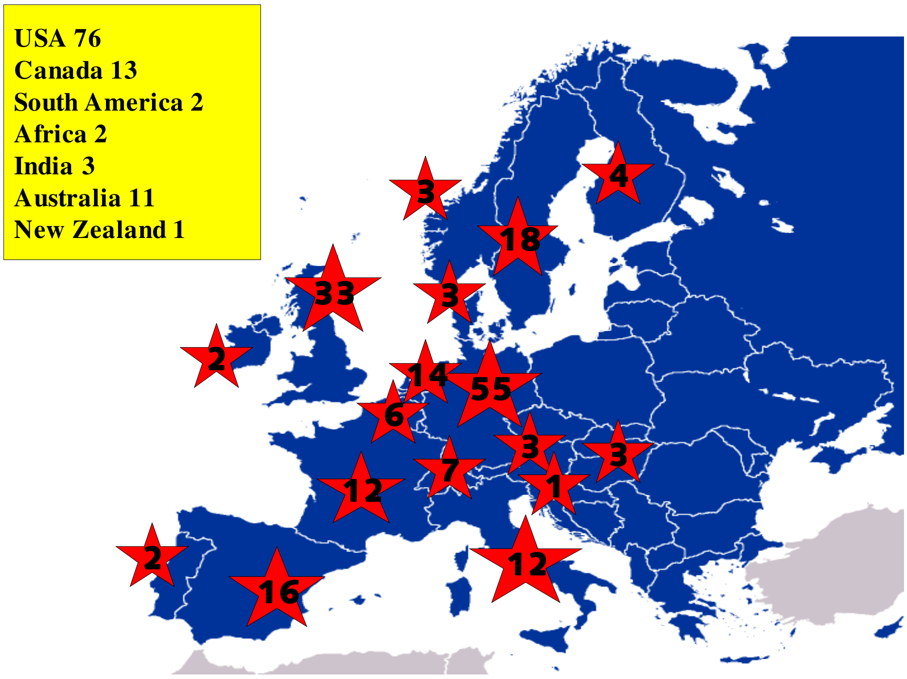


Figure 2: Institutional repositories covered in BASE

Focusing on the harvesting process, the first challenge is how to find relevant academic OAI servers. We monitor the well-known registries of openarchives.org, Eprints, the experimental registry of the University of Illinois, DSpace, and, since January 2006, the Directory of Open Access Repositories (DOAR). These resources provide different numbers of listed servers and a different quality of stability and status of their data. The map in Figure 2 shows the geographical distribution of repositories covered in BASE with the main focus on academic repositories in Europe. The map illustrates the strong position of Sweden, Germany, the U.K. and the Netherlands. Additionally BASE integrates a large number of repositories from the U.S., accompanied by Australian and Canadian repositories.

The harvesting procedure itself proved to be complicated in detail and posed a number of problems. To handle these problems, and to make the process more efficient, we adopted and developed a small collection of software tools. Firstly, as our core system we are using the Perl-based open source *harvester* delivered by the U.S. company FS Consulting. While harvesting we faced some minor error situations which we were able to solve by adapting the source code. Relatively often the delivered OAI data contains XML errors. This is a serious problem to deal with, because XML parsing is then impossible for the whole file. Therefore, we wrote an *XML validator and repairer script* which removes the invalid records and saves the correct ones. A so-called *Harvest Watcher* monitors the harvesting processes and reports the results (count of records, time stamp). A cronjob script, the *OAI Resource Updater*, automatically requests repository servers in defined intervals. Finally, the *Registry Watcher* takes the valuable HTML or XML files which are delivered from some of the registries, compares them with the BASE harvesting configuration file and lists the servers already covered, and, much more interestingly, the resources unprocessed up to now. All these additional tools have been written in Perl.

While the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) defines the OAI harvesting process very clearly, the daily routine brings up a broad set of serious problems and challenges. Among the long list of problems are non-responding servers, document links which do not work, invalid XML files and OAI data which only contains references either without any full text behind or where full text access is restricted for specific access. This situation requires a lot of observation and a lot of detailed configuration work in response. Some short examples will deliver a deeper insight into the problems. Some installations deliver the URL for access in the “source” field. Sometimes authors’ names are inserted as a list with different separators in one Dublin Core (DC) “creator” field. On the other hand, one can also find author names split into two different fields. The content of the DC field “subject” contains classification codes, classification terms or true subject headings without any qualifying. In some cases one can find author and title information in this field as well. A bitter experience is that fields with rather standardized content, such as the DC fields “date” and “language”, vary in a very broad way. In particular, the “language” field should determine the language of a resource correctly, because this is the basis for several linguistic processing steps. Another significant quality problem is the fact that correct citation information for journal articles is missing among many OAI servers. To handle all these problems it has been necessary to put a lot of effort into registration and configuration. This has taken more time than expected.

As a conclusion of our experiences we can establish a list of personal rules derived from the harvesting activities. Firstly, standard repository software is very useful, not only for the work of system administrators, but for the OAI harvesting procedure as well. The delivered results of those systems are strongly standardized, which makes the integration process much easier. Besides that, small collections generally only bring up small problems, probably because the content is more basic and more homogeneous. A serious problem is combining metadata and corresponding full text via OAI because the linking method and the presentation of the documents vary in practice. This is the reason why we only succeeded in realising this approach for a few installations. Another point shown by experience is that libraries as data providers produce a higher level of data quality, probably because they have much broader experience with bibliographic metadata. An important point for improving the quality is participating in the OAI community. Writing e-mails to the repository administrators helps, but sadly only sometimes. In 60 percent of our e-mail contacts we got a response, and in half of those cases the problems were repaired within a week. Sometimes there was no response at all but some weeks later the problem was solved, perhaps an internal reaction to the e-mail contact. As a last point we have to mention that OAI data aggregation, and, in particular, using aggregator services, may produce problems. In some cases we faced duplicates, updating delays and loss of individual information on the way from the original repository to the aggregator service.

As a last topic, the aspect of integrating BASE in other services has necessitated an ambitious approach to technology, especially the idea of embedding the system in external repository-based environments. The easiest way of integration is to include a search form for BASE. This feature works already and there is HTML code available which can be incorporated into any user-defined web page. In cooperation with a German library software company we have developed an HTTP-based interface to integrate BASE retrieval in a more flexible way. This technique will be improved with a more comfortable interface. Besides this, we are working on a web services-based technology which accepts and responds with XML files, including search queries and result pages. In relation to the German project Vascoda we have discussed a concept for a federated search of different search engines with a high level of result-merging based on IT standards. Hopefully this approach will support the development of another type and quality of search environments in the future.

Notes

1. www.fastsearch.com
2. www.base-search.net
3. http://base.ub.uni-bielefeld.de/about_sources_english.html

References

Crow, R. (2002), *The Case for Institutional Repositories: A SPARC Position Paper*, available at: www.arl.org/sparc/IR/IR_Final_Release_102.pdf (accessed 9 May 2006).

Lossau, N. (2004), “Search engine technology and digital libraries, libraries need to discover the academic Internet”, *D-Lib Magazine*, Vol. 10, No. 6, available at: http://dx.doi.org/10.1045/june2004-lossau (accessed 9 May 2006).

Lossau, N. and Summann, F. (2004), “Search engine technology and digital libraries: moving from theory to practice”, *D-Lib Magazine*, Vol. 10, No. 9, available at: http://dx.doi.org/10.1045/september2004-lossau (accessed 9 May 2006).