New Developments in Electronic Publishing
AMS/SMM Special Session, Houston, May 2004
ECM4 Satellite Conference, Stockholm, June 2004
p. 91-97.

Bielefeld Academic Search Engine: a (Potential Information-)BASE for the Working Mathematician

Michael Höppner

Bielefeld University Library

Summery. A modern search engine based approach to scientific information retrieval is described as the consequent next step after building up digital libraries. Bielefeld University Library has just established two demonstrators for possible retrieval services, one of them from mathematics.

1. The customer's point of view

Services of an university library do not end in itself, but the "general user", i.e. the "customer", is in the centre of interest. Customer's demands can easily be described by

- simple and easy search facilities
- relevant and complete lists of results
- direct and transparent access to all retrieved items,

which can best be summarized as a "one-stop-shop" to meet the customer's demand for information. While customer's demands are legitimate and very easy to understand, customer's own skills very often are poor and inappropriate: So in practice Google has become the first choice for information retrieval and seldom there is an awareness of the really relevant scientific sources and the reliability of retrieved items.

Beyond traditional library services, the first step to keep the customer satisfied was building up "digital libraries", also called portals or subject gateways, which provide a simultaneous search in different databases by a single user entry, i.e. a meta-search, provide availability information for each retrieved item and provide access to the shelves, the electronic full-texts, or document delivery services with respect to the libraries holdings or licensed materials. In fact, digital libraries integrate printed and electronic materials, monographic and journal-article information, local and remote holdings, so they are called hybrid libraries also. A well-known example is the "Digital Library of North Rhine-Westphalia", which started in

© 2005 FIZ Karlsruhe

June 1999 and is now completely and transparently integrated in Bielefeld University Library's web-based information services.

While digital libraries are milestones to better information services in principle, some obstacles emerge very soon. Performance often is very poor, due to the performance of different target systems, and stability problems occur for the same reason. With respect to the content, digital libraries are restricted to databases and pay no attention to full-texts and web-pages in general. Last but not least, digital libraries tend to be either too complicated for the customer or are completely neglected because often they are secondary offers besides the usual library's information services only, contrary to the successful Bielefeld approach of integrated services.

From the customer's point of view, it is obvious that Google has become so successful. This search engine can be handled very easily, results are obtained in a split second and the relevance ranking seems to be reasonable in most cases, ignoring the fact that it is annoying to drill down huge lists of results in the remaining cases. So, why Google should not be the next step to better scientific information services? Obviously, a commercially driven system cannot guarantee long term accessibility, for commercial reasons indexing is done automatically without respect to different data types and structures, moreover relevance ranking of results is influenced by commercial interests, too. With respect to the relevance and completeness of the provided content, Google and other commercial search engines focus on the visible web, i.e. neglect 90% of information available at the web, especially database content.

2. The BASE-service, still a demonstrator

Although the objections against Google are serious, it is not true that search engines provide simple search facilities only, that search engines cannot handle structured, i.e. high-quality, data, and that search engines need a monolithic index structure which can be achieved at significant costs only. Modern search engines cope with heterogeneous data, i.e. with different types and formats like full-texts, structured meta-data, images, and binary data, they provide advanced navigation and browsing facilities like scientific taxonomies, thesauri, and cross-referencing, they offer flexible ranking and ordering schemes for list of results, and they can be based on distributed and federated indices. One of the most advanced product is Fast Data Search (current version 4.0.2) by Fast Search & Transfer in Oslo, Norway, which has proved its potential by being the platform for AllTheWeb, Scirus, FirstGov and lots of other well established web-sites. Furthermore, based on a cooperation with Prof. Guenthner's Centrum für Informations- und Sprachverarbeitung in Munich, Germany, Fast is going to implement linguistic tools like approximate search and cross language information retrieval which might even be used for an automatic extraction of meta-data.

Bielefeld University Library has been a pioneer in electronic information services since it was founded in 1967, so after developing

- the first OPAC on CD-ROM in Germany together with one of the first CD-ROM-networks in 1988,
- the document delivery service JASON together with the interdisciplinary journal-article database JADE in 1993, see [1],

- the integrated library information system IBIS in 1997, see [2],
- the Digital Library of North Rhine-Westphalia in 1999, see [5], which was transparently integrated in the library's service in 2000,
- an "intelligent search assistant" based on fuzzy logic in 2003, see [3],

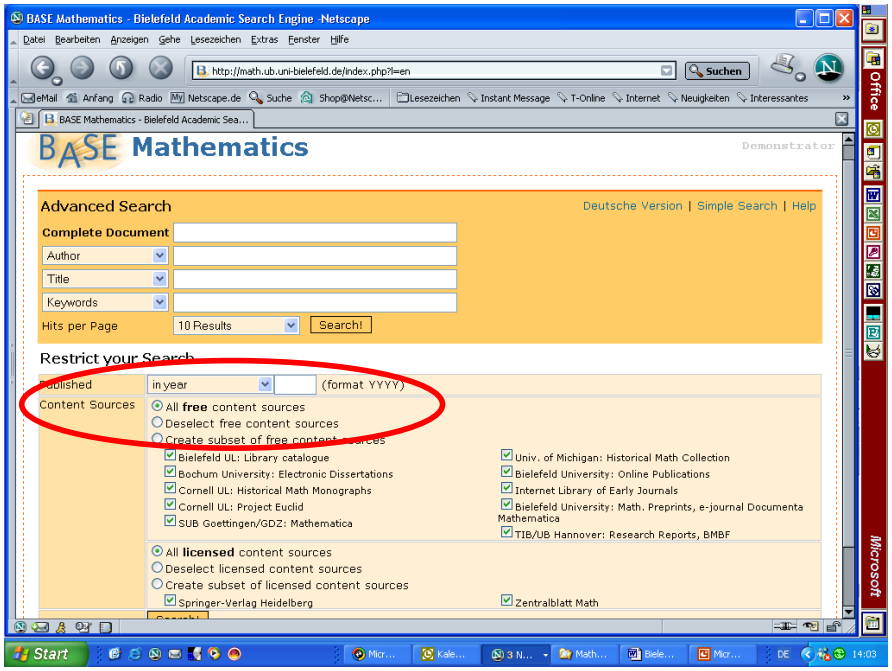
it was quite reasonable to exploit modern search engine technology for further service improvements. This is based on a cooperation agreement with Fast Search & Transfer.

Since it would have been too ambitious to cover all subjects from the very beginning, Bielefeld University Library has restricted its BASE-activities, i.e. the Bielefeld Academic Search Engine activities, to two fields up to now: The first field is a set of "Digital Collections" from various subject fields, most of them from the humanities including a core set of journals from the German age of enlightenment, actually retro-digitised by ourselves and indexed by the Göttinger Akademie der Wissenschaften long time ago. The second field is a representative sample of electronic sources from "Mathematics". This is reasonable because mathematics is covered bibliographically very well and there are lots of "eic-projects", i.e. electronic information and communication projects, including retro-digitising projects in this field. Moreover, Bielefeld has been a former regional subject library for mathematics, and one of the first scholarly published open access e-journals, the "Documenta Mathematica", is published in Bielefeld. Both applications are not a real service at the moment, but still demonstrators of modern search engine applications for scientific information retrieval. Both are online since the beginning of this Conference on "New Developments in Electronic Publishing of Mathematics" at `http://base.ub.uni-bielefeld.de`. Comments and criticism on both applications are invited.

"BASE : Mathematics" covers lots of well-known content sources which can be found on the BASE pages and need not be introduced to a mathematical audience, see Picture 1.

BASE comes along with a German and an English user interface and provides a simple Google-like search line for basic search as well as an advanced search interface (see Picture 1) for different search terms. Search can be restricted to free (accessible) content sources, i.e. sources not only provided to certain users by license agreements, and can be further restricted to specified sources. Search results can be refined by different aspects such as author, MSC-classes, document type, and language and search results can be improved by a similarity search with respect to a single retrieved item, see:

BASE search results are presented by their meta-data if there are any, or by a standardized description. Full-texts can be accessed directly from the list of results by standard browser and plug-in facilities. The BASE interface can be customized for different purposes, up to now there is a completely different experimental view prepared in the corporate design of the University of Michigan's Library.



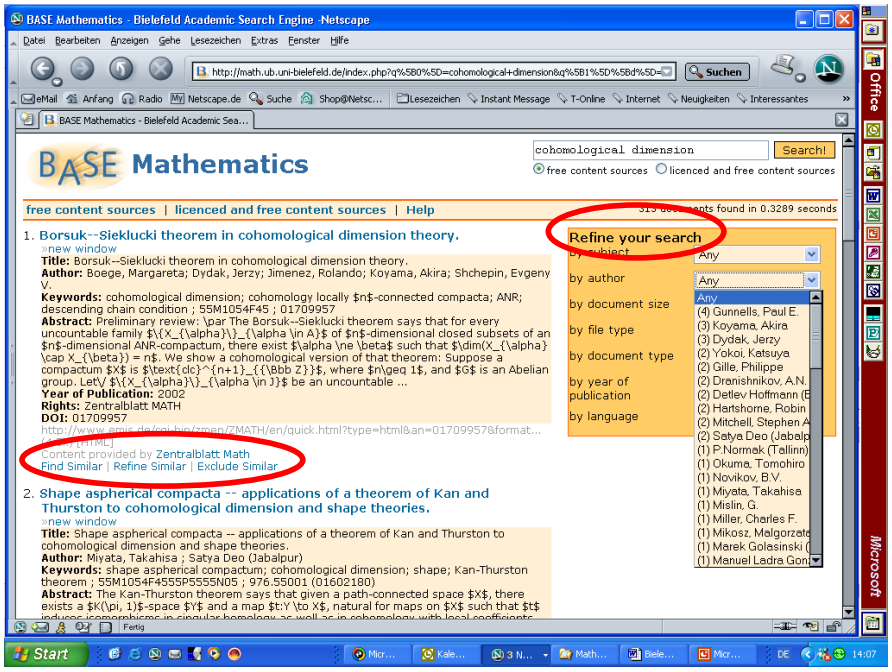
Picture 1. BASE advanced search and content sources

3. The BASE-architecture

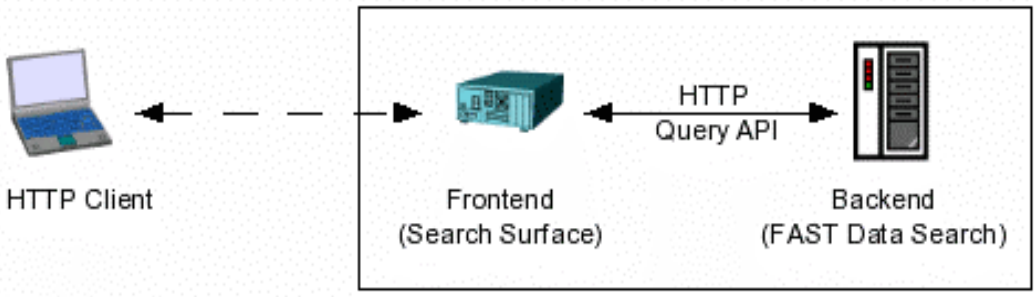
The system-architecture is based on a client-server-structure, moreover there are different frontend and backend servers (see Picture 3) which both can be enhanced to multi-node-systems. While the client needs nothing more than a web-browser, the frontend server is a standard high-performance web-server (Bielefeld: 2 CPU Pentium III/800 MHz with 1 GB RAM at the moment, based on the usual LAMP-structure, i.e. SUSE-Linux 9.0, Apache 1.3, and PHP 4.3 with PostgreSQL instead of My-SQL at the moment) running PHP (with C++ and JAVA-APIs as an alternative recommended by Fast), and the backend server should be a high-performance multiprocessor system with fast hard disk arrays (Bielefeld: 2 CPU Pentium IV/2.8 GHz, 270 GB RAID 5 at the moment, based on SUSE Linux 9.1, with Linux Red Hat Enterprise as an alternative recommended by Fast) running Fast Data Search 4.0.2.

The frontend server is used for providing the user interface, for processing the search results, and presenting the lists of results only. The frontend server communicates via http and a query API with the backend server which is used not only for query and result processing with respect to actual user-requests, but for crawling and harvesting data, for file traversing, for pre-processing data into an internal XML-format by Fast as well as for document processing and indexing, i.e. for building the system's backbone, as shown in the following picture.

Up to now, we have implemented http- and OAI-interfaces to databases from our well-understood meta-search environment as well as a bundle of SGML-interfaces for proprietary formats from various sources. Further we have realized the appropriate database calls and prepared the retrieved items for Fast's internal processing. Our index is based on the 15 well-known fields of the Dublin Core Meta-Data Set with 5 additional index fields, e.g. dcisb for



Picture 2. Refinement of results

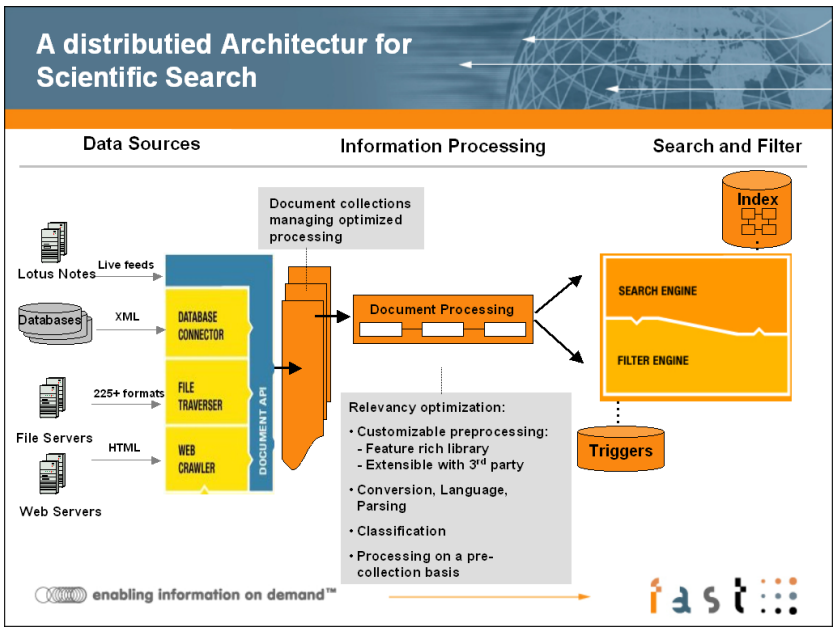


Picture 3. BASE client-server-architecture

ISBN and ISSN, dcdoi for DOI or similar document identifiers, dcyear for processing a year as an integer, dctype for data types like meta-data, full-texts etc., and dcrights for the names of sources by copyright reasons.

4. The vision of a service

So far, BASE uses basic tools of Fast Data Search only. Even this can be done more advanced in the future by extending the index structure and by including more different document and data types. The quality of information retrieval can be improved further by using the more advanced features of Fast, some of them are under development. Besides intelligent user interfaces, an integrated citation analysis, and improved, maybe personalized ranking-algorithms, the implementation of linguistic tools are on the agenda. This includes approximate search and a cross-language information retrieval by the help of appropriate dictionaries. The quality of the search facilities of BASE can be complemented by push services. Progress towards



Picture 4. Fast data search architecture

more advanced feature depends heavily on the consolidation of our current workflow by automating more and more parts.

Developing "BASE : Mathematics" to a real service for the mathematical community does not only depend on further exploitation and development of software tools, but on adding content, content, and even more content. This can be done by a cooperative approach, in fact this seems to be the only chance to reach the goal. So, one of our next interests is to build up and test linking with external indices to realise federated searches. Moreover, Bielefeld University Library is interested to build up partnerships for the development of BASE with respect to technical improvements and to the enrichment of content for making "BASE : Mathematics" not only a potential, but a real base for the working mathematician.

A more detailed description of BASE, including motivation, technical aspects, and workflow, is given in [4] and [6].

References

[1] Höppner, Michael: *Zeitschriftenschnellbestell- und -liefersystem JASON-NRW*. In: Mitteilungen der Deutschen Mathematiker Vereinigung **1** (1995), 47–50.

[2] Höppner, Michael: IBIS - *ein internetbasiertes Informationssystem für Bibliotheken*. In: Der Weg in die Informationsgesellschaft – Teil II /Martin Grötschel. Mitteilungen der Deutschen Mathematiker Vereinigung **4** (1997), 46–47.

[3] Homann, Ingo; Binder, Wolfgang: *Ein Fuzzy-Rechercheassistent für Bibliographische Datenbanken*. In: Informatik: Forschung und Entwicklung **19** (2004), 97–108.

[4] Lossau, Norbert: *Search Engine Technology and Digital Libraries: Libraries Need to Discover the Academic Internet*. In: D-Lib Magazine **10** (2004), `http://www.dlib.org/dlib/june04/lossau/06lossau.html`.

[5] Pieper, Dirk; Summann, Friedrich: *Die Entwicklung des Zugangssystems der Digitalen Bibliothek NRW*. In: Nachrichten für Dokumentation **50** (1999), 397–405.

[6] Summann, Friedrich; Lossau, Norbert: *Search Engine Technology and Digital Libraries: Moving from Theory to Practice*. In: D-Lib Magazine **10** (2004), `http://www.dlib.org/dlib/september04/lossau/09lossau.html`.

Received December 30, 2004