

A Multidisciplinary Search Engine for Scientific Open Access Documents

Matthias Lösch, University Library, Bielefeld

Abstract

There is no doubt that the increased visibility of publications deposited in freely accessible documents servers is an important motivation for scholars to publish their works Open Access. Bielefeld Academic Search Engine (BASE) is a service provider that indexes scientific content stored in both institutional and subject repositories. After a short introduction to the technical and conceptual background, this article describes the features of BASE and finally projects future developments to be taken out to further improve the system.

1 Introduction

In the field of Digital Libraries, one of the most striking success stories is that of institutional and subject repositories-servers which store scientific documents of a particular academic institution or subject. Over the past decade, they have increasingly gained importance for the scholarly communication (Horstmann 2007), which is mostly due to the efforts made by the Open Archives Initiative (OAI) to develop common standards for them.

However, repositories are just one part of a bigger picture envisioned by the OAI. In 2001, the Protocol for Metadata Harvesting (OAI-PMH) was released, which standardises the exchange of metadata between repositories and so-called service providers (Lagoze and Van de Sompel 2001). Based on this protocol, the service providers can aggregate the metadata of several repositories on a regular basis, a process referred to as harvesting in the language of OAI-PMH. On top of that metadata, the service providers can build value-added services for their users, for example federated search.

Subsequent to the release of the protocol, service providers began to emerge, one of the first being OAIster (Hagedorn 2003). In 2001, Bielefeld University Library started with the conceptual development of an academic search engine based on OAI-PMH. After several tests and beta releases, Bielefeld Academic Search Engine (BASE, <http://base-search.net>) finally went productive in 2004, at that time indexing 500,000 documents coming from 15 servers. In the meantime, the system has been constantly growing, and by now provides access to almost 25,000,000 documents from more than 1,700 data sources. The BASE index is continuously updated by regularly harvesting the known sources, but also by adding new repository servers (Pieper and Summann 2006).

2 Motivations for Developing BASE

Six years ago, Lossau (2004) asked provocatively whether Google, Yahoo or Microsoft would be the only portals to global knowledge by 2010. Thankfully we can now deny this question, but his diagnosis of libraries being challenged by commercial internet services is still up to date.

This diagnosis was based on the observation that commercial players were increasingly outperforming library services in the field of digital information. Their user interfaces are more convenient, and often they provide access to the actual full texts while library systems are limited to bibliographic data. On the other hand, regarding sustainability and academic quality of the indexed information, library databases are still unmatched.

BASE was developed to get the best of both worlds, that is to index only high-quality content coming from intellectually selected resources, and to present it through a streamlined user interface that would also provide access to full texts wherever possible.

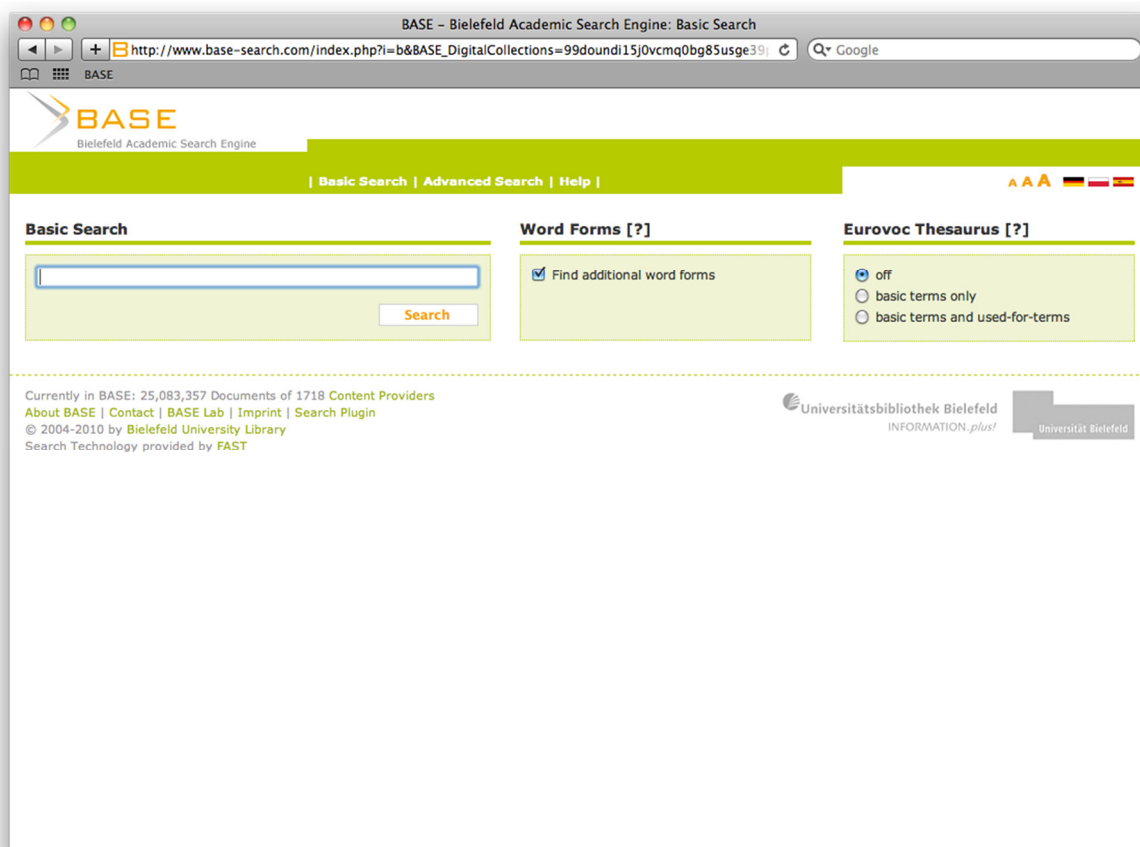
Another important aspect is that with the rise of Open Access, articles published with commercial publishers suffer from a loss of visibility, as pre- or postprints deposited in repositories are gaining importance. However, these Open Access publications are scattered across the different repository servers lacking a central entry point. It is thus crucial for scholarly communication that the repositories are accessible through central services like BASE.

3 Using BASE

3.1 Searching

Figure 1 shows the BASE search page. The first thing to notice is its simplistic layout—an important lesson learned from the commercial search engines. Besides the large search field and the search button, there are only two further options to take: A checkbox for query expansion in order to find additional synonyms, and a selection for whether the system should use the Eurovoc Thesaurus to do a multilingual search. Both options are preset to reasonable settings so they will not unnecessarily bother the average user. Nevertheless, there is also an advanced search mode that offers a lot more options to satisfy power users as well.

Figure 1: The BASE search page

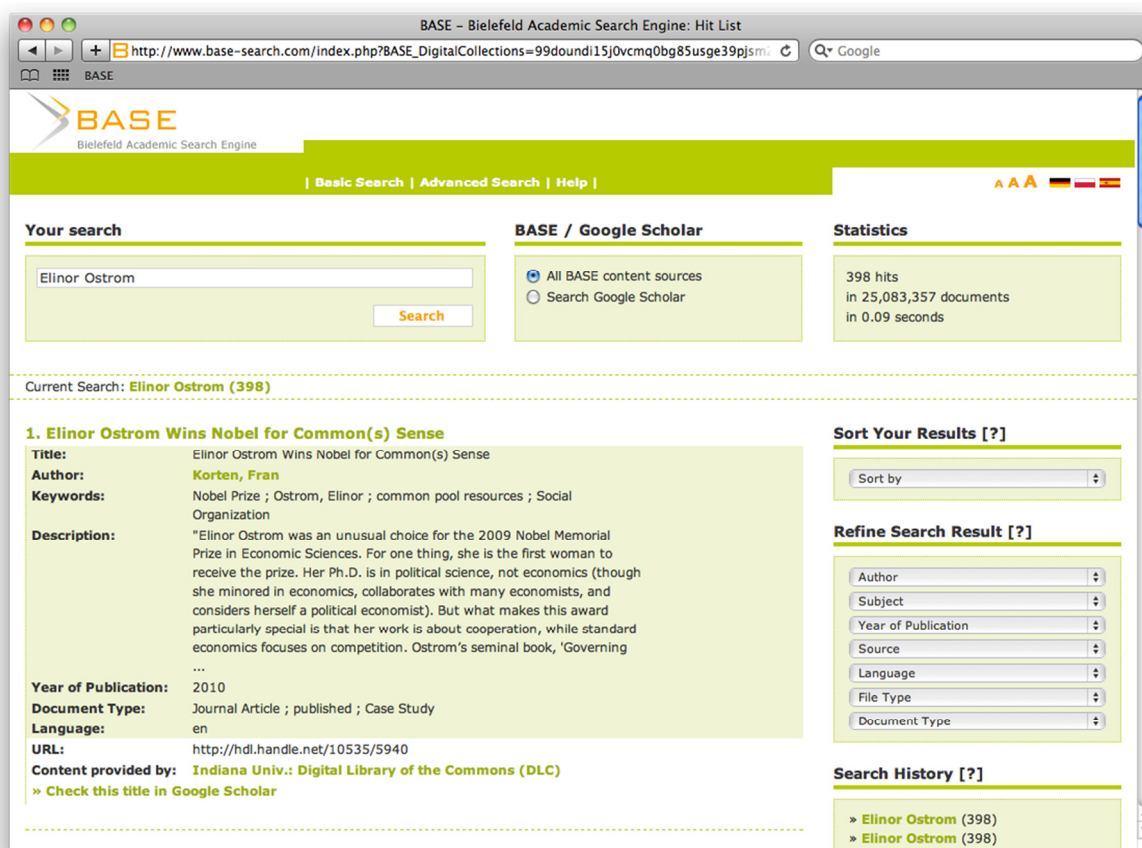


3.2 Results and Search Refinement

The results view (Figure 2) features a detailed and clearly structured display of the rich bibliographic data provided by the OAI format. By clicking on the title of a result entry, the user gets redirected to this particular document at the hosting repository, from where in many cases he or she can access the actual full text. A click on one of the authors of a result triggers a new search for that author.

Furthermore, users can also refine their current search by using the drop-down menus on the right. Here one can gradually narrow the result set simply by limiting the results to those featuring a particular author, language, or subject, or even only to those coming from a particular repository.

Figure 2: The results view



By using the drop-down menus on the right, users can narrow the result list until they find what they need.

3.3 API


For third-party organisations, an application programming interface (API) exists which allows for integrating the BASE index into their own infrastructures. The interface is currently being used by several library catalogues and meta catalogues like the Karlsruher Virtueller Katalog (KVK), and by meta search engines like MetaGer.

4 Future Perspectives

Of course there is always room for improvements. Although the bibliographic information in the OAI data already feature a high degree of detail, there is a lack of consistent subject indexing. This currently prevents even more valuable services, like, for example, subject-based browsing across all documents as shown in Figure 3.

To adress this problem, we have teamed up with researchers from the department of Text Technology at Bielefeld University and from the department of Computer Science at Leipzig University, and launched a project called "Automatic Enhancement of OAI Metadata". Within this project, we are going to automatically assign Dewey Decimal Classification numbers (Dewey, Mitchell, and Alex 2005) to documents indexed by BASE.

Figure 3: Prototypical DDC-based browsing interface



BASE Lab » Browsing (Menu | List)

0 Computer science, information & general works	30 Social sciences, sociology & anthropology	330 Economics
1 Philosophy & psychology	31 Statistics	331 Labor economics
2 Religion	32 Political science	332 Financial economics
3 Social sciences	33 Economics	333 Economics of land & energy
4 Language	34 Law	334 Cooperatives
5 Science	35 Public administration & military science	335 Socialism & related systems
6 Technology	36 Social problems & social services	336 Public finance
7 Arts & recreation	37 Education	337 International economics
8 Literature	38 Commerce, communications & transportation	338 Production
9 History & geography	39 Customs, etiquette & folklore	339 Macroeconomics & related topics

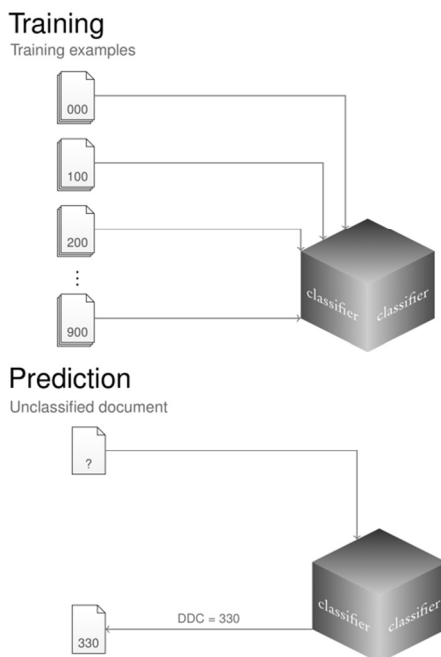
How to browse

This browsing tool is subdivided into 3 levels. Example: Main class 5 (Science), Division 54 (Chemistry), Section 547 (Organic chemistry). Click an entry to get to the next sub-level. **Click** an entry to start searching BASE for documents within this main class, clicking for a main class, the divisions and sections are automatically searched as well, if you search for a division the sections are automatically searched as well.

Currently, only about 150,000 documents carry a Dewey number and are therefore accessible by browsing.

This automatic subject indexing is going to be realised by machine learning-driven text classification based on the OAI data (Mehler and Waltinger 2009). More specifically, a so-called supervised classifier is going to be used, which learns how to classify by looking at correctly labeled training examples before. This process is depicted in Figure 4.

Figure 4: Automatic assignment of Dewey numbers to documents using a supervised classifier



5 Conclusion

With BASE, a performant and user-friendly search engine for digital scientific information was created and has been well-established over the past years. The system is being continuously improved to further ease its usage. As Open Access publishing is constantly getting more important in the academic world, the visibility of pre- or post-prints deposited in institutional repositories will become even more vital in the future. Therefore we see a growing need for independent and reliable search services like BASE.

References

- Dewey, M., Mitchell, J. S., and Alex, H. (2005): *Dewey Dezimalklassifikation und Register: DDC 22*, 22nd ed., Saur: München.
- Hagedorn, K. (2003): OAIster: A “No Dead ends” OAI Service Provider, *Library Hi Tech*, 21 (2): 170-181.
- Horstmann, W. (2007): Open Access International – Lokale Systeme, kooperative Netzwerke und visionäre Infrastrukturen, *Zeitschrift für Bibliothekswesen und Bibliographie*, 54 (4/5): 230-233.
- Lagoze, C. and Van de Sompel, H. (2001): The Open Archives Initiative: Building a Low-Barrier Interoperability Framework, in: *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, 54-62.
- Lossau, N. (2004): Search Engine Technology and Digital Libraries, *D-Lib Magazine*, 10 (6).
- Mehler, A. and Waltinger, U. (2009): Enhancing Document Modeling by Means of Open Topic Models: Crossing the Frontier of Classification Schemes in Digital Libraries by Example of the DDC, *Library Hi Tech*, 27(4): 520-539.
- Pieper, D. and Summann, F. (2006): Bielefeld Academic Search Engine (BASE): An End-User Oriented Institutional Repository Search Service, *Library Hi Tech*, 24 (4): 614-619.