

# From Theory to Practice



# Bielefeld Academic Search Engine

Friedrich Summann Bielefeld UL



# **Presentation Overview**

### Part 1:

General introduction to the BASE Digital Collections

objectives, content, potential

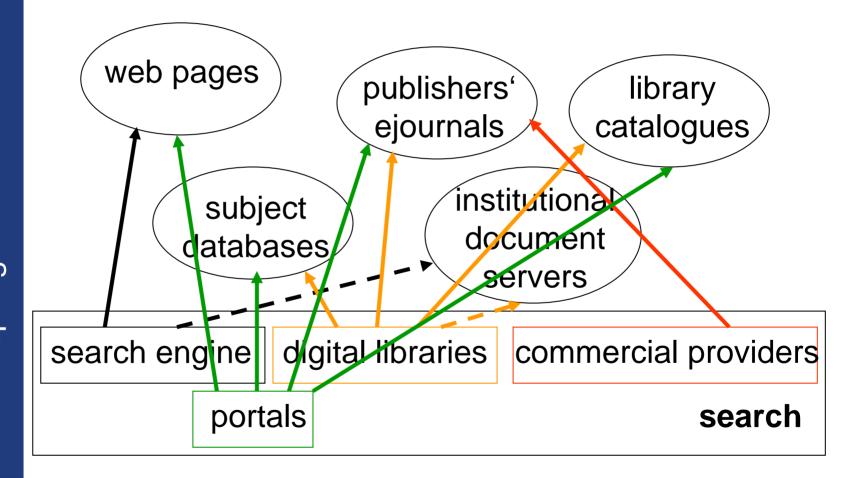
### Part 2:

Technical report about the BASE Digital Collections

- backend: harvesting and preprocessing, processing and indexing
- frontend: search interface, search and result presentation



# Academic Online Information: the Reality





# **Starting Point:**

Digital Library North-Rhine Westphalia (1998-2001)

### **Problems:**

- Users want search engine look and feel
- Search functionality too slow
- No integration of fulltext resources
- No integration of web pages



# Academic Online Information: the Vision

web pages

publishers' ejournals

subject databases

institutional document servers

search engine for academic online information



# Evaluation of Search Technology

(02/2002 - 08/2002)

- Convera
- Google
- mnoGo
- FAST
- -Lucene (2003)



# From Theory to Practice (1)

 Proof of concept with FAST Data Search (Math Demonstrator)

Math-Demonstrator			
Mathematic Basic Search - Search for	Deutsche Version   Advanced Search   Help		
Search!			
€ free content sources			
© 2003 by Bielefeld University Lib	prary		

Extension: Bielefeld Academic Search Engine: search service for digital collections

asic Search:		Deutsche Version   Advanced Search   Help
	Search!	
free content sources 🔘	icenced and free content sources	



# From Theory to Practice (2)

## The concept of BASE:

- collecting and making accessible in a single index a representative and heterogeneous set of academic online content
  - different document types
  - different data formats
  - fulltext and/or metadata
  - content from the "visible" and "invisible" web



# From Theory to Practice (3)

# Objectives of the BASE Search Engine:

- Real-life testing and implementation of FAST Data Search
- working with interoperability standards (OAI, XML)
- developing prototypes of intelligent and flexible user interfaces



# From Theory to Practice (4)

## Some general information:

- work on BASE in progress at Bielefeld UL since summer 2003
- team of 2 software developers
- software: FAST Data Search 3.2
- pilot project for national proposal "Using Search Engine Technology in Digital Libraries and Scientific Information Portals" by Bielefeld UL and North Rhine-Westphalian Service Centre for Academic Libraries (part of VDS in vascoda)



# OLF Spring Forum 2004

# The Content (1)

 about 600,000 documents indexed up to now in 15 collections:

## a)Metadata

- Project Euclid (6,516 articles):
  - harvested using OAI-protocol
- Bielefeld UL: Journals of the German Enlightenment (57,690 articles)
  - database output



# The Content (2)

# b)Fulltext without metadata / web content

- Documenta Mathematica (e-journal)
- preprint servers at Bielefeld University (together 18,301 documents)
- TIL/UL Hannover: project reports of BMBF (64 documents)



# The Content (3)

## c)Fulltext with metadata

- Springer journals (224,387 articles):
  - > metadata indexed, fulltext in preparation
- Bochum UL: electronic dissertations of Ruhr-University Bochum (1908 documents):
  - harvested using OAI-protocol
- Bielefeld UL: Online publications (369 documents)
  - harvested using OAI-protocol, metadata indexed, fulltext in preparation
- Oxford Internet Library of Early Journals (104,516 pages)
  - SGML export



# The Content (4)

## c)Fulltext with metadata (cont.)

- University of Michigan Historical Math Collection (772 documents)
- Cornell University Library Historical Math Monographs (630 documents)
  - metadata harvested using OAI-protocol, metadata and fulltext indexed separately
- SUB Göttingen/GDZ: Mathematica (427 documents)
  - harvested using OAI-protocol, up to now only metadata indexed



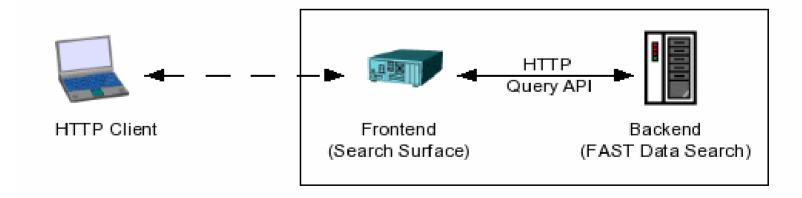
# The Potential

- making accessible different kinds of content sources in one index: web content and databases/catalogues
- indexing metadata and fulltext with or without metadata
- enhancing fulltext data by metadata extraction
- flexible and customizable frontend
- transferring performance and scalability of search engine technology to digital library world



# System Components

- separate frontend and backend server
- currently one frontend server
- can be easily enhanced with more servers
- currently one backend server (single node)
- can be enhanced to multi node system





# Dispatching

### Frontend:

- search interface (basic, advanced) (PHP)
- result processing and result presentation

### Backend:

- harvesting (Perl OAI harvester)
- preprocessing and conversion from BRS,
   OAI-DC and other DB formats with Perl
- filetraverser, crawler
- document processing and indexing of data
- query and result processing

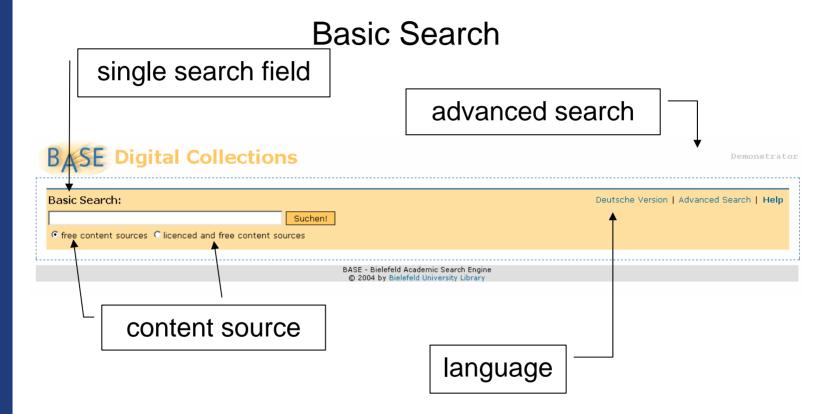


# The Frontend

- Siemens Primergy, 2 x 800MHz CPU
   1.28 GB RAM
- RAID1, Adaptec SCSI, 36GB
- SuSE Linux 9.0, Kernel 2.4.21-smp
- Apache web server with PHP 4

Bielefeld University Library web server







# Search Interface (2)

### **Advanced Search**

		search fi	ield selection			
Advand	ced Search:		che Version   Basic Search   H			
Complet	t <mark>e Document</mark>					
Author	+ -	I	Index			
Title						
Keyword			Index			
Hits per F	Page 10 Results	Search!	Index			
·						
rree c	yea	ır limit				
Restric	t your search		source selection			
ublished	in year (format Y	YYY)	Source Selection			
Content	☑ Bielefeld University: Online Publications		☐ Springer-Verlag Heidelberg (licenced)			
Sources	☑ Cornell UL: Project Euclid	<b>\</b>	☑ TIB/UB Hannover: Research Reports, BMBF			
☑ Cornell UL: Historical Math Monograp			☑ Bochum University: Electronic Dissertations			
	☑ SUB Goettingen/GDZ:Mathematica		☑ Univ. of Michigan: Historical Math Collection			
	☑ Bielefeld UL: Journals of the German Enligh		☑ Oxford University: Internet Library of Early Journals			
	▼ Bielefeld University: Math. Preprints, e-jour Mathematica	rnal Documenta				
	Search!					
	Scarcin					
	History					
<ul><li>linear</li></ul>	algebra src:allfree algebra src:allfree sj:numerische mathematik					
<ul><li>linear</li></ul>	algebra src:allfree au:[ralph. abraham]					
			Search history			
			Ocaron History			



# Result Presentation (1)

### **BASE Digital Collections** dccreator:[abel] Search! • free content sources • licenced and free content sources 169 documents found in 0,3060 seconds free content sources | licenced and free content sources | Help 1. Documenta Math 23 The Abel-Jacobi Map ... Refine your search Author: A. Iliev and D. Markushevich by content source All Title: The Abel-Jacobi Map for a Cubic Threefold and Periods of Fano Threefolds of Degree by author ΑII Abstract: The Abel-Jacobi maps of the families of elliptic quintid by subject All lying on a smooth cubic threefold are studied. It is proved that by document size ΑII dimensional projective space for quintics, and a smooth 3-dimensional variety pirational to the cubic itself for quartics... by file type All Subject: 1991 Digital Collections Subject Classification: 14J30,14J60,14J45 by year of publication http://www.mathematik.uni-bielefeld.de/DMV-J/vol-05/03.pdf (243.9k) [PDF] Content provided by Bielefeld University: Math. Preprints, e-journal Documenta Mathematica Find Similar | Refine Similar | Exclude Similar by document type ΑII 2. Oeuvres complètes de Niels Henrik Abel result change Author: Abel, Niels Henrik by language All Title: Oeuvres complètes de Niels Henrik Abel Publisher: publiée aux frais de l'Etat Norvégien par L. Sylow et S. Lie Search History **Year:** 1973 linear algebra src:allfree Pages: VIII, 621, 338 S. linear algebra src:allfree sj:numerische Remarks: Nachdr. d. 2-baendigen Ausg. Christiania 1881 in 1 Bd mathematik Further Authors: SYLOW, LUDVIG [Hrsq.] linear algebra src:allfree au:[ralph. v.ub.uni-bielefeld.de/netacgi/nph-brs?Sect5=C00P&d=C00P8 linear algebra src:allfree yr:1970-1980 /opac1.html%f=S&r=0&l=20&s18=HT005205602 (1.6k) [HTML] search history linear algebra src:allfree yr:1970-1980 Content provided by Bielefeld UL library catalogue Find Similar | Refine Similar | Exclude Similar linear algebra src:allfree yr:1970-1980 lg:eng au:[barker, g. p.] 3. Abel's equation and regular growth: variations on a theme by Abel linear algebra src:allfree lg:de Author: George Szekeres linear algebra src:allfree lg:de type:pdf Title: Abel's equation and regular growth: variations on a theme by Abel linear algebra src:allfree lg:de src:fulltext Source: Experiment, Math. 7 (1998), no. 2, 85-100 linear algebra src:umh umc



metadata display

# Result Presentation (2)

### 8. Nonlinear Abel integral equations

Find Similar | Refine Similar | Exclude Similar

Gorenflo, Rudolf; Nonlinear **Abel** integral equations; applications, analysis, numerical methods; Preprint / Universitaet (Berlin, West) / Fachbereich Mathematik; 248; Berlin; 1987

http://www.ub.uni-bielefeld.de/netacgi/nph-brs?Sect5=COOP&d=COOP&p=1&u=/netahtml /opac1.html&f=S&r=0&l=20&s18=HT005546358 (1.7k) [HTML] Content provided by Bielefeld UL library catalogue

9. Die Periodicitätsmoduln der Abel'schen Integrale erster Gattung als Functionen

eines Parameters aufgefasst für den Fall ...

Broecker, Hermann; Die Periodicitätsmoduln der Abel'schen Integrale erster Gattung als
Functionen eines Parameters aufgefasst für den Fall ...; 1893

http://gdz-srv2.sub.uni-goettingen.de:8091/cgm/servlet/dienstreq? verb=Display&Service=Repository&version=1&identifier=PPN314377646 (1.4k) [HTML] Content provided by SUB Goettingen/GDZ:Mathematica Find Similar | Refine Similar | Exclude Similar

### 10. Abel integral equations

Gorenflo, Rudolf; Vessella, Sergio; **Abel** integral equations; applications-motivated solution concepts; Preprint / Universitaet (Berlin, West) / Fachbereich Mathematik; 209.245; Berlin

http://www.ub.uni-bielefeld.de/netacgi/nph-brs?Sect5=COOP&d=COOP&p=1&u=/netahtml /opac1.html&f=S&r=0&l=20&s18=HT005417129 (1.6k) [HTML]

Content provided by Bielefeld UL library catalogue Find Similar | Refine Similar | Exclude Similar fulltext

Results: 1-10 | 11-20 | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | » New Search»

BASE - Bielefeld Academic Search Engine © 2004 by Bielefeld University Library



# DLF Spring Forum 2004

# The Backend

### Live-System:

- Intel PC Dual Pentium, 2 x 2.8 GHz CPU, 2GB RAM
- RAID5 + Hotspare 290 GB
- Suse Linux 9.0
- System report: 344.8GB total, 24.8GB used, 320GB free
  - > 15 Collections, > 600000 documents
  - FAST Search 3.2 (PHP 4, Python 2.2)

### **Test-System**

### (provided by FAST Search & Transfer):

- Dell PowerEdge, 1 x PIII 730 MHz CPU, 1.2 GB RAM
- RAID5, Adaptec SCSI, 66GB
- RedHat Linux 7.3, Kernel 2.4.22-pre5



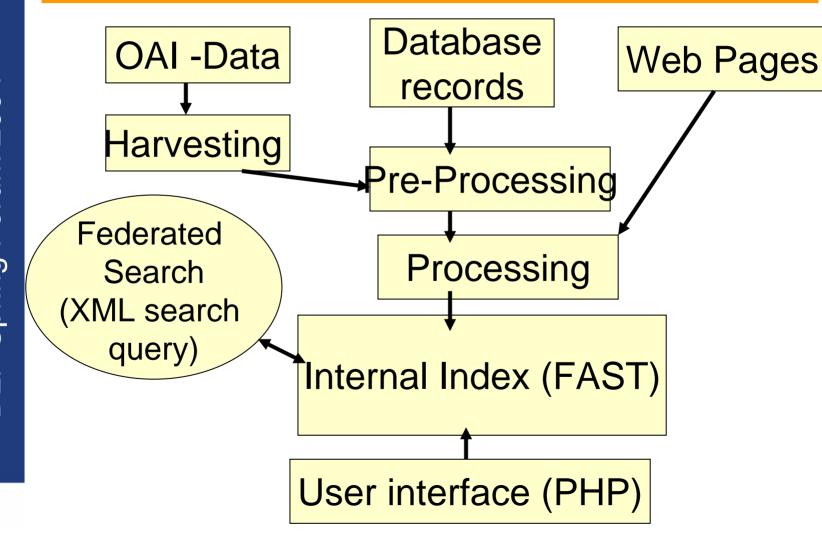
# Our current routines:

Integrating resources in a "Metasearch environment"

- Analyzing database interface (HTTP, Z39.50)
- Implementing search database calls
- Scanning results and preparing internal display format



# BASE data workflow Samples





# BASE data workflow tools

	FAST	Additional developments	
Data Loading	Crawler, File Traverser, DB Connector	OAI Harvester DB export	
Pre-processing		Perl, XSLT crosswalks	
Processing	Standard stages	Python stages	
Indexing	Indexer		
Access, Navigation	Search API	PHP scripts	



# Bielefeld Development Environment

- OAI Harvester tool (harvesting metadata) (open source)
- Perl, XSLT (Preprocessing)
- Python (Processing)
- PHP (frontend programming)
- Apache
- Linux



# Spring Forum 2004

# Harvesting

### harvested OAI-DC data

### assumed:

<date>1991</date>

### reality:

```
<date>-set=math&amp;until=2003-11-03&amp;metadataPrefix=oai_dc-
</date>
```

```
<date>1903;1903-09-02</date>
<date>[c1911] </date>
```

```
<date>1906-1928 [v.1, &apos;28] </date>
<date>[192-?] </date>
```

```
<date>C. Gerolds sohn,</date>
<date>28 cm.</date>
```



# Preprocessing (1)

```
in:
<date>[c1915]</date>
out:
<element name="dcdate"><value>[c1915]</value></element>
<element name="dcyear"><value>1915</value></element>
in:
<language>ENG</language>
out:
<element name="dclanguage"><value>eng</value></element>
<element name="language"><value>en</value></element>
```



# LF Spring Forum 2004

# Preprocessing (2)

in (binary data from CDROM database):

... Japanese. Esperanto summary ...

out:

```
<element name="dclanguage"><value>Japanese. Esperanto
summary</value></element>
<element name="language"><value>jp</value></element>
```

<element name=,,secondarylanguage"><value>eo</value>

</element>



# Preprocessing (3)

### Summary

- language code (text and ISO 639-2 to ISO 639-1)
  - > ISO 639-1 (de,fr)
  - > ISO 639-2/B (ger,fre), ISO 639-2/T (deu,fra)
- date filtering
- XML encoding and conversion
- generating unique document id (doi, document number, ...)
- general filtering and error correction
- building of body content (author, title, description, ...)
- fulltext link extraction



# F Spring Forum 2004

# Processing (1)

### Filetraverser sources

- loading of preprocessed content with filetraverser
- processing with self created pipelines
  - language detection from title and description
  - setting of mime type
  - generate teaser based on description, meta data or body
  - tokenize selected fields
  - lemmatize (run, runs, running, ran) and synonyms (security, safety) → dictionary based
  - vectorizer (for analyzing similarities between docs)



# **JLF Spring Forum 2004**

# Processing (2)

### Crawler sources

- crawling of selected web sites according to rules
- crawling of fulltext link lists
- system processing with stages (partly self developed)
  - deleting format (mime type)
  - format detection
  - uncompressing (zip, gzip, tar, ...) and setting of new format
  - set content type (metadata, fulltext, mixed, unknown)
  - Postscript conversion (Ghostscript)
  - PDF conversion (XPDF)
  - SearchMLConverter (FAST)
  - language and encoding detection (FAST)



# Indexing

### Current indexstructure

- enhanced by 15 DC fields
- additional 5 index fields
  - dcisbn (ISBN, ISSN)
  - dcdoi (DOI or similar identifier)
  - dcyear (filtered year as integer)
  - dcstype (metadata, fulltext, ...)
  - rights (name of source)



# Further Development (1)

## **Frontend**

- Templating (customised views)
- search interface (search API)
- combining metadata record and corresponding fulltext in result display
- truncation



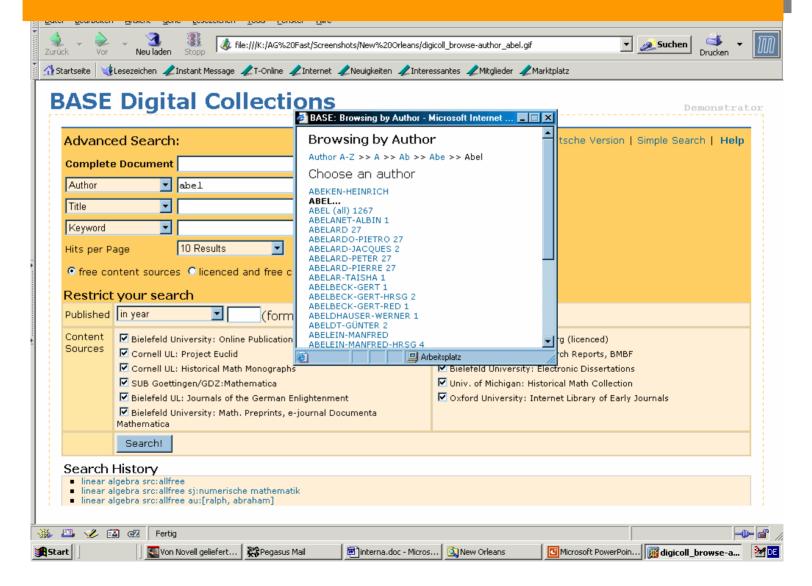
# **Customised views**

UNIVERSITY LIBRARY						
MIRLYN VISITORS	MyLibrary   SITEMAP   SY	STEM OUTAGES   HELP				
LIBRARY NEWS	SEARCH FOR DIGITAL CO	LLECTIONS				
Budget Impacts & the Library's Collections			Search			
New and Improved MIRLYN on the Way!						
Manage Citations with RefWorks						
Brown v. Board Digital Archive						
EMPLOYMENT OPPORTUNITIES						
U-M Alumni and .ibrary Friends						
MIRLYN VISITORS	MyLibrary   SITEMAP   S	YSTEM OUTAGES   HELP				



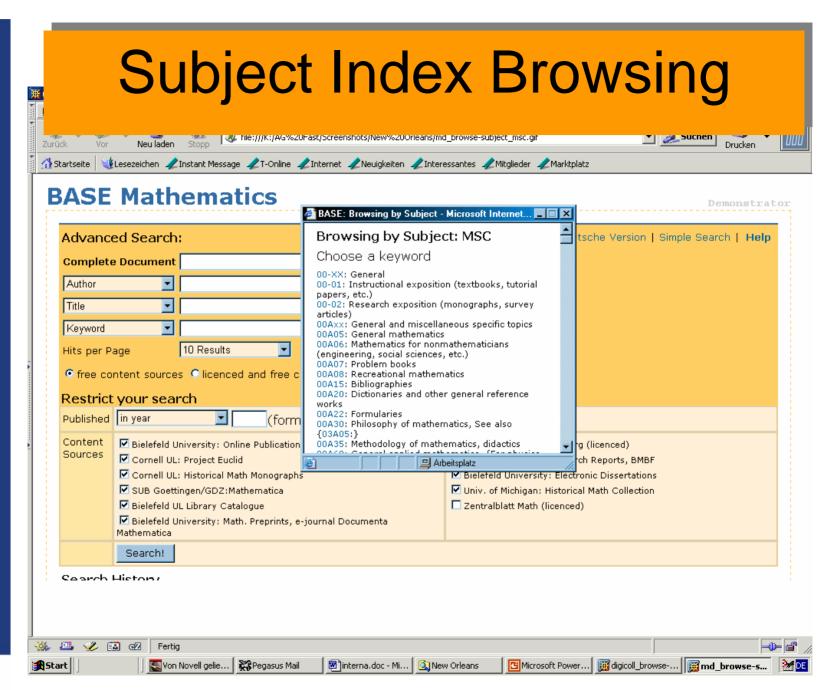


### **Author Index Browsing**









### Further Development (2)

### Backend

- automation of harvesting and content preprocessing
- Federated search, linking with external indexes
- search result improvement (ranking, boosting, linguistics)
- performance optimisation
- support of standard protocols (Z39.50, OAI, SOAP) as a target system



### **Further Visions**

- Anchor analysis
- Link topology analysis
- Citations analysis
- Automatic linguistic analysis of anchor texts
- Push services
- Personalized ranking
- Cross-language information retrieval



## DLF Spring Forum 2004

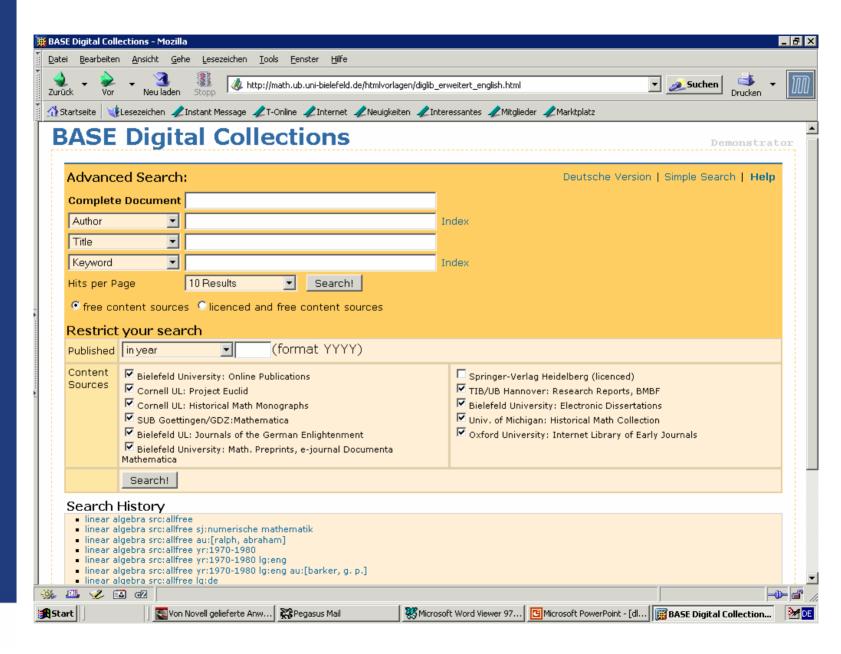
Thank you!



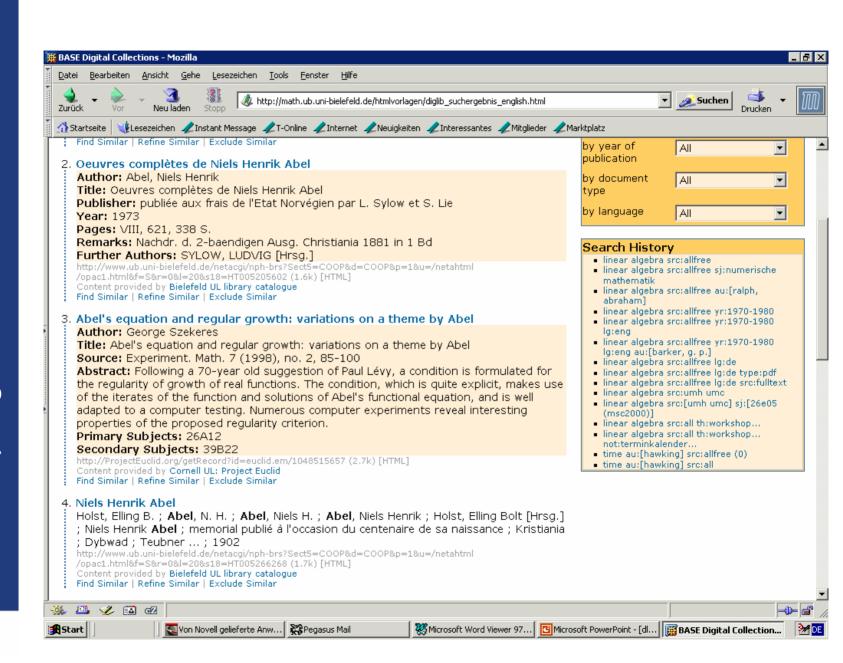


### **Bielefeld University Library** DLF Spring Forum 2004

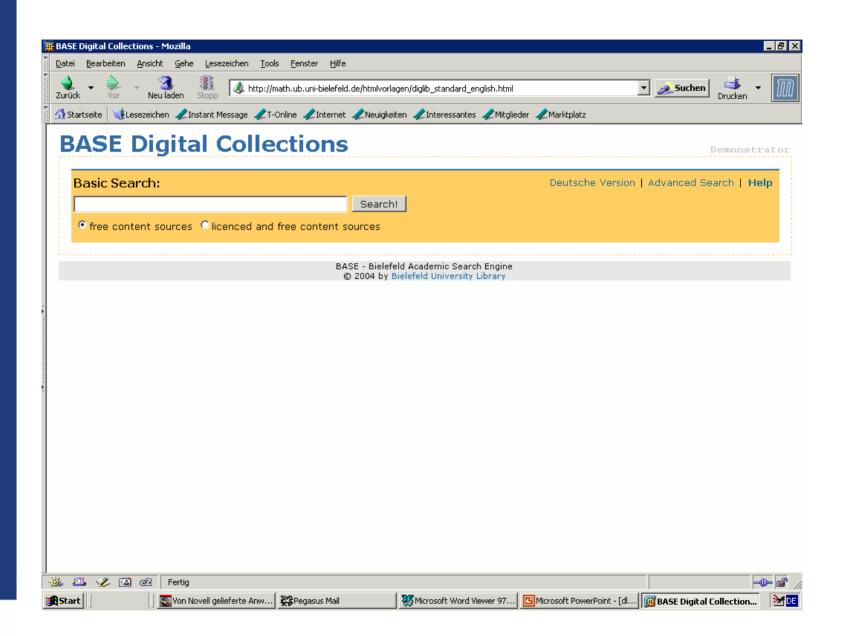




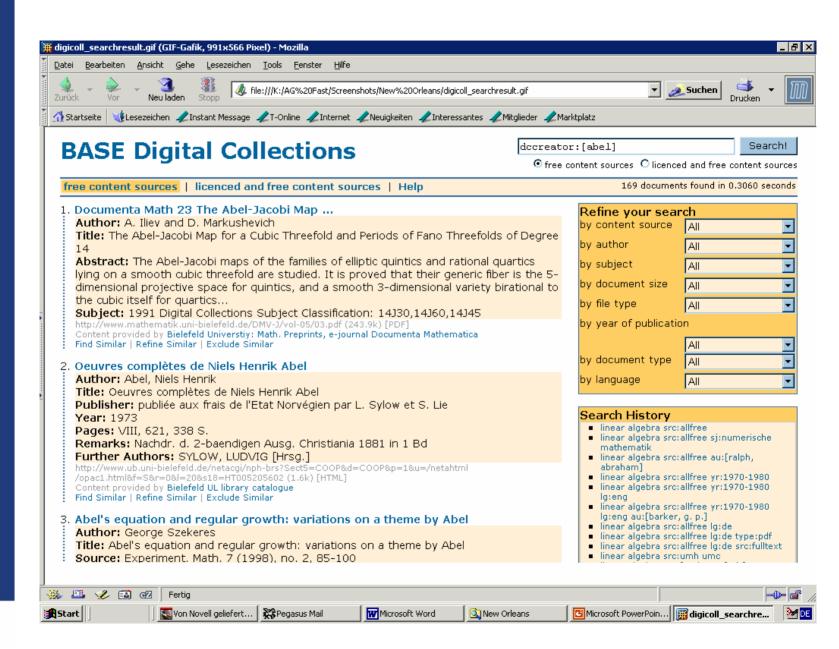




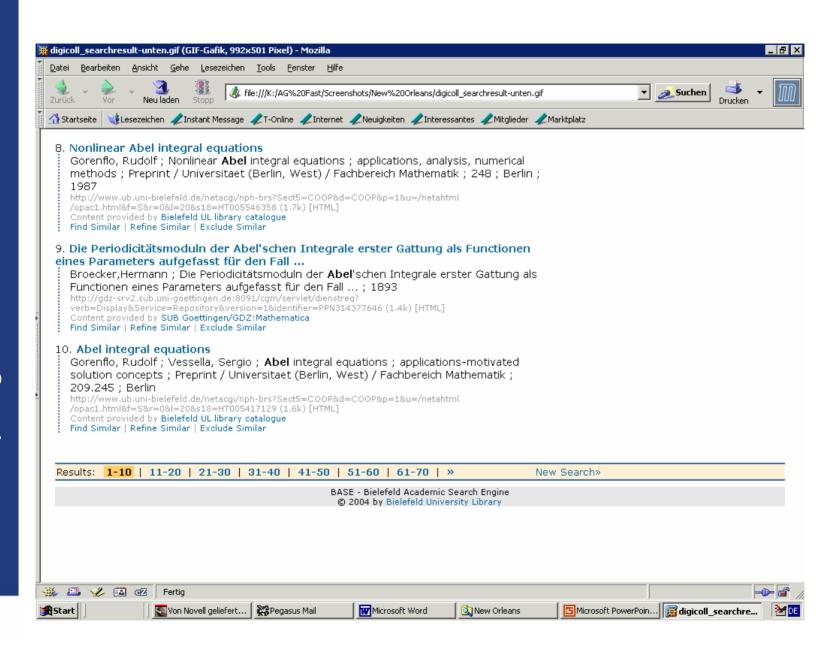




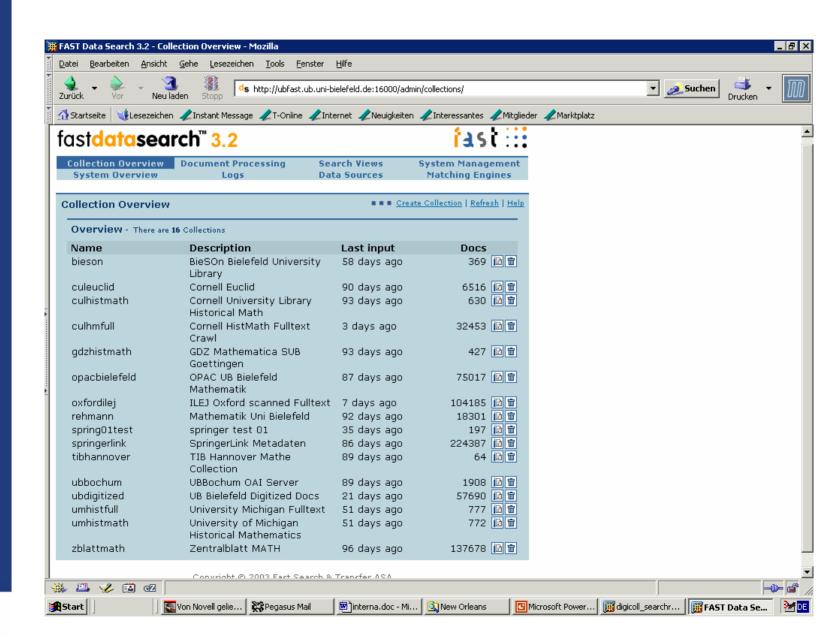










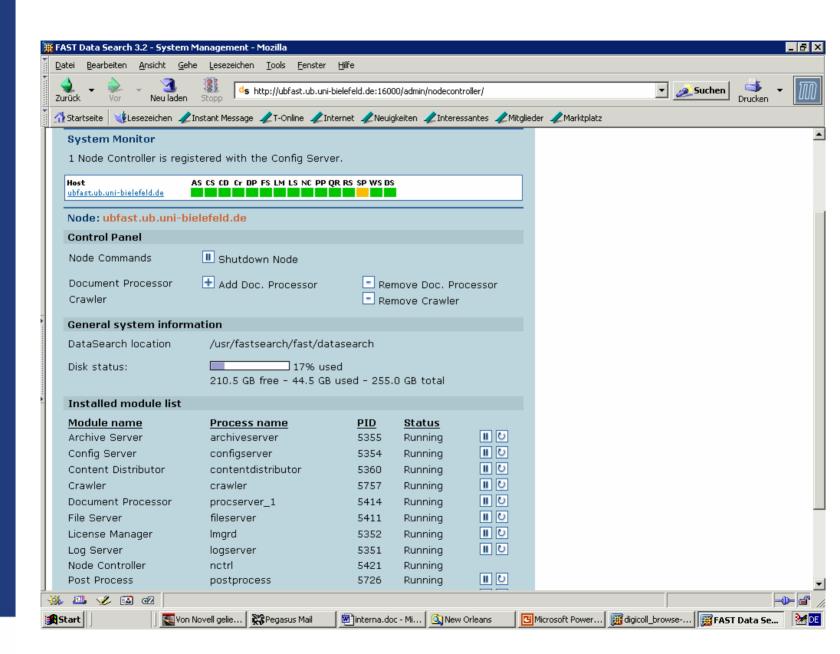




### **Bielefeld University Library** DLF Spring Forum 2004

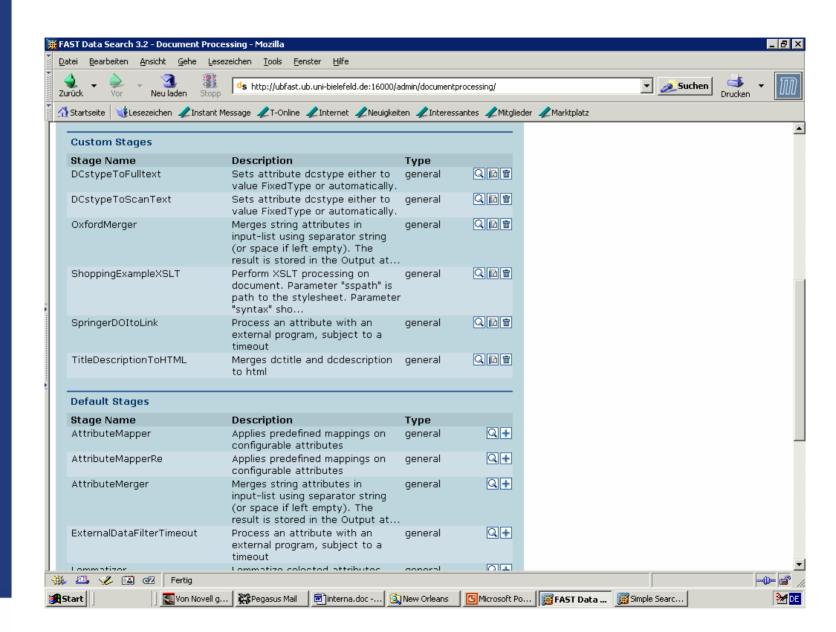
## LF Spring Forum 2004





## LF Spring Forum 2004





## **Bielefeld University Library**

BASE	Digital	Coll	lectio	ns

Demonstrator

lvanced Search	Deutsche Version   Simple Search   Help
mplete Document	
uthor 🔽	
tle 🔽	
eywords 👤	
s per Page 10 Results Search!	
free content sources Clicenced and free content sources	
estrict your Search	
olished in year (format YYYY)	
ntent Bielefeld University: Online Publications	☐ Springer-Verlag Heidelberg (licenced)
urces Cornell UL: Project Euclid	▼ TIB/UB Hannover: Research Reports, BMBF
Cornell UL: Historical Math Monographs	Bochum University: Electronic Dissertations
SUB Goettingen/GDZ: Mathematica	☑ Univ. of Michigan: Historical Math Collection
☑ Bielefeld UL: Journals of the German Enlightenment	🔽 Oxford University: Internet Library of Early Journals

BASE - Bielefeld Academic Search Engine © 2004 by Bielefeld University Library

## **Bielefeld University Library**

# DLF Spring Forum 2004

	BASE Digital Collections	Demonstrator
	Basic Search:	Deutsche Version   Advanced Search   Help
	Suchen!  • free content sources • licenced and free content sources	
i.		BASE - Bielefeld Academic Search Engine © 2004 by Bielefeld University Library