CERN Search Engine Status

CERN IT-OIS

Tim Bell, Eduardo Alvarez Fernandez, Andreas Wagner

HEPiX Fall 2010 Workshop

3rd November 2010, Cornell University

Outline

- Enterprise Search
  - What is Enterprise Search?
  - Requirements for protected search
  - Enterprise Search solution providers
- CERN Search
  - Background & Objectives
  - Architecture, Document Workflow
  - Search Relevancy, Ranking algorithms
- Improving TWiki Search
  - Indexing TWiki Topics
- Google Comparison
  - What about Google Search Appliance ?
  - Comparison with FAST
- Future Steps
  - FAST Search Server 2010

Enterprise Search

- **Components of Enterprise Search:**
  - Document retrieval
    - Not only web pages
    - Database/XML data (CDS, Indico, Phone data)
  - Search Engine with ranking
  - Integration within existing infrastructure
    - Authentication
    - Authorization
  - Protected documents
    - Getting access to document data
    - Recording ACLs as well
- Enterprise Search is not only a question about the search technology used!

Protection Requirements

- Protected information must not ‘leak’ from search
  - Search engine only presents data you can read

- To obtain full results, authentication is required
  - Results filtered by your access rights

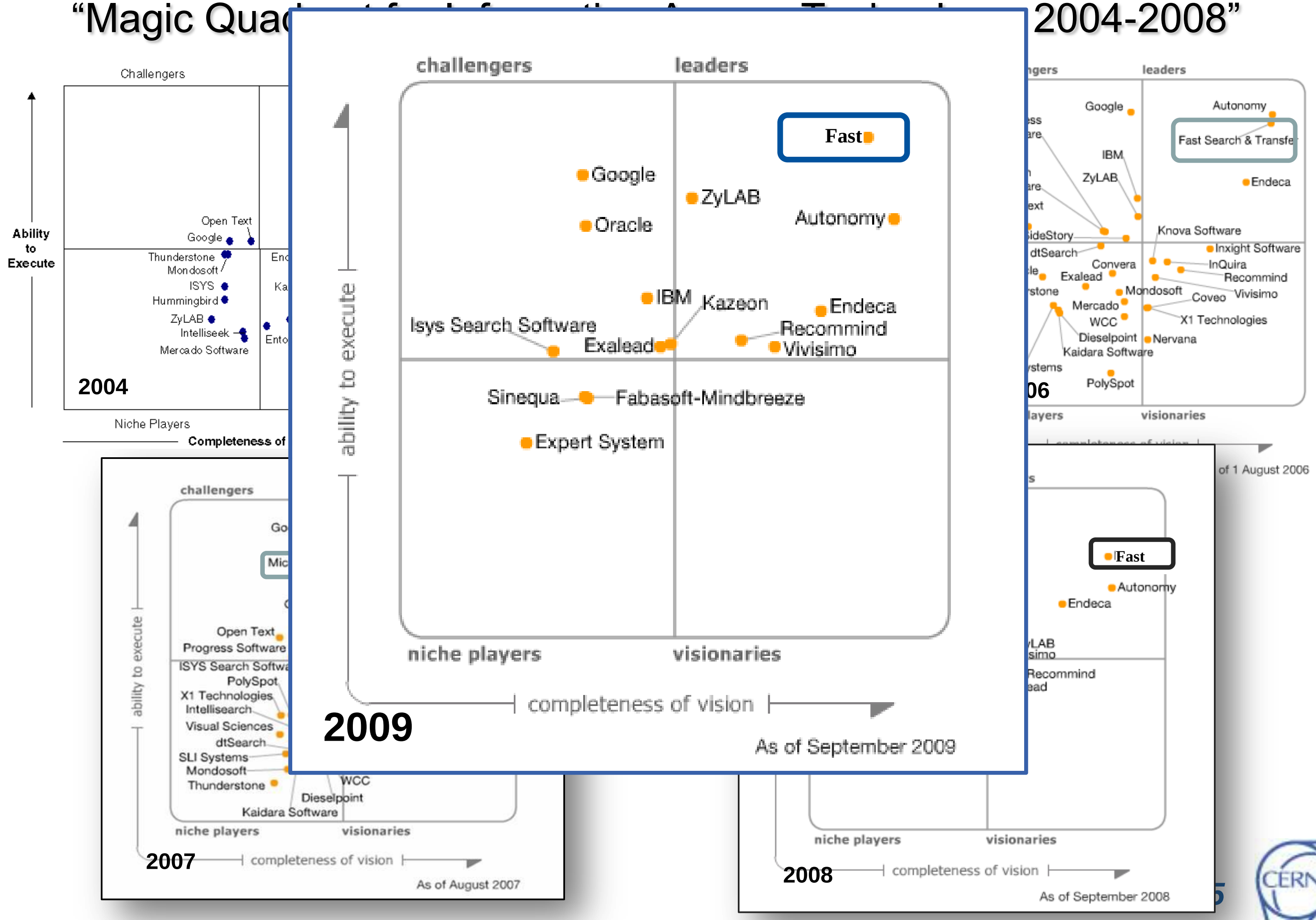
- Authentication models can be based on
  - Document ACL at time of indexing
  - Callback to the application

- Dependent on role based model for the site
  - Ideally only one role model

OIS

Enterprise Search Providers

- Gartner Report:
  “Magic Qua… 2004-2008”

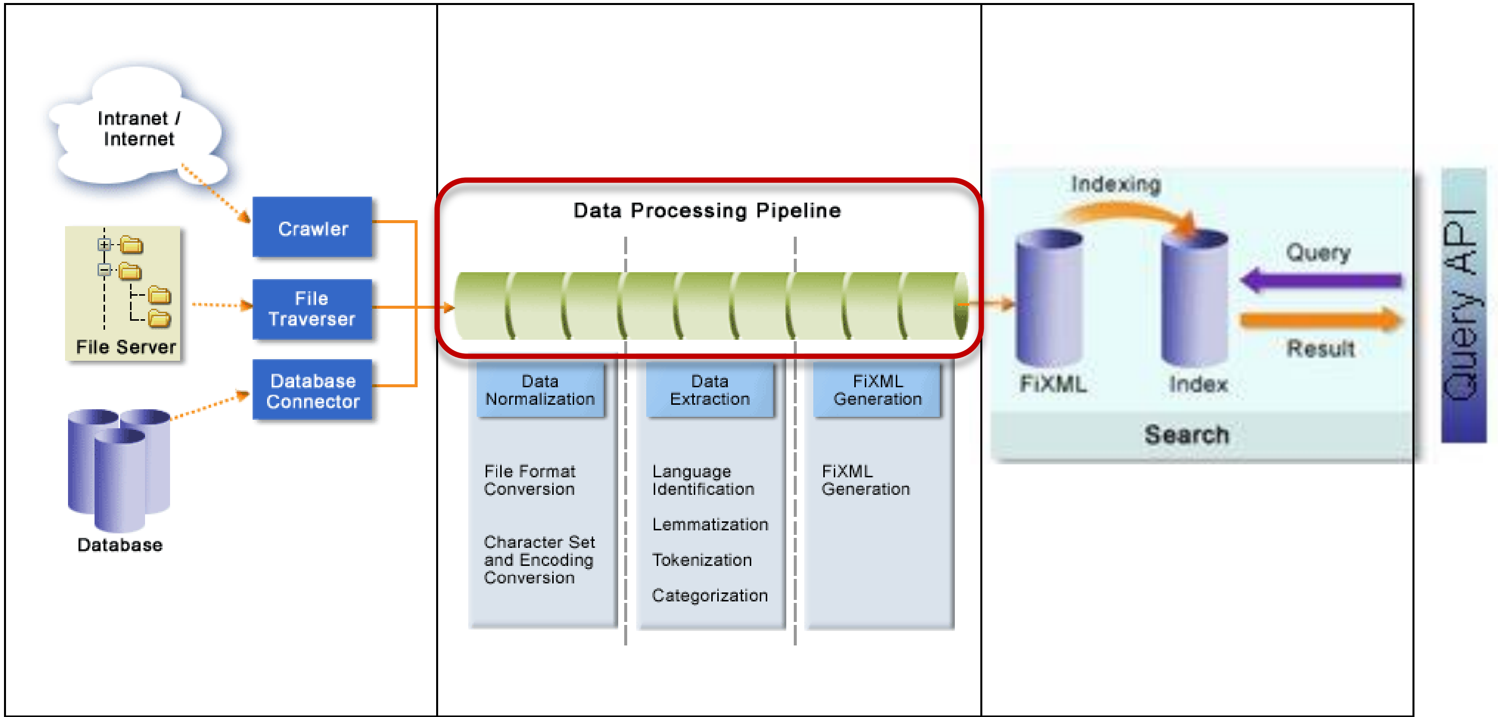


CERN Search

- A CERN Search page for the whole site
  - www.cern.ch search for public data
  - Central IT services
  - Experiment web sites
  - Infrastructure / HR / Administrative workflow sites
- Start of project in February 2006
  - Based on FAST as one of market leaders
  - Present resources 1 Project Associate and small share of an engineer
- In production since 2007

FAST ESP Architecture

Document Content Flow

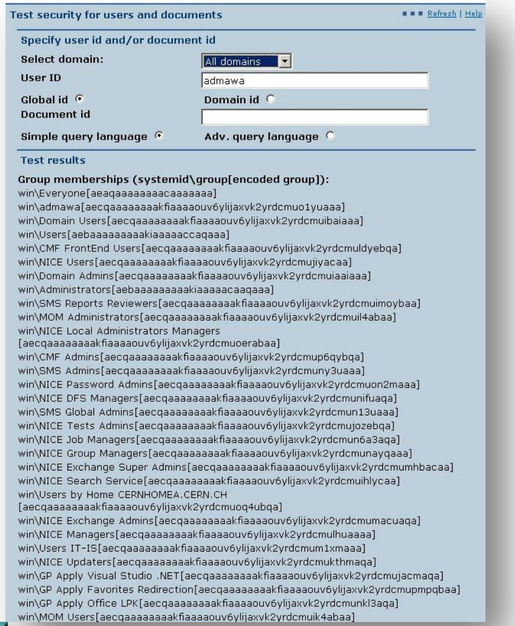


Document retrieval	Document processing	Document indexing

Indexing Protected Content

- Document Processing
  - Resolve ACLs to text strings
  - Sent to Indexer with document

- Security Access Module of FAST
  - Active Directory integration based on CERN accounts and e-groups



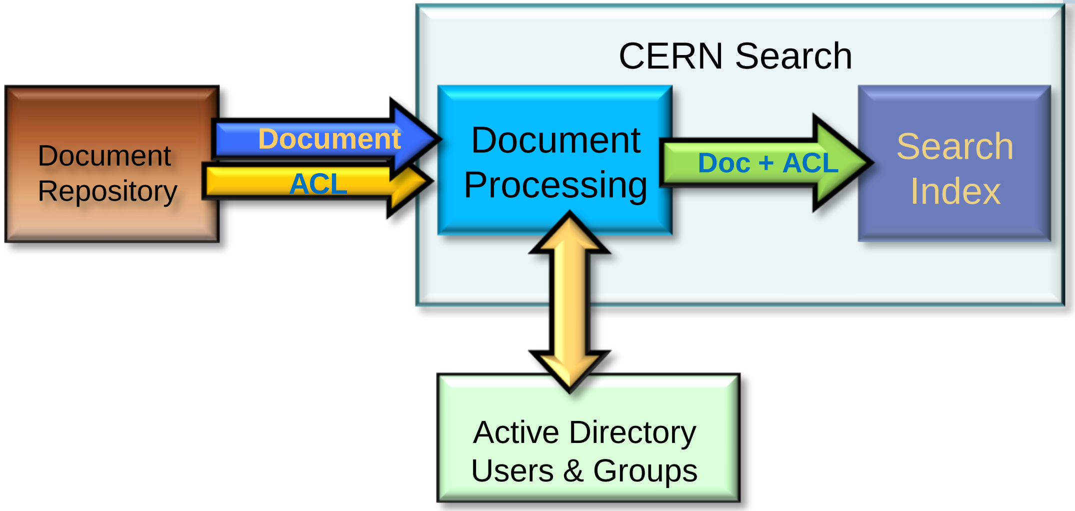
Test security for users and documents

Specify user id and/or document id

Select domain: All domains
User ID admawa
Global id ● **Domain id** ○
Document id
Simple query language ● **Adv. query language** ○

Test results

```
Group memberships (systemid\group[encoded group]):
win\Everyone[aeaqaaaaaaaacaaaaaaa]
win\admawa[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuo1yuaaa]
win\Domain Users[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuibaiaaa]
win\Users[aebaaaaaaaaakiaaaaaccaqaaa]
win\CMF FrontEnd Users[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuldyebqa]
win\NICE Users[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmujiyacaa]
win\Domain Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuiaaiaaa]
win\Administrators[aebaaaaaaaaakiaaaaacaaqaaa]
win\SMS Reports Reviewers[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuimoybaa]
win\MOM Administrators[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuil4abaa]
win\NICE Local Administrators Managers
[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuoerabaa]
win\CMF Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmup6qybqa]
win\SMS Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuny3uaaa]
win\NICE Password Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuon2maaa]
win\NICE DFS Managers[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmunifuaqa]
win\SMS Global Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmun13uaaa]
win\NICE Tests Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmujozebqa]
win\NICE Job Managers[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmun6a3aqa]
win\NICE Group Managers[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmunayqaaa]
win\NICE Exchange Super Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmumhbacaa]
win\NICE Search Service[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuihlycaa]
win\Users by Home CERNHOMEA.CERN.CH
[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuoq4ubqa]
win\NICE Exchange Admins[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmumacuaqa]
win\NICE Managers[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmulhuaaaa]
win\Users IT-IS[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmum1xmaaa]
win\NICE Updaters[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmukthmaqa]
win\GP Apply Visual Studio .NET[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmujacmaqa]
win\GP Apply Favorites Redirection[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmupmpqbaa]
win\GP Apply Office LPK[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmunkl3aqa]
win\MOM Users[aecqaaaaaaaakfiaaaaouv6ylijaxvk2yrdcmuik4abaa]
```



CERN Search

Document Repository

Document

ACL

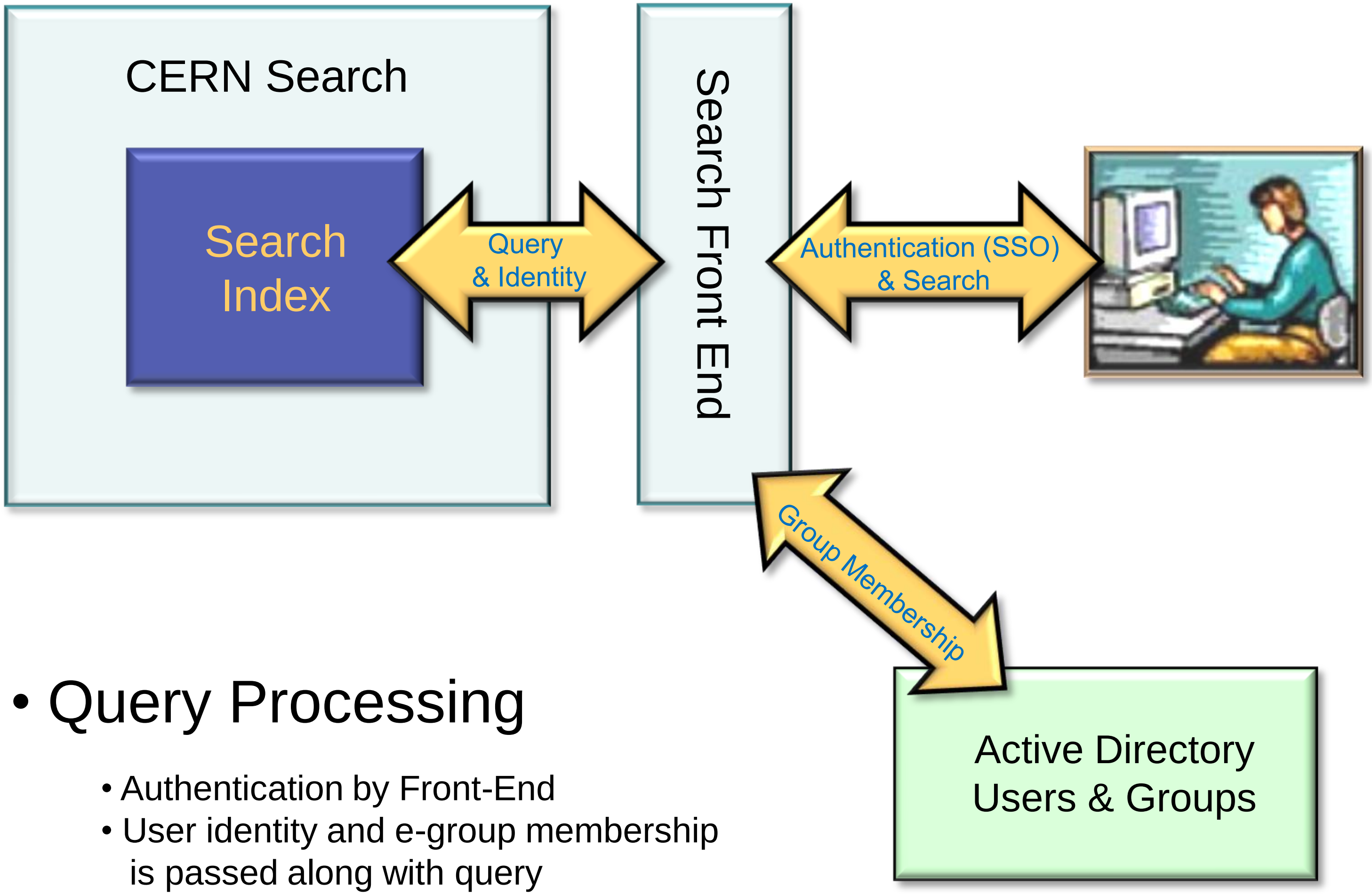
Document Processing

Doc + ACL

Search Index

Active Directory Users & Groups

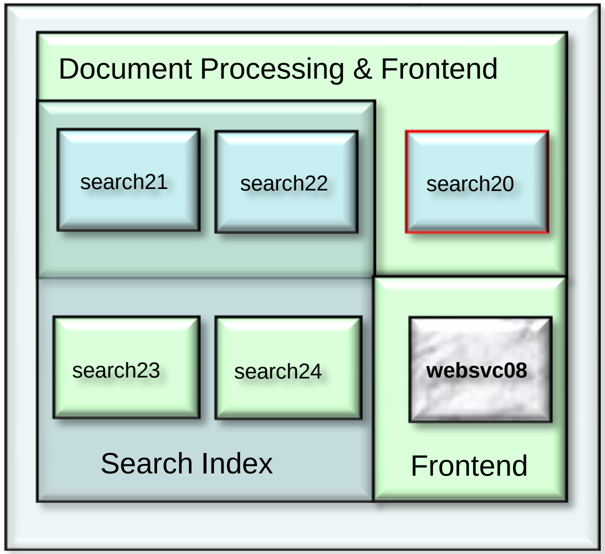
Authentication / Authorisation



- Query Processing
  - Authentication by Front-End
  - User identity and e-group membership is passed along with query

CERN Search - System Layout

Production System



Document Processing & Frontend

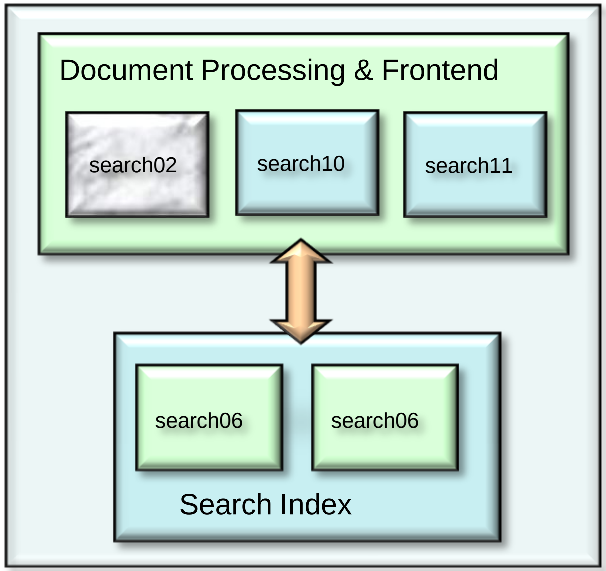
search21
search22
search20
search23
search24
websvc08

Search Index

Frontend

Search01	**- Index &Search - Document processing**
Search02	**- Index &Search - Document processing**
Search03	**- Admin node - Crawler / Webalyzer - Database connector**
Search04	**- Index - Document processing**
Search05	**- Index - Document processing**

Development System



Document Processing & Frontend

search02
search10
search11
search06
search06

Search Index

Search10	**- admin node - database connector - document processing**
Search11	**- Crawler / Webalyzer - document processing**
Search06	**- indexer - search engine**
Search02	**- dev Search frontends (EDMS, CFU, etc.)**

Indexed Documents

Documents indexed by CERN Search

	2010	2009	2008
CERN Websites	1537483	1787805	829542
CDS	1078094	1040694	936018
TWiki Pages	61277	---	---
Indico (Public)	328538	255365	432339
Joint Accelerator Conferences	157566	---	---
Phonebook	31198	25629	23982

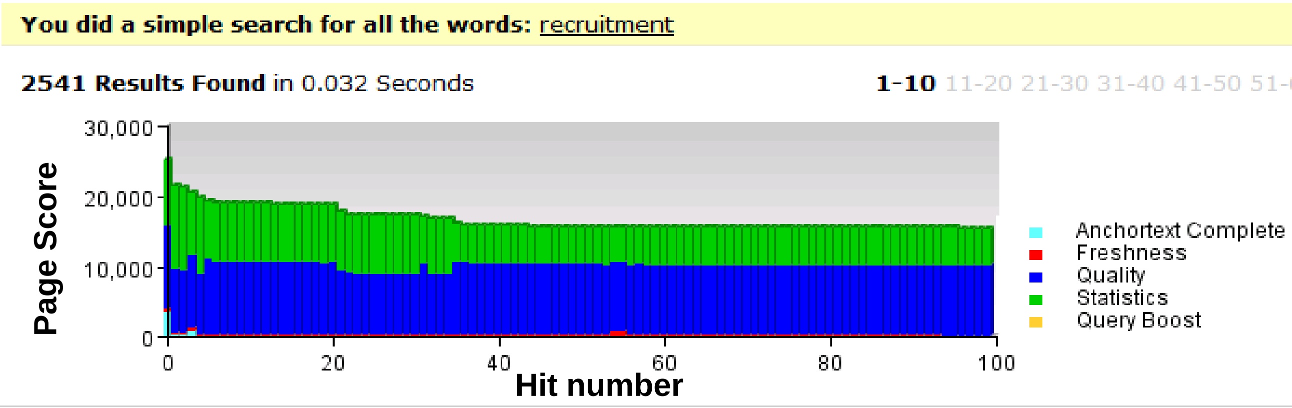
- **Currently >3 million documents**
- **Estimated 10 million in total if all sites indexed**

Result Ranking – Relevancy

- Order search with most interesting document first in list
- Ranking Metrics:
  - Search Terms:
    - Occurrence in URL, page title and page contents.
    - Proximity of terms in document
  - Quality of a page:
    - Relevance of page in the Web space of all indexed pages (how many other pages link to the page)
    - How deep inside a Website a page is located
  - Freshness of document
    - Generally the newer the document, the more interesting
  - Anchortext
    - Text of a link pointing to a page

Ranking Issues at CERN

- Flat Web space
  - ~10,000 Web sites just one level down
    http://www.cern.ch/site1
    http://www.cern.ch/site2
  - No consistent structure and navigation (apart from back-links to CERN home page)
- Keyword distribution
  - Small number of significant words in large number of pages



You did a simple search for all the words: recruitment

2541 Results Found in 0.032 Seconds

1-10 11-20 21-30 31-40 41-50 51-

Result Ranking – Improvements

- How to improve ranking?
  - Manual Tuning of results
    - to assure expected results during important events
      - LHC first physics; Angels & demons
  - Usage analysis
    - e.g. review of "zero result" queries
    - user tracking – "what links users follow"
- Best results obtained with hints to search engine and effort by content authors
  - Add keyword and author meta data tags at minimum

TWiki Search

- Request from experiments to index protected TWiki content and to improve ranking
  - Built in TWiki search functionality was weak
- Pages are protected so access requires CERN SSO step
  - Not natural for web crawlers
- URLs are not words so break of topic name improved ranking
  - ‘Example Topic Template’ from https://twiki/TWiki/ExampleTopicTemplate
- Get changed pages only
  - Twiki ‘find’ for modified documents to be re-indexed
  - Could increase frequency to hourly
- In production since June 3rd 2010
  - Users reporting substantial improvements compared to built in TWiki search

What about Google?

- What makes Google Web search work well
  - The whole web for analysis
    - who links to your site
  - Huge usage data used for “voting” for results
    - most popular results swim up
  - Substantial resources to tune and correct results
    - usage data analysis
    - taking into account popular events
    - hand edited results for popular single key word searches

- Above is valid for all public search engines
  - Yahoo!, Bing, …

Google Search Appliance

- Google make a packaged offering
  - Hardware
  - Software
  - 2 year license and then need to replace
- Priced by number of documents
  - CERN has around 10 million documents
- Black box solution
  - Management GUI
  - Alerting
  - Does retrieval, analysis and indexing
  - Single-sign on support (but see later…)

ATLAS Comparison

- Test
  - BNL have a Google Search Appliance which they use to index ATLAS public pages at CERN
  - Performed sampling comparisons with CERN FAST Search for sample common terms
- Results
  - Google Search Appliance did better job at ranking according to content owners
  - Indexing of protected pages did not work
    - Issues with Single Single On javascript
    - Google engineers could not find a solution
  - GSA cost would have been substantially higher

Looking Ahead

- Include additional protected content
  - e.g. Indico, EDMS, Sharepoint, Drupal, …
- Migrate to FAST Search 2010
  - Improved web selection filtering
    - – Show documents from past X months
    - – Show documents written by author Y
  - Partition web space
    - – Official content
    - – Personal sites
  - Feedback based on previous user choices
    - – Put higher if often selected
  - Allow content managers to adjust rankings themselves
- – Repeat comparisons with other solutions in 2011 such as GSA
  - Interested to see what other sites are doing

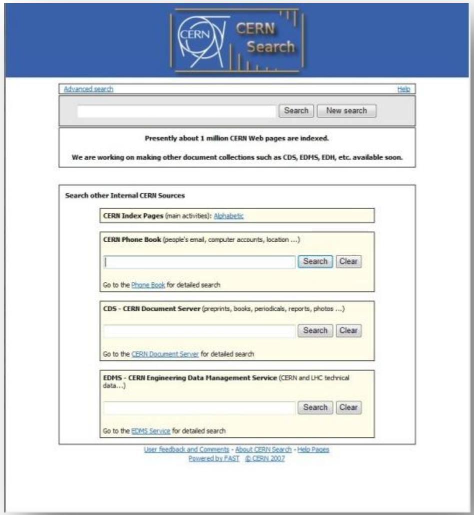
Questions ?

OIS

CERN Search

- CERN Search:
  http://cern.ch/search

- and also via:
  - CERN Intranet & Public Pages
  - TWiki
  - IT, HR,
    PH Websites
  - JACOW



CERN Search

Advanced search | Help

Search | New search

Presently about 1 million CERN Web pages are indexed.

We are working on making other document collections such as CDS, EDMS, EDH, etc. available soon.

Search other Internal CERN Sources

CERN Index Pages (main activities): Alphabetic

CERN Phone Book (people's email, computer accounts, location ...)

Search | Clear

Go to the Phone Book for detailed search

CDS - CERN Document Server (preprints, books, periodicals, reports, photos ...)

Search | Clear

Go to the CERN Document Server for detailed search

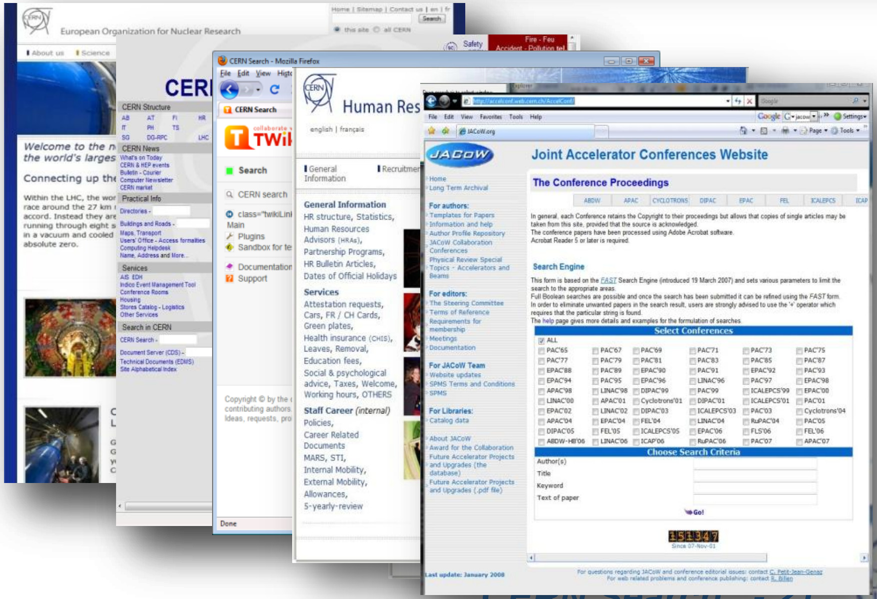
EDMS - CERN Engineering Data Management Service (CERN and LHC technical data...)

Search | Clear

Go to the EDMS Service for detailed search

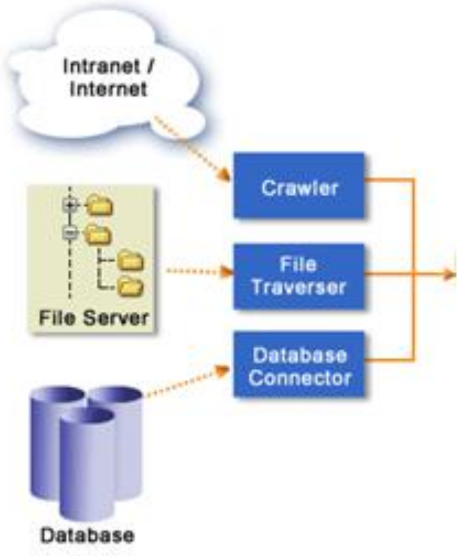
User feedback and Comments - About CERN Search - Help Pages

Powered by FAST © CERN 2007


European Organization for Nuclear Research
CERN Search
TWiki
Human Res
JACoW
Joint Accelerator Conferences Website
The Conference Proceedings

Enterprise Search

- Wide range of document sources:



  - Web Pages
  - File systems
  - Databases
  - Directories (People and Places)
  - Document repositories (CDS, EDMS, Indico, …)
- Variety of meta data
- Different access protection schemes
- Different retrieval methods and frequencies