

CERN Spring Campus 2014



# Information Retrieval and Search Engines: Introduction to Solr

Patrick GLAUNER

CERN

[patrick.oliver.glauner@cern.ch](mailto:patrick.oliver.glauner@cern.ch)

April 14, 2014

# About me

- Graduated from Karlsruhe University of Applied Sciences, Germany in 2012
- Software Engineer at CERN since 2011
- R&D focus on document management and search engines

```
...an_done) {  
    if(plant[0].bar_id == 1)  
        temp_mode = mode;  
        mode = watch_bars;  
        file_output("newlac", c  
an_done = on;  
}
```

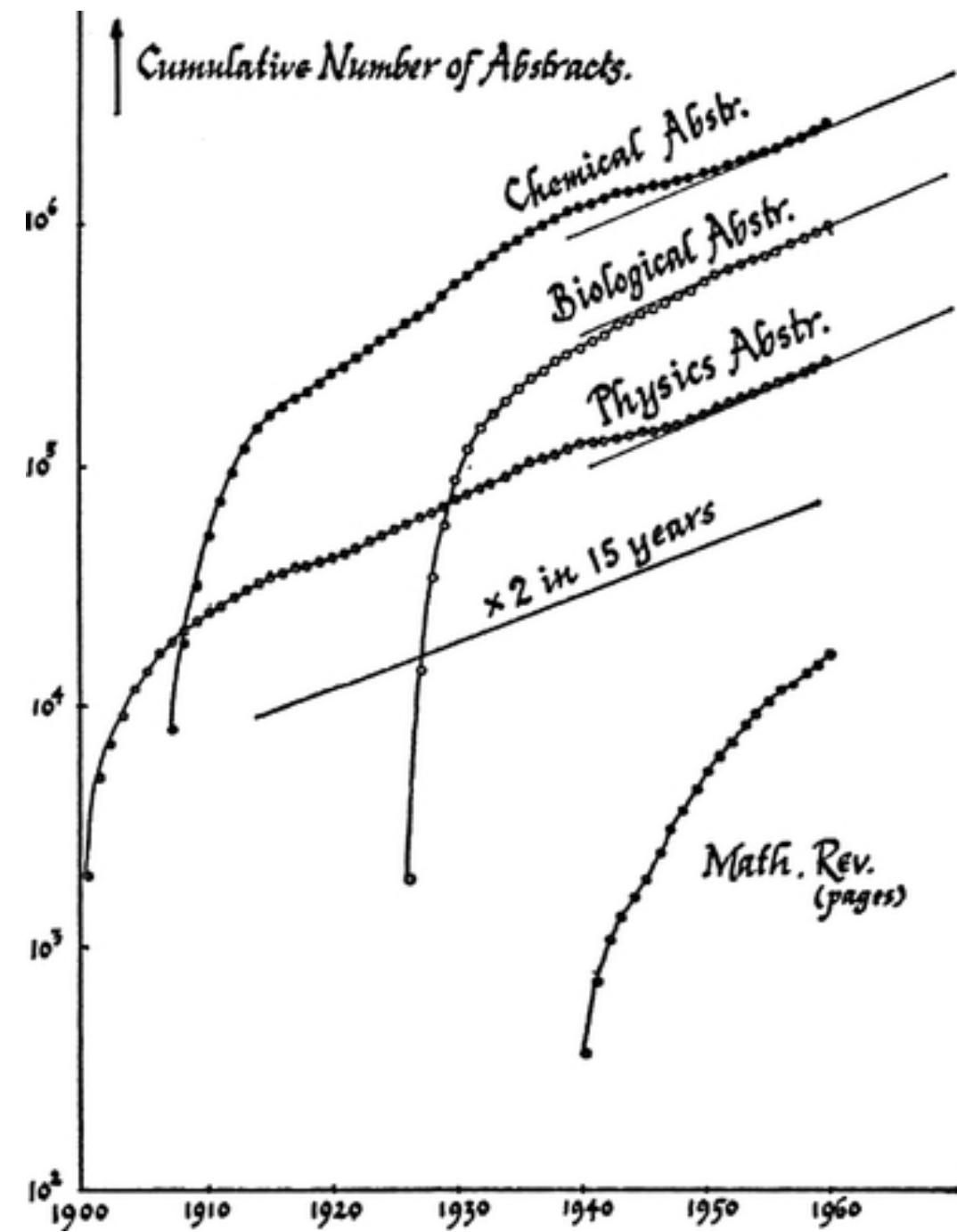
# Agenda

- The Importance of Search
- Information Retrieval
- Introduction to Solr
- Solr-based Search and Ranking in Invenio
- Conclusions
- Q&A



# The Importance of Search

- Why is search so important?
- Amount of data increases exponentially
- Different kinds of search engines: WWW, desktop, repository, etc.
- Effectivity: finding correct results
- Efficiency: performance, ranking, usability, etc.



*P. Glauner: Information Retrieval and Search Engines: Introduction to Solr*



# The Importance of Search: History

The screenshot shows the Yahoo! homepage with the following elements:

- Navigation icons: My, Finance, Travel, **YAHOO!**, Mail, Messenger, HotJobs, and a Help link.
- Travel promotion: [Yahoo! Travel](#) - Flights, Hotels, Cars, Vacations, Last-Minute Getaways, Cruises, Summer Travel Sale.
- Search bar: "Search the Web:" with a "Yahoo! Search" button and links to "Advanced" and "Preferences".
- Left sidebar menu:
  - New!** [Yahoo! Mail](#) - now with 100MB of free storage. [Sign up.](#)
  - Shop** [Auctions](#), [Autos](#), [Classifieds](#), [Real Estate](#), [Shopping](#), [Travel](#)
  - Find** [HotJobs](#), [Maps](#), [People Search](#), [Personals](#), [Yellow Pages](#)
  - Connect** [Chat](#), [GeoCities](#), [Greetings](#), [Groups](#), [Mail](#), [Messenger](#), [Mobile](#)
  - Organize** [Addresses](#), [Briefcase](#), [Calendar](#), [My Yahoo!](#), [PayDirect](#), [Photos](#)
  - Fun** [Games](#), [Horoscopes](#), [Kids](#), [Movies](#), [Music](#), [Radio](#), [TV](#)
  - Info** [Finance](#), [Health](#), [News](#), [Sports](#), [Weather](#), [More Yahoo!...](#)
- Bottom left: [Make Yahoo! your home page](#) - [Yahoo! Toolbar with Pop-Up Blocker](#)
- Personal Assistant section:
  - Personal Assistant** [Sign In](#)
  - Yahoo! Mail** - now with 25 times more free storage. [Sign up now!](#)
  - Advertisement: "Save \$50" - List your site in search results on Yahoo! [Find out how](#) - [Take a tour](#)
- In The News** 11:06am, Wed Jun 23
  - [Bush claimed right to waive torture laws](#)
  - [Fugitive vows to assassinate Iraqi leader](#)
  - [Iran orders release of British sailors](#)
  - [U.S. forces lead airstrike after beheading](#)
  - [Prisons seek help to cope with mentally ill](#)
  - [Digital bugle plays taps when GIs can't](#)
  - [Wimbledon](#) - [Baseball](#) - [Euro 2004](#) - [NBA](#)
- Marketplace**
  - [Free shipping from Dell Home](#)
- Bottom: [Yahoo! Business Services](#) | [Yahoo! Premium Services](#)



# The Importance of Search: History

The screenshot shows the Yahoo! homepage with the following elements:

- Navigation icons: My, Finance, Travel, YAHOO!, Mail, Messenger, HotJobs, and a Help button.
- Travel promotion: Yahoo! Travel - Flights, Hotels, Cars, Vacations, Last-Minute Getaways, Cruises, Summer Travel Sale.
- Search bar: Search the Web: [input field] with buttons for Yahoo! Search, Advanced, and Preferences.
- Left sidebar: New! Yahoo! Mail (100MB free storage), Shop (Auctions, Autos, Classifieds, Real Estate, Shopping, Travel), Find (HotJobs, Maps, People Search, Personals, Yellow Pages), Connect (Chat, GeoCities, Greetings, Groups, Mail, Messenger, Mobile), Organize (Addresses, Briefcase, Calendar, My Yahoo!, PayDirect, Photos), Fun (Games, Horoscopes, Kids, Movies, Music, Radio, TV), Info (Finance, Health, News, Sports, Weather, More Yahoo!...).
- Personal Assistant: Sign In, Yahoo! Mail (25 times more free storage), Save \$50 (List your site in search results on Yahoo!).
- In The News: 11:06am, Wed Jun 23. News items include: Bush claimed right to waive torture laws, Fugitive vows to assassinate Iraqi leader, Iran orders release of British sailors, U.S. forces lead airstrike after beheading, Prisons seek help to cope with mentally ill, Digital bugle plays taps when GIs can't, Wimbledon, Baseball, Euro 2004, NBA.
- Marketplace: Free shipping from Dell Home.
- Bottom: Yahoo! Business Services, Yahoo! Premium Services.

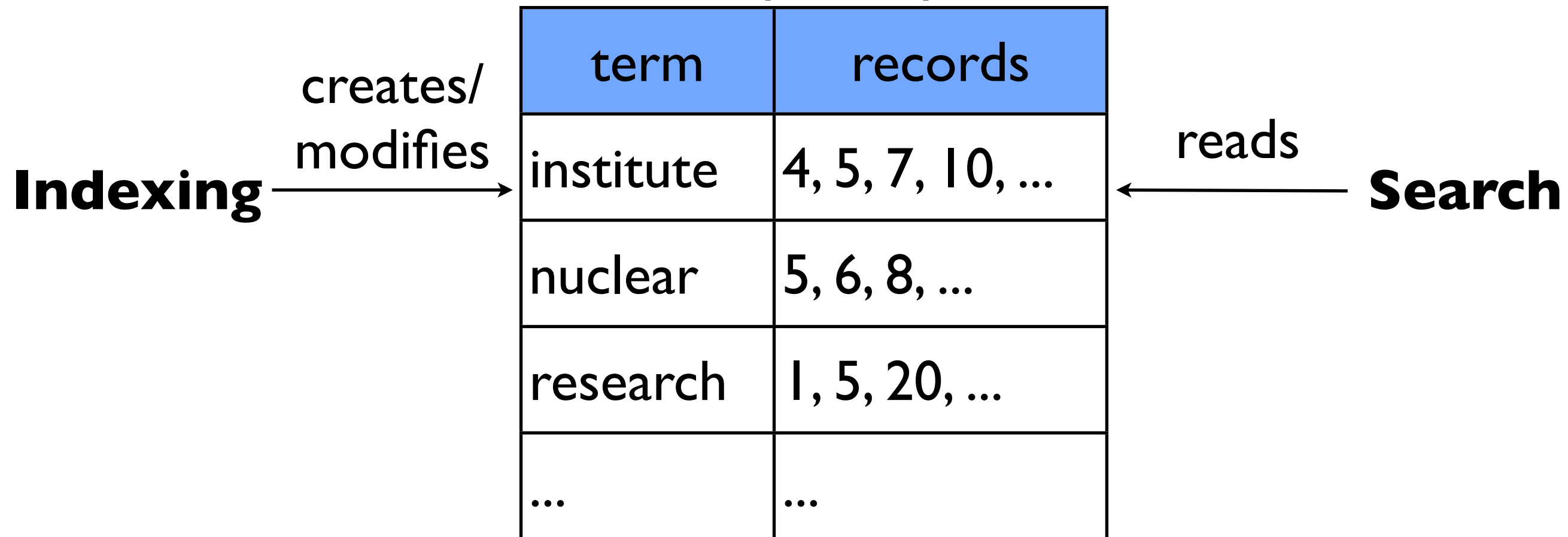
The screenshot shows the Google homepage with the following elements:

- Google logo.
- Navigation links: Web, Images, Video, News, Maps, more ».
- Search bar: [input field] with buttons for Google Search and I'm Feeling Lucky.
- Right sidebar: iGoogle | Sign in, Advanced Search, Preferences, Language Tools.
- Bottom: Advertising Programs - Business Solutions - About Google - Go to Google Deutschland, ©2007 Google.



# The Importance of Search: Fundamental Capabilities

- Harvesting of data: submissions, crawling, etc.
- Construction of indexes to speed up search



# The Importance of Search: Fundamental Capabilities

- **Case insensitivity:** `computer`  $\equiv$  `Computer`
- **Stemming:** `computer`  $\equiv$  `computers`
- **Stop words:** removal of terms such as `at`, `of`, `at` among others
- **Synonyms:** `because`  $\approx$  `as`
- ...



# The Importance of Search: Advanced Capabilities



# The Importance of Search: Advanced Capabilities

The image shows a screenshot of the Google search interface. On the left, the Google logo is displayed with 'Canada' underneath. Below the logo is a search bar containing the letter 'a'. A dropdown menu lists suggestions: 'air canada', 'amazon', 'access d', 'apple', 'aldo', 'amt', 'amazon.ca', 'aeroplan', 'air transat', and 'addicting games'. Below the suggestions are two buttons: 'Google Search' and 'I'm Feeling Lucky'. To the right of the search bar, there are links for 'Advanced Search' and 'Language Tools'. The main search results area shows the Google logo, a search bar with the text 'Search', and 'SafeSearch moderate'. Below the search bar, it says 'About 28 results (0.10 seconds)' and 'Advanced search'. The results are categorized under 'Everything', 'Images', and 'More'. The 'Images' category is selected, showing a grid of six similar images of the Colosseum. Each image has its dimensions displayed below it: 848 x 565, 700 x 467, 700 x 467, 640 x 480, 679 x 451, and 600 x 451. There are also links for 'Similar images', 'More sizes', 'All images', 'Similar to...', 'Visually similar', 'More sizes', and 'Reset tools'.

# Information Retrieval

- “The tracing and recovery of specific information from stored data.”
- Relevance ranking based on term frequency (word similarity ranking):

Query	nuclear research	Relevance
Result 1	I like nuclear research, computer science and mathematics	Low
Result 2	Employed in nuclear research	High

- Querying for near words
- Handling of misspellings
- ....



# Introduction to Solr

- Wide-spread open source search server
- Written in Java
- Communication through HTTP
- Configuration in XML files
- Flexible data schema, NoSQL-like
- Rich indexing, search and IR capabilities
- <http://lucene.apache.org/solr/> (Solr 4.7.1)



# Introduction to Solr: Setup

- Prerequisites: a Linux/Mac OS X environment containing Java 6, Python 2.7 and solrpy:

```
$ easy_install solrpy
```

## 1. Download Solr (~150MB):

```
$ curl -O http://mirror.switch.ch/mirror/apache/dist/lucene/solr/4.7.1/solr-4.7.1.zip
```

## 2. Unzip Solr:

```
$ unzip solr-4.7.1.zip
```

## 3. Start Solr:

```
$ cd solr-4.7.1/example
```

```
$ java -jar start.jar
```



# Introduction to Solr: Indexing

## 4. Index some documents:

```
import solr

# create a connection to the Solr server
s = solr.SolrConnection('http://localhost:8983/solr')

# add documents to the index
s.add(id=1, title='The best solr book ever written', author='Author1')
s.add(id=2, title='The fastest computer', author='Author2')
s.add(id=3, title='Introduction to Solr', author='P. Glauner')

s.commit()
```



# Introduction to Solr: Search and Ranking

## 5. Search documents:

```
import solr

# create a connection to the Solr server
s = solr.SolrConnection('http://localhost:8983/solr')

# query the index
response = s.query('title:Solr')

for hit in response.results:
    print '%s: %s' % (hit[u'score'], hit[u'title'][0])
```

## 6. Output:

```
0.5: Introduction to Solr
0.375: The best solr book ever written
```



# Introduction to Solr: Configuration

- `schema.xml` defines the document structure and properties:

```
<field name="title" type="text_general" indexed="true" stored="true" multiValued="true"/>  
<field name="subject" type="text_general" indexed="true" stored="true"/>  
<field name="description" type="text_general" indexed="true" stored="true"/>  
<field name="comments" type="text_general" indexed="true" stored="true"/>  
<field name="author" type="text_general" indexed="true" stored="true"/>  
<field name="keywords" type="text_general" indexed="true" stored="true"/>
```

- `text_general` defines tokenizer, stemmer and synonym filter among others
- Changes require a Solr restart (and reindexing in most cases)



# Introduction to Solr: Further Topics

- **Delete all data in Solr:**

```
$ cd solr-4.7.1/example/exampledocs
```

```
$ java -Ddata=args -jar post.jar '<delete><query>*:*</query></delete>'
```

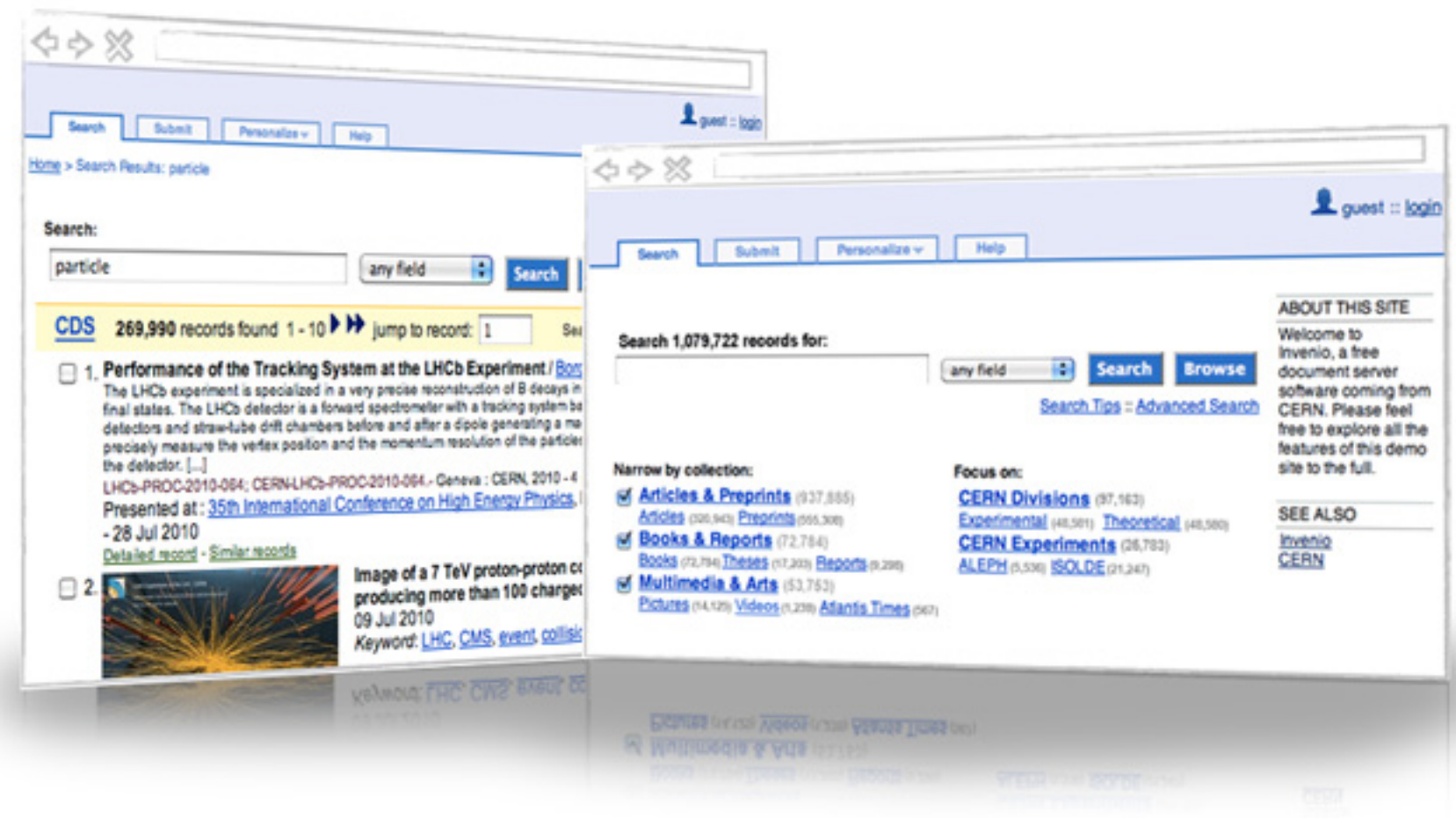
- solrpy tutorial: <https://code.google.com/p/solrpy/>

- Solr tutorial: <https://lucene.apache.org/solr/tutorial.html>



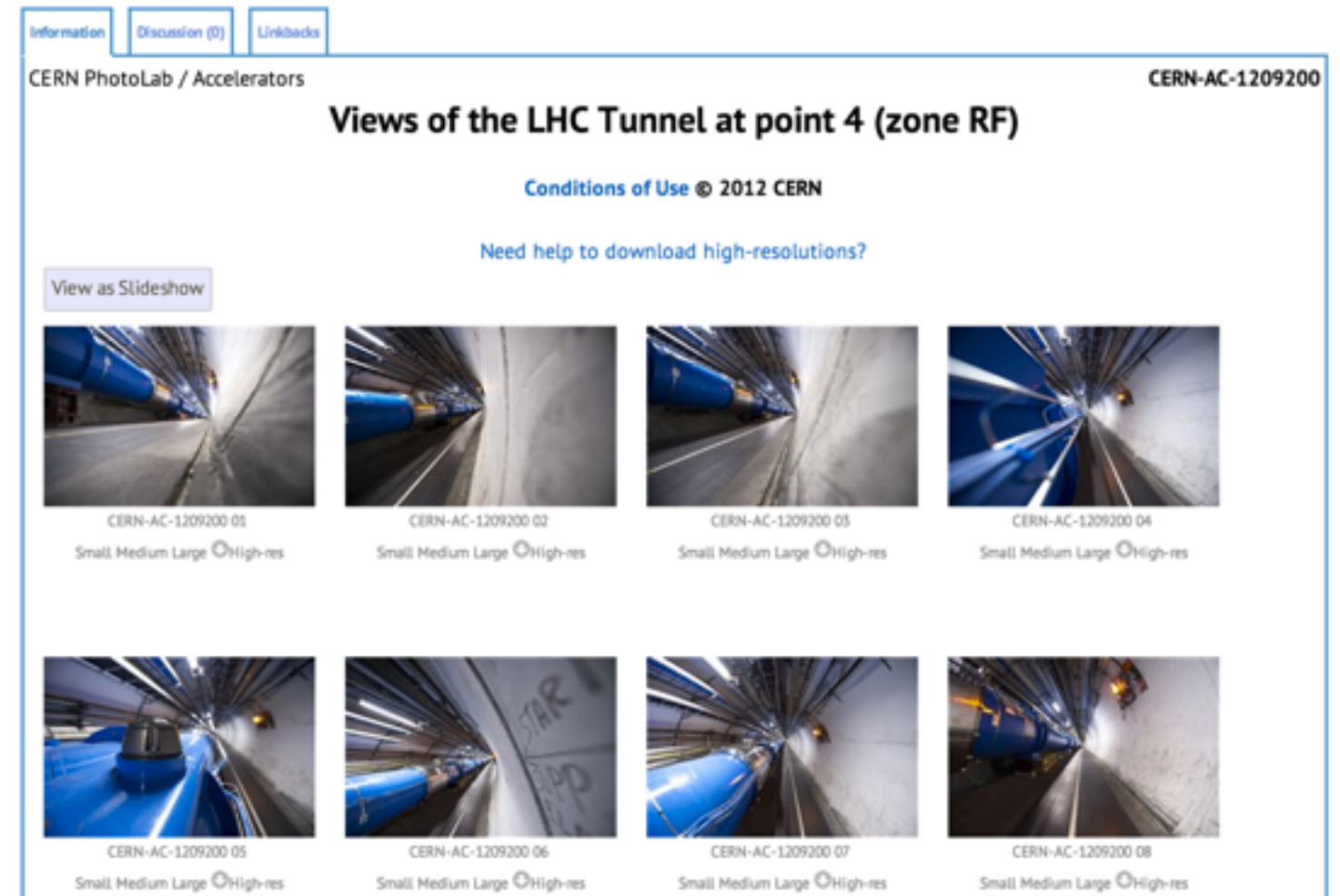
# Solr-based Search and Ranking in Invenio

- Invenio is a free document management system
- Originally developed at CERN, nowadays by an international collaboration
- More than 30 instances around the globe
- <http://invenio-software.org/>

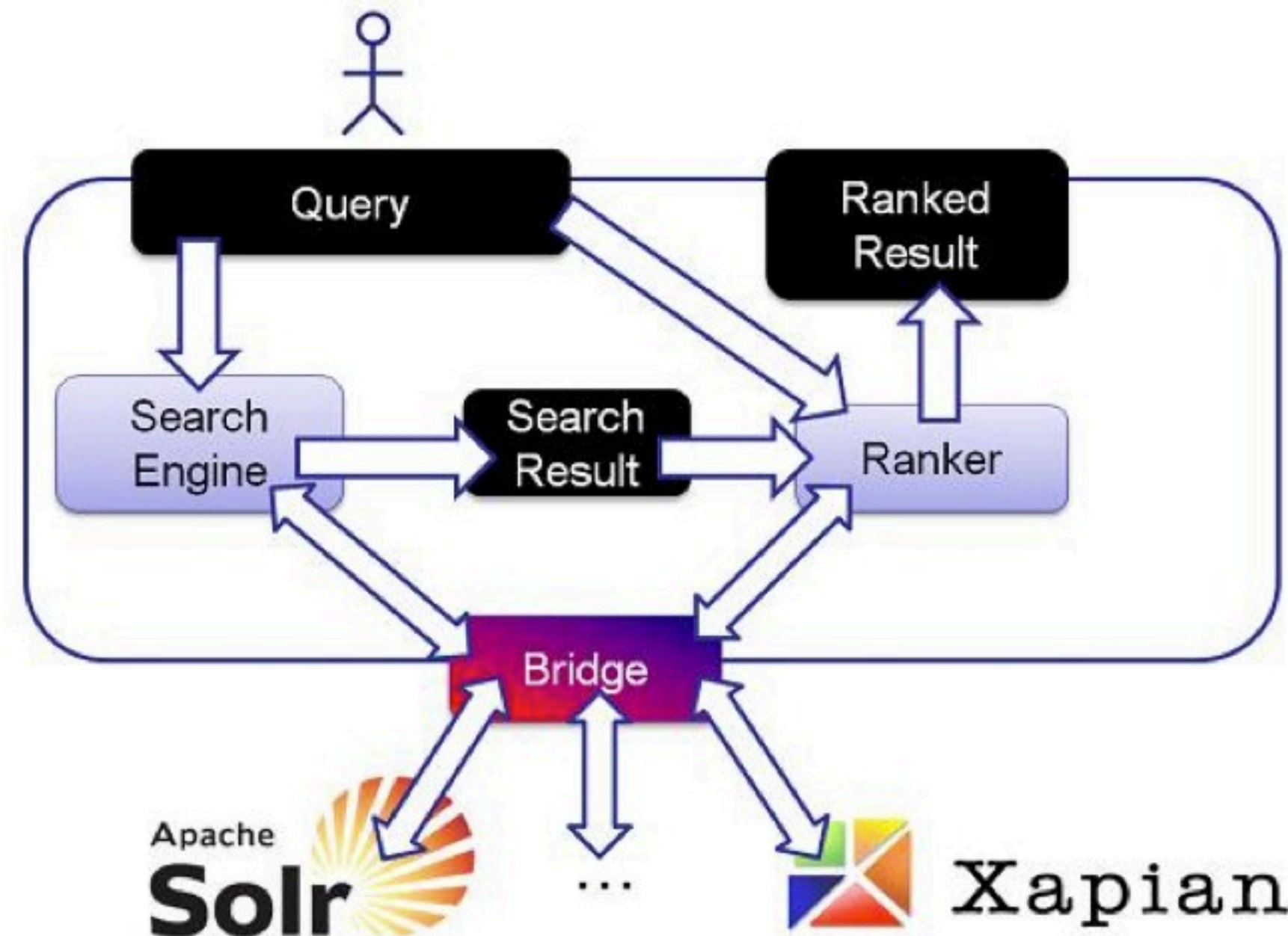


# Solr-based Search and Ranking in Invenio

- The CERN Document Server (CDS) is one of the largest Invenio instances
- More than 1.5M records and 400K full-text documents in the high energy physics domain
- Receives more than 10K unique visits per day with occasional peaks of up to 350K visits per day during important physics announcements
- <http://cds.cern.ch/>



# Solr-based Search and Ranking in Invenio



# Solr-based Search and Ranking in Invenio

Search:

Sort by:

**Results overview:** Found **51,090** records in 1.32

1. **Is it the Standard Model Higgs ? / Dührss**  
(100) Once an excess of events in one of the H Model was found. This talk shows what n  
ATL-SLIDE-2007-015; CERN-ATL-SLIDE-2  
Is it the **Standard Model** Higgs ? Michael Due  
Precision tests of the **Standard Model** ... M. D

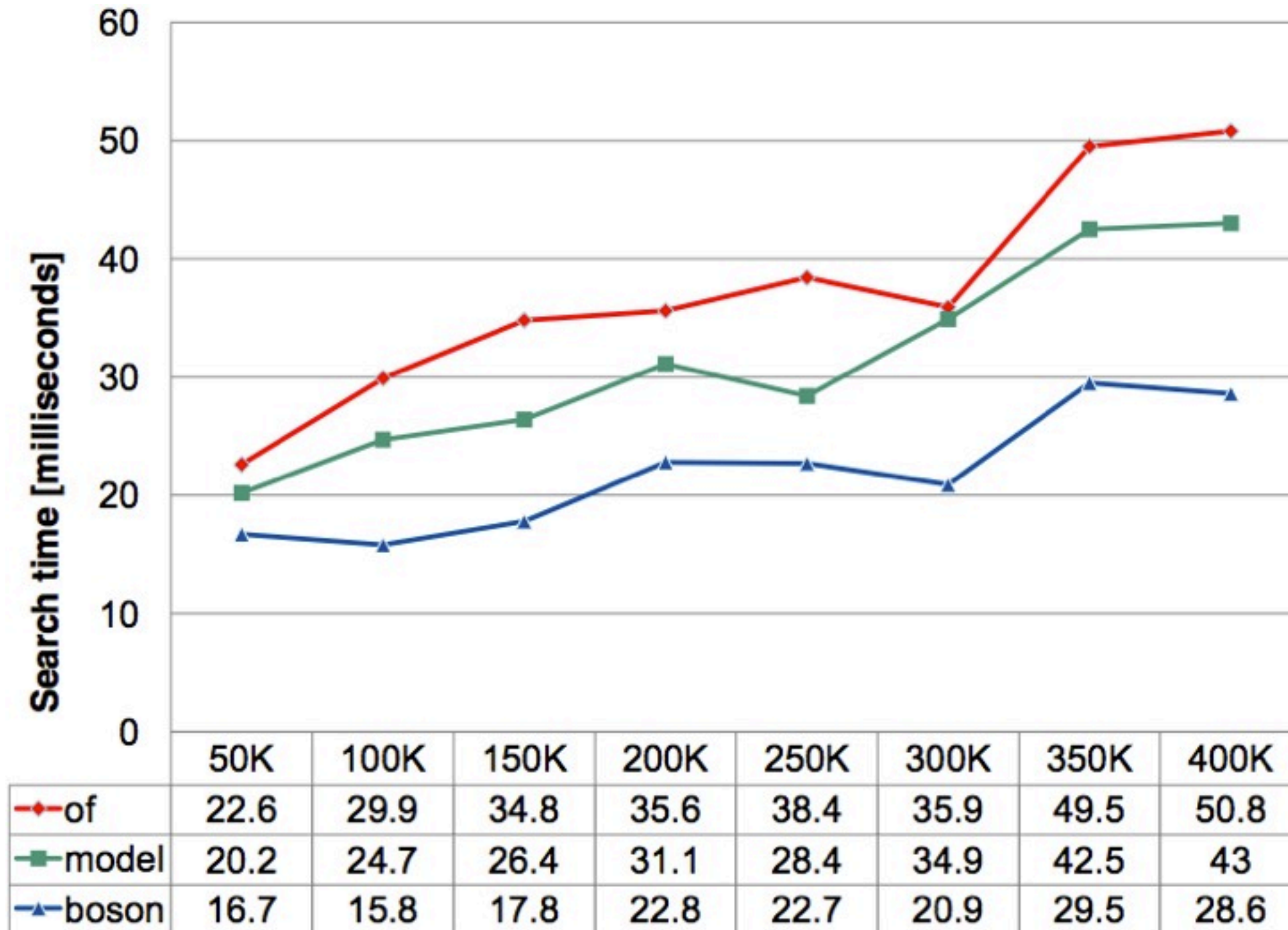
2. **Status and recent results from the LHC /**  
(93) Summary of the most recent physics resu

Figure 2: Solr-based word similarity result by percentage

- To benefit from Solr's efficient and scalable full-text search and word similarity ranking, it was integrated into Invenio
- Full-text search
- Word similarity ranking
- Snippet generation
- Finding records similar to a specific one
- Query correction



# Solr-based Search and Ranking in Invenio



*P. Glauner: Information Retrieval and Search Engines: Introduction to Solr*



# Conclusions

- Search engines build on top of indexes, algorithms and information retrieval theory
- Solr is a wide-spread open source search server
- It can be easily integrated into applications, e.g. Invenio
- Solr offers a rich set of features including search, ranking, snippets and finding records similar to a specific one
- It scales well for large corpuses





# References

- <http://lucene.apache.org/solr/>
- <http://invenio-software.org/>
- P. Glauner et al. Use of Solr and Xapian in the Invenio document repository software. Open Repositories Conference 2013. Charlottetown, Canada. arXiv:1310.0250.
- P. Glauner, T. Simko. Enhancing Invenio Digital Library With An External Relevance Ranking Engine. BSc thesis, Karlsruhe University of Applied Sciences, Germany. 2012. arXiv:1211.0689.

