Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

PKP Scholarly Publishing Conference 2013

The new OJS search features with Lucene/Solr

DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Overview

Why new search features at all?

Goals and Achievements

Decisions to be made

Installation, Configuration and Maintenance

New search features live

DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

OJS search and Solr

OJS search

- across article metadata, full texts and supplementary files
- simple text analysis for indexing and search (white space “tokenization”, conversion of words to lowercase and elimination of “stopwords”)
- doesn’t work for the languages that use logographic notational systems

Solr

- an open source search server using Lucene (indexing and search software library)
- provides more sophisticated tokenizers, token filters and language-specific analysis components “out of the box”
- supports many other advanced search features, e.g. auto-suggestion, alternative spelling proposal, paging, highlighting, ordering, faceting, “more like this” feature, improved ranking, etc.

→ Search optimization part of the project "Functional extensions and value added services for OJS", founded by the Deutsche Forschungsgemeinschaft (DFG)

DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Goals

Support multilingualism and all OJS languages

Ensure backward compatibility

Benefit from additional search features provided by Lucene/Solr

Consider different deployment scenarios (ranging from single journal installations to large OJS provider deployments):

- search across articles of a single journal
- search across multiple journals of a single OJS installation
- search across various OJS installations
- search across various applications

Design simple, flexible and user friendly

→ Concept: http://pkp.sfu.ca/wiki/index.php?title=OJSdeSearchConcept

→ Result: Lucene plug-in for OJS

DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Decisions to be made

OJS or Lucene/Solr search?

- Are journal languages supported by OJS search?
- Is the use of the additional search features desired?
- Is the support of additional document formats or faster indexing needed?
- Is there the required knowledge and the possibility to maintain another software (Solr)?
- Is PostScript required?

Deployment requirements and possibilities - Embedded or Remote Server Mode?

- one or more independent OJS installations on one server → embedded mode, fully preconfigured and supported by the plug-in and thus easier to use and maintain
- several servers, search across OJS installations or even across different applications → further configuration, implementation and understanding of Solr needed

Push or Pull indexing?

- Push: up-to-date index and simple configuration, works out-of-the-box
- Pull: server load can be balanced and it is more resilient against Solr server downtime

→ Use the simpler push configuration by default but check its performance and reliability

DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Installation, Configuration and Maintenance

Embedded scenario:

- install Java
- download the appropriate version of Jetty (servlet container Solr runs with) and Solr binaries, extract them, create symbolic links
- check the installation
- secure server
- start Solr server
- enable and setup the preconfigured plug-in
- rebuild indexes

Remote scenario:

- can be installed and deployed in many different ways
- there is no one best deployment

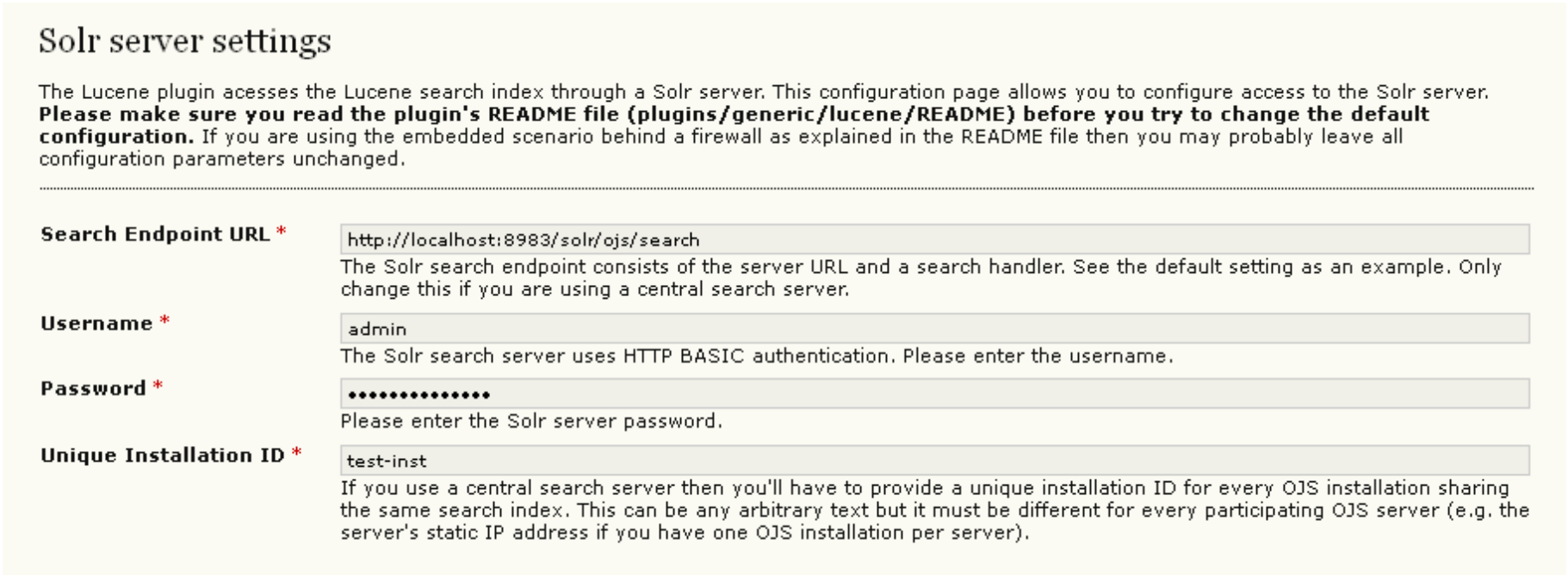
→ See README:
https://github.com/pkp/ojs/blob/master/plugins/generic/lucene/README

DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Solr server settings



Solr server settings

The Lucene plugin acesses the Lucene search index through a Solr server. This configuration page allows you to configure access to the Solr server. **Please make sure you read the plugin's README file (plugins/generic/lucene/README) before you try to change the default configuration.** If you are using the embedded scenario behind a firewall as explained in the README file then you may probably leave all configuration parameters unchanged.

Search Endpoint URL *
http://localhost:8983/solr/ojs/search
The Solr search endpoint consists of the server URL and a search handler. See the default setting as an example. Only change this if you are using a central search server.

Username *
admin
The Solr search server uses HTTP BASIC authentication. Please enter the username.

Password *
••••••••••••••
Please enter the Solr server password.

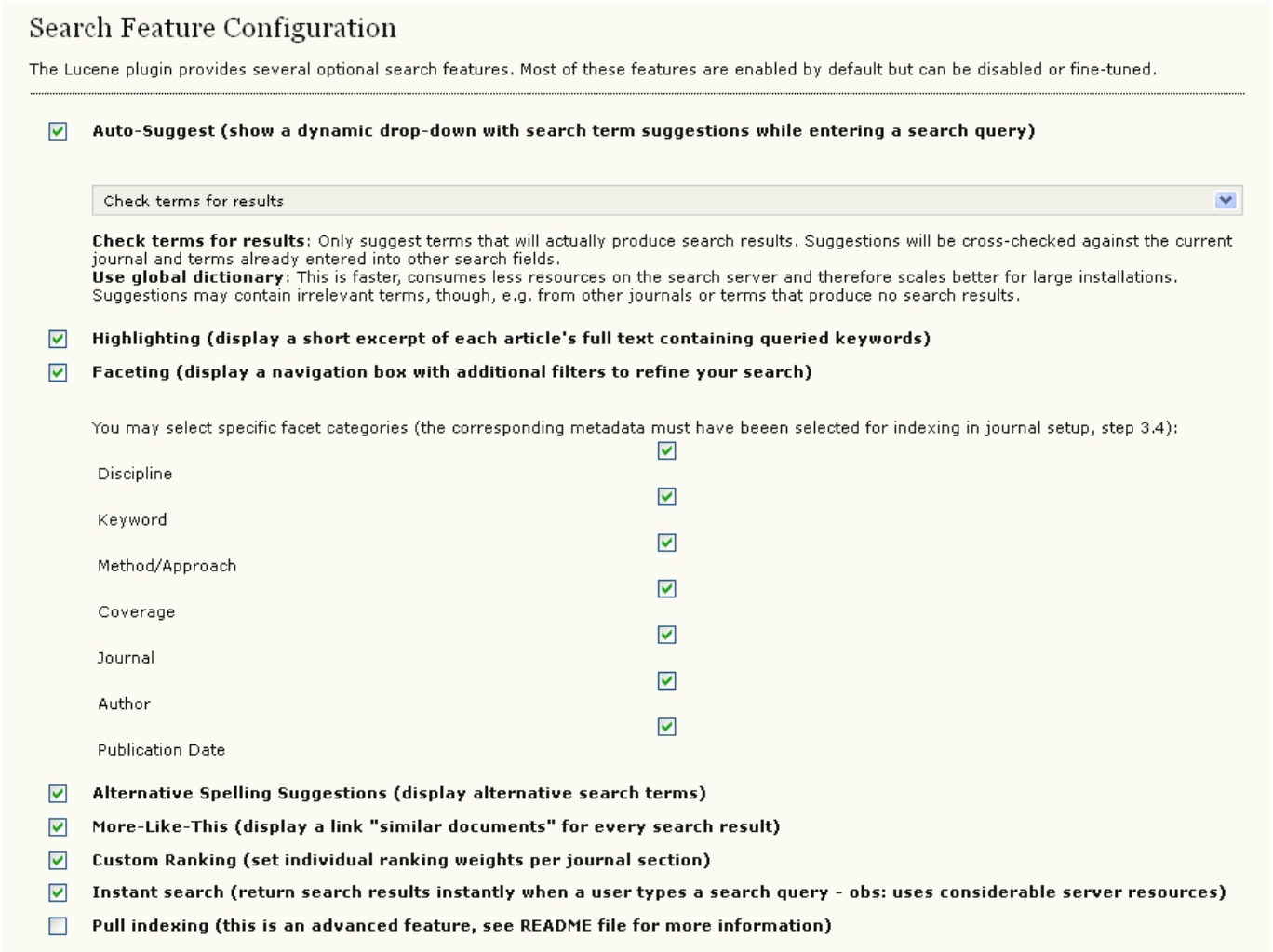
Unique Installation ID *
test-inst
If you use a central search server then you'll have to provide a unique installation ID for every OJS installation sharing the same search index. This can be any arbitrary text but it must be different for every participating OJS server (e.g. the server's static IP address if you have one OJS installation per server).

DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Search Feature Configuration

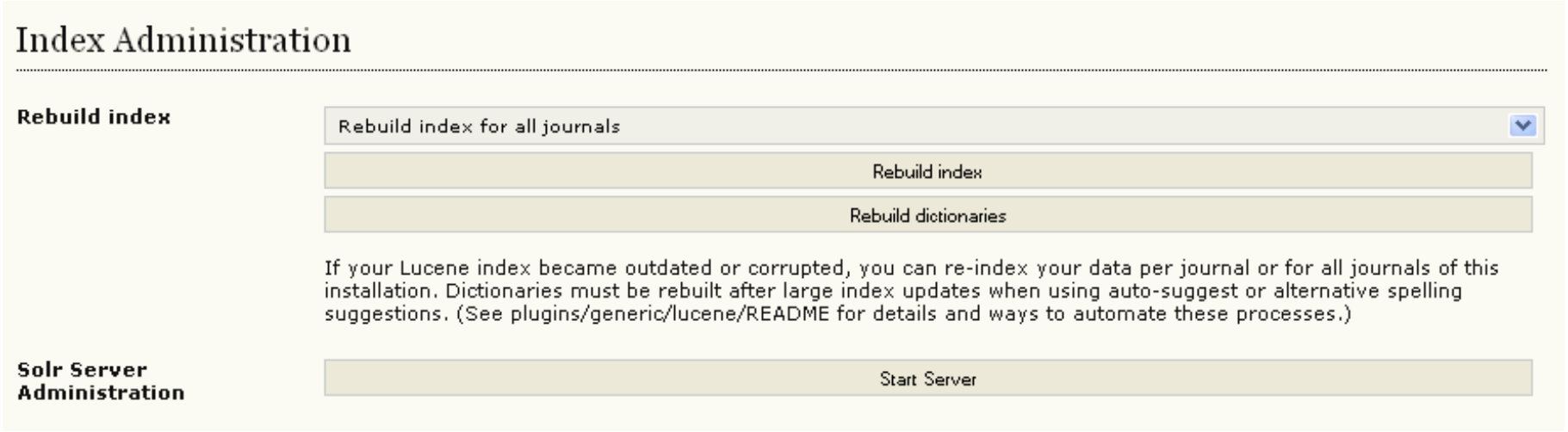


DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Index Administration

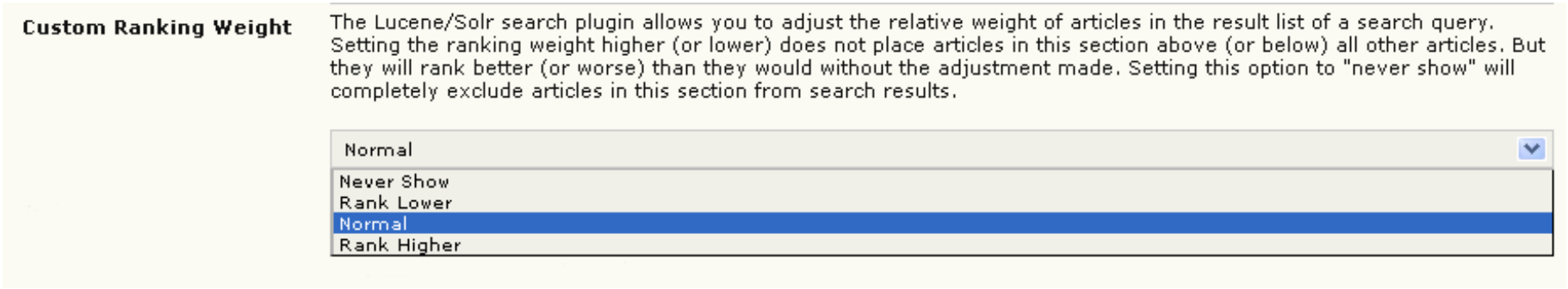


DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Custom Ranking Weight

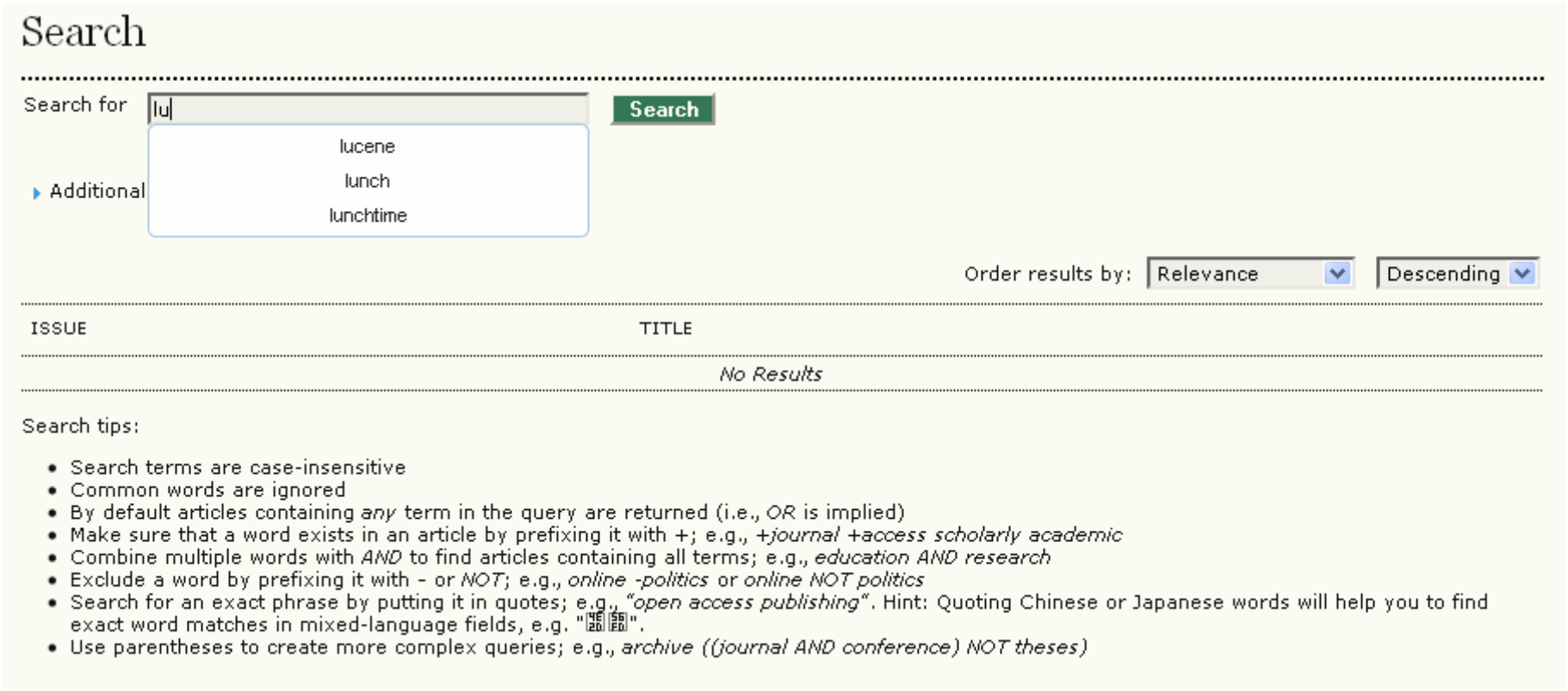


DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

Auto-suggestion



DFG

Božana Bokan
Center for Digital Systems (CeDiS)

Freie Universität Berlin

REFINE YOUR SEARCH

- Discipline
- Keyword
- Method/Approach
- Coverage
- Author
  - mcautomatic, arthur (3)
  - author, another (1)
  - author, some (1)
  - authorname, second a (1)
- Publication Date

HOME ABOUT USER HOME SEARCH CURRENT ARCHIVES

Home > Search

Search

Search for lucene test Search

Additional Search Options (click to show)

Order results by: Relevance Descending

ISSUE	TITLE	
Vol 1	Lucene Test Article 1	ABSTRACT HTML SIMILAR DOCUMENTS
Some Author, Second A Authorname "... **Lucene Test** Article 1 Abstract ..."		
Vol 1	Lucene Test Article 2	ABSTRACT PDF SIMILAR DOCUMENTS
Another Author "... **Lucene Test** Article 2 Abstract ..."		
Vol 1	Ranking Test Article 3	ABSTRACT SIMILAR DOCUMENTS
Arthur McAutomatic "... Ranking **Test** Article 3 abstract ..."		
Vol 1	Ranking Test Article 2	ABSTRACT SIMILAR DOCUMENTS
Arthur McAutomatic "... Ranking **Test** Article 2 abstract ..."		
Vol 1	Ranking Test Article 1	ABSTRACT SIMILAR DOCUMENTS
Arthur McAutomatic "... Ranking **Test** Article 1 abstract ..."		

1 - 5 of 5 Items

DFG

Freie Universität Berlin

Thank you!

Thanks Florian Grandel for realizing this!

Božana Bokan
bozana.bokan@fu-berlin.de
Freie Universität Berlin
Center for Digital Systems

DFG