Goldschmiede

anderScore
trust in competence

Solr Grundlagen

20.11.2015, Marcel Becker

Schedule

anderScore trust in competence

1. Was ist Solr?
2. Grundlagen Suchen
3. Schema
4. Solr Features / Solr Cloud
5. Hands On
6. Fragen
7. Links

Über mich

- Software Entwickler (Dipl.-Inf.) bei der anderScore GmbH, Köln
- Einige Jahre Erfahrung im Umfeld
  - Java
  - Web
  - Agile Methoden

Suche / Suchmaschine

Suche / Suchmaschine

anderScore trust in competence

- Wenig Struktur
- Natürliche Sprache
- Viele Treffer
- Große Datenmengen

Apache Lucene

High Performance

Index & Search

Apache Software License

Java Bibliothek

Apache Solr

anderScore
trust in competence

Stand-alone Server

HTTP / XML /
JSON / CSV

Apache Software License

Cluster / Cloud

Grundlagen

Invertierter Index

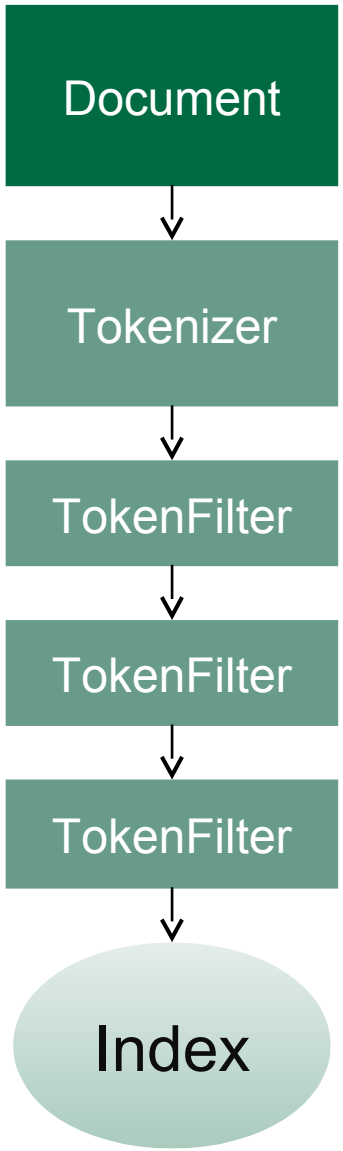
anderScore trust in competence

- **Dokumente**
  - Dokument 1: Die Hauptstadt von Frankreich heißt Paris.
  - Dokument 2: Paris ist die Hauptstadt von Frankreich.
  - Dokument 3: Die Hauptstädte von England bzw. Frankreich heißen London bzw. Paris.

- **Invertierter Index:**
  - bzw. -> 3
  - die -> 1, 2, 3
  - England -> 3
  - Frankreich -> 1, 2, 3
  - Hauptstadt -> 1, 2
  - Hauptstädte -> 3
  - heißen -> 3
  - heißt -> 1, 2
  - ist -> 2
  - London -> 3
  - Paris -> 1, 2, 3
  - von -> 1, 2, 3

Quelle: http://wikis.gm.fh-koeln.de/wiki_ir/InformationRetrieval/Invertierte-Liste

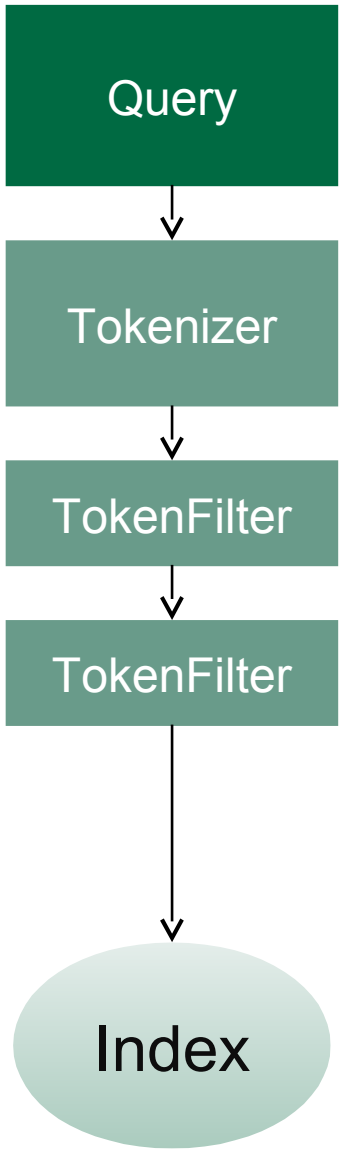
Indexer Pipe



Beispiele

- Whitespace Tokenzier
- LowerCaseFilter
- StopFilter
- SynonymFilter

Query Pipe

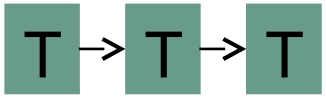


Beispiele

- Whitespace Tokenzier
- LowerCaseFilter
- StopFilter

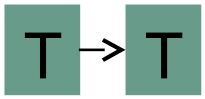
Beispiel

Dokument: Die Froscon findet in der Hochschule Bonn-Rhein-Sieg statt.



Tokens: froscon, findet, hochschule, fh, bonn-rhein-sieg, brs, statt

Query: FH BRS FrosCon 2015



Tokens: fh, brs, froscon, 2015

Match: fh, brs, froscon

Ranking / Score

- **tf** - Term Frequency
- **idf** - Inverse Document Frequency
- **coord** - Coordination Factor
- **fieldNorm** - Field length

- **distance** - bei Fuzzy Search

- **boost**
  - **Index**
  - **Query:** `title:"foo bar"^10`

schema.xml

anderScore trust in competence

```
<fieldType name="text_general" class="solr.TextField">
   <analyzer type="index">
      <charFilter class="solr.MappingCharFilterFactory" mapping="ascii.txt"/>
      <tokenizer class="solr.WhitespaceTokenizerFactory" />
      <filter class="solr.LowerCaseFilterFactory" />
      <filter class="solr.StopFilterFactory" words="stopwords.txt" />
      <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" />
   </analyzer>
   <analyzer type="query">
      <charFilter class="solr.MappingCharFilterFactory" mapping="ascii.txt"/>
      <tokenizer class="solr.WhitespaceTokenizerFactory" />
      <filter class="solr.LowerCaseFilterFactory" />
      <filter class="solr.StopFilterFactory" words="stopwords.txt" />
   </analyzer>
</fieldType>


<field name="text" type="text_general" indexed="true" stored="true" />


<dynamicField name="*_txt" type="text_general" indexed="true" stored="true" />
```

Solr Features

anderScore trust in competence

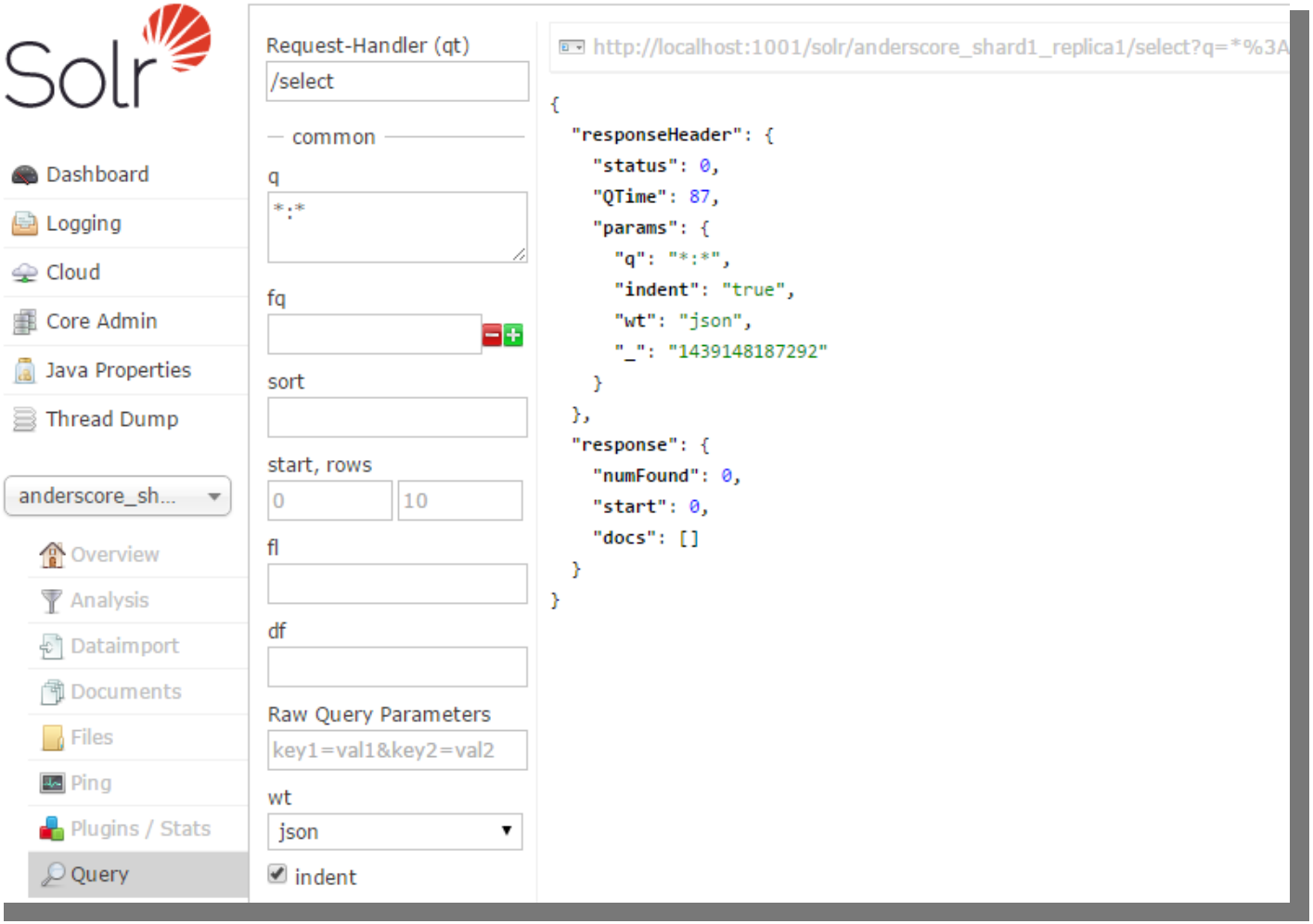
- Java API

```
String collection = "anderscore";

Book book = new Book();
book.id = "1";
book.author_txt = "Eric Ries";
book.title_txt = "The Lean Startup";

Address address = new Address();
address.id = "2";
address.addressLine_txt = "Grantham-Allee 20";
address.postalCode_txt = "53757";
address.city_txt = "Sankt Augustin";
address.country_txt = "Germany";

CloudSolrClient c = new CloudSolrClient("localhost:2001");
c.addBean(collection, book);
c.addBean(collection, address);

c.close();
```

Solr Features

anderScore trust in competence

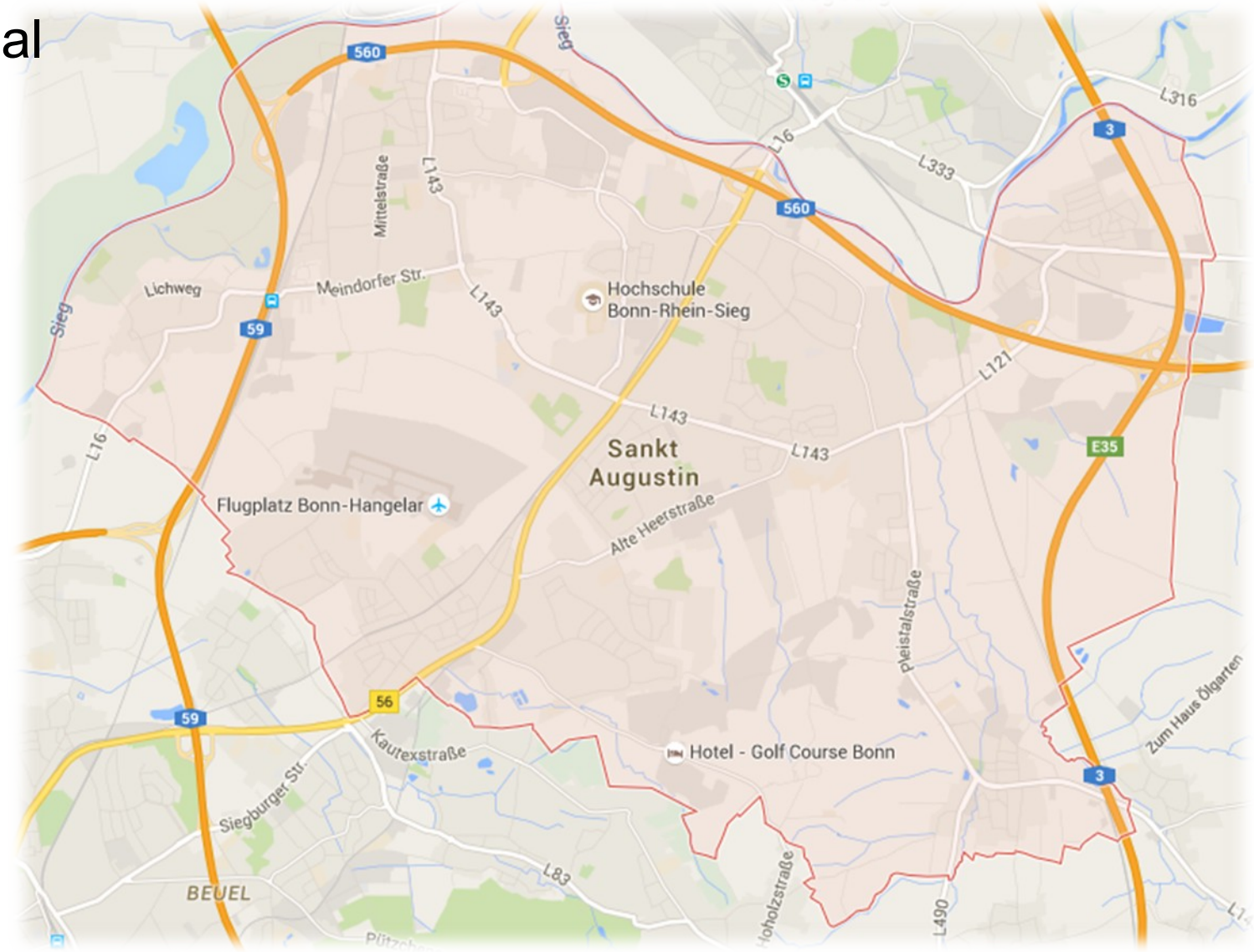
- Admin UI



Solr Features

anderScore
trust in competence

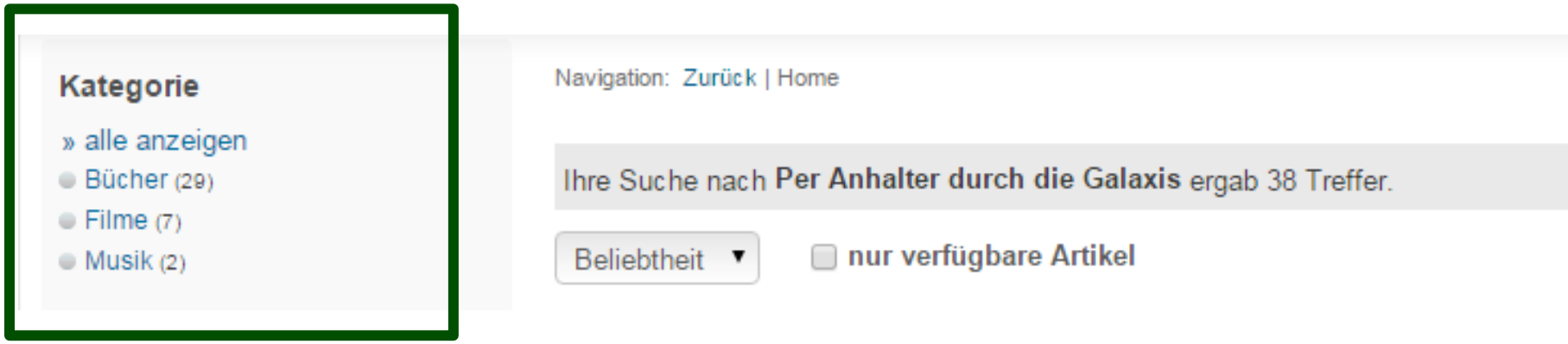
- Geospatial Support



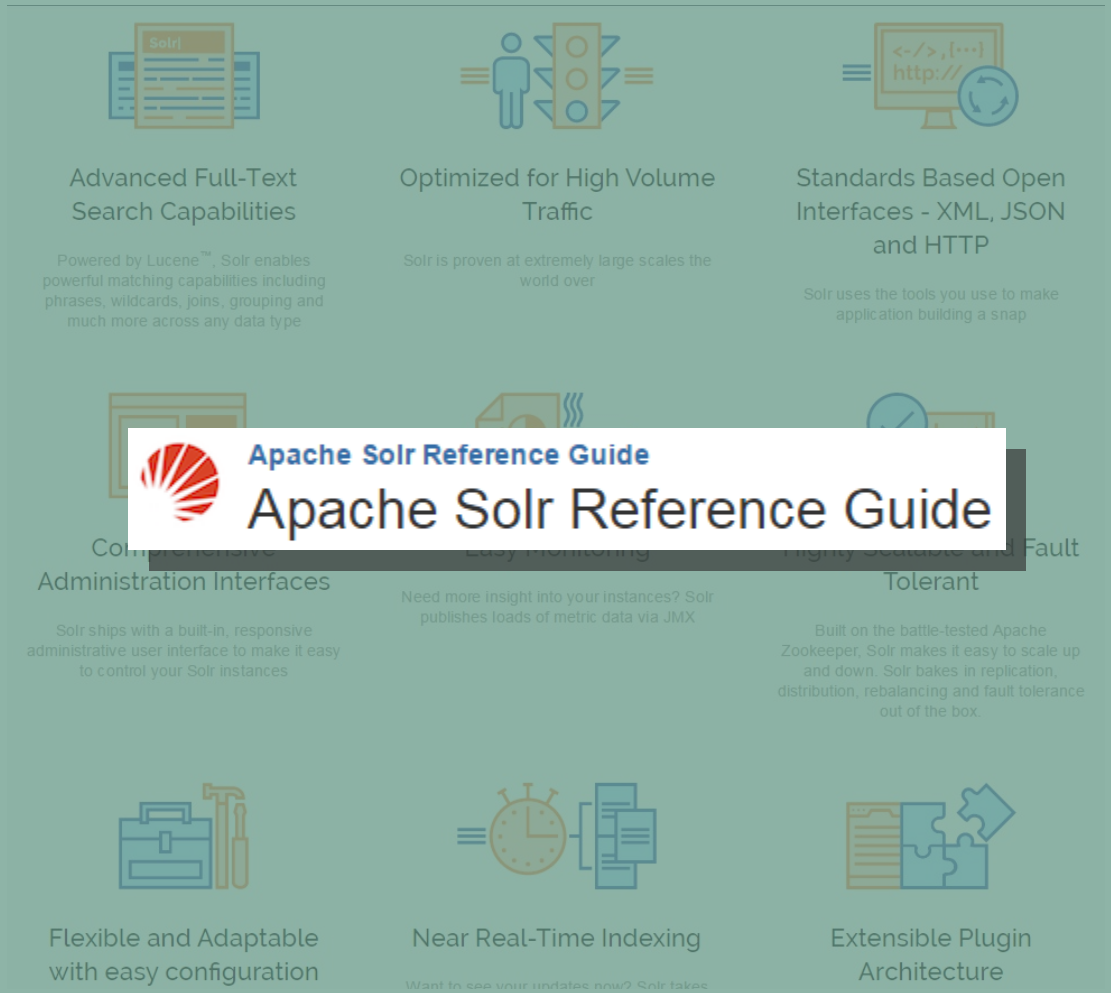
Quelle: Google Maps

Solr Features

- Facets:



Solr Features



Quelle: http://lucene.apache.org/solr/features.html

SolrCloud

- Menge von Features, die helfen einen Index horizontal zu skalieren (verteilen):
  - Sharding
  - Replikation
- Ziele:
  - Skalierbarkeit
  - Performance
  - Hoch-Verfügbarkeit
  - Einfachheit / Komfort

SolrCloud - Terminologie

- ZooKeeper
  - Zentralisierte Konfiguration
  - Cluster-State Management
  - Leader Election
- Node
- Core
- Collection
  - Shard
  - Replication Factor
  - Hash Range
  - Leader
- Replica
- Leader

Quelle: http://de.slideshare.net/thelabdude/solr-exchange-introtosolrcloud

Hands On

Fragen

Danke

Quellen und Links

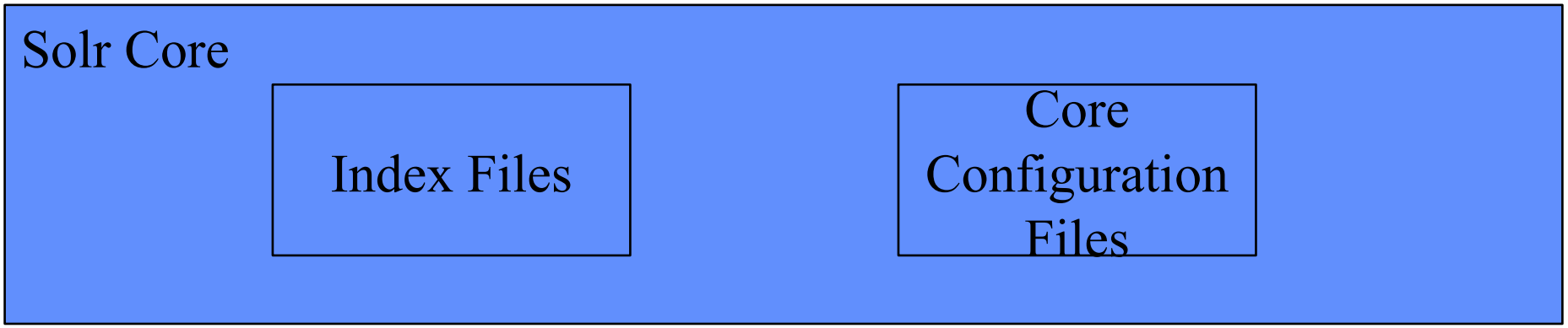
- http://lucene.apache.org/solr/
- http://lucene.apache.org/index.html

- http://wikis.gm.fh-koeln.de/wiki_ir/InformationRetrieval/Invertierte-Liste
- http://de.slideshare.net/thelabdude/solr-exchange-introtosolrcloud

Anhang

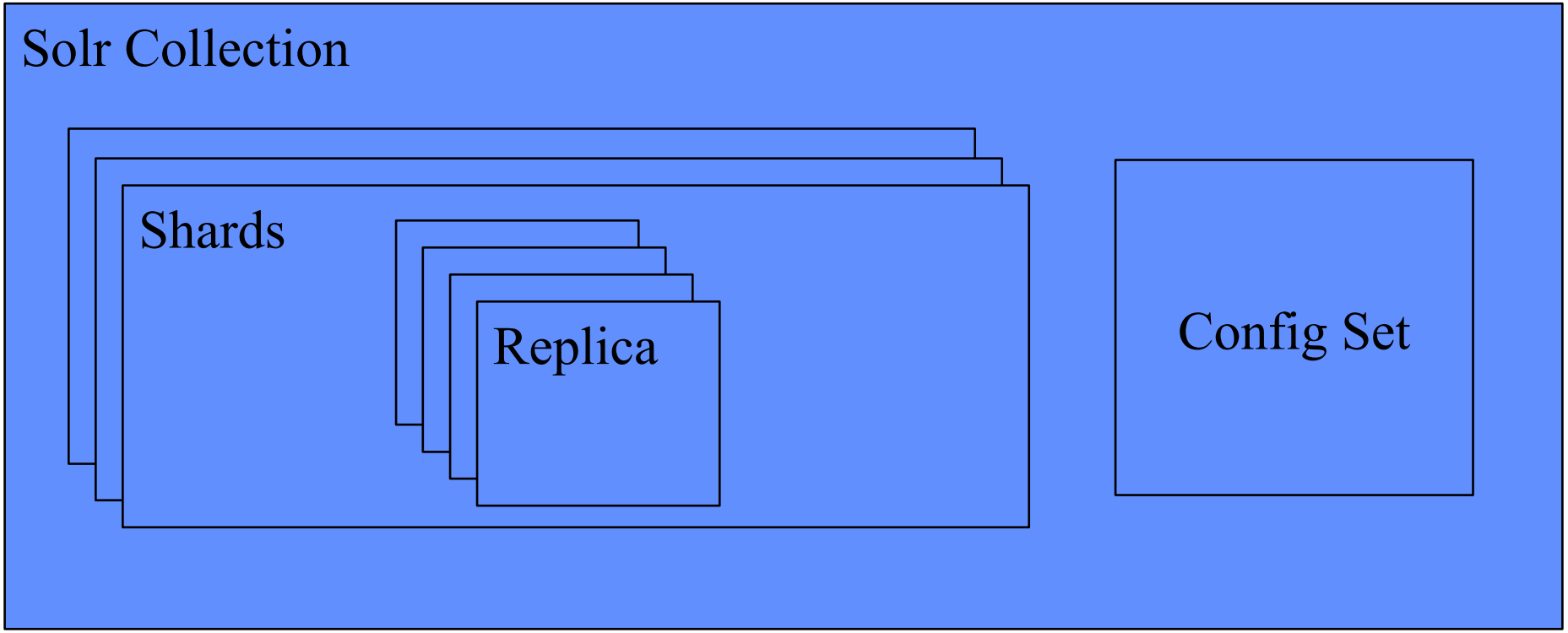
Solr Indexes

- Classic Solr



Distributed Indexes

- Solr Cloud (Logical View)



Distributed Indexes

anderScore trust in competence

- Solr Cloud (Physical View)