Facilitating the Retrieval of OPM Health Plan Guidance Documents With an Open Source Search Engine

LHNCBC Maria Yang & Lauren Evoy

NIH U.S. National Library of Medicine

About OPM

- The Office of Personnel Management
  - “HR Department” of the government
- Provide healthcare options for all Federal employees
  - Offer guidance documents and brochures
  - Answer questions about coverage of specific conditions

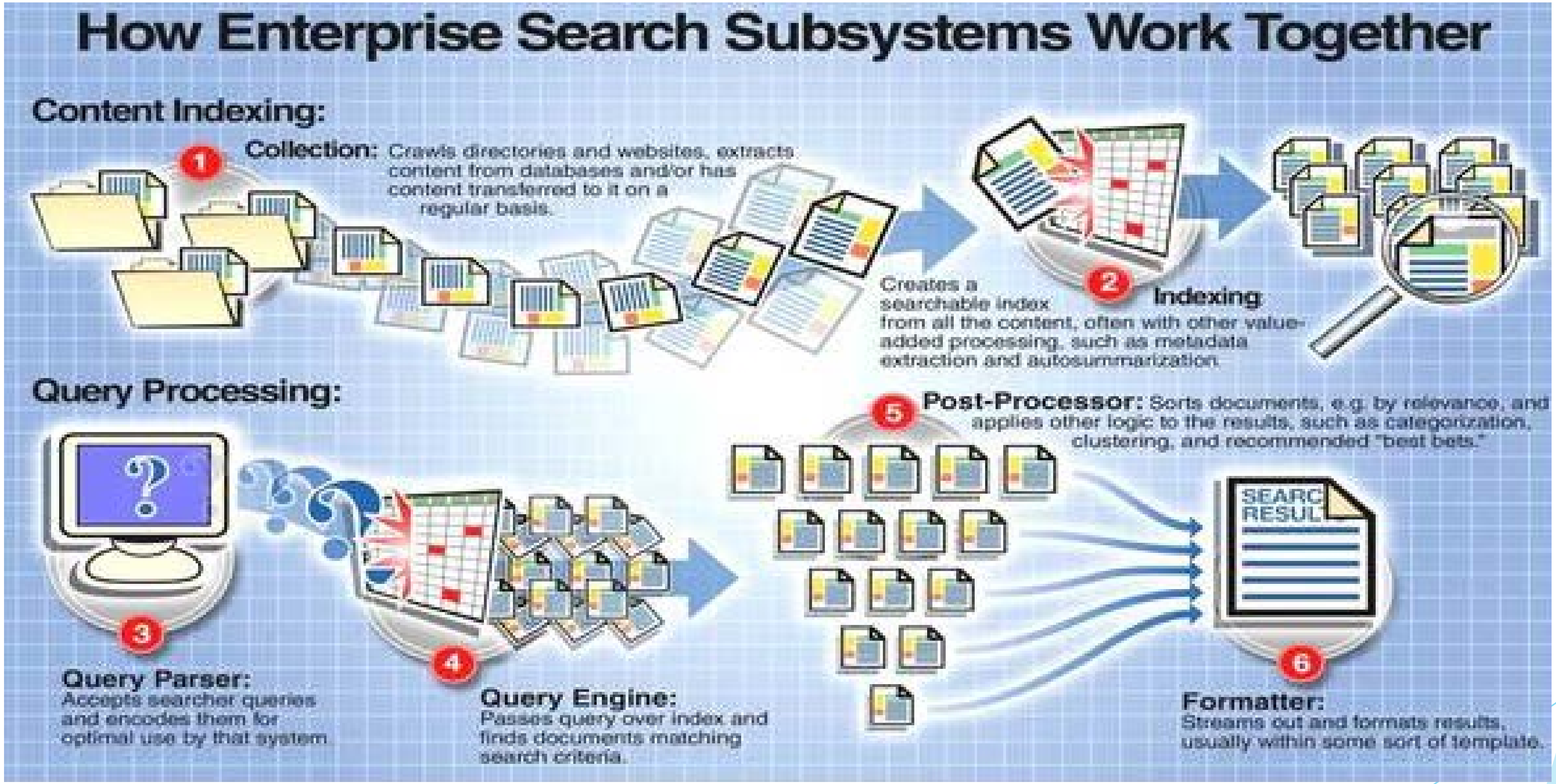
Motivation & Goal

- **Motivation**:
  - Guidance information scattered across hundreds of PDF documents
  - Difficult to search and retrieve information from these documents

> - What guidance has OPM issued regarding the use of tobacco?
>   *Keywords: tobacco, cessation counseling, quit attempts*
> - When does federal preemption affect program governance?
>   *Keywords: state mandate, preemption, preempt, FEHBA*

- **Goal**: to make the archive of letters and brochures searchable

Background



How Enterprise Search Subsystems Work Together

Content Indexing:

1 **Collection:** Crawls directories and websites, extracts content from databases and/or has content transferred to it on a regular basis.

2 **Indexing** Creates a searchable index from all the content, often with other value-added processing, such as metadata extraction and autosummarization.

Query Processing:

3 **Query Parser:** Accepts searcher queries and encodes them for optimal use by that system.

4 **Query Engine:** Passes query over index and finds documents matching search criteria.

5 **Post-Processor:** Sorts documents, e.g. by relevance, and applies other logic to the results, such as categorization, clustering, and recommended "best bets."

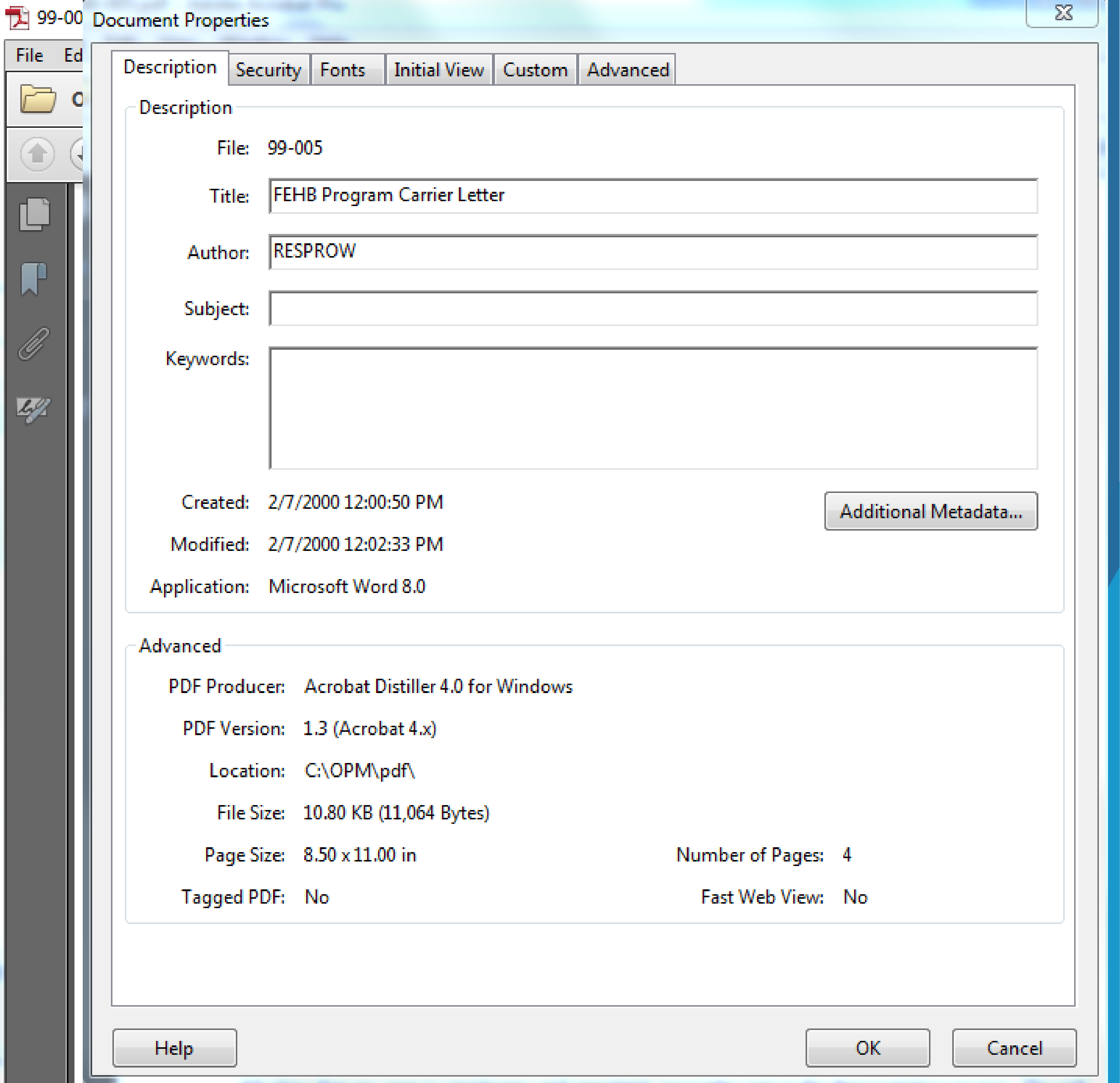
6 **Formatter:** Streams out and formats results, usually within some sort of template.

Solr

- Solr 5.2.1 (latest version): a standalone enterprise search server
- REST-like API, allowing for application building
- Important features:
  - Full-text search
  - Real-time indexing
  - Rich document handling (PDF, Word, Excel)
    - Tika module: detects and extracts metadata and text from different file types
  - Comprehensive admin interface

Materials

- 969 PDF files from OPM's archive (carrier letters, brochures)
- Data
  - body of document
- Available metadata
  - creation date
  - last modification date
  - title
  - author
  - file name
  - size



Document Properties

Description

File: 99-005

Title: FEHB Program Carrier Letter

Author: RESPROW

Subject:

Keywords:

Created: 2/7/2000 12:00:50 PM

Modified: 2/7/2000 12:02:33 PM

Application: Microsoft Word 8.0

Advanced

PDF Producer: Acrobat Distiller 4.0 for Windows

PDF Version: 1.3 (Acrobat 4.x)

Location: C:\OPM\pdf\

File Size: 10.80 KB (11,064 Bytes)

Page Size: 8.50 x 11.00 in

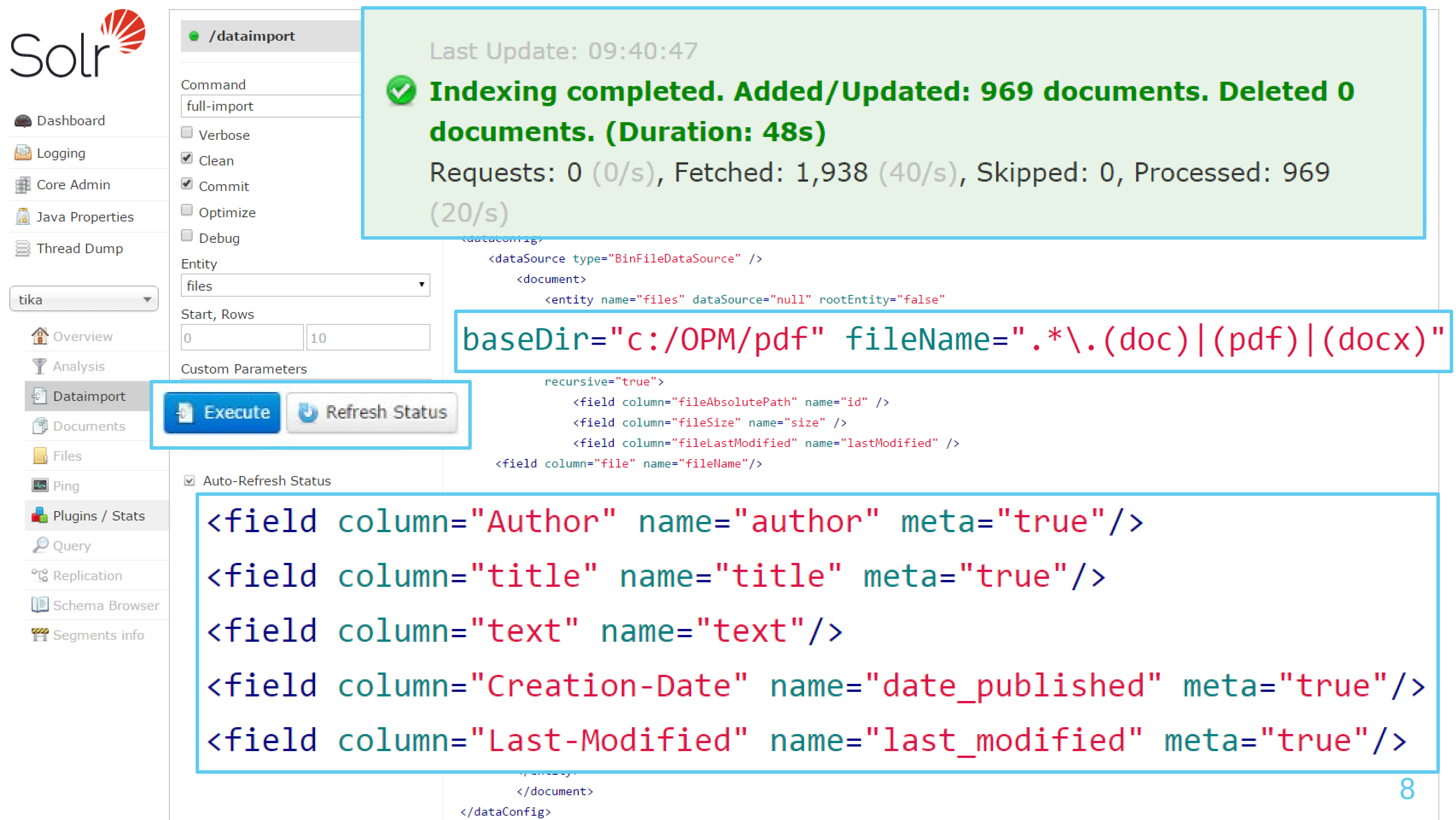
Number of Pages: 4

Tagged PDF: No

Fast Web View: No

Methods

- Acquiring collection of files
  - Web crawler written to download all PDF files into a specified folder
- Indexing of documents
  - Add to Solr using dataimport
  - There are fields for data (text) and metadata (e.g., author, dates)
  - Make all fields searchable
- Querying
  - Boolean operations between keywords
  - Synonyms
  - Search by field



Command: full-import

☐ Verbose
☑ Clean
☑ Commit
☐ Optimize
☐ Debug

Entity: files

Start, Rows: 0 10

Custom Parameters

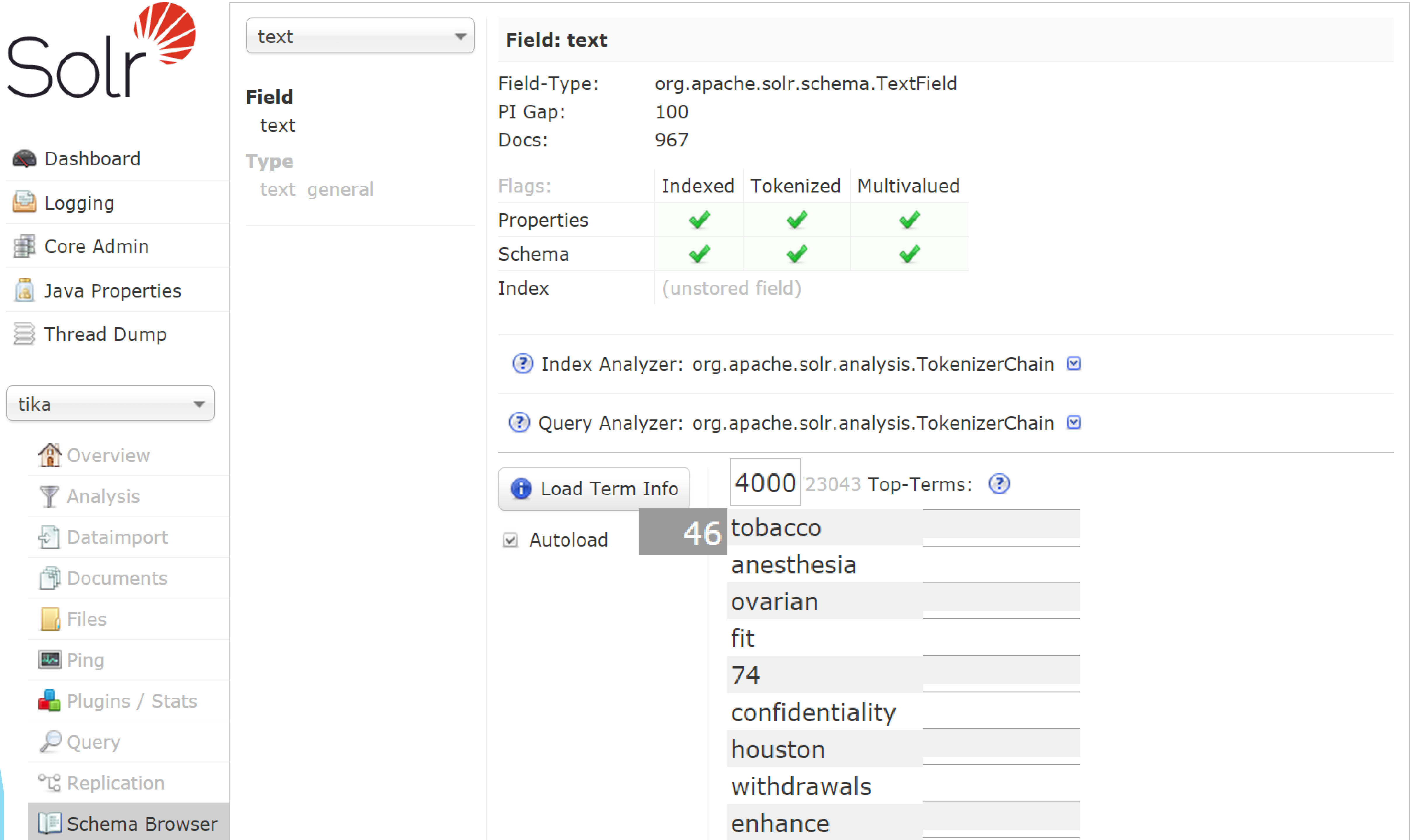
☑ Auto-Refresh Status

Last Update: 09:40:47

✓ **Indexing completed. Added/Updated: 969 documents. Deleted 0 documents. (Duration: 48s)**

Requests: 0 (0/s), Fetched: 1,938 (40/s), Skipped: 0, Processed: 969 (20/s)

```
<dataconfig>
    <dataSource type="BinFileDataSource" />
        <document>
            <entity name="files" dataSource="null" rootEntity="false"
baseDir="c:/OPM/pdf" fileName=".*\.(doc)|(pdf)|(docx)"
            recursive="true">
                <field column="fileAbsolutePath" name="id" />
                <field column="fileSize" name="size" />
                <field column="fileLastModified" name="lastModified" />
      <field column="file" name="fileName"/>
<field column="Author" name="author" meta="true"/>
<field column="title" name="title" meta="true"/>
<field column="text" name="text"/>
<field column="Creation-Date" name="date_published" meta="true"/>
<field column="Last-Modified" name="last_modified" meta="true"/>
            </entity>
        </document>
</dataConfig>
```



Field
text

Type
text_general

Field: text

Field-Type: org.apache.solr.schema.TextField
PI Gap: 100
Docs: 967

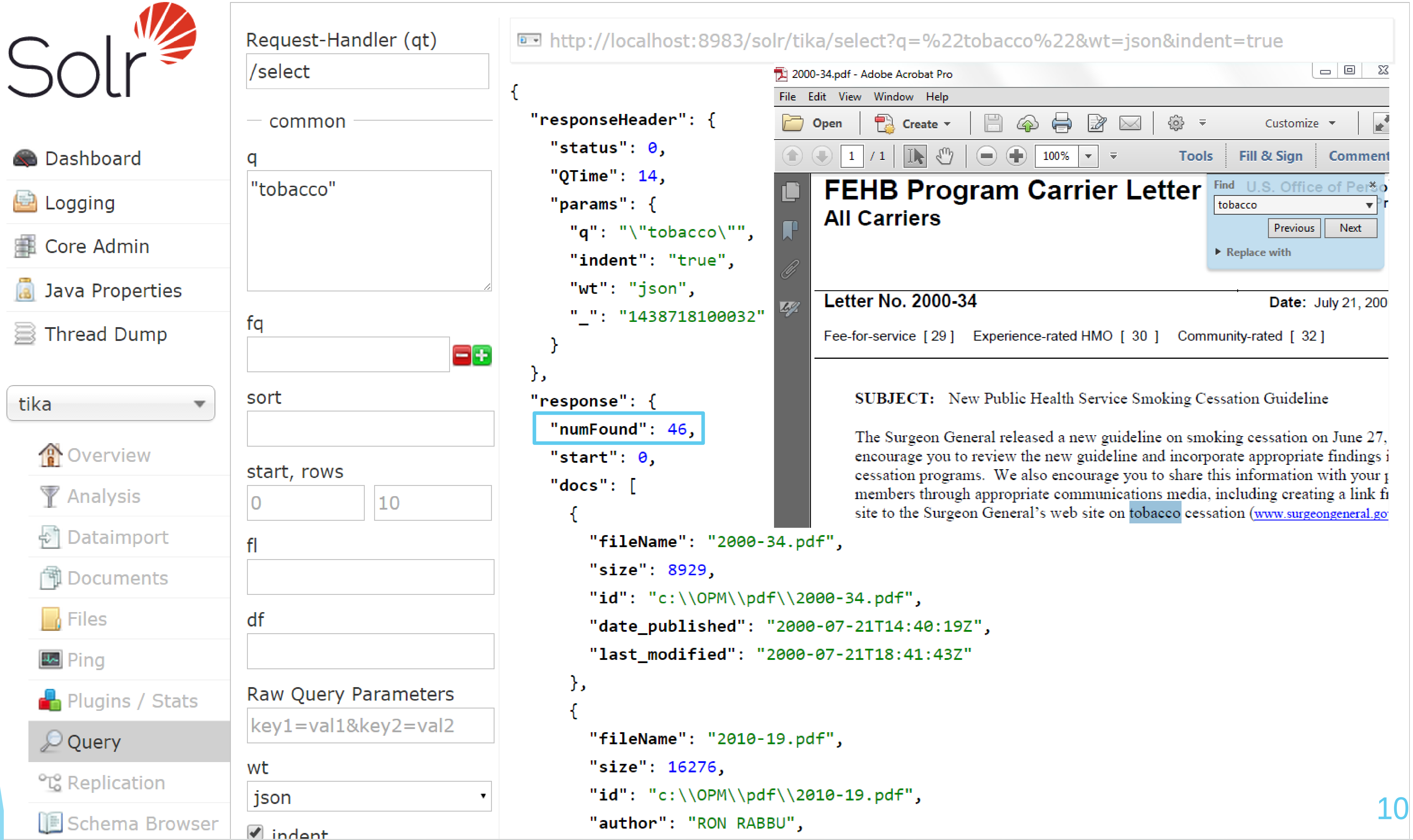
Flags:	Indexed	Tokenized	Multivalued
Properties	✓	✓	✓
Schema	✓	✓	✓
Index	(unstored field)		

Index Analyzer: org.apache.solr.analysis.TokenizerChain

Query Analyzer: org.apache.solr.analysis.TokenizerChain

4000 23043 Top-Terms:

46	tobacco
	anesthesia
	ovarian
	fit
	74
	confidentiality
	houston
	withdrawals
	enhance



Request-Handler (qt): /select

common

q: "tobacco"

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

http://localhost:8983/solr/tika/select?q=%22tobacco%22&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 14,
    "params": {
      "q": "\"tobacco\"",
      "indent": "true",
      "wt": "json",
      "_": "1438718100032"
    }
  },
  "response": {
    "numFound": 46,
    "start": 0,
    "docs": [
      {
        "fileName": "2000-34.pdf",
        "size": 8929,
        "id": "c:\\OPM\\pdf\\2000-34.pdf",
        "date_published": "2000-07-21T14:40:19Z",
        "last_modified": "2000-07-21T18:41:43Z"
      },
      {
        "fileName": "2010-19.pdf",
        "size": 16276,
        "id": "c:\\OPM\\pdf\\2010-19.pdf",
        "author": "RON RABBU",
```

Find: tobacco

U.S. Office of Perso

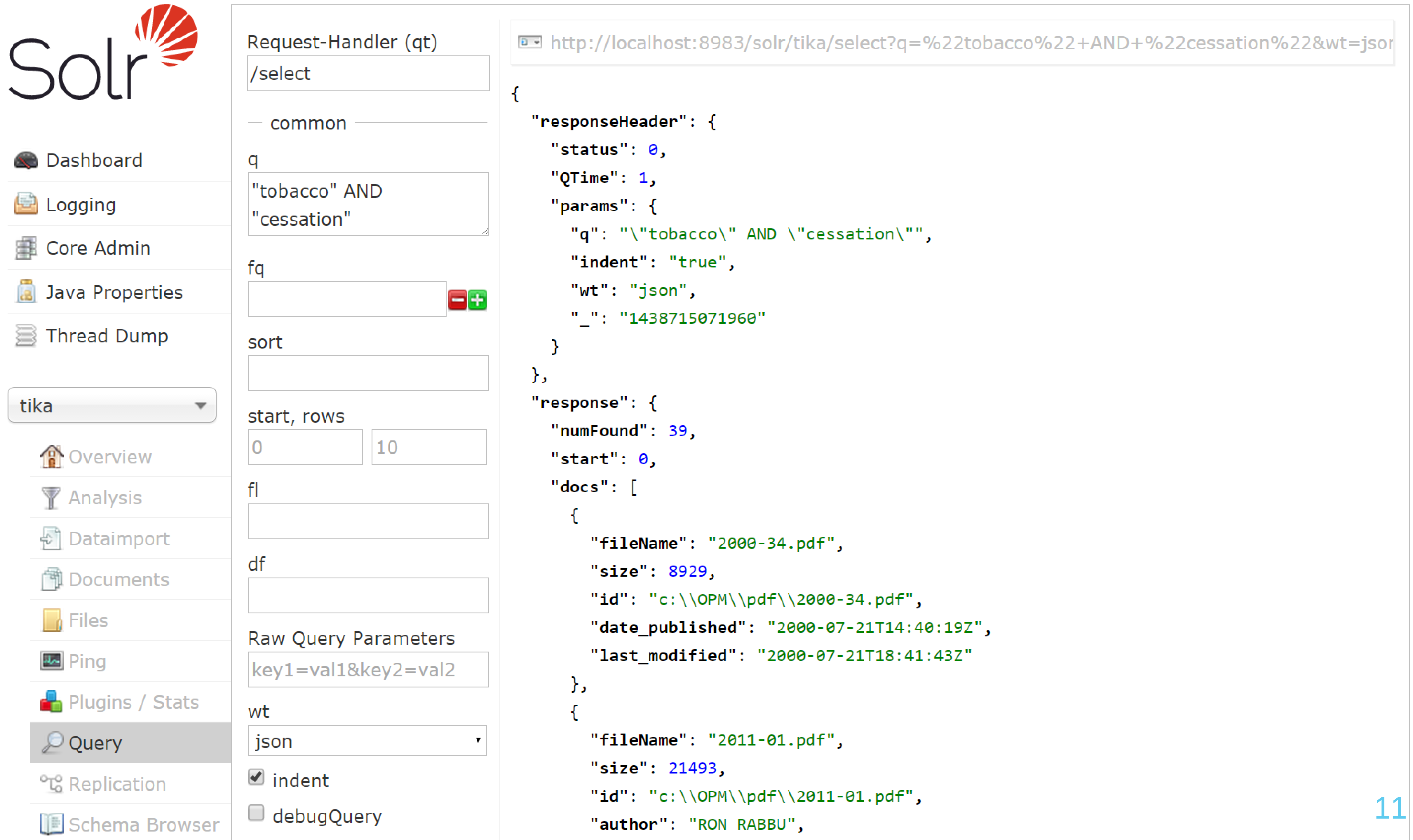
FEHB Program Carrier Letter
All Carriers

Letter No. 2000-34 **Date:** July 21, 200

Fee-for-service [29] Experience-rated HMO [30] Community-rated [32]

SUBJECT: New Public Health Service Smoking Cessation Guideline

The Surgeon General released a new guideline on smoking cessation on June 27, encourage you to review the new guideline and incorporate appropriate findings i cessation programs. We also encourage you to share this information with your p members through appropriate communications media, including creating a link fr site to the Surgeon General's web site on tobacco cessation (www.surgeongeneral.go



Request-Handler (qt): /select

q: "tobacco" AND "cessation"

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

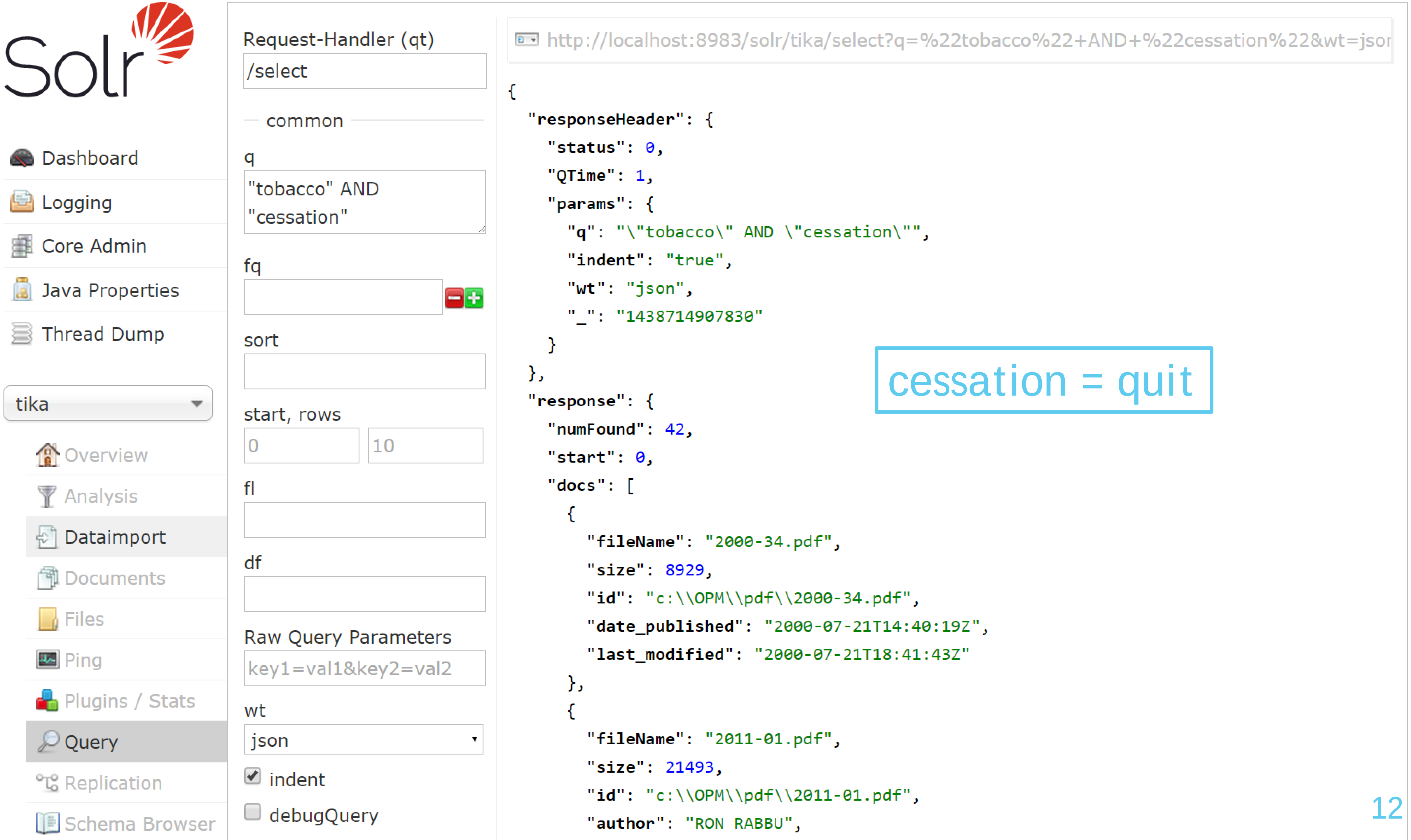
wt: json

☑ indent

□ debugQuery

http://localhost:8983/solr/tika/select?q=%22tobacco%22+AND+%22cessation%22&wt=jsor

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "q": "\"tobacco\" AND \"cessation\"",
      "indent": "true",
      "wt": "json",
      "_": "1438715071960"
    }
  },
  "response": {
    "numFound": 39,
    "start": 0,
    "docs": [
      {
        "fileName": "2000-34.pdf",
        "size": 8929,
        "id": "c:\\OPM\\pdf\\2000-34.pdf",
        "date_published": "2000-07-21T14:40:19Z",
        "last_modified": "2000-07-21T18:41:43Z"
      },
      {
        "fileName": "2011-01.pdf",
        "size": 21493,
        "id": "c:\\OPM\\pdf\\2011-01.pdf",
        "author": "RON RABBU",
```



Request-Handler (qt): /select

common

q: "tobacco" AND "cessation"

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

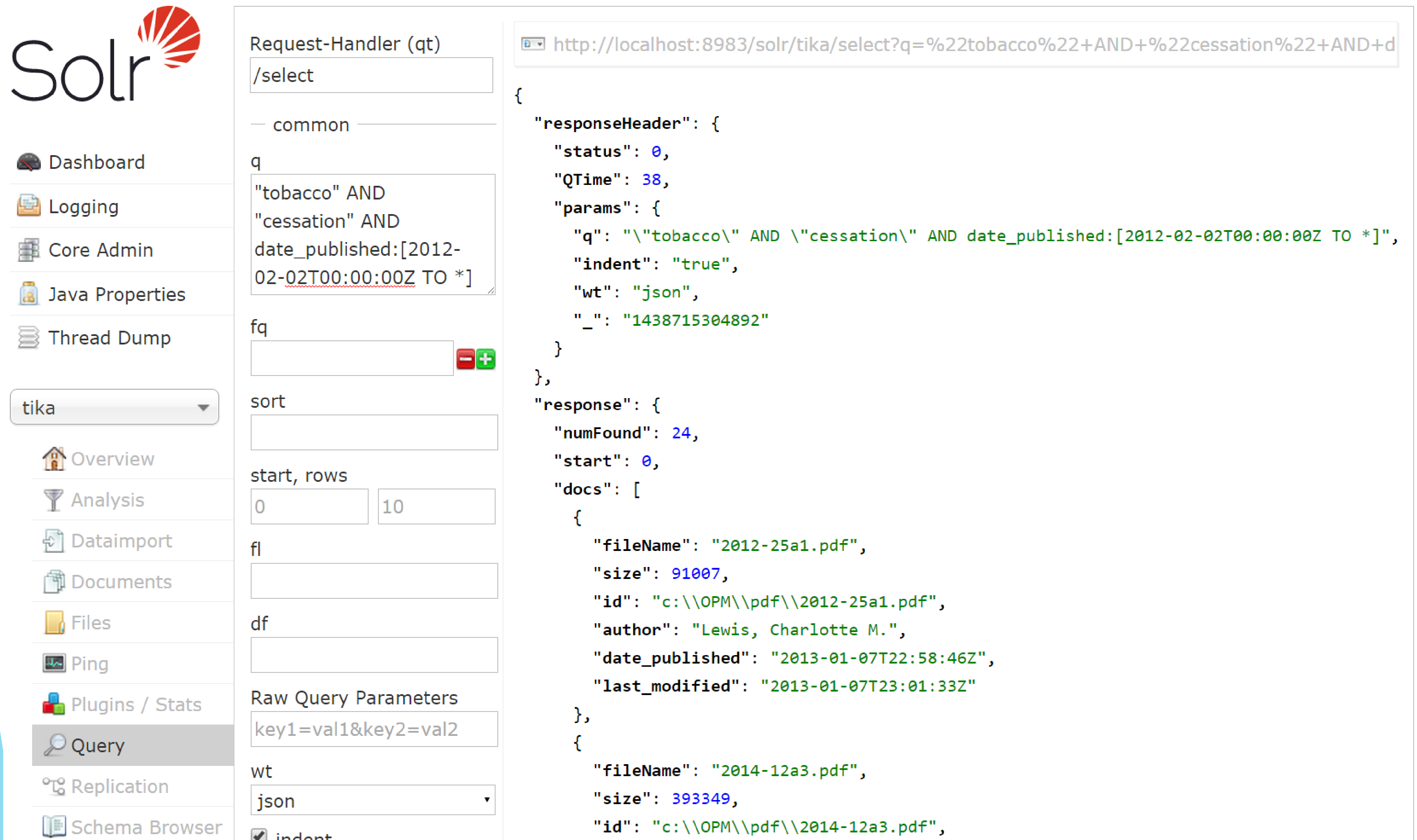
☑ indent

□ debugQuery

http://localhost:8983/solr/tika/select?q=%22tobacco%22+AND+%22cessation%22&wt=jsor

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "q": "\"tobacco\" AND \"cessation\"",
      "indent": "true",
      "wt": "json",
      "_": "1438714907830"
    }
  },
  "response": {
    "numFound": 42,
    "start": 0,
    "docs": [
      {
        "fileName": "2000-34.pdf",
        "size": 8929,
        "id": "c:\\OPM\\pdf\\2000-34.pdf",
        "date_published": "2000-07-21T14:40:19Z",
        "last_modified": "2000-07-21T18:41:43Z"
      },
      {
        "fileName": "2011-01.pdf",
        "size": 21493,
        "id": "c:\\OPM\\pdf\\2011-01.pdf",
        "author": "RON RABBU",
```

cessation = quit



Request-Handler (qt)
/select

common

q
"tobacco" AND "cessation" AND date_published:[2012-02-02T00:00:00Z TO *]

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

☑ indent

http://localhost:8983/solr/tika/select?q=%22tobacco%22+AND+%22cessation%22+AND+d

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 38,
    "params": {
      "q": "\"tobacco\" AND \"cessation\" AND date_published:[2012-02-02T00:00:00Z TO *]",
      "indent": "true",
      "wt": "json",
      "_": "1438715304892"
    }
  },
  "response": {
    "numFound": 24,
    "start": 0,
    "docs": [
      {
        "fileName": "2012-25a1.pdf",
        "size": 91007,
        "id": "c:\\OPM\\pdf\\2012-25a1.pdf",
        "author": "Lewis, Charlotte M.",
        "date_published": "2013-01-07T22:58:46Z",
        "last_modified": "2013-01-07T23:01:33Z"
      },
      {
        "fileName": "2014-12a3.pdf",
        "size": 393349,
        "id": "c:\\OPM\\pdf\\2014-12a3.pdf",
```

Limitations & Challenges

- No evaluation for the queries
  - Did not have a gold standard
- Not user-friendly interface
  - Using the admin interface as user interface
- Difficult to use advanced features such as hit highlighting
- Limited Solr documentation quality and availability

Conclusions

- Successfully developed a proof of concept
  - Presented to OPM colleagues last week
- Successfully indexed all documents, taking roughly a minute
- Able to provide a (minimal) searching interface
- Found that it is easy to add to this collection
  - Collection can be expanded and re-indexed

Acknowledgements

NLM

- Dr. Olivier Bodenreider
- Mr. Phill Wolf
- Mr. Lee Peters
- Ms. Keyla Cooper
- Dr. Paul Fontelo
- Dr. Clem McDonald

OPM

- Rhoda Schulzinger, Senior Policy Analyst
- Meredyth Hindsley, Program Analyst
- Anne Easton, Deputy Director of Planning & Policy Analysis
- Christine Hunter, Chief Medical Officer
- Merle Townley, III, Healthcare Program Manager