



An On-premise Search Service

updated

Joe Doupnik

jrd@netlab1.net

Prof (retired) Univ of Oxford

MindworksUK

Version 3, Feb 2025

Introduction



I have previously talked about the search service topic.
Two aspects have urged me to discuss it again.

- a) The many technical changes by Apache Solr and Linux folks.
Mysterious system conflicts arose, major components had evolved separately, Solr had failures. The machinery won that contest.
But I have returned with better techniques, and now we win.
- b) We should consider offering this facility to groups at our site as useful assistance and gestures of good will to our colleagues.
Think about this because the step can help a site.



What is this about?

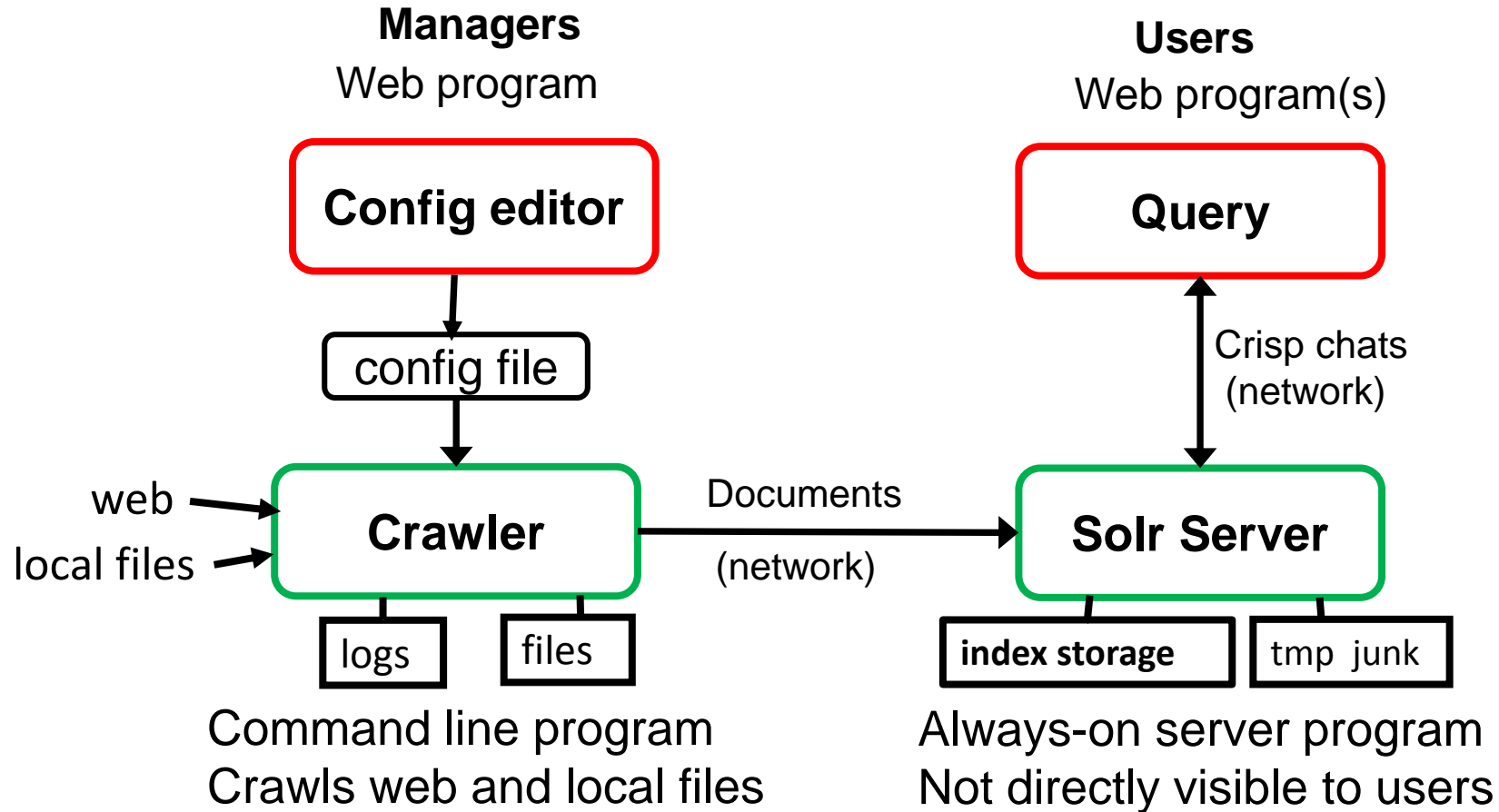
- Providing file content Search facilities for our environment
- Users expect it, diverse groups can benefit from it
- Different indices for different audiences
- After construction local groups can operate their own installation or feed into centralized indexers
- Data remains local, control remains local, we own it
- Modern web appearance, mobile device friendly, changable
- Scales to hundreds of thousands of documents

(On my main OES server I have indexed 250,000 files)

- Components (all are free) can reside in production Linux servers

This is Big Data territory but domesticated to work nicely in our existing file servers. Expertise is not required.

Topology of the search service



The system allows for multiple Solrs, crawlers, query programs, and multiple Solr schema. Parts may be located on many machines.

Aspects to note

The system is designed to allow diversified deployment, spread throughout a site as one or more Solr *indexers*, many *crawler* agents feeding files into indexers, and many different *query* agents.

Communication amongst agents is over the network, plus share a config file.

Ownership of parts can be diverse, and they are usable by many kinds of people (expertise is not required).

Ease of operation is a goal.

Security and privacy are essential ingredients.

It runs in Linux machines, including our OES servers and as appliance(s).

Installation steps are in distribution file [Installing when using Solr v9.7.0.txt](#)

In the beginning

We think about the tasks to manually create and update one or more indices. Hmm, that can be a lot to remember easily.

The current system tries to cluster these chores into a simple web interface which regular workers can deal with. Advice, but no training course is needed. Routine updates can be automated.

The real complexities remain within the worker apps, available for a system manager to adjust. Modification work is easily accomplished by reviewing the commented PHP code.

We start with the all-in-one configuration application `config.php`.

Crawler configuration files, config.php

Web and file crawler configuration editor

Choose or create a crawler configuration file

Files in /home/search/index. Click a filename in the list to choose:

k12list.cfg opennovell.cfg provotalk.cfg slctalk.cfg tfolder.cfg ttplist.cfg
ttpweb.cfg

Create new file

These files control the Crawler.

The menu shows existing crawler config files, such as my set above, and allows creation of new ones as shown in the next slide.

Details appear after choosing or while creating a file.

Configuration files reside in subdirectory “index” beneath config.

Web and file crawler configuration editor

Edit configuration file

* means a required field

* Solr index

Must start with a letter, then alphanumerics, hyphen and minus. Keep short

Description

History in the making

* Crawl base Where to crawl: [http\(s\)://example.com/goodies/](http://example.com/goodies/) or </some/local/place>Show base as

In the Solr index replace prefix Crawl base with this text. Changes index field url

Username Password

Credentials for authentication to a web source.

Not before

digits YYYY-MM-DD or phrase n days ago where n is the count of whole days.

Ignore files before this date. Leave empty to omit the date test.

Replace index Yes ☐ No ☒

Yes - empty and replace Solr index. No - add this data to the index

Index files having no filename extension Yes ☐ No ☒

Yes - index files not having a filename extension. No - read but do not include in the index.

* File filter (can edit it here)

+/- accept/reject result upon a pattern match and exit test. PHP regex syntax follows.

^/\$ start/end of line indicators, . (dot) is any char, .+ (dot plus) is 1 or more chars,

(a|b|c) groups OR items, \char treats char as literal, ; (semicolon) starts a comment.

```

; Example filter to select files with common extensions,
; reject other extensions, allow those with no extension.
; Normally place exceptions above the main accept(+) line.
; If no match is found then accept the filename.
+^.\+\. (pdf|ppt|pptx|doc|docx|xls|xlsx|txt) $
+^.\+\. (htm|html|sww|odp|stw|xml|rtf|odt|ods) $
-^.\+\.+$

```

Logging Yes ☒ No ☐

Enable results log.

Solr delay (ms)

Delay (milliseconds) after sending a file to Solr. Suggest 100 or larger.

Crawl depth

Limit exploration to this many levels below the starting point. Empty defaults to 8.

* Solr address Address of Solr server. Format is <http://localhost:8983/solr>Solr username Password

Credentials for authentication to Solr.

Solr configset Solr internal configuration set. Leave empty unless Solr has been configured for it.Save as

Cancel

Fresh edit of a new file

Whole edit screen page:

14 options,

4 are required,

2 if use defaults

Many details are here when needed. We use only 2-3 for ordinary work. Thank goodness.

Cheat sheet: edit an existing file and save results under a new name. Saves typing.

Detailed views follow, as top, middle, bottom

Create screen top, what to do portion

Web and file crawler configuration editor

Edit configuration file

* means a required field

* Solr index

Must start with a letter, then alphanumerics, hyphen and minus. Keep short

Description

History in the making

* Crawl base

Where to crawl: **http(s)://example.com/goodies/** or **/some/local/place**

Show base as

In the Solr index replace prefix Crawl base with this text. Changes index field **url**

Username Password

Credentials for authentication to a **web source**.

Not before

digits **YYYY-MM-DD** or phrase **n days ago** where **n** is the count of whole days.

Ignore files before this date. Leave empty to omit the date test.

Replace index Yes ☐ No ☒

Yes - empty and replace Solr index. No - add this data to the index

Index files having no filename extension Yes ☐ No ☒

Yes - index files not having a filename extension. No - read but do not include in the index.

Common ajustables are
Index, Crawl base and Replace

In the middle, file filter rules, it's simpler than it looks

* File filter (can edit it here)

+/- accept/reject result upon a pattern match and exit test. PHP regex syntax follows.

^/\$ start/end of line indicators, . (dot) is any char, .+ (dot plus) is 1 or more chars,

(a|b|c) groups OR items, \char treats char as literal, ; (semicolon) starts a comment.

```
; Example filter to select files with common extensions,  
; reject other extensions, allow those with no extension.  
; Normally place exceptions above the main accept(+) line.  
; If no match is found then accept the filename.  
+^.+\. (pdf|ppt|pptx|doc|docx|xls|xlsx|txt) $  
+^.+\. (htm|html|sxw|odp|stw|xml|rtf|odt|ods) $  
-^.+\.+.$
```

The filter applies to the URL component following the Crawl base.
Rules are examined from the top down. The first match ends testing.

Rules are accept(+) or ignore (-) files matching the regular expression.
Note the assistance message at the top.

A default filter, such as this example, can be preset into config.php.

Another example file filter

A filter for Mailman message archives.

Accept html messages and subject.html pages, ignore all else.

Boilerplate Office selections remain intact (saves typing).

```
; Mailman archive material
-(database|attachments|pipermail.pck)
-(author|date|thread)\.html$
-\.txt$
+^\.+(pdf|ppt|pptx|doc|docx|xls|xlsx|txt|htm|html|sxw|odp|stw|xml|rtf|odt|ods)$
-^\.+\.+$
```

A comment (semicolon to end of line)

Several ignore (-) lines, two using OR grouping parens (a | b | c)

Standard Office et al acceptance (+) of what remains

Standard “ignore all other files with an extension”

Solr details are at the page bottom

“O Solr, Solr wherefore art thou?”

Logging ☐ No ☒ Yes

Enable results log.

Solr delay (ms)

← Pace sending files to Solr

Delay (milliseconds) after sending a file to Solr. Suggest 100 or larger.

Crawl depth

← Avoid endless loops

Limit exploration to this many levels below the starting point. Empty defaults to 6.

* Solr address

Address of Solr server. Format is `http://localhost:8983/solr`

Solr username Password

← Optional

Credentials for authentication to Solr.

Solr configset

← “myconf” is the default configset

Solr internal configuration set. Leave empty unless Solr has been configured for it.

Save as

Cancel

A resultant configuration file, <700 bytes

```
solrindex="ttplist"  
description="TTP HE (ttp) list server messges"  
crawlbase="/usr/local/mailman/archives/private/ttpfiles"  
showas="/novttp/ttpmail"  
credentials="myself:secret"  
notbefore="1 day ago"  
replace="no"  
filter[]=""; Mailman archive material"  
filter[]="-(database|attachments|pipermail.pck)"  
filter[]="-(author|date|subject|thread|index)\.html$"  
filter[]="-.+\.txt$"  
filter[]="+^\.+\. (pdf|ppt|pptx|doc|docx|xls|xlsx|txt|htm|html|sxw|odp|stw|xml|rtf|odt|ods)$"  
filter[]="-.+\.+ $"  
logging="yes"  
solrdelay="100"  
depth="6"  
solraddress="http://netlab1.net:8983/solr"  
solrcreds=":"  
solrconfigset=""
```

What the Edit menu provided, with equals signs & quotes for parsing.
The default solrconfigset name “*myconf*” is built into the crawler.

“Choose” existing file shows more



Web and file crawler configuration editor

Choose or create a crawler configuration file

Files in /home/search/index. Click a filename in the list to choose:
k12list.cfg opennovell.cfg provotalk.cfg slctalk.cfg tfolder.cfg **ttplist.cfg**
ttpweb.cfg

Create new file

Choose what to do with file ttplist.cfg

Solr index = ttplist
Description = TTP HE (ttp) list server messges
Crawl base = /usr/local/mailman/archives/private/tpfiles
Show as = /novttp/tpmail
Username = jrd
Not before date = 1 day ago Replace index = no
File filter:

```

; Mailman archive material
- (database|attachments|pipermail.pck)
- (author|date|subject|thread|index) \.html$
- .+\.txt$
+^.\. (pdf|ppt|pptx|doc|docx|xls|xlsx|txt|htm|html|sxw|odp|stw|xml|rtf|odt|ods) $
-^.\.++$

```

Logging = yes
Solr delay (ms) = 100 Crawl depth = 6
Solr address = http://netlab1.net:8983/solr

Edit Run test crawl Visit Solr

Delete config file Permit delete ☐

Delete Solr index Permit delete ☐

Summary

Five actions are available

Edit Run test crawl Visit Solr

Delete config file Permit delete ☐

Delete Solr index Permit delete ☐

Tick boxes are safety measures

A test crawl, watch it work

Web and file crawler configuration editor

Perform test crawl

Configuration file `ttplist.cfg` to Solr index `ttplist`

Execution stops when exiting this page

Return use this, not browser's back button

Configuration file `index/ttplist.cfg`, Solr index `ttplist`

Description: TTP HE (ttp) list server messges

Post `/usr/local/mailman/archives/private/ttpfiles/index.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/subject.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130966.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130967.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130968.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130971.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130972.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130973.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130974.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130975.html`

Post `/usr/local/mailman/archives/private/ttpfiles/2017-February/130976.html`

Crawler internal guidelines

Both file and web crawls:

- Stay within the starting place

- Solr indexing is the slowest link in the chain thus
pace file insertions to avoid overwhelming Solr

File crawl:

- skip files with leading dot, they are hidden on purpose

- skip symbolic links (one copy of a file is enough, thank you)

Web crawl:

- Try to deal with complex web layouts

- Avoid indexing the same URL multiple times

- Accept href's with `../` but stay within bounds

- Accept href's with `blah#foo` but remove **#foo**

Configuration footnote: incrementals

Two controls can work together to support incremental updates:

Not before
digits **YYYY-MM-DD** or phrase **n days ago** where **n** is the count of whole days.
Ignore files before this date. Leave empty to omit the date test.
Replace index Yes ☐ No ☒
Yes - empty and replace Solr index. No - add this data to the index

Yes, **1 day ago** is fine, but **yesterday** is not

Example: in /etc/cron.daily place a file holding these lines to catch daily email

```
#!/bin/bash
```

```
cd /home/search
```

```
php crawler.php index/ttplist.cfg      (uses Not before 1 day ago, Replace No)
```

A handy crawler control, Show base as

* **Crawl base**

Where to crawl: **http(s)://example.com/goodies/** or **/some/local/place**

Show base as

In the Solr index replace prefix Crawl base with this text. Changes index field **url**

103 [\[ttp\] Is there any documentation available on how to PROPERLY upgrade from SLES11SP4](#)

[ttp] Is there any documentation available on how to PROPERLY upgrade from SLES11SP4

Penris, SJ (Bas) rect mailto:ttpfiles%40netlab1.net?Subject=Re%3A%20%5Bttp

%5D%20Is%20there%20any%20documentation%20available%20on%20...

url: https://netlab1.net/novttp/ttpmail/2017-February/130896.html

date: 2017-02-22 00:00:00 UTC **rank:** 25%

Perform a local file system crawl, but show users the results in web browser (https://) form. The change is stored in the index.

Both the URL line and the href= underlying the clickable title use this rewrite.

Crawling nuances

Solr delay (ms)

Delay (milliseconds) after sending a file to Solr. Suggest 100 or larger.

Crawl depth

Limit exploration to this many levels below the starting point. Empty defaults to 6.

Solr delay allows Solr time to do processing after accepting a file from the crawler. It needs time to parse the file and create index material. 100 is good, 200 is better but slightly slower. Check with the web site owner about fast & furious crawls.

Crawl depth is to help break loops: page A points to page B which points to page C which points to A, resulting in repeating path of .../A/B/C/A/B/C/A/B/...

Crawler's command line

Simple form of

```
php crawler.php -h -v config_file
```

where

- h produces just this short help message

- v (verbose) echoes log entries to the screen

config_file is the crawler configuration filename

Example of manual usage :

```
su -
```

(become all powerful root)

```
cd /home/search
```

(where I keep Search apps)

```
php crawler.php index/ttplist.cfg
```

(do the work)

A visit to Solr's admin web page

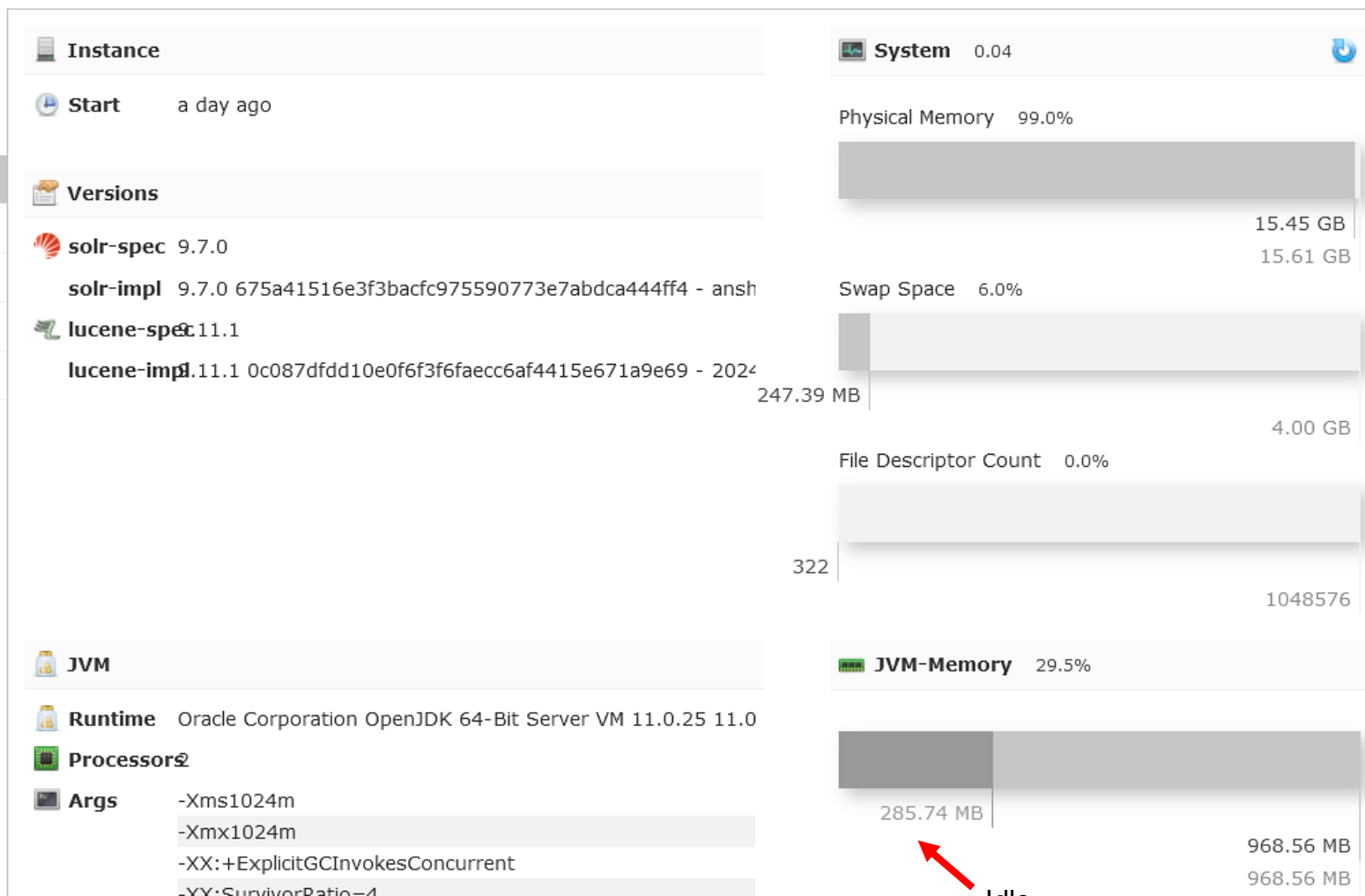


Dashboard

- Logging
- Security
- Core Admin
- Java Properties
- Thread Dump

Core Selector

Lots in this pull down menu



Idle,
Up to 1GB while indexing

Solr refers to “indices” as “cores”

A typical query web page

Search interface for a query web page. The search bar contains the text "OES 24.x". Below the search bar are buttons for "Index", "ttplist", "k12list", "provotalk", "slctalk", "opennovell", and "Novell-FAQ". Below these are buttons for "Sort", "normal", "date", "title", "Mode", "or", "and", "Text", "less", and "more". Below the search bar are buttons for "Search" and "Help". Below the search bar are buttons for "Results: 1 to 20 of 414", "First", "Next", "Previous", and "Last". Below the search bar is a slider control labeled "Choose a starting value:". Below the search bar is a list of search results.

Results: 1 to 20 of 414 [First](#) [Next](#) [Previous](#) [Last](#)

Choose a starting value:

- [\[TTP\] OT machines incorrectly announced end of support for OES 24.x](#)
[TTP] OT machines incorrectly announced end of support for **OES 24.x** [TTP] OT machines incorrectly announced end of support for **OES 24.x** Richard Williams via TTP Forum list@thettp.org Wed Nov 6 12:28:32 GMT 2024 Previous message (by thread): [TTP] OT machines incorrectly announced end of support ...
url: /novttp/ttpmail/2024-November/142766.html (7282 bytes)
date: 2024-11-12 06:38:35+00:00 UTC **rank:** 59%
- [\[TTP\] OT machines incorrectly announced end of support for OES 24.x](#)
[TTP] OT machines incorrectly announced end of support for **OES 24.x** [TTP] OT machines incorrectly announced end of support for **OES 24.x** Girish KS via TTP Forum list@thettp.org Tue Nov 12 06:38:30 GMT 2024 Previous message (by thread): [TTP] OT machines incorrectly announced end of support for ...
url: /novttp/ttpmail/2024-November/142772.html (7519 bytes)
date: 2024-11-12 16:46:45+00:00 UTC **rank:** 59%
- [\[TTP\] OT machines incorrectly announced end of support for OES 24.x](#)
[TTP] OT machines incorrectly announced end of support for **OES 24.x** [TTP] OT machines incorrectly announced end of support for **OES 24.x** Joe Doupnik via TTP Forum list@thettp.org Wed Nov 6 11:12:44 GMT 2024 Previous message (by thread): [TTP] NetIQ Identity Manager force sync Next message ...
url: /novttp/ttpmail/2024-November/142765.html (4432 bytes)
date: 2024-11-11 15:11:39+00:00 UTC **rank:** 59%

Notice multiple indices, plus other controls

Slider for easy navigation

Clickable titles

Highlighting of terms matching the query

Help while in the middle of things

OES 24.x ←

Index

Sort Mode Text

Results: 1 to 20 of 420

Choose a starting value:

Some hints about using this facility:

1. Query terms containing other than just letters or digits may be placed within double quotes so that those other characters do not separate a term into many terms. A dot (period) and white space are neither letter nor digit. Examples: "Now is the time for all good men" (spaces, quotes impose ordering too), "goods.doc" (a dot). Ordinary search terms (not in item 3 below) are case insensitive.
2. Mode button "**or**" (the default) means match one or more terms, perhaps scattered about. Mode button "**and**" means must match all terms, scattered or not.
3. A one word query term may be prefixed by **title:** or **url:** to search on those fields. A space must follow the colon, and the search term is case sensitive. Examples: **url:** .ppt or **title:** Goodies. Many docs do not have a formal internal title field, thus prefix **title:** may not work well.
4. Compound queries can be built by joining terms with **AND OR && ||** and group items with (|). Not is expressed as a minus sign (-) prefixing a term; prefix + means must have. A bare space means use the current Mode (or, and). Example: Nancy AND Mary AND -Jane AND -(Robert|Daniel) which means both the first two and not Jane and neither of the two guys.
5. A query of asterisk/star (*) means match everything (zero or more characters). A query mark ? matches one character, backslash \ escapes the next character. Examples: * for everything (zero or more characters). Fussy, show all without term .pdf -.pdf" or -\.pdf

1 [\[TTP\] OT machines incorrectly announced end of support for OES 24.x](#)

[TTP] OT machines incorrectly announced end of support for **OES 24.x** [TTP] OT machines incorrectly announced end of support for **OES 24.x** Richard Williams via TTP Forum list@thettp.org Wed Nov 6 12:28:32 GMT 2024 Previous message (by thread): [TTP] OT machines incorrectly announced end of support ...

url: /novttp/tpmail/2024-November/142766.html (7282 bytes)

date: 2024-11-12 06:38:35+00:00 UTC **rank:** 59%

The Help button works without disturbing other settings.

The content discusses query syntax expressions.
Good quiz material.

Typos: did you mean?



TTP emeb Regansburg

Index

Sort Mode Text

Results: 1 to 20 of 48163

Choose a starting value:

Suggested query spelling changes:

emeb did you mean

regansburg did you mean

1 [\[ttp\] OT: Moving to MS Storage, opinions ?](#)

[ttp] OT: Moving to MS Storage, opinions ? **[ttp]** OT: Moving to MS Storage, opinions ?
 Turner, Sean Sean.Turner@wbs.ac.uk Mon Oct 5 22:14:28 BST 2015 Previous message
 (by thread): **[ttp]** OT: Moving to MS Storage, opinions ? Next message (by thread): **[ttp]**
 OT: Moving to MS Storage, opinions ...
 url: /novttp/tpmail/2015-October/128090.html (27476 bytes)
 date: 2024-11-04 10:09:19+00:00 UTC rank: 10%

Clicking an offering yields

a new search

The number of alternative spellings is adjustable in the query program file.

Change words are selected from the current index.

TTP emeb regensburg

Index

Sort Mode Text

Results: 1 to 20 of 48569

Choose a starting value:

Suggested query spelling changes:

emeb did you mean

1 [\[novttp\] Betr.: TTP EMEA Meeting 2011 Regensburg](#)

[novttp] Betr.: **TTP** EMEA Meeting 2011 **Regensburg** [novttp] Betr.: **TTP** EMEA Meeting 2011 **Regensburg** Bas Penris b.penris@ettyhillesumlyceum.nl Tue May 10 18:51:43 BST 2011 Previous message (by thread): [novttp] Betr.: **TTP** EMEA Meeting 2011 **Regensburg** Next message (by thread): [novttp] ...
 url: /novttp/tpmail/2011-May/096880.html (4286 bytes)
 date: 2024-11-04 10:00:22+00:00 UTC rank: 57%

Some points about this user interface

- Controls remain on-screen, at the top, all the time
- Scrolling is internal to the result display list
- Index choice and paging controls are visible only if multiple choices
- Multiple indices may be queried, one by one, just by clicking a button
- Switching indices does not modify query settings
- Help is a toggled visual chunk, not upsetting searches
- The visual design works well with both mobile phones and desktops
- Search syntax is roughly Google-like, with button control of search term combinations (OR, AND) plus using “a quote string”
- The design is people, network and device friendly, with crisp responses.
- A button press does a new search, ~10 packets to repaint the screen

A nuance about Query's slider control

Some browsers by default may not support the slider



One reason, in Firefox's *about:config* screen, is that
 javascript.enabled=true
 security.csp.enable=false
need to be set as shown here.

The Query web page uses several lines of simple internal
(no external files) javascript to provide the slider.
(This example has only one index, thus no index buttons)

Query, a manager's adjustables

The manager selects which indices this particular query program file offers and show them in our desired order.

```
$indexlist = array("ttplist", "k12list", "opennovell");
```

Different program copies for different audiences. Diversity.

This is also a security feature. Access to the web page is applied outside of the program (Apache etc). The set of allowed indices is fixed inside the program (user tinker proof).

Each index result also can have its reported URL prefix line be rewritten by manager-selected text (original, replacement):

```
$trans["ttpfiles"] =  
    array("/home/user/novttp/files", "https://netlab1.net/novttp/files");
```

Query invoked from other web pages

The query program can be invoked with URL arguments of the form
?q=a_question&index=where_to_look

Such as this item in a web page:

```
<a href="query5.php?q=chocolates&index=munchies">yummy</a>
```

where q= a query, and index= which index to explore of those configured into the program.

If no index= then use the first displayable index

This permits another web program's Search box to invoke Query in a controlled manner. Name the Query file to be whatever you wish.

The *myconf* Solr engine indexing configset directory

A configset defines parsing rules etc when creating an index. It is a bundle of complex files to control Solr's indexing. Tricky stuff.

This distribution defaults to using my configset *myconf*, which is a near copy of Solr example *techproducts* configset. More sets are allowed.

Two particular files in myconf are modifications: *managed-schema.xml* and *solrconfig.xml*.

In *managed-schema.xml* are lines added to define “filedate” and “filesize” attributes.

solrconfig.xml has several changes which add spell checking and enable three RequestProcess workers “dedupe”, “langid”, and “script”.

Look for my initials JRD near changes in both files.

Ensure files in /home/search/solr are owned by solr:solr

Crawler timing performance results

Simple HTML files, such as HTML email, index about 5-10 docs/sec

Posted 71860 documents in 152.42 minutes

Office and PDF files, index about 1-2 docs/sec (pace slowly please)

Posted 3953 documents in 22.60 minutes

Indexing a 130K docs archive of html email takes about six hours

Posted 131339 documents in 379.52 minutes

Indexing yesterday's additions to it (incremental) takes ~30 seconds

Posted 8 documents in 0.47 minutes (examine all, index new)

Often an incremental addition is even quicker

Posted 24 documents in 0.17 minutes (same index as above)

A large archive done as a single crawl

Posted 644621 documents in 1,963.70 minutes

Clicking a Query button: response time is a small fraction of a second

Resources

Config web program: 28KB of PHP, writes ~650 byte config files to subdir “index”, can invoke crawler for a test run, can invoke Solr admin web page, etc. A one stop shop.

Crawler program: 40KB of PHP, writes to subdirs “logs” & “files”.

Uses 15-50MB memory while crawling.

Requires a control filename on its command line to perform its tasks.

Supports authentication to web input sources and to Solr.

Paces files into Solr to limit large peak memory and CPU usage.

Query web program: 32KB of PHP, no disk files.

Talks to Solr via the network, supports login to Solr.

Solr engine from Apache.org: 310MB of complex Java, <1GB memory

Locating indices outside of Solr

If the index-holding directory is not within Solr, unlike my examples, then ensure its location is stated correctly in file **/etc/default/solr.in.sh**, property SOLR_HOME=.

User *solr:solr* will need to own that new directory.

A configsets directory needs to be adjacent to the indices.

That area receives many changes due to index creation so it would be best located in a Posix file system rather than encounter NSS Salvage operations.

Allow for growth. Example: an email archive directory, 2.6GB source size, resulted in a 420MB index.

Testing Solr start

Start service solr by

`cd /etc/init.d` and then say `./solr start`

or for systemd: `systemctl daemon-reload`

`systemctl enable solr`

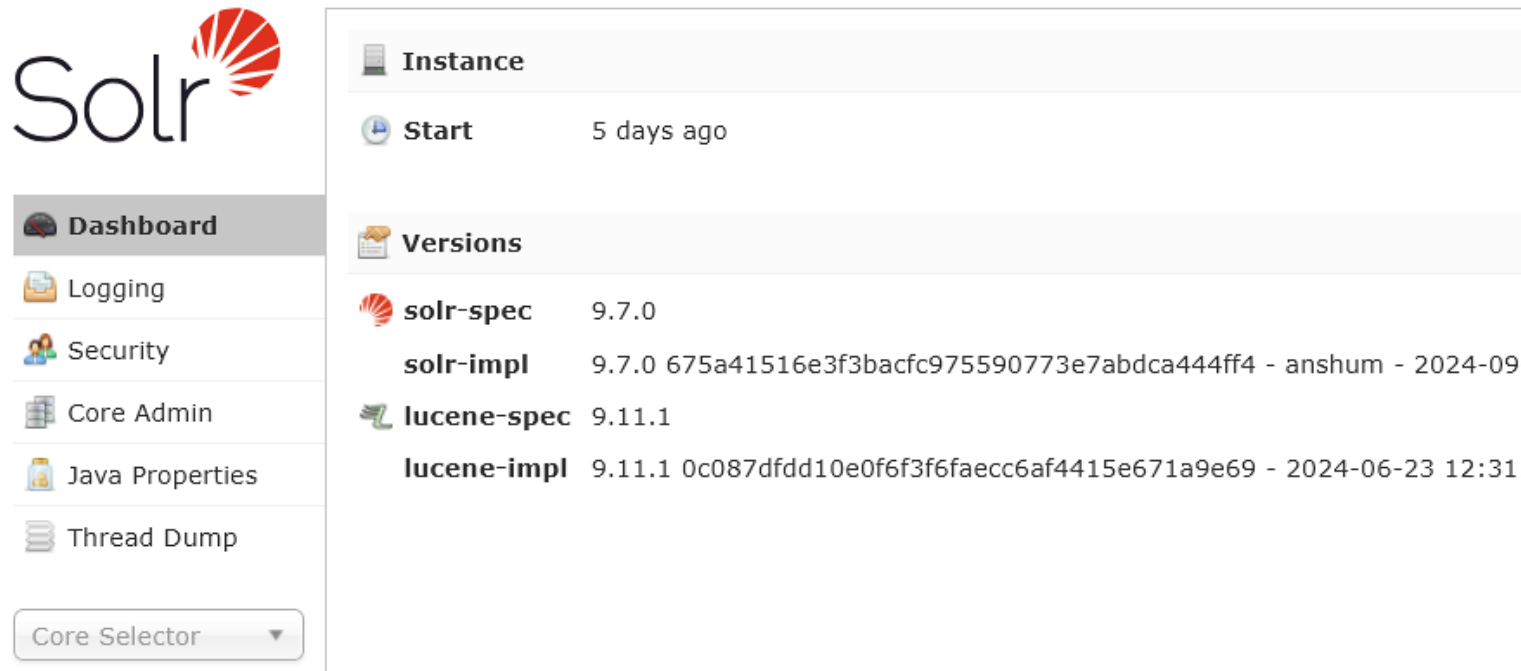
`systemctl start solr`

update systemd with our script

enable Solr start at boot time

Observe a simple startup response

Double check by visiting Solr's admin web page, at <http://localhost:8983/solr>



The screenshot displays the Solr Admin web interface. On the left is a sidebar with navigation links: Dashboard (selected), Logging, Security, Core Admin, Java Properties, and Thread Dump. Below these is a 'Core Selector' dropdown menu. The main content area is divided into two sections. The 'Instance' section shows a 'Start' button and the text '5 days ago'. The 'Versions' section lists the installed components and their versions:

Component	Version
solr-spec	9.7.0
solr-impl	9.7.0 675a41516e3f3bacfc975590773e7abdca444ff4 - anshum - 2024-09
lucene-spec	9.11.1
lucene-impl	9.11.1 0c087dfdd10e0f6f3f6faecc6af4415e671a9e69 - 2024-06-23 12:31



Failure to start Solr

Six common causes of failure to start:

1. Permissions not correctly set on Solr (/home/search/solr and /home/search/tmp), all of the contents thereof
2. Ownership of solr and tmp not assigned to the Solr user
Ownership needs to be `solr:solr`
3. Solr version has changed its schema requirements (yet again, sigh)
4. Java JDK not pointed to correctly. Use Java 11 with Solr v9
5. Typos, of course
6. You are not root. Please find a grownup to help.

Now for the fun part, web query & crawler

- These are PHP v8.x programs. The code has many comments to help us when modifying the material to suit local requirements.
- Many PHP modules are used (but not the database items). Best is we install most PHP v8 modules. See following slides about supporting PHP with Apache web server.
- Web PHP apps may show a white screen if a module is missing. View the Apache error log for the cause (typically /var/log/apache2/error_log). Add the missing PHP module(s).
- The query and crawler's config program are written in PHP v8 and are run by a web server.
Crawler (PHP v8) and Solr (Java) are stand alone executable programs.
We may run crawler manually to do some updates.

Query program internal adjustables

\$solrhost = "http://localhost:8983/solr";	How to reach Solr/Lucene's jetty
\$solrusername = "myself";	Use if Solr access is credential protected
\$solrpassword = "secret";	
\$pagesize = 20;	Number of responses in each query web page
\$spellcheck.count = 5;	Number of "did you mean" results
\$indexlist = array();	List of which indices (<u>empty means all</u>), and their order, which this particular program copy will touch.

Example: **\$indexlist = array("sales", "marketing", "pricing");**

This is an important localization and security feature.

\$trans["index"] = array("from", "to"); Modify the URL prefix ("from") of each query's result to be URL prefix ("to") for index "index". A dynamic localization option.

Indices are not changed, just the program's returned results. Thus a query program may do post-indexing adjustment of the displayed URLs.

PHP with Apache MPM *worker/event*

If using Apache MPM *worker* or *event*, not default *pre-fork*, then employ module **mod_fcgid** or standalone program **php_fpm**, and do not load mod_php.

The mod_fcgid approach (which I prefer with MPM worker, four steps) –

1. Obtain the module from https://httpd.apache.org/mod_fcgid
2. Build mod_fcgid, as shown in its distribution file README-FCGID
- 3 . In file /etc/sysconfig/apache2, add fcgid (omit php, put ssl near start).

Apache details in file /etc/sysconfig/apache2 in my OES systems:

```
APACHE_MODULES="actions alias ssl version auth_basic authn_core  
blah blah blah ...  
speling status substitute suexec unique_id userdir fcgid wsgi"
```

Prefork creates a new program image for each connection. Expensive.
Worker is a threaded approach, one image then a small memory area for each connection.

mod_fcgid approach, cont'd

4. My file /etc/apache2/conf.d/mod_fcgid.conf, at the bottom:
(I have commented out the original FilesMatch clause)

```
PHP_Fix_Pathinfo_Enable 1
<FilesMatch "\.php$">
    Options +ExecCGI
    AddHandler fcgid-script .php
    FCGIWrapper /srv/www/php-fcgi-scripts/php-fcgi-starter .php
</FilesMatch>
##JRD addition, do not buffer output
    OutputBufferSize 0
    MaxRequestLen 40000000
    MaxRequestsPerProcess -1
    BusyTimeout 86400
    IdleScanInterval 3
</IfModule>
# End of <IfModule fcgid_module>
```

This configures a CGI wrapper for .php files with settings to permit smooth operation of the config.php program and others.

Shielding Solr via a web proxy server

We can use a web server (say Apache) to shield Solr from unauthorized contacts. Below is an example configuration:

```
<Location "/solr">                                # Solr engine access is via URI /solr
<RequireAll>                                       # For multiple checking criteria
    Require IP 1.2.3.4/24 127.0.0.1 # First criteria: be in the IP list
    other Require conditions to suit local requirements
</RequireAll>                                     # end of Who are you? checking
ProxyPass http://localhost:8983/solr             # proxy to Solr's jetty interface
ProxyPassReverse http://localhost:8983/solr # rewrite the responses
</Location>
```

Important: Config and Query use this proxy's address plus optional credentials.

Thus edit `$solraddress` to be `https://ourproxybox.net/solr`, optionally with credentials for username: `$solrusername` and `$solrpassword`

→ **Protect Apache proxy** from the bad guys. See file [SSL/TLS for system admins](#), slide 26, on <https://netlab1.net>. Proxing needs that protection.

Apache web server modules `proxy_*` are required.

Sample Apache conf files for general access to Config and Query programs

Alias /config /home/search/config.php

Alias /query /home/search/query.php

I place these apps in

/home/search on my gear

<Directory /home/search>

Require ip 11.22.33.0/24

other Require clauses as you prefer

</Directory>

Gives Apache permission to enter

Tickets please...

See <https://netlab1.net/longterm/Configure%20Apache.pdf>
and Apache web server docs about “Require”

Summary

This project is designed to be used and run by independent groups.

Crawlers and indices here&there, many query agents tailored for specific usage situations. Local groups can control indexing their way.

Complexity of normal use has been minimized. Ordinary users need no special training to do indexing, just some introductory information.

A working system can be built within 30 minutes or so.

Building a number of them for a site is reasonable.

Tailoring is simple yet it and building are best done by IT staff.

This becomes a service broadly available for a variety of local groups.

We can easily modify the apps to be what we want them to be.

It is free, open source, and our valuables are kept safely local to our site.

Our good deed, done. Be proactive, share this project with colleagues.

Appendix: obtaining the project's materials

The project's offerings are within one bundle held on

<https://netlab1.net/long-term/SearchService/>

→ **Distribution bundle** [SearchService3.tar.gz](#) (~8MB) has items:

Open Horizons [article](#) from 2019

[SearchService2.pdf](#) - overall description

[Installing Search Service v2.1.pdf](#) - illustrated installation instructions

[Installing Search Service v2.1-updated.pdf](#) - updated for Solr v9

[Installing Search Service v2.2.pdf](#) - updated further, Aug 2024

and the new v3 files below:

[Search Service v3.pdf](#) which is this presentation

mybundle-Solrv9 = building blocks of the v3 service for Solr v9.7.0,

holding config.php, crawler.php, query.php, and files

[Installing when using Solr v9.7.0.txt](#), schema, control et al.

The free Solr engine solr-9.7.0.tgz, 283MB, is available from <https://solr.apache.org>



MindWorks Inc. Ltd
210 Burnley Road
Weir
Bacup
OL13 8QE UK

Telephone: +44 (0) 170 687 1900
Fax: +44 (0) 170 687 8203
Web: www.mindworksuk.com
Email: training@mindworksuk.com