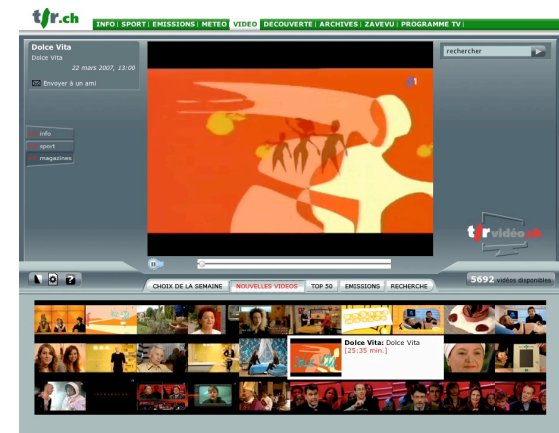


Beyond full-text searches With Lucene and Solr

Bertrand Delacrétaz
ApacheCon EU 2007, Amsterdam
bdelacretaz@apache.org
www.codeconsult.ch

slides revision: 2007-05-03



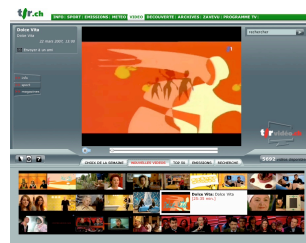
Slides at

<http://wiki.apache.org/apachecon/Eu2007OnlineSessionSlides>

How to graft a Lucene-based
dynamic navigation system
on an search-challenged CMS
using Solr.

As seen from the “Solr integrator” point of
view.

Beyond full-text?



Solr

tsrvideo.ch - a Solr client

The screenshot displays the website interface for tsrvideo.ch. At the top, the logo 'tsr.ch' is visible, followed by a navigation bar with links: INFO | SPORT | EMISSIONS | METEO | VIDEO | DECOUVERTE | ARCHIVES | ZAVEVU | PROGRAMME TV. The main content area features a video player for 'Dolce Vita', which is currently paused. The video player includes a search bar labeled 'rechercher' and a 'tsrvidéo.ch' logo. Below the video player, there are navigation tabs: CHOIX DE LA SEMAINE, NOUVELLES VIDEOS (highlighted), TOP 50, EMISSIONS, and RECHERCHE. A button indicates '5692 vidéos disponibles'. The bottom section shows a grid of video thumbnails, with the selected video 'Dolce Vita: Dolce Vita [25:35 min.]' highlighted in the center.

tsr.ch INFO | SPORT | EMISSIONS | METEO | VIDEO | DECOUVERTE | ARCHIVES | ZAVEVU | PROGRAMME TV |

Dolce Vita
Dolce Vita
22 mars 2007, 13:00
✉ Envoyer à un ami

rechercher

info
sport
magazines

tsrvidéo.ch

CHOIX DE LA SEMAINE | **NOUVELLES VIDEOS** | TOP 50 | EMISSIONS | RECHERCHE | 5692 vidéos disponibles

Dolce Vita: Dolce Vita
[25:35 min.]

tsrvideo.ch - a Solr client

tsrvideo.ch interface showing a search bar (rechercher) and navigation options (info, sport, magazines). The main content area displays a grid of video thumbnails with titles and descriptions:

- Régimes, balances**
Evitez les pièges!
[11:44 min.]
A Bon Entendeur
- Dolce Vita**
Le meilleur de la semaine [50:39 min.]
Dolce Vita
- AI**
Economiser ou démanteler?
[01:32 min.]
Infrarouge
- Euro 2008**
L'engagement financier de la Suisse
[07:19 min.]
Classe éco
- Entretien**
Jean-Claude Carrière
[25:24 min.]
Pardonnez-moi
- Trafic**
Le GPS nargue les radars [04:55 min.]
Nouvo

Additional interface elements include: "Envoyer à un ami" button, "6205 vidéos disponibles" indicator, and navigation tabs: CHOIX DE LA SEMAINE, NOUVELLES VIDEOS, TOP 50, EMISSIONS, RECHERCHE.

The Project

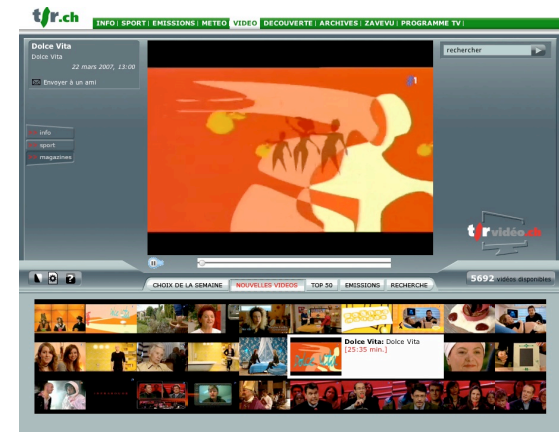
Deliver a rich video player experience

Users explore much more than they search

Existing content stored in two separate CMSes
with very different content models
(and http/XML interfaces)

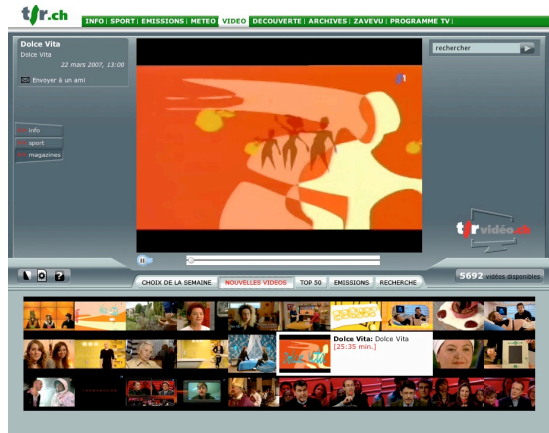
Solr

Lucene



Client system overview

Ajax + HTML



HTTP/JSON

Apache
HTTP server

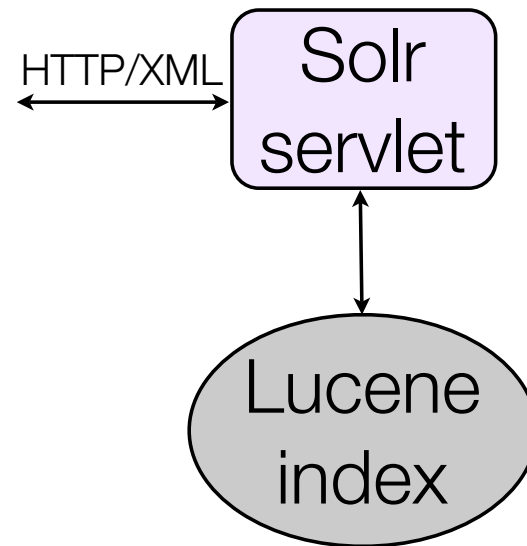
Solr
Search server

Lucene
index

replicated index
for backup

the Solr search server

What is Solr?



See also <http://wiki.apache.org/apachecon/Eu2007OnlineSessionSlides>

Solr architecture

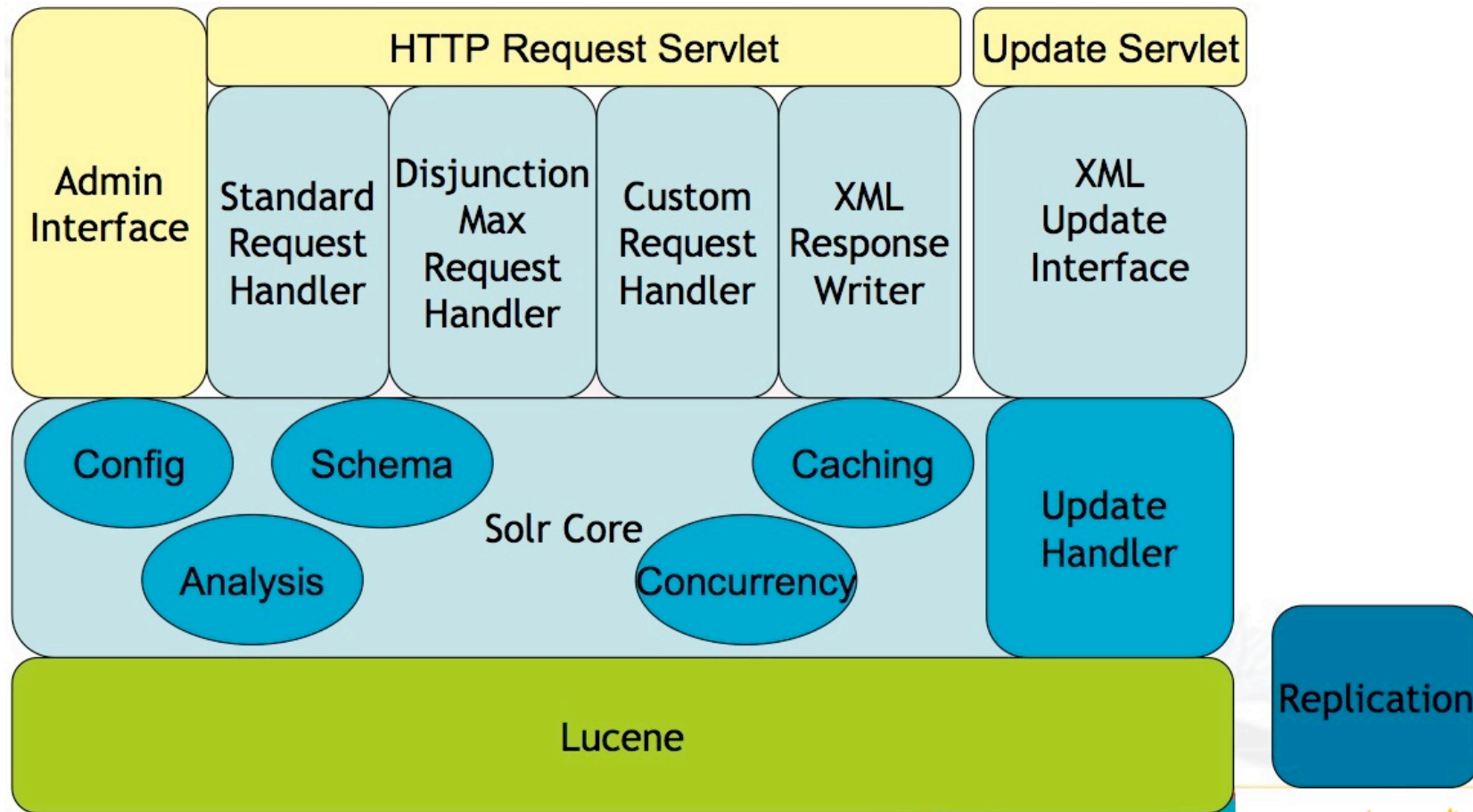


Diagram by Yonik Seeley



Indexing in Solr

```
<add>
  <doc>
    <field name="id">9885A004</field>
    <field name="name">Canon PowerShot SD500</field>
    <field name="category">camera</field>
    <field name="features">3x optical zoom</field>
    <field name="features">aluminum case</field>
    <field name="weight">6.4</field>
    <field name="price">329.95</field>
  </doc>
</add>
```

→ HTTP POST

“Solr XML” documents are POSTed to Solr via HTTP
Field names and types are defined in the Solr schema

Solr indexing schema

```
<field name="id" type="string" indexed="true" stored="true"/>
```

```
<field name="category" type="text_ws" indexed="true" stored="true"/>
```

```
<dynamicField name="*_tws" type="text_ws" indexed="true" stored="true"/>
```

```
<dynamicField name="*_dt" type="date" indexed="true" stored="true"/>
```

```
<fieldtype name="sfloat"
```

```
  class="Solr.SortableFloatField" sortMissingLast="true"/>
```

```
<uniqueKey>id</uniqueKey>
```

```
<copyField source="cat" dest="text"/>
```

```
<copyField source="name" dest="text"/>
```

```
<copyField source="name" dest="nameSort"/>
```

```
<copyField source="manu" dest="text"/>
```

Field content analysis

```
<fieldtype name="text_fr" class="Solr.TextField">
  <analyzer>
    <tokenizer
      class="Solr.StandardTokenizerFactory"/>
    <filter
      class="Solr.ISOLatin1AccentFilterFactory"/>
    <filter
      class="Solr.LowerCaseFilterFactory"/>
    <filter
      class="Solr.StopFilterFactory"
      words="french-stopwords.txt"
      ignoreCase="true"/>
    <filter
      class="Solr.SnowballPorterFilterFactory"
      language="French"/>
  </analyzer>
</fieldtype>
```

Le Châtelain
et ses chevaux



chatelain
cheval

Solr Field Analysis test page

Field Analysis

Field name	abstract
Field value (Index)	Le <u>Châtelain</u> et <u>ses</u> <u>chevaux</u>
verbose output <input type="checkbox"/>	
highlight matches <input checked="" type="checkbox"/>	
Field value (Query)	<u>une</u> <u>belle</u> <u>chatelaine</u> <u>à</u> <u>chevals</u>
verbose output <input type="checkbox"/>	
<input type="button" value="Analyze"/>	

Index Analyzer

le	châtelain	et	ses	chevaux
le	chatelain	et	ses	chevaux
le	chatelain	et	ses	chevaux
le	chatelain	et	ses	chevaux
chatelain	chevaux			
chatelain	cheval			
chatelain	cheval			

Query Analyzer

une	belle	chatelaine	à	chevals
une	belle	chatelaine	a	chevals
une	belle	chatelaine	a	chevals
une	belle	chatelaine	a	chevals
belle	chatelaine	chevals		
bel	chatelain	cheval		
bel	chatelain	cheval		

Index Analyzer

le	châtelain	et	ses	chevaux
le	chatelain	et	ses	chevaux
le	chatelain	et	ses	chevaux
le	chatelain	et	ses	chevaux
chatelain	chevaux			
chatelain	cheval			
chatelain	cheval			

Query Analyzer

une	belle	chatelaine	à	chevals
une	belle	chatelaine	a	chevals
une	belle	chatelaine	a	chevals
une	belle	chatelaine	a	chevals
belle	chatelaine	chevals		
bel	chatelain	cheval		
bel	chatelain	cheval		

Solr queries

http://solr.xy.com/select?q=apache & fl=solr_id,title

```
<result numFound="2" start="0">
  <doc>
    <str name="solr_id">tsr.ch/story/4336075</str>
    <str name="title">ApacheCon Amsterdam</str>
  </doc>
  <doc>
    <str name="solr_id">tsr.ch/story/1715414</str>
    <str name="title">Geeks are upon us</str>
  </doc>
</result>
```

Enhanced Lucene query
language as standard

The Solr logo is rendered in a bold, orange, sans-serif font.

Play it again, JSON!

http://solr.xy.com/select?q=apache&fl=solr_id,title&wt=json

```
{"response":  
  {"numFound":54,"start":0,  
   "docs":[  
     {"solr_id":"tsr.ch/story/4336075",  
      "title":"ApacheCon Amsterdam"  
     },  
  
     {"solr_id":"tsr.ch/story/4336032",  
      "title":"Geeks are upon us"  
     },  
   ]  
  }
```

...

Solr live statistics

Category	[CORE] [CACHE] [QUERY] [UPDATE] [OTHER]
	Current Time: Thu May 03 09:50:31 CEST 2007
	Server Start Time: Tue May 01 15:37:32 CEST 2007
CORE	
name:	Searcher@122cbaf main
class:	org.apache.solr.search.SolrIndexSearcher
version:	1.0
description:	index searcher
stats:	caching : true numDocs : 166391 maxDoc : 166483 readerImpl : MultiReader readerDir :
QUERY HANDLERS	
name:	standard
class:	org.apache.solr.request.StandardRequestHandler
version:	1.0
description:	The standard Solr request handler
stats:	requests : 79959 errors : 1602
name:	dismax
class:	org.apache.solr.request.DisMaxRequestHandler
version:	\$Revision:\$
description:	DisjunctionMax Request Handler: Does relevancy based on combination of fields using configured boosts
stats:	requests : 0 errors : 0

Solr Function Query

A Function query influences the score by a function of a field's numeric value or ordinal.

```
// OrdFieldSource  
ord(myfield)
```

```
// ReverseOrdFieldSource  
rord(myfield)
```

```
// LinearFloatFunction on numeric field value  
linear(myfield,1,2)
```

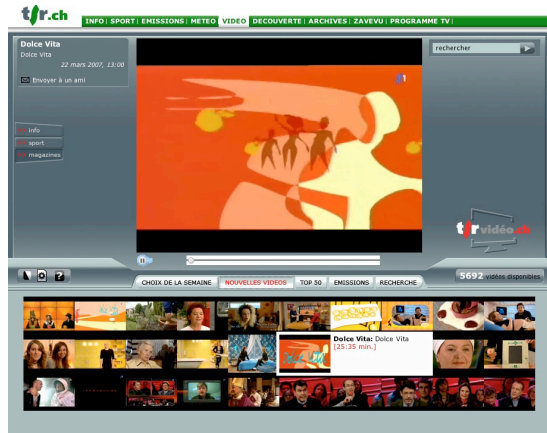
```
// MaxFloatFunction of LinearFloatFunction on numeric field value or constant  
max(linear(myfield,1,2),100)
```

```
// ReciprocalFloatFunction on numeric field value  
recip(myfield,1,2,3)
```

```
_val_:"linear(recip(rord(broadcast_date),1,1000,1000),11,0)"
```

That's our client

Ajax + HTML



HTTP/JSON

Apache
HTTP server

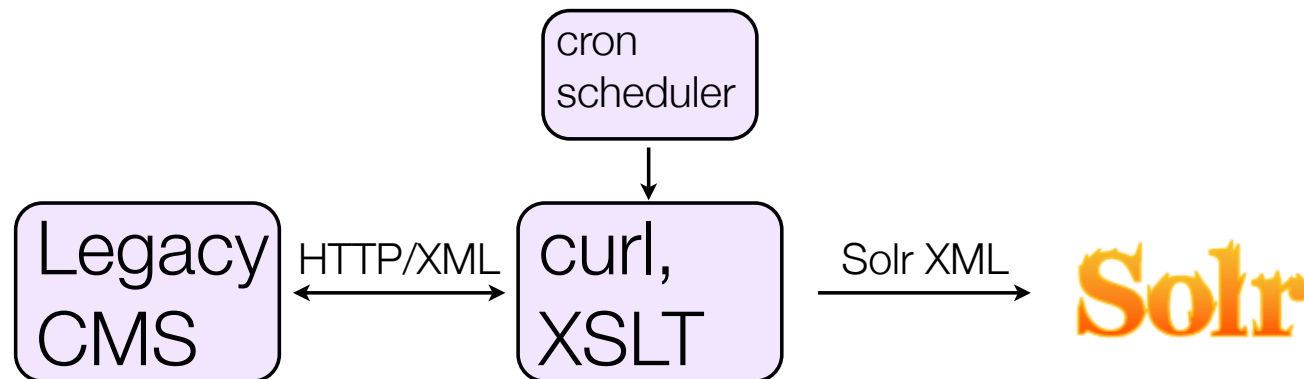
Solr
Search server

Solr schema
and analyzers

Lucene
index

Indexing

Indexing Process



Issues:

Change/delete signals?

Polling? RSS feeds?

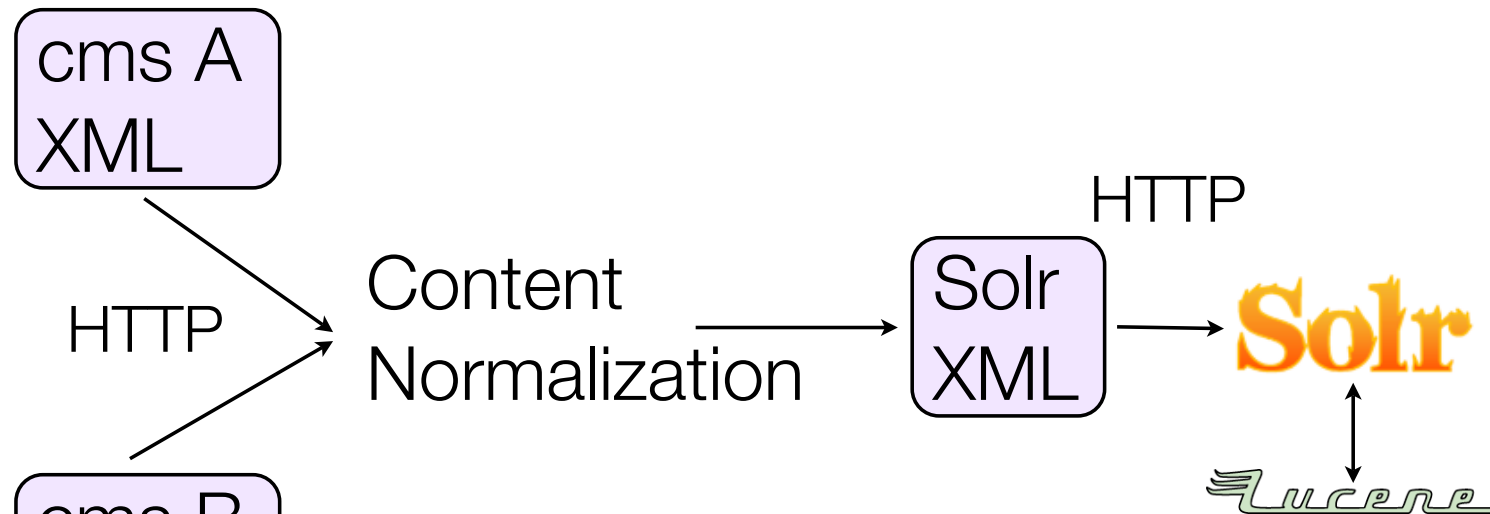
Legacy content structure and consistency

Indexing delay

Deleted/retired documents

Non-transactional behavior

Content Normalization



Convert to "Solr XML".
Common field names.
Normalized values.
-> unified acces



Normalized and unified values

<field name="solr.id">story.cmsA.12129</field>

<field name="role">story</field>

<field name="topic">news</field>

<field name="topic">sports</field>

<field name="author">Bob S. Ponge</field>

<field name="author.id">person.438</field>

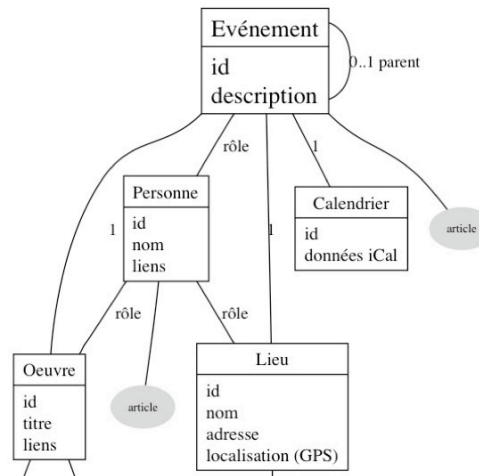
<field name="link.related">story.cmsB.73-1</field>



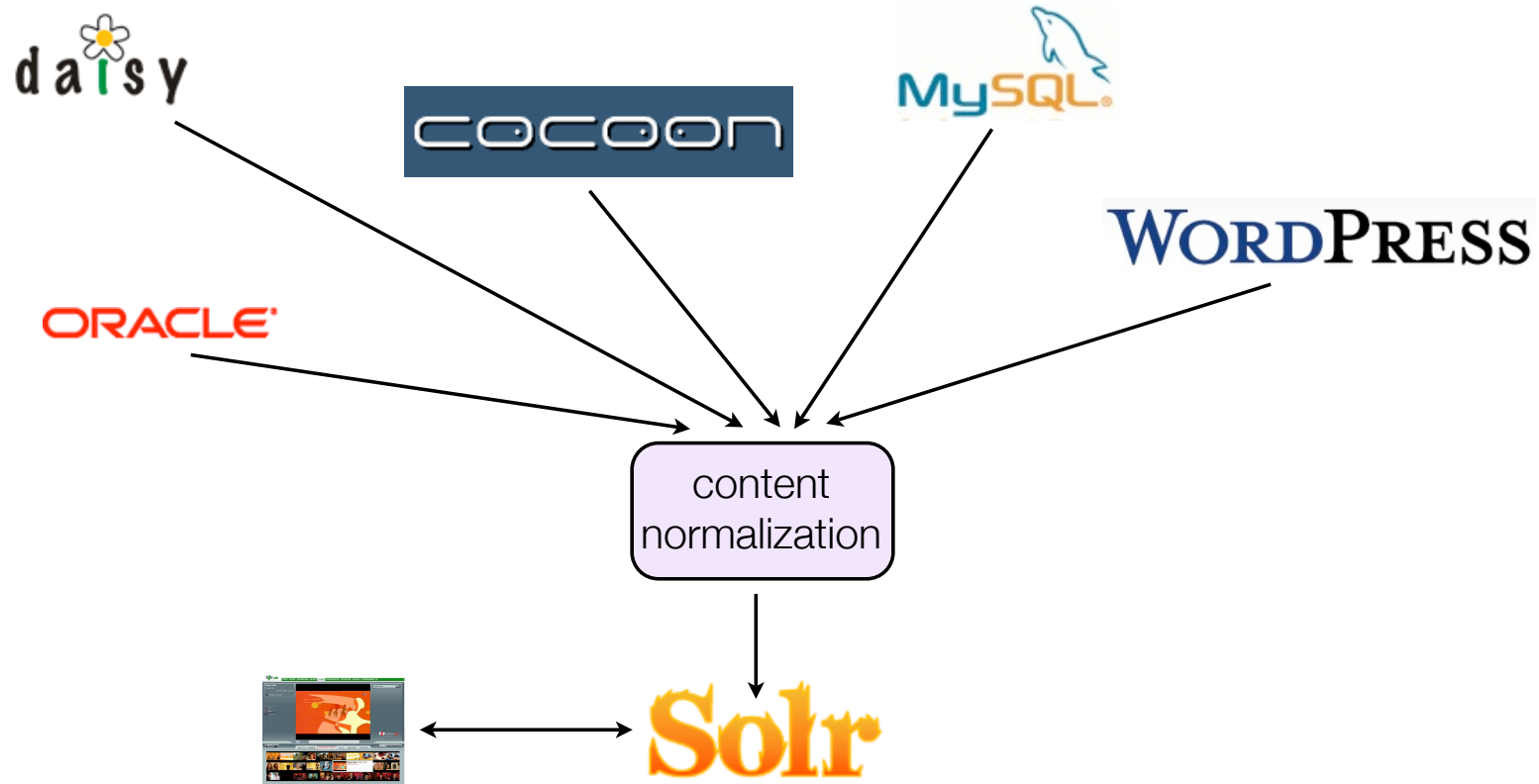
More than “just” full-text searches

<field name="author.id">person.438</field>

<field name="link.related">story.cmsB.73-1</field>



Content Mining?



Unified navigation
and queries

Testing

“How do I break this thing before it breaks by itself?”

Use-cases based testing

Do I find “cheval” when searching for “chevaux”?

Is document 98.345 found when searching for “+montreux -casino AND role:story”?

etc...

Reference data required for such tests: Solr indexes are collection of files that can easily be saved

Why not automate these? read on...



Automated functional testing

Scenarii are executed by our auto-test tool, based
on htmlunit (<http://htmlunit.sourceforge.net/>)

test a query that returns no results

request : /solr/select?wt=xslt&q=thismustnotbefound

match : /response/result/@numFound : 0

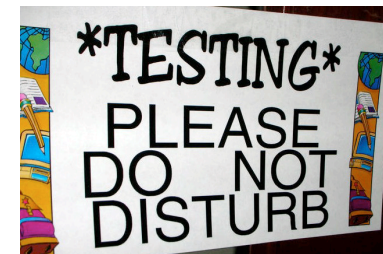
dontMatch : /response/result/@numFound : 1

test a title query

request : /solr/select?q=title%3Afootball

match : contains(/response//doc[1]/str[@name='title'],'ootball') : true

Test are run as JUnit tests, against a Solr instance.



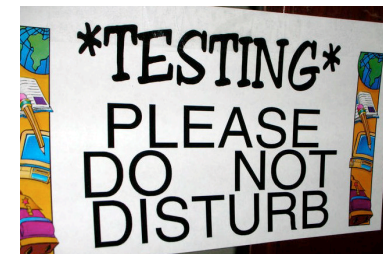
Stress tests

Generate heaps of semi-random query URLs, and replay them in many HTTP clients simultaneously using httpstone

HttpStone performance test utility		
Written by Bertrand Delacretaz, bdelacretaz@codeconsult.ch, see http://code.google.com/p/httpstone/ Each of the workers shown below makes requests continuously to a given URL. If the page is not retrieved within a configurable time, an error report is generated. The colors of the Workers titles indicate the result of the last request. See file config/httpstone-default.properties for configuration info.		
Worker #1 Google homepage (max 150 msec) 87/2973/290 (min/max/avg msec) 23/17/6/0 (total requests/ok/late/failure) waiting 1439 msec for next iteration	Worker #2 Yahoo homepage (max 150 msec) 418/3125/650 (min/max/avg msec) 19/0/19/0 (total requests/ok/late/failure) waiting 3180 msec for next iteration	Worker #3 Bertrand's weblog (max 200 msec) 80/660/224 (min/max/avg msec) 14/9/5/0 (total requests/ok/late/failure) Retrieving URL...
Worker #4 Bertrand's weblog, RSS feed (max 250 msec) 111/441/189 (min/max/avg msec) 16/13/3/0 (total requests/ok/late/failure) waiting 3141 msec for next iteration		

<http://code.google.com/p/httpstone/>

[http://solr...&q="attirer" role:audio "enfants" "fidéliser"](http://solr...&q=)
[http://solr...&q="fidéliser" "carottes" role:story "enfants..."](http://solr...&q=)
[http://solr...&q="surtout" "adultes" "histoire" "L'avis"](http://solr...&q=)
[http://solr...&q="Résultats" "enfants..." "publicité" role:video](http://solr...&q=)
[http://solr...&q="lunettes" "différences?" "Résultats" "fabrications,"](http://solr...&q=)
[http://solr...&q="attirer" role:story "solaires:" "rend-t-on"](http://solr...&q=)
[http://solr...&q=role:audio "quelles" "Mêmes" "Mêmes"](http://solr...&q=role:audio)



Test outcomes

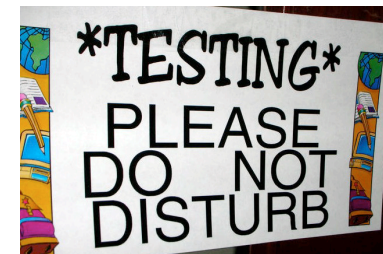
Explain search features with use cases

Avoid regressions with automated tests

Tune the index and analyzers with automated functional tests

Get a feel for scalability with stress tests

Build confidence before launches!



Lessons Learned

Lessons Learned (a.k.a “conclusion”)

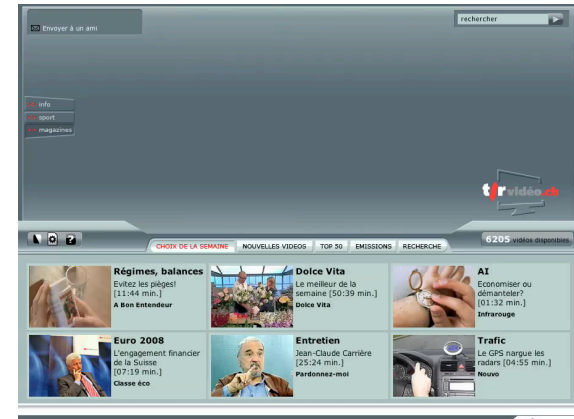
Solr opens the doors to Lucene!

Designing the “right” indexing content model takes time.

Do not hesitate to duplicate fields with different indexing parameters, denormalized, aggregated, etc.

Content unification enables “content mining”.

Tune and run automated tests.
Repeat. Repeat. Repeat...



↓
Solr

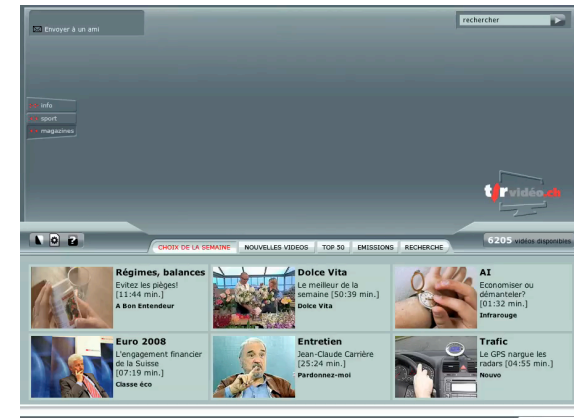
References

<http://lucene.apache.org/solr>

<http://wiki.apache.org/solr/SolrResources>

<http://lucene.apache.org/java>

<http://lucenebook.com/>



“Modern Information Retrieval”, Ricardo Baeza-Yates

<http://www.ischool.berkeley.edu/~hearst/irbook/>

Other ApacheCon EU 2007 presentations:

<http://wiki.apache.org/apachecon/Eu2007OnlineSessionSlides>