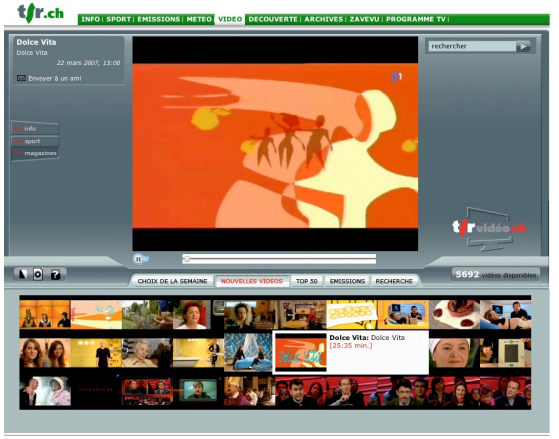
Beyond full-text searches With Lucene and Solr

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

slides revision: 2007-05-03



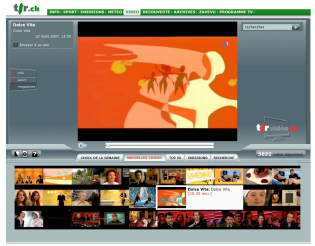
ApacheCon
leading the wave of open source

Slides at
http://wiki.apache.org/apachecon/Eu2007OnlineSessionSlides

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

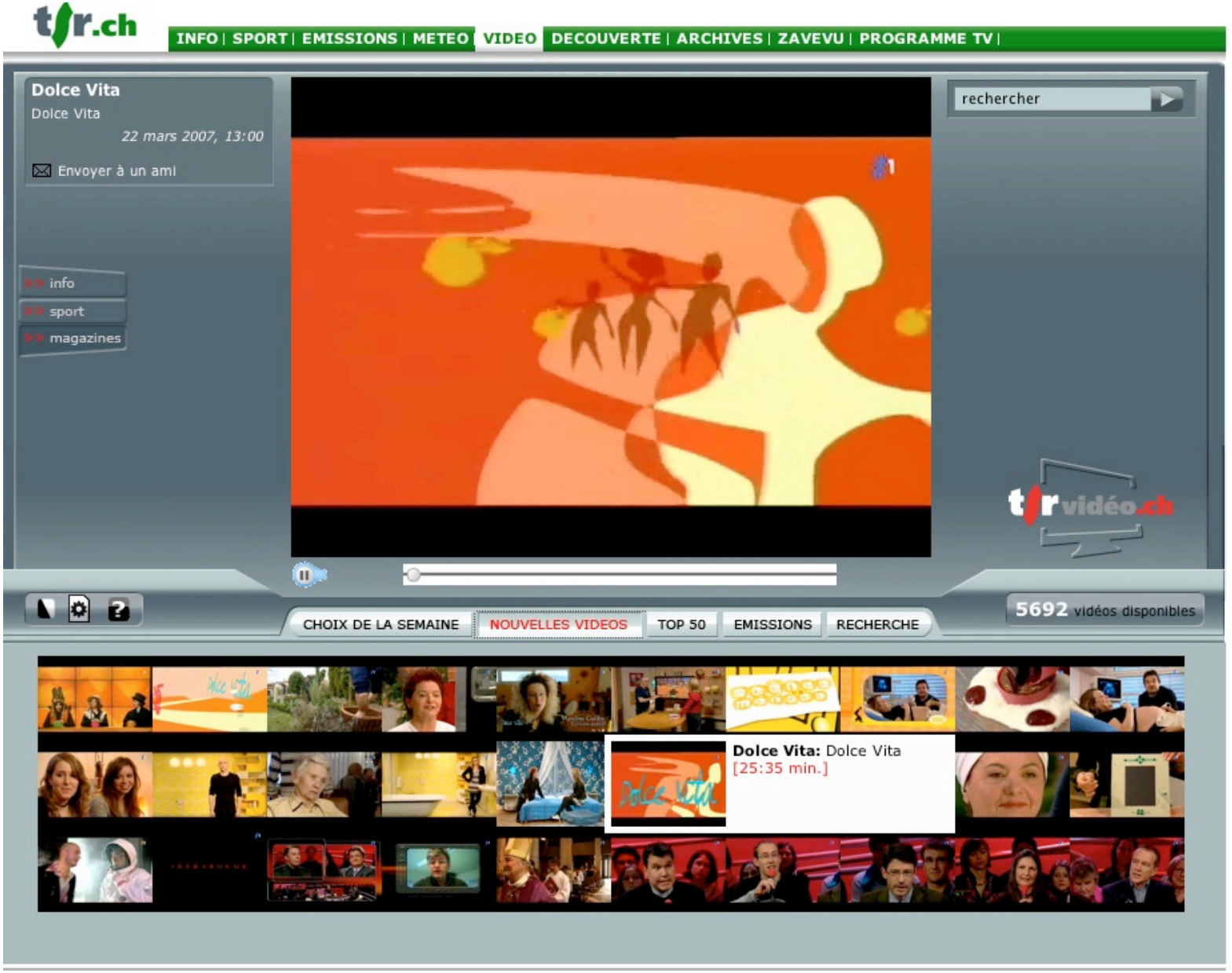
As seen from the “Solr integrator” point of view.

Beyond full-text?

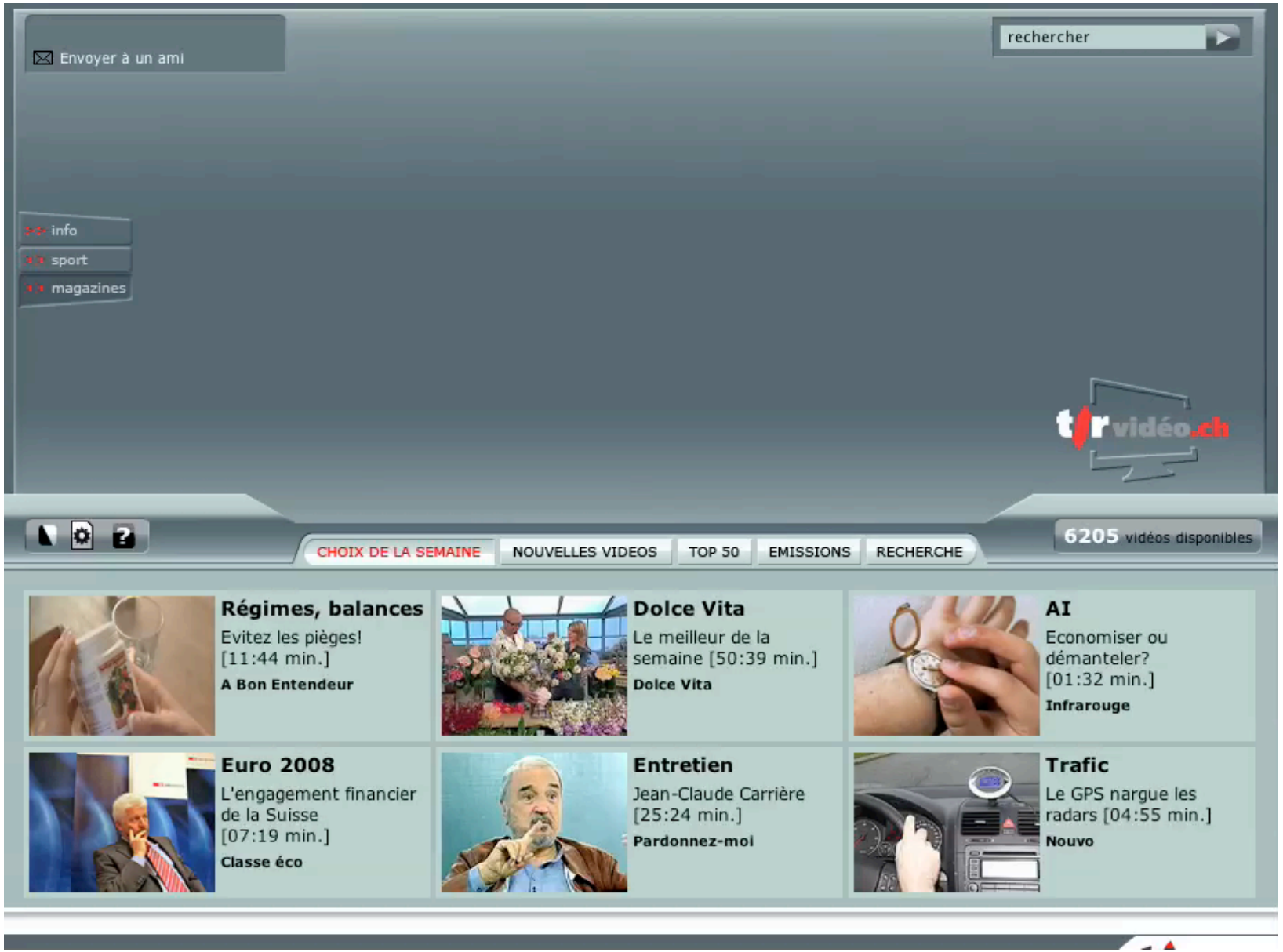


Solr

tsrvideo.ch - a Solr client



tsrvideo.ch - a Solr client

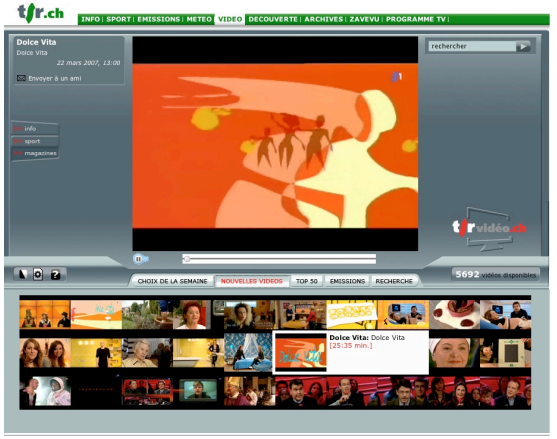


The Project

Deliver a rich video player experience

Users explore much more than they search

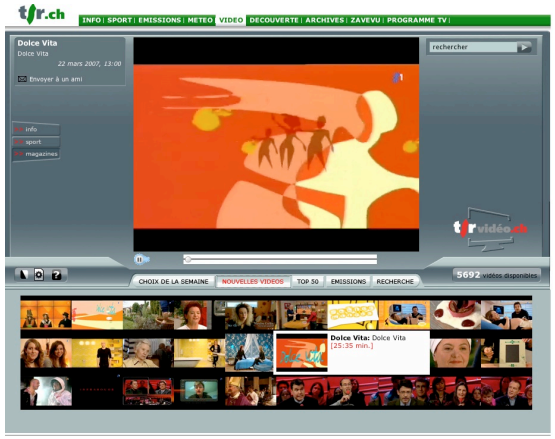
Existing content stored in two separate CMSes with very different content models (and http/XML interfaces)



Solr Lucene

Client system overview

Ajax + HTML



Apache HTTP server

Solr Search server

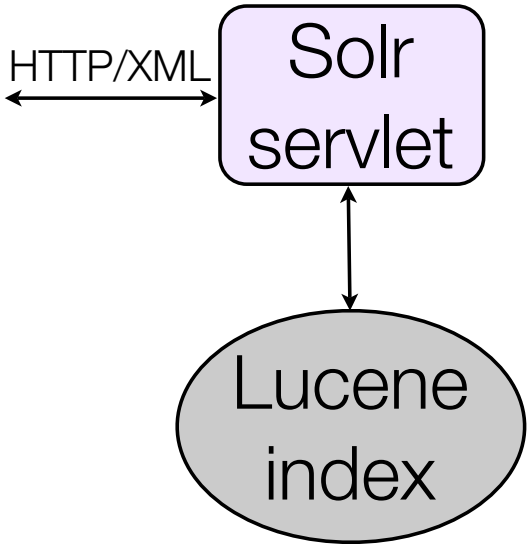
HTTP/JSON

Lucene index

replicated index for backup

the Solr search server

What is Solr?



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

Solr architecture

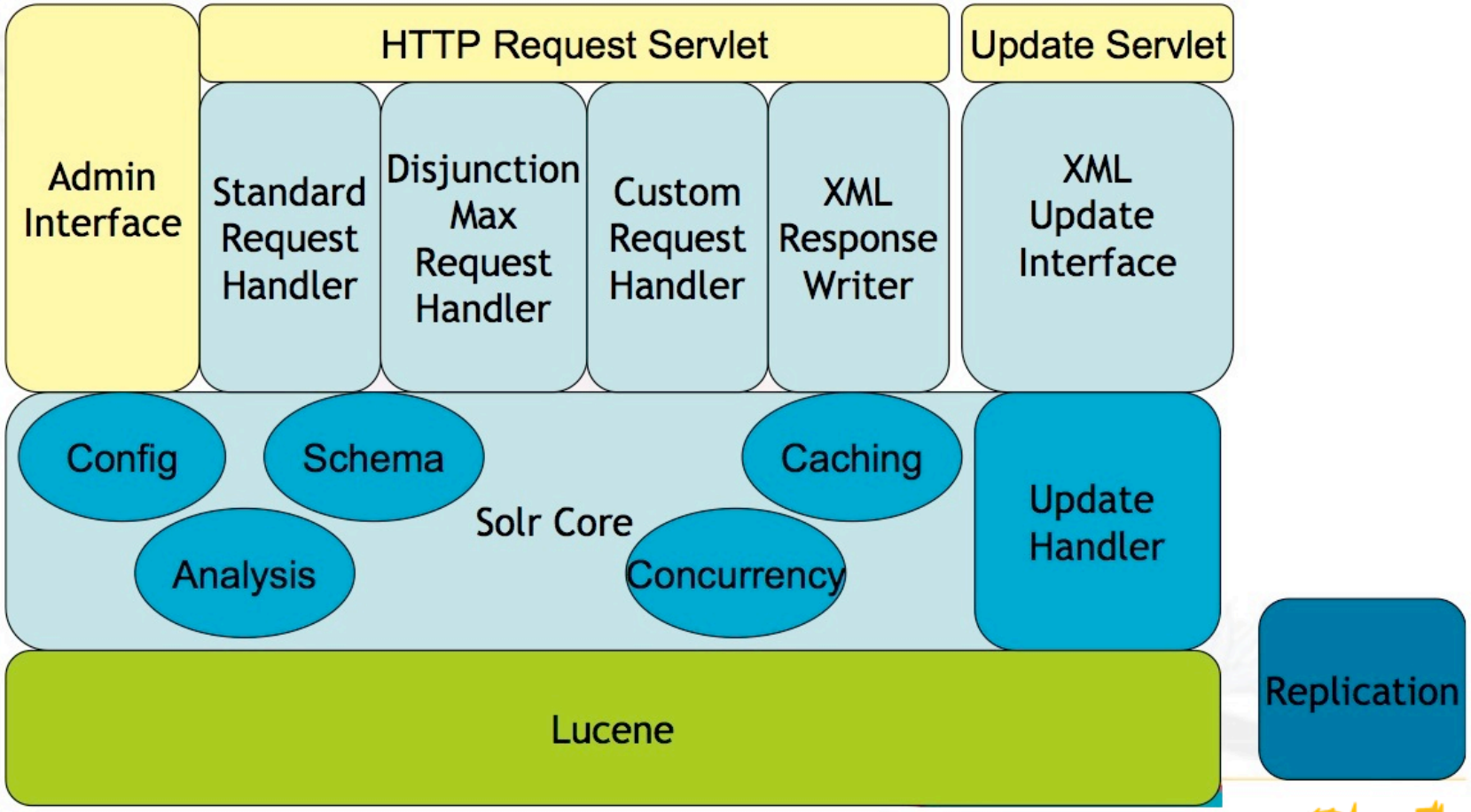


Diagram by Yonik Seeley

Solr

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>
```


"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"/>
```

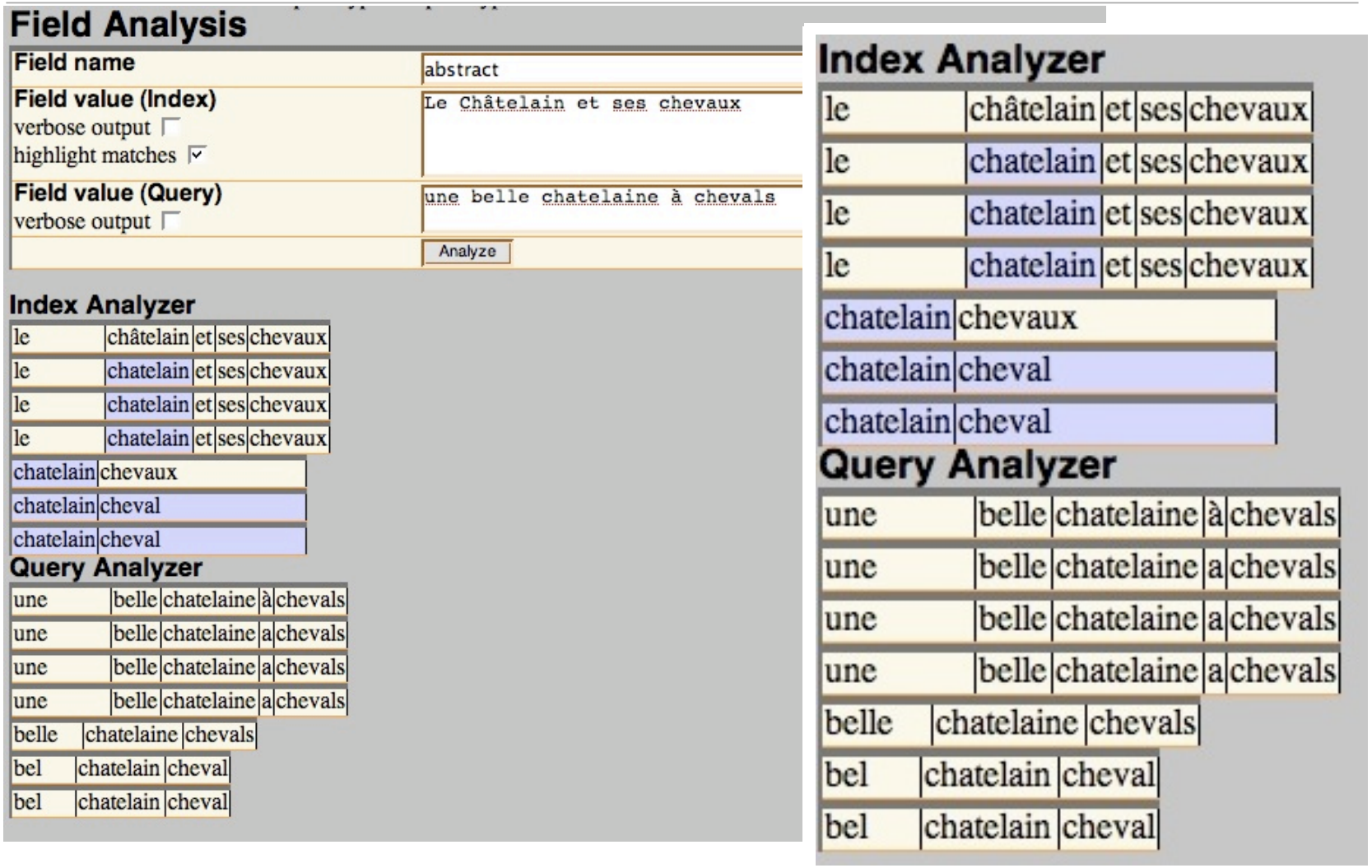
Field content analysis

```
<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



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>
```

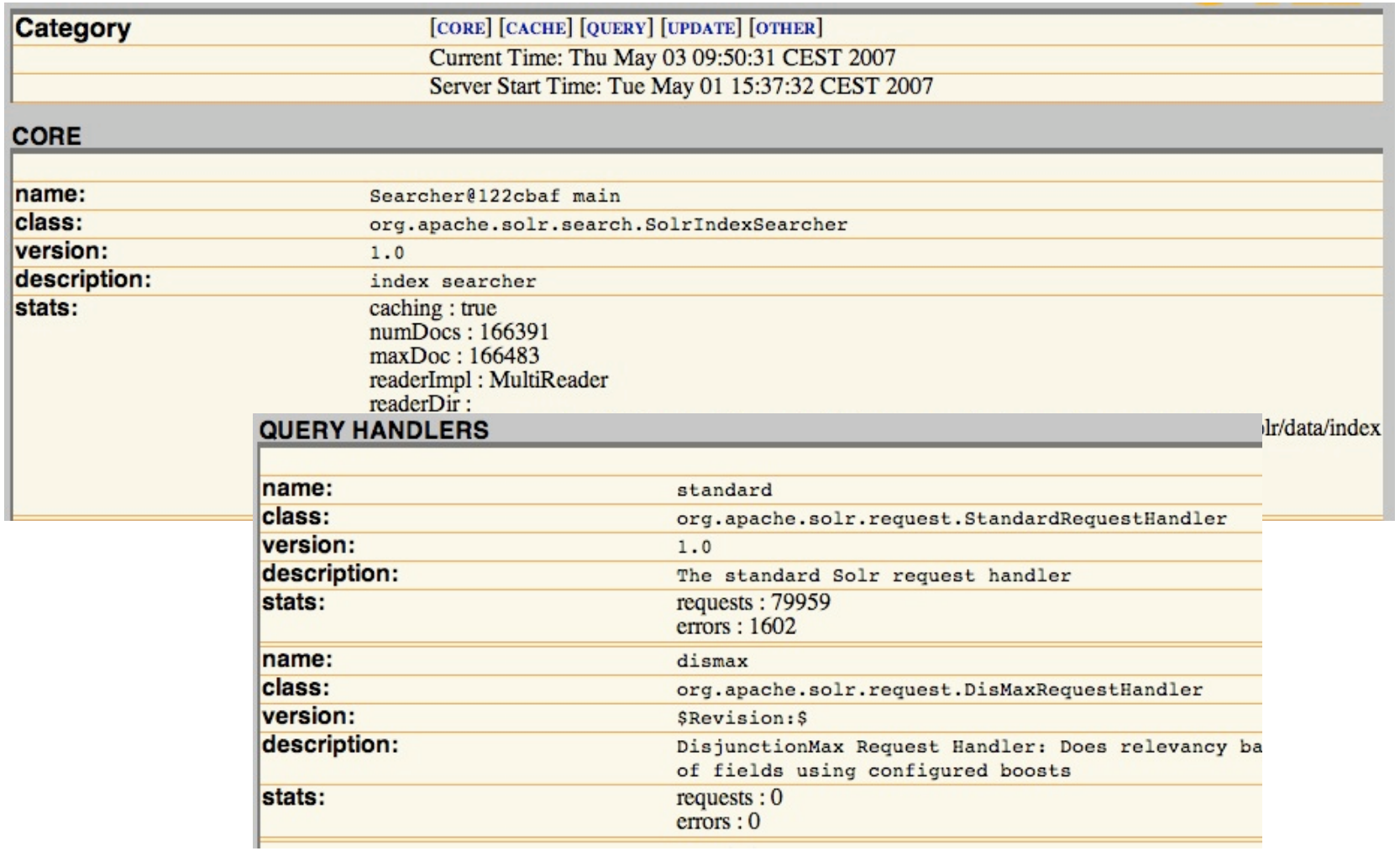
Enhanced Lucene query language as standard

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 :

lr/data/index

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 ba of fields using configured boosts`
stats:	requests : 0 errors : 0

Solr Function Query

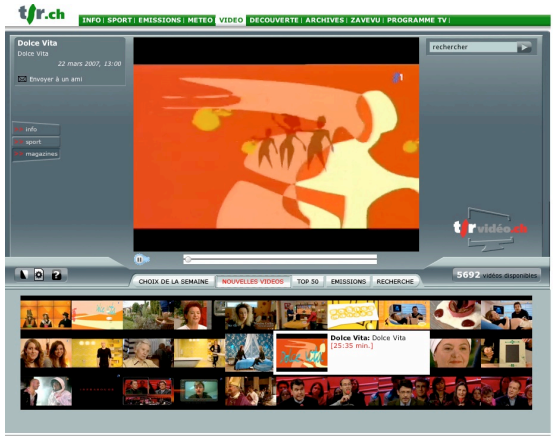
A Function query influences the score by a function of a field's numeric value or ordinal.

```
// 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



Apache HTTP server ⟷ Solr Search server

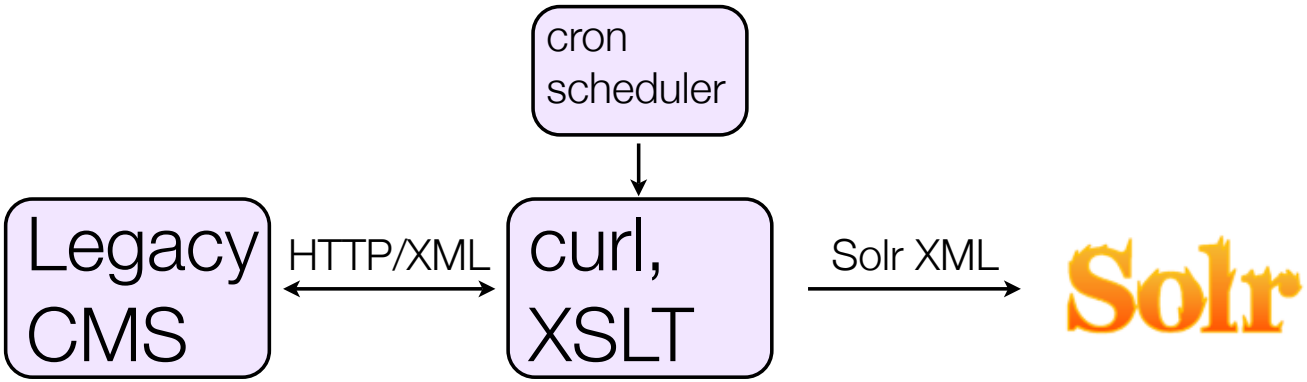
Solr schema and analyzers

HTTP/JSON

Lucene index

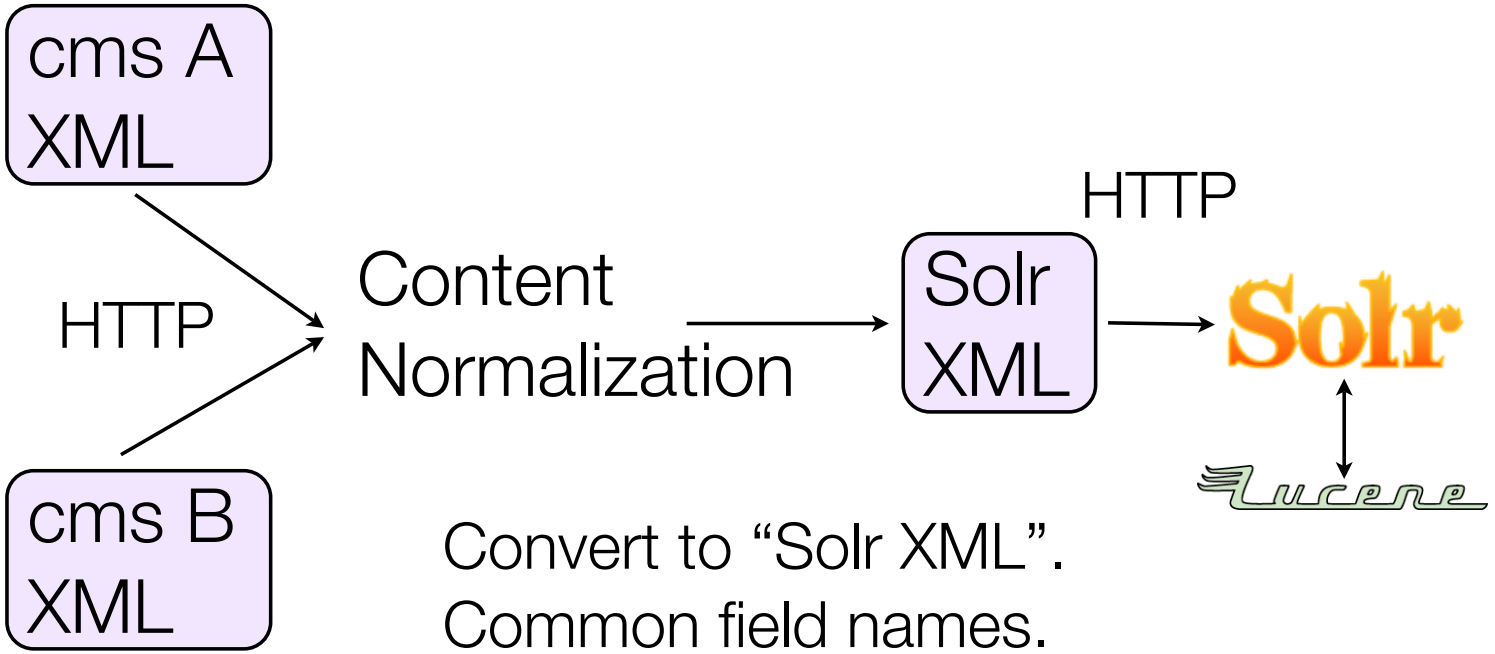
Indexing

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>

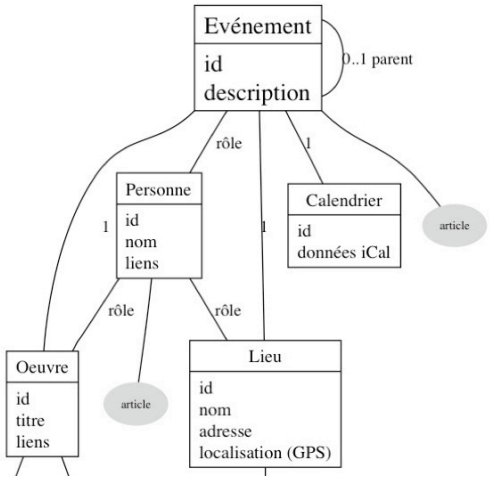
<field name="author">Bob S. Ponge</field>
<field name="author.id">person.438</field>

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

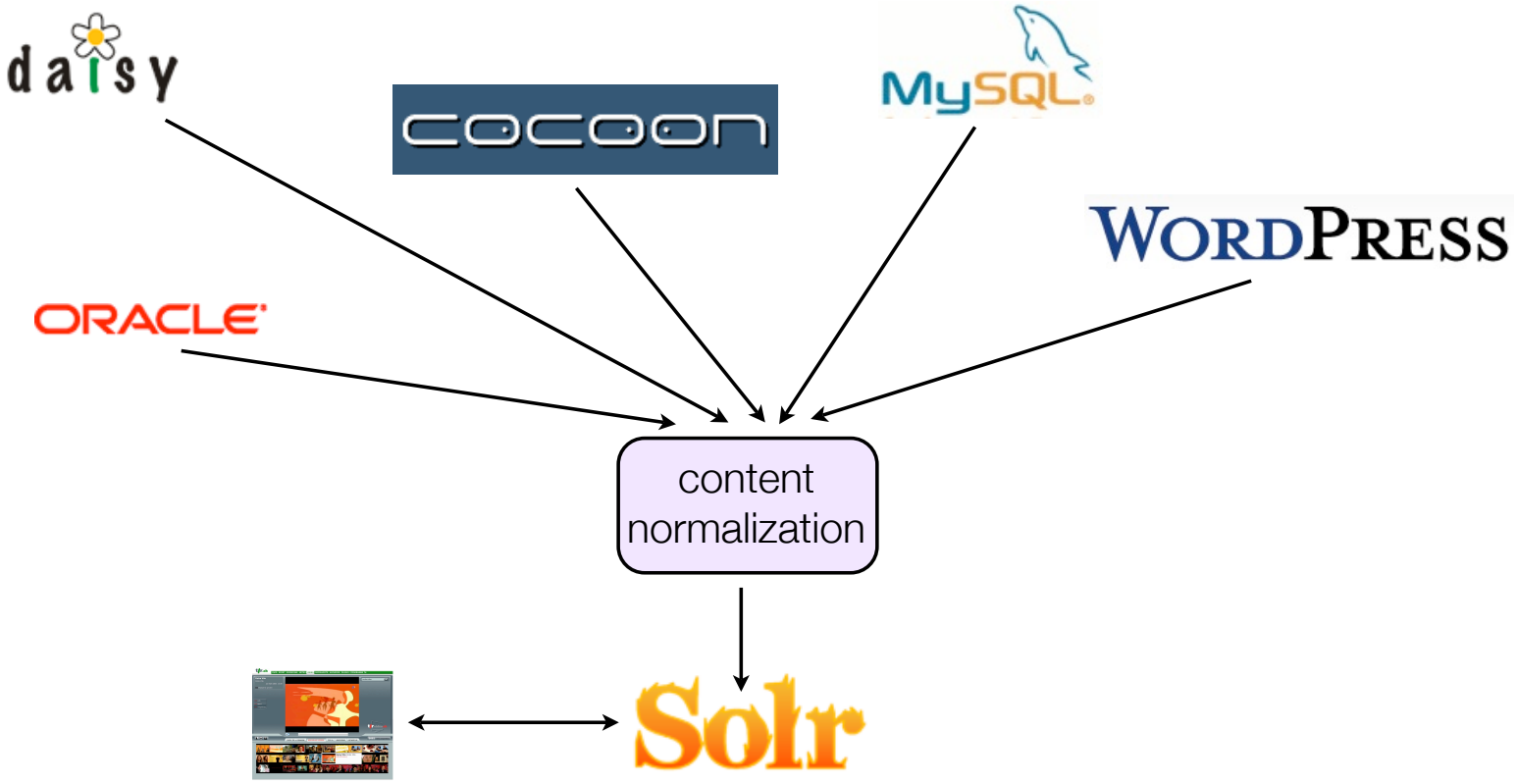
More than “just” full-text searches

<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?"

"testing" picture: taliesin on morguefile.com

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/)

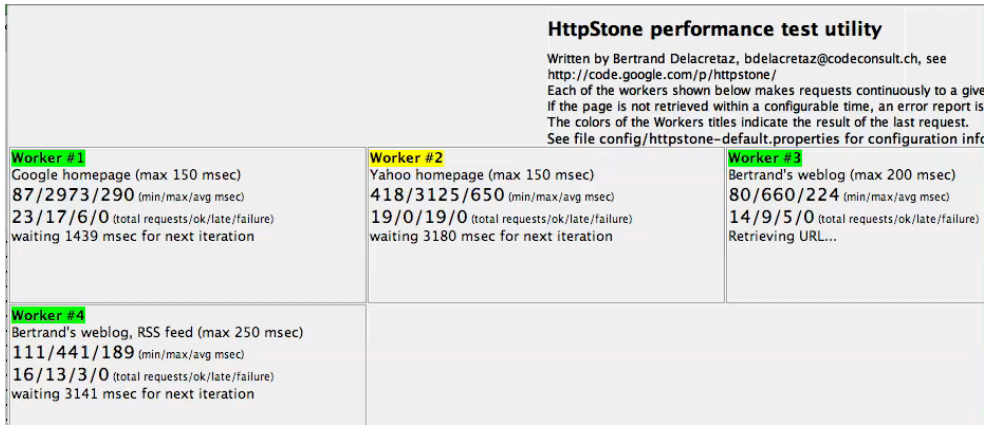
test a query that returns no results
request : /solr/select?wt=xmlt&q=thismustnotbefound
match : /response/result/@numFound : 0
dontMatch : /response/result/@numFound : 1

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



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

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

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

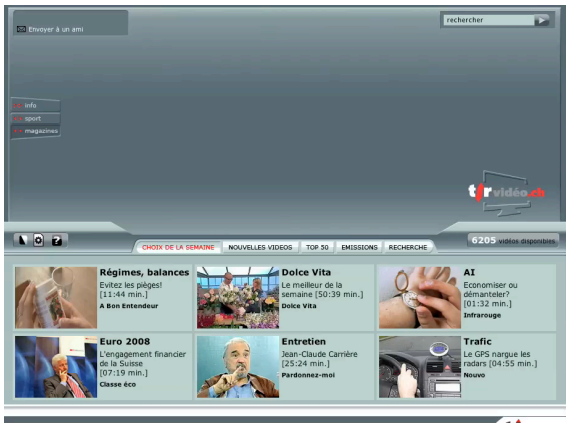
Build confidence before launches!

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.

Solr

Content unification enables “content mining”.

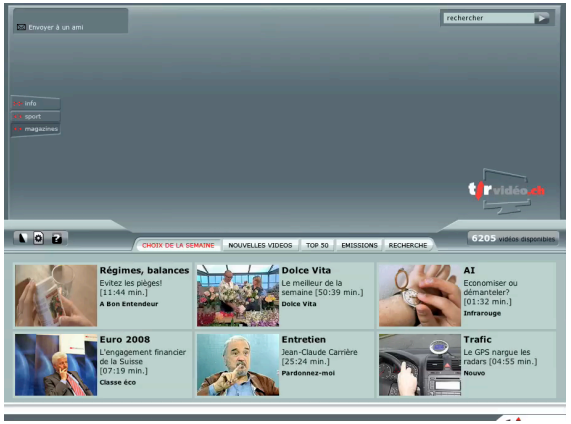
Tune and run automated tests.
Repeat. Repeat. Repeat...

ApacheCon
leading the wave of open source

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

ApacheCon
leading the wave of open source