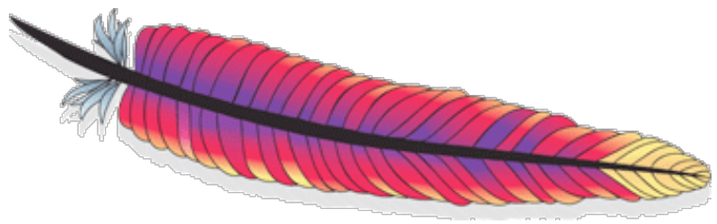
Rapid Prototyping with Solr

Presented by Erik Hatcher, Lucid Imagination

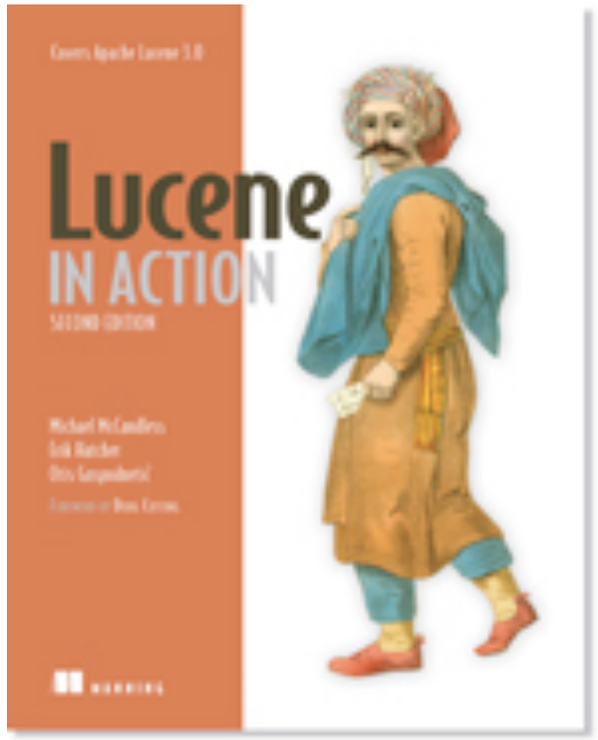
ApacheCon
NORTH AMERICA 2010



lucid
IMAGINATION

About me...

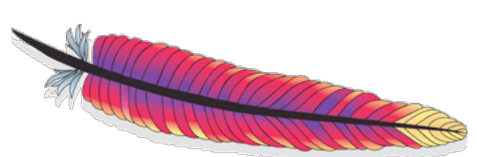
- Co-author, "Lucene in Action"
- Commiter, Lucene and Solr
- Lucene PMC and ASF member
- Member of Technical Staff / co-founder, Lucid Imagination



lucid IMAGINATION

Abstract

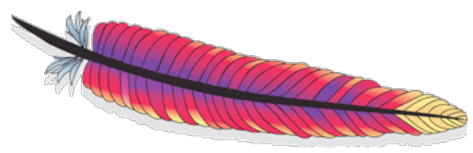
Got data? Let's make it searchable! This presentation will demonstrate getting documents into Solr quickly, will provide some tips in adjusting Solr's schema to match your needs better, and finally will discuss how to showcase your data in a flexible search user interface. We'll see how to rapidly leverage faceting, highlighting, spell checking, and debugging. Even after all that, there will be enough time left to outline the next steps in developing your search application and taking it to production.



lucid IMAGINATION

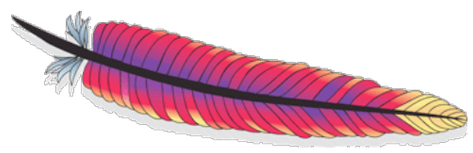
Why prototype?

- Demonstrate Solr can handle your data and searching needs
- It's quick and easy
- An immediate functional user interface impresses decision makers and target users
  - The User Interface ***IS*** the App



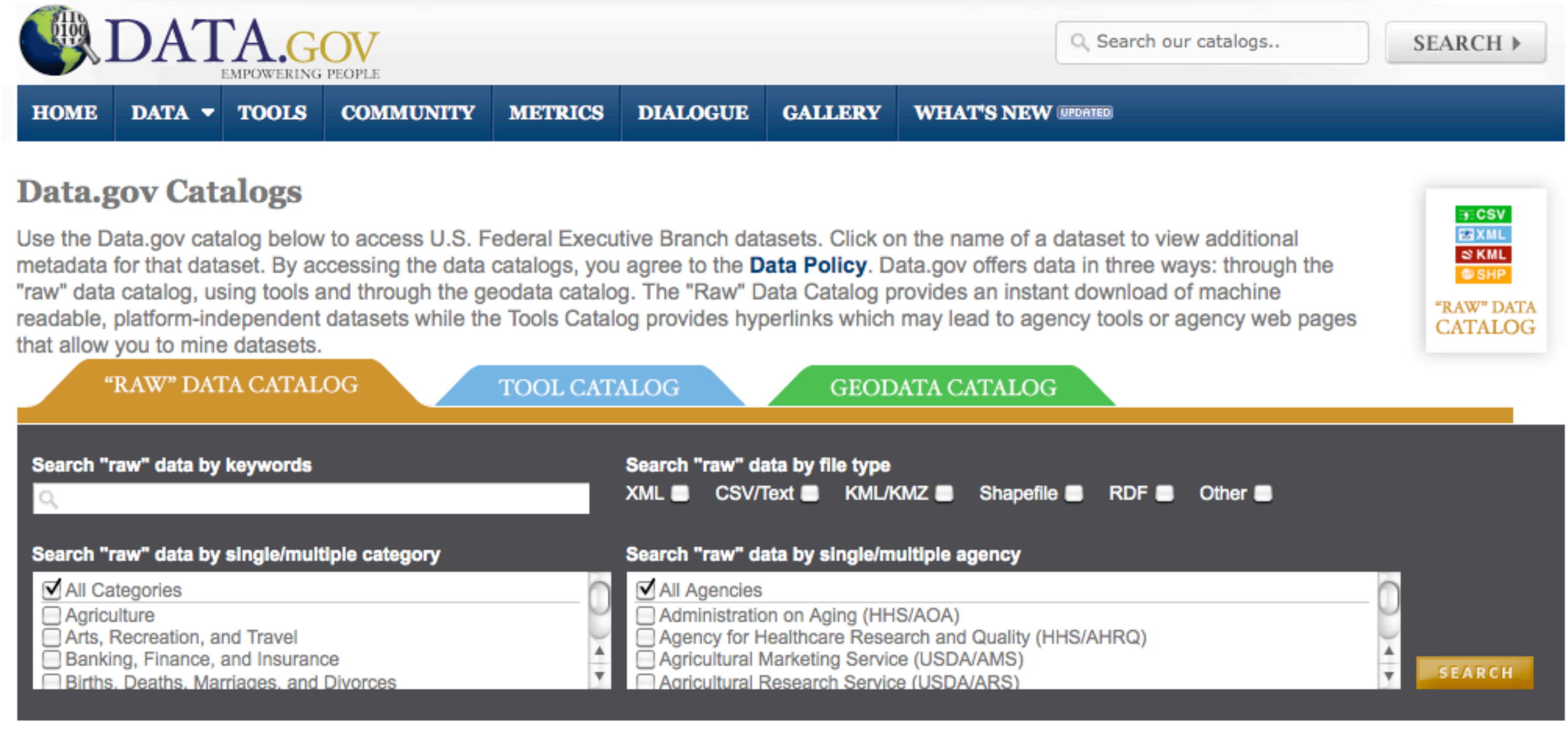
Got data?

- Files (Word, PDF, HTML, etc)?
  - Solr Cell
- Databases? Feeds?
  - Data Import Handler
- 3rd party repositories? Web crawl?
  - Solr HTTP API, Manifold Connectors Framework, Nutch
- CSV?
  - see below!



lucid IMAGINATION

The Data(.gov)



Data.gov CSV catalog

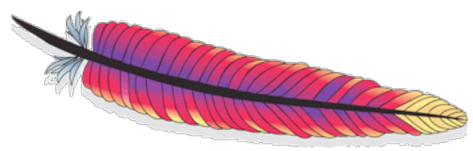
URL,Title,Agency,Subagency,Category,Date Released,Date Updated,Time Period,Frequency,Description,Data.gov Data Category Type,Specialized Data Category Designation,Keywords,Citation,Agency Program Page,Agency Data Series Page,Unit of Analysis,Granularity,Geographic Coverage,Collection Mode,Data Collection Instrument,Data Dictionary/Variable List,Applicable Agency Information Quality Guideline Designation,Data Quality Certification,Privacy and Confidentiality,Technical Documentation,Additional Metadata,FGDC Compliance (Geospatial Only),Statistical Methodology,Sampling,Estimation,Weighting,Disclosure Avoidance,Questionnaire Design,Series Breaks,Non-response Adjustment,Seasonal Adjustment,Statistical Characteristics,Feeds Access Point,Feeds File Size,XML Access Point,XML File Size,CSV/TXT Access Point,CSV/TXT File Size,XLS Access Point,XLS File Size,KML/KMZ Access Point,KML File Size,ESRI Access Point,ESRI File Size,Map Access Point,Data Extraction Access Point,Widget Access Point

"http://www.data.gov/details/4","Next Generation Radar (NEXRAD) Locations","Department of Commerce","National Oceanic and Atmospheric Administration","Geography and Environment","1991","Irregular as needed","1991 to present","Between 4 and 10 minutes","This geospatial rendering of weather radar sites gives access to an historical archive of Terminal Doppler Weather Radar data and is used primarily for research purposes. The archived data includes base data and derived products of the National Weather Service (NWS) Weather Surveillance Radar 88 Doppler (WSR-88D) next generation (NEXRAD) weather radar. Weather radar detects the three meteorological base data quantities: reflectivity, mean radial velocity, and spectrum width. From these quantities, computer processing generates numerous meteorological analysis products for forecasts, archiving and dissemination. There are 159 operational NEXRAD radar systems deployed throughout the United States and at selected overseas locations. At the Radar Operations Center (ROC) in Norman OK, personnel from the NWS, Air Force, Navy, and FAA use this distributed weather radar system to collect the data needed to warn of impending severe weather and possible flash floods; support air traffic safety and assist in the management of air traffic flow control; facilitate resource protection at military bases; and optimize the management of water, agriculture, forest, and snow removal. This data set is jointly owned by the National Oceanic and Atmospheric Administration, Federal Aviation Administration, and Department of Defense.","Raw Data Catalog",...

lucid IMAGINATION

lucidworks for Solr

- great starting point
- built-in and pre-configured:
  - Clustering, via Carrot2
  - Search UI
    - Solritas
    - Server includes root context, handy for serving static files
  - Better stemming, via KStem
  - Tomcat, optionally



~/LucidWorks: start.sh

.
.
.
INFO: using system property solr.solr.home: /Users/erikhatcher/LucidWorks/lucidworks/jetty/../solr
Nov 4, 2010 11:14:40 AM org.apache.solr.servlet.SolrUpdateServlet init
INFO: SolrUpdateServlet.init() done
2010-11-04 11:14:40.709::INFO: Started SocketConnector @ 0.0.0.0:8983
Nov 4, 2010 11:14:40 AM org.apache.solr.core.SolrCore execute
INFO: [] webapp=null path=null params={start=0&event=firstSearcher&q=solr+rocks&rows=10} hits=0 status=0 QTime=31
Nov 4, 2010 11:14:40 AM org.apache.solr.core.SolrCore execute
INFO: [] webapp=null path=null params={event=firstSearcher&q=static+firstSearcher+warming+query+from+solrconfig.xml} hits=0 status=0 QTime=2
Nov 4, 2010 11:14:40 AM org.apache.solr.core.QuerySenderListener newSearcher
INFO: QuerySenderListener done.
Nov 4, 2010 11:14:40 AM org.apache.solr.handler.component.SpellCheckComponent$SpellCheckerListener newSearcher
INFO: Loading spell index for spellchecker: default
Nov 4, 2010 11:14:40 AM org.apache.solr.core.SolrCore registerSearcher
INFO: [] Registered new searcher Searcher@4ff217ec main

Getting started...

```
&fieldnames=id,title
```

HTTP ERROR: 400 CSVLoader: input=file:/Users/erikhatcher/dev/prototyping/ApacheCon2010/data_gov_catalog.csv, line=0,expected 2 values but got 53

```
curl "http://localhost:8983/solr/update/csv
?commit=true
&stream.file=data_gov_catalog.csv
&header=true&fieldnames=id,title,agency_s,su
bagency_s,category_s,,,,,description,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,"
```

First look

http://localhost:8983/solr/itas

Query:

Facets

cat

Clusters

- Other Topics

3575 results found in 21 ms

id : http://www.data.gov/details/4

title : Next Generation Radar (NEXRAD) Locations

agency_s : Department of Commerce

subagency_s : National Oceanic and Atmospheric Administration

category_s : Geography and Environment

description : This geospatial rendering of weather radar sites gives access to an historical archive of Terminal Doppler Weather Radar data and is used primarily for research purposes. The archived data includes base data and derived products of the National Weather Service (NWS) Weather Surveillance Radar 88 Doppler (WSR-88D) next generation (NEXRAD) weather radar. Weather radar detects the three meteorological base data quantities: reflectivity, mean radial velocity, and spectrum width. From these quantities, computer processing generates numerous meteorological analysis products for forecasts, archiving and dissemination. There are 159 operational NEXRAD radar systems deployed throughout the United States and at selected overseas locations. At the Radar Operations Center (ROC) in Norman OK, personnel from the NWS, Air Force, Navy, and FAA use this distributed weather radar system to collect the data needed to warn of impending severe weather and possible flash floods; support air traffic safety and assist in the management of air traffic flow control; facilitate resource protection at military bases; and optimize the management of water, agriculture, forest, and snow removal. This data set is jointly owned by the National Oceanic and Atmospheric Administration, Federal Aviation Administration, and Department of Defense.

score : 1.0

id : http://www.data.gov/details/6

title : EPA Geospatial Data Download: Facility and Site Information

agency_s : Environmental Protection Agency

category_s : Geography and Environment

description : Contains informaiton about facilities or sites subject to environmental regulation, including key facility information along with associated environmental interests for use in mapping and reporting applications.

score : 1.0

Solritas

- Pronounced: so-LAIR-uh-toss
- Celeritas is a Latin word, translated as "swiftness" or "speed". It is often given as the origin of the symbol c, the universal notation for the ***speed*** of light - http://en.wikipedia.org/wiki/Celeritas
- VelocityResponseWriter - simply passes the Solr response through the Apache ***Velocity*** templating engine
- • http://wiki.apache.org/solr/VelocityResponseWriter

First facets

http://localhost:8983/solr/itas?facet.field=category_s

Facets

category_s

- Geography and Environment (1901)
- National Security and Veterans Affairs (214)
- Health and Nutrition (177)
- Federal Government Finances and Employment (141)
- Labor Force, Employment, and Earnings (134)
- Other (104)
- Energy and Utilities (102)
- Information and Communications (88)
- Social Insurance and Human Services (86)
- Agriculture (82)
- Science and Technology (81)
- Law Enforcement, Courts, and Prisons (77)
- Population (74)
- Business Enterprise (69)
- Education (47)
- Banking, Finance, and Insurance (42)
- Transportation (39)
- Income, Expenditures, Poverty, and Wealth (31)
- Construction and Housing (14)
- Elections (14)

Iterations

- Cleaned up field names:
  - s/\t/,/ and s/\ /_/
- Mapped some field values
- Copied Solr trunk Velocity templates and CSS, polished layout and content templates
- Schema and config adjustments

lucidworks DATA.GOV EMPOWERING PEOPLE

Find: weather Submit Query

Field Facets

Agency

- Department Of Commerce (59)
- National Aeronautics And Space Administration (6)
- Department Of Transportation (2)
- Department Of Labor (1)

Designation

- Administrative (35)
- Geospatial (22)
- Research (6)
- Surveillance (3)
- Other (1)
- Statistical (1)

Category

- Geography And Environment (59)
- Science And Technology (6)
- Births, Deaths, Marriages, And Divorces (1)
- Labor Force, Employment, And Earnings (1)
- Transportation (1)

Category Type

- Tool Catalog (40)
- Raw Data Catalog (28)

FGDC Compliance

- Yes (11)

Privacy & Confidentiality

- Yes (68)

Data Quality Certification

- Yes (68)

68 results found in 18 ms Page 1 of 7

... Geospatial display of current ***weather*** radar images (RIDGE ***Weather*** Radar) ...

... Provides GIS overlays for current ***weather*** radar results ...

... RSS Feed for Marine ***Weather*** Discussions ...

... The National ***Weather*** Service (NWS) National Hurricane Center uses regularly updated RSS feeds ... ... to disseminate Marine ***Weather*** Discussions. ...

... RSS Feeds for Severe ***Weather*** Advisories ...

... The National ***Weather*** Service (NWS) uses RSS feeds to disseminate severe ***weather*** advisories ... ... ***weather*** and storm ; high wind ; small craft; hurricane; tornado; fog ; red flag ; freeze and frost ... ... ; special ***weather*** statements, etc. ...

... RSS Feed for Atlantic Tropical ***Weather*** Discussion ...

... The National ***Weather*** Service (NWS) National Hurricane Center uses regularly updated RSS feed ... ... to disseminate an Atlantic Tropical ***Weather*** Discussion. ...

... RSS Feed for East Pacific Tropical ***Weather*** Discussion ...

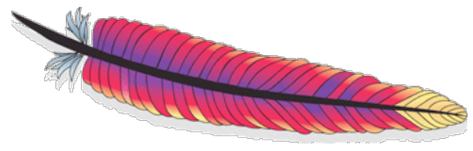
... The National ***Weather*** Service (NWS) National Hurricane Center uses regularly updated RSS feed ... ... to disseminate an East Pacific Tropical ***Weather*** Discussion. ...

RSS Feeds for Central Pacific Hurricanes and Tropical Cyclones

... The National ***Weather*** Service (NWS) National Hurricane Center uses RSS feeds to disseminate Central ... ... : Tropical ***Weather*** Outlooks; Central Pacific hurricane advisories , outlooks and discussions; Hawaiian surf ... ... and land forecasts; and Hawaiian ***weather*** news. ...

... RSS Feeds for Storm Prediction Center Severe ***Weather*** Outlooks, Watches, Discussions, and Status Reports ...

... The National ***Weather*** Service (NWS) Storm Prediction Center uses RSS feeds to disseminate severe ... ... ***weather*** outlooks and watches, mesoscale discussion, and status reports. These RSS feeds are regularly ... ... ; convective outlooks; and fire ***weather*** outlooks. ...



Debugging

http://localhost:8983/solr/data.gov?q=searching&debugQuery=true

toggle parsed query

```
+DisjunctionMaxQuery((text:search^0.5 | description:search)) ()
```

Department of Education Discretionary and Formula Grant Database

... -agreement awards for the most recent fiscal years. Five Ways to ***Search*** for Grant Awards: Pick List ***Search*** ... ... - ***search*** for grant awards by selecting (picking) from lists of actual data (up to 10 selections per list ... ...). Text ***Search*** - ***search*** for awards using a particular text string, such as zip code Date ***Search*** - ***search*** ...

toggle explain

```
1.7863293 = (MATCH) sum of:
  1.7863293 = (MATCH) max of:
    0.28246897 = (MATCH) weight(text:search^0.5 in 1900), product of:
      0.39765334 = queryWeight(text:search^0.5), product of:
        0.5 = boost
        4.32969 = idf(docFreq=127, maxDocs=3575)
        0.18368675 = queryNorm
      0.7103398 = (MATCH) fieldWeight(text:search in 1900), product of:
        3.0 = tf(termFreq(text:search)=9)
        4.32969 = idf(docFreq=127, maxDocs=3575)
        0.0546875 = fieldNorm(field=text, doc=1900)
    1.7863293 = (MATCH) fieldWeight(description:search in 1900), product of:
      3.0 = tf(termFreq(description:search)=9)
      5.444051 = idf(docFreq=41, maxDocs=3575)
      0.109375 = fieldNorm(field=description, doc=1900)
```

Data analysis

http://localhost:8983/solr/admin/luke?wt=xslt&tr=luke.xsl

privacy_and_confidentiality

type	string
schema	Indexed - Stored - Multivalued - Omit Norms - Sort Missing Last -
dynamicBase	*
index	
docs	
distinct	

topTerms

Yes	
Not Relevant	
YES	

data_quality_certification

type	string
schema	Indexed - Stored - Multivalued - Omit Norms - Sort Missing Last -
dynamicBase	*
index	Indexed - Stored - Omit Norms - Lazy -
docs	3571
distinct	3

topTerms

Yes	3537
YES	33
Not Relevant	1

Mapping field values

- CSV update handler can map field values
- &f.privacy_and_confidentiality.map=YES:Yes

&f.data_quality_certification.map=YES:Yes

privacy_and_confidentiality

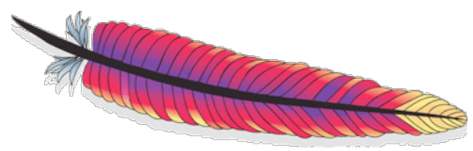
Yes (3265)

Not Relevant (306)

data_quality_certification

Yes (3570)

Not Relevant (1)



Making it all searchable

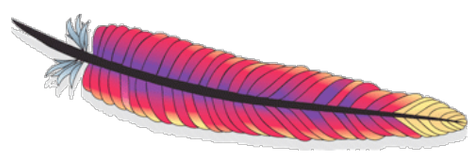
- To quickly bring in arbitrary data
- Make everything searchable

```
<dynamicField name="*" type="string" multiValued="true"/>
...
<copyField source="*" dest="text"/>
```

Splitting keywords

- CSV handler: f.keywords.split=true
  - stored values are split, multivalued
- Or via schema
  - Stored value remains as in original, single valued

```
<fieldType name="comma_separated" class="solr.TextField" omitNorms="true">
  <analyzer>
    <tokenizer class="solr.PatternTokenizerFactory" pattern="\s*,\s*"/>
  </analyzer>
</fieldType>
...
<field name="keywords" type="comma_separated" indexed="true" stored="true"/>
```



Suggest

- Suggest terms as user types in search box
- Technique: jQuery autocomplete, Solr's TermsComponent, Velocity template

```
http://localhost:8983/solr/terms
?terms.fl=suggest
&terms.prefix=sola&terms.sort=count
&wt=velocity&v.template=suggest
```

```
#foreach($t in $response.response.terms.suggest)
$t.key
#end
```

Suggest schema

```
<fieldType name="suggestable" class="solr.TextField"
positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.PatternReplaceFilterFactory"
            pattern="([^a-z])"
            replacement="" replace="all"/>
    <filter class="solr.StopFilterFactory"
            ignoreCase="true" words="stopwords.txt"
            enablePositionIncrements="true" />
  </analyzer>
</fieldType>

...

<field name="suggest" type="suggestable"
       indexed="true" stored="false" multiValued="true"/>
```

Custom pages

- Document detail page
- Multiple query intersection comparison with Venn visualization

Document detail

```
http://localhost:8983/solr/data.gov/document
?id=http%3A%2F%2Fwww.data.gov%2Fdetails%2F61
```

lucidworks DATA.GOV EMPOWERING PEOPLE

http://www.data.gov/details/61

id:	http://www.data.gov/details/61
title:	Geospatial display of current weather radar images (RIDGE Weather Radar)
agency:	Department of Commerce
subagency:	National Oceanic and Atmospheric Administration
category:	Geography and Environment
date_released:	Unknown
date_updated:	near real time
time_period:	near real time
frequency:	near real time
description:	Provides GIS overlays for current weather radar results
data_category_type:	Raw Data Catalog
specialized_data_category_designation:	Geospatial
keywords:	radar integrated display with geospatial elements, ridge radar, doppler radar, geographic overlay, weather radar, warning, enhanced radar images

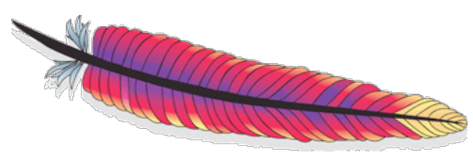
Document detail detail

solrconfig.xml

```
<requestHandler name="/data.gov/document" class="solr.SearchHandler">
  <lst name="defaults">
    <str name="wt">velocity</str>
    <str name="v.template">document</str>
    <str name="v.layout">layout</str>
    <str name="title">Data.gov data set</str>
    <str name="q">{!raw f=id v=$id}</str>
  </lst>
</requestHandler>
```

document.vm

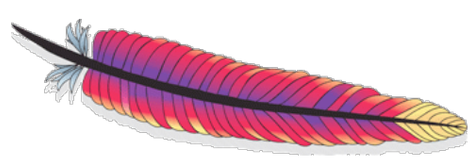
```
#set($doc= $response.results.get(0))
<span><a href="$doc.getFieldValue('id')">$doc.getFieldValue('id')</a></span>

<table>
#foreach($fieldname in $doc.fieldNames)
  <tr>
    <td>$fieldname:</td>
    <td>
     #foreach($value in $doc.getFieldValues($fieldname))
       $esc.html($value)
     #end
     </td>
  </tr>
#end
</table>
```

lucid
IMAGINATION

Query intersection

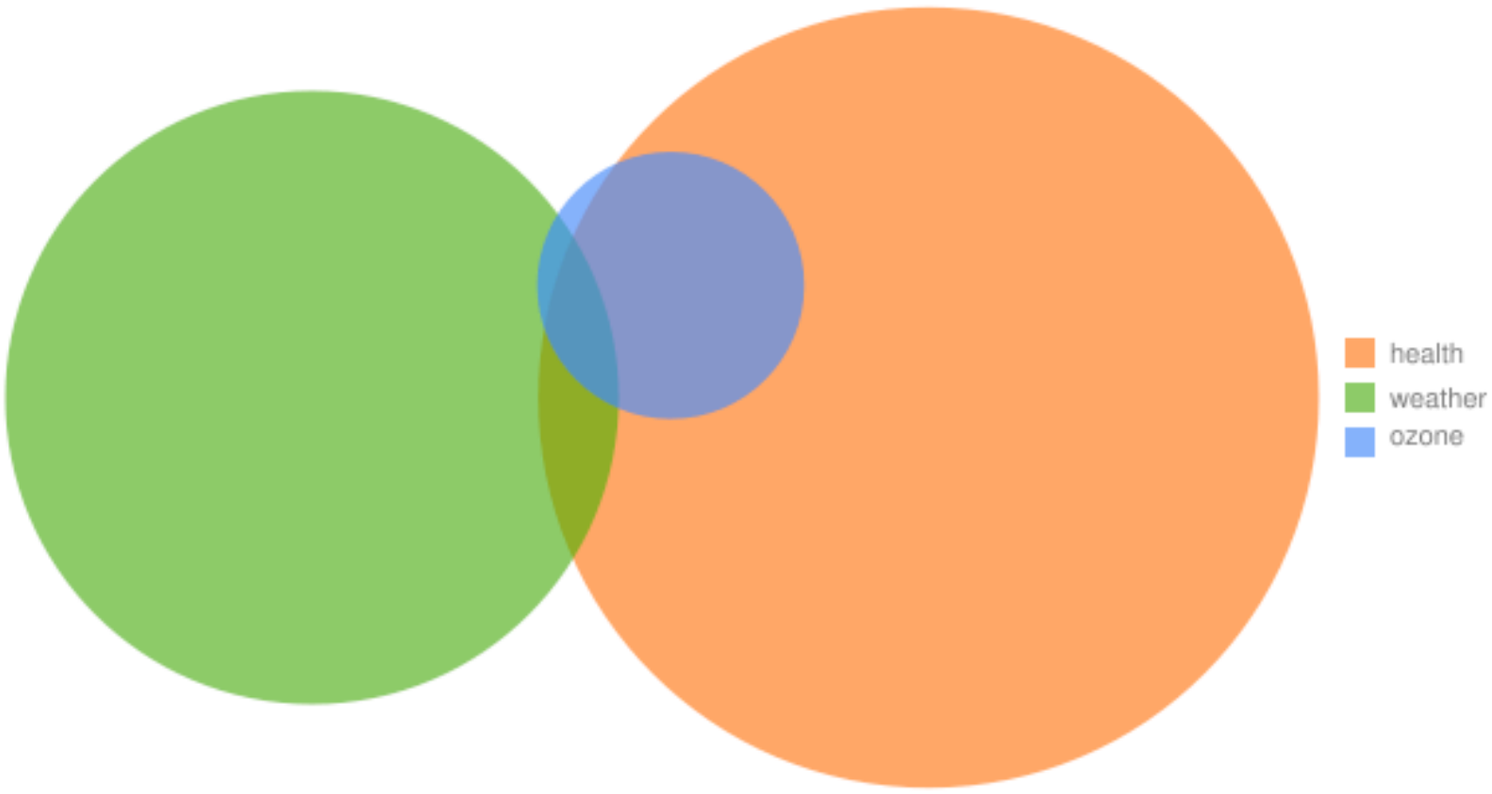
- Just showing off.... how easy it is to do something with a bit of visual impact
- Compare three independent queries, intersecting them in a Venn diagram visualization



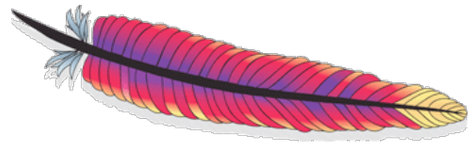
lucid IMAGINATION

lucidworks DATA.GOV EMPOWERING PEOPLE

A: health B: weather C: ozone Submit Query



- A: health (376)
- B: weather (68)
- C: ozone (13)
- A&B: (health) AND (weather) (2)
- A&C: (health) AND (ozone) (6)
- B&C: (weather) AND (ozone) (1)
- A&B&C: (health) AND (weather) AND (ozone) (1)



Compare static page

solrconfig.xml

```
<requestHandler name="/data.gov/compare" class="solr.DumpRequestHandler">
  <lst name="defaults">
    <str name="wt">velocity</str>
    <str name="v.template">compare</str>
    <str name="v.layout">layout</str>
    <str name="title">Data.gov Query Comparison</str>
  </lst>
</requestHandler>
```

compare.vm

```
<script type="text/javascript">
  function generate_venn() {
    var a=encodeURIComponent($("#a").val());
    var b=encodeURIComponent($("#b").val());
    var c=encodeURIComponent($("#c").val());
    var ab='('+a+')+AND+('+b+')';
    var ac='('+a+')+AND+('+c+')';
    var bc='('+b+')+AND+('+c+')';
    var abc='('+a+')+AND+('+b+')+AND+('+c+')';
    $('#venn').load('/solr/select?
q=*:*&wt=velocity&v.template=venn&rows=0&facet=on&facet.query={!key=a}'+a+'&facet.query={!key=b}'+b
+'&facet.query={!key=c}'+c+'&facet.query={!key=intersect_ab}'+ab+'&facet.query={!key=intersect_ac}'+ac
+'&facet.query={!key=intersect_bc}'+bc+'&facet.query={!key=intersect_abc}'+abc+'&q_a='+a+'&q_b='+b+'&q_c='+c
+'&q_ab='+ab+'&q_ac='+ac+'&q_bc='+bc+'&q_abc='+abc);
    return false;
  }
</script>
<form action="#" id="compare_form" onsubmit="return generate_venn()">
  A: <input type="text" name="a" id="a" value="health"/>
  B: <input type="text" name="b" id="b" value="weather"/>
  C: <input type="text" name="c" id="c" value="ozone"/>
  <input type="submit"/>
</form>

<div id="venn"></div>
```

Venn chart

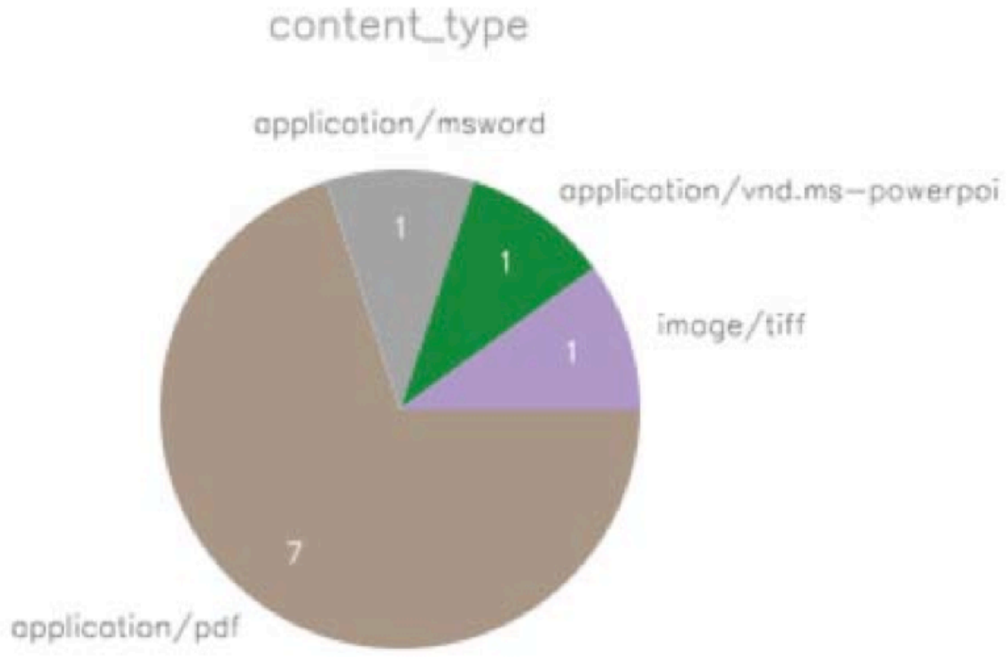
venn.vm

```
#set($values = $response.response.facet_counts.facet_queries)
#set($params = $response.responseHeader.params)

<img src="http://chart.apis.google.com/chart?
chs=600x400&cht=v&chd=t:$values.a,$values.b,$values.c,
$values.intersect_ab,$values.intersect_ac,$values.intersect_bc,
$values.intersect_abc&chdl=$esc.url($params.q_a)|$esc.url
($params.q_b)|$esc.url($params.q_c)"/>

<ul>
  <li>A: <a href="/solr/data.gov?q={!lucene}$params.q_a">$params.q_a</a> ($values.a)</li>
  <li>B: <a href="/solr/data.gov?q={!lucene}$params.q_b">$params.q_b</a> ($values.b)</li>
  <li>C: <a href="/solr/data.gov?q={!lucene}$params.q_c">$params.q_c</a> ($values.c)</li>
  <li>A&B: <a href="/solr/data.gov?q={!lucene}$params.q_ab">$params.q_ab</a>
($values.intersect_ab)</li>
  <li>A&C: <a href="/solr/data.gov?q={!lucene}$params.q_ac">$params.q_ac</a>
($values.intersect_ac)</li>
  <li>B&C: <a href="/solr/data.gov?q={!lucene}$params.q_bc">$params.q_bc</a>
($values.intersect_bc)</li>
  <li>A&B&C: <a href="/solr/data.gov?q={!lucene}$params.q_abc">$params.q_abc</a>
($values.intersect_abc)</li>
</ul>
```

Other fun and quick viz

Treemap

Pie



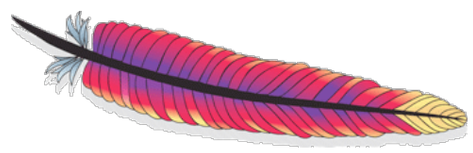
http://lucene-eurocon.org/slides/Rapid-Prototyping-with-Solr_Erik-Hatcher.pdf

http://www.lucidimagination.com/solutions/webinars/Rapid-Prototyping-Search-Applications-with-Solr

lucid IMAGINATION

Prototyping Tools

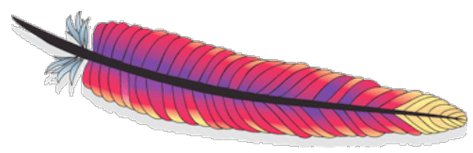
- CSV update handler
- Schema Browser - /solr/admin/schema.jsp
  - Luke request handler - /solr/admin/luke
- Solritas



lucid IMAGINATION

Then what?

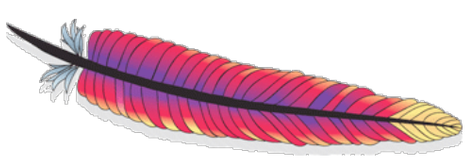
- Script the indexing process: full & delta
- Work with real users on actual needs
- Integrate with production systems
- Iterate on schema enhancements, configuration tweaks such as caching
- Deploy to staging/production environments and work at scale: collection size, real queries and performance, hardware and JVM settings



lucid IMAGINATION

Test

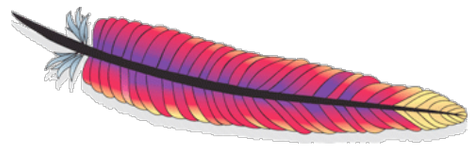
- Performance
- Scalability
- Relevance
- Automate all of the above, start baselines, avoid regressions



lucid IMAGINATION

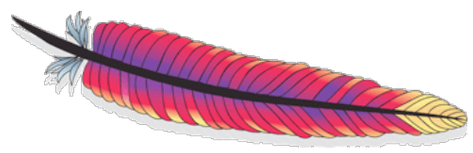
The Code

https://github.com/erikhatcher/solr-rapid-prototyping



For more information...

- http://www.lucidimagination.com
- LucidFind
  - search Lucene ecosystem: mailing lists, wikis, JIRA, etc
  - http://search.lucidimagination.com
- Getting started with LucidWorks Enterprise:
  - http://www.lucidimagination.com/enterprise-search-solutions



lucid

IMAGINATION