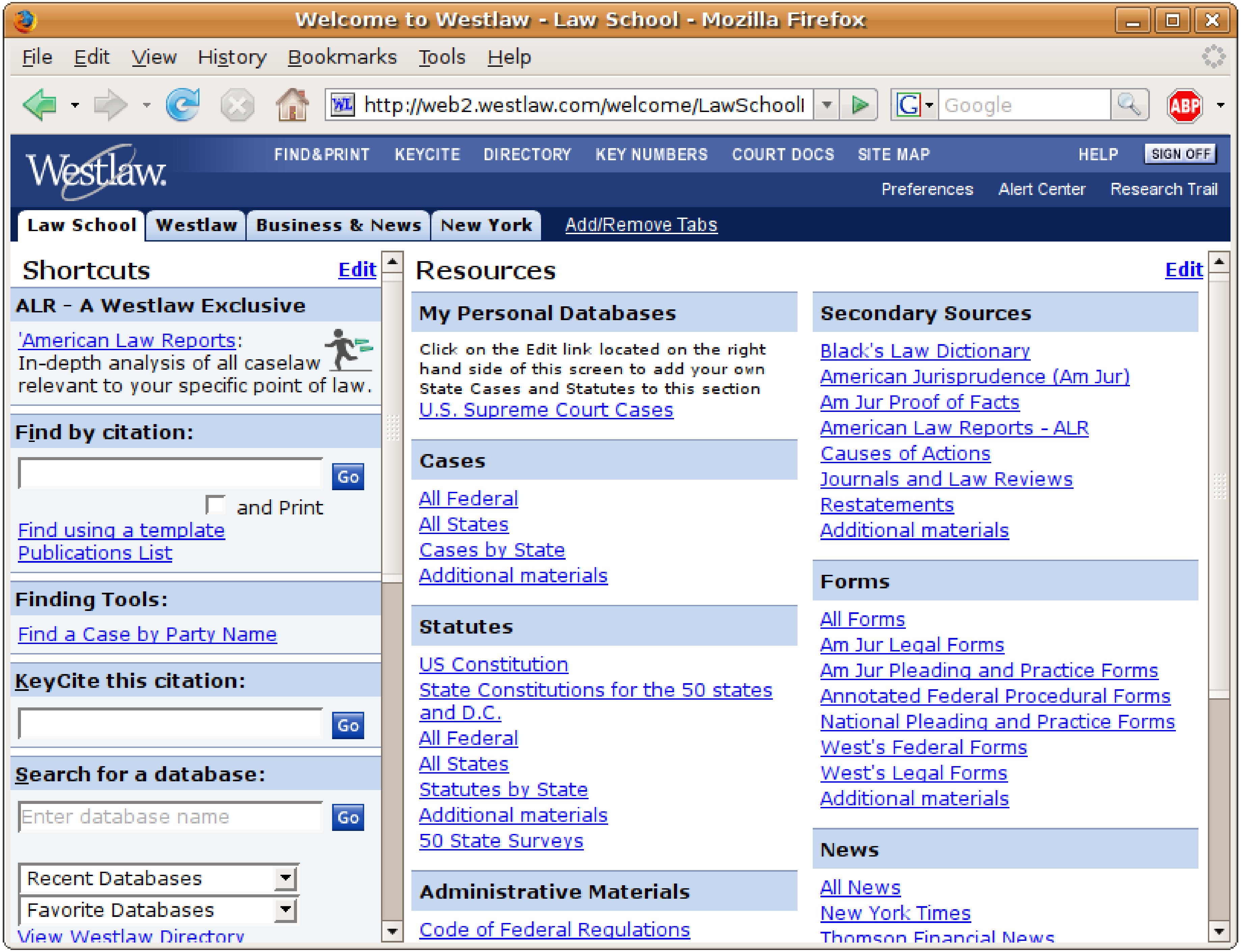
Hadoop, Clojure, and the Properties Pattern

NoSQL NYC
Monday, October 5, 2009

Stuart Sierra, AltLaw.org


Welcome to Westlaw - Law School - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://web2.westlaw.com/welcome/LawSchoolI
Google
Westlaw.
FIND&PRINT KEYCITE DIRECTORY KEY NUMBERS COURT DOCS SITE MAP HELP SIGN OFF
Preferences Alert Center Research Trail
Law School Westlaw Business & News New York Add/Remove Tabs

Shortcuts

Edit

ALR - A Westlaw Exclusive

'American Law Reports: In-depth analysis of all caselaw relevant to your specific point of law.

Find by citation:

Go

□ and Print

Find using a template
Publications List

Finding Tools:

Find a Case by Party Name

KeyCite this citation:

Go

Search for a database:

Enter database name Go

Recent Databases
Favorite Databases
View Westlaw Directory

Resources

Edit

My Personal Databases

Click on the Edit link located on the right hand side of this screen to add your own State Cases and Statutes to this section

U.S. Supreme Court Cases

Cases

All Federal
All States
Cases by State
Additional materials

Statutes

US Constitution
State Constitutions for the 50 states and D.C.
All Federal
All States
Statutes by State
Additional materials
50 State Surveys

Administrative Materials

Code of Federal Regulations

Secondary Sources

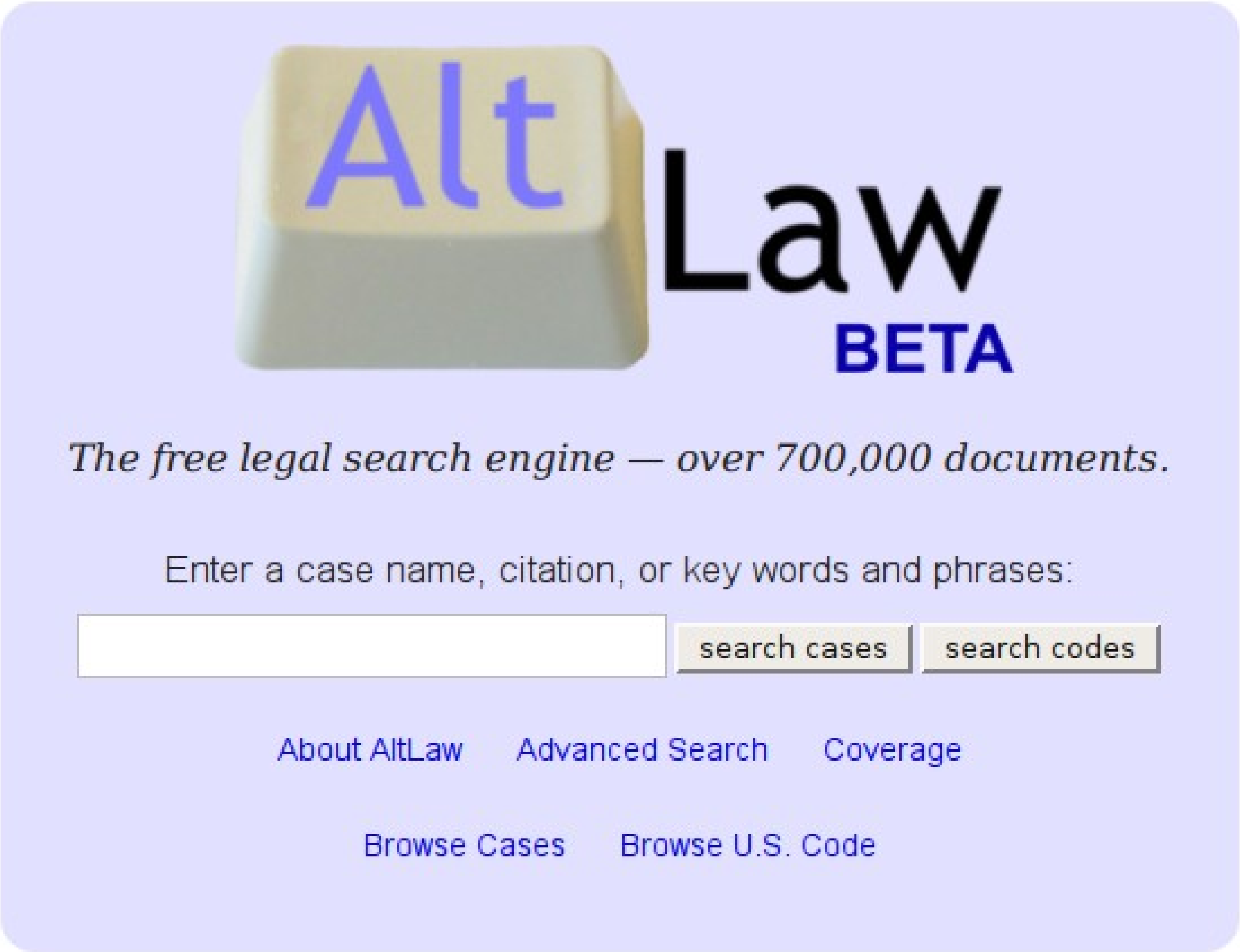
Black's Law Dictionary
American Jurisprudence (Am Jur)
Am Jur Proof of Facts
American Law Reports - ALR
Causes of Actions
Journals and Law Reviews
Restatements
Additional materials

Forms

All Forms
Am Jur Legal Forms
Am Jur Pleading and Practice Forms
Annotated Federal Procedural Forms
National Pleading and Practice Forms
West's Federal Forms
West's Legal Forms
Additional materials

News

All News
New York Times
Thomson Financial News



The free legal search engine — over 700,000 documents.

Enter a case name, citation, or key words and phrases:

Data Sources – Large Corpora

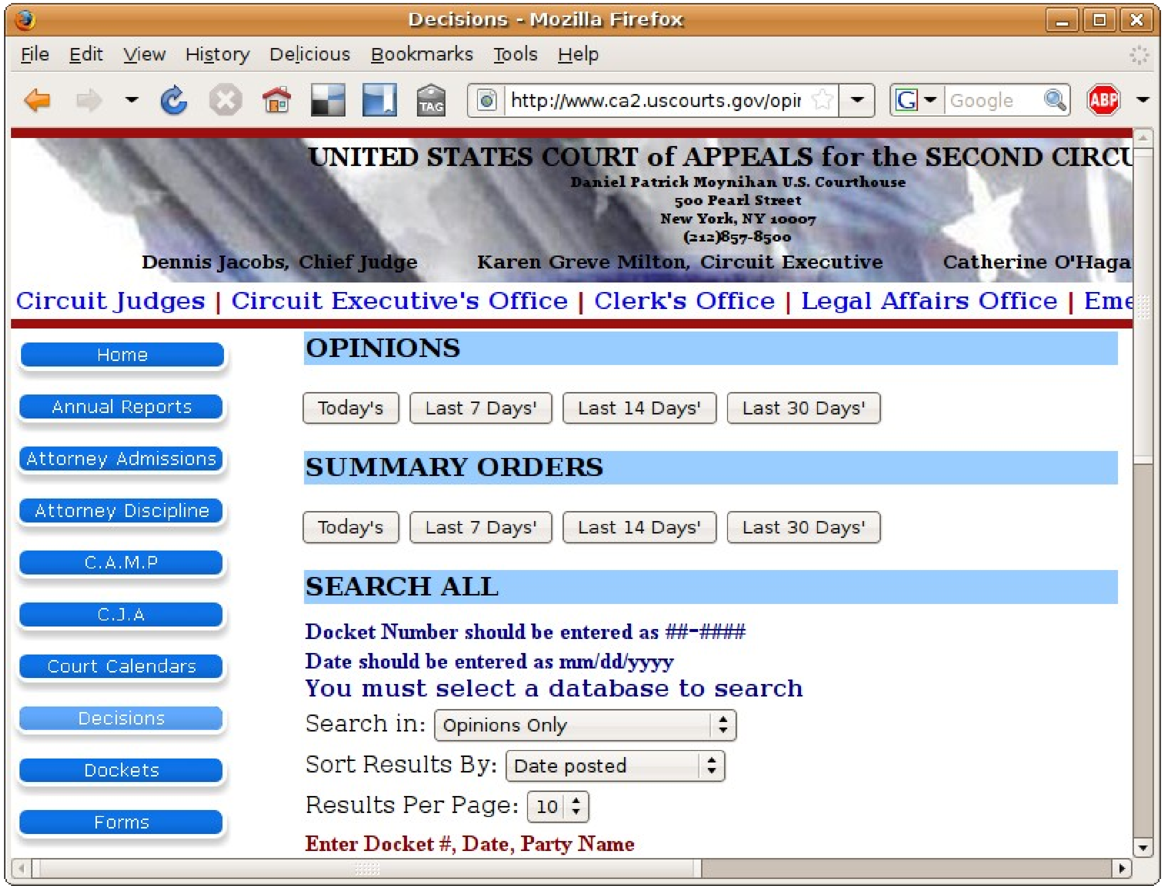
- Paul Ohm's corpus, http://bulk.altlaw.org/
  - 7 GB, 200,000+ files harvested from court web sites
- Cornell U.S. Code
  - 748 MB of XML
- http://bulk.resource.org/courts.gov/c/
  - 2 GB, 700,000+ federal cases, XHTML
- http://pacer.resource.org/
  - 736 GB, 2.7 million PDFs, 1.8 million HTML files
- Federal Register XML

Data Sources – Court Web Sites

www.supremecourtus.gov
www.ca1.uscourts.gov
www.ca2.uscourts.gov
www.ca3.uscourts.gov
www.ca4.uscourts.gov
www.ca5.uscourts.gov
www.ca6.uscourts.gov
. . .

14 appeals courts total
94 district courts
?? state courts
?? local/other courts

- 20-40 new cases daily
- PDF, WordPerfect, HTML, plain text


Decisions - Mozilla Firefox
File Edit View History Delicious Bookmarks Tools Help
http://www.ca2.uscourts.gov/opir
Google
Circuit Judges | Circuit Executive's Office | Clerk's Office | Legal Affairs Office | Eme
Home
Annual Reports
Attorney Admissions
Attorney Discipline
C.A.M.P
C.J.A
Court Calendars
Decisions
Dockets
Forms
Today's
Last 7 Days'
Last 14 Days'
Last 30 Days'

UNITED STATES COURT of APPEALS for the SECOND CIRCU
Daniel Patrick Moynihan U.S. Courthouse
500 Pearl Street
New York, NY 10007
(212)857-8500

Dennis Jacobs, Chief Judge | Karen Greve Milton, Circuit Executive | Catherine O'Haga

OPINIONS

SUMMARY ORDERS

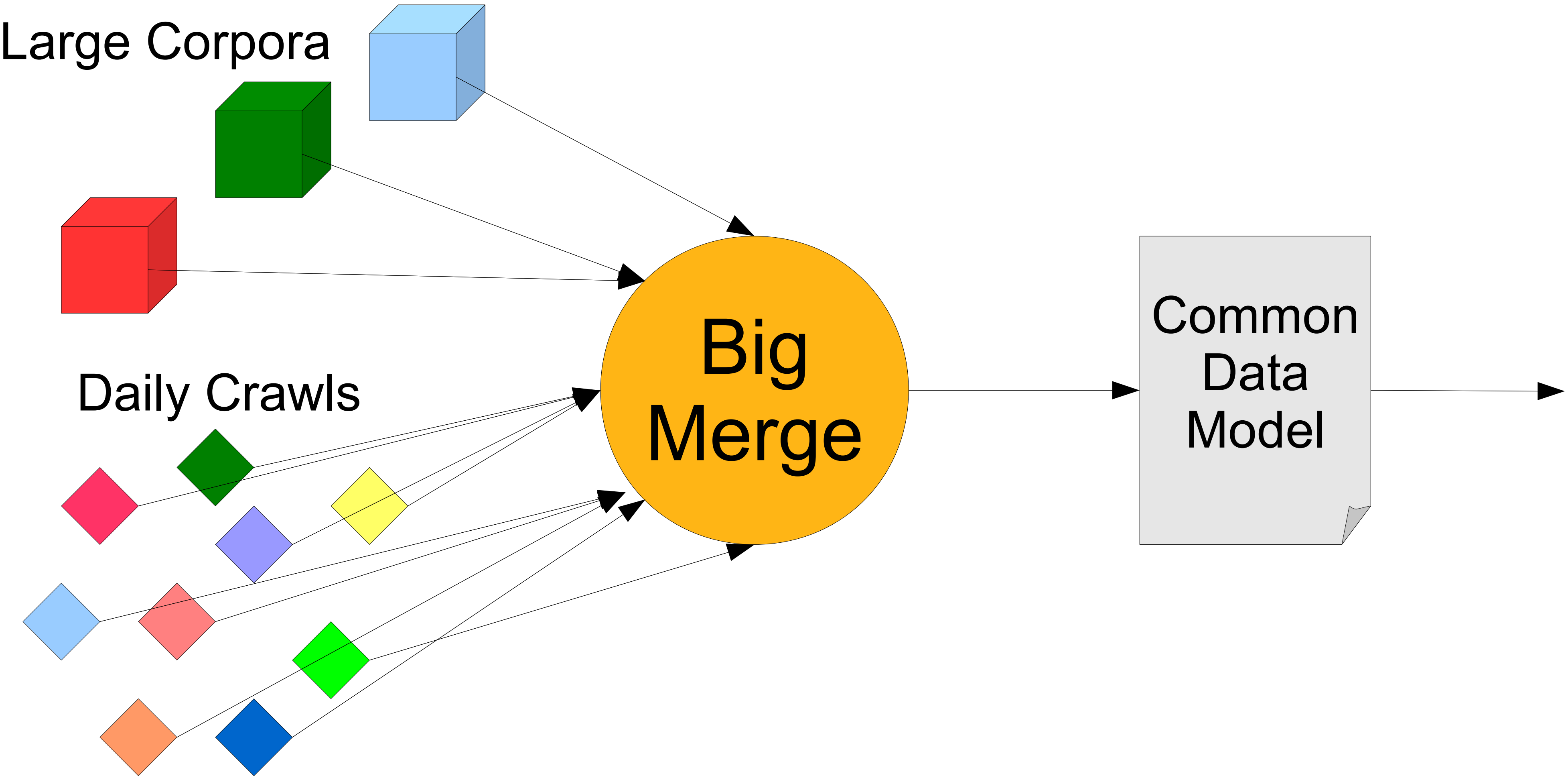
SEARCH ALL

Docket Number should be entered as ##-####
Date should be entered as mm/dd/yyyy
You must select a database to search

Search in: Opinions Only
Sort Results By: Date posted
Results Per Page: 10

Enter Docket #, Date, Party Name

AltLaw (1)



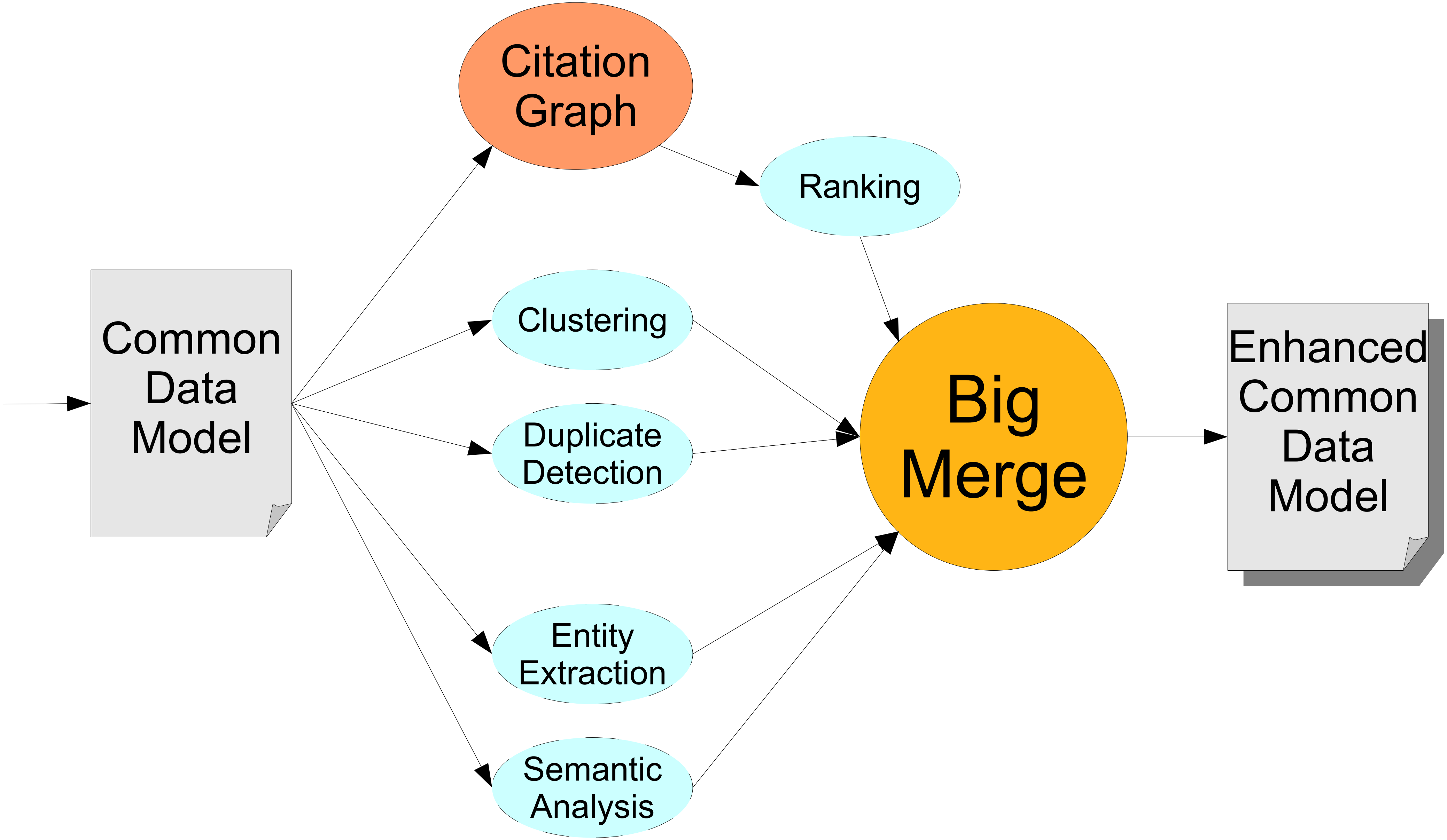
Large Corpora

Daily Crawls

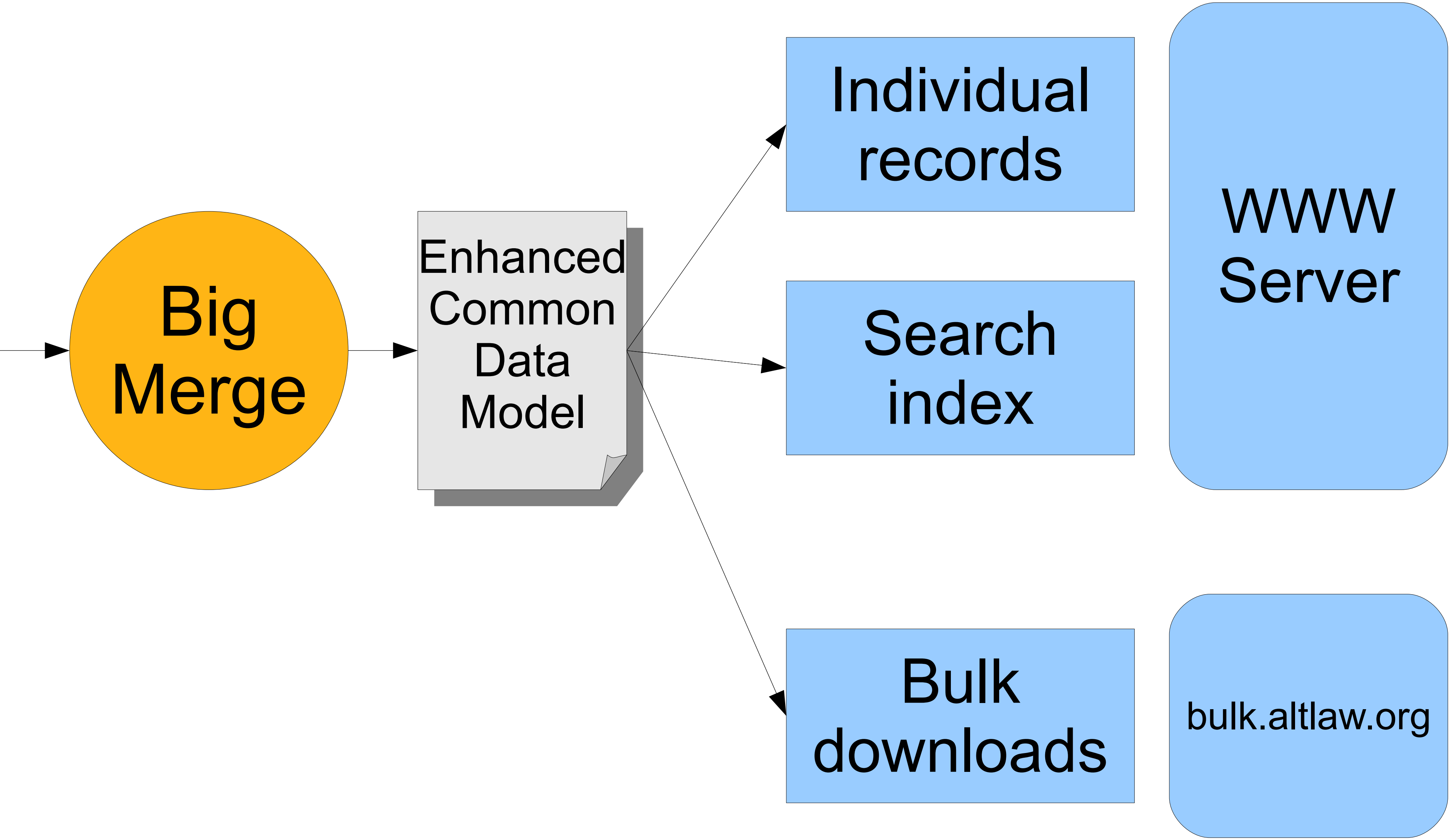
Big Merge

Common Data Model

AltLaw (2)


Citation Graph
Ranking
Common Data Model
Clustering
Duplicate Detection
Big Merge
Enhanced Common Data Model
Entity Extraction
Semantic Analysis

AltLaw (3)



Big Merge

Enhanced Common Data Model

Individual records

Search index

Bulk downloads

WWW Server

bulk.altlaw.org

The Grand Unified Data Model

- Key-value pairs? (files, Berkeley DB)
- Documents? (Solr/Lucene, CouchDB)
- Trees? (XML, JSON, Objects)
- Graphs? (RDF, triple stores)
- Tables? (SQL)

hadoop

- “Disk is the new tape.”
  - NO random access
  - NO disk seeks
  - Run at full disk transfer rate, not seek rate
- Data must be splittable
- Process each record in isolation

```java
public static class MapClass extends MapReduceBase
  implements Mapper<LongWritable, Text, Text, IntWritable> {

  private final static IntWritable one = new IntWritable(1);
  private Text word = new Text();

  public void map(LongWritable key, Text value,
                  OutputCollector<Text, IntWritable> output,
                  Reporter reporter) throws IOException {
    String line = value.toString();
    StringTokenizer itr = new StringTokenizer(line);
    while (itr.hasMoreTokens()) {
      word.set(itr.nextToken());
      output.collect(word, one);
    }
  }
}
```

```java
public static class Reduce extends MapReduceBase
  implements Reducer<Text, IntWritable, Text, IntWritable> {

  public void reduce(Text key, Iterator<IntWritable> values,
                     OutputCollector<Text, IntWritable> output,
                     Reporter reporter) throws IOException {
    int sum = 0;
    while (values.hasNext()) {
      sum += values.next().get();
    }
    output.collect(key, new IntWritable(sum));
  }
}
```

Clojure

- a new Lisp,
  neither Common Lisp nor Scheme
- Dynamic, Functional
- Immutability and concurrency
- Hosted on the JVM
- Open Source (Eclipse Public License)

Clojure Collections

List	`(print :hello "NYC")`
Vector	`[:eat "Pie" 3.14159]`
Map	`{:lisp 1  "The Rest" 0}`
Set	`#{2 1 3 5 "Eureka"}`

Homoiconicity

```
public void greet(String name) {
  System.out.println("Hi, " + name);
}

greet("New York");
Hi, New York
```

```
(defn greet [name]
  (println "Hello," name))

(greet "New York")
Hello, New York
```

```
(mapper key value)
```

↳ *list of key-value pairs*

```
(reducer key values)
```

↳ *list of key-value pairs*

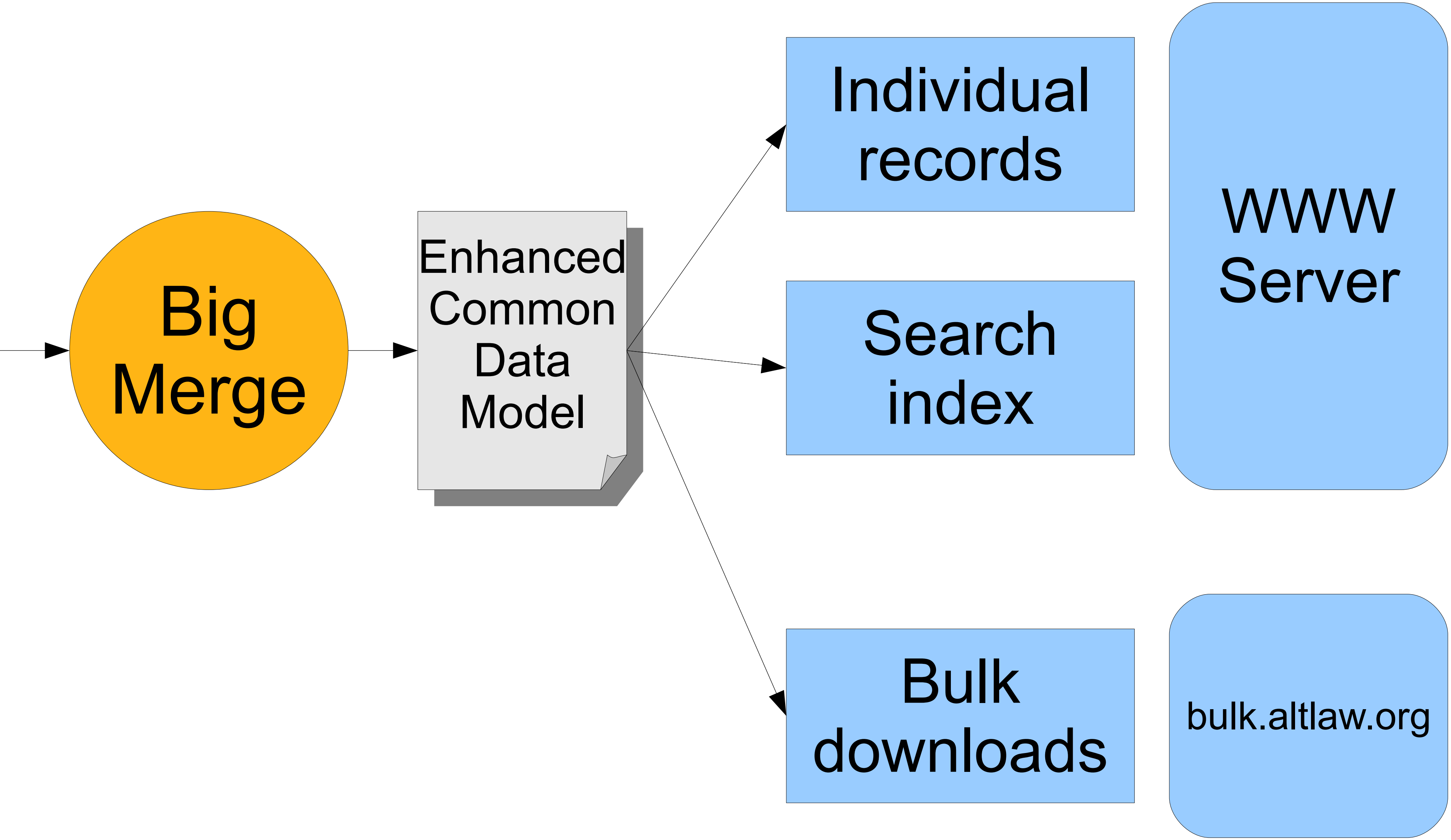
```
public static class MapClass extends MapReduceBase
  implements Mapper<LongWritable, Text, Text, IntWritable> {

  private final static IntWritable one = new IntWritable(1);
  private Text word = new Text();

  public void map(LongWritable key, Text value,
                  OutputCollector<Text, IntWritable> output,
                  Reporter reporter) throws IOException {
    String line = value.toString();
    StringTokenizer itr = new StringTokenizer(line);
    while (itr.hasMoreTokens()) {
      word.set(itr.nextToken());
      output.collect(word, one);
    }
  }
}
```

```
public static class Reduce extends MapReduceBase
  implements Reducer<Text, IntWritable, Text, IntWritable> {

  public void reduce(Text key, Iterator<IntWritable> values,
                     OutputCollector<Text, IntWritable> output,
                     Reporter reporter) throws IOException {
    int sum = 0;
    while (values.hasNext()) {
      sum += values.next().get();
    }
    output.collect(key, new IntWritable(sum));
  }
}
```

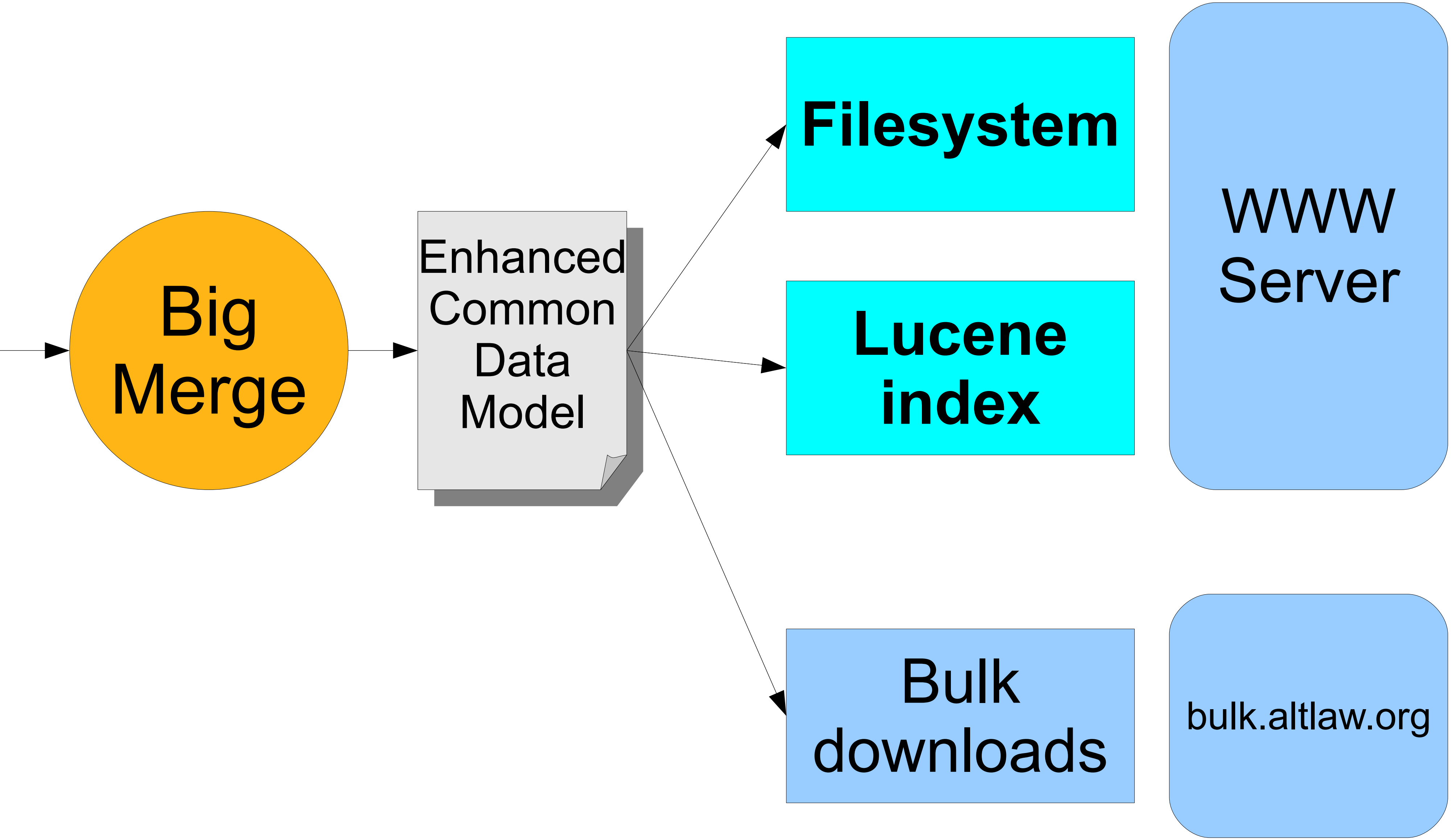
Clojure-Hadoop

```
(defn my-map [key val]
 (map (fn [token] [token 1])
      (enumeration-seq (StringTokenizer. val))))

(defn my-reduce [key values]
  [[key (reduce + values)]])

(defjob job
  :map my-map
  :map-reader int-string-map-reader
  :reduce my-reduce
  :inputformat :text)
```

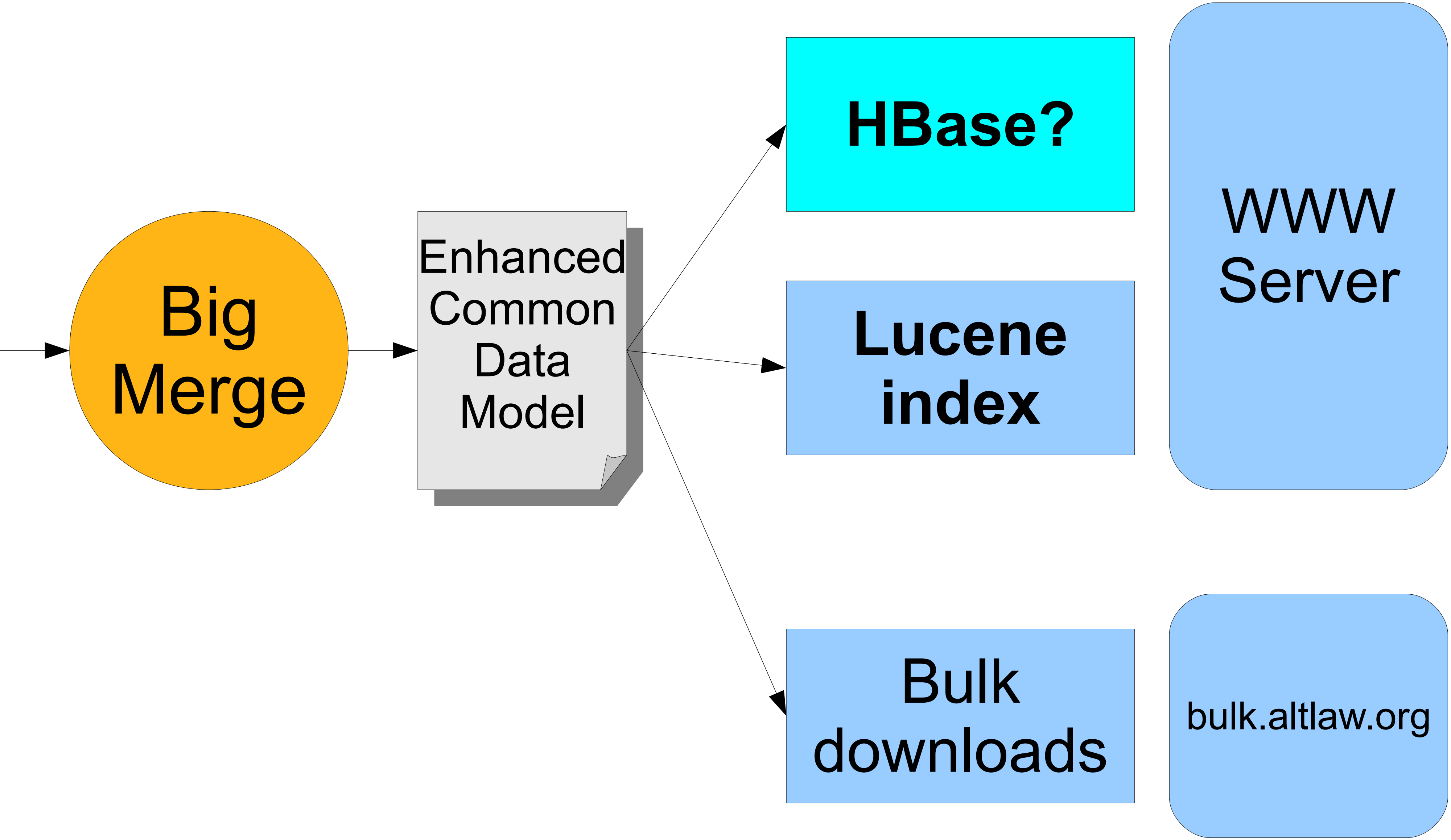
AltLaw (3)


Big Merge
Enhanced Common Data Model
Individual records
Search index
Bulk downloads
WWW Server
bulk.altlaw.org

AltLaw (3)


Big Merge
Enhanced Common Data Model
Filesystem
Lucene index
Bulk downloads
WWW Server
bulk.altlaw.org

AltLaw (3)



Big Merge

Enhanced Common Data Model

HBase?

Lucene index

Bulk downloads

WWW Server

bulk.altlaw.org

The Grand Unified Data Model

- Key-value pairs? (files, Berkeley DB)
- Documents? (Solr/Lucene, CouchDB)
- Trees? (XML, JSON, Objects)
- Graphs? (RDF, triple stores)
- Tables? (SQL)

Properties & RDF

```
{:uri     "http://id.altlaw.org/doc/101"
 :type    :Document
 :docid   101
 :title   "National Bank v. U.S."
 :cite    #{"101 U.S. 1" "25 L.Ed. 979"} }
```

```
<http://id.altlaw.org/doc/101>
       <rdf:type>   <alt:Document> ;
       <alt:docid>  "101"^xsd:integer ;
       <alt:title>  "National Bank v. U.S." ;
       <alt:cite>   "101 U.S. 1" ;
       <alt:cite>   "25 L.Ed. 979" .
```

The Properties Pattern:
http://steve-yegge.blogspot.com/2008/10/universal-design-pattern.html

More

- http://clojure.org/
- Google Groups: Clojure
- #clojure on irc.freenode.net & Twitter

- http://stuartsierra.com/
- @stuartsierra on Twitter
- http://github.com/stuartsierra
- http://www.altlaw.org/
- http://lawcommons.org/