Enterprise Search

Solr

BLUELID TECHNOLOGIES

Outline

- Overview
  - Problem Statement
  - Solution Process
- Live End to End Demo
- Technical Details
- Question and Answers

BLUELID TECHNOLOGIES

Problem Statement

The Iowa Legislative Services Agency (LSA) is responsible for the delivery and dissemination of numerous publications. Many of these publications are produced in multiple formats (PDF, Word, HTML) and information from one publication is often repeated in another publication. The LSA was looking for a way to repurpose their content to avoid rekeying information from one publication in to another, as well as organize their content in such a manner that consumers of the publications (citizens, members, staff) could quickly find content across all publications by subject matter or keyword search and better understand the interrelationships between each of the publications.

BLUELID TECHNOLOGIES

Solution Process

- Convert the content in to a structured XML format (author in xml moving forward)
- Create a unified taxonomy
- Perform interactive and automated indexing of all content
- Attach additional metadata to content (Collections, Notes)
- Publish content and all metadata to Solr

BLUELID TECHNOLOGIES

Solution Process

Convert the content in to an XML format

- Identify parts of the content that hold special meaning or need to be repurposed
  - Headnotes (Give user a summary of content)
  - Identifiers (Provide quick navigation to parts of content)
  - Paragraphs
  - References (Link to other content)
- Define a schema that isolates each of those pieces in to separate tags
- Apply unique id attributes to tags that need to support metadata

Solution Process

Need for unified taxonomy

- Multiple publications: Code of Iowa, Administrative Code, Acts, Bills, Amendments, Resolutions, etc…
- Independent taxonomies for each publication (index terms)
- Inconsistent classifications (Pig vs. Swine)
- Difficult and time consuming to maintain

BLUELID TECHNOLOGIES

Solution Process

Create a unified taxonomy

Together with Access Innovation, a leading company in taxonomy creation and term classification, the Iowa LSA Indexing staff reviewed all of their content and created a unified master taxonomy for use across all publications.

Reference to pigs in the Code of Iowa or in the Iowa Acts for example would now be categorized as “Porcine Animals” with synonyms, and related terms defined allowing users to quickly find the correct categorization upon searching for any of the popular terminologies associated with pigs such as hogs, swine, razorbacks, boars, pineywoods, rooters, etc…).

BLUELID TECHNOLOGIES

Solution Process

Interactive and automated indexing

Due to budget constraints, publication deadlines, and growing size of published content, manual indexing of all content quickly became unfeasible.

Using tools developed by BlueLid Technologies the Indexing staff was able to quickly submit all content, including legacy content dating back to the 1840s, to the Access Innovation's *Maistro* application for machine aided indexing.

Maistro analyzed the text of each document, and returned classifications chosen from the master taxonomy based on the documents text. These index terms where then automatically related to the appropriate locations in the content via the id attributes present in the xml.

The Indexing staff was then able to quickly review the applied terms and make corrections to the classifications for those documents that required a slightly refined taxonomy classification due to nuances in the language that *Maistro* did not detect.

BLUELID TECHNOLOGIES

Solution Process

Attach additional metadata to content

Since each important piece of content is nested in its own xml tag with an associated id any related metadata can now be attached to granular levels of the document without modifying or injecting that information directly in to the document.

These ids allow countless users, whether staff or citizens of Iowa, to annotate (add personal/shared notes) and organize content (create collections) in ways most meaningful to them.

Content chunks can also be flagged as belonging to other publications and those flagged pieces can automatically be pulled in and repurposed from their master source without re-keying or even copy paste.

BLUELID TECHNOLOGIES

Solution Process

Publish content and metadata to Solr

- Define a Solr document schema
  - Fields that define your documents
  - How those fields are parsed
- Define Solr data import handlers
  - Invoke the index build via simple get requests
- Define Solr search handlers

Outline

- Overview
  - Problem Statement
  - Solution Process
- Live End to End Demo
- Technical Details
- Question and Answers

BLUELID TECHNOLOGIES

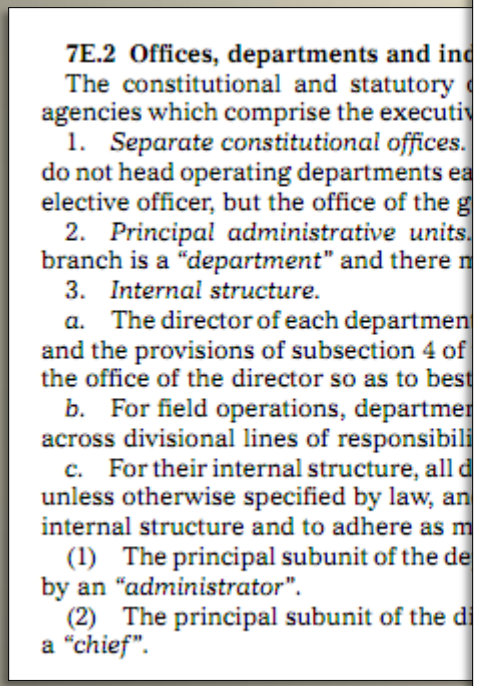
Outline

- Overview
  - Problem Statement
  - Solution Process
- Live End to End Demo
- Technical Details
- Question and Answers

BLUELID TECHNOLOGIES

Technical Details

Convert the content in to an XML format



7E.2 Offices, departments and ind

The constitutional and statutory
agencies which comprise the executiv

1. *Separate constitutional offices.*

do not head operating departments ea
elective officer, but the office of the g

2. *Principal administrative units.*

branch is a "*department*" and there n

3. *Internal structure.*

a. The director of each departmen
and the provisions of subsection 4 of
the office of the director so as to best

b. For field operations, departmer
across divisional lines of responsibili

c. For their internal structure, all d
unless otherwise specified by law, an
internal structure and to adhere as m

(1) The principal subunit of the de
by an "*administrator*".

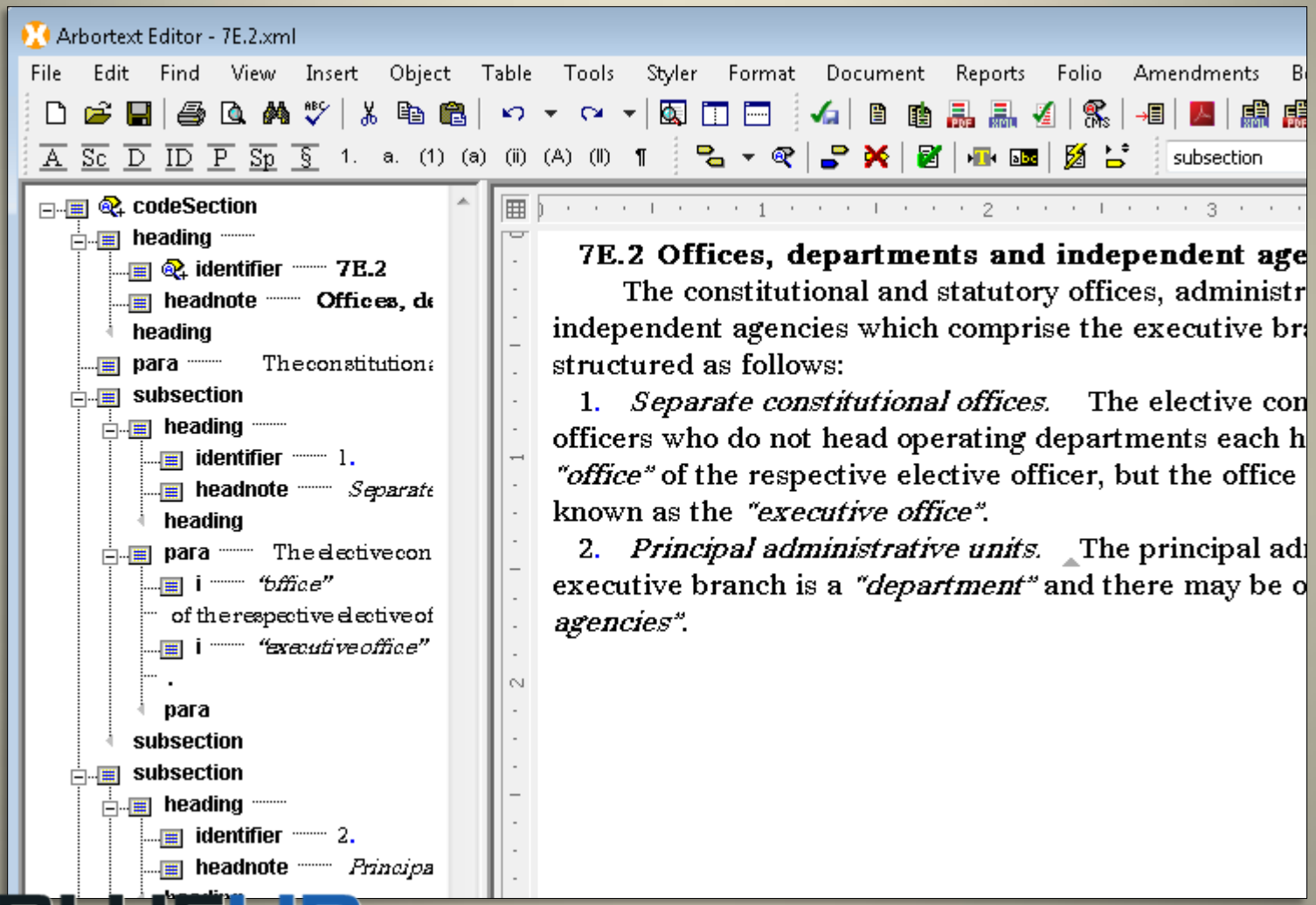
(2) The principal subunit of the di
a "*chief*".

```
<codeSection cms-id="42C2F55F-78F8-4D0F-93F5-A1D8F260EE91">
  <heading>
    <identifier>7E.2</identifier>
    <headnote>Offices, departments and independent agencies.</headnote>
  </heading>
  <para cms-id="B6ADD04B-3DB1-4C06-B3E4-BF2F9571F2D7">
    The constitutional and statutory offices, administrative departments,
    and independent agencies which comprise the executive branch of state
    government are structured as follows:
  </para>
  <subsection cms-id="819D4BE9-8E7E-48EF-91D9-E9EAB9B3124A">
    <heading>
      <identifier>1</identifier>
      <headnote>Separate constitutional offices.</headnote>
    </heading>
    <para cms-id="719CD8B5-63AD-4812-A6BA-4790B847C7A5">
      The elective constitutional and statutory officers who do not head
      operating departments each head a staff to be termed the <i>"office"</i>
      of the respective elective officer, but the office of the governor
      shall be known as the <i>"executive office"</i>.
    </para>
  </subsection>
  <subsection cms-id="9D1C062C-A63C-4FF7-9CA3-E3F22C625713">
    <heading>
      <identifier>2</identifier>
      <headnote>Principal administrative units.</headnote>
    </heading>
```

BLUELID TECHNOLOGIES

Technical Details

Author content in XML moving forward



BLUELID TECHNOLOGIES

Technical Details

Association of terms to content

	xml_id	content_as_xml	content_size
13	E42D1228-D932-42DC-A328-0000AF22BB06	<codeSection cms-id="E6635E36-A30A-48FE-9303-B64B...	5904
14	77E210FF-790A-4EF7-8503-0000BFEDEB09	<codeSection cms-id="708A02E4-BD17-4CD3-A9A5-3AA9...	5904
15	39D3EE27-5A05-4AE7-A01C-000125CA93A5	<bill xmlns:atipl="http://www.arbortext.com/namespace/P...	7455
16	211373A4-31FC-4170-9270-000178C735C8	<slSummary xmlns:"	
17	10B2BFBE-87C1-4D70-89AE-0001926FEBA5	<amendment xmln	
18	D066FB7B-28B6-45DF-8881-0001F87F9261	<bill xmlns:xsi="htt	
19	A53005A2-665A-466D-A8F7-000257460EEA	<codeSection cms	
20	1CEB1847-3074-482C-AD25-0002A7E4C190	<codeSection cms	
21	CBD05335-C0DF-4678-8CC1-0002DBE08FBC	<codeSection cms	
22	5B83FE6C-7F88-4C47-A972-00034246EB7B	<bill xmlns:xsi="htt	
23	0EE6A97F-C822-46ED-977B-00034C8452CE	<bill xmlns:atipl="h	
24	CFE51A28-378D-4		
25	7A341B95-3BD2		
26	BA700099-787E-		
27	E0F93162-BEDD		
28	9A2206FB-5E1B-		

	xml_id	cms_id
1	56263C92-CE88-41F3-9533-00001645C46B	405B3235-E766-45D6-B3A0-E82A3300CEA9
2	56263C92-CE88-41F3-9533-00001645C46B	83E0E9F7-46D6-427F-9C0B-E8F2C5B91CAD
3	9BB7A8C0-F87E-46B5-8FDA-00001B1A45FF	DAC824BE-6C1E-4520-ABCE-4F8EF1A6EAB0
4	9BB7A8C0-F87E-46B5-8FDA-00001B1A45FF	B4B7E00F-78E9-4F11-9311-DFE62EE1F29E
5	26D02D60-7DB3-475E-BA97-000035D80311	4EA7131C-A9B6-0B32-D7A4-5992FBAF1AB7
6	26D02D60-7DB3-475E-BA97-000035D80311	B749B2BE-BB4A-454A-2742-85BE1C8E2401
7	EDC65575-B874-4C0B-AB63-00003EA6CA11	D47192BB-A8DB-4700-AF90-793ECE69CA4C
	)3EA6CA11	4BDABCED-8B36-4712-B104-C3781FA9286F
	3F49F28C	D1AD4E47-13B4-4E3B-89D8-017D8CBC8541
	3F49F28C	00784D09-4C0A-428A-8A14-02EC7BE1B5DB
	3F49F28C	7E1438FD-FEB4-4A37-B8B2-10AF52A70102
	3F49F28C	5E5AFECB-B99A-49A0-8BB2-1540098E1236
	3F49F28C	58D935F9-2601-405B-A2A1-16752FD4BA29
	3F49F28C	277CD187-5C34-43CA-AC39-192E995793EE

	cms_id	term
10	60920241-53C0-472D-8812-AC618B6491E3	Commerce Department
11	60920241-53C0-472D-8812-AC618B6491E3	Sales
12	60920241-53C0-472D-8812-AC618B6491E3	Violations
13	1A1170EA-986A-41E1-9FF0-02D1F8EF1321	Alcoholic beverages
14	1A1170EA-986A-41E1-9FF0-02D1F8EF1321	Sales
15	1A1170EA-986A-41E1-9FF0-02D1F8EF1321	Wholesale
16	1A1170EA-986A-41E1-9FF0-02D1F8EF1321	Breweries
17	B8122B0E-A7BD-42CB-8231-76F9A4C49A63	Violations
18	B8122B0E-A7BD-42CB-8231-76F9A4C49A63	Controlled substances
19	B8122B0E-A7BD-42CB-8231-76F9A4C49A63	Pharmacy

BLUELID TECHNOLOGIES

Technical Details

Define Solr schema

```
    <field name="id" type="string" indexed="true" stored="true" required="true" />
    <field name="content_id" type="string" indexed="true" stored="true" required="true" />
    <field name="version" type="int" indexed="false" stored="true" required="true" />
    <field name="major_version" type="int" indexed="false" stored="true" required="true" />
    <field name="label" type="string" indexed="true" stored="true"/>
    <field name="type" type="string" indexed="true" stored="true" required="true"/>
    <field name="subtype" type="string" indexed="true" stored="true" required="true"/>
    <field name="type_id" type="int" indexed="true" stored="true" required="true"/>
    <field name="status" type="string" indexed="true" stored="true" required="true"/>
    <field name="name" type="string" indexed="true" stored="true" required="true"/>
    <field name="identifier" type="string" indexed="true" stored="true" required="true"/>
    <field name="headnote" type="text" indexed="true" stored="true"/>
    <field name="created_by" type="textgen" indexed="true" stored="true"/>
    <field name="created_date" type="date" indexed="true" stored="true"/>
    <field name="modified_by" type="textgen" indexed="true" stored="true"/>
    <field name="modified_date" type="date" indexed="true" stored="true"/>
    <field name="cache_path" type="string" indexed="false" stored="true"/>
    <field name="path" type="string" indexed="true" stored="true"/>
    <field name="index_term" type="string" indexed="true" stored="true" multiValued="true"/>

<!-- Mainly used by ACO for storing divisionHeadings for Chapters and titleHeading for agency -->
    <field name="subtitles" type="text" indexed="true" stored="false" />

<!-- catchall field, containing all other searchable text fields (implemented
    via copyField further on in this schema  -->
    <field name="text" type="text" indexed="true" stored="true" multiValued="true"/>
```

BLUELID TECHNOLOGIES

Technical Details

Define Solr data import handler

```
<dataConfig>
    <dataSource type="JdbcDataSource" name="aco" driver="com.microsoft.sqlserver.jdbc.SQLServerDriver"
    <dataSource type="JdbcDataSource" name="ico" driver="com.microsoft.sqlserver.jdbc.SQLServerDriver"
    <document name="content">

        <entity name="content" pk="doc_name" dataSource="aco" query="SELECT * FROM View_IndexChapter"
            <field column="doc_name" name="id" />
            <field column="doc_name" name="content_id" />
            <field column="version" name="version" />
            <field column="version" name="major_version" />
            <field column="content_type_id" name="type_id" />
            <field column="content_subtype" name="subtype" />
            <field column="content_type" name="type" />
            <field column="doc_name" name="name" />
            <field column="identifier" name="identifier" />
            <field column="status" name="status" />
            <field column="path" name="path" />
            <field column="headnote" name="headnote" />
            <field column="created_by" name="created_by" />
            <field column="created_date" name="created_date" />
            <field column="last_updated_by" name="modified_by" />
            <field column="last_updated_date" name="modified_date" />
            <field column="text" name="text" />
            <field column="subtitles" name="subtitles" />
            <field column="agency_number" name="agency_number" />
            <field column="agency_name" name="agency_name" />
            <field column="pub_date" name="pub_date" />
            <field column="cache_path" name="cache_path" />

            <entity name="terms" dataSource="ico" query="select term from View_SearchIndex where other
                <field column="term" name="index_term" />
            </entity>
```

BLUELID TECHNOLOGIES

Technical Details

Define Solr search handlers

```
<requestHandler name="/search" class="linc.servlet.SearchServlet">
    <lst name="defaults">
        <str name="echoParams">explicit</str>

                <!-- VelocityResponseWriter settings -->
        <str name="wt">velocity</str>
        <str name="v.contentType">text/html;charset=UTF-8</str>
        <str name="v.template">search</str>
        <str name="v.layout">layout</str>
        <str name="v.properties">velocity.properties</str>
        <str name="title">Iowa Legislature - Advanced Search</str>

        <str name="defType">edismax</str>
        <str name="group">true</str>
        <str name="group.field">type</str>
        <str name="group.limit">5</str>
        <str name="q.alt">*:*</str>
        <str name="rows">10</str>
        <str name="fl">*,score</str>
        <str name="qf">text^1 name^3.5 headnote^2.5 type^2.0 id^10.0 modified_by^0.2 modified_date^0.2</str>

        <str name="facet">on</str>
        <str name="facet.field">index_term</str>
```

BLUELID TECHNOLOGIES

Outline

- Overview
  - Problem Statement
  - Solution Process
- Live End to End Demo
- Technical Details
- Question and Answers

BLUELID TECHNOLOGIES