Record Indexer Data Format

Introduction

In order to develop and test your Record Indexer, you will need to populate your system with project data, including user accounts, project definitions, batch images, known field data, and field help files. To help you in this process, we provide some pre-made data that you can import into your system. The provided data is in a ZIP file that contains several different types of files and directories. This document describes the format of the provided data. Understanding the data format will allow you to write a program that imports the data into your Record Indexer database.

ZIP File Contents

A data ZIP file contains a single top-level directory that is named after the data set contained in the ZIP. For example, the record-indexer-data.zip file, which you can find in the **Download** section of our 240 web site, contains the single, top-level directory *Records*. The top-level directory contains a single XML file (which is named after the data set, e.g. Records.xml), and three sub-directories named `fieldhelp`, `images`, and `knowndata`. The XML file contains information about all of the user accounts and projects in the data set. In turn, each project section in the XML file contains information about its batch images, records, and fields. The XML file contains references to the files in the `fieldhelp`, `images`, and `knowndata` sub-directories.

The `fieldhelp` sub-directory contains help files for any field (i.e. named column in an image we are extracting from) in any project in the data set. As an example, in the Records data set there are three kinds of images, for the three projects in the dataset: the 1890 images, the 1900 images, and the draft images. Each 1890 image has the fields: Last Name, First Name, Gender, and Age. The 1900 images have the fields: Gender, Age, Last Name, First Name, and Ethnicity. The draft images have the fields: Last Name, First Name, Age, and Ethnicity. To cover all five types of fields in the three kinds of images, the *fieldhelp* directory contains help files for: age, ethnicity, first name, gender, and last name. The field help files are in HTML format. Each file contains HTML-encoded instructions that explain to the user the meaning of a particular field, and how to properly enter the values for that field.

The `images` sub-directory contains batch image files for all of the projects. The batch image files are in PNG format. Each file contains the image for a single batch.

The *knowndata* sub-directory contains text files that list the known values for the various project fields in the data set. As an example, the *knowndata* sub-directory in the top-level *Records* directory contains the following files: 1890_first_names.txt, 1890_last_names.txt, 1900_first_names.txt, 1900_last_names.txt, draft_first_names.txt, draft_last_names.txt, ethnicities.txt, and genders.txt. As you can deduce, there is a different set of known first name for the 1890 and 1900 census and for the Draft project. Similarly, there is a different set of known last names. The set of known values for ethnicities and genders is the same for all three projects.

Each known data file is a text file containing the known values for a single field. The values are in a comma-separated list, as shown in the following BNF:

```
KNOWN_DATA_FILE ::= \s*VALUE\s*(,\s*VALUE\s*)*

VALUE ::= String
```

For example, both of the following files contain the values ‘Red’, ‘Green’, ‘Blue’, and ‘Bright Yellow’.

```
Red,Green,Blue,Bright Yellow
```

```
Red,     Green   ,

    Blue

    ,

Bright Yellow
```

XML File Format

The XML file contains information about all of the user accounts and projects in the data set. Each project may or may not already be indexed. If a project is already indexed, in addition to basic information about the project, the XML file also contains field values for the records in the project’s batches. This allows you to test your Record Indexer on projects that have already been indexed without having to do the indexing manually.

The overall format of the XML file is as follows:

```
<indexerdata>
     <users>
          <user> ... </user>
          ...
     </users>
     <projects>
          <project> ... </project>
          ...
      </projects>
</indexerdata>
```

The root element is <indexerdata>, which contains two sub-elements: <users> and <projects>. The <users> element contains one or more <user> elements. The <projects> element contains one or more <project> elements.

Users

Each <user> element contains the information for a single user of the system. The contents of the <user> element are depicted below:

```
<user>
      <username>sheila</username>
      <password>parker</password>
      <firstname>Sheila</firstname>
      <lastname>Parker</lastname>
      <email>sheila.parker@gmail.com</email>
      <indexedrecords>0</indexedrecords>
</user>
```

The <username>, <password>, <firstname>, <lastname>, and <email> elements have obvious meanings, and contain non-empty strings. The <indexedrecords> element contains a non-negative integer specifying the number of records that have been previously indexed by this user.

Projects

Each <project> element contains the information about a single project. The format of the <project> element is as follows:

```
<project>
      <title>Draft Records</title>
      <recordsperimage>7</recordsperimage>
      <firstycoord>195</firstycoord>
      <recordheight>65</recordheight>
       <fields>
            <field> ... </field>
            ...
      </fields>
       <images>
            <image> ... </image>
            ...
      </images>
</project>
```

The <title> element contains the project's title, and is a non-empty string.

The <recordsperimage> element contains a positive integer that indicates the number of records that appear on each batch image.

The <firstycoord> element contains a non-negative integer that specifies the y-coordinate of the top of the first record on the project's images. Coordinates are relative to the top-left corner of the image, zero-based, and measured in pixels.

The <recordheight> element contains a positive integer that specifies the height of each record in the project's images, measured in pixels. All records in the project's images have the same height.

The <fields> element contains information about the project's fields. It contains one or more <field> elements (described below).

The <images> element contains information about the project's images. It contains one or more <image> elements (described below).

Fields

Each <field> element contains the information about one of the project's fields. The contents of the <field> element are depicted below:

```
<field>
      <title>Last Name</title>
      <xcoord>75</xcoord>
      <width>325</width>
      <helphtml>fieldhelp/last_name.html</helphtml>
      <knowndata>knowndata/draft_last_names.txt</knowndata>
</field>
```

The <title> element contains the field's title, and is a non-empty string.

The <xcoord> element contains a non-negative integer that specifies the x-coordinate of the field on the project's images. Coordinates are relative to the top-left corner of the image, zero-based, and measured in pixels.

The <width> element contains a positive integer that specifies the width of the field in the project's images, measured in pixels.

The <helphtml> element specifies the location of an HTML file that contains end-user help for this field. The file's location is a relative path to a file in the ZIP file's `fieldhelp` sub-directory. The path is relative to the directory containing the XML file.

The <knowndata> element, which is OPTIONAL, specifies the location of a text file that contains known values for this field. The file's location is a relative path to a file in the ZIP file's `knowndata` sub-directory. The path is relative to the directory containing the XML file.

Images

Each <image> element contains information about one of the project's images. The contents of the <image> element are depicted below:

```
<image>
      <file>images/draft_image0.png</file>
      <records>
            <record> ... </record>
            ...
      </records>
</image>
```

The <file> element specifies the location of a PNG file that contains a single image for the project. The file's location is a relative path to a file in the ZIP file's images sub-directory. The path is relative to the directory containing the XML file.

The <records> element is OPTIONAL. If present, it contains a <record> element for each record on the image. Each <record> element contains the field values for a single record. (This is used to represent images that have already been indexed. For a given project, some images may have record values, and others may not. If an image has record values, it has already been indexed, and should not be given to another user for indexing. If an image does not have record values, it has not yet been indexed, and should be given to a user for indexing.)

The format of the <record> element is as follows:

```
<record>
    <values>
        <value>FOX</value>
        <value>RUSSELL</value>
        <value>19</value>
        <value>ALASKA NATIVE</value>
    </values>
</record>
```

The <values> element contains values for all of the fields in the record. Each field value is contained in a <value> element. Each value can be any string (including empty).

Sample XML File

```
<indexerdata>
  <users>
    <user>
      <username>test1</username>
      <password>test1</password>
      <firstname>Test</firstname>
      <lastname>One</lastname>
      <email>test1@gmail.com</email>
      <indexedrecords>0</indexedrecords>
    </user>
    <user>
      <username>test2</username>
      <password>test2</password>
      <firstname>Test</firstname>
      <lastname>Two</lastname>
      <email>test2@gmail.com</email>
      <indexedrecords>0</indexedrecords>
    </user>
  </users>
  <projects>
```

```
<project>
  <title>Draft Records</title>
  <recordsperimage>7</recordsperimage>
  <firstycoord>195</firstycoord>
  <recordheight>65</recordheight>
  <fields>
    <field>
      <title>Last Name</title>
      <xcoord>75</xcoord>
      <width>325</width>
      <helphtml>fieldhelp/last_name.html</helphtml>
      <knowndata>knowndata/draft_last_names.txt</knowndata>
    </field>
    <field>
      <title>First Name</title>
      <xcoord>400</xcoord>
      <width>325</width>
      <helphtml>fieldhelp/first_name.html</helphtml>
      <knowndata>knowndata/draft_first_names.txt</knowndata>
    </field>
    <field>
      <title>Age</title>
      <xcoord>725</xcoord>
      <width>120</width>
      <helphtml>fieldhelp/age.html</helphtml>
    </field>
    <field>
      <title>Ethnicity</title>
      <xcoord>845</xcoord>
      <width>450</width>
      <helphtml>fieldhelp/ethnicity.html</helphtml>
      <knowndata>knowndata/ethnicities.txt</knowndata>
    </field>
  </fields>
  <images>
    <image>
      <file>images/draft_image0.png</file>
      <records>
        <record>
          <values>
            <value>FOX</value>
            <value>RUSSELL</value>
            <value>19</value>
            <value>ALASKA NATIVE</value>
          </values>
        </record>
        <record>
          <values>
            <value>BARTLETT</value>
            <value>DAVE</value>
            <value>21</value>
            <value>HISPANIC</value>
```

```
            </values>
          </record>
          <record>
            <values>
              <value>ACOSTA</value>
              <value>JEROME</value>
              <value>28</value>
              <value>ALASKA NATIVE</value>
            </values>
          </record>
          <record>
            <values>
              <value>HALL</value>
              <value>JAY</value>
              <value>20</value>
              <value>NATIVE HAWAIIAN</value>
            </values>
          </record>
          <record>
            <values>
              <value>MILES</value>
              <value>LUTHER</value>
              <value>28</value>
              <value>NATIVE HAWAIIAN</value>
            </values>
          </record>
          <record>
            <values>
              <value>JARVIS</value>
              <value>LEROY</value>
              <value>24</value>
              <value>WHITE</value>
            </values>
          </record>
          <record>
            <values>
              <value>DUNN</value>
              <value>TED</value>
              <value>29</value>
              <value>BLACK</value>
            </values>
          </record>
        </records>
      </image>
    </images>
  </project>
 </projects>
</indexerdata>
```