EDA095
Uniform Resource Locators (URLs) and HTML Pages

Pierre Nugues

Lund University
http://www.cs.lth.se/home/Pierre_Nugues/

April 22, 2010

Covers: Chapter 7, pages 184-222, and Chapter 8, pages 248-266, *Java Network Programming*, 3rd ed., Elliotte Rusty Harold

The Web

Everybody knows what the web is. No need to present it.
When it started, it was a combination of three main features:

- A document format in plain text: HTML derived from the older and clumsy SGML
- A communication protocol to transfer data: HTTP
- A way to name and address objects on the network: URLs

We will review ways to handle URLs and HTML directly from Java.

Uniform Resource Locators (URL)

A URL is the name of a "resource" on the Internet and an access mode.
It goes beyond data that can be exchanged using HTTP.
It can extend to FTP, telnet, mail protocols, and many more.
In its simplest form, a URL is a string that consists of three parts:
`protocol://hostname/path/object`

- The `protocol` can be `http`, `ftp`, `telnet`, etc.
- The `hostname` is an Internet address such as `www.cs.lth.se` or `130.235.16.34`
- The `path/object` corresponds for instance to `/home/pierre/file.html`

URLs are defined by RFC 2396 and RFC 2732.
(`http://www.rfc-editor.org/`)

URL (Continued)

Protocols have a default port.
The URL can specify a new one:
`http://www.cs.lth.se:80/EDA095/index.html`
A URL can be absolute or relative to a base URL
The link `labs.html` in a document whose URL is
`http://www.cs.lth.se:80/EDA095/index.html`
has the URL
`http://www.cs.lth.se:80/EDA095/labs.html`

The URL Class

The URL class defines a way to build and parse URL strings Constructors:

- `URL(String spec)`
- `URL(String protocol, String host, String file)`
- `URL(URL context, String spec)`

The last one creates a URL by parsing the given spec within a specified context.

The algorithm is complex, read the reference:
`http://java.sun.com/javase/6/docs/api/java/net/URL.html#URL(java.net.URL,%20java.lang.String)`

The constructor throws a `MalformedURLException` if it fails to create the URL

A First Program Using the URL Class

Creating URLs

The constructor does not check the syntax. See MyDoc3

```
import java.net.*;
public class MyURL {
  public static void main(String[] args) {
    try {
      URL myDoc1 = new URL("http://www.cs.lth.se/index.html");
      URL myDoc2 = new URL("http", "www.cs.lth.se", "/index.html");
      URL myDoc3 = new URL("http", "www.cs.lth.se", "index.html");
      URL myDoc4 = new URL(new URL("http://www.cs.lth.se/EDA095/"),
        "labs.shtml");
    } catch (Exception e) { }
  }
} //MyURL.java
```

Client and Server Communications Using HTTP

The communication between a server and a client in its simplest form starts with a TCP connection from the client to port 80.
Then we have the exchange:

- The client sends a request
- The server sends a response.

Both request and response messages feature a header that consists of parameter-value pairs
In addition:

- The request consists of a command word (HTTP method) to identify the request, parameters, and possibly other data.
- The response consists of a response code and objects to be displayed or rendered by the client

HTTP Request

The request behind the URL
`http://www.cs.lth.se/~pierre/index.html` consists of:

1. HTTP method, URL, version

   `GET /~pierre/index.html HTTP/1.0`

2. Sequence of parameter names (46 types) followed by ':' and values – pairs `Name: Value`

   ```
   Accept: text/plain
    ...
   Host: www.cs.lth.se
   User-Agent: Mozilla/4.0
   ```

3. Empty line: `\r\n`
4. Possibly a message body (data) whose size is given by the `Content-Length` attribute

RFC 2616 (`http://www.ietf.org/rfc/rfc2616.txt`)

HTTP Response

Servers send a response: header followed by data

1. Protocol, status code, textual phrase

   ```
   HTTP/1.0 200 OK
   ```

2. Sequence of parameter names followed by ':' and values

   ```
   Date: Wed, 28 Mar 2007 12:12:54 GMT
   Server: Apache/2.0.52 (sparc-sun-solaris2.8)
   ...
   Connection: close
   ```

3. Empty line: `\r\n`
4. Data

   ```
   <html>
   ...
   </html>
   ```

Getting Data from a Server

In its simplest form, the client does not send data
It just requests the content of the URL file
For HTTP, the corresponding methods are `GET` or `POST`.
The method is sent automatically by the URL class and is hidden to the programmer.
We will review the steps to get data from simple to more complex
The URL class uses `GET` by default, hides the header details, and receives data

Getting Data from a Server

The `URL` class has methods to open a connection from a URL:

- `InputStream openStream()` //Opens a connection and returns an `InputStream` for reading from that connection.
- `URLConnection openConnection()` //Returns a `URLConnection` object.

The opening methods will connect to the resource using the specified protocol

Once open, the stream can be read and contain data sent by the server

The protocol machinery (http, ftp, etc.) is invisible to the programmer

We will first consider examples with `openStream()`

The combination of the `URL` constructor and `openStream()` uses the GET method

Getting Data Using openStream()

```
//ViewHTML.java
try {
  URL myDoc = new URL("http://www.cs.lth.se/index.html");
  InputStream is = myDoc.openStream();
  BufferedReader bReader =
    new BufferedReader(new InputStreamReader(is));
  String line;
  while ((line = bReader.readLine()) != null) {
    System.out.println(line);
  }
} catch (Exception e) {
  e.printStackTrace();
}
```

Relative URLs

Using the 3rd constructor, we can create URLs relatively to a context (MyURL2.java)

```
URL myDoc1 = new URL("http://www.cs.lth.se/index.html");
URL myDoc2 = new URL(myDoc1, "pierre.html");
URL myDoc3 = new URL(myDoc1, "/home/pierre.html");
URL myDoc4 = new URL(myDoc1, "http://www.lu.se/pierre.html");
URL myDoc5 = new URL(myDoc1, "mailto:pierre@cs.lth.se");
```

URL Exceptions

The description in the URL class exceptions

> *Throws:* `MalformedURLException` – *If the string specifies an unknown protocol.*

is not consistent with that of `MalformedURLException`:

> *Thrown to indicate that a malformed URL has occurred. Either no legal protocol could be found in a specification string or the string could not be parsed.*

The URL constructor only checks the protocol.

Exceptions

```
//MyURL3.java
package url;
import java.net.*;
public class MyURL {
  public static void main(String[] args) {
    try {
      URL myDoc1 = new URL("http://www.cs.lth.se/index.html");
      URL myDoc2 = new URL("http", "www.cs.lth.se", "index.html");
      URL myDoc3 =
        new URL(new URL("http://www.cs.lth.se/EDA095/"),
          "labs.shtml");
      URL myDoc4 = new URL("http:////www.cs;lth.se\\^/index.html");
    } catch (Exception e) {
      e.printStackTrace();
    }
```

Exceptions II

```
    try {
      URL myDoc5 = new URL("htp:////www.cs;lth.se\\^/index.html ");
    } catch (Exception e) {
      e.printStackTrace();
    }
    try {
      URL myDoc6 = new URL("htp", "www.cs.lth,se", "index.html");
    } catch (Exception e) {
      e.printStackTrace();
    }
  }
}
```

Methods of the URL Class

Some useful methods are:

- `String getProtocol()` //Gets the protocol name.
- `String getHost()` //Gets the host name.
- `int getPort()` //Gets the port number.
- `String getPath()` //Gets the path part and the file name.
- `String getFile()` //Gets the path, the file name, and the query if it exists. In fact nearly the same as `getFile()`, see documentation

Using URL Methods

```
//myURL4.java
URL myDoc1 = new URL("http://www.cs.lth.se/index.html");
URL myDoc2 = new URL("http", "www.cs.lth.se", "/index.html");
URL myDoc3 = new URL("http:////www.cs;lth.se\\^/index.html");
System.out.println("Protocols: "+ myDoc2.getProtocol() + " " +
  myDoc3.getProtocol());
System.out.println("Files: "+ myDoc2.getFile() + " " +
  myDoc3.getFile());
System.out.println("Ports: "+ myDoc2.getPort() + " " +
  myDoc3.getPort());
System.out.println("Hosts: "+ myDoc2.getHost() + " " +
  myDoc3.getHost());
```

Testing Available Protocols

Using the URL class, can we find protocols supported by a machine
Idea: Create a set of URL objects and call `getProtocol()`
Try: http, https, ftp, tftp, mailto, telnet, file, gopher, ladp, jar, nfs, jdbc, rmi, etc.
and catch the errors...
This is the idea of ProtocolTester.java, Elliotte Rusty Harold, *Java Network Programming*, page 186.
The code available here:
`http://www.cafeaulait.org/books/jnp3/examples/index.html`

Uniform Resource Identifiers (URI)

URIs are name conventions close to URLs (RFC 2396)
`[scheme:]scheme-specific-part[#fragment]`
However, URIs do not provide a method to access a network resource.
URIs can be absolute (with a scheme) or relative (without a scheme)
URIs can also be:

- Opaque: The scheme-specific part does not begin with a /. They are not parsed as: `mailto:java-net@java.sun.com`, `news:comp.lang.java`, `urn:isbn:096139210x`
- Hierarchical: They are absolute and begin with a / or they are relative and have no scheme as: `http://java.sun.com/j2se/1.3/` or `docs/guide/collections/designfaq.html#28`

Parsing a URI

A hierarchical URI is parsed according to the syntax

`[scheme:][//authority][path][?query][#fragment]`

If the authority is a server, its syntax is:

`[user-info@]host[:port]`

The user info can consist of a user name and a password, `anonymous:pierre@cs.lth.se` on an anonymous ftp server

URI Methods

The Java implementation of the `URI` class is more conformant than the URL one.
The URI class has methods to build a URI from its components.
Constructors:

- `URI(String str)`
- `URI(String scheme, String ssp, String fragment)`
- `URI(String scheme, String userInfo, String host, int port, String path, String query, String fragment)`

It has Booleans to determine whether it is:

- opaque or hierarchical, `isOpaque()`
- relative or absolute, `isAbslolute()`

URI Methods

The URI class has methods to parse a string (a hierarchical URI):

`[scheme:][//authority][path][?query][#fragment]`

- `String getScheme()`
- `String getAuthority()`
- `String getPath()`
- `String getQuery()`
- `String getFragment()`
- `URI parseServerAuthority()`, parses the authority, (`String getUserInfo()`, `String getHost()`, `int getPort()`)

URISplitter.java from Elliotte Rusty Harold, Java Network Programming, page 218.

```
java url.URISplitter http://www.cs.lth.se:80/pierre/index.html
```

Sending Data from the HTTP Client

Clients can send data to HTTP servers using a list of parameter-value pairs:

- book=Java Network Programming
- author=Harold

This is used when you fill in HTML forms, for instance
`GET` and `POST` are the two methods to carry this out
We review `GET` now

Sending Parameters with GET

GET sends the list of parameter-value pairs in the URL in the query part:

```
[scheme:][//authority][path][?query][#fragment]
```

The parameter list must comply with a specific format and encoding for the accents:

```
Arg1=Value1&Arg2=Value2
```

as in

```
book=Java+Network+Programming&author=Harold
```

The preferred encoding is UTF8
The URLEncoder class carries this out with the method:

```
static String encode(String s, String enc)
```

HTML Forms

Web clients send data to servers using HTML forms
Forms define the possible parameters that will be filled by the user with values
Input components are defined by `<input>` elements with a type attribute.
Possible types are hidden, text (default), radio, scroll down menus, submit buttons, reset, etc.
Data under the form of pairs (parameter name, value) are sent using GET and POST methods.
GET uses the URL string to send the parameters
POST sends a MIME message

Google

The input box of the Google page

```
<form action=/search name=f>...
  <input type=hidden name=hl value=sv>
  <input maxLength=256 size=55 name=q value="">
  <input type=submit value="Google-sökning" name=btnG>
  <input type=submit value="Jag har tur" name=btnI>
  <input id=all type=radio name=meta value="" checked>
    <label for=all> webben</label>
  <input id=lgr type=radio name=meta value="lr=lang_sv" >
    <label for=lgr> sidor på svenska</label>
  <input id=cty type=radio name=meta value="cr=countrySE" >
    <label for=cty>sidor från Sverige</label>...
</form>
```

Queries

The query Nugues to Google is a sequence of pairs (name, value)

```
http://www.google.se/search?hl=sv&q=Nugues&
  btnG=Google-sökning&meta=
```

`URISplitter` extracts the query:

```
hl=sv&q=Nugues&btnG=Google-sökning&meta=
```

We can create a `GET` request using the URL constructor and send it to Google `openStream()`
Google returns a 403 error: Forbidden. AltaVista is nicer.
//GoogleQuery.java
POST would send

```
Content-Type: application/x-www-form-urlencoded
Content-Length: 29
Connection: close

hl=sv&q=Nugues&btnG=Google-sökning&meta=
```

Encoding Characters in URLs

From the Java specifications:

- The alphanumeric characters `a` through `z`, `A` through `Z` and `0` through `9` remain the same.
- The special characters `.`, `-`, `*`, and `_` remain the same.
- The space character ' ' is converted into a plus sign `+`.
- All other characters are unsafe and are first converted into one or more bytes using some encoding scheme. Then each byte is represented by the 3-character string `%xy`, where `xy` is the two-digit hexadecimal representation of the byte. The recommended encoding scheme to use is UTF-8.

You have to encode the string explicitly with UTF-8
We will review UTF-8 and other types of encoding in a next lecture

Encoding Strings

To demonstrate `URLEncoder`, we will use the program EncodeTest.java from Elliotte Rusty Harold, *Java Network Programming*, page 210.
We call:

```
URLEncoder.encode("This string has spaces", "UTF-8")
```

In the textbook, there is a mistake with the encoding option of javac
It assumes that the editor saves the program in UTF-8.
By default, most editors use the machine's encoding, MacRoman on a MacIntosh

Running the Encoder

```
pierre% javac url/EncodeTest.java
pierre% java url/EncodeTest
This%C3%A9string%C3%A9has%C3%A9non-ASCII+characters
This%C3%A9string%C3%A9has%C3%A9non-ASCII+characters
pierre% javac -encoding UTF8 url/EncodeTest.java
pierre% java url/EncodeTest
This%EF%BF%BDstring%EF%BF%BDhas%EF%BF%BDnon-ASCII+characters
This%C3%A9string%C3%A9has%C3%A9non-ASCII+characters
```

Encoded strings can subsequently be sent to a search engine for instance

URL Decoders

URLDecoder decodes a string encoded using UTF-8 in the machine's encoding
It has one method:

```
static String decode(String s, String enc)
```

You have to decode the string explicitly with UTF-8

```
System.out.println(URLDecoder.
  decode("This+string+has+spaces", "UTF-8"));
System.out.println(URLDecoder.
  decode("This*string*has*asterisks", "UTF-8"));
System.out.println(URLDecoder.
  decode("This%25string%25has%25percent%25signs", "UTF-8"));
```

(DecodeTest.java)

HTML

HTML is the page description language used by the web.
Derives from SGML
Very messy:

- Sloppy syntax
- Many flavors, many versions

There are tons of sites where to learn it
We will review basic Java tools to analyze HTML pages to display and parse HTML

Displaying HTML Tags

Most Swing components understand HTML tags as JLabel:

```
public class SimpleGUI {
  public static void main(String[] args) {
    JFrame frame = new JFrame();
    JLabel label = new JLabel("<html>
      <p>Hello! This is a multiline label with <b>bold</b>
      and <i>italic</i> text<hr></p></html>");
    frame.setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
    frame.getContentPane().add(label);
    frame.setSize(300, 300);
    frame.setVisible(true);
  }
}// SimpleGUI.java
```

Displaying HTML Pages: JEditorPane

```
public static void main(String[] args) {
  JFrame f = new JFrame("LTH");
  f.setDefaultCloseOperation(WindowConstants.DISPOSE_ON_CLOSE)
  JEditorPane jep = new JEditorPane();
  jep.setEditable(false);
  try {
    jep.setPage("http://cs.lth.se");
  } catch (IOException ex) {
    jep.setContentType("text/html");
    jep.setText("<html>Could not load http://? </html>");
  }
  JScrollPane scrollPane = new JScrollPane(jep);
  f.setContentPane(scrollPane);
  f.setSize(512, 342);
  f.setVisible(true);
} //LTHHomePage.java
```

Parsing HTML Pages

Java has a class to implement a primitive HTML parser: `HTMLEditorKit.Parser` in the `javax.swing.html.text` package

It is an inner class of `HTMLEditorKit`

`Parser` is an abstract class that is instantiated using an obscure and complex sequence of operations. See textbook, page 248.

The code is idiosyncratic: hardwired and impossible to explain

We get the parser using the `getParser()` method that needs to be made public

Making getParser() Public

Java Network Programming, chapter 8, page 250

```
import javax.swing.text.html.*;

public class ParserGetter extends HTMLEditorKit {
    // purely to make this method public
    public HTMLEditorKit.Parser getParser(){
        return super.getParser();
    }
}
```

Parsing

Once we have the parser, we can parse a page using the method:

```
parse(Reader r, HTMLEditorKit.ParserCallback cb,
  boolean ignoreCharSet)
```

cb will react to HTML tags, text, or comments using the undocumented methods:

```
void handleText(char[] data, int pos) //TagStripper.java
void handleStartTag(HTML.Tag t, MutableAttributeSet a,
  int pos)
void handleEndTag(HTML.Tag t, int pos)
void handleSimpleTag(HTML.Tag t, MutableAttributeSet a,
  int pos)
void handleComment(char[] data, int pos)
void handleEndOfLineString(String eol)
void handleError(String errorMsg, int pos)
```

Printing the Text of a Web Page

TagStripper.java by Elliotte Rusty Harold, Java Network Programming, page 251.

```
public class TagStripper extends
    HTMLEditorKit.ParserCallback {
  private Writer out;
  public TagStripper(Writer out) {
    this.out = out;
  }
  public void handleText(char[] text, int position) {
    try {
      out.write(text);
      out.flush();
    } catch (IOException ex) {
      System.err.println(ex);
    }
  }
```

Printing the Text of a Web Page

```
  public static void main(String[] args) {
    ParserGetter kit = new ParserGetter();
    HTMLEditorKit.Parser parser = kit.getParser();
    HTMLEditorKit.ParserCallback callback =
      new TagStripper(new OutputStreamWriter(System.out));
    try {
      URL url = new URL("http://cs.lth.se/EDA095/");
      InputStream in =
        new BufferedInputStream(url.openStream());
      InputStreamReader r = new InputStreamReader(in);
      parser.parse(r, callback, true);
    } catch (IOException ex) {
      ex.printStackTrace();
      System.err.println(ex);
    }
  }
}
```

Parsing Tags

HTML tags:

```
static HTML.Tag.A
static HTML.Tag.ADDRESS
static HTML.Tag.APPLET
static HTML.Tag.AREA
static HTML.Tag.B
static HTML.Tag.H1
```

etc.

Formatting instructions:

```
boolean breaksFlow()
boolean isBlock()
boolean isPreformatted()
```

Outlining the Titles of a Web Page

The HTML source:

```
<h1>Text header 1</h1>
<h2>Text header 21</h2>
<h3>Text header 3</h3>
<h2>Text header 21</h2>
```

will be displayed as:

Text header 1
    Text header 21
        Text header 3
    Text header 21

Outlining the Titles of a Web Page

```
public class OutlinerSimple extends HTMLEditorKit.
    ParserCallback {
  private Writer out;
  private boolean inHeader = false;
  private int level = 0;
  public void handleStartTag(HTML.Tag tag,
      MutableAttributeSet attributes, int position) {
    if (tag == HTML.Tag.H1 || tag == HTML.Tag.H2 ||
        tag == HTML.Tag.H3 || tag == HTML.Tag.H4) {
      inHeader = true;
      if (tag == HTML.Tag.H1) level = 1;
      if (tag == HTML.Tag.H2) level = 2;
      if (tag == HTML.Tag.H3) level = 3;
      if (tag == HTML.Tag.H4) level = 4;
    }
  }
```

Outlining the Titles of a Web Page

```
public void handleEndTag(HTML.Tag tag, int position) {
  if (tag == HTML.Tag.H1 || tag == HTML.Tag.H2 ||
      tag == HTML.Tag.H3  || tag == HTML.Tag.H4) {
    inHeader = false;
    level = 0;
  }
}
public void handleText(char[] text, int position) {
  if (inHeader == true) {
    if (level == 2) System.out.print("\t");
    if (level == 3) System.out.print("\t\t");
    if (level == 4) System.out.print("\t\t\t");
    System.out.println(text);
  }
}
```

Outlining the Titles of a Web Page

```
  public static void main(String[] args) {
    ParserGetter kit = new ParserGetter();
    HTMLEditorKit.Parser parser = kit.getParser();
    HTMLEditorKit.ParserCallback callback =
      new OutlinerSimple();
    try {
      URL url = new URL("http://www.cs.lth.se/EDA095/");
      InputStream in =
        new BufferedInputStream(url.openStream());
      InputStreamReader r = new InputStreamReader(in);
      parser.parse(r, callback, true);
    } catch (IOException ex) {
      ex.printStackTrace();
      System.err.println(ex);
    }
  }
} // OutlinerSimple.java
```

Printing the Links of a Web Page

HTML links have the form:

```
<a href="blackhole.html">click me!</a>
```

where

- `A` is called the tag or element in XML
- `HREF` is an attribute: `HTML.Attribute.HREF`

HTML frames have the form: `<frame src="myframe.html"/>`
HTML images have the form: `<img src="mypics.jpg"/>`
To build absolute URLs from relative URLs, we extract the `BASE` tag from the current web page and its `HREF` attribute, if it exists, as in

```
<base href="http://cs.lth.se/" />
```

otherwise we use the address of the page.

Printing the Links of a Web Page

We extract them using the code:

```
public void handleStartTag(HTML.Tag tag,
    MutableAttributeSet a, int position) {
  if(tag == HTML.Tag.A) {
    String href = (String)
      a.getAttribute(HTML.Attribute.HREF);
    System.out.println("Link: " + href);
  }
  if(tag == HTML.Tag.FRAME) {
    String href = (String) a.getAttribute(HTML.Attribute.SRC);
    System.out.println("Frame: " + href);
  }
}
```

Printing the Links of a Web Page

```
public void handleSimpleTag(HTML.Tag t,
    MutableAttributeSet a, int pos) {
  if (tag == HTML.Tag.BASE) {
    String href =
      (String) attributes.getAttribute(HTML.Attribute.HREF);
    baseURL = href;
    System.out.println("Base URL: " + href);
  }
  if(t == HTML.Tag.IMG) {
    String href = (String) a.getAttribute(HTML.Attribute.SRC);
    System.out.println("Image: " + href);
  }
} // LinkGetter.java
```

Absolute Links

Some links are absolute, while others are relative.
We can use this piece of code to create absolute links in `handleStartTag()`:

```
try {
  if (!new URI(href).isAbsolute()) {
    System.out.println(
     "\tAbsolute link: " + new URL(new URL(baseURL), href));
  }
} catch (Exception e) { }
```

or just

```
new URL(new URL(baseURL), href)
```