Universal Resource Locator (URL)

Mendel Rosenblum

Hypertext

- Text with **links** to other text
  - Click on links takes you somewhere else
  - Old idea:
    - Ted Nelson coined the term (early '60s), built Xanadu system
    - Doug Englebart: "Mother of all demos" in 1968
    - HyperCard for the Macintosh: 1987
- Web adapted the idea, link specification:
  - Uniform Resource Locators (URL) - Provided **names** for web content

```
<a href="https://en.wikipedia.org/wiki/URL">URL</a>
```

Parts of an URL

http://host.company.com:80/a/b/c.html?user=Alice&year=2008#p2

Scheme (**http:**): identifies protocol used to fetch the content.

Host name (**//host.company.com**): name of a machine to connect to.

Server's port number (**80**): allows multiple servers to run on the same machine.

Hierarchical portion (**/a/b/c.html**): used by server to find content.

Query parameters (**?user=Alice&year=2008**): provides additional parameters

Fragment (**#p2**): Have browser scroll page to fragment (html: p2 is anchor tag)

Used on the browser only; not sent to the server.

URL: schemes (e.g. `http`)

`http`: is the most common scheme; it means use the HTTP protocol

`https`: is similar to http: except that it uses SSL encryption

`file`: means read a file from the local disk

`websocket`: means create a TCP connection

`mailto`: means open an email program composing a message

There are many (~350) other schemes: https://www.iana.org/assignments/uri-schemes/

Example: mongodb: points to a MongoDB database

URL: Hierarchical portion (/a/b/c.html)

- Passed to the web server for interpretation. Early web servers:
  - Path name for a static HTML file.
  - Path name of a program that will generate the HTML content (e.g., `foo.php`).
- Web server programmed with **routing** information
  - Map hierarchical position to function to be performed and possibly the function's parameters
- API design, Example:
  - `/user/create`
  - `/user/list`
  - `/user/0x23490`
  - `/user/0x23433`
  - `/user/delete/0x23433`

Query Parameters (e.g. ?user=Alice&year=2008)

- Traditionally has been to provide parameters to operation:

  `http://www.company.com/showOrder.php?order=4621047`

- For modern apps has implications of when the browser switches pages

Links

- Browser maintains a notion of current location (i.e. URL)
- Links: content in a page which, when clicked on, causes the browser to go to URL
- Links are implemented with the <a> tag:

```
<a href="http://www.company.com/news/2009.html">2009 News</a>
```

Different types of links

Full URL: `<a href="http://www.xyz.com/news/2009.html">2009 News</a>`

Absolute URL: `<a href="/stock/quote.html">`
same as `http://www.xyz.com/stock/quote.html`

Relative URL (intra-site links): `<a href="2008/March.html">`
same as `http://www.xyz.com/news/2008/March.html`

Define an anchor point (a position that can be referenced with # notation):
`<a name="sec3">`

Go to a different place in the same page: `<a href="#sec3">`

Uses of URLs

- Loading a page: type the URL into your browser
- Load a image:
  ```
  <img src="..." />
  ```
- Load a stylesheet:
  ```
  <link rel="stylesheet" type="text/css" href="...">
  ```
- Embedded a page:
  ```
  <iframe src="http://www.google.com">
  ```

URL Encoding

- What if you want to include a punctuation character in a query value?

  `http://www.stats.com/companyInfo?name=C&H Sugar`

- Any character in a URL other than `A-Z, a-z, 0-9,` or any of `-_.~` must be represented as %xx, where xx is the hexadecimal value of the character:

  `http://www.stats.com/companyInfo?name=C%26H%20Sugar`

- Escaping is a commonly used technique and also a source of errors

Miscellaneous Topics

- Computer scientists take on hypertext:

  Need to have **referential integrity**

- The web (done by physicists):

  `Error 404`

- URI (Uniform Resource Identifier) vs. URL (Uniform Resource Locator)