*:96 Overheads

Part 2c: URL, Media types

More about this course about Internet application protocols can be found at URL:

`http://www.dsv.su.se/~jpalme/internet-course/Int-app-prot-kurs.html`

Last update: 05-01-14 11.54

URL, Uniform Resource Locator

An URL identifies a resource, such as a document, as stored in one particular location, and an access protocol to connect to the resource or, in the case of a document, to retrieve it.

References:

RFC 1738: Uniform Resource Locators (URL), by T. Berners-Lee, L. Masinter and M. McCahill. URL
`ftp://ftp.sunet.se/pub/Internet-documents/rfc/rfc1738.txt`

RFC 1808: Relative Uniform Resource Locators, by R. Fielding, URL
`ftp://ftp.sunet.se/pub/Internet-documents/rfc/rfc1808.txt`

Examples:

`http://dsv.su.se/~jpalme`	identifies my personal home page, as retrieved with the HTTP (WWW) protocol.
`ftp://ftp.sunet.se/pub/Internet-documents/rfc/rfc1738.txt`	identifies the copy of RFC1738 stored at FTP.SUNET.SE for retrieval using FTP.
`http://ftp.sunet.se/pub/Internet-documents/rfc/rfc1738.txt`	identifies the copy of RFC1738 stored at FTP.SUNET.SE for retrieval using HTTP.

URL schemes standardized in RFC 1738

ftp	File Transfer protocol
http	Hypertext Transfer Protocol
gopher	The Gopher protocol
mailto	Electronic mail address
news	USENET news
nntp	USENET news using NNTP access
telnet	Reference to interactive sessions
wais	Wide Area Information Servers
file	Host-specific file names
prospero	Prospero Directory Service

Character set in URLs (not in referenced document)

Only US-ASCII allowed

Unsafe characters: `space < > " # % { } | \ ^ ~ [ ] ` `

Reserved characters in some URL schemes: `; / ? : @ = &`

Unsafe characters must be encoded in transport.

Reserved characters not used in their reserved meaning must be encoded or may not be used.

Safe characters is the rest of US-ASCII, i.e. `A-Z, a-z, 0-9, $ - _ + ! * ' ( ) ,`

Encoding of unsafe characters in URL-s

In addition, octets may be encoded by a character triplet consisting of the character "%" followed by the two hexadecimal digits (from "0123456789ABCDEF") which forming the hexadecimal value of the octet. (The characters "abcdef" may also be used in hexadecimal encodings.)

Examples:

The string

```
"Donald Duck"
```

is encoded as

```
%22Donald%20Duck%22
```

Top-level URL Syntax:

```
<scheme>:<scheme-specific-part>
```

Common Internet Scheme Syntax

```
//<user>:<password>@<host>:<port>/<url-path>
```

Examples of three URL-s referring to the same document

```
ftp://ftp.dsv.su.se/users/Jacob.Palme/draft-ietf-mailext-new-fields-05.txt
ftp://anonymous:@ftp.dsv.su.se/users/Jacob.Palme/draft-ietf-mhtml-info-01.txt
ftp://anonymous:@ftp.dsv.su.se:21/users/Jacob.Palme/draft-ietf-mhtml-info-01.txt
```

Examples of five URL-s referring to the same directory

```
ftp://jpalme:password@ester.dsv.su.se
ftp://jpalme:password@ester.dsv.su.se/home0/ester/dsv/dsv-jp
ftp://jpalme:password@ester.dsv.su.se:21/home0/ester/dsv/dsv-jp
ftp://jpalme:password@ester.dsv.su.se:21/home0/ester/dsv/dsv-jp/
ftp://jpalme:password@ester.dsv.su.se:21
/home0/ester/dsv/dsv-jp/;type=d
```

Relative URLs

Based on this Base URL:

`URL:http://a/b/c/d;p?q#f`

The following URLs are resolved as shown:

URL	Resolved URL
g:h	g:h
g	http://a/b/c/g
./g	http://a/b/c/g
g/	http://a/b/c/g/
/g	http://a/g
//g	http://g
?y	http://a/b/c/d;p?y
g?y	http://a/b/c/g?y
g?y/./x	http://a/b/c/g?y/./x
#s	http://a/b/c/d;p?q#s
g#s	http://a/b/c/g#s
g#s/./x	http://a/b/c/g#s/./x

URL	Resolved URL
g?y#s	http://a/b/c/g?y#s
;x	http://a/b/c/d;x
g;x	http://a/b/c/g;x
g;x?y#s	http://a/b/c/g;x?y#s
.	http://a/b/c/
./	http://a/b/c/
..	http://a/b/
../	http://a/b/
../g	http://a/b/g
../..	http://a/
../../	http://a/
../../g	http://a/g

HTTP URL syntax

```
http://<host>:<port>/<path>?<searchpart>
```

Example of an HTTP Query URL

A search for "Donald Duck" to Alta Vista is encoded as:

```
http://altavista.digital.com/cgi-
bin/query?pg=q&what=web&fmt=.&q=%22Donald+Duck%22
```

Reference to fragments of an HTML document

Relative reference:

```
#anchor1003017
```

Absolute reference:

```
http://www.dsv.su.se/~jpalme/ietf/jp-ietf-home#anchor1003017
```

Markup in the referenced HTML document

```
<A NAME="anchor1003017"></A><BR>
```

Part of the URL?

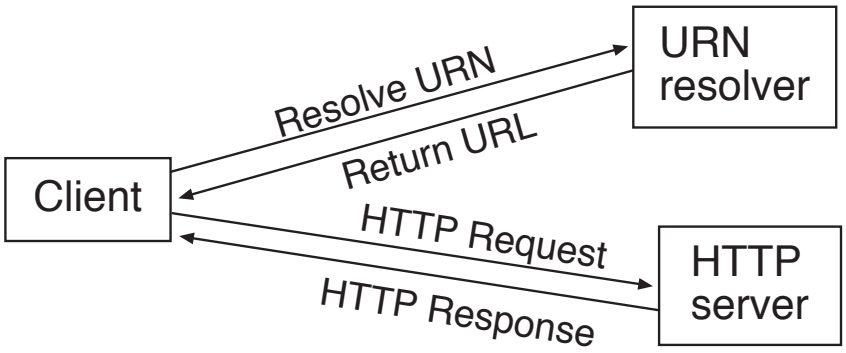
Section preceded by ? are regarded as part of the URL itself, but section preceded by # are not regarded as part of the URL itself.

URL, URI, URN, URC

URI = Uniform Resource Identifier

URL = Uniform Resource Locator

An URI, which refers to a particular copy of a document stored on a particular host. May have to be changed if the document is moved. Belongs to registered URI/URL schemes.



URN = Uniform Resource Name

A URI scheme for various namespaces. Refers to a document, wherever it is stored. Can be resolved into an URL by an URN resolver. An URN resolver may locate the copy of a mirrored document which is clostest to the requestor. Begins with “urn:” followed by the namespace, followed by the value, for example “urn:isbn:

URI = URL or URN

URC = Uniform Resource Characteristics

The purpose or function of a URC is to provide a vehicle or structure for the representation of URIs and their associated meta-information.

Media types

Initially defined in RFC 1521. New media types can be registered with IANA, usually an RFC defining the type is provided. Registered media types are listed in `ftp://ftp.isi.edu/in-notes/iana/assignments/media-types`.

Format:

```
<type> "/" <subtype> [ ";" *(<para> "=" <value>) ]
```

Primary media types

Type	Description
text	Mainly text. Can always have "charset" parameter. Default for charset is sometimes US-ASCII.
multipart	Consists of several parts, which each may be of different type.
message	An encapsulated message (usually includes message heading).
application	Executable code. Note that postscript is application/postscript, not text/postscript.
image	Still picture.
audio	Sound.
video	Moving picture (may include sound).

Some important subtypes:

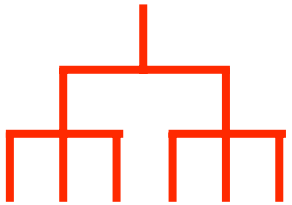
The *text* media type

Mainly text. Can always have “charset” parameter. Default for charset is US-ASCII in e-mail, ISO 8859-1 in HTTP 1.0.

Type	Description
text/plain	Text without other formatting codes than horizontal tab, CRLF and form feed.
text/richtext, text/enriched	Two simpler formatting schemes than HTML and SGML.
text/html	Hypertext Markup Language, The main document format in the WWW. Version 2.0 is defined in RFC 1866. HTML is an application of SGML.
text/sgml	Standard Generalized Markup Language. ISO standard.
text/rfc822-headers	Headers from a mail message (returned in a delivery status notification).

HTML (.HTML, .HTM) = mjuk formattering,
Postscript (.PS) och Adobe Acrobat (.PDF) = hård formattering

The *multipart* media type



Contains several parts, which may be of different types

Type	Description
multipart/mixed	A sequence of parts to be displayed in sequence.
multipart/alternative	Several versions of the same information, simplest first, recipient displays the most advanced version it can handle. Example: plain text versus richtext versus html.
multipart/digest	A set of messages, headings sometimes abbreviated.
multipart/parallel	Parts to be shown at the same time. Example: Image and sound.

multipart/related	Related parts, such as an HTML document and inline images, RFC 1872.
multipart/report	Delivery status and other notifications RFC 1892
multipart/form-data	Using HTML forms to upload files from the client, RFC 1867.
multipart/header-set	Set of data, some of which is system-specific and some of which is in MIME standard types
multipart/appledouble	Binary Macintosh files
multipart/voice-message	Using Internet mail for communication between voice-mail machines, RFC 1911

Content-Disposition

Content-Disposition: ("Inline" / "Attachment")

The *message* media type

Usually contains a message

Type	Description
message/rfc822	Internet e-mail message.
message/partial	One of a set of messages which are to be combined.
message/external-body	Message, whose body is referenced and not included. Access types: FTP, ANON-FTP, TFTP, AFS, LOCAL-FILE, MAIL-SERVER, Content-ID, URL.
message/news	A usenet news article.
message/http	A document which has been transmitted through HTTP.
message/delivery-status	Delivery status report.

The *application* media type (not complete)

Type	Description
octet-stream	Any binary data.
postscript	Adobe postscript page description language.
rtf	Rich Text Format, Microsoft standard for exchange of documents between word processing software, also supported by other vendors than Microsoft.
pdf	Adobe Acrobat.
activemessage	How to connect to an Active Mail application at a remote host.
mac-binhex40	Macintosh binary-to-text conversion method
remote-printing	RFC 1486: Printing in a remote location
msword, cybercash, wordperfect5.1, vnd-framemaker, etc.	Vendor-specific formats.

The *audio* media type

Contains sound.

Type	Description
audio/basic, audio/32kadpcm	Two different sound encoding methods

The *video* media type

Contains moving pictures, can include sound.

Type	Description
mpeg, quicktime, vnd.vivo	Two different video encoding methods

The charset attribute

US-ASCII	Plain US-ASCII, not other ISO 646 variants. 7 bits. Default in e-mail.
ISO-8859-1	Also known as ISO latin 1. 8 bits. Default in WWW.
ISO-8859-?	Other variants of ISO 8859 for different language groups.
UTF-8	Unicode/ISAO 10646 with the UTF-8 encoding