Le WEB

Gérard Florin

Laboratoire Cedric
Conservatoire National
des Arts et Métiers

Désignation
et liaison dans le WEB

I **Les localisateurs**
URL 'Uniform Resource Locators'

II **Les noms**
URN 'Uniform Resource Names'

III **Les informations de niveau méta**
URC 'Uniform Resource Characteristics'

IV **Les identificateurs**
URI 'Uniform Resource Identifiers'

I Désignation de base dans le WEB: les localisateurs

URL ‘Uniform Resource Locators’ RFC 1738 Décembre 1994

- La désignation avec les URL est placée dans le cadre d'un ensemble de **schémas** associés à des applications Internet:

ftp	File Transfer protocol
http	Hypertext Transfer Protocol
gopher	Gopher protocol
mailto	Adresse de courrier électronique
news	USENET news
nntp	USENET news avec NNTP
telnet	Référence session interactive
wais	Wide Area Information Servers
file	Noms de fichiers spécifiques
….	D’autres schéma sont proposés.

- Un schéma définit une construction et une interprétation des noms ou adresses.

Syntaxe commune aux schémas

Forme générale des URLs

<nom_schema>:
<partie-spécifique-schéma>

Forme générale des URL Internet

- Pour Internet la partie spécifique du schéma commence par // **(en fait pas d'autre architecture de réseau utilisée)**

- En Internet on définit un accès à distance en **TCP/IP en tant qu'usager d'un OS**.

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

Signification des champs d’une URL Internet

User : Un nom d'usager optionnel.

Password : Un mot de passe optionnel.

Host : Le nom DNS d'un hôte (FQDN fully qualified domain name) ou une adresse IP

Port : Un numéro de port TCP (optionnel les protocoles ont un port par défaut).

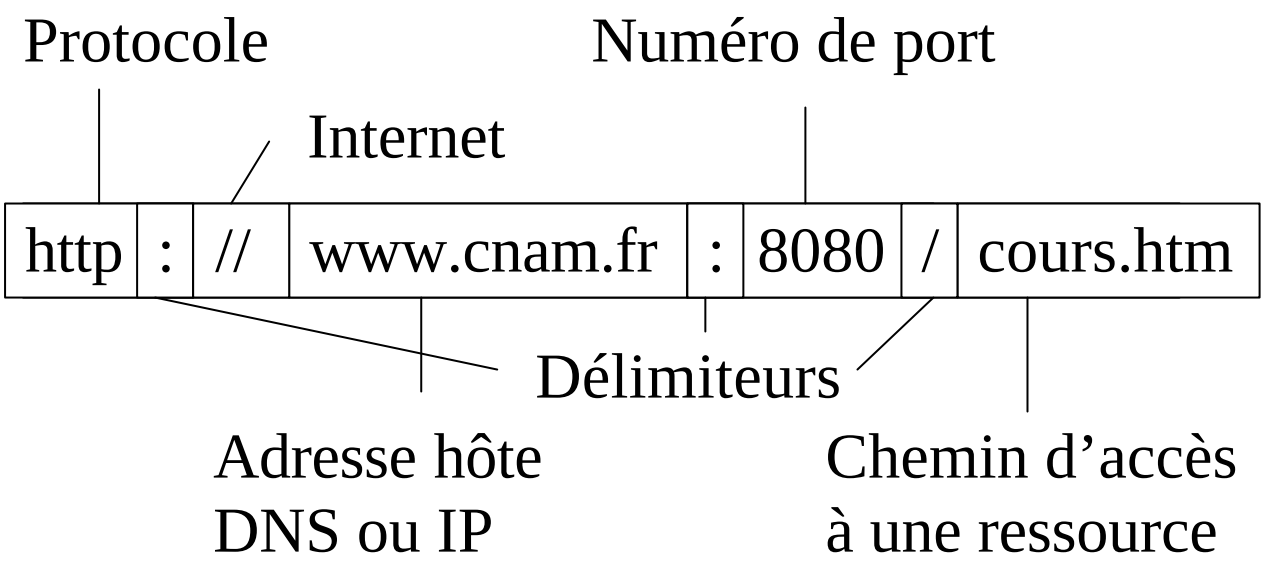
url-path: Le reste du nom est spécifique du schéma de désignation. Il s’agit d’un chemin d’accès à la ressource.

Remarque

Le "/" entre host (ou port) et url-path ne fait pas partie de l'url-path.

Le schéma d'adressage HTTP

Schéma très tourné vers la localisation



Aspects spécifiques HTTP de la partie définition du chemin d'accès

url-path = <path>?<search-part>

<path> : un chemin d'accès absolu à la ressource demandée sur le site serveur.

<search-part> : une chaîne de caractère qui complète le nom par des paramètres complémentaires définissant une question.

Complément :
Accès interne aux documents

- Complément d'adressage à l'intérieur d'un document.

- Non défini dans la RFC 1738 URL mais dans le cadre des URI.

url-path = <path>#<fragment>
?<search-part>

<fragment> : le nom d'une ancre locale d'un document permettant un accès interne.

- Balise A présente dans le document HTML pour définir une ancre locale :
<A NAME="etiquette_fragment"></A>

- Localisateur du fragment
http://<host>/<path>#etiquette_fragment

Exemples d'URL

- Une URL pour une **question posée à un moteur de recherche** (partie question de la forme symbole=valeur).

http://www.google.com/search
?q=URL+syntax

- Une URL **avec question et accès interne** dans un document.

http://home.netscape.com/assist/extensions.html#topic1?x=7&y=2

Remarques complémentaires

- La chaîne question contient des couples (**nom de variable, valeur)** séparés par des 'esperluettes' (ampersand &).
- Si des caractères réservés dans la création des URL (";"/"?":"@"&"="+"$","")doivent figurer dans les **données une procédure d'échappement** est définie:

escape = "%" HEX HEX

Réalisation de la liaison avec les URLs
Mise en œuvre dans le protocole HTTP

- L'analyse syntaxique de l'URL permet **de délimiter les éléments** de la référence**.**
- La liaison est très facile car les éléments sont presque tous des **adresses**
=>directement utilisables pour la liaison.

Détermination de l'adresse IP du serveur
- Soit elle est **en clair**.
- Soit la partie **nom de domaine DNS** est soumise au serveur d'annuaire DNS qui retourne l'adresse IP de l'hôte distant

Ouverture de connexion TCP sur le serveur
- Au moyen de **l'adresse IP** et **du numéro de port** (ou du port par défaut **80**).

Utilisation par le serveur de la partie chemin d'accès de l'URL
- Chemin **d'accès absolu** dans le système de fichier du site distant.

Conclusion: Avantages Inconvénients du schéma d'adressage de base URL HTTP

- Les URL HTTP définissent un adressage absolu **pratique pour la localisation des ressources** mais uniquement pour cela.

Limitations

- Manque d'outils d'adressage relatif (pour définir une ressource relativement à une autre et éviter des problèmes de relocation des liens lors du déplacement d'une page HTML).

- Les URL comportent essentiellement des informations **d'adresses ('codées en dur').**
Protocole, Adresse DNS, No Port
Accès en échec (destruction/migration de ressource). Seul mécanisme possible de migration: **les liens de poursuite**.

- Les URL n'offrent pas de description du **contenu (informations méta).**

II Désignation WEB : Les noms symboliques

URN ‘Uniform Resource Names’

- Les URN ont été définis au départ pour résoudre le problème **de persistance** des noms que ne possèdent pas les URLs.
- Les noms choisis doivent donc être **indépendants de la localisation** des ressources: les URN sont en fait les **noms symboliques** du WEB.
- La résolution doit faire appel à **des services d'annuaire**s qui font correspondre noms symboliques et adresses.
- Si on ajoute aux noms symboliques des **meta-informations** (des attributs comme "titre", "auteur" or "sujet") on relie le problème des URN à celui de la recherche documentaire (voir URC).
- Plusieurs **structurations** ont été proposées.

II.1 URN selon la RFC 2141 Mai 1997

- Définition d'un ensemble de **noms symboliques, uniques, persistants**.

- Forme générale retenue similaire aux URL:
  Schéma_d'adressage: Chaîne opaque

- Un nouveau schéma d'adressage est introduit pour les noms symboliques.
- Il est baptisé '**urn**' (insensible à la casse).

- La structure générale d'un urn est

<URN> ::= "urn:" <NID> ":" <NSS>

<NID> '**Namespace IDentifier**' ou autorité.
Le NID détermine la structure syntaxique de la partie NSS 'Namespace Specific String'.
Partie autorité responsable de la sructuration d'un espace de noms.

<NSS> '**Namespace Specific String**'.
La partie NSS définit un objet proprement dit.

Identification de l'autorité de nommage
NID 'Namespace Identifier'

- Symboles de **32 caractères au plus** (alphanumériques avec _).
  Exemple: urn:cnam_fr
- La possibilité d'utiliser . + semble en débat.
  Exemple: urn:cnam.fr:CoursInfos

- L'unicité des noms repose principalement sur une autorité de contrôle des NID. Ici on trouve un nom de domaine DNS.

Identification locale de l'objet
NSS 'Namespace Specific String'

- Symboles alphanumériques avec les caractères spéciaux "(" | ")" | "+" | "," | "-" | "." |":" | "=" | "@" | ";" | "$" |"_" | "!" | "*" | "'"
- Une procédure de transparence est définie au moyen du caractère %.

Les URN RFC 2141 sont pour l'instant encore peu utilisés. Ils devraient prendre de l'extension avec les services d'annuaires.

2 Les URN selon l'approche des PURL: URL persistantes

- Une PURL a la forme d'une URL mais son interprétation sémantique est différente.

- Elle comporte trois parties:

(1) Une définition de **protocole** (de schéma d'adressage)
Exemple: http

(2) une **adresse de serveur d'annuaire**,
Exemple: //purlserv.cnam.fr

(3) un **nom symbolique** que le résolveur transforme en une URL effective par consultation d'annuaire.
Exemple: /chemin/symbole

`http://purlserv.cnam.fr/chemin/symbole`

Bibliographie sur les URN

RFC 2141, May 1997 - URN Syntax

RFC 2288 February 1998 - R.Moats 'Using Existing Bibliographic Identifiers as Uniform Resource Names'
RFC 1737, December 1994 - K. Sollins, L. Masinter - C. Lynch, R. Daniel 'Functional Requirements for Uniform Resource Names'
RFC2611, June 1999 - L.Daigle, D.van Gulik, R. Iannella, P. Falstrom 'A URN Namespace for IETF Documents'
RFC 2483, January 1999 - M. Mealling URI 'Resolution Services Necessary for URN Resolution'
RFC2168 June 1997 - 'R. Daniel, M. Mealling Resolution of Uniform Resource Identifiers using the Domain Name System'
RFC 2276, January 1998 - K.Sollins 'Architectural Principles of Uniform Resource Name Resolution'

III Désignation dans le WEB: les informations de niveau méta

URC ‘Uniform Resource Characteristics’

Objectif: Rajouter aux ressources WEB des descriptions de caractéristiques (contenu, date, auteur, ….).

- Les URC **codent les informations de niveau meta** concernant les ressources.
- Actuellement le plus souvent sous la forme **d'attributs (nom, valeur)**.

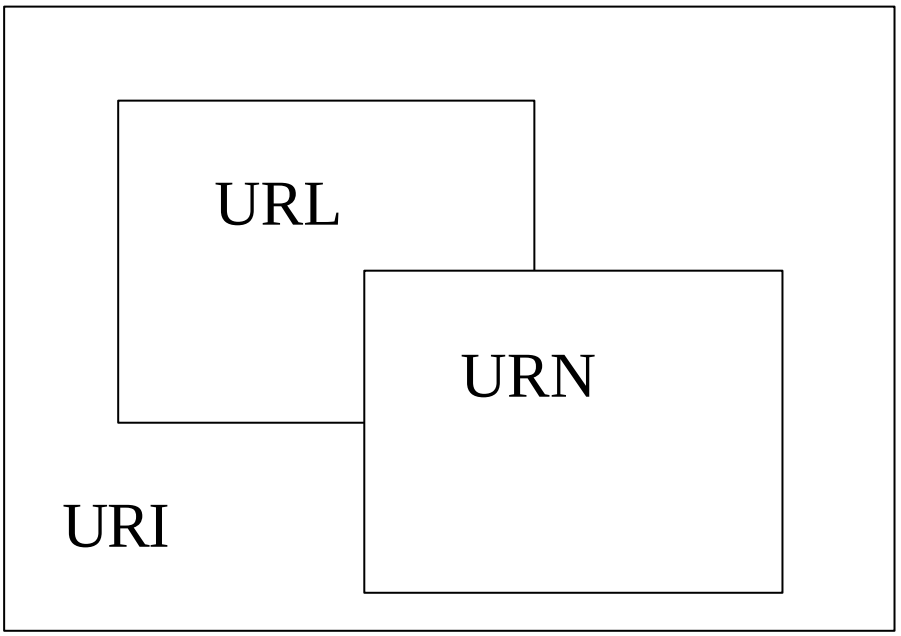
- Différentes expériences ont fonctionné sur l'Internet . **IAFA** ‘Internet Anonymous FTP Archives’ formats,
  . **WHOIS++** templates,
  . **SOIF** 'Summary Object Interchange Format' HARVEST

- Le groupe de travail W3C sur le thème**: RDF** '**Resource Description Framework**' a produit un modèle et une syntaxe de description XML.

IV Désignation WEB: Les identificateurs

URI ‘Uniform Resource Identifier’ RFC 2396 Août 1998

Une formulation de la désignation WEB destinée à définir dans le même cadre toutes les améliorations des mécanismes d'adressage proposés dans des RFCs.



- **URI: la norme actuellement en vigueur** dans le WEB de nommage et d’adressage au moyen de chaînes ASCII.

Les trois aspects de l'adressage WEB

URL ‘Uniform Resource Locators’.

L’ensemble des schémas de désignation comportant un moyen explicite d’accès à une ressource pour un protocole donné.

URN ‘Uniform Resource Names’.

1. Un URI tel que la ressource est garantie disponible, persistante (un URN peut prendre la forme d’un URL ex les PURL).
2. Un schéma particulier de désignation spécifié par la RFC2141 pour des noms symboliques (identificateurs de ressources persistants, indépendants de la localisation).

URI ‘Uniform Resource Identifiers’.

La réunion des **deux systèmes de nommage** précédents **et d'un nommage des ressources relatif.**

Syntaxe des URI

- Les URIs HTTP peuvent être définis selon deux approches: absolue ou relative.

URI = (absoluteURI | relativeURI)
["#" fragment]

- Les **URI absolus** sont indépendants du contexte dans lequel ils sont utilisés.
- Ils commencent toujours comme les URLS par un nom de schéma d'adressage.

absoluteURI = scheme ":"
***(uchar | reserved)**

- Les **URI relatives** commencent par la partie chemin d'accès réseau ou par un chemin d'accès fichier absolu ou relatif.

relativeURI= net_path|abs_path|rel_path
net_path = "//" net_loc [abs_path]
abs_path = "/" rel_path
rel_path = [path] [";" params] ["?" query]

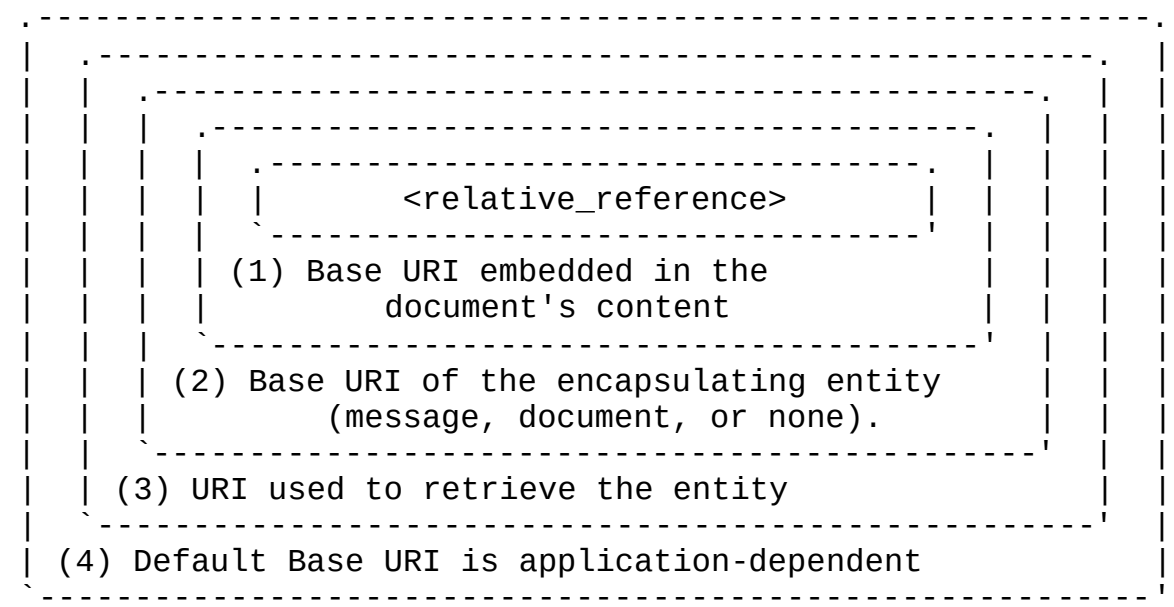
Remarques**:** a) Adresse absolue/relative au niveau des fichiers**.**

b) **[";" params]** paramètres complétant l'adresse (auth=mandatory).

URI d'adressage relatif

- La notion d'URI d'adressage "relatif" implique qu'il existe une **base de départ** pour déterminer un URI absolu de base.

- Quatre propositions pour définir la base:



- Des mécanismes d'adressage relatif doivent être définis: **ils sont très similaires aux mécanismes de désignation des systèmes de fichier** UNIX
. le répertoire courant, .. le père etc

Exemples d'URI relatives

- Pour l'URI absolue de base http://a/b/c/d;p?q les URIs relatives = donnent le résultat:

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/?y
g?y	=	http://a/b/c/g?y
#s	=	(current document)#s
g#s	=	http://a/b/c/g#s
g?y#s	=	http://a/b/c/g?y#s
;x	=	http://a/b/c/;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/

Conclusion adressage WEB

- De très loin **le système d'adressage en univers réparti le plus utilisé**.

- Reste marqué par les motivations d'origine de **désignation de documents**. Mais des évolutions successives lui ont **apporté des caractéristiques plus générales** d'un système de désignation.

Les limitations

- Certaines caractéristiques récemment introduites demandent à être réellement implantées et utilisées (noms symboliques, informations méta, …)

- Le système de désignation WEB est corienté point à point**: pas de notions de relation 1 vers N (groupes).** => Avec XML: **Xlink**

- Tous les éléments d'un document ne sont pas adressables => Avec XML: **Xpath**

Bibliographie: Désignation WEB

RFC 1630 Universal Resource Identifiers in WWW

RFC 1736 Functional Recommendations for Internet Resource Locators

RFC 1737 Functional Requirements for URNs.

RFC 1738 Uniform Resource Locators (URL)

RFC 1808 Relative URLs

RFC 1959 An LDAP URL Format

RFC 2016 Uniform Resource Agents (URAs)

RFC 2056 URL for Z39.50

RFC 2111 Content-ID and Message-ID URL

RFC 2122 VEMMI URL Specification

RFC 2141 URN Syntax

RFC 2168 Resolution of Uniform Resource Identifiers using the Domain Name System

RFC 2276 Architectural Principles of Uniform Resource Name Resolution