World Wide Web

Florence Zara

LIRIS – Université Lyon 1

http://liris.cnrs.fr/florence.zara

E-mail : florence.zara@liris.cnrs.fr

Objectifs du cours

- Découvrir ce qu'est le Web
- Savoir comment ça marche
- Première notions sur la conception de page Web
- Connaître les différentes sources d'information

Le Web : historique rapide

- 1946 : invention de la notion d'hypertexte
- 1970 : développement de l'Internet
- 1989-91 : application de la notion d'hypertexte à Internet (CERN, Tim Berners-Lee)
- 1993 : diffusion universitaire (navigateur Xmosaic, 50 serveurs dans le monde)
- 1993 : création du W3C pour normaliser le Web
- 1994-95 : premiers navigateurs privés (Netscape puis IE)
- 1998 : 2,2 millions de sites
- 2000 : 20 millions de sites (2,5 milliards de pages)
- 2002 : 3 / 550 milliards de pages (web de surface / web profond)
- 2004 : première conférence « Web 2.0 »
- 2005 : Google prétend indexer 8 milliards de pages
- 2020 : Google prétend indexer 130 000 milliards de pages

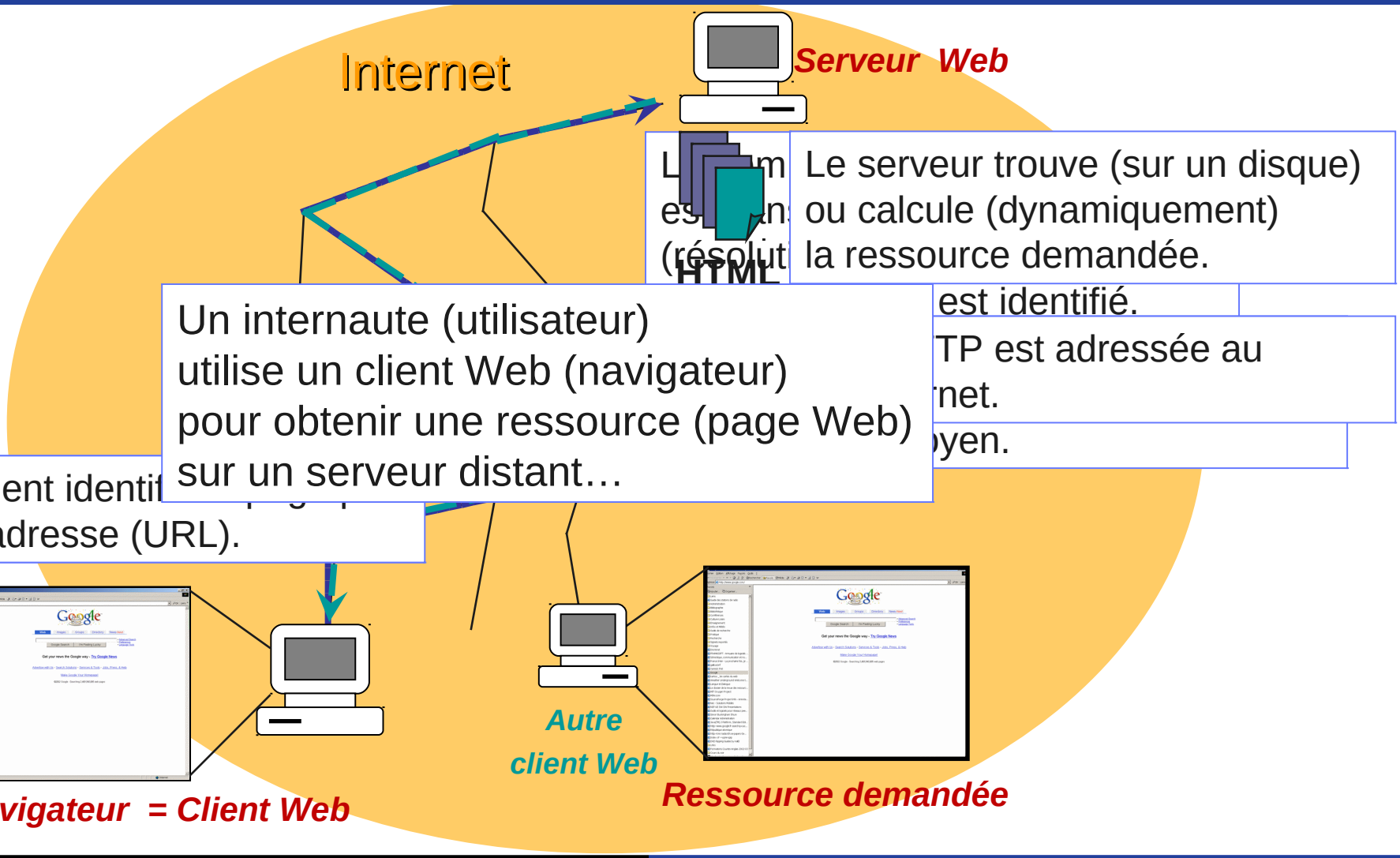
Le Web : première définition

- World Wide Web (WWW, Toile)

- Définition
  - ensemble de données disponibles sur les serveurs accessibles sur le réseau Internet
  - pouvant être visualisées et/ou utilisées avec un navigateur Web

- Attention : Web ≠ Internet
  - Internet = support de communication général
  - Web = une partie des contenus circulant sur l'Internet

Le Web en un schéma

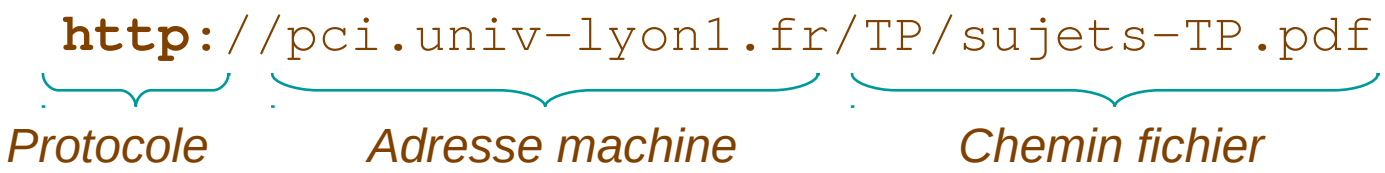


URL (Uniform Resource Locator)

- Permet d'identifier une ressource sur le réseau, c'est-à-dire :
  - une page Web
  - une image (seule ou utilisée dans une page Web)
  - un programme
  - un fichier à télécharger…
- Indique
  - un protocole (langage de communication entre deux programmes sur deux machines).
    - Exemple: FTP (File Transfert Protocol), HTTP (HyperText Transfert Protocol)…
  - une adresse et un chemin
  - forme générale : `protocole://adresse`
  - exemple : `http://www.univ-lyon1.fr/`

Différents types d'URL

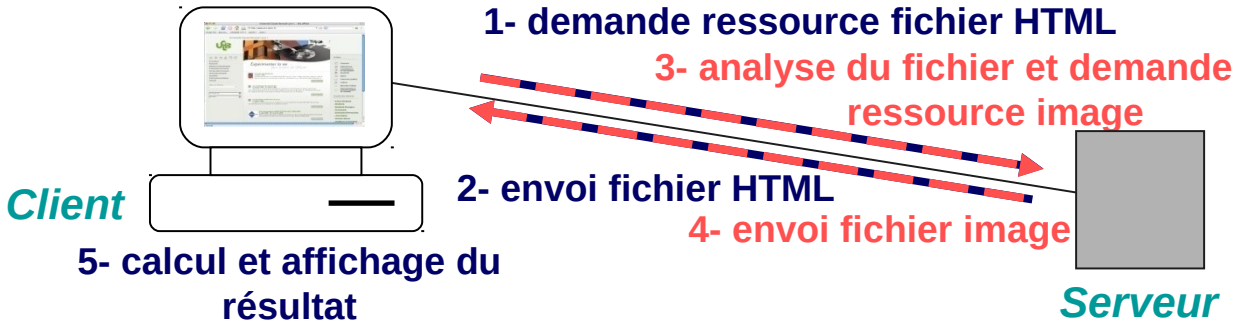
- Forme principale (protocole *HTTP*)



- Forme pour désigner les fichiers *locaux*
  - Chemin relatif : `fichier.html` ou `dossier/toto.html`
  - Sur un disque : `file://C:/chemin/fichier.htm`
- Forme pour le *transfert* de fichiers
  - `ftp://ftp.inria.fr/INRIA/tech-reports/RR-5645.pdf`
- Forme pour l'envoi de *courrier* électronique
  - `mailto:jean.durand@univ-lyon1.fr`

Le protocole HTTP (HyperText Transfer Protocol)

- Protocole de type client-serveur
- Protocole de niveau applicatif
- Mécanisme de requête / réponse HTTP
  - Requête du client : URL de la ressource
  - Réponse du serveur : code de statut / d'erreur + ressource (si « OK »)
    - tout s'est bien passé : « OK » (code de statut : 200)
    - pas le droit d'accéder : « forbidden » (code d'erreur : 403)
    - la ressource n'existe pas : « not found » : (code d'erreur : 404)
- Exemple de transaction avec plusieurs ressources

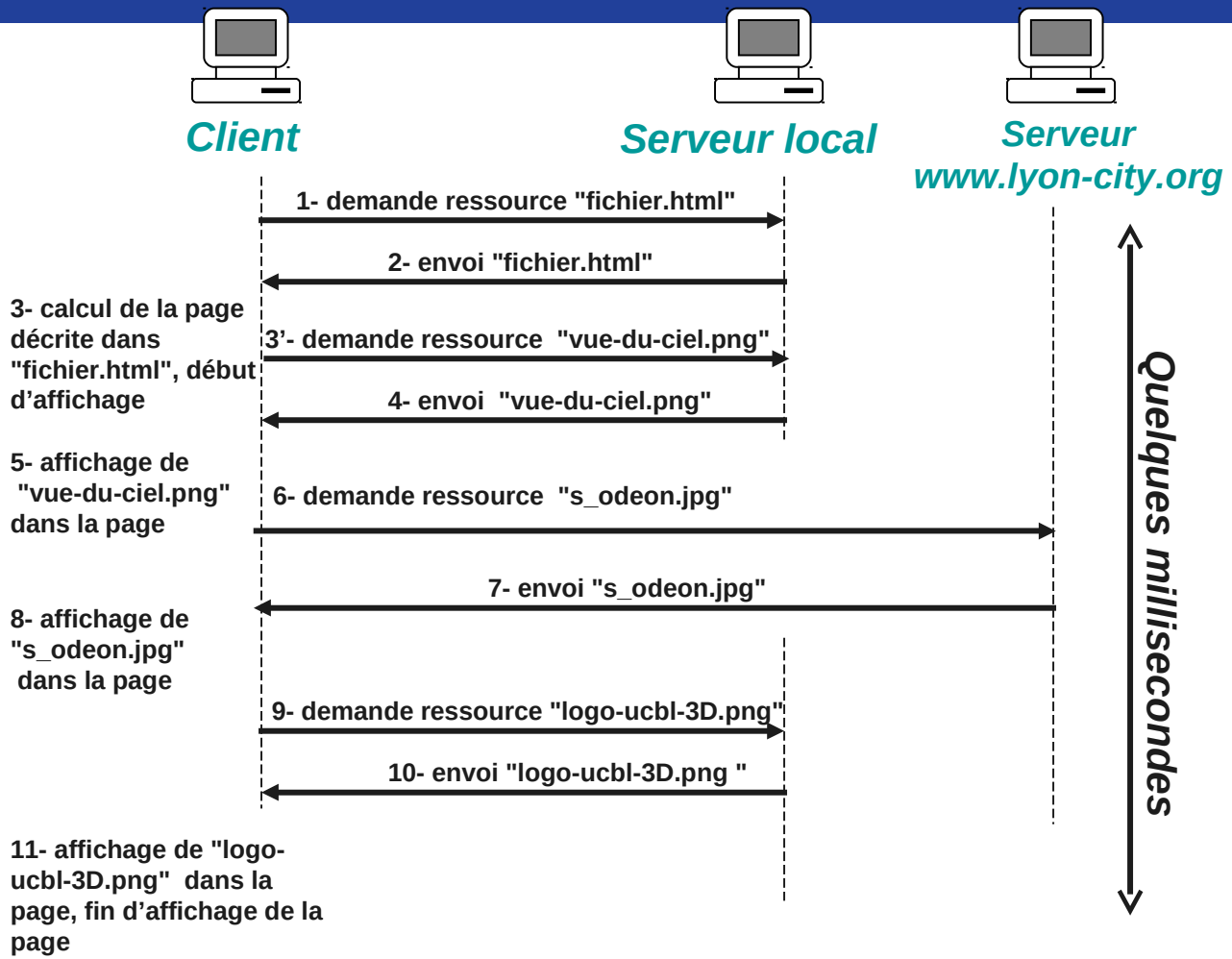


- Remarque : le protocole HTTPS (sécurisé) fonctionne de la même façon

Transactions pour une page Web

fichier.html

```
<html>
 <head>
 <title>INSERTION
    IMAGES</title>
 </head>
 <body bgcolor="#ffffff">

<p>Voici une photo aérienne de
    Lyon (chemin relatif) :
    <img src= "vue-du-
    ciel.png" alt="vue du
    ciel" />.</p>

<p>Voici une autre photo (URL
    distante) : <img
    src="http://www.lyon-
    city.org/static/vdl/conten
    u/decouverte/au_fil_des_qua
    rtiers/s_odeon.jpg"
    alt="Odeon" />..</p>

<p>Cliquer sur le logo <a
    href="http://www.univ-
    lyon1.fr/"><img
src="logo-ucbl-3D.png"
    alt="Logo UCBL" /></a>
    pour visiter le site Web
    de Lyon 1.</p>
</body>
</html>
```



Vitesse d'une transaction HTTP

- Temps d'affichage d'une page Web lié
  - à la complexité de la page à afficher
    - nombre et volume des ressources composant la page
    - traitements éventuellement nécessaires pour générer dynamiquement ces ressources
  - au débit et à l'encombrement des réseaux parcourus
  - à la possibilité de présenter une page avant que toutes les données soient arrivées
    (ex. texte avant images, chargement asynchrone…)
  - à la charge du serveur
    - réponse à de nombreux clients en même temps
    - cryptage ou non des données (HTTP / HTTPS)
    - autres applications tournant sur la même machine
  - à la vitesse de la machine cliente

Serveur Web

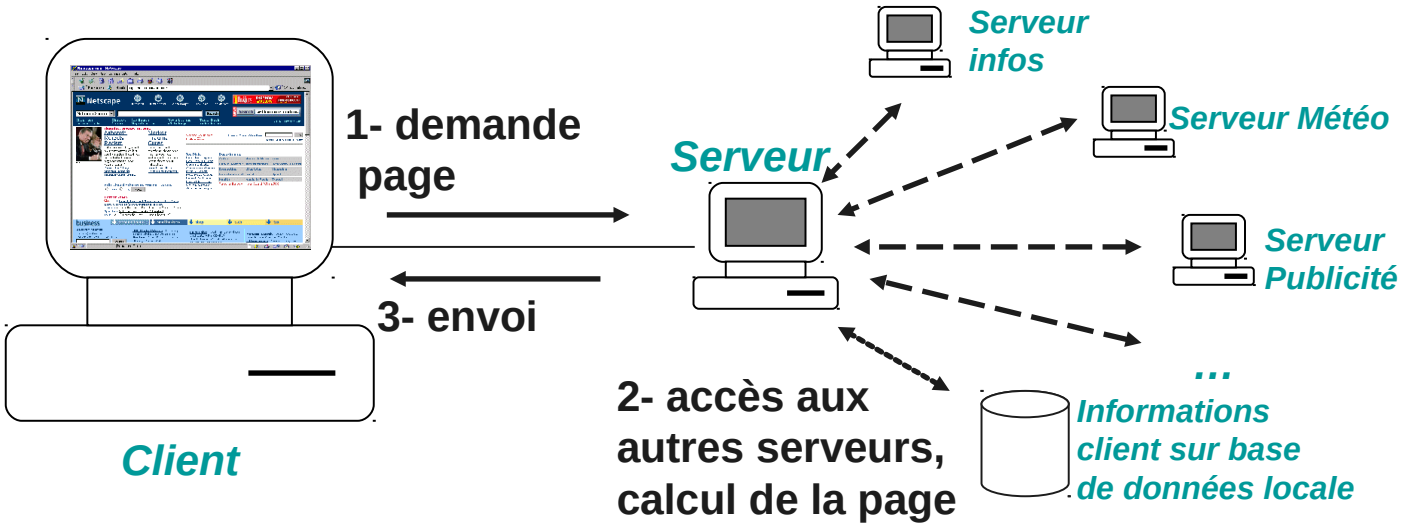
- Définition
  - Application dont le rôle est de fournir des ressources Web aux clients qui les demandent
    - par le protocole HTTP(S)
    - sur Internet ou sur un réseau local
  - Par extension, machine sur laquelle s'exécute cette application

- Deux façons de fournir des ressources
  - statique : à partir de fichiers existants sur le serveur
  - dynamique : en générant la ressource au moment de la requête

Pages générées dynamiquement

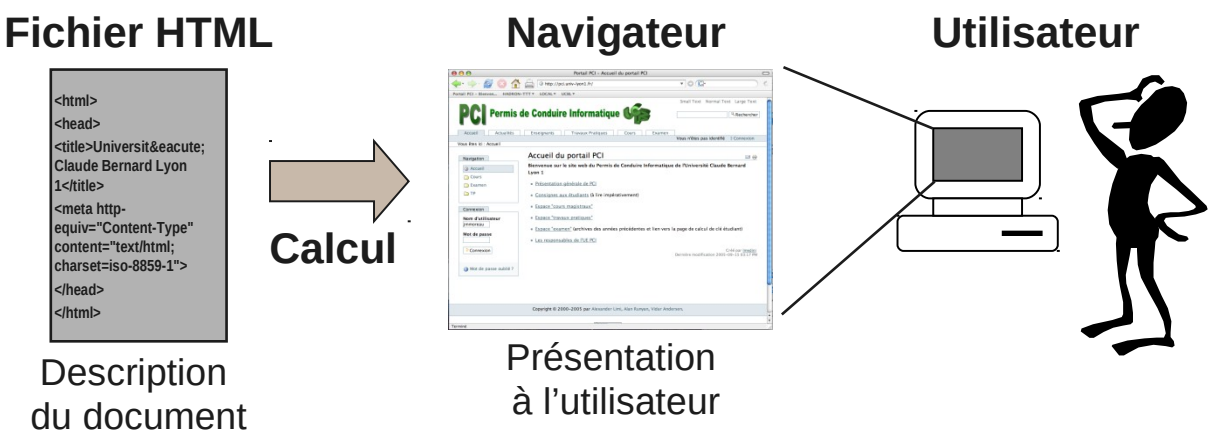
- Le serveur doit calculer la ressource avant de la renvoyer
  - en fonction des données
    - indiquées par le client (formulaires, cookies)
    - présentes sur le serveur (fichiers de données, pages statiques)
    - présentes sur d'autres serveurs (web ou bases de données) : le serveur considéré est le client d'autres serveurs
  - avec un langage de programmation côté serveur (PHP, ASP, Java, Python…)

- Le mécanisme de requête / réponse reste le même



Navigateur Web (browser, butineur)

- Caractéristiques
  - logiciel client pour les protocoles HTTP et HTTPS
  - demande des fichiers (X)HTML à un serveur
  - interprète ces fichiers
  - demande éventuellement d'autres ressources
  - les présente à l'utilisateur
  - peut aussi réaliser quelques traitements (scripts)
- Principaux navigateurs
  - Mozilla Firefox, Safari, Google Chrome, etc.



Différents types de documents accessibles sur le Web

- Pages Web personnelles
  - document libre
- FAQ
  - Foire Aux Questions, Frequently Asked Questions
  - ensemble de questions/réponses classiques sur des sujets très variés (science, vie pratique...)
  - lié à des communautés
- How-to
  - description de la manière de faire une certaine action (ex. comment installer un logiciel)
- Articles (cf. journaux, documents d'information)
  - forme ± standard et figée (auteur, titre, date...)
- Tableaux synoptiques, listes
  - réponse à une requête (offres d'emplois...)
- Cartes et plans

Différents médias accessibles sur le Web

- Formes du point de vue de l'utilisateur
  - texte
  - image (photo, dessin, graphique)
  - son (musique, voix)
  - vidéo (images animées + son)

- Formats de fichiers correspondants
  - texte simple : ASCII (.txt)
  - images : JPEG (.jpg), GIF (.gif)
  - sons : WAVE (.wav), MPEG3 (.mp3)
  - vidéo : MPEG (.mpg), AVI (.avi), Quicktime (.qt)…

- Applications permettant de « lire » les fichiers (ouvrir les documents)
  - éditeur de texte
  - visualisateur d'images
  - lecteur de son, de vidéo

Document multimédia

- Combinaison de différents médias
  - texte, image, son, vidéo…
- Deux caractéristiques principales
  - « ça bouge » : le document se déroule temporellement (vidéo, son, objets animés)
  - « c'est interactif » : on peut cliquer sur un bouton, suivre un lien hypermédia, remplir une zone texte...
- Exemples d'utilisations
  - cours, jeux interactifs, logiciels éducatifs, encyclopédies interactives, bornes dans les lieux publics, pages Web, films DVD…

Page Web

- Une page Web, c'est
  - un document multimédia
    - images, textes, possibilité d'interaction, liens
  - décrit élément par élément
    - titre, morceaux de texte, images…
  - avec un langage de description
    - HTML (1992) ou XHTML (2002)
  - stocké dans un fichier
    - .html (ou .htm, .xhtml)
- Une page Web
  - est calculée et affichée par un navigateur
  - est localisée sur Internet à l'aide d'une adresse (URL)
  - permet d'accéder à d'autres pages en suivant des liens

Site Web

- Regroupement de pages Web
  - autour d'une thématique commune
  - reliées entre elles par des liens hypertextes
  - émanant d'une même entité (organisation, entreprise, particulier…)
  - accessibles à partir d'une page d'accueil commune
  - accessibles à partir d'une URL de base commune
  - en général, localisées sur un même serveur

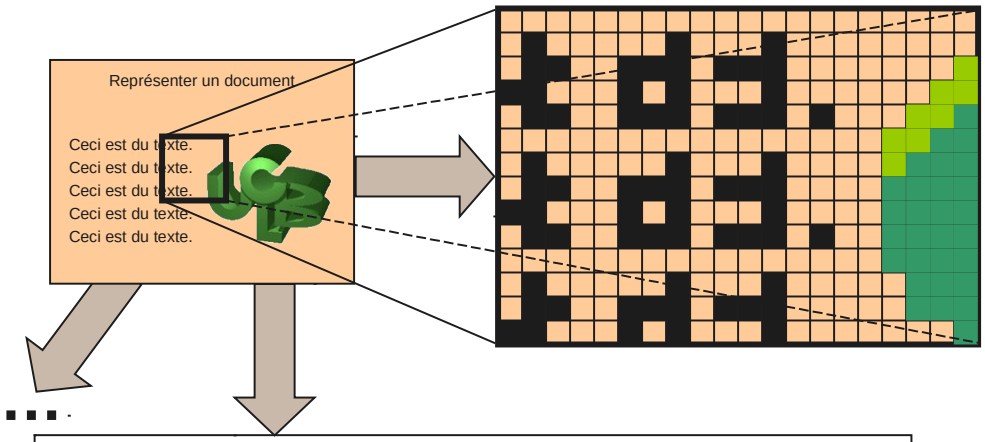
- Organisation hiérarchique
  - structure sous forme d'arborescence de dossiers et de fichiers (comme un disque local)
  - permet de définir des chemins relatifs
  - **exceptions** (de plus en plus fréquentes) : pages générées dynamiquement

- Liens vers et à partir d'autres sites
  - permettent de naviguer d'un site à un autre
  - composent la « toile » mondiale (World Wide Web)

Le « Web 2.0 »

- Principe du Web « classique »
  - ce sont les organisations qui détiennent des sites qui décident de l'information qui y figure
    - Communication de type « one-to-many » (= diffusion)
- Principe du « Web 2.0 »
  - donner le contrôle de l'information aux utilisateurs
  - faire émerger des « réseaux sociaux »
    - chacun peut déposer des contenus
    - pour donner son avis sur un sujet donné (blogs)
    - pour partager ses documents (images, vidéos…)
    - pour étiqueter (« tagger ») des contenus existants
- Aspects techniques
  - outils (protocoles, clients, serveurs) identiques à ceux du Web classique
  - nécessite plus de puissance de calcul (pages dynamiques)
  - nécessite plus d'espace de stockage (contenus envoyés par les utilisateurs)
- Exemples de sites « Web 2.0 »
  - Wikipédia, Del.icio.us, Technorati, Flickr, Picasa Web album, Dailymotion, YouTube, Kartoo…

Plusieurs façons de décrire un document



Description « par points »

- coordonnées
- couleur

```
Document = titre + image + paragraphe
 Titre = Texte + Position
         Texte = "Le titre"
         Position = Haut ET Centré
 Paragraphe = Texte + Position
         Texte = "Blablah… "
         Position = Gauche
 Image = Fichier + Position
         Fichier = "logo-ucbl.gif"
         Position = ADroite Paragraphe
```

Description « par éléments »

document décomposé en zones,

puis zones décrites en détaillant petit à petit les éléments

Un document comme composition d'éléments

- Un élément
  - est une partie significative d'un document
    - TITRE, AUTEUR, PARAGRAPHE, IMAGE, LIEN, TABLEAU…
  - peut être composé d'autres éléments
    - un DOCUMENT est composé d'un TITRE et d'une suite de PARAGRAPHE
    - un TABLEAU est composé de LIGNEs
    - une LIGNE est composé de CASEs…
  - peut avoir des attributs
    - un attribut "nom du fichier" pour un élément IMAGE
    - un attribut "URL de destination" pour un élément LIEN
- Pour les pages Web
  - Environ 40 éléments de description dans les langages (X)HTML

HTML (HyperText Markup Language)

- HTML permet de décrire des pages Web dans des fichiers textuels (ASCII) en utilisant des **balises**
  - mots-clés simples pour délimiter des descriptions d'éléments (`title`, `img`, `p`...)
  - balises ouvrantes (`<title>`), fermantes (`</title>`), vides (`<hr />`)
  - balises avec attributs (`<a href="http://pci.univ-lyon1.fr">`)

- **Un élément**
  - est soit délimité par deux balises…
    - entre lesquelles se situe son contenu : `<p>Un paragraphe...</p>`
    - dans de cas, il peut contenir d'autres éléments (et ainsi de suite)

      `<p>Un paragraphe avec un mot en <b>gras</b> (bold).</p>`

      (= l'élément **p** contient l'élément **b**)
  - …soit entièrement décrit par une balise vide : `<br />`
    - avec éventuellement des attributs : `<img src="zebre.jpg" alt="image de zèbre" />`

Exemple de structuration en (X)HTML

Balise ouvrante →

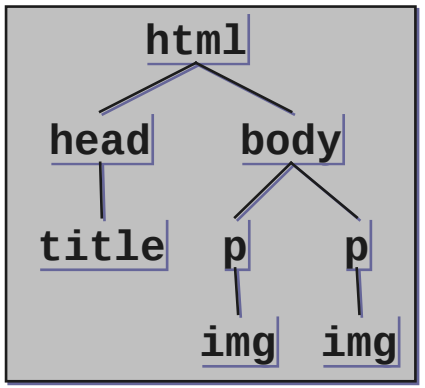
```
<html>
   <head>
    <title>Merveilles du monde</title>
   </head>

   <body>
    <p>Dans l'image ci-après
   <img src="zebre.jpg" alt="Image de zèbre"
   />
   nous présentons un zèbre vivant
   dans la savane.
    </p>
    <p>
   <img src="lion.gif" alt="Image de lion" /
   >    Le lion n'est pas vraiment son ami
    </p>
   </body>
</html>
```

Élément avec contenu

Attributs de l'élément Image

Balise d'élément vide (une seule balise) →

Balise fermante →



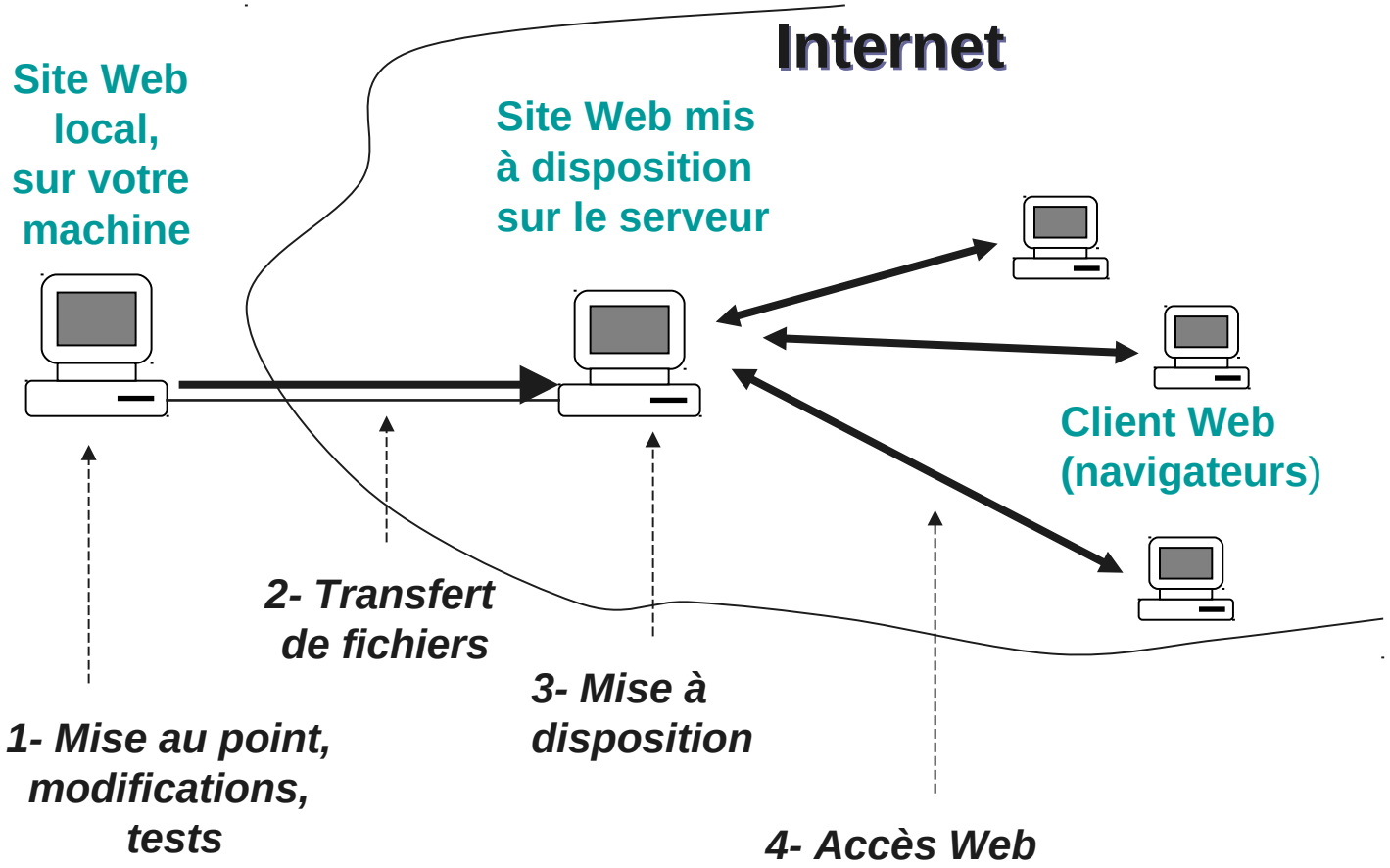
Arborescence

Plusieurs situations type

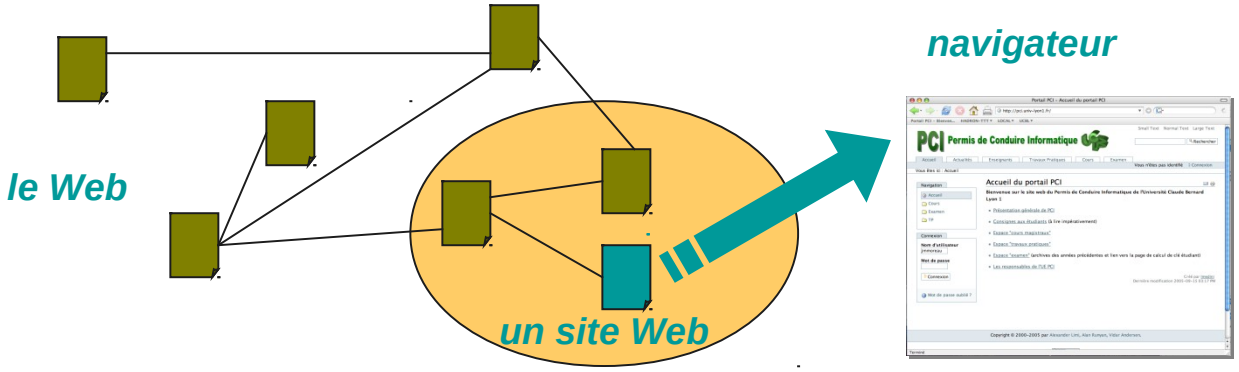
- Cas d'un particulier
  - site chez le Fournisseur d'Accès Internet
    - conception en local à la maison
    - transfert des fichiers chez le FAI par FTP
  - site sur machine à la maison
    - conception en local sur une partie du disque
    - copie de fichiers sur le site publié
- Cas d'une entreprise
  - site externalisé (conception, mise à disposition)
  - site conçu en interne et mis à disposition sur un serveur de l'entreprise

Conception, transfert et accès

- Dans tous les cas
  - Cycle de conception et de publication d'un site Web



World Wide Web : une description plus formelle



- Ensemble de pages Web formant un réseau (ou une toile : *web* en anglais) sur lequel il est possible de naviguer
- WWW = World Wide Web
- Navigateur
  - outil permettant d'afficher des pages HTML et de passer d'une page à l'autre au moyen de liens hypertextes → navigation

Qu'est-ce que naviguer sur le Web ?

- Navigation proprement dite
  - aller directement à une page en tapant son URL
  - aller d'une page à une autre en suivant un lien
  - remplir et envoyer des formulaires

- Activités liées à la navigation
  - arrêter des chargements trop longs (bouton Stop)
  - recharger une page qui a été modifiée (bouton Reload)
  - mener plusieurs navigations en parallèle (plusieurs fenêtres / onglets)
  - définir et organiser ses signets
  - envoyer des courriers électroniques en cliquant sur des liens
  - télécharger des fichiers

- Attention à la communication de vos données personnelles
  - l'échange de données peut être sécurisé (ex. paiement par Carte Bleue)
  - mais vos informations arrivent toujours quelque part

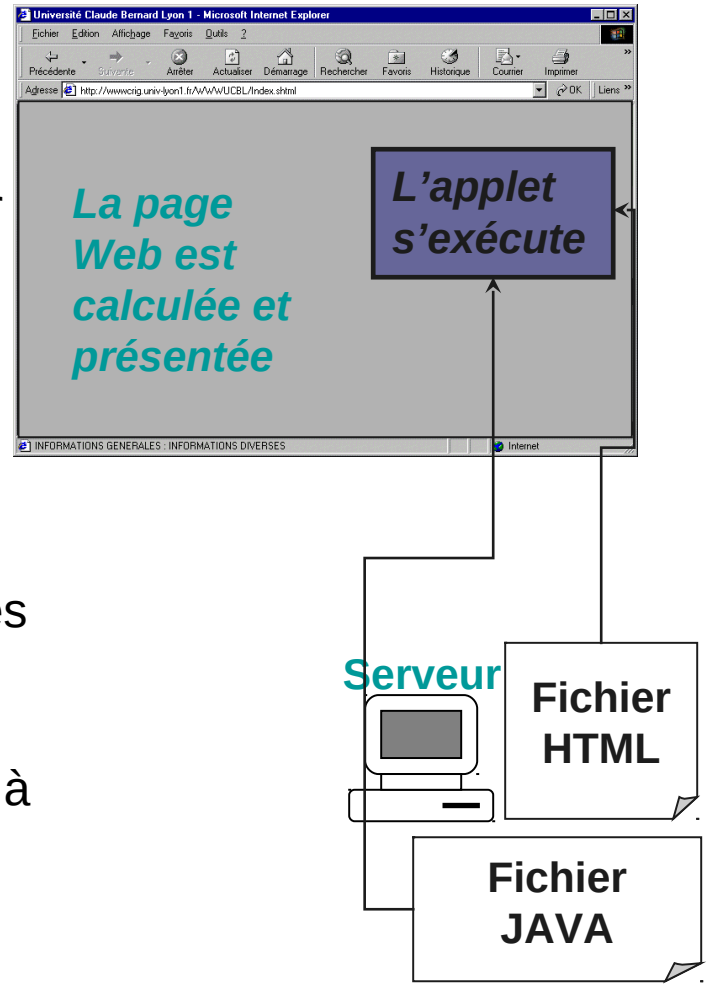
    avez-vous confiance en la personne ou l'entreprise qui les réceptionne ?

Interpréteurs de scripts

- Scripts
  - éléments de programmes (identifiés par l'élément **<script>**)
    - à l'intérieur des pages HTML ou dans des fichiers séparés
    - écrits dans des langages de programmation spécifiques (JavaScript, VBScript)
    - exécutés (interprétés) par le navigateur de l'ordinateur client
  - soumis à des restrictions de sécurité (pas d'accès aux fichiers du client, pas d'exécution d'autres programmes…)
- Exemples
  - gestion du navigateur (lancer des fenêtres, créer des info-bulles…)
  - gestion de la navigation (vérifier qu'un formulaire est bien rempli avant de l'envoyer…)
  - documents plus interactifs et dynamiques qu'avec (X)HTML seul
- Il est possible – mais pas recommandé – de désactiver l'interpréteur Javascript du navigateur

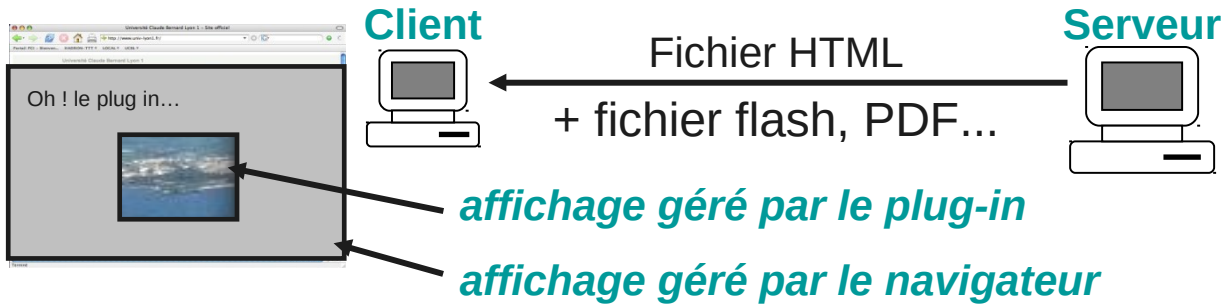
Exécution d'applets Java

- Java
  - langage de programmation qui peut fonctionner sur tout type de machine (matériels et systèmes d'exploitation différents)
  - doit être installé sur la machine avant de pouvoir exécuter des programmes
- Applet
  - élément de programme Java
  - restrictions de sécurité au niveau des fonctionnalités autorisées
- Utilisation
  - dans une page Web, on spécifie le nom de l'applet à exécuter dans un élément <applet>
  - l'applet s'exécute dans une portion de la page présentée à l'utilisateur



Plug-in

- Principe
  - étendre les possibilités du navigateur pour présenter des éléments HTML (décrits par l'élément **<object>**) non gérés par le navigateur
  - le navigateur donne toute ou une portion de la page Web à un programme « branché » (plug-in) qui prend en charge l'exécution ou l'affichage (cf. applet Java)
  - S'il n'est pas disponible, on peut en général le télécharger et l'installer
- Exemples
  - Flash / Shockwave : documents multimédias
  - Real Audio / Vidéo : sons ou vidéos
  - Acrobat : documents PDF

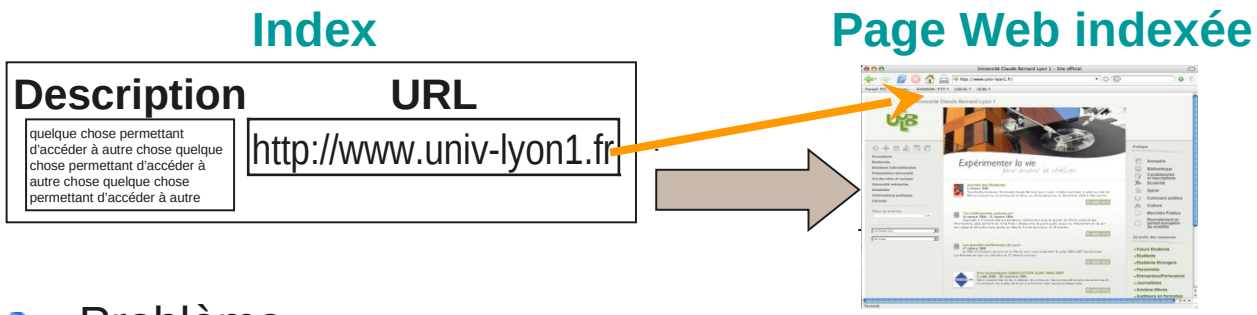


Recherche d'information sur le Web

- Recherche d'information
  - navigation simple (plus ou moins au hasard)
    - en accédant directement à des URL connues
    - en naviguant de lien en lien
  - assistance à la recherche : outils
    - annuaires
    - moteurs de recherche
    - portails

Indexation (1/2)

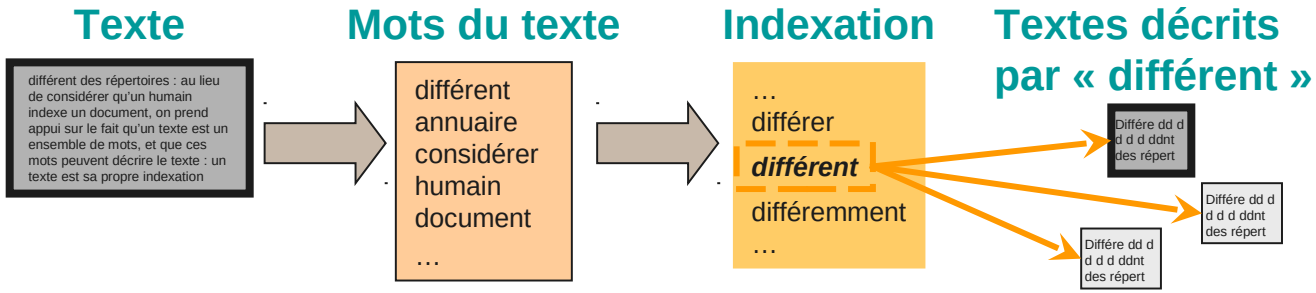
- Index
  - quelque chose permettant d'accéder à autre chose
  - marque-page, table des matières
  - ex. des bibliothèques : Auteur, Titre, ISBN, Mot-clés...
- Sur le Web, un index c'est
  - une description du site
    - ex. « Ce site est le site officiel de l'UCBL, il présente les diverses composantes de la faculté... »
  - un identifiant qui permet l'accès au site : son URL



- Problème
  - sur le Web, comment indexer (décrire) des documents pour aider un utilisateur à les trouver ?

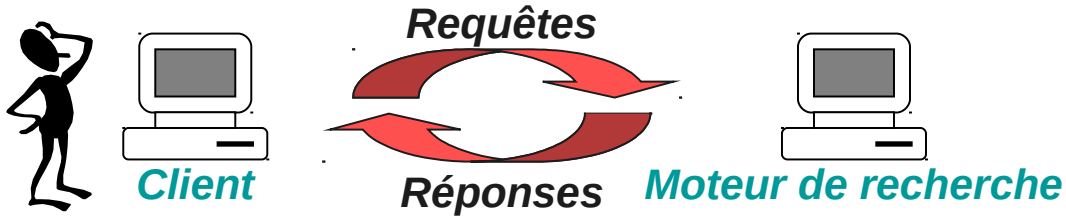
Indexation (2/2)

- Indexation « en texte intégral » (utilisée dans les moteurs de recherche)
  - Principe
    - un texte est un ensemble de mots
    - ces mots peuvent décrire le texte
    - donc un texte contient ses propres index
  - Application
    - chercher automatiquement dans un texte les mots-clés qui vont le décrire, puis les stocker comme index du texte
  - Traitements statistiques
    - chercher les mots les plus pertinents (ceux qui n'apparaissent pas trop souvent)
    - éliminer les « mots vides » (est, au, le, la…)



Moteurs de recherche

- Principe
  - des **robots** (logiciels spécifiques) parcourent le Web, lisent le texte des pages, les indexent automatiquement et stockent les résultats
  - un site Web comportant un **formulaire de recherche** est proposé aux internautes
  - pour interroger le moteur, un utilisateur soumet une **requête** (*i.e.* l'ensemble des termes à rechercher dans les documents)
  - lorsqu'une requête arrive, le moteur de recherche
    - retrouve dans les index l'ensemble des documents contenant les termes de la requête
    - classe ces documents par ordre de pertinence selon un algorithme particulier (**ranking**)
    - génère dynamiquement des **réponses** comportant des descriptions des documents retrouvés et des liens vers ces documents, et les présente à l'utilisateur
- Cycle de recherche
  - S'il obtient trop de réponses ou des réponses non satisfaisantes, l'utilisateur peut modifier sa requête pour affiner les résultats



Position du problème

- Nous passons notre temps à rechercher de l'information
  - exemple : s'inscrire à la fac
    - trouver les formulaires, les bureaux…
- Le degré de confiance accordé aux informations trouvées varie
  - suivant les enjeux
    - plus ou moins grand besoin de fiabilité
  - suivant la source de l'information
    - document officiel
    - prospectus publicitaire
    - opinion d'un internaute (blog, forum)…

Sources d'information sur Internet

Pages personnelles

Un particulier met un ensemble de connaissances à disposition du reste du monde (coin à champignons, mécanique, comment faire une page Web...)

Pages d'organisations

Un groupe d'individus met à disposition des informations.

Pages institutionnelles

Une institution de l'État joue son rôle en mettant des informations à disposition des citoyens (Ministère des Finances, Journal Officiel...)

Organisations à but non lucratif

(Amicale des Boulistes de Villeurbanne, Médecins du Monde…)

Organisations à but lucratif

(Casse-auto 69, Infogrames, Total-Elf)

Confiance en l'information disponible sur le Web

- Quelle information sur le Web ?
  - à peu près tout, sur tout…
  - …et son contraire (fiabilité discutable)

- Il faut analyser l'information en fonction
  - de sa source
  - de sa forme (type de document)
  - de la confiance que vous lui accordez

- Questions
  - auteur identifié ?
  - auteur identifiable ?
  - institution ?
  - importance de la validité (puis-je utiliser l'information comme preuve) ?
  - date de rédaction ?
  - durée de validité de l'information ?