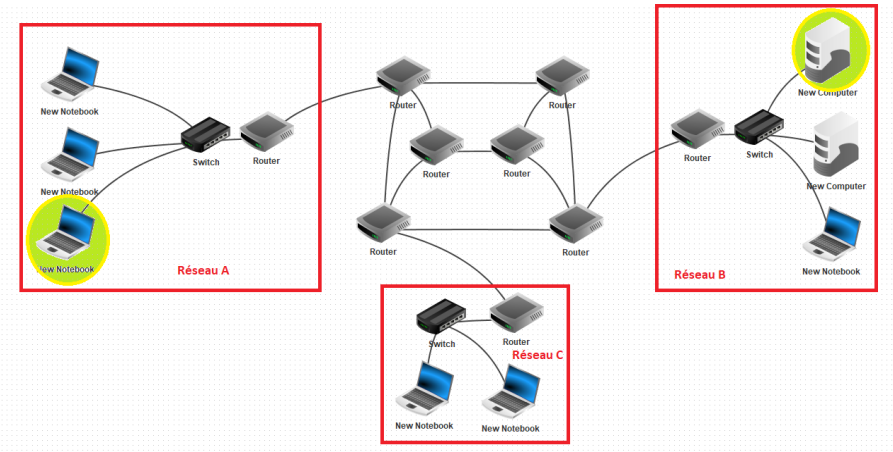
02 – Décoder les URL

Qu'est-ce que le Web ? Qu'est-ce qui le caractérise le plus ?
Nous avons vu qu'il s'agissait de la notion de documents hypertextes accessibles via des liens (ou hyperliens). Nous allons voir comment on parvient à trouver une page précise.

Reprenons l'exemple de l'ordinateur de réseau A qui veut accéder à une ressource hypertexte situé sur le serveur d'un réseau B. Comment le trouve-t-il ?



A – URL

01 - Ouvrir un navigateur (qui est un client HTTP) et copier cela dans la barre d'adresse du navigateur :

`http://info.cern.ch/hypertext/WWW/TheProject.html`

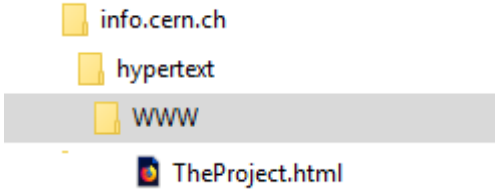
On voit que l'adresse est composée de plusieurs parties :

- `http` : on commence par signaler le **protocole** qu'on désire utiliser pour communiquer avec le serveur. On doit faire suivre le nom du protocole par le caractère « deux points » « `:` » pour signaler qu'on a fini de transmettre le nom du protocole.
- `//` veut dire que l'adresse qu'on donne ensuite est l'**adresse du service qu'on tente de joindre si le protocole le nécessite.** C'est le cas pour le protocole HTTP qui nécessite de se connecter à un serveur HTTP. Ici, on devra donc le faire suivre du nom de domaine du site.
- `info.cern.ch` : cette partie identifie le site sur lequel on veut se connecter.
- `/` veut dire que l'adresse qui suit est une adresse **absolue** sur le site : on va donner un chemin partant du « début », **la racine du site**.
- `info.cern.ch/hypertext/WWW/TheProject.html` veut donc dire d'aller
- sur le serveur http gérant `info.cern.ch`
- d'aller dans le répertoire `hypertext` *qui se trouve à la racine du site*
- puis dans le répertoire `WWW` *qui se trouve dans le répertoire* `hypertext`
- puis de demander le document `TheProject.html` *qui devrait y être stocké.*

URL veut dire ***Uniform Resource Locator***, soit « localisateur uniforme de ressource ». C'est un système normé (possédant des règles) permettant d'effectuer une demande sans créer

d'ambiguïté sur la demande elle-même.

Si on résume l'adresse visuellement avec une structure de dossiers imbriqués, on peut avoir :



La requête est donc décomposable en plusieurs parties :

- Nom du protocole : `http:`
- Nom du service à joindre : `//info.cern.ch`
- Adresse du document : `/hypertext/WWW/TheProject.html`

Les navigateurs Web sont néanmoins créés avec une volonté de robustesse : ils sont capables de compléter des URL incomplets.

02 – Tapez `info.cern.ch/hypertext/WWW/TheProject.html`
03 – Tapez `info.cern.ch/hypertext/WWW/`
04 – Tapez `info.cern.ch`

Conclusion ?

05 – Tapez `info.cern.ch/hypertext/WWW/TheProgect.html` **avec un g!**

La plupart des sites proposent des pages 404 personnalisées. Vous pouvez tenter celle-ci : on se connecte sur le site de blizzard et on demande `totovaalaplage.html` qui n'existe pas :

`http://eu.blizzard.com/fr-fr/totovaalaplage.html`

Reste une question : **comment parvient-on à trouver le bon serveur dans l'immensité d'Internet ?**

B – Moteur de recherche

Au tout début du Web, il y avait des sites qui tentaient de maintenir une liste des sites. Au fur et à mesure de l'évolution du Web, il est vite devenu évidemment que tenir à jour une telle liste était illusoire.

C'est ainsi qu'est née la nécessité du **moteur de recherche**.

Qu'est-ce qu'un moteur de recherche ?

C'est une application qui attend qu'on lui fournisse une liste de mots clés et qui fournit en retour un document hypertexte permettant d'obtenir une liste de sites liés aux mots clés.

Comment le moteur de recherche parvient-il à fournir un résultat ?

La société qui possède le **moteur de recherche** utilise en permanence des clients HTTP qui visitent les sites et suivent les liens qui s'y trouvent. Ces clients ne sont pas des navigateurs manipulés par des employés. Non, il s'agit de simple programmes qu'on nomme **robots d'indexation** (ou web crawlers).

Le robot d'indexation analyse le texte présent sur la page. Il va ensuite fournir ses résultats et la société associe cette page à un certain nombre de mots-clés. C'est l'un de critères de la recherche.

Mais imaginons qu'on ai 20 000 pages parlant d'un sujet.

06 – A votre avis, comment un moteur de recherche fait-il pour classer les résultats obtenus du plus pertinent au moins pertinent ?

07 – Faire une recherche sur +SNT+Informatique sur Google, Qwant et DuckDuckGo. Comparer les résultats. Obtient-on les mêmes résultats ? Qu'est-ce qu'un lien sponsorisé ?

Si vous ne payez pas, vous obtenez un résultat lié à un référencement naturel. S'il ne s'agissait que d'explorer les pages et de récupérer les mots, les moteurs devraient rendre un résultat plus ou moins similaire. Mais c'est plus complexe que cela.

La société Google a été créée en 1998, à une époque où il y avait déjà plus d'un million de sites disponibles. Elle a rapidement écrasé ses concurrents.

Pourquoi ? Google donnait des résultats beaucoup plus pertinents.
Comment ? Avec un **algorithme** créé par Larry Page, cofondateur de Google, inspiré de la manière dont les scientifiques classent les articles scientifiques.

Cet algorithme est devenu depuis un marque déposé : **PageRank : Google** mettait des notes aux pages. Plus une page semblait importante, plus elle apparaissait tôt dans la liste. La note finale est comprise entre 0 au minimum et 1 au maximum.

Le principe tient en quatre points :
Point 1 : On effectue un fonctionnement aléatoire : on prend une page au hasard et on regarde les liens qu'elle fournit et suit l'un des liens.
Point 2 : Plus on détecte de liens vers un site, plus sa note augmente.
Point 3 : Plus le lien provient d'un site important, plus le lien apporte de points.
Point 4 : Plus un site crée de liens, moins on donne d'importante aux liens qu'il fournit (cela permet de lutter contre les sociétés tentant d'augmenter artificiellement les notes de leurs clients).

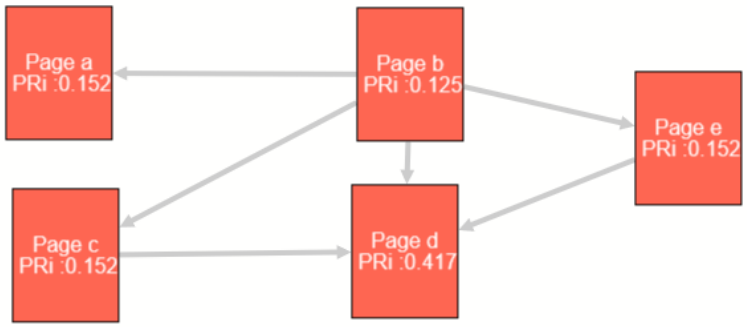
Après la théorie, passons à la pratique. Voici plusieurs sites proposant une simulation de PageRank **:**

http://faculty.chemeketa.edu/ascholer/cs160/WebApps/PageRank/

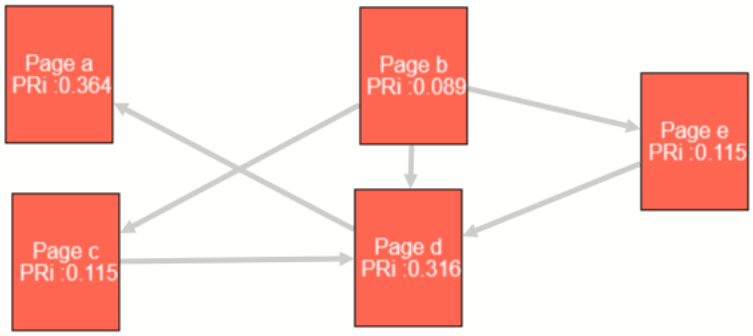
https://www.seoquantum.com/test-pagerank

08 – Réaliser la simulation correspond au graphe ci-dessous : aucun lien. Pourquoi obtient-on une note de 0,2 pour chaque page ?

09 – Modifier la configuration pour obtenir un site qui est cité 3 fois et un site qui fournit 3 liens.Qui est le site le mieux noté ? Le moins bien noté ?



10 – Modifier le lien vers la page a : ce n'est plus la page b (mal notée) mais la page d (bien notée) que la propose. Est-ce important d'être cité par un grand site ?

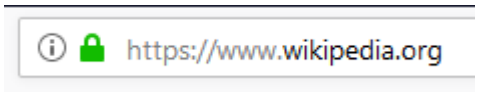


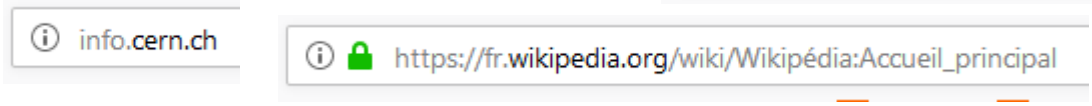
C – Nom de domaine

Lorsqu'on possède un nom de domaine (`cern.ch`, `wikipedia.org` …), on peut se créer autant de sous-domaine qu'on le désire. Par contre, ils seront toujours **à gauche** dans le nom permettant d'accéder au site, pas à droite.

Le sous-domaine le plus courant est `www` pour world wide web. Il était utilisé très couramment pour montrer que cette adresse menait à un site Web.

Sous Firefox, les sous-domaines apparaissent grisés.





Remarque : un nom de domaine est toujours composé d'un nom suivi d'une extension (.com, .fr, .org …) qui correspond au domaine de premier niveau. Les sous-domaines, le domaine et l'extension sont tous séparés par un point (dot en anglais).

11 – Remplir les cases pour les adresses proposées dans le tableau ci-dessous.

URL	Sous-domaine	Domaine	Extension (domaine de premier niveau)
`www.wikipedia.fr`			
`fr.wikipedia.org`			
`www.openoffice.org`			
`www.qwant.com`			

12 – Un e-commerce vous propose de payer un magnifique achat en ligne sur

`www.paypal.paiement.securise.com`. C'est sérieux ou pas cette histoire ?

13 – A quoi fait référence le cadenas vert ?

Remarque : que faire face à un triangle jaune ?

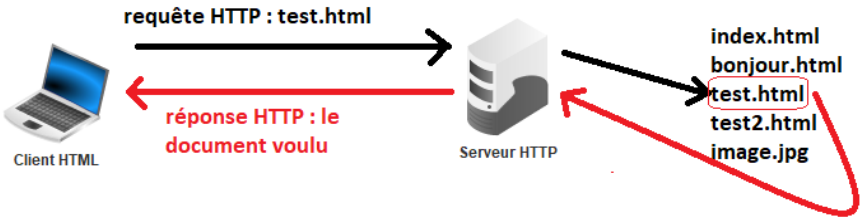
14 – Pourquoi tous les sites ne sont-ils pas passés en https ?

Mais parfois on voit des urls bizarres, avec des ?, des & et plein de chiffres. Et pas d'extention .html ou autre. C'est un bug ?

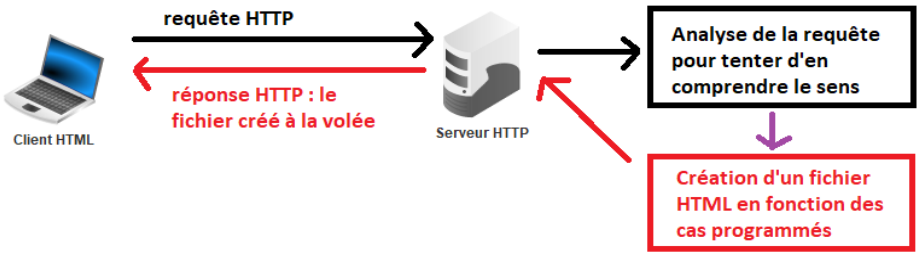
D – SITE DYNAMIQUE et les URL bizarres qu'on y rencontre

Un **site <u>statique</u>** est un site dont le serveur HTTP ne peut faire qu'une chose : fournir des documents déjà présents : les requêtes doivent contenir des URL faisant référence à un chemin

menant à un document existant. L'exception reste la page d'accueil **index.html** qu'on peut atteindre en tapant simplement le nom du site.



Un **site dynamique**, c'est tout autre chose au niveau des URL : le serveur HTTP reçoit la requête et celle-ci pourra être analysée par un programme. Le programme peut alors décider de renvoyer n'importe quel contenu en fonction de ce qu'il y a dans la requête.
C'est notamment pratique pour les recherches !



15 – Allez sur Wikipédia et lancez une recherche sur « HTML ». La page reçue est-elle une page html ? En cas de doute, vous pouvez visualiser le code source (le code que le navigateur interprète pour créer le visuel visible à l'écran) : clic-droit sur la page et sélectionner **CODE SOURCE DE LA PAGE**.

16 – Allez sur Wikipédia et lancez une recherche sur « page HTML ». La page reçue est-elle une page html ? En cas de doute, vous pouvez visualiser le code source (le code que le navigateur interprète pour créer le visuel visible à l'écran) : faire un clic-droit sur la page puis sélectionner CODE SOURCE DE LA PAGE.

```
https://fr.wikipedia.org
/w/index.php
?search=page+html
&title=Sp%C3%A9cial%3ARecherche
&go=Continuer
&ns0=1
```

17 – En modifiant uniquement manuellement les **paramètres** sur l'URL de votre dernière page Wikipédia, tentez de lancer une recherche sur **page+php+parametre**.

Les URL obtenues sur les sites dynamiques sont souvent longues et incompréhensibles. Avec un peu de programmation, on peut faire mieux : on peut fournir des URL « propres ». Cette technique a longtemps été utilisée car on disait que les moteurs de recherche indexaient mieux les sites qui avaient des URL « compréhensibles ».

18 – Utiliser le lien suivant qui ne fait référence à aucun document .html ou .php en particulier : on donne simplement le titre de l'article.

https://www.larecherche.fr/astrophysique/la-première-image-dun-trou-noir-révélée

Nous en savons maintenant bien plus sur les sites Web et leurs URL.

Il reste maintenant à voir comment on les utilise pour générer les fameux hyperliens et les fichiers HTML dans lequels les URLs apparaissent d'une façon ou d'une autre.

Pour aller plus loin : les difficultés juridiques liées aux hyperliens :

http://eduscol.education.fr/internet-responsable/ressources/legamedia/liens-hypertextes.html