Département
Éducation
et Technologie

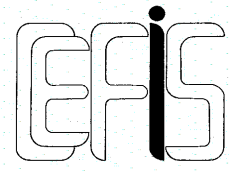
- Internet
- Naviguer
- Principes de navigation
- La recherche

Utiliser Internet et ses services

Guy Vastersavendts

5.49

Octobre 1998



Centre pour la Formation à
l'Informatique dans le Secondaire

Introduction

Jusqu'aux environs de 1993, l'Internet n'était connu que des seuls initiés, essentiellement des universitaires. Depuis 1994, l'Internet s'est ouvert aux activités commerciales et son existence a été portée à la connaissance du grand public.

HTTP signifie HyperText Transfer Protocol

HTML signifie HyperText Markup Language

Ces deux termes sont largement explicités dans la suite.

Actuellement, les média de l'information et de la publicité nous submergent d'articles et de reportages qui font référence à "Internet", "Web", "E-mail" ; les conversations sont remplies de "naviguer", "surfer" … pire encore de "HTTP", "HTML" et autres sigles ésotériques mais, finalement, que signifie tous ces termes, quel est ce monde bizarre habité par un vocabulaire étrange ?

Nous allons pénétrer dans cette jungle et essayer de nous faire une idée, la plus claire possible, sur ce qu'est réellement Internet et les services disponibles.

Nous allons donc tenter de répondre le plus clairement possible à trois questions fondamentales :

Internet,

- qu'est-ce que c'est ?
- à quoi ça sert ?
- comment on s'en sert ?

Les questions sont ambitieuses, les réponses ne seront peut-être pas toujours complètes. Néanmoins, nous nous efforcerons d'aller à l'essentiel et de fournir les éléments de base permettant une approche réfléchie de ce sujet.

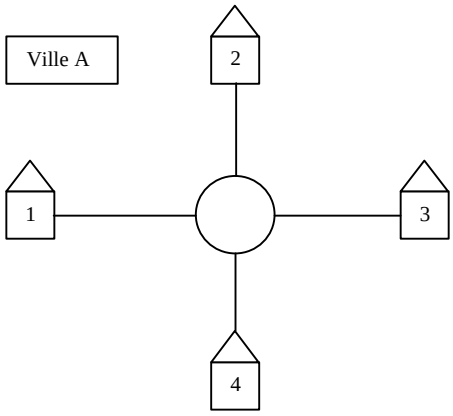
Chapitre I
Internet

Qu'est ce qu'Internet ?

Utilisons une analogie pour mieux comprendre ce qu'est Internet, tout en gardant bien à l'esprit qu'une analogie n'est qu'une image comparative, qui nous aide à mieux comprendre, à mieux se représenter l'objet qu'on veut décrire. Une analogie reste une analogie avec ses limites, il ne faut pas la pousser trop loin sous peine de faire des erreurs.

L'analogie utilisée est celle d'un réseau téléphonique.

Prenons une ville A où les habitants possèdent chacun un téléphone.

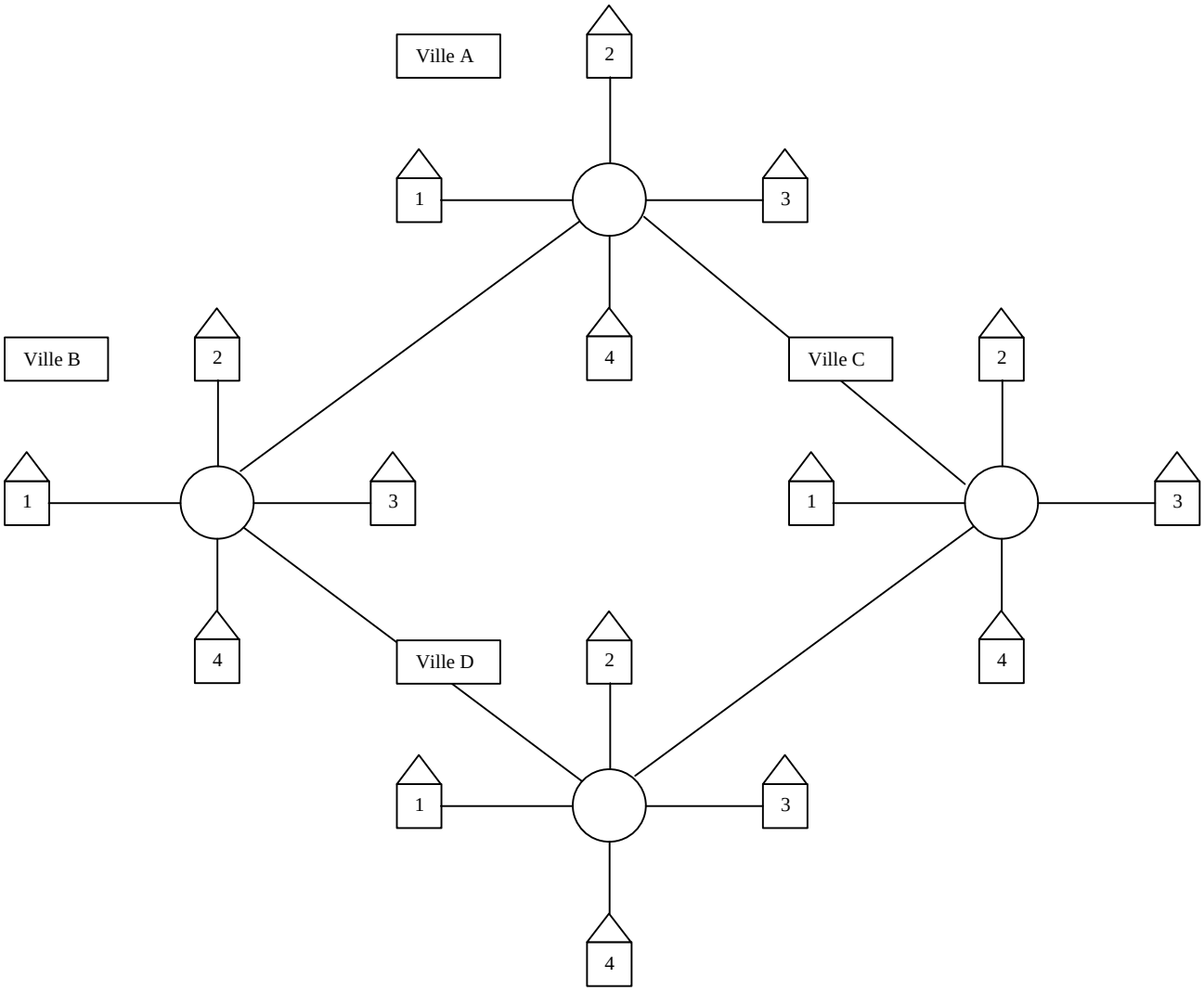


Chaque maison est reliée au central téléphonique grâce à des lignes (fils métalliques) et les habitants peuvent communiquer entre eux. Pour cela, ils doivent respecter certaines règles : décrocher, attendre la tonalité, former le numéro, attendre qu'on décroche de l'autre côté, … Ce système leurs rend un certain nombre de services dont les plus courants sont : se parler, s'envoyer des fax.

Ajoutons dans notre schéma trois autres villes (B, C et D), les habitants de B étant reliés au central téléphonique de cette ville peuvent se téléphoner, tout comme les habitants de la ville A, même chose pour les habitants des villes C et D.

Que faut-il pour qu'un habitant de la ville A puisse entrer en contact avec un habitant de la ville B, C ou D ? Réponse évidente : il faut que les centraux téléphoniques soient reliés entre eux. Bien, posons un câble (métallique ou

de fibre optique) entre ces villes. Notre réseau s'étend et peut s'étendre aux villes, E, F, G, … jusqu'à recouvrir la terre entière. Nous devrons être attentif au fait que chaque habitant possède bien un numéro d'appel unique, pour éviter toute confusion.



Nous pouvons retirer 3 renseignements importants de cette analogie :

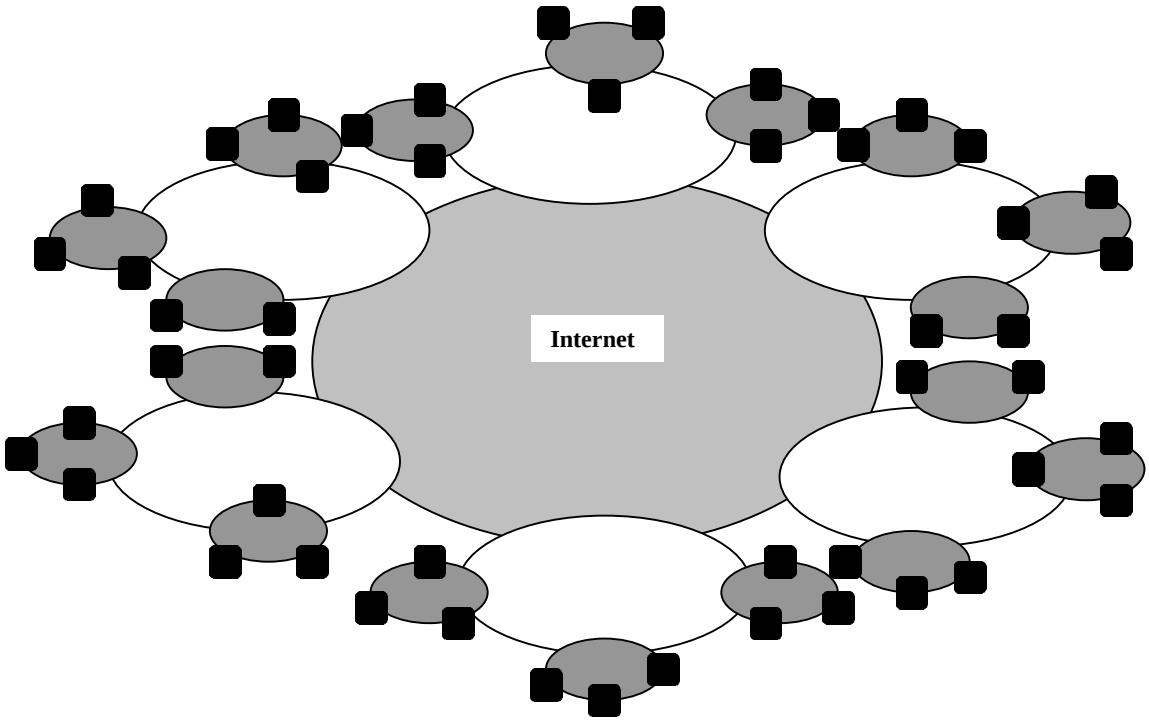
- la notion de **support**, les fils qui relient les maisons aux centraux.
- la notion de **convention**, les règles à respecter lors de l'utilisation du téléphone.
- la notion de **service**, l'utilisation de ce système pour se parler et se "faxer".

Notre schéma ressemble de plus en plus à celui qui représenterait Internet.

Remplaçons les maisons et les centraux téléphoniques par des ordinateurs, reliés physiquement entre eux afin de pouvoir échanger de l'information. Et

nous obtenons des réseaux d'ordinateurs reliés entre eux, c'est-à-dire un réseau de réseaux. Toutes ces liaisons sont matérielles (fils, fibres, …) ; on parle de support du réseau. Pour que ces ordinateurs puissent s'échanger de l'information, ils doivent pouvoir communiquer. Cette communication va respecter certaines règles, utiliser un certain langage ; techniquement, on parlera de protocole de communication. Et enfin, ce réseau nous offrira un certain nombre de services, tous basés sur la communication et l'échange d'informations.

Multiplions cette idée de réseaux interconnectés utilisant un protocole de communication commun et nous obtiendrons une représentation assez correcte d'Internet.



Internet est l'équivalent anglais de inter-réseau. Net(work) signifie réseau.

Du point de vue technique, l'Internet est un réseau international (réseau de réseaux ou inter-réseau) d'ordinateurs communiquant entre eux grâce à un protocole d'échange de données standardisé (TCP/IP – Transport Control Protocol/Internet Protocol). Chaque ordinateur du réseau possède une adresse, appelée adresse IP ou adresse Internet), qui est unique (dans le monde).

Les ordinateurs connectés au réseau Internet peuvent communiquer entre eux de façon transparente pour l'utilisateur, indépendamment du type d'ordinateurs utilisés, mais en utilisant des logiciels appropriés, c'est-à-dire utilisant les protocoles reconnus sur Internet.

Les adresses Internet (adresse IP)

Dans un réseau reliant des millions d'ordinateurs, un des problèmes fondamentaux est celui de l'identification de chacun de ceux-ci.

Chaque ordinateur faisant partie d'Internet possède une adresse unique, qui lui permet d'être identifié de manière spécifique sur le réseau, quelle que soit sa situation géographique. Par analogie, on peut comparer cette adresse à un numéro de téléphone, qui identifie de manière absolument unique un abonné dans le monde entier.

Octet : nombre binaire constitué de 8 bits.

Bit : chiffre binaire (0 ou 1)

Un octet a donc une valeur comprise entre 0 et 255.

Les adresses Internet se composent de quatre octets, chacun étant compris entre 0 et 255. Les quatre nombres sont séparés par des points, par exemple 138.48.38.67. Cette adresse désigne un ordinateur bien précis au Département Education et Technologie, celui qui possède l'adresse IP 67 dans ce réseau.

Ce système de numérotation à quatre octets permet d'identifier 4.294.967.296 de machines différentes.

Le système de noms de domaine (DNS)

La désignation des ordinateurs par une adresse numérique (adresse IP) est une bonne chose pour une machine communiquant avec d'autres machines, mais est nettement moins pratique pour des humains. Aussi a-t-on donné des noms aux machines d'Internet. Le fait d'utiliser des noms pour désigner des machines introduit des problèmes spécifiques, du genre : comment être sûr que deux machines ne possèdent pas le même nom, surtout que cela concerne des millions de machines.

Le système de noms de domaine (Domain Name System – DNS) est une méthode d'administration des noms qui répartit la responsabilité d'attribution à différents niveaux. Chaque niveau dans le système est appelé un domaine.

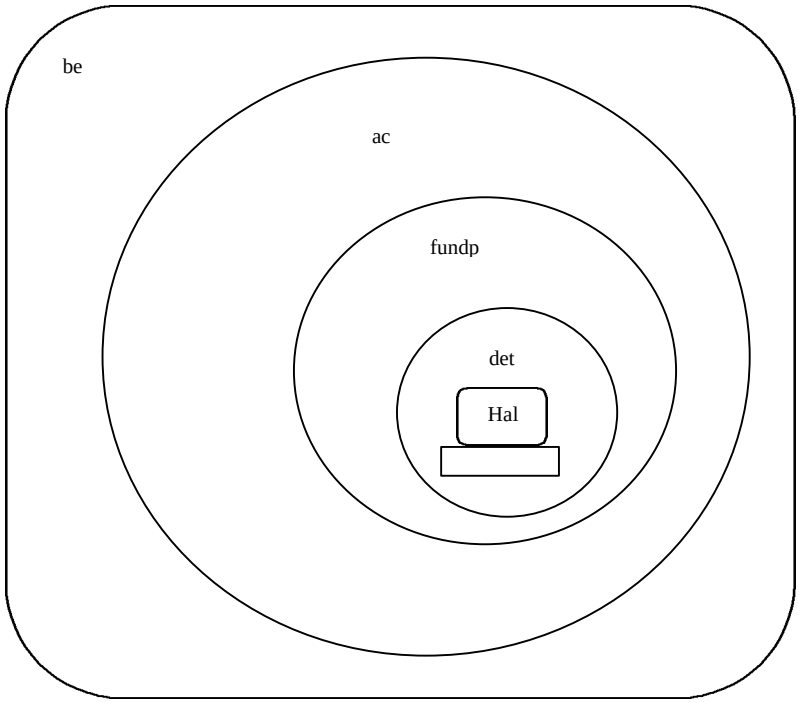
Prenons un exemple, celui d'une machine appelée "hal" au Département Education et Technologie des Facultés Universitaires Notre-Dame de la Paix.

Cette machine appartient au domaine "det" (Département Education et Technologie), qui lui-même appartient au domaine "fundp" (Facultés Universitaires Notre-Dame de la Paix), qui lui-même appartient au domaine "ac" (Académique), qui lui-même appartient au domaine "be" (Belgique). Ouf !

Le DNS de cette machine est donc "hal.det.fundp.ac.be".

L'attribution du nom de la machine "hal" est sous la responsabilité de l'administrateur du domaine "det", l'attribution du domaine "det" est sous la responsabilité de l'administrateur du domaine "fundp", etc.

Figure 1 : Domain Names System



Les principaux domaines de base sont :

Domaine	Usage	Exemple
com	Organisations commerciales	http://altavista.digital.com/
edu	Organisations éducatives (universités, …)	http://www.ctr.columbia.edu/
gov	Organisations gouvernementales non militaires	http://www.nasa.gov/
mil	Armée	http://www.navy.mil/
org	Autres organisations	http://www.oceano.org/
net	Services réseaux	http://rs.internic.net/

Comme Internet est rapidement devenu international, il était nécessaire de trouver un moyen pour donner aux autres pays l'autorité sur leurs noms de domaine. Pour cela, il existe la liste des domaines à deux lettres qui correspond au plus haut niveau de domaine pour chaque pays (be pour Belgique, fr pour France, …).

Un dernier petit problème …

Les machines préfèrent les chiffres, adresse IP (ex. 138.48.38.67), les hommes préfèrent les … noms (ex. hal.det.fundp.ac.be). Qui fait la transformation des noms en adresse IP ? Un ordinateur bien sûr, appelé serveur DNS, qui interroge d'abord le DNS local, qui lui-même peut

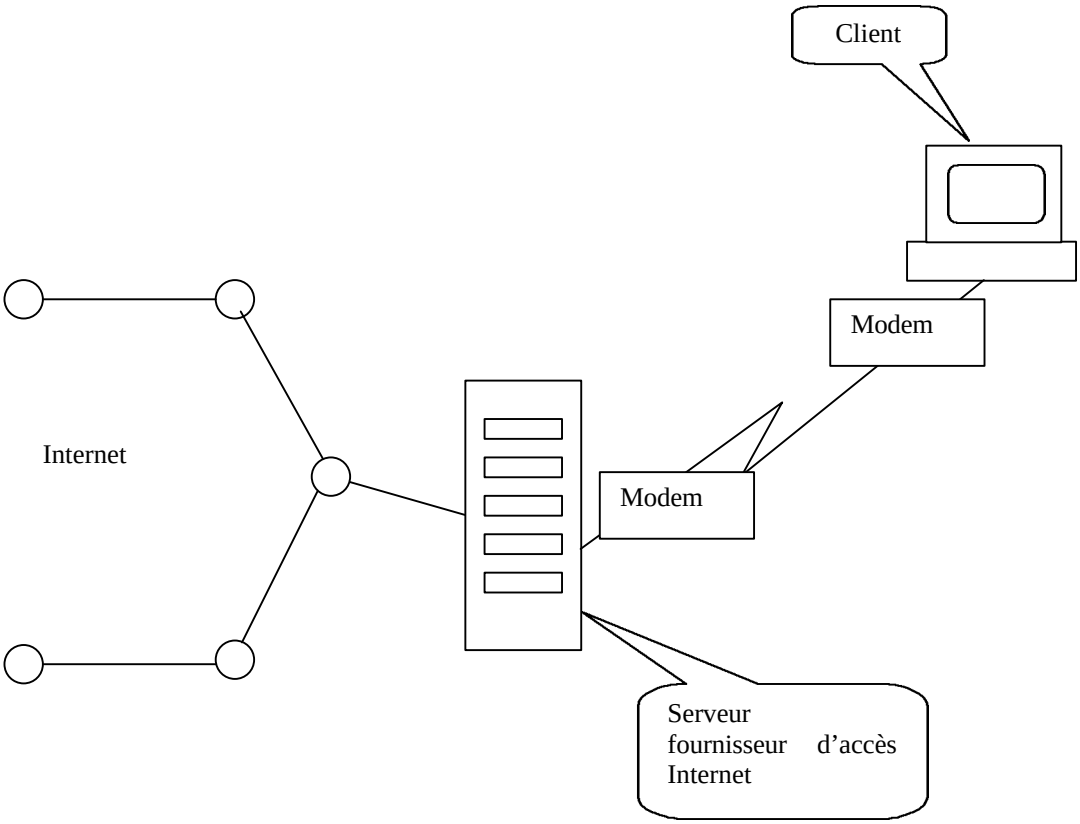
interroger le serveur racine (c'est le serveur qui connaît les adresses du plus haut niveau).

L'analogie de l'annuaire téléphonique peut nous servir encore une fois, il est plus facile pour nous de retenir le nom d'une personne et de consulter l'annuaire pour lui téléphoner que de retenir directement son numéro de téléphone.

Et moi, et moi, et moi ...

Les utilisateurs (privés) d'Internet n'ont pas (besoin d') une connexion permanente au réseau et ne possèdent pas non plus une adresse IP permanente.

Pour accéder à Internet, ils utilisent les services d'un fournisseur d'accès



(ISP, Internet Services Provider).

Le fournisseur d'accès possède des ordinateurs qui font partie d'Internet et auxquels les abonnés peuvent se connecter via une ligne téléphonique.

Voici le principe d'une connexion. L'ordinateur-client lance un logiciel d'accès réseau à distance et, grâce à son modem, se connecte au serveur du fournisseur d'accès. Une fois la connexion établie, le serveur du fournisseur d'accès attribue (provisoirement) à l'ordinateur-client une adresse IP et ce dernier fait donc partie (provisoirement) du réseau Internet.

A partir de ce moment, l'ordinateur-client peut accéder à n'importe quel ordinateur-serveur qui est connecté à Internet, quelle que soit le localisation géographique de celui-ci. Pour cela, l'ordinateur-client devra exécuter un logiciel-client approprié, par exemples un logiciel de courrier électronique, un logiciel de navigation, …

A quoi sert Internet ?

Du point de vue pratique, l'Internet, qui interconnecte des millions d'ordinateurs, est un formidable moyen de communication à travers le monde.

Grâce à cette possibilité de communication, Internet est donc un moyen d'accès à une masse indescriptible d'informations, un outil de collaboration permettant l'apprentissage et aussi le support des compétences de personnes à travers le monde.

Concrètement, nous pouvons considérer qu'Internet est un outil capable de nous rendre un certain nombre de services.

Voici quelques uns des services auxquels l'utilisateur peut accéder :

- **Echanger du courrier et des documents :**

  E-mail (electronic mail ou courrier électronique) : Il permet d'échanger (quasi instantanément) du courrier (et des documents) avec toute personne possédant une adresse électronique. Protocole utilisé : Simple Mail Transfer Protocol (SMTP).

- **Transférer des fichiers d'une machine à une autre :**

  Transfert de fichiers : Ce service, comme son nom l'indique, permet de transférer directement des fichiers d'une machine à une autre. Protocole utilisé : File Transfer Protocol (FTP)

- **Participer à des groupes de discussion :**

  News (Group News ou forum de discussion) : Il s'agit d'un immense ensemble de forum. Les débats s'organisent sous forme de questions et de réponses animées par les abonnés à ces forums. Protocole utilisé : News Network Transfer Protocol (NNTP).

- **Accéder à des pages hypertexte et hypermédia :**

  World Wide Web : Ce service permet d'accéder à des pages, appelée page Web. Une page Web est écrite en langage HTML (Hyper Text Mark-up Language) et peut contenir du texte, des images (statiques ou animées), des séquences vidéos, du son et des (hyper)liens. Ces liens

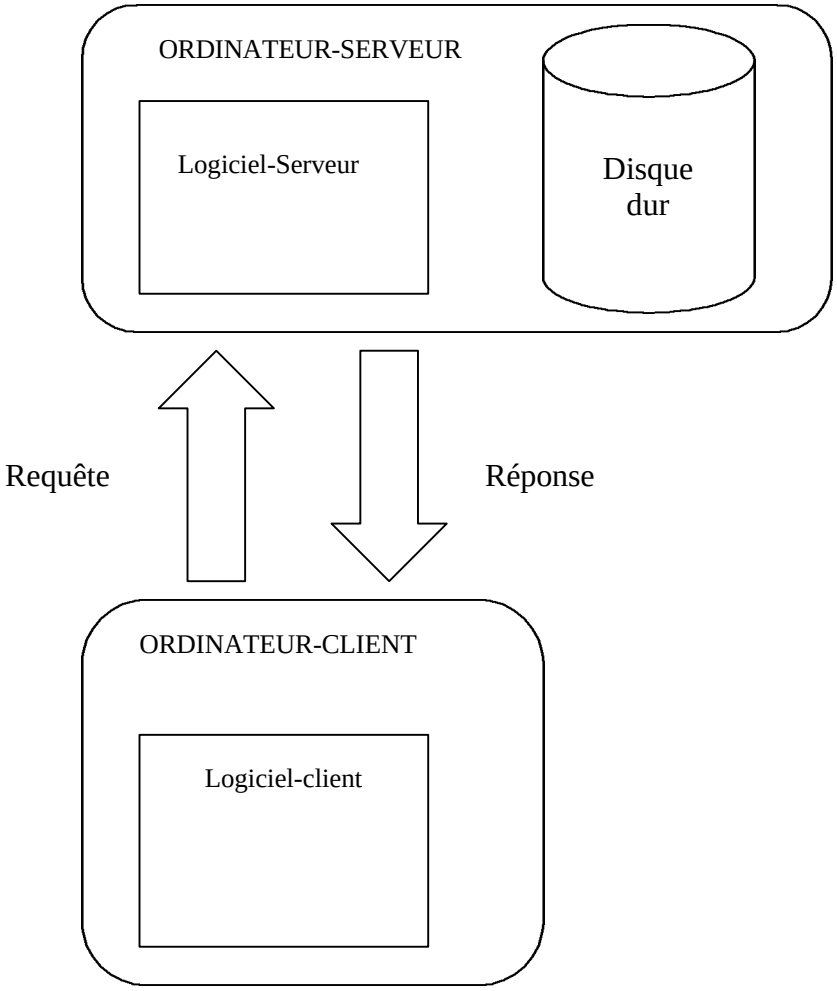
HTML est un langage de description de texte.

HTTP est un protocole de transfert de documents hypertextes.

permettent de passer d'une page Web à une autre (située éventuellement à l'autre bout du monde). Protocole utilisé : Hypertext Transfer Protocol (HTTP).

Le système fonctionne selon un modèle client-serveur. Le logiciel-client émet des requêtes (demandes) et le logiciel-serveur répond aux requêtes. La requête est formulée d'une manière standardisée (uniforme) au moyen des URL (Uniform Resource Locator).

L'utilisation de chacun de ces services repose sur un logiciel-serveur et un logiciel-client approprié.



Qu'est-ce que le World Wide Web

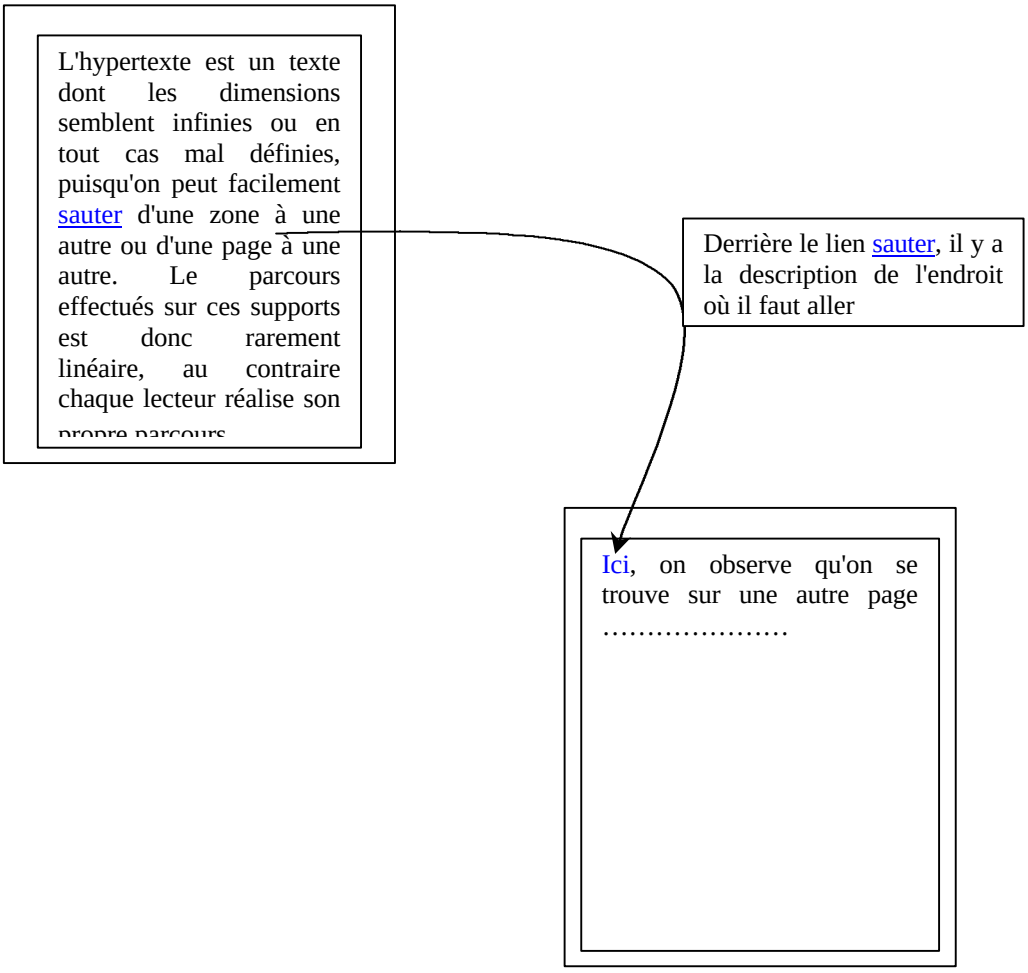
World Wide Web : traduction littérale Toile Mondiale.

Le World Wide Web (WWW) est un système d'information ouvert, conçu spécifiquement pour simplifier l'utilisation et l'échange de documents. Les documents publiés reposent sur le principe de l'hypertexte.

Liens : les liens se présentent sous forme de texte ou d'image.

L'hypertexte est une manière d'organiser et de présenter l'information dans laquelle certains éléments du texte, appelés liens, permettent de se déplacer vers d'autre zone du texte ou vers une autre page.

L'hypertexte est un texte dont les dimensions semblent infinies ou en tout cas mal définies, puisqu'on peut facilement sauter d'une zone à une autre ou d'une page à une autre. Le parcours effectués sur ces supports est donc rarement linéaire, au contraire chaque lecteur réalise son propre parcours.



Les CD ROM éducatifs et les encyclopédies électroniques sont deux exemples classiques de support utilisant le principe de l'hypertexte.

L'hypertexte n'est pas l'apanage du Web. En effet, celui-ci est utilisé à profusion sur les CD ROM et dans les aides logicielles en ligne.

Le Web donne accès à bon nombre de services Internet et ceci au travers d'une interface facile à utiliser et uniforme pour les différents services

En outre, le Web permet de publier facilement ses propres données et/ou informations. Il suffit d'installer un serveur Web et un éditeur de liens hypertexte.

Les pages HTML sont en réalité de fichiers de texte qui contiennent la description de la page à visualiser. Les images, les liens, les sons et la plupart des informations ne sont pas contenus dans ce descripteur de pages, mais bien dans des fichiers séparés. Il est donc possible de disposer d'outils qui génèrent dynamiquement le contenu de ces pages avant que le serveur

HTTP ne les distribuent vers les postes clients. De même, le mode de dialogue entre serveur et client peut intégrer des composants dynamiques, qui ne font pas partie du langage de description de page HTML. Ces parties sont présentées sous la forme de programmes et sont appelées (selon la structure et les langages utilisés) Javascript, applets Java, Active Server Page ou encore JavaBeans.

Chapitre II
Naviguer sur le WEB

Navigateurs, explorateurs,...

Quelques adresses, telles que vous pouvez les trouver dans des articles, sur des publicités,…

http://www.rtbf.be
http://www.wallonie.org
http://altavista.digital.com

Le chapitre précédent évoque le différents services qu'Internet peut intégrer. Parmi ceux-ci, le courrier électronique, le transfert de fichiers, les groupes de discussion, la vidéo-conférence,… Mais le service le plus connu et le plus apprécié actuellement est sans conteste la navigation sur le Web. C'est essentiellement par ce biais que le grand public découvre Internet. C'est aussi pourquoi le présent ouvrage y est entièrement consacré. Les journaux, les revues, les magazines sont aujourd'hui truffés d'adresses au format très particulier qui sont pour les lecteurs autant de points d'attaque de l'océan de données que constitue Internet.

Notre propos sera donc d'examiner ce que signifie concrètement "naviguer sur le WEB", mais particulièrement d'un point de vue technologique. En effet, malgré le niveau actuel de développement du traitement de l'information, il est manifeste que les limites des systèmes sont toujours aussi présentes. Nous pensons qu'il est important de vous les faire percevoir et nous espérons qu'une meilleure connaissance du fonctionnement de ces systèmes vous conférera davantage d'autonomie lors de l'utilisation de ces services.

Globalement, nous pouvons affirmer que pour tirer profit des ressources disponibles via Internet, un ordinateur connecté doit posséder dans sa mémoire un programme permettant d'y **envoyer des requêtes à destination d'autres ordinateurs serveurs** (cfr chapitre 1), puis d'**afficher de manière correcte les documents envoyés par ces derniers**.

Navigator de la firme Netcape, et Internet Explorer de la firme Microsoft sont les plus connus. Ces deux concepteurs se livrent d'ailleurs à une concurrence commerciale hors du commun.

Les programmes créés à cette intention sont destinés à rendre plusieurs services à la fois. Mais comme le plus souvent, ils sont utilisés pour rechercher et ramener de l'information, ils portent des noms évocateurs tels: navigateurs, explorateurs, fouineurs,... En ce qui nous concerne, nous emploierons uniquement le mot "navigateur" pour désigner ce type de programme. Nous préférons d'ailleurs utiliser le verbe "naviguer" plutôt que le verbe "surfer" pour suggérer la nécessité d'un contrôle de la manœuvre. Malgré tout, cette appellation risque d'être restrictive dans la mesure de l'extension des capacités actuelles et futures des logiciels de ce type.

Les documents au format HTML

La description qui suit peut paraître un peu technique. Notez que les détails de syntaxe sont peu importants. Toutefois, les exemples donnés sont destinés à illustrer le principe selon lequel les hypertextes sont exploités sur Internet. Il ne s'agit donc pas ici d'apprendre à connaître un langage (HTML), mais plutôt d'acquérir une compréhension plus profonde et plus éclairante du fonctionnement du système.

Le discours courant personnalise souvent les ordinateurs. Lorsque c'est le cas, il faut toujours y voir un ordinateur géré par un ou plusieurs programmes.

Les requêtes les plus classiques, dans l'esprit de ceux qui considèrent Internet comme une source inépuisable de données, sont donc des demandes de fourniture de documents. On peut comprendre qu'un document sans forme particulière soit aisément transmissible. La numérisation des caractères ne pose en effet pas trop de problèmes. Mais qu'en est-il de la mise en page? L'idée qui a prévalu est que le programme navigateur peut localement réaliser lui-même cette mise en page sur base de consignes, pourvu que celles-ci soient formalisables et normalisées. Ainsi, dans une suite de caractères, certains groupes peuvent être interprétés non comme du texte, mais comme des consignes de mise en forme et en page.

> Lorsque le navigateur reçoit le texte suivant:
>
> PRINCIPES DE <B>NAVIGATION</B>
>
> il interprète les groupes de symboles <B> et </B> comme des consignes de "mise en gras" (B pour bold) des caractères qu'elles encadrent.
>
> Le texte qu'il affichera ressemblera donc à ce qui suit:
>
> PRINCIPES DE **NAVIGATION**
>
> S'il reçoit le texte:
>
> <H1>Les moteurs de recherche</H1>
>
> il affichera le texte en utilisant un style correspondant à un titre de premier niveau (heading 1) du genre de ce qui suit.
>
> # Les moteurs de recherche

Une des conséquences d'un tel principe, c'est qu'il est possible et même certain que tous les navigateurs n'affichent pas toujours tous les documents exactement de la même manière.

Ces groupes de caractères sont appelés **balises** (en anglais: *tags*). Ils ne servent d'ailleurs pas qu'à la mise en page, puisqu'ils permettent aussi la définition des liens dans les hypertextes (cfr chapitre 1).

> La chaîne de caractères qui suit contient deux balises, dont une assez longue.
>
> **<A HREF="http://agora.unige.ch/ctie/">**CTIE**</A>**
>
> Lorsque le navigateur rencontre dans un document de telles balises, il affiche la chaîne de caractères CTIE dans un style particulier (souvent

souligné et dans une autre couleur) et il mémorise que cette chaîne est liée à un autre document. En d'autres termes, si le lecteur du document clique sur ce lien, le navigateur devra faire le nécessaire pour lui afficher le document en question.

Le langage HTML a été mis au point dans les laboratoires du CERN à Genève par Tim Berners-Lee.

Il existe un nombre important de possibilités de mise en page. L'ensemble des balises interprétables par les navigateurs en termes de commandes de mise en page, de création de liens et autres opportunités est appelé **langage HTML** (*HyperText Markup Language*).

Il est possible de modifier un document au format HTML avec le Bloc-notes de Windows, même s'il existe d'autres outils plus pratiques.

Un document est dit au format HTML, s'il contient du texte dont une partie constitue des balises du langage. En conséquence, un tel document peut être affiché, voire modifié et imprimé par un programme de traitement de texte, ou un simple éditeur de texte. Toutefois, il est difficile à lire comme tel, puisque les balises sont mélangées au texte normal.

Un document au format HTML mis à disposition sur un ordinateur connecté à Internet grâce à un programme serveur Web est appelé couramment **page Web**. Contrairement à ce qui se passe dans le domaine de l'édition sur papier, la taille d'une *page Web* est très variable. Son affichage à l'écran peut nécessiter un défilement.

Malgré l'existence de plusieurs versions du langage, les commandes essentielles sont standardisées. Cela signifie que si le rédacteur d'un document au format HTML se limite à l'utilisation de commandes reconnues par de nombreuses versions, les navigateurs pourront afficher le document sans trop de problèmes. En revanche, l'utilisation de commandes très récentes risque de poser des problèmes aux "vieux" navigateurs.

En résumé, nous pouvons dire que le travail le plus attendu des navigateurs consiste à demander à des ordinateurs connectés à Internet, des copies de documents au format HTML, puis à les afficher dans une mise en page correcte.

Nous allons maintenant examiner d'un peu plus près comment peut s'effectuer cette demande.

Principes de navigation

Il existe des balises de toutes sortes associées à des styles de mise en forme et des commandes de mise en page des documents à l'écran. Si certains styles correspondent à des niveaux de titres, d'autres sont liés au statut particulier de certaines informations (citations, exemples,…). Quant aux commandes de mise en page, elles concernent notamment les énumérations avec des possibilités de numérotation et la disposition des informations en tableaux.

Ce n'est pas (pour l'instant) votre problème de créer des documents au format HTML, mais ces explications vous permettent de comprendre comment les choses se passent, et aussi pourquoi certains problèmes d'affichage peuvent surgir.

Etablir la connexion

Pour faire fonctionner un programme de navigation présent en mémoire, il faut que vous lui fournissiez une requête. En d'autres termes, et dans les cas les plus fréquents, vous lui donnez l'adresse d'un ordinateur serveur et (éventuellement) le nom d'un document se trouvant sur l'un des supports d'information qu'il gère. De la sorte, le navigateur envoie cette requête sur Internet et attend de recevoir le document en question pour l'afficher.

Une requête envoyée sur Internet n'a de chance d'être satisfaite que si votre ordinateur est connecté. Il faut distinguer plusieurs situations:

- votre ordinateur n'est pas connecté à un réseau local et l'accès à Internet est un accès à distance (via modem, carte RNIS,…)
- votre ordinateur fait partie d'un réseau local dont un serveur possède un accès à Internet

Le programme peut très bien mémoriser les trois informations, mais en ce qui concerne la dernière, c'est peu recommandé, sinon n'importe quel utilisateur de votre ordinateur aura accès à Internet sans formalités.

Dans le premier cas, un programme d'accès au réseau à distance est nécessaire. Habituellement, vous devrez lancer ce programme et fournir au moins trois informations: le numéro de téléphone du fournisseur d'accès à Internet, votre nom d'utilisateur et votre mot de passe. Ces informations vous sont fournies par le fournisseur lui-même au moment de vous abonner à ce service. A terme, il n'est pas impensable que vous puissiez paramétrer le programme navigateur pour qu'il établisse lui-même la connexion en cas de besoin.

Dans le second cas, il faut encore distinguer plusieurs situations:

- l'accès du serveur à Internet est permanent
- l'accès du serveur à Internet est un accès à distance

Dans la première situation, la connexion de votre ordinateur au réseau local suffit. Evidemment, si au démarrage votre ordinateur ne s'est pas connecté au réseau local, la connexion à Internet est du domaine de l'impossible.

Dans la seconde, la connexion doit être établie au niveau du serveur, soit manuellement par la personne qui en a la charge, soit automatiquement, sur sollicitation d'un des ordinateurs clients du réseau local.

Notez qu'il existe encore d'autres solutions intermédiaires, notamment celles qui mettent en jeu des systèmes de routage. Il est donc difficile d'être complet.

Vous le voyez, les raisons d'une absence d'accès peuvent être nombreuses et il convient que dans chaque situation, vous soyez informés de la manière dont la connexion s'établit en fonction de la configuration particulière liée au système utilisé.

Le port d'attache

Souvent, lors de son chargement, le navigateur envoie automatiquement une requête, ce qui se traduit par l'affichage d'un document par défaut. Généralement, il s'agit d'un document d'accueil hébergé sur le disque dur d'un ordinateur appartenant à votre fournisseur d'accès à Internet. Si vous avez omis d'établir la connexion avec votre fournisseur d'accès avant de lancer le programme de navigation, celui-ci risque d'afficher un message

d'erreur correspondant à l'impossibilité de satisfaire cette première requête (qui est une requête par défaut).

Si le navigateur est configuré de manière à ce qu'il n'existe pas de document par défaut, la zone de travail du programme est vierge et il s'agira, pour commencer à naviguer, de lui fournir une adresse cohérente.

Si le document par défaut est local, ce qui est possible et que nous expliquons plus loin, la connexion ne sera pas nécessaire.

Sites WEB

Nous avons déjà parlé de page WEB, mais on parle aussi souvent de **site Web**. Sans entrer dans trop de détails, nous admettrons qu'un site WEB est un ensemble d'informations, accessibles sur un ordinateur faisant tourner un programme « serveur WEB », organisé ou mieux encore, tissé de liens dont certains peuvent d'ailleurs donner accès à des sites extérieurs. Les types de liens seront examinés dans la suite. La notion de site est aussi liée à l'existence d'un auteur (au sens large) en ce qui concerne la collecte et l'organisation de ces informations.

Comment accéder à un site, ou à un endroit particulier d'un site? La question pourrait devenir: comment rédiger correctement une adresse? En effet, ce qu'il y a lieu de préciser lors d'une requête, c'est à quel ordinateur elle s'adresse et éventuellement, quel est le document demandé.

Adresses distantes

http est une abréviation qui signifie **HyperText Transfer Protocol**.

Le transfert des documents ne peut se faire sans respecter un certain nombre de règles. On parle de **protocole**. Ce sont d'autres programmes qui prennent le relais et donc, prennent en charge ce transfert. Ceci dit, le navigateur a besoin de savoir à quel type de ressource on lui demande d'accéder. Le transfert de documents HTML n'est en effet pas le seul type de service qu'il peut offrir. Pour demander ce service, la chaîne de caractères que constitue l'adresse doit commencer par **http://**. Pour un autre service, il faudra utiliser d'autres caractères. C'est l'annonce au navigateur que le document à afficher doit être transféré selon une technique propre à Internet.

Par défaut, les navigateurs récents utilisent http comme protocole par défaut. Il est donc facultatif d'écrire *http://* dans l'adresse.

Cette chaîne doit être suivie d'une autre chaîne précisant le nom de cet ordinateur (ce qui permettra de le localiser) et, éventuellement, le nom et la localisation du document sur les supports d'information gérés par cet ordinateur. Il faut noter que si ce nom est absent, le document renvoyé sera un document par défaut, souvent un document d'accueil.

La structure la plus complète d'une adresse ressemble donc à ce qui suit:

protocole://ordinateur/chemin/document

> Si vous fournissez comme adresse:
>
> http:// www.det.fundp.ac.be /cefis/index.html
>
> vous annoncez au navigateur que le travail à réaliser est la récupération, dans la mémoire centrale de l'ordinateur local, en utilisant le protocole *http* propre à Internet, d'une copie d'un document au format HTML appelé *index.html* et se trouvant sur un disque dur géré par un

ordinateur identifié www.det.fundp.ac.be dans un dossier reconnu par cet ordinateur sous le nom *cefis*.

Comment un ordinateur peut-il être identifié, à partir d'un nom se composant de chaînes de caractères séparées par des points? C'est essentiellement grâce à un système de gestion de domaines qui évite les confusions. Un domaine est identifié par quelques caractères et possède généralement plusieurs sous-domaines.

Il y a certainement des sites hébergés en Belgique par des ordinateurs dont le nom ne se termine pas par *be*.

Il existe quelque part sur Internet un ordinateur équipé d'un programme serveur de domaine qui gère tous les sites voulant se faire identifier comme des sites belges (*be*). La personne responsable de l'exécution de se programme gère les noms des sous-domaines reconnus du domaine *be*. Elle autorise ou non la création de nouveaux sous-domaines. Ceux-ci sont extrêmement nombreux. Presque tous les sites ont leur sous-domaine propre: *rtbf.be*, *wallonie.be*, *segec.be*,... Il en existe quelques-uns qui ont d'autres sous-domaines comme: *ac*, *cec*, *nato*,… Il existe quelque part sur Internet un autre ordinateur équipé d'un programme serveur de domaine qui gère tous les sites voulant se faire identifier comme des sites académiques (*ac*). La personne responsable de l'exécution de se programme gère les noms des sous-domaines reconnus du domaine *ac*. Elle autorise ou non la création de nouveaux sous-domaines. Les sous-domaines de ac sont par exemple: *fundp*, *ulg*, *ucl*, *umh*, *ulb*, *vub*, *kuleuven*, *rug*…

be fait référence à des sites (ordinateurs) situés en **Be**lgique.

ac fait référence à des sites (ordinateurs) situés dans les universités (**ac**adémique)

fundp fait référence à des sites (ordinateurs) situés aux **F**acultés **U**niversitaires **N**otre-**D**ame de la **P**aix à Namur

det fait référence à des sites (ordinateurs) situés au **D**épartement **E**ducation et **T**echnologie

www c'est d'abord l'abréviation de *World Wide Web*. Cela signifie littéralement: toile d'araignée (*web*) grande (*wide*) comme le monde (*world*). C'est l'image qu'on utilise pour désigner ce réseau de réseaux d'ordinateurs. Mais plus concrètement, cela ne désigne pas un domaine comme les chaînes de caractères qui suivent, mais le nom par défaut de l'ordinateur capable de servir des documents. Cela signifie qu'il est possible de s'adresser à un autre ordinateur de ce domaine, pour autant que celui-ci soit capable de servir des documents. Cette capacité à l'existence d'un programme, appelé *serveur Web*, exécuté sur l'ordinateur concerné.

Une variante:

http://www.det.fundp.ac.be/~eva/

Cette adresse fait référence à un dossier particulier du disque dur du même ordinateur serveur. Dans ce cas, c'est le document par défaut situé dans ce dossier qui sera chargé. Dans le cas présent, ce document est identifié par le serveur Web comme devant s'appeler *index.html*. On obtiendrait donc exactement le même résultat en précisant le nom du document comme ci-dessous.

http://www.det.fundp.ac.be/~eva/index.html

Le navigateur acceptent certains paramètres par défaut de précision de ceux-ci. Ainsi, il est possible qu'il reconnaisse *http* comme protocole par défaut. De la sorte, l'adresse suivante suffirait:

www.det.fundp.ac.be/

De même, si votre ordinateur fait partie du même domaine que l'ordinateur serveur auquel vous vous adressez, le nom de cet ordinateur suffit.

www

Si vous voulez vous adresser à un autre ordinateur serveur du même domaine appelé *forum*, il vous suffit d'écrire comme adresse:

forum

Quoi qu'il en soit, une adresse complète garantit toujours un bon résultat, même si l'augmentation de la longueur de cette adresse va de paire avec l'augmentation du risque d'erreur. Les adresses sont en effet généralement peu compréhensibles et les fautes de frappe sont fréquentes.

L'adresse fournie au navigateur est en quelque sorte, un appel à une ressource. Il n'y a d'ailleurs pas que le chargement de documents HTML qui soit possible. Il existe d'autres ressources mais qui utiliseront une nomenclature assez semblable. C'est pourquoi on parle d'**URL** (*Uniform Ressource Location*).

Adresses locales

La chaîne de caractères *file://* (fichier) annonce au navigateur que la ressource se trouve sur l'ordinateur local. Il ne s'agit donc plus de naviguer sur le web, mais sur un disque dur de la machine locale. Cette chaîne doit être suivie d'une autre chaîne précisant au minimum le **nom de l'unité**:

file://c:

aura pour conséquence de faire générer par le navigateur un document au format HTML qui sera affiché en présentant la liste des documents et des dossiers présents à la racine du disque dur C.

Ce nom peut être éventuellement suivi du **nom du dossier**:

file://c:/fp/algo/

aura pour conséquence de faire générer par le navigateur un document au format HTML qui sera affiché en présentant la liste des documents et des dossiers présents dans le dossier *fp/algo* du disque dur D.

Vous pouvez aussi faire suivre le **nom d'un document**. Si ce document est au format HTML, le navigateur l'interprétera comme il le fait avec un document provenant d'un autre ordinateur:

file://c:/fp/algo/exemple.htm

aura pour conséquence de faire afficher par le navigateur le document au format HTML appelé *exemple.htm* et présent dans le dossier *fp/algo* du disque dur D.

Si le document est d'un autre format, le navigateur vous demandera ce que vous voulez en faire (généralement de l'ouvrir, c'est-à-dire de lancer l'application et ouvrir le document si c'est un fichier de données):

file://d:/fp/algo/tiroir.wpd

aura pour conséquence de faire afficher par le navigateur une boîte de dialogue. Si vous demandez d'ouvrir le document, le navigateur demande au système d'exploitation de lancer le programme de traitement de texte WordPerfect et d'ouvrir le document *tiroir.wpd* présent dans le dossier *fp/algo* du disque dur D.

Attention: tous les navigateurs ne réagissent pas de la même façon si toutes les règles de syntaxe ne sont pas respectées. A titre d'exemple, si on doit admettre que la chaîne de caractères *file://c:/* soit toujours bien interprétée, il n'en est pas forcément de même des chaînes *file://c:* ou *file:/c:* et même *file:c:* Certains navigateurs s'en contenteront, d'autres vont « caler ».

Retenez que le navigateur vous permet d'explorer le contenu des disques durs de votre ordinateur, mais aussi d'afficher les documents au format HTML qu'ils contiennent. Tout cela est possible, même si votre ordinateur n'est pas connecté.

Pages WEB

Un site est composé d'une ou plusieurs pages reliées entre elles de différentes manières, voire reliées à d'autres pages sur d'autres sites.

Une page WEB correspond à un document HTML présent sur le disque dur d'un ordinateur qui fait tourner un programme « serveur WEB ». Lorsque le navigateur affiche une page WEB, c'est qu'il a reçu le document HTML demandé et qu'il en a interprété le contenu en termes de mise en forme.

La longueur d'une page WEB est très variable. Elle n'a pas de dimensions fixes, si ce n'est qu'elle est affichée à l'intérieur de la fenêtre du navigateur. Quelquefois, son contenu intégral sera visible à l'écran, quelquefois il faudra user du défilement.

Une page WEB contient généralement des liens vers d'autres pages, bien que ce ne soit pas une obligation. On peut très bien créer une page WEB qui ne donne accès à aucune autre information que celles qu'elle contient. Mais ce n'est pas vraiment dans ce but que les navigateurs ont été développés. Ces liens sont de plusieurs types, mais ils ont tous en commun de « pointer » soit vers un endroit précis du document, soit vers une ressource extérieure au document (un autre document HTML présent sur le même site), soit vers une ressource extérieure au site en cours d'exploration (un document HTML d'un autre site)...

Liens hypertextes

Certaines parties du texte affiché sont donc sensibles au clic de la souris, celui-ci provoquant le déclenchement d'un lien et par-delà, l'affichage d'autres informations.

Ces parties de textes font évidemment l'objet d'un balisage en HTML, ce qui fait que le navigateur est capable de les afficher dans une couleur particulière. Cette couleur est un paramètre modifiable au niveau de la configuration du navigateur. Mieux, cette couleur peut changer lorsque le lien a déjà été activé. La durée du « souvenir » de cette activation est aussi un paramètre modifiable qui s'exprime un nombre de jours.

Pour être complet, il faut signaler que les couleurs des liens peuvent aussi être fournies par le document et que le navigateur peut être paramétré pour

accepter ou refuser ces couleurs « étrangères ». Quoi qu'il en soit, **une des qualités d'une bonne page WEB est de faire apparaître clairement les liens hypertextes à l'utilisateur**. Si ce n'est pas le cas, le navigateur modifie de toute façon la forme du pointeur (une main au lieu d'une flèche) lorsque la souris est déplacée sur un lien hypertexte.

Liens images

Il n'y a pas que du texte qui soit sensible au clic de la souris. Des images peuvent aussi l'être. Un clic sur une image peut donc provoquer un effet de navigation.

Le problème des images (comme du texte d'ailleurs) c'est d'avoir un pouvoir d'évocation suffisant pour que l'utilisateur ne soit trompé sur la marchandise.

Exemple:

Si le texte *Bruxelles* d'une page WEB est l'objet d'un lien, cela peut signifier qu'un clic de la souris sur ce texte va déclencher l'affichage de la page d'accueil d'un site sur Bruxelles (reste à savoir lequel...), l'apparition d'un plan de la ville de Bruxelles, ou même tout autre chose.

Une image peut être très suggestive, voire contenir un texte évocateur. Elle peut aussi ne rien suggérer du tout, sinon la curiosité, et par-delà, la déception.

La qualité d'une page WEB se mesure donc aussi à la puissance suggestive de ses liens, qu'ils soient de type hypertexte ou des images (dessins, photos, icônes,...)

Liens boutons

Votre navigateur est également capable d'afficher des boutons qui sont également des déclencheurs de liens. Un bouton joue un rôle assez semblable à celui des autres déclencheurs, à ceci près, qu'il n'y aurait pas beaucoup de sens à faire afficher des boutons qui seraient sans effet. Ce n'est pas le cas pour les textes et les images.

Naviguer: voyager de sites en sites et de pages en pages

C'est donc bien cela la navigation:

- passer d'un endroit à l'autre d'une page,
- passer d'une page à une autre,
- passer d'un site à un autre.

Tout cela est rendu possible par une technique de création de liens exploitée par les éditeurs de chacune de ces pages.

Il faut préciser que le navigateur est capable de bien d'autres choses telles l'exécution de programmes interactifs qui véhiculent des données (formulaires d'inscription, de prise de renseignements) voire qui génèrent eux-mêmes de nouvelles pages de WEB. Mais nous n'irons pas jusqu'à entrer dans des explications à ce niveau.

Rien ne vous empêche d'adresser une requête à votre propre ordinateur où à un ordinateur du réseau local. Il suffit de lui préciser la localisation du

document HTML que vous voulez voir affiché. En conséquence, **pour utiliser un programme de navigation, votre ordinateur ne doit pas nécessairement être connecté**. Il vous est tout à fait possible de naviguer sur un disque dur de votre propre ordinateur ou de votre propre réseau.

Exercices

Chargez le programme navigateur en mémoire. Notez l'adresse (URL) de la page d'accueil par défaut.

Chargez le document dont l'adresse est http://www.ciger.be . Sans cliquer, repérez le nombre de liens et leur type (texte, image).

Texte: Images:

Parmi ceux-ci, combien y en a-t-il qui sont susceptibles de fournir « des informations sur l'eau »? Lesquels?

Choisissez une des solutions. Les informations affichées proviennent-elles du même site?

Qu'est-ce qui vous permet de le dire?

Devinez-vous quatre moyens différents d'afficher le document précédent?

Testez-les tous les quatre. Créez un signet pour la page d'accueil du site du Ciger.

Connectez-vous au site du CeFIS: http://www.det.fundp.ac.be/cefis. Enregistrez un signet pour ce site. Recherchez-y des informations sur une fiche pédagogique traitant du système d'exploitation. A chaque étape, lorsque vous activez un lien, pouvez-vous préciser de quel type est la référence (autre document ou autre endroit du même document)?

Voyez-vous deux moyens d'obtenir une réponse à cette question? Lesquels?

Le document actuellement affiché est trop volumineux pour être affiché complètement à l'écran. Recherchez-y un lien vers le site des Facultés et activez-le. Enregistrez un nouveau signet. Observez à nouveau l'évolution de l'adresse.

Prenez des renseignements sur la Faculté de Droit. Quel est l'adresse de son site?

Un document traite de l'Association des Juristes Namurois. Faites-le afficher. A ce stade, de quels moyens disposez-vous pour faire afficher à nouveau la page d'accueil du site des Facultés?

S'il vous reste du temps (ou si vous vous ennuyez le soir), connectez-vous aux sites suivants et **naviguez à votre gré en faisant l'effort de vous poser toutes les questions que nous vous avons suggérées**: types de liens, types de références, choix du moyen pour revenir à un affichage déjà réalisé auparavant (boutons, historique, signets,...), observation de la zone d'adresse et de la barre de statut.

http://www.wallonie.org/

http://www.segec.be/

http://www.restode.cfwb.be/

http://www.cndp.fr/

http://www.profor.be/

http://forum.swarthmore.edu/

Chapitre III
La recherche sur le Web
Les moteurs de recherche

Introduction

Internet nous donne accès une masse considérable d'informations, des dizaines de millions de pages. Les documents publiés embrassent à peu près tous les sujets imaginables et sont diffusés aussi bien par des institutions que par des associations, des entreprises ou des individus. Ils sont de présentation et de qualité très inégales.

Ce système d'information et de communication est :

- distribué, l'information n'est pas centralisée, ni hiérarchisée,
- hétérogène, des ressources d'origine très différente coexistent,
- non certifié, l'information n'est pas toujours validée,
- instable, les sites d'information évoluent, naissent et disparaissent quotidiennement.

Les pages personnelles sont un bel exemple d'informations non validées. N'importe qui, ayant accès à Internet, peut publier n'importe quoi.

Attendez-vous donc à rencontrer le pire et le meilleur sur Internet. Pour y trouver de l'information intéressante et valable, vous devrez faire preuve d'ingéniosité dans vos recherches et de beaucoup de sens critique pour sélectionner les documents. Il est donc intéressant, non seulement de savoir comment rechercher de l'information, mais de savoir aussi comment la sélectionner.

Moteurs de recherche

Comment retrouver des documents dans cette masse (apparemment) inextricable ? Il existe sur Internet une série d'outils permettant de faire des recherches. Une manière classique consiste à utiliser un (des) moteur(s) de recherche proposé sur le Web. Dans son approche la plus simple, un moteur de recherche est une base de données qu'on peut interroger à distance. Plus précisément, un moteur de recherche est un serveur spécialisé dans la compilation et l'indexation des informations et qui, en outre, possède un

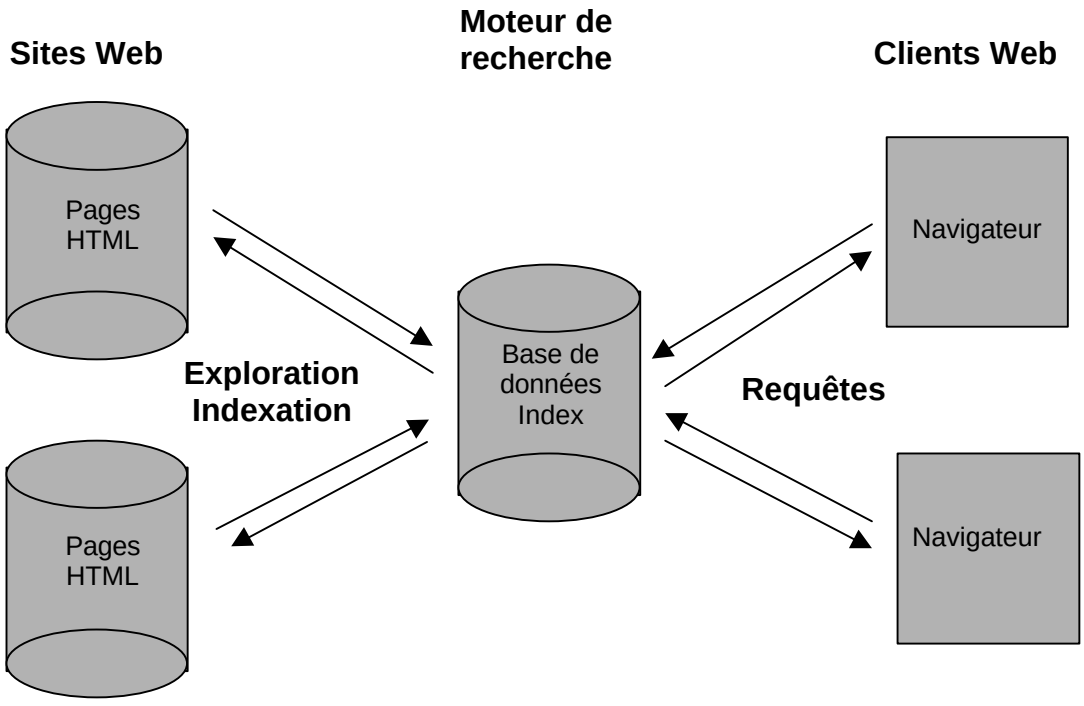
module pour répondre aux requêtes (demandes de recherche) des utilisateurs.

Les moteurs de recherche possèdent, en général, trois modules principaux.

- le module de mise à jour : son rôle est d'explorer l'Internet à la recherche de nouveaux sites ou de nouvelles informations sur des sites déjà répertoriés. Cette exploration peut se faire automatiquement par des programmes (robots), qui rapatrient des informations prédéterminées, ou manuellement par des personnes spécialisées. Dans ce deuxième cas, l'information recueillie est moins importante en quantité, mais elle est plus sûre.

Le module d'indexation est à la fois le point fort et le point faible des moteurs de recherche.

- le module d'indexation : c'est la phase de structuration et de classification de l'information rapatriée. En général, les informations répertoriées sont : l'adresse du site, le titre des pages, les mots clés, voire même, l'intégralité des pages. Les méthodes d'indexation varient très fort d'un moteur à l'autre et ceci influence donc les résultats que l'on peut obtenir lors d'une requête.
- le module de requête et de présentation des résultats : il s'agit d'une part de l'interface qui est présenté à l'utilisateur pour établir sa requête et visualiser les résultats et d'autre part d'un programme qui interprète et applique la requête de l'utilisateur.



Selon les moteurs, il y a actuellement deux approches de base :

Les répertoires validés.

Il s'agit de constituer des répertoires organisés par thème ou sujet avec une validation, une classification humaine des ressources et une indexation automatique sur certaines zones. YAHOO est un exemple de moteur de ce type.

Les index.

Il s'agit d'une collecte automatique de pages Web par un robot, suivie de l'indexation automatique des ressources. ALTAVISTA est un exemple de moteur de ce type.

Recherche d'informations

L'existence de deux types fondamentaux de moteurs de recherche nous amène assez naturellement à envisager deux formes de recherche d'informations : une **recherche par thème** basée sur l'exploitation des répertoires validés et une **recherche par mot clé** basée sur les index construits par les robots.

Recherche par thème

La page d'accueil des moteurs de recherche à répertoires validés présente, généralement, les principaux thèmes qui sont répertoriés par ce moteur de recherche. Chaque thème est subdivisé en sous-thèmes et ainsi de suite. La manœuvre de recherche consiste donc à naviguer de thème en sous-thème jusqu'à aboutir aux (éventuels) documents qui nous intéressent.

> Prenons, par exemple, comme objectif de recherche le thème *de l'impressionnisme en peinture*. Nous pourrions faire le parcours suivant : dans la page d'accueil, choisir *Art et culture*, dans la page Art et culture, choisir *Histoire de l'art*, dans la page Histoire de l'art, choisir *Périodes et mouvements* et dans cette page, choisir *Impressionnisme*. L'arboresecence se termine sur une série de documents qui concernent l'impressionnisme. Il nous reste à consulter ces documents pour analyser leur contenu et voir s'ils correspondent à ce qu'on recherche.

Arboresence d'une recherche par thème

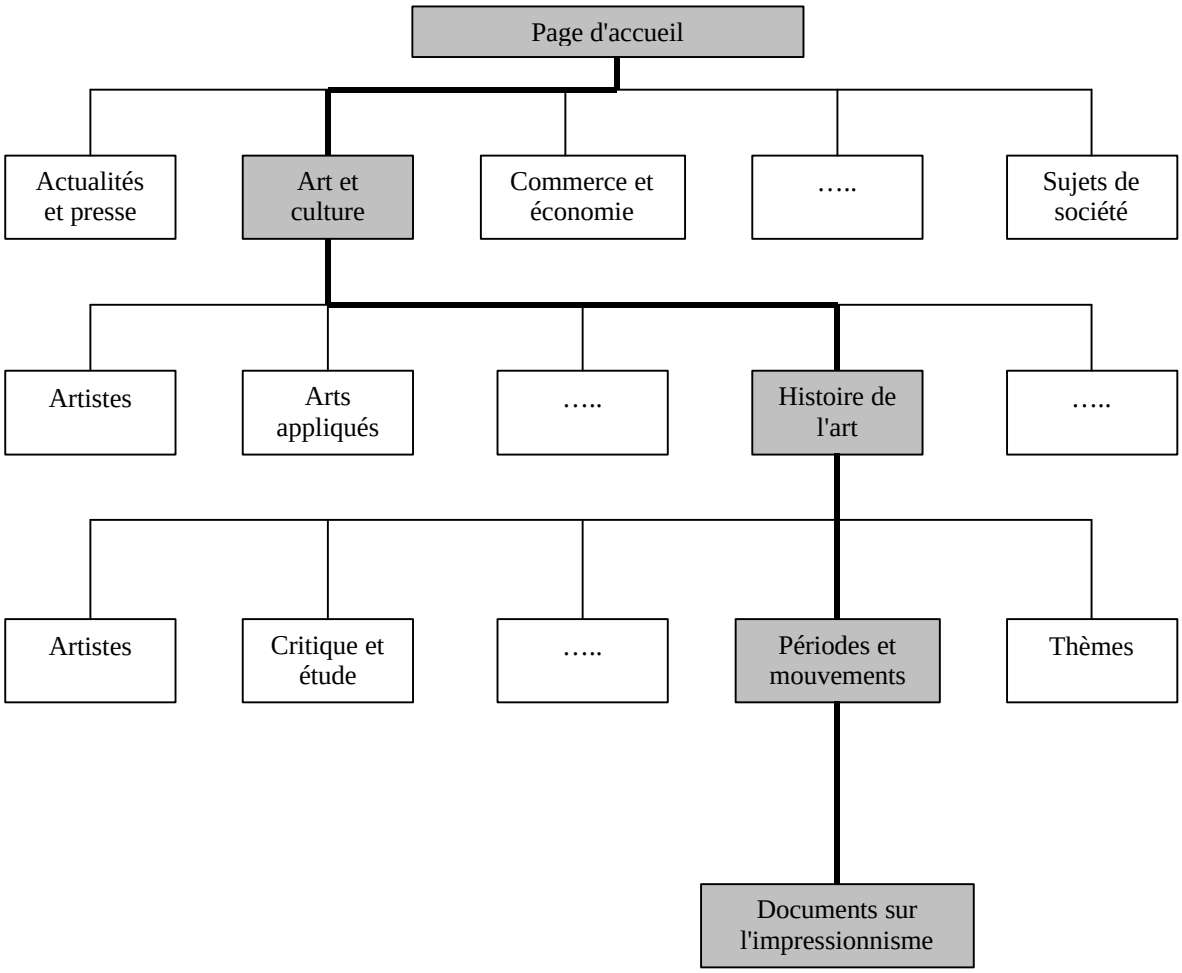


Illustration : rechercher ces documents sur YAHOO (http://www.yahoo.com).

Exercice :

Recherche par thème :

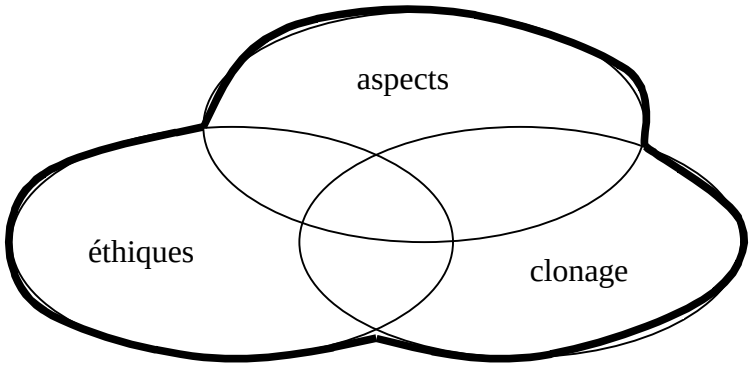
en utilisant le moteur de recherche YAHOO (http://www.yahoo.fr/), faites une recherche de documents sur le thème de "*la vache folle*".

Recherche par mot-clé

Fixons, d'abord, l'objectif de notre recherche. Nous désirons recueillir des informations sur le sujet suivant : *les aspects éthiques du phénomène de clonage*.

Tous les moteurs de recherche de type robot travaillent à partir de mot clé et donnent à l'utilisateur la possibilité de formuler sa requête avec un ou plusieurs mots clés.

La première idée qui pourrait venir à l'esprit est donc d'écrire *: aspects éthiques du clonage*. Cette recherche donne quelques dizaines de milliers de résultats. Pourquoi ? Parce que les moteurs de recherche sont conçus de telle sorte que lorsqu'on leur fournit plusieurs mots clés, ils recherchent les documents contenant un ou plusieurs de ces mots clés. En terme de logique et si on néglige les articles, la requête revient à rechercher les documents contenant *aspects* ou *éthiques* ou *clonage*.



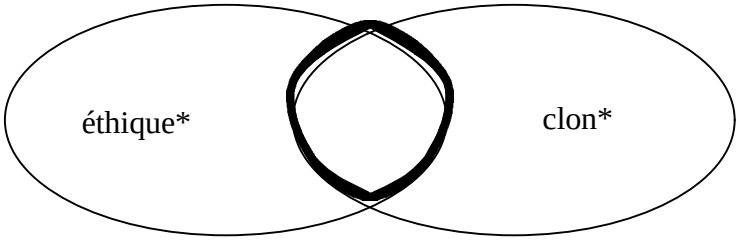
Requête : *aspects* ou *éthiques* ou *clonage*

Or notre idée première était, évidemment, de rechercher des documents contenant à la fois *aspects*, *éthiques* et *clonage*. Les moteurs de recherche incluent la possibilité de grouper des mots clés pour une recherche. Pour cela, il faut utiliser des marqueurs de début et de fin de groupe, par exemple les guillemets. La requête peut donc s'écrire "*aspects éthiques du clonage*". Le résultat risque d'être assez décevant : du style, aucun document trouvé. Pourquoi ? Tout simplement parce qu'il est très improbable qu'un document contienne la suite exacte "*aspects éthiques du clonage*" quelque part dans le titre ou au milieu du texte.

Le caractère * permet de désigner ces deux mots-clés en même temps, clon* représente tous les mots qui commencent par clon et se terminent par n'importe quel(s) caractère(s).

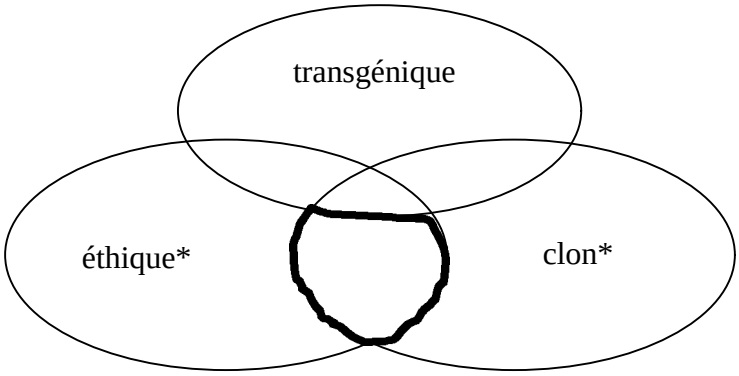
Pour résoudre notre problème, il vaut mieux considérer des mots clés uniques ou multiples. Prenons d'abord le cas des mots clés uniques, ceci veut dire prendre un mot clé **significatif** par rapport à notre recherche. Nous pouvons, par exemple, choisir le mot clé *clonage*, ce qui nous conduira à obtenir, probablement, quelques milliers de documents, c'est-à-dire beaucoup trop. Le résultat serait à peu près le même si nous avions choisi le mot clé : *clone*.

Pour restreindre notre recherche, nous allons être obligé de combiner plusieurs mots clés (mots clés multiples), mais en spécifiant qu'ils doivent tous être présents en même temps dans le document. Nous utiliserons donc une expression du type : *clon** et *éthique**. Cette requête nous donnera quelques centaines de documents.



Requête : *éthique** et *clon**

Si nous voulons encore affiner notre recherche, nous avons deux possibilités : premièrement, ajouter un ou deux mots clés avec le connecteur et ; deuxièmement, éliminer certains mots parasites ou gênants. Pour éliminer un mot, nous utiliserons le connecteur et-pas. Par exemple, si nous voulons éliminer le mot *transgénique*, nous écrirons : *clon** et *éthique** et-pas *transgénique*. Le résultat probable de cette recherche correspondra à environ une centaine de documents.



Requête : *éthique** et *clon** et-pas *transgénique*

Illustration : rechercher ces documents sur Altavista (http://altavista.digital.com/).

Tableau de synthèse sur la formulation d'une requête de recherche

Mot-clé		**Exemple**	**Altavista (recherche simple)**
unique		clonage	clonage
groupe		vache folle	"vache folle"
multiple	ou	clone ou clonage	clone clonage
	et	éthique et clonage	+éthique +clonage
	non (et_pas)	éthique et clonage et_pas religion	+éthique +clonage -religion

Remarque : la plupart des moteurs de recherche autorise l'utilisation d'un caractère spécial, généralement l'étoile (*) pour désigner n'importe quel caractère. Exemple, jardin* désigne aussi bien jardin que jardins ou jardiner ou jardinage ou jardinier.

Il existe aussi des fonctions spéciales qui permettent de limiter la recherche au titre, au texte, à l'URL ou à d'autres zones particulières.

Fonction spéciale	**Exemple**	**Altavista (recherche simple)**
titre	clonage dans le titre	title:clonage
texte	clonage dans le texte	text:clonage
URL	fundp dans l'URL	url:fundp

Pomper le Web!

Les pages de Web accessibles sur Internet représentent une quantité monstrueuse d'informations. Ceci ne signifie pas que toutes ces informations soient utiles et pertinentes.

Si on s'intéresse exclusivement aux techniques de capture de ces informations, on peut donner quelques lignes directrices intéressantes en ce qui concerne les textes et les images, ceci en vue de les intégrer dans d'autres documents, et essentiellement des documents réalisés avec un traitement de texte.

Du texte

La technique du copier-coller est bien connue. Elle s'applique aussi aux navigateurs. La sélection du texte se fait dans la fenêtre d'affichage. La commande coller ayant pour but d'envoyer la sélection dans le presse-papiers, seuls les caractères sont pris en considération. Les balises de mise en page sont négligées car inexploitables (dans l'état actuel des choses) par les programmes de traitement de texte.

Conséquences

- Les styles ne passent pas.
- Les effets de mise en page sont quelquefois convertis en espaces, ce qui demande un petit nettoyage.

Des images

Deux techniques sont à prendre en considération:

- le copier-coller
- l'enregistrement de l'image dans un format choisi

NB: cette dernière est aussi envisageable pour du texte, mais sans doute moins intéressante.

Copier-coller

Cette technique est la plus simple. Toutefois, sa conséquence est que l'image capturée, quel que soit son format initial, risque d'être intégrée dans le nouveau document dans un format qui ne permet plus beaucoup de transformations par la suite (format simplifié).

Enregistrer l'image sous…

La technique de l'enregistrement est plus intéressante en ce sens qu'elle permet à d'autres outils logiciels de les retravailler au besoin.

Les formats les plus courants pour des images du Web sont GIF (Graphic Interface Format) et JPEG (Joint Photographic Experts Group) qui présentent l'avantage d'une compression efficace. Mais le format BMP (bitmap) est aussi possible (bien que peu recommandé).

Il est possible d'enregistrer l'image au format référencé dans le document, mais aussi au format bitmap, ce qui permet de faciles transformations par la suite (avec un outil simple de dessin par exemple).

Dans un cas comme dans l'autre, les informations, une fois capturées, peuvent être collées ou insérées dans un document géré par un logiciel qui les accepte.

Exercices :

Recherche par mot-clé : en utilisant Altavista (http://altavista.digital.com/)

- Recherchez le nom d'une toile de Cézanne ou si vous préférez une reproduction d'une toile de Braque.
- Recherchez des renseignements, parmi lesquels une photo, sur le Vésuve.
- En vue d'une activité culturelle, recherchez des informations sur un musée de la bière que vous pourriez visiter avec vos élèves.
- Recherchez la liste des pays membres de la communauté européenne.
- Créez un document Word traitant des volcans dans lequel vous ajouterez à une petite production personnelle des petits bouts de texte et deux images de volcan récupérés sur le Web.

Chapitre IV
Le courrier électronique (E-mail)

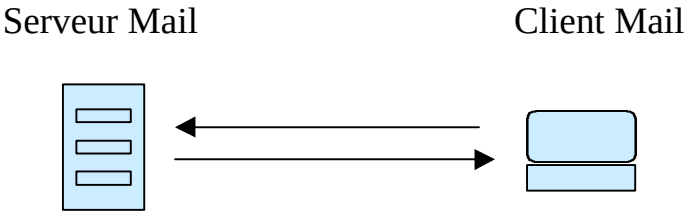
Introduction

Le courrier électronique est le service le plus utilisé sur l'Internet, après s'être répandu dans le monde professionnel, il est devenu l'un des outils favoris des particuliers. Quoi de plus normal, puisque ce service permet de communiquer en différé facilement et rapidement. De plus, il offre la possibilité d'envoyer et de recevoir des documents attachés au message, ce qui en fait un outil d'échange performant et intéressant. Ajoutons à cela qu'on peut tout aussi facilement échanger des messages entre 2 personnes qu'à l'intérieur de tout un groupe.

Nous allons nous efforcer, dans ce chapitre, de comprendre le fonctionnement et l'utilisation du courrier électronique. Nous verrons que ce service fonctionne suivant le modèle client-serveur déjà abordé par ailleurs. Il faudra découvrir les particularités du dialogue de ces 2 partenaires pour bien comprendre ce service. Une bonne compréhension de ce fonctionnement devrait permettre une meilleure maîtrise des logiciels-clients associés à ce service. Le but n'est donc pas de décortiquer les logiciels-clients existant sur le marché, mais bien de rechercher les principes de base et les invariants de ce service, qui sont applicables et utilisables quelque soit le logiciel-client.

Principes de base

Les applications de courrier électronique fonctionnent, elles aussi, selon le modèle client-serveur déjà abordé dans les chapitres précédents.



Les 2 partenaires sont donc bien un client d'un côté, représenté par un logiciel-client, et un serveur de l'autre, représenté par un logiciel-serveur. Ils sont reliés, comprenez plutôt connectés, par l'intermédiaire de l'Internet (ou d'un autre réseau, mais ne compliquons pas les choses). Ils peuvent donc dialoguer, bien sûr dans leur jargon (appelé plus académiquement protocole).

Mail signifie courrier en anglais. E-mail veut dire Electronic Mail, c'est-à-dire Courrier Electronique. L'expression anglaise s'est quasi imposée en français, sauf pour les puristes.

Rappelons-nous bien que le terme serveur-mail désigne la machine mais aussi, et surtout, le programme qui gère le service de courrier électronique et qui est installé (qui tourne) sur cette machine.

De même, le terme client-mail désigne la machine, mais aussi et surtout, le programme qui tourne sur la machine d'un utilisateur et qui permet à ce dernier de gérer son courrier électronique.

Pour accéder au service de courrier électronique, l'utilisateur doit, la plupart du temps (mais là aussi ne compliquons pas les choses), avoir un accès à l'Internet via un fournisseur.

Le terme courrier entrant désigne le courrier qui est envoyé à l'utilisateur. Le terme courrier sortant désigne le courrier que l'utilisateur envoie.

Le grand principe de fonctionnement du courrier électronique ressemble à celui d'un autre service que nous utilisons quotidiennement, la banque. Pour profiter des services offerts par une banque, il faut avoir un compte dans cette banque et pour réaliser ses opérations bancaires, du moins avec une carte de banque, il faut un code secret, qui protège le client.. De même, pour accéder au service de courrier électronique sur un serveur, l'utilisateur doit posséder un compte de courrier électronique sur ce serveur. Ce compte est assorti d'un mot de passe et donne à l'utilisateur la possibilité de recevoir et d'envoyer du courrier et met à sa disposition une boite aux lettres, pour le courrier entrant et le courrier sortant.

Remarque : Il faut bien distinguer ce compte de courrier électronique de l'éventuel compte d'utilisateur sur le serveur. Pour rappel ce compte permet à l'utilisateur de se connecter au serveur et via cette connexion d'accéder, par exemple, à l'Internet. De même, il ne faut pas confondre les serveurs d'accès et de courrier électronique. Il peut s'agir de 2 machines différentes. Ceci n'a d'ailleurs pas d'implications pour l'utilisateur, qui ne doit pas se soucier de savoir quel est son serveur d'accès, ni quel est son serveur de courrier électronique.

Le compte de courrier électronique aura la forme compte@serveur (par ex. : gvasters@hermes.fundp.ac.be mailto:gvasters@hermes.fundp.ac.be). Tout utilisateur, possédant un compte de courrier électronique de ce type se verra attribuer une adresse de courrier électronique (appelée aussi adresse E-mail) plus explicite du type alias@domaine (par exemple guy.vastersavendts@fundp.ac.be). C'est cette adresse qui apparaît dans le courrier électronique et qui doit être utilisée par une personne désirant envoyé un courrier à cet utilisateur. A ce compte est aussi associé un mot de passe, permettant à l'utilisateur d'avoir accès à son courrier entrant. C'est une manière de préserver la confidentialité du courrier.

Remarque : il est important de noter que l'adresse E-mail comporte un nom de domaine, alors que le compte de courrier électronique comporte un nom de serveur.

Une fois en possession de son compte de courrier électronique et après avoir installé, configuré et lancé son logiciel de courrier électronique, l'utilisateur pourra composer des messages et les envoyer, recevoir des messages, les lire et les classer. Ce sont là les principales fonctions (primitives) du logiciel-client de courrier électronique.

La distance qui sépare ces 2 serveurs n'a pas d'importance, puisqu'il communique entre eux via l'Internet.

Voyons maintenant comment s'effectue l'échange de courrier entre 2 personnes, possédant chacune un compte de courrier électronique. Supposons que ces 2 comptes se trouvent sur 2 serveurs différents, ce qui n'est pas obligatoire, mais rend l'explication plus claire.

L'utilisateur A va composer un message avec son logiciel de courrier électronique. Pour le moment cet utilisateur ne doit pas être connecté sur l'Internet, il compose simplement son message comme il taperait n'importe quel texte avec son logiciel de traitement de texte. Il doit aussi indiquer l'adresse E-mail de son correspondant, pour le message puisse lui parvenir.

Le serveur du fournisseur d'accès n'est pas nécessairement le même que le serveur de courrier électronique.

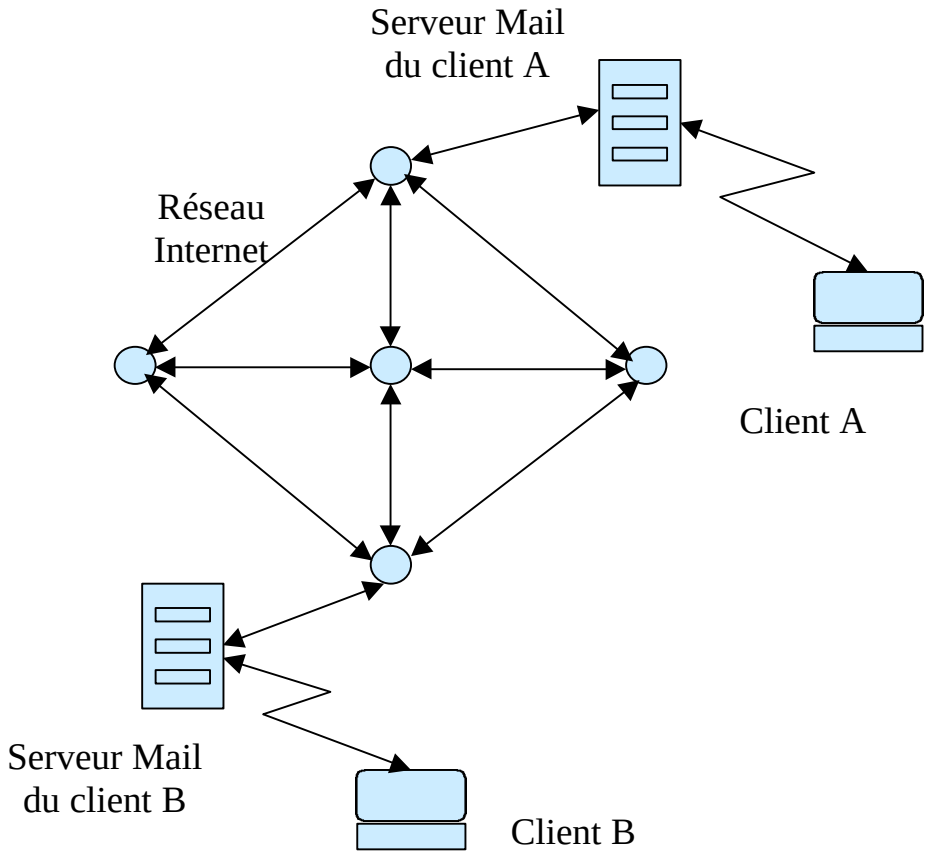
Maintenant que l'utilisateur a composé son message, il désire l'envoyer à son correspondant, pour cela il faut qu'il soit connecté sur le serveur de son fournisseur d'accès. Soit il possède une connexion permanente, auquel cas il n'a rien à faire, soit il ne possède pas cette connexion et il doit l'établir, en général, via le réseau téléphonique.

Une fois la connexion établie, l'utilisateur envoie son message qui aboutit dans sa boite aux lettres de courrier sortant sur son serveur de courrier. De là, ce message sera envoyé sur le serveur de courrier du destinataire (utilisateur B), plus précisément dans sa boite aux lettres de courrier entrant.

Ce message va rester sur le serveur jusqu'au moment où le destinataire (B) lance son logiciel-client de courrier électronique, se connecte et "demande" au serveur s'il y a du courrier dans sa boite de courrier entrant.

Si effectivement il y a du nouveau courrier, ce qui est le cas dans notre exemple, le serveur-mail du client B va envoyer une copie du (des) nouveau(x) message(s) au logiciel-client de B.

Selon le paramétrage fait par l'utilisateur B, le logiciel-client B enverra (ou n'enverra pas) un ordre d'effacement du message au logiciel-serveur. Ceci a une conséquence très importante : si l'ordre d'effacement est envoyé, seule la machine-client B possèdera une copie de ce message et celui-ci ne pourra plus être lu à partir d'une autre machine-client, même par le propriétaire du compte de courrier.



L'utilisateur peut aussi paramétrer son logiciel pour qu'une copie des messages reste sur la machine-serveur. Dans ce cas, l'utilisateur pourra consulter ses messages à partir d'autant de machines différentes qu'il le désire, à condition que les logiciels-clients de chaque machine soit configuré de cette manière et avec ses paramètres personnels.
Du fonctionnement global du service de courrier électronique, on retiendra une chose importante, il s'agit d'un service asynchrone, c'est-à-dire qui ne demande pas la présence simultanée des correspondants, un petit peu comme l'est le service postal par opposition au service téléphonique.

Configuration d'un logiciel de courrier électronique

"Paramètres de configuration" signifie toutes les informations nécessaires au logiciel pour fonctionner correctement et de manière adaptée à l'utilisateur.

La description du fonctionnement du service de courrier électronique, qui précède, met en évidence les différents paramètres de configuration d'un logiciel-client de courrier électronique. Une bonne compréhension de ce fonctionnement est nécessaire pour configurer correctement votre logiciel-client de courrier électronique, selon l'utilisation que vous voulez en faire. De la discussion précédente, il ressort que l'utilisateur devra fournir au logiciel-client les renseignements suivants :

Paramètres utilisateur

Ces paramètres concernent le compte que l'utilisateur possède.

POP signifie Post Office Protocol, il s'agit d'un protocole couramment utilisé pour le courrier entrant.

- le compte de courrier électronique, appelé aussi compte POP (ou POP account dans les versions anglaises), par exemple gvasters.
- l'adresse de courrier électronique (adresse E-mail), appelé aussi adresse de retour (ou return address dans les versions anglaises), par exemple guy.vastersavendts@fundp.ac.be.

Paramètres serveur

Ces renseignements concernent la machine-serveur sur laquelle l'utilisateur possède son compte de courrier électronique.

- l'adresse du serveur POP (protocole pour le courrier entrant), par exemple hermes.fundp.ac.be.

SMTP signifie Simple Mail Transfer Protocol, il s'agit d'un protocole couramment utilisé pour le courrier sortant.

- l'adresse du serveur SMTP (protocole pour le courrier sortant), par exemple hermes.fundp.ac.be.

Paramètres de connexion

Pour configurer complètement le logiciel de courrier électronique, il faudra encore préciser le type de connexion au serveur, soit par l'intermédiaire d'un réseau local, soit par l'intermédiaire d'une ligne téléphonique.

Remarque : le logiciel-client de connexion est un autre logiciel, il suffit de préciser ici quel est le type de connexion et non pas comment cette connexion doit s'établir.

Paramètres du courrier

Concernant les messages, le paramètre le plus important est celui de la conservation du courrier sur le serveur.

Il existe 2 possibilités :

- laisser une copie des messages sur le serveur, éventuellement pendant x jours.
- supprimer les messages sur le serveur quand ils sont copiés sur le client.

Rappelons la conséquence importante de ce choix, si l'utilisateur choisit la première possibilité, il pourra lire et relire son courrier à partir de n'importe quelle machine connectée à l'Internet et configurée avec ses paramètres personnels. Si, au contraire, l'utilisateur choisit la deuxième possibilité, ceci ne sera plus possible.

Composition d'un message

Parmi les primitives de base d'un logiciel-client de courrier électronique, on trouve assez logiquement une fonction qui permet de créer (composer) un nouveau message. Notons bien que durant cette phase, il n'est pas nécessaire d'être connecté puisqu'il n'y a pas d'échange entre le client et le serveur.

Chaque message comporte 4 parties bien distinctes jouant chacune un rôle bien défini :

- l'entête du message qui permet d'indiquer quel(s) est (sont) le(s) destinataire(s) et quel est le sujet du message.
- le corps du message qui contient le texte du message.
- la signature du message qui permet à l'auteur de se personnaliser.
- les objets attachés (joints) au message, par exemple des documents qu'on veut envoyer.

L'entête du message

L'entête comporte :

- une zone "A" ou "TO", dans laquelle on introduit la (les) adresse(s) du (des) destinataire(s), très précisément l'adresse E-mail, par exemple "guy.vastersavendts@fundp.ac.be" ou un alias qui désigne l'adresse de courrier électronique d'une ou de plusieurs personnes. (cfr. Carnet d'adresses)
- une zone "CC" (copie carbone ou carbon copy) dans laquelle on peut introduire une ou plusieurs adresses de courrier électronique de personnes qui recevront aussi le message.
- une zone "CCI" (copie carbone invisible) ou "BCC" (blind carbon copy), dans laquelle on peut mettre une ou plusieurs adresses de courrier électronique qui recevront aussi le message, mais sans que les autres destinataires ne le sachent.
- une zone "OBJET" ou "SUBJECT", dans laquelle on peut écrire le sujet ou l'objet du message, ce qui permettra au destinataire de se faire une idée du contenu sans devoir le lire et aussi, par la suite, de classer ses messages par sujet.

Le corps du message

Il s'agit du texte écrit pour le destinataire, le message qu'on veut lui envoyer. Cette zone est "contrôlée" selon le logiciel de courrier électronique par un éditeur de texte, éventuellement par un "mini-traitement" de texte. Le contenu et la présentation du texte est libre et n'est soumise à aucune règle spécifique, si ce n'est celle de la présentation, de la politesse, …

La signature du message

L'utilisateur peut personnaliser la signature de son message et fournir ainsi des renseignements personnels plus précis que sa simple adresse de courrier électronique. L'insertion de la signature peut se faire automatiquement ou manuellement.

Les objets attachés (joints) au message

L'utilisateur a la possibilité, très intéressante et très utile, de joindre des fichiers à son message. Les fichiers peuvent être de n'importe quelle nature, documents provenant d'un traitement de texte, d'un tableur, une image, un fichier compressé, un exécutable, …

L'utilisateur peut ainsi communiquer rapidement à son correspondant des documents. Bien sûr le destinataire doit posséder les logiciels nécessaires à la relecture de ces documents. Il est donc prudent de se renseigner sur les logiciels possédés par le destinataire.

Envoi et réception de courrier

Chaque logiciel-client de courrier électronique possède une primitive d'envoi et une primitive de rapatriement du courrier. Souvent ces 2 primitives sont combinées, l'envoi et le rapatriement se font en même temps.

Autant il n'était pas nécessaire d'être connecté pour rédiger son courrier, autant cette connexion est indispensable pour envoyer et "rapatrier" son courrier sur sa machine. En effet l'envoi et le rapatriement nécessitent un dialogue entre le logiciel-client de l'utilisateur et le logiciel-serveur. Avant d'utiliser les fonctions d'envoi et de rapatriement, il faut donc vérifier que la machine-client est bien connectée à l'Internet.

Rappelons une chose importante, l'envoi du courrier se fait sur la machine-serveur sur laquelle l'utilisateur possède un compte, de même le rapatriement des messages se fait à partir de la même machine-serveur et ces opérations n'ont rien à voir avec la machine-serveur de l'envoyeur (cas du courrier entrant) ou du destinataire(cas du courrier sortant).

Gestion de son courrier électronique

Boites de courrier

Tous les logiciels de courrier électronique possèdent une boite de réception, une boite d'envoi et une poubelle, comme un(e) secrétaire bien organisée. La boite de réception contient tous les messages qui sont entrés (qui ont été lus sur le serveur). Un signe distinctif permet de reconnaître les messages lus par l'utilisateur des messages non lus.

L'utilisateur peut créer des boites, en fait il s'agit de dossiers, dans lesquelles il classera son courrier en fonction des sujets ou de tout autre critère spécifique. Après la création des boites, la technique du glisser-coller est utilisée pour le classement.

Parmi tous les messages reçus, certains ne doivent pas être gardés, d'autres sont inintéressants, il faut donc pouvoir s'en débarrasser. La poubelle est là pour ça.

Carnet d'adresses

Faut-il que je retienne ou que je note (sur un bout de papier) les adresses électroniques de tous mes correspondants ? Question angoissante, qui, heureusement, a une réponse plutôt rassurante : le carnet d'adresses.

Les logiciels-clients de courrier électronique possèdent une fonction, appelée très souvent carnet d'adresses, qui permet de mémoriser les adresses E-mail des correspondants en les associant avec un nom clair, par exemple charles.duchateau@fundp.ac.be associé à Charles Duchâteau (Directeur CeFIS -– FUNDP). On peut alors utiliser ce nouvel alias pour envoyer un message.

Mieux encore ! On peut créer, dans le carnet d'adresses, un alias qui correspond à un groupe de personnes, c'est-à-dire lui associé plusieurs adresses électroniques. Par exemple, dans notre unité le CeFIS, chacun peut se créer un alias appelé CeFIS auquel il associe, dans le carnet d'adresses, toutes les adresses électroniques des personnes qui travaillent dans cette unité.

Répondre et rediriger

Les logiciels-clients de courrier électronique possèdent encore 2 autres facilités intéressantes : répondre à un message et rediriger un message. Le terme facilité est utilisé à dessein, car ces fonctionnalités peuvent être réalisées à partir des primitives déjà vues, essentiellement "envoyer du courrier" et des primitives de base de l'environnement, comme par exemple copier-coller.

Première situation : j'ai reçu un message auquel je voudrais répondre point par point, dois-je créer un nouveau message, copier-coller chaque point du message original, taper ma réponse et l'envoyer ?

Non, il suffit d'utiliser la fonction "répondre" qui créera automatiquement un nouveau message, placera l'adresse du destinataire et le texte du message original. Il suffira d'insérer les réponses ou commentaires à chaque point du message.

Deuxième situation : j'ai reçu un message qui ne me concernait pas directement ou je voudrais faire prendre connaissance de ce message à quelqu'un d'autre, dois-je créer un nouveau message, copier-coller le message, inscrire l'adresse E-mail du destinataire et l'envoyer ?

Non, il suffit de rediriger le message. L'appel à cette fonction créera automatiquement un nouveau message contenant déjà le texte. Il suffira d'ajouter l'adresse du destinataire ou mieux de la reprendre dans le carnet d'adresses. Le destinataire final saura que le message a été redirigé.

Ces 2 exemples sont assez éloquents sur la manière dont les logiciels fonctionnent. Il existe des primitives (fonctions de base) qui peuvent être regroupées d'une certaine manière pour obtenir une fonction (action) plus élaborée.

Une autre utilisation du courrier électronique : les listes de diffusion

Dans un des paragraphes précédents, nous avions abordé un fonction assez utile, le carnet d'adresses. Celui-ci permet, rappelons-le, d'associer à un alias une série d'adresses E-mail, c'est-à-dire de se créer une facilité pour travailler dans un groupe d'échange. Il subsiste cependant une lourdeur et une limitation à ce système : chaque membre du groupe doit encoder l'alias et les adresses associées, de plus les membres extérieurs à ce groupe ne savent pas qu'il existe.

Un certain nombre de serveurs offre un service appelé liste de diffusions, qui centralise des groupes de discussion, autour d'un thème défini. Les utilisateurs peuvent s'inscrire (et se déinscrire) à une liste, de manière plus ou moins automatique. Chaque membre de la liste de diffusion peut déposer un message, qui sera diffusé automatiquement à tous les autres membres du groupe.

La différence avec l'utilisation du carnet d'adresses d'un logiciel-client de courrier électronique, c'est que la gestion des inscriptions et de la diffusion des messages est gérée par le logiciel-serveur de liste de diffusion. La constitution du groupe peut évoluer dynamiquement sans intervention des membres du groupe.

Exercices

1. Configurez les logiciels-clients de courrier électronique, que vous trouverez sur votre machine, avec vos paramètres. N'oubliez pas d'activer l'option "laisser une copie sur le serveur".
2. Envoyez un message à
   - 1 personne du groupe
   - 3 personnes du groupe
     - 3 personnes dans le champ "A"
     - 1 personne dans le champ "A", 1 personne dans le champ "CC", 1 personne dans le champ "CCI"
   - toutes les personnes du groupe
3. Répondez à une personne du groupe qui vous a envoyé un message.
4. Faites une redirection d'un message reçu.
5. Créez un petit document Word et attachez-le à un message que vous envoyez à une personne du groupe.
6. Abonnez-vous à l'agent de recherche Informant http://informant.dartmouth.edu et fixez les paramètres de votre recherche.
7. Abonnez-vous à une liste de diffusion proposée par Majordomo@fundp.ac.be. Dans votre premier message, laissez la zone subject vide, tapez help suivi (à la ligne suivante) de end.