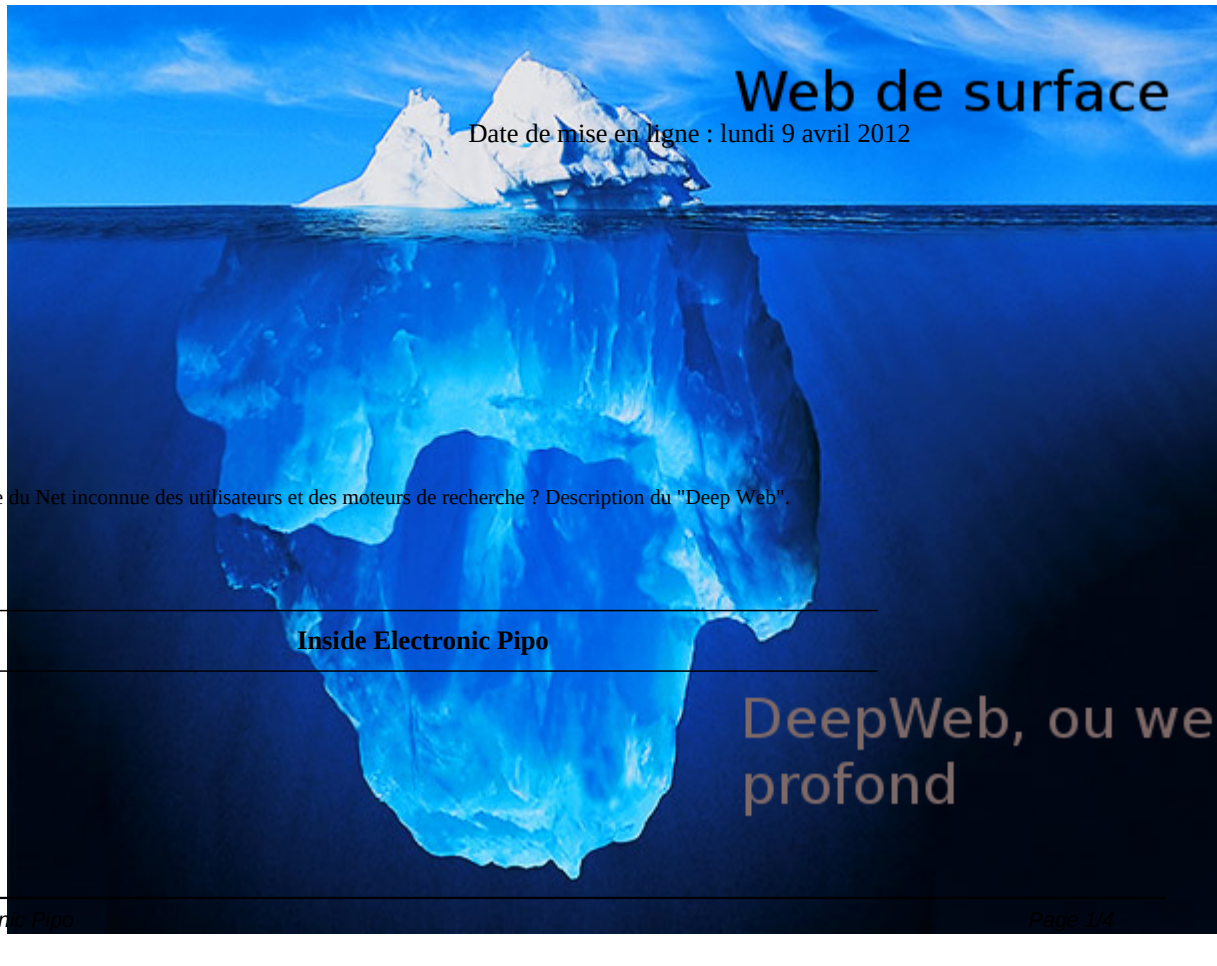
Dans les profondeurs du Web

- Espace culturel - Technologie -

Date de mise en ligne : lundi 9 avril 2012



Description :

Et s'il existait toute une partie du Net inconnue des utilisateurs et des moteurs de recherche ? Description du "Deep Web".

Inside Electronic Pipo

Le Web peut, à bien des égards, être considéré comme une forêt. L'internaute y rentrant se munit d'une carte et emprunte les sentiers qui y sont dessinés, de plus bien déblayés par tous ceux qui sont passés avant lui. Mais l'itinéraire qu'il suit ne lui révèle qu'une partie infime de tous les arbres de la forêt. Ceux-ci, pour leur immense majorité, restent dans l'ombre et le silence loin du regard des internautes.

Le Web profond

<a href="http://img.chan4chan.com/img/2009-03-30/26.jpg" title='JPEG - 80.5 ko' type="image/jpeg">

Cette métaphore d'introduction est évidemment aussi frappante qu'elle est incorrecte, incomplète et pédante. Elle a cependant le mérite de faire une bonne introduction au concept de Web invisible, ou Web opaque, ou Web profond. Bien qu'il soit presque impossible de s'en rendre compte en naviguant de manière classique, seule une fraction de toutes les données en ligne sont accessibles facilement par les internautes.

Cette fraction est nommée le Web visible, ou **Web surfacique**. Il consiste en toutes les pages aspirées puis indexées par les moteurs de recherche et stockées dans les bases de données de leurs serveurs. A l'inverse, les pages mal ou non-répertoriées par les moteurs de recherche 'conventionnels' ne peuvent être trouvée par le biais de ces moteurs de recherche, et constituent le **Web profond**. Elles sont pourtant bien présentes, mais les moteurs de recherche qui constituent l'interface d'entrée sur le Web pour la plupart des internautes ne les ayant pas en mémoire, leur accès est donc restreint.

Ce défaut d'indexation a de nombreuses origines, dont voici les principales :

- Certaines bases de données sont tout simplement trop grosses pour être entièrement répertoriées. C'est le cas par exemple de l'Internet Movie Database, qui possède plusieurs millions de pages. Dans certains cas ce sont les pages elles mêmes qui sont trop volumineuses pour être archivées par les moteurs de recherche.

- Les formats des documents en lignes ne sont parfois pas supportés par les moteurs de recherche. Ce phénomène se résorbe progressivement, puisque les formats Pdf, .doc, .xls et autres sont peu à peu acceptés et indexés par les moteurs de recherche : aucun d'entre eux ne l'était avant le début des années 2000.

- Une autre raison enfin prend sa source tant dans le fonctionnement des moteurs eux mêmes que dans la **structure du Web**. La recherche et l'indexation est faite par des robots, qui pour ce faire naviguent de pages en pages en utilisant les liens présents sur celles-ci pour passer des unes aux autres. Or le réseau de sites qui compose le Web n'a - comme son nom ne l'indique pas - pas la forme d'une toile d'araignée mais plus d'un « noeud papillon » : un centre auquel se rattachent deux ailes constituées de pages sources et pages destinations (donc des pages avec des liens qui ne renvoient que vers le "coeur" ou qui ne sont accessibles que depuis des liens du "coeur"), ainsi que des filaments qui ne sont accessibles que depuis des zones très précises du Web. Enfin pour couronner le tout certains sites Internet sont de véritables ilots au milieu de la mer de l'information : ne possédant aucun liens externes et aucun site ne redirigeant vers ces endroits, ces zones isolées ne sont

accessibles que si l'on connait leur existence.
<a href="http://www.vol-de-papillon.com/images/papweb.gif" title='GIF - 11.1 ko' type="image/gif">

Schéma de la structure du Web

Au vu de cette structure particulière du Web, les robots d'indexations sont donc parfois dans l'impossibilité de répertorier des groupes entiers de données mises en lignes, leur accès étant rendu impossible par le manque de passerelles entre les pages internet.

Les caractéristiques du Web profond

Ceci est d'autant plus dommage que la taille, le contenu et la qualité du Web profond recèlent de nombreuses surprises.

Ainsi on estime que **le Web profond est de très loin plus volumineux que le Web surfacique**. Les chiffres varient légèrement mais l'ordre de grandeur le plus couramment admis est que le premier est 500 fois plus volumineux que le second, sachant que le Web invisible croit plus rapidement : les estimations sont de 900% par an... Par ailleurs seuls 5% de ces pages ne sont pas consultables librement, ce qui signifie que la quasi-totalité de ces informations sont disponibles **gratuitement**.

Le contenu du Web profond est quand à lui assez spécifique : il est dans sa plus grande majorité constitué de bases de données concernant des sujets précis et majoritairement scientifiques, de bibliothèques en lignes, et de publications diverses (autour de 70% pour ces trois composantes). Bref il semblerait que de fait le Web profond soit avant tout **un immense réservoir de connaissances variées** !
<a href="http://academics.smcvt.edu/sburks/Image7.gif" title='GIF - 6.5 ko' type="image/gif">

Répartition du contenu du Web profond

Mais là ou l'affaire vient encore plus intéressante c'est que **le Web invisible se distingue aussi par la qualité de ses pages**... Il est d'ailleurs assez simple de comprendre pourquoi : la pertinence des pages du Web profond est beaucoup plus forte car il est massivement constitué de sites spécialisés rédigés par des chercheurs, experts ou professionnels : un des meilleurs exemple étant sans doute celui de la National Library of Medicine qui est considérée comme la plus grande base de données médicale du Net. Certaines agences de recherche vont jusqu'à estimer que la qualité des pages est environ **trois fois supérieure à celles du Web surfacique**, et même si la qualité est une notion toute relative ce résultat est assez solidement établi.

Man vs Wild : explorer le Web invisible

Bref, on l'aura compris, le Web profond peut servir pour des occasions très spécifiques, comme de la recherche par exemple, et l'étudiant a tout intérêt à savoir explorer le Web invisible dans le cadre de ses études, ou par curiosité.

Une première méthode est tout simplement de **se servir de bases de données spécialisées** ou de sites relayant l'indexation du contenu de ces bases de données.

Quelques moteurs et sites permettant d'accéder au Web profond :

www.incywincy.com

www.completeplanet.com

scholar.google.fr

www.archive.org

L'autre solution est de formuler ses requêtes sur les moteurs traditionnels de telle façon que ceux-ci donnent accès à des répertoires ouvrant eux sur des pans spécialisés du Web, appelés des **méta-ressources**, qui contiendront les précieux documents recherchés. En plus de chercher des mots en rapport avec la discipline ou la question étudiée, il est préférable par exemple d'adjoindre le format dans lequel on désire trouver des documents ou des termes en rapport avec des bases de données : *Pdf, links, directories, resources*, etc..

L'exploration manuelle du Web profond est encore hélas très fastidieuse, la technologie ne permettant pas encore de couvrir le véritable océan informationnel qu'il constitue.