Le web invisible

1. Qu'est-ce que c'est ?

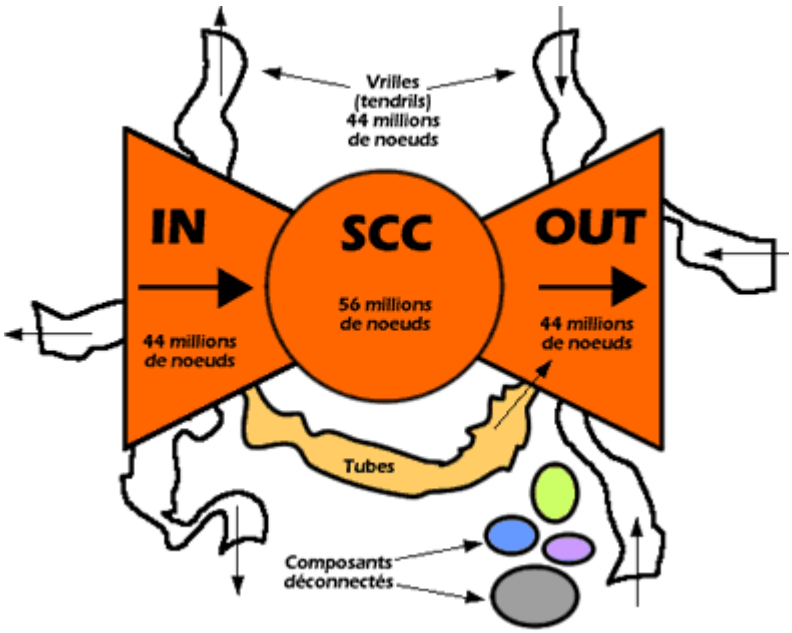
- **Le web invisible est l'ensemble des documents accessibles en ligne mais qui ne sont ni lus, ni indexés par les moteurs de recherche traditionnels.**

On l'appelle aussi « web caché » ou « web profond » (*invisible web* ou *deep web* en anglais). En 2008, on estimait que le web invisible représentait 70% des ressources totales du web.

2. Comment ça marche ?

Internet est une architecture de multiples réseaux dont le web n'est qu'un exemple parmi d'autres. Or le système du web repose sur l'image de la toile d'araignée (on appelle d'ailleurs aussi le web « la Toile ») : dans le web visible les pages sont reliées entre elles, décrites par des mots-clés et interrogeables avec un moteur de recherche. Mais **une part importante de la documentation de recherche échappe encore au Web** ou n'y est pas disponible directement et gratuitement.

En effet, certaines pages sont générées au vol par l'interrogation de bases de données, d'autres sont interdites aux robots de collecte des moteurs, d'autres encore sont trop « profondes » depuis la page d'accueil… La visite d'indexation des robots se réduit donc aux seules pages accessibles par des adresses web fixes.



Le web invisible peut s'illustrer par la "théorie du nœud papillon".

Le nœud, ce sont les pages du Web liées entre elles. L'aile gauche, ce sont les pages qui peuvent être jointes via le nœud mais qui ne contiennent pas de liens vers le nœud. L'aile droite, c'est l'inverse.

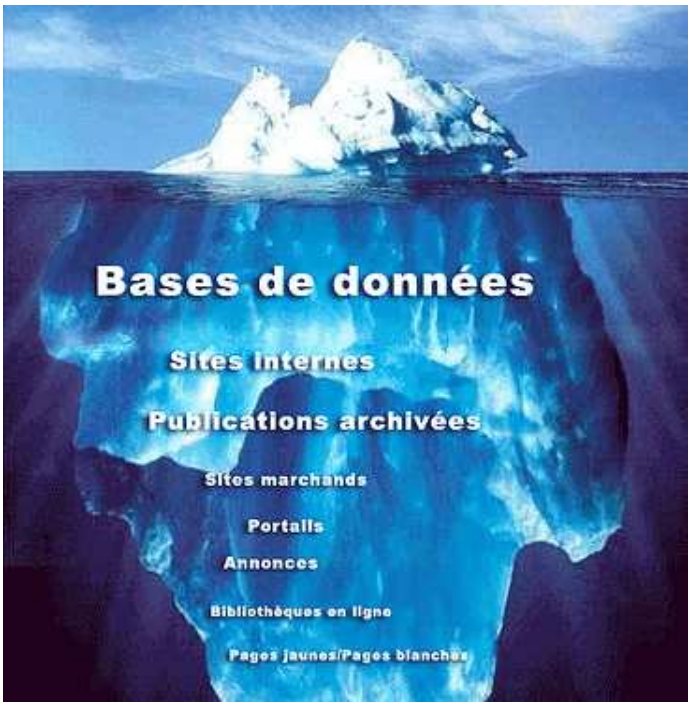
Les pages situées dans le nœud, c'est-à-dire au cœur du Web, sont bien référencées dans les moteurs de recherche. Elles pointent vers d'autres pages et d'autres pages pointent vers

elles. En revanche les parties situées dans l'aile gauche ou dans l'aile droite sont moins bien reliées, et donc moins bien intégrées à la Toile.

3. Qu'est-ce qu'on y trouve ?

Les principales informations composant le web invisible sont :

- Les sites protégés par mots de passe et les pages interdites de référencement (interdiction posée par l'auteur des pages ou le gestionnaire du site en utilisant un fichier Robots.txt)
- Les contenus dynamiques : pages web dynamiques (structure des bases de données : la forme est fixe et le contenu variable, ce qui permet ainsi de s'adapter aux critères de recherche), information de presse diffusée en temps réel.
- Les fichiers dont le format n'est pas indexable : documents dans des formats de données non supportés par les robots d'indexation (mais de plus en plus de formats sont désormais indexés)
- Les pages sans liens hypertextes : pages qui ne sont pas liées par d'autres pages et qui ne peuvent donc pas être découvertes par les robots d'indexation. D'autres pages ne sont accessibles qu'à travers des liens produits par l'exécution de programmes, par exemple en JavaScript.



On utilise souvent l'image de l'iceberg pour représenter les proportions web visible/web invisible .

"Distribution des sites du Deep Web par types de contenu" de l'étude Bright Planet.

4. A quoi ça sert ?

Dans un parcours universitaire, il est indispensable savoir que toutes les informations ne seront pas trouvables par un moteur de recherche ! Se limiter à Google ou Yahoo, c'est prendre le risque de passer à côté d'une grande masse d'information, souvent d'excellente qualité. Votre Université, par l'intermédiaire de son Service Commun de la Documentation, consacre un budget important à l'abonnement aux revues, bouquets de revues et bases de données qui fournissent une documentation scientifique de haut niveau. Renseignez-vous sur les ressources disponibles et leurs conditions d'accès.