Web invisible

written by Françoise Laugée | 21 décembre 2009

Egalement Web caché ou Web profond (*deep Web*), le terme désigne la partie du Web regroupant l'ensemble des ressources (documents, pages, sites Web) non indexées ou partiellement indexées par les moteurs de recherche ou les annuaires classiques. Les résultats proposés par ces derniers porteraient seulement sur 10 % du Web. Ces documents ou ces sites, qui échappent au balayage opéré par les outils de recherche généralistes, ne sont pas inaccessibles pour autant. Il est possible de les interroger directement, comme c'est le cas pour une base de données. De nombreux répertoires ou portails spécialisés sur Internet recensent les bases de données et les centaines d'outils de recherche spécialisés dans l'indexation des documents «cachés», permettant ainsi d'identifier des gisements d'informations spécialisées. La partie invisible du Web serait 500 fois plus importante que la partie visible.

Le Web invisible se compose des sites web trop volumineux pour être indexés dans leur intégralité, comme des bibliothèques en ligne, des bases de données ; les documents constitutifs d'une banque de données accessibles par un moteur de recherche interne a`celle-ci ; les pages accessibles uniquement avec un identifiant et un mot de passe ; les pages volontairement interdites aux robots d'indexation par leur créateur ; les pages écrites dans un format propriétaire encore mal indexé ; les pages dites dynamiques, c'est-a`-dire générées par une requête, ainsi que les intranets et extranets.

Les experts américains Chris Sherman et Gary Price, auteurs d'un ouvrage intitulé *The Invisible Web*, divisent le Web invisible en quatre catégories : *The Opaque We*b regroupant les pages non indexées par défaillance des moteurs de recherche classique, *The Private Web* dont les pages sont protégées volontairement, *The Proprietary Web* pour lequel l'usage d'un identifiant est nécessaire et, enfin, *The Truly Invisible Web*, inaccessible pour des raisons d'incompatibilité technique.

Une autre terminologie est proposée par la société américaine Bright Planet, leader de la recherche dans ce domaine, adoptant le critère d'accessibilité plutoˆt que celui de visibilité pour distinguer les ressources reconnues par les moteurs de recherche classiques, le Web de surface (*surface Web*), de celles exploitables uniquement par des outils d'un autre type, le Web profond (*deep Web*). Selon Bright Planet qui édite un annuaire du Web invisible (http://aip.completeplanet.com), les 60 premiers sites de ce Web profond sont des sites scientifiques, des bases de données, des sites universitaires, des sites de médias, de commerce en ligne et des intranets de grandes entreprises : ensemble, ils représentent plus de 40 fois le volume de tous les sites du Web de

surface.

Un répertoire des outils de recherche pour explorer le Web invisible est réalisé et mis a` jour par la Bibliothèque nationale de France (http://signets.bnf.fr).