

Web invisible

Description

Egalement Web cache? ou Web profond (*deep Web*), le terme de?signe la partie du Web regroupant l'ensemble des ressources (documents, pages, sites Web) non indexe?es ou partiellement indexe?es par les moteurs de recherche ou les annuaires classiques. Les re?sultats propose?s par ces derniers porteraient seulement sur 10 % du Web. Ces documents ou ces sites, qui e?chappent au balayage ope?re? par les outils de recherche ge?ne?ralistes, ne sont pas inaccessibles pour autant. Il est possible de les interroger directement, comme c'est le cas pour une base de donne?es. De nombreux re?pertoires ou portails spe?cialise?s sur Internet recensent les bases de donne?es et les centaines d'outils de recherche spe?cialise?s dans l'indexation des documents «cache?s», permettant ainsi d'identifier des gisements d'informations spe?cialise?es. 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 e?tre indexe?s dans leur inte?gralite?, comme des bibliotheq?ues en ligne, des bases de donne?es ; les documents constitutifs d'une banque de donne?es 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 cre?ateur ; les pages e?crites dans un format proprie?taire encore mal indexe? ; les pages dites dynamiques, c'est-a?-dire ge?ne?re?es par une reque?te, ainsi que les intranets et extranets.

Les experts ame?ricains Chris Sherman et Gary Price, auteurs d'un ouvrage intitule? *The Invisible Web*, divisent le Web invisible en quatre cate?gories : *The Opaque Web* regroupant les pages non indexe?es par de?faillance des moteurs de recherche classique, *The Private Web* dont les pages sont prote?ge?es volontairement, *The Proprietary Web* pour lequel l'usage d'un identifiant est ne?cessaire et, enfin, *The Truly Invisible Web*, inaccessible pour des raisons d'incompatibilite? technique.

Une autre terminologie est propose?e par la socie?te? ame?ricaine Bright Planet, leader de la recherche dans ce domaine, adoptant le crite?re d'accessibilite? pluto?t que celui de visibilite? 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 e?dite un annuaire du Web invisible (<http://aip.completeplanet.com>), les 60 premiers sites de ce Web profond sont des sites scientifiques, des bases de donne?es, des sites universitaires, des sites de me?dias, de commerce en ligne et des intranets de grandes entreprises : ensemble, ils repre?sentent plus de 40 fois le volume de tous les sites du Web de surface.

Un re?pertoire des outils de recherche pour explorer le Web invisible est re?alise? et mis a? jour par la Bibliothe?que nationale de France (<http://signets.bnf.fr>).

Categorie

1. A retenir
2. Repères & tendances

date créée

21 décembre 2009

Auteur

francoise