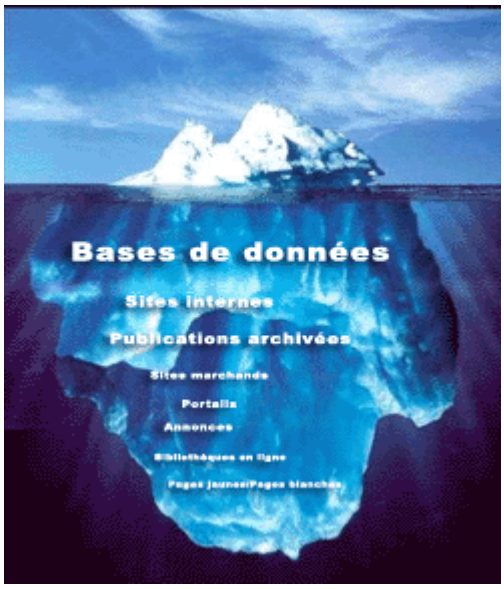
WEB INVISIBLE ET BASES DE DONNEES



Présenté par : Manuel MOLLI, Stéphane COLAS
Avec la collaboration de : Farouk DJELLOUL

Qu'est ce que le web invisible ?

Le web invisible désigne l'immense partie du web non accessible aux moteurs de recherche conventionnels. Il regorge d'informations de haute qualité et de véritables trésors cachés, trop souvent ignorés par les nombreux professionnels qui ont encore le seul « réflexe Google ».

On y trouve des bibliothèques en lignes (articles scientifiques, essais, etc...), des bases et banques de données (payantes ou gratuites), des documents protégées ou trop volumineux pour être indexés, ou encore des pages dynamiques (générées par requêtes).
Il s'agit d'une manne très souvent ignorée par les robots des moteurs de recherche classiques qui naviguent de liens en liens pour indexer les pages web qu'ils rencontrent.

Estimation de la taille du web invisible : des chiffres impressionnants.

Certains parlent de web « de surface » et de web « profond » (ou « deep web »), comparant le web à un gigantesque iceberg en perpétuelle expansion. Les estimations sur la taille du web profond sont très variables (et difficiles par définition).

A ce sujet, des chiffres impressionnants ont été avancés par certaines études (*Brightplanet*, *Cyveillance*) : la partie immergée du web serait jusqu'à 500 fois plus vaste que le web visible que nous connaissons tous !
L'étude Brightplanet appuie ses hypothèses sur le fait que seulement quelques dizaines de sites représenteraient à eux seuls plus de 40 fois le volume du web de surface, ou web visible (celui sur lequel nous surfons tous les jours).

Quelques exemples : les sites scientifiques (comme la *NASA*), les bases de données (*Lexis Nexis* et *Dialog*), les sites universitaires (*Berkeley*, etc...), les sites de commerce en ligne (*e-bay*).
Autre exemple parlant, dans le secteur de la Santé, on trouve des millions d'articles (base *PubMed* de la « *National Library of Medicine* »), les notices de la *Banque de Données Santé Publique* (*Direction Générale de la Santé*), les 45.000 médicaments du marché français répertoriés dans *Theriaque*, etc... Et ce parmi des milliers d'autres sources.

Liens intéressants :

http://www.intelligence-center.com
http://c.asselin.free.fr/french/theses_IE.htm
http://www.brightplanet.com/pdf/deepwebwhitepaper.pdf

Recherche personnelle :

Comme nous l'avons vu ci-dessus, le web profond peut être une véritable mine d'or et aujourd'hui, la difficulté est plus d'analyser une information que de la rechercher. La moindre requête peut vous faire crouler sous des millions de résultat, sans parler du fait que les informations ne sont pas indexées de la même façon selon les moteurs de recherche utilisés.

L'objectif de cette recherche était de trouver un maximum de bases de données (articles, données financières ou scientifiques, bibliothèques en ligne, etc...) accessibles si possible gratuitement sur Internet.

Pour cela, nous nous sommes basé sur les *bluesheets* de *Dialog*, qui fait parti, avec la bibliothèque du Congrès Américain, des plus grandes sources d'informations formelles du monde (600 bases de données, ce qui représente un volume équivalent à 20% du web mondial). Les *bluesheets* nous ont servi à utiliser un maximum de mots clés dans nos requête, avec des résultats pas toujours heureux mais quelques bonnes surprises.

[Source : http://library.dialog.com/bluesheets/html/bln.html]

Le résultat est (très) loin d'être exhaustif et ne concerne qu'une infime partie des bases disponibles sur *Dialog*, mais l'intérêt de ces recherches résident surtout dans une certaine méthodologie.

Vous trouverez donc ci-dessous quelques liens pointant vers des portails dédiés, des sites permettant un accès gratuit à certaines bases de données, des bibliothèques et des librairies en ligne, etc...

- **Bases de données gratuites :** http://dadi.enssib.fr/index.php?page=search
- **Bases de données bibliographiques :**
  http://www-bu.univ-paris8.fr/Ref/DocBdd.html#cataloguesarticles
- **Panorama des outils de recherche sur le web :**
  http://www.adbs.fr/site/repertoires/outils/outils-recherche.php#bases
- **Medline by Scirus - scientific information :** http://www.scirus.com/srsapp/
- **Pubmed (National Library of Medicine) :**
  http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
  http://www.nlm.nih.gov
  http://pubmedcentral.nih.gov
  http://www.medportal.com
- **Internet Law Library :** http://www.lawguru.com/ilawlib/index.html
- **Company Research - Financial informations – News :**
  http://www.library.csuhayward.edu/staff/Faust/company_research.htm
- **Free Online Annual Reports (with login) :** http://www.annualreportservice.com
- **BIAM - molécules, substances, indications et contre-indications :**
  http://www.biam2.org/accueil.html
- **Veille.com - Bases de données :**
  http://www.veille.com/fr/annuaire.php3?id_rubrique=106
- **ERIC :** http://www.eric.ed.gov/

- **SEC Online :** http://www.sec.gov/edgar/searchedgar/webusers.htm
  http://www.secinfo.com/
  http://www.sec.gov/info/edgar/siccodes.htm

- **Investext :** http://www.cas.org/ONLINE/DBSS/investextss.html
  http://stneasy.fiz-karlsruhe.de/html/french/login1.html?service=STN
  http://www.alacra.com/cgi-bin/alacraout.exe/alacra/help/databasealpha.htm

- **ICC British Company Directory :**
  http://www.cd-rom-directories.co.uk/index.html?target=p_3758.html&lang=en-gb&gclid=CI209uC5u4ICFS2gEAodJj-BXw

- **ICC British Company Financial Datasheets :** http://www.hrzone.co.uk/databases/
  http://www.datasheet.in/
  http://www.alldatasheet.com/

- **ICC British Company Annual Reports :**
  http://www.solent.ac.uk/library/subject/page95.stm
  http://www.cd-rom-directories.com/index.html?target=p_3701.html&lang=en-us&gclid=CKr6kvryu4ICFRBREgodlU_LAA
  http://www.innovation.gov.uk/

- **Population Demographics™ :**
  http://www.anywho.com/cgi-bin/webdrill?catkey=gwd/Top/Science/Social_Sciences/Demography_and_Population_Studies/Europe
  http://infotree.library.ohiou.edu/bysubject/general/statistics/

- **Kompass USA :** http://www.chemindustry.com/category/11.html
  http://www.industryweek.com/research/iw1000/2005/IW05Enter.asp

- **Civil Engineering Abstracts :** http://www.pubs.asce.org/cedbsrch.html

- **GPO Monthly Catalog :** http://www.gpoaccess.gov/cgp/index.html

- **World Textiles™ :** http://www.sciencedirect.com

- **EMBASE® :** http://stneasy.fiz-karlsruhe.de

- **CA SEARCH® - Chemical Abstracts® (1967- present) :** http://www.cas.org/

- **LC MARC – Books :** www.dimdi.de

- **Fort Worth Star-Telegram :** www.dfw.com/

- **Current Contents Search® :** http://www.ovid.com/site/catalog/DataBase

- **Prous Science Drugs of the Future™ :** www.prous.com/

- **Hospital Inpatient Profiles (HIP) :** www.tdrdata.com

- **Hospital Outpatient Profiles Database (HOP) :** http://www.marketresearch.com

- **The Irish Times :** http://www.ireland.com/newspaper/articleindex/2006/0111/index.html