Le Web invisible…

Plan

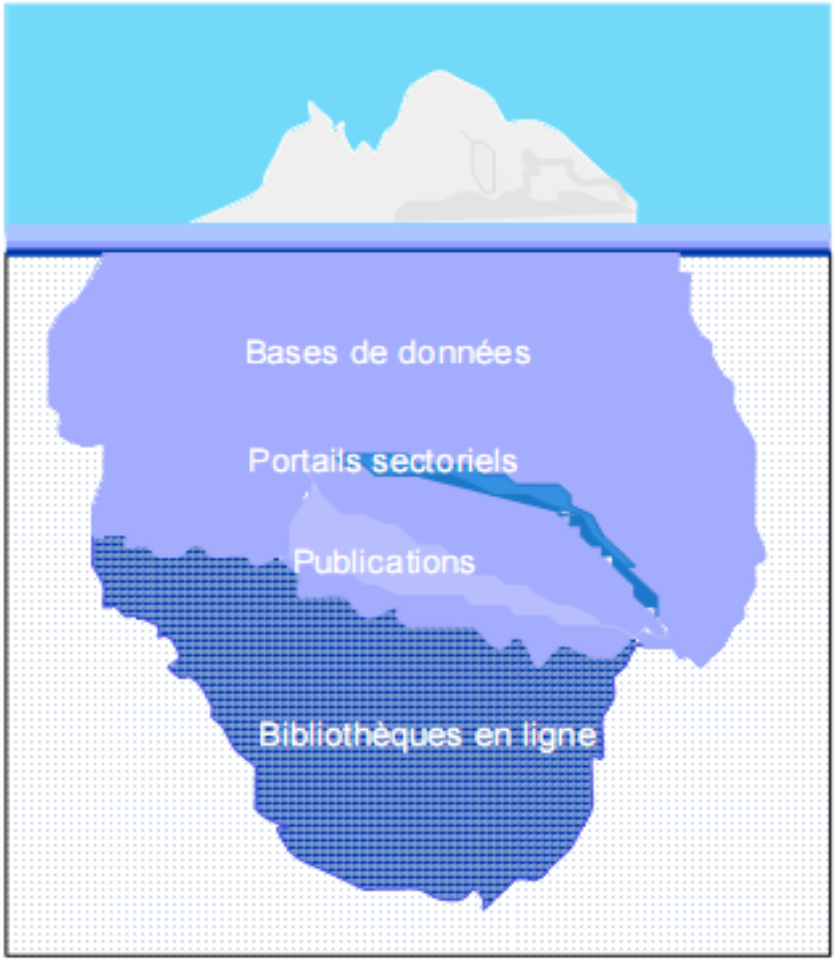
- **Définition**
- **Les 4 Types du web invisible**
- **Les bases de données**
- **Les outils et Moteurs de recherche**
- **Bibliothèques en ligne**
- **Portails Sectoriels**
- **Web invisible Vs Web Visible**

Web invisible, web caché, web profond

Définition :

- Le "web invisible" (*deep web, hidden web*) désigne la partie du web non accessible aux moteurs de recherche classiques.

- Le web invisible comprend des bases, banques de données et bibliothèques en ligne gratuites ou payantes.



- Des moteurs comme Google, MSN/Live Search, Yahoo! Search ou des répertoires tels que Yahoo! Directory ne vous donnent accès qu'à une petite partie (inférieure à 10%) du web, le Web Visible.

- La technologie de ces moteurs conventionnels ne permet pas d'accéder à une zone immense du web, le Web Invisible est un espace beaucoup plus important que le web visible.

Pourquoi ?

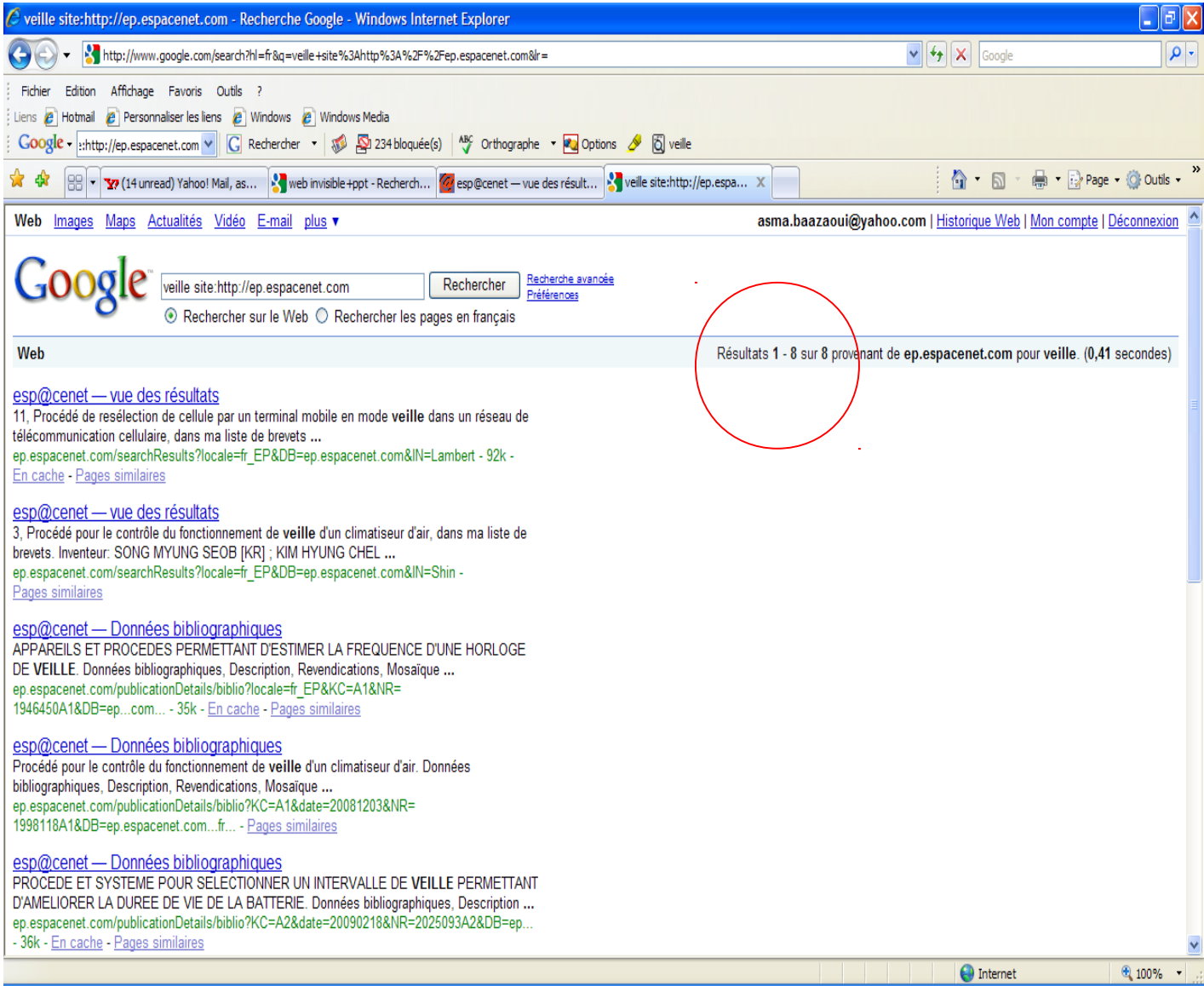
- Parce que la majeure partie des sites du web invisible sont des sites spécialisés, dédiés à une activité, une technologie, un métier et que leur contenu émane ou est validé par des professionnels, spécialistes et experts.

Exemple illustrant la différence entre web « visible » et web « invisible »

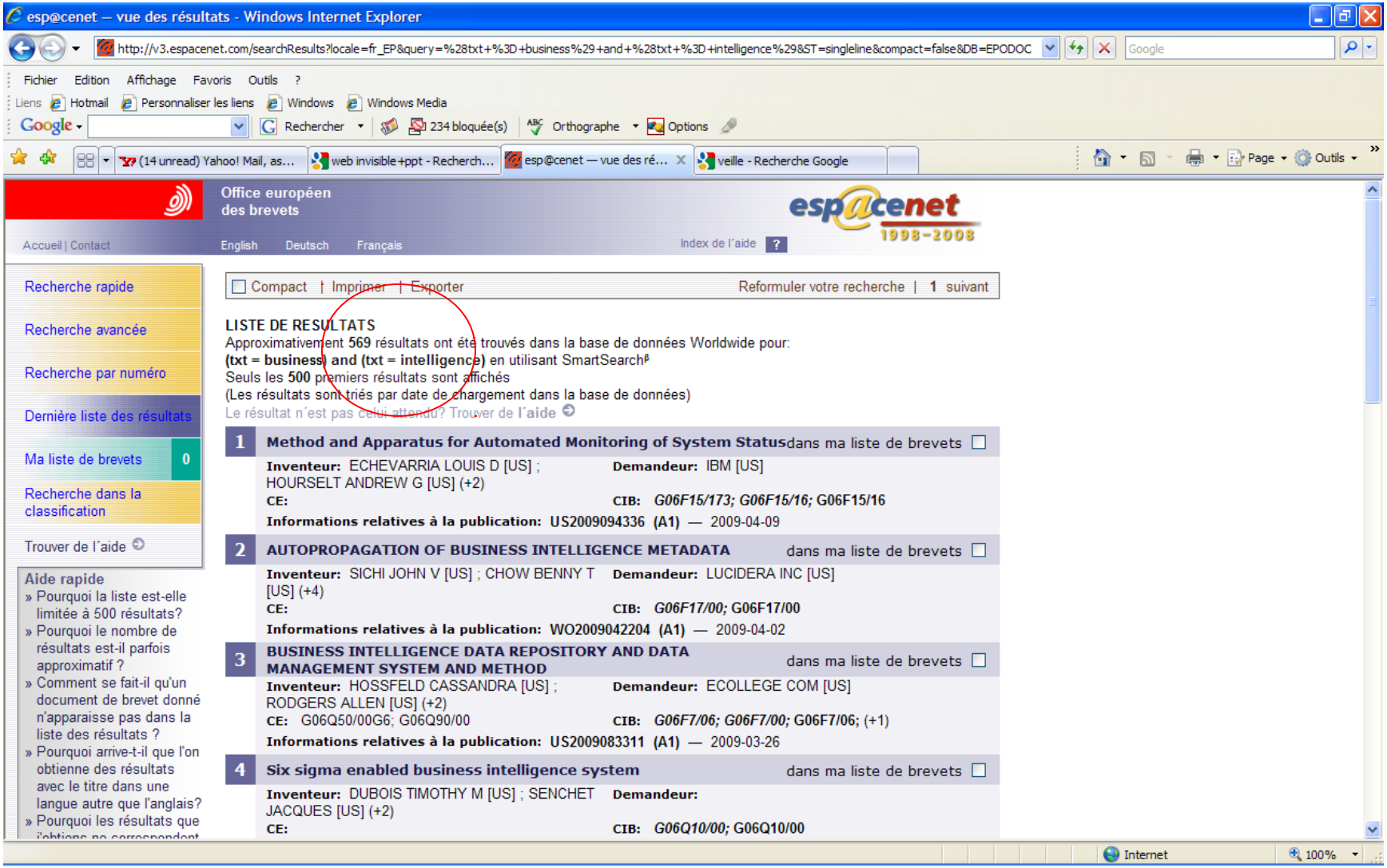
Recherche du mot « veille » en limitant la recherche sur le site ep.espacenet.com, à partir de Google.

La requête correspondante est donc : veille site:http://ep.espacenet.com

8 résultats sont obtenus dans le « web visible »



Recherche du mot « veille » à partir du formulaire de recherche sur Espacenet.com **:569 résultats dans le web « invisible »**



Google n’est donc pas capable de trouver tous les documents stockés.

L’étude "The Deep Web: Surfacing Hidden Value" réalisé par Michael K. Bergman propose des ordres de grandeur permettant de mieux mettre en perspective le web profond à l'égard du web de surface :

- l'information publique sur le web profond est considérée comme de 400 à 550 fois plus volumineuse que le web de surface (web visible)

- le web profond est constitué de plus de 200 000 sites web.

- 60 % des sites les plus vastes du web profond représentent à eux seuls un volume qui excède de 40 fois le web de surface.

- le web profond croît plus vite que le web visible.
- plus de la moitié du Web Profond est constitué de Bases de données spécialisées.
- 95% du contenu du web profond est accessible à tous (gratuit ou à accès non restreint)

Une partie du web est non accessible aux moteurs parce que :

- *Les documents, pages et sites web ou bases de données sont trop volumineux* pour être entièrement indexés.

  Exemple : L'Internet Movie Database, une base de donnée en libre accès consacrée au cinéma répertorie plus de 7 millions de pages descriptives consacrées aux films et acteurs, représentant chacune une page web. Soit plus de 7 millions de pages. Les moteurs conventionnels n'indexent pas la totalité de ce contenu (son indexation varie entre 5 et 60 % selon les moteurs).

- *des pages sont protégées par l'auteur* (balise Meta qui stoppe le robot) :

  Certains sites sont protégés par leur créateur ou gestionnaire (webmaster), qui, grâce à un fichier *robot.txt* inséré dans le code des pages, interdit leur accès aux robots des moteurs.

  Exemple : le site du journal *Le Monde* interdit aux robots des moteurs de recherche l'accès à ses pages payantes.

- *les pages sont protégées avec une authentification par identifiant (login) et mot de passe* :

  De nombreux sites, qu'ils soient payants ou gratuits, protègent tout ou partie de leur contenu par mot de passe. Les robots de moteurs n'ayant pas la faculté de taper des mots dans des formulaires complexes, ces pages ne leur sont pas accessibles.

- *le format des documents n'est pas reconnu par les moteurs* (de moins en moins vrai aujourd'hui) :

  Il y a quelques années, on incluait dans le Web Invisible toutes les pages aux formats autres que le html, seul format reconnu et indexé par les moteurs. Aujourd'hui, les moteurs indexent les documents Word, Excel, Power Point, PDF....Seul le Flash restent assez mal indexé de par sa nature.

Les 4 types de web distingués par Chris Sherman et Gary Price :

- Chris Sherman et Gary Price, "**search engines' US experts**", proposent dans leur ouvrage ***"The Invisible Web"*** de distinguer***4 types de web invisible***:

- **The *Opaque Web*** :

  les pages qui pourraient être indexées par les moteurs mais qui ne le sont pas (limitation d'indexation du nombre de pages d'un site, fréquence d'indexation, liens absents vers des pages ne permettant donc pas un crawling)

- **The *Private Web*** :

  les pages web disponibles mais volontairement exclues par les webmasters (mot de passe, metatags ou fichiers dans la page pour que le robot du moteur ne l'indexe pas).

- **The *Proprietary web*** :
  pages seulement accessibles pour les personnes qui s'identifient. Le robot ne peut donc pas y accéder.

- **The *Truly Invisible Web*** :
  contenu qui ne peut être indexé pour des raisons techniques. Ex : format inconnu par le moteur (Google est l'un des rares moteurs à reconnaître autant de formats), pages générées dynamiquement (incluent des caractères comme ? et &).

Web invisible : Les Bases de Données

- Ce sont des ressources en pleine mutation. Encore payantes en totalité il y a quelques années, de plus en plus d'informations de qualité, notamment à travers les bases de données, deviennent gratuites.

Les bases de données gratuites :

Sites de références scientifiques gratuits ou payants (Université de Bordeaux I) :

Ce site recense des centaines de ressources (sites, base de données) gratuites ou payantes dans le domaine scientifique :

Bibliographies générales et ressources pluridisciplinaires, Bibliographies spécialisées, Anthropologie, Astronomie et astrophysique, Agriculture, Biologie, Botanique, Brevets, Chimie, Energie, Géologie, Informatique, etc.

Les bases de données gratuites (URFIST de Nice) :

L'URFIST de l'Université de Nice-Sophia Antipolis propose elle aussi une sélection de bases de données gratuites.

The Internet Archive :

The Internet Archive est une bibliothèque digitale destinée à conserver tous les documents numériques issus de l'internet pour les préserver d'une disparition complète.

The IA fournit des documents créés à partir de 1996 (**10 milliards de pages web**). Accessible au public depuis le 24 octobre 2001.

Google News Archives :

Google News Archive, qui permet de rechercher parmi les archives des actualités de ces 200 dernières années :

Google a passé des accords avec de prestigieuses sources de presse telles que le Time, le Wall Street Journal, le New York Times, la BBC, le Guardian ou le Washington Post (archives gratuites ou payantes) et de grands services d'agrégateurs de presse, comme Factiva, LexisNexis, Thomson Gale et HighBeam Research (payants), afin d'indexer le plein texte de leurs articles sur les 2 siècles passés.

2 types de recherches sont disponibles :

- *Search Archives* : classiquement, il faut taper un mot clé pour obtenir tous les articles en relation avec la requête.

- *Show Timeline* : permet d'afficher la chronologie d'un événement ou l'actualité d'une personne à travers les années.

Les bases de données payantes :

Questel-Orbit:

Service français de plus de 80 bases de données dédiées à la **Propriété Industrielle** (Europe et Internationale):

brevets, marques et modèles, informations scientifiques et techniques, marques et noms de domaine Internet; également affaires (fichiers et profils d'entreprises, défaillances, presse internationale, congrès...), actualités, sciences humaines et sciences sociales.

Les produits : QWEB v.2, QPAT, Imagination, Trademark Explorer.

Dialog Datastar (groupe Thomson) :

Plus de 700 bases de données de nature variée:

Intelligence Economique, Economie, Gestion, Chimie, Marché et Produits, Biomédical, Santé, Pharmacie, Ingénierie et Technologies, Environnement, Données gouvernementales, Sciences de la terre,…Informations scientifiques et Economiques, Sociétés européennes, Biomédical, Pharmacie, Actualité de l'Europe de l'Ouest et Europe de l'Est, Informations Techniques, Economie et Affaires Internationales….
http://www.datastarweb.com/
http://www.dialog.com/

Factiva :

Il permet d'obtenir des informations personnalisées à travers la définition d'un profil de reherche concernant des acteurs ,des marchés ,des concurrents..

Factiva donne accès à des publications en 22 langues provenant de plus de 110 pays. 8 000 grandes **publications**, 8 500 sites Internet, plus de 20000 profils de sociétés et photos récentes.

Lexis Nexis :

Lexis Nexis permet d'obtenir des informations personnalisées à travers la définition d'un profil de recherche concernant des acteurs, des marchés, des technologies ou des concurrents **à partir de 35.000 sources** (journaux, rapports, brevets etc.) indexées dans tous les domaines et pour de nombreux pays (plus de 90)

- **Lexis.com : archives juridiques**
- **Nexis.com : archives journaux**

Web invisible : Outils et moteurs de recherche

Turbo10

- **Turbo10, le métamoteur britannique**, utilise des moteurs de recherches spécialisés permettant de rechercher dans des bases de données ou des documents du "web profond" dans des domaines spécialisés.

- Il offre la possibilité de se connecter à plus de 1000 moteurs spécialisés ou généralistes:Turbo10 interroge par défaut altavista.com, dogpile.com, google.com, hotbot.com, lycos.com, metacrawler.com, search.msn.com et yahoo.com, donc des moteurs plutôt anglophones.

- L'internaute peut choisir de rajouter, grâce au module "My Collection", les moteurs figurant dans une liste assez impressionnante (1170 à ce jour). Il peut choisir un moteur généraliste comme voila.fr ou des moteurs spécialisés.

Ce qui permet à Turbo10 d'explorer (un peu) le "web invisible" ou "web profond": Car cette liste propose des moteurs internes de portails, de vastes bases de données universitaires ou sites de e-commerce :

zdnet.com, **amazon.com**, **europages.net** (annuaire de sociétés), **imdb.com** (cinéma), **dictionary.com** (dicos et thésauri), **eea** (european environment agency), encyclopedia.com, findarticles.com (archives d'articles depuis 1998), etc.

Xrefer :

Moteur de recherche britannique **spécialisé dans les ouvrages de référence** :encyclopédies, dictionnaires et recueils de citations.

Thèmes : art, santé, langues, philosophie, musique, sciences, technologies, géographie et littérature anglaise...

Adobe PDF Search :

Permet de rechercher parmi plus d'1 million de documents au format Adobe PDF (Portable Document Format).

Wondir :

- Wondir associe les possibilités d'un métamoteur et d'un service de recherche humain.

- Wondir est différent des autres outils de recherche:

  D'abord parce qu'il est géré par une organisation à but non lucratif. Ensuite, parce que le but de cette fondation est de fournir de l'information de haute qualité à tous.

- Wondir combine la technologie d'un métamoteur à des technologies propriétaires qui permettent d'utiliser les ressources du **web invisible**.

- Lorsque l'on tape une requête, la page de résultats se divise en plusieurs parties :

  - les résultats web.
  - les résultats issus des newsgroups (forums de discussion) et mailing lists (listes de diffusion)
  - les propositions de service d'experts de la communauté Wondir pouvant potentiellement répondre à votre question.
  - des questions et leurs réponses en relation avec la requête.
  - des dépêches d'actualités liées au domaine de la requête.

- Wondir dispose d'une communauté de volontaires qui répondent aux questions trop complexe pour le moteur de recherche.

Web invisible : bibliothèques en ligne :

- Il s’agit de sites donnant accès à des catalogues d'ouvrages, périodiques. Portails fédérant des magazines en ligne.

Les catalogues de la Bibliothèque Nationale de France (BnF) :

- Ces catalogues décrivent les **documents et objets conservés à la BNF** (documents imprimés, documents audiovisuels, cartes et plans, monnaies et médailles, manuscrits).

- Certains sont numérisés et/ou microfilmés.

LibDex :

- Créé par Peter Scott :**University of Saskatchewan (Canada)**

- Répertoire de plus de **17000 bibliothèques publiques** mais aussi privées à travers le monde (133 pays: de l'Albanie au Zimbabwe).

- La recherche peut s'effectuer par pays mais également par Open Access Catalogs (OPAC) c'est-à-dire les catalogues informatisés signalant les ouvrages et les périodiques présents dans la bibliothèque.

Web invisible : Les portails sectoriels :

- Leur approche est verticale : Ce sont des portails spécialisés dans un secteur d'activité, une technologie.

- On peut aussi parler de "Vortail" (contraction de "vertical" et "portail").

- Ils sont nombreux. Ainsi pour le secteur de la chimie, il existe (entre autres...) France-Chimie, Chemindustry, Chem.com, Chemscope, Chemweb, ...

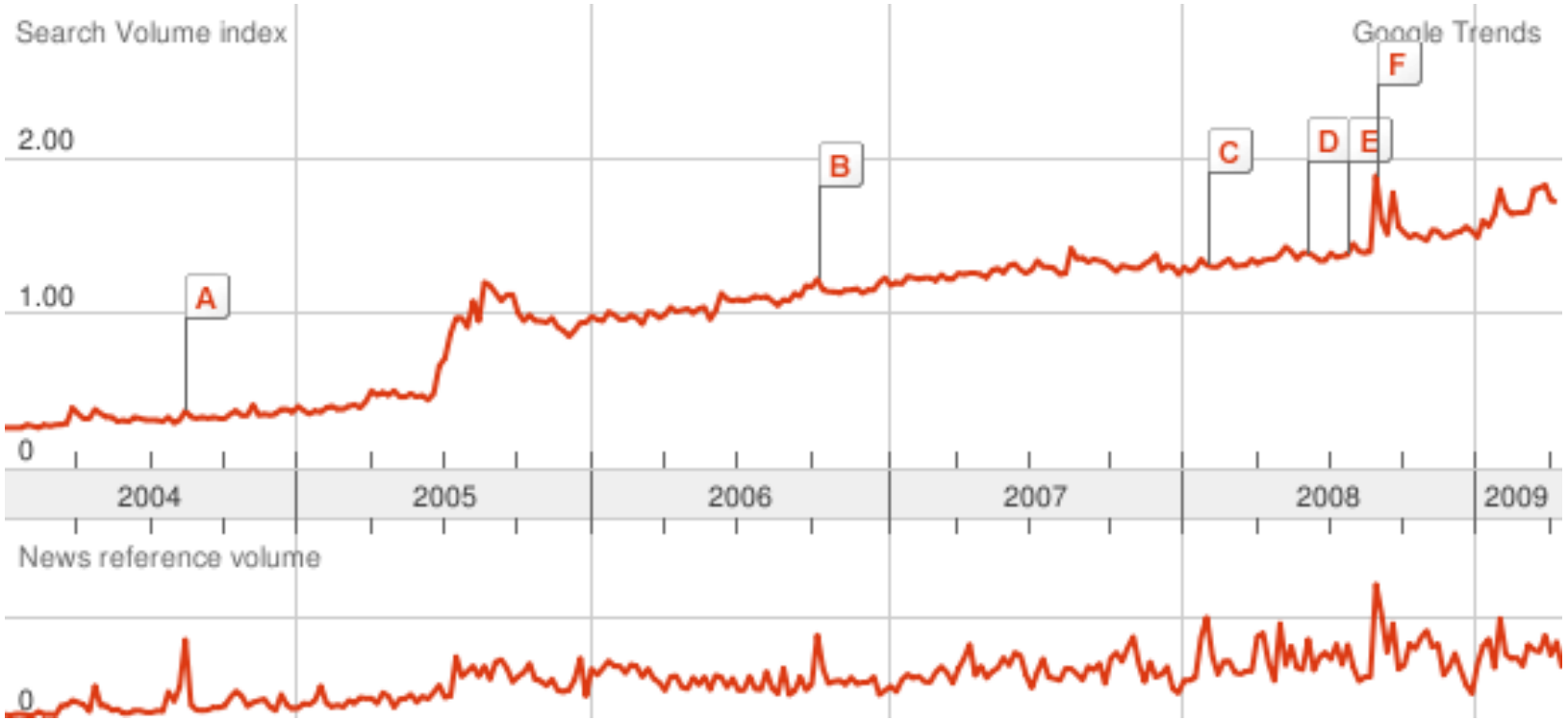
Web invisible VS Web visible:

web visible 0.04 vs **web invisible** 1.00

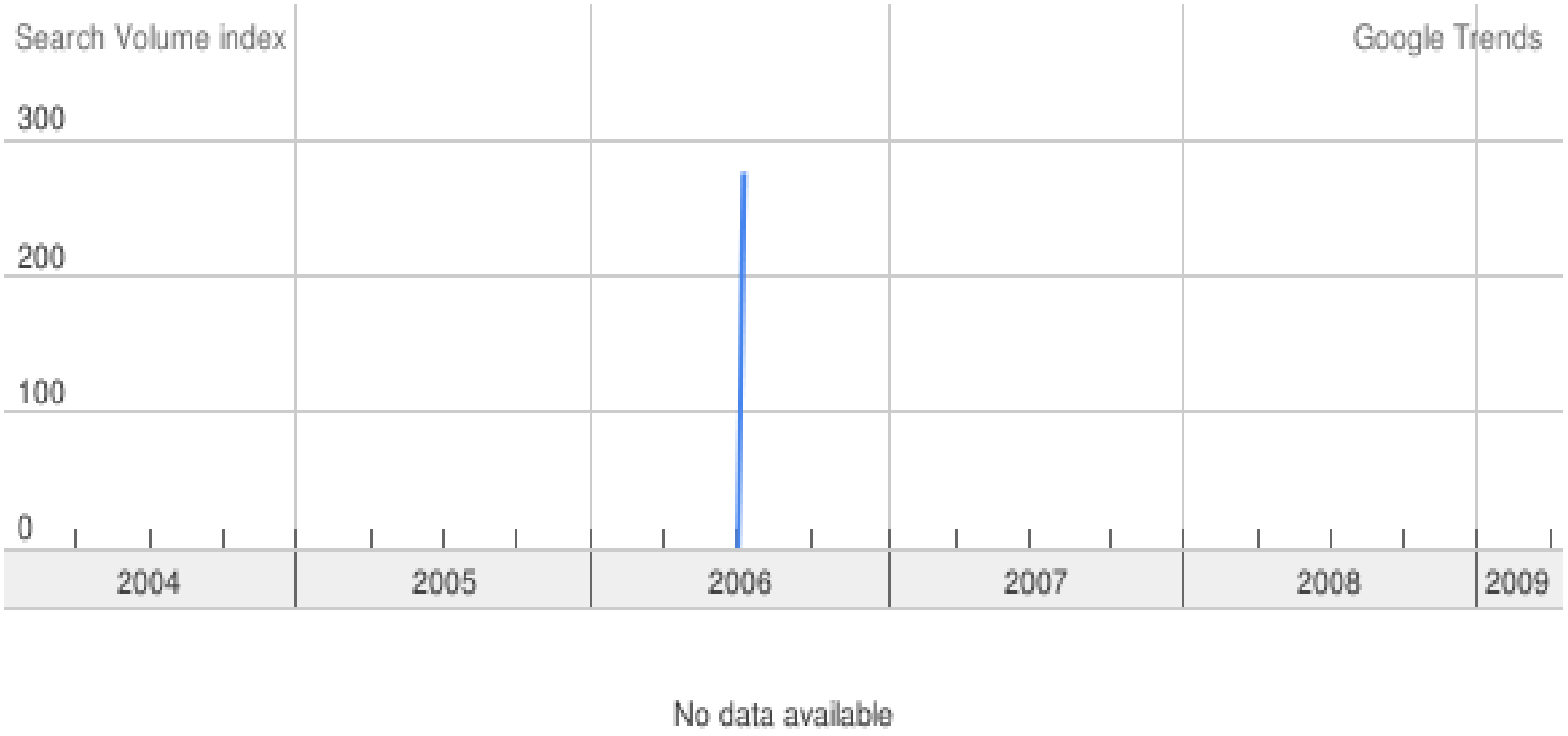
Source **: Google Trends**

Google Vs Turbo10:

turbo 10 0 vs **Google** 1.00



Wondir vs Turbo10



Merci pour votre attention