
A collection of historical artifacts is arranged on a light-colored surface. On the left, a portion of a chessboard with a checkered pattern and several chess pieces is visible. Below it, a blue ribbon with a circular emblem is attached to a large, ornate silver star-shaped medal. To the right, another similar star-shaped medal is shown. In the bottom left corner, a circular compass with a white face and black markings is visible. A pair of gold-rimmed glasses with thin temples lies across the center of the image, partially overlapping the medals and the compass.

**Recherche d'informations  
sur Internet :  
nouveaux outils,  
nouveaux usages.**

Véronique MESGUICH  
INFOTHEQUE  
POLE UNIVERSITAIRE  
LEONARD DE VINCI

7 mars 2006



# La recherche d'information sur Internet : un art plutôt qu'une science

Abondance de l'information

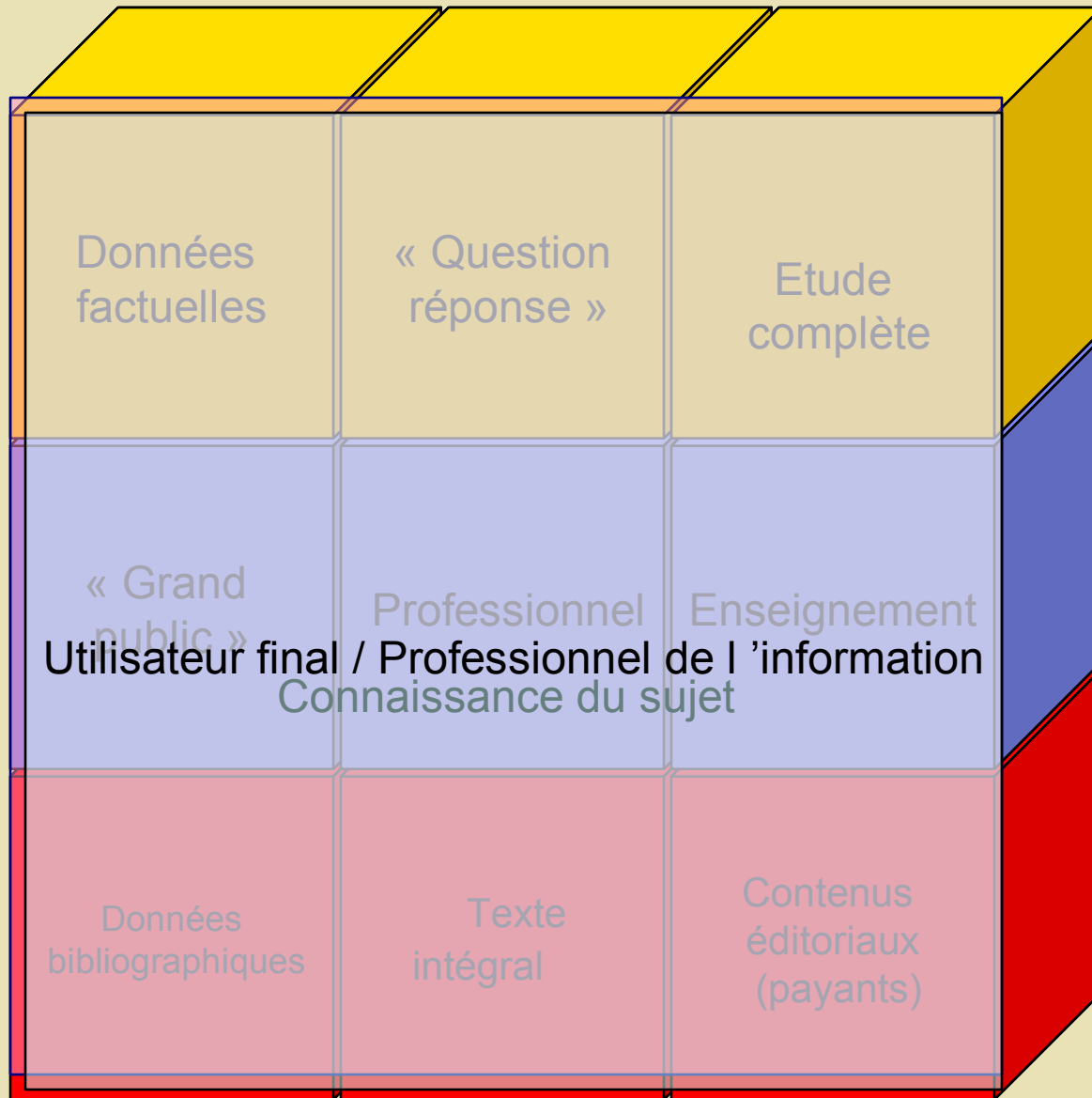
Hétérogénéité et fragmentation de l'information

Coexistence de contenus structurés et non structurés

Renouvellement continu

Multilinguisme

Internet, outil documentaire ou outil de communication ?






# Deux approches méthodologiques

L'approche « mots clés » : recherche par mots clés sur texte intégral des pages web. La qualité de la recherche dépendra du choix des mots clés : nombre de mots clés, degré de précision, langue, combinaison avec opérateurs booléens...

*Inconvénient : le manque d'exhaustivité des moteurs et méta-moteurs (« web invisible »)*

L'approche « exploration des sources » : identifier les sources d'information les plus pertinentes par rapport à la requête, utiliser ensuite les outils de recherche intégrés à ces sources, l'exploration de liens...

*Inconvénient : suppose une bonne connaissance des sources*



# Recherche d'information sur Internet : se méfier des idées reçues

Les moteurs de recherche, même les plus puissants, n'indexent qu'une partie du web (notion de pages dynamiques, « web invisible »)

Les moteurs de recherche n'indexent pas le web en temps réel et ne sont pas à jour

L'outil n'est pas tout : rechercher l'information « à la source » : portails spécialisés, portails géographiques...



# Les nouvelles tendances de la recherche d'information sur le web

Regroupement des acteurs. Simplification  
de la syntaxe

Représentation cartographique des  
résultats (Kartoo )

Thématisation et Génération de  
« thésaurus » dynamique (Exalead, Teoma,  
Vivisimo...)

Développement des portails verticaux  
(accès au web invisible)

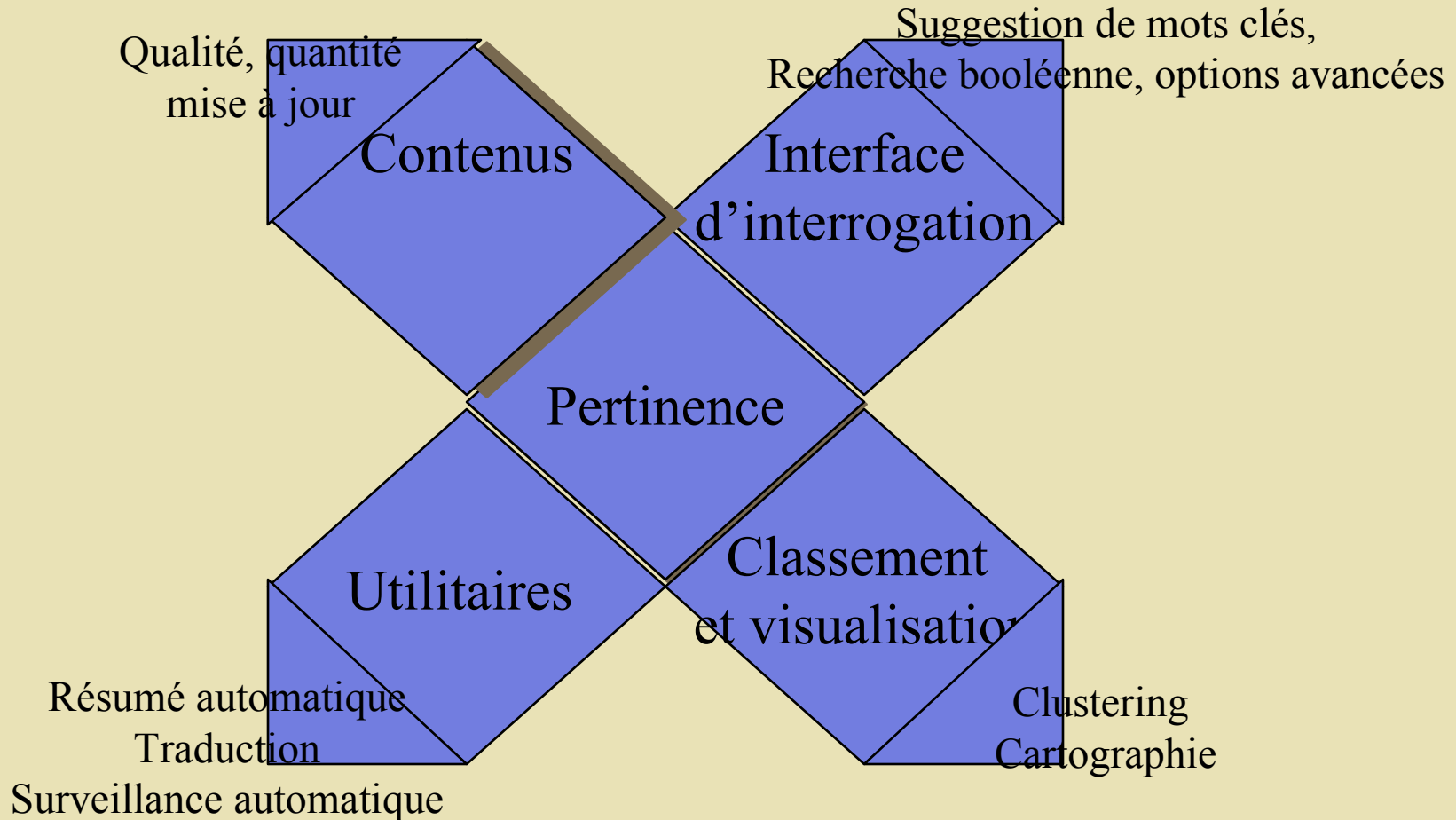
Régionalisation (Mirago)

Personnalisation (Yahoo, Google, Ujiko)

# 3 générations de moteurs de recherche

<b>1<sup>ère</sup> génération (apparus en 95-96)</b>	<b>2<sup>ème</sup> génération (apparus en 98-99)</b>	<b>3<sup>ème</sup> génération (apparus à partir de 2001)</b>
Altavista, Lycos, Hotbot, Excite	Google, Fast/Alltheweb, Yahoo Search Technology, Northern Light	Exalead, Wisenut, Teoma, Ujiko, Mozbot
« Vieillesse » de l'index. Algorithmes de pertinence pas toujours précis Orientation « grand public »	Index important Simplicité d'utilisation	Simplicité d'utilisation Nouvelles options : clustering, personnalisation...

# Portrait robot d'un moteur idéal...





# Les principaux critères de pertinence des moteurs

- Occurrence et densité des mots-clés
- Présence dans l'URL, dans le titre ou positionnement dans la page
- Proximité et ordre des mots-clés
- Taille et styles de polices
- Présence dans les méta-données (meta-keyword, meta-description)

*Critères « off the page » :*

- Indice de popularité (page rank)



# La troisième génération

Kartoo *www.kartoo.com* pour la  
représentation cartographique

Exalead : pour l'analyse statistique des  
**mots clés par occurrence**  
*www.exalead.com*

Mozbot : pour la **suggestion de mots clés**,  
les définitions et les archives des pages  
*www.mozbot.fr*

Teoma : pour la **catégorisation par  
dossiers** *www.teoma.com*



# La nouvelle donne...

L'exploration du web invisible

*Bases de données, sites peu ou mal indexés  
par les moteurs et méta-moteurs classiques*

Les portails spécialisés

Regroupement thématique de ressources :  
*annuaire de sites, newsletter, forum,  
agenda...*

Les weblogs et fils RSS

Le « web 2.0 » ou « web social »




# Web invisible

Pages non localisables et/ou non indexables par les moteurs de recherche web

Accéder au contenu de bases de données diversifiées

Exploiter le contenu des pages « à identification », ou « confidentielles »

Découvrir des pages peu ou mal indexées (isolées, ou d'un format « original »).




# Le web invisible : comment y accéder

Bonne connaissance des ressources. Veille sur un domaine (portails thématiques, listes de diffusion...)

Répertoires de « web invisible »

ex : [www.completeplanet.com](http://www.completeplanet.com)  
[www.invisible-web.net](http://www.invisible-web.net)

Méta-moteurs spécialisés



# Internet versus bases de données

Intérêt d 'Internet :

- . Multiplicité des sources d 'information
- . Interactivité
- . Couverture internationale

*A utiliser pour :*

- . *Actualité immédiate*
- . *Analyse sites des entreprises*
- . *Infos sur pays*
- . *Fédérations professionnelles - portails spécialisés*

Intérêt des bases de données :

- . Fiabilité de l 'information
- . Données à valeur ajoutée
- . Forme structurée

*A utiliser pour :*

- . *Archives de presse*
- . *Bilans entreprises*
- . *Etudes de marché*

# Connaissance des sources

Données factuelles	Encyclopédies dictionnaires Sites officiels
Données académiques : publications de chercheurs, cours en ligne, thèses	Moteurs spécialisés
Presse généraliste	Agrégateurs de presse Services d'actualité des moteurs
Description bibliographique d'ouvrages	Catalogues de bibliothèques Catalogues collectifs Catalogues d'éditeurs ou libraires en ligne
Revue spécialisées	Bases bibliographiques
Ouvrages en texte intégral	Bibliothèques numériques
Données commerciales	Sites d'entreprises
Infos financières	Fournisseurs d'infos financières
Données scientifiques et techniques	Brevets, rapports, prépublications, thèses, sites universitaires




# Méta-moteurs : quand les utiliser

Les méta-moteurs « on-line » (Ixquick, Profusion...) parfois trop aléatoires.  
Privilégier les unitermes.

Les méta-moteurs « clients » (Copernic)  
Plus de paramétrages mais souvent manque  
de finesse.

Certains méta-moteurs (Jux2, Releton...) comparent les résultats des  
« grands moteurs »



# Avantages et inconvénients des méta-moteurs

Permet de cumuler la puissance de plusieurs outils.

N'intègre pas la syntaxe de chaque moteur,

Récupère un nombre limité de résultat par moteur (10 premiers résultats de chaque outil)

Réponses pas toujours pertinentes.

**Usage** : pour une recherche large de premier niveau.



# Une tendance : les méta-moteurs spécialisés

Recherche simultanée sur des corpus spécialisés (web invisible). Mélange d'outil humain et automatique

Les rubriques spécialisées de Profusion ou Copernic Pro

Les méta-moteurs spécialisés (le problème des « liens profonds »):

Emploi *[www.keljob.com](http://www.keljob.com)*

*Recherche sur des sites prédéfinis :*

*[www.goshme.com](http://www.goshme.com)*

Des outils personnalisables : Rollyo

*[www.rollyo.com](http://www.rollyo.com)*



# De nouveaux types d'annuaires

Les annuaires « contributifs » ou « ouverts »

ex : Open Directory [www.dmoz.fr](http://www.dmoz.fr)

Les annuaires « professionnels »

ex : Indexa [www.indexa.fr/](http://www.indexa.fr/)

Les annuaires de portails

ex : Mediaveille [www.mediaveille.com/outil/](http://www.mediaveille.com/outil/)

[outil.htm](http://outil.htm) Objectif Grandes écoles

[www.objectifgrandesecoles.com](http://www.objectifgrandesecoles.com)

# Les techniques spécifiques utilisables pour la recherche de sources

Trouver des listes de liens

Trouver des sites  
« pointant » sur  
une source déjà  
connue

Trouver des  
portails / sites  
fédérateurs

Trouver des sites « similaires » à  
une source connue





# Identifier des portails spécialisés

Attention à l'exhaustivité et à la mise à jour

Répertoires ouverts (*dmoz*)

Répertoires d'outils de recherche (*enfin,, beaucoup, mediaveille, Objectifs grandes écoles...*)

Sites d'associations professionnelles, sites de référence

Recherche par mots clés sur moteurs




# L'évaluation des sites web

Identifier l'origine d'un site (Alexa)

Identifier la date de dernière mise à jour d'une page

Remonter dans le temps : *www.archive.org*

Identifier un nom de domaine : les annuaires WHOIS (*www.indomco.com*)



# Le phénomène des weblogs : un outil pour la veille ?

« Journal en ligne » sur internet ou intranet tenu par une ou plusieurs personnes.

Possibilité d'insérer des liens, noter ses commentaires, ses points de vue, ses activités...

Suivi de l'évolution d'une idée, d'un thème de projet



# Les fils RSS

Un fil RSS (RDF Site Summary ou Really simple syndication) est un format de transmission de données fondé sur le langage XML qui permet de décrire les nouveautés mises en ligne sur un site et de les transmettre sous forme de flux d'information

Abonnement gratuit aux fils RSS via :

- . Un lecteur RSS à télécharger (voir panorama ADBS)
- . Les options personnalisées Yahoo, Google, MSN
- . Marque page dynamique dans Firefox

Agrégateurs de fils RSS presse francophone : *Alertinfo*  
*www.geste.fr*



# Les agents d 'alerte

Signalent les modifications à l 'intérieur d 'une page

Agents d 'alerte « on line »

ex : *www.infominder.com*

Agents d 'alerte « clients »

ex : Kbcrawl *www.kbcrawl.com*

*www.websitewatcher.com*

Parfois, aspirateurs et agents d 'alerte

ex : Wysigot *www.wysigot.com*




# Agents d'alerte : fonctions avancées

Des critères de modification avancés permettent de limiter les alertes non pertinentes.

Critères de modification avancés (nombre de phrases modifiées, lien modifié, pourcentage de contenu modifié, images, page disparue)

Extraction des modifications de données



# Automatiser une requête récurrente avec Google

**Google alert** : permet d'automatiser jusqu'à 5 requêtes récurrentes. Possibilité de recherche avancée. Envoi des résultats par mail.

*[www.googlealert.com](http://www.googlealert.com)*

**Google newsalert** : veille sur l'actualité

*[www.google.fr/newsalerts](http://www.google.fr/newsalerts)*

*Les deux types d'alertes ont été fusionnées  
en sep 04*



# Les aspirateurs de sites

Capture d'un site complet et consultation off line

Permet l'archivage d'anciennes versions d'un site

Exemples : **Memoweb**

*[www.memoweb.com](http://www.memoweb.com),*

**Wysigot** *[www.wysigot.com](http://www.wysigot.com)...*



# En guise de conclusion...

## les 10 règles d'or

Savoir questionner

Savoir utiliser les outils de navigation et de recherche

Savoir choisir les bons mots-clés

Savoir sélectionner les bons points de repère

Savoir analyser

Savoir poser des balises

Savoir se limiter dans le temps

Savoir rester clair sur ses objectifs

Savoir conjuguer recherche outils et navigation

Savoir être agile et « rebondir »