# PROBLEMATIQUE GENERALE DE LA RECHERCHE D'INFORMATION



---

### Notions de base et typologies de la recherche d'information

Pour la recherche d'information, nécessaire de bien connaître :
- les types d'informations : **primaire et secondaire**
- la nature des informations recherchées
- les modes de recherche
- le contexte de la recherche
- les principes de base de la recherche d'information
- les sources
- les outils de recherche

---

### Questions de terminologie

- **Recherche documentaire :**

Définition officielle :

- **Action, méthodes et procédures ayant pour objet de retrouver dans des fonds documentaires les références des documents pertinents (Vocabulaire de la documentation, AFNOR, 1987).**
  **Ensemble des techniques et modalités permettant de sélectionner l'information dans un fonds documentaire structuré en fonction de critères de recherches propres à l'utilisateur.**

- **Recherche Documentaire Informatisée :**
  (ou RD Automatisée)
  Recherche documentaire utilisant un logiciel documentaire, sur ordinateur ; implique l'élaboration d'une démarche, d'une stratégie de recherche :
  - définition des mots-clés, des clés d'accès...
  - élaboration d'une équation de recherche...
  >> RDI plutôt assimilée à l'interrogation des banques de données.

- **Recherche de l'information :**

- Définition officielle :
  - **Action, méthodes et procédures ayant pour objet d'extraire d'un ensemble de documents les informations voulues (d'après l'AFNOR, 1979). Dans un sens plus large, toute opération (ou ensemble d'opérations) ayant pour objet la recherche, la collecte et l'exploitation d'informations en réponse à une question sur un sujet précis.**

**Quels problèmes aujourd'hui pour ces définitions ?**
Toutes ces définitions sont bousculées par Internet, qui permet à la fois :
- de rechercher des références de documents : recherche bibliographique ou recherche documentaire classique ("secondaire")
- de rechercher des documents entiers, sous forme numérique : recherche de type "primaire"
- de rechercher des informations (exploitation des documents)

>> ce qui explique que l'on parle surtout de **recherche de l'information sur Internet,** pour signifier la numérisation du document et la primauté accordée désormais à l'information sur le document, au contenu sur le contenant

## Typologies des recherches d'informations

Il existe une grande diversité de modalités, de stratégies et de types de recherches d'informations : recherches en accès direct, recherches hypertextuelles, recherches par requêtes, recherches multicritères, recherche factuelle, etc. On peut distinguer les recherches d'information notamment selon :
- **la nature des informations** recherchées : qu'est-ce qui est recherché ?
- **les modalités de la recherche**, les modes d'accès : comment rechercher ?

### A/ Typologie selon la nature des informations recherchées :

- **Recherche factuelle :**
  Recherche d'informations précises, en réponse à une question précise. Recherche d'informations structurées, numériques ou alphanumériques, généralement contenues dans des banques de données, des annuaires, des répertoires ; ou bien recherche d'informations non structurées, dans une base textuelle.

- **Recherche documentaire :**
  recherche de documents à partir de leurs références ; recherche d'informations secondaires, aboutissant au document primaire.

- **Recherche bibliographique :**
  sous-ensemble de la recherche documentaire, limitée aux références bibliographiques des documents ou aux adresses des sites web.

- **Recherche contextuelle sur le contenu :**
  recherche à partir d'un mot ou d'un groupe de mots, pour

aboutir à un texte (à l'information primaire). Recherche en texte intégral.

## B/ Typologie selon les différents modes de recherche :

Cette typologie est plus importante que la précédente, car les modes d'accès induisent des compétences, des comportements, des pratiques et des résultats très différenciés. Par ailleurs, la recherche d'information sur Internet permet aujourd'hui de combiner tous les modes de recherche, auparavant très distincts, voire exclusifs les uns des autres.

Le tableau ci-dessous cherche à présenter une vue d'ensemble de ces différents modes de recherche, résumés ici à quatre grandes modalités, correspondant chacune à des principes de recherche différents, à des figures ou des métaphores du savoir distinctes, à des "gestes cognitifs" ou des démarches intellectuelles également distinctes.

### LES QUATRE PRINCIPAUX MODES DE RECHERCHES D'INFORMATIONS

| ***Modes de recherche*** | ***Principe, démarche intellectuelles*** | ***Type d'information concernée en priorité*** | ***Exemples d'outils*** |
|---|---|---|---|
| **Recherche par navigation arborescente** | **Figure de l'Arbre**<br><br>Démarche **systématique**, du général au particulier<br><br>Recherche **par menus successifs** | **Information structurée**, organisée en plan de classement<br><br>Information **secondaire** | **Tables des matières des livres**<br><br>**Classifications documentaires** (CDU, Dewey)<br><br>**Annuaires thématiques du web** (Yahoo, Nomade…)<br><br>**Page d'accueil d'un site web** |
| **Recherche par** | **Figure du Réseau** | **Information non** | **Renvois** dans une |

| | | | |
|---|---|---|---|
| **navigation hypertextuelle** | Démarche **associative**, d'une notion à l'autre.<br><br>Navigation dans un réseau de noeuds et de liens | **structurée**<br><br>Information **primaire**, brute | encyclopédie<br><br>**Hypertextes** sur CD-ROM<br><br>**Sites web** |
| **Recherche par requête sur les métadonnées du document** | **Principe de l'Index**<br><br>Démarche **d'indexation** de l'information<br><br>Recherche **par champs, logique booléenne** | Information **structurée en champs**.<br><br>Information **secondaire** | **Index des livres**<br><br>**Banques de données**<br><br>**Catalogues de bibliothèques** |
| **Recherche par requête sur le texte intégral** | **Texte**<br><br>Démarche **d'analyse linguistique**<br><br>Recherche **contextuelle sur le contenu** | **Information non structurée**<br><br>Information brute, **primaire** | Outils de **TALN** (***Traitement Automatique du Langage Naturel***)<br><br>**Moteurs de recherche**<br><br>**Outils linguistiques** |

Pour simplifier, **deux grandes méthodes** de recherche documentaire, actuellement juxtaposées :

- **la recherche dans les bases de données, c.a.d. dans des champs structurés :**
  > recherche booléenne, par mots-clés
  **la "RDI" (Recherche Documentaire Informatisée) classique**

- **la recherche dans les corpus de textes, dans le texte intégral :**
  > recherche linguistique
  **la recherche d'information sur Internet**

Mutations accélérées des méthodes de recherche et d'accès avec Internet : essor des recherches linguistiques.

## Indexation humaine / indexation automatisée

Pour l'analyse et la description des documents, deux grandes réponses apportées par la documentation :
- **l'indexation par un langage documentaire**
**- l'indexation automatisée du texte intégral**

- **Langage documentaire** :
  langage artificiel élaboré pour éliminer les problèmes d'ambiguïté du langage naturel. Un langage documentaire vise à fournir une représentation univoque et formalisée des documents.

  Exemples de langages documentaires : classifications décimales (Dewey, CDU...), listes de vedettes-matières (RAMEAU), thésaurus.

- **Indexation automatisée :**
  traitement linguistique du texte intégral.

## Notions sur l'indexation automatisée

L'indexation automatisée repose sur les techniques de **TALN** :
Traitement Automatique du Langage Naturel.
Au fondement de la recherche d'information sur Internet, notamment dans les moteurs de recherche.

Plusieurs niveaux d'analyse du texte intégral :

- **niveau morphologique** : reconnaissance du mot
- **niveau lexical** : réduction du mot à sa forme canonique > lemmatisation
- **niveau syntaxique** : niveau d'utilisation de la grammaire
- **niveau sémantique** : niveau de la reconnaissance des concepts

**Sur Internet : l'indexation morphologique est généralement le seul niveau d'indexation utilisé par les moteurs de recherche. La**

**plupart des moteurs n'éliminent pas les mots-vides et prennent en compte tous les mots.**

- **Indexation en texte intégral ou plein texte (full text) :**

  - Fondée sur l'analyse morphologique des mots : leur **forme**.

  - Méthode fondée sur **l'élimination des mots-vides** (à partir d'un dictionnaire de termes) et la **constitution d'un index des termes non éliminés**, considérés comme des chaînes de caractères.

  - Recherche se fait selon logique booléenne ou avec des opérateurs de proximité ; méthode de recherche en full text quotidiennement utilisée dans les banques de données, pour les recherches portant sur les chaînes de caractères, dans les résumés par exemple.

    Exemple : dans la phrase "*Prolétaires de tous les pays : unissez-vous",* l'indexation en full text éliminera les termes : de, tous, les, vous, et gardera "prolétaires", "pays" et "unissez"
    A la recherche, il suffira de taper l'un de ces termes, ou une combinaison des termes, pour retrouver la phrase.

    **Quels problèmes et quels défauts de l'indexation morphologique ?**
    Inconvénients évidents de cette méthode très fruste d'indexation :
    - tous les mots non vides mis sur le même plan :
    - pas de prise en compte de l'ordre des mots
    - apparition des différentes formes d'un mot : par ex. un verbe va apparaître plusieurs fois sous des formes différentes
    - l'analyse porte seulement sur des mots isolés (des unitermes), et délaisse toutes les expressions (les syntagmes), souvent porteurs de sens :
      - ex. : pomme de terre donnera deux mots "pomme" et "terre", analysés séparément
    - polysémie, synonymie du langage naturel pas prise en compte :
    > vol = aussi bien vol d'avion que vol à la tire

- stricte équivalence entre termes de recherche et termes indexés : pas de faute de frappe…

>> limites très sérieuses de l'analyse morphologique, qui peut générer beaucoup de « bruit ou de silence documentaire »

---

## Principes de la recherche d'information

Deux notions transversales de la recherche d'information (surtout dans la "RDI") :

**le bruit et le silence documentaire :**

- **bruit** : **documents retrouvés non pertinents**
- **silence** : **documents pertinents non retrouvés**

> ces deux indices définissent la **pertinence** (*relevance*) d'un système documentaire :

---

## Notions de base de la RDI

- **Structuration de l'information :**

  - une notice est découpée en **champs**
  - les champs sont **interrogeables** (indexés) ou **non interrogeables** (non indexés)
  - les champs indexés sont interrogeables par des **mots-clés**
  - pour l'interrogation d'une base de données, utilisation d'opérateurs de recherche

- **Notion de zones, champs ou rubriques :**

  Toutes les informations présentes dans un document sont regroupées dans des **champs** :
  - appelés aussi zones ou rubriques
  - chaque élément de description sera isolé dans un espace précis, identifiable : un champ.

Ex. : champ Auteur, Titre, Editeur...
Ensemble des champs définis constitue la « structure » de la base de données, ou du moins de l'enregistrement, de la notice.

- **Notice, enregistrement :**

  **Notice ou enregistrement, ou fiche, ou référence = champs + données**

- **Les clés de recherche : les mots-clés**

  Une fois saisies les informations, le problème posé est celui de leur recherche, de leur accès.
  Deux possibilités :
  - **accès séquentiel** (par balayage : l'ordinateur devrait lire toutes les notices une par une...) : > lenteur inacceptable !
  - **accès direct** : spécificité des systèmes de recherche documentaire, possibilité d'accéder directement à une notice, par l'intermédiaire d'une clé de recherche, contenue dans un index
  > notion de « **mot-clé** » (*keyword*)

  Tous les mots-clés sont regroupés dans des fichiers séparés des notices : les **fichiers index** (parfois appelés lexiques selon les logiciels)
  Ex. : index des auteurs, des titres, des dates de publication...

---

**Les opérateurs de recherche (booléens, troncature, proximité…)**

**TABLEAU DES OPERATEURS DE RECHERCHE**

| *Catégories d'opérateurs* | *Définition* | *Symboles* | *Types* | *Fonction* | *Exemples* |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **roncatur(**<br><br>**Wildcards, oker** | Substitution d'un symbole à des caractères | • *<br>• +<br>• ? | • **Troncature à droite** | Recherche sur tous les mots contenant la même racine ou le même préfixe | franco* : >> francophone, francophonie, francophobe |
| | | | • **Troncature à gauche** | Recherche à partir d'un suffixe | *phobe : >> technophobe, agoraphobe, etc. |
| | | • ?<br>• # | • **Masque ou troncature centrale** | Remplace un ou plusieurs caractères dans un mot. | francopho?e > francophobe et francophone |
| **pérateurs ooléen** | Logique booléenne, permettant des relations entre les termes de la recherche | • **ET**<br>• **AND**<br>• + | • **Opérateur d'intersection** | Relie deux mots-clés se trouvant ensemble dans le même document | Pêche ET Bretagne :<br><br>>> les documents traitant de la pêche en Bretagne |
| | | • **OU**<br>• **OR** | • **Opérateur d'union** | Relie deux mots-clés dont l'un ou l'autre, ou les deux doivent se trouver dans le document | Pêche OU Bretagne :<br><br>>> les documents traitant ou bien de pêche ou bien de Bretagne, ou des deux à la fois |
| | | • **SAUF**<br>• **NOT**<br>• **-** | • **Opérateur d'exclusion** | Relie deux mots-clés dont le premier doit être présent et le second absent dans un document | Pêche SAUF Bretagne :<br><br>>> les documents traitant de la pêche sauf en Bretagne |
| **pérateurs umérique** | Permettent des recherches selon **des critères quantitatifs** | - égal : =<br><br>- supérieur : > | | Recherche sur les dates, par exemple | • = 1995 : en 1995<br><br>• > 1995 : depuis |

| | | | | | |
|---|---|---|---|---|---|
| | | - supérieur ou égal : >=<br><br>- inférieur : <<br><br>- inférieur ou égal : <=<br><br>- intervalle entre deux dates : : | | | 1995<br>• >= 1995 : depuis ou en 1995<br>• <= 1995 : avant ou en 1995<br>• 1991:1995 : de 1991 à 1995 |
| **pérateurs de roximit** | Permettent des recherches sur le texte intégral selon la proximité des termes | • **ADJ** | • **Opérateur d'adjacence** | Recherche sur des termes adjacents, dans l'ordre donné | Fibre ADJ optique :<br><br>>> texte contenant l'expression " fibre optique " |
| | | • **nAV** | • **Opérateur de distance** | Recherche sur des termes séparés par une distance *n* | Ecole 1AV privée :<br><br>>> école primaire privée, ou école technique privée |
| | | • **NEAR**<br>• **M (pour Mot) ou**<br>• **W (Word)** | • **Opérateur de présence** | Recherche sur des termes présents dans le texte, quelle que soit leur distance | Fibre NEAR optique<br><br>>> texte contenant les deux termes, mêmes séparés |

A noter : **l'universalité des opérateurs de recherche**, présents (partiellement ou en totalité) dans la totalité des systèmes de recherche d'information : catalogues de bibliothèques, bases et banques de données,

moteurs de recherche, outils d'indexation du texte intégral, annuaires thématiques...

---

**Différences entre la recherche sur Internet et la RDI "classique"**

| *Critères* | *Systèmes documentaires classiques* | *Internet* |
|---|---|---|
| **Caractéristiques générales de l'information** | **Homogénéité : des sources, des informations, des domaines, de la description, de la présentation, de la recherche...** | **Hétérogénéité ...** |
| **Nature de l'information** | **Informations secondaires** | **Informations primaires mélangées aux informations secondaires** |
| **Structuration de l'information** | **Information structurée, découpée en champs** | **Information plutôt non structurée. Utilisation réduite des métadonnées** |
| **Validité de l'information** | **Informations doublement validées : sources connues, responsabilité du professionnel de la documentation** | **Informations non validées : sources non connues, qualification sous la responsabilité de l'utilisateur** |
| **Cohérence de l'indexation** | **Indexation par un langage documentaire homogène, adapté à la base et au domaine** | **Pas de langage d'indexation.** |
| **Etendue de l'indexation** | **Indexation de la totalité de la base** | **Indexation limitée du web visible (30 %), web "invisible" non indexé...** |

| | | |
|---|---|---|
| **Niveau de l'indexation** | **Indexation humaine (niveau sémantique) ou automatisée (plusieurs niveaux)** | **Indexation automatisée, limitée souvent au plus bas niveau (analyse morphologique) ; mais développement de l'indexation lexicale et syntaxique.** |
| **Représentation des documents** | **Normalisée** : notices textuelles, de description bibliographique.<br>Format unique de représentation des informations | **Non normalisée :** multiplicité des formats, liberté des auteurs...<br>Exemple des 3 formes possibles d'un article :<br>- une seule page HTML<br>- plusieurs pages HTML<br>- " chapeau " + document en format PDF (ou DOC) |
| **Information accessible en recherche** | **Notices** structurées, portant sur les différentes caractéristiques des documents (auteurs, titre, mots-clés...), et normalisées selon le système documentaire | Dans les annuaires : **Adresses de sites**.<br><br>Dans les moteurs : **texte des pages HTML, critères de date** |
| **Volume d'information** | **Limité** (qqs centaines de milliers de documents), maîtrisé par le responsable du système.<br>Croissance linéaire de la base | **Très grande quantité d'informations ; volume redondant, non maîtrisé ;** pas de filtrage.<br>Croissance exponentielle du web |
| **Enrichissement de l'information** | Ajout de documents, mais **stabilité du contenu et des notices des documents** | Information **très changeante**, modifications constantes du contenu et de l'organisation des sites |
| **Responsabilité de l'alimentation du système** | **Base documentaire alimentée par les professionnels de l'information : souci de cohérence de l'information** | **Sites web alimentés par les webmasters : souci de présentation, de diffusion et d'impact de l'information** |
| **Outils de recherche** | **Outil d'interrogation unique (logiciel documentaire),** | **Diversité des outils de recherche, à adapter selon** |

| | | |
|---|---|---|
| | **adapté à la base et au langage documentaire.** | **les types de recherche. Mélange de tous les modes de recherche.** |
| **Compétences ou connaissances requises pour l'utilisateur** | Connaissance minimale :<br>- **du domaine documentaire**<br>**- de la struture des notices, des mots-clés** (et du langage d'indexation utilisé) | Connaissance :<br>- **des sites pertinents**<br>> l'offre<br>- **des outils de recherche**<br>**- des méthodes de recherche** |
| **Qualité de la recherche** | Pour une recherche par mots-clés :<br><br>- dépend directement de la **qualité de l'indexation manuelle**<br><br>- **aucune intelligence de l'outil informatique** : retrouve ***exactement*** les notices contenant les mots-clés demandés | Qualité de recherche : **aléatoire**.<br>Dépend :<br>- **de la couverture du domaine** par des sites accessibles en recherche (> inégalité des domaines couverts sur Internet)<br>- **de l'expertise de l'utilisateur**<br>- **des performances de l'outil**<br><br>**Intelligence distribuée dans les outils** ; recherche retourne la question qui ***ressemble*** le plus à la question posée<br>> recherche par approximations successives...<br>Notion de "**serendipité**" ou "*découverte, par chance ou sagacité, de résultats que l'on ne cherchait pas*". |
| **Multilinguisme** | **Pas à la charge de l'utilisateur** : mots-clés dans les bdd en une ou deux langues, même si documents en plusieurs langues.<br>> pas de prise en compte de la langue des documents dans la recherche | **A la charge de l'utilisateur**:<br>- choix de la langue des documents dans le moteur, tri... |

**Date de dernière mise à jour :**
**31 janvier 2004**