Chapitre 7

Terminologie et accès à l'information en santé

1 Introduction

Un système de Recherche d'Information (SRI) crée une interface entre une base documentaire, ou collection de documents, et des utilisateurs qui recherchent des informations contenues dans cette base. La figure 7.1 illustre l'architecture d'un tel système, et met en évidence le rôle central du langage d'indexation utilisé dans un SRI. En effet, c'est l'élément qui conditionne par la suite la description des documents de la base de connaissance lors de l'indexation et la méthode d'interrogation de la base par les utilisateurs en quête d'information lors de la rédaction d'une requête ou de l'élaboration d'un algorithme de recherche à implanter dans le moteur de recherche. À ce titre, le choix d'un langage d'indexation doit être murement réfléchi et motivé par le contexte particulier du SRI considéré.

L'avancée des connaissances dans le domaine de la santé dépasse désormais les capacités de la mémoire humaine. De fait, un large public, y compris les professionnels de santé, rencontre des besoins d'information en santé fréquents et diversifiés. Pour trouver des réponses, la littérature concernant un point donné peut-être dispersée, et difficile à synthétiser. Par ailleurs, l'Internet est devenu une source incontournable d'information dans tous les domaines. Pour toutes ces raisons, il est pertinent, et même nécessaire, de développer des SRI spécifiques au domaine médical.

Chapitre rédigé par Aurélie NÉVÉOL et Stéfan DARMONI.

Les informations stockées dans les bases de données médicales peuvent être classées en deux catégories : il s'agit soit d'informations spécifiques aux patients (telles que les données personnelles figurant dans le Dossier Electronique du Patient) soit d'informations génériques concernant les connaissances médicales. A l'heure actuelle, les Systèmes de Recherche d'Information en santé à la disposition du public, comme PubMed®[1] pour l'anglais ou Doc'CISMeF[2] pour le français, s'intéressent en premier lieu aux informations génériques. On peut cependant remarquer que la loi récente sur le droit d'accès des patients à leur dossier électronique et l'utilisation accrue de documents électroniques dans les dossiers, en particulier pour ce qui est de l'imagerie médicale, laisse présager une croissance de l'intérêt pour des SRI dédiés ou même des SRI couplant données personnelles et données génériques [CIM 03].

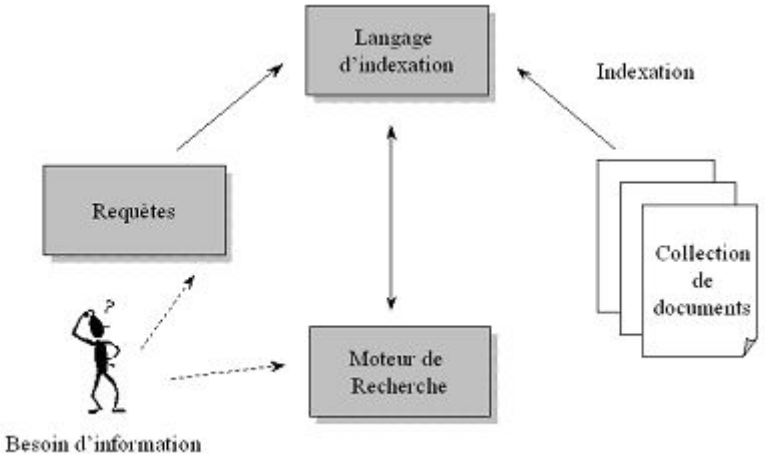

Figure 7.1. *Architecture d'un Système de Recherche d'Information*

De nombreuses ressources terminologiques sont disponibles dans le domaine de la santé et peuvent intervenir à plusieurs niveaux dans le cadre des SRI : au moment de la construction du système, lors du choix d'un langage d'indexation, lors de l'évaluation du système par l'utilisation de mesures sémantiques ou encore pour faciliter l'interopérabilité du système et étendre l'accès initial à l'information. Nous commencerons ce chapitre par une réflexion sur les critères de choix d'un langage d'indexation pour un SRI. Puis, nous nous intéresserons à l'évaluation des SRI, et en particulier à l'évaluation de l'indexation des documents à l'intérieur d'un SRI.

[1] http://www.pubmed.gov
[2] http://www.cismef.org

Finalement, nous présenterons quelques défis de la Recherche d'Information que les ressources terminologiques permettent de relever.

2 Choix d'un langage d'indexation

2.1 Définitions

Les différentes définitions[3] de l'activité d'indexation s'accordent pour décrire cette tâche comme un *travail d'analyse* consistant à *extraire des concepts* d'un document, puis à *traduire* ces concepts en utilisant un « langage documentaire »[4]. Il est intéressant de remarquer que cette dénomination particulière (utilisée dans la norme NF Z 47-102) semble exclure d'office le langage naturel pour cet usage. Une divergence plus significative émerge au niveau de la présentation des résultats de l'indexation. Dans le cas des index de fin de livre, il s'agit d'une liste très détaillée comprenant des références aux passages où les concepts de l'index apparaissent dans le document. Dans le cas des SRI, que nous abordons dans ce chapitre, il s'agit d'une liste synoptique composée de quelques mots clés.

2.2 Langages d'indexation

Le rôle des unités descriptives attribuées à un document lors de la phase d'indexation est double [SAL 83]. Il doit à la fois être descriptif (c'est-à-dire représentatif du contenu du document) et discriminant (c'est-à-dire qu'il doit mettre en évidence ce qui différencie le document des autres documents de la collection). Cette idée est également reprise par [SOE 94] qui distingue deux approches de l'indexation : l'une recentrée sur le document («entity-oriented») qui cherche à en faire une description précise et l'autre centrée sur la recherche d'information («request-oriented»), qui cherche à employer le vocabulaire des utilisateurs du système pour l'indexation, afin de faire ressortir le document à l'intérieur de la collection. Ces approches ne sont pas incompatibles, mais elles peuvent jouer sur le choix des unités descriptives utilisées.

De nombreuses unités descriptives peuvent être utilisées pour la représentation d'un document. Avant de détailler ces unités, remarquons qu'il existe deux types d'indexation, utilisant un langage différent :

- l'indexation libre utilise à loisir tous les mots d'une langue naturelle donnée, voire même des groupes de caractères appelés n-grammes [HAL 97]. Ce type

[3] Voir par exemple le Trésor de la Langue Française ou la norme NF Z 47-102.

[4] [JAC 97] définit les langages documentaires comme une liste contrôlée de termes d'indexation ayant fait l'objet d'une validation humaine.

d'indexation est notamment utilisé par les moteurs de recherche procédant à une indexation entièrement automatique comme Google. Dans ce cas, l'index d'une ressource est une liste de tous les mots (définis ici comme des suites de caractères séparés par un espace ou un signe de ponctuation) contenus dans le document, après filtrage ou normalisation [SAL 83]. Dans l'indexation libre, l'ensemble des unités descriptives qui peuvent être utilisés n'est pas connu.

- l'indexation contrôlée utilise de manière contrainte les entités répertoriées dans une liste pré-définie. Le nombre de termes d'indexation susceptibles d'être utilisés est connu, il s'agit des termes contenus dans la liste de référence (terminologie) choisie. Cette terminologie définit également la forme des termes d'indexation utilisés. Il peut s'agir de termes ou d'expressions de la langue naturelle, de termes d'un méta-langage dit « langage documentaire », employé pour souligner le caractère normatif attribué au descripteur, ou bien de symboles choisis pour représenter un concept de manière normative et unique.

On peut dire qu'indexation libre et indexation contrôlée se distinguent par la connaissance *a priori* ou non des unités descriptives à utiliser. En pratique, les termes contrôlés ne sont pas nécessairement observables directement dans les textes à indexer. L'utilisation d'unités descriptives relevant de l'indexation libre peut s'avérer une étape intermédiaire nécessaire.

2.3 Incidence de la représentation

Après une brève description de quelques unes des unités les plus couramment employées dans un SRI, nous illustrerons l'incidence des différentes représentations sur la recherche d'information dans un système utilisant les langages d'indexation présentés grâce à l'indexation de deux énoncés.

- mots : groupes de caractères séparés par un espace ou un signe de ponctuation. Ainsi, « mal de tête » comporte trois mots et « céphalée » un seul.

- mots formes : le découpage prend en compte du sens des unités de la langue. Ainsi, « mal de tête » et « céphalée » correspondent à deux mots formes distincts.

- termes : expression normalisée d'un concept dans un domaine de spécialité. Ainsi, dans le domaine médical, « mal de tête » et « céphalée » correspondent à un seul et même concept.

- racines : forme primitive d'où dérivent les mots d'une même famille. Ainsi, « mal » a pour racine *mal*. Il faut remarquer que ce type d'unité peut s'avérer ambigu. En effet, « male» – qui n'est pas de la même famille que « mal » a également pour racine *mal*.

- lemmes : forme standard à laquelle ramener les formes fléchies des unités de la langue (infinitif, masculin singulier, etc.). Ainsi, « mal » et « maux » peuvent être ramenés à la forme *mal*.

- n-grammes : groupes de n caractères (ou même n mots) utilisés pour créer des modèles statistiques de la langue.

REMARQUE – Seuls les langages d'indexation issus d'une représentation par n-grammes ou par termes peuvent donner lieu à une indexation contrôlée. Les mots ou mots formes en tant qu'éléments constitutifs d'une langue sont indénombrables.

Afin d'illustrer les langages d'indexation contenant les unités présentées ci-dessus, considérons l'énoncé 1 ci-dessous, et ses représentations, indiquées dans le tableau 7.1.

Le diabète de type 1 représente 20% des cas de diabète sucré. (énoncé 1)

Unité descriptive	Représentation
Tri-grammes	Le ;dia;bèt;e d;e t;ype; 1 ;rep;rés;ent;e 2;0% ;des; ca;s d;e d;iab;ète; su;cré;
Mots	Le;diabète;de;type;1;représente;20%;des;cas;sucré
Mots formes	Le;diabète de type 1;représente;20;%;des;cas;de;diabète sucré
Racines	Le;diabèt;de;typ;1;représent;20;%;de;cas;de;diabèt;sucr
Lemmes	Le;diabète_de_type_1;représenter;20;%;de;le;cas;de;le;diabète_sucré
Termes MeSH®[5] 2005	Diabète de type i; diabète

Tableau 7.1. *Représentation de l'énoncé 1* à *l'aide de plusieurs langages d'indexation*

Considérons à présent un deuxième énoncé (énoncé 2), et comparons sa représentation à celle de l'énoncé 1 : le tableau 7.2 indique le nombre d'unités en commun pour chaque type d'unité. On peut observer que le nombre d'unités descriptives en commun dépend du langage d'indexation choisi. Dans notre exemple, il peut aller de zéro (termes MeSH) à cinq unités (tri-grammes ou mots).

L'incidence de ce type de maladie est de 1 pour 100 000 habitants contre 1,6 chez les patients atteints de diabète insipide. (énoncé 2)

[5] Le MeSH (Medical Subject Headings) est le thesaurus de référence du domaine biomédical développé et mis à jour depuis les annees 60. Il est en particulier utilisé pour indexer les documents de la base MEDLINE®.

Unité descriptive	Unités communes avec (énoncé 1)
Tri-grammes	e d;s d;e d;iab;ète; → *5 unités*
Mots	Le;diabète;de;type;1; → *5 unités* (dont 2 mots grammaticaux)
Mots formes	Le;de → *2 unités* (dont 2 mots grammaticaux)
Racines	Le;diabèt;de;typ;1→ *5 unités* (dont 2 mots grammaticaux)
Lemmes	Le;de; → *2 unités* (dont 2 mots grammaticaux)
Termes MeSH 2005	*Aucune unité.*

Tableau 7.2. *Nombre d'unités communes aux énoncés 1 et 2*

2.4 Critères de choix

Dans un contexte de recherche d'information, on peut par exemple s'interroger sur la pertinence des énoncés 1 et 2 lors d'une recherche sur le « diabète ». Selon que l'on considère la représentation par mots ou par termes MeSH, l'énoncé 2 sera retenu ou non. Bien que conceptuellement l'énoncé 2 évoque le diabète insipide et non le diabète (mellitus), il peut être pertinent de le retenir s'il s'avère que les utilisateurs sont en fait intéressés par les deux types de diabète ou que la collection ne contient pas d'autre document plus ciblé sur ce thème.

Ainsi, comme l'indique [LAN 91], le choix de la politique d'indexation doit prendre en compte les attentes des utilisateurs ainsi que le contenu de la collection. A cela s'ajoutent des considérations pratiques. En effet, dans le cadre d'une indexation automatique, les n-grammes ou les mots offrent une simplicité de mise en œuvre très supérieure aux autres unités. La question qui se pose ensuite est donc de déterminer si ces méthodes « simples » ont pour conséquence de limiter l'efficacité de la recherche d'information dans la collection. Les différentes études de la littérature indiquent qu'il y a peu de différences au niveau de la recherche d'information selon que l'indexation soit effectuée à l'aide de mots ou de mots formes [SAL 89]. De même, il a été montre que la racinisation et la lemmatisation sont des méthodes quasi-équivalentes pour les langues à morphologie simple comme l'anglais [HUL 96], mais que la lemmatisation est significativement plus efficace pour les langues à morphologie complexe tel que le français.

Nous ne pouvons que généraliser la conclusion de [PIN 04] et constater qu'il n'existe pas *a priori* de langage d'indexation supérieur à tous les autres, qu'il soit contrôlé ou libre. Le choix d'un langage d'indexation doit se faire en considérant la tâche d'indexation dans son contexte spécifique et repose principalement sur:

- le caractère descriptif ou discriminant de l'indexation : quel trait privilégier ?
- la recherche d'information doit elle être plutôt pertinente ou exhaustive?
- le domaine des documents traités est-il connu, délimité? Si oui, est-il convenablement couvert par des terminologies spécialisées?

2.5 Choix d'un langage d'indexation pour un SRI médical

Compte tenu du nombre de plus en plus élevé de documents disponibles pour tout type de recherche d'information dans le domaine médical, qu'il s'agisse d'une recherche factuelle ou personnelle, on peut estimer que, sans négliger l'aspect descriptif, le besoin concernant l'annotation des documents est principalement discriminant : les résultats d'une recherche d'information doivent être ciblés, pertinents. Par ailleurs, le domaine de la sante est l'un des mieux pourvus en terminologies spécialisées. A lui seul, le méta-thésaurus UMLS®[6] rassemble plusieurs dizaines de terminologies. Ces ressources sont un formidable outil pour la recherche d'information. L'enjeu devient alors d'utiliser au mieux cet outil et de l'améliorer si nécessaire.

Avant d'arrêter un choix sur l'utilisation d'une terminologie, il est essentiel de bien cerner le domaine ciblé mais aussi le besoin auquel elle répond [ZWE 99]. A titre d'exemple, le tableau 7.3 présente quelques terminologies spécifiquement créées pour indexer des informations médicales génériques ou personnelles sur l'ensemble du domaine.

Besoin	**Terminologie**
Description d'informations : connaissance médicale	MeSH (Medical Subject Headings)
Caractérisation « orientée » de données : statistiques hôpitaux	CIM-10 (Classification statistique Internationale des Maladies et problèmes de santé connexes)
Caractérisation « ouverte » de données : dossier patient	SNOMED (Nomenclature systématique des médecines humaine et vétérinaire)

Tableau 7.3. ***Adéquation des terminologies disponibles avec les besoins d'information***

La terminologie choisie détermine également la granularité de la description qui sera faite. Le MeSH comporte environ 23 000 mots clés, la CIM-10 contient 18 000

[6] Unified Medical Language System [McR 89]

codes et la SNOMED 109 000 concepts. La pertinence du choix effectue pourra ressortir de l'évaluation du SRI.

3 Evaluation d'un Système de Recherche d'Information

Un SRI comporte essentiellement deux modules, fondés sur l'utilisation d'un langage pivot (le langage d'indexation). Il s'agit de l'*indexation* et de l'*extraction de documents*. Ainsi, il semble naturel d'évaluer chacun de ces composants et de définir des critères de qualités pour réaliser ces évaluations.

3.1 Critères de qualité de l'indexation

Qu'est-ce qu'une « bonne » indexation ? Ou, en se replaçant dans le contexte d'un SRI, comment déterminer si un ensemble de descripteurs attribués à un document en constitue une description « utile » pour la recherche d'information ? Par définition, l'indexation doit traduire le contenu conceptuel d'un document. Une bonne indexation devrait donc recenser les thèmes abordés de manière précise et exhaustive. Par ailleurs, une recherche d'information sera plus aisément menée à bien si l'utilisateur peut disposer d'une méthode lui permettant d'anticiper dans une certaine mesure les résultats de ses actions. La régularité de l'indexation peut alors s'avérer un critère important. En d'autres termes, l'évaluation de l'indexation dans le cadre d'un SRI doit se fixer le triple objectif de déterminer :

- dans quelle mesure les descripteurs attribués à un document correspondent à un thème effectivement traité (mesure de *précision*)

- si l'ensemble des thèmes traités dans le document sont évoqués par un descripteur adéquat (mesure de *rappel*)

- s'il y a bien une correspondance régulière observable entre thèmes et descripteurs (mesure de *consistance*)

Indexation de référence. Le problème principal de l'évaluation de l'indexation, souligné par [LAN 91] est qu'il n'existe pas d'indexation « de référence » à laquelle confronter l'indexation à évaluer, qu'elle soit humaine ou automatique. Ainsi, les méthodes utilisées dans la littérature pour évaluer l'indexation en tant que telle sont au nombre de deux :

- la comparaison de l'indexation à un « gold standard », une indexation particulière prise comme référence, élaborée par un indexeur expert. Dans ce cas, les documents du corpus d'évaluation sont soumis à l'expert sans qu'il ait connaissance de l'indexation à évaluer.

- la validation de l'indexation par un indexeur expert. Dans ce cas, le corpus d'évaluation indexé est soumis à l'expert qui effectue une validation de chaque

descripteur proposé (l'expert détermine si le descripteur est adéquat ou non) et dresse l'inventaire des descripteurs qui auraient dus être sélectionnés selon lui.

La première méthode permet de réaliser une évaluation « en aveugle » dans la mesure où, à aucun moment, l'expert n'a connaissance des descripteurs à évaluer. D'un point de vue pratique, cette méthode permet de réaliser plusieurs évaluations sur un même corpus de test sans effort supplémentaire de la part de l'expert mis en cause. Cependant, il faut rappeler que l'indexation est un problème *ouvert* : pour tout document, il n'existe pas un seul et unique jeu de descripteurs constituant une indexation idéale. Au contraire, plusieurs solutions sont possibles et acceptables. L'utilisation d'une indexation de référence comme base de l'évaluation implique donc de rejeter – ou du moins de pénaliser toute autre solution, y compris des solutions acceptables. Afin de palier ce problème, on peut envisager de nuancer l'évaluation en prenant en compte la similarité sémantique des descripteurs grâce à diverses mesures comme celles passées en revue dans [PED 05]. Une alternative plus directe – mais plus coûteuse – serait de prendre en compte non pas un mais plusieurs jeux de descripteurs acceptables établis par plusieurs experts.

La seconde méthode permet de réaliser une évaluation adaptée au jeu de descripteurs proposé. En effet, l'expert évaluateur aborde le problème de manière ouverte et peut tout à fait valider plusieurs jeux descripteurs. D'un point de vue pratique, cette méthode se révèle couteuse en temps si plusieurs évaluations sont à réaliser. Par ailleurs, la méthode introduit un biais notable en faveur de l'indexation évaluée. L'expert est confronté à une solution déjà établie, et peut être tenté de s'adapter à la représentation du document qui lui est proposée.

Dans les deux cas, l'évaluation revient finalement à mesurer la consistance entre l'indexation étudiée et l'indexation prise comme référence, si on considère l'ensemble des descripteurs attribués à un document comme un tout indissociable. Cependant, dans le cas de l'indexation automatique, la plupart des systèmes proposent en fait une liste ordonnée de descripteurs susceptibles d'être utilisés pour l'indexation. Dans ces listes, le premier descripteur proposé est considéré par le système comme plus pertinent que le second et ainsi de suite. On voit ici que l'ordre d'apparition des descripteurs dans la liste est important. En conséquence, les mesures de qualité utilisées doivent également prendre en compte le rang des descripteurs extraits automatiquement. Dans la section suivante, 3.3, nous donnons une définition précise de la consistance et présentons plusieurs mesures de consistance. Nous introduisons ensuite en 3.4 les mesures de précision et de rappel et montrons comment ces dernières peuvent être utilisées pour évaluer une liste ordonnée de descripteurs.

3.2 Critères de qualité de l'extraction de documents

L'extraction de documents est effectuée à partir d'une requête formulée par un utilisateur du SRI. Dans le cadre de SRI utilisant un langage d'indexation autre que

la langue naturelle, on peut distinguer deux étapes dans le processus d'extraction de document[7].

- la première étape est la compréhension de la requête de l'utilisateur. La figure 7.1 représente à juste titre le langage d'indexation comme un pivot entre la requête de l'utilisateur et les documents de la base. L'extraction de documents ne sera possible que grâce à une traduction de la requête dans le langage d'indexation qui a servi à annoter les documents de la base.

- la deuxième étape est l'analyse de la pertinence du jeu de documents apportés en réponse à une requête. Comme dans le cas de l'indexation, cette *pertinence* peut être évaluée en estimant a/ dans quelle mesure les documents retournés correspondent à la requête (précision), b/ si l'ensemble des documents correspondant à la requête ont bien été retournés (rappel), et c/ la reproductibilité d'une recherche (consistance). Alors, tout comme pour l'évaluation de l'indexation, se pose la question de la référence : y a-t-il un seul jeu universel de documents pertinents étant données une requête et une base documentaire? Bruandet et Chevallet [BRU 03] prennent acte de cette question en évoquant une pertinence dite « système » qui reflète l'évaluation faite par le système et qui conduit au résultat de la recherche. La pertinence dite « utilisateur » reflète alors le jugement de l'utilisateur sur ce résultat. L'optimisation d'un SRI consiste alors à réduire au maximum l'écart moyen entre pertinence système et pertinence utilisateur.

Par ailleurs, contrairement à l'indexation, l'ensemble des documents retournés peut difficilement être appréhendé comme un tout. Chaque document sera consulté séparément par l'utilisateur. L'ordre dans lequel sont classés les documents retournés a donc une importance non négligeable pour éviter à l'utilisateur le temps de consultation de documents non pertinents. Les mesures d'évaluation de l'extraction de document doivent tenir compte de ce paramètre. Nous exposons en 7.3.4 les mesures utilisées à cet effet.

REMARQUE – L'évaluation de l'extraction de documents peut également servir de mesure de la qualité de l'indexation [KIM 01]. En effet, le résultat de cette étape dépend forcement de l'indexation des documents qui a été effectuée au préalable.

3.3 Consistance

Consistance de l'indexation. La consistance de l'indexation est une notion qui vise à apprécier la concordance entre des indexations proposées pour un même document par deux indexeurs ou deux méthodes d'indexations différentes. Idéalement, si les règles d'indexation sont bien définies, deux indexeurs différents devraient produire la même indexation pour un même document : c'est la

7 Remarquons que ces étapes sont par exemple traitées conjointement dans les systèmes de type LSI [DEE 90].

consistance inter-indexeur. De même, un même indexeur devrait produire la même indexation pour un même document à deux moments donnés : c'est la consistance intra-indexeur. Dans les faits, on observe des écarts entre les indexations réalisées dans ces deux situations (deux indexeurs différents à un même moment, un même indexeur à deux moments différents). La consistance inter-indexeur semble meilleure dans le cas d'une indexation contrôlée, par opposition à une indexation libre: l'étude de [BER 02] sur l'indexation libre de 3 chapitres de livre rapporte une consistance moyenne[8] de 35% contre 50% pour les études de consistance réalisée par [FUN 83] et [LEI 00] sur l'indexation d'articles scientifiques à l'aide de deux vocabulaires contrôlés, le MeSH et le Thesaurus of Psychological Index Terms. Par ailleurs, [LAN 03] montrent que les différences d'indexation pour un même ouvrage entre deux indexeurs ayant reçu les mêmes instructions peuvent s'expliquer par une différence de formation (technique *vs.* généraliste) et d'expérience (indexation d'ouvrages techniques vs. généralistes). [LEI 00] évoque également une série d'autres facteurs susceptibles d'influer sur la variabilité de l'indexation, tels que les outils (logiciels, manuels…) à disposition des indexeurs ou l'environnement dans lequel l'indexation à lieu. Par exemple, le dernier document indexé peut avoir une influence sur l'indexation en cours car l'indexeur pourra avoir tendance à mieux repérer les similitudes ou différences entre les deux documents. Pour résumer, on peut dire que les facteurs de variabilité de l'indexation sont de deux types :

- des facteurs internes : il s'agit des connaissances propres à l'indexeur, acquises au cours de la formation ou de l'expérience, ainsi que de son jugement propre ou ses préférences personnelles.

- des facteurs externes : il s'agit des règles d'indexation, du vocabulaire contrôlé (le cas échéant), des contraintes temporelles imposées pour l'indexation (le cas échéant), des outils disponibles, de l'environnement de travail…

Mesures de consistance. Les nombreuses études réalisées sur la consistance[9] - que ce soit pour l'indexation ou pour d'autres problèmes cognitifs tels que la formulation de requêtes [SAR 88] ou la construction de documents hypertextes [FUR 99], utilisent un spectre de mesures de la consistance assez large. Les mesures les plus usitées à ce jour semblent être la mesure de Hooper [HOP65], la mesure de Rolling [ROL 81] et le taux de recouvrement [SAR 88]. Nous les détaillons ci-dessous.

Soient A et B deux listes de descripteurs attribués à un document (par exemple, par deux indexeurs différents). Soient M le nombre de descripteurs appartenant exclusivement à la liste A (i.e. n'appartenant pas à la liste B) et N le nombre de descripteurs appartenant exclusivement à la liste B (i.e. n'appartenant pas à la liste

[8] Ces évaluations sont fondées sur la mesure de Hooper, définie ci-dessous.
[9] Cependant, nous n'avons connaissance d'aucune étude récente dans le domaine de la santé.

A[10]) - *a priori*, N≠M. Enfin, soit C le nombre de descripteurs communs aux listes A et B. La figure 7.2 illustre cette répartition.

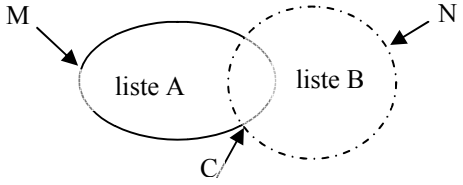

Figure 7.2. ***Répartition des descripteurs attribués pour un même document***

On peut alors définir:

- la mesure de Hooper [HOP 65], qui évalue la proportion de descripteurs proposés par les deux indexeurs à la fois, sur l'ensemble des descripteurs proposés par l'un ou l'autre des indexeurs:

$$C_H = \frac{100 * C}{M + N + C} \qquad [7.1]$$

- la mesure de Rolling [ROL 81], qui accorde un poids supplémentaire aux descripteurs témoignant d'un consensus entre les indexeurs (descripteurs proposés par les deux indexeurs à la fois), par rapport aux descripteurs témoignant d'une divergence d'appréciation (descripteurs proposés par l'un des indexeurs seulement).

$$C_R = \frac{100 * 2C}{M + N + 2C} \qquad [7.2]$$

- le taux de recouvrement [SAR 88], qui ne place pas les deux listes de descripteurs au même niveau. Elle permet d'évaluer le taux de recouvrement d'une liste par rapport à l'autre. Ainsi, la mesure S_A est utilisée si la liste A est prise comme référence et la mesure S_B si la liste B est prise comme référence. On a :

$$S_A = \frac{100 * C}{M + C} \text{ et } S_B = \frac{100 * C}{N + C} \qquad [7.3]$$

REMARQUE – Une étude portant sur la recherche d'information [SAR 88] a montré que les termes utilisés par différentes personnes pour rechercher la même information ne concordent que rarement (taux de recouvrement moyen de 27%). Ainsi, le problème de consistance n'est pas propre à l'indexation. Il survient

[10] Ces ensembles ne tiennent pas compte de l'éventuelle proximité sémantique entre les descripteurs. Cependant, des variantes des mesures présentées introduisent la notion de proximité sémantique dans le calcul [PED 05].

également dans d'autres tâches faisant appel à une traduction conceptuelle dans un langage contrôlé telle que la formulation de requêtes dans un moteur de recherche.

3.4 Mesures d'évaluation

Nous nous plaçons à présent strictement dans le cadre d'une évaluation fondée sur une *référence*, que ce soit pour l'indexation ou pour l'extraction de documents. Ainsi, comme dans le cas de la consistance, nous disposons de deux listes de descripteurs ou de documents qui doivent être comparées. Cependant, l'une ces listes constitue la référence à laquelle l'autre liste doit être confrontée. Nous pouvons alors définir les notions de vrai/faux positif et de vrai/faux négatif. On appelle « vrai positif » un descripteur (resp. document) qui figure à la fois dans la liste évaluée et dans la liste de référence. Plus précisément, on parle de « positif » pour les descripteurs (resp. documents) qui figurent dans la liste de référence et de « vrai » pour indiquer que la décision reflétée par la liste à évaluer concernant ce descripteur (resp. document) est correcte. Par suite, un « faux positif » désigne un descripteur (resp. document) qui figure dans la liste de référence mais pas dans la liste évaluée – la décision reflète par la liste à évaluer est erronée. Un « faux négatif » désigne un descripteur (resp. document) qui ne figure pas dans la liste de référence mais figure dans la liste évaluée. Un « vrai négatif » désigne un descripteur (resp. document) qui ne figure ni dans la liste de référence ni dans la liste évaluée. La distribution des descripteurs (resp. documents) ainsi définie peut être représentée dans une matrice de confusion (ou tableau de contingence) présentée par la figure 7. 3

Liste évaluée	**∈ Référence**	**∉ Référence**
Sélectionné	VP: Vrai Positif	F P : Faux Positif
Non sélectionné	FN: Faux Négatif	VN : Vrai Négatif

Figure 7.3. *Matrice de confusion*

Soient P la précision, R le rappel, et F_α la F-measure. En utilisant les notations introduites dans la matrice de confusion (figure 7.3) on a :

$$P = \frac{VP}{VP + FN} \text{ et } R = \frac{VP}{VP + FP} \quad [7.4]$$

$$F_\alpha = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \text{ avec } 0 \leq \alpha \leq 1 \quad [7.5]$$

où α représente le poids attribué à la précision. Si précision et rappel sont considérés comme étant d'importance égale, on a α = 0,5.

La précision représente les termes correctement extraits par le système. On peut aussi évoquer le silence (*S=1-R*) pour évaluer la proportion de termes attendus n'ayant pas été extraits. Le rappel représente la couverture du système. On peut aussi évoquer le bruit (*N=1-P*) pour évaluer la proportion de termes erronés extraits par le système (faux positifs) ou la pureté (*Pureté=VN/(VN+FP))* pour évaluer la proportion d'erreurs d'indexation (extraction de termes erronés) évitées par le système [SOE 94]. Il faut remarquer que cette dernière mesure n'est utilisable que dans le cas de l'indexation contrôlée. Évaluer le nombre de vrais négatifs implique de connaitre le nombre total de descripteurs qui peuvent être utilisés. Le choix d'une mesure d'évaluation dépend donc du langage d'indexation utilisé dans le SRI.

Au delà des performances quantitatives des systèmes qu'il est possible d'évaluer comme nous venons de le décrire[11], des éléments *qualitatifs* peuvent entrer en ligne de compte. Ainsi, entre deux systèmes offrant des performances équivalentes, la différence pourra se faire sur la facilite d'utilisation ou la disponibilité de fonctionnalités supplémentaires. Par exemple, l'un des atouts du système d'indexation semi-automatique de littérature médicale MTI [ARO 04] est de permettre aux indexeurs qui l'utilisent de sélectionner les propositions qui les intéressent en quelques clics. Ce principe simple permet un gain de temps significatif apprécié. Nous présentons à la section suivante quelques fonctionnalités qui, en s'appuyant sur des ressources terminologiques, constituent une valeur ajoutée innovante dans les SRI modernes.

4 Enjeux de la recherche d'information en santé

Etant donné une collection de documents, un des enjeux de la Recherche d'Information dans tous les domaines, et en particulier en santé, est de proposer un accès « intelligent » à l'information. Cette « intelligence » doit s'exercer à plusieurs niveaux. D'une part, elle doit permettre la *compréhension* des besoins d'information de l'utilisateur. Celle-ci s'effectue avec l'analyse de la requête soumise qui permet de situer le besoin d'information dans le réseau sémantique du domaine, puis d'établir une correspondance avec les documents de la collection les plus proches. D'autre part, le développement des outils terminologiques doit également permettre d'*approfondir* le service rendu à l'utilisateur en lui proposant la consultation d'informations connexes à sa recherche. Ce type d'information permet par exemple d'étendre une requête à une autre langue, un autre thème ou un autre type de

[11] Nous invitons le lecteur à consulter [MAN 00] pour approfondir la revue des mesures d'évaluation en Recherche d'Information proposée ici.

documents. Dans les sections suivantes, nous présentons plus en détail quelques unes de ces fonctionnalités en montrant le rôle central de la terminologie dans chacune d'elles.

4.1 Approfondir la recherche au sein du même SRI

Recherche de proches voisins. Une fois identifié un document correspondant au besoin d'information de l'utilisateur, on peut supposer que des documents sémantiquement proches du document initial, c'est-à-dire traitant par exemple des mêmes thèmes, auront également un intérêt pour l'utilisateur. Dans le cadre d'un SRI, cela revient à considérer le document initial comme une expression fidèle du besoin d'information. L'avantage de cette hypothèse est de donner accès à une interprétation de la requête de l'utilisateur dans le langage d'indexation du système grâce à l'indexation du document telle qu'elle figure dans la collection. Cependant, l'indexation peut également être combinée à d'autres éléments porteurs de sens dans le document, comme le titre. Dans le domaine de la santé, l'algorithme « Pubmed Related Citations » (PRC) [KIM 01] est un exemple d'outil permettant la recherche de proches voisins dans la base documentaire MEDLINE. PRC recherche les proches voisins à partir du titre, du résumé et de l'indexation MeSH d'un document.

Extension de requête. Les SRI de type LSI [DEE 90] que nous évoquions en 7.3.2 sont particulièrement remarquables par leur capacité à prendre en compte des expressions *latentes* des concepts évoqués dans un document ou une requête. Concrètement, cela signifie par exemple qu'une requête sur les hypoglycémiants pourra proposer des documents évoquant la metformine, qui est un hypoglycémiant particulier, sans contenir explicitement « hypoglycémiants ». Ce rapprochement est possible si, statistiquement, les mots « hypoglycémiants » et « metformine » partagent un grand nombre de contextes. De la même manière, les ressources terminologiques disponibles peuvent permettre d'*interpréter* voire d'*étendre* une requête, comme l'illustre la figure 7.4. Ainsi, dans le cadre d'un SRI en santé, le lien de synonymie qui existe dans le thesaurus MeSH entre les expressions « maladies cardiaques » et « cardiopathies » permet d'interpréter une requête sur les maladies cardiaques en recherchant les documents indexés à l'aide du mot clé MeSH <*cardiopathies*> ou de mots clés MeSH reliés à <*cardiopathies*> par un lien « EST-UN » dans la terminologie, comme par exemple <*arrêt cardiaque*>. La construction et l'exploitation de telles ressources terminologiques fait l'objet de nombreux travaux aussi bien en santé [GRA 04], [SOU 04] que dans d'autres domaines [BRU 03]. Les liens morphologiques ainsi que les relations de synonymie et de parenté directe (relations « EST-UN » ou « PARTIE-DE ») entre les termes sont privilégiées pour interpréter les requêtes formulées en langage naturel dans le langage d'indexation du SRI.

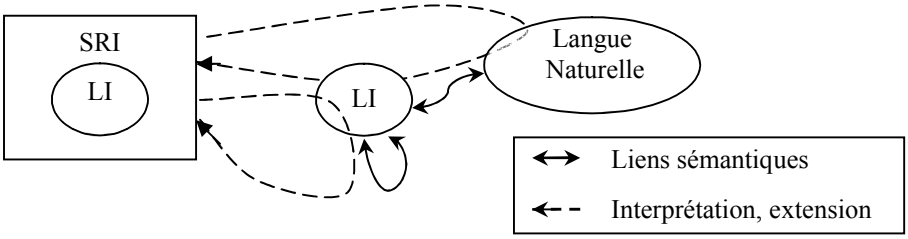

Figure 7.4. *Approfondissement d'une Recherche d'Information au sein d'un Système de Recherche d'Information (SRI) utilisant un Langage d'Indexation (LI) donné : les liens sémantiques entre éléments du langage d'indexation ou entre le langage d'indexation et la langue naturelle permettent d'interpréter et d'étendre les requêtes au sein du SRI.*

Cependant, d'autres types de relations présentent un intérêt pour approfondir une recherche en cours ou pour rediriger une recherche infructueuse. Ce sont les relations de type « VOIR AUSSI » qui indiquent des liens sémantiques entre les termes et permettent de rapprocher des concepts proches. Certaines de ces relations sont explicitées dans les ressources terminologiques. Une telle relation existe dans le MeSH entre les termes <*mammographie*> et <*échographie mammaire*> qui désignent deux procédures différentes ayant en commun d'être des examens du sein. Il semble alors tout à fait pertinent d'indiquer ce lien aux utilisateurs effectuant une recherche sur la <*mammographie*> et de leur proposer d'étendre leur recherche à l'<*échographie mammaire*>. L'adaptation de méthodes de fouille de données (analyse formelle de treillis de concepts) au corpus indexé CISMeF a permis d'extraire de nouvelles connaissances - inexistantes dans le MeSH- exprimées sous la forme de règles d'association entre concepts: par exemple <*tumeurs du sein/prévention et contrôle*> R <*mammographie*>. En appliquant cette règle pour la recherche d'information, une requête sur le terme "mammographie" permet de proposer à l'utilisateur des documents traitant de la "prévention du cancer du sein" [Sou 04]. Parallèlement, cela a permis à un expert du domaine médical (B. Thirion) de modéliser d'autres règles telles que <*fœtus/échographie*> R < *échographie prénatale*> qui peuvent également être exploitées pour l'indexation des documents.

4.2 Elargissement de la recherche à d'autres SRI

Recherche d'Information trans-langue. Bien que de nombreuses langues soient représentées sur l'Internet certaines informations sont parfois disponibles dans un nombre réduit de langues. L'anglais est par exemple très prégnant dans les publications scientifiques de tous les domaines, y compris en santé. Dans ce cadre, de nombreux utilisateurs qui possèdent les rudiments d'une langue étrangère

peuvent potentiellement profiter d'une source supplémentaire d'information complétant celle constituée par les documents rédigés dans leur langue maternelle. Cependant, formuler des requêtes précises dans une langue autre que sa langue maternelle peut s'avérer difficile. De plus, l'utilisateur multilingue peut également souhaiter éviter la répétition de requêtes dans les différentes langues qu'il maîtrise. L'utilisation de terminologies multilingues permet d'effectuer une recherche d'information trans-langue, soit à l'intérieur d'un même SRI soit entre SRI distincts. La figure 7.5 illustre ces deux cas. Dans le domaine de la sante, le thesaurus MeSH, disponible dans plusieurs langues, peut être utilisé à cet effet. Par exemple, dans le catalogue francophone CISMeF, les requêtes peuvent être indifféremment formulées en français ou en anglais. Grâce aux liens d'équivalence entre les termes français et anglais dans le MeSH, une requête sur <*heart arrest*> sera automatiquement interprétée comme une recherche sur les documents indexés à l'aide du mot clé <*arrêt cardiaque*>. Pour aller plus loin, il est particulièrement intéressant de faire le lien entre différents SRI construits autour du même langage d'indexation. Le MeSH, qui est le thesaurus de référence pour la description d'informations biomédicales, est utilisé dans de nombreuses bases documentaires telles que MEDLINE, CISMeF ou OMNI. Ainsi, une requête formulée en français dans le moteur de Recherche Doc'CISMeF peut être traduite dans la langue d'autres bases documentaires connues, et donner lieu à une recherche dans ces bases.

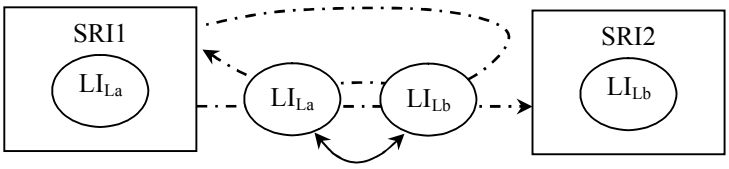

Figure 7.5. *Recherche d'Information trans-langue au sein de Systèmes de Recherche d'Information (SRI1 et SRI2) utilisant un Langage d'Indexation (LI) dans deux langues, La et Lb respectivement : les liens sémantiques (flèche pleine) entre les versions de LI disponibles dans La et Lb permettent d'effectuer une recherche trans-langue (flèches en pointillés) au sein de SRI1 ou de SRI1 vers SRI2.*

Dans des travaux récents [NEV 06] nous avons adapté spécifiquement cette méthode au vocabulaire « patient » afin d'effectuer une recherche d'information trans-langue dans la base documentaire patient américaine MedlinePlus à partir de CISMeF.

Recherche d'Information contextuelle. La Recherche d'Information Contextuelle est une recherche connexe correspondant à un besoin d'information qui apparaît au cours de la recherche initiale, c'est-à-dire *dans le contexte* d'une autre recherche. Un exemple simple d'un tel besoin est la rencontre d'un terme dont le sens est inconnu à

la lecture d'un document. La nécessité de consulter un dictionnaire pour connaitre le sens du terme en question est survenu pendant la lecture du document et n'était *a priori* pas prévisible[12]. De la même façon, des besoins d'information d'un type différent de celui fourni par la base consultée peuvent survenir au cours d'une recherche. Dans le domaine de la santé, un médecin peut par exemple avoir besoin d'information sur une pathologie ou un traitement particulier lors de la consultation du dossier d'un patient y correspondant. Dans ce contexte, il est nécessaire de faire le lien entre une base de données médicales personnelles (les dossiers patient) et une base de données médicales factuelles (par exemple, des cours de médecine ou des recommandations de bonne pratique clinique). Pour cela, Cimino et Li [CIM 03] proposent de doter les dossiers électroniques patient de boutons contextuels permettant d'effectuer une recherche d'information factuelle sur une base de connaissance externe. Ces recherches contextuelles sont possibles s'il existe des liens connus entre les différentes terminologies médicales utilisées comme langage d'indexation dans les deux SRI. Nous illustrons cette situation sur la figure 7.6 :

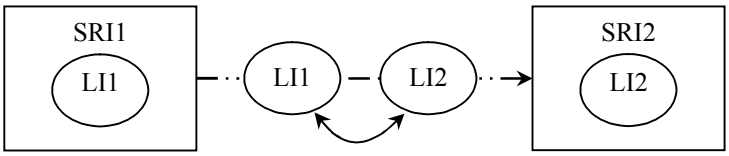

Figure 7.6. *Elargissement d'une Recherche d'Information d'un Système (SRI1) utilisant un Langage d'Indexation (LI1) à un autre Système (SRI2) utilisant un autre Langage d'Indexation (LI2) grâce aux liens sémantiques entre les termes de LI1 et LI2.*

De tels liens sont par exemple recensés dans l'UMLS pour plus de 70 terminologies médicales. Dans les expériences de Cimino et Li, les dossiers patients sont codés avec la CIM10 et les données factuelles sont indexées avec le MeSH. Dans ce contexte, il est par exemple possible d'appliquer la correspondance entre le code CIM10 <*syndrome abdominal aigu*> et le mot clé MeSH <*urgence abdominale*> afin d'interpréter un besoin d'information sur le <*syndrome abdominal aigu*> par une requête factuelle sur <*urgence abdominale*>.

5 Conclusion

Dans ce chapitre, nous avons rappelé la structure globale d'un système de recherche d'information et montré le rôle central du langage d'indexation choisi. Le

[12] Remarquons que le système Alexandria de la société Memodata (http://www.memodata.com/) permet une consultation de dictionnaire par simple clic à partir d'un document html. L'équipe CISMeF a récemment contribué à l'ajout de définitions de termes médicaux dans le dictionnaire français d'Alexandria.

domaine de la santé, riche en vocabulaires contrôlés dispose de ressources terminologiques considérables pour permettre l'archivage et l'accès aux documents de santé. Ces ressources terminologiques peuvent également être utilisées afin d'étendre la portée des systèmes de Recherche d'Information à d'autres langues ou d'autres types de documents, ouvrant la voie vers une véritable interopérabilité des systèmes.

6 Remerciements

La rédaction de ce chapitre a été réalisée dans le cadre de la participation d'A. Névéol au programme postdoctoral du NIH financé par la National Library of Medicine et administré par ORISE (Oak Ridge Institute for Science and Education).

7 Bibliographie

[ARO 04] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004: 268-72.

[BER 02] Berrios, D. C., Cucina, R. J., Fagan, L. M. (2002). « Methods for semi-automated indexing for high precision retrieval », JAMIA, 9 (6): 637-651

[BRU 03] Bruandet, M-F., Chevallet J-P. (2003) «Utilisation et construction de bases de connaissances pour la Recherche d'Informations», in Assistance Intelligente à la Recherche d'Information , M.-H. Stefanini, E. Gaussier, Hermes, chapter 3, pp. 85-118.

[CIM 03] Cimino J.J., Li J. (2003) « Sharing infobuttons to resolve clinicians' information needs », AMIA Annu Symp Proc. 2003:815.

[DEE 90] Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R. (1990) «Indexing by latent semantic analysis». JASIS, 6 (41) :391-407.

[FUN 83] Funk, M. E., Reid, C. A., McGoogan, L. S. (1983). « Indexing consistency in MEDLINE », Bull. Med. Libr. Assoc., 2 (71) :176-183.

[FUR 99] Furner, J., Ellis, D., Willett, P. (1999). « Inter-linker consistency in the manual construction of hypertext documents », ACM Computing Surveys, 4es(31). (On-line supplement : article no. 18)

[GRA 04] Grabar, N. (2004). « Terminologie medicale et morphologie : acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique. » Thèse de l'Université Paris 6.

[HAL 97] Halleb, M., Lelu, A. (1997). « Hypertextualisation automatique multilingue à partir des frequences des n-grammes », Hypertextes et hypermedias, 1 (2-3-4) : 275-287

[HOP 65] Hooper, R. S. (1965). « Indexer consistency tests : origin, measurement, results and utilization », (Tech. Rep.). IBM Corporation. (Bethesda, MD)

[HUL 96] Hull, D. A., Grefensette, G. (1996). « Experiments in multilingual information retrieval », In Proc. of the 19th Annual International ACM SIGIR Conference.

[JAC 97] Jacquemin, C. (1997). « Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus », HDR de l'Université de Nantes.

[KIM 01] Kim, W., Aronson, A. R., Wilbur, W. J. (2001). Automatic MeSH term assignment and quality assessment. In *Proc. AMIA Symp. 2001:* 319-323.

[LAN 91] Lancaster, F. W. (1991). « Indexing and abstracting in theory and practice », University of Illinois : Champaign, IL.

[LAN 03] Landes, D., Spidal, D. (2003). « An index comparison project : The effects of two indexers' diverse backgrounds on creating an index from a software manual », In Proc. ASI-IASC/SCAD.

[LEI 00] Leininger, K. (2000). « Interindexer consistency in psycINFO », Journal of Librarianship and Information Science, 1 (32).

[MAN 00] Manning, C. D., & Shütze, H. (2000). Foundations of Statistical Natural Language Processing. Cambridge, Massashusetts : MIT Press.

[McR 89] Mc Cray, A. T. (1989). « The UMLS semantic network », SCAMC - Washington D.C., 503-507

[NEV 06] Névéol A., Pereira S., Soualmia L.S., Thirion B., Darmoni S.J. (2006) « A method of cross-lingual consumer health information retrieval » Soumis à MIE 2006.

[PED 05] Pedersen, T., Pakhomov, S., Patwardhan S. (2005) « Measures of Semantic Similarity and Relatedness in the Medical Domain » *University of Minnesota Digital Technology Center Research Report* DTC 2005/12.

[PIN 04] Pincemin, B. (2004). Compte rendu du n°2 de la revue Corpus sur « la distance inter-textuelle ». Texto. (Disponible sur : http://www.revue-texto.net/Parutions/CR/Pincemin CR.html. (Consulté le 11 /06/05))

[ROL 81] Rolling, L. (1981). « Indexing consistency, quality and efficiency », Information Processing and Management, 2 (17): 69-76

[SAL 83] Salton, G., McGill, M. J. (1983). « Introduction to modern information retrieval », New York :McGraw-Hill.

[SAL 89] Salton, G. (1989). « Automatic text processing : The transformation, analysis, and retrieval of information by computer », Reading, MA : Addison-Wesley.

[SAR 88] Saracevic, T., Kantor, P. (1988). « Study of information seeking and retrieving : Part iii. searcher, searches and overlap », JASIS, 3 (39): 197-216.

[SOE 94] Soergel, D. (1994). « Indexing and retrieval performance : the logical evidence », Journal of American Society for Information Science (JASIS)

[SOU 04] Soualmia, L. S. (2004) « Etude et évaluation d'une approche multiple pour la projection de requêtes sur une terminologie normalisée.» Thèse de l'INSA de Rouen.

[ZWE 99] Zweigenbaum, P. (1999). « Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances », *ISIS*.