Assistance intelligente à la recherche d'information : élaboration d'un projet de moteur de recherche au service de la connaissance dans l'organisation

Michail Frontère

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

École Management et Société-Département CITS

INTD

MÉMOIRE pour obtenir le Titre enregistré au RNCP
« Chef de projet en ingénierie documentaire »
Niveau I

Présenté et soutenu par
Mikhaïl Frontère

le 17-12-2015

Assistance intelligente à la recherche d'information : élaboration d'un projet de moteur de recherche au service de la connaissance dans l'organisation

Jury :
Nadia Raïs, Directrice de mémoire CNAM-INTD
Sylvie Dalbin, Co-directrice de mémoire
Eric Debonne, Co-directeur de mémoire, CNAM-INTD, SOLACI
Sophie Longuet, Co-jury de mémoire, OSTEOBIO

Promotion 45

Remerciements

Merci à ma directrice de mémoire, Nadia Raïs, pour son écoute, sa bienveillance et son exigence.

Merci à mes Co-directeurs de mémoire, Sylvie Dalbin et Eric Debonne, pour leurs observations précieuses.

Merci à l'équipe d'Ostéobio pour sa disponibilité et la confiance dont elle a su faire preuve à mon égard.

Merci à ma mère pour son regard aiguisé.

Merci à la direction de la scolarité et à l'administration de l'INTD pour sa disponibilité et sa réactivité tout au long de l'année.

Notice

Frontère Mikhaïl. Assistance intelligente à la recherche d'information : élaboration d'un projet de moteur de recherche au service de la connaissance dans l'organisation. 2015. 169 p. Mémoire professionnel INTD, Titre I, Ingénierie documentaire. CNAM - INTD, 2015.

L'éparpillement des connaissances dans une organisation fait peser des risques importants sur son fonctionnement. Perte de temps pour accéder, souvent partiellement, aux informations existantes, faible valorisation des connaissances pouvant entraîner des pertes de valeur scientifiques et donc financières, mauvais choix stratégiques, non prise en compte des risques juridiques et industriels, etc. Face à ces risques, il est important que les organisations se dotent d'outils permettant la mise à disposition et l'exploitation « intelligente » des ressources informationnelles qu'elles produisent ou qu'elles consultent. C'était l'objet de la mission qui m'a été confiée par la société Ostéobio, école et centre de recherche en ostéopathie, que d'élaborer un cahier des charges pour un projet de moteur de recherche qui permette de promouvoir l'utilisation d'un modèle de connaissance dans la recherche et l'exploitation des ressources scientifiques de l'entreprise. Après avoir présenté le contexte de cette mission et initié une réflexion sur ce qui définit la recherche d'informations, ce mémoire se propose de présenter les façons dont un moteur de recherche peut être au « service de la connaissance » dans une organisation. Il insiste particulièrement sur les notions d'indexation automatique des documents et de référentiel terminologique qui sont, pour l'auteur, les fondements de l'assistance intelligente à la recherche et l'exploitation d'informations.

Accès à l'information ; recherche d'information ; indexation automatique ; métadonnées ; moteur de recherche ; référentiel ; ontologie ; thésaurus ; traitement automatique des langues ; aide utilisateur ; cahier des charges.

The scattering away of knowledge in an organization poses significant risks to its operation.
Waste of time to reach, often partially, the existing information; low valuation of the knowledge leading to scientific and hence financial losses; bad strategic choices; legal and industrial risks not being taken into account, etc.
Faced with these risks, it is important that the organizations should adopt tools that provide and use "with intelligence" the information they produce or consult.

It was the purpose of the mission that I have been entrusted with by Ostéobio, school and center for research in osteopathy. "The mission consisted in designing a search engine that can promote a model of knowledge *dedicated to* the research and *the use* of the company's scientific resources."
After having presented the context of this mission and initiated a reflection on what defines the search for information, the following work aims at presenting the ways in which a search engine can be at the service of knowledge within an organization.
It particularly insists on the notions of automatic indexing and terminology repository that are, according to the author, the foundations for intelligent research assistance and exploitation of information.

Table des matières

Liste des figures

Introduction

Une activité primordiale

La recherche d'information est une activité fondamentale qui se situe au cœur des autres activités humaines. Il est en effet difficile de trouver une tâche qui ne nécessite pas de rechercher de l'information, que cette tâche soit de l'ordre du loisir, de la consommation, de la connaissance ou du travail. Et de plus en plus, voire exclusivement parfois, cette recherche d'information se déroule dans un environnement numérique. Ce constat se vérifie d'autant plus si l'on se situe dans le cadre d'activités professionnelles [3, Dinet].

En effet dans les organisations (entreprises privées, associations, administrations, etc.), la prise de décision ne peut se faire qu'après un important travail de synthèse et de communication de l'information. Que ces décisions concernent des choix d'investissement, des orientations marketing ou des axes de recherche elles nécessitent toutes que soient regroupé, ordonné et communiqué un ensemble d'informations relatives à ces décisions, aux contextes dans lesquels ces décisions se prennent et aux conséquences qu'elles peuvent avoir. Or, avant d'effectuer ces opérations sur ces informations, il faut les trouver et donc les chercher [3, Dinet].

Par ailleurs, la question de l'innovation (veille technologique et concurrentielle, décisions d'axes de recherche, développements de nouveaux produits) est directement liée à la question de la gestion des connaissances dans l'organisation, à leur accessibilité, leur utilisation, leur exploitation, leur partage et leur enrichissement. En effet, l'intelligence dune organisation repose en grande partie sur l'identification, la collecte, l'analyse, la validation, la valorisation, le stockage et la recherche d'informations nécessaires à son fonctionnement et à son développement. Accéder à l'information est une chose, savoir la comprendre, l'interpréter et l'utiliser en est une autre, c'est là que se situe la véritable valeur ajoutée de l'exploitation de l'information dans une organisation. [2, Vuillequiez]

Une activité mal exploitée…

Lors d'une enquête réalisée en 2009 pour Delphi Group (www.delphigroup.com), qui opère à la jonction du marketing et des technologies, 1030 employés de quinze entreprises américaines de taille moyenne ont répondu qu'ils passaient plus du tiers de leur temps de travail à rechercher de l'information, sans pour autant qu'ils appartiennent à des services de veille documentaire ou informationnelle. Ces résultats se rapprochent de ceux obtenus par Mscdermott[1] qui montrent que 38 % des employés des grandes entreprises passent l'essentiel de leur temps à rechercher des informations [3, Dinet].

[1] McDermott M., « Knowledge workers : you can gauge their effectiveness », Leadership excellence, 22(10), 15-17, 2005.

Pourtant, tout ce temps de travail consacré à la recherche d'information semble manquer considérablement d'efficacité. En effet, selon un rapport interne de la firme Google datant de 2008, seul un quart des recherches d'information menées au sein d'une grande entreprise serait couronnées de succès. De plus, presque la moitié de ces activités seraient non productives (et donc non rentables) puisqu'elles consisteraient à recréer des informations déjà existantes, à convertir des informations sous d'autres formats, à collecter des documents ou des informations sans les analyser ou encore à effectuer des recherches sans résultats probants [3, Dinet].

Outre leur faible efficacité, et alors qu'elles représentent un temps d'activité de plus en plus important, on estime que pour la plupart d'entre elles, les activités de recherche d'information sont superfétatoires. On constate en effet que le temps passé à rechercher des informations inutiles est de plus en plus important. Les notions d'utilité et d'inutilité pour des recherches d'information, ainsi que le temps qui leur est consacré, sont certes difficiles à évaluer de façon précise, mais certains organismes économiques (par exemple l'IDC : www.idc.com/) estiment que 90 % des documents et informations produits par l'entreprise existeraient déjà ailleurs [3, Dinet].

Qui engendre des coûts importants

Beaucoup d'organisations sociales ont à gérer et maintenir des connaissances, que ce soit leur raison d'être (réseaux d'intérêt, équipes de recherche, écoles, etc.) ou un résultat de leur fonctionnement (entreprises, administrations, associations...). Comme nous l'avons vu, la capacité de ces organisations à répondre au monde extérieur (innovation des recherches, temps de réponse au marché, qualité des formations, etc.) dépend de leur capacité à détecter, mémoriser, se remémorer et activer leurs connaissances. Dans ce contexte où la connaissance est un capital, un système d'information performant est un atout primordial [9, Gandon].

Or la faible efficacité, voire l'inutilité des activités de recherche d'information telles que nous les avons exposées plus haut, engendre des coûts considérables pour les organisations dans lesquelles elles se déroulent. Ces coûts sont de diverses natures et susceptibles d'affecter tous les aspects de la vie d'une organisation :

- Perte de valeur liée à la non-disponibilité d'une information potentiellement innovante ;
- Perte de valeur liée à une mauvaise diffusion des connaissances ;
- Temps en termes de travail humain et donc perte de productivité ;
- Échecs commerciaux dus à de mauvais choix stratégiques relatifs aux :
- Investissements ;

- Axes de recherches ;
- Campagnes marketing ;
- Risques juridiques ;
- Risques d'accidents et d'incidents sur les sites industriels.

Ces coûts peuvent avoir des conséquences néfastes importantes pour une entreprise qui évolue dans un environnement concurrentiel exacerbé, un environnement normatif complexe et contraignant, un environnement industriel complexe et porteur de risques et un contexte technologique de plus en plus évolutif.

Conjointement aux coûts économiques, la mauvaise gestion de l'information peut engendrer un stress important chez les agents de l'organisation. Stress quant à la difficulté d'accéder à l'information, stress quant à la difficulté de l'exploiter pertinemment en regard des objectifs de l'organisation et stress enfin quant aux conséquences de cette mauvaise gestion. Or le stress, la tension, l'inadéquation entre les objectifs assignés aux salariés et les moyens dont ils disposent pour atteindre ces objectifs ne peuvent avoir que des conséquences négatives sur le fonctionnement de l'organisation pour laquelle ils travaillent.

De l'importance des Systèmes de recherche d'information

Cette augmentation des pressions liées à la concurrence, aux contraintes techniques, juridiques et temporelles ainsi que l'importance et la dispersion des masses informationnelles (l'information pertinente ne se trouve pas dans un document, mais dans un ensemble de documents hétérogènes et éparpillés) ont rendu encore plus prépondérante la phase de recherche d'information dans les organisations.

Face aux écueils que connaissent les activités de recherche d'information précédemment évoquées, il est logique que les entreprises soient demandeuses de systèmes de recherche d'information qui assistent et aident leurs employés à trouver et exploiter l'information la plus pertinente possible en regard des objectifs de leurs recherches et de leurs activités professionnelles (nous reviendrons sur la notion de pertinence dans la deuxième section de ce mémoire).

Cette demande correspond précisément à celles que nous avons rencontrée lors de notre mission de stage de fin d'étude. La société Ostéobio, qui regroupe des pôles de recherche et une école

d'ostéopathie, souhaitait que soit élaboré un « Cahier des Charges » pour la mise en place d'une solution de recherche et d'exploitation de l'information qui lui permette :

- De stocker et de rendre accessible un ensemble de ressources scientifiques produites en son sein ou présentes sur des bases de données externes (fournisseurs de contenu scientifique) ;
- De rechercher et d'exploiter ces ressources selon un modèle de connaissance qui lui est propre afin de :
    - Promouvoir ce modèle auprès de ses utilisateurs ;
    - Utiliser ce modèle pour exploiter les informations contenues dans les ressources concernées par le projet afin d'extraire et d'exploiter de façon innovante (selon la logique du modèle) les connaissances qui y sont contenues de manière formelle ou de manière implicite.

Pour concevoir ce Cahier des Charges qui, comme vous le verrez plus loin, a été élaboré de façon à être réutilisable et évolutif, il a fallu mener deux investigations conjointes dans des contextes scientifique et informationnel spécifiques :

- D'une part, faire émerger et comprendre les attentes de futurs utilisateurs qui évoluent dans un domaine jusqu'alors inconnu de l'auteur : la recherche et les études dans le domaine de l'ostéopathie ;
- D'autre part, dans le cadre de ce domaine, comprendre les attentes propres aux membres d'Ostéobio relativement à la recherche et l'exploitation de l'information, considérant qu'ils voulaient que ces opérations se fondent sur un système d'organisation des connaissances transdisciplinaire en Ostéopathie, nommé « modèle Princeps ».

Dans le cadre j'ai pu recueillir toutes les informations nécessaires à l'élaboration des documents qui ont été livrés, à savoir :

- Un document nommé « Éléments pour un Cahier des Charges », qui propose de découper le projet de moteur de recherche en trois phases allant d'un moteur de recherche « basique » exploitant une terminologie de référence à une solution sophistiquée comportant de nombreuses fonctionnalités et fondée sur l'implémentation d'une ontologie (libre à l'équipe projet de choisir à partir de quelle phase ils souhaitent démarrer le projet) ;
- Un document intitulé « Considérations sur les référentiels » qui présente la façon dont les référentiels terminologiques peuvent être mis au service de la recherche et l'exploitation de l'information en étant implémentés dans un moteur de recherche.

Les attentes et besoins relatifs au projet constitueront en quelque sorte le fil rouge de ce mémoire. Les thématiques qui y seront abordées le seront en référence à la problématique centrale d'Ostéobio : comment un moteur de recherche peut être au service de la connaissance au sein d'une organisation en fournissant à ses utilisateurs une assistance intelligente à la recherche d'information.

Nous entendons par être au service de la connaissance :

- Soutenir la diffusion et l'appropriation de la connaissance dans l'organisation ;
- Soutenir les activités de recherche dans l'organisation.

Cela signifie que la société Ostéobio est à la recherche d'un outil permettant :

- De valoriser une culture commune à travers la compréhension et la valorisation du modèle Princeps ;
- D'exploiter cette culture dans le cadre des activités de recherche d'information de la société.

Nous entendons par assistance intelligente les stratégies destinées à fournir aux utilisateurs une aide dans l'expression et la satisfaction de leurs besoins d'information [19, Gaussier]. Ces stratégies pouvant concerner aussi bien les fonctionnalités de recherche d'information que l'organisation et l'apparence de l'interface, nous décidons pour ce mémoire de nous concentrer sur les premières.

Pour répondre à cette problématique, nous allons, dans une première partie, examiner le contexte professionnel à l'origine du sujet de ce mémoire. Dans une deuxième partie, nous tenterons de comprendre comment se définit et se structure une recherche d'information du point de vue des usagers. Dans une troisième partie, nous nous intéresserons aux processus d'indexation permettant de « soutenir » les pratiques et les fonctionnalités de recherche d'information. La quatrième partie du mémoire sera consacrée aux référentiels terminologiques que peut exploiter un moteur de recherche et qui permettent certaines fonctionnalités d'assistance intelligente à la recherche et à l'exploitation de l'information. La cinquième partie portera sur les autres fonctionnalités que peut comporter un moteur de recherche, toujours dans la perspective d'une assistance intelligente à la recherche d'information. Enfin, nous conclurons par l'évocation des « points de vigilance » identifiés lors de notre mission de stage et au long de la rédaction de ce mémoire.

Chaque fois que cela sera pertinent, le contenu de ce mémoire sera mis en perspective avec la question de l'assistance intelligente à la recherche d'information dans le contexte professionnel à l'origine de cette réflexion.

La notion de pertinence en recherche d'information, bien qu'abordée en elle-même dans la section 2, sera également une notion transverse de ce mémoire. Parce que rechercher une information, c'est implicitement exprimer un besoin d'information, et donc rechercher une information pertinente en regard de ce besoin.

Note préliminaire : dans ce mémoire les termes de « ressource » et de « document » sont utilisés pour désigner un même objet. Cela pourrait engendrer des confusions. Nous tenons donc à préciser que nous avons maintenu l'utilisation du terme ressource dans la première partie qui expose le cas pratique à l'origine de ce mémoire parce que c'est avec ce terme qu'ont été rédigés les documents relatifs à la mission professionnelle correspondante (Fiche de projet, Note de Cadrage, Cahier des Charges, etc.). Pour la suite, nous avons privilégié le terme de « document », parce qu'il était majoritairement utilisé dans les ressources qui ont été consultées pour l'élaboration de ce mémoire.

Pour la clarté du propos, nous tenons ici à préciser ce que nous entendons par l'une et l'autre notion. Le concept de « document » est compris comme un terme désignant un support contenant une ou des informations. Le concept de « ressource » est entendu comme une acception de document qui en définit des usages : une ressource est un document dont les informations sont extraites et exploitées pour atteindre un ou des objectifs.

Première partie :
Le contexte

1 Le contexte professionnel à l'origine de ce mémoire

Avant d'entamer notre réflexion sur la manière dont un moteur de recherche peut aider à la diffusion et à l'émergence de connaissances au sein d'une organisation, il nous paraît indispensable de revenir sur la mission professionnelle qui a été à l'origine de cette réflexion. En effet, les notions qui seront abordées dans ce document seront mises en perspective avec les objectifs de la mission et la problématique qui en a découlé.

Dans un premier temps, nous allons décrire le contexte du déroulement de la mission et la nature de la demande initiale. Dans une deuxième partie, nous présenterons le projet tel qu'il a été finalisé et retranscrit dans un cahier des charges.

1.1 Les spécificités du contexte

Décrire le contexte dans lequel s'est déroulée la mission à l'origine de ce mémoire est important dans la mesure où tout contexte professionnel produit ses propres contraintes et attentes, ici à l'égard d'un projet de moteur de recherche. Ces contraintes et attentes vont déterminer la définition du projet. Dans le cas considéré, trois contextes propres à la société Ostéobio sont à prendre en compte : le contexte organisationnel, le contexte informationnel et le contexte scientifique.

1.1.1 Le contexte organisationnel

Ostéobio a été créée en 1988.
La société regroupe trois entités :

- Une école privée d'ostéopathie biomécanique ;
- Des pôles de recherche en ostéopathie répartis par spécialités ;
- Des activités cliniques au sein et hors de la structure.

Elle a établi un partenariat privilégié avec Cogitobio, dont l'objet est le transfert vers le monde industriel de technologies issues de la recherche fondamentale sur le système musculo-squelettique.

Ces quatre entités sont concernées par le projet de moteur de recherche. De ce fait, ce projet aura plusieurs publics dont les attentes et les objectifs d'utilisation du moteur peuvent connaître quelques distinctions :

- Les étudiants ;
- Les enseignants ;
- Les chercheurs ;
- Les partenaires de recherche ;
- Les internautes (à terme, lorsque le moteur de recherche sera accessible depuis l'Internet).

1.1.2 Le contexte informationnel

L'essentiel de la production documentaire d'Ostéobio est constitué des mémoires des étudiants de quatrième et cinquième année dont les sujets sont déterminés en fonction d'objectifs de recherche. Les documents correspondant à chaque mémoire comprennent le texte du mémoire proprement dit, ses ressources bibliographiques, ses fichiers de formule, des images scientifiques (scanners, IRM, schémas), un fichier de présentation et un poster.

Les mémoires sous forme numérique sont actuellement stockés dans des répertoires de dossiers sur deux postes informatiques. Ils sont classés par années ou par ordre alphabétique de leurs auteurs. Ils ne sont accessibles directement que par les personnes détenant ces postes informatiques. Les mémoires sous forme papier sont stockés dans des armoires et classés par années. Les sujets des mémoires correspondent en général à des axes de recherche validés par l'école et contiennent donc potentiellement des informations scientifiques à forte valeur ajoutée, voire innovantes. Il est donc important qu'ils soient faciles d'accès.

Outre les mémoires, les ressources propres à Ostéobio sont les documents pédagogiques fournis par les enseignants. Ces documents sont sous la forme de fichiers texte, PDF, PowerPoint, formules et images (nous n'entrons pas ici plus avant dans le détail des formats).

Les autres ressources concernées par le projet sont des articles issus de revues spécialisées principalement accessibles en ligne par l'intermédiaire des bases de données de fournisseurs de contenus scientifiques (Pubmeb, Thèses.fr, Cochrane, etc.). Ces bases de données pourront être interrogées par le moteur de recherche, car ces fournisseurs de contenu ont prévu la mise en place de protocoles d'interopérabilité des données permettant un tel « moissonnage ».

1.1.3 Le contexte scientifique

Il paraît ici important de préciser la méthode d'appréhension des troubles musculo-squelettiques (TMS) développée au sein d'Ostéobio et sur laquelle une partie de l'indexation et de la recherche de documents du moteur de recherche sera fondée.

Ce modèle fonctionne comme un langage de référence pour l'ensemble de la communauté scientifique d'Ostéobio.

Cette méthode est connue sous le nom de « modèle princeps ». La démarche du modèle princeps est transdisciplinaire, c'est-à-dire qu'il cherche à dépasser les cloisonnements entre les disciplines de l'ostéopathie en développant une approche qui repose sur :

- La définition de quatre niveaux d'observation (tissulaire, articulaire, segmentaire et plurisegmentaire) ;
- L'étude de la géométrie de ces quatre niveaux (formes, agencements et orientations spatiales) ;
- L'étude de l'interdépendance des géométries de ces quatre niveaux ;
- L'étude des forces qui s'exercent sur les paramètres géométriques et qui entraînent des contraintes sur les structures, et donc des déformations et des mouvements à ces quatre niveaux.

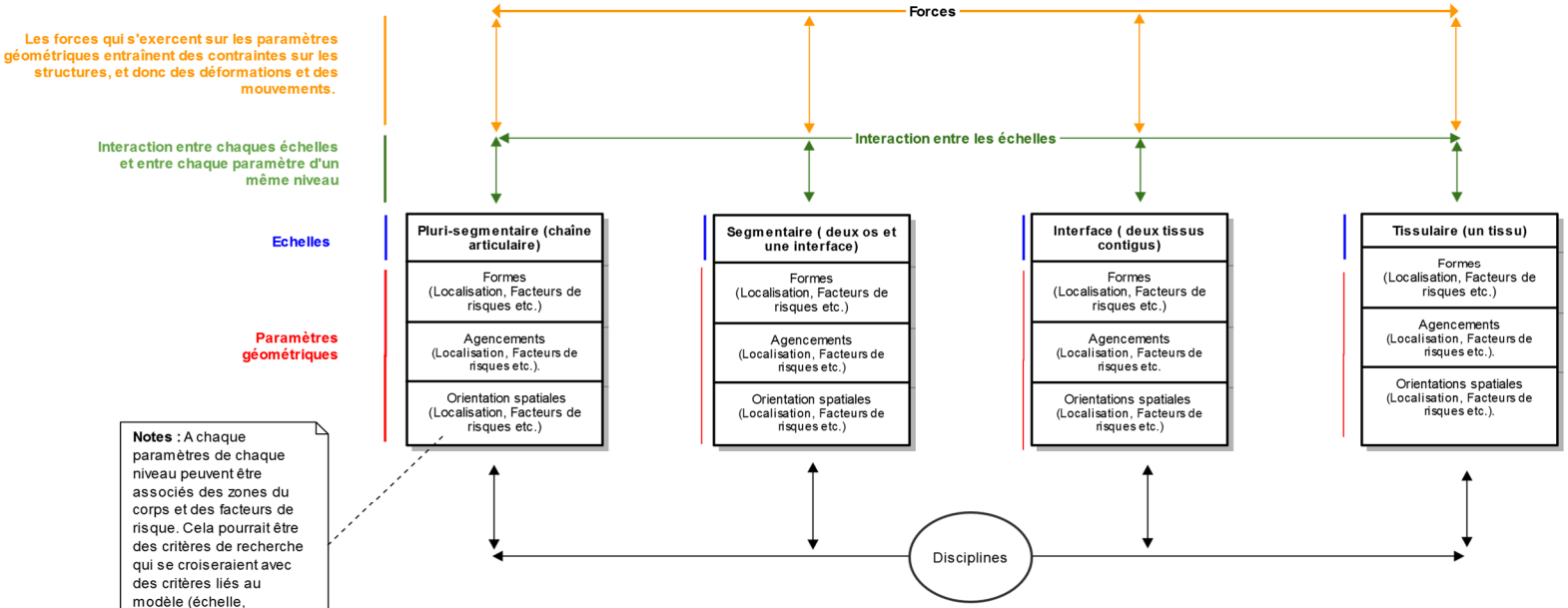
Cette méthode a pour but d'étudier les paramètres nécessaires et suffisants au bon déroulement des déformations des tissus et de leurs retours à une position « normale ».

Les disciplines « traditionnelles » sont prises dans le modèle et appartiennent à un ou plusieurs niveaux d'observation.

Le modèle est donc, de fait, disciplinaire et interdisciplinaire. Le schéma ci-dessous permettra au lecteur de mieux appréhender la logique du modèle princeps.

SCHEMA DU MODELE TRANSDISCIPLINAIRE :

Le modèle (ou méthode) princeps, dit aussi transdisciplinaire, repose sur :
- la définition de quatre niveaux d'observation (tissulaire, articulaire, segmentaire et pluri-segmentaire) ;
- l'étude de la géométrie de ces quatre niveaux (formes, agencements et orientations spatiales) ;
- l'étude de l'interdépendance des géométries de ces quatre niveaux ;
- l'étude des forces qui s'exercent sur les paramètres géométriques et qui entraînent des contraintes sur les structures, et donc des déformations et des mouvements.



Les disciplines de l'ostéopathie interviennent dans un ou plusieurs niveaux du modèle (ex, le viscéral concerne le niveau tissulaire, mais le crânien concerne le tissulaire et l'interface. La plupart des disciplines interviennent à plusieurs niveaux).
L'interaction entre les disciplines est prise en compte par le modèle transdisciplinaire.

Figure 1 — schéma du modèle princeps

1.2 Le projet

Après avoir mené une série d'entretiens avec les membres de l'équipe projet ainsi que des représentants des futurs utilisateurs, il a été possible de définir le projet selon trois attendus principaux :

Rendre accessibles les documents scientifiques produits au sein d'OSTEOBIO afin de :

- Permettre de savoir ce qui a déjà été produit sur un thème de recherche ;
- Faciliter l'accès des acteurs scientifiques d'ostéobio (chercheurs, enseignants et étudiants) aux ressources scientifiques.

Mettre en place un moteur de recherche qui permettra de rechercher et d'exploiter les documents scientifiques produits ou importés au sein d'OSTEOBIO selon les logiques du modèle princeps afin de :

- Systématiser l'approche transdisciplinaire dans la recherche et l'exploitation des documents ce qui permettra de :
- Révéler le caractère interdisciplinaire d'un document ;
- Faire le lien entre la pratique clinique et la recherche ;
- Faire émerger des documents des relations de causalité scientifique qui n'y sont pas exprimées de manière formelle.

Permettre une veille scientifique selon les critères du modèle princeps dans les domaines considérés :

- Dans les bases de données internes à Ostéobio ;
- Dans les bases de données des fournisseurs de contenus interrogées par la solution.

Nous verrons un peu plus bas en détail la façon dont il a été convenu de répondre à ces attendus.

Il nous paraît d'abord important de préciser que c'est par sa structure transdisciplinaire que le modèle princeps, s'il est utilisé comme principe d'indexation et d'interrogation des ressources, est susceptible de faire émerger des connaissances qui n'y sont pas formulées de manière explicite, mais y sont implicitement contenues. En effet, celui-ci n'aborde plus le fonctionnement du corps selon les logiques des disciplines ou des relations d'interdépendance entre les disciplines, mais selon une logique transversale qui lui est propre et qui intègre de nombreuses disciplines à chacun de ses niveaux d'organisation. Cette « vision des choses » permet de mettre en relation les concepts des disciplines dans le cadre de la logique propre au modèle. S'il exploite un référentiel terminologique reprenant le

fonctionnement du modèle, le moteur de recherche sera capable d'y rattacher les ressources interrogées et les concepts qu'elles contiennent et de les traiter selon sa logique.

Ainsi, lorsque le moteur de recherche trouvera dans un document deux concepts dont l'interaction produit un résultat, il sera capable de rattacher ces concepts à d'autres concepts, certes non présents dans la ressource, mais qui leur sont liés dans le modèle princeps. Cette mise en relation sera susceptible de mettre en évidence des liens de causalité qui n'apparaissent pas explicitement dans la ressource. Cette mise à jour de liens de causalité pourra représenter, pour les acteurs scientifiques d'Ostéobio, des pistes de recherche et donc d'innovations techniques ou thérapeutiques intéressantes.

Exemple : Dans la ressource, l'interaction de A valeur x et de B valeur y est cause de C valeur z (pathologique). Dans le modèle, la valeur de A dépend du facteur V. Donc il apparaît que la pathologie C valeur z est dépendante du facteur V.

1.2.1 Le projet tel qu'il a été finalisé

Face à l'ambition initiale du projet et compte tenu des développements technologiques (extraction de concepts, associations de concepts et de ressources, etc.) et sémantiques (référentiel linguistique de type thésaurus ou ontologie), et donc des investissements financiers qu'implique une telle solution, il nous a paru pertinent de le découper en trois niveaux de développement distincts :

- Un premier niveau d'usage où la recherche de ressources est envisagée sous le simple prisme d'une recherche par mots clefs contenus dans un glossaire et par critères de recherche classique de type bibliographique ou typologique. À ce niveau, il n'est pas envisagé d'exploitation de documents de type suggestion de termes de recherche associés, suggestions de ressources associées, extraction de concept des ressources et mise en relation de ces concepts avec le modèle princeps, etc.
- Un deuxième niveau d'usage ou la recherche de ressources est envisagé sous le prisme d'un thésaurus comportant relativement des relations hiérarchiques, des relations associatives de type « voir aussi » et la prise en compte de la synonymie des termes. L'implémentation de ce référentiel va permettre une recherche et une exploitation plus élaborées des ressources (possibilité d'utiliser certains critères issus du modèle princeps fonctionnant selon le principe hiérarchique dans la recherche avancée, par exemple). Il comporte aussi des fonctionnalités supplémentaires.

- Un troisième niveau d'usage où l'ensemble des fonctionnalités d'exploitation des ressources que nous avons évoquées est envisagé ainsi que de multiples fonctionnalités de type alerte (pour faire de la veille scientifique), recommandations et commentaires de ressources entre utilisateurs, etc. Ce niveau de développement doit être soutenu par un thésaurus portant des relations nombreuses et fines entre concepts, voire une ontologie. Outre l'exploitation de ressources, ce troisième niveau comporte de nombreuses fonctionnalités qui vont en enchérir le coût, car elles supposent des développements informatiques supplémentaires.

Les trois niveaux d'usage fonctionnent selon le principe « d'héritage des usages et fonctionnalités ». Le deuxième niveau reprend les usages attendus et les fonctionnalités envisagées du premier niveau, et le troisième niveau fait de même avec ceux des niveaux un et deux.

Ces niveaux sont présentés ci-dessous.

1.2.2 Projet niveau 1 : accéder aux ressources

1.2.2.1 Présentation du projet et des attentes relatives au niveau 1

Ce niveau d'usage correspond à la mise en place d'une solution de recherche « basique » dont les fonctionnalités de recherche et d'exploitation de documents sont communes à tous les types de moteurs de recherche. Elle vise à répondre à deux types de besoins :

- Permettre d'accéder facilement aux ressources scientifiques produites au sein d'Ostéobio ;
- Permettre de centraliser l'accès aux fournisseurs de contenu régulièrement consultés par les acteurs scientifiques d'Ostéobio en permettant à la solution de moissonner leurs bases de données selon :
    - Les termes scientifiques qui les décrivent (métadonnées) ou ceux qu'elles contiennent (plein texte) ;
    - Les critères bibliographiques, typologiques et techniques qui les décrivent.

Ces besoins concernent l'ensemble des acteurs scientifiques d'Ostéobio.

En regard des usages attendus qui seront faits des ressources et compte tenu des contraintes financières liées au projet, nous avons pu déterminer quels devaient être les modes et les critères de recherche des ressources.

1.2.2.2 Les critères de recherche du projet niveau 1

Les acteurs qui vont chercher un document en fonction de ces attentes vont principalement le faire selon trois types de critères :

- Critères liés à la description scientifique (métadonnées) ou aux contenus scientifiques des documents (expressions scientifiques) ;
- Critères liés au type de ressource (articles, documents pédagogiques, etc.) ;
- Critères bibliographiques (titre, date, auteur, etc.) ;
- Critères relatifs aux formats des documents (fichier texte, fichier de formule, fichier image, etc.).

Les critères de recherche les plus importants pour les utilisateurs seront les critères scientifiques, car c'est essentiellement sous le prisme de la science qu'ils ont besoin d'accéder aux documents et aux informations qu'ils contiennent.

Dans le cadre de la recherche scientifique ou pédagogique, il est très important que la solution prenne en compte les relations de synonymie entre les termes de recherche. En effet, les acteurs scientifiques peuvent passer à côté d'informations importantes parce qu'un concept de leur périmètre de recherche n'est pas nommé de la même manière selon qu'il est présent dans un document faisant référence à telle ou telle discipline.

1.2.2.3 Les modes de recherche du projet niveau 1

En conséquence des critères de recherche précédemment définis, les modes de recherche disponibles à ce niveau de développement de la solution sont :

- Une recherche simple, dans un champ de recherche, sans autocomplétion avec prise en compte des synonymes du terme de recherche et qui soit :
    - Insensible à la casse des termes de recherche ;
    - insensible à l'accentuation des termes de recherche.
    - Insensible à l'accord des termes de recherche.
- Une recherche avancée par critères, limitée aux critères typologiques, bibliographiques et de format et pouvant se croiser avec une recherche simple telle que décrite précédemment, le croisement de ces deux types de recherches devant pouvoir se faire :
    - Suite à une première recherche simple (l'utilisateur décide de passer en mode « recherche avancée » en conservant sa requête initiale).
    - Directement sans passer par une recherche simple (l'utilisateur n'effectue pas de recherche simple, mais compose une requête simple directement dans le formulaire de recherche avancée et la croise avec des critères de recherche avancée).

1.2.2.4 Les classements des résultats de la recherche principale du projet niveau 1

Le classement des documents en liste résultat doit tenir compte de :

- La présence des termes de requête dans la ressource ;
- La coprésence des termes de requête dans la ressource ;
- La proximité des termes de requête dans la ressource ;
- Le nombre d'occurrences des termes de la requête par rapport à la longueur des documents ;
- La similarité entre la requête et l'expression dans la ressource ;
- La position des termes de requête dans la ressource ;
- La proximité entre les termes de requête dans les documents.

1.2.3 Projet niveau 2 : la recherche avancée

1.2.3.1 Présentation du projet niveau 2

Le projet de moteur de recherche dans son niveau 2 de sophistication doit permettre une recherche plus fine des documents, la suggestion de concepts et de documents liés aux termes de la requête et un début d'utilisation du modèle princeps dans le mode avancé de recherche.

Les attentes liées à la recherche scientifique varient selon le profil des futurs utilisateurs de la solution.

À ces attentes scientifiques correspondent des fonctionnalités de recherche et d'exploitation des documents.

1.2.3.2 Les attentes et fonctionnalités correspondantes du projet niveau 2

1.2.3.2.1 Composer des équations de recherche

Les acteurs scientifiques d'Ostéobio ont besoin, dans le cadre de leurs recherches scientifiques, d'effectuer des recherches fines dans les documents qui seront indexés par la solution. Ces recherches doivent pouvoir leur permettre d'étendre, de restreindre ou de préciser une recherche. Ainsi ils peuvent avoir besoin de :

- Rechercher des documents dans lesquels des concepts scientifiques sont présents ;
- Rechercher des documents dans lesquels des concepts scientifiques sont conjointement présents ;
- Rechercher des documents dans lesquels certains concepts scientifiques sont présents alors que d'autres ne le sont pas ;

- Rechercher des documents dans lesquels des concepts sont mentionnés de manière rapprochée ou éloignée ;
- Rechercher des mesures scientifiques selon leurs valeurs ;
- Rechercher des expressions exactes ;
- Rechercher des termes commençant par un même préfixe ou contenant un même radical.

Les acteurs scientifiques d'Ostéobio ont besoin de pouvoir croiser ces modes de recherche afin de trouver des informations ou des processus scientifiques précis dans le cadre de leurs activités scientifiques. La solution doit donc permettre, dans l'idéal, d'enchâsser les opérateurs de recherche.

1.2.3.2.2 Accéder à des concepts associés et des documents relatifs à ces concepts

Accéder à des concepts associés aux termes de requêtes et à des documents relatifs à ces concepts va permettre aux utilisateurs de commencer à faire de l'inter et de la transdisciplinarité (modèle princeps) dans leurs recherches de documents et donc dans leurs réflexions scientifiques. En effet, en se voyant suggérer des concepts associés et des concepts hypo et hypéronimiques, les utilisateurs, non seulement verront quels sont les liens scientifiques pertinents pour leurs recherches, mais ils pourront aussi exploiter ces liens en consultant des documents suggérés. Les contenus des documents pourront alors être exploités relativement aux liens entre concepts qui ont été faits ou à la recherche de l'utilisateur.

1.2.3.2.3 Visualiser les concepts contenus dans un document

La visualisation des concepts scientifiques contenus dans un document sans que ces concepts soient mis en relation permettra aux utilisateurs :

- D'avoir une première approche synthétique du contenu scientifique du document ;
- D'associer les concepts entre eux sans passer par un système de connaissance ou une hiérarchie préétablie.

L'approche synthétique du contenu d'un document sera pour l'utilisateur un moyen de voir rapidement quels sont les concepts mobilisés dans le raisonnement ou l'exposé contenu dans la ressource.

Présenter les concepts contenus dans un document sans que soient établies de relations entre eux permettra aussi aux utilisateurs de faire des raisonnements ou des « découvertes » par association « libre ». En effet, si les concepts extraits sont présentés selon un modèle de connaissance préexistant, ou selon la hiérarchie ou la logique dans laquelle ils figurent dans la ressource, cela peut certes permettre de faire des associations, mais uniquement dans le cadre du modèle dans lequel ils

sont situés. Les présenter en dehors de toute relation préétablie peut permettre aux utilisateurs d'établir des relations de causalité (internes au document ou appelant des concepts non mentionnés dans le document) qui ne sont pas envisagées dans la ressource. Cela peut être source d'innovation scientifique.

1.2.3.2.4 Utiliser les échelles d'observation du modèle princeps comme critère de recherche.

À ce niveau de sophistication de la solution, il est aussi possible d'envisager une utilisation du modèle princeps proprement dit pour rechercher des documents. C'est en effectuant une recherche avancée utilisant des critères de recherche scientifique issus du modèle que cette première approche pourra se faire. Cela se fera à partir d'une représentation simplifiée du modèle, mais permettra tout de même de le décrire en partie et donc d'utiliser la logique d'échelle, qui est en partie hiérarchique, pour rechercher des documents (voir *supra* « schéma du modèle princeps » et « contexte scientifique du projet »).

Ensuite, l'utilisateur doit pouvoir utiliser des sous-critères de recherche dépendants des critères d'échelle précédemment cités. La valeur de ces sous-critères doit être conditionnée par le choix du premier critère de recherche avancé.

Dans ce contexte, l'utilisation de critères et sous-critères de recherche scientifique ne peut pas se croiser avec la composition d'une requête libre. Il faut prévoir que l'utilisation de l'un empêche l'utilisation de l'autre.

Outre le fait de pouvoir chercher des documents à partir de la logique du système princeps, ces fonctionnalités de recherche avancée permettront aux utilisateurs qui ne maîtrisent pas le modèle de se familiariser avec sa logique et son utilisation. C'est une manière de faciliter son appréhension et son appropriation que d'inciter les utilisateurs à réfléchir d'abord selon une logique d'échelle

1.2.4 Projet niveau 3 : exploitation du modèle princeps

Ce niveau 3 de sophistication du projet correspond à la mise en place d'une solution qui vise à permettre de rechercher et d'exploiter les documents selon la logique du modèle princeps, notamment par la mise en relation des documents et des concepts qui y sont contenus avec des concepts du modèle ou d'autres concepts, mais toujours selon la logique du modèle.

1.2.4.1 Utiliser le modèle princeps pour rechercher des documents scientifiques

L'utilisation du modèle princeps va permettre aux utilisateurs d'utiliser la logique du modèle pour rechercher des documents.

Même si le modèle n'est pas décrit dans les documents, cette utilisation est possible, car le modèle intègre, selon sa propre logique, les concepts contenus dans les documents.

Ex. : la ressource contient les concepts « rupture des ligaments croisés » et « torsion ». Ces concepts sont présents dans le modèle et rattachés à des concepts propres au modèle.

L'utilisation du modèle princeps pour rechercher des documents scientifiques doit pouvoir se faire au niveau :

- De la recherche simple ;
- *De la recherche avancée.*

1.2.4.1.1 Utilisation du modèle princeps au niveau de la recherche avancée

Au niveau de la recherche avancée, l'utilisateur doit pouvoir utiliser des critères et sous-critères de recherche propres au modèle princeps.

Ces critères et sous-critères de recherche doivent être liés par une relation de dépendance, c'est-à-dire que :

- La sélection d'un critère de recherche doit pouvoir conditionner la valeur d'un sous-critère de recherche ;
- La sélection d'un critère de recherche doit pouvoir activer la disponibilité d'un sous-critère de recherche (et inversement, la non-sélection d'un critère doit pouvoir empêcher la sélection d'un sous-critère).

Suite à la formulation d'une première requête, l'utilisateur doit retrouver ces critères et sous-critères de recherche sous forme de facettes et les sélectionner afin de discriminer les résultats obtenus.

1.2.4.1.2 Utilisation du modèle princeps au niveau de la recherche simple.

Dans le cadre de la recherche simple, il ne s'agit pas d'imposer à l'utilisateur des critères de recherche relatifs au modèle princeps, mais plutôt de lui en suggérer l'utilisation.

Ainsi, la solution devra pouvoir bénéficier d'une fonctionnalité « d'autocomplétion tolérante » dont certains termes relèveront du modèle princeps.

Le modèle reprenant par ailleurs, comme nous l'avons dit précédemment, les concepts des différentes disciplines de l'Ostéopathie en les associant à ses propres concepts, il sera important que l'autocomplétion proposée par la solution s'applique à plusieurs termes de la requête, éventuellement séparés par un opérateur de recherche et à des expressions composées de plusieurs termes.

1.2.4.1.3 Élargir une thématique de recherche

Élargir une thématique de recherche, c'est, pour les chercheurs d'Ostéobio, réfléchir selon le modèle princeps, soit de façon « transversale » à partir d'un thème de recherche initial. Ici, il est demandé à ce que ce soit la solution qui « produise de la transversalité ». Il faut donc que, à partir d'une requête utilisateur, la solution :

- Propose à l'utilisateur des thèmes de recherche liés à sa requête, selon la logique du modèle princeps ;
- Suggère la consultation de documents en rapport avec les thèmes de recherche qui ont été proposés à l'utilisateur.

Grâce à cette fonctionnalité, les utilisateurs pourront :

- Être plus efficaces dans leurs recherches scientifiques, notamment en ce qui concerne les enseignants-chercheurs ;
- Bien saisir la logique du modèle princeps en ce qui concerne les étudiants.

1.2.4.1.4 Exploiter les résultats d'une recherche selon le modèle princeps

Comme nous l'avons dit *supra* dans ce document, les documents scientifiques contiennent parfois des connaissances qui n'y sont pas mentionnées de manière explicite, parce qu'elles n'ont pas été « identifiées comme telles » ni par les lecteurs, ni même par les auteurs du document. Utiliser le modèle princeps pour exploiter les documents permettra d'en faire émerger des relations de causalité scientifique qui n'y sont pas exprimées de manière formelle.

Pour cela, il faut que la solution permette :

- D'extraire les concepts contenus dans les documents ;
- De mettre les concepts extraits des documents en relation avec d'autres concepts selon la logique du modèle princeps ;
- De suggérer la consultation de documents en rapport avec les concepts qui ont été suggérés.

Là aussi d'une certaine manière, il s'agit que la solution « produise de la transversalité » à la place des utilisateurs, mais à partir du contenu des documents.

Cela permettra aux utilisateurs d'exploiter tout le potentiel conceptuel et applicatif que contient un document.

Une autre manière d'extraire des connaissances non identifiées des documents est de permettre aux utilisateurs de modifier les critères par défaut de hiérarchisation des résultats. Rechercher un concept rare dans un document plutôt que son occurrence peut faire émerger des liens de causalité non encore découverts.

1.2.4.1.5 Éviter les confusions dues à la polysémie de certains concepts

Certains concepts des différentes disciplines prises par le modèle princeps peuvent avoir plusieurs sens selon le contexte dans lequel ils sont employés.

Cette polysémie peut entraîner :

- De la confusion si les utilisateurs n'en ont pas pleinement conscience ;
- Du bruit dans les résultats de recherche.

Il est donc important que la solution dispose d'une fonctionnalité qui signale à l'utilisateur que son terme de requête renvoie à plusieurs significations et lui propose de choisir celle qu'il souhaite prendre en compte dans sa recherche.

1.2.4.1.6 Faire de la veille scientifique

Ostéobio est une école, mais comprend aussi des pôles de recherche et travaille en partenariat avec une société dont l'objectif est le transfert des connaissances vers des applications industrielles.

Il est donc important pour ses acteurs scientifiques de se tenir au courant de la production scientifique relative à certains thèmes de recherche.

Pour permettre ces opérations de veille, la solution doit disposer de fonctionnalités d'alertes paramétrables par les utilisateurs et qui doivent pouvoir s'appliquer :

- Aux documents provenant d'Ostéobio ;
- Aux documents proposés par les fournisseurs de contenu dont la solution « moissonne » les bases de données.

Deuxième partie :
La recherche d'information

2 Qu'est-ce que la recherche d'information ?

Avant d'aborder l'étude des fonctionnalités d'assistance à la recherche d'information et des bases de connaissance sous-jacentes aux moteurs de systèmes de recherche d'information (SRI) qui permettent certaines de ces fonctionnalités, il nous a paru intéressant de nous pencher sur ce qui définit une recherche d'information pour les utilisateurs. Comprendre en effet ce qui la motive, ce qu'elle vise et les étapes qui la constituent nous paraît fondamental si l'on souhaite trouver les meilleurs moyens pour l'assister.

2.1 Plusieurs définitions

Certains organismes professionnels francophones tels que l'ADBS dans « Le Vocabulaire de la documentation » (Paris, ADBS, 2004) proposent une double définition en distinguant la recherche d'information et la recherche de l'information[2] :

- Recherche d'information : « Ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'usager, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés. Toute recherche d'information suppose trois phases successives : a) une recherche bibliographique des références de documents pertinents ; b) une recherche documentaire, c'est-à-dire une recherche bibliographique complétée par la recherche (l'acquisition) des documents eux-mêmes ; c) et enfin le repérage de l'information dans les documents sélectionnés (recherche de l'information). »
- Recherche de l'information : « Ensemble des méthodes, procédures et techniques ayant pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes. »

La première définition met en avant la recherche des documents contenant l'information dont a besoin celui qui cherche et les techniques pour accéder à ces documents. Elle se rapproche de la recherche documentaire définie par l'ADBS comme l'« ensemble des méthodes, procédures et techniques ayant pour objet de retrouver des références de documents pertinents (répondant à une demande d'information) et les documents eux-mêmes ».

[2] http://www.techno-science.net/?onglet=glossaire&definition=11203

Il nous semble que la « recherche de l'information », la deuxième définition, constitue l'objectif principal de la majorité des « recherches d'informations », les utilisateurs souhaitant la plupart du temps s'affranchir, ou tout au moins simplifier au maximum les étapes qui la jalonnent. Et c'est précisément l'objet de ce mémoire que d'étudier la façon dont les SRI, et plus particulièrement les moteurs de recherche, peuvent aider les utilisateurs à franchir ces étapes, voire s'en affranchir.

Il importe ici de signaler que, comme le font remarquer Nicole Boubée et André Tricot, la recherche d'information n'est pas une fin en soi pour l'utilisateur d'un moteur de recherche : elle s'effectue dans le cadre d'une autre activité ou tâche, dans le cadre d'un contexte. Il est essentiel de prendre en compte ce contexte dès lors que l'on souhaite étudier « comment les humains recherchent »[3]. C'est en tenant compte de ce contexte de la recherche d'information que l'on pourra le mieux assister les chercheurs d'information. Cette notion de contexte est large, mais si l'on s'abstrait de son traitement par le système (que nous aborderons dans la dernière partie), on peut définir le contexte de la recherche d'information comme : les objectifs et les métaobjectifs assignés à la recherche d'information dans le cadre d'une organisation productrice de normes explicites et implicites, d'un système de références intellectuelles (raisonnements partagés) et culturelles (significations partagées) et d'un système sociotechnique d'interrogation de sources et de documents.

Mais au-delà du contexte, ou plutôt, quel que soit le contexte, est-il possible de définir la recherche d'information en tant que telle ?

2.2 La recherche d'information du point de vue de l'utilisateur

La recherche d'information a été longuement étudiée et de nombreuses modélisations en ont été proposées. Il ne s'agit pas ici d'opposer ces modèles, ni de chercher lequel est le plus juste, mais plutôt de voir en quoi ils sont complémentaires et ce qu'ils peuvent nous dire sur la place de l'utilisateur dans le processus de recherche d'information.

En effet, selon que l'on se place du point de vue de l'utilisateur ou du point de vue technique, la recherche d'information ne se modélise pas de la même façon. Il nous intéresse ici de nous placer du point de vue de l'utilisateur, et plus précisément du point de vue de ses objectifs lorsqu'il procède à une recherche d'information.

---

[3] Arlette Boulogne, Sylvie Dalbin à à propos l'ouvrage de Nicole Boubée et André Tricot. Qu'est-ce que rechercher de l'information ? – Villeurbanne : Presses de l'Enssib, 2010. – 286 p. – (Papiers. Série Usages des documents). – ISBN 978-2-910227-83-8 http://www.adbs.fr/qu-est-ce-que-rechercher-de-l-information--101173.htm?RH=ACCUEIL.

De ce point de vue, on peut dire que la recherche vise à répondre à une question plus ou moins précise, à découvrir de nouvelles connaissances, à réduire l'incertitude liée à un déficit ou à une surcharge d'informations, à palier un état de connaissance transitoire ou insatisfaisant, ou encore à confirmer ou infirmer une croyance (entendue ici comme une connaissance non étayée). [2, Vuillequiez].

Cette définition assez complète de la recherche d'information du point de vue de l'utilisateur permet de l'envisager comme une activité de type « résolution de problème ».

2.2.1 La recherche d'information vue comme une activité de résolution de problème

La recherche d'information définie comme une activité de type « résolution de problème » peut se définir ainsi : l'utilisateur part d'un état initial, généralement mal défini, et doit atteindre un état final, c'est-à-dire un but en respectant des contraintes matérielles et temporelles et en passant par des sous-objectifs intermédiaires [1, Hérigault].

Cette définition, très « formelle » (qui s'attache aux procédures) de la recherche d'information vue comme une activité de type « résolution de problème » peut être complétée par la définition de Sharit *et al.* qui l'assimilent à une situation de résolution de problème « flou » dont les contours sont mal définis et dans laquelle l'utilisateur élabore une représentation mentale du problème à résoudre sous l'influence des connaissances qu'il a du domaine et de la nature du besoin informationnel (complexe ou non, structuré ou flou). On retrouve là, en partie, la notion de contexte évoquée et définie précédemment à travers la notion de « connaissance » [3, Dinet].

Ces deux définitions complémentaires présentent toutefois l'inconvénient de n'envisager qu'un seul « état initial » de l'utilisateur relativement à son objectif de recherche d'information : « flou » ou encore « mal défini ».

Marchionini distingue, quant à lui, trois types de recherche d'information qui dépendent précisément de l'état initial du besoin de l'utilisateur.

2.2.2 Les trois recherches d'information selon Marchionini

Les trois types de recherche selon Marchionini sont discriminés selon que ces recherches sont non dirigées, semi-dirigées ou dirigées [14, Desfriches].

2.2.2.1 Le « browsing » (« navigation » en français soit une recherche non dirigée)

Dans une activité dite de « browsing », l'usager consulte de nombreuses sources, plus ou moins superficiellement, dans une stratégie dite « par butinage » [3, Dinet]. Il s'investit dans « une exploration associative » et laisse les informations émerger. Son objectif n'est pas clairement défini.

S'il est peu probable que les utilisateurs que nous considérons dans notre cas pratique cherchent à résoudre des « problèmes flous » (en effet, les recherches des chercheurs et des étudiants ont plutôt des objectifs de recherche précis), c'est la deuxième partie de la définition qui nous intéresse ici. Car même dans l'optique d'une recherche précisément définie, les utilisateurs sont susceptibles de naviguer à travers des liens en utilisant les suggestions de concepts (associés aux termes de requêtes ou aux concepts contenus dans les documents) et de documents. Et cette navigation par liens, qui introduit une forme de sérendipité[4] associative, est potentiellement source de découvertes scientifiques (un peu à la façon dont Flemming a découvert les effets antibactériens de la pénicilline).

2.2.2.2 Le « searching » (recherche semi-dirigée)

Dans une recherche « semi-dirigée », dite de « searching », l'usager est dans une démarche de type « exploratoire », et consulte plusieurs sources (web, BDD, blog, forums, annuaires, etc.), en utilisant plusieurs systèmes d'organisation des connaissances (moteurs de recherche, classifications, langages documentaires, etc.) et en mobilisant ses connaissances préalables du domaine, pour « raisonner par inférences et abduction » dans une stratégie analytique.

Ce type de recherche d'information peut correspondre à une recherche scientifique qui s'initie, dont l'objet, le périmètre et le contenu ne sont pas encore totalement définis.

2.2.2.3 Le « queyring » (« requête » en français soit une recherche dirigée)

Dans une activité dite de « querying », l'usager a tendance à interroger un moteur de recherche en formulant une requête de type « chaîne de caractère », issue du langage naturel ou d'un système

[4] Ici, la sérendipité est le fait de trouver information de façon inattendue à la suite d'un concours de circonstances fortuit et très souvent dans le cadre d'une recherche concernant un autre sujet. La *sérendipité* est le fait de « trouver autre chose que ce que l'on cherchait »

d'organisation des connaissances (thésaurus, ontologie ou taxonomie). Le queyring correspond au modèle computationnel classique autorisé par les systèmes documentaires (opérateurs booléens, langages documentaires, etc.), il se situe dans une logique de consultation et utilise des raisonnements hypothético-déductifs.

Ici, les recherches sont très définies. Elles visent à répondre à des questions précises dans le but de valider ou invalider des hypothèses ou des résultats d'expérience. Elles demandent des connaissances expertes du domaine telles qu'elles sont par exemple envisagées dans notre cas pratique et dans l'hypothèse de la recherche avancée qui utilise les critères du modèle de connaissance.

La distinction de ces trois types de recherche d'information nous permet de valider à postériori nos hypothèses de travail. Ce modèle nous permettra aussi, si un tel cas se représente à nous, de disposer d'une grille d'analyse sur laquelle fonder les différents cas d'utilisation à considérer et donc les fonctionnalités à envisager pour y répondre au mieux.

On voit bien à travers ces différentes définitions que la recherche d'information ne peut être considérée en tant que telle sans prendre en compte celui qui la conduit (l'utilisateur) et le contexte dans lequel il évolue tel que nous l'avons défini. Pour aller plus loin et comprendre comment se structurent ces recherches d'information et à quelles lois elles obéissent, de nombreux auteurs ont voulu en modéliser le déroulement**.**

2.2.3 La recherche d'information modélisée

Parce qu'elle est au cœur des activités humaines et qu'elle comporte des dimensions stratégiques liées à ces activités, la recherche d'information intéresse beaucoup de disciplines qui ont tenté d'en rendre compte depuis leur point de vue. Parmi celles-ci nous pouvons citer [3, Dinet] :

- Les approches conceptuelles et méthodologiques qui cherchent à en restituer les processus objectifs et subjectifs ;
- Les approches psycho-ergonomiques qui se concentrent sur l'"interaction entre les utilisateurs et le système sociotechnique ;
- Les approches économiques qui visent à en maximiser la rentabilité et en minimiser les risques ;
- Les approches informatiques qui abordent la notion de pertinence de l'information d'un point de vue mathématique ;
- Les approches issues de la robotique qui ont pour objectif d'améliorer les comportements d'une machine et ses interactions avec l'environnement pour améliorer son activité de recherche d'information.

Si toutes ces approches sont intéressantes et riches d'enseignements quant aux besoins en assistance des utilisateurs de SRI, nous ne pouvons toutes les aborder dans ce mémoire. Nous avons donc décidé de nous intéresser aux approches conceptuelles et méthodologiques. Ce sont ces types d'approches qui nous paraissent être les plus proches de nos préoccupations. En effet, il nous semble que c'est en étudiant les objectifs et les étapes qui constituent les processus de RI ainsi que le rôle de la subjectivité des utilisateurs dans ces processus que nous pourrons être le plus à même de répondre à la question de l'assistance intelligente à la recherche d'information.

Il est important de préciser ici que, de manière générale dans ce mémoire, nous nous situons dans le cadre de recherches d'informations qui se déroulent dans des environnements numériques.

Les approches conceptuelles et méthodologiques s'intéressent en priorité aux utilisateurs, c'est-à-dire à ceux qui effectuent des tâches de recherche d'information. Elles ont pour objectif de rendre compte de leurs comportements, de leurs difficultés et des stratégies qu'ils déploient dans le cadre de cette activité. Nous ne pouvons pas ici présenter l'ensemble, ni même un grand nombre de modèles théoriques de la recherche d'information. Nous avons décidé d'en sélectionner quatre avec, comme critère de sélection, le fait qu'ils soient pertinents et instructifs relativement à notre cas pratique.

2.2.3.1 Le modèle itératif de Marchionini

Comme nous l'avons vu précédemment, le modèle itératif de Marchionini distingue trois types de recherche d'information [14, Desfriches,] :

- Dirigée. L'individu a alors un but précis et ses comportements sont orientés pour atteindre ce but selon une procédure stricte.
- Semi-dirigée. L'individu n'a qu'une idée approximative de l'information qu'il recherche et ses comportements visent à trouver l'information la plus proche possible de ce qu'il pense devoir obtenir.
- Non dirigée. L'individu n'a pas de but possible et va se contenter de butiner au gré des liens qu'il rencontre et de l'idée qu'il se fait de son besoin.

Selon Marchionini, quel que soit le type de RI auquel il procède, l'activité de l'utilisateur passe par huit étapes successives [3, Dinet] :

- La reconnaissance et l'acceptation qu'une information est nécessaire pour résoudre un problème ;
- La définition et la délimitation du problème à résoudre ;
- Le choix d'un système et/ou d'une source pour rechercher l'information ;
- La formulation d'une requête ;

- La réalisation effective de la recherche d'information ;
- L'examen des résultats et des informations proposées par le sri ;
- L'extraction de l'information jugée utile parmi les informations retournées par le sri ;
- L'évaluation du résultat obtenu et la reprise d'une étape antérieure du processus si ce résultat n'est pas jugé satisfaisant.

Selon Dinet les principaux apports du modèle de Marchionini sont [3, Dinet] :

- D'avoir posé l'activité de recherche d'information comme étant une activité humaine primordiale, notamment lorsqu'elle est réalisée dans des environnements numériques ;
- D'avoir insisté sur le fait que la ri est un processus interactif mettant en jeu, d'une part, un système technologique et, d'autre part, un système humain via une interface au sein d'un environnement informationnel partagé ;
- D'avoir considéré que la ri était un processus essentiellement itératif et que l'interaction entre le système technologique et le système humain y jouait un rôle important.

Il est donc fondamental pour Marchionini d'appréhender l'environnement informationnel dans lequel interagissent les deux systèmes, technologique et humain, si l'on veut comprendre les difficultés des usagers et concevoir des moyens de les assister.

La dernière étape du modèle de Marchionini nous intéresse particulièrement en ce qu'elle pose que la recherche d'information est un processus itératif qui se répète jusqu'à l'obtention d'une réponse au besoin. Or l'itération, si elle est importante en recherche d'information, l'est tout autant en recherche scientifique. Cette parenté peut s'inscrire autour de ces questions :

- Qu'est-ce que l'utilisateur cherche ?
- Est-ce qu'il manipule les bons concepts pour chercher ce qu'il cherche ? Quels sont les concepts qui vont lui permettre d'obtenir l'information qu'il souhaite ?
- Si sa première requête n'aboutit pas, comment faut-il qu'il la reformule, avec quels nouveaux concepts, avec quelles restrictions et/ou extensions, pour obtenir les résultats escomptés ?

Mais aussi, et cela nous rapproche de la notion de sérendipité :

- Pourquoi est-ce que cette requête donne ces résultats ?
- Que disent ces résultats, même si ce ne sont pas ceux auxquels s'attendait l'utilisateur ?

Or, si l'on peut concevoir que s'il y a parenté de cheminement cognitif entre les deux processus de recherche, scientifique et d'information, il est possible de penser que le système de l'un peut soutenir l'activité de l'autre.

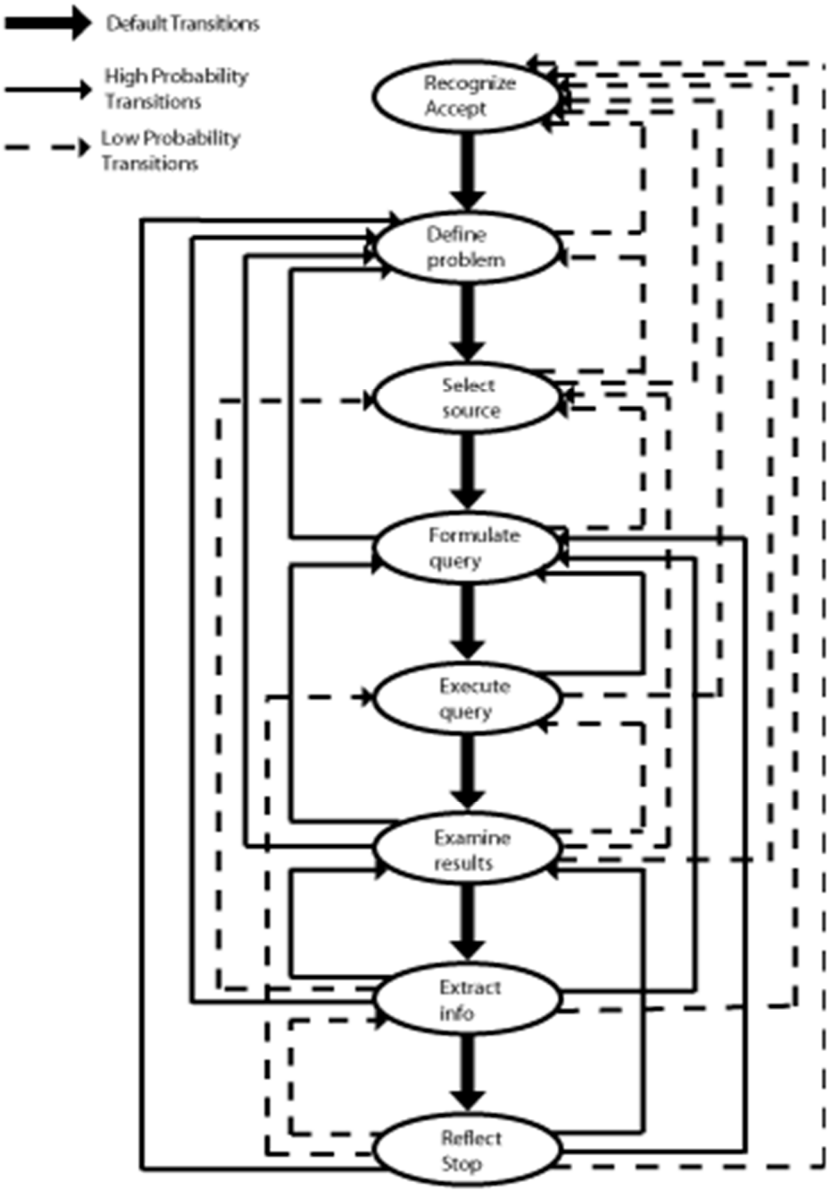


Figure 2 — Les huit étapes essentielles de recherche d'information selon Marchionini[5].

[5] http://www.jonasfransson.com/6-the-search-process-and-the-search-query/

2.2.3.2 Le modèle holistique de Kuhlthau

Le modèle holistique de Kuhlthau a pour ambition de décrire les comportements et les processus internes lors d'une recherche d'information. Il vise à rendre compte de la totalité des dimensions impliquées et/ou affectant les étapes d'une recherche d'information.
Selon Kuhlthau, il faut distinguer deux axes indépendants [3, Dinet] :

- D'une part les sept étapes présentes dans une recherche d'information :
    - L'identification d'un besoin informationnel ;
    - La sélection d'une source et/ou d'un document d'information ;
    - L'exploration de ces sources et documents en vue d'extraire des précisions afin de mieux délimiter les contours du besoin informationnel ;
    - La formulation de concepts, généralement traduits en mots clefs, comme critères de recherche ;
    - La collecte d'informations ;
    - La restitution des informations collectées (leur présentation) :
    - L'évaluation du produit final (présentation orale, rapport, etc.) Ainsi que de la procédure et de l'activité de recherche d'information ;
- D'autre part, trois types de facteurs qui peuvent influencer ces sept étapes :
    - Les aspects physiques, c'est-à-dire, plus précisément, les aspects moteurs ;
    - Les affects, c'est-à-dire les sensations et les sentiments ressentis par l'usager lors des différentes étapes (peur de ne pas trouver l'information pertinente, découragement ou au contraire enthousiasme et sentiment d'être sur la bonne voie, etc.) ;
    - La cognition (les mécanismes « rationnels », ou tout du moins perçus comme tels, de la pensée).

Ce modèle présente deux intérêts majeurs :

- L'activité de RI est considérée de la phase d'initiation (définition d'un besoin informationnel) à la phase d'utilisation-restitution des informations collectées et à leur évaluation ;
- Des facteurs relevant de la motricité, des affects ou des capacités cognitives sont jugés susceptibles d'influer sur le processus de RI.

On peut même, avec Kuhlthau, considérer que ces trois derniers facteurs doivent être pensés en interaction les uns avec les autres (le surgissement d'affect influençant bien évidemment les capacités

cognitives, par exemple) ainsi qu'en étroite relation avec les profils des utilisateurs (dans notre situation, un chercheur en phase d'initiation de recherche n'aura pas les mêmes capacités cognitives qu'un étudiant en phase de fin rédaction de mémoire ou de préparation d'examen).

C'est tout l'intérêt de ce modèle holistique, qui se propose d'appréhender la totalité des comportements et processus intellectuels durant toutes les phases de RI [3, Dinet]. Indirectement, il affirme l'importance du contexte dans l'analyse des processus de recherche d'information. Les aspects physiques, affectifs et cognitifs sont de l'ordre du contexte utilisateur, l'identification du besoin informationnel ou l'évaluation de la restitution font partie du contexte organisationnel.

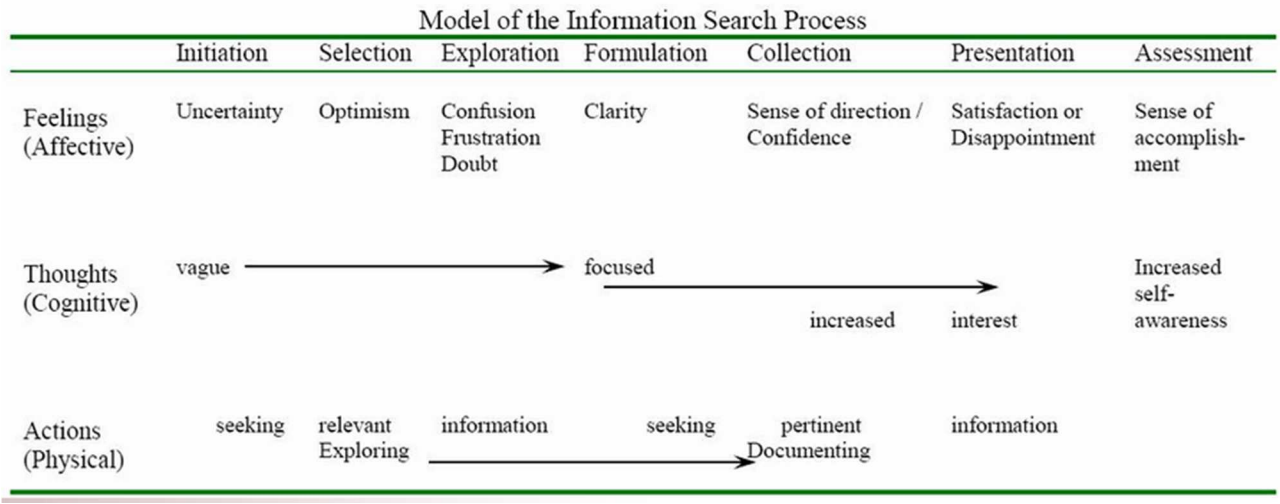


Figure 3 — Les sept informations et trois dimensions de la RI selon Kuhlthau[6]

Les deux modèles que nous venons de présenter sont essentiellement descriptifs et qualitatifs. Quand ils sont transposés dans des observations *in situ*, ils fournissent des renseignements intéressants sur la façon dont se déroulent les recherches d'information. Toutefois, certaines notions qu'ils utilisent restent floues. Ainsi, les mécanismes mentaux ou les affects dont il est question dans le modèle de Kuhlthau ne sont pas clairement définis. Enfin, d'un point de vue méthodologique, très peu de données expérimentales sont venues en étayer les conceptions théoriques.

[6] https://comminfo.rutgers.edu/~kuhlthau/information_search_process.htm

D'autres modèles visent à expliquer les comportements des individus qui recherchent des informations en mettant l'accent sur les mécanismes mentaux sous-jacents qui les influence.

2.2.3.3 Le modèle centré sur les compétences de Brand-Gruwel

Ce modèle propose une approche centrée sur les compétences. Il a trouvé un écho favorable auprès des pédagogues et des didacticiens. En cela, il nous intéresse puisque l'un des objectifs du projet de moteur de recherche présenté en première partie est notamment d'encourager, voire généraliser l'utilisation d'un modèle de connaissance au sein de l'école.

Brand-Gruwel et ses collaborateurs proposent de se focaliser sur les compétences nécessaires à la recherche d'information. Pour ce faire, il en définit d'abord les étapes :

- La définition de la tâche ;
- La définition et le choix de la « meilleure » stratégie de recherche à adopter ;
- La localisation et l'accès à l'information ;
- L'utilisation des informations prélevées ;
- La synthèse des informations trouvées avec les informations déjà possédées en mémoire ;
- L'évaluation du résultat global (synthèse).

Dans ce modèle, la recherche d'information est conçue comme une activité de type « résolution de problème ». L'usager part d'un état initial, généralement mal défini et doit atteindre un état final (le but) en respectant des contraintes matérielles (les outils de recherche à sa disposition) et temporelles, tout en passant par des sous-buts grâce à la mobilisation de compétences ainsi que de processus cognitifs et métacognitifs :

- Gérer la réalisation de la tâche de RI en fonction des conditions matérielles et temporelles ;
- Gérer et ajuster ses comportements de recherche et d'analyse d'information (par exemple en modifiant ses comportements si les résultats ne sont pas satisfaisants) ;
- Évaluer la pertinence des informations trouvées ;
- Évaluer le produit issu de la RI après avoir traité les informations (pour rédiger un plan, une partie de plan, une présentation, etc.).

L'intérêt de ce modèle est d'aborder la recherche d'information sous l'angle des compétences requises pour sa menée à bien. Sa formalisation permet de cibler les compétences que l'on cherche à faire acquérir ou à développer chez les utilisateurs de systèmes de recherche d'information.

Ainsi, dans notre cas, certaines compétences minimales doivent être détenues par les futurs utilisateurs du SRI si l'on veut qu'ils puissent l'exploiter dans le cadre de leur activité. Il faut aussi envisager de quelle façon le SRI peut les aider à disposer de ces compétences lors de leur recherche d'information. La fonctionnalité d'autocomplétion tolérante, par exemple, qui va suggérer des termes de recherche aux utilisateurs, peut être considérée comme une aide à la gestion et à l'ajustement des comportements de recherche.

Enfin pour Brand-Gruwel, lors d'une recherche d'information, trois compétences de base sont déterminantes que nous pouvons rattacher à notre cas pratique :

- Les capacités en lecture que l'on peut rapprocher de la capacité des utilisateurs à comprendre le modèle de connaissance tel qu'il est proposé dans le SRI envisagé. Il faut alors s'assurer que les utilisateurs, soit puissent rapprocher les résultats qu'ils trouvent du modèle princeps (grâce à leurs connaissances ou grâce à des fonctionnalités d'association), soit puissent utiliser le modèle princeps tel qu'il leur est proposé par le SRI pour rechercher des informations (à travers l'utilisation de critères et de sous-critères de recherche).
- Les capacités d'évaluation, par exemple pour juger de la pertinence d'une information. Notre SRI devra alors aider l'utilisateur à comprendre en quoi un document s'intègre à sa recherche dans le cadre de l'utilisation du modèle princeps. Cela peut être le cas lorsque le SRI répond à la demande de transdisciplinarité des utilisateurs en suggérant des liens entre concepts scientifiques qui sont propres au modèle de connaissance de l'organisation.
- Les capacités liées à la maîtrise des ordinateurs et des environnements numériques. Cette capacité n'est pas réellement problématique si l'on considère le public du projet de SRI. Cependant, le moteur de recherche tel qu'il est envisagé dans sa version la plus élaborée comporte de nombreuses fonctionnalités qui viendront considérablement « encombrer » et complexifier l'interface de recherche. Cet encombrement et cette complexité peuvent « gêner » les futurs utilisateurs, leur donner précisément le sentiment d'une « incompétence » et les inciter à n'utiliser que la recherche simple comme mode d'accès à l'information. Il sera alors sans doute pertinent de prévoir de recourir à des ergonomes afin de dessiner au mieux les interfaces utilisateurs.

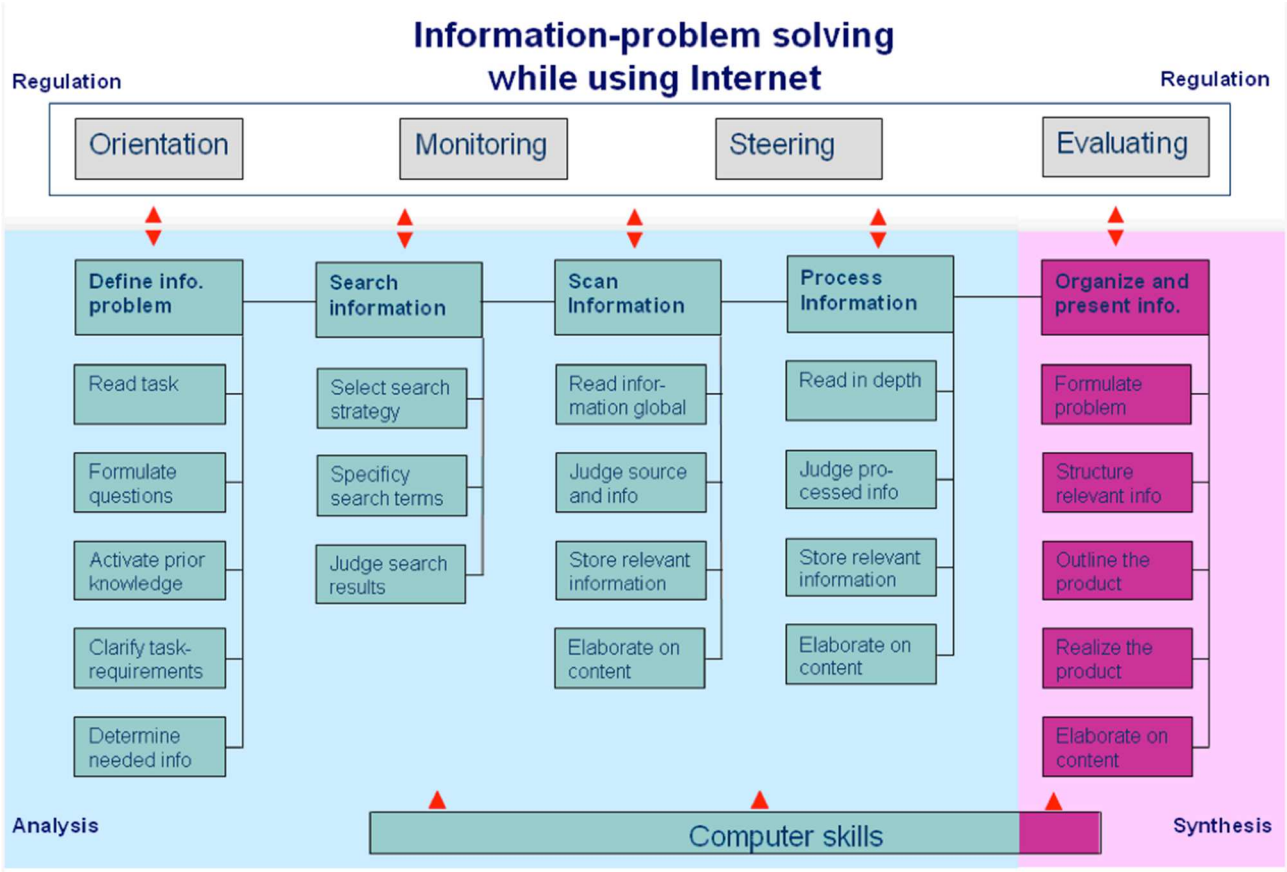


Figure 4 — Le modèle centré sur les compétences selon Brand-Gruwel, Wopereis et Waldaren[7]

2.2.3.4 Le modèle d'Ingwersen (1992)

Ingwersen considère « les comportements humains liés à la recherche d'information comme une série de processus dynamiques et interactifs entre l'espace cognitif de l'individu et l'espace informationnel offert par un système numérique » [3, Dinet]. Cependant, l'originalité de son modèle tient dans le fait qu'il envisage non pas une interaction entre deux entités, soit l'humain et le système technique, mais entre trois entités :

- L'espace cognitif de l'individu, qui comprend la représentation du besoin informationnel, de ses buts et sous buts et de la compréhension des consignes ;
- L'espace technique lié au sri, qui comprend les langages documentaires implémentés, la structure des informations, les règles d'indexation ou encore la logique formelle sous-jacente ;

[7] http://portal.ou.nl/web/topic-informatievaardigheden/home/-/wiki/Main/Cognitieve+processen

- Les informations stockées dans la mémoire de l'individu (ses connaissances) et les informations contenues dans le système technique (c'est-à-dire les documents auxquels il peut avoir accès par l'intermédiaire du sri).

Dans ce modèle, les trois entités s'influencent réciproquement lors de la recherche d'information. Ingwersen envisage aussi l'influence de l'environnement social et organisationnel sur la RI.

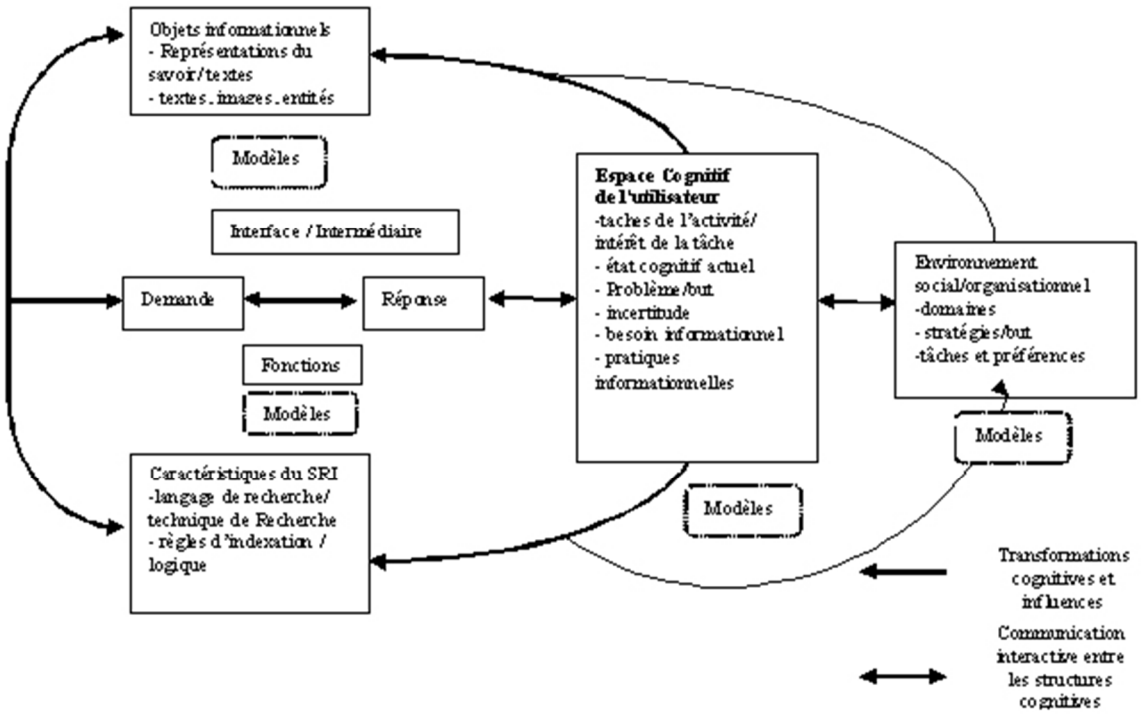


Figure 5. Le modèle des processus interactifs de recherche d'information selon Ingwersen[8]

L'intérêt de cette approche, notamment dans le cas qui a été exposé dans la première section de ce mémoire, nous paraît être la prise en compte de l'interaction forte entre les points de vue utilisateur (les connaissances de l'individu, son espace cognitif ainsi que son environnement social et intellectuel) et les points de vue techniques.

[8] http://campus.hesge.ch/ressi/Numero_10_decembre2009/articles/HTML/RESSI_061_Sicomor.htm

Il nous paraît évident que ces paramètres et leurs interactions ont par ailleurs une grande influence sur les capacités des utilisateurs à rechercher des informations et à évaluer la pertinence des résultats obtenus.

Or la pertinence est une notion centrale dans le domaine des systèmes d'information. On peut même affirmer qu'elle en constitue le principal objectif, à savoir : trouver l'information pertinente relativement au besoin informationnel qui a été identifié. C'est pour cela qu'avant de nous attacher à étudier comment un système d'information peut répondre aux besoins qui ont été formalisés dans la première partie de ce mémoire, il nous paraît important de revenir sur la notion de pertinence qui est attachée aux résultats d'une recherche d'information.

2.2.4 La notion de pertinence

Définir ce qu'est la pertinence d'un résultat de recherche d'information est fondamental dès que l'on se penche sur les systèmes de recherche d'information. C'est en effet en fonction de cette définition que l'on jugera si un système d'information est « performant », c'est-à-dire s'il retourne des résultats qui répondent aux besoins d'information de ses utilisateurs. C'est donc à partir de cette notion que l'on va penser, élaborer, paramétrer et ajuster les fonctionnalités des systèmes de recherche d'information.

Comme dans la section précédente, nous ne pourrons évoquer toutes les notions de pertinence telles qu'elles ont été étudiées dans les sciences de l'information. Nous nous concentrerons sur celles qui nous paraissent utiles à notre réflexion sur la façon dont les SRI peuvent être au service de la connaissance en organisation.

La façon dont est calculée la pertinence dans un SRI permet au système de sélectionner les documents répondant à la requête de l'utilisateur, mais aussi à les classer de façon hiérarchique dans la liste des résultats présentée à l'utilisateur.

Là encore, il est important de préciser que nous nous situons dans le cas d'une recherche d'information avec un moteur de recherche d'entreprise, c'est-à-dire dédié à une organisation, ses ressources informationnelles et ses objectifs.

En effet, sur un moteur de recherche qui indexe et interroge l'Internet, la pertinence d'une page dépend certes de son contenu (et de sa correspondance avec la requête de l'utilisateur), mais aussi de sa fréquentation et de sa popularité (qui se mesurent au nombre et à la popularité des liens

entrants). Parallèlement, le classement des résultats est lié à des principes commerciaux. L'objectif des grands moteurs de l'Internet est de satisfaire les intérêts des annonceurs. Leurs critères de pertinence ne sont pas transposables directement aux moteurs de recherche d'entreprise.

2.2.4.1 La pertinence pour un moteur de recherche d'entreprise

Pour un moteur de recherche d'entreprise, il existe deux types de pertinence : la pertinence « système », et la pertinence « utilisateur ».

La pertinence système est objective, elle est calculée par l'ordinateur. La pertinence utilisateur est subjective et correspond à la façon dont l'utilisateur juge de la pertinence des documents qui lui ont été retournés par le système. L'enjeu pour un moteur de recherche est de réduire la distance qui peut exister entre l'approche système, qui montre vite ses limites en regard des attentes utilisateurs et l'approche utilisateur, souvent imprécise.

2.2.4.1.1 La pertinence système

La pertinence système du moteur de recherche d'entreprise est fondée sur l'analyse statistique du contenu. Elle dépend d'un calcul de similarité entre les termes de la requête et les termes de l'index représentant le document. Pour calculer cette similarité, l'application de recherche va mettre en œuvre un modèle de pertinence qui calcule à chaque requête et pour chaque information le « meilleur » score de pertinence. La similarité est déterminée par quatre critères principaux [2, Vuillequiez] :

- L'occurrence et la cooccurrence des termes de la requête (présence ou non de l'ensemble des termes de la requête dans les documents) ;
- La correspondance d'expression fondée sur la similarité entre la question (la requête) et l'expression dans la page ;
- Le nombre d'occurrences des termes de la requête par rapport à la longueur des documents ;
- La position des termes de la requête dans les documents (titre, corps, liste de mots clefs, résumé) ;
- La proximité des termes de la requête dans les documents.

Ce calcul de pertinence permet aussi d'améliorer les taux de rappel et de précision qui sont des notions importantes en recherche d'information :

- Le rappel est la mesure du nombre de documents pertinents retrouvés en fonction du nombre de documents pertinents présents dans les bases de données interrogées par le moteur de

recherche. Plus le taux de rappel est élevé, plus l'utilisateur a accès à l'ensemble des documents correspondant à sa requête.

- La précision est la mesure du nombre de documents pertinents retrouvés rapportée à un nombre de documents renvoyés par le moteur de recherche pour une requête donnée (par exemple sur la première page de liste résultat, ou les deux premières). Plus le taux de précision est élevé, plus les premiers documents sont pertinents.

Le problème de la pertinence système est qu'elle est définie en termes absolus : soit un document est pertinent, soit il ne l'est pas. Or, la pertinence dépend aussi du contexte dans lequel évolue l'organisation et de la capacité du moteur à correctement interpréter et contextualiser la requête de l'utilisateur [2, Vuillequiez].

À cette pertinence « système » se juxtapose donc une pertinence « utilisateur ».

2.2.4.1.2 La pertinence utilisateur

La pertinence d'un résultat n'est pas seulement une question de correspondance entre une information et une requête calculée statistiquement. Elle est au contraire fortement dépendante du jugement et du contexte de jugement de celui qui a obtenu ce résultat. La pertinence est donc un jugement personnel fondé sur la valeur qu'un utilisateur accorde à un élément d'information dans un contexte donné [1, Hérigault].

Nous avons vu à travers le modèle de Marchionini que le processus type d'une recherche d'informations met en évidence l'importance du rôle de l'utilisateur qui, en procédant par itérations, interprète et évalue en permanence l'adéquation entre les documents renvoyés par le moteur et son besoin d'information initial. La consultation des résultats amène l'utilisateur à leur donner du sens en fonction de ses besoins et de ses attentes. Bref, l'utilisateur procède à des jugements de pertinence qui font pleinement partie du processus itératif de recherche d'information. Et c'est bien parce que « l'individu est au cœur de la notion de pertinence que celle-ci est particulièrement complexe » [1, Hérigault].

En pratique, ce qui est pertinent pour un utilisateur ne le sera pas pour un autre, à fortiori s'ils évoluent dans un contexte métier diffèrent et même si tous deux ont effectué une requête identique. Chaque utilisateur possède sa propre vision des contenus, son propre système de valeurs et interprète une information en fonction de son contexte de recherche d'information.

Les trois principaux critères de la pertinence utilisateur sont [1, Hérigault] :

- La relativité (l'information a-t-elle un rapport avec le besoin exprimé ou en cours d'élaboration ?) ;
- L'utilité (l'information aide-t-elle l'utilisateur à atteindre le but de sa recherche d'information (lui apporte-t-elle un nouvel élément de connaissance pertinent ?) ;
- L'utilisabilité (l'information peut-elle être facilement utilisée par le sujet ?).

2.2.4.2 La pertinence selon Gabrielli et Mizzaro

La définition du degré de pertinence par Gabrielli et Mizzaro se rapproche de ces trois critères et permet de mieux les préciser. Selon les deux auteurs, le degré de pertinence d'une information dépend de l'interaction de trois dimensions :

- La représentation du besoin de l'usager (axe Infeeds) : si l'on considère la recherche d'information comme une activité de résolution de problème, plus une information trouvée est proche du besoin de l'usager et plus sa pertinence est élevée, que ce besoin soit réel ou perçu. Cette dimension se rapproche du critère de relativité.
- La ressource informationnelle : certains documents sont plus pertinents que d'autres pour répondre aux besoins de l'usager (un document, un ensemble de documents, les métadonnées liées aux documents). Cette dimension se rapproche du critère utilisabilité.
- Le contexte de l'activité de recherche d'information (axe infuseco) : cet axe englobe la recherche d'information, la tâche à réaliser et les attributs physiques et psychologiques de l'utilisateur (expertise domaine, préférences, capacités, etc.). Cette dimension se rapproche en partie du critère d'utilité lorsqu'elle considère la tâche à réaliser pour laquelle a été initiée la recherche d'information.

L'autre intérêt de cette définition de la pertinence d'une information est d'insister sur l'interaction de ces trois dimensions. Tâche à réaliser, représentation du besoin et des documents à même d'y répondre sont évidemment interdépendants.

2.2.4.3 Les pertinences selon Saracevic (modèle d'interaction stratifié, 1996)

Saracevic a conçu un modèle de recherche d'information dit « modèle d'interaction stratifié » que nous ne présenterons pas ici, mais qui a pour intérêt de distinguer différents types de pertinence [6, Simmonot] :

- La pertinence système, ou algorithmique, qui est issue de l'évaluation par le système de l'adéquation entre les informations proposées et la requête produite par l'usager ;

- La pertinence sujet, ou thématique, qui concerne la distance sémantique entre le thème produit par la requête de l'usager et le thème des informations proposées ;
- La pertinence cognitive, qui concerne la « distance cognitive » dans l'espace problème donné entre l'état des connaissances stockées en mémoire chez l'usager et son besoin d'information ;
- La pertinence situationnelle, qui correspond à l'utilité des informations proposées par le système pour atteindre le but final de la ri ;
- La pertinence affective, ou « motivationnelle », qui correspond à la capacité des informations à satisfaire ou à plaire à l'utilisateur.

La distinction opérée par Saracevic n'en reste pas moins intéressante en ce qu'elle introduit des notions, certes subtiles et difficiles à opérationnaliser, mais qui permettent de penser des situations et donc éventuellement de répondre à des problématiques. Nous pensons notamment aux pertinences cognitives et affectives.

La notion de pertinence cognitive peut faire réfléchir le concepteur d'un système de recherche d'information à la nécessité de l'adapter au contexte cognitif des futurs utilisateurs. Cette pertinence est inférée à partir de la correspondance avec les connaissances de l'utilisateur (l'information doit être adaptée à son niveau de compréhension) de « l'informativité », de la nouveauté, de la qualité de l'information et des goûts de l'utilisateur [6, Simmonot]. Comment le SRI peut-il aider à combler une potentielle « distance cognitive » ? C'est précisément l'un des objectifs de notre cas pratique que la solution de recherche d'information puisse s'adapter à différents « publics cognitifs » (les chercheurs, les étudiants, les partenaires de recherche).

La notion de pertinence affective évoque aussi une éventuelle distance ; une distance entre l'utilité et la pertinence thématique d'une information, d'une part, et la satisfaction ressentie par l'utilisateur, d'autre part. À priori, on penserait que cette distance n'existe pas, qu'une information thématiquement pertinente et utile doit « plaire » à l'utilisateur. Or la pertinence affective est inférée à partir de son sentiment de satisfaction, de succès et de réussite ainsi que de ses goûts [6, Simmonot]. Autant de notions qui dépendent beaucoup de la subjectivité de l'utilisateur.

On voit bien à partir de ces exemples que la notion de pertinence peut être difficile à définir. Cette difficulté n'est pas sans conséquence sur la conception SRI puisqu'ils visent précisément à répondre à une demande de pertinence. C'est notamment l'enjeu de l'indexation des documents dont l'objet est de définir les critères d'appariement des documents relativement aux requêtes utilisateur.

Troisième partie :
L’indexation automatique des documents

3 Les processus d'indexation automatique des documents

Si nous nous situons dans la perspective d'une indexation automatique des documents, c'est essentiellement pour deux raisons : d'une part, les résultats de nombreuses expériences ont prouvé que l'indexation automatique était tout aussi performante qu'une indexation manuelle fondée sur l'utilisation d'un vocabulaire contrôlé, notamment lorsqu'il s'agit de rechercher des documents « pertinents » [12, Bruandet] ; d'autre part, il s'agit pour nous dans ce mémoire de voir comment un système peut fournir des fonctionnalités d'assistance intelligente à la recherche d'information. C'est donc l'indexation des documents par le système qui nous intéresse.

La norme Afnor NF Z 47-102 décrit l'indexation comme « une représentation condensée d'un document en s'appuyant sur les mots ou les concepts qu'il contient », c'est-à-dire une transcription en langage documentaire des informations après les avoir extraites du document par une analyse.

L'ADBS propose une définition plus complète de l'indexation vue comme un « processus destiné à représenter, au moyen des termes ou indices d'un langage documentaire ou au moyen des éléments d'un langage libre, les notions caractéristiques du contenu d'un document (ressource, collection) ou d'une question, en vue d'en faciliter la recherche, après les avoir identifiées par l'analyse. Les combinaisons possibles des notions identifiées sont représentées explicitement (indexation précoordonnée) ou non (indexation post-coordonnée) en fonction des possibilités du langage documentaire utilisé[9]. »

Ces deux définitions nous permettent d'en déduire l'objectif principal de l'indexation, à savoir : référencer, donc décrire, des documents et leurs contenus dans des bases de données afin de permettre à l'utilisateur du système de recherche d'information de rechercher et trouver ces documents et les informations qu'ils contiennent. Bref, l'indexation a pour objectif fondamental le signalement optimal du contenu des documents.

[9] http://www.adbs.fr/indexation-1--17361.htm?RH=ACCUEIL

Parce qu'une indexation est principalement évaluée par rapport à sa performance pour la recherche rétrospective des documents [21, Chartron], elle doit répondre à un certain nombre de critères permettant d'atteindre les objectifs ci-dessus mentionnés.

3.1 Les critères d'une « bonne » indexation

Pour atteindre les objectifs de recherche de document et d'information qui lui sont assignés, l'indexation doit [21, Chartron] :

- Offrir une description pertinente. On mesure la pertinence de la description d'une indexation par :
    - L'exhaustivité de l'indexation, c'est-à-dire sa capacité à signaler toutes les notions importantes d'un document. Cette exhaustivité peut être propre à un domaine considéré et/ou relative à des notions connexes évoquées dans les documents, mais non spécifiques du domaine considéré. On parle d'exhaustivité interne et externe.
    - La spécificité de l'indexation, qui est caractérisée par la capacité du système à rendre compte de l'information telle qu'elle est citée dans les textes, sans effet de généralisation. La spécificité conserve le contexte de l'information considérée.
    - Le degré d'ambiguïté de l'indexation qui représente la polysémie possible des termes d'indexation. Le système doit être capable de restituer la potentielle pluralité de sens d'un terme, notamment en fonction de son contexte d'énonciation.
- Rendre accessibles les documents et l'information qu'ils contiennent. Un système d'indexation doit rendre l'information la plus accessible possible. Pour cela, l'utilisateur peut bénéficier de catégories intermédiaires de mots lui permettant de disposer de différents niveaux de description des informations et donc d'affiner progressivement sa requête.
- Être cohérente ; la cohérence de l'indexation repose essentiellement sur la régularité de la pratique. Les documents étudiés peuvent être comparables uniquement si la description utilisée pour représenter leur contenu présente une certaine invariabilité. Cette constance est une prévention contre les silences[10] éventuels qu'engendrerait au contraire une variabilité de description des documents.
- Être évolutive. Pour l'interrogation des bases, l'évolutivité de l'indexation permet d'actualiser la description des documents en fonction des nouveaux problèmes qui sont posés, et ainsi de prendre en compte rapidement et rétrospectivement des évolutions des contenus scientifiques et techniques. Ce concept recouvre à notre avis deux notions :

[10] Silence : un système de recherche d'information produit du silence lorsqu'il ne retourne pas tous les documents pertinents correspondant à la requête de l'utilisateur.

- L'évolution des vocabulaires d'indexation ;
- La réindexation de chaque document (introduction, suppression, remplacement de termes).

Pour répondre à ces attentes relatives à l'indexation automatique, il existe deux techniques d'indexation automatique : l'indexation contrôlée et l'indexation libre.

Pour des raisons de clarté, nous distinguons ici volontairement ces deux techniques, mais il importe de savoir que ces deux techniques peuvent aujourd'hui facilement être associés dans le cadre d'une solution de recherche d'information.

L'indexation automatique contrôlée est fondée sur l'implémentation dans la solution d'un référentiel terminologique précoordonné (liste préalablement définie de descripteurs) pour normaliser la représentation des sujets.

L'indexation automatique libre repose sur la description linguistique du contenu des documents à l'aide des termes qu'ils contiennent (liste post-coordonnée de candidats descripteurs). Elle permet à l'utilisateur d'effectuer une recherche plein texte des informations contenues dans les documents.

Dans la section suivante, nous nous intéresserons à l'indexation automatique libre (l'indexation automatique contrôlée sera brièvement abordée dans la section consacrée aux référentiels terminologiques).

3.2 L'indexation automatique libre

L'indexation automatique libre, ou dite « plein texte », consiste à ce que le système extraie automatiquement des documents les termes les plus pertinents pour en représenter le contenu informationnel. Ces termes sont ensuite enregistrés dans un index inversé qui, pour chaque entrée, inclut le mot, sa forme de base, son lemme, sa catégorie grammaticale, ses informations morphologiques, sa position et son poids dans le document ou dans la page, et son document d'appartenance [17, Chaudiron].

Cet index inversé va associer les termes d'accès susceptibles d'être sollicités pendant une recherche à tous les documents présents dans l'index inversé [1, Vuillequiez].

Pour créer automatiquement cet index inversé, il existe deux approches très différentes dans leur mise en œuvre, mais complémentaires :

- La première approche repose sur des technologies linguistiques proposant « des solutions qui s'attachent à traiter le contenu informationnel des documents en prenant en compte les niveaux morphologiques, syntaxiques et sémantiques de la langue » [17, Chaudiron]. Le défi pour ces techniques est de devoir composer avec les ambiguïtés du langage naturel.

- La seconde approche repose sur l'extraction automatique des unités linguistiques ou mots du document. Par un calcul de fréquence, le système d'indexation choisit ensuite les plus significatifs et les plus représentatifs [1, Hérigault]. L'inconvénient de cette approche, fondée essentiellement sur des techniques statistiques, est qu'elle n'est capable que d'identifier des chaînes de caractères et non des concepts qui font sens.

Le problème de ces opérations de traitement du contenu des documents est donc qu'elles se heurtent aux difficultés que présente le traitement automatique du langage naturel.

3.2.1 Les obstacles du langage naturel

Les moteurs de recherche qui disposent de fonctionnalités d'interrogation des documents en texte intégral analysent des textes qui utilisent un langage naturel. Or ce langage, que nous utilisons au quotidien, est ambigu, redondant et implicite.

Le tableau ci-dessous présente les caractéristiques problématiques du langage naturel relativement à son exploitation par un moteur de recherche.

Caractéristiques du langage naturel	**Les difficultés pour l'indexation (et la R.I.)**	**Définitions**	**Exemples**
1/L'implicite	**La pragmatique** Impossible à prendre en compte par des logiciels ou des langages	Liée au contexte du message, aux connaissances sur le monde, à l'usage... domaine de la pragmatique : étude	« Paul donna le billet à la jeune femme » : transaction commerciale : billet de banque ? spectacle : billet d'entrée ?

	documentaires	du « langage en action »	Relation amoureuse : billet doux ? Espionnage : message chiffré ?
2/La redondance	**La synonymie :**	Mots ou expressions différents ayant le même sens, ou des sens voisins.	Voiture et automobile ; tremblement de terre et séisme ; train et chemin de fer…
	La paraphrase :	Expressions équivalentes, mais de structure ou de termes différents	Mon fils a cessé de fumer Jean a renoncé au tabac
	Le glissement de sens :	La **dénotation** : sens propre d'un mot La **connotation** : sens d'un mot dans un contexte particulier	Il prend un bain Il est dans le bain
3/L'ambiguïté	**L'homonymie :** **ambiguïté lexicale**	Mots ayant la même forme, la même graphie, mais des sens différents.	Je porte la porte Les poules du couvent couvent

	La polysémie : **ambiguïté lexicale et sémantique**	Mots ou expressions ayant plusieurs sens ; phénomènes de dérivation, par métonymie et métaphore	Mémoire humaine, mémoire d'ordinateur, le mémoire de maîtrise… Métonymie : Policier (personne et roman) Métaphore : La racine de tous les maux
	L'homotaxie : **Ambiguïté syntaxique** > problèmes pour les logiciels de TALN	Une même syntaxe recouvrant des réalités différentes	Jean est facile à convaincre Jean est habile à convaincre

Les Principaux défis de l'indexation du langage naturel selon Alexandre Serres[11]

Si les contenus des documents interrogés par un moteur de recherche spécialisé, tel que nous l'envisageons par exemple dans la première partie de ce mémoire, sont certainement, du fait d'être issus d'un domaine spécialisé, moins ambigu, redondant et implicite que ceux interrogés par un moteur de recherche généraliste, ils n'en sont pas pour autant totalement exempts.

La question est donc de savoir comment les moteurs de recherche peuvent procéder à une indexation « intelligente » des documents, c'est-à-dire qui représente le sens du texte pour le domaine considéré en levant les ambiguïtés du langage naturel qui y sont contenues.

[11] http://www.sites.univ-rennes2.fr/urfist/Supports/Indexation/Indexation2Defis.html

C'est l'objet des technologies de traitement automatique de la langue (TAL) qui sont intégrées aux moteurs de recherche ayant cette ambition.

3.2.2 Les principes du TAL

Les systèmes de traitement automatiques du langage prennent en entrée un texte ou un ensemble de textes qu'ils transforment pour obtenir en sortie une ou plusieurs représentations du sens de ce(s) texte(s). Comme nous venons de le voir, la tâche essentielle de l'opération de transformation consiste à traduire des documents potentiellement ambigus en représentations non ambiguës.

C'est donc la question de la « compréhension » d'un document texte par le système qui est au cœur des tâches du traitement automatique de la langue. Cette question renvoie à deux problématiques majeures :

- La représentation du sens du texte ;
- La prise en compte du monde de connaissance de référence.

Les technologies de TAL « visent à faire manipuler, interpréter ou générer par les machines le langage naturel écrit ou parlé par les humains » [1, Hérigault]. Il s'agit notamment d'analyser les contenus afin d'en faire émerger, par l'analyse de la langue, les concepts pertinents porteurs de sens.

Pour effectuer une tâche de TAL, on distingue classiquement six niveaux de traitement [17, Chaudiron] :

- Le niveau de la segmentation, qui consiste à découper le texte en « unités élémentaires » (mots et phrases) ;
- Le niveau morphologique, qui traite des variations de formes entre mots d'une même famille, de la manière dont sont constituées les unités lexicales (flexion, dérivation, composition, etc.) Et vise à déterminer la catégorie de discours de l'unité considérée ;
- Le niveau syntaxique, qui détermine la structure des phrases (relations entre les mots et relations entre les groupes de mots) en fonction de la grammaire de référence ;
- Le niveau sémantique, qui traite du sens des mots et des phrases en fonction du contexte ou du domaine de référence ;
- Le niveau du discours qui vise à identifier la structure discursive et argumentative du document ;
- Le niveau pragmatique qui traite du monde de connaissance de référence, c'est-à-dire qui prend en compte les informations extralinguistiques qui peuvent contribuer à la compréhension du texte.

Cette décomposition en six niveaux est bien sûr théorique. Elle ne correspond pas nécessairement au mode de fonctionnement réel de tous les logiciels de TAL. Certains groupent les niveaux 2, 3 et 4 en une seule étape du traitement, alors que d'autres ne prennent pas en compte certaines des étapes mentionnées (par exemple, le niveau pragmatique est rarement pris en compte en tant que tel, mais des connaissances de nature pragmatique peuvent être intégrées dans les dictionnaires de référence, en particulier les connaissances métiers) [17, Chaudiron].

Dans la section qui suit, nous présentons le fonctionnement des quatre premiers niveaux qui correspondent actuellement à l'état de l'art des systèmes commerciaux les plus avancés fondés sur les technologies linguistiques [17, Chaudiron].

3.2.3 Application du TAL à l'indexation automatique

Le processus de l'indexation en texte intégral des documents comprend trois étapes :

- L'analyse des termes du document ;
- La sélection des termes représentatifs du contenu informationnel du document ;
- L'enregistrement de ces termes dans l'index après les avoirs transformés et enrichis par des traitements linguistiques.

Les deux premières étapes relèvent des technologies de traitement automatique du langage. Elles peuvent se décomposer en quatre types d'opérations [17, Chaudiron] :

- L'élimination des mots vides[12] ;
- La lemmatisation des formes fléchies ;
- L'identification des syntagmes[13] comme candidats descripteurs ;
- La pondération des mots, syntagmes ou descripteurs retenus.

[12] L'expression « mot vide » ne signifie pas que le mot n'a pas de sens (il en a toujours, particulièrement pour l'utilisateur) mais qu'il n'a pas de valeur pour le système au moment de l'indexation.

[13] Syntagme : groupe de mots qui se suivent avec un sens.

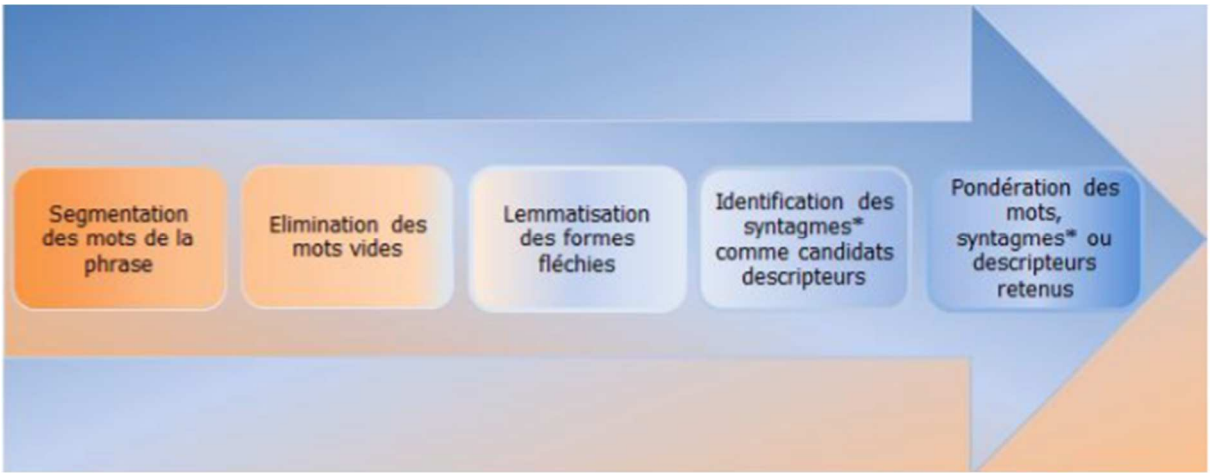


Figure 6 — Les 5 étapes de l'indexation automatique selon Stéphane Chaudiron[14]

Il est possible de rajouter une sixième étape qui relève du traitement sémantique des syntagmes retenus comme candidats descripteurs. Cette sixième étape se situe à l'interface de l'indexation automatique et de l'élaboration de référentiels terminologiques, élaboration qui utilise pour une bonne part les technologies du TAL.

3.2.3.1 La segmentation

Fondée sur l'analyse de chaînes de caractères, la première étape n'est pas à proprement parler linguistique. Elle vise à identifier les mots puis les phrases constitutifs du texte et à former des unités lexicales élémentaires. Elle consiste en la reconnaissance et le regroupement des chaînes de caractères alphabétiques, numériques et typographiques en identifiant préalablement les signes séparateurs entre ces unités lexicales [17, Chaudiron].

Le découpage du texte en mots (tokens) et en phrases est donc piloté par « des signes qui jouent le rôle de séparateurs entre les unités lexicales » [17, Chaudiron]. Cette première tâche effectuée par le système n'est pas sans poser de problèmes dans la mesure où :

- Les signes séparateurs n'ont pas forcément le même rôle dans une phrase selon leur utilisation : par exemple, le point et le tiret, utilisés dans un sigle ou dans un mot composé, ne sont pas des marqueurs de fin de phrase. Pour éviter ce genre de problème, le système distingue les contextes

[14] HERIGAULT Myriam. Moteur de recherche d'entreprise : déploiement du moteur sémantique Exalead à la R&D de Diagnostica Stago. 2012, 90 p. Mémoire, Sciences de l'information, INTD - CNAM, 2012.

dans lesquels un caractère joue le rôle de séparateur. Ainsi, il obtient une liste de séparateurs sans condition (virgule, point-virgule, point d'exclamation et d'interrogation, etc.) Et une liste de caractères dont le rôle varie en fonction du contexte (apostrophe, point, tiret, etc.). Il est aussi possible de fournir au système la liste des formes pour lesquelles le caractère ne joue pas le rôle de séparateur (comme dans aujourd'hui).

- Un texte n'est pas une suite d'énoncés isolés les uns des autres, mais une suite d'énoncés coréférencés, c'est-à-dire qui s'articulent et « font sens » les uns par rapport aux autres : par exemple, « Paul n'a pas apprécié ce roman. Il goûte peu la littérature irlandaise. » Résoudre ces problèmes posés par les relations anaphoriques entre les phrases n'est pas aisé.

La sortie du traitement par segmentation est un texte segmenté en phrases, elles-mêmes segmentées en unités qui sont appelées formes de surface (tokens en anglais).

3.2.3.2 Élimination des mots vides

L'élimination des mots vides, ou « Stop Words » en anglais, vise à écarter de l'indexation certains mots ou termes considérés comme trop génériques, qu'ils soient d'un usage trop courant (les auxiliaires être ou avoir, par exemple) ou à faible valeur sémantique et donc informative (le, de, pour, par exemple). L'objectif de cette étape est de constituer une liste générique dépendante de la langue d'origine. À cette liste générique fournie par le moteur peut être ajoutée une liste spécifique incluant des termes de domaine peu discriminants (en ostéopathie, le terme de « corps » ne sera peut-être pas assez discriminant alors qu'il peut l'être en criminologie).

La liste des « mots vides » est donc constituée d'une liste générique dépendante de la langue et d'une liste spécifique au domaine de spécialité concerné.

Ces deux premières étapes de traitement des documents ne relèvent pas spécifiquement des techniques de traitement du langage naturel. Les techniques du TAL sont utilisées dans les trois prochaines étapes.

3.2.3.3 L'analyse morphologique

Après avoir identifié les termes pertinents pour représenter le contenu d'un document, le moteur fait appel à un analyseur morphosyntaxique.

L'analyse morphologique consiste à reconnaître la structure des formes de surface telles qu'elles ont été segmentées précédemment, soit la lemmatisation, puis à affecter à ces lemmes une catégorie grammaticale.

3.2.3.3.1 La lemmatisation

Ce traitement consiste à mettre sous leur forme canonique les unités lexicales élémentaires issues de la segmentation et à reconnaître les mots simples ou composés. La forme canonique d'un mot, le lemme, est l'infinitif pour un verbe et le masculin-singulier pour les autres formes [36, Balmisse]. En sortie de module, le texte apparaît dans l'index sous la forme d'une liste de lemmes avec leur catégorie syntaxique (ou grammaticale) et les informations morphologiques nécessaires.

La lemmatisation correspond à une analyse lexicale des termes qui permet de regrouper les mots d'une même famille. C'est un regroupement par lemme.

Chaque mot à une forme canonique (forme racine) et des formes fléchies (ou formes de surface) qui correspondent aux différentes occurrences possibles de ce terme. Ces dernières sont toutes les déclinaisons que peut prendre ce terme : verbes à l'infinitif/conjugué, mots au singulier/pluriel, déclinaisons masculin/féminin, etc.). Le travail de l'analyseur morphologique est donc de procéder à la lemmatisation des formes fléchies. [17, Chaudiron].

Pour cela, à partir des formes fléchies du texte, le lemmatiseur va identifier la forme de base et le lemme de référence et la flexion qui lui est associée.

Ainsi, par exemple, pour chaque mot (ou forme fléchie) tel que « saluait », il détermine :

- Sa forme de base : salut ;
- Sa flexion : ait ;
- Son lemme (ou forme canonique) : saluer.

À l'issue de cette analyse, l'index du moteur de recherche aura répertorié pour chaque candidat descripteur :

- Son lemme (nom au singulier, adjectif au masculin singulier, verbe à l'infinitif, adverbe) ;
- Ses dérivations morphologiques : flexions (nombre, genre, temps et mode) et dérivation (suffixe, préfixe, composition).

En recherche d'information, éliminer les variations morphologiques flexionnelles permet de réduire le silence. Ainsi, si un document possède le mot « torsions », il est raisonnable de penser que la requête « torsion » permette d'y accéder.

3.2.3.3.2 L'attribution de catégories grammaticales

La seconde tâche de l'analyseur morphologique consiste à attribuer une catégorie grammaticale ou étiquette syntaxique (nom, verbe, adjectif, etc.) à chacune des formes fléchies identifiées. Cette opération soulève plusieurs difficultés, car le choix des catégories syntaxiques (on parle également de partie du discours, ou de « part of speech » en anglais) est une question délicate (Chaudiron) :

- S'il existe un accord de fait concernant l'emploi des catégories principales (comme nom, verbe, adjectif, etc.), il n'existe néanmoins pas de norme ni de standard concernant le nombre, la nature ou l'intitulé de ces catégories :
    - La finesse des catégories dépend des objectifs poursuivis. Ainsi, dans certains cas, il sera nécessaire de différencier les types de pronoms au sein de la catégorie générale des pronoms personnels alors que, dans d'autres cas, ce ne sera pas utile.
    - La nécessité ou non de segmenter en composants élémentaires certaines expressions (par exemple, « faire fausse route » ou « faire route commune ») est une autre question. Dans certaines situations (l'indexation d'un texte, par exemple), il peut être utile de considérer comme expression figée ou semi-figée un multi-terme (par exemple, « ligaments croisés » ou « membre inférieur »).

Par ailleurs, le français, comme d'autres langues, possède également une morphologie dérivationnelle. Celle-ci définit les règles permettant d'associer un affixe (suffixe ou préfixe) à une forme de base. Par exemple, le préfixe « re » peut être utilisé avec de nombreux verbes comme refaire ou rejouer ; le préfixe « in » est, quant à lui, utilisé pour les adjectifs, comme dans « injuste » ou « insatisfait ». De même, un grand nombre de suffixes existent en français, comme « isme », « ité » ou « iste ». Les règles de morphologie dérivationnelle sont alors utilisées pour retrouver une forme de base et son lemme à partir d'une forme de surface, et pour aider ainsi à son analyse (attribution d'une catégorie syntaxique par exemple).

À la sortie du module d'analyse morphologique, le texte apparaît sous la forme d'une liste de lemmes avec leur catégorie syntaxique et les informations morphologiques nécessaires.

3.2.3.4 L'analyse syntaxique

L'analyse syntaxique a d'abord pour objet d'identifier les différents éléments constitutifs de la phrase que l'on appelle syntagmes. Les syntagmes sont des groupes de mots qui se suivent avec un sens. Par exemple, « courrier électronique » est un syntagme.

Les étapes d'identification des syntagmes sont réalisées par la mise en œuvre d'un analyseur syntaxique dont le but est d'identifier les différents éléments constitutifs de la phrase, et particulièrement les syntagmes nominaux et les expressions idiomatiques [2, Vuillequiez]

L'enjeu de cette analyse est important, car ces identifiants, particulièrement les syntagmes nominaux, vont être utilisés comme candidats descripteurs pour représenter le contenu informationnel d'un texte. Pour déterminer ceux qui possèdent les propriétés pour devenir les descripteurs, le système peut recourir à un calcul de fréquence ou à une comparaison avec un vocabulaire contrôlé (une liste d'autorité ou un thésaurus).

Dans un deuxième temps, l'analyse va construire la structure globale de l'énoncé. Pour ce faire, l'analyse est régie par une grammaire de la langue qui est utilisée au niveau local pour la construction des syntagmes et au niveau global pour l'attribution des rôles syntaxiques à chacun des syntagmes (groupe sujet, groupe verbal, groupe complément, etc.).

Pour constituer ces syntagmes, deux grandes familles de méthodes peuvent être utilisées [17, Chaudiron] :

- Les méthodes fondées sur l'utilisation de « patrons » (« patterns » en anglais), où la structure syntaxique est définie à l'avance (par exemple, les groupes nominaux constitués de la suite < Nom Adjectif Adjectif >, comme encéphalopathie spongiforme bovine). Cette méthode est efficace, car le traitement effectué prend en compte le contexte immédiat. Inversement, elle risque d'exclure des informations qui pourraient être importantes et qui se trouvent, par exemple, dans un constituant non identifié par le patron, ou dans le verbe de la phrase s'il s'agit d'un patron uniquement destiné à extraire les syntagmes nominaux.
- Les méthodes qui reposent sur des grammaires à base de règles de réécriture. Elles permettent à la fois de rendre compte de manière souple des différentes façons de composer un même syntagme et d'exprimer les diverses structures de constituants qui sont acceptables pour une phrase. Le pouvoir d'expression de ces grammaires est beaucoup plus important que la méthode des patrons. En effet, ces grammaires de constituants (dont il existe de très nombreuses versions) permettent de dériver plusieurs constituants à partir d'une seule règle. Ces règles de réécriture

sont constituées de deux parties : une partie gauche, qui correspond à l'un des symboles utilisés pour designer les constituants, et une partie droite, qui indique la suite de constituants ou de catégories syntaxiques attendus. Par exemple, GN (qui signifie groupe nominal) pourra se réécrire par la suite < Déterminant Nom Adjectif > ou < Déterminant Adjectif Nom > ou < Nom propre > ; GV (groupe verbal) se réécrit < Verbe suivi de GN >.

Dernière étape, une fois les syntagmes identifiés et extraits des documents, il est nécessaire de les normaliser. Cette normalisation permet de s'assurer que les syntagmes qui indexent le document et ceux qui seront identifiés lors de l'analyse de la requête sont homogènes. En effet, un même syntagme peut connaître des variations lexicales, morphologiques et/ou syntaxiques.

3.2.3.5 La pondération des syntagmes retenus, l'apport des traitements statistiques

La dernière étape du processus d'indexation consiste à affecter un indice d'importance aux termes et syntagmes pressentis pour indexer le document. Le poids affecté à chaque entrée de l'index dépend de leur importance relative pour décrire le document.
La pondération des syntagmes retenus par l'indexation automatique s'effectue grâce à des analyses de type statistique effectuées au moyen d'algorithmes. Elle permet au système d'effectuer un classement par pertinence des documents retournés à l'utilisateur suite à sa requête par comptabilisation et localisation des termes [12, Bruandet]. Elle tient principalement compte de deux critères :

Tout d'abord, **le nombre d'occurrences** du terme dans le document. L'hypothèse est que l'importance d'un sujet traité dans un texte est reflétée par la fréquence des termes ou syntagmes exprimant le sujet en question [17, Chaudiron].

Mais ce critère de pondération ne suffit pas. Au-delà d'un certain seuil, plus un terme est fréquent, moins il est pertinent pour décrire le document dans lequel il figure. Dans certains documents, la fréquence d'un terme peut être si élevée que celui-ci n'est plus discriminant pour représenter le document. Un deuxième critère est alors celui de **la fréquence documentaire**, soit le nombre d'occurrences du terme dans l'ensemble du fonds documentaire auquel appartient le document concerné. Cet indice est fourni par la mesure du rapport entre la fréquence du terme dans le

document et sa fréquence dans l'ensemble de la collection (ce critère est plus facile à mettre en place dans le document en cours d'indexation, mais plus difficile dans les documents de la collection).

À partir de cette formule de base, d'autres critères de pondération ont ensuite été proposés :

- La longueur des documents dans lesquels le terme apparaît ;
- La partie du document dans laquelle le terme apparaît (titre, résumé, introduction, etc.) ;
- Le pouvoir discriminant des termes (plus un terme est rare, plus il est discriminant). C'est le groupe intermédiaire entre les termes qui apparaissent fréquemment et ceux qui apparaissent rarement qui contient les termes caractéristiques du thème du document ;
- La probabilité d'apparition des termes dans les documents pertinents et non pertinents.

Certains moteurs de recherche peuvent appliquer des algorithmes de pondération complémentaires pour prendre en compte des termes peu cités et à haute valeur significative.

Une autre solution est de remplacer les termes et les syntagmes extraits des documents par leurs équivalents choisis dans un thésaurus de référence. Cette fonction permet donc de rapporter la question de l'indexation libre à celle de l'indexation contrôlée que nous verrons ultérieurement.

3.2.4 Ressources linguistiques pour l'indexation automatique

Les traitements et l'enrichissement des entrées indexées s'appuient sur des dictionnaires lexicographiques et des grammaires. Ces dictionnaires recensent les termes et les expressions (idiotismes) en indiquant les formes canoniques, grammaticales, morphologiques ainsi que les synonymes. Les plus élaborés fournissent des informations conceptuelles telles que les termes spécifiques et les termes associés, constituant la base du réseau sémantique du domaine.

Des règles de découpage du texte et de repérage de structures de phrases types peuvent être déduites de référentiels phraséologiques.

Ces solutions contiennent aussi des « anti-dictionnaires » (liste de mots vides). Il est aussi possible de faire appel à des anti-dictionnaires de domaine afin d'éliminer les mots vides, non pas de la langue, mais du domaine considéré.

L'ensemble de ces documents détermine la performance d'une solution de recherche et explique aussi son coût. Les documents sont convoqués dans l'ordre suivant :

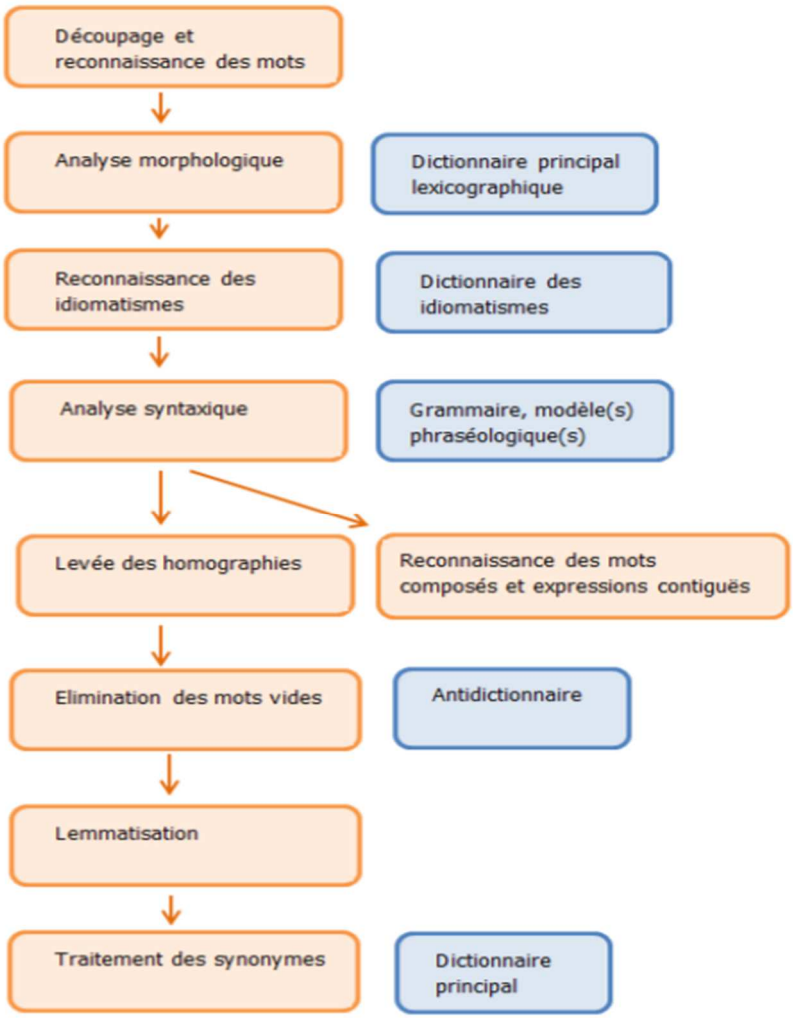


Figure 7 — Description des traitements linguistiques de l'indexation automatique des ressources linguistiques[15]

[15] Source : CHAUMIER Jacques et DEJEAN Martine. Recherche et analyse de l'information textuelle. Tendances des outils linguistiques. Documentaliste-Sciences de l'information, 2003/1, vol. 40, pp. 14-24. ISSN 0012-4508. In Hérigault

3.2.5 Indexation de la question/traitement de la requête

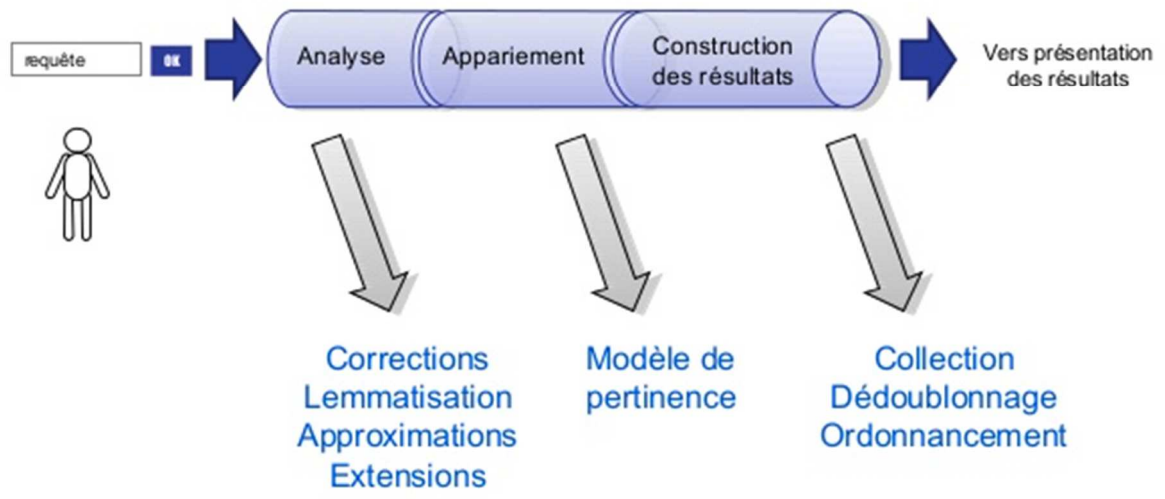


Figure 8 — Pipeline de traitement des requêtes[16]

Comme on le voit dans le schéma ci-dessus, les traitements linguistiques de l'indexation sont également opérés, au moment de la recherche d'information, sur les termes de la requête. Le « pipeline de traitement de la requête » s'effectue en plusieurs étapes :

- Le moteur de génération de requête applique des traitements linguistiques : élimination des mots vides, normalisation de la casse et de l'accentuation, transformation des mots en lemmes.
- Le module analyse la requête, la traduit en équation logique à l'aide d'opérateurs booléens entendus (c'est-à-dire essentiellement une pondération des termes de la requête) permettant ainsi d'obtenir un classement par pertinence. Les opérateurs booléens utilisés sont :
    - L'opérateur « et avec contraintes », qui sert à spécifier une contrainte de proximité pour obtenir les documents contenant tous les mots côte à côte ou dans la même phrase ou dans le même paragraphe (plus les mots sont proches, meilleur sera le classement du document) ;

[16] http://fr.slideshare.net/knowledgeconsult/faciliter-laccs-linformation-qrce-un-moteur-de-recherche. Gilles Balmisse

- L'opérateur « ou cumulatif » qui permet de rechercher les documents contenant n'importe quel sous-ensemble de mots parmi ceux demandés et classe en tête de liste les documents qui en contiennent le plus (c'est le plus largement utilisé) ;

- À cette étape du traitement de la requête, le système peut enrichir automatiquement la requête par des mots proches orthographiquement et phonétiquement, ou des synonymes identifiés dans l'index principal.
- Une opération de fusion des résultats est effectuée par « collection, dédoublonnage et ordonnancement » [36, Balmisse]. Enfin, le composant en charge de la présentation des résultats procède à une mise en forme avant affichage.

Nous pouvons voir ci-dessous les opérations effectuées par un moteur de recherche en analyse et en interrogation plein texte des documents.

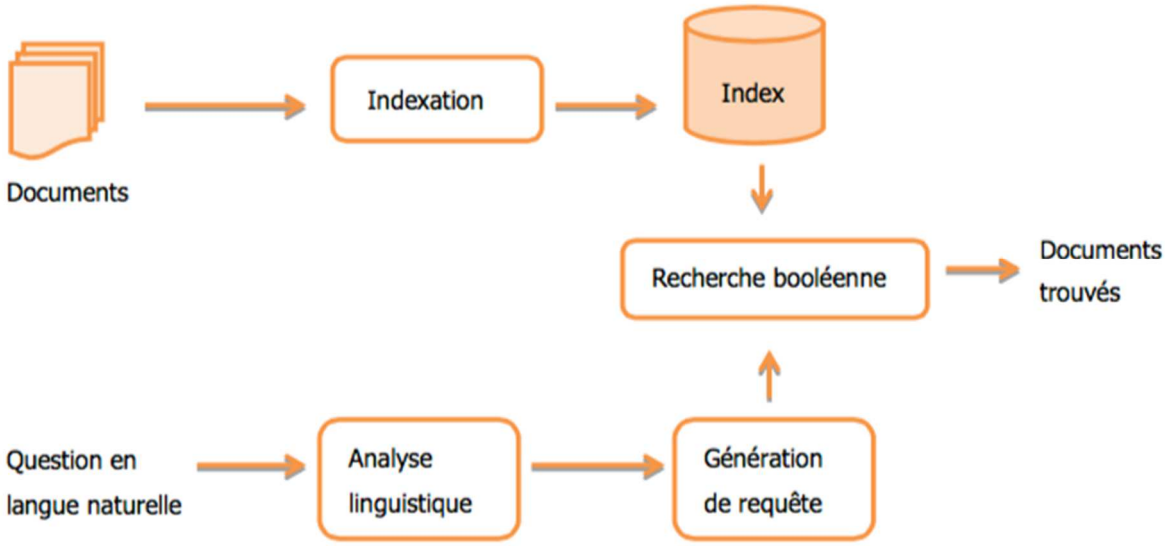


Figure 9 — Architecture générale d'un système de recherche en langue naturelle[17]

[17] LALLICH-BOIDIN Geneviève, MARET Dominique et CHAMBAUD Serge. Recherche d'information et traitement de la langue : fondements linguistiques et applications. Villeurbanne, Presses de l'enssib, 2005. Collection Les cahiers de l'enssib. 281 p. ISBN 2- 910227-60-X. **In** HERIGAULT Myriam. Moteur de recherche d'entreprise : déploiement du moteur sémantique Exalead à la R&D de Diagnostica Stago. 2012, 90 p. Mémoire, Sciences de l'information, INTD - CNAM, 2012.

Ce schéma souligne l'importance de l'analyse linguistique pour l'indexation des documents et des de la reformulation des questions des utilisateurs en requêtes dans un langage formel.

Elle passe sous silence l'acquisition des données, l'appariement des index et la présentation des résultats. Nous en déduisons que c'est aux étapes d'indexation en analyse et en recherche que se joue l'intelligence du moteur.

3.2.6 Les apports de l'indexation morphosyntaxique à l'indexation et à la recherche d'information

L'indexation morphosyntaxique permet au moteur de se référer à un dictionnaire et à des grammaires afin de :

- Lever les ambiguïtés d'homographies. Par exemple, l'analyseur établit que le mot « cabot » correspond au couple cabot-grade ou cabot-comédien et parvient à le désambiguïser ;
- Faire l'apprentissage de toutes les flexions et dérivations associées, des mots composés et des synonymes (le mot est indexé avec ces différentes informations).

En recherche, grâce à l'index enrichi ainsi qu'à des calculs (ou instructions) algorithmiques, statistiques et linguistiques, le moteur de recherche va être capable d'analyser et reformuler les requêtes utilisateurs et ainsi de :

- Gérer les extensions de requêtes de manière transparente pour l'utilisateur : reformulation de la requête avec les lemmes, extension de la requête aux singuliers et pluriels, masculins et féminins, aux formes fléchies et dérivées, à des mots proches orthographiquement et phonétiquement. Cela contribue à réduire le silence
- Améliorer le taux de précision dans la recherche d'information grâce à l'utilisation des syntagmes, moins ambigus que les termes simples, comme entrées d'index (l'expression pierre à fusil est en effet plus précise que les deux mots pris isolément).
- Restituer des résultats de manière automatique ou avec sollicitation de l'utilisateur (fonction « vouliez-vous dire ? ») malgré des fautes d'accord, d'orthographe. Recherche floue.
- Produire un résumé du document sur la base des phrases « significatives » qui contiennent les termes jugés représentatifs du contenu du document considéré (nous verrons cela plus en détail dans la partie consacrée aux fonctionnalités d'assistance à la recherche d'information).

Au-delà de ces techniques de traitement automatique du langage, le moteur devient sémantique :

- En indexation lorsqu'il permet la désambiguïsation de certains termes ;
- En recherche avec l'interprétation des termes de la question et leur extension à des termes proches ou liés par une relation sémantique.

Pour cela, il faut que les outils d'indexation automatique procèdent à une analyse sémantique des documents textes.

Le moteur de recherche est capable de trouver des documents sur le sujet sans que ceux-ci contiennent les termes demandés (lorsque le moteur étend la requête en ajoutant des synonymes, on parle de recherche sémantique : la recherche ne porte plus sur les mots de la question, mais sur des mots dont la signification est équivalente ou proche).

3.2.7 L'analyse sémantique

Le dernier niveau de l'analyse linguistique concerne le traitement sémantique du document et vise à en identifier le sens intrinsèque. Elle a pour objectif d'en « calculer » leur sens en utilisant :

- Soit un système de relations (graphe conceptuel, réseau sémantique) ;
- Soit un système de traits sémantiques ;
- Soit une représentation conceptuelle pivot.

Nous n'aborderons pas ci-après la représentation conceptuelle pivot qui est souvent utilisée dans des applications multilingues

3.2.7.1 Analyse sémantique par graphe conceptuel ou réseau sémantique

L'analyse sémantique par graphe conceptuel vise à établir des relations de significations entre les lemmes. C'est donc la place du lemme dans le réseau qui détermine son sens et non la description sémantique qui en est faite. Le thésaurus est un exemple connu qui décrit les relations existant entre les termes par leurs positions (relations de synonymie, d'hyponymie, d'hyperonymie, etc.).

Le graphe et le tableau ci-dessous donnent l'exemple des relations sémantiques pour le lemme « car » (voiture en anglais).

Cette approche est très efficace pour décrire des mondes conceptuels fermés (domaines de spécialités), mais sa généralisation à la langue générale pose de nombreux problèmes :

- La polysémie de la plupart des termes, les glissements de sens, les nouvelles acceptions (évolutions) rendent difficilement « maintenable » un réseau de cette importance (sauf à simplifier les relations au risque de perdre la capacité du réseau à représenter les concepts et leurs relations) ;
- Le problème de l'universalité de la représentation du monde qui est sous-jacente à l'idée d'un réseau sémantique dans :
    - Le choix des lemmes signifiants ;
    - La nature des relations entre les lemmes ;
    - La place des lemmes les uns par rapport aux autres dans le réseau.

3.2.7.2 Analyse sémantique par trait sémantique

La deuxième approche consiste à décrire les lemmes au moyen de traits sémantiques (ou sèmes) qui correspondent à des étiquettes. De même que le lemme est décrit, sur le plan syntaxique, par sa catégorie morphologique et le modèle flexionnel qui lui est associé, il est décrit, sur le plan sémantique, par les sèmes qui le caractérisent.

Ainsi, le terme « avocat » sera affecté des traits sémantiques indiquant qu'il peut s'agir d'un fruit ou d'un homme de loi. Si, dans la même phrase, on rencontre le lemme « plaider » affecté des traits sémantiques indiquant qu'il s'agit d'une prise de parole pour défendre un accusé, seule l'acception homme de loi sera retenue.

La compatibilité des traits sémantiques entre les lemmes d'une même phrase est vérifiée dans un processus d'unification. L'unification vérifie qu'il existe un même trait (ou ensemble de traits) commun aux différents lemmes de la phrase pour conclure à la validité de celle-ci. Par exemple, le syntagme « l'avocat marron » est accepté, car marron comporte le sème malhonnête, alors que le syntagme « l'avocat bleu » ne sera pas accepté.

Si le principe de fonctionnement de l'approche par traits sémantiques est simple, sa mise en œuvre s'avère délicate notamment parce que :

- Il est impossible de déterminer à priori tous les sèmes qui seront nécessaires pour les différentes applications. Par exemple, définir le terme « caviar » uniquement avec les sèmes indiquant qu'il s'agit d'œufs d'esturgeon salés est insuffisant, et il conviendrait d'ajouter le sème luxueux (ou passe si l'on est dans le domaine sportif). Mais, avec cet exemple, on voit bien que les sèmes sont dépendants du type de représentation que l'on donne du monde de référence et du contexte d'usage de l'application qui va manipuler ces connaissances. On retrouve donc d'une certaine

manière les objections adressées à l'approche par relations sémantiques pour la question de l'universalité des sèmes.

- Enfin, se pose également la question de l'adaptation du système de traits à des domaines de spécialités nouveaux ainsi que celle de la maintenance du dictionnaire comportant la description sémantique des termes. Les définitions peuvent évoluer ou de nouveaux termes apparaître.

La décision de recourir à ces approches dépend de plusieurs critères :

- La délimitation conceptuelle du domaine : plus le domaine est spécialisé et bien délimité, meilleurs sont les résultats ;
- L'évolutivité du domaine : plus le domaine est stable, moins le système de représentation sémantique devra évoluer, moins la maintenance sera problématique ;
- Le volume des données à traiter : on ne peut guère envisager un traitement sémantique complexe pour l'indexation du web, mais traiter un intranet d'entreprise est tout à fait envisageable.

L'exemple de ces techniques de traitement automatique du langage visant à produire une analyse sémantique des documents d'un système d'information nous montre qu'à partir d'un certain niveau d'exigence de représentation et d'exploitation du contenu des documents, il est intéressant de recourir à des référentiels linguistiques pré-coordonnés pour rechercher et exploiter l'information présente dans un SRI (même si la création de ces référentiels recourt largement aux techniques de TAL telles que nous les avons vues).

Quatrième partie :
Les référentiels terminologiques et l'assistance à la recherche d'information

4 L'apport des référentiels terminologiques à l'assistance à la recherche d'information

L'un des problèmes clefs des systèmes de recherche d'information est la définition d'une fonction de correspondance entre la représentation du contenu sémantique des documents et la requête de l'utilisateur. Cette fonction sert à modéliser la pertinence d'un document pour l'utilisateur [12, Bruandet].

Comme nous l'avons vu précédemment, il existe deux types de pertinence, une pertinence système et une pertinence utilisateur. Améliorer la qualité d'un SRI consiste à réduire la distance entre ces deux pertinences ou, pour le dire plus simplement, à les faire correspondre autant qu'il est possible.

Pour réduire cette distance, une première approche a consisté à étudier la façon dont les utilisateurs composent leurs requêtes, les compétences dont ils bénéficient ou pas pour cela, la manière dont ils définissent leurs objectifs et stratégies de recherche. Nous avons certaines d'entre elles à travers les modélisations de la recherche d'information. Cette approche est essentiellement théorique, mais il ne faut pas négliger les effets de la recherche théorique sur les pratiques et, dans ce cas, les réflexions sur les fonctionnalités d'aide à la recherche qu'ont très certainement engendrés ces travaux de modélisation).

La deuxième approche est liée intrinsèquement à la problématique de la recherche d'information. En effet, outre la question de la capacité de l'utilisateur à définir clairement et précisément son besoin d'information, il existe une autre question fondamentale, celle de l'expression de ce besoin d'information. Les utilisateurs de SRI utilisent une grande variété de termes pour exprimer le même concept. Ce problème est fondamental en recherche d'information, d'autant plus que les utilisateurs expriment leurs besoins par des requêtes brèves et « incomplètes » comme cela est souvent le cas sur l'Internet et dans les modes « recherche simple » des moteurs spécialisés ou d'entreprise [12, Bruandet].

Par exemple, les termes « voiture », « automobile », « auto », voire « bagnole », désignent le même concept de « véhicule automobile à roues propulsé par un moteur et destiné au transport terrestre de personnes, de leurs bagages et de petits objets[18] ».

[18] https://fr.wikipedia.org/wiki/Automobile

Par ailleurs, un même terme peut être utilisé pour exprimer différents concepts. Nous avons vu que « caviar » peut désigner des œufs d'esturgeons ou une passe permettant de marquer facilement des points dans un sport de ballon.

Il a été montré dans la section précédente qu'une façon de résoudre cette difficulté d'expression de requête est de procéder à une indexation des documents en utilisant des technologies de traitement automatique du langage.

Une autre façon de trouver une solution à ce problème consiste à implémenter dans le SRI un référentiel terminologique qui servira, entre autres, à définir les termes d'indexation et donc de recherche des documents.

Tout au long de l'évolution des systèmes d'information, on s'est ingénié à concevoir des outils de description du contenu des documents à des fins de recherche documentaire. Ces outils ont pris de nombreuses formes : taxinomies, classifications, thésaurus [13, Chaumier].

Du référentiel qui sera utilisé pour soutenir les opérations d'indexation dépendent les opérations d'interrogation et d'exploitation des documents que permettra le moteur de recherche. Les plus simples se contentent d'établir une liste de mots clefs et/ou d'identités nommées, sans établir de relation entre les termes dudit référentiel. Ce sont les terminologies ou les listes d'autorité (l'une et l'autre pouvant se confondre). L'inconvénient principal de ces langages documentaires à mots clefs est qu'ils ne précisent pas la sémantique de l'association des mots qu'ils contiennent, et donc de la requête dès qu'il s'agit de dépasser les opérateurs booléens classiques.

D'autres référentiels proposent une mise en œuvre des relations de sens entre des catégories de termes par l'expression des liens sémantiques qui décrivent les composants et l'organisation de la connaissance du domaine qu'ils expriment [13, Chaumier]. Dans cette section, nous nous intéresserons aux référentiels les plus élaborés parmi ceux qui expriment des relations sémantiques et qui sont couramment utilisés dans les systèmes de recherche d'information, à savoir : les thésaurus et les ontologies.

4.1 Les thésaurus

L'origine des thésaurus est ancienne et les documentalistes l'utilisent depuis longtemps pour indexer manuellement les documents ou les livres de façon à les retrouver rapidement selon d'autres critères que les critères bibliographiques (titre, auteur, date, éditeur, etc.).

4.1.1 Définition

Le thésaurus est un outil classique en recherche d'information. Il consiste en la définition de termes du domaine généralement appelés concepts, et la représentation de relations sémantiques entre ces termes. [12, Bruandet]. Un thésaurus a pour objectif de rendre explicite l'organisation conceptuelle d'un domaine. C'est une forme de modélisation des connaissances. On peut donc le définir comme un réseau de concepts, représentés par des termes et articulés selon trois grands types de relations : relations hiérarchiques, d'équivalence (synonymie) ou d'association. Ces relations permettent de définir les concepts et leur(s) position(s) dans le champ sémantique décrit par le thésaurus [16, Keller].

4.1.2 Fonctionnement

Un thésaurus possède trois grands types de relations lui permettant d'organiser les concepts qui le composent :

- Les relations d'équivalences ;
- Les relations hiérarchiques ;
- Les relations associatives.

4.1.2.1 Relations d'équivalence

La relation d'équivalence lie tous les termes d'une même langue qui expriment le même concept. Elle recouvre trois types d'équivalents intralinguistiques (ou considérés comme tels) :

- La synonymie véritable ;
- La quasi-synonymie ;
- L'antonymie.

4.1.2.1.1 Synonymie véritable

Il peut d'agir de :

- La forme complète et son abréviation ou sigle ;
- Les termes d'origines linguistiques ou culturelles différentes mais utilisés couramment pour désigner le même concept ;

- Le nom populaire et le nom scientifique ;
- Les variantes orthographiques ;
- Les appellations ancienne et moderne.

4.1.2.1.2 Quasi-synonymie

La quasi-synonymie concerne principalement deux types de cas :

- Des termes de niveaux hiérarchiques différents nécessaires à l'exhaustivité du thésaurus, mais trop spécifiques pour être utilisés comme descripteurs. Un terme plus général leur sera alors préféré. Les termes trop spécifiques seront alors définis comme « employés pour ». Exemple : « Ours » employé pour « Ours blanc » et « Ours brun »
- Des concepts très proches et difficiles à distinguer pour des non-spécialistes. Ex. : « Entorse » et « foulure ».

4.1.2.1.3 Antonymie

Il s'agit dans ce cas des deux pôles d'un même axe sémantique [16, Keller]. Exemple : « Imposition » et « Non-imposition ».

Ces deux aspects d'un même concept n'ont pas besoin d'être deux descripteurs indépendants, mais ils sont importants à relier dans la mesure où un document portant sur l'un fera fréquemment référence au second aspect du concept [16, Keller].

4.1.2.2 Relations hiérarchiques

On distingue trois types de relations hiérarchiques :

- La relation hiérarchique générique (de division à subdivision de même nature). Il s'agit du type de relation hiérarchique privilégié. Ex. : Humain -> Femme.
- La relation hiérarchique partitive (de tout à partie). Ex. : Système digestif -> Estomac.
- La relation hiérarchique d'instance (de classe générique à spécimen). Ex. : Os-> Humérus.

En outre, un thésaurus peut permettre la polyarchie des concepts, ceux-ci pouvant alors renvoyer à des concepts de différentes catégories. Par exemple :

- Doigt/Auriculaire (relation générique de division à subdivision de même nature) ;
- Main/Doigt (relation partitive, de tout à partie).

4.1.2.3 Relations associatives d'un thésaurus

Les relations associatives d'un thésaurus permettent de lier des concepts généralement associés mentalement par les membres de la communauté dont le thésaurus représente le domaine de connaissance. C'est une relation entre concepts qui ne sont pas reliés hiérarchiquement, mais qui partagent une connexion sémantique. On associe de préférence des termes de niveaux de spécificité comparables.

Selon son niveau de sophistication, un thésaurus va permettre des relations associatives plus ou moins nombreuses et fines.

Les thésaurus les plus simples se contentent de mentionner les « termes associés » par une relation de type « voir aussi ».

Ex. : Manipulation -> Position de travail.

Une association conceptuelle étant par nature très subjective, il est important de toujours s'interroger sur son utilité et sa pertinence. Elle doit aider, selon les cas, à percevoir la signification du descripteur et/ou appréhender son aire sémantique globale.

Les thésaurus les plus élaborés peuvent exprimer un plus grand nombre de relations sémantiques entre les concepts qu'il contient. Le nombre et la nature de ces relations dépendent du domaine que le référentiel doit représenter et des applications envisagées en recherche d'information. Ce type de relation permet que le système oriente l'utilisateur vers des concepts liés à sa requête initiale afin de l'enrichir. Cela augmente ainsi ses chances d'accéder aux documents attendus et pertinents. C'est aussi un « soutien » à sa réflexion scientifique. Dans « *Guide pratique pour l'élaboration d'un thésaurus documentaire* [19] », M. Hudon recense treize principaux types de relations associatives :

- La cause et l'effet : Décalcification et Arthrose. ·
- Un tout et une composante essentielle : Corps et Eau ;
- Une action et son agent : Manipulation et Ostéopathe ;
- Une action et son produit : Jardinier et Jardin ;
- Une action et son objet : Manipulation et Craquement ;
- Une action et le lieu de son déroulement : Soins et Clinique ;
- Une science et son objet : Paléontologie et Fossile ;

[19] In *Guide pratique pour l'élaboration d'un thésaurus documentaire*. Danièle Dégez, Dominique Ménillet, collab. 2e édition [1994]. Montréal, Les Éditions ASTED (diffusion France : ADBS), 2009. 274 p. ISBN 978-2-923563-17-6

- Un objet et sa propriété : Poison et Toxicité ;
- Un objet et son application : Ordinateur et Traitement des données ;
- Un objet et un de ses matériaux constitutifs : Os et Calcium.

En outre un thésaurus permet d'associer :

- Des concepts de sens proche : Bateau et Navire ;
- Des antonymes : Tolérance et Préjugé ;
- Des concepts complémentaires : Enseignement et Apprentissage.

4.1.3 Utilisation d'un thésaurus dans un système de recherche d'information

En recherche d'information, l'implémentation d'un thésaurus permet de nombreuses fonctionnalités d'assistance à l'indexation et de recherche d'information.

4.1.3.1.1 Utilisation d'un thésaurus lors de l'indexation des documents

Lors de l'indexation, le thésaurus joue le rôle de liste d'autorité et autorise une pré-coordination des termes d'indexation. Cette pré-coordination va permettre de :

- S'assurer que chaque concept distinct est exprimé par une forme linguistique unique et qu'une forme linguistique ne revêt pas plusieurs concepts ;
- Indexer des documents dans une base de données en décrivant précisément leurs contenus, grâce à un langage validé et partagé ;
- Lier les termes d'indexation et les documents qui les contiennent selon les relations sémantiques établies dans le thésaurus.

Ces opérations d'indexation vont permettre à l'utilisateur de disposer d'un ensemble de fonctionnalités d'assistance à la recherche d'information.

4.1.3.1.2 Utilisation d'un thésaurus en recherche d'information

L'utilisation d'un thésaurus dans un système de recherche d'information va permettre à l'utilisateur de bénéficier de fonctionnalités d'assistance :

- À la composition de la requête ;
- À l'exploitation des connaissances du domaine.

4.1.3.1.2.1 Fonctionnalités d'assistance à la composition de la requête

4.1.3.1.2.1.1 Expansion de requête de l'utilisateur au moment de l'interrogation

L'expansion de requête de l'utilisateur lors du processus d'interrogation est l'usage le plus fréquent qui est fait d'un thésaurus dans un système de recherche d'information.

L'expansion de requête consiste à ajouter des termes liés aux termes des requêtes (synonymes, quasi-synonymes, éventuellement hyponymes et hyperonymes selon les usages) ou à remplacer ces termes par ceux du thésaurus. C'est une forme de post-coordination des termes d'indexation qui fait correspondre des termes exprimés lors d'une requête à l'indexation des documents qui a été réalisée avec d'autres termes. Cette extension de requête peut être :

- manuelle : le système présente à l'utilisateur le thésaurus du domaine et lui permet de naviguer entre les termes ;
- automatique et transparente pour l'utilisateur par insertion ou remplacement d'un ensemble de termes à partir de ceux fournis par l'utilisateur.

L'extension de requête est déjà une assistance « intelligente » à l'information en ce que le thésaurus permet de prendre en compte tous les termes de recherche qui désignent un même concept sans que l'utilisateur ait à composer tous ces termes. Dans notre cas, où le modèle de connaissance est transdisciplinaire (il incorpore plusieurs disciplines), cette fonction est très importante. Un même concept peut être désigné de façon différente selon la discipline dans laquelle il est envisagé. En ayant accès à des documents issus de différentes disciplines, mais abordant un même concept, le chercheur peut par exemple :

- savoir comment une pathologie dépend de plusieurs facteurs qui ne sont pas habituellement rapprochés parce qu'ils sont considérés dans des disciplines différentes ;
- savoir comment la pathologie est abordée en termes de soins dans des disciplines différentes.

4.1.3.1.2.1.2 Reformulation de requête

La reformulation de requête a pour objectif de jouer sur la fréquence documentaire et d'améliorer la pertinence des réponses du système en regard de la pertinence utilisateur. Cette idée part de trois hypothèses maintenant classiques [12, Bruandet] :

- Les termes significatifs ont une distribution inégale dans les documents (ils sont très occurrents dans certains, peu dans d'autres, sans que cela diminue nécessairement la pertinence de ces derniers pour l'utilisateur). Les termes non significatifs (que l'on appelle aussi les termes vides) sont, quant à eux, plus également distribués.

- Les termes utiles à la description des documents sont peu redondants entre eux. Les termes que l'on retrouve dans tous les documents ne sont donc pas de bons termes d'indexation, car ils décrivent un ensemble proche de documents qui ne sont pas nécessairement tous pertinents pour l'utilisateur (production de bruit).
- Les termes utiles à l'indexation des documents doivent en décrire un minimum (ne pas être trop sélectifs). Les termes qui ont une fréquence documentaire trop faible doivent donc être éliminés du thésaurus (production de silence).

La reformulation de la requête fonctionne alors selon ces principes :

- Elle fait décroître la fréquence documentaire en combinant deux termes fréquents de requête ou en les remplaçant par un terme plus précis et donc plus rare. Ici, le thésaurus permet de réduire le bruit documentaire.
- Elle augmente la fréquence documentaire en assemblant des termes peu fréquents par un terme générique plus fréquent. Le thésaurus permet dans ce cas de réduire le silence produit par certains termes de requête trop peu présents dans les documents et pourtant pertinent du point de vue de l'utilisateur.

4.1.3.1.2.1.3 Autocomplétion

Implémenter un thésaurus dans un moteur de recherche permet une fonctionnalité d'assistance à la recherche de type autocomplétion des termes de requête.

L'autocomplétion est un système de suggestion de termes de requêtes. Lorsque l'utilisateur commence à composer une requête dans un champ prévu à cet effet, le moteur de recherche lui suggère des termes qui commencent par les mêmes lettres que celles de son début de requête. De fait, les termes suggérés changent au fur et à mesure que l'utilisateur compose sa requête.

Outre que l'autocomplétion permet de saisir des requêtes plus rapidement, elle a aussi une fonction d'assistance à la requête par suggestion de termes. Dans notre perspective d'assistance intelligente à la recherche d'information, cette fonctionnalité a pour utilité de :

- Guider l'utilisateur dans la recherche d'information en lui suggérant un ensemble de termes appartenant à une même thématique de recherche. Exemple :

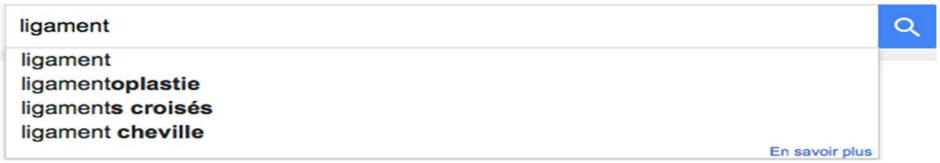


Figure 10 — Exemple d'autocomplétion

- Par le même effet, proposer à l'utilisateur des thèmes de recherche connexes à sa recherche d'information et ainsi lui suggérer d'autres pistes de recherche potentiellement en lien avec la recherche initiale. On est alors à mi-chemin entre l'assistance à la recherche d'information et l'assistance à la recherche scientifique.

On voit ici que l'autocomplétion peut être un outil de recherche ou d'étude. Elle peut aussi être un outil pédagogique en renseignant les étudiants sur les concepts pertinents à utiliser relativement à une thématique dans le domaine considéré.

4.1.3.1.2.1.4 Critères et sous-critères de recherche

Nous nous situions jusque-là dans un mode de recherche simple, effectuée dans un champ de recherche libre. Si nous nous projetons dans un mode de recherche avancée, le thésaurus peut aussi permettre une fonctionnalité d'assistance intelligente à la recherche d'information.

Nous entendons par recherche avancée une recherche qui s'effectue par la sélection de critères et éventuellement de sous-critères de recherche.

Dans un premier temps, un thésaurus permet de définir quels sont les critères et sous-critères pertinents pour une recherche dans le domaine considéré en reproduisant sa structure dans l'organisation sémantique des critères.

Dans un deuxième temps, en faisant dépendre la valeur des sous-critères de la valeur des critères maîtres sélectionnés, le thésaurus reproduit l'organisation du modèle de connaissance qu'il représente. Il guide la recherche et permet ainsi à l'utilisateur de toujours se situer dans la logique de son domaine.

Enfin, dans un troisième temps, suite à la formulation d'une première requête et l'obtention d'une première liste de résultats, l'utilisateur peut retrouver ces critères et sous-critères de recherche sous forme de facettes et les sélectionner afin de discriminer ou dé-discriminer les résultats obtenus (nous reviendrons sur les facettes de recherche dans la cinquième partie).

4.1.3.1.2.2 Fonctionnalités d'assistance à l'exploitation des connaissances du domaine

4.1.3.1.2.2.1 Suggestion de concepts associés

Toujours dans le cadre d'une assistance intelligente à la recherche d'information, le thésaurus peut aussi être un outil pour suggérer à l'utilisateur des concepts associés à ses termes de requête (entendus ici comme étant des concepts) ou aux concepts contenus dans les documents résultats obtenus (ce qui suppose que ces concepts aient été extraits des documents, bien évidemment).

Selon son degré de sophistication, c'est-à-dire selon le nombre et le type de relations entre les concepts qu'il embarque, le thésaurus peut suggérer des associations de différente nature.

Ainsi, si l'utilisateur compose, par exemple, une requête qui exprime un concept, il peut se voir suggérer :

- La cause d'un concept pathologie : Infection cause de fièvre.
- Le nom populaire du concept : Fièvre se dit aussi Température.
- L'ensemble auquel appartient le concept : l'estomac fait partie du système digestif.
- La classe à laquelle appartient le concept : l'estomac fait partie des organes.
- Etc.

Le système peut en outre proposer des documents relatifs aux concepts qu'il a suggérés à l'utilisateur.

Nous ne repréciserons pas l'ensemble des relations sémantiques qui peuvent être exploitées à partir d'un thésaurus (il est possible, pour cela, de se référer aux relations sémantiques que peut contenir un thésaurus détaillé dans la partie 4.1.1.1.3). Ce qui est important, c'est de comprendre qu'à partir de ces relations, un thésaurus qui reprend la structure d'un domaine de connaissance peut suggérer des concepts associés aux termes de requête et ainsi faire du SRI :

- Un outil de recherche d'information/recherche scientifique ;
- Un outil d'acquisition/transmission des connaissances d'un domaine.

4.1.3.1.2.2.2 Rattacher les concepts à un domaine de connaissance

Comme nous l'avons vu *supra* dans la première partie de ce document, les documents scientifiques contiennent parfois des connaissances qui n'y sont pas mentionnées de manière explicite, parce qu'elles n'ont pas été « identifiées comme telles » ni par les lecteurs, ni même par ses auteurs. Utiliser la logique transdisciplinaire pour exploiter les documents permettra d'en faire émerger ces relations de causalité scientifique qui n'y sont pas exprimées de manière formelle.

Pour cela, il faut que la solution permette :

- D'extraire les concepts contenus dans les documents.
- De mettre les concepts extraits des documents en relation avec d'autres concepts selon la logique du modèle princeps.

Ces fonctionnalités d'extraction/association permettent aux utilisateurs d'exploiter tout le potentiel conceptuel et applicatif que contiennent les documents.

4.2 Les ontologies

4.2.1 Définition

Il nous paraît ici intéressant d'envisager deux façons complémentaires de définir une ontologie soit, en la comparant avec un thésaurus, et en elle même.

4.2.1.1 Les ontologies par rapport aux thésaurus

Les ressemblances entre un thésaurus et une ontologie sont évidentes. Dans les deux cas, il s'agit d'un vocabulaire contrôlé, représentant des concepts utilisés et validés par les acteurs d'un domaine. Dans les deux cas, ce vocabulaire représente des concepts du domaine, et est structuré selon des relations sémantiques hypéronymiques, hyponymiques et associatives qui reprennent la structuration du domaine.

Mais si l'ontologie reprend la structuration du thésaurus ainsi que les relations entre concepts, elle le fait avec beaucoup plus de finesse et de façon plus formelle.

Cela s'explique parce que, contrairement à un thésaurus, qui sert principalement à indexer et rechercher des documents, voire parfois lier des concepts et des documents entre eux, une ontologie va aussi servir à instancier et raisonner, notamment en utilisant conjointement plusieurs règles d'associations ainsi que les propriétés qui sont attachées aux concepts. On peut dire qu'une ontologie est une extension des thésaurus en ce sens qu'elle incorpore des règles de déduction.

Par exemple :

Thésaurus :

Le terme « Voiture » est un terme spécifique de « Véhicule terrestre », lui-même spécifique de « Véhicule ». Le terme « Voiture » peut aussi se dire « Automobile », ou « Bagnole ». En anglais on dira « Car ». Si l'on s'intéresse à « Voiture », on peut également allez voir le terme « Autoroute ».

Ontologie :

« Voiture » est un type particulier de « Véhicule terrestre » (instanciation), lui-même un type particulier de « Véhicule ». La propriété « milieu de déplacement du véhicule » (terre, air, mer, espace) s'applique aux « Véhicules » ; les « Véhicules terrestres » sont tous les « Véhicules » pour lesquels « milieu de déplacement » vaut « terre » (raisonnement).

De plus, une ontologie se rapproche plus de la notion de « langage » que de celle de « vocabulaire contrôlé » en ce sens qu'elle embarque aussi une grammaire (règles de la langue) et une culture (partage de significations attachées aux objets)

4.2.1.2 Les ontologies en elles-mêmes

Si l'on cherche à définir les ontologies, non pas par rapport aux thésaurus, mais en elles-mêmes on peut dire qu'une ontologie définit les concepts d'un domaine (principes, idées, catégorie d'objet, notions potentiellement abstraites) et les relations entre ces concepts ainsi que des règles et axiomes qui les contraignent. Elles sont orientées vers l'expression des connaissances [12, Bruandet] et fournissent le vocabulaire spécifique à un domaine [13, Chaumier]. Par ailleurs, dans le domaine de l'ingénierie des connaissances, elles sont des « artefacts » élaborés dans le cadre d'une modélisation conceptuelle apte à jouer un rôle de référentiel conceptuel calculable par une machine. C'est pour cela qu'elles incluent des définitions lisibles en machine des concepts de base de ce domaine et de leurs relations [13, Chaumier].

Une ontologie correspond donc à un vocabulaire associé à un langage formel, c'est-à-dire une grammaire qui définit la façon dont les termes peuvent être utilisés relativement les uns aux autres.

4.2.2 Le fonctionnement d'une ontologie

Chaque concept d'une ontologie a des propriétés qui en décrivent les caractéristiques ou les attributs et auxquels il peut être assigné des valeurs. Les relations qui lient les concepts d'une ontologie sont contraintes par :

- Des axiomes ou des règles d'inférence permettant de définir les propriétés de ces relations ;
- Des règles de constructions possibles/interdites ;
- Des règles de déductions et de conditions de déclenchement de ces règles de déduction.

C'est l'ensemble de ces concepts auxquels sont attachées des caractéristiques et de propriétés, des relations entre ces concepts et des règles qui contraignent ces relations qui vont permettre à un SRI de tenir des raisonnements à partir d'une ontologie

Ainsi une ontologie peut se décrire comme un réseau de concepts avec un nombre potentiellement élevé de relations sémantiques et permettant :

- D'attribuer des propriétés aux concepts et des valeurs à ces propriétés ;
- D'établir des commentaires sur ces concepts ;
- D'introduire des règles permettant de tenir des raisonnements et d'inférer des informations nouvelles.

4.2.2.1 Les concepts

Dans une ontologie, les intentions sont organisées, structurées et contraintes pour représenter la conception du monde et ses contraintes (par exemple, une voiture est forcément un véhicule). C'est ce que nous avons appelé culture un peu plus avant (un partage de significations). L'ontologie capture les intentions et les lois qui les régissent, afin de rendre compte des aspects de la réalité. Ces aspects sont choisis pour leur pertinence dans les scénarios d'application considérés en recherche d'information.

Ainsi les concepts d'une ontologie peuvent contenir [15, Gandon] :

- Des connaissances de composition : en médecine, les catégories anatomiques, en chimie, les catégories d'éléments, etc. ;
- Des définitions complètes, par exemple : un regroupement de cellules est un tissu si et seulement si elles concourent à une même fonction ;
- Des contraintes d'intégrité : un parent ne peut pas être plus jeune que ses enfants ;

- Des fonctions de calcul : le rythme cardiaque conseillé pour une personne lors d'un effort cardio-vasculaire est égal à (220 – son âge) x 0.65 ;
- Des propriétés algébriques : la relation « est marié avec » est symétrique, cela signifie que si thomas est marié avec Stéphanie, alors le système peut aussi déduire que Stéphanie est mariée avec thomas, et vice-versa ;
- Des connaissances par défaut : par défaut, une articulation comprend un tendon ;
- Des règles spécifiques au domaine considéré : en biologie, pour chaque récepteur qui active une fonction moléculaire, si cette fonction joue un rôle dans le fonctionnement de l'organisme, alors le récepteur joue le même rôle.

4.2.2.2 Les relations entre les concepts

Comparativement à un thésaurus, une ontologie dispose de plus de types de relations sémantiques. Elle contient par exemple :

- Des relations inverses, par exemple : « faire partie de » est l'inverse de « inclure », c'est-à-dire que si un cartilage fait partie d'une articulation, alors l'articulation inclut le cartilage, et vice-versa ;
- Des relations d'évocation (« rupture des ligaments croisés » évoque « pratique sportive ») ;
- Des relations d'inversion (par exemple, si un vin a été produit par un établissement vinicole, alors on peut dire que l'établissement vinicole produit ce vin) ;
- Des relations morphologiques (« chloration » a pour morphème « chlore »).

Par ailleurs, ces relations entre concepts et leurs propriétés peuvent être utilisées conjointement par le système de recherche d'information.

Enfin, l'auteur de l'ontologie peut aussi définir lui-même des relations du référentiel, autrement dit il peut intervenir sur le langage d'indexation et donc, dans notre cas, sur l'exploitation du moteur de recherche, des documents et de leurs contenus.

Ainsi, dans l'exemple ci-dessous, l'auteur a décrit des relations telles que « employé par », « nomme », « situé à… », etc., qui permettront lors de la recherche d'identifier la nature précise du lien entre les concepts.

Figure 11 — Exemple d'une ontologie dont les relations ont été définies par l'auteur[20]

Pour être susceptibles d'être exploitées automatiquement par un moteur de recherche, les métadonnées doivent être entièrement explicites et exprimées dans un vocabulaire clairement et formellement défini. Les ontologies sont le réceptacle de ces définitions. On y représente les « valeurs » que l'on peut donner aux métadonnées et l'interprétation que les systèmes peuvent en faire, c'est-à-dire les concepts d'un domaine, les relations qu'ils entretiennent, la sémantique de ces relations et les règles de raisonnement qui leur sont applicables [13, Chaumier].

Enfin, et cela est très important, les relations sémantiques peuvent être utilisées conjointement et avec les propriétés attachées aux concepts.

C'est cette combinaison de concepts auxquels sont attachées des propriétés et qui entretiennent des relations cumulables et contraintes par des règles qui va permettre au système d'instancier et de raisonner.

4.2.3 Utilisation d'une ontologie dans un système de recherche d'information

Il nous est difficile de présenter ici l'ensemble des fonctionnalités d'assistance intelligente à la recherche d'information que permet l'implémentation et l'exploitation d'une ontologie dans un moteur de recherche. Ce domaine est en effet encore l'objet de nombreuses recherches et ce qui est évoqué dans les publications sur ce sujet n'est pas toujours transposable à la réalité. Là encore, nous avons

[20] http://www.dia-logos.net/ressources/schemas-de-classification-thesaurus-taxonomie-ontologie

décidé de nous intéresser aux exploitations d'une ontologie qui pourraient répondre aux attentes et besoins exprimés dans la première partie de ce mémoire.

Pour présenter ce à quoi peut servir une ontologie implantée dans un moteur de recherche en termes d'assistance intelligente à la recherche d'information, nous allons partir du principe que les ontologies héritent des qualités des thésaurus. Nous n'allons donc pas revenir sur ce que nous avons dit dans la section précédente, mais juste mentionner ce qu'elles peuvent faire de plus dans notre perspective d'assistance intelligente à la recherche d'information.

L'introduction d'une ontologie dans un moteur de recherche vise plusieurs objectifs [15, Gandon] :

- Réduire, voire éliminer, la confusion conceptuelle et terminologique et tendre vers une compréhension partagée pour améliorer la communication entre les utilisateurs et le système ;
- Déclarer formellement un certain nombre de connaissances utilisées pour caractériser les informations gérées par le système et se fonder sur ces caractérisations et la formalisation de leur signification pour automatiser des tâches de traitement de l'information ;
- Implanter des mécanismes de raisonnement déductif, de classification automatique et de recherche dans les systèmes de recherche d'information.

Dans un moteur de recherche, c'est par exemple :

- Améliorer la précision de la recherche d'information, en évitant des ambiguïtés au niveau terminologique (provenant de l'homonymie ou de la polysémie) ;
- Améliorer le taux de rappel de cette recherche d'information, en intégrant des notions plus précises ou équivalentes (en utilisant la synonymie, l'hyperonymie, l'hyponymie) ou en déduisant des connaissances implicites (par exemple, des règles d'inférence) ;
- Relâcher des contraintes trop strictes en cas d'échec de la requête (par généralisation) ;
- Regrouper des résultats trop nombreux selon leur similarité pour les présenter de façon plus conviviale (regroupement ou clustering conceptuel).

4.2.3.1 Désambiguïsation des requêtes

Dans un système d'information, l'ajout de la connaissance contenue dans une ontologie permet d'améliorer considérablement les capacités de désambiguïsation des requêtes du système. Prenons l'exemple d'un utilisateur qui cherche des livres écrits par un certain « Hugo ». Si le moteur de recherche se contente de faire de la fouille plein texte avec les mots clefs « Hugo » et « Livre », vont apparaître plusieurs problèmes :

- Le bruit : le système ne saura pas faire la différence entre le nom de famille « Hugo », le prénom « Hugo » ou le nom de rue « Hugo » ;
- Le silence : le système, s'il rencontre le terme « R-o-m-a-n », ne saura pas qu'il est pertinent pour votre requête, car il cherche le mot « L-i-v-r-e ».

Mais si, grâce à l'ontologie, le système connaît quelques aspects formels du monde sur [15, Gandon] :

- Les humains : Homme et Femme sont des sous-types d'Humain, qui est lui-même un sous-type d'Être Vivant ;

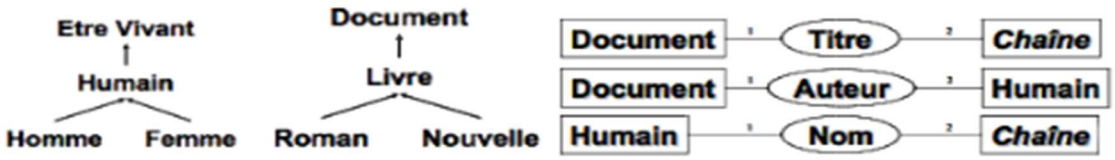


Figure 12 — Désambiguïsation d'un terme de requête par implémentation d'une ontologie 1/3[21]

- Les documents : Roman et Nouvelle sont des sous-types de Livre, qui est lui-même un sous-type de Document ;

Figure 13 — Désambiguïsation d'un terme de requête par implémentation d'une ontologie 2/3

- Les relations entre les deux, avec leurs signatures : par exemple, il existe une relation Auteur, qui peut s'établir entre un Document et un Humain.

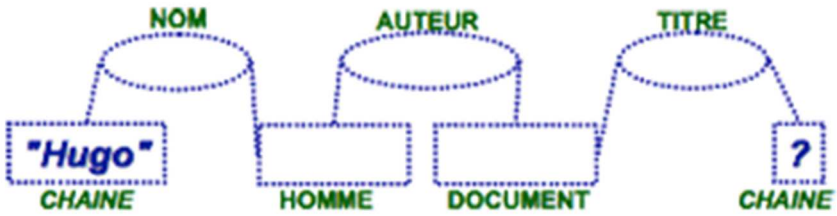


Figure 14 — Désambiguïsation d'un terme de requête par implémentation d'une ontologie 3/3

[21] https://interstices.info/jcms/c_17672/ontologies-informatiques

Et si l'on utilise ce vocabulaire pour décrire la réalité : un homme dont le nom est « Hugo » a écrit un roman intitulé « Notre-Dame de Paris ».

Alors on peut formuler une requête non ambiguë avec ce même vocabulaire pour rechercher les documents écrits par un certain « Hugo ».

Et en utilisant la logique du langage, le système peut inférer qu'un roman est un livre, un livre est un document, donc un roman est un document, et que la réponse « Hugo a écrit le roman Notre-Dame de Paris » est valide. (Voir les figures.)

Cette fonction de désambiguïsation n'est pas très pertinente dans le cas d'un SRI fermé qui n'interrogerait que des bases de données spécialisées avec une ontologie du domaine. Il est en effet improbable que des ambiguïtés de cette nature surviennent dans ce type de contexte informationnel. Mais la fonction peut-être utile si le moteur de recherche est aussi en mesure d'interroger l'Internet et donc de rechercher des documents spécialisés parmi tous les documents indexés de l'Internet [15, Gandon]. Par exemple :

Figure 15 — Exemple de l'intérêt de l'utilisation d'une ontologie de domaine

On voit dans l'exemple de recherche présenté ci-dessus à quoi aurait pu servir une ontologie des espèces animales implémentée dans un moteur de recherche dédié à des naturalistes.

4.2.3.2 Guider une recherche

Dans le cadre d'un nouveau projet de recherche scientifique qui a des similitudes avec un projet précédent, une ontologie permet de guider la recherche de documents afin d'identifier et de réutiliser les documents traitant de concepts proches de ceux contenus dans le nouveau projet envisagé [15, Gandon] ;

Une ontologie peut aussi guider le chercheur dans l'expression de sa requête, par exemple en en montrant les contraintes des signatures des relations, c'est-à-dire les instances qui peuvent être reliées et les règles de ces liens.

4.2.3.3 Annoter des documents

En couplant les ontologies avec des outils d'analyse de la langue naturelle, il est possible d'annoter formellement des textes avec les connaissances qu'ils décrivent. On peut, par exemple, extraire d'un article une relation de causalité entre un gène et une maladie et annoter cet article avec cette connaissance. Celle-ci pourra être ensuite utilisée afin de :

- Garder une trace de la source de chaque connaissance (et donc retrouver le document concerné à partir de la connaissance) ;
- Utiliser cette connaissance dans des raisonnements d'analyses d'expériences et pour la confrontation des résultats avec ceux contenus dans le document.

Pour ce faire ; l'ontologie pilote l'analyse de texte en fournissant au système :

- Les termes à chercher (représentations linguistiques associées aux concepts et aux relations et le sens qui peut leur être associé, soit les intentions) ;
- Les structures qui peuvent être extraites (les signatures des relations) des documents. Ces contraintes dirigent et focalisent l'analyse.

Dans un deuxième temps, l'ontologie est utilisée dans des inférences de recherche d'information (le système va « raisonner » pour mettre en relation et donc proposer à l'utilisateur des documents annotés avec des connaissances qui entretiennent des relations scientifiques).

4.2.3.4 Faire de la veille informationnelle

Un système d'information peut aussi utiliser une ontologie pour rendre plus performante la veille de l'utilisateur. Si un utilisateur identifie et annote un document intéressant pour sa recherche ou son entreprise dans le cadre d'une activité de veille, l'ontologie est utilisée pour :

- Caractériser le document (annotation automatique par repérage des concepts du domaine présent dans le document ou le décrivant) ;
- Archiver l'annotation avec des annotations similaires ;
- Identifier des documents qui ont été annotés avec les mêmes concepts (clustering) ;
- Identifier les profils de personnes potentiellement intéressées par ces documents et leur envoyer un message de notification (ce qui suppose que la solution dispose de fonctionnalités de profilage, nous y reviendrons plus tard).

4.2.3.5 Effectuer des raisonnements

Une ontologie permet d'établir des règles d'inférence, des raisonnements à partir de l'utilisation des relations établies entre les concepts.

Ex. : Concept de « boîte de vitesse » (BV)

« Boîte de vitesse automatique » (BVA) et » boîte de vitesse manuelle » (BVM) sont des spécifiques et les seuls spécifiques de « boîte de vitesse ».

Si j'ai une boîte de vitesse et que ça n'est pas une boîte de vitesse manuelle, alors c'est une boîte de vitesse automatique.

Ou encore : si dans tel document, la pathologie n'est pas mentionnée comme ayant été causée par tek facteur, c'est que c'est cet autre facteur qui en est la cause (extraction de connaissances induites).

4.2.3.6 Exploiter les modélisations

Comme nous l'avons vu, le but d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné. Bien sûr, il ne s'agit pas de modéliser pour modéliser, mais de modéliser pour répondre aux besoins de représentation et d'exploitation de ces connaissances. Voici un exemple de la façon dont les connaissances peuvent être exploitées à plusieurs niveaux de modélisation :

Homme — Patient

Femme → Patiente

Patient et patiente sont deux types de concept, sur lesquels le système ne peut rien « dire » sauf qu'ils sont différents, soit Hommes, soit Femmes.

Si je lie les concepts Hommes, Femmes, Patients et Patientes ainsi que leurs relations au concept Arthrose, je peux dire que les patients souffrant d'arthrose sont soit des hommes, soit des femmes.

Ou : j'ai un concept « primitif » (Patient), qui a une caractéristique Sexe. Les caractéristiques d'Homme et de Femme différencient ainsi les deux concepts primitifs selon la valeur qui leur est attribuée.

Patient → caractéristique — Féminin

Patient → caractéristique → Masculin

Si je lie les concepts Patients et Patientes au concept Arthrose, je peux dire que le fait pour un patient de posséder la caractéristique « homme » ou « femme » a une influence sur la probabilité de souffrir d'arthrose. Le système pourra alors raisonner à partir de ces caractéristiques, par exemple en suggérant un ensemble de pathologies similaires à l'arthrose et dont la prévalence dépend entre autres de la caractéristique Homme ou Femme.

Ou : même chose que dans le cas précédent, mais en plus la différence est explicitement liée au sexe puisque c'est une valeur de l'attribut.

Patient → attribut ->sexe → valeur — Femme

Patient → attribut → sexe → valeur → Homme

Ajoutons à cela deux contraintes : « La caractéristique Sexe du concept patient a nécessairement une valeur » et « La caractéristique Sexe du concept patient ne peut avoir qu'une seule valeur ».

Si je lie mes patients à la pathologie d'arthrose, je peux dire le risque de souffrir d'arthrose est lié à la valeur de l'attribut sexe du patient.

Si on ajoute d'autres attributs tels que le poids ou l'âge et que je leur assigne des valeurs, l'utilisateur peut commencer à :

- Se voir suggérer des liens conceptuels et donc des documents en fonction de cette modélisation (la solution voyant qu'il cherche selon un attribut facteur de risque peut lui proposer de le croiser avec d'autres attributs facteurs de risque, notamment si la valeur de ces attributs est définie pour qu'ils deviennent facteurs de risque) ;
- Exploiter le contenu des documents selon le même principe d'associations.

On voit à travers ces possibilités d'exploitation qu'une ontologie de domaine peut être un réel outil de recherche d'information et de recherche scientifique.

4.2.3.7 Faire des commentaires

Dans l'optique d'une utilisation par des humains, une ontologie permet de faire des commentaires sur les concepts ou leurs propriétés. Cela peut être intéressant notamment lorsque certaines ambiguïtés n'ont pu être levées sur la définition de certains concepts ou de leurs relations avec d'autres concepts. Ces commentaires peuvent en outre être accessibles depuis la solution de moteur de recherche.

4.2.3.8 Polysémie contextualisée

Une ontologie permet de faire varier le sens d'un concept selon son contexte d'expression. Ainsi, selon son contexte d'expression, le concept d'hypoglycémie peut représenter :

- Un symptôme ;
- Une maladie ;
- Un motif d'admission aux urgences.

Grâce à l'utilisation d'une ontologie, le moteur de recherche va savoir que si, dans un document, le concept d'hypoglycémie est associé à tel ou tel concept décrivant tel contexte, il aura tel sens. Par exemple, on peut penser que « insuline » est un concept qui attribue le contexte de Maladie au concept d'Hypoglycémie.

La possibilité de faire varier le sens d'un concept, notamment selon son contexte, permet que l'utilisateur d'une solution de moteur de recherche décide du sens qu'il souhaite voir attribuer au terme de sa requête (de sa propre initiative ou parce que la solution le lui propose). Cette contextualisation de la recherche permet d'être plus précis :

- Quant aux résultats de la requête ;
- Quant aux concepts suggérés à l'utilisateur ;
- Quant aux documents suggérés à l'utilisateur.

4.2.3.9 Enrichissement des documents des associations de concept et suggestions de documents

Grâce aux nombreux types de relation que peut embarquer une ontologie, le système va pouvoir « d'enrichir » les documents, c'est-à-dire étendre l'indexation d'un document et donc son utilisation, notamment transversale, par association de concepts.

Exemple : le « chlore » est un « désinfectant » ; le « désinfectant » a un lien morphologique avec la « désinfection » ; la « désinfection » est un « traitement de finition dans le traitement des eaux usées ». Ainsi, si « chlore » est un concept présent dans un document :

- « Désinfectant » et « traitement de finition » seront des concepts ajoutés aux métadonnées du document (permettant de la rechercher).
- « Désinfectant » et « traitement de finition » pourront aussi être suggérés à l'utilisateur comme concepts associés au document.
- Des documents relatifs à « désinfectant » et « traitement de finition » pourront être suggérés à l'utilisateur qui a composé la requête « chlore ».

L'ensemble des fonctionnalités permises par une ontologie que nous venons de présenter constitue à la fois des assistances intelligentes à la recherche d'information et des assistances à la recherche scientifique. Cela tient à la nature d'une ontologie qui vise à représenter les définitions, l'organisation et les règles d'organisation d'un domaine. Il est alors intéressant de constater qu'un outil de recherche d'information peut aussi être un outil de recherche scientifique.

Cinquième partie :
Autres fonctionnalités d'assistance intelligente à la recherche d'informations

5 Les fonctionnalités d'un moteur de recherche qui permettent de répondre à ces attentes d'assistance intelligente à la recherche d'information

Comme précisé précédemment, nous ne reviendrons pas, dans cette section, sur les fonctionnalités liées à l'indexation automatique des documents ni à l'exploitation de référentiels telles qu'elles sont présentées dans les sections précédentes. Cependant, certaines des fonctionnalités présentées ci-dessous nécessitent que l'une ou l'autre de ces techniques aient été employées.

Ces fonctionnalités sont de quatre types :

- Les fonctionnalités liées aux requêtes ;
- Les fonctionnalités de traitement des documents résultat ;
- Les fonctionnalités de présentation des résultats ;
- Les fonctionnalités d'extraction d'information des documents.

À la fin de cette section, dans une démarche prospectiviste, nous aborderons certaines pistes quant à la recherche d'information contextuelle qui laissent entrevoir quelques perspectives relatives à l'assistance intelligente à la recherche d'informations. Nous rattacherons ces pistes à l'objectif de valorisation et d'exploitation des connaissances dans l'organisation.

5.1 Les fonctionnalités liées aux requêtes

Les fonctionnalités liées aux requêtes sont de deux types :

- Celles concernant la capture de l'expression du besoin d'information, ici sous la forme d'une requête libre (il existe d'autres moyens, que nous n'aborderons pas, pour que l'utilisateur n'ait pas à formuler une requête. Ce sont, par exemple, les interfaces graphiques interactives zoomables dans lesquelles les concepts occupent une place correspondant à leur place dans le champ sémantique de leur domaine).
- Celles concernant le traitement des requêtes par le système.

5.1.1 Fonctionnalités relatives à la capture du besoin d'information : le modèle booléen

Tout moteur de recherche, si ambitieux et spécialisé soit-il, se doit d'intégrer un mode de recherche « simple » dans lequel la requête se compose librement et sans contrainte. L'absence de ce mode de recherche est un facteur d'échec connu des projets de moteur de recherche. Les utilisateurs réguliers

de moteurs de recherche sont en effet habitués à débuter leurs recherches de cette façon. Et pour beaucoup d'entre eux, ils ne passent pas à une requête experte (qui utilise des critères préétablis de recherche), même si les résultats qui leur sont retournés suite à une « recherche simple » sont insatisfaisants.

Mais parce qu'il n'impose aucune contrainte de composition de requête, et parce que ces requêtes sont la plupart du temps exprimées en langage naturel, les requêtes simples sont susceptibles de produire à la fois :

- Du bruit, du fait d'une trop faible précision sémantique des termes de la requête des requêtes ;
- Du silence, du fait d'une trop forte précision sémantique des termes de la requête des requêtes.

Tout l'enjeu pour le système de recherche est alors de retourner des réponses pertinentes à l'utilisateur, c'est-à-dire celles qui bénéficient d'un bon :

- Taux de rappel : les réponses du système correspondent à l'ensemble des documents pertinents contenus dans les bases de données interrogées par la solution (réduction du silence) ;
- Taux de précision : les réponses du système contiennent le minimum de documents non pertinents en regard du besoin informationnel de l'utilisateur (réduction du bruit).

Même s'il n'y répond pas parfaitement, le modèle booléen de recherche d'information a été la première réponse à ces problématiques. Aujourd'hui, il reste un modèle largement utilisé par les usagers. Directement dans un formulaire de recherche simple, ou à travers la sélection de critères de discrimination.

5.1.1.1 Le modèle booléen standard de recherche d'information.

Le modèle booléen est fondé sur la théorie des ensembles. Un document est représenté par un ensemble de termes qu'il contient ou qui le décrivent. Une requête correspond alors à une série d'opérations logiques qui permettent au système de retourner ou de ne pas retourner des documents en fonction de leur contenu.

Ces opérations logiques d'appariement s'effectuent à partir du calcul d'une requête composée avec les opérateurs de recherche booléens. Les opérateurs booléens sont des mots ou des symboles utilisés pour spécifier des relations entre deux termes de recherche. Voici comment fonctionne le modèle :

- **ET**, de ne retourner que les documents dans lesquels les termes A et B sont présents ;
- **OU**, de retourner les documents dans lesquels A ou B sont présents, conjointement ou non ;

- **SAUF**, de ne pas retourner les documents dans lesquels le terme qui suit l'opérateur est présent, même si les autres termes de la requête le sont.

Dans le modèle booléen [4, Poincot] :

- Les requêtes doivent être composées avec les termes exacts d'indexation et combinées avec les opérateurs booléens : dans la figure 16 ci-dessous c'est la requête A ET B sauf C qui a été composée.
- Les documents retournés sont ceux qui correspondent exactement aux requêtes

Par exemple, dans la figure ci-dessous, la requête « A ET C SAUF B » retourne à l'utilisateur les documents contenant les termes A et/ou C, mais ne contenant pas B.

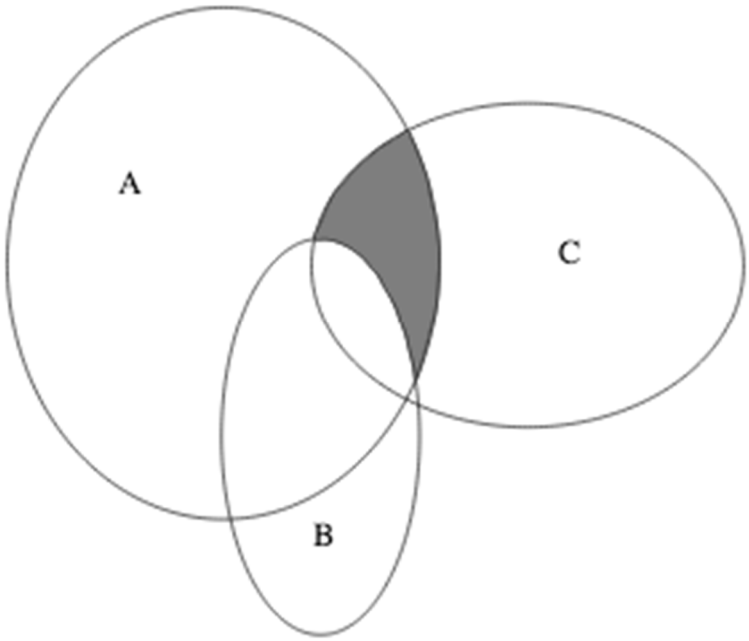


Figure 16 - Schéma correspondant à la requête « A ET C SAUF B »[22]

[22] http://www.classification-society.org/csna/mda-sw/inform/these-philippe-poincot/chap4.pdf

Dans beaucoup de systèmes de recherche, l'opérateur ET est implicite dès lors que la requête comporte plusieurs termes d'exploitation.

Les avantages du modèle booléen sont :

- D'opérer une sélection des documents retournés par le système plus fine que si la requête utilisateur ne comportait qu'une suite de termes sans opérateurs de sélection (réduction du bruit) ;
- D'être efficace en termes de temps de réponse, même pour les collections volumineuses ;
- D'être facile à expliquer et à comprendre ;
- D'être prévisible dans les résultats qu'il retourne.

Mais le système booléen comporte aussi quelques inconvénients :

- Il ne fonctionne bien que si l'on connaît exactement quels sont les termes à écrire pour retrouver les documents que l'on cherche.
- Les réponses du système ne peuvent pas être retournées dans un ordre qui mettrait en avant les documents susceptibles d'intéresser davantage les utilisateurs (pas de classement de pertinence).
- La sélection d'un document repose sur une décision binaire.
- Le modèle ne peut pas dépister des documents pertinents qui ne correspondraient qu'à une partie de la requête.
- La formulation de la requête peut être difficile pour les utilisateurs.
- L'emploi des opérateurs ET et OU ne correspond pas toujours à son emploi dans la langue naturelle. Par exemple, un besoin d'information concernant *la rotule et les tendons* ne se traduit pas par la requête « rotule ET tendon » mais « rotule OU tendon ».
- Pour les collections volumineuses, le nombre de documents retournés peut être considérable.
- Si la requête est trop stricte, le nombre de documents retournés est au contraire trop faible (silence).

L'une des conséquences de ces inconvénients est que, pour obtenir des résultats pertinents et un nombre suffisant de documents, il faut que l'utilisateur trouve un équilibre entre une requête trop stricte et une requête trop large. C'est un équilibre qui peut être difficile à trouver et demander un temps de recherche supplémentaire.

Pour pallier certains des inconvénients du modèle booléen et faciliter l'écriture d'un besoin d'information, le modèle booléen a été étendu à d'autres opérateurs de recherche.

5.1.1.2 Le modèle booléen étendu

Les opérateurs de recherche du modèle booléen étendu permettent de restreindre ou au contraire d'étendre les champs couverts par les équations de recherche, toujours dans le but de pallier les problèmes de bruit et de silence produits par une équation booléenne.

Les opérateurs de restriction de requête sont :

- **L'opérateur d'adjacences**, souvent noté « ADJ », qui donnent un sens plus strict à l'opérateur « ET » en ne retournant que les documents dans lesquels les mots de la requête sont côte à côte. Cet opérateur peut imposer que les deux termes de la requête soient présents, ou non, dans l'ordre de la requête.
- **L'opérateur de proximité**, souvent noté « NEAR » qui permet de retrouver des mots « géographiquement » proches les uns des autres, sans pour autant être obligatoirement adjacents. Il est possible de compléter cet opérateur par un nombre entier indiquant le nombre de mots pouvant s'intercaler entre les deux termes de la requête.

Les opérateurs d'extension de requête sont :

- **les opérateurs de troncature**, souvent représentée par l'astérisque « * », ou par le « ? » qui permettent de remplacer une ou plusieurs lettres d'un mot, et dont il existe trois formes :
    - **La troncature à droite** : placée à la fin d'un mot, la troncature à droite permet d'élargir la requête à l'ensemble des termes qui commencent par les mêmes caractères (par exemple : chev* donnera cheval, chevaux). La troncature à droite doit être utilisée avec parcimonie, car elles entraîner du ***bruit*** dans les résultats : biblio* donnera bibliothécaire, bibliographie, bibliophile, etc.
    - **La troncature à gauche** : placée au début, elle permet la recherche sur les préfixes (*phobe donnera anglophobe, agoraphobe, etc.).
    - **La troncature centrale** (ou « masque ») : placée au milieu d'un mot, ou d'une expression, elle permet d'élargir une requête aux variantes de ce mot ou de cette expression. Ex. : francopho*e donnera francophone ou francophobe.

5.1.1.3 Le modèle booléen hybride

Les réponses du modèle booléen standard sont fondées sur la théorie des ensembles, elles correspondent donc à un ensemble de documents qui satisfont la requête. Or, il est important pour les

utilisateurs que ces réponses soient classées selon le degré de satisfaction qu'elles apportent par rapport à la requête qu'ils ont composée. C'est l'un des objectifs du modèle booléen hybride que de retourner les documents dans une liste ordonnée dont l'ordre reflète ce degré de satisfaction [12, Bruandet].

Le modèle booléen hybride intervient principalement au moment de l'indexation.

Le système inclut une pondération pour chacun des termes retenus pour l'indexation d'un document. Ainsi, le système peut attribuer un poids différent aux termes selon qu'ils correspondent plus ou moins au contenu d'un document en lui associant un poids unitaire. Par exemple 1 si le terme correspond parfaitement au contenu du document, 0,5 s'il n'y correspond que partiellement, et 0,2 s'il n'y correspond que marginalement. Grâce à cette pondération, qui revient donc à pondérer le poids des termes de requête, les documents pourront être ordonnés selon qu'elles correspondent plus ou moins à la requête de l'utilisateur [12, Bruandet].

Certains opérateurs n'ont pas pour fonction d'étendre ou de restreindre le champ d'une requête, mais de lui apporter des spécificités particulières :

- les opérateurs arithmétiques :
    - l'opérateur égal à, « = », qui permet une recherche sur le nombre exact ;
    - les opérateurs supérieur à « > », supérieur ou égal à « >= », les opérateurs inférieur « < », inférieur ou égal à « <= », qui permettent une recherche sur des périodes ou des séquences de nombres.

Si ces opérateurs additionnels permettent de pallier certains inconvénients du modèle booléen strictement entendu, ils ne résolvent pas l'un de ses inconvénients majeurs, celui de ne pas permettre de classer les documents retournés par le système selon leurs degrés de correspondance avec la requête, c'est-à-dire selon leurs degrés de pertinence [12, Bruandet]. La recherche par critères, en offrant à l'utilisateur la possibilité d'affiner considérablement sa recherche, va augmenter la pertinence des réponses du système et lui permettre de les classer selon cette même exigence.

5.1.2 Fonctionnalités relatives à la capture du besoin d'information : la recherche par critères

La recherche par critères correspond en général, dans les SRI, au mode de recherche avancée. Elle consiste à fournir à l'utilisateur des critères de recherche prédéterminés. Ces critères peuvent concerner :

- Des catégories du domaine de connaissance considéré (squelette, organe, tissu) ;
- Des catégories relatives aux métadonnées bibliographiques (source, format, date) ;
- Des catégories relatives au type de documents (imagerie, cours, cas cliniques) ;
- Etc.

Il est possible d'imaginer autant de types critère que nécessaire tant que ceux-ci sont renseignés dans les métadonnées des documents.

L'intérêt de la recherche par critère est de guider l'utilisateur dans sa recherche d'information en lui fournissant les catégories pertinentes par rapport à son activité.

Les critères peuvent être organisés en critères et sous-critères de recherche et être interdépendants, c'est-à-dire que la sélection de la valeur de l'un peut modifier les valeurs disponibles de l'autre (par ex. la sélection d'une valeur dans le critère « zone corporelle » va modifier les valeurs disponibles à la recherche dans la catégorie « tissus »).

Les critères, même si leurs valeurs ne sont pas interdépendantes, peuvent se croiser afin d'affiner les réponses du système à la requête. C'est même l'un des principal intérêt de ce type de recherche (par ex. l'utilisateur peut croiser la valeur « estomac » du critère organe avec la valeur « cours » du critère « type de document » si l'objectif de sa recherche est de préparer des examens).

Enfin lorsqu'ils sont propres au domaine considéré et entretiennent des relations de dépendance, les critères tendent à en reproduire la logique. Ils peuvent alors être considérés comme des outils d'appropriation et d'utilisation de la connaissance de l'organisation.

Les outils linguistiques d'assistance à la recherche d'information du modèle booléen sont certes très utiles pour composer des requêtes fines dans le but d'obtenir des réponses très précises à un besoin informationnel, mais ils restent difficiles et quelque peu fastidieux à utiliser pour des utilisateurs « pressés » et qui ne sont pas des spécialistes de la recherche d'information. De plus, le « modèle

Google » de la recherche d'information a fait s'habituer les utilisateurs à composer des requêtes simples, courtes, sans opérateur de recherche et en langage naturel sans contrainte syntaxique liée à un langage de requête, ni contrainte terminologique liée à un langage documentaire [2, Vuillequiez]. Les modes de recherche avancée par critère rencontrent les mêmes écueils, ils sont souvent délaissés par les usagers, même experts du domaine, parce que leur utilisation leur paraît trop fastidieuse (ceux-ci se reportant plus facilement, suite à une recherche simple, sur les facettes, point que nous abordons un peu plus loin).

Dans ce contexte, c'est le système lui-même qui doit directement intervenir sur la requête utilisateur. L'une des façons d'intervenir sur la requête est de la traiter d'un point de vue linguistique et morphosyntaxique.

5.2 Le traitement des requêtes

Outre l'aide à la composition de requête avec l'utilisation d'opérateurs booléens ou de fonctions d'autocomplétions (comme nous l'avons vu dans la partie consacrée aux référentiels), le système de recherche d'information peut aussi assister les utilisateurs en opérant des traitements automatiques sur leurs requêtes. Ces traitements sont de deux nature, linguistiques et booléens. Une autre façon de traiter les requêtes est la fonctionnalité de « recherche floue ».

5.2.1 Le traitement linguistique et booléen des requêtes

Les traitements linguistiques et booléens de l'indexation sont également opérés, au moment de la recherche d'information, sur les termes de la requête. Ce traitement de la requête » s'effectue en plusieurs étapes [1, Hérigault] :

- Le moteur de génération de requête applique des traitements linguistiques. Les traitements linguistiques des requêtes sont similaires à ceux opérés lors de l'indexation des documents (élimination des mots vides, normalisation de la casse et de l'accentuation, transformation des mots en lemmes, etc.). Ils consistent en ce que le système intervienne sur la forme et les déclinaisons des termes de requête afin d'en étendre le champ d'interrogation. Ils ont pour but, à partir d'un terme de requête A, que toutes les déclinaisons linguistiques et morphosyntaxiques ainsi que les équivalents sémantiques de ce terme soient pris en compte par le système de recherche d'information. C'est un enrichissement de requête.

- Le moteur analyse la requête, la traduit en équation logique à l'aide d'opérateurs booléens entendus (c'est-à-dire essentiellement une pondération des termes de la requête) permettant ainsi d'obtenir un classement par pertinence. Les opérateurs booléens utilisés sont :
    - L'opérateur « Et avec contraintes », qui sert à spécifier une contrainte de proximité (plus les mots sont proches, meilleur sera le classement du document).
    - L'opérateur « Ou cumulatif », qui permet de rechercher les documents contenant n'importe quel sous-ensemble de mots parmi ceux demandés et classe en tête de liste les documents qui en contiennent le plus (c'est le plus utilisé).
- Le moteur réalise une opération de fusion des résultats par « collection, dédoublonnage et ordonnancement » [12, Vuillequiez].

Ce n'est que parce que ces opérations de traitement linguistiques sont effectuées aux bouts de la chaîne d'information qu'un SRI peut améliorer la pertinence de ses appariements de document.

5.2.2 Les apports des traitements linguistiques à la recherche d'information

En recherche, les traitements linguistiques et sémantiques qui ont été opérés à l'indexation et à la composition de requête, le moteur de recherche va être capable de :

- Procéder à des extensions linguistiques de requêtes de manière transparente pour l'utilisateur : reformulation de la requête avec les lemmes, extension de la requête aux singuliers et pluriels, masculins et féminins, aux formes fléchies et dérivées, à des mots proches orthographiquement et phonétiquement. Cela contribue à réduire le silence.
- Procéder à des extensions sémantiques de requête par l'interprétation des termes de la question et leur extension à des termes proches ou liés par une relation sémantique.
- Améliorer le taux de précision dans la recherche d'information grâce à l'utilisation des syntagmes, moins ambigus que les termes simples, comme entrées d'index (l'expression « heure de départ » est en effet plus précise que les deux mots pris isolément).

5.2.3 Recherche floue

La recherche floue se définit comme une recherche tolérante vis-à-vis d'approximations ou d'erreurs commises au moment de la composition de la requête (fautes d'orthographe, par exemple).

La recherche floue utilise des techniques statistiques, phonétiques ou de reconnaissance de forme. Elle restitue des résultats de manière automatique ou avec sollicitation de l'utilisateur (fonction « voulez-vous dire ? »).

Pour restituer des résultats à partir de recherches comportant des fautes d'orthographe, il faut que le moteur de recherche étende la requête à des mots ressemblants à ceux de la requête ou comportant des variantes orthographiques identifiées à l'indexation via des techniques statistiques de reconnaissance de forme.

5.3 Le traitement des documents résultats

Une autre façon d'assister les utilisateurs dans leur recherche d'information est d'intervenir, non pas sur les requêtes qu'ils composent, mais sur les documents résultats qu'ils obtiennent après une première requête. Parmi ces fonctionnalités, nous en avons retenu deux :

- La rétroaction de pertinence ;
- L'expansion de requête par cooccurrence.

5.3.1 Rétroaction de pertinence

La rétroaction de pertinence est une technique qui étend la portée de la recherche en incluant les termes issus des documents retournés par le système suite à une première recherche.

La pertinence des documents peut-être calculée de façon :

- Automatique et implicite par enregistrement des clics utilisateurs sur la première liste résultat ;
- Automatique et explicite en considérant que les documents en tête de la liste résultat sont les plus pertinents ;
- Manuelle et explicite par enregistrement de « notes de pertinence » attribuées par l'utilisateur aux documents qu'il a obtenus.

Lorsqu'un document est jugé pertinent par le système, celui-ci va étendre la recherche aux termes qu'il contient, autres que ceux de la requête. Ces termes, il va les chercher :

- Dans le contenu des documents (mais il faut alors qu'ils soient enregistrés dans un index ou référentiel pour que le système sache quels « termes d'expansion » choisir dans l'ensemble du document) ;
- Dans les métadonnées qui décrivent les documents (titre, résumé, liste de mots clefs, etc.)

Une fois qu'il aura collecté ces termes, le système retournera les documents qui les contiennent pour une même requête que la requête initiale.

5.3.2 Expansion de requête par cooccurrence

Nous avons vu dans le chapitre précédent comment un système pouvait générer une expansion de requête automatique en utilisant un référentiel linguistique.

Ici, il s'agit de générer une expansion de requête par repérage des cooccurrences de termes dans les documents ou dans les métadonnées qui les décrivent.

Le système va considérer que si pour un terme de requête A, il apparaît que le terme B est souvent co-présent dans les documents, alors B sera automatiquement pris en compte comme terme de requête associé à A.

Par exemple, si l'utilisateur compose une requête avec le terme « métacarpe », le système, constatant que le terme « phalange » est souvent cooccurrent à « métacarpe », va aussi retourner les documents ne contenant que le mot phalange.

L'expansion automatique de requête par cooccurrence fonctionne comme l'opérateur « OU » du modèle booléen.

Cette fonctionnalité d'expansion de requête est intéressante dans le cas de recherches scientifiques tels que nous l'avons présenté *supra*. Mais il est aussi susceptible de produire du bruit informationnel.

5.4 Fonctionnalités de présentation des résultats

5.4.1 Hiérarchisation des résultats

Suite à la composition d'une requête par l'utilisateur, la hiérarchisation des documents présentés en liste résultat est la première fonctionnalité d'assistance à la recherche d'information mise en œuvre par les moteurs de recherche. En effet, les critères de hiérarchisation peuvent être considérés comme

des aides à la sélection puisque les utilisateurs des SRI ont naturellement tendance à privilégier les résultats qui viennent en tête de liste, voire à se contenter de ceux présents dans la première page.

Nous ne reviendrons pas ici sur les critères de hiérarchisation d'un moteur de recherche puisqu'ils ont déjà été abordés dans la section consacrée aux pertinences système et utilisateur. En effet, ces critères de pertinence sont les critères de base que va utiliser le système pour hiérarchiser les documents qu'il retourne.

Cependant, il est important de noter que dans un contexte scientifique, il peut être intéressant pour un utilisateur de pouvoir modifier ces critères de hiérarchisation :

- Un critère de rareté plutôt qu'un critère d'occurrence pour privilégier les documents où l'occurrence d'un terme est faible peut aider à mettre à jour des rapports de causalité scientifique ;
- Un critère d'éloignement des termes plutôt qu'un critère de proximité ;
- Un critère de position des termes inversé ;
- Un critère inversé d'occurrence relative à la longueur du document.

Ces inversions de critères visent à trouver des termes qui n'occupent pas de position centrale dans les documents, mais dont la présence « périphérique » peut révéler l'implication de concepts à laquelle on ne s'attendait pas dans des phénomènes scientifiques.

5.4.2 Regroupement (clustering)

Les clusters sont des agrégats de documents considérés comme proches d'un point de vue lexical, et donc conceptuel. La proximité des documents est calculée par des outils de traitement automatique du langage tels que nous les avons vus (l'extraction des termes des documents ou des syntagmes issus d'un texte) et de « text-mining » (analyse du nombre d'occurrences d'un terme, du nombre de cooccurrences de plusieurs termes et de la fréquence d'apparition des termes dans un ensemble de documents) [1, Hérigault]. Ce sont ces technologies qui permettent la constitution automatique et « libre » d'agrégats (« cluster » en anglais) de termes ou de concepts. Cette construction automatique est donc indépendante de tout « plan de classement » préexistant [2, Vuillequiez].

Compris de ce point de vue conceptuel, le clustering a pour objectif de regrouper des documents selon des similarités de contenu.

Mais le clustering peut aussi regrouper les documents selon d'autres critères tels que :

- Le type de documents (documents pédagogiques, articles scientifiques, mémoires) ;
- Le format de documents (textes, vidéos, images, etc.).

Cette technique de regroupement permet d'atteindre les objectifs suivants [7, Nie] :

- Les réponses du système sont regroupées plutôt que mises en liste individuellement. L'avantage de cette présentation de résultats est que l'utilisateur peut avoir une idée globale des résultats que le système a trouvés assez rapidement.
- Si un document est pertinent à une requête, alors les documents similaires ont plus de chance d'être pertinents aussi. Ainsi, le clustering peut être aussi vu comme un moyen d'expansion de requête par similarité de contenu.
- Le nombre de clusters, par rapport au nombre de documents, est beaucoup plus petit. Ainsi, on peut accélérer le processus de recherche et de sélection : si l'utilisateur, relativement à un terme de recherche, sait qu'il a besoin d'accéder à des résumés d'articles scientifiques ou à des documents pédagogiques, il ira beaucoup plus vite en passant par la fonctionnalité de regroupement.

Par ailleurs, en construisant automatiquement des classes à partir des agrégats de termes, la méthode de regroupement est un potentiel outil de découverte scientifique permettant de mettre à jour la coprésence non repérée de concepts scientifiques dans les documents. Le clustering est une forme d'aide intelligente à la recherche d'information par traitement sémantique des documents.

5.4.3 Les facettes

Souvent, après avoir lancé une recherche, les utilisateurs, s'ils ne sont pas satisfaits des premiers résultats, ont trois options possibles :

- Trier la liste de résultats avec des fonctions de classement des résultats de recherche (par date, par auteur ou par ordre alphabétique, par langue) ;
- Aller de document en documents dans la liste résultats au risque de se « perdre » ;
- Reformuler la requête.

Or, sachant que le modèle le plus répandu de recherche d'information chez les utilisateurs repose sur une requête comprenant « un ou deux mots clefs » en langage naturel, il est important de leur donner les moyens d'affiner le résultat d'une première recherche.

Les facettes de recherche permettent précisément à l'utilisateur de filtrer les résultats de sa recherche selon un ensemble de critères propres à son organisation. Ainsi, on peut distinguer [1, Hérigault] :

- Les facettes propres à l'organisation et à son modèle de connaissance éventuellement représenté dans le système par un référentiel. Dans notre cas par exemple :
    - Échelle d'observation (segmentaire, plurisegmentaire, etc.) ;
    - Discipline (crânien, viscéral, etc.) ;
    - Etc.
- Les facettes génériques (métadonnées descriptives) :
    - Date de publication (ou modification) des résultats ;
    - Type de documents ;
    - Source d'origine ;
    - Langue ;
    - Etc.
- Les facettes relatives aux processus de l'organisation :
    - Code projet de recherche ;
    - Code discipline ;
    - Code produit ;
    - Etc.
- Les facettes sémantiques dont les valeurs sont automatiquement extraites du document et de ses métadonnées :
    - Termes voisins ;
    - Mots clefs ;
    - Mots ou d'expressions clefs liés à la recherche ;
    - Entité nommée ;
    - Etc.

Les facettes sont un moyen pour l'utilisateur de naviguer dans un grand nombre de sources d'informations avec une grande fluidité. Elles sont généralement présentées en complément ou à côté d'une liste de résultats.

Les utilisateurs peuvent alors les sélectionner, contraindre ou élargir le champ initial de leur recherche (en cela, elles reprennent souvent partiellement les critères de recherche avancée).

Les facettes sont interactives en ce sens que :

- Le nombre de résultats répondant à la fois à la requête effectuée et à la facette, est affiché.
- L'utilisateur peut directement observer les résultats de sa sélection sur la liste résultat attenante. Elles sont une aide d'autant plus efficace à l'affinage des résultats qu'ils indiquent le nombre de résultats associés à chacune de leur valeur. L'utilisateur peut ainsi progresser avec moins d'incertitude et anticiper sur l'effort de filtrage qu'il lui restera à fournir.
- La sélection d'une facette supplémentaire va affiner la requête courante en appliquant le critère lié à la facette en plus des critères déjà sélectionnés, pour ne présenter que les résultats pertinents par rapport à la réunion de toutes ces facettes ;
- L'utilisateur peut désélectionner les facettes qu'il a préalablement sélectionnées.

La recherche par facette est un compromis intéressant entre la recherche totalement « libre » et l'utilisation parfois fastidieuse des critères de recherche. C'est une façon de faire faire une recherche avancée à un utilisateur, sans qu'il ne s'en rende compte, en lui faisant sélectionner ou désélectionner les facettes (ou filtres) de recherche. En rendant les conséquences de ces sélections immédiatement observables dans la liste résultat, les facettes peuvent constituer un tableau de bord intéressant pour l'utilisateur.

L'intérêt des facettes est aussi de refléter l'organisation de l'entreprise, son activité, son langage, ses objectifs constituant ainsi un univers familier et donc propice à la recherche pour les utilisateurs.

Ces facettes enrichissent la navigation en liant les concepts selon d'autres aspects, et elles permettent d'améliorer la pertinence des résultats en augmentant le niveau de spécificité des requêtes. Pour y parvenir, les facettes peuvent elles-mêmes être subdivisées en sous-facettes pour enrichir les critères de filtrage. Par exemple, la facette « Organe » d'un domaine peut se diviser en « Organe selon la région du corps » ou « Organes selon la fonction.

5.5 Fonctionnalités d'extraction d'informations

Si la recherche d'information rapporte des documents ou des passages de documents, l'extraction d'information a pour objectif d'extraire des informations spécifiques et structurées d'un texte portant sur un domaine particulier. L'extraction d'informations documents utilise également largement les techniques de traitement automatique du langage.

Nous présentons ci-dessous deux applications qui nous paraissent pertinentes dans l'optique d'une assistance intelligente à la recherche d'information :

- Le résumé automatique parce que son résultat constituera un indice important de la pertinence d'un document pour l'utilisateur ;
- Le système « questions-réponses » parce qu'il nous paraît être une perspective prometteuse et complémentaire aux fonctionnalités que nous avons vu jusque-là.

5.5.1 Résumé automatique

Le résumé automatique consiste en la génération, par le système, à partir d'un document texte, d'un texte plus court permettant à l'utilisateur de se faire une idée du contenu du document. Un bon résumé doit indiquer à l'utilisateur les informations principales contenues dans le document d'origine afin de lui permettre de décider si elle est pertinente.

Les processus d'indexation automatique et de résumé sont très proches. L'indexation a pour objectif de décrire le contenu d'un texte au moyen de descripteurs ; résumer un texte consiste également en la description d'un contenu, mais sous forme textuelle et en appliquant un taux variable de réduction. En outre les « résumeurs automatiques » utilisent souvent les mêmes techniques d'extraction que pour l'indexation [17, Chaudiron].

Il existe trois méthodes pour générer un résumé automatique à partir d'un document (18, Dalbin) :

- L'abstraction, fondée sur la compréhension du texte, produisant un résumé basé sur la reformulation ou la fusion des idées ou des phrases du texte ;
- La compression qui, en supprimant des éléments jugés non essentiels, aboutit à une réduction des phrases (pouvant aller jusqu'à 33) ;
- L'extraction d'unités textuelles (phrases ou parties de phrases) fournissant, après leur réorganisation, un aperçu du contenu du document.

Quelle que soit la méthode envisagée, le schéma général du fonctionnement d'un système de résumé automatique comprend trois étapes :

- L'analyse du texte source pour identifier l'information pertinente (prétraitement et sélection) ;
- La représentation de l'information par généralisation ou par extraction ;
- La génération d'un nouveau texte (résumé).

5.5.1.1 Les méthodes fondées sur la compréhension du texte

Cette méthode est fondée sur la compréhension du texte source. Elle consiste à construire une représentation du sens du texte d'origine, puis à extraire les informations essentielles de cette représentation à partir desquelles le système va générer le résumé automatique.

On y retrouve les techniques qui visent à représenter le contenu du texte en s'appuyant sur la structure argumentative de celui-ci et sur les différents niveaux linguistiques, notamment sémantique.

Cette méthode comporte trois étapes :

- La compréhension du texte source, c'est-à-dire la construction d'une représentation de son sens ;
- L'extraction des informations essentielles de cette représentation ;
- La génération à partir de cette extraction d'un texte en langue naturelle.

Comme pour l'indexation automatique, cette méthode fait face aux difficultés du traitement du langage naturel, notamment en ce qui concerne la construction d'une représentation du sens, car elle nécessite un besoin important de connaissances sur le domaine (pour la compréhension et l'extraction des informations essentielles) et sur la langue (nécessaire en génération de texte).

Parmi les techniques utilisées pour construire une représentation du sens du texte d'origine, on retrouve les techniques qui visent à représenter le contenu du texte en s'appuyant sur sa structure argumentative et sur ses différents niveaux linguistiques, notamment sémantique [12, Bruandet]. Ces techniques sont au nombre de trois :

- La première technique consiste à s'appuyer sur un référentiel terminologique préalablement défini pour construire une représentation du sens du texte.
- La deuxième technique consiste à extraire automatiquement du document des unités qui sont considérées comme des candidats-descripteurs. Le système d'indexation choisit ensuite les plus significatifs comme descripteurs du document. Cette technique est essentiellement fondée sur des techniques de traitement linguistique telles que celles que nous avons vues dans la section consacrée à l'indexation automatique (traitement des niveaux morphologique, syntaxique et sémantique de la langue).
- La troisième technique consiste à structurer les sources d'informations en rendant explicites les relations sémantiques qui peuvent exister entre les différentes unités informationnelles que contiennent les documents.

5.5.1.2 La méthode par extraction

La seconde méthode consiste à extraire du texte initial les phrases les plus importantes et à les mettre bout à bout pour construire le résumé. C'est une méthode par extraction d'unités linguistiques jugées représentatives [12, Bruandet]. Elle est dite superficielle [18, Dalbin] parce qu'elle s'intéresse aux formes linguistiques et non aux contenus sémantiques.

La « méthode par extraction » vise à repérer les unités textuelles supposées informatives puis à sélectionner les plus significatives en fonction du taux de réduction souhaité. Les phrases significatives d'un texte peuvent repérées par :

- Leur position : une phrase en début de paragraphe sera jugée plus pertinente qu'une phrase en milieu de paragraphe.
- La fréquence relative des mots qu'elles contiennent : une phrase contenant certains mots clefs ou cooccurrences de mots clefs sera jugée plus pertinente.
- Un calcul de similarité lexicale (si plusieurs phrases ont une similarité lexicale, c'est qu'elles portent un sens qui mérite de figurer dans le résumé).
- La comparaison à des patrons de phrase : certains schémas de phrases correspondront, ou non, à des types à extraire : l'analyse s'appuie sur la structure discursive du texte à résumer (sa logique, ce qui procède de sa structure de raisonnement) en identifiant, par exemple, certains mots ou expressions qui jouent un rôle clef dans la structuration du document (*en premier lieu, en conclusion, il est important de..., l'idée essentielle de ce texte est...,* etc.). Ces formes de surface donnent des indices utiles pour repérer les passages importants d'un document.
- Par des chaînes de référence lexicale (p. ex. : les anaphores) qui permettent d'identifier les phrases porteuse de sens.

On ne peut pas vraiment considérer comme un véritable système de résumé automatique au sens linguistique du terme. Avec cette technique par extraction de syntagmes ou groupes nominaux plus ou moins complexes, le « résumé automatique est plus une suite d'extraits significatifs qu'une véritable « condensation » du texte d'origine [17, Chaudiron]. Néanmoins, ces outils donnent des résultats intéressants en fournissant des « clefs de lecture » pour l'accès au texte [12, Bruandet]. Ils sont par ailleurs les plus utilisés pour produire des résumés automatiques.

5.5.2 Les systèmes de question-réponse (SQR).

Les SQR exploitent des requêtes formulées à l'aide du langage naturel et non plus en se fondant uniquement sur des mots clefs. Leur objectif est de fournir une réponse précise à une question posée.

Cette tâche diffère de la recherche d'information, car, au lieu de retourner un document entier, le système retourne l'information pertinente que contient le document [12, Bruandet].

Comme nous l'avons vu précédemment, des moteurs de recherches « classiques » proposent une suite de documents classés selon l'estimation de leur pertinence. L'utilisateur doit ensuite effectuer un tri dans ces documents et parfois en consulter plusieurs afin d'obtenir l'information complète dont il a besoin. Dans le cas d'un système de réponse aux questions, le système reconstruit une réponse en langage naturel à partir des contenus des documents qui contiennent partiellement ou en totalité les réponses à la question formulée.

Le SQR a des applications dans de nombreux domaines [12, Bruandet] :

- l'accès à des bases de données pour retrouver des valeurs associées à des champs ou pour accéder à de l'information contenue dans des documents textes exprimés en langage naturel ;
- l'accès à des informations « fines » contenues dans les documents ;
- l'accès fin à de grandes collections de documents.

Cela permet donc de gagner énormément en efficacité lors de la consultation du document pour obtenir l'information désirée.

Ce type de recherche nous semble être très intéressant, notamment dans le secteur scientifique où elle constituerait un complément remarquable des processus d'exploitation des connaissances contenues dans les documents tels que nous les avons vus dans le chapitre consacré aux référentiels. C'est pourquoi nous allons y consacrer un peu plus d'espace que pour les fonctionnalités précédemment exposées.

La recherche d'une information dans un SQR peut se décomposer en trois tâches [12, Bruandet] :

- l'analyse de la question (type de réponse attendue, extraction de termes clefs et rapprochement du terme avec un référentiel) ;
- la recherche d'une sous-partie d'un document susceptible de contenir la réponse (zone du document, champs d'une base de données, document appartenant à une collection) ;
- la sélection des passages contenant la réponse ;
- l'extraction de la réponse (réalisation des inférences nécessaires, accès à plusieurs sources si nécessaire, formulation ou reformulation de la réponse.).

Prenons l'exemple d'une question pour laquelle la réponse est une entité nommée et l'espace de recherche une collection d'articles de journaux[23] :

- Les questions sont analysées selon leur type :
    - Questions factuelles : « Où a été signé l'armistice ? »
    - Questions booléennes (oui ou non) : « François Hollande est-il marié ? »
    - Définitions : « Que signifie le sigle RATP ? »
    - Causes/conséquences : « Pourquoi le ciel est-il bleu ? »
    - Procédures : « Comment aller à Vierzon ? »
    - Listes : « Citer 3 présidents français »
    - Requêtes évaluatives/comparatives : « Quelle est la plus grande ville de France ? »
    - Opinions : « Que pensent les Français du concubinage de François Hollande ? »

- En fonction de l'objet de la question, des réponses-types sont élaborées. Le type de la réponse attendue correspond à l'identification de l'objet de la question ou du type de la phrase attendu. Par exemple :
    - Exemple de type d'objets :
        - Personne : « Quel président français... » → « Qui... »
        - Organisation : « Quelle compagnie... » → « Qui... »
        - Lieu : « Dans quelle ville... » → « Où... »
        - Date : « En quelle année... » → « Quand... »
    - Exemple de type de phrase :
        - Explication : « Pour quelle raison le ciel est-il bleu ? » → « Pourquoi... »,
        - Procédure : « Quelles sont les étapes pour renouveler un passeport ? » → « Comment... »

- Les questions sont analysées (les termes peuvent en être lemmatisées) et transformées en requêtes. Des documents sont associés aux requêtes. Exemple de transformation de questions en requêtes par extraction de mots clefs :
    - « Qui est Clint Eastwood ? » → requête : Clint Eastwood.
    - « Que fabrique l'entreprise Majorette ? » → requête : Majorette; entreprise; fabriquer.
- Les documents appariés sont analysés par un traitement automatique des langues « superficielles » (indexation sur des termes complexes, reconnaissance des entités nommées).

[23] https://fr.wikipedia.org/wiki/Systèmes_de_questions-réponses

Une première série de passages candidats est sélectionnée. Ceux-ci peuvent être composés de simples phrases, de paragraphes ou de documents entiers. Ces passages contiennent ou ne contiennent pas de réponses candidates. Par exemple :

- Question : Qui est le père de Mazarine Pingeot ? → Type de réponse « personne » :
  - Passage candidat : paragraphe « François Hollande et Mazarine Pingeot inaugurent la Bibliothèque François Mitterrand. »
  - Réponse candidate dans le passage : « Mazarine Pingeot a évoqué l'action de son père, François Mitterrand, en faveur de la culture. »

- Le système évalue la qualité des passages afin de, si nécessaire, réajuster les mots clefs utilisés par la requête. Il tiendra aussi compte du nombre de passages obtenus. S'il en obtient trop, il restreindra le nombre de mots clefs. S'il n'en obtient pas assez, il étendra la requête.
- Une fois une série de passages sélectionnés, il leur attribue un score afin de les classer.
- Les documents sont ensuite analysés par des techniques de TAL plus fines, permettant d'établir une similitude entre leur contenu et la question par résolution des variations linguistiques. Exemple de variations que traitent les TAL :
  - Variation morphologique : « Où se trouve la *capitale de l'Europe* ? » ou « Où se trouve la *capitale européenne* ? »
  - Variation lexicale : « Comment s'appelle la reine de *Hollande* ? » ou « Comment s'appelle la reine des *Pays-Bas* ? »
  - Variation syntaxique : « Moscou compte 9 millions d'habitants » ou « Les 9 millions d'habitants de Moscou ».
  - Variation sémantique : « Comment Adolf Hitler est-il *mort* ? » où la réponse peut être « Adolf Hitler *s'est suicidé.* »

- En fonction de cette similitude linguistique, et pour extraire les documents candidats, le système attribue un score aux réponses candidates selon des critères prenant en compte :
  - Le contexte global : on essaie d'évaluer la pertinence du passage d'où provient la réponse candidate. Pour cela, on se fonde sur :
    - Le nombre de mots clefs présents dans le passage ;
    - Le nombre de mots communs à la question et au passage ;
    - Le nombre de mots de la réponse candidate qui ne sont pas des mots clefs de la question ;

        - La distance moyenne entre la réponse candidate et les mots clefs présents dans le passage.
    - La justesse du type sémantique : on s'assure que le type de la réponse candidate est soit du même type, soit un sous-type du type de réponse attendue (adjectif, verbe, nom propre).
    - La redondance : on sélectionne les réponses présentes dans le plus possible de passages candidats possibles.
    - Les relations syntaxiques : on analyse la syntaxe de la question et des passages candidats. Par exemple :
        - Questions : « Qui a écrit Les Misérables ? » Relation : [X, écrire], [écrire, Les Misérables] ;
        - Réponse candidate : « Victor Hugo a écrit Les Misérables » Relations : [Victor Hugo, écrire], [écrire, Les Misérables] ;
        - À partir de cette analyse, le système peut répondre « Victor Hugo ».
- Le système peut aussi introduire une analyse sémantique supplémentaire en tenant compte du type de relations liant les éléments par construction d'un arbre de dépendance à partir de la question.

On voit à travers la description du processus d'un système de question-réponse que celui-ci est prometteur, mais complexe. Les traitements en langage naturel qu'il demande peuvent nécessiter de faire appel à des ontologies d'entités nommées ou des ontologies du domaine considéré afin de permettre l'extraction d'informations sémantiques riches [12, BRUANDET].

Par ailleurs, les SQR rencontrent encore de nombreuses limites liées notamment aux difficultés de traitement de la langue naturelle et d'interrogation des documents :

- Quand la réponse à une question est répartie sur plusieurs documents, il est difficile de construire une réponse à partir de passages issus de ces différents documents.
- Certaines questions doivent être décomposées. Par exemple, « Le président français est-il marié ? » implique « Qui est le président de la France ? », puis « François Hollande est-il marié ? ».
- Le traitement de la langue n'est pas encore parfait, certains traitements ne fonctionnent pas ou de manière non optimale : le traitement des anaphores, des synonymes, des paraphrases, des métonymies, de la négation, des quantifieurs (unités), la reconnaissance des figures de style.
- Certaines questions demandent que le système fasse appel à des systèmes d'inférences complexes pour élaborer les réponses : « Lille est la 2e plus grande ville de France » ; « Londres

est plus grand que Lille »; « Paris est plus grand que Londres » ; « Paris est en France ». Donc Paris > Londres > Lille, etc.

- Les questions portant sur les relations entre objets d'information ne sont pas encore résolubles : « Quelles sont les relations entre ostéopathes et son patient ? »
- Il est difficile de faire de l'implication textuelle, c'est-à-dire de faire en sorte que le système soit capable de reconnaître qu'un passage peut en impliquer un autre : par exemple, « Nicolas Sarkozy a fait refaire un avion destiné à ses transports officiels » implique « Un avion est destiné aux voyages du président français ».

Ces limites, ainsi que la complexité de leur implémentation font que les SQR se situent encore à la frontière entre les technologies d'usage et les technologies à venir. De nombreuses autres fonctionnalités d'assistance à la recherche d'information sont envisagées dans un futur proche. Parmi celles qui nous intéressent le plus, il y a celles qui tiennent compte du contexte utilisateur.

5.6 La recherche contextuelle

La valeur dune information dépend d'un grand nombre de paramètres contextuels ou de « points de vue. L'un des défis de la recherche d'information, c'est de considérer chaque utilisateur de manière singulière, avec sa propre perception de ce qu'est un contenu pertinent ou pas. Dans l'environnement informationnel de l'entreprise, cela consiste à trouver un moyen de répondre aux attentes individuelles à l'aide d'un outil « générique » (qui est le même pour tous les utilisateurs) [2, Vuillequiez].

5.6.1 Contexte et pertinence

L'aspect contextuel fait ici référence à la connaissance implicite ou explicite des intentions de l'utilisateur, de son environnement et du système lui même. La recherche contextuelle vise à modéliser les différents aspects de ces contextes pour les intégrer dans un processus de recherche de l'utilisateur. L'hypothèse est que si l'on rend explicites certains éléments du contexte de la recherche d'information, cela pourrait en améliorer sensiblement les performances, c'est-à-dire la pertinence de ses réponses aux requêtes utilisateur [12, Bruandet].

On retrouve ici la notion de pertinence qui peut s'aborder du point de vue du contexte. Une réponse du système à une requête n'aura pas la même pertinence selon le contexte de cette requête. Dans « La recherche d'information dans les environnements numériques », Dinet souligne que les cinq

dimensions de la pertinence en recherche d'information définies par Saracevic[24] sont fortement liées à la notion de contexte. Ces dimensions sont [3, Dinet] :

- La pertinence système ;
- La pertinence thématique ;
- La pertinence de situation ;
- La pertinence cognitive ;
- La pertinence affective.

On voit là aussi que, mis à part la pertinence système, la nation de pertinence est très liée au contexte utilisateur. Dans l'idéal, les paramètres contextuels devraient pouvoir être modélisés et calculés par le système afin d'adapter ses réponses aux requêtes en fonction du contexte dans lequel elles ont été émises.

5.6.2 Une modélisation difficile

La notion de contexte recouvre de nombreux aspects et peut se comprendre comme [11, Mothe] :

- Le sens qui est attribué aux termes (termes de requête ou termes contenus dans les documents) : dans une même organisation ou au sein d'une même communauté de chercheurs, le sens attribué à un terme peut différer selon plusieurs facteurs :
    - Un sous-domaine du domaine dans lequel travaille l'ensemble des utilisateurs du système : de quel point de vue l'utilisateur cherche-t-il ?
    - La genèse cognitive particulière de l'utilisateur considéré : quelles sont ses références ? Comment se sont-elles construites ?
- Le besoin d'information auquel vise à répondre la recherche d'information. Selon son besoin, l'utilisateur ne cherchera pas à obtenir le même type ni le même nombre de documents. En fonction de l''objectif de la recherche d'information, veut-il :
    - Connaître la réponse à une question précise ?
    - Réaliser une étude par rapport à un domaine ?
    - Vérifier une information ?

De plus, ces aspects du contexte dépendent parfois eux-mêmes d'autres aspects contextuels : par exemple, le besoin d'information auquel vise à répondre la recherche d'information dépend de l'organisation dans laquelle évolue utilisateur, des objectifs de cette organisation et de la

[24] Saracevic T. « Modeling interaction in information retrieval (IR) : A review and proposal », Journal of the American Society fo Information Science, 33, 3-9, 1996.

correspondance de ces objectifs avec ceux de l'utilisateur au moment de sa recherche (le paramétrage du système pourra-t-il permettre de répondre à ces deux objectifs ?).

C'est en partie cette variété et cet enchevêtrement qui rend ces aspects difficiles à modéliser et à exploiter dans un système de recherche d'information. L'autre difficulté vient du fait que le système doive objectiver des critères de détermination du contexte parfois subjectifs, c'est-à-dire :

- Apprendre et modéliser les contextes et les traitements associés à chaque contexte ;
- Reconnaître un contexte lors d'une nouvelle recherche.

Nous présentons ci-après des pistes de recherche relatives à des fonctionnalités de contextualisation de la recherche d'information. Ces fonctionnalités n'ont pas toutes été rendues opérationnelles, loin s'en faut, notamment du fait de la difficulté d'extraire et de représenter la connaissance concernant les utilisateurs. Néanmoins, il est intéressant de penser les futurs systèmes de recherche d'information comme capables d'utiliser la notion de contexte pour inférer des caractéristiques sur les besoins d'information et utiliser ces caractéristiques dans les modèles de recherche et les réponses produites.

5.6.3 Les fonctionnalités de recherche contextuelle

Nous ne pouvons pas aborder dans ce mémoire l'ensemble des éléments potentiellement associés au contexte dans un SRI. Ceux-ci sont trop nombreux et parfois trop éloignés de potentielles applications concrètes. Nous allons donc nous attacher à deux aspects de la prise en compte du contexte qui nous paraissent pertinents dans le cas de l'assistance intelligente à la recherche d'information au service de la connaissance en organisation et qui connaissent ou devraient rapidement connaître des traductions applicatives, à savoir :

- le traitement des requêtes ;
- la recherche collaborative.

5.6.3.1 Le traitement des requêtes

Le traitement contextuel des requêtes est un moyen d'améliorer la pertinence des réponses qui leur seront retournées, on non, par traitement linguistique et sémantique, mais en traitant le contexte de l'émetteur de la requête. Nous en avons retenu deux types [11, Mothe] :

- Le traitement contextuel des requêtes populaires ou répétées ;
- Le traitement contextuel des requêtes reformulées ou de l'historique des requêtes.

5.6.3.1.1 Les requêtes populaires ou répétées

Les moteurs de recherche ont développé des outils permettant de repérer les requêtes les plus « populaires », soit celles qui sont le plus composées par les utilisateurs du système.

Une étude[25] a montré que pour les besoins de recherches spécialisés (dans notre cas, celles considérées), la répétition correspond à environ 55-60 % des requêtes. Une autre étude indique que 40 % des requêtes du corpus ont pour objectif de répondre à une recherche déjà effectuée auparavant et que plus de 33 % des requêtes sont identiques à une requête déjà posée par le même utilisateur. Enfin, plusieurs travaux dans le domaine de l'analyse des requêtes montrent que les requêtes récurrentes sont fréquentes et peuvent être identifiées [8, Denos]. Ces études sont principalement menées dans des SRI « ouverts », c'est-à-dire qui effectuent leurs recherches sur l'Internet. On peut aisément imaginer que leurs résultats seraient confirmés dans des SRI utilisés par une communauté partageant les mêmes références et objectifs de recherche et interrogeant des bases de données spécifiques et spécialisées.

Ces résultats sont à mettre en relation avec ce que nous indiquions dans l'introduction de ce mémoire ainsi qu'avec l'une des problématiques de la société Ostéobio : il y a potentiellement une perte importante de temps et donc de valeur, dans les organisations, liée à la méconnaissance et donc à la non-exploitation d'un travail qui a déjà été effectué.

Pour répondre à cette problématique, plusieurs pistes de recherche sont envisagées :

- Une première piste consiste en la filtration d'informations par rapport à un profil utilisateur. La requête récurrente est considérée comme un profil qui permet de filtrer les documents.
- Une deuxième piste concerne les principes de recommandation. Il s'agit de suggérer à un utilisateur des documents par rapport à ceux qui ont satisfait d'autres utilisateurs en exploitant les similarités et les différences entre les profils utilisateurs afin de leur recommander un objet particulier. Cette technique est pertinente pour les requêtes récurrentes, notamment si on les croise avec les retours de pertinence des utilisateurs (implicites, comme les clics, ou explicites, comme les tags, commentaires, notes, etc.).

[25] SMYTH B, BALFE E, FREYNE J, BRIGGS P, COYLE M, BOYDELL O. « Exploiting Queyring Repetition and Regularity in an Adaptive Communuty-Based Web Search Engine ». User Model. User Adapt. Interact, 14(5), p. 383-423, 2004. **In** Recherche d'information contextuelle, assistée et personnalisée. Lavoisier, Paris, Hermès-Science, 2011, 302 pages. Recherche d'information et web. Page 27 – 70. ISBN 978-2-7462-2583-1 ISNN 1968-8008.

5.6.3.1.2 Les requêtes reformulées ou historiques des requêtes

L'utilisation des requêtes reformulées et de l'historique des requêtes pour améliorer la réponse du système aux requêtes de l'utilisateur, et donc réduire la distance entre la pertinence système et la pertinence utilisateur, est un bon exemple de traitement contextuel de la recherche d'information.

Ici, le contexte est constitué de l'historique des requêtes et des reformulations de requête qui y sont enregistrées.

L'historique des requêtes reformulées par l'utilisateur pour un même besoin d'information permet de :

- désambiguïser certains termes : les requêtes précédentes fournissent un contexte d'utilisation du terme de requête ;
- disposer des documents retrouvés pour les requêtes précédentes : les documents communs à plusieurs listes résultats pour un même besoin d'information sont alors considérés comme les plus pertinents.

Si, en plus, on utilise les « clics » des utilisateurs sur les documents présentés en liste résultat comme retour implicite de pertinence en les combinant avec l'historique des requêtes reformulées, on constate une nette amélioration de la précision moyenne des résultats.

5.6.3.2 La recherche d'information collaborative

La dimension sociale de la recherche d'information est très importante, notamment dans les organisations. Partage de document, recommandation de sources, liens de citation de documents, critère de popularité, on consulte plus facilement une information si elle nous a été recommandée par un utilisateur dans le jugement duquel on a confiance ou si l'on sait qu'elle a été consultée par un pair.

La recherche d'information collaborative c'est-à-dire « l'ensemble des approches qui permettent ou facilitent la collaboration dans le processus d'information » vise à exploiter cette dimension sociale de la recherche d'information [8, Denos]. Ces perspectives peuvent être très intéressantes dans un contexte de recherche et de collaboration scientifique.

On distingue deux types de recherche d'information collaborative :

- La recherche synchrone ;
- La recherche asynchrone.

Les situations de recherche synchrone étant potentiellement plus rares dans une organisation, nous consacrerons la suite de cette section aux recherches asynchrones telles que :

- Le filtrage collaboratif et le système de recommandation ;
- Le partage des traces ;
- Le reclassement collaboratif des résultats de recherche.

5.6.3.2.1 Filtrage collaboratif et système de recommandation

Le filtrage collaboratif fait partie des fonctionnalités de recherche d'information contextuelle les plus avancées et utilisées dans les SRI. Il consiste à rapprocher les utilisateurs sur la base des traces qu'ils ont laissées lors de l'usage des documents d'une base de données partagée. Le système recommande à chaque utilisateur les documents que d'autres utilisateurs ont utilisés [8, Denos].

Ce filtrage, qui fonctionne comme un principe de recommandation, peut s'appuyer sur des similarités de profils entre les utilisateurs pour prédire un score personnalisé pour chaque document et chaque utilisateur. Les profils utilisateurs eux-mêmes peuvent être définis en partie par les traces que ceux-ci ont laissées lors de leurs recherches d'informations (requêtes, documents consultés, documents sauvegardés, etc.).

Dans le cas de recherches scientifiques, il peut être très intéressant pour des utilisateurs-chercheurs de savoir quels sont les documents qui ont été consultés par d'autres utilisateurs-chercheurs travaillant sur des recherches similaires ou proches. D'une part, ils peuvent bénéficier de l'expertise des autres utilisateurs, d'autre part, ils économisent un temps de recherche considérable, ce qui est toujours bénéfique pour l'organisation qui les emploie, tout en découvrant des documents « pertinents » dont il n'aurait pas eu forcément connaissance. De plus, cette approche est vraiment personnalisée.

5.6.3.2.2 Partage des traces

Le partage des traces est une autre forme de recherche d'information collaborative qui consiste à montrer aux utilisateurs d'un SRI les traces que les autres utilisateurs ont laissées au cours de leurs interactions avec le système. Celui-ci peut alors se laisser guider, rester dans les traces d'un parcours de recherche déjà effectué, ou s'en inspirer pour créer le leur propre [8, Denos].

Les traces des utilisateurs attachées à un document peuvent être :

- Sa popularité (nombre de fois où elle a été consultée) ;
- Les annotations laissées par les autres utilisateurs lors de la consultation (tags, commentaires, notes associées) ;
- La requête qui a amené à trouver le document.

On voit bien ici l'intérêt du partage des traces en matière d'assistance intelligente à la recherche d'information dans le contexte scientifique d'un moteur de recherche d'entreprise :

- Une trace de popularité peut indiquer qu'un document est reconnu comme pertinent par les membres de l'organisation. Et même si cette pertinence dépend de nombreux facteurs subjectifs — contextuels —, elle est déjà une assurance quant à une certaine « valeur » scientifique du document.
- Les traces d'annotations sont un moyen d'identifier très rapidement le contenu d'un document, son domaine, ou mieux encore les associations scientifiques que l'on peut faire à partir d'un document (il est alors important de se poser la question des utilisateurs qui ont le droit d'annoter les documents du système).
- La trace de requête est une indication intéressante sur le « chemin cognitif » qui a amené tel chercheur à tel résultat. La trace de requête est une trace de raisonnement scientifique, une trace de pensée qu'il est toujours intéressant de partager.

5.6.3.2.3 Reclassement collaboratif des résultats de recherche

Le reclassement collaboratif des résultats d'une recherche consiste à améliorer le classement des réponses à une requête en exploitant le retour de pertinence des utilisateurs qui ont précédemment formulé la même requête. Ce retour de pertinence est ici considéré comme explicite, c'est-à-dire qu'il est demandé aux utilisateurs d'indiquer expressément si le document retourné est pertinent en regard de leurs requêtes [8, Denos]. L'intérêt est alors que l'utilisateur bénéficie des avis des utilisateurs émis précédemment.

Le reclassement collaboratif est une forme de partage d'expertise. Il permet de :

- Gagner du temps dans la sélection des documents en liste résultat (il importe là aussi de définir pertinemment les utilisateurs ayant le droit d'effectuer un retour de pertinence sur les documents).
- Bénéficier d'une recommandation qui n'occulte pas les documents qui n'ont pas bénéficié d'un retour de pertinence (cela permet de faire conserver une part de subjectivité propre à la

pertinence utilisateur). Le retour de pertinence est particulièrement adapté à une communauté scientifique qui partage des mêmes objectifs de recherche.

L'ensemble des fonctionnalités d'assistance à la recherche d'information que nous venons d'exposer sont intéressantes à envisager dans la perspective de recherches scientifiques dans une communauté particulière. Toutefois, il nous semble que leur utilisation et leur accumulation peuvent amener à rencontrer deux écueils :

- D'une part, il faut préserver les systèmes de recherche d'information, même ceux à l'intention d'experts d'un domaine, d'un trop grand nombre de fonctionnalités et d'une trop grande complexité quant à leur utilisation. Les utilisateurs pourraient se sentir dépossédés de la maîtrise du système, avoir l'impression de ne pas comprendre comment il opère. Le risque serait alors de les voir ne pas s'approprier, et donc ne pas utiliser l'outil.
- D'autre part, trop assister la recherche d'information, trop la guider en quelque sorte, c'est potentiellement empêcher l'utilisateur de trouver des documents qu'ils ne recherchaient pas, mais qui s'avèrent source de découverte et d'innovation. C'est empêcher une forme de sérendipité alors qu'elle a toujours joué un rôle important dans les processus de découverte scientifique. Et si l'on ne peut empêcher cette tension entre ces deux mouvements, l'assistance à la recherche et la sérendipité, il faut aussi donner à l'utilisateur les moyens de se départir des fonctionnalités qui cloisonnent sa recherche.

Conclusion

Dans la conclusion de ce mémoire, je voudrai aborder deux points : d'une part le cahier des charges en lui même, la façon dont il a été élaboré, perçu par l'équipe d'Ostéobio et les suites qu'il conviendra de lui donner dans la perspective de la réalisation du projet. D'autre part, les similarités que l'on peut trouver entre recherche d'information et recherche scientifique et comment la prise en compte de ces similarités peut aider à concevoir un projet de moteur de recherche dans un contexte scientifique.

Le projet

L'élaboration du cahier des charges

Le recueil des informations nécessaires à l'élaboration du cahier des charges s'est effectué à partir d'entretiens semi-directifs menés avec des membres de l'équipe projet ainsi que d'autres futurs utilisateurs du système d'information (principalement des étudiants des différentes promotions de l'école et certains enseignants). Ces entretiens ont permis de :

- Comprendre la logique du système de pensée promu au sein de l'école ;
- Comprendre les attentes relatives au futur système d'information dans le cadre de la recherche scientifique, notamment en regard de l'exigence d'exploitation du modèle de pensée promu au sein de l'école ;
- Comprendre comment le système d'information, en plus de permettre de rechercher et d'exploiter un certain nombre de ressources scientifiques, devait aider les étudiants à appréhender et utiliser le modèle Princeps.

Ces investigations se sont déroulées dans un contexte organisationnel peu évident, du fait que les membres de l'équipe projet, comme c'est souvent le cas dans ce genre de situation professionnelle n'avaient pas toujours la disponibilité que nécessite l'élaboration d'un cahier des charges, surtout si l'on tient compte des délais qui m'étaient impartis. De plus, une partie du stage s'est déroulé pendant les vacances scolaires, période ou n'étaient présents dans la société que des membres du personnel administratifs.

Parallèlement à ces entretiens, j'organisais des réunions afin de valider la définition des besoins relatifs au projet. Ces réunions, pour lesquelles je préparais en amont des documents de travail, étaient aussi l'occasion de communiquer à l'équipe projet sur la façon dont un moteur de recherche pourrait répondre à leurs attentes.

C'est en menant ce processus itératif, complété de mes lectures et découvertes dans les domaines des moteurs de recherche et des référentiels, soutenu en outre par mes co-directeurs de mémoire, que j'en suis arrivé à produire les documents qui ont été livrés à l'équipe projet.

Les retours de l'équipe projet

Le projet de moteur de recherche tel qu'il a été formalisé a rencontré l'approbation de l'équipe projet d'Ostéobio. Les documents finaux qui ont été livrés ont permis aux membres de l'équipe de « mieux comprendre[26] » leur projet, sa nature, sa portée, ainsi que les ressorts techniques et linguistiques sous-jacents.

La demande de l'équipe projet et ma volonté de a été de proposer un cahier des charges mettant en valeur toutes les options de recherche et d'exploitation de l'information correspondant aux attentes et besoins qui ont été identifiés. A partir de cette proposition, il conviendra à l'équipe projet de faire des choix pour éviter une solution surdimensionnée ou impliquant des modalités de mise en place inadaptées au contexte organisationnel (ressources financières et humaines).

C'est précisément l'un des intérêts du document « Éléments pour un Cahier des Charges » qui a été livré en fin de mission. Son caractère « modulaire » le rend durable et réexploitable selon les options qui seront privilégiées. Lors des prochaines phases d'avancement, les personnes concernées pourront finaliser le projet en décidant quelles options des trois phases envisagées devront y être intégrées. D'ores et déjà, il a été acté lors de la dernière réunion de travail que c'est l'hypothèse deux qui serait dans un premier temps envisagée, mais avec un thésaurus moins « basique » que celui qui est proposé dans le document (c'est-à-dire un thésaurus qui comporterait des relations typées entre les concepts du domaine).

Les suites du projet

Le choix de se concentrer sur la réponse aux besoins dans un projet « modulable » et qui ne soit pas dépendant de contraintes techniques et donc de la volatilité des solutions technologiques, tout cela dans un laps de temps limité et dans un contexte projet particulier tel que je l'ai décrit plus avant, font que ce travail nécessite qu'une deuxième phase soit menées à bien pour que ce travail soit complet. Les études complémentaires nécessaires à la complétion du projet sont :

[26] David Dessauge, Directeur Général de Ostéobio lors d'une réunion de travail.

- Une étude des limites des méthodes linguistiques dans l'indexation et la recherche des documents, et de la façon dont les techniques statistiques peuvent y répondre ;
- Une étude sur la conception des bases de données (champs, standards, langages d'expression, interopérabilité, etc.) et benchmark des prestataires ;
- Benchmark des solutions propriétaires et open-source existantes pouvant répondre aux besoins exprimés dans le projet ;
- Étude des moteurs de recherche existants et fonctionnant sur des principes similaires ou partiellement similaires (par exemple le moteur de recherche en sciences sociales « isidore »).

Ces études devront être menées par la prochaine personne qui prendra en charge le projet. Celle-ci devra aussi formaliser un Cahier des Charges en fonction des choix qui seront faits à partir du document « Éléments pour un Cahier des Charges ». Elle pourra éventuellement pour cela approfondir l'étude des usages qui seront faits du moteur de recherche.

Recherche d'information et recherche scientifique

Tout au long de ma mission de stage puis de la rédaction de ce mémoire, il m'est apparu de que en plus clairement que, même si les processus ne sont pas en tout points identiques, la recherche d'information et la recherche scientifique comportaient certaines caractéristiques communes. En effet, les deux démarches :

- Sont des processus itératifs ;
- Utilisent un langage en partie naturel ;
- Utilisent des modèles conceptuels (cadre de connaissance, ensemble de références), objectivés ou implicites ;
- Utilisent des relations sémantiques entre les concepts du modèle (hiérarchiques, d'équivalence, d'association) ;
- Procèdent des raisonnements logiques dans le cadre du modèle ;
- Utilisent des processus d'association d'idées, qui passent par le langage, dans ou en partie hors du domaine de connaissance considéré :
    - Dans le domaine si le processus associatif ne dépasse pas les limites conceptuelles du modèle ;
    - En partie hors du domaine si le processus associatif fait intervenir des concepts qui ne

sont pas pris par le modèle (innovation potentielle).

La similitude de ces deux démarches me fait penser qu'une solution de recherche d'information, si elle s'élabore dans un contexte de recherche scientifique, doit impérativement se penser dans l'optique d'un soutien à la recherche et à l'innovation scientifique.

Comme nous l'avons vu, une façon de soutenir conjointement la recherche d'information et la recherche scientifique est d'implémenter un référentiel terminologique dans le système afin de représenter et d'exploiter les connaissances des membres de la communauté qui va être amenée à l'utiliser. C'est une représentation des connaissances (RC).

L'implémentation de ce que l'on peut appeler un Système d'Organisation des Connaissances (SOC), c'est-à-dire un système organisant une représentation des connaissances relatives à un domaine (modèle entité-relation, modèle conceptuel, réseau sémantique, taxinomies associées à un système de classification, thésaurus, ontologie formelle ou sémiotique, folksonomie, etc.) [23, Mahé], va permettre au système de contribuer à remplir un certain nombre de tâches relatives à la gestion des connaissances:

- Conservation des connaissances de l'organisation (d'autant plus si celles-ci sont amenées à évoluer et si cette évolution est reportée dans le SOC) ;
- Transmission et appropriation des connaissances au sein de l'organisation ;
- Promotions des connaissances aux utilisateurs de la solution extérieurs à l'organisation ;
- Exploitation automatique des connaissances par le système au service de ses utilisateurs et, in fine, de l'organisation.

Cette dernière tâche est particulièrement intéressante en ce qu'elle suppose que le système soit capable d'une forme d'intelligence. Difficile en effet d'imaginer que l'on puisse exploiter de façon pertinente des connaissances sans une forme d'intelligence, ici entendu comme la capacité à interpréter, raisonner et associer. Mais il convient alors de s'interroger sur ce que signifie « représenter les connaissances » à partir desquelles le système devra faire preuve d'intelligence. Car c'est précisément de cette représentation que dépendra son intelligence. À ce sujet, et à partir d'une comparaison entre les modélisations des comportements humains en psychologie cognitive et les traitements automatiques des informations par les techniques de TALN, Maria Caterina Manes-Gallo distingue trois objectifs que l'on peut assigner à la représentation des connaissances [22, Manes-Gallo] :

- « Se représenter » la connaissance pour comprendre ;
- « Représenter » la connaissance pour imiter la compréhension ;
- Un niveau intermédiaire pour *avoir l'air de comprendre.*

« Se représenter » la connaissance pour comprendre

Nous nous situons là dans la modélisation cognitiviste. Il convient de distinguer trois formes de mémoires : la Mémoire à Long Terme (MLT), la Mémoire à Court Terme (MCT), et la Mémoire de Travail (MT).

Pour résumer : la MLT sert à stocker les données de connaissances acquises par le sujet au cours de son existence. Par exemple, les connaissances relatives aux savoirs et aux savoir-faire. La MCT et la MT permettent au sujet, à partir de l'activation de ses connaissances en MLT, de se construire des représentations circonstancielles (en contexte) des situations avec lesquelles il doit interagir. Par exemple, formuler des anticipations sur un raisonnement qui lui est énoncé pour envisager des réparties.

Dans ce cadre, la représentation des connaissances sert à indiquer deux phénomènes différents : d'un côté les connaissances, considérées comme des structures stabilisées et stockées en MLT, c'est-à-dire les savoirs de base acquis et éventuellement modifiés par l'expérience. D'un autre côté les représentations considérées comme des états provisoires de connaissance en MCT, résultants de l'activité d'attribution de signification à l'information en entrée. L'acquisition de nouvelles *connaissances* ou la modification de connaissances acquises en MLT se faisant par stabilisation des représentations les plus fréquentes en MCT.

Cette relation entre RC et activité mentale permet de différencier les caractéristiques spécifiques de la *cognition* humaine, par rapport à la fonction de la RC en TALN. Par exemple pour un humain, concernant le *traitement des significations* qui dépendent de la gestion des *connaissances* et des représentations, on distingue deux grands types de processus : les processus automatiques et les processus contrôlés. Les *processus automatiques* impliquent la mise en œuvre automatique et non-consciente des données de *connaissance.* Les *processus contrôlés* impliquent une mise en œuvre consciente des données de *connaissance* soit par échec des automatismes, soit par nouveauté des problèmes à affronter. Par exemple, la compréhension d'un énoncé ironique est considérée comme le résultat d'un processus contrôlé mis en œuvre par le sujet après l'échec du calcul automatique de sa signification littérale. Or les systèmes informatiques, même équipés de SOC représentant les

connaissances d'un domaine, ne peuvent à priori pas déterminer de façon autonome si leur interprétation d'un message est erronée. Et s'ils le pouvaient, il leur serait malgré tout difficile de savoir à quelle(s) autre(s) interprétation(s) ils devraient procéder. Les systèmes informatiques ne disposent que d'un seul type de processus de traitement des significations. Il est à la fois automatique et contrôlé. Il ne peut donc pas s'évaluer et se corriger lui même. Sauf à envisager, comme nous l'avons vu dans la dernière partie de ce mémoire, une recherche d'information contextuelle. Dans le cas d'un énoncé ironique, le retour de pertinence de l'utilisateur, explicite et implicite, serait pour le système une manière de « comprendre » qu'il a mal interprété sa requête (ici l'énoncé ironique, MCT). La connaissance qu'a l'individu de l'interlocuteur et de sa propension à utiliser l'ironie peut être comprise comme l'identification du profil utilisateur. La situation, perçus comme étant susceptible de susciter un énoncé ironique peut-être apparentée à l'identification d'un contexte de recherche notamment par la prise en compte de recherches similaires précédentes (ici les interactions que l'individu a déjà eues avec cet interlocuteur ou ce profil d'interlocuteur, MLT). Enfin, la prise en compte de la valorisation du » point de vue ironique » dans milieu social dans lequel se déroule l'interaction peut-être comprise comme la prise en compte du contexte organisationnel, de ses objectifs et de sa culture. On voit ici comment la recherche contextuelle peut contribuer à réduire la distance entre les processus cognitifs humains et les processus automatiques lors de la recherche d'information.

« Représenter » la connaissance pour imiter la compréhension

Mc Carthy et Hayes (1969) distinguent deux aspects conceptuellement distincts de la conception de programmes de résolution de problèmes : un aspect heuristique et un aspect épistémologique. L'*aspect heuristique* correspond à la définition d'algorithmes efficaces pour une recherche automatique de la solution. L'*aspect épistémologique* concerne la modalité à travers laquelle on représente les problèmes que le système doit résoudre.

Les difficultés soulevées par la formulation de programmes efficaces sont proportionnelles à la complexité de l'aspect épistémologique, complexité qui rejaillit en retour sur l'aspect heuristique. Par exemple, la RC nécessaire pour un programme qui joue aux échecs est beaucoup moins complexe que la représentation des *connaissance*s (linguistiques et extra-linguistiques) nécessaires pour la reconnaissance du sens implicite d'une phrase ou d'un texte. Le programme pour un jeu d'échecs, suppose la définition d'un ensemble d'états correspondants aux possibles configurations de l'échiquier et un ensemble d'opérateurs simples permettant de passer d'un état à un autre. Dans la communication humain/machine en langage naturel, un des principaux problèmes qui se pose est

celui de définir ce que signifie pour une machine de comprendre et produire un texte en langage naturel en effectuant des traitements en analyse pour passer de la séquence de surface au sens implicite véhiculé. L'efficacité d'un système dépend de sa capacité de résolution des ambiguïtés et de la reconstruction du sens implicite. Cette activité de résolution des ambiguïtés, les êtres humains les effectuent continuellement et « naturellement » lors de leurs interactions communicationnelles. Nous y reviendrons.

Un niveau intermédiaire pour *avoir l'air de comprendre*

La principale différence entre la perspective informatique et la perspective psychologique tient au fait qu'un système traite de l'information en manipulant des séquences de caractères qui n'ont aucun sens pour lui. En effet, pour les systèmes automatiques, l'*attribution d'une signification* à l'information en entrée dépend des données de *connaissance* qui ont été représentées dans la machine. Pour le sujet humain au contraire, comme nous l'avons vu précédemment, il y a acquisition de nouvelles *connaissances* en MLT par stabilisation des représentations qu'il a le plus fréquemment construites en MCT.

La RC est donc à la base de la capacité du système d'inférer et/ou de dériver des données informatives nouvelles. Dans le cas d'une interface en langage naturel, ces données coïncident avec la reconnaissance et la reconstruction du sens implicite des requêtes de l'utilisateur. Ce sens implicite qui renvoie d'un côté à l'interprétation de la signification véhiculée par les marques linguistiques de surface (le repérage de ce qui est dit), et de l'autre au but communicationnel qu'elles permettent de réaliser (le repérage des intentions de l'utilisateur).

La résolution des ambiguïtés du langage naturel et la restitution du sens implicite d'un énoncé vue dans les deux sections précédentes peuvent être en partie traitées par des fonctionnalités de recherche contextuelle. Mais les ambiguïtés sont principalement résolues, et le sens implicite des requêtes « révélé », à partir des « intentions » qui sont véhiculées par exemple dans une ontologie. Nous entendons par « intention » la culture au sens ethnologique du terme, c'est-à-dire pour paraphraser Howard.S Becker[27] : un ensemble de significations partagées attachées aux objets sociaux (les objets sociaux étant compris comme des concepts réels et/ou symboliques qui sont l'objet de discours et d'usages).

[27] BECKER Howard (1985), « Outsiders : études de sociologie de la déviance. Métailié .Paris. Leçon de choses. 1985 (1ère éd. 1963), 248 p. ISBN-10: 2864249189. ISBN-13: 978-2864249184.

Or c'est précisément ce que véhicule une ontologie à travers les concepts qui la composent, la définition de ces concepts, les propriétés qui leur sont attachées, les relations sémantiques qui les lient et la grammaire qu'elle embarque.

L'ontologie va donc permettre au système de prendre connaissance du sens implicite de la requête de l'utilisateur. Elle va par ailleurs participer au maintien et à la transmission des connaissances, soit à travers son utilisation dans le cadre du système de recherche d'information (suggestions de concepts associés, etc.), soit directement si elle est aussi élaborée dans la perspective d'une utilisation par les êtres humains.

Cette comparaison des utilisations de représentations des connaissances par les sujets humains et les systèmes informatiques nous paraît pertinente pour trois raisons :

- Elle permet de comprendre quelles sont les limites des systèmes informatiques dans leurs rôles d'assistance intelligente à la recherche d'informations ;
- Elle indique quelles sont les pistes de recherche envisageables pour améliorer lesdits systèmes.
- Elle est un moyen de rassurer les futurs utilisateurs du système quant à leurs compétences pour rechercher de l'information, surtout si ledit système est relativement complexe à utiliser (nombreuses fonctionnalités dont certaines peuvent être peu « intuitives »).

Enfin, elle a le mérite de confirmer l'hypothèse selon laquelle l'implémentation d'un référentiel terminologique adéquat et l'utilisation de fonctionnalités de recherche contextuelle sont de bons moyens pour doter les systèmes de recherche d'information de fonctionnalités d'assistance intelligente à la recherche d'information.

Bibliographie

La bibliographie analytique ci-dessous a été arrêtée au 15 novembre 2015.

Elle a pour objectif d'indiquer au lecteur les références qui ont été utilisées pour la rédaction de ce mémoire.

La rédaction des références bibliographiques respecte les normes ci-dessous, quand les informations étaient disponibles :

– Z44-005. décembre 1987. Documentation. Références bibliographiques : contenu, forme et structure et à la norme·

– NF ISO 690-2 février 1998 Information et documentation. Références bibliographiques Documents électroniques, documents complets et parties de documents·

Les références bibliographiques sont classées par thème. Elles sont précédées d'un chiffre entre crochets qui correspond à leur ordre d'apparition. Elles sont également suivies d'un texte en italique qui présente leurs contenus.

Le classement thématique est le suivant :

Moteurs de recherche
Recherche d'information : modèles et théories
Recherche d'information : fonctionnalités
Référentiels terminologiques
Techniques d'indexation automatique
Traitement automatique du langage naturel
Gestion des connaissances

Dans le corps de ce mémoire, certains sites Internet ou sources externes sont mentionnés en notes de bas de page. Les liens sont tous actifs à la date du 30 novembre 2015.

Moteurs de recherche

[1] HERIGAULT Myriam. Moteur de recherche d'entreprise : déploiement du moteur sémantique Exalead à la R&D de Diagnostica Stago. 2012, 90 p. Mémoire, Sciences de l'information, INTD — CNAM, 2012.

Ce mémoire décrit les modalités de mise en place d'un accès unique à l'information par le biais d'une application de recherche associée à des fonctionnalités sémantiques. Il en explore l'interface et les mécanismes linguistiques sous-jacents afin d'établir en quoi le moteur est sémantique.

[2] VUILLEQUEZ Jean-Yves. Le moteur de recherche d'entreprise : quels enjeux organisationnels et technologiques ? 2013. 125 p. Mémoire professionnel, Titre I, Ingénierie documentaire. INTD - CNAM, 2013.
Ce mémoire propose une réflexion sur les enjeux et les usages des moteurs de recherche d'entreprise et à montrer en quoi ils peuvent être considérés comme un projet managérial à travers leurs aspects organisationnel et technologique.

[3] DINET, Jérôme. La recherche d'information dans les environnements numériques. ISTE éditions Londres, Hermès Science, 2014, 134 pages. Systèmes d'information, web et informatique ubiquitare. ISBN 978-1-78405-018-4 (print), 978-1-78406-018-3 (ebook).
Cet ouvrage dresse un panorama subjectif des enjeux de la recherche d'information en organisation et des modèles théoriques qui ont été élaborés pour tenter d'en comprendre et modéliser les processus.

[4] POINCOT Philippe. Classification et recherche d'information bibliographique par l'utilisation des cartes auto-organisatrices, applications en astronomie. 1999. 144 p. Thèses, Université Louis pasteur, Observatoire Astronomique de Strasbourg, 1999.
Cette thèse s'interroge sur les possibilités d'appliquer les techniques de recherche d'information par cartes bibliographiques auto-organisatrices dans le domaine de l'astronomie.

[5] SAVOY Jacques. Modèles en recherche d'information. In GAUSSIER Eric, Stéfanini Marie-Hélène. Assistance intelligente à la RI. Lavoisier. Paris, Hermes, 2003. P 31 – 59. ISBN 2-7462-0726-5.
Ce chapitre présente les modèles booléens, vectoriels et probabilistes en recherche d'information ainsi que leurs variantes et évolutions.

[6] SIMMONOT Brigitte. La pertinence en sciences de l'information : des modèles, une théorie ? dans PAPY F. Problématiques émergentes dans les sciences de l'information, p161-182, Hermès-Lavoisier, Paris, 2008.
A travers certaines des modélisations qui en ont été faites, ce chapitre interroge le concept de pertinence en recherche d'information, notion centrale pour comprendre la façon dont les individus interagissent avec les systèmes de recherche d'informations et donc concevoir ces derniers.

Recherche d'information : fonctionnalités

[7] *NIE Jian-Yun,* Université de Montréal « Clustering et classification des documents » In Université de Montréal. Site de l'Université de Montréal, [En ligne]. http://www.iro.umontreal.ca/~nie/IFT6255/Clustering.pdf [consulté le 23/10/2015].
Cette ressource, dont l'auteur n'a pu être identifiée, explique quels sont les avantages de l'utilisation de Clusters en recherche d'information ainsi que les problématiques soulevées par leur élaboration.

[8] DENOS Nathalie. Recherche d'information collaborative. In Recherche d'information contextuelle, assistée et personnalisée. Lavoisier, Paris, Hermès-Science, 2011, 302 pages. Recherche d'information et web. Page 153 – 184. ISBN 978-2-7462-2583-1 ISNN 1968-8008.
Ce chapitre définit les différents types de recherche d'information collaborative et explique en quoi ils peuvent être de précieux soutiens aux utilisateurs de SRI.

[9] GANDON Fabien L. Graphes RDF et leur Manipulation pour la Gestion de Connaissances. 2008, 298 pages. Mémoire d'Habilitation à Diriger les Recherches. Soutenue le mercredi 5 novembre 2008 INRIA Sophia Antipolis — Méditerranée. Université de Nice. Sophia Antipolis.
Cette thèse porte sur la représentation des connaissances en graphe et l'utilisation de ces représentations pour permettre de nouvelles formes d'inférences dans l'exploitation des documents et des informations qu'ilsl contiennent.

[10] MORICEAU Véronique. Les systèmes de questions-réponses. 2010, 77 pages. Master 2 de recherche, Université Paris Sud 11.
Ce mémoire présente les objectifs et les principes de fonctionnement des systèmes de recherche d'information dits « systèmes de questions-réponses ». Il comporte de nombreux schémas qui permettent d'en saisir rapidement les principaux éléments.

[11] MOTHE Josiane. Recherche d'information contextuelle, le cas des requêtes. In Recherche d'information contextuelle, assistée et personnalisée. Lavoisier, Paris, Hermès-Science, 2011, 302 pages. Recherche d'information et web. Page 27 – 70. ISBN 978-2-7462-2583-1 ISNN 1968-8008.

Ce chapitre présente comment un SRI peut opérer à un traitement contextuel des requêtes des utilisateurs afin de les assister dans leur recherche d'information.

Référentiels terminologiques

[12] BRUANDET Marie-France, CHEVALLET Jean-Pierre. Utilisation et construction des bases de connaissances pour la recherche d'information. In GAUSSIER Eric, Stéfanini Marie-Hélène. Assistance intelligente à la RI. Lavoisier. Paris, Hermès, 2003. P 99 – 132. ISBN 2-7462-0726-5.
Ce chapitre explique comment un référentiel terminologique peut servir la recherche d'information. Ils indiquent aussi quelles sont les grandes étapes de construction des référentiels.

[13] CHAUMIER Jacques. Les ontologies. Antécédents, aspects techniques et limites. Documentaliste-Sciences de l'Information, A.D.B.S, 2007/1 (Vol. 44). Mis en ligne non disponible, [consulté le 13 octobre 2015], page 81-83. http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-81.htm.DOI 10.3917/docsi.441.0081 DOI 10.3917/docsi.441.0081.
Dans cet article, l'auteur définit la notion d'ontologie et en donne les règles de construction, fonctionnement et d'expression.

[14] DESFRICHES Doria Orélie et ZACKLAD Manuel. Améliorer la recherche d'information à l'aide de thésaurus « ad hoc ». Expérimentations et réflexions méthodologiques, *Document numérique* 2010/2 (Vol. 13), p. 13-40. ISBN : 9782746232334.
Cet article présente un nouveau type d'outil d'indexation et de recherche d'information, les thésaurus ad hoc, leurs fondements théoriques, ce qu'ils impliquent et comment il est possible de les concevoir grâce aux outils *de traitement automatique du langage. Il contient en outre une description intéressante du modèle de Marchionini.*

[15] GANDON Fabien. Ontologies informatiques In https://interstices.info/. [En ligne]. Mis en ligne le 22/05/2006 [consulté le 23/10/2015]. https://interstices.info/jcms/c_17672/ontologies-informatiques.
Cet article explique et illustre les rôles que peuvent tenir les ontologies informatiques dans les systèmes de recherche d'information.

[16] KELLER Loraine. Encadrer la réingénierie d'un thésaurus : méthode, enjeux et impacts pour l'équipe d'un service de veille et documentation en entreprise. 2013, 149 pages. Mémoire, Sciences de l'information, INTD — CNAM, 2013.

Ce mémoire, à partir de la présentation dune mission de réingénierie de thésaurus documentaire spécialisé, examine les spécificités d'un thésaurus spécialisé et sa place dans l'entreprise. Il fournit des clefs méthodologiques pour la gestion de ce type de projets et en analyse l'impact en termes de montée en compétences pour les documentalistes impliquées et pour le management.

Traitement automatique du langage naturel

[17] CHAUDIRON Stéphane. Technologies linguistiques et modes de représentation de l'information textuelle. Documentaliste — Sciences de l'information 2007, ADBS, vol. 44, n ° 1, p. 30-39. Mise en ligne non connue, [consulté le 17 octobre 2015], http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2007-1-page-30.htm. DOI : 10.3917/docsi.441.0030.
Dans cet article, l'auteur dresse un panorama assez approfondi des techniques de traitement automatique du langage et de ses utilisations en représentation de l'information.

[18] DALBIN Sylvie. Le résumé automatique aujourd'hui. Documentaliste — Sciences de l'information, A.D.B.S 2012, vol. 49, n ° 3, p.14-15. ISSN : 0012-4508.
A travers la présentation d'un ouvrage consacré au résumé automatique, « résumé automatique des documents » de Juan-Manuel Torres-Moreno, l'auteur fait une présentation synthétique des différentes techniques de résumé automatique.

[19] GAUSSIER Eric, STEFANINI Marie-Hélène. Traitement automatique des langues et recherche d'informations. In GAUSSIER Eric, Stéfanini Marie-Hélène. Assistance intelligente à la RI. Lavoisier. Paris, Hermes, 2003. P 99 – 92. ISBN 2-7462-0726-5.
Dans cet article, les auteurs, après avoir expliqué comment elles sont mises en œuvre, expliquent en quoi les technologies de traitement automatique du langage naturel contribuent à l'indexation et donc à la recherche d'information.

[20] TROUVILLIEZ Benoît. Lemmatisation et Racinisation en Français : Flexion, Lemme et Racine d'un mot In http://blog.onyme.com [En ligne]. Mise en ligne le 13/07/2010 [consulté le 12/10/2015].
http://blog.onyme.com/lemmatisation-et-racinisation-en-francais-flexion-lemme-et-racine-dun-mot/
Cet article explique clairement les notions de lemmatisation et de racinisation ainsi que les techniques qui permettent d'en effectuer les opérations.

Techniques d'indexation

[21] CHARTRON Ghislaine, DALBIN Sylvie, MONTEIL Marie-Gaëlle, VERILLON Monique. Indexation manuelle et indexation automatique : Dépasser les oppositions. . Documentaliste - Sciences de l'information, juillet-octobre 1989, vol. 26, n° 4-5, p.181-187.
Après avoir défini les critères de qualité d'une « bonne » indexation, cet article compare les avantages et inconvénients respectifs des indexations manuelles et automatiques à travers l'évocation d'un cas pratique.

Gestion des connaissances

[22] MANES-GALLO Maria Caterina, PAGANELLI Céline, « La recherche d'information assistée par ordinateur : quelle représentation des connaissances ? », Les Enjeux de l'information et de la communication 1/2003 (Volume 2003), p. 36-51.
Cet article traite de la représentation des connaissances dans les systèmes de recherche d'information en comparant les modélisations issues de la psychologie cognitive et les des techniques de traitement automatique du langage naturel.

[23] MAHE Sylvain et al, « Gestion des connaissances et systèmes d'organisation de connaissances. Premier modèle et retours d'expérience industriels », Document numérique 2010/2 (Vol. 13), p. 57-73.
A travers l'exemple d'un cas d'expérience, cet article traite de la gestion des connaissances dans l'organisation notamment à travers l'utilisation d'un Système d'Organisation des Connaissances.

Annexes

Annexe 1 : Points de vigilance

Points de vigilance :

Lors de ma mission de stage ainsi que lors de la préparation de ce mémoire, j'ai pu identifier trois principaux points pour lesquels je serai particulièrement vigilant lors de mes prochaines expériences professionnelles. Le premier point concerne la définition du périmètre du projet, le deuxième est relatif à l'utilisation être donc la conception des documents qui jalonnent le déroulement d'un projet informationnel, le troisième point porte sur la façon dont les lectures peuvent nourrir les pratiques professionnelles.

La question du périmètre

Parmi les premières questions qu'il faut se poser lorsque l'on commence à travailler sur un projet informationnel est, je pense, la définition du périmètre du projet. La question du périmètre se pose sur plusieurs plans :

- Le plan des personnes qui seront concernées par l'élaboration et la mise en œuvre du projet : l'équipe projet.
- Le plan des services et personnes au sein de ces services qui seront amenés à utiliser : les utilisateurs du projet.
- Le plan des services et personnes au sein de ces services dont l'activité sera impactée par la mise en route du projet
- Le plan des ressources informationnelles qui sont concernées par le projet.

Cette définition doit se faire en étroite collaboration avec les personnes de l'organisation qui sont à l'origine du projet ainsi que celles qui en seront les signataires. Bien sûr le périmètre est amené à évoluer au cours du projet, mais il importe d'en bien définir rapidement les principaux contours.
Relativement à cette question, j'ai connu quelques écueils qui, je pense, ne m'ont pas aidé dans le déroulement du projet.

La constitution de l'équipe projet

Le premier périmètre sur lequel j'ai éprouvé quelques difficultés est celui de l'équipe projet. Il a été simple de définir les personnes qui faisaient évidemment partie de cette équipe à savoir :

- Le fondateur de la société et initiateur du projet ;
- Le futur (et actuel à ce jour) directeur de la société ;
- Les référents scientifiques ;
- Ma tutrice de stage ;
- Le responsable des bases de données ;
- Les personnes qui avaient déjà produit en interne une réflexion relative au projet.

Mais paradoxalement, il a été plus difficile pour moi de définir qui ne faisait pas partie de l'équipe projet. Lors de la phase initiale de recueil des attentes et de définition des besoins, j'ai été amené à m'entretenir avec des futurs utilisateurs du système d'information. Pour certains d'entre eux, ils avaient des responsabilités importantes dans l'organisation (Responsables de pôle de recherche, directeur des études, enseignants). Il était difficile pour moi de décider s'ils devaient ou non intégrer l'équipe projet. Outre que cela ne les aurait pas nécessairement intéressés, les y intégrer aurait été risqué de rendre l'équipe trop pléthorique et donc incapable de prendre rapidement des décisions. Sans compter la plus grande difficulté à synchroniser les emplois du temps respectifs pour l'organisation de réunions de travail. Mais ne pas les intégrer me posait le problème de ma légitimité à prendre cette décision. Cet atermoiement dans la définition du périmètre de l'équipe projet n'a pas eu de conséquences dommageables parce que les membres d'Ostéobio étaient plutôt bienveillants à mon égard et que le projet n'avait pas de dimensions stratégiques immédiates pour la société. Mais cela peut-être plus dommageable dans d'autres situations et être à l'origine de dysfonctionnements dans la menées à bien du projet.

La définition du périmètre du projet

L'autre périmètre dont la définition m'a causé quelques préjudices est celle du périmètre du projet.
Lors de mes premiers entretiens, notamment avec Sophie Longuet, enseignante-chercheuse, directrice de mémoire et tutrice de mon stage au sein d'Ostéobio, il m'a été dit que celui-ci portait principalement sur la réflexion autour d'un moteur de recherche et accessoirement sur l'amélioration de la communication entre les différents membres de l'école. Il y avait donc un projet principal, le

moteur de rechercher, et un projet secondaire, une Gestion Electronique des Documents (GED) collaborative.
Au début de mon stage, et pour plusieurs raisons, j'ai eu tendance à plus travailler sur le projet secondaire que le projet principal. Ces raisons étaient multiples :

- Diagnostic de difficultés de communication entre les élèves d'une part et l'administration et le corps enseignant d'autre part (Accès aux documents pédagogiques, accès aux emplois du temps et à leurs modifications, etc.)
- Intérêt pour ces problématiques de communication et de collaboration en organisation.
- Sentiment que ces problématiques seraient plus faciles à aborder que celles concernant le moteur de recherche.

Il m'est vite apparu que je ne pourrai pas mener à bien ces deux projets dans le temps qui m'était impartis et compte tenu de ma faible expérience dans ces deux domaines ainsi que ma découverte de l'entreprise. Lors d'un entretien informel avec Sophie Longuet où j'ai exprimé mes doutes quant à ma capacité à mener de front ces deux projets, il a été convenu qu'il fallait prioritairement que je me consacre à celui de moteur de recherche. Or cette information.

Cette expérience m'amène à penser qu'il ne faut pas perdre de vue les différents périmètres et leurs hiérarchies. Définir les priorités et s'y tenir doit faire partie des premières tâches dans la menée d'un projet informationnel, même si sa définition est encore incertaine (la définition de priorités, même temporaires, permettant précisément de pallier à la définition incertaine du périmètre). Il faut aussi ne pas trop s'écouter, c'est-à-dire de mettre ses inclinations de côté pour privilégier celles du demandeur.

Usages et fonction des documents

Tout au long d'un projet informationnel, il se produit des documents qui jalonnent son déroulement. Ces documents sont à destination de l'équipe projet, des futurs prestataires qui seront chargés de sa mise en œuvre ou des futurs utilisateurs de la solution. Ils visent à communiquer des informations ou préparer des réunions de travail. Ce sont ces deux dimensions que je voudrais ici évoquer.

Le document comme un outil de communication

La première fonction d'un document, c'est de communiquer. Communiquer sur la définition du projet, ses objectifs, son périmètre, son état d'avancement ou ses fonctionnalités. Dans ce cadre, on pourrait

dire que communiquer signifie « énoncer de façon intelligible pour les lecteurs ». Il importe donc de réfléchir avant tout à la qualité des lecteurs du futur document. Seront-ils les membres de l'équipe projet ? Les signataires du projet ? Les futurs prestataires ? Les futurs utilisateurs ? Ceux dont l'activité sera impactée par le projet ?

Plusieurs éléments constitutifs du document dépendent de la qualité de ses lecteurs :

- Le point de vue qui y est développé : on ne décrit pas un projet de la même façon selon que l'on s'adresse à ceux qui en sont les commanditaires ou les bénéficiaires.
- Les thèmes qui y sont abordés : toutes les facettes d'un projet n'intéressent pas également toutes les parties qui y sont prenantes.
- Le niveau de précision : un document à destination des prestataires va aborder des questions techniques de mise en œuvre qui n'intéressent pas nécessairement, par exemple, les référents d'un projet.
- Le registre lexical qui y est développé : Un document peut relever d'un registre spécialisé ou d'un registre commun selon qu'il est destiné à un public familier ou non du domaine des sciences de l'information.
- L'utilisation qui en est demandée. Ce document peut-être purement informatif ou demander à ses lecteurs un travail de réflexion et de retour à l'auteur quant à son contenu.

Le dernier point évoqué nous permet de nous intéresser à la dimension « travail » de certains des documents qui jalonnent la vie d'un projet.

Le document comme outil de travail

Outre qu'il permet de communiquer, le document est aussi un outil de travail. Il sert de base pour des présentations (de fonctionnalités possibles par exemple), des discussions et des arbitrages à effectuer. À ce titre, il se doit d'être précis dans les notions qu'il aborde, exhaustif quant aux fonctionnalités qu'il propose et clair quant aux conséquences des décisions qu'il envisage. Par principe, avant de proposer de choisir parmi des orientations ou de prendre des décisions de fonctionnalité, il est souhaitable qu'il récapitule les arbitrages préalablement effectués afin que ceux qui le consultent puissent les resituer dans une vision d'ensemble du projet.

En conclusion de cette section consacrée aux documents produits lors d'un projet, il m'apparaît opportun de conseiller une méthode : avant de rédiger un document, il est intéressant de composer une brève note introductive qui en précisera le contenu ainsi que ses objectifs. Cette note sera utile

aux futurs lecteurs dudit document bien sûr, mais elle le sera aussi à son rédacteur. Elle détermine un cadre préalable au document, à son contenu et sa structuration qui permet son l'auteur de toujours s'y référer pour évaluer s'il est pertinemment élaboré. Si au départ, les lecteurs et les objectifs du document sont bien définis, son élaboration n'en est que plus aisée et son efficacité s'en trouve renforcée.

Annexe 2 : Fiche de projet

Cette fiche de projet a été un élément important de la conduite du projet, permettant à son auteur et aux membres de l'équipe projet de s'y reporter pour se remémorer les aspects « essentiels » du projet.

FICHE DE PROJET

TITRE DU PROJET :

OSTEOSEARCH : projet de moteur de recherche pour la valorisation transdisciplinaire et intelligente des ressources scientifiques d'OSTEOBIO.

CONTEXTE DU PROJET :

Contexte organisationnel :

La société OSTEOBIO a été créée en 1988 par Camille GOSSARD
Elle regroupe trois entités :

- Une école privée d'ostéopathie biomécanique (environs 350 étudiants);
- Des pôles de recherche en ostéopathie répartis par spécialités (6 pôles, 1 responsable/pôle);
- Des activités cliniques au sein et hors de la structure (1 clinique intra-muros, 2 cliniques extra-muros).

Elle a établi un partenariat privilégié avec COGITOBIO, société également fondée par Camille GOSSARD, dirigée par Pol LEBORGNE, et dont l'objet est le transfert de technologies issues de la recherche fondamentale sur le système musculo-squelettique vers le monde industriel.

Contexte scientifique :

Au sein d'OSTEOBIO, Camille GOSSARD a développé une approche originale des troubles musculo-squelettiques (TMS) et du fonctionnement bio-mécanique du corps en général. Ce modèle a été lauréat du prix ANVAR en 2000. C'est le modèle princeps.

Cette démarche est transdisciplinaire, c'est-à-dire qu'elle cherche à dépasser les cloisonnements entre les disciplines de l'ostéopathie en développant une approche qui repose sur :

- la définition de quatre niveaux d'observation (tissulaire, interface, segmentaire et pluri-segmentaire) ;
- l'étude de la géométrie de ces quatre niveaux (formes, agencements et orientations spatiales) ;
- l'étude de l'interdépendance des géométries de ces quatre niveaux ;
- l'étude des forces qui s'exercent sur les paramètres géométriques, entraînant des contraintes sur les structures, et donc des déformations et des mouvements.

Les disciplines « traditionnelles » sont prises dans le modèle et appartiennent à un ou plusieurs niveaux.

Le modèle est de fait interdisciplinaire.

Le modèle n'est pas pour le moment décrit de manière formelle dans un document. Ce travail est en cours d'élaboration.

Contexte informationnel :

L'essentiel de la production scientifique d'OSTEOBIO est constitué des mémoires des étudiants de cinquième année dont les sujets sont déterminés en fonction d'objectifs de recherche. Ils sont composés des textes des mémoires, des articles bibliographiques, des fichiers de formule, d'images scientifiques (scanners, IRM, schémas), d'un fichier de présentation et d'un poster.

Les mémoires sous forme numérique sont actuellement stockés dans des répertoires de dossiers sur deux postes informatiques. Ils sont classés par années ou par ordre

alphabétique de leurs auteurs. Ils ne sont accessibles directement que par les personnes détenant ces postes informatiques. Les mémoires sous forme papier sont stockés dans des armoires et classés par années.

L'essentiel des ressources informationnelles des acteurs d'OSTEOBIO est constitué d'articles issus de revues spécialisées accessibles en ligne ou sur support papier, à travers des portails d'information scientifique (PubMed, Sciences Directe, BIUM) ou des abonnements papiers à des revues (« Gate and Posture », « The Spine Journal », « Journal of Bioméchanics »).

FINALITÉ DU PROJET :

Rendre accessibles les documents scientifiques produits au sein d'OSTEOBIO.
Mettre en place un moteur de recherche qui permettra de rechercher et d'exploiter les documents scientifiques produits ou importés au sein d'OSTEOBIO selon les logiques du modèle princeps.

Permettre une veille scientifique du point de vue du modèle dans les domaines considérés.

IMPACTS ATTENDUS :

- Systématiser l'appréhension du fonctionnement bio-mécanique du corps sous le prisme du modèle princeps ;
- Développer, dans les secteurs industriels et cliniques, tout le potentiel applicatif de cette approche ;
- Définir de nouveaux axes de recherche ;
- Eviter une perte de valeur scientifique et donc économique dues à une méconnaissance des activités de recherche déjà produites sur certains sujets.
- Prendre connaissance des innovations scientifique dans les domaines considérés.

PRINCIPALES EXIGENCES LIÉES AUX FINALITÉS ET AUX IMPACTS ATTENDUS :

- Systématiser l'approche transdisciplinaire dans la recherche et l'exploitation des documents ;
- Faire émerger des documents des connaissances qui n'y sont pas exprimées de manière formelle ;
- Permettre de savoir ce qui a déjà été produit sur un thème de recherche ;

- Faciliter l'accès des acteurs scientifiques d'osteobio aux documents scientifiques.

PUBLICS VISÉS PAR LE PROJET :

- Les chercheurs d'osteobio ;
- Les enseignants d'osteobio ;
- Les étudiants d'osteobio ;
- Le personnel de COGITOBIO ;
- Les partenaires de recherche ;
- Les institutions qui voudront « moissonner » les bases de données d'osteobio ;
- À terme les internautes qui voudront consulter les bases de données d'osteobio depuis le site d'osteobio.

ORGANISATION DU PROJET :

- **Commanditaires du projet** : Camille GOSSARD, David DESSAUGE.
- **Signataire du projet :** David DESSAUGE.
- **Chef de projet AMOA :** Mikhaïl FRONTERE.
- **Encadrement :** Sophie LONGUET.
- **Référents scientifiques :** Philippe MAS, Pol LEBORGNE, Amélie BEAU.

Annexe 3 : Note de cadrage

Cette note de cadrage a été établie à l'attention de l'équipe projet et éventuellement des futurs prestataires. Il y manque les informations concernant les ressources informationnelles et le périmètre technique du projet.

NOTE DE CADRAGE

OBJECTIFS DU PROJET :

Rendre accessibles les documents scientifiques produits au sein d'OSTEOBIO afin de :

- Permettre de savoir ce qui a déjà été produit sur un thème de recherche ;
- Faciliter l'accès des acteurs scientifiques de la semev (chercheurs, enseignants et étudiants aux documents scientifiques) ;

Mettre en place un moteur de recherche qui permettra de rechercher et d'exploiter les documents scientifiques produits ou importés au sein d'OSTEOBIO selon les logiques du modèle princeps afin de :

- Systématiser l'approche transdisciplinaire dans la recherche et l'exploitation des documents ce qui permettra de ;
- Faire émerger des documents des relations de causalité scientifique qui n'y sont pas exprimées de manière formelle ;
- Révéler le caractère interdisciplinaire d'un document ;
- Faire le lien entre la pratique clinique et la recherche ;

Permettre une veille scientifique du point de vue du modèle dans les domaines considérés.

PUBLICS PRIORITAIRES :

- Les enseignants et les chercheurs d'Ostéobio ;
- Les étudiants de quatrième et cinquième année.

OBJECTIFS DU MOTEUR DE RECHERCHE ET CRITÈRES DE RÉUSSITE :

Ce que doit permettre la solution	Utilisateurs types	Critères de réussite en regard des usages attendus :
Accéder aux mémoires.	Etudiants Enseignants Chercheurs	Chercher et trouver les mémoires selon des critères de recherche scientifiques, bibliographiques, typologique et de format.
Accéder aux documents composites d'une même unité documentaire.	Etudiants Enseignants Chercheurs	Lier à un résultat de recherche les documents composites de la même unité documentaire (les fichiers de formule avec les mémoires par exemple).
Accéder aux articles de certains fournisseurs de contenus	Etudiants Enseignants Chercheurs	Interroger les bases de données des fournisseurs de contenus et consulter les contenus.
Savoir ce qui a déjà été produit sur un thème de recherche.	Etudiants Enseignants Chercheurs	Trouver les documents existants relatifs à un thème de recherche.
Utiliser le modèle princeps pour rechercher des informations scientifiques.	Etudiants Enseignants Chercheurs	Rechercher et trouver des documents à partir de critères notamment issus du modèle princeps.
Comprendre comment un thème de recherche s'intègre au modèle princeps (niveau thème de recherche).	Etudiants	Mettre le thème de recherche en relation avec des concepts issus du modèle princeps. Suggérer la consultation de documents en rapport avec les concepts issus du modèle princeps.
Elargir une thématique de recherche (niveau thème de recherche).	Etudiants Enseignants Chercheurs	Suggérer d'autres échelles de recherche pour un même thème (Note au prestataire : voir la présentation du modèle princeps dans la fiche de projet). Que les résultats d'une recherche en fassent apparaître les dimensions transdisciplinaires à travers l'utilisation du modèle princeps pour suggérer d'autres liens conceptuels et d'autres documents.

Exploiter les résultats d'une recherche selon le modèle princeps afin d'en extraire des connaissances (niveau document résultat).	Etudiants Enseignants Chercheurs	Extraire des documents les principaux concepts et mesures scientifiques qui y sont contenus. Mettre ces concepts en relation avec des concepts issus du modèle princeps. Suggérer la consultation de documents en rapport avec les concepts issus du modèle princeps.

PÉRIMÈTRE FONCTIONNEL DU PROJET PERMETTANT DE RÉPONDRE AUX ATTENTES ET OBJECTIFS/

Concernant l'indexation des documents :

- Exploitation d'un référentiel de vocabulaire ;
- Indexation des documents selon les critères du modèle princeps ainsi que d'autres critères scientifiques qui restent à déterminer ;
- Indexation des documents selon des critères bibliographiques, typologiques et techniques (format).

Concernant la recherche des documents :

- Recherche simple avec suggestion de termes de recherche ;
- Recherche des documents par critères de recherche (au début de la recherche) et facettes de recherche (après une première liste de résultats de recherche) ;
- Recherche des documents selon les critères du modèle princeps ainsi que d'autres critères scientifiques qui restent à déterminer ;
- Recherche des documents selon des critères bibliographiques, typologiques et techniques ;
- Possibilité de localiser un terme de recherche dans une représentation schématique du modèle princeps.
- Possibilité d'effectuer une recherche à partir d'une représentation schématique du modèle princeps.
- Recherche de documents dont le contenu et les métadonnées sont exprimées en anglais.
- Possibilités d'accéder aux documents pédagogiques par l'intermédiaire d'un plan de classement reprenant l'organisation des cours tels qu'ils sont délivrés.

Concernant l'exploitation des documents :

- Exploitation des termes de recherche selon le modèle princeps (voir usages attendus et critères de réussite) ;

- Exploitation des documents selon le modèle princeps (voir usages attendus et critères de réussite) ;
- Prise en compte de la synonymie, de la polysémie et des acronymes dans la recherche et l'exploitation des documents ;
- Exploitation de documents en langue anglaise ou dont les métadonnées sont exprimées en anglais ;
- Extraction et exploitation selon le modèle princeps des mots clefs d'un document résultat.
- Possibilité de localiser un document dans une représentation schématique du modèle princeps.

Concernant la relation avec d'autres bases de données :

- Interopérabilité des bases de données d'Oétéobio (qu'elles soient interrogeables par d'autres moteurs de recherche) ;
- Possibilité d'interroger d'autres bases de données (compatibilité de la solution avec des protocoles d'échange dont OAI-PMH).

Concernant l'interrogation d'autres bases de données :

- Il est prioritaire de permettre l'interrogation de bases de données extérieures à Ostéobio.
- Les documents provenant de ces autres bases de données doivent pouvoir être recherchés selon les critères du modèle princeps (cela est faisable en pensant qu'un terme issu du modèle, par exemple une échelle, croisé avec d'autres termes non spécifiquement issus du modèle, renverra à des termes contenus dans les articles provenant des autres bases de données).
- Les documents provenant de ces autres bases de données doivent pouvoir être exploités selon les critères du modèle princeps (extraction de concepts du document et mise en relation avec des concepts issus du modèle).

Concernant l'accessibilité :

- Possibilité d'accès distant au moteur de recherche (postes extérieurs à Ostéobio) ;
- Possibilité de permettre l'accès temporaire à des partenaires de recherche ;
- La solution doit-être « responsive design "(s'adapter au support de consultation, tablette, téléphone, etc.).

Autres exigences :

Exigence de réversibilité (que les formats des documents puissent être pris en charge par d'autres systèmes d'information, que les métadonnées puissent être extraites du système en même temps que les documents et prises en charge par d'autres systèmes d'information. Important en cas de changement de système) ;

HORS PÉRIMÈTRE FONCTIONNEL :

Concernant l'accessibilité :

La solution ne doit pas, dans un premier temps, être accessible depuis le web.

Concernant l'interface :

- La solution ne doit pas avoir de fonctionnalité de type 'portail d'information' (permettre d'accéder aux applications pédagogiques de l'école [Moodle, Doodle, Hyperplanning, emplois du temps des cliniques]

PROCESSUS MÉTIERS IMPACTÉS :

- **Les étudiants :**
- Appropriation du modèle princeps ;
- Structuration systématique des mémoires selon les normes définies [structure globale du document].
- Standardisation de la forme des contenus des mémoires selon les normes définies [forme des tables des matières, des résumés, typographie, etc.].

Les directeurs et tuteurs de mémoire :

- Surveillance du respect des normes de structuration des mémoires par les étudiants.

Les enseignants :

- Fourniture des cours afin qu'ils soient indexés ;
- Si possible, structuration des cours selon des normes à définir ;
- Si possible, standardisation de la forme et de la présence de certains contenus [bibliographie, sommaires etc.]
- Si possible, enregistrement des cours sous des formats communs.

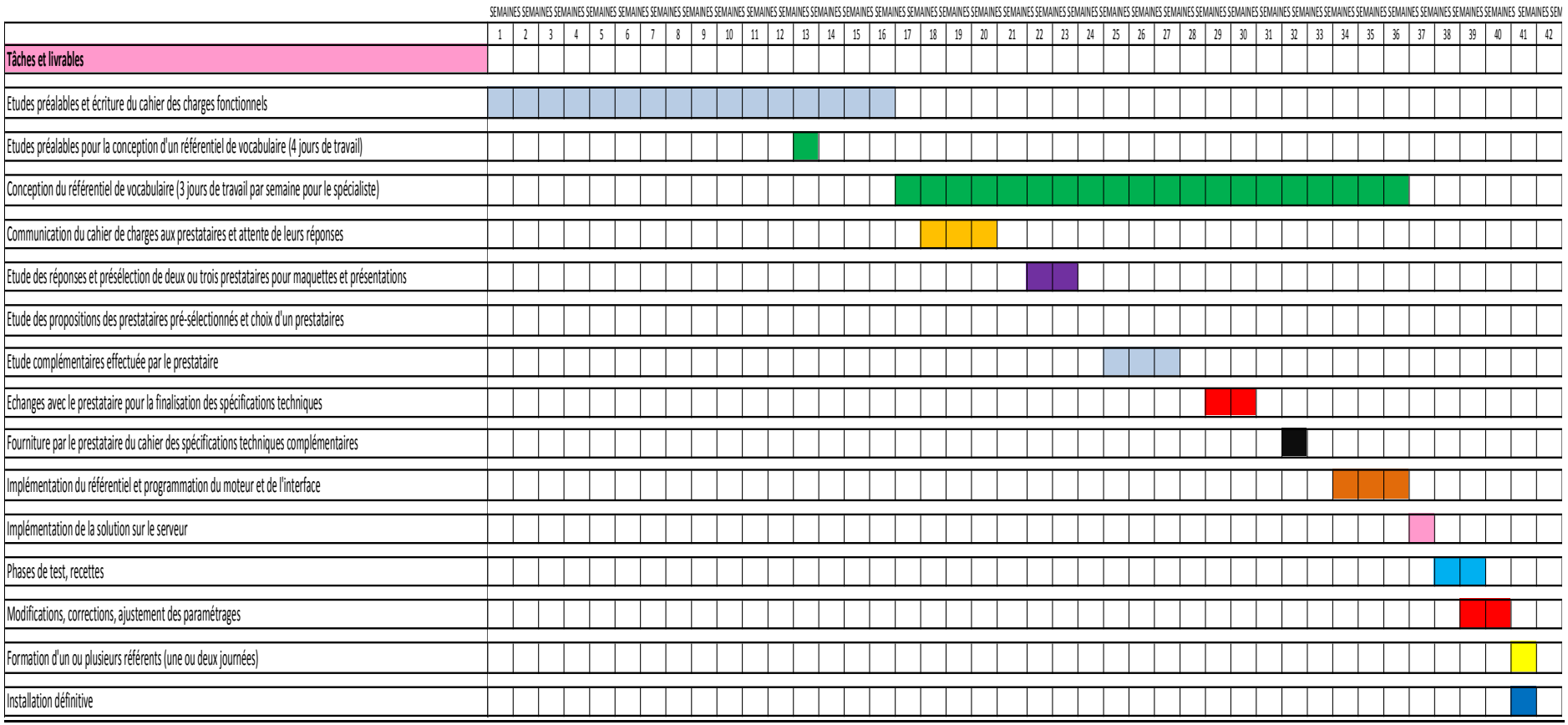
Les enseignants et les chercheurs concernés :

- Maîtrise des procédures techniques d'indexation ;
- Indexation de l'existant ;
- Indexation 'au fil de l'eau.

La société :

- Abonnement aux fournisseurs de contenu

PLANNING PREVISIONNEL DU PROJET :



PÉRIMÈTRE INFORMATIONNEL DU PROJET :

Ressources internes à Ostéobio :

Documents	**Langue**	**Production annuelle**	**Existant à importer**
Unités documentaires « mémoires »		30	300
• Textes mémoires	Français	30	300
• Articles des mémoires	Anglais	50/mémoire	50/mémoire
• Fichiers de formule	Français	10/mémoire	10/mémoire
• Images scientifiques	Français	3/mémoire	3/mémoire
• Présentations	Français	1/mémoire	1/mémoire
• Posters	Français	1/mémoire	1/mémoire
Unités documentaires « documents pédagogiques » :			
• Les cours magistraux	Français	À déterminer	À déterminer
• Les TD	Français	À déterminer	À déterminer
• Les schémas	Français	À déterminer	À déterminer
• Les animations	Français	À déterminer	À déterminer
• Les images scientifiques	Français	À déterminer	À déterminer
• Les bonnes pratiques	Français	À déterminer	À déterminer
• Vidéos (cours et manip)	Français	À déterminer	À déterminer
Les articles produits par les chercheurs.	Français	À déterminer	À déterminer
Dossiers patients (à anonymiser)	Français	À déterminer	À déterminer
Documents portalis.	Français	À déterminer	À déterminer
Certains programmes d'OstéoTV.	Français	À déterminer	À déterminer

Ressources externes à Ostéobio :

Sources	Intérêt	Langue	Accès Bases gratuites dps moteur recherche	Télécharger articles gratuites	Diffuser articles gratuits dans système fermé	Diffuser articles gratuits dans système semi-ouvert	Diffuser articles gratuits dans système ouvert	Télécharger à articles payants	Diffuser articles payants dans système fermé	Diffuser articles payants dans système semi-ouvert	Diffuser articles payants dans système ouvert
Pub Med	Oui	Anglais	Oui	Non sauf licence cc	Non sauf licence cc (minorité des articles)	Non sauf licence cc (minorité des articles)	Non sauf licence cc (minorité des articles)	Voir conditions abonnements	Voir conditions abonnements	Voir conditions abonnements	Voir conditions abonnements
Science Directe	Oui	Anglais	Oui	Non sauf licence cc	Non sauf licence cc	Non sauf licence cc	Non sauf licence cc	Voir conditions abonnements	Voir conditions abonnements	Voir conditions abonnements	Voir conditions abonnements
Thèses	Oui	Français	???	???	???	???	???	???	???	???	???
Chiropractis & Manual Thérapies	Oui	Anglais	???	???	???	???	???	???	???	???	???
BioMedCentral	???	Anglais	???	???	???	???	???	???	???	???	???
DOAJ	???	Anglais	???	???	???	???	???	???	???	???	???
Chochrane	???	???	???	???	???	???	???	???	???	???	???
H.A.L	???	Fr/Ang	???	???	???	???	???	???	???	???	???
OCLC	???	Anglais	???	???	???	???	???	???	???	???	???

Système fermé : accès restreints à Ostéobio

Système semi-ouvert : accès Ostéobio + partenaires

Système ouvert : accès ouvert aux internautes

Licence CC : Licence « Créatives Communs » permettant, selon la licence, d'accéder et de diffuser un article scientifique

PÉRIMÈTRE TECHNIQUE :

Informations non obtenues auprès du prestataire de la société demandeuse.

Annexe 4 : Planning du projet de la mission de stage

Ce planning est un planning prévisionnel qui a été élaboré en juillet 2015.

VP1… VPn : Version soumise à l'encadrement pédagogique
Vint1… Vintn : versions soumises à l'équipe projet/VF : version finale