CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département CITS

INTD

MEMOIRE pour obtenir le

Titre professionnel "Chef de projet en ingénierie documentaire" INTD

niveau I

Présenté et soutenu par

Nathalie DADOU

le 10 novembre 2011

Indexation pour le Web : usages et application au fonds documentaire des Editions Techniques de l'Ingénieur

Jury

Mme Muriel AMAR
Mme Maud BUISINE

Promotion XLI

Remerciements

J’adresse mes remerciements à l’ensemble des personnes rencontrées aux Editions Techniques de l’Ingénieur pour leur accueil chaleureux et plus particulièrement à Maud Buisine pour sa gentillesse et sa disponibilité pour me fournir informations et explications tout au long de mon stage, ainsi qu’à Marie Lesavre pour m’avoir également fait bénéficier de son expérience.

Un grand merci à Muriel Amar pour avoir accepté de suivre mon travail ainsi que pour ses conseils et apports à la réflexion et à l’organisation de ce mémoire.

Je remercie également toute l’équipe de l’INTD, les professionnels qui nous ont transmis leur savoir et leur passion tout au long de l’année ainsi que les élèves de cette promotion (et plus particulièrement le groupe 3) pour leur soutien, le partage des connaissances et tous les moments passés ensemble.

Notice

DADOU Nathalie. Indexation pour le Web : usages et application au fonds documentaire des Editions Techniques de l'Ingénieur. 2011. 105 p. Mémoire de Titre professionnel niveau I « Chef de projet en ingénierie documentaire », CNAM-INTD, 2011.

Les pratiques d'indexation de documents évoluent avec le passage au numérique et l'amélioration des techniques de recherche. Ce mémoire propose une étude des possibilités de mise en valeur d'articles scientifiques et techniques et d'apport de compléments d'informations à travers la création et l'exploitation de différents types de mots-clés. La réflexion menée à partir de l'analyse des besoins des Editions Techniques de l'Ingénieur a permis de déterminer une typologie de mots-clés applicables sur leur site Web et d'expliciter leurs caractéristiques ainsi que les contraintes de mise en œuvre.

Indexation ; Mot-clé ; Internet ; Site Web ; Recherche d'information ; Moteur de recherche ; Information scientifique et technique ; Navigation ; Catégorisation ; Diffusion de l'information ; Valorisation

Table des matières

Liste des tableaux

Liste des figures

Introduction

Les Editions Techniques de l'Ingénieur proposent une documentation scientifique et technique en français, destinée principalement aux ingénieurs et cadres techniques des bureaux d'études et de l'industrie ainsi qu'aux étudiants et enseignants de l'enseignement supérieur technique et scientifique.

La collection est volumineuse et intégralement disponible sur le Web depuis 2001. Elle est organisée selon la même structure depuis sa création en 1946, conçue pour la diffusion de contenu sur support papier et donc moins adaptée au site Web.

Une réflexion a donc été lancée au niveau de la direction pour réorganiser l'information et s'adapter aux nouvelles demandes de la clientèle.

Ce mémoire s'inscrit dans cette perspective d'évolution et la réflexion menée porte ici sur les possibilités d'utilisation de mots-clés pour améliorer la perception du site Web et l'accès au contenu.

Afin de pouvoir proposer différentes possibilités de création et d'exploitation de mots-clés dans le cadre du site Web des Techniques de l'Ingénieur, j'ai tout d'abord été amenée à étudier, dans la première partie, les différents types d'indexation et de langages documentaires existants. Je me suis également intéressée à la manière dont sont utilisés les mots-clés par les internautes et à leur exploitation par les moteurs de recherche dans le cadre de la recherche d'informations.

Dans la deuxième partie, l'exploitation actuelle du fonds documentaire et l'analyse des besoins sont détaillées ainsi qu'une étude des différents mots-clés déjà attribués ou utilisés sur les moteurs de recherche pour l'accès au contenu du site. A partir de toutes ces informations et données internes et externes, j'ai pu dresser une typologie de mots-clés applicables aux Editions Techniques de l'Ingénieur.

Enfin, en troisième partie, des préconisations sont données pour la création et l'implantation de ces différents types de mots-clés sur le site Web.

Première partie
Terminologie et problématique

1 Types d'indexation et langages documentaires

1.1 Concepts clés relatifs à l'indexation

Nous allons, dans un premier temps, étudier comment définir l'indexation et observer les différentes formes qu'elle peut prendre.

1.1.1 Définitions et objectifs de l'indexation

Selon la définition de la norme AFNOR NF Z 47-102 (1978),

> « *L'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse.* »

Cette définition est très stricte car elle implique que les termes soient extraits du document puis que l'on utilise un langage documentaire défini auparavant pour les valider ou trouver un concept équivalent.

Toujours selon la norme AFNOR NF Z 47-102,

> *« La finalité de l'indexation est de permettre une recherche efficace des informations contenues dans un fonds de documents et d'indiquer rapidement, sous forme concise, la teneur d'un document.* »

Pour Jacques Maniez,

> *« L'objectif premier de l'indexeur est de créer l'outil de recherche privilégié, l'idéal étant qu'il permette de fournir à la demande toutes les adresses (exhaustivité) et seulement les adresses (précision) répondant aux besoins des usagers.* » ([7], Maniez, p.147]

Enfin selon Suzanne Waller,

> « *Une définition dynamique de l'indexation pourrait être de dire que lorsque l'on indexe, on cherche dans un texte des réponses à des questions susceptibles d'être posées.* » ([15], Waller, p.150]

Dans toutes ces définitions, on note que l'indexation doit être faite dans le but de retrouver des documents et doit pouvoir servir à l'utilisateur qui fait sa recherche. L'écueil serait de

faire une indexation minutieuse du contenu sans se préoccuper du type de personnes qui utiliseront cette indexation et du type de recherches qu'elles effectueront.

Le résultat de l'indexation est un ensemble de mots ou expressions que l'on peut nommer selon les cas :

- Mot-clé : Selon la définition de la norme AFNOR NF Z 47-102, le mot-clé est un « *mot ou groupe de mots choisi soit dans le titre ou le texte d'un document, soit dans une recherche documentaire, pour en caractériser le contenu.* »

Issu du langage naturel, le mot-clé est libre et souvent extrait du document analysé. Dans ce mémoire, j'emploierai cependant le terme mot-clé pour qualifier tous les mots ou groupes de mots caractérisant un document, à partir du moment où ils sont choisis librement. Ils peuvent donc être contenus dans le document ou en être absents et apporter un complément d'information.

- Descripteur : c'est un mot ou une expression qui appartient à un vocabulaire contrôlé.
- Tag : c'est une forme totalement libre d'indexation qui mélange des mots ou groupes de mots de toute nature, voir d'autres formes d'expressions que des mots. L'exploitation en est donc délicate si aucune règle n'est donnée.

L'utilité de l'indexation est reconnue par tous mais peut, selon les cas, représenter un investissement très important, essentiellement pour le temps passé au traitement de l'information. Il faut donc pouvoir justifier son utilisation par des améliorations significatives par rapport au but recherché. Elle demande de comprendre le domaine, d'avoir un esprit d'analyse et une certaine pratique. Le résultat a toujours une composante variable en fonction de la personne qui indexe le document.

1.1.2 Langage libre vs langage contrôlé

Afin d'étudier ensuite les possibilités d'indexation, il est important de comparer indexation libre et indexation utilisant un langage contrôlé.

L'indexation libre est beaucoup moins contraignante mais elle ne prend pas en compte les problèmes de synonymie (mots différents ayant le même sens ou des sens très proches) et de polysémie (mot ayant plusieurs sens).

Le vocabulaire contrôlé permet une représentation univoque du contenu mais il nécessite un travail important de création ou d'adaptation pour établir les listes de termes et leurs relations afin que ceux-ci correspondent à l'environnement dans lequel ils sont utilisés. Lors de l'indexation, il faut ensuite trouver dans la liste le terme correspondant à la notion

souhaitée. Le risque est de choisir un terme avec un sens légèrement différent ou d'utiliser des termes qui ne correspondent pas au langage des utilisateurs.

Le choix des termes retenus dans la liste et qu'il faudra ensuite utiliser est arbitraire et ne correspond pas toujours à toutes les utilisations ou tous les utilisateurs même s'ils relèvent d'un seul domaine.

Pour tirer tout le bénéfice souhaité de l'utilisation de vocabulaire contrôlé, il faut que celui-ci soit suffisamment riche, mis à jour régulièrement, d'accès rapide, avec des règles d'utilisation simples.

L'utilisation des listes de vocabulaire contrôlé lors de la recherche est également une contrainte pour l'internaute qui veut souvent aller le plus vite possible. Il est également possible d'utiliser le vocabulaire contrôlé en assistance automatique de recherche mais là encore, le risque est de biaiser la demande en associant des termes qui n'ont pas strictement le même sens.

Les études cherchant à comparer l'efficacité de différents dispositifs d'indexation à base de langages contrôlés ou de langage naturel sur les résultats de recherche d'informations obtiennent des résultats hétérogènes.

Certaines montrent de meilleurs résultats avec une indexation par sélection d'unitermes dans le texte des documents. « *Tout semble se passer comme si « les anomalies » de la langue étaient moins préjudiciables à la recherche d'informations que les inévitables imperfections et approximations des langages contrôlés.* » ([3], Menon, p.21). D'autres études arrivent à la conclusion que le langage contrôlé est plus performant. Les résultats dépendent des méthodes d'évaluation, les mêmes paramètres n'étant pas toujours pris en compte.

La solution idéale pourrait être d'effectuer la recherche sur le texte intégral (les moteurs de recherche bénéficiant de progrès constants) qui permet de faire sortir des références pertinentes non indexées avec le mot utilisé par l'internaute, avec en option la possibilité d'utiliser un langage documentaire (ou une assistance terminologique fournie par le langage documentaire) pour étendre ou restreindre le recherche.

1.1.3 Evolution de l'indexation pour le numérique (enrichissement avec les métadonnées)

Avec l'avènement de l'information numérique, « *on ne parle plus seulement d'indexation, mais également d'enrichissement, d'annotation et de marquage, de métadonnées et de balises.* » ([8], Menon, p.340]).

Les métadonnées sont des ensembles de données structurées qui décrivent, expliquent, localisent des ressources et en facilitent la recherche, l'usage et la gestion ([12], NISO).

> «*On peut considérer que les métadonnées peuvent fournir toutes sortes d'informations relatives à une ressource ou à son usage :*
> - *des métadonnées descriptives (du contenu, de l'origine de l'information, bibliographique…),*
> - *des métadonnées administratives (juridiques, commerciales…),*
> - *des métadonnées structurelles (relations entre composants d'une collection, fractionnement…).* » ([14], Richy, Despres, p.3)

On ajoute au document des informations caractérisant son contenu, mais de manière à le rendre manipulable par des machines.

L'indexation, incluse dans les métadonnées, peut alors se rapporter à un document dans son intégralité mais également à des portions de celui-ci, qui sont séparées et repérées numériquement et que l'on rend ainsi accessibles isolément.

Les métadonnées d'un document peuvent avoir plusieurs origines, certaines étant intégrées automatiquement lors de sa création, d'autres pouvant être ajoutées par différents intervenants au fil du temps.

On arrive ainsi à une définition élargie de l'indexation, qui inclus différentes façons de décrire un document, et dans laquelle on peut également inclure la catégorisation, c'est-à-dire le rattachement à un groupe prédéfini.

1.1.4 Synthèse des types d'indexation (manuelle, automatique, collaborative)

Les origines principales de l'indexation sont donc :

- les auteurs qui donnent des mots-clés relatifs à leurs articles après rédaction ;
- des professionnels qui utilisent en général les langages contrôlés, procédé onéreux, avec la difficulté de choisir une terminologie en rapport avec les utilisateurs ;
- les traitements informatiques qui génèrent des mots-clés automatiquement grâce à l'analyse linguistique et statistique de texte, procédé considéré comme moins coûteux mais moins efficace (mais ils peuvent également être aidés par des langages contrôlés);
- les utilisateurs finaux qui taggent les articles après publication.

Il s'agit de différentes façons de procéder mais qui peuvent avoir une partie des résultats communs, surtout pour des articles avec un sujet précis qui ne prête pas à interprétation.

Je vais maintenant détailler un peu plus indexation automatique et tags et voir ce qu'ils apportent par rapport à l'indexation « traditionnelle ».

- Indexation automatique

L'indexation assistée par ordinateur sélectionne les sujets significatifs des documents par des méthodes statistiques ou sémantiques.

Les méthodes statistiques partent du principe que la fréquence d'utilisation d'un mot dans un article, son positionnement, sa typographie ou son appartenance à une liste d'autorité permettent de mesurer sa valeur significative (les mots-vides étant écartés). Comme pour l'indexation manuelle libre, le problème principal de cette indexation est l'absence de prise en compte de la synonymie.

Des traitements linguistiques permettent de lever certaines ambiguïtés en utilisant l'analyse morpho-lexicale (segmentation, lemmatisation, identification dans un dictionnaire) et l'analyse syntaxique.

La plupart des outils utilisent ces méthodes conjointement, il faut cependant les adapter au fonds documentaire, aux pratiques et aux besoins des utilisateurs, ce qui nécessite presque toujours une intervention humaine complémentaire ([9], Mercuri).

- Tags

Le tag est une étiquette reliée à une ressource et fondée sur la notion d'accès. Le choix des termes est libre. Le tag a cependant un certain encadrement car il faut définir qui a le droit de tagger (les personnes ayant accès), quelles sont les parties sur lesquelles on peut tagger et quels sont les moyens pour le faire ([10], Monnin).

Les avantages du tag sont la rapidité, la facilité et la souplesse d'utilisation ainsi que le côté évolutif. Les termes employés sont en général plus courants et proches des utilisateurs, ils donnent un point de vue original ou une connaissance particulière de l'internaute.

Les tags présentés sur le Web (sous forme de nuages par exemple) vont induire l'utilisation de ces termes par les autres utilisateurs, ils créent de nouveaux chemins d'accès à la ressource. Ils sont représentatifs de l'intérêt porté à la ressource mais sont aussi liés à une utilisation personnelle qui sert aux autres et n'ont pas forcément un but collaboratif. Ils peuvent également révéler des minorités qui ne ressortiraient pas du volume d'informations autrement.

Le tag manque cependant de précision, ne traite pas la synonymie et ne décrit pas de manière exhaustive. Le nuage de tags ne donne pas de relations entre les termes ([2], Broughton).

> « *L'ensemble des tags assignés par les internautes constitue une folksonomie.* […] *La construction des folksonomies étroites repose sur l'indexation par l'utilisateur de ses propres ressources uniquement, tandis que les folksonomies générales sont le résultat de l'attribution de tags par les internautes à leurs propres ressources, mais également à celles d'autrui.* » ([4], Durieux, p.71).

On voit ici que le tag devient un moyen de communication. L'utilisateur final qui appose le tag peut finalement être l'auteur ou le producteur qui veut faire passer des messages par rapport au contenu.

Les tags montrent surtout leur potentiel pour décrire autre chose que du texte (images…). Ils sont ensuite utilisés pour la recherche, la navigation ou le filtrage.

Comment choisir ?

> « *L'indexation est prise dans un dilemme bien connu : on la souhaite objective, fiable, régulière, économe en moyens d'une part, conceptuelle et intuitivement accessible d'autre part. Automatisée, elle satisfait les premiers réquisits, mais pas les seconds ; humaine, c'est le contraire.* » ([13], Polity, Henneron, Palermiti, p.140).

En parcourant les interventions de la Journée d'étude ADBS de 2004 sur l'indexation à l'heure du numérique, on remarque que sont utilisées : de l'indexation libre par unitermes ou par groupes de mots, de l'indexation avec un vocabulaire contrôlé, de l'indexation avec des rubriques de classement, voir une combinaison de plusieurs types d'indexation, inscrits dans des schémas standardisés de métadonnées ou suivant des modèles spécifiques.

Le recours à l'indexation automatique du texte intégral, ou à partir d'éléments tels que les titres et le résumé est en général complété par une part manuelle d'indexation comme des rubriques de classement ou de caractérisation supplémentaire ou un contrôle de l'indexation automatique. Rares sont les systèmes entièrement basés sur l'automatisation.

On passe également « *de l'analyse unitaire des documents à la nécessité d'une vision globale et structurée des besoins et des ressources de l'organisation.* » ([8], Menon, p.342).

Nous avons remarqué que l'indexation pouvait prendre de multiples formes et nous allons explorer par la suite les différentes possibilités en prenant le terme au sens large, qui est de donner des informations sur un article que ce soit en terme de contenu, de descriptif ou d'appartenance à une entité plus large.

1.2 Langages documentaires

Comme nous l'avons vu, l'indexation libre pose un certain nombre de problèmes pour la recherche. Les spécialistes de l'information ont donc eu recours à des listes de vocabulaire contrôlé de plus en plus riches et structurées pour améliorer les résultats de recherche.

La définition des langages documentaires selon Jacques Maniez est la suivante :

> « *Code sémantique de représentation des sujets, permettant à un système documentaire de repérer les documents par une formulation rigoureuse de leur contenu et aux utilisateurs d'ajuster leurs interrogations à ces formulations.* » ([7], Maniez, p.207).

Toujours selon Jacques Maniez, les fonctions que l'on peut attendre d'un langage documentaire sont de normaliser la représentation des sujets pour une recherche optimale et de permettre à l'usager de naviguer entre des sujets voisins.

Le vocabulaire contrôlé utilisé à la fois pour l'indexation et pour la recherche doit permettre d'améliorer l'identification entre les deux. Un terme sera sélectionné parmi un groupe de synonymes, avec une forme d'écriture (masculin, singulier, sigle, ...).

Les règles étant établies préalablement, ceci simplifie la tâche de l'indexeur ou de la personne qui effectue la recherche.

1.2.1 Notions sur l'organisation des connaissances

Organiser un ensemble de documents en classes selon des caractéristiques communes est quelque chose de très répandu. La difficulté vient du fait que tous les documents ont plusieurs caractéristiques qui peuvent être utilisées pour faire des regroupements différents. Il est ainsi impossible de créer une classification qui convienne à tous les domaines et les classifications sont d'autant mieux adaptées à une collection que le domaine d'application est étroit.

Dans un domaine précis, les langages documentaires utilisés doivent être construits en analysant les documents relevant du domaine pour en extraire les concepts. L'analyse de ces concepts permet de ramener le nombre théoriquement illimité des termes utilisés dans les documents à un nombre restreint de catégories pertinentes. L'organisation ainsi établie n'est cependant pas fixée pour toujours mais doit évoluer en même temps que le domaine.

La représentation du domaine qui sera retenue dépend également du contexte et des besoins auxquels on doit répondre par la construction de ce langage documentaire. Le domaine d'étude définit l'ensemble des termes à modéliser. Les acteurs qui l'utiliseront et les

actions réalisées déterminent comment les termes doivent être représentés dans la modélisation. Chaque contexte génère un langage documentaire différent.

En résumé, pour extraire et organiser les informations afin de caractériser de manière efficace chaque document, il faut tenir compte des types de recherches effectuées et des requêtes utilisées (recherches globales ou recherches pointues, par exemple). On peut aussi définir des niveaux de description (par exemple technique, fonctionnel et applicatif pour des logiciels) qui utiliseront chacun un langage documentaire spécifique ([13], Polity, Henneron, Palermiti, chap.3).

> « *L'utilisation d'un système de représentation des connaissances permet de mieux traduire le besoin exprimé en recherche d'information. Le système est alors en mesure de fournir des informations complémentaires à l'utilisateur pour lui permettre de préciser ou de reformuler sa requête s'il n'obtient pas directement des résultats satisfaisants.* » ([13], Polity, Henneron, Palermiti, p.173)

L'organisation des documents à l'aide de langages documentaires permet donc de les retrouver plus facilement ce qui a entraîné le développement de nombreux types de langages, allant du plus simple au plus complexe.

1.2.2 Listes simples, hiérarchisées et/ou relationnelles

Pour avoir un panorama des langages documentaires, je vous invite à vous reporter au numéro spécial de Documentaliste - Sciences de l'information sur les langages documentaires et outils linguistiques, très complet et ou l'on peut trouver des explications sur classifications, thésaurus, taxonomies ou ontologies ([3]). Détailler toutes ces notions pourrait faire l'objet d'un chapitre à part entière, je n'évoquerai donc que certains aspects liés à ma problématique.

D'après Suzanne Waller, pour être de bons auxiliaires de recherche, les langages documentaires doivent être élaborés à partir du vocabulaire courant des utilisateurs, autant sinon plus que des textes à indexer. *« Leur adaptation au milieu est le meilleur garant de leur qualité.* » ([15], Waller).

On ne peut pas dire d'un type de langage documentaire qu'il soit bon ou mauvais : il faut les évaluer selon leur pertinence et leur efficacité en fonction de l'objectif.

Les listes simples reposent sur une action de regroupement : on rapproche les entités qui ont au moins une caractéristique commune (de contenu comme le sujet, ou externe comme le format par exemple).

Les listes hiérarchisées impliquent un classement en plus du regroupement. Or, chaque entité possède plusieurs caractéristiques essentielles et appartient donc potentiellement à plusieurs groupes distincts. Il faut donc faire des choix qui valorisent un aspect du document au détriment des autres possibilités. De plus, certains documents entrent difficilement dans l'un des groupes définis préalablement. ([6], Hudon, Mustafa El Hadi).

Certains langages documentaires sont très rigides et conviendront uniquement pour des domaines stables, d'autres sont un peu plus souples avec une structure permettant d'introduire des évolutions. Ils peuvent être mono-hiérarchique (un seul terme générique) ou accepter la poly-hiérarchie (rattachement à plusieurs termes génériques).

Les classifications et autres listes hiérarchisées sont des outils de travail lourds et complexes, difficiles à créer et à utiliser. Il faut suivre les évolutions du domaine traité, intégrer de nouveaux sujets mais aussi de nouvelles façons de considérer les sujets existants en modernisant la terminologie ([5], Hudon).

La difficulté reste toujours le lien entre indexation et requête utilisateur. Même avec un langage documentaire, l'indexation peut être multiple : plusieurs profondeurs d'indexation ou niveaux de précision peuvent être pris en compte.

1.2.3 Adaptation au web : avantages et inconvénients

L'utilisation d'un langage documentaire structuré pour accéder à une collection présente plusieurs avantages :

- elle autorise le furetage et la navigation ;
- elle facilite l'élargissement ou le rétrécissement des recherches ;
- elle fournit un contexte aux termes de recherche et aux ressources ;
- elle garantit une assistance aux personnes non spécialistes du domaine.
  ([5], Hudon)

En effet, souvent, la recherche d'information ne résulte pas d'une demande clairement établie mais d'un besoin dont on ne cerne pas bien les contours et pour lequel on ne sait pas à l'avance s'il existe des documents ciblés. Les langages documentaires en consultation Web ou en assistance à la recherche vont permettre d'accéder à une terminologie et de naviguer dans les termes spécialisés, fournissant une aide précieuse.

Les grandes classifications sont cependant peu utilisées sur le Web, seuls les thésaurus sont assez bien représentés car ils sont très bien adaptés à la logique de recherche booléenne des moteurs de recherche. La lourdeur de la mise à jour entre également en ligne de compte.

Mise à jour rare et complexe (maintien de la cohérence)	Mise à jour systématique régulée par l'évolution du fonds	Mise à jour progressive et négociée	Mise à jour fréquente, facile, immédiate
• Classification universelles • Ontologies formelles	• Thésaurus	• Annuaire internet • Ontologie sémiotique • Cartes conceptuelles	• Folksonomie

D'après Zacklad, Fréquence et facilité de mise à jour ([16], Zacklad, p.13)

Tab. 1 - Fréquence et facilité de mise à jour de langages documentaires

La recherche est de plus en plus multidisciplinaire, avec des passerelles entre les domaines. Les grandes classifications qui présentent une hiérarchie stricte ne peuvent pas répondre au besoin de transversalité et de navigation entre les sciences. Il faut donc abandonner l'idée d'une indexation fine utilisant ce type de langage mais elles peuvent être utiles pour la structuration d'un système ou d'une recherche ([15], Waller).

Plus souples et plus conviviales, les catégories de sujets ou de types de documents (taxonomies) sont omniprésentes sur les sites web, utilisées pour la navigation ou pour le filtrage.

L'objectif des listes de vocabulaire contrôlé devient marketing : on ne se soucie plus de normalisation et de stabilité, les termes sont plus intuitifs mais l'ensemble manque de cohérence et de logique car non hiérarchisé. Le manque d'uniformité peut dérouter les internautes et ne convient pas à la navigation.

L'essor des moteurs de recherche sur le texte intégral a fait craindre la fin des langages documentaires et des pratiques d'indexation classiques, au profit de l'automatisme. Cependant, malgré les performances actuelles des moteurs de recherche, ils sont souvent associés à des systèmes de classification qui apportent une dimension de repérage de l'information supplémentaire.

D'après Manuel Zacklad, les évolutions viendront certainement de solutions hybrides d'aide à la recherche d'information avec la participation de plus en plus aisée des experts et des utilisateurs, en complémentarité avec l'usage des moteurs de recherche ([16], Zacklad).

Nous allons maintenant voir comment les utilisateurs procèdent pour leur recherche d'information, quelles sont les aides mises à leur disposition ainsi que les différentes utilisations des mots-clés.

2 Usages des mots-clés pour la recherche d'information sur Internet

Interrogation des moteurs de recherche et navigation sont deux approches complémentaires pour retrouver des informations sur le Web.

Les stratégies de présentation de l'information conditionnent les performances de recherche des usagers. Ces derniers vont d'autant mieux trouver l'information qu'ils recherchent si on leur fournit une aide, du vocabulaire adéquat ou s'ils sont avertis des thématiques ou sujets traités. On peut, par exemple, proposer à l'utilisateur les documents ayant un lien avec sa requête, une catégorisation qui fasse émerger les documents sur le même sujet, voire lui conseiller d'autres documents que ceux de la même catégorie mais ayant un autre point commun.

La difficulté est de se positionner par rapport au contexte de la recherche, une même requête pouvant correspondre à des besoins différents.

Attribuer des mots-clés aux documents va permettre de proposer à l'utilisateur différentes voies d'accès à l'information, qu'il pourra utiliser en fonction de ses pratiques (en s'assurant de la facilité d'utilisation et sans surcharger le site).

2.1 Evolution des usagers et de leur utilisation des moteurs de recherche

On peut schématiser la démarche de l'internaute face à un besoin d'information de la manière suivante :

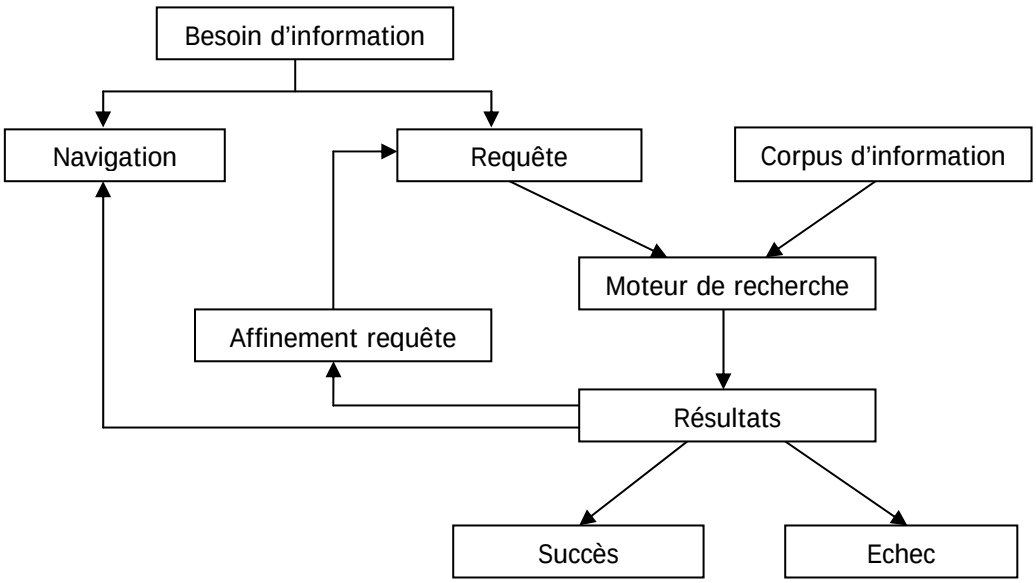

Fig. 1 - Etapes d'une recherche d'information sur un site Web

L'objectif est bien sûr de limiter au maximum les échecs, qui se traduiront par une sortie du site, l'internaute allant chercher ailleurs, et de l'amener au contraire vers ce qu'il recherche.

2.1.1 Perte des intermédiaires professionnels de l'info-doc au profit des utilisateurs finaux, impact sur le comportement

Lorsque sont apparues les premières bases de données spécialisées, la complexité des requêtes utilisées nécessitait l'intervention de spécialistes de l'information. Il y avait alors dissociation entre la personne qui avait le besoin et la personne qui conduisait la recherche.

Avec la simplification des accès à l'information, les demandeurs font directement leurs recherches. On s'aperçoit alors que l'exploration va entraîner l'utilisateur à s'interroger sur son besoin et que celui-ci va parfois évoluer au-cours de son interrogation.

> « *Ce sont les différents résultats des recherches entreprises qui vont modifier les buts, préciser le besoin en information de l'usager et donc agir sur sa représentation du problème.* » ([26], Kolmayer, p.124).

L'activité professionnelle des utilisateurs n'est plus centrée sur la recherche d'information. La finalité de la recherche peut être très variée (sollicitation ponctuelle, constitution d'un dossier,...), de même que la nature des informations recherchées (explications, données statistiques, normes) ([19], Dalbin).

L'usager est donc moins centré sur la requête en elle-même, qu'il simplifie au maximum, et s'attache plus aux possibilités d'obtenir une réponse rapide et de naviguer facilement.

2.1.2 Types de requêtes et moteurs de recherche

Les moteurs de recherche ont un rôle central sur le Web : faciles à utiliser, ils fournissent des résultats à n'importe quelle requête. Mais quels résultats ?

La recherche directe par requête nécessite une connaissance précise de ce que l'on cherche par rapport au butinage et à la navigation par hyperlien.

On doit comprendre le fonctionnement du moteur et le positionnement des résultats pour pouvoir effectuer la recherche de façon pertinente. Or, sur le Web, l'utilisateur est habitué à Google et va donc considérer que tous les moteurs qu'il utilise fonctionnent de la même manière sans s'intéresser de plus près aux particularités de chacun. Seul devant son ordinateur, il ne bénéficie pas de conseils d'utilisation. D'où encore une fois, la nécessité de simplicité et de clarté ([28], Simonnot).

> « *Les algorithmes d'indexation et de classement constituent le cœur technologique des moteurs de recherche. Derrière la liste des résultats se donnent à lire des principes de classification et d'organisation de l'information et des connaissances : l'affichage d'une liste de résultats est l'aboutissement de l'itération de principes non plus seulement implicites mais invisibles et surtout dynamiques.* » ([20], Ertzscheid).

Les résultats d'enquêtes reprises par Madjid Ihadjadene montrent que les requêtes des usagers ne comportent qu'un, voire deux mots et qu'en grande majorité elles ne contiennent pas d'opérateurs booléens. Les utilisateurs n'ont presque jamais recours aux fonctionnalités avancées. Ils vont rarement au-delà des trois premières pages de résultats. Ils utilisent les catégories disponibles de suite mais vont rarement dans les sous-catégories d'une arborescence. Les propositions un peu sophistiquées comme les cartographies sont souvent délaissées, car associées à des temps de traitement plus longs sans améliorer les performances. Ces comportements sont observés aussi bien sur les moteurs de recherche généralistes que sur les intranets ([25], Ihadjadene).

Les faiblesses que l'on peut trouver dans les requêtes (fautes d'orthographe, imprécisions, langage naturel,...) sont partiellement traitées par les moteurs. Les possibilités de choix d'autres termes données à l'usager, utilisables par simple clic pour rebondir dans sa recherche vont donc lui permettre de mieux satisfaire sa demande.

2.1.3 Adéquation des termes et satisfaction de l'usager

Lorsque le sujet de recherche est explicite, la détermination automatique des ressources pertinentes n'est pas simple pour autant. Les principaux inconvénients résident dans la difficulté d'exprimer la requête mais aussi dans la nécessité de disposer d'un « étiquetage » des ressources permettant d'établir un lien entre la requête de l'utilisateur et le document. Cet étiquetage peut se présenter sous la forme de mots-clés représentatifs du document, ou exploiter la notion de catégories ou de thématiques des documents.

- **concept de pertinence en recherche d'information ([28], Simonnot)**

La pertinence de la recherche d'information va dépendre de la relation qui s'établit entre la collection et l'utilisateur.

Dans cette relation, il faudrait pouvoir prendre en compte le domaine concerné par l'étude, le but de la recherche (à quoi vont servir les informations trouvées) et le contexte, qui correspond à tout ce qui peut influencer le déroulement de la recherche et de l'évaluation des résultats (par exemple, les documents que l'utilisateur connait déjà, le temps dont il dispose pour faire sa recherche, ...).

En réalité, tous ces paramètres sont très difficiles à intégrer et la qualité des résultats de la recherche dépend essentiellement de la formulation de la requête.

Ceci est également constaté par Anne Boyer dans son « Analyse des usages pour améliorer l'accès aux ressources » :

> « *Traditionnellement, les requêtes sont évaluées indépendamment de l'utilisateur, de son profil, de son historique et de son contexte, les seuls critères étant la sélection par le contenu, la popularité, la confiance et la disponibilité des ressources.* » ([17], Boyer).

Comme nous l'avons vu au paragraphe précédent, l'utilisation des opérateurs booléens, des caractères + et – devant les termes pour les rendre obligatoire ou les exclure, ou les autres outils mis à disposition par les moteurs de recherche sont loin d'être évidents à manipuler pour la plupart des utilisateurs. Les requêtes sont souvent limitées à un ou deux mots ce qui est insuffisant pour définir précisément le besoin. Les formulaires de recherche avancée ne rencontrent pas vraiment de succès car ils demandent un effort de compréhension et d'adaptation. De plus, les possibilités de combinaison de termes sont souvent limitées. On peut ainsi arriver à des résultats éloignés, voire inverses de ce qui est souhaité.

Les documents n'ont pas une valeur intrinsèque. Leur valeur est variable dans le temps (certaines technologies deviennent vite obsolètes tandis que d'autres, mêmes anciennes présentent toujours un intérêt) et va dépendre de l'utilité qu'ils vont avoir pour l'usager.

Or, de plus en plus, devant le nombre de ressources disponibles, les internautes passent beaucoup de temps à chercher, naviguer sur les sites ce qui les rend moins disponibles pour la lecture et l'analyse des documents trouvés. Il devient donc indispensable de leur faciliter la tâche en leur proposant des aides qui leur permettent de mieux visualiser les types d'informations disponibles et de les sélectionner.

2.2 Aide à la recherche et à la navigation

> « *Les méthodes d'aide à la navigation reposent sur un certain nombre d'éléments pouvant guider l'utilisateur, comme des termes liés provenant d'un thésaurus ou extraits automatiquement des textes, des catégories prédéfinies, des documents proches de celui qui est regardé, la visualisation d'éléments importants ou petits résumés.*
>
> *L'utilisation des moteurs de recherche sur les documents en plein texte a conduit à travailler sur l'interaction automatique entre utilisateur et système, avec propositions d'expansion ou d'affinement de requêtes, procédés visant à résoudre le problème de vocabulaire.* » ([22], Grivel, chap.5).

Nous allons maintenant voir quelles sont les aides à l'utilisateur les plus fréquentes sur les sites Web.

2.2.1 Filtrage des résultats par catégories

Le filtrage consiste à affiner les résultats d'une recherche par mot-clé, lorsqu'ils sont dispatchés dans des catégories par facettes, en sélectionnant une ou plusieurs de ces catégories. Cette utilisation est particulièrement développée dans le commerce électronique.

La difficulté porte surtout sur :

- le choix et l'identification des facettes, qui doivent être caractéristiques des unités documentaires de la collection,
- la création d'une hiérarchie ou d'une liste de valeurs pour chaque facette, le nombre et l'ordre d'affichage,
- la classification des éléments de la collection dans les facettes ([22], Grivel, chap.5).

Un document peut être classé dans plusieurs facettes, contrairement à d'autres langages contrôlés.

Le design sur le site a un impact majeur sur l'utilisation et les performances des sélections proposées. Les consignes pour l'utilisation des facettes sont :

- de les présenter dans un panneau latéral à gauche de l'écran, à côté de la liste des résultats,
- de les nommer de façon intuitive et univoque (la longueur des noms des catégories a également une importance car ils risquent d'être tronqués sur le site pour ne pas avoir une colonne trop large),
- de ne présenter que les plus importantes si elles sont nombreuses, avec possibilité de cliquer pour voir les catégories suivantes,
- d'avoir la possibilité de sélectionner une ou plusieurs valeurs et de les croiser entre elles.
- de supprimer dans la liste la catégorie qui a été sélectionnée et de l'afficher dans le fil d'Ariane.

Evaluer la qualité d'un système de facette est difficile. On regarde en général si tous les éléments du fonds entrent bien dans les catégories définies et le nombre de clics pour atteindre un élément. On peut également suivre leur utilisation à partir des traces des utilisateurs sur le site.

Des études montrent que l'on retrouve plus de documents pertinents sur un temps fixe et que l'on consulte moins de documents non pertinents en utilisant les facettes ([22], Grivel, chap.5), ([23], Hearst), ([24], Hearst).

L'idéal serait de pouvoir activer des filtres différents selon l'intérêt de l'utilisateur.

Les filtres rencontrés le plus souvent utilisent des tris par concepts (discipline, fonctionnalité, …), par date ou par entités nommées.

2.2.2 Extension de requête par suggestion cliquable

L'indexation des articles permet d'afficher, en complément des résultats de requête, des termes auxquels l'internaute ne pense pas spontanément, qui peuvent l'intéresser et lui permettre de relancer sa recherche.

L'utilisation de listes de synonymes ou de parenté directe permet de faire des suggestions pertinentes de termes (voir aussi…) à partir des mots saisis dans le moteur de recherche.

En enregistrant les historiques de requête rattachés à un article, on peut également donner des propositions telles que « les personnes qui sont allées voir cet article ont également vu ou recherché… ».

Pour les tags, on affiche souvent les mots les plus utilisés mais on peut également faire ressortir, par exemple, les derniers mots entrés (les plus récents).

Grâce aux métadonnées, on peut associer un document contenant une liste de mots-clés aux autres documents partageant les mêmes mots-clés. Ces derniers peuvent apparaître en suggestion pour l'utilisateur.

Les métadonnées permettent également de faire des propositions sur des documents liés tels que fiche bibliographique, autres articles du même auteur ou commentaires sur l'article.

2.2.3 Accès à des langages contrôlés

Comme nous l'avons mentionné page 21, les thésaurus sont les langages documentaires les plus utilisés sur le Web.

> « *Le thésaurus a été largement utilisé à la recherche d'information en miroir du thésaurus à l'indexation : les documents indexés avec un thésaurus étaient utilisés avec le même thésaurus à la recherche.* » ([19], Dalbin).

Avec l'expansion de la recherche en texte intégral et du traitement automatique de l'indexation, les choses ont évolué. De plus, les internautes ne veulent plus perdre de temps à chercher des mots dans une liste complexe.

Associés aux moteurs de recherche, les langages contrôlés sont maintenant utilisés de manière transparente plutôt que mis à disposition des internautes. On peut ainsi associer automatiquement à la requête les synonymes ou termes définis comme équivalents. On peut également proposer à l'internaute d'élargir ou de restreindre sa requête en prenant en

compte termes génériques ou spécifiques. Ceci permet d'améliorer les résultats de recherche sans demander d'efforts à l'utilisateur.

Les classifications peuvent aussi être utilisées sur le Web pour la navigation. Elles apportent alors une possibilité agréable de parcourir la collection. L'inconvénient est que la structure est figée et ne donne qu'une seule possibilité de parcours.

Nous venons de voir comment l'indexation peut améliorer la recherche de l'internaute lorsqu'il est sur un site Web. Dans le prochain sous-chapitre, nous montrons que les mots-clés doivent aussi permettre d'amener l'utilisateur sur le site.

2.3 Positionnement sur les moteurs de recherche

D'après Olivier Andrieu, les moteurs de recherche sont un passage obligé sur Internet (ils génèrent en moyenne un tiers des visites sur un site Web). Il est donc très important de bien se placer dans les pages de résultats, sur Google en particulier, qui représentait près de 90% du marché en France en 2009.

Travailler sur le positionnement de son site permet d'arriver sur les premières pages de résultats pour les recherches sur des notions représentatives du contenu du site, ce qui est d'autant plus important que les internautes ne consultent que les premières pages.

Il faut donc définir les mots-clés qui correspondent à la fois à ce que l'on veut montrer et à ce que les internautes veulent voir. Ces mots-clés doivent ensuite être positionnés dans les pages du site pour qu'elles soient retenues par rapport à l'algorithme de pertinence du moteur et bien classées dans la liste des résultats ([29], Andrieu).

2.3.1 Approche marketing

L'approche marketing est basée sur le principe que la mission de toute organisation est d'adapter l'offre (un produit, un service) à la demande du client (consommateur ou entreprise) ([32], Marck).

Cette approche consiste donc à choisir les mots-clés en fonction du trafic généré. Les termes utilisés doivent bien sûr refléter le contenu du site mais également être adaptés à la cible visée.

Le vocabulaire employé sera fonction de celui utilisé par les clients potentiels : il faut établir des profils d'utilisateurs et regarder s'ils utilisent plutôt des termes techniques ou du vocabulaire plus familier, des sigles ou autres spécificités. Pour être placé sur un concept, il ne faut pas hésiter à utiliser toutes les terminologies et variantes existantes.

Cette démarche demande de bien réfléchir aux concepts à travailler car le positionnement ne peut se faire que sur une liste assez restreinte de termes pour arriver à se placer sur la première page de résultats. Les mots-clés choisis doivent répondre à des attentes ciblées.

Un bon positionnement est un moyen de communication qui permet d'assurer du trafic sur un site Web, de prendre des parts de visibilité et des clients à la concurrence et de développer la fidélisation des internautes. Il donne une bonne image du site Web et donc de la société à qui il appartient. La notoriété du site renforce également sa crédibilité.

2.3.2 Mise en œuvre

Quels sont les points à prendre en compte pour choisir ses mots-clés ?

Selon Olivier Andrieu, il faut regarder :

- « *le potentiel du mot-clé : est-il souvent saisi sur les moteurs de recherche par les internautes ?*
- *la faisabilité technique du positionnement : est-il possible de positionner une page de votre site sur ce mot-clé ?* » ([29], Andrieu).

La première chose à faire est donc de regarder les statistiques de saisie du mot choisi, ce qui est possible notamment grâce au générateur de mots-clés de Google, disponible à l'adresse suivante : <https://adwords.google.fr/select/KeywordToolExternal>.

Pour savoir si l'on peut se positionner correctement sur ce terme, il faut ensuite regarder le nombre de résultats générés lorsqu'on le saisit sur Google. Plus le nombre de réponses est élevé, plus le nombre de concurrents potentiels est important et plus il sera difficile de se placer. Le nombre de réponses au-delà duquel le positionnement sera très difficile dépend du domaine d'activité et de l'agressivité des concurrents.

> « *Bien choisir les mots-clés pour un référencement consiste donc en un arbitrage entre le potentiel des termes choisis et la faisabilité technique d'un positionnement sur ceux-ci.* » ([29], Andrieu).

La dernière étape sera d'inclure les termes dans les zones stratégiques du site (titre, URL,...) en utilisant les techniques de SEO (Search Engine Optimisation).

Enfin, il faut réaliser un bilan régulier des mots-clés qui ont généré le plus de trafic sur le site pour adapter la liste des mots choisis pour le positionnement en fonction des résultats.

Afin de compléter notre investigation, nous allons observer, dans le chapitre suivant, les dernières évolutions dans la mise à disposition de l'information scientifique sur le Web.

3 Evolutions de l'information scientifique sur le Web

3.1 Notions de fragmentation et de structuration des documents

La recherche d'information devient d'autant plus complexe que le volume d'information disponible augmente. Ceci est encore amplifié par le fait que, grâce aux métadonnées, les articles ou documents qui formaient un tout et donc une unité documentaire par le passé sont maintenant découpés en « morceaux » disponibles individuellement. On parle alors de « fragmentation de l'unité documentaire ».

> « *Les écrits sont de plus en plus difficilement identifiables en tant que tels car fragmentés, déstructurés, sortis du contexte qui les légitime, difficilement localisables également à cause de la multiplicité des plates-formes d'hébergement.* » ([30], Chaudiron, Ihadjadene, Maredj).

Pour s'adapter à ce nouveau fonctionnement, les textes scientifiques sont ainsi de plus en plus structurés, et rédigés suivant une modélisation adaptée au Web, qui permet de les rendre utilisables et manipulables par des programmes, en donnant l'accès à l'ensemble des composants.

Les documents et données qu'ils contiennent sortent donc des collections, des revues ou des actes de colloques auxquels ils appartiennent pour avoir une vie par eux-mêmes.

Ceci est facilité par les nombreuses possibilités de dépôt sur les archives ouvertes, les archives institutionnelles, les sites personnels ou les sites commerciaux. On peut ainsi accéder par exemple à un seul chapitre d'une thèse, à condition que ce découpage ait été déterminé lors de l'intégration du document sur le Web. On peut également accéder à la même information de plusieurs façons et sous des conditions différentes.

La fragmentation des documents numériques entraîne de nouveaux besoins descriptifs avec des liens hiérarchiques d'appartenance ainsi que des liens relationnels, que ce soit au niveau de l'article ou au niveau de la collection.

L'indexation des documents est donc toujours d'actualité et correspond même à de nouvelles applications. Avec les avancées technologiques en informatique, on peut même décrire les textes selon plusieurs structurations et logiques différentes.

De plus en plus, les portails et plates-formes scientifiques vont permettre aux utilisateurs de reconstituer un ensemble qui leur est propre, à partir d'informations glanées au fil de leur recherche et de leur navigation. De nouvelles unités documentaires personnalisées vont ainsi être créées ([30], Chaudiron, Ihadjadene, Maredj).

3.2 Notion d'e-science

Avec la notion d'e-science, apparue ces dernières années, on va encore plus loin dans la fragmentation en mettant à disposition des données de recherche sans publication d'articles.

Cette notion est très bien expliquée dans le mémoire de stage d'Emilie Manon [27], dont je m'inspire pour le paragraphe suivant.

Les progrès informatiques ont permis à la fois d'exploiter de beaucoup plus grandes quantités de données et de les partager avec les autres chercheurs du monde entier presque instantanément en utilisant les nouveaux outils du Web tels que blogs, wikis, réseaux sociaux ou archives ouvertes.

On va ainsi trouver aussi bien des données brutes de résultats d'expériences que des nouvelles formes de diffusion comme des présentations vidéo ou audio. Du travail collaboratif est mis en place, avec des interconnexions entre plusieurs bases de données. Les outils de partage permettent des collaborations internationales et multidisciplinaires.

> « *L'exemple de NanoHub, plateforme* **e-science** *dédiée aux nanosciences, fait ici figure de référence : le chercheur y a accès à différentes ressources (données expérimentales, publications, notes, présentations, vidéos, etc.) qu'il peut noter, indexer et commenter dans les nombreux forums proposés.* » ([27], Manon).

Ce type de fonctionnement va donc également mettre à contribution les outils de recherche et d'indexation, avec utilisation de requêtes complexes car elles doivent s'appliquer à des bases de données et formats multiples ([27], Manon).

Pour conclure, il me parait important de souligner la diversité des approches et possibilités d'utilisation de l'indexation, multipliées avec les évolutions de la mise à disposition des informations sur le Web. Nous allons donc dans la deuxième partie de ce mémoire, nous attacher à définir le contexte dans lequel nous évoluons et les besoins des utilisateurs pour ensuite proposer des axes de développement d'utilisation de mots-clés qui soient pertinents pour l'entreprise.

Deuxième partie
Contexte, analyse et résultats

4 Contexte, particularités du fonds documentaire et exploitation actuelle sur le site Web

4.1 Présentation des Editions Techniques de l'Ingénieur

4.1.1 Historique

Les Editions Techniques de l'Ingénieur ont été fondées en 1946, dans le but de créer une encyclopédie de documentation scientifique et technique en langue française, éditée sous forme de fascicules mobiles qui permettent sa mise à jour en permanence. Après le succès du premier volume « Généralités », cinq autres volumes ont rapidement suivi : Mécanique, Construction, Electricité, Chimie et Métallurgie.
Tous les grands domaines scientifiques et techniques sont maintenant couverts : des matériaux à l'énergie, en passant par l'environnement ou les technologies de l'information.

En 1995, les Editions Techniques de l'Ingénieur ont intégré le groupe d'édition européen Weka, qui a continué à faire vivre la collection dans le même esprit.

Les articles de Techniques de l'Ingénieur ont commencé à être numérisés en 1997 lors de la création du site Internet. Ils sont disponibles sur le Web dans leur intégralité depuis fin 2001. Entièrement en français, ils sont largement diffusés dans tous les pays francophones.

La clientèle visée a toujours été regroupée en deux principales catégories :

→ les ingénieurs et cadres techniques des bureaux d'études et de l'industrie (principalement dans des fonctions de recherche et développement ou en production),

→ les étudiants et enseignants de l'enseignement supérieur technique et scientifique.

Actuellement, Techniques de l'ingénieur en quelques chiffres, c'est :

→ 53 bases documentaires (sous-thème correspondant à une unité d'achat)

→ 4 000 articles fondamentaux

→ 60 000 pages d'information

→ 3 000 auteurs reconnus pour leur rigueur scientifique et technique

→ 30 000 abonnés

→ 300 000 visiteurs par mois sur le site www.techniques-ingenieur.fr

→ 1,5 million de pages vues par mois

4.1.2 Services proposés

- **la collection**

Cette importante collection documentaire scientifique et technique fait appel à plus de trois mille spécialistes et propose une information fiable, précise et régulièrement actualisée.

Deux types d'accès sont proposés aux clients :

→ la formule Web,

→ la formule Duo qui contient l'abonnement papier plus l'accès sur Internet.

Les tarifs sont étudiés en fonction des bases documentaires souhaitées ainsi que du nombre d'accès. Les titres, sommaires et introductions des articles sont en accès libre sur le site Web. L'accès au contenu complet de l'article (une vingtaine de pages en moyenne) se fait par abonnement. Les clients ont la possibilité de télécharger les fichiers des articles des bases documentaires auxquelles ils sont abonnés (voir annexe 1, Liste des bases documentaires).

Les dernières bases documentaires créées ne sont disponibles que sur le web. La possibilité de souscrire un abonnement uniquement papier a été supprimée en 2009. L'abandon du papier n'est cependant pas à l'ordre du jour, de nouvelles formes d'abonnement sont même en développement : les « sélections » qui seront des sous-parties des bases documentaires et qui seront brochées.

Les archives (anciennes versions des articles ainsi que les articles traitant de technologies qui ne sont plus utilisées) sont toujours disponibles sur le site Web.

- **les services complémentaires**

En parallèle des bases documentaires traditionnelles, une offre de services d'information a été développée :

→ des fiches pratiques : une collection de fiches et d'outil est disponible sur des thématiques précises, pour accompagner les responsables dans leur travail quotidien.

→ des contenus journalistiques : les Instantanés Techniques Online proposent des actualités permanentes sur les produits et les dernières innovations technologiques. Chaque mois, selon un planning éditorial qui suit les évènements du moment, un dossier complet de 8 à 10 articles est mis en ligne par la rédaction.

→ un espace emploi : plateforme pour échanger, intégrer de nouvelles communautés, cet espace propose également de nombreuses offres d'emploi ainsi que la possibilité de déposer son CV en ligne.

→ des formations et des journées techniques sont également proposées sur les bonnes pratiques et outils nécessaires aux industriels.

4.1.3 Fonctionnement

- **le processus éditorial**

Il est réalisé et validé par des spécialistes francophones :

→ **acteurs internes :**

Les responsables éditoriaux : de culture scientifique (3ème cycle, doctorat, ingénieurs), ils coordonnent le travail des conseillers (experts scientifiques) et des auteurs.

→ **acteurs externes :**

Les conseillers : plus de cent cinquante experts scientifiques et techniques forment les comités éditoriaux.

Les auteurs : plus de trois mille spécialistes industriels (directeurs, ingénieurs, recherche et développement, production, technique) ou académiques (chercheurs, enseignants-chercheurs, universitaires) rédigent les articles.

Les conseillers et les responsables éditoriaux déterminent les grandes orientations des bases documentaires. Ils décident des actualisations nécessaires et des nouveaux sujets à traiter. Les nouveaux sujets sont proposés en fonction des tendances mais également selon les demandes clients remontées par les commerciaux.

Tous les articles sont soumis à des comités de lecture composés de plusieurs conseillers.

- **la fabrication**

Les chargés de production réalisent la maquette des articles avec mise en forme et balisage, puis envoient celle-ci à la composition. Après relecture des épreuves, renvoi à l'auteur et corrections, l'article est validé et d'une part, envoyé à l'impression pour les feuillets papier et d'autre part, stocké pour l'alimentation du site Web.

- **la publication Web**

Le site Web est géré avec le CMS (Content Management System) eZ Publish.

Les articles sont stockés en SGML (Standard Generalized Markup Language) qui permet de structurer un document et de le décrire à l'aide de métadonnées.

Des fichiers zip XML (eXtensible Markup Language) sont ensuite générés pour charger l'article sur le site Web. XML offre une syntaxe souple et un ensemble de règles. C'est une technologie de base qui permet de partager et de structurer à la fois les documents et les données sur le Web ([14], Richy, Despres).

Le site est supervisé par le chef de projet web pour les mises à jour, évolutions et projets. Un webdesigner extérieur intervient pour les changements importants de graphisme.

Le webmaster et les développeurs assurent la partie technique (maintenance, corrections, évolutions du site).

4.2 Présentation des articles

Les articles répondent à un cahier des charges précis, aussi bien pour le contenu que pour la structure.

4.2.1 Type de contenu

Chaque auteur reçoit des consignes pour l'aider à rédiger son article afin que celui-ci soit conforme à la politique éditoriale de la collection des Techniques de l'Ingénieur.

La ligne éditoriale permet à l'auteur d'adapter son futur texte au lectorat et à l'esprit de la collection qui, en dépit d'une grande variété de sujets et de collaborateurs, doit être homogène.

- **un contenu adapté au lecteur**

Les lecteurs des articles constituant la collection des Techniques de l'Ingénieur (T.I.) sont, en majorité, des ingénieurs et des techniciens supérieurs. En principe, ce ne sont pas des spécialistes du domaine traité. S'ils consultent les T.I., c'est essentiellement pour éclaircir un sujet, s'initier à une technique qu'ils ne connaissent pas ou peu, voire se rafraîchir la mémoire. La lecture d'un dossier des T.I. doit, par exemple, leur permettre de discuter avec un sous-traitant, de s'orienter vers une solution technique en faisant un pré-choix ou de bien définir un problème lorsqu'ils s'adressent à un spécialiste.

Autrement dit, les T.I. doivent constituer un outil de travail et d'aide à la décision ainsi qu'un support de formation permanente.

Le niveau visé est celui du 2e cycle universitaire (master). Toutefois, une structure des dossiers « en pyramide inversée » (l'auteur allant du plus simple au plus compliqué) doit rendre la collection accessible aux décideurs, qui n'ont pas obligatoirement une formation scientifique, aux élèves des IUT ou aux techniciens.

Le vocabulaire utilisé est technique et scientifique avec une approche encyclopédique.

- **une ligne de conduite pour le déroulement de l'article**

Les articles des Techniques de l'Ingénieur doivent comporter :

→ une brève présentation théorique évitant les longues démonstrations et limitant les aspects historiques,

→ des aspects pratiques largement développés avec applications, cas d'entreprise, retours d'expérience, critères de choix, comparatifs, formulaires,

→ des considérations relatives à la sécurité et à l'environnement lorsque le sujet s'y prête,

→ des informations chiffrées.

4.2.2 Structure de l'article

L'article comprend les éléments suivants :

- un résumé

Le résumé doit contenir la problématique (définition du sujet) et la description des différentes parties de l'article. Il est fourni en français et en anglais.

- des mots-clés

Six mots-clés au moins, en français et en anglais, doivent être indiqués immédiatement après le résumé.

- une introduction

L'introduction présente les domaines d'application, les atouts et les limites de la technique traitée. C'est un avant-propos permettant au lecteur de comprendre l'enjeu du sujet traité, sans pour autant en résumer le plan. Ce préambule doit permettre au lecteur de savoir quel est l'objectif de l'article et sous quel angle il est écrit. Il replace le sujet dans l'environnement technico-économique.

- le corps de l'article

L'article est découpé avec une hiérarchie décimale démarrant après l'introduction (cinq niveaux de titres peuvent être utilisés). Il contient également des encadrés, illustrations et tableaux ainsi qu'une conclusion qui ouvre sur l'avenir et les évolutions technologiques possibles.

- une rubrique « Pour en savoir plus »

Il s'agit d'une annexe de l'article, rédigée sur une fiche séparée et particulièrement importante pour le lecteur qui y trouvera des renseignements complémentaires tels que bibliographie, liens vers d'autres articles T.I., normes et réglementation, brevets ou renseignements économiques.

L'auteur reçoit une feuille de style reprenant les différentes parties de l'article et doit remettre son document sous Word.

Les documents sont ensuite balisés selon la DTD (Document Type Definition) en vigueur, description qui définit le système d'encodage du document SGML. Celle-ci est fixée pour l'ensemble de la collection et est rarement modifiée.

Les principaux éléments de la DTD sont :

→ l'entête avec titre / sous-titre / auteur / commentaire / nota,

→ le sommaire avec les libellés de différents niveaux,

→ l'introduction,

→ les niveaux hiérarchiques du corps de l'article,

→ les tableaux,

→ les images avec figure et légende,

→ les formules mathématiques et chimiques,

→ les mots-clés, sigles, sociétés,

→ les bibliographies,

→ les normes,

→ les adresses Internet.

4.2.3 Volume et diffusion

Environ 400 articles sont produits chaque année. Les nouvelles parutions et mises à jour papier sont groupées et envoyées chaque trimestre. Une mise à jour est diffusée en ligne en moyenne tous les 2 jours.

Les articles archivés sont supprimés de la table des matières et réintégrés manuellement dans une table d'archives. Aucun article n'est supprimé.

Le nombre total d'articles est d'environ 6.000. Ceci est important pour une collection très technique, mais faible par rapport au volume d'informations trouvées sur le Web.

L'accès au contenu des articles est payant et réservé aux clients : la diffusion au plus grand nombre est donc concentrée sur les parties gratuites que sont les titres, sommaires, introductions, noms des auteurs.

L'étude des articles montre qu'ils possèdent tous la même structure et la même typologie de contenu : ils forment ainsi une population homogène qui sera difficile à différencier en-dehors du sujet traité (seuls les articles Recherche et Innovation ont un style particulier, plus pointu).

4.3 Classification actuelle

4.3.1 Thèmes, bases documentaires et rubriques

L'unité documentaire la plus fine est actuellement l'article.

Les articles sont regroupés en unités de plus en plus larges :
Thèmes (12 + Archives) > Bases documentaires (61) > Rubriques (400) > Sous-rubriques (540) > Articles (6000) soit 110.000 pages actives + 50.000 pages archives.
(voir annexe 2, Extrait de la classification pour le thème Mesures et Analyse)

La collection s'articule aujourd'hui autour de douze thématiques principales, dans lesquelles se répartissent les soixante et une bases documentaires (ces bases documentaires, formant autant de sous-thématiques aux thématiques principales).

Thèmes :

1. Mesures et Analyses
2. Procédés Chimie – Bio – Agro
3. Construction
4. Energies
5. Environnement – Sécurité
6. Génie Industriel
7. Mécanique
8. Sciences fondamentales
9. Technologie de l'Information
10. Electronique - Photonique
11. Nanotechnologies
12. Matériaux

Le format papier de la collection compte environ cent quatre vingt reliures à feuillets mobiles. Chaque reliure est identifiée par le nom de la base documentaire à laquelle elle se rattache, avec une codification particulière permettant de suivre l'ordre de classement.

Un article étant composé d'une vingtaine de pages (dans sa mise en forme actuelle), l'encyclopédie complète compte environ cent mille pages.

Chaque année de nouvelles bases documentaires viennent enrichir le fonds (entre deux et quatre par an) et certains articles existant sont mis à jour, supprimés ou remplacés suivant leur degré d'obsolescence.

Cette classification est reprise sur le site Web, et on peut la suivre par navigation successive dans les sommaires ou par le fil d'Ariane qui propose la liste des choix pour chaque niveau.

Elle n'est pas remise en cause car elle correspond à un historique et une habitude de la clientèle, mais présente cependant quelques points faibles : les thèmes regroupent par exemple des disciplines et des secteurs d'activité. Ils sont peu nombreux (12 thèmes) donc peu précis. Les bases documentaires, au niveau suivant, sont nombreuses (61 bases documentaires) mais leur titre n'est pas toujours très expressif.

La question se pose donc de trouver d'autres formes d'accès que ce soit par catégorisation ou par d'autres moyens, qui s'ajouteront à la classification actuelle et permettront aux usagers d'arriver plus facilement à obtenir des articles répondant à leur besoin.

Les mots-clés peuvent par exemple apporter une meilleure visibilité du contenu car les titres des bases documentaires ne sont pas toujours explicites.

4.3.2 Navigation sur le site

- **descriptif**

Créé en 1997 et très vite alimenté, le site Internet Techniques de l'Ingénieur a été rapidement très bien positionné dans les moteurs de recherche généraux. C'est un souci constant de la part de l'ensemble de l'équipe (pas de termes anglais apparents pour apparaître en meilleure place, introduction d'un glossaire avec les mots les plus recherchés). Le nombre de visites est en progression constante. C'est la vitrine de l'entreprise qui permet d'attirer de nouveaux clients mais également l'outil principal de consultation des abonnés au détriment des collections papier.

Le site, géré par le CMS eZ Publish, propose un accès aux articles :

→ par navigation : thème > base documentaire > rubrique > sous-rubrique > articles.

→ par moteur de recherche (Exalead, mis en place depuis moins d'un an) avec recherche simple ou recherche avancée, sur le titre, par auteur, par thème ou en texte intégral.

La classification, qui était adaptée à la collection papier, devient très complexe à suivre sur le site Web. Des articles présentant des contenus pluridisciplinaires sont rattachés à plusieurs bases documentaires. Ceci entraîne des difficultés de compréhension, par exemple entre le fil d'Ariane qui ne présente qu'un seul rattachement principal et les filtres qui tiennent compte de tous les positionnements. A l'inverse, on trouve des articles qui traitent d'une discipline mais avec un ou plusieurs domaines d'application (par exemple, un article peut traiter de chimie appliquée à l'environnement) et qui ne sont classés que dans un seul thème.

La création des nouvelles bases documentaires ainsi que l'évolution des bases existantes peuvent répondre à un besoin marketing plus qu'à un nouveau domaine traité, ce qui entraîne parfois un manque d'homogénéité sur l'ensemble du fonds.

Sur le site, les pages principales pour notre étude sont les suivantes :

→ la page d'accueil qui donne accès aux différents espaces du site ainsi qu'à la liste des thèmes

Fig. 2 - Copie d'écran - Page d'accueil du site www.techniques-ingénieur.fr

→ les pages de présentation des bases documentaires

Sur la copie d'écran de la page suivante, nous avons accès, par exemple, à la liste des rubriques de la base documentaire « Agroalimentaire ».

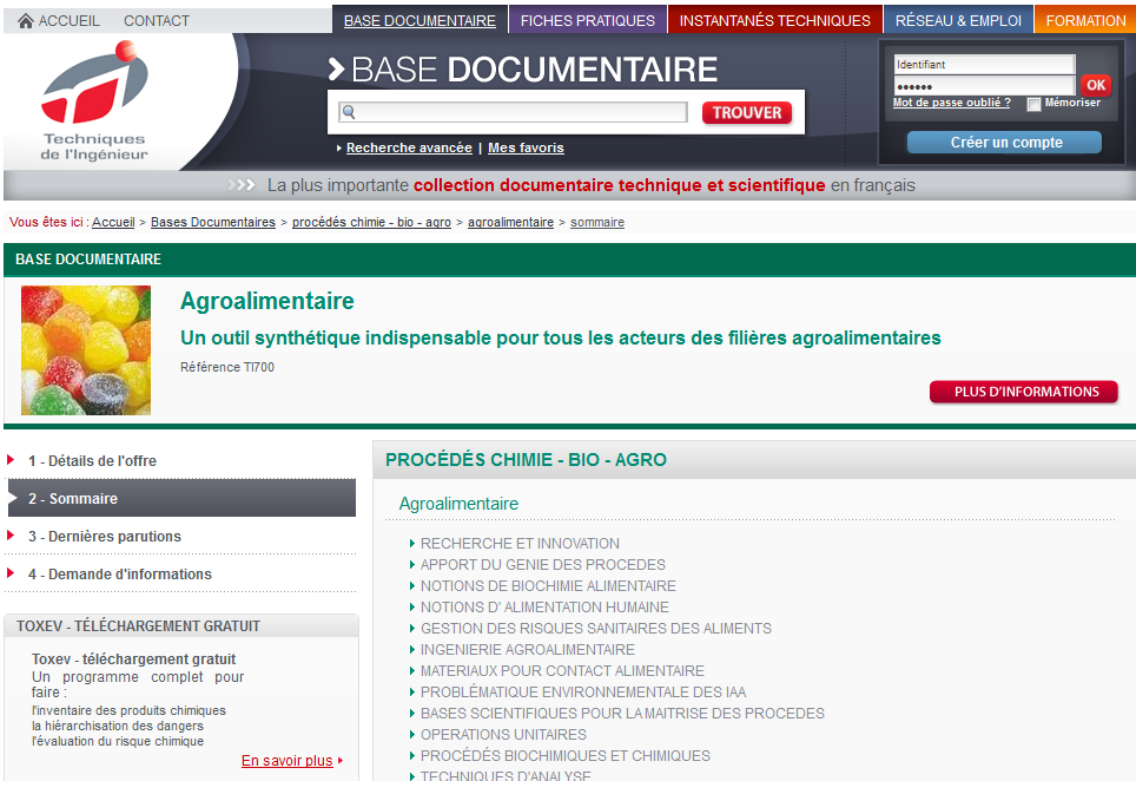

Fig. 3 - Copie d'écran - Page de sommaire d'une base documentaire

→ la page article avec accès au sommaire et à l'introduction de l'article, la suite étant réservée aux abonnés

Fig. 4 - Copie d'écran - Page de présentation d'un article

→ la page de résultats de recherche

Elle sera analysée dans le paragraphe suivant (4.4) consacré au moteur de recherche, avec une copie d'écran p. 46.

- **analyse**

L'évolution des technologies amène à être plus transversal et non figé dans une classification hiérarchique arbitraire. Les articles devraient pouvoir être sélectionnés et regroupés selon des critères indépendants de la classification, même si celle-ci ne peut pas être abandonnée puisque c'est la base des abonnements (les clients abonnés uniquement à certaines bases doivent pouvoir à tout moment visualiser ce qui est compris dans leur abonnement).

4.4 Fonctionnement du moteur de recherche Exalead

Afin de pouvoir ensuite donner de nouvelles orientations, il parait important de connaître le fonctionnement du moteur de recherche implanté sur le site des Techniques de l'Ingénieur.

4.4.1 Extraction et affichage des résultats

Exalead utilise un dictionnaire propre (non modifiable) pour la lemmatisation, une table de synonymes (acronymes) rajoutée en interne et un fichier de mots vides (interne). Les résultats sont affichés selon une pondération très complexe paramétrée dans le moteur de recherche.

- **lors d'une recherche simple**

Exalead effectue une recherche sur le texte intégral de l'article avec les mots ou expressions saisis par l'internaute et utilise :

→ la lemmatisation (prise en compte du mot et de ses dérivés selon le radical) (dictionnaire propre à Exalead)

→ l'approximation et la phonétique (selon un calcul propre à Exalead)

→ la liste de mots vides (fournie mais modifiable)

→ une liste de synonymes (actuellement seule une liste d'acronymes est utilisée)

- **lors d'une recherche avancée**

Exalead donne la possibilité de cibler la recherche :

- dans le titre
- avant ou après une date
- par auteur
- par thème
- par référence

Les réponses sont ensuite affichées par ordre de pertinence : les termes retenus pour la recherche sont affectés d'un poids (de 0 à 8) selon leur catégorie et leur emplacement. Les résultats de recherche sont actuellement très larges car ils prennent en compte toutes les variations possibles autour du mot recherché.

Le classement sera également fonction du nombre d'occurrence dans le texte, ce qui dépend du style des auteurs et n'est pas toujours le plus pertinent.

Les suggestions du moteur de recherche :

→ au niveau du masque de saisie : lors de la saisie des termes, une liste de propositions est donnée à partir des titres, auteurs et identifiants.

Remarque : le lien avec le mot demandé est souvent difficile à comprendre.

→ Essayez avec cette orthographe : Exalead propose parfois un terme d'orthographe voisine de celui saisi.

4.4.2 Affinage des résultats par filtres

Les filtres utilisent un marquage des documents au niveau de l'article.

Sur la page d'accueil, le moteur prend en compte les bases documentaires et les actualités d'Instantanés Techniques. Les filtres, affichés sur la partie gauche de l'écran (voir fig. 5 page suivante) reprennent donc l'ensemble de ces informations et proposent une sélection :

- par vue (comprendre&savoir, tendances&innovation, l'actu des secteurs…)
- par thèmes
- par espace du portail (base doc / IT)
- par ouvrages (titres des bases documentaires)
- par type d'information (essentiels, archives, innovation, vite s'informer, comprendre…)
- par abonnements (les clients peuvent choisir d'avoir dans la liste des résultats uniquement les articles auxquels ils sont abonnés)

Les filtres « Thèmes » et « Espace du portail » sont explicites mais les autres catégories sont plus difficiles à appréhender pour une personne qui ne connait pas le fonctionnement des Techniques de l'Ingénieur.

4.4.3 Proposition d'articles complémentaires

Sur la partie droite de l'écran (voir fig. 5 page suivante), le moteur de recherche propose un accès aux compléments bibliographiques ainsi qu'aux bases documentaires, rubriques ou cahiers d'instantanés techniques traitant du sujet.

Fig. 5 - Copie d'écran - Page de résultats de recherche

Le site Techniques de l'Ingénieur est donc particulièrement riche en informations et nous allons maintenant analyser, grâce notamment aux résultats d'une enquête clientèle, quels sont les besoins d'évolution qui ressortent.

5 Analyse des besoins

5.1 Evolutions prévues dans l'organisation et la commercialisation du fonds

5.1.1 Nouvelle segmentation du fonds

La collection se présente toujours sur le même modèle depuis sa création et a connu peu d'évolutions marquantes. Les processus de traitement du contenu, de la création d'un nouvel article à sa mise à jour, ont également très peu évolué, et sont principalement orientés autour de la diffusion du contenu sur support papier (avec son équivalent en fichier PDF). Les processus de diffusion sur le site Web héritent donc des contraintes liées au processus de diffusion papier.

Il y a donc un risque que l'image et la notoriété des Techniques de l'Ingénieur subissent l'érosion du temps faute d'innovation et de nouveauté (hormis de nouvelles bases documentaires). Face à la montée d'Internet et des contenus gratuits, il faut réussir à maintenir l'intérêt des clients. Une réflexion sur l'image des Editions Techniques de l'Ingénieur à travers son produit a ainsi mis en évidence la nécessité de faire évoluer ce dernier de manière importante, afin de redynamiser les ventes.

- **réorganisation de l'information**

La commercialisation par bases documentaires offre une granularité de classement des articles insuffisante pour refléter leur richesse et assurer un accès simple et rapide à l'information. L'objectif est de conserver les bases documentaires actuelles, mais de les subdiviser en autant d'unités élémentaires et indépendantes que nécessaire, chacune traitant d'un sujet particulier au travers des articles spécifiques à ce sujet. Ces unités sont appelées des « Sélections ». Il s'agit de réorganiser les bases documentaires et de les présenter de telle sorte que les sujets qu'elles traitent soient davantage mis en avant par rapport à l'existant. Ce découpage des soixante et une bases documentaires devrait conduire à la création d'environ quatre cent vingt sélections. A noter que l'utilisation d'un même article dans plusieurs sélections est possible et souhaitable, afin de donner tout le corps et l'autonomie nécessaire à chacune (on ne doit pas être obligé de systématiquement passer d'une sélection à une autre pour avoir les articles les plus pertinents sur un sujet).

Le découpage de la collection en Sélections, s'accompagnera d'une évolution du site Internet afin que les Sélections deviennent des produits à part entière, au même titre que les bases documentaires, consultables et achetables directement sur le site. Il devra parallèlement

conduire à un changement de format de l'encyclopédie papier. Autant le concept du classeur à feuillet mobile avec mise à jour régulière a fait ses preuves de nombreuses années dans le passé, autant celui-ci parait désormais dépassé par Internet et ses mises à jour en temps réel. Par ailleurs, les clients ont de moins en moins de temps à consacrer à la mise à jour de leurs classeurs. Les mises à jour non classées qui s'empilent renforcent le sentiment d'inutilité de l'abonnement et de produit dépassé faute de mise à jour.

- **avantages escomptés**

L'objectif général est donc de rénover le produit afin de le rendre plus en phase avec son temps, de meilleur qualité, plus interactif et ainsi susciter encore d'avantage d'intérêt auprès du client. L'accès à l'information sera plus évident (plus rapide, intuitif et visible) du fait de la segmentation.

Ceci augmentera également la réactivité pour aborder de nouveaux sujets. Contrairement à une nouvelle base documentaire qui nécessite de très nombreux articles, une sélection se contente d'une vingtaine d'articles. La capacité d'innovation éditoriale s'en trouve renforcée par des délais de rédaction plus courts, et donc une communication plus fréquente sur de nouveaux sujets, avec également davantage de mises à jour sur l'ensemble de la base.

Dans un proche avenir, l'idéal serait de proposer au client, visiteur du site, de construire sa propre sélection en piochant parmi les 6000 références d'articles et de lancer l'impression à la demande.

5.1.2 Intégration de services

Parallèlement à la restructuration du fonds, les Editions Techniques de l'Ingénieur souhaiteraient apporter de nouveaux services à leurs clients.

Nous avons vu dans la présentation des articles page 37 que l'utilisation des ouvrages se fait généralement dans le cadre d'une première approche, pour aborder un nouveau sujet ou avant de se lancer dans un nouveau projet de développement technique.

Les clients sont des personnes habituées à rechercher de l'information mais pas expertes du domaine. La recherche d'experts ou de partenaires scientifiques est donc l'étape suivante pour la mise en œuvre de leurs projets.

Il parait ainsi important d'apporter des services complémentaires permettant d'exploiter encore mieux les ressources d'information fournies, tels qu'une hotline avec des experts, un contact avec les auteurs (pour interagir ou rendre le site plus vivant) ou un accès aux spécialistes du sujet.

Ainsi, un partenariat est en cours d'étude avec la société Expernova qui a développé une véritable expertise dans la création de cartographies et de bases de compétences scientifiques grâce à des approches de datamining et d'analyse sémantique. Expernova cartographie et met à jour automatiquement chaque mois les compétences de plus de 15.000 centres de recherche européens.

Un autre objectif serait de réaliser des partenariats pour être diffusé plus largement car les recherches des internautes se font de plus en plus sur des packages qui interrogent de multiples éditeurs et bases de données. Rester isolé présente le risque d'être noyé parmi la multitude d'informations.

5.2 Présentation de l'enquête de satisfaction réalisée par un cabinet extérieur auprès de la clientèle

5.2.1 Contexte et objectifs de l'enquête

Les Editions Techniques de l'Ingénieur ont décidé de lancer une étude auprès de la société ECOFFENSIVE, spécialisée en études marketing, pour répondre aux questions suivantes :

- Comment les abonnés de « l'Encyclopédie » utilisent-ils ce produit aujourd'hui ?
- Quelle est leur évaluation du produit, de son fonctionnement et quelles évolutions souhaitent-ils ?

L'étude s'est déroulée en 2 phases :

→ Une première phase de pilotage par entretiens en face à face qui a permis à la fois de calibrer les entretiens et d'ajuster le guide de questionnement. Pour les Grands Comptes abonnés, différents niveaux d'interlocuteurs impliqués dans le produit ont été interrogés : un ou deux ingénieurs utilisateurs ainsi que la documentaliste en charge de la gestion des Techniques de l'Ingénieur (T.I.).

→ Une seconde phase de terrain par entretien téléphonique à laquelle a été associé l'outil internet, afin de commenter en live le site des Techniques de l'Ingénieur. Ces entretiens ont eu lieu sur rendez-vous et ont été enregistrés sous format numérique.

L'étude a été réalisée auprès de 27 ingénieurs ou documentalistes, clients ou anciens clients.

Les résultats repris ici ne concernent que la partie sur l'utilisation du site Web.

5.2.2 Présentation des clients et de leur utilisation du site Web

- Les grandes entreprises (disposant d'un service de documentation)

Le service documentation est en charge de l'abonnement et du produit, et il fait le lien avec les besoins des ingénieurs et techniciens (droit d'accès ou recherche pour le compte des

ingénieurs). Les ingénieurs consultent essentiellement en passant par l'accès Web. La navigation se fait via le moteur de recherche par mot-clé. Ils consultent pour conforter des connaissances fraichement acquises (jeunes) ou défricher un domaine dont ils ne sont pas spécialistes (projets transversaux).

- Les PMI (pas de service de documentation)

L'utilisation est faite directement par les intéressés, sur le Web ou par consultation papier, pour valider un choix technique, découvrir une technologie, répondre à un appel d'offre ou assurer la gestion transversale d'un projet.
Sur le site, ils recherchent également par mot-clé ou titre de l'article.

- Les sociétés de conseil, bureaux d'études et experts (pas de service documentation)

Ils utilisent majoritairement le Web, pour découvrir des disciplines, des questions techniques, des process ou des machines mais également en support pour rédiger leurs rapports (car l'expertise des auteurs T.I. est reconnue par leurs clients).

- Les écoles d'ingénieurs et collectivités (présence d'un service documentation)

T.I. est un référentiel de base pour les élèves ingénieurs et un support pour les cours et projets pratiques des professeurs.

Les besoins sont multiples pour les collectivités : législation, management, nouvelles technologies, approfondissement de sujets pour rédiger des appels d'offres.

5.2.3 Attentes des clients et problèmes rencontrés

Lorsque la connaissance du domaine est hésitante ou sur des sujets larges, la recherche par mot-clé est moins efficace que sur des sujets cernés et pointus. Dans ce cas, les utilisateurs déplorent un manque de visibilité des bases accessibles et de l'étendue des contenus disponibles : ils font difficilement le tour de la question.

Leur problème est de savoir s'ils ont bien accédé à toute l'information disponible. Parfois, ils ne trouvent pas du tout l'information recherchée et l'arborescence ne leur permet pas de vérifier si elle existe ou non.

La page d'accueil est trop chargée, ce qui ne permet pas l'identification rapide des chemins possibles. L'arborescence du site ainsi que le contenu des bases documentaires ne sont pas clairement perceptible, ce qui amène certains interviewés :

- soit à tâtonner un certain temps pour trouver enfin la bonne information,

- soit à abandonner purement et simplement la recherche, jugée alors trop longue et fastidieuse.

Les interviewés souhaiteraient avoir sur le site, en se connectant :

- une visibilité immédiate de tout ce à quoi l'entreprise est abonnée,
- l'accès à une arborescence permettant de vérifier si le contenu que l'on cherche se trouve dans la base documentaire à laquelle l'entreprise est abonnée.

Finalement, ils n'utilisent que les mots-clés sur le moteur de recherche car ils se retrouvent perdus dans la navigation, mais les mots-clés ne marchent pas toujours. Sur certains thèmes, le nombre d'articles proposés est très important et le tri s'avère long et fastidieux.

Ils souhaitent ou sont favorables à :

- une réorganisation des contenus,
- plus de lisibilité et d'ergonomie,
- une organisation plus logique des informations par rubriques,
- une meilleure visibilité du contenu et de l'abonnement,
- un regroupement autour de questions qui sont abordées dans différents articles,
- des rubriques qui fassent le tour d'une question,
- un sommaire qui permette de voir rapidement si l'information est présente,
- une offre à la carte avec un forfait annuel pour x articles ou rubriques autonomes.

Les résultats de l'enquête confirment donc la nécessité de réorganiser l'information et de clarifier l'accès au contenu de la collection.

5.3 Besoins et objectifs de l'entreprise

5.3.1 Clarifier le positionnement des articles

Lors de la navigation sur le site, l'unité de lecture est l'article. Les thèmes, bases documentaires, sélections, ne sont que des aides pour donner un contexte à l'article, au même titre que le fil d'Ariane.

L'enquête clientèle montre que ces repères ne sont pas clairs et qu'il est donc nécessaire de réfléchir aux moyens de donner ou montrer des liens ou une appartenance à des groupes d'articles, avec un environnement d'affichage pertinent.

Ceci est d'autant plus important que les articles sont multi-positionnés dans les bases documentaires et le seront encore plus dans les sélections.

5.3.2 Valoriser l'information

La collection représente un contenu riche et de qualité qu'il faut valoriser en améliorant son accessibilité.

- **objectifs**

→ fidéliser les clients en leur fournissant rapidement une information pertinente

→ amener l'internaute à découvrir la richesse du site pour qu'il y passe plus de temps et qu'il ait envie de revenir et d'accéder à des données payantes

Lorsque le client arrive sur une page du site, l'objectif commercial est :

→ pour les non abonnés,

- de renvoyer vers un essai gratuit ou une demande d'information,

- d'orienter vers les articles de la base ou de la sélection qui correspondent le mieux à l'utilisateur,

→ pour les abonnés,

- de générer du trafic, de l'intérêt vers les autres articles ainsi que vers les autres produits (autres sélections ou bases documentaires),

- de générer un partage d'information pour améliorer la visibilité de la marque et de rendre visible le bouquet de services liés.

- **démarche**

La valorisation du contenu peut se faire :

→ en optimisant les résultats de recherche

En partant du principe que les internautes ne consultent que les premières pages de résultats, il est important que les documents les plus représentatifs du mot-clé saisi sur le site arrivent bien les premiers.

→ en proposant des possibilités de sélection ou de regroupement d'articles sur des sujets précis, avec un accès rapide à l'information et un affichage clair des données

→ en proposant de nouveaux services pour donner plus d'attrait à l'information proposée

Nous allons maintenant étudier plus en détail les différents types de mots-clés à mettre en œuvre pour répondre à ces besoins.

6 Typologie des mots-clés potentiellement utilisables

L'étude est basée sur l'observation des sites du domaine scientifique, sur l'analyse de l'existant et prend en compte les différentes possibilités d'indexation et d'utilisation des mots-clés déterminées dans la première partie.

6.1 Panorama des sites Web dans le domaine scientifique et technique

Avant d'étudier en détail les possibilités d'indexation pour le site Techniques de l'Ingénieur, je suis allée à la découverte des sites proposant de l'information scientifique et technique afin de voir quels sont les types de recherche et de navigation proposés.

- **Access Science from McGraw-Hill**

Access Science est un site avec abonnement qui propose le contenu de la dixième édition de l'Encyclopédie de McGraw-Hill ainsi que les nouvelles tendances et développements en sciences et technologies, un dictionnaire des termes scientifiques et techniques ainsi que des biographies des chercheurs. <http://accessscience.com/>

Sur ce site, il est possible de faire une recherche simple avec autosuggestion (proposition de tous les termes de l'index commençant par les lettres tapées, qui se réajuste au fur et à mesure que l'on tape le mot recherché).

On peut également faire une recherche avancée avec possibilité de choisir :

- le type de document (articles, images, multimédia, biographies,...).
- le thème (agriculture, forestry & soils, anthropology & archeology,...).

L'index alphabétique des termes autosuggérés est également consultable.

Exemple d'extrait de liste :
-Nanochemistry
-Nanoelectronics
-Nanomaterials in the forest products industry
-Nanometer magnets
-Nanometrology
-Nanoparticles
-Nanoprint lithography
-Nanosatellites

- **INIST**

L'Institut de l'Information Scientifique et Technique (INIST) a pour but de faciliter l'accès aux résultats de la recherche mondiale. Il produit les bases de données bibliographiques, multilingues et multidisciplinaires, PASCAL et FRANCIS qui, avec 20 millions de références, recensent l'essentiel de la littérature scientifique internationale.

Les notices des documents sont accessibles sur <http://www.refdoc.fr/> selon titre, auteur, etc. On peut ensuite filtrer par langue, par format papier ou électronique et sélectionner uniquement les références avec résumé.

TermSciences est le portail terminologique de l'INIST sur lequel on accède aux concepts du thésaurus qu'ils utilisent. <http://www.termsciences.fr/>

Les différents portails de l'INIST (sur <http://www.inist.fr/>) permettent également d'accéder à un ensemble de ressources par :

- mots du titre,
- ordre alphabétique du titre,
- plateforme de diffusion,
- base de données
- disciplines scientifiques classées sur deux niveaux et donnant accès à la liste des revues, bases de données et ouvrages.

- **Science Direct Elsevier**

ScienceDirect est une base de données couvrant plus de deux mille cinq cent revues et plus de neuf millions d'articles en texte intégral de revues ou d'ouvrages en littérature scientifique. <http://www.sciencedirect.com/>

Le site propose en plus de la recherche directe sur tous les champs, sur l'auteur ou sur le titre (avec volume, numéro et page), une recherche avancée avec :

- sélection sur toutes les sources ou seulement sur les revues ou les livres,
- recherche sur des champs particuliers tels que résumé, mots-clés ou référence de l'article
- recherche par sujet :
  - Agricultural and Biological Sciences
  - Arts and Humanities
  - Biochemistry, Genetics and Molecular Biology
  - ...
- sélection sur la date (périodes)

La liste alphabétique des titres de revues et de livres est également disponible par lettre de l'alphabet.

La navigation par thèmes puis sous-thème renvoie ensuite sur les titres de revues ou d'ouvrages appartenant au thème choisi.

Exemple :

Physical Sciences and Engineering

Chemical Engineering
Chemistry
Computer Science
Earth and Planetary Sciences
Energy >Advanced Energy Conversion
>Advanced Well Completion Engineering
>...
Engineering
Materials Science
Mathematics
Physics and Astronomy

- **L'usine nouvelle et Industrie & Technologies**

Sites de la revue « L'usine nouvelle » spécialisée en information professionnelle avec l'actualité économique et industrielle ciblée par secteurs et de la revue « Industries & Technologies » ciblée actualités en R&D et nouveautés des produits de l'industrie. <http://www.usinenouvelle.com/> et <http://www.industrie.com/it/>

Le site L'usine nouvelle propose uniquement un menu par secteur :

Aéronautique
Agroalimentaire
Armement
Automobile
...

Le site Industries & Technologies propose d'autres moyens de navigation avec un accès :

- aux derniers articles
- aux articles les plus lus
- à l'expo permanente, avec un nuage de mots-clés présentant les catégories les plus consultées et une navigation par catégories sur deux niveaux.

- **Autres sites**

Sur les sites d'éditeurs consultés tels que **EDP Sciences**, on accède d'une manière générale à une recherche simple ainsi qu'à une recherche avancée sur les revues, les livres, les actes de conférence avec une sélection possible par date et des regroupements par thématiques. <http://publications.edpsciences.org/>

D'autres sites donnent une idée des catégories utilisées dans le domaine scientifique :

CNISF <http://www.cnisf.org/>
Le site du Conseil National des Ingénieurs et Scientifiques de France est organisé en menus. On trouve également un classement en comités sectoriels :

Aéronautique
Agroalimentaire
Défense
Eau
...

Le **portail de la science du Ministère de l'enseignement supérieur et de la recherche** donne un autre exemple de liste de thèmes <http://www.science.gouv.fr/> :

Archéologie
Astronomie / Aéronautique
Biologie
Biologie / Sciences du vivant
Biologie intégrative
Biologie moléculaire
...

Kompass <http://fr.kompass.com/>

Sur ce site, nous avons un exemple de classement d'entreprises par secteur d'activité :

Agroalimentaire
Industries d'extraction (mines, carrières, pétrole, gaz)
Bâtiment et génie civil
Informatique
Education, formation, R&D, services techniques (essais, analyses, sécurité industrielle)
Services environnementaux, gestion des déchets (électricité, chauffage urbain, gaz, eau)
...

Conclusion

On remarque sur l'ensemble des sites du domaine scientifique, la prédominance des catégories en support de navigation. Des index de termes ou de titres sont assez souvent proposés. On dispose fréquemment d'une recherche avancée avec possibilité de choisir des champs d'interrogation.

Il faut noter que l'étude des sites permet uniquement de voir les types de recherche, de navigation ou de filtres proposés côté utilisateur mais ne permet pas de découvrir le traitement mis en œuvre pour obtenir les résultats.

6.2 Analyse des données existantes

Avant de faire des propositions, une dernière étude doit être réalisée : celle des mots-clés utilisés actuellement, que ce soit ceux fournis par les auteurs ou ceux saisis par les utilisateurs ainsi que l'analyse des résultats de recherche sur le site Web.

6.2.1 Analyse des mots-clés auteur actuels

Les mots-clés donnés par les auteurs proviennent d'un nombre important de personnes différentes puisque la plupart des auteurs n'écrivent qu'un seul article pour Techniques de l'Ingénieur. Ils utilisent un vocabulaire libre, il n'y a pas de liste de référence, ni de consignes particulières en-dehors du nombre de termes à donner (six) et du fait que ces mots-clés vont servir à l'indexation par les moteurs de recherche. Ils n'ont jusqu'à ce jour pas été intégrés sur le site, la nouvelle application permettant de les intégrer vient juste d'être mise en place.

Il a été décidé de récupérer l'historique des mots-clés auteur stockés sur le serveur pour les deux dernières années. Environ 300 articles sont concernés par cette récupération.

L'analyse de ces données montre que les auteurs fournissent en moyenne 5,2 mots-clés par article. En regardant l'occurrence des mots, il apparait que :

- 81% des mots ne concernent qu'un seul article,
- 13% sont attribués à deux articles,
- seulement 6% se retrouvent sur 3 à 8 articles.

Les mots-clés sont donc majoritairement très ciblés sur le contenu spécifique de l'article (exemples : huile cristallisable, microscopie de fluorescence). Les termes que l'on retrouve sur plusieurs articles sont plus généraux (exemples : Sécurité, Environnement, Energie).

6.2.2 Analyse des logs sur les moteurs de recherche

- **moteurs de recherche donnant l'accès au site T.I.**

Le moteur dominant du marché étant Google à environ 90%, nous avons étudié les mots tapés sur Google qui amènent les visiteurs sur le site Technique de l'Ingénieur.

L'extraction porte sur la période allant de janvier à octobre 2009 et sur les 2.000 expressions apportant le plus de visites (ce qui correspond aux expressions ayant occasionné au moins 24 visites).

Nous avons classé les requêtes en deux catégories :

- expressions du type « techniques de l'ingénieur » qui sont tapées par les personnes connaissant le site et souhaitant y accéder directement : elles représentent de l'ordre de 39% des requêtes.
- autres requêtes sur des sujets qui amènent les internautes sur le site.

Requête	Nombre de visites	Pages par visite	Taux de rebond
Techniques de l'ingénieur	94.670	12,89	15,41%
Sujet divers	146.200	2,49	72,19%
Total	240.870	6,58	49,87%

Tab. 2 - Analyse des visites sur le site via Google

On remarque que le nombre de pages visualisées est nettement supérieur et le taux de rebond faible pour les utilisateurs ayant saisi « Techniques de l'Ingénieur » ce qui est normal puisque ce sont des personnes qui connaissent le site et qui souhaitent l'interroger.

Nous nous sommes ensuite intéressés uniquement aux requêtes sur des sujets pour voir quels types de recherche sont effectués :

Nombre de mots tapés	Nombre de visites	%
1	37.421	25,60
2	61.022	41,74
3	36.135	24,72
4	9.429	6,45
5	1.659	1,13
6	430	0,29
7	77	0,05
8	27	0,02
Moyenne : 4,5	Total : 146.200	Total : 100%

Tab. 3 - Analyse du nombre de visites en fonction du nombre de mots tapés sur Google

Il ressort de cette étude que les internautes utilisent à 92% entre un et trois mots pour leur requête, ce qui est en accord avec les informations recueillies en première partie.

Les requêtes les plus fréquentes comportent deux mots. En effet, l'utilisation d'un seul mot est souvent source d'imprécision et deux mots sont souvent nécessaires pour définir une notion dans son contexte (exemple le plus recherché : électricité bâtiment).

En analysant les requêtes comportant trois mots, on s'aperçoit que le plus souvent le troisième mot n'apporte pas de sens supplémentaire mais qu'il s'agit d'un mot de liaison. Les requêtes les plus fréquentes sont par exemple : gestion de production, contrôle non destructif ou électronique de puissance.

Les requêtes avec un nombre de mots supérieurs confirment qu'au-dessus de deux mots les internautes tapent leur requête dans un langage proche du langage naturel (exemple : lutter contre la pollution de l'eau).

Enfin les mono-termes utilisés sont souvent des technologies avec, par exemple, dans les mots les plus recherchés : lyophilisation ou chromage.

- **moteur de recherche interne au site**

L'extraction des mots-clés employés par les internautes sur le moteur de recherche du site Techniques de l'Ingénieur a été réalisée sur la période de janvier à août 2011.

Nous avons analysé les mille termes les plus recherchés.

Comme pour les recherches sur Google, il nous a semblé intéressant de voir combien de mots sont saisis par requête.

Nombre de mots saisis par requête	Nombre d'occurrences	%
1	67556	80,1
2	9933	11,8
3	5920	7,0
4	747	0,9
5	47	0,06
6	42	0,05
8	51	0,06

Tab. 4 - Répartition du nombre de mots-clés dans les requêtes les plus utilisées

La prédominance des unitermes est nettement plus importante sur le site que sur Google.

Exemple des requêtes les plus utilisées :

Unitermes : corrosion, adsorption, photovoltaïque, acier, soudage,

Deux mots : fibre optique, osmose inverse, smart grid, tension superficielle,

Trois mots : traitement de surface, mécanique des fluides, gestion de projet.

Cette extraction nous donne un bon aperçu des termes les plus utilisés. Très peu d'expressions longues ressortent car elles ont peu de chance d'être tapées à l'identique par plusieurs personnes et ne présenteront donc pas une occurrence élevée.

A noter également que les internautes utilisent toujours les sigles plutôt que la forme développée (exemple : ICPE).

L'analyse des cheminements sur le site a permis de voir que les filtres ne sont utilisés que pour 2,5% des requêtes.

6.2.3 Analyse de résultats de recherche sur le site Web

Nous avons sélectionné des termes représentant des notions tendances (sur lesquelles Techniques de l'Ingénieur souhaite montrer son positionnement et avoir une bonne image) et étudié les réponses obtenues via le moteur de recherche.

Nous obtenons du **silence** (des articles pertinents ne sont pas dans la liste de résultats) :

→ lorsque le terme employé a des synonymes ou d'autres façons de s'écrire,

→ lorsque le terme représente une notion globale qui n'est pas mentionnée dans l'article.

Exemple :

L'expression « chimie verte » donne 18 articles. Or, il existe une base documentaire intitulée « chimie verte » qui contient 43 articles qui devraient donc ressortir pour cette notion.

Nous remarquons que certains articles qui ne ressortent pas mentionnent « chimie durable » ou « chimie dite verte » ou « procédés respectueux de l'environnement » ou « méthodes respectueuses de l'environnement ». Nous trouvons également dans cette base documentaire des articles sur la production d'énergie renouvelable tels que les biocarburants pour lesquels la notion de chimie verte est absente du texte.

Nous obtenons du **bruit** :

→ lorsque le terme employé a d'autres significations.

Exemple :

Le terme photovoltaïque donne 278 réponses mais Exalead lui associe l'abréviation PV qui est à la fois utilisée pour photovoltaïque mais aussi pour le facteur pression de contact /

vitesse de glissement. Ce facteur, utilisé de nombreuses fois dans les articles qui traitent de cette notion, fait donc remonter les articles concernés dans les premières pages de résultats.

D'autres articles sortent car ils mentionnent bien le terme photovoltaïque mais ne traitent pas de cette notion de façon pertinente comme, par exemple, un article intitulé « Bois énergie ».

D'autre part, même en utilisant la recherche avancée, le nombre d'opérateurs est limité et il est difficile de sélectionner l'ensemble des réponses pour des expressions dont on souhaite avoir différentes possibilités d'écriture. On ne peut pas utiliser l'opérateur de proximité « AV » ni le caractère « ? » pour remplacer une lettre. La seule possibilité est le « NEXT » mais pour celui-ci les termes se suivent obligatoirement.

Exemple :

La seule solution pour obtenir l'ensemble des réponses correspondant à « distribution d'énergie », « distribution de l'énergie » et « distribution énergétique » est de saisir ces trois expressions avec des guillemets et avec l'opérateur OR entre chaque. Il est impossible de penser à chaque fois à toutes les variantes d'écriture et les réponses ne seront donc pas exhaustives sur le sujet.

A partir de toutes ces données, nous allons maintenant pouvoir proposer différentes indexations ou production de mots-clés répondant aux besoins déterminés précédemment.

6.3 Recensement des types de mots-clés envisageables selon les objectifs

- **réflexions préliminaires**

L'étendue des domaines traités par Techniques de l'Ingénieur fait écarter un travail sur une liste de termes contrôlés telle que le thésaurus, car le volume de termes à étudier entrainerait un travail considérable. Il serait nécessaire de récupérer des listes pour chacune des thématiques et ensuite de les adapter au vocabulaire utilisé dans les articles et par les utilisateurs, avec une amélioration des résultats qui ne justifierait pas les coûts générés.

La possibilité d'utiliser l'indexation automatique n'a également pas été retenue en raison du nombre d'articles restreint appartenant à un même domaine (en comparaison du volume d'informations brassées sur le Web, dans la presse par exemple). En effet, les traitements automatiques d'extraction de termes ou de catégorisation sont en général basés sur un apprentissage à partir d'un corpus qui est souvent lui-même composé de plusieurs milliers d'articles pour permettre ensuite un traitement statistique. Ce type de traitement ne peut

donc pas s'appliquer à notre cas. De plus, les traitements automatiques nécessitent quasiment toujours une intervention humaine à postériori et demandent donc un investissement aussi long que le traitement humain direct si les règles sont simples et souples. En effet, les premiers mots qui ressortent à l'extraction par les moteurs de recherche par exemple, sont principalement les mots des titres auxquels sont mêlés des mots non significatifs. Une personne peut facilement les repérer avec plus de pertinence.

Les usagers n'étant pas des personnes internes à l'entreprise, il n'est pas possible de les former à l'utilisation du site. Il faut donc leur donner quelque chose de très simple (accès rapide et facile) et qui respecte les usages, avec des moyens de cibler les résultats après une première recherche puisque les internautes font des requêtes très simples qui ne peuvent donc pas être suffisamment précises.

- **principes et méthodologie**

Les mots-clés doivent permettre de faire le lien entre le contenu de l'article et la recherche d'information.

Ils peuvent être à la fois descriptifs en cherchant à montrer l'ensemble du contenu (recensement des thèmes abordés de manière précise et exhaustive), sélectifs en cherchant à mettre en évidence les particularités de l'article, orientés utilisateurs en cherchant à utiliser les termes qu'ils emploient.

L'indexation multiple permet de prendre en compte plusieurs niveaux de précisions ou plusieurs angles d'approche pour la sélection d'articles.

Des propositions ont été formulées à partir des différentes possibilités d'utilisation des mots-clés découvertes dans les études préalables, en tenant compte de notre objectif de valoriser les articles, soit en proposant des suggestions de recherche ou de navigation sur le site, soit en donnant la possibilité de trier, extraire ou réorganiser le fonds selon des thématiques.

Elles ont ensuite été discutées lors de réunions avec les responsables éditoriaux et la responsable online, pour analyser la faisabilité et l'intérêt pour l'entreprise. Les personnes non présentes dans la discussion sont les utilisateurs, difficilement interrogeables de façon représentative.

6.3.1 Mots-clés de mise en valeur du contenu de l'article

La première catégorie de mots-clés à laquelle nous avons naturellement pensé en premier est la plus traditionnelle, à savoir les mots-clés représentant le contenu d'un article.

Ce sont les auteurs qui connaissent le mieux le texte, son environnement (contexte), le message à faire passer (ce qu'ils ont voulu montrer), ce sont des spécialistes du sujet. Ils sont donc normalement les mieux placés pour choisir des mots pertinents. De plus, cela leur prend peu de temps d'ajouter les mots-clés à leur texte avant de le transmettre tandis qu'un traitement par le personnel des Editions Techniques de l'Ingénieur serait plus long et entraînerait un coût supplémentaire.

L'objectif des mots-clés donnés par les auteurs actuellement est l'indexation pour la recherche (c'est ce qui est noté dans les consignes actuelles). Les auteurs ne disposent d'aucun autre repère pour le choix des termes.

Les principales questions soulevées sont :

→ leur vocabulaire correspond-il à celui utilisé en recherche ?

→ rencontre-t-on beaucoup de synonymie ? Dans ce cas, quel terme choisir ?

Une enquête très intéressante sur la production de mots-clés a été réalisée auprès d'auteurs par Sophie Assal, dont les résultats sont communiqués dans son mémoire.

> « *Pour construire les mots-clés, les auteurs rencontrés se posent un certain nombre de questions : faut-il mettre les mots du titre dans les mots-clés ? Doit-on donner des termes généraux, spécifiques, des expressions-clés ? De plus, le nombre de mots-clés demandés pose parfois un problème car, selon les auteurs interrogés, il faudrait pouvoir en donner plus que cinq ou six car ce n'est pas suffisant pour rendre bien compte du contenu d'un texte.* » ([1], Assal).

Plus on veut être précis et plus le nombre de mots-clés nécessaires sera important car il faut également mettre des mots-clés généraux permettant de situer le terme spécifique.

La répétition des mots du titre dans les mots-clés va dépendre des auteurs, il peut donc être intéressant de donner des consignes également à ce niveau.

Il apparaît clairement que les mots-clés auteur présentent un intérêt mais qu'ils doivent être encadrés pour avoir une certaine homogénéité entre les articles et pouvoir les exploiter.

La question est également de savoir jusqu'où il est possible d'aller dans notre demande : peut-on leur faire scinder les mots-clés en plusieurs catégories (domaine général, thèmes spécifiques, …) sans trop compliquer la tâche ? Pour cela, doit-on mettre à leur disposition des listes de vocabulaire contrôlé en plus des consignes ? Doit-on effectuer un contrôle en interne à réception ? Dans ce cas, qui le fera ?

6.3.2 Mots-clés de regroupement en catégories

Comme nous l'avons remarqué dans l'étude des autres sites, le regroupement des articles en catégories est très fréquent.

Nous avons fait ressortir deux besoins potentiels concernant la catégorisation :

- Etablir une classification intermédiaire entre thèmes et bases documentaires

Actuellement, le filtre par thème est trop large pour sélectionner un nombre restreint d'articles et le choix de bases documentaires est difficile car leur titre n'est pas toujours explicite. Par exemple, il existe un thème « Environnement – Sécurité », à l'intérieur duquel on trouve une base documentaire intitulée « Environnement » et une autre intitulée « Innovations – Environnement ».

- Etablir une classification complètement différente

L'utilisation de nouvelles classifications permettrait d'accéder de manière différente aux articles. Elles pourraient faire l'objet de nouveaux filtres ou paramètres de recherche avancée.

Les différents types de catégories que nous avons déterminés sont les suivants :

→ Approche par discipline (ex : INIST, ScienceDirect)
→ Approche par secteur industriel (ex : Kompass)
→ Approche par centre d'intérêt (ex : qualité / sécurité / environnement / droit / économie / matériau / équipement)
→ Approche par process (ex : recherche et innovation / conception / production / contrôle / distribution)
→ Approche par structure d'emploi (ex : entreprise industrielle / bureau d'études / enseignement)
→ Approche par fonction (ex : directeur industriel / architecte / informaticien / chimiste)
→ Approche par type de contenu (ex : théorique / appliqué / descriptif / réglementaire / statistique)
→ Approche par type d'information (ex : introduction / résumé / tableau / bibliographie)

Ces différentes approches sont à étudier par rapport au contenu de la collection Techniques de l'Ingénieur et à leur pertinence par rapport aux demandes des utilisateurs.
Chaque catégorisation est un angle d'approche parmi d'autres ; il est très difficile de déterminer s'il existe un type de cheminement majoritairement utilisé par les internautes, chaque personne ayant une logique qui lui est propre.

Remarques :

→ les choix proposés par catégories doivent permettre de couvrir une majorité d'articles,

→ le rattachement à une catégorie peut être obligatoire ou facultatif,

→ la catégorisation peut s'appliquer à un article, une partie d'article ou une sélection d'articles.

Les listes de catégories peuvent être d'un seul niveau ou sous forme d'arborescence hiérarchique à plusieurs niveaux (maximum 2 à 3 niveaux conseillés). Elles correspondent à une liste restreinte de termes car elles deviennent inutilisables lorsqu'elles sont trop complexes, elles sont donc assez générales.

Elles peuvent être utilisées en filtre (affinage d'une première recherche) ou en recherche avancée (guide). Pour ces utilisations et l'affichage sur le site, l'arborescence paraît difficilement utilisable.

Au final, la collection Techniques de l'Ingénieur étant par principe très homogène dans la structuration des articles et la façon d'aborder le sujet, la seule catégorisation qui nous semble refléter la diversité de la collection est celle reposant sur le secteur d'activité ou domaine d'étude.

⇨ Une première liste a ainsi été proposée, qui serait plus précise que la liste des thèmes actuels et reprend l'ensemble des domaines traités par Techniques de l'Ingénieur.

AGRO-ALIMENTAIRE
ANALYSES / ESSAIS / METROLOGIE
AUTOMATIQUE / ROBOTIQUE
BATIMENT / GENIE CIVIL
BIOTECHNOLOGIES
CHIMIE
COMPOSITES
ELECTRICITE
ELECTRONIQUE
EMBALLAGE / CONDITIONNEMENT
ENERGIE (hors électricité, nucléaire)
ENVIRONNEMENT
INFORMATIQUE
MATERIAUX (hors métaux, plastiques, composites)
MATHEMATIQUES / MODELISATION
MECANIQUE
METAUX
NANOTECHNOLOGIES
NUCLEAIRE
OPTIQUE
PHYSIQUE
PLASTIQUES
RESEAUX – TELECOMS
TRAITEMENT DU SIGNAL
TRANSPORTS
DIVERS
TOUS SECTEURS

Remarques :

Pour Energie et Matériaux, exclure certaines thématiques peut entraîner des problèmes pour sélectionner l'ensemble du domaine dans le cas des filtres.

Avantages :

→ La liste est simple et facile à mettre en œuvre.

→ Chaque article peut appartenir à autant de catégories que nécessaire, ce qui donne une meilleure représentation du contenu de l'article.

Inconvénients :

→ Pour une utilisation pertinente, il est nécessaire de catégoriser l'ensemble des articles, ce qui demande une charge de travail très importante.

→ Il y a un risque de confusion entre cette liste et les thèmes actuels qui continueront à être utilisés dans la classification principale des articles.

→ Les domaines retenus sont peut-être encore trop larges pour être vraiment utiles.

La pertinence du contenu d'un document par rapport à la requête fait aussi intervenir l'état des connaissances de l'utilisateur sur le sujet de sa recherche. Il pourrait donc être intéressant de tagger les articles également en fonction de leur type, tel que généraliste ou pointu.

⇨ D'autres listes de catégories ont ainsi également été proposées :

- selon le type d'article

MISE EN ŒUVRE
PANORAMA
REGLEMENTATION

- correspondant à une fonction dans les entreprises

PRODUCTION
RECHERCHE / DEVELOPPEMENT
QUALITE

Celles-ci ne seraient « cochées » que pour les articles qui présentent une particularité par rapport à l'ensemble de la collection. Or, les catégories utilisées en filtre doivent couvrir l'ensemble. De plus, la multiplication des possibilités de filtres entrainerait une nouvelle surcharge à l'affichage de la page de résultat. Il faudrait donc voir s'il est possible de les utiliser d'une autre manière.

6.3.3 Mots-clés de mise en valeur de notions innovantes

Nous avons déjà souligné l'importance de pouvoir repérer tous les articles répondant à un sujet (question récurrente des abonnés : « a-t-on vraiment fait le tour de la question ? »).

L'idée serait de regrouper les articles dans des ensembles différents des sélections. En effet, celles-ci correspondent à une thématique précise mais l'appartenance à un groupe répond aussi à des objectifs marketings et commerciaux. Tous les sujets ne font pas l'objet d'une sélection. La création de nouveaux ensembles apporterait donc une amélioration certaine.

Par exemple, la recherche portant sur « énergie NEXT solaire » donne 84 résultats répartis dans les sélections suivantes :

- Thermique et conditionnement de l'air dans le bâtiment 10
- Systèmes électriques pour énergies renouvelables 8
- Sources d'énergie hors nucléaire 7
- Environnement et construction 6

Les articles concernés sont donc dispatchés dans plusieurs sélections où l'on trouve également des articles qui n'ont pas de lien avec l'énergie solaire. En apposant un tag « énergie solaire » par exemple sur les dix articles qui traitent le sujet le plus en profondeur et selon différents aspects (choix effectué par le responsable éditorial concerné), on pourrait les faire ressortir des autres résultats et proposer à l'internaute une première approche du sujet plus accessible. Libre à lui d'aller ensuite voir d'autres articles si ceux-là ne lui suffisent pas.

La possibilité d'attribuer des mots-clés aux articles est une façon simple et très évolutive de les faire ressortir sur des thèmes précis. Les mots évoluent avec les technologies.

Il faut arriver à exposer les ressources récentes, les sujets tendances qui ne ressortiraient pas de l'ensemble du fonds autrement.

Etant donné l'orientation de la collection, utilisée majoritairement en recherche et développement, choisir les mots-clés parmi les nouvelles tendances de l'industrie et des technologies permettrait une mise en valeur visible et appréciée de notre clientèle.

L'objectif est dans ce cas de se placer sur les mots-clés les plus utilisés pour les technologies innovantes, en suivant régulièrement les évolutions. Il ne s'agit plus de traiter toute la collection mais de cibler les articles à promouvoir.

Pour avoir un bon aperçu des tendances, on peut se référer au ministère de l'industrie qui donne la liste des Technologies clés. <http://www.industrie.gouv.fr/tc2015/index.php> [Mise en ligne le 23 mars 2011]

> « *L'étude technologies clés 2015 a pour objectif d'identifier des segments stratégiques de notre économie et de mener une analyse des forces et faiblesses du développement de ces technologies en France. […]*
>
> *La quatrième édition de l'étude de prospective technologique « Technologies clés 2015 » présente 85 technologies clés qui trouvent leurs applications dans sept secteurs économiques. Les organismes les plus pertinents ont été associés dans chacun des domaines investigués pour faire de l'étude Technologies clés 2015 une analyse stratégique et un outil structurant.* »

Les mots-clés à retenir peuvent également provenir des extractions faites sur les moteurs de recherche car il est également important de présenter des résultats clairs sur les termes les plus demandés par les internautes. Les termes pourront également être testés sur Google adwords (voir p. 30) <https://adwords.google.fr/select/KeywordToolExternal>.

Les éditeurs joueront alors le rôle d'internautes qui taggent les articles à posteriori pour apporter leur point de vue et faire ressortir des notions un peu masquées dans l'ensemble du fonds documentaire.

L'avantage par rapport aux tags apposés par les internautes est une réflexion sur le vocabulaire employé même si celui-ci est libre.

6.3.4 Mots-clés permettant le lien à des services

Dans le cadre de la mise en place de nouveaux services proposés aux clients, un partenariat est en-cours avec la société Expernova, dont la vocation est de présenter des experts (majoritairement des chercheurs du secteur public mais également du privé). Pour un sujet donné, Expernova propose les centres de recherche, chercheurs, publications, projets, brevets par pays et suggestions d'autres mots-clés liés au sujet. Leur base de données est mise à jour chaque mois à partir des archives libres, universités, bases institutionnelles et agrégateurs sur l'Europe. Une base sur l'Asie est également en création. Leurs clients sont des groupes industriels et des cabinets de conseil, leur cible est donc sensiblement la même que celle des Techniques de l'Ingénieur.

Deux niveaux d'intégration d'applications Expernova sont prévus sur le site Techniques de l'Ingénieur :

- **Niveau 1 : intégration d'un widget**

Un widget sur les pages articles du site permettra un lien à expernova.com. A partir d'un ou plusieurs concepts scientifiques en anglais, l'application renverra des statistiques issues de la base d'expernova.com. C'est une incitation à demander plus d'informations.

Pour que cela fonctionne, Expernova a besoin des mots-clés en anglais pour chaque document sur lequel on veut poser le widget.

La pertinence des mots-clés choisis n'est pas forcément un problème à ce niveau d'intégration, car ce n'est pas le détail des résultats de recherche qui est affiché mais uniquement des résultats chiffrés (nombre de centres d'étude, nombre d'experts,...). Il faut cependant tester pour vérifier que l'expression employée existe et fournit un nombre de réponses cohérent.

Fig. 6 - Copie d'écran - Résultat d'une requête sur Expernova

- **Niveau 2 : intégration d'un service Expernova restreint**

Ce service sera accessible uniquement pour les abonnés. Au niveau du contenu, l'internaute aura accès à une partie du service Expernova. Le service doit cependant rester restreint avec renvoi vers Expernova pour obtenir des informations complètes.

A ce stade, les mots-clés choisis doivent fournir des résultats liés à la thématique générale de l'article pour montrer aux clients la pertinence du service.

6.4 Evaluation des mots-clés

Comme nous l'avons déjà souligné, l'indexation va dépendre de nombreux paramètres et doit en particulier prendre en compte le contexte de recherche. Que recherchent les utilisateurs : plutôt une information pertinente ou plutôt l'exhaustivité ? Connaissent-ils le domaine ? Existe-t-il une terminologie spécialisée ? Quelle granularité faut-il employer ?

6.4.1 Définition des types d'évaluation

La qualité de l'indexation se mesure par l'utilité des mots-clés attribués à un document pour la recherche d'information.

« L'évaluation de l'indexation doit permettre de déterminer :

→ *la précision : les mots-clés utilisés correspondent-ils au thème traité ?*

→ *le rappel : l'ensemble des thèmes traités sont-ils traduits par un mot-clé ?*

→ *la consistance : correspondance entre thèmes et mots-clés dans le temps (intra-indexeur) et selon les personnes qui indexent (inter-indexeur) ?* » ([11], Névéol).

La difficulté de l'évaluation vient du fait qu'il n'y a pas une seule indexation valable mais plusieurs possibilités donnant de bons résultats.

Les mots-clés attribués à un document vont varier d'un indexeur à un autre selon ses connaissances mais également pour un même indexeur qui va évoluer dans sa manière d'indexer en fonction de son expérience passée ([11], Névéol).

On ne va donc pas juger sur les mots utilisés mais sur les résultats de recherches effectuées sur les documents indexés.

Les définitions des mesures à effectuer deviennent :

Précision = nombre de documents fournis pertinents / nombre total de documents fournis

Rappel = nombre de documents fournis pertinents / nombre total de documents pertinents

Un autre problème surgit alors qui est de vérifier la pertinence d'un document par rapport à une recherche. De nouveau nous sommes face à un jugement humain qui va dépendre de la façon dont la personne perçoit le sujet.

Il est donc difficile de déterminer objectivement quel est le nombre total de documents pertinents pour une recherche. Parfois, un seul document peut suffire pour satisfaire le besoin de l'utilisateur.

On trouve également les notions fréquemment utilisées de bruit et de silence, qui posent exactement les mêmes problèmes de mesure.

Bruit = nombre de documents fournis non pertinents / nombre total de documents fournis

Silence = nombre de documents non fournis et pertinents / nombre total de documents pertinents

Ce type de mesure, pour permettre d'évaluer la qualité de l'indexation, doit être réalisé avec un système de recherche qui n'évolue pas entre les différentes mesures afin de ne pas influencer les résultats de recherche. Une méthode de recherche sera définie et utilisée pour toutes les mesures : par exemple, on saisit le mot-clé dans le moteur de recherche et on étudie les résultats.

La performance est mesurée par rapport à la capacité du système à retrouver les documents qui correspondent à la requête.

Une autre possibilité, pour laquelle ne se posera plus le problème de connaître le nombre total de documents pertinents, est de demander à un utilisateur de retrouver un nombre de documents pertinents sur un temps donné en utilisant les moyens qu'il souhaite. Par exemple, il devra retrouver les dix meilleurs documents possibles en cinq minutes.

Dans ce cas, il faut noter le cheminement de chaque utilisateur pour pouvoir analyser et comparer les différents résultats obtenus. L'étude devient complexe et les résultats difficiles à interpréter.

Les limites de l'évaluation viennent du fait que, lors de celle-ci, le besoin d'information est bien cerné alors que c'est souvent beaucoup plus flou lors des recherches réelles des utilisateurs ([21], Fluhr), ([18], Chaudiron).

6.4.2 Application aux mots-clés définis dans notre étude

Les résultats de recherche sur le site des Techniques de l'Ingénieur dépendent avant tout de la recherche en texte intégral du moteur de recherche Exalead.

Des tests pourront être réalisés seulement à partir du moment où les différents mots-clés seront intégrés et le paramétrage du moteur de recherche modifié.

Les modifications de paramétrage devront permettre de prendre en compte avec un coefficient plus ou moins important les différents types de mots-clés, leur attribuant ainsi une pondération qui influencera le classement des résultats de recherche.

Nous allons donc seulement donner une marche à suivre pour le futur mais nous n'avons pas pu tester l'efficacité des préconisations.

Les tests seront réalisés avec recherche sur l'intégralité du fonds documentaire car celui-ci est assez stable (peu d'articles nouveaux sont introduits par an sur un même sujet).

Les tests sur les mots-clés auteur seront difficiles à mettre en œuvre car le pourcentage d'articles traités sera faible les premières années. L'amélioration ne pourra se voir qu'à long terme. Au début, la seule possibilité sera de regarder si les consignes sont respectées.

Les mots-clés de catégorie servent à filtrer mais ne peuvent pas être testés sur les résultats de recherche.

Pour les mots-clés de mise en valeur, il faudra surtout définir comment ils seront exploités au niveau du site mais ils ne seront pas testés par rapport à une pertinence puisque celle-ci

sera définie arbitrairement lors du rattachement. Des tests peuvent cependant être réalisés sur le site pour les termes les plus interrogés ou sur les termes innovants pour permettre d'évaluer le besoin de tagger les articles sur ces thèmes.

L'évaluation par rapport aux critères de pertinence des résultats de recherche est donc difficilement applicable pour les types de mots-clés que nous avons définis. Il parait plus adéquat d'évaluer ceux-ci par rapport à des critères de performance.

- **Choix d'indicateurs de performance**

Un indicateur de performance doit être informatif, fiable, fidèle, adéquat, applicable, comparable. Il faut également définir la fréquence de relevé : faire un relevé par exemple trimestriel ou annuel.

Les indicateurs de performance que nous avons pu déterminer dans notre étude sont les suivants :

→ **Indicateurs d'alimentation**

Il faudra étudier à l'usage :

- si l'alimentation est bien réalisée en temps voulu,
- si les consignes sont respectées,
- si les opérateurs éprouvent des difficultés,
- si des mises à jour sont effectuées.

→ **Indicateurs de perception par la clientèle**

A évaluer par enquêtes auprès des clients :

- est-ce que l'image du site en clientèle évolue ?
- quelle est la satisfaction des clients par rapport à leurs résultats de recherche ?
- est-ce qu'ils utilisent les nouvelles fonctionnalités ?

→ **Indicateurs de résultats et de productivité**

- temps passé à l'alimentation des mots-clés
- nombre de documents traités
- coût des développements informatiques engendrés par la mise en place

Un tableau de bord pourra être mis en place, permettant de donner les axes d'amélioration et de planifier des actions de progrès.

Les différents types de mots-clés qui nous paraissent utilisables et source d'amélioration ayant été listés, nous allons maintenant donner les préconisations de mise en œuvre pour chacun d'entre eux.

Troisième partie
Préconisations pour la création et la mise en place des différents types de mots-clés

7 Explications sur la mise en œuvre du projet

Avant de détailler les particularités de chaque type de mot-clé proposé, nous avons cherché à définir quelques consignes communes à tous les développements.

7.1 Intégration dans le CMS eZ Publish

L'intégration des mots-clés peut se faire soit au niveau de la DTD soit au niveau du CMS eZ Publish. Nous avons donc regardé quelles sont les implications dans chaque cas.

7.1.1 Métadonnées et CMS

Philippe Lahaye décrit parfaitement les fonctionnalités des CMS dans son mémoire qui m'a servi pour la rédaction des paragraphes suivants ([31], Lahaye).

- **alimentation**

Les métadonnées sont soit partie intégrante d'un document, soit à l'extérieur du document. Il est important de définir comment gérer le processus de description de contenu par rapport au contenu proprement dit. En effet, il peut s'agir de tâches gérées par une seule personne ou par plusieurs, avec des droits donnés à des groupes de travail auxquels on affecte un rôle lié à une tâche.

Il faut également se poser la question des possibilités d'automatisation pour l'intégration des métadonnées, question indispensable par rapport à la productivité. En effet la principale critique adressée aux CMS est la lourdeur de la saisie et du renseignement des métadonnées. La saisie des métadonnées se fait généralement de manière traditionnelle avec l'aide d'un formulaire.

- **échanges de données**

L'échange de données informatisées (EDI) concerne la capacité du système à importer des ressources ou à en exporter. Il faut prendre en compte les possibilités d'intégrer des métadonnées déjà existantes dans d'autres systèmes et de transmettre les métadonnées saisies pour des utilisations externes ou un changement de système de gestion de contenu.

Même lorsque les formats d'échange sont harmonisés, il faut veiller à l'harmonisation des modèles de chaque métadonnée, certaines devant être éditées spécifiquement.

- **utilisation par le moteur de recherche**

Afin d'exploiter au mieux le contenu des métadonnées, celles-ci doivent être renseignées avec des termes non ambigus et des valeurs cohérentes.

Une recherche effectuée sur des champs de métadonnées correspond à une requête paramétrée, qui est bien plus efficace qu'une recherche plein texte sur toutes les valeurs possibles de toutes les propriétés d'un document. Encore faut-il que les documents soient systématiquement indexés et référencés selon une procédure générale, commune et respectée par tous.

Des règles peuvent être appliquées aux métadonnées, associant un document contenant une liste de mots-clés pour le décrire aux autres documents partageant par exemple les mêmes mots clés. Elles peuvent également indiquer à l'utilisateur la dépendance d'une ressource à une autre.

L'utilisateur peut faire le choix non seulement d'une métadonnée pour ordonner sa navigation, mais aussi d'une valeur (ou d'un domaine de valeur) de la métadonnée pour filtrer sa recherche d'informations.

Ces facilités de navigation (classification personnalisée, fenêtre de liens complémentaires, exploration dynamique) sont d'autant plus importantes qu'elles peuvent être complétées de métadonnées propres à l'utilisateur. Potentiellement, cela lui permet de réorganiser et classifier le contenu pour répondre à ses besoins et ses exigences et de partager ces nouvelles organisations et classifications avec d'autres ([31], Lahaye).

7.1.2 Application aux Editions Techniques de l'Ingénieur

Un espace mots-clés est réservé dans la DTD, qui pourrait être attribué aux nouveaux termes créés. Un premier problème viendrait du nombre de catégories de mots-clés que nous avons décidé de mettre en place. Chaque type de mot-clé devrait avoir un espace défini. Il faudrait donc modifier la DTD qui ne prévoit pas les emplacements nécessaires. C'est une procédure qui ne se fait que rarement et est assez lourde à mettre en place. De plus, l'alimentation se fait au moment de l'intégration de l'article. Or, certains types de mots-clés seront attribués à posteriori. Il est important que le pôle éditorial puisse les modifier et en intégrer de nouveaux quand ils le souhaitent. Ceci serait difficile à faire si les mots-clés étaient intégrés dans la DTD.

La décision a donc été prise d'intégrer les nouveaux mots-clés via le CMS eZ Publish et un premier développement a déjà été réalisé pour intégrer le résumé et les mots-clés auteur en français et en anglais. Les développements à venir porteront sur la création d'espaces distincts pour les différents types de mots-clés définis dans cette étude.

Les avantages portent sur la possibilité d'intégrer massivement des données à partir de tableaux Excel pour reprendre un historique ou alimenter une grande quantité de contenus, ce qui permet de travailler de façon beaucoup plus souple pour les personnes qui vont définir les mots-clés. Elles pourront renseigner progressivement leur tableau de données et l'intégration se fera ensuite en une seule fois.

L'autre avantage est la possibilité de donner l'accès à de multiples personnes qui peuvent alimenter ponctuellement ou modifier les données, ce qui entraîne une beaucoup plus grande réactivité, nécessaire pour s'adapter aux évolutions de contenu et d'objectifs.

7.2 Points à prendre en compte pour les évolutions

7.2.1 Clarté et simplicité

Sur les sites Web, il est important de toujours penser à la cohérence des données publiées et de respecter une logique dans leur affichage ou leur mise à disposition pour les utilisateurs.

Cette logique sera conservée sur les différentes pages du site afin de faciliter la mémorisation et l'identification, de façon à ne pas désorienter l'internaute.

Afin que le site soit clair et les informations visibles, il est conseillé d'éviter les surcharges cognitives. En effet, tous les outils mis à disposition, même très performants, ne seront pas utilisés s'ils sont perdus au milieu d'une multitude d'informations.

Le site étant déjà très chargé, une attention particulière sera portée sur le positionnement des nouveautés et la possibilité de proposer de nouvelles navigations sans perdre l'utilisateur.

7.2.2 Souplesse d'utilisation et de mise à jour

Lorsque les outils sont mis à la disposition des utilisateurs finaux, ils doivent être le plus simple, le plus explicite et le plus intuitif possible car les internautes ne lisent pas les explications éventuelles d'utilisation, surtout si celles-ci sont longues ou complexes.

L'alimentation doit également être simple à mettre en œuvre pour être bien acceptée des opérateurs et ne pas être abandonnée au fil du temps.

Le recours à l'intégration automatique doit se faire dans tous les cas où cela est possible, l'intervention manuelle étant à prévoir uniquement s'il n'y a pas d'autre solution.

Enfin, les solutions développées seront adaptables. Les principales évolutions à prendre en compte portent sur le contenu, avec la création de nouvelles bases documentaires, couvrant des domaines scientifiques plus vastes et sur les types de mots-clés utilisés qui peuvent être amenés à changer, avec l'introduction de nouvelles catégorisations par exemple.

7.2.3 Rapport coût / efficacité

Les solutions doivent tenir compte du temps nécessaire à leur mise en place, à leur alimentation ultérieure ainsi qu'aux développements informatiques nécessaires. Les décisions d'implantation seront prises par rapport aux améliorations apportées et à l'impact au niveau clientèle.

Les évaluations doivent porter sur l'adéquation du projet avec l'utilisation qui en sera faite et la satisfaction qui en découlera pour les prospects et les clients.

La performance des nouveautés apportées sera difficilement quantifiable du point de vue usager. Certains indicateurs de performance pourront être mis en place (voir page 72).

7.2.4 Personnel impliqué et gouvernance

La mise en œuvre du projet doit être progressive afin que le personnel de l'entreprise puisse avoir le temps d'intégrer les changements. Les acteurs impliqués dans la création et l'alimentation des mots-clés doivent recevoir suffisamment d'informations pour que leur tâche soit claire. Il leur sera précisé si la mise en place se fait sur les nouveaux articles ou sur le corpus existant, et dans ce cas quel sera l'historique traité.

Les personnes qui vont superviser la mise en place puis faire les tests ou audits de fonctionnement des nouvelles applications seront nommées et un planning sera communiqué à l'ensemble des personnes concernées.

8 Développements proposés

Les différents types de mots-clés retenus ont été classés en fonction de leur possibilité d'intégration rapide ou de la nécessité d'un traitement plus conséquent et donc d'une mise en place à plus long terme.

Afin d'aider à la prise de décision, les spécifications suivantes ont été étudiées pour chaque type de mot-clé :

- Caractéristiques
- Production et exploitation
- Contrôles et évolutions

Dans tous les cas de figure, le préalable à toute implantation est de définir les espaces concernés dans eZ Publish. Dans la suite du mémoire, nous partirons du principe que ces zones sont créées et nous donnerons seulement les préconisations spécifiques à chaque mise en place.

A court terme

8.1 Nouvelles consignes aux auteurs et exploitation de leurs mots-clés

Des mots-clés sont déjà donnés par les auteurs. Leur exploitation doit donc pouvoir se faire assez rapidement.

8.1.1 Caractéristiques

Lorsque l'utilisateur se trouve sur un article, les mots-clés donnés par les auteurs permettent la relance de la recherche et donc son affinage ou son élargissement selon les cas.

Ils sont libres et représentatifs du **contenu** de l'article. Leur nombre ne doit pas être trop élevé sous peine de bruit et pas trop faible sous peine de silence. Ils doivent réellement correspondre à des notions développées dans l'article ou demandées en consigne.

La distinction entre sujet principal et sujet secondaire est difficile à exploiter lors de la recherche. Nous laissons donc l'auteur juger de l'intérêt de reprendre ou non toutes les notions traitées dans l'article.

La question se pose également de reprendre ou non les mots du titre lorsqu'ils correspondent à un concept clé ([15], Waller). Afin de faciliter la tâche de l'auteur et ne pas

le noyer sous des consignes trop longues, nous avons décidé de le laisser libre dans son choix, sachant que les titres sont de toute façon indexés par le moteur de recherche mais que la présence des mots dans l'indexation donne un poids supplémentaire le cas échéant.

Les conseils que l'on peut donner sont de choisir des noms plutôt que des adjectifs ou adverbes, de les écrire au masculin et au singulier si possible, de privilégier les sigles et de ne pas hésiter à employer des expressions lorsqu'elles sont plus explicites que des unitermes.

Après réflexion, nous avons choisi de demander aux auteurs de scinder les mots-clés donnés en différentes catégories afin de pouvoir les exploiter ensuite pour différentes utilisations.

Les consignes données sont les suivantes :

6 mots-clés en français, au moins, doivent être indiqués immédiatement après le résumé ainsi que leur traduction en anglais. Ces mots-clés serviront à l'indexation par les moteurs de recherche.

Ils sont scindés en 3 catégories :

→ **Domaines technologiques** (technologies ou matériaux étudiés, sujet général de l'article) (1 à 2 mots-clés)
(ex : spectrométrie, optoélectronique, biomatériaux, thermoplastiques, traçabilité, éolien, systèmes embarqués, toxicologie, photovoltaïque)

→ **Secteurs d'activités** (à indiquer lorsque le sujet traité s'applique à des secteurs précis) (0 à 2 mots-clés)
(ex : logistique, transport, bâtiment, environnement)

→ **Caractéristiques du contenu** (2 à 4 mots-clés)

- type d'article si celui-ci traite le sujet avec un angle particulier (ex : panorama, réglementation, mise en œuvre)
- contenu ciblé de l'article (ex : époxydation, flux laminaire, ISO 17025, nanocornet, piézomètre, turbine à gaz, gradient thermique, polyphénol, CO2)

Séparer les domaines technologiques et les secteurs d'activité pourra permettre ensuite de voir s'il est possible de les utiliser pour faire des tris et les généraliser à l'ensemble du fonds documentaire.

8.1.2 Processus de production

- **Création Auteurs**

Les auteurs reçoivent des consignes pour la rédaction des articles. La partie réservée à l'attribution de mots-clés doit donc être modifiée afin qu'ils intègrent les nouveaux axes que

nous souhaitons leur donner. Ils recevront ainsi les instructions dans les consignes de rédaction et la feuille de style dont ils disposent systématiquement.
Les auteurs transmettent déjà des mots-clés avec leur article (sous Word), il n'y a donc pas de modification de la procédure mais seulement une définition plus précise des termes à attribuer.

Etant donné la demande de répartition des mots-clés en plusieurs « zones », un regard interne semble nécessaire pour vérifier que les consignes ont été respectées et effectuer éventuellement des modifications avant l'intégration dans le système.

- **Intégration informatique**

o Pour les nouveaux articles

Les mots-clés seront intégrés au cours du processus de fabrication, en même temps que les résumés pour lesquels un nouvel espace a également été prévu.

o Pour l'historique

Il est possible de récupérer sur le serveur les mots-clés attribués aux articles de 2009 à 2011. Ceux-ci n'auront pas été indexés selon les nouvelles consignes, mais nous pouvons les intégrer dans la zone « caractéristiques du contenu » afin de ne pas perdre complètement l'information (la plupart des mots-clés attribués correspondent de toute façon au contenu spécifique de l'article). La récupération a déjà été faite de façon automatique dans un fichier Excel qui servira pour une alimentation également automatique.

8.1.3 Exploitation sur le site Web et incidence client

o Affichage

Les mots-clés auteur seront affichés sur la page article (fig. 4) et sur la page de résultats du moteur de recherche (fig. 5), où ils seront cliquables.

⇨ ils donnent une information de contenu,

⇨ ils fonctionnent comme suggestion pour relancer la recherche.

o Pondération Exalead

Les mots-clés seront pris en compte dans la grille d'analyse d'Exalead, avec un coefficient à déterminer.

⇨ ils amélioreront la pertinence des résultats avec un classement qui fera ressortir en premier les articles qui ont le terme recherché dans les mots-clés,

⇨ ils feront ressortir des articles qui ne contiennent pas le terme recherché.

8.1.4 Contrôle et évolutions

- Modifications

Un accès doit être donné aux éditeurs dans eZ Publish pour des modifications ponctuelles sur des erreurs d'attribution ou des manques sur certaines zones.

- Contrôle d'intégration

Un contrôle peut être réalisé en éditant la liste des articles ne possédant pas de mots-clés sur une période donnée :

- à partir de la date de mise en œuvre, ceci permet de voir si les consignes de répartition en différentes zones sont suivies et toutes les zones alimentées,
- entre deux dates plus anciennes, cela peut permettre de compléter l'intégration.

- Evaluation qualitative

Tous les ans par exemple, il serait intéressant d'extraire la liste des mots-clés par types avec un tri par base documentaire puis un tri alphabétique ou un tri par occurrence pour que les éditeurs puissent vérifier les termes utilisés et faire évoluer les consignes si nécessaire.

- Possibilités d'évolutions

Nous pourrions envisager de donner aux auteurs un accès à des listes de termes contrôlés pour certains types de mots-clés (à définir s'il s'agirait d'une liste fermée avec des mots imposés ou d'une liste ouverte avec possibilité d'ajouter de nouveaux termes).
Ceci permettrait d'avoir une meilleure cohérence et homogénéité entre les différents articles, éviterait l'emploi d'un certain nombre de synonymes et par conséquent entraînerait une amélioration des résultats de recherche.

8.1.5 Coût de mise en œuvre et contraintes

La mise en place de ces mots-clés nouvelle version implique de passer un peu plus de temps lors de l'intégration pour vérifier si les mots-clés ont été répartis dans les différentes catégories et les consignes respectées.

8.2 Création de mots-clés pour l'application Expernova

Le partenariat avec Expernova étant décidé, la création des mots-clés permettant son fonctionnement est prioritaire (voir paragraphe 6.3.4 pages 68 et 69). La mise en place de la première phase est prévue en octobre et la deuxième phase en janvier.

8.2.1 Caractéristiques

Expernova traite les mots-clés uniquement en anglais, avec autosuggestion lorsqu'on tape un mot sur le site. Ils sont issus de la littérature scientifique (algorithme d'extraction développé par le CNRS) : le champ lexical de chaque expert est extrait à partir de ses écrits.

Par défaut le moteur relie les mots par un OU, on doit mettre des + devant les termes pour qu'ils soient reliés par un ET, des – pour exclure des termes (SAUF) et on doit mettre les expressions entre guillemets.

Pour les acronymes, il est conseillé de mettre à la fois le sigle et l'expression développée.

Le mot-clé utilisé n'est pas trop spécifique afin d'obtenir un nombre de réponses suffisant pour attirer les clients. Il est défini pour une ou plusieurs sélections et attribué à l'ensemble des articles contenus dans ces sélections. Pour les articles appartenant à plusieurs sélections il y en a une qui sera définie comme prioritaire.

Les mots-clés seront attribués de façon à être représentatifs dès la mise en place de la première phase, afin de pouvoir être utilisés à l'identique lors de la deuxième phase.

Exemples :

Titre de la sélection	**Terme retenu**	**Nb d'experts**
Gestion et traitement des déchets	« waste treatment »	41
Réglementation ICPE et droit environnemental	« environmental regulation »	70
Gestion de l'eau par les industriels	« water management »	280

8.2.2 Processus de production

- **Création Editeurs**

Pour être certain que les termes retenus renvoient bien des résultats et de manière cohérente, il est nécessaire de les tester sur le site Expernova. Les seules personnes connaissant suffisamment le fonds documentaire pour choisir les expressions appropriées sont les responsables éditoriaux.

Un terme doit être attribué à chaque article, cependant la précision requise par Expernova n'est pas très importante : les résultats doivent traiter du sujet global pour donner envie aux clients d'en savoir plus mais ils ne sont pas là pour fournir des informations précises à ce stade. Les tests montrent qu'il est plus difficile de trouver des termes pour chaque article, qui risquent finalement d'être à côté du sujet réel, plutôt que des termes plus généraux qui s'appliqueront à un ensemble.

Après concertation, il a donc été décidé d'attribuer un terme à chaque sélection, ce terme étant ensuite appliqué à tous les articles contenus dans la sélection. La méthode d'alimentation la plus simple est de créer un tableau Excel pour chaque éditeur, avec les sélections dont il a la charge. Ce tableau servira pour l'alimentation initiale et pour les grosses mises à jour si besoin. Les éditeurs choisissent les termes sur le site Expernova et les reportent dans le tableau. Les consignes sont indiquées dans le fichier Excel au-dessus du tableau (voir annexe 3).

- **Intégration informatique**

L'intégration des termes se fera à partir des fichiers Excel lorsqu'ils seront entièrement remplis par les éditeurs. Elle se fera automatiquement (programme à développer) avec attribution du terme à tous les articles du fonds documentaire rattachés à une sélection (ce qui couvre quasiment l'ensemble du fonds à l'exception des articles archivés).
Certains articles peuvent être rattachés à plusieurs sélections. Dans ce cas, le programme doit tenir compte du rattachement « prioritaire » défini pour l'article.

8.2.3 Exploitation sur le site Web et incidence client

- Première étape

Un lien vers le widget Expernova est établi sur le site, donnant les résultats du mot-clé attribué à l'article. Il faut définir les pages sur lesquelles le widget apparait ainsi que son emplacement dans la page. Un programme permettra de renvoyer vers ces pages les mots-clés Expernova.

- Deuxième étape

Seuls les clients auront accès aux informations plus complètes.

8.2.4 Contrôle et évolutions

- Evolution

Les termes attribués ne seront normalement pas soumis à des modifications, sauf si certaines expressions étaient mal saisies et ne donnaient aucun résultat. Il faut de toute façon prévoir une possibilité d'intervention mais qui sera ponctuelle avec un accès qui peut être limité à une ou deux personnes.
Par contre, il faut pouvoir alimenter cette zone lors de la création de nouvelles sélections ou de modification majeure de leur contenu. Dans ce cas, il sera possible d'utiliser de nouveau les tableaux Excel qui ont servi lors de l'alimentation initiale.
La mise à jour peut être réalisée une fois par an.

- Contrôle d'intégration

Tous les articles doivent avoir un mot-clé à l'exception des archives. Un contrôle peut donc être effectué en éditant la liste des articles sans mot-clé.

- Evaluation qualitative

Elle est faite lors du test sur le site Expernova : un relevé du nombre de centres et d'experts trouvés permet de vérifier que le choix du terme est cohérent. Si le nombre est trop élevé, il donne l'impression que les résultats ne sont pas sélectionnés et sont donc peu pertinents. Si le nombre est trop faible, le risque est de ne pas intéresser l'utilisateur.

8.2.5 Coût de mise en œuvre et contraintes

Le temps passé à alimenter les tableaux Excel est estimé à dix minutes par mot-clé attribué. Etant donné qu'il y a quatre cent sélections, le temps global estimé sera d'environ soixante sept heures à répartir entre les quatre responsables éditoriaux.

Le développement de cette application entre dans les travaux de maintenance pour les premières étapes (les résultats statistiques sont donnés à tous les internautes).

Un développement spécifique devra être réalisé pour intégrer l'application en janvier 2012 pour la dernière étape. En effet, pour celle-ci les clients auront accès à des données plus complètes. Il faudra donc gérer les droits d'accès.

Un point particulier à surveiller : que le widget Expernova ne ralentisse pas le chargement de la page sur le site Web, car lors des tests sur le site Expernova, le temps de réponse est parfois assez long.

A moyen terme

8.3 Création des mots-clés « phares »

8.3.1 Caractéristiques

Ils servent à annoter les articles indépendamment de la classification Techniques de l'Ingénieur, ce qui permet un accès simplifié et plus de souplesse pour valoriser certains articles ainsi que choisir et faire évoluer les termes utilisés.

Ils sont donnés avec un objectif **marketing** et ciblés « nouvelles technologies ». Il s'agit de chercher les mots sur lesquels on peut être visible, les mots-clés émergents, liés à l'innovation, aux avancées et nouveautés technologiques et utilisés par les usagers.

Ils permettent de rassembler des contenus sous une même étiquette et de les valoriser par rapport au reste du fonds.

Pour chaque domaine, une liste des mots-clés phares est établie, qui peut être mise à jour chaque année en fonction des nouveautés introduites dans les bases documentaires.

Le but est ensuite d'indexer les 10 à 20 articles correspondant le mieux à chaque mot-clé et de leur donner une forte pondération dans Exalead.

Un article peut avoir de 0 à plusieurs mots-clés phares, qui seront affichés au niveau de l'article avec les mots auteur.

Exemple de liste correspondant à l'énergie :

Batterie
Biocarburant
Efficacité énergétique
Energie éolienne
Energie hydrogène
Energie nucléaire
Energie solaire
Géothermie
Hydroélectricité
Photovoltaïque
Pile(s) à combustible
Récupération d'énergie
Smart grid
Stockage d'énergie
Véhicule électrique

Remarques :

- Les listes peuvent également être proposées en support d'interrogation.
- Les mots-clés phares attribués aux articles peuvent être répercutés aux sélections auxquelles ils appartiennent.

8.3.2 Processus de production

- **Création Editeurs**

Les mots-clés « phares » sont définis par les responsables éditoriaux qui connaissent bien les parties de la collection dont ils sont responsables et sont entourés d'experts avec lesquels ils font une veille sur les nouveautés. Ils définiront des listes de mots par thèmes (pour être cohérent dans les termes choisis) :

- sur la base des technologies clés et tendances,
- à partir des nouveautés introduites dans les bases documentaires,

- en comparant avec la liste transmise par le responsable SEO,
- en testant les mots-clés sur Google adwords.

Un tableau Excel sera rempli avec le thème, la liste des termes retenus et pour chaque terme la liste des numéros d'articles à rattacher (10 à 20 articles).

- **Intégration informatique**

L'alimentation de l'espace « mot-clé phare » dans eZ Publish se fera à partir du fichier Excel avec rattachement à l'article (prévoir jusqu'à 3 mots-clés) et éventuellement à la sélection (prévoir jusque 10 mots clés) si l'on souhaite une mise en valeur à ce niveau également.

8.3.3 Exploitation sur le site Web et incidence client

o Affichage

Les mots-clés phares peuvent être affichés au niveau de l'article avec les mots-clés auteurs. Il serait intéressant de les mettre plus en avant sur certaines pages sous forme de nuage de tags ou de proposition de liste de mots à l'utilisateur.

o Pondération Exalead

La pondération dans Exalead sera très forte sur ces termes pour que les articles rattachés apparaissent toujours en tête des résultats.

8.3.4 Contrôle et évolutions

o Evolution

On peut estimer qu'une mise à jour annuelle des listes de mots phares avec transmission des nouveaux mots-clés et de leurs rattachements serait pertinente.
Il faut prévoir une possibilité d'intervention à tout moment par les éditeurs dans eZ Publish pour ajouter ou enlever des rattachements afin d'être réactif sur certains termes lorsqu'ils le jugeront nécessaire.

o Contrôle d'intégration

Les mots-clés phares ne sont pas attribués à tous les articles, il n'y a donc pas de contrôle d'alimentation.

o Evaluation qualitative

Annuellement et avant mise à jour, il serait intéressant d'éditer la liste complète des mots utilisés (tri alphabétique) avec leurs rattachements pour vérifier l'homogénéité des mots-clés entre éditeurs.

Pour tester l'impact de la mise en œuvre de ces mots-clés, il faut tester la recherche sur ces termes avant de les intégrer et noter les dix ou vingt premiers résultats donnés par le moteur de recherche, puis refaire la même chose après intégration.

8.3.5 Coût de mise en œuvre et contraintes

Le coût sera fonction du temps passé à définir les mots-clés et leurs rattachements ainsi que de développements informatiques éventuels pour leur exploitation.
L'avantage de ce type d'indexation est que sa mise en place peut se faire progressivement avec un bénéfice dès les premiers termes intégrés.

A long terme

8.4 Catégorisation

Les catégories ont l'avantage de permettre l'accès aux ressources par différents points de vue mais l'inconvénient d'être trop figées et pas assez proche du contenu. Il s'agit toujours de groupes définis de façon très large car on ne peut pas multiplier le nombre de catégories sinon elles deviennent inutilisables.

La difficulté est donc de définir la granulométrie des catégories par rapport aux domaines :

- si celle-ci est trop large, cela ne présentera pas d'intérêt pour l'utilisateur,
- si elle est trop fine, la liste sera très longue et posera donc des problèmes de conception et d'utilisation.

Pour pallier à cet inconvénient, il faudrait avoir plusieurs types de catégories, ou plusieurs rattachements pour chaque article afin de pouvoir les croiser et mieux cibler les réponses.

Il faut alors faire face à un nouveau problème qui est celui de la multiplication des filtres : cela devient trop complexe et on risque de perdre l'internaute.

Ainsi, la mise en place de la nouvelle liste de domaines proposée ne fait pas l'unanimité car on note un risque d'incompréhension de la part des clients entre les thèmes existants (qu'il faut maintenir car utilisés pour la classification) et ces domaines qui sont pour certains identiques.
Il faudrait ensuite rattacher tous les articles du fonds à une ou plusieurs catégories pour que l'exploitation soit ensuite significative car si une partie du fonds n'est pas traitée, les résultats sont faussés.

L'intérêt serait, par exemple pour le rattachement à des secteurs d'activité, de pouvoir ensuite facilement extraire tous les articles d'un secteur pour les présenter sur un portail thématique ou pour créer des sélections d'articles par type de clientèle.

La consigne donnée aux auteurs d'indexer le secteur d'activité en le séparant des autres mots-clés vient de cette réflexion et permettra d'observer si cette information est exploitable. Le processus de catégorisation ne sera donc envisagé qu'à long terme.

Synthèse

8.5 Tableau récapitulatif des développements

Afin de résumé les différentes propositions, je vous propose un tableau de synthèse selon la méthode QQOQCCP.

	Mots-clés auteurs	**Mots-clés Expernova**	**Mots-clés phares**
Qui ?	Auteurs	Responsables éditoriaux	Responsables éditoriaux
Quoi ?	Contenu d'un article	Sujet général d'une sélection	Nouvelles technologies
Où ?	Externe	Interne	Interne
Quand ?	A la création de l'article	A la création de nouvelles sélections	Annuellement ou lors de nouveautés
Comment ?	Indexation libre, avec des consignes	Par test sur le site Expernova	Veille et choix marketing
Combien ?	6 à 8 mots-clés	1 à 3 mots-clés	0 à 3 mots-clés
Pourquoi ?	- Information de contenu - Suggestion de recherche - Amélioration de la pertinence des résultats de recherche	- Apport d'un service aux clients	- Valorisation des articles - Communication innovante

Tab. 5 - Synthèse des propositions de mise en place de mots-clés

Conclusion

Cette étude a permis de montrer les multiples intérêts que peuvent avoir les mots-clés lors de la diffusion d'information sur le Web et les nombreuses applications qui en résultent, que ce soit pour permettre d'apporter une aide à l'utilisateur, pour clarifier le contenu des articles ou pour valoriser les informations disponibles. Nous avons cependant remarqué que pour que ces différentes fonctions demandées aux mots-clés soient optimales, l'indexation doit être faite séparément pour chacune d'elles, avec des consignes spécifiques aboutissant à l'objectif recherché.

On peut donc définir autant de mots-clés que l'on veut pour caractériser un article, mais pour qu'ils soient adaptés à l'usage que l'on souhaite en faire, il faut mener une réflexion préalable permettant de donner le plus d'indications possible sur la façon de les créer.

Cependant, une trop grande complexité dans les recommandations peut entraîner une augmentation importante du temps de traitement pour l'attribution des mots-clés. L'enjeu est alors de trouver un compromis afin de donner des consignes qui soient suffisamment simples et précises à la fois. De même, il faut tenir compte de la nécessité ou non de traiter l'ensemble du fonds documentaire pour exploiter certains types de mots-clés avant de se lancer dans des projets qui peuvent s'avérer démesurés en temps passé par rapport aux bénéfices apportés.

On s'attachera également à définir quelles sont les personnes les mieux placées pour la création des mots-clés et à les faire intervenir dans l'établissement des règles et la mise en place des nouveaux processus, tout en réfléchissant aux possibilités d'automatisation de certains traitements.

Il ne faut jamais perdre de vue que l'objectif est de satisfaire les utilisateurs. Or, les besoins sont spécifiques à chaque internaute consultant le site Web et à chaque recherche effectuée. On ne peut donc pas espérer répondre à tous les besoins mais seulement proposer des outils suffisamment diversifiés pour apporter une amélioration sur les points que l'on aura définis comme importants pour l'entreprise.

Les préconisations effectuées répondent à une recherche d'équilibre entre la satisfaction d'un besoin et la charge de travail à réaliser afin de rester dans des solutions économiquement viables.

Parmi les pistes qui se sont dégagées, les propositions pouvant être menées à court terme se mettent en place :

- de nouvelles consignes sont maintenant diffusées aux auteurs. Il faudra bien sûr étudier leurs réactions et voir dans quelle mesure ils appliquent ces nouvelles règles. Les réflexions émises par les auteurs et l'étude des termes après un an permettront

- les tableaux de mots-clés pour l'application Expernova sont en cours de finalisation et le lancement est prévu sur les prochains mois.

L'utilisation de mots-clés représentatifs des nouvelles technologies est également à l'étude. La difficulté sera d'évaluer l'impact qu'ils pourront avoir sur la clientèle.

Seule la catégorisation n'est pas décidée pour l'instant car elle demande beaucoup plus d'investissement en temps, pour une amélioration difficilement quantifiable et incertaine.

Pour le futur, un outil plus élaboré et plus performant d'aide à la recherche serait la création d'une liste de synonymes, permettant un lien entre le vocabulaire employé par les internautes et celui utilisé dans les articles.

Enfin, ce travail pourrait être complété par une étude plus approfondie sur l'exploitation des différents mots-clés sur le site Web relativement à leur affichage ou à la mise à disposition d'index comme support d'interrogation en recherche avancée par exemple ainsi que sur la pondération au niveau du moteur de recherche afin d'améliorer la pertinence des résultats.

Bibliographie

Les références bibliographiques sont présentées selon les normes de référence :

- Z44-005. Décembre 1987. Documentation. Références bibliographiques : contenu, forme et structure.

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

Les notices sont classées selon les principaux thèmes présentés dans ce mémoire, dans l'ordre suivant :

- Indexation, langages documentaires et métadonnées
- Recherche d'information et analyse des usages
- Gestion de contenu et valorisation de l'information sur le Web

A l'intérieur de chaque thème, elles sont classées par ordre alphabétique du nom du premier auteur cité ou de l'organisme lorsqu'il n'y a pas d'auteur. Elles sont numérotées en continu.

Quelques références complémentaires, non citées dans le mémoire mais permettant d'alimenter la réflexion sont ajoutées à la suite dans la rubrique « pour en savoir plus ».

Il s'agit d'une bibliographie analytique, arrêtée à la date du 10 octobre 2011.

Indexation, langages documentaires et métadonnées

[1] ASSAL Sophie. Mots-clés d'auteurs et langages documentaires. Réflexions sur la valorisation des revues du pôle éditorial de la Maison René-Ginouvès. Mémoire INTD. 2009. 128 p.

Mémoire qui présente les langages documentaires, l'indexation des documents, la recherche d'information ainsi que sa valorisation (accessibilité, visibilité et diffusion). Une enquête sur la fabrication des mots-clés et les usages qui en sont faits a été réalisée auprès de chercheurs/auteurs. Les résultats de cette enquête ont permis de proposer quelques produits et services documentaires pour valoriser les revues et les informations qu'elles contiennent.

[2] BROUGHTON Vanda. Emergent vocabulary control in Web 2.0. Comparisons with conventional LIS theory and practice. Les Cahiers du numérique, 2010, vol. 6, n°3, p. 49-75. ISSN 1622-1494

Article qui compare tagging et vocabulaire contrôlé, avec présentation des méthodes de contrôle du vocabulaire dans les langages documentaires traditionnels, comme les thésaurus, ainsi que les avantages qu'ils offrent lors de la recherche d'information. Le langage d'indexation et d'accès qui résulte du tagging est ensuite comparé aux langages documentaires plus formels avec discussion sur les façons de combiner vocabulaires contrôlés et étiquettes (tags).

[3] Documentaliste – Sciences de l'information. Dossier Langages documentaires et outils linguistiques, 2007, vol. 44, n° 1. ISSN 0012-4508

Numéro spécial consacré aux langages documentaires avec une première partie axée sur la représentation des contenus et une deuxième partie s'attachant aux normes, standards et à l'interopérabilité.

[4] DURIEUX Valérie. Collaborative tagging et folksonomies. L'organisation du web par les internautes. Les Cahiers du numérique, 2010, vol. 6, n°1, p. 69-80. ISSN 1622-1494

Article sur le tagging, ses forces et ses faiblesses.

[5] HUDON Michèle. Le passage au XXIe siècle des grandes classifications documentaires. Documentation et bibliothèques, 2006, vol. 52, p. 85-97. ISSN 0315-2340

Article présentant les avantages et inconvénients des classifications sur le Web.

[6] HUDON Michèle, MUSTAFA EL HADI Widad. Organisation des connaissances et des ressources documentaires. De l'organisation hiérarchique centralisée à l'organisation sociale distribuée. Les Cahiers du numérique, 2010, vol. 6, n°3, p. 9-38. ISSN 1622-1494.

Article sur l'organisation des connaissances, les classifications documentaires et l'environnement Web 2.0.

[7] MANIEZ Jacques. Actualité des langages documentaires : fondements théoriques de la recherche d'information. Paris, ADBS Editions, 2002, 395 p. ISBN 2-84365-060-7

Ouvrage sur le rôle et l'importance des langages d'indexation et de recherche à l'heure d'internet.

[8] MENON Bruno. L'indexation à l'heure du numérique. Journée d'étude ADBS. Documentaliste - Sciences de l'information. 2004, vol. 41, n° 6, p. 340-342. ISSN 0012-4508

Présentation des interventions de la journée d'étude ADBS-INTD sur l'indexation et la diversification des pratiques.

[9] MERCURI Carole. Indexation automatique et enrichissement documentaire : le cas de la documentation de presse. Mémoire INTD. 2006, 93 p.

Mémoire qui présente le fonctionnement des systèmes d'indexation automatique et montre leur impact sur les pratiques des professionnels de la documentation, dans le domaine de la presse.

[10] MONNIN Alexandre. Qu'est-ce qu'un tag ? Entre accès et libellés, l'esquisse d'une caractérisation [en ligne]. In « Connaissance et communautés en ligne », 20e Journées Francophones d'Ingénierie des connaissances (IC2009), Hammamet, 2009. [consulté le 10/08/2011]. <http://ic2009.inria.fr/docs/paper/Monnin_IC2009_41.pdf>

Article définissant le tag par rapport aux descripteurs, mots-clés et vedettes matières.

[11] NEVEOL Aurélie, DARMONI Stéfan. Terminologie et accès à l'information en santé. In Mustafa El Hadi (éd.), Terminologie et accès à l'information. Hermès Lavoisier, Paris, 2006, p. 141-161. ISBN 2-74621-295-1

Chapitre présentant le choix d'un langage d'indexation et les critères de qualité, dans le cadre de l'information en santé.

[12] NISO – National Information Standards Organization. Understanding Metadata [en ligne]. Bethesda, NISO, 2004. 20 p. [consulté le 25/08/2011]. <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

Brochure présentant les métadonnées : leur création, l'interopérabilité et les échanges, la structuration en schémas, les perspectives de développement.

[13] POLITY Yolla, HENNERON Gérard, PALERMITI Rosalba. L'organisation des connaissances : approches conceptuelles. L'Harmattan. 2005, 272 p. ISBN 2-74758-274-4

Chapitres 3 sur la multi-dimensionnalité des termes, la multi-représentation et la multiplicité des niveaux de description, à usage de la recherche.

[14] RICHY Hélène, DESPRES Sylvie. Métadonnées, ontologies et documents numériques. Editions Techniques de l'Ingénieur. 2007, 19 p.

Article qui définit les concepts de métadonnées, de ressources sur le Web et d'ontologies.

[15] WALLER Suzanne. L'analyse documentaire : une approche méthodologique. Paris, ADBS Editions. 1999, 319 p. ISBN 2-84365-030-5

Ouvrage proposant une méthodologie d'analyse des documents pour en extraire le sens et le transmettre.

[16] ZACKLAD Manuel. Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information [en ligne]. In Actes de la Conférence CAIS/ACSI, Montréal, 2007. [consulté le 10/08/2011]. <http://www.cais-acsi.ca/search_fr.asp?year=2007>

Article comparant les systèmes d'organisation des connaissances (classifications, thésaurus, ontologies formelles, ontologies sémiotiques, folksonomies) selon différents critères pour évaluer leur pertinence en regard de la Recherche Ouverte d'Information.

Recherche d'information et analyse des usages

[17] BOYER Anne. Analyse des usages pour améliorer l'accès aux ressources. In Calderan Lisette, Bernard Hidoine et Jacques Millet (coord.), Métadonnées : mutations et perspectives. Paris, ADBS Editions. 2008, p. 89-111. ISBN 2-84365-104-2

Chapitre sur la recherche documentaire et l'amélioration de la pertinence des réponses par modélisation des comportements de l'utilisateur et la personnalisation de services numériques à partir des usages.

[18] CHAUDIRON Stéphane. La place de l'usager dans l'évaluation des systèmes de recherche d'informations. In Chaudiron Stéphane (dir.), Evaluation des systèmes de traitement de l'information. Paris, Hermès. 2004, p. 287-310. ISBN 2-74620-862-8

Chapitre présentant une analyse de la place de l'usager dans les protocoles d'évaluation des systèmes de recherche d'informations avec différentes approches destinées à placer l'usager au centre de la pratique évaluative.

[19] DALBIN Sylvie. Thésaurus à la recherche. In Journée d'Etude ADBS, Optimiser l'accès à l'information, une opportunité pour les langages documentaires ? Paris, 2007, 10 p.

Présentation des évolutions de la recherche documentaire et de l'utilisation de thésaurus pour améliorer la recherche.

[20] ERTZSCHEID Olivier. Moteurs de recherche : des enjeux d'aujourd'hui aux moteurs de demain. In Calderan Lisette, Bernard Hidoine et Jacques Millet (coord.), Métadonnées: mutations et perspectives. Paris, ADBS Editions. 2008, p. 59-88. ISBN 2-84365-104-2

Chapitre sur les moteurs de recherche, la notion de recherche universelle et l'approche sémantique.

[21] FLUHR Christian. L'évaluation des systèmes de recherche d'informations textuelles. In Chaudiron Stéphane (dir.), Evaluation des systèmes de traitement de l'information. Paris, Hermès. 2004, p. 27-45. ISBN 2-74620-862-8

Chapitre présentant les campagnes TREC (Text REtrieval Conference) en décrivant les différentes tâches qui sont évaluées avant de souligner les limites inhérentes à cette approche.

[22] GRIVEL Luc (dir.). La recherche d'information en contexte. Outils et usages applicatifs. Paris, Hermès Science Publications. 2011, 278 p. ISBN 978-2746225817.

Chapitre 3 intitulé « L'utilisation de la sémantique dans les applications basées sur la recherche d'information » expliquant le fonctionnement analytique des moteurs de recherche (Exalead) et chapitre 5 intitulé « La navigation à facettes dans les moteurs de recherche » avec une réflexion sur le choix et l'utilisation des facettes pour un site web.

[23] HEARST Marti A. Design Recommendations for Hierarchical Faceted Search Interfaces [en ligne]. In the *ACM SIGIR Workshop on Faceted Search*, Seattle, WA, 2006. [consulté le 16/08/2011]. <http://people.ischool.berkeley.edu/~hearst/publications.html>

Article présentant des recommandations pour l'utilisation de facettes sur un site Web.

[24] HEARST Marti A. UIs for Faceted Navigation : Recent Advances and Remaining Open Problems [en ligne]. In the *Workshop on Human-Computer Interaction and Information Retrieval, HCIR*, Redmond, WA, 2008. [consulté le 16/08/2011]. <http://people.ischool.berkeley.edu/~hearst/publications.html>

Article présentant une réflexion sur l'utilisation de la navigation à facettes tout en conservant flexibilité dans la recherche et bonne compréhension.

[25] IHADJADENE Madjid. Usages des moteurs de recherche. In Journée d'Etude ADBS, Optimiser l'accès à l'information, une opportunité pour les langages documentaires ? Paris, 2007, 40 p.

Présentation sur les enjeux et fonctionnalités des moteurs de recherche, la catégorisation des résultats ainsi que les aspects cognitifs de la recherche d'information.

[26] KOLMAYER Élisabeth. Démarche d'interrogation documentaire et navigation [en ligne]. In Quatrième colloque hypermédias et apprentissages. Poitiers, 1998. [consulté le 11/08/2011]. <http://hal.inria.fr/docs/00/00/26/63/PDF/HyperAp4p121.pdf>

Présentation sur les modélisations d'interrogation documentaire et les types d'aide à l'interrogation qui correspondent à chaque modélisation.

[27] MANON Émilie. E-science et professionnels de l'IST. Mémoire de stage Master 2 Spécialité Art, culture et médiations techniques. UPMF Grenoble. 2010, 54 p.

Mémoire présentant la notion d'e-science, ses enjeux dans les nouveaux environnements de recherche ainsi que les domaines d'intervention.

[28] SIMONNOT Brigitte. De la pertinence à l'utilité en recherche d'information : le cas du Web [en ligne]. In Recherches récentes en Sciences de l'information - convergences et dynamiques, Actes du colloque international MICS-LERASS, Toulouse, 2002. [consulté le 11/08/2011].
<http://hal.archives-ouvertes.fr/docs/00/06/26/04/PDF/sic_00001410.pdf>

Présentation qui analyse dans quelle mesure l'usager est pris en compte dans la conception des systèmes de recherche documentaire, les différentes étapes du processus de recherche et la complexité du concept de pertinence.

Gestion de contenu et valorisation de l'information sur le Web

[29] ANDRIEU Olivier. Optimisation d'un site web en vue de son référencement. Editions Techniques de l'Ingénieur. 2009, 16 p.

Article sur le référencement des sites Web, qui nous intéresse pour la partie utilisation des mots-clés.

[30] CHAUDIRON Stéphane, IHADJADENE Madjid, MAREDJ Azzeddine. La fragmentation et l'unité documentaire en question [en ligne]. In Actes du 16e Congrès de la SFSIC, Compiègne, 2008. [consulté le 10/08/2011].
<http://hal.archives-ouvertes.fr/00/46/87/96/PDF/Chaudiron-Ihadjadene-Maredj_SFSIC-2008.pdf>

Article sur le processus de déconstruction / reconstruction de l'unité documentaire sur internet et plus spécifiquement dans le champ de l'information scientifique et technique.

[31] LAHAYE Philippe. Les systèmes de gestion de contenu : description, classification et évaluation. Mémoire CNAM. 2004, 130 p.

Mémoire décrivant les fonctionnalités des trois sous-systèmes fondant l'architecture d'un CMS : le système de collecte (acquisition et édition), le système de gestion de contenu (référentiel et fichiers de configuration) et le système de publication (transformation, recherche d'informations).

[32] MARCK Adeline. Référencement : stratégie documentaire versus stratégie marketing. Le cas des sites web des cyberlibrairies et maisons d'édition. Mémoire INTD. 2005, 170 p.

Mémoire présentant de façon générale les moteurs de recherche et leur fonctionnement ainsi que les méthodes d'optimisation des pages web en vue de leur référencement. Puis sont définis les concepts fondamentaux liés à cette opération ainsi que son double objectif documentaire et marketing.

POUR EN SAVOIR PLUS

AMAR Muriel. Nouvelles pratiques d'indexation, nouveaux enjeux documentaires ? [en ligne]. Support de cours URFIST, Paris, 2008. [consulté le 11/07/2011].
<http://urfist.enc.sorbonne.fr/ressources/>
Support de formation sur la diversité des modes d'indexation des ressources numériques : de l'indexation documentaire à l'indexation sociale en passant par l'indexation automatique et l'indexation sémantique.

AMAR Muriel. Taxonomies, ontologies et folksonomies : situer les différences et identifier les usages [en ligne]. Support de cours URFIST, Paris, 2008. [consulté le 11/07/2011].
<http://urfist.enc.sorbonne.fr/ressources/>
Support de formation présentant les définitions et comparaison entre taxonomies, ontologies et folksonomies ainsi que l'identification de leurs usages.

BAUDRY DE VAUX Marie, DALBIN Sylvie. Métadonnées et valorisation de l'information : journée d'étude ADBS-INTD. Documentaliste – Sciences de l'Information. 2006, vol. 43, n°2, p. 144-147, ISSN 0012-4508.
Présentation des interventions de la journée d'étude ADBS-INTD sur les métadonnées, de leur constitution à leur rôle sur le web.

DALBIN Sylvie. Représentation et accès à l'information : transformation à l'œuvre. In Calderan Lisette, Bernard Hidoine et Jacques Millet (coord.), Métadonnées : mutations et perspectives. Paris, ADBS Editions. 2008, p. 9-57. ISBN 2-84365-104-2
Chapitre qui présente le processus de création de métadonnées ainsi que l'importance des modélisations et la prise en compte de l'interopérabilité.

SITES WEB

<http://www.abondance.com/>
L'actualité et l'information sur le référencement et les moteurs de recherche.

<http://dossierdoc.typepad.com/descripteurs/>
Site dédié aux thésaurus et autres vocabulaires contrôlés pour l'accès à l'information.

<https://adwords.google.fr/select/KeywordToolExternal>
Générateur de mots-clés de Google, outil donnant les statistiques de recherche d'un mot-clé donné ainsi que des mots-clés connexes à celui-ci.

Annexes

Annexe 1 Liste des bases documentaires

1) Agroalimentaire
2) Bâtiment et environnement
3) Bâtiment et travaux neufs
4) Bioprocédés
5) Bruit et vibrations
6) Chimie verte
7) Conception et production
8) Constantes physico chimiques
9) Construction – Réglementation et planification
10) Convertisseurs et machines électriques
11) Corrosion Vieillissement
12) Documents numériques Gestion de contenu
13) Elaboration et recyclage des métaux
14) Electronique
15) Emballages
16) Environnement
17) Etude et propriétés des métaux
18) Fonctions et composants mécaniques
19) Formulation
20) Génie civil
21) Génie énergétique
22) Génie nucléaire
23) Informatique industrielle
24) Innovations – Construction
25) Innovations – Electronique et TIC
26) Innovations – Environnement
27) Innovations – Management Organisation
28) Innovations – Mesure Analyse
29) Innovations – Procédés chimie bio agro
30) Innovations Matériaux avancés
31) Instrumentation et méthodes de mesure
32) Le traitement du signal et ses applications
33) Logistique
34) Machines hydrauliques, aérodynamiques et thermiques
35) Maintenance
36) Management industriel
37) Matériaux fonctionnels
38) Mathématiques pour l'ingénieur
39) Mesures et tests électroniques
40) Mesures mécaniques et dimensionnelles
41) Mesures physiques
42) Mise en forme des métaux et fonderie
43) Nanotechnologies
44) Opérations unitaires. Génie de la réaction chimique
45) Optique Photonique
46) Pathologie – Réhabilitation / Démolition – Déconstruction
47) Physique chimie
48) Plastiques et composites

49) Qualité et sécurité au laboratoire
50) Réseaux électriques et applications
51) Réseaux Télécommunications
52) Sécurité des systèmes d'information
53) Sécurité et gestion des risques
54) Structure et gros œuvre
55) Techniques d'analyse
56) Technologies de l'eau
57) Technologies logicielles Architectures des systèmes
58) Traçabilité
59) Traitements des métaux
60) Travail des matériaux – Assemblage
61) Tribologie

Annexe 2 Extrait de la classification

Mesures – Analyses (thème)

Informatique industrielle (base documentaire)

Recherche et innovation (rubrique)
Automatique
- *Rappels mathématiques* (sous-rubrique)
- *Généralités. Analyse des systèmes asservis*
- *Régulation*
- *Commande des systèmes asservis*
- *Technologie des asservissements*

Informatique temps réel
Robotique

Innovations – Mesure Analyse

Mesures
Analyse et caractérisation

Instrumentation et méthodes de mesure

Organisation. Méthodes de mesure
- *Unités de mesure*
- *Etalons*
- *Organisation française*
- *Organisations internationales*
- *Terminologie. Méthodes*
- *Eléments de statistique*
- *Capteurs*
- *Mesures dans des conditions particulières*

Mesures et tests électroniques

Techniques de mesure
- *Traitement du signal*
- *Techniques analogiques et numériques*
- *Visualisation*
- *Filtrage*

Temps
Mesures électriques
- *Etalons et références*
- *Instrumentation*
- *Grandeurs électriques*
- *Choix d'un système d'enregistrement*
- *Métrologie de l'Internet et des réseaux*
- *Mesures sur les matériaux*
- *Mesures en radiofréquences*
- *Mesures en optoélectronique*

Mesures mécaniques et dimensionnelles

Mesures dimensionnelles
Contrôle non destructif
Grandeurs mécaniques
Recherche et innovation
Acoustique. Vibrations
Métrologie optique et photonique

Mesures physiques

Recherche et innovation
Masses
Grandeurs hydrauliques et pneumatiques
Volumes
Niveaux
Pressions
Vitesse des fluides
Débits
Métrologie dans les fluides
Détection de gaz
Grandeurs thermiques
Thermométrie
Mesure des grandeurs thermophysiques

Qualité et sécurité au laboratoire

Recherche et innovation
Qualité. Validation
Qualité des essais et analyses au laboratoire
Mise en œuvre de la norme ISO 17025
Sécurité au laboratoire
Risques industriels
Analyses chimiques et biologiques
Sécurité des personnes
Sécurité d'utilisation des produits
Sécurité des locaux
Réglementation. Organisation de la prévention

Techniques d'analyse

Recherche et innovation
Notions de base en chimie analytique
Préparation du dosage
Caractérisations
Imagerie
Etudes de structure. Granulométrie
Structure et caractérisation des macromolécules biologiques
Méthodes thermiques d'analyse
Instrumentation. Métrologie
Chromatographie et techniques séparatives
Méthodes électrochimiques
Méthodes nucléaires d'analyse
Analyses de surface
Spectrométries
Spectrométrie de masse et techniques associées
Spectrométries des rayonnements électromagnétiques
Analyse organique. Chimie structurale
Analyse des macromolécules biologiques
Analyse des matériaux
Analyses dans l'environnement
Documentation générale en chimie analytique

Annexe 3 Tableau pour mots-clés Expernova

Consignes

Les mots-clés sont définis pour une sélection et attribués à l'ensemble des articles contenus dans cette sélection.

Ils ne sont pas affichés sur le site mais servent à alimenter le futur widget Expernova.

Donner 1 à 2 mots ou expressions en anglais, testés sur le site Expernova avec un résultat compris entre 30 et 400 centres de recherche ou experts.

Expressions encadrées de " ", mots-clés séparés par ; avec un + devant les mots pour une recherche en ET (par défaut le moteur attribue un OU entre les mots)

Ex 1 : nanomaterial / **Ex 2 :** "security management" / **Ex 3 :** +"air pollution" ; +industrial / **Ex 4 :** "acoustic analysis" ; "acoustic measurement"

Thématique	Base mère	code produit	mnémo	Titre du broché / produit	Mot-clé Expernova
Environnement	Bruit et vibrations	**42422210**	mbacse	**Acoustique des salles et de l'environnement**	
Environnement	Bruit et vibrations	**42423210**	mbacov	**Applications de l'acoustique**	
Environnement	Bruit et vibrations	42mbactr210	mbactr	**Acoustique des transports**	
Environnement	Bruit et vibrations	42mbnfav210	mbnfav	**Notions fondamentales en acoustique et vibratic**	
Environnement	Bruit et vibrations	**42424210**	mbbrvi	**Vibrations en milieu industriel, mesures, surveilla**	
Environnement	Environnement	**42436210**	mbaien	**Air, bruit et odeurs dans l'environnement**	
Environnement	Environnement	**42438210**	mbeaen	**Eaux industrielles**	
Environnement	Environnement	**42437210**	mbdech	**Gestion des déchets**	
Environnement	Environnement	**42440210**	mbsopo	**Gestion des sites et sols pollués**	
Environnement	Environnement	**42434210**	mbrine	**Recherche et innovation en environnement**	
Environnement	Environnement	**42439210**	mbicpe	**Réglementation ICPE et droit environnemental**	
Environnement	Environnement	**42442210**	mbmepa	**Systèmes de management environnemental**	