lingway
Vertical search solutions

Présentation détaillée

de

Lingway KM version 3.6

technologie de recherche sémantique, multilingue, prête à l'emploi et adaptable aux métiers

Les noms, lieux ou événements cités dans cette publication ne visent aucune personne, assemblée ou association existante ou ayant existé. Toute similitude ou ressemblance serait une coïncidence absolument fortuite.

AUCUNE GARANTIE, DE QUELQUE NATURE QUE CE SOIT, N'EST FOURNIE PAR CE DOCUMENT. La description des produits documentés et plus généralement l'information présentée sont placées sous les termes et conditions des contrats en cours de validité concernant l'achat ou la location de matériels et les droits de licence d'utilisation des logiciels dont il est question. Les seules garanties données par Lingway figurent le cas échéant dans ces contrats. Lingway ne pourra être tenu pour responsable, financièrement ou autrement, d'aucune conséquence de l'usage de l'information présentée dans la documentation ou dans les logiciels eux-mêmes, incluant tout dommage direct ou indirect.

Il est de la responsabilité du lecteur de s'assurer que l'usage des informations fournies et celui des logiciels décrits entrent bien dans le cadre de la Loi et des règlements émis par les juridictions compétentes.

L'information présentée ici peut être modifiée à tout moment et sans préavis. Des rééditions ou mises à jour peuvent être diffusées par Lingway pour la modifier ou la compléter.

Édition juillet 2007. La correspondance concernant cet ouvrage peut être adressée à : Lingway, Immeuble Paritalie - 18, Rue Pasteur 94270 LE KREMLIN BICETRE, France Tel.+33 (0)1 58 46 12 40

Conception et rédaction : ADB Réalisations ~ www.adbrealisations.com

Des moteurs de recherche qui comprennent votre metier

Dans tous les domaines – personnel, public, secteurs professionnels, l'utilisation croissante d'internet et de l'informatique en général entraîne une véritable explosion des volumes d'informations disponibles sous forme numérique… et ceci dans toutes les langues ! Dans le même temps, la recherche d'efficacité et la concurrence prévalent à tous les niveaux. « Connaissance » est devenue synonyme de « puissance », et celui qui « trouve » la bonne information « tout de suite » dispose d'un avantage considérable.

Dans ce contexte, tous les outils de recherche devraient être de plus en plus performants. Il n'en est rien. Il suffit pour s'en convaincre de taper tour à tour, sur un moteur classique, un couple de mots puis leurs synonymes, et de constater les écarts considérables dans les résultats.

Fruit de plus de 100 années-hommes de recherche et développée par des linguistes et scientifiques expérimentés, la plate-forme Lingway KM met en œuvre des composants sémantiques puissants et complémentaires pour réaliser des moteurs :
→ qui se basent, non pas sur les mots, mais sur leur sens ;
→ qui, de ce fait, sont capables d'exploiter simultanément plusieurs langues ;
→ qui présentent des résultats consolidés cohérents sous une forme facile à exploiter, en outre adaptable à un environnement déterminé.

Toutes les solutions de recherche de Lingway (moteur généraliste prêt à l'emploi ; plate-forme de construction de moteurs « sur mesure » ; moteurs spécialisés pour le recrutement, le médical, la propriété intellectuelle, le commerce électronique…) « *comprennent votre métier* » *!*

Leur approche sémantique du document, couplée à divers outils de catégorisation et d'aide à la lecture, structure et organise virtuellement toutes les sources d'information – email, web, documents bureautique, corpus en partie structurés – et permet la localisation rapide des documents intéressants, accélérant leur interprétation et leur utilisation directe.

La technologie de Lingway KM constitue une innovation qui a été primée. La plate-forme se cache aujourd'hui derrière les interfaces de recherche avancée proposées par des sociétés bien connues, dans divers secteurs d'activité.

Sommaire

Introduction

Dans tous les domaines, la dématérialisation des documents et messages, l'accélération des échanges, la numérisation des données existantes, accroissent considérablement la masse d'informations sous toutes formes –structurées ou non, corps des pages web, emails, documents bureautiques… Et dans le même temps, la concurrence omniprésente impose à chacun de disposer au plus tôt de la bonne information.

Apports de la technologie

Dans cette optique d'efficacité, la plate-forme Lingway KM constitue avec sa technologie émergente un atout considérable. En comparaison des outils de recherche classiques, elle apporte des fonctionnalités novatrices qui deviennent sans aucun doute de plus en plus nécessaires dans tous les domaines d'activité :

- ◊ recherche non plus basée sur des chaînes de caractères, mais sur *le sens des mots*
- ◊ extraction automatique ou contrôlée des *termes significatifs* contenus dans les documents en retour, permettant de filtrer les résultats
- ◊ exploitation simultanée de corpus documentaires *multilingues*, permettant de trouver des documents écrits dans une langue différente de celle de la requête
- ◊ cartographie des documents, alertes…
- ◊ aides à la lecture avec fiches résumés, colorisation, fonction de comparaison, etc.

Objectif du document

Cette « Présentation détaillée » a pour objectif de faire comprendre aux décideurs de différents domaines métier, sans trop entrer dans les détails techniques, quelles sont les fonctionnalités des gammes de solutions proposées par Lingway et de quelle façon celles-ci peuvent être intégrées aux systèmes existants afin d'améliorer ceux-ci.

Fil conducteur

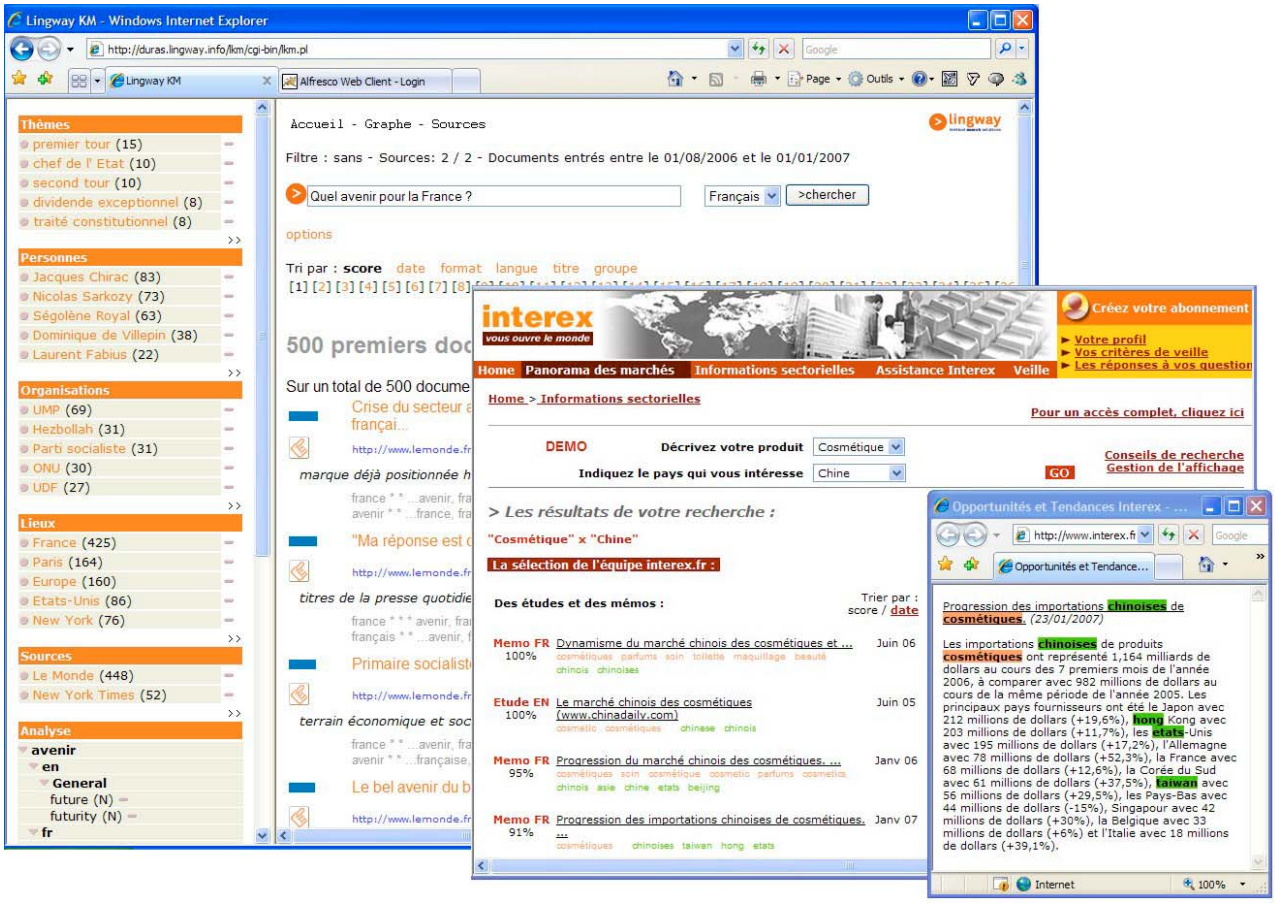
Pour rendre les explications et illustrations plus concrètes, ce document commence par décrire la face visible du système (l'interface utilisateur) pour aborder ensuite ses principes de fonctionnement, et aller jusqu'aux questions touchant à son intégration dans un environnement existant. Plus précisément :

- Ce document montre tout d'abord comment les fonctionnalités du système peuvent apparaître à l'utilisateur, au travers de l'exemple de l'interface fournie par défaut.
- Puis il détaille tour à tour chacune des fonctions sous-jacentes, relativement d'abord à la préparation des corpus documentaires, puis à leur exploitation : recherche d'information, aide à la lecture et gestion documentaire.
- Il présente ensuite l'architecture du système avec ses composants sémantiques et dictionnaires avant de décrire le fonctionnement interne d'un flux de recherche, depuis l'analyse de la requête à son interprétation, et couvrant la fonctionnalité de cross-language.
- S'appuyant sur la compréhension générale de ce fonctionnement, il traite alors de l'administration de la plate-forme et du réglage des dictionnaires.
- Il aborde enfin les aspects liés à la mise en œuvre : choix de la version de produit (packaging), installation, intégration en fonction du contexte, méthodologie.

Sont ensuite fournies les caractéristiques détaillées de la plate-forme, suivies du glossaire des termes utilisés dans ce document.

Une interface utilisateur transparente

La plate-forme Lingway KM propose une interface par défaut pouvant être utilisée telle quelle ou personnalisée ; alternativement, la plate-forme est conçue pour s'intégrer directement à un contexte existant, via diverses API.



Une interface simple et puissante, combinant sélection (via *index thématiques*) et recherche sémantique. En arrière plan : l'interface de démonstration fournie par défaut ; devant : un exemple de personnalisation pour un contexte métier particulier

Pour chaque document listé en réponse à une requête est fournie, sous son titre et en regard d'un indicateur de pertinence (score), la liste des mots qui ont permis de le sélectionner. En cliquant sur le titre, on ouvre le document proprement dit.

Recherche « intelligente » et « croisée »

En bref, la plate-forme Lingway KM est capable de doter tout type d'application de recherche professionnelle utilisant des *collections documentaires* volumineuses, de fonctions étendues :

- recherche sémantique, basée sur le sens de mots simultanément dans plusieurs langues
- requêtes en langage naturel ou par opérateurs logiques
- sélection croisée des documents selon leurs contenus relatifs à des *descripteurs* déterminés automatiquement ou imposés (*index thématiques*, colonne gauche).

Des résultats cohérents, exploitables

Les résultats, beaucoup plus efficaces qu'avec les moteurs de recherche traditionnels, sont présentés de sorte à faciliter l'accès aux documents et leur lecture. Différents outils et fonctions accélèrent l'exploitation :

- Au niveau de l'ensemble des résultats, pour localiser un document – des indicateurs de pertinence, des fonctions de tri (par score, date, format, langue, titre), de regroupement (*clustering*)…, la limitation de l'affichage aux documents dont le score dépasse un niveau de pertinence défini
- Au niveau de chaque document, pour estimer au-delà du score son intérêt au regard de la recherche – une liste des *hits* (mots de la requête ou dérivés sémantiques trouvés dans le document, justifiant sa sélection), une fiche analytique avec diverses aides à la lecture, document colorisé, fiche de méta-données (normalisée ou paramétrable), etc.

Fonctionnalités essentielles

Les fonctionnalités de Lingway KM sont basées sur des composants sémantiques sophistiqués et complémentaires que la plate-forme regroupe autour de dictionnaires (personnalisables). Justifiant les performances constatées, ce « dictionnaire » est en fait une base de données de connaissances linguistiques de 150.000 concepts décrivant finement les sens des différents mots dans 6 langues, et les mettant en correspondance via un *réseau sémantique*.

- Acquisition de données multiformats et activité de veille
- Structuration automatique des documents au format XML
- Recherche sémantique en langage naturel et ou via opérateurs logiques
- Recherche multilingue simultanée (cross-language)
- Recherche ciblée en base de données (multi-champs), pondérée et multi-critères
- Sélection croisée des documents par index thématiques
- Indexation plein-texte et indexation croisée par descripteurs
- Extraction automatique des descripteurs ou sélection à partir de listes ou thésaurus
- Recherche approximative de noms et correction orthographique sur les mots d'index
- Création automatique de fiches de synthèse (au standard DCMI et personnalisées)
- Navigation facile entre fiches documentaires et documents
- Personnalisation du dictionnaire standard et ajout de dictionnaires métiers
- Paramétrage du mécanisme d'analyse sémantique et des relations entre vocables
- Gestion de flux documentaires (workflow).

Avantages de Lingway KM

1. La combinaison de la recherche sémantique (particulièrement performante) et de la sélection par index thématiques accélère la localisation des informations sur des corpus volumineux.

2. La qualité des résultats fournis, mesurée en termes de « précision » et de « rappel » par divers tests (benchmarks), est largement supérieure à la qualité habituellement constatée.

3. Les résultats sont étendus aux documents rédigés dans une autre langue que celle de la question.

Applications de la plate-forme

La technologie Lingway KM a été conçue pour s'intégrer aux environnements-métier les plus variés, qu'ils soient généralistes ou très spécialisés. La gamme de solutions Lingway apporte sous une forme ou une autre une réelle valeur ajoutée, partout où il est important de retrouver rapidement une information précise au milieu de grandes quantités de documents, à l'origine non ou peu structurés.

Moteurs généraliste, « métier » et « sur-mesure »

La technologie Lingway KM est disponible sous différentes formes :

- ◊ prête à l'emploi pour un usage généraliste tel que par exemple la localisation d'informations d'origine journalistique
- ◊ applications-métier existantes (lire ci-dessous)
- ◊ offre intégrée de construction de moteurs de recherche.

Usage généraliste

Lingway KM utilise en standard un dictionnaire générique (pouvant au besoin, ponctuellement, être modifié par un « dictionnaire d'adaptation ») ; l'interface standard peut être soit utilisée telle quelle (éventuellement habillée, changement de logo…), soit remplacée par un accès direct aux fonctions de Lingway KM, via API, depuis l'application externe.

Dans tous les cas, la mise en œuvre est rapide.

Domaines métier existants

Des applications professionnelles spécifiques ont déjà été développées pour :

- ◊ le marché de l'emploi (mise en adéquation des offres et des curriculum vitaes – grâce, notamment, à la structuration XML des documents sous forme standardisée)
- ◊ le domaine médical (analyse et recherche de textes médicaux s'appuyant sur un puissant dictionnaire médical ; applications de pharmacovigilance avec codage automatique des phrases, etc.)
- ◊ la propriété intellectuelle (nouvelles méthodes d'accès aux bases de données brevets)
- ◊ e-commerce (applications dans le domaine des annuaires ; recherche dans les catalogues et nomenclatures, facilitation du processus de localisation et achat).

Dans chacun de ces secteurs, des entreprises renommées ont fait confiance à Lingway KM. [1] D'autres contextes professionnels peuvent bénéficier des fonctionnalités de la plate-forme, profitant des possibilités d'adaptation de ses fonctions et dictionnaires, de la facilité d'emploi de ses API, sans oublier l'aide proposée par les Services d'intégration de Lingway.

Développement sur mesure

De façon générale, les ressources linguistiques de Lingway KM – standard ou existantes – peuvent être adaptées à une application particulière.

L'offre Lingway Custom Search permet la mise en œuvre rapide de solutions de recherche personnalisées, s'articulant autour de deux pôles : plate-forme de développement et méthodologie éprouvée (détaillée plus loin dans ce document).

---

[1] e-commerce ou annuaires : AFNOR, Pages jaunes… ; Emploi : APEC, BNP Paribas, Groupe PSA, Peugeot-Citroën, Keljob, La Mondiale, MAIF, Opteaman, Siemens, Talent Up, Vivarte, Würth France … ; Medical : CEN Biotech, Fournier Pharma, Institut Gustave Roussy, Laboratoire de Fractionnement Biologique (LFB), Umanis… ; Propriété intellectuelle : FIST SA, INPI, Questel.Orbit… ; Généralistes ou autres : Bibliothèque Nationale de France, Géopost, INSEE, Interex, MINEFI.

Gestion électronique de documents

Lingway KM intègre un noyau de gestion documentaire qui permet, via une API ou un web service, d'intégrer simplement des bases documentaires déjà en partie structurées.

Il est ainsi possible d'effectuer – via le même protocole de communication – des recherches combinées plein-texte et par descripteurs sur la base documentaire ; d'ajouter, supprimer et modifier des documents ; de gérer un cycle de publication des documents (workflow).

Gamme et formules Lingway KM

Pour s'adapter à tout type d'application, la plate-forme Lingway KM est proposée en trois versions –Light, Standard et Premium–,déterminant les fonctionnalités incluses (pour plus de détail voir « Versions de la plate-forme », page 39).

Au choix, le logiciel peut être acquis sous licence ou loué à l'année.

Personnalisation et intégration métier

Personnalisation

La plate-forme est conçue pour une personnalisation à la fois des interfaces et des mécanismes d'indexation et de recherche.

Le développement d'interfaces personnalisées est particulièrement simple grâce à l'architecture en « web services ».

Pour l'*indexation de type 2* (autre que plein-texte), il est possible de charger les descripteurs à partir de listes définies ou de thésaurus.

Pour orienter le mécanisme de recherche et l'analyse sémantique, le dictionnaire peut être édité afin de définir des associations valides dans un domaine particulier ou d'autres relations entre termes ou locutions.

Intégration

L'architecture Lingway KM, exploitant certaines briques open source, est totalement ouverte via des API standard. Lingway KM se présente sous forme de serveur communiquant par défaut en XML, via le protocole SOAP. Une API Java est également disponible.

Globalement, la plate-forme a été conçue pour s'intégrer facilement à un contexte m étier particulier et un environnement logiciel existant.

INFORMATION

Les mots en italiques dans le texte renvoient au glossaire situé page 43.

Comment cela fonctionne-t-il, globalement ?

Les fonctionnalités de Lingway KM s'inscrivent dans l'une ou l'autre de deux phases distinctes, qui concernent des populations différentes :

- La préparation des "données" dans le sens large - personnes chargées de mettre en œuvre les *applications*
- L'exploitation de celles-ci par les usagers des dites applications – professionnels, généralistes ou navigateurs web.

Enchaînement des fonctions. A gauche la préparation, à droite l'exploitation

La préparation

Acquisition

Des fonctions d'acquisition et de veille basées sur différents outils et un « crawler web » alimentent les applications depuis diverses sources, pages web et fichiers des formats courants (Office, pdf, etc.)... filtrant les données et gérant les versions de documents (*vdocs*).

Indexation et *catégorisation* thématique

Outre l'indexation « plein texte » de ces vdocs, une structuration XML de leurs contenus permet –de façon asynchrone– de les indexer par *descripteur* (défini ou extrait) et de les catégoriser selon leurs contenus.

L'extraction des contenus correspondant à ces descripteurs – entités nommées (personnes, organisations, produits, lieux, etc.) ou *termes* – est rendue possible par l'analyse linguistique. Celle-ci permet le marquage de la structure générale d'un texte à l'origine non-structuré, et le marquage des termes significatifs ou des phrases importantes sous un angle de vue donné.

L'exploitation

Recherche croisée

Le moteur de recherche *sémantique* multilingue permet d'interroger, en langage naturel ou par opérateurs logiques, divers formats de bases textuelles :

- bases « full text »
- bases indexées par descripteurs (générés automatiquement ou imposés) ou *méta-données*.

Ces différents modes de recherche peuvent être utilisés séparément ou combinés entre eux, se renforçant mutuellement.

Aide à la lecture

Diverses fonctions permettent d'appréhender rapidement d'abord le contenu d'un ensemble de documents trop important pour être lu intégralement (fonctions de *clustering* et de cartographie), puis le contenu d'un document donné (fonctions de résumé et de colorisation).

Préparation…

Cette phase initiale correspond aux processus d'acquisition et d'analyse des sources de documents et des documents qu'il s'agit de structurer, annoter et indexer pour permettre leur exploitation pour la recherche et l'aide à la lecture.

Le travail de mise en œuvre (définition de l'application, de ses sources, du paramétrage) est effectué dans la plupart des cas une fois pour toutes, par les personnes chargées d'administrer les applications, éventuellement aidées par l'équipe dédiée Lingway.

Sources

Organisation des collections documentaires

La notion de « *collection documentaire* » se définit dans Lingway KM par une « Application » qui correspond à un ensemble de sources de données, éventuellement multilingues ; et éventuellement à des méta-données utilisateur s'appliquant à ces données : *plan de classement* où ranger les documents, *thésaurus* utilisateur, *dictionnaire d'adaptation* permettant de particulariser le dictionnaire générique inclus dans le produit.

Remarque : « *Thesaurus* » et « *Dictionnaire* » sont deux notions différentes dans Lingway KM. Le « thesaurus » est une *liste d'autorité* documentaire qui définit des descripteurs à retenir prioritairement pour l'indexation thématique des documents, alors que le « dictionnaire » est une ressource linguistique interne servant à interpréter les questions (pour produire une requête expansée) et à analyser les textes (pour extraire des index ou méta-données).
Voir Glossaire à la fin de ce document.

A une application correspond une base de données MySQL indexée en standard par l'indexeur Lucene [2], alimentée par une ou plusieurs sources.

Acquisition de données et crawling

Les sources peuvent être simultanément, pour une même application :

- « web » : les documents sont acquis (régulièrement ou non) en parcourant un site à partir d'une URL
- arborescence de fichiers: les documents résident sur un système de fichiers accessible par le serveur Lingway KM
- arborescence de fichiers historisés: les documents sont stockés sur le système de fichiers, mais représentent des images de sites (ou de systèmes de fichiers) à des dates différentes. Lingway KM reconstitue l'historique.

Les formats acceptés (autres que les formats basiques txt et html) sont rtf, pdf, tout format Microsoft Office (doc, xls, etc.) et xml.

Pour toute source, l'alimentation peut être complétée via un web service ajoutant les documents un par un.

Suivant les types d'alimentation, les paramètres d'acquisition peuvent être différents, mais ils se regroupent en :

---

[2] Le moteur d'indexation « plein texte » est en standard Lucene (open source) mais peut être remplacé par un indexeur propriétaire. Voir « Standards d'interfaçage » page 42.

- ◊ filtres permettant de restreindre les types de fichiers retenus, la profondeur des parcours dans les sites web, les sites à utiliser, etc.
- ◊ filtres de veille fixant la périodicité de l'aspiration, la synchronisation de la base avec les nouveaux contenus et le traitement des documents indexés ayant disparu de la source.

Dans le cas du HTML, le crawler élimine les parties non pertinentes des pages récupérées sur des sites web (telles que les barres de navigation, les publicités, etc.) et repère les documents au contenu identique.

Versionnement et structure des documents

Dans des contextes ne prenant pas en compte l'historique des données, un document correspond à un fichier, qui possède un contenu textuel et que l'on peut lire, afficher, modifier.

Si l'on fait intervenir la notion d'évolution du fonds documentaire, un même document (au sens Lingway KM) peut avoir plusieurs versions au cours du temps. Typiquement, les pages Web sont remises à jour régulièrement.

Les vdocs

Quand le système détecte un contenu différent pour une même adresse, il crée une nouvelle version de document sous la forme de « vdoc ». On utilise au sens Lingway KM le terme « document » pour l'unité abstraite correspondant à une adresse, et le terme « vdoc » pour une version de document à une date donnée, ou un ensemble de dates données.

Cependant, les modes avec et sans historique utilisent le même modèle de données. Dans le mode non historisé, chaque document est associé à une seule vdoc.

Remarque : Dans la suite de ce document de présentation, le terme de « document » doit être assimilé à « vdoc », version de document. Dans le cas où une application ne gère pas le versionnement des fichiers, la vdoc est unique.

Structure du document

Composition

L'unité documentaire est une liste de champs, textuels ou non textuels, qui sont remplis soit par extraction du contenu d'un fichier physique, soit par fourniture directe de la valeur des champs (lors de l'ajout de documents par l'API uniquement).

Cela permet une intégration simple de bases documentaires déjà en partie structurées. Par exemple, une base bibliographique en XML avec champs « description, origine, date, titre etc. », sans stockage des fichiers physiques source ; ou encore une base contenant des documents physiques (fichiers) associés à des méta-données non stockées dans les documents originaux.

Structuration

Pour baliser la structure des vdocs, Lingway KM reconnaît les titres, sous-titres, sections, paragraphes et phrases, et produit une version XML du document avec des balises correspondant à cette structure.

Cette reconnaissance n'est pas uniquement basée sur les balises du document en entrée (<title>, <h1>, etc.) ou d'autres marques de formatage, mais aussi sur la position, le format des caractères et un certain nombre d'autres éléments.

Champs

Les champs d'un vdoc abritent :

- du contenu textuel
- un identifiant avec emplacement physique (pour gestion interne de la vdoc)
- les méta-données, qui se répartissent en deux catégories :

- ◊ les méta-données « standard » présentes sur le contenu lui-même, déductibles automatiquement en dehors de tout processus « intelligent » (par exemple, date de publication ou source) ou encore celles positionnées directement lors de l'analyse
- ◊ les méta-données ou entités, extraites automatiquement par l'analyse linguistique.

Modèle documentaire et standard

Standard « Dublin Core » étendu

Concernant les champs de méta-données standard, Lingway KM implémente la norme Dublin-Core et renseigne ces champs automatiquement, reprenant les contenus lorsqu'ils sont présents ou les inférant à partir de règles.

Le modèle standard (Dublin Core) de la base Lingway KM peut être étendu à volonté pour prendre en compte des contraintes documentaires particulières.

Paramétrage des champs

Le mode d'utilisation des champs de vdoc par les divers traitements (détection de langue, analyse linguistique, détection de nouvelle version de document, etc) est entièrement paramétrable dans Lingway KM.

Définition des descripteurs et méta-données

Parallèlement à l'indexation "plein-texte", Lingway KM indexe les documents à partir des contenus des champs correspondant à des descripteurs, pour faciliter l'accès aux informations.

Ces descripteurs sont des termes et/ou des entités nommées (en standard, des noms de personnes, de lieux et d'organisations) qui peuvent soit être automatiquement extraits des vdocs par analyse linguistique, soit imposés à partir de listes ou thésaurus chargés.

Après la phase de balisage de la structure du texte (voir paragraphe précédent), le moteur d'extraction choisit et normalise les bons éléments parmi les résultats, qui peuvent correspondre à ces descripteurs et servir à catégoriser les documents.

Les documents en entrée, de divers formats (Office, HTML, PDF, TXT, etc.) sont convertis en XML, un grand nombre d'éléments (structure du document, entités nommées, terminologie, etc.) étant alors balisés.

Entités nommées extraites

Lingway KM extrait en standard les entités nommées suivantes :

- ◊ noms de personnes (en isolant nom, prénom et fonction quand cela est possible)
- ◊ noms de lieux (typés en pays, villes, régions, etc.)
- ◊ noms d'organisations (nom complet et sigle quand c'est possible)
- ◊ dates et autres marqueurs temporels.

Cependant, des jeux de règles plus spécifiques peuvent être rapidement développés pour des besoins précis. A titre d'exemple, ont été développées des règles de marquage de :

- ◊ types et caractéristiques de matériels informatiques
- ◊ niveaux de formation
- ◊ événements scientifiques et techniques (workshops, conférences, ateliers, etc.)
- ◊ autres événements (culturels, sportifs, ...).

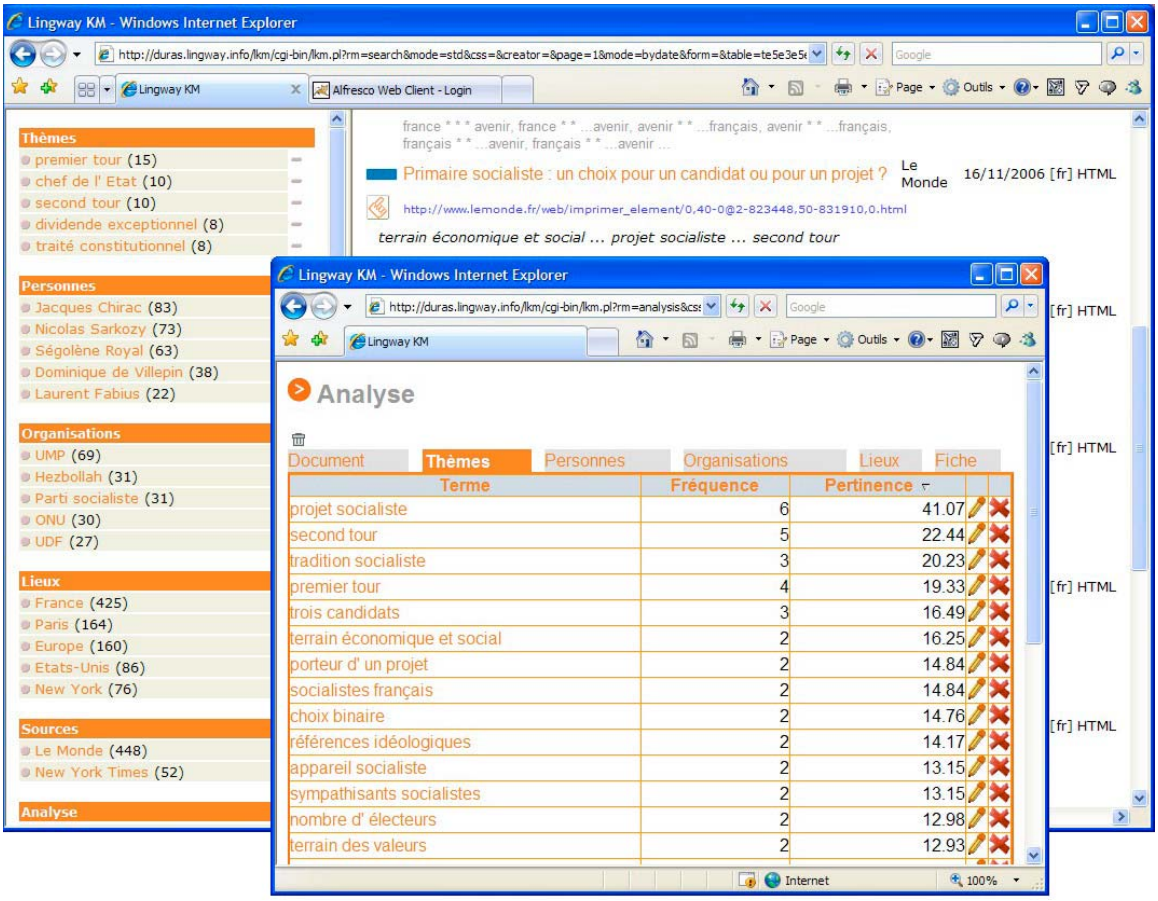
Paramétrage de listes d'entités

Aux fins de personnalisation, il est non seulement possible de définir des règles d'extraction d'entités de type non standard, mais aussi de charger directement des listes externes d'entités nommées à extraire, de type standard (listes d'organisations, personnes...).

Extraction des descripteurs thématiques

Un extracteur de terminologie retrouve les termes présents dans le document sur la base d'une analyse linguistique (structures Nom-Adjectif, Nom-Préposition-Nom, etc.). Ces termes deviennent des descripteurs potentiels du contenu du document. Le choix effectif des termes retenus comme descripteurs associés au document se fait sur la base de critères statistiques complémentaires. Ces descripteurs peuvent être validés via l'interface standard.

Les descripteurs retenus sont considérés comme méta-données, et indexent automatiquement le document.



L'indexation automatique « libre », à partir de descripteurs non normalisés

Imposition de termes via thésaurus

Autre personnalisation possible, l'indexation automatique peut se faire à partir des termes d'un thésaurus paramétrable.

Indexations croisées

Deux types d'indexation coexistent, *indexation de type 1* dite "full-text" et *indexation de type 2*, à partir de descripteurs extraits ou imposés.

Lingway KM désynchronise l'indexation 1 (qui est très rapide) et l'indexation 2 (qui activant une analyse linguistique plus profonde, est nécessairement plus consommatrice de ressources). Ceci permet de réagir très vite pour les opérations de veille et de recherche simple, et de réaliser en tâche de fond les extractions qui sont essentiellement utiles pour les tâches d'analyse et de recherche assistée (par l'utilisation des méta-données générées).

De plus, il est possible de lancer manuellement l'indexation de documents spécifiques, pour gérer par exemple un cycle de publication.

Indexation Full-text, open source ou propriétaire

Cette opération consiste à produire, à partir d'un ensemble de documents, une liste "inversée" de mots avec les liens vers les documents dans lesquels chaque mot apparaît.

L'indexation plein texte dans Lingway KM concerne toutes les vdocs actives d'une application ; elle s'effectue sur le contenu des champs de vdocs paramétrés « à indexer ».

Par défaut, cette tâche fait appel à l'indexeur Open Source Lucene, cependant l'architecture ouverte de Lingway KM lui permet de se connecter pour tenir ce rôle, à des moteurs de recherche du marché tels que Exalead, Hummingbird, Oracle…

Indexation à partir de termes et entités nommées

L'indexation 2 consiste à affecter des descripteurs à des documents. Lingway KM permet d'indexer automatiquement les documents (en base de données) à partir de termes et entités nommées : soit informations extraites (entités sémantiquement typées, termes de thésaurus), soit informations externes chargées.

Thésaurus externe
Utilisé pour indexer les documents, le thésaurus déclaré doit être mis au préalable au format d'entrée de Lingway KM. Ce format permet de prendre en compte la relation de synonymie.

Personnalisation des types d'entités
La liste des entités est paramétrable. Des règles d'extraction d'entités de type non standard peuvent, sous certaines conditions, être définies.

Indexation manuelle et gestion de cycle de publication

Allant de pair avec l'association d'états de workflow aux documents, l'indexation par descripteurs peut être complétée ou révisée manuellement via une interface d'indexation manuelle.

Grâce à un paramétrage spécifique, un cycle de publication (« workflow ») peut être activé. Les documents reçoivent au départ une indexation « brouillon », puis peuvent être validés ou refusés.

Classification de documents

Disponible dans la plate-forme à partir de la version 3.6.1, la fonction de *classification* permet de mettre en œuvre un type d'organisation documentaire hiérarchique à partir de plans de classement externes. Le « rangement » des documents dans un plan de classement est effectué automatiquement, par apprentissage à partir des documents déjà classifiés.

Exploitation

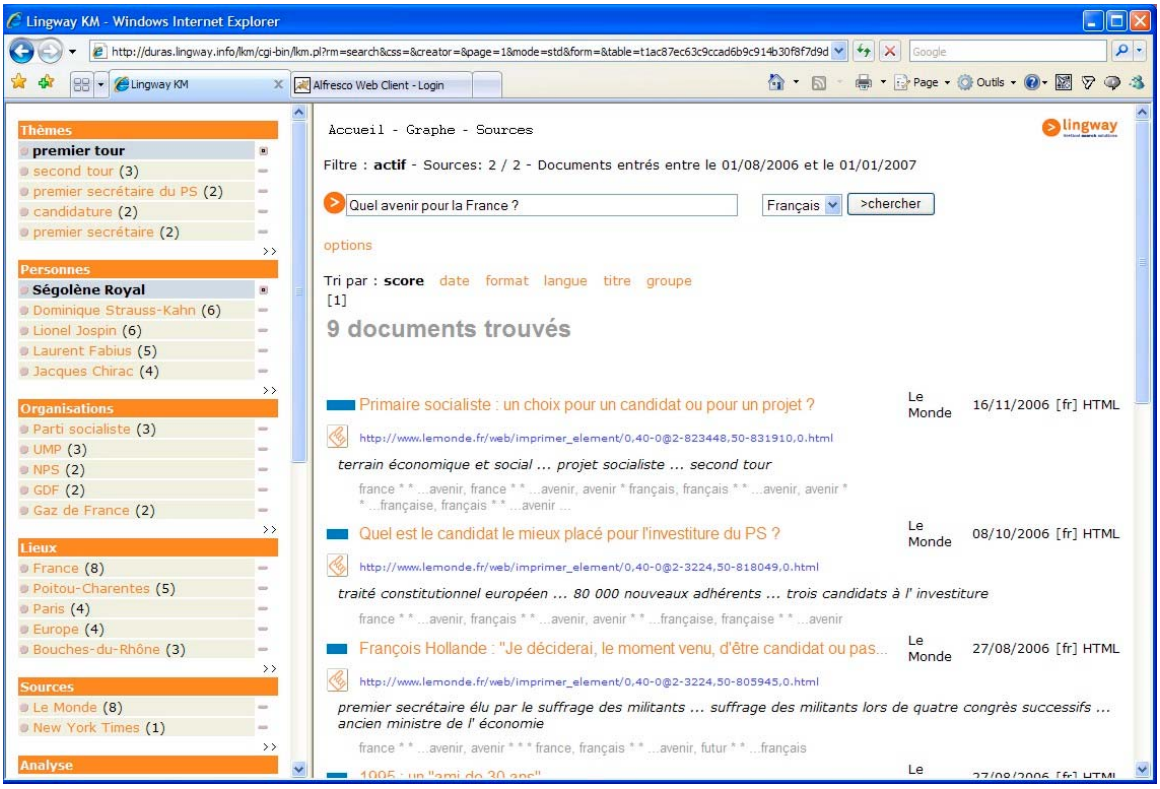
Ce chapitre présente les possibilités du système en exploitation, relativement aux fonctions de recherche, catégorisation et aide à la lecture. Il est considéré que la préparation des données et leur rafraîchissement périodique sont mis en place par un personnel dédié (ainsi qu'exposé au chapitre précédent). Complétant cette vue fonctionnelle « de l'extérieur », le paragraphe « Fonctionnement interne du flux de recherche » expose page 31 le processus d'analyse et d'interprétation des requêtes –dont la compréhension nécessite une présentation préalable de l'architecture de la plate-forme et des fonctions des composants sémantiques (page 27).

Recherche "croisée"

C'est la combinaison de la recherche sémantique (basée sur l'indexation de type 1, « plein texte ») et de la sélection par descripteurs (basée sur l'indexation de type 2, de termes, entités nommées ou autres descripteurs, en base de données), qui confère au système sa puissance, accélérant la localisation des informations sur des corpus volumineux.

Utilisation des descripteurs à l'interrogation

Lingway KM propose (colonne gauche de l'interface standard) des *index thématiques* – listes de descripteurs ou méta-données. Les descripteurs sont extraits du corpus, regroupés par type (descripteurs libres ou de thésaurus, entités nommées standard ou personnalisées), chaque liste pouvant être triée alphabétiquement ou par nombre d'occurrences. En cliquant sur un des descripteurs, on sélectionne toutes les vdocs indexées par ce descripteur, c'est-à-dire les vdocs pour lesquelles le descripteur a été considéré comme pertinent par le système.



Exemple d'une double sélection « premier tour » et « Ségolène Royal »

Il est possible d'affiner la sélection, réduisant le résultat, en activant de nouveaux termes dans la même liste ou dans une autre, puis de relancer la recherche en langage naturel.

Requêtes sémantiques multilingues

Le fait de pouvoir poser des questions en langage naturel rend le moteur très simple à utiliser et particulièrement bien adapté aux applications destinées à des non-spécialistes. Et cette facilité d'emploi s'accompagne d'une amélioration sensible de la qualité des résultats, par rapport aux moteurs traditionnels.

En effet, les mécanismes d'*expansion sémantique* et de dégradation exposés au paragraphe « Fonctionnement interne du flux de recherche » page 31, génèrent à partir d'une question en langage naturel toutes les requêtes correspondant à des formulations équivalentes ; ceci diminue le silence sans nuire à la précision des résultats.

Lorsqu'il travaille en mode sémantique, le moteur Lingway est capable de recherches « cross-language » : il peut trouver des documents écrits dans une autre langue que celle de la question (les langues actuellement traitées sont l'anglais, le français, l'allemand, l'espagnol et le néerlandais).

De façon paramétrable, les requêtes seront considérées comme formulées « en langage naturel » (par défaut) ou comme des requêtes booléennes à transmettre telles quelles à l'indexeur (sans traitement sémantique) ; alternativement, le choix entre ces deux modes peut être laissé à l'analyseur. Celui-ci bascule alors en mode logique s'il détecte un opérateur booléen dans la requête.

Recherche « ciblée et pondérée », multi-critères

S'appuyant sur l'analyse linguistique des documents et l'extraction de termes signifiants, ainsi que sur une logique combinatoire des requêtes, la recherche en language naturel de Lingway KM est à la fois souple et puissante. Globalement, les fonctionnalités de base sont les suivantes :

Recherche « ciblée et pondérée »
Exécution des requêtes sur plusieurs champs choisis des enregistrements en base de données (contenant le titre, le texte, des méta-données standard telles que date, source du document, etc. et des entités extraites par analyse) tout en accordant un poids sémantique à chaque champ (privilégiant par exemple les « matchs » sur le titre).

Les champs à interroger et leur poids sont entièrement paramétrables, que ce soit pour une interrogation via l'interface standard ou via l'API.

Interrogation multi-critères
Génération d'une sous-requête par champ de saisie, chaque sous-requête ayant les mêmes caractéristiques de recherche « ciblée et pondérée » qu'une requête. Ainsi des requêtes composites pouvent comprendre plusieurs sous-requêtes indépendantes reliées par des conditions "AND" (contrainte conjointe) et/ou "OR" (contrainte alternative).

Remarque : Cette fonctionnalité est utile pour mieux guider l'utilisateur dans sa saisie. Elle est pleinement utilisable via l'API lors de la mise en œuvre d'une application particulière. Elle est implémentée sous forme simplifiée dans l'interface standard de Lingway KM.

Aides à la lecture

Des indicateurs de pertinence (score), des fonctions de tri, l'affichage des *hits* sous chaque document listé … tous ces éléments aident à interpréter les résultats d'une recherche (voir paragraphe « Des résultats cohérents, exploitables », page 9).

De plus, divers outils d'aide à la lecture viennent accélérer notablement l'interprétation des résultats, dès lors que la collection documentaire est volumineuse et que, de ce fait, ceux-ci sont nombreux.

Ces outils répondent à un double besoin :

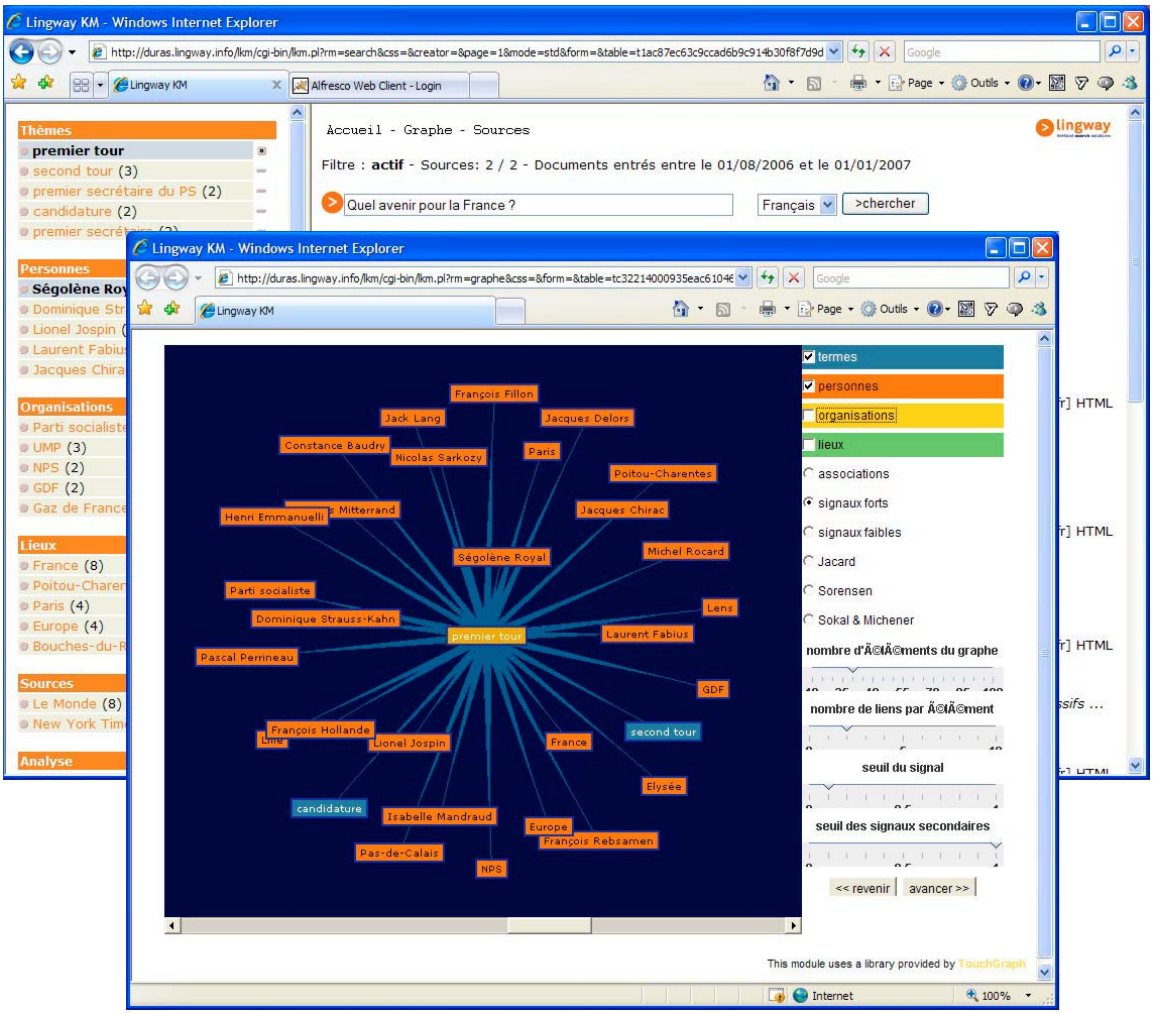
- Estimer rapidement le contenu global d'un ensemble de documents
- Appréhender rapidement le contenu d'un document donné.

La présentation des résultats, les fonctions de clustering et de cartographie de Lingway KM répondent au premier point ; les fonctions de marquage des phrases typiques, de colorisation du texte, de résumé et de navigation intra-document répondent au deuxième.

Clustering de termes et cartographie

Ces fonctions permettent d'obtenir une vue synthétique des contenus d'un ensemble de documents en les groupant selon leur sens. L'application d'algorithmes de clustering (agrégation) détermine une *distance sémantique* entre tous les descripteurs, termes et entités extraits du document lors du balisage XML.

Il en résulte un graphe qui montre pour un terme donné quels sont ceux qui lui sont proches dans l'ensemble des documents traités. On peut ainsi repérer des associations entre produits, personnes, sociétés, etc.

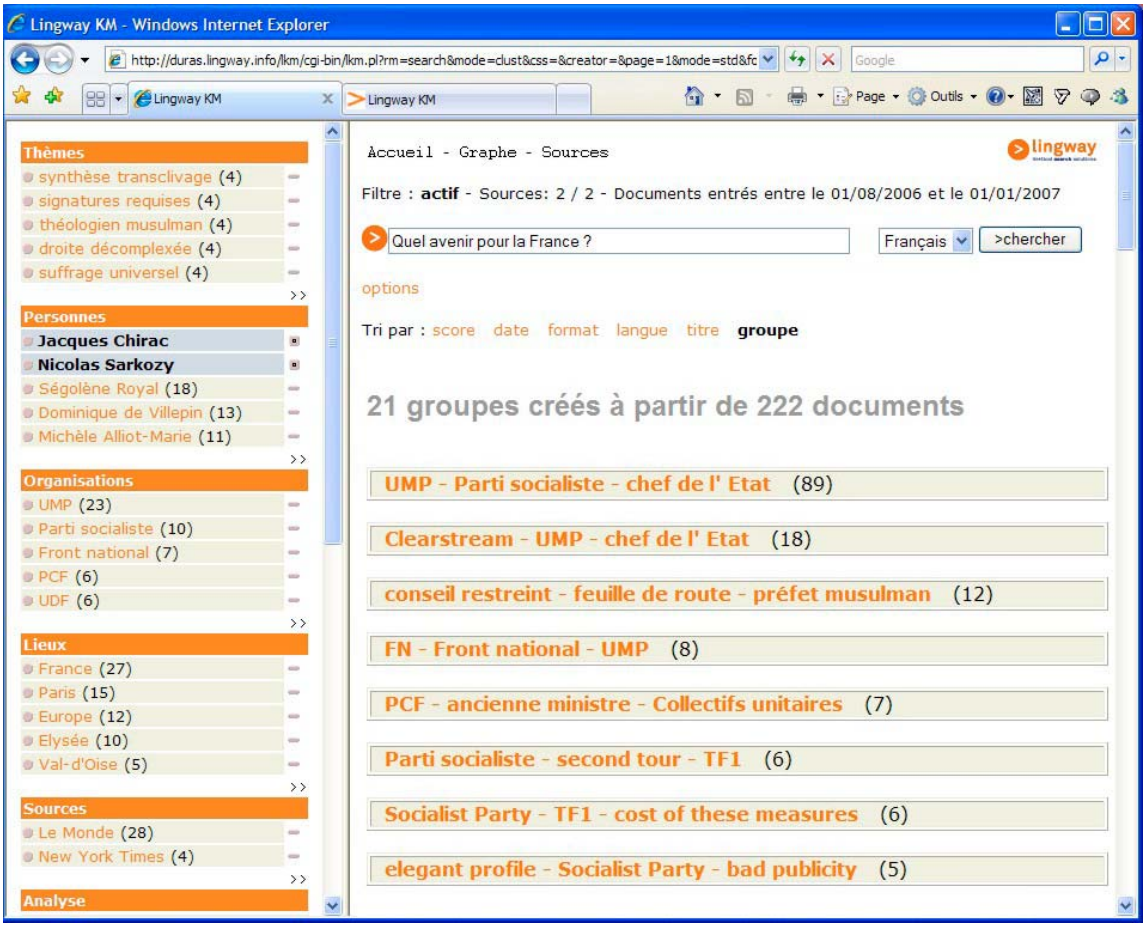


Exemple tiré de l'interface standard, du résultat de l'analyse des liens entre les termes extraits de l'ensemble des documents répondant à une question

Cette fonctionnalité est particulièrement utile pour des fonctions de veille, donnant une vue synthétique d'un ensemble donné de réponses à un instant t.

Clustering des résultats de recherche

De même que l'on peut construire un *clustering de termes* (cf. paragraphe précédent), on peut construire un clustering à partir des documents eux-mêmes, qui sont alors regroupés par similitude de termes et entités nommées indexant le document.



Des clusters construits sur les réponses à la question « Quel avenir pour la France ? »

Dans l'exemple de la figure ci-dessus, vingt et un clusters ont été constitués sur les documents répondant à la question « Quel avenir pour la France ? » citant à la fois Jacques Chirac et Nicolas Sarkozy, ceci sur la base de la proximité de leurs descripteurs. Le titre de chaque cluster est automatiquement généré, sous forme de lien. En cliquant sur celui-ci, le cluster s'ouvre et fait apparaître la liste des documents qu'il contient.

Fiche analytique de document

Pour répondre au deuxième besoin, « appréhender rapidement le contenu d'un document », une fiche analytique est associée à chaque document en réponse.

Un simple clic dans la liste, sur l'icône en regard d'un document, donne accès à une fenêtre proposant divers onglets. Notamment :

- **L'onglet Document** affiche le texte original en colorisant les entités nommées reconnues
- **L'onglet Fiche** donne accès à la fiche documentaire, qui regroupe : la liste des versions de document avec la possibilité d'effectuer des comparaisons ; l'affichage des méta-données sous forme normalisée et paramétrable ; et l'identification du document (Lire ci-dessous le paragraphe « Fiches documentaires et comparaison de documents »).

Autres fonctions

Certaines fonctions d'aide à la lecture qui ont été spécialement mises au point pour des secteurs professionnels spécifiques, sont susceptibles d'intéresser d'autres applications. Par exemple :

Marquage des phrases typiques
Cette fonction permet de reconnaître des séquences dénotant des types d'information tels que annonces, conclusions, liens causaux, phrases soulignées, phrases définitoires et causales dans les articles scientifiques, etc.

Ces informations sont repérées par le biais de marqueurs définis par des expressions linguistiques et par des règles contextuelles vérifiant l'emplacement et l'enchaînement des expressions. Il est ainsi possible de créer automatiquement des fiches de synthèse, facilement paramétrables.

Résumés, colorisation et navigation
Ces fonctions, basées sur une visualisation particulière du même format XML que celui utilisé pour l'indexation de type 2 et la catégorisation, sont ici exécutées en temps réel.

- Construction automatique de résumés de documents par assemblage de parties extraites telles que entités nommées, phrases importantes, etc.
- Colorisation du texte selon les types d'entités ou la nature des passages, attirant l'attention sur les parties jugées importantes d'un document
- Navigation rapide entre le résumé et le texte complet.

Besoins spécifiques
Selon les types de corpus et les attentes des utilisateurs du système de recherche, qu'ils soient professionnels, spécialistes ou néophytes, des règles spécifiques peuvent être mises au point pour faciliter la localisation des informations et leur lecture.

Attention :
Certaines fonctions ne sont pas directement accessibles via l'interface standard de Lingway KM, mais sont utilisables à travers l'API lors de la mise en œuvre d'une application intégrée.

Fiches documentaires et comparaison de documents

Lingway KM permet l'affichage d'une fiche documentaire associée à chaque document.

Cette fiche a trois parties, présentant :

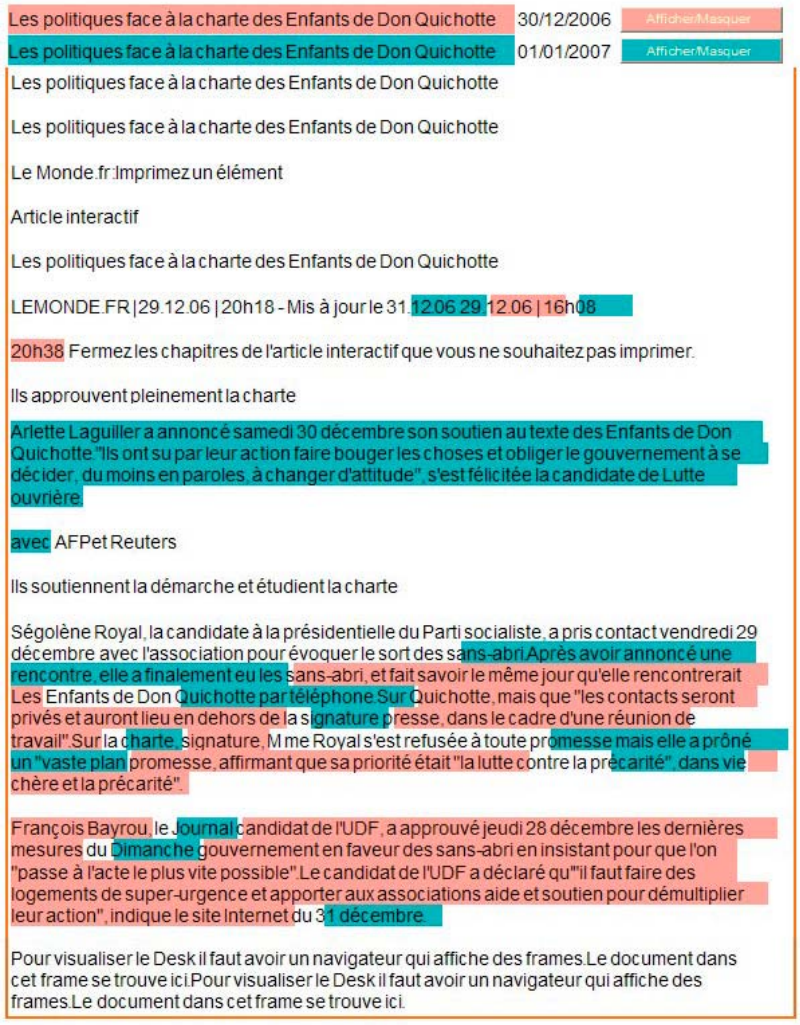
- les versions du document (vdocs), avec comparaison possible entre les versions
- la table des méta-données du document (au style Dublin Core)
- l'identification du document.

Comparaison textuelle des versions d'un document

Chaque vdoc est affichée avec sa date, son créateur et son titre.

La sélection de deux versions permet leur comparaison

L'illustration suivante montre le résultat de la comparaison textuelle entre les deux versions de documents cochées ci-dessus, colorisant les passages attribués à l'une ou l'autre des versions.



Les boutons Afficher/Masquer permettent d'appréhender immédiatement les changement apportés au texte initial

Comparaison graphique des versions d'un document

Le choix « comparer les fichiers source » (plutôt que « comparer le contenu textuel ») permet la visualisation graphique, parallèle, des deux versions sélectionnées.

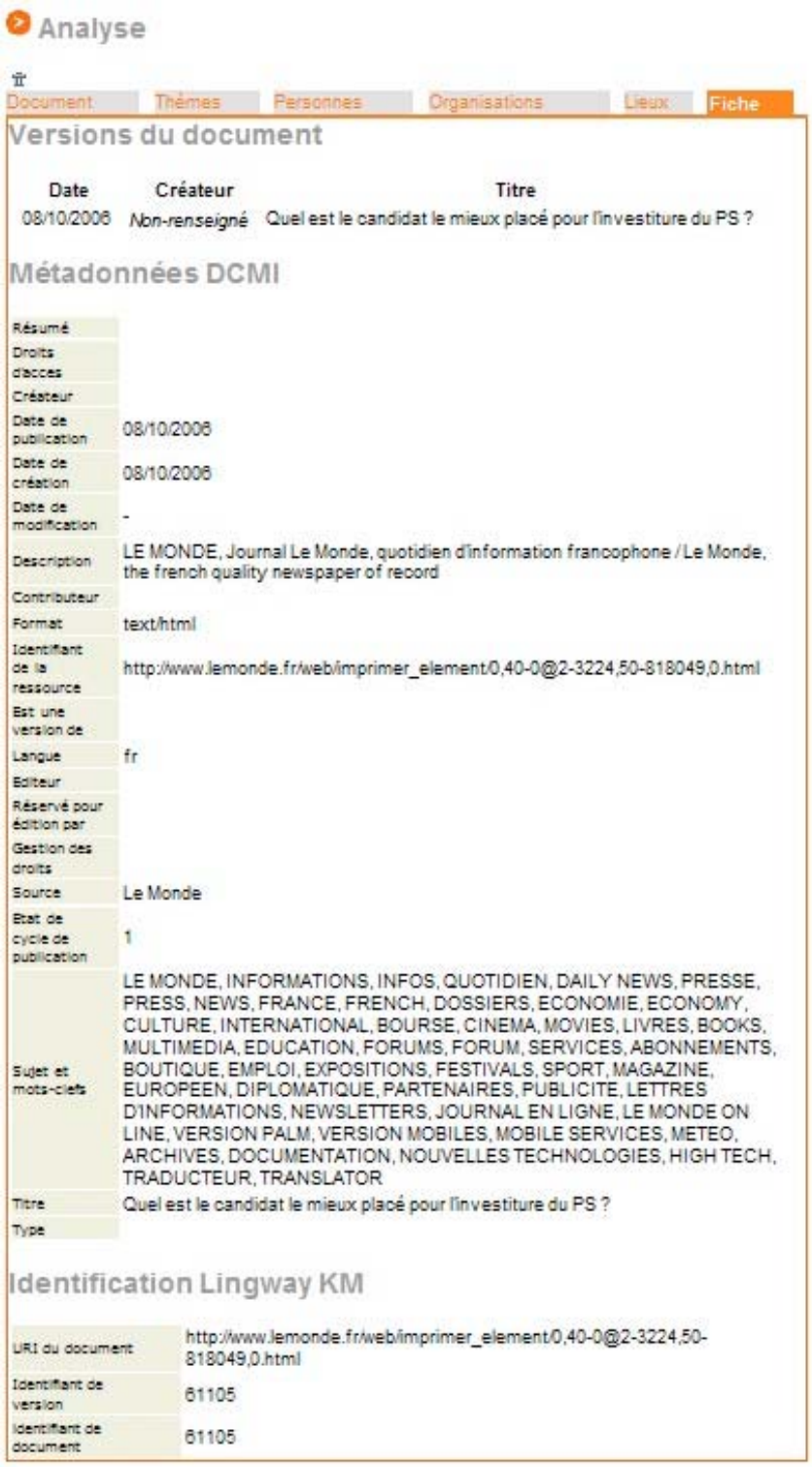
Le choix « comparer les fichiers source » est sélectionné

Après validation, le double affichage suivant apparaît.

Ce type de présentation visualise clairement non seulement les modifications effectuées, mais aussi le contexte des textes modifiés

Méta-données Dublin Core et identification document

La fiche de synthèse d'une vdoc, affichable à partir d'un simple clic depuis la liste des résultats d'une recherche, contient le tableau des méta-données au standard Dublin-Core (DCMI) et l'identification du document.



Analyse

Document | Thèmes | Personnes | Organisations | Lieux | Fiche

Versions du document

Date	Créateur	Titre
08/10/2006	Non-renseigné	Quel est le candidat le mieux placé pour l'investiture du PS ?

Métadonnées DCMI

Résumé	
Droits d'acces	
Créateur	
Date de publication	08/10/2006
Date de création	08/10/2006
Date de modification	-
Description	LE MONDE, Journal Le Monde, quotidien d'information francophone / Le Monde, the french quality newspaper of record
Contributeur	
Format	text/html
Identifiant de la ressource	http://www.lemonde.fr/web/imprimer_element/0,40-0@2-3224,50-818049,0.html
Est une version de	
Langue	fr
Editeur	
Réservé pour édition par	
Gestion des droits	
Source	Le Monde
Etat de cycle de publication	1
Sujet et mots-clefs	LE MONDE, INFORMATIONS, INFOS, QUOTIDIEN, DAILY NEWS, PRESSE, PRESS, NEWS, FRANCE, FRENCH, DOSSIERS, ECONOMIE, ECONOMY, CULTURE, INTERNATIONAL, BOURSE, CINEMA, MOVIES, LIVRES, BOOKS, MULTIMEDIA, EDUCATION, FORUMS, FORUM, SERVICES, ABONNEMENTS, BOUTIQUE, EMPLOI, EXPOSITIONS, FESTIVALS, SPORT, MAGAZINE, EUROPEEN, DIPLOMATIQUE, PARTENAIRES, PUBLICITE, LETTRES D'INFORMATIONS, NEWSLETTERS, JOURNAL EN LIGNE, LE MONDE ON LINE, VERSION PALM, VERSION MOBILES, MOBILE SERVICES, METEO, ARCHIVES, DOCUMENTATION, NOUVELLES TECHNOLOGIES, HIGH TECH, TRADUCTEUR, TRANSLATOR
Titre	Quel est le candidat le mieux placé pour l'investiture du PS ?
Type	

Identification Lingway KM

URI du document	http://www.lemonde.fr/web/imprimer_element/0,40-0@2-3224,50-818049,0.html
Identifiant de version	61105
Identifiant de document	61105

Fiche de synthèse avec méta-données DCMI
et identification de la vdoc

Alertes

A partir de la version 3.6.1, la fonction Alertes permet d'enregistrer une requête sur la base, avec les mêmes possibilités d'interrogation qu'à partir de l'interface (expression en langage naturel ou booléen, choix des sources, etc.). Il est possible de "rejouer" cette requête régulièrement (tous les jours, toutes les semaines ou à une autre périodicité).

Si de nouveaux documents correspondant à la requête enregistrée arrivent dans la base, un email proposant de les consulter est envoyé au possesseur de l'alerte. Un utilisateur peut définir plusieurs alertes.

Gestion documentaire

Afin de faire bénéficier certaines applications particulières, notamment de workflow, des fonctions de recherche sémantique, de catégorisation et de présentation des résultats, Lingway KM intègre un noyau de gestion documentaire.

Via une API Java ou un web service, il est possible d'intégrer facilement dans Lingway KM des bases documentaires en partie structurées, pour effectuer des recherches combinées plein-texte et par descripteurs ; ajouter, supprimer et modifier des documents ; gérer un cycle de publication des documents.

Intégration de bases documentaires

La vdoc, version de document, est une liste de champs textuels ou non textuels, qui sont dans ce cas remplis (plutôt que extraits du contenu d'un fichier physique) en fournissant directement la valeur des champs lors de l'ajout des documents via l'API.

Il est ainsi facile d'intégrer une base bibliographique en XML avec champs description, origine, date, titre, etc., sans stockage des fichiers physiques source ; ou encore une base contenant des documents physiques (fichiers) associés à des méta-données non stockées dans les documents originaux.

Edition de documents via API ou Web Service

Le noyau de gestion documentaire de Lingway KM met à disposition au travers d'une API ou d'un web service, des fonctionnalités :

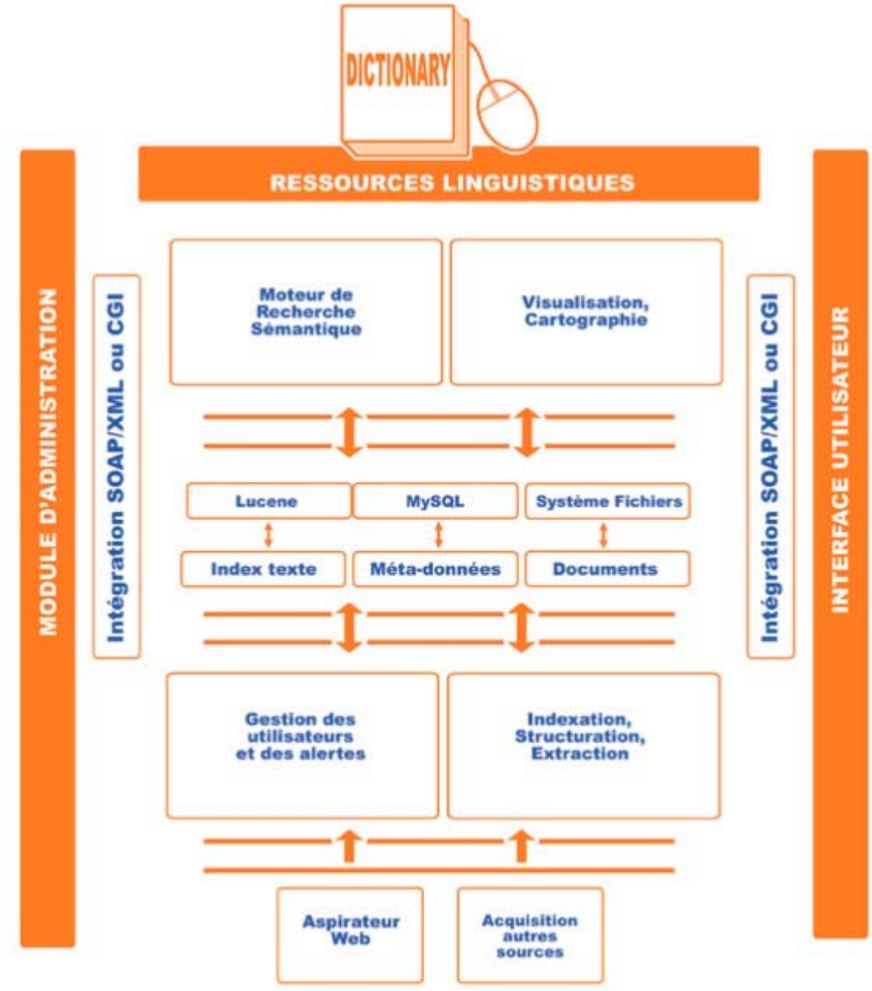
- de déclaration, paramétrage, suppression d'applications et de sources d'applications
- d'ajout, suppression et modification de documents
- de recherche combinée plein-texte et par descripteurs (cf. Indexation manuelle, page 17).

Cycle de publication

Les vdocs sont associées à des états de workflow, dont la liste est paramétrable (par défaut : Brouillon, A Valider, En ligne, Refusé, Archivé). Cette fonctionnalité est utilisable de manière complètement paramétrable via l'API. Un composant externe peut ainsi utiliser Lingway KM pour stocker / gérer une base documentaire avec workflow.
Pour une utilisation sans composant tiers, l'interface peut être paramétrée pour :

- afficher les états associés aux documents
- intégrer l'état comme critère de recherche (boutons pour chercher dans les documents ayant tel ou tel état)
- fournir une interface de validation / indexation manuelle, qui permet de faire avancer les documents dans le cycle de publication.

Architecture de la plate-forme



Vue conceptuelle montrant les différentes fonctions de Lingway KM, ses ressources linguistiques, moteurs et base de données, ainsi que les interfaces - humaines ou API

Pour les briques « base de données » et « indexation full-text », la plate-forme Lingway KM utilise respectivement, en standard, les Open Sources MySQL et Lucene. Cependant le logiciel peut se connecter à d'autres bases de données et aussi s'interfacer avec divers moteurs d'indexation du marché. Lire à ce sujet le paragraphe « Intégration à un contexte existant », page 40.

Le moteur de recherche Lingway ne pose pas de contraintes fortes sur le modèle d'indexation. Il peut s'adapter soit à des bases indexées de manière traditionnelle (liste inverse classique), soit au contraire à des textes ayant subi un traitement plus élaboré lors de l'indexation (*lemmatisation*, indexation par entités nommées, par termes extraits, etc.).

Pour les ressources linguistiques, Lingway KM s'appuie sur de puissants composants sémantiques dont les fonctionnalités principales sont présentées ci-après.

Composants sémantiques

Cœur de la technologie Lingway, les composants d'analyse automatique de la langue Lingway sont au nombre de 5. Grâce à leur architecture ouverte, ils s'intègrent facilement aux applications métier externes.

Lingway **Proxy**

Composant de recherche approximative de noms. Il permet l'accès à une liste de noms ou de libellés en orthographe approximative, corrigeant les erreurs typographiques, phonétiques, de soudure ou de coupure.

Lingway **Fulty**

Composant de recherche sémantique multilingue en « texte intégral ». Il peut être utilisé de manière autonome, mais est conçu pour s'interfacer avec tout indexeur en « texte intégral » disposant d'un langage de recherche booléen.

Des connecteurs sont actuellement disponibles pour Exalead, Hummingbird et Oracle.

Lingway **Xtirp**

Composant d'analyse transformant des textes non structurés en structures XML. Il enrichit les vdocs par des marqueurs indiquant la structure générale du texte, repérant des entités nommées (personnes, organisations, produits, lieux ou autre entité paramétrée) ; des index thématiques (correspondant à des descripteurs libres ou contrôlés) ; des phrases importantes sous un point de vue donné, pour en faciliter la lecture. Il permet également de constituer une base des données extraites.

Possibilités : créer un document à partir de plusieurs ; créer des fiches de synthèse ; mettre des documents à la norme Dublin Core (web sémantique).

Lingway **Tacsy**

Composant de recherche et de classification de phrases, questions ou énoncés courts, verbatims par exemple, dans une *taxonomie* (*plan de classement*). Exemple d'application : codification dans une nomenclature métier pour un site de commerce électronique.

Ce composant sémantique, qui fonctionne dans certaines applications métier bâties sur Lingway KM, est en cours d'intégration dans la plate-forme standard.

Lingway NLP

Module de gestion des ressources linguistiques (terminologies, *lexiques*, etc.), Lingway NLP[3] gère les données des modules précédents. Les services linguistiques qu'il délivre s'appuient sur une vingtaine de composants de traitement du langage naturel.

Soubassement de la plate-forme, Lingway NLP regroupe un ensemble de dictionnaires et grammaires électroniques, dont un dictionnaire multilingue de plus de 400.000 mots représentant 150.000 *concepts*, et de puissants services d'analyse automatique de la langue. Les services linguistiques s'articulent autour d'une vingtaine de composants de traitement du langage naturel.

---

[3] NLP en anglais, pour Natural Language Processing – TAL en français, pour Traitement Automatique de la Langue.

Dictionnaires et réseau sémantique

Tout système basé sur le traitement du langage naturel nécessite des connaissances linguistiques pour fonctionner. Ces connaissances peuvent être amenées à évoluer en fonction des applications et au cours du temps, pour combler d'éventuelles lacunes.

Les dictionnaires de Lingway ont une couverture suffisamment large pour aborder sans aucune modification la plupart des applications visées. De plus, le système est suffisamment robuste pour fonctionner en présence de mots inconnus (non définis dans les dictionnaires).

Cependant, Lingway dispose de nombreux moyens pour adapter et faire évoluer ses dictionnaires, permettant d'intervenir directement sur la qualité finale des résultats.

Rôle des dictionnaires dans Lingway KM

Sans entrer dans le détail (car le sujet est très vaste), les tâches dépendant de l'usage des dictionnaires et ressources sémantiques Lingway sont les suivantes, par type de ressource utilisée.

Tâche	Ressource
Reconnaissance des noms de personnes, d'organisations, de lieux (entités nommées)	Listes
Reconnaissance de *thèmes* (pour l'extraction de termes)	Patrons linguistiques
Interrogation en langage naturel (monolingue ou cross-language)	Réseau sémantique de dictionnaires.

Nous décrivons ci-après les fonctions des différents dictionnaires Lingway, leur structure à trois niveaux (dictionnaires de base et spécialisés), ainsi que les outils permettant de les mettre au point, selon leur type.

Types de dictionnaires

Les dictionnaires de Lingway KM sont de 3 sortes complémentaires :

- **Les dictionnaires de base.** Dictionnaire de base livré pour chaque langue avec le produit, correspondant à une description assez générale dans plusieurs langues. Il comprend environ 80.000 mots par langue.
- **Les dictionnaires spécialisés.** Dictionnaires propres à un domaine, comme par exemple le droit, le sport, les sciences de la vie, etc. Ces dictionnaires viennent compléter le dictionnaire standard, comprenant en général quelques milliers de mots.
- **Les dictionnaires d'adaptation.** Mots, locutions et expansions propres à une application donnée, venant compléter ou modifier les dictionnaires précédents. Ils comprennent en général quelques centaines de définitions.

Les deux premiers types de dictionnaires sont construits sur le modèle à trois niveaux décrit ci-après, et sont régulièrement mis à jour mis à jour par Lingway.

Modèle à trois niveaux

Les données lexicographiques des dictionnaires de base et dictionnaires de spécialités sont codées selon trois niveaux :

1. **Au niveau morphologique** sont codées les informations relatives à la catégorie syntaxique du mot, ses caractéristiques flexionnelles (genre, nombre, mode de flexion) et son poids

sémantique, calculé en fonction de sa polysémie (plus un mot a de sens possibles, moins il est pertinent dans une application de recherche).

2. **Au niveau sémantique**, on distingue les sens possibles des mots. Ainsi une même entrée de la couche morphologique peut être reliée à plusieurs sens (les sens sont dépendants de la langue).
3. **Au niveau conceptuel**, les différents sens sont reliés à un concept, cette fois indépendant de la langue. Les concepts sont également porteurs d'attributs sémantiques (appartenance à un domaine ou une classe) et de relations sémantiques (liens de type générique-spécifique, associé, dérivé).

Ce modèle est « nativement » multilingue : les correspondances entre langues découlent de la connexion des sens à la couche conceptuelle.

Dictionnaire d'adaptation, personnalisable

Le dictionnaire générique standard de Lingway KM peut être complété ou adapté ponctuellement par un « dictionnaire d'adaptation ». Ce dernier utilise un modèle très simplifié qui permet de modifier l'utilisation des dictionnaires sans nécessiter de connaissances particulières en linguistique ou en ingénierie documentaire. La mise en œuvre des modifications, immédiate, ne nécessite aucun traitement particulier –compilation ou autre.

Les adaptations réalisées sont compatibles avec les nouvelles versions du dictionnaire générique de Lingway KM. Ceci signifie qu'aucune révision n'est nécessaire lors de la livraison d'une nouvelle version de dictionnaire générique.

La taille importante du dictionnaire générique et la puissance des règles d'extraction de termes permettent dans la plupart des cas de démarrer une application Lingway KM immédiatement. Le dictionnaire d'adaptation permet ensuite d'affiner le dictionnaire par rapport aux spécificités d'une application. Voir le paragraphe « Réglages des dictionnaires », page 36.

Fonctionnement interne du flux de recherche

S'appuyant sur les fonctions et modules précédemment exposés, ce chapitre détaille le fonctionnement d'un flux de recherche – de la requête saisie par l'utilisateur à la présentation des résultats, passant par les étapes d'analyse de la requête, de construction booléenne, d'expansion ou de dégradation de la requête via l'exploitation du *réseau sémantique* multilingue.

Une compréhension globale de ce fonctionnement est notamment utile au paramétrage des stratégies de recherche, abordé page 35, et au réglage des dictionnaires, décrit page 36.

Analyse de la requête

Le module linguistique produit une représentation interne de la requête de l'utilisateur en réalisant une analyse complète, en trois niveaux successifs :

- **Le niveau morphologique** reconnaît les mots simples et mots composés sous toutes leurs formes (genre, nombre, verbes conjugués, etc.). La reconnaissance de mots composés met en œuvre des traitements linguistiques, comme la dé-coordination (par exemple, les mots composés « ligne à haute tension » et « ligne à basse tension » sont identifiés dans la requête « lignes à haute et basse tension »).
- **Le niveau syntaxique** identifie la structure des requêtes et les rôles des mots les uns par rapport aux autres. Une des conséquences de ce traitement est l'attribution de « poids » aux différents mots en fonction de leur rôle syntaxique dans la question.
- **Le niveau sémantique** identifie le sens des mots. Ceci permet par la suite la recherche par synonymes ou mots proches ou encore dérivés (par exemple *«généticien»* - *«génétique»*). Ces mots constituent « l'expansion sémantique » du mot initial. Cette phase d'identification du sens de chaque mot est importante, car l'expansion de la question vers d'autres mots dépend du sens reconnu. Par exemple, le mot *«avocat»* dans *«avocat spécialiste de la fiscalité»* pourra être expansé *«juriste»* ou *«lawyer»* (si l'on valide la recherche multilingue), alors que *«recette de salade d'avocat»* sera expansé vers *«fruit exotique», «cuisine»,* etc.

Ce module exploite les 150.000 concepts, les 6 langues et le réseau sémantique de la base de données linguistiques.

Interprétation des requêtes

Sur la base de l'analyse linguistique, Lingway KM propose une série d'interprétations de la requête de l'utilisateur. Ces interprétations comportent :

- soit une expansion en une des séries de mots de sens équivalent ou proche des mots de l'utilisateur (en fonction de l'analyse sémantique), classées en fonction des sens du mot initial, puis de leur distance sémantique à ce mot
- soit une dégradation (ou « généralisation ») de la requête initiale. Par exemple, on propose une requête analogue à «production de fraises» à partir de la requête «production de fraises en Limousin», ou «RTT et emploi des jeunes» à partir de «conséquence de la RTT sur l'emploi des jeunes».

L'interprétation des questions dépend des modes de recherche possibles :

- ◊ en langage naturel
- ◊ en mode « avancé » ou « pseudo booléen ».

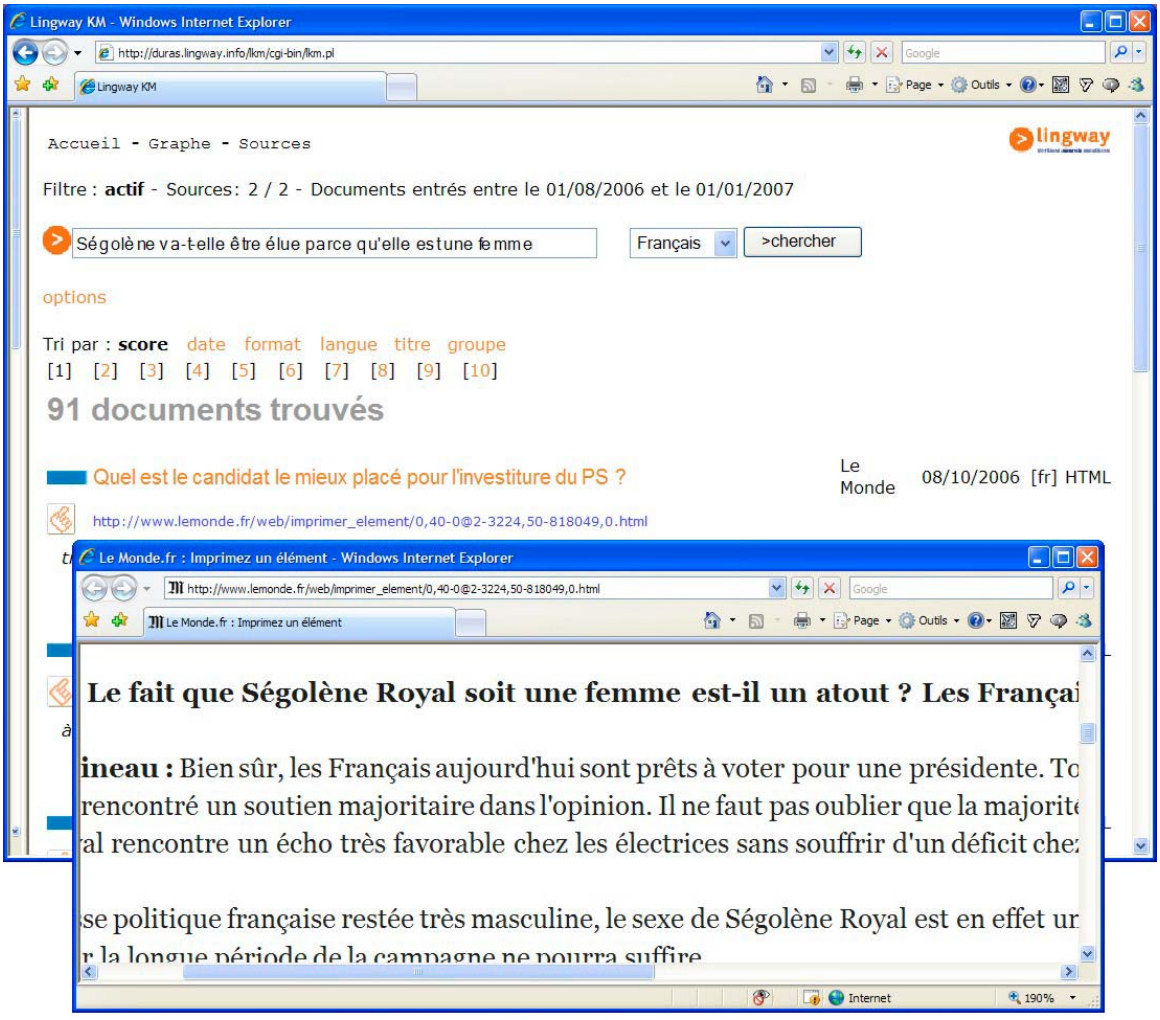
Interprétation en langage naturel

Dans ce mode, l'utilisateur entre une phrase en langage naturel et le moteur la traduit dans une requête booléenne en utilisant les mécanismes d'expansion et de dégradation (en clauses décrites ci-dessus). Cela permet de réduire considérablement le silence, en élargissant la question à des formulations voisines.

Conséquence, la saisie de « *veille concurrentielle en ligne* » peut faire apparaître des résultats répondant à « *veille Internet* » par expansion de « en ligne » sur « *internet* » et abandon de l'adjectif « *concurrentiel* ».

A partir de l'analyse de la question, le moteur construit une requête documentaire qui dépend du modèle d'indexation et de la syntaxe particulière de l'indexeur (dans le cas où l'on utilise un connecteur vers un autre système de gestion documentaire).

L'image suivante montre le résultat d'une recherche en langage naturel sur une base d'articles de presse, la question posée étant *« Ségolène va-t-elle être élue parce qu'elle est une femme ».*



91 documents sont proposés en réponse. Certains hits sont dérivés de la requête initiale et des articles en anglais apparaissent aussi

Dans l'exemple de cette recherche, le document affiché, qui a le score maximal dans la liste, apporte précisément un élément de réponse à la question posée, l'exprimant dans des termes différents ; notamment, le mot « élue » de la requête a permis d'utiliser le mot « voter » parmi les résultats possibles ; « chance » a généré « atout ».

Interprétation en mode avancé ou « pseudo booléen »

Dans ce mode, l'utilisateur a plus de contrôle sur la formulation de la question, en interdisant notamment la dégradation. Le moteur passe automatiquement en mode avancé dès qu'un opérateur booléen est entré (AND pour ET logique, OR pour OU logique, guillemets "..." pour garder un mot à l'identique). Les parties séparées par un opérateur logique ne sont pas dégradées, mais l'expansion est en revanche réalisée.

Dans ce mode, la requête **«** veille AND concurrentiel AND "en ligne" » gardera toujours les 3 termes, mais on pourra la transformer par exemple en **«** veille AND concurrent AND "en ligne" **».**

Fonctionnalité « cross-language »

La fonction de CLIR (pour « Cross-Language Information Retrieval ») permet de générer des interprétations dans une langue différente de la langue de la requête. Cela permet d'interroger un fonds documentaire à partir d'une langue différente de la langue utilisée par les utilisateurs pour leurs requêtes. Cette fonction peut s'utiliser avec plusieurs langues cibles.

Par exemple, on peut interroger en français une base contenant des documents français et anglais. Lingway KM interroge alors les documents français avec des mots français, les documents anglais avec des mots anglais de même sens.

Ainsi, à une requête formulée en français, certains résultats en anglais peuvent apparaître de façon tout à fait cohérente, apportant souvent des informations inédites dans notre langue.

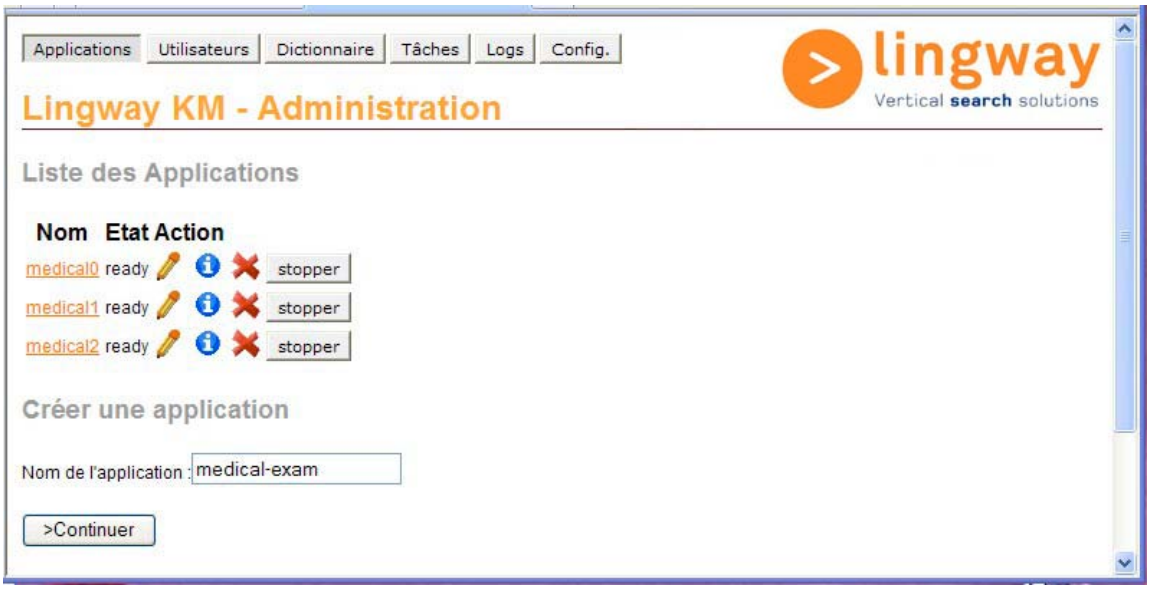
Paramétrages administrateur

Gestion des applications de recherche

Sur la page d'accueil de l'interface d'administration est visible l'ensemble des applications créées avec, pour chacune d'elle, son état : « ready », lorsque l'application est prête ; « declared », si l'initialisation n'est pas terminée ; « frozen », si elle a été interrompue.

Trois types d'actions peuvent être effectués sur ces applications :

- en cliquant sur l'icône , on peut accéder aux propriétés de l'application
- en cliquant sur l'icône , on accède au paramétrage avancé de l'application
- en cliquant sur l'icône , il est possible de supprimer une application.



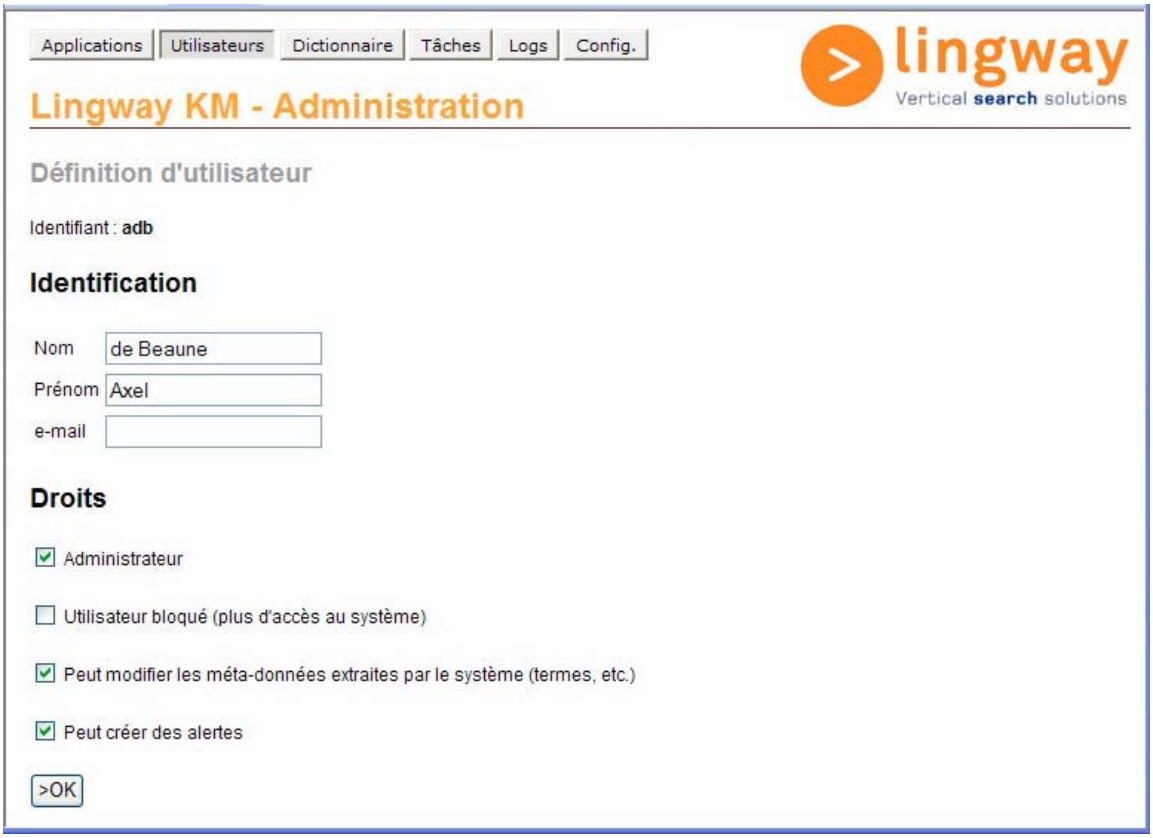
Interface et fonctions d'administration des applications de recherche

Ce qui est paramétrable :

- les sources locales à traiter
- l'aspiration de sources externes pour le crawler, précisant :
  - ◊ la fréquence d'aspiration
  - ◊ les limites éventuelles (profondeur de recherche, etc.)
  - ◊ les sites et domaines associés à indexer
  - ◊ les filtres sur les types de fichiers acceptés, etc.
- les langues à utiliser par le système de recherche (interrogation cross-language)
- l'extraction :
  - ◊ types d'entités nommées souhaitées
  - ◊ filtres sur la forme des descripteurs thématiques (proportion d'unitermes, de bigrammes, trigrammes).

Gestion des utilisateurs

Lingway KM inclut un module gérant l'identification et les droits des utilisateurs, notamment ceux de modifier des méta-données générées par l'analyse linguistique et de créer des alertes.



Edition des propriétés associées à un utilisateur

Paramétrage des stratégies de recherche

De nombreux paramètres sont disponibles pour agir sur le comportement du moteur et les interprétations : privilégier les sens de domaines donnés, fixer des distances maximales pour les mots proches, parcourir ou non certaines relations du réseau sémantique, etc. Ces stratégies peuvent être différentes selon les champs à interroger. Typiquement, on peut vouloir élargir le champ de la requête (c'est-à-dire utiliser des mots plus génériques) pour le titre, alors que l'on peut privilégier la recherche de mots plus précis pour le corps du document.

Accès aux logs

L'interface d'administration permet d'accéder aux logs de questions pour chaque application (onglet "Questions" du menu spécifique à une application donnée), ainsi qu'au log des tâches en cours/programmées de Lingway KM (onglet "Logs" du menu principal de l'interface).

Le sous-menu applicatif

Réglages des dictionnaires

Deux interfaces sont disponibles, « simple » et « complète », pour agir à différents niveaux sur les dictionnaires :

- L'interface complète –généralement utilisée par Lingway– donne accès à un mode expert qui sert à modifier le fonctionnement des dictionnaires génériques en jouant de façon précise sur les mots, expansions et relations du réseau sémantique.
- L'interface simple –généralement utilisée par l'administrateur d'une application– permet d'effectuer très facilement des actions locales pour adapter le dictionnaire générique et l'affiner par rapport aux spécificités d'une application. Les modifications se font en temps réel et ne nécessitent aucune intervention sur le serveur. Le niveau d'information linguistique à entrer est minimal.

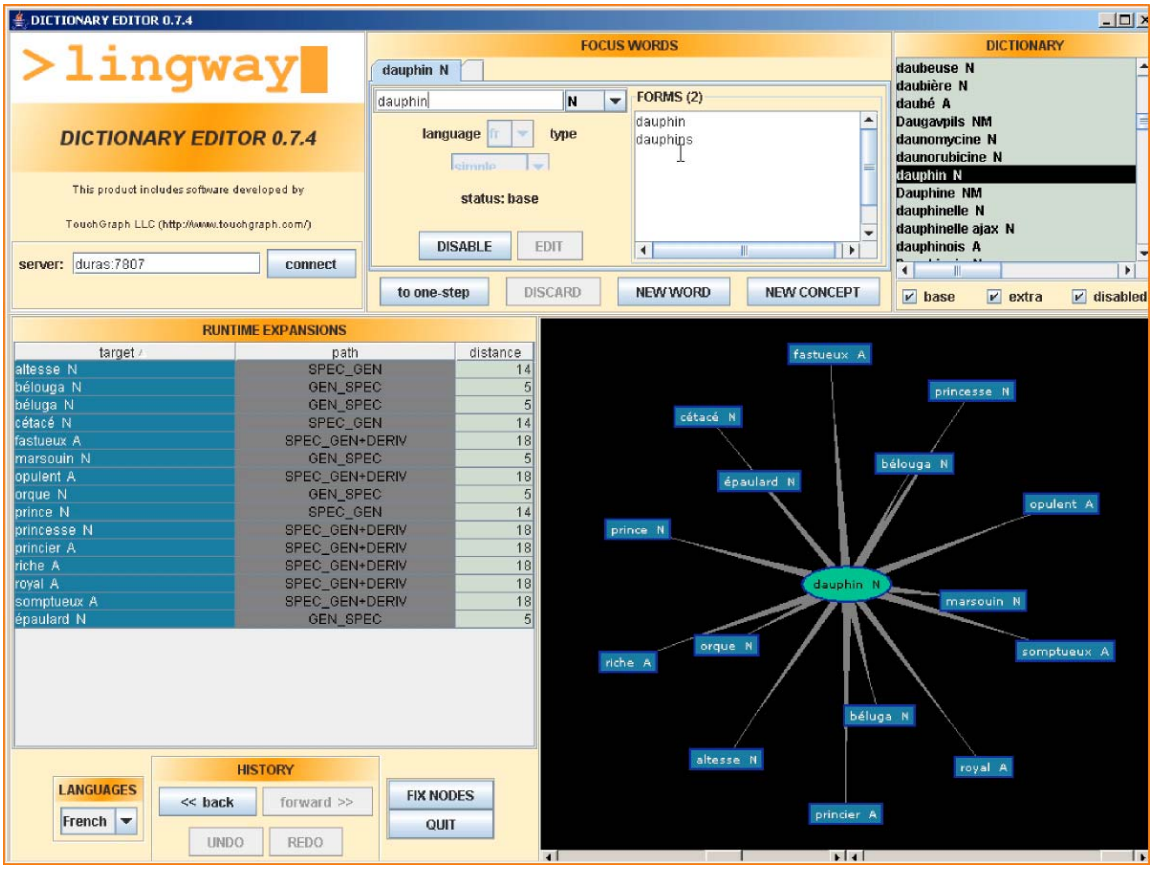
Interface complète - mode expert

Des vues différentes proposent deux façons d'agir sur le dictionnaire :

◊ La vue « expansions » permet d'influer sur le dictionnaire en fixant les expansions résultant des cheminements dans le réseau sémantique (action aval)

◊ La vue « réseau » permet de modifier directement les relations sémantiques entre concepts et mots du dictionnaire (action amont).

La vue « expansions »



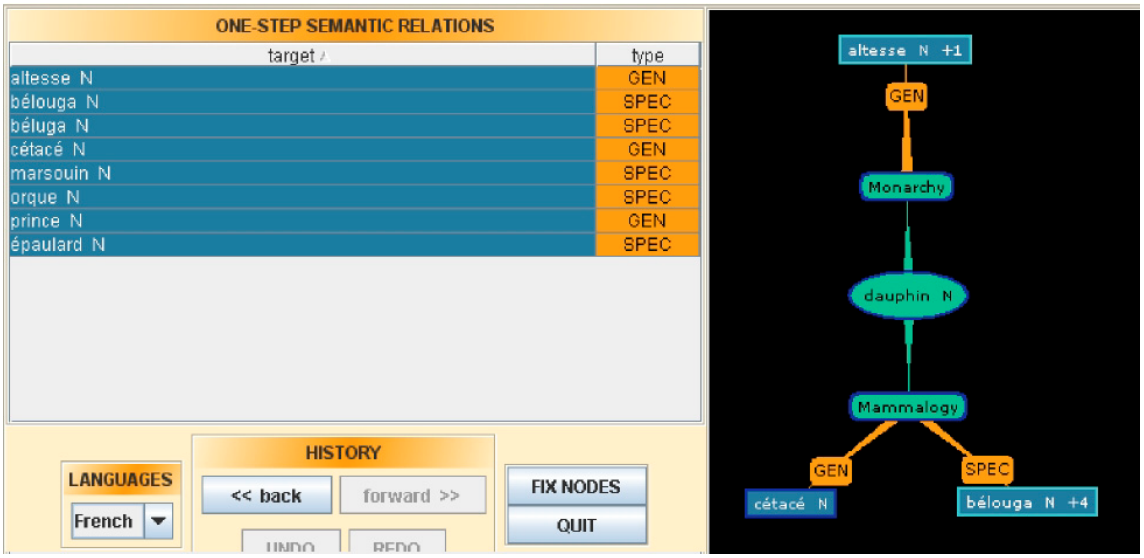
Le mot sélectionné dans le dictionnaire apparaît au centre du graphique, alors que les distances sémantiques aux « mots liés » sont listées à gauche

Les relations entre un mot sélectionné et ses expansions sont représentées graphiquement par des rayons dont la taille est fonction de la distance sémantique.

Il est possible d'agir en cliquant sur les liens du graphique ou sur un élément de la liste, par exemple pour désactiver des expansions. Cependant, ce mode d'action très rapide n'est pas conçu pour modifier précisément le réseau sémantique, les mots seuls étant pris en compte et non pas les liens entre concepts.

Pour travailler sur la structure du dictionnaire, il faut passer en mode « réseau » (ce qui se fait en cliquant sur le bouton « to one-step » de la figure précédente).

La vue « réseau »



Exemple de vue « réseau » avec liste des types de relations sémantiques

Cette vue donne accès au réseau sémantique proprement dit, visualisant à la fois les liaisons entre mots et concepts, et les relations sémantiques (typées) entre concepts.

Pour assurer une bonne lisibilité, l'affichage des expansions est limité à un pas dans le réseau du dictionnaire. La figure ci-dessus montre pour ce « pas de réseau » les principales relations entre mots et concepts sur lesquelles on peut agir :

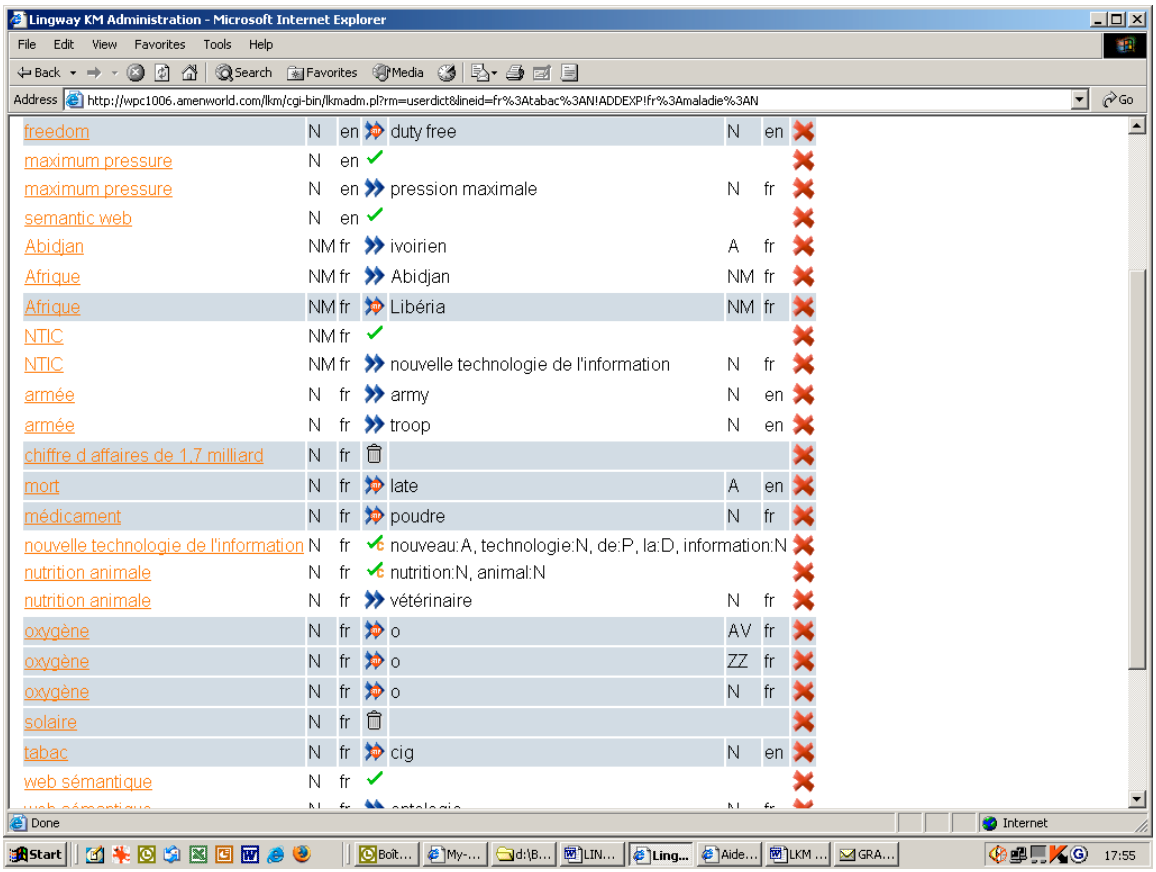
- la relation de synonymie (SYN)
- la relation d'association (ASSOC)
- la relation générique-spécifique (SPEC)
- la relation de dérivation (DERIV).

Ainsi l'interface complète du mode expert permet-elle –en cliquant sur les liens des graphiques– de créer/supprimer des sens, de connecter/déconnecter des mots à d'autres sens et de définir tout type de relation entre vocables.

Les mots et relations supprimés apparaissent en rouge (non physiquement détruits, mais seulement désactivés logiquement) ; les éléments ajoutés apparaissent en jaune… D'autres couleurs facilitent la lecture et l'interprétation des vues de l'interface expert.

Interface simple - mode rapide

Cette interface permet d'adapter un dictionnaire générique à une application donnée. Elle est accessible via le bouton "Dictionnaire" de l'interface d'administration, livrée avec Lingway KM.



Interface d'adaptation du dictionnaire. On note que la création d'expansion peut être multilingue

Le dictionnaire d'adaptation permet les actions suivantes, au seul niveau des mots :

Action	Effet
Ajouter au dictionnaire	Crée un nouveau mot dans le dictionnaire d'adaptation.
Ajouter une expansion	Ajoute un mot cible dans l'expansion de l'entrée. Ce mot cible sera donc considéré lors de la recherche à partir de cette entrée.
Définir la composition	Décrit un mot composé : pomme:N, de:P, terre:N Cette description indique que le mot composé *pomme de terre* est formé des noms *pomme* et *terre* reliés par la préposition *de*.
Supprimer du dictionnaire	Supprime un mot.
Supprimer une expansion	Le mot cible n'apparaîtra plus dans l'expansion du mot de départ.
Ne plus "*expanser*"	Le mot de départ est conservé, mais n'est plus du tout expansé.

Choix, mise en œuvre, adaptation métier et intégration

Versions de la plate-forme

La version choisie de la plate-forme Lingway KM –Light, Standard ou Premium– détermine les fonctionnalités disponibles.

Fonctions	dont	Version Light	Version Standard	Version Premium
Acquisition	(filtres, crawler web…)	✓	✓	✓
Recherche	(langage naturel, opérations booléennes, adaptation dictionnaire…)	✓	✓	✓
Extraction et navigation	(méta-données applicatives, entités nommées, extraction de thèmes…)		✓	✓
Aide à la lecture	(résumé termes significatifs, clustering des réponses, cartographie des thèmes…)			✓
Profils de veille et alertes				✓

Au choix, le logiciel peut être acquis sous licence ou loué à l'année.

Installation standard

La mise en œuvre de Lingway KM est immédiate si l'on utilise l'interface standard.

Bases de données et moteur « full-text »

Les bases de données (base relationnelle et base d'index) sont gérées en standard par des logiciels open-source (Lucene et MySQL). En cas de copie locale des sites, ceux-ci sont gérés par le système de fichiers du système d'exploitation.

Paramétrage fin

Un certain nombre de paramètres sont modifiables au niveau Administrateur ; d'autres réglages, plus fins, sont possibles.

Adaptation métier

Changement d'interface

Le développement d'interfaces personnalisées est particulièrement simple grâce à l'architecture en « web services » de Lingway KM.

Au delà de la simple adaptation de la feuille de style, des exemples d'applications utilisant des interfaces particulières au dessus de Lingway KM sont visibles sur les sites web de certains clients de Lingway.

Adaptation du dictionnaire

Le dictionnaire standard, prêt à l'emploi, est conçu pour répondre à des besoins généraux tels que par exemple la recherche d'informations journalistiques. Ainsi qu'il est décrit au chapitre précédent, un dictionnaire « d'adaptation » paramétrable peut être utilisé pour modifier et compléter le dictionnaire standard.

Il existe déjà des dictionnaires spécialisés pour des secteurs professionnels déterminés, tels que le secteur de l'emploi, le domaine médical, celui de la propriété intellectuelle et le e-commerce.

Chargement de listes, thésaurus et plans

Selon les besoins de catégorisation des documents d'un corpus, liés à un secteur professionnel particulier ou à une application particulière, il est possible de paramétrer les listes d'entités à extraire afin de générer des index thématiques (voir page 16) ou d'imposer des termes via thésaurus (voir page 16).

De plus, des plans de classement peuvent être chargés à partir de la version 3.6.1 de Lingway KM.

Intégration à un contexte existant

API

L'intégration dans un système d'information s'effectue de manière particulièrement souple et standard, car Lingway KM se présente sous forme de serveur communiquant par défaut en XML, via le protocole SOAP. L'intégration est donc facilitée dans le cas d'applications Web (ou autres) existantes, quelle que soit la technologie sur laquelle elles sont bâties : ASP, JSP/Servlet, PHP ou CGI.

En parallèle, une API Java est disponible.

Indexeur externe

Le composant de recherche sémantique, multilingue, peut s'interfacer à tout indexeur « texte intégral » disposant d'un langage de recherche booléen, ou fonctionner en mode autonome.

Kits d'intégration

Afin de faciliter la tâche d'intégration du moteur de Lingway KM à un contexte existant, des kits sont prévus – manuels et autres ressources – regroupant l'ensemble des moyens nécessaires.

Méthodologie

Pour compléter sa gamme de solutions, Lingway a inclus à son offre une prestation de service visant à accélérer la mise en œuvre.

Cette prestation, qui repose sur une expérience enrichie par de nombreuses interventions et sur l'utilisation d'une méthodologie éprouvée, peut se décliner selon les besoins, pour réaliser rapidement de nouveaux moteurs de recherche spécialisés.

Concrètement, cette méthodologie –associée à la plate-forme– met en jeu des fonctions et des outils qui adressent 3 domaines et modélisent trois types de connaissances nécessaires à une application :

- la connaissance des documents
- la connaissance des usages
- la connaissance des domaines.

1- Adaptation aux types de documents

- Les modules d'acquisition, indexation et extraction sont paramétrés en fonction des sources (URL Internet, Intranet, fichiers, pièces attachées à des e-mails, numérisation OCR), des formats (Word, PowerPoint, Excel, PDF, etc.), de la nature textuelle (textes longs, courts, bien ou mal structurés), de la nature linguistique (niveau de langue, précision terminologique), etc.

2- Adaptation aux conditions d'utilisation

- Le back-office est paramétré : contrôle humain de l'indexation ou indexation complètement automatique, utilisation éventuelle d'un référentiel d'indexation de type thésaurus, types de méta-données à extraire…
- Le front-office est paramétré : type de recherche (mono-champ « à la Google » ou structurée), expansion sémantique, cross-language, dialogue contextuel (« colonne de gauche »), etc.

3- Adaptation des connaissances du domaine

- Utilisation des dictionnaires électroniques généraux (plusieurs dizaines de milliers de mots simples, connectés à des centaines de milliers de sens, et autant de termes composés)
- Construction de dictionnaires métiers grâce à de puissants outils d'analyse des textes et de gestion des dictionnaires.

Caractéristiques détaillées

Données générales

Architecture	Client / Serveur
Serveur	Linux ou Windows
Client	Toute plate-forme
Mode de lancement du serveur	Automatique (*service* Windows ou *démon* Unix)

Standards d'interfaçage

Formats supportés en entrée	TXT, XML, HTML, PDF, formats Office (.doc, .xls et .ppt)
Modèle documentaire	Standard Dublin-Core
Intégration du module Client	Web service SOAP/XML
	API JAVA
Module de recherche sémantique multilingue en « texte intégral »	Utilisable en mode autonome ou interfaçable avec tout indexeur en « plein texte » disposant d'un langage de recherche booléen.
Indexeur « plein texte »	Lucene (open source) en standard Connecteurs pour Exalead, Hummingbird et Oracle.

Dictionnaires et langues

Dictionnaire standard et dictionnaire personnalisable (dictionnaires métier sur demande).

Ontologie	150.000 concepts mis en correspondance avec 6 langues (enrichissable)
Taille du dictionnaire	Environ 40 Mo (version monolingue)
	plus 20 Mo supplémentaires par langue additionnelle.
Langues actuellement traitées	L'allemand, l'anglais, l'espagnol, le français, le néerlandais, le portugais.
Mises à jour	2 modes : expert et rapide pour la personnalisation.
	Compatibilité ascendante du dictionnaire personnalisable

Pré-requis et systèmes d'exploitation

Plates-formes disponibles	Linux, Windows Plate-forme spécifique : nous consulter
Configuration minimum	2 GHz, 1 Go RAM, espace disque selon taille de la base

Performances

Recherche	De quelques requêtes par seconde à plusieurs dizaines, selon le type d'application
Indexation	2 Go / heure
Extraction	50 à 200 Mo/h selon la nature de l'extraction

Glossaire

A

Application

- Au sens Lingway KM, une application correspond au résultat de l'analyse linguistique sur un ensemble de sources de données (des fichiers sur une directory ou une liste d'URL Internet). Voir *collection documentaire*.

C

Catégorisation

Opération consistant à placer un document à une ou plusieurs places d'un *plan de classement* prédéfini.

[à distinguer de Classification ci-après].

Classification

Opération consistant à placer un document à une ou plusieurs places d'un *plan de classement*.

- Lingway KM peut effectuer une classification automatique des documents dans un plan de classement défini par l'utilisateur, en utilisant des calculs à partir de documents déjà classifiés (apprentissage).

Clustering

Le clustering ou "clusterisation" consiste à construire des groupes ("clusters") homogènes d'objets.

- Dans Lingway KM, il est possible d'effectuer un *clustering de documents* ou un clustering de termes.

Clustering de documents

Clustering de documents à partir d'un ensemble de documents non classifiés. Autrement dit, il s'agit d'une organisation automatique d'un ensemble de documents en sous-groupes.

Cette opération repose généralement sur un calcul de proximité entre documents.

Le clustering est une méthode de recherche d'informations de type bottom up (ascendante).

- Lingway KM réalise le clustering de l'ensemble de documents retrouvés suite à une requête (bouton "groupe").

Clustering de termes

Clustering de termes à partir de l'analyse de leurs co-occurrences dans un ensemble de documents.

- Dans Lingway KM, un clustering de termes est possible et le résultat est visualisable sous forme d'une carte (bouton « graphe »).

Collection documentaire

Ensemble de documents pouvant évoluer dans le temps (ajout / édition / suppression de documents), regroupés dans le but d'une activité humaine.

- Dans Lingway KM, la notion de « collection documentaire » se définit par une « Application » et correspond à un ensemble données acquises de différentes *sources*.

Concept

Un concept est un objet qui représente l'abstraction d'un terme ou plus généralement d'un ensemble de termes synonymes dans une ou plusieurs langues, cet objet étant indépendant de sa réalisation linguistique (la façon de le nommer). Cet objet sert à décrire des propriétés indépendantes de la langue (des propriétés conceptuelles). Par exemple, le concept <marteau> appartient à la classe <instruments>, et ce, indépendamment de la langue.

- Dans Lingway KM, le *dictionnaire* décrit 150.000 concepts qui sont reliés à 6 langues. Les concepts sont reliés entre eux par un ensemble de relations formant un *réseau sémantique.* Exemple : le concept n° 344 est une sorte de concept n° 765 – Le concept n° 344 se dit « fauteuil » en français et « armchair » en anglais. Le concept n° 765 se dit « meuble » en français et « furniture » en anglais.

D

Descripteur

- *Terme* ou *Entité nommée* utilisé pour caractériser (*indexation2*) un document. Un descripteur peut être *libre* ou *contrôlé*.

Descripteur libre

Descripteur choisi indépendamment de toute *liste d'autorité*.

- Lingway KM effectue une indexation en descripteurs libres et de plusieurs natures différentes : « thèmes » ou « entités nommées », tous deux extraits du texte.

Descripteur contrôlé

Descripteur choisi obligatoirement dans une *liste d'autorité,* généralement un *thésaurus*.

- Dans Lingway KM il est possible de définir une liste d'autorité. Voir *Thésaurus*, *Indexation contrôlée*.

Dictionnaire [électronique]

Base de données regroupant l'ensemble de l'information linguistique et conceptuelle nécessaire pour l'analyse des textes et des questions. Le dictionnaire comprend la description morphologique des mots, leur décomposition en sens, leur rattachement aux concepts et le réseau sémantique entre les concepts.

- Le dictionnaire générique de Lingway, comporte plus de 400.000 mots de 6 langues différentes, rattachés à 150.000 concepts.

Dictionnaire d'adaptation

Dictionnaire venant modifier le *dictionnaire générique* pour l'adapter à une application particulière.

- Lingway KM permet de gérer un dictionnaire d'adaptation dont le modèle est très simplifié par rapport au dictionnaire générique, afin d'en rendre l'utilisation très facile.

Distance sémantique

Formalisation de la proximité / distance du sens de deux mots.

- Dans Lingway KM, la distance sémantique mesure la « longueur » du parcours nécessaire dans le *réseau sémantique* pour relier deux concepts. Cette distance est fonction du

nombre et du type de relations sémantiques à parcourir pour relier les deux concepts. Voir *réseau sémantique*, *expansion sémantique*.

Domaine sémantique

Activité ou discipline dans laquelle un mot est utilisé avec un sens donné. Par exemple, le mot « souris » a un certain sens dans le domaine « informatique » et un autre dans le domaine «zoologie».

- Dans le dictionnaire Lingway, il existe environ 350 domaines sémantiques qui sont essentiellement utilisés pour déterminer le sens des mots dans les questions. Les domaines apparaissent dans la fenêtre « analyse » qui donne l'interprétation de la question.

E

Entité nommée

Descripteur particulier désignant un objet (typiquement une personne, un lieu, une organisation) par son nom. Par extension, on inclut également les valeurs et les dates dans les entités nommées.

- Lingway KM extrait ces différents types d'entités nommées.

Expanser

Par abus de langage, réaliser une expansion sémantique.

Expansion sémantique

Opération consistant à construire à partir d'un terme de départ donné (en utilisant le *réseau sémantique* du *dictionnaire)* une liste de *termes* dont le sens est voisin, généralement en vue de la construction d'une équation de recherche dans la base de données documentaire.

- Dans Lingway KM, l'expansion sémantique est paramétrable par une « distance sémantique » permettant d'aller plus ou moins loin dans l'expansion.

H

Hit

Mot retenu pour la sélection du document, issu de la requête ou dérivé sémantique, indiqué sous la description d'une vdoc dans la liste des résultats de la recherche.

- Dans Lingway KM, la liste des hits apparaît sous le document et sa description dans la liste des résultats afin de renseigner l'utilisateur sur les raisons de la sélection du document.

I

Index thématique

- Dans Lingway KM, liste de descripteurs extraits des vdocs (colonne de gauche de l'interface standard) regroupés en différents types : terme extrait librement par le système, terme extrait sur la base d'un thésaurus externe, *entité nommée* standard (personne, lieu, organisation) ou personnalisée. En cliquant sur un des descripteurs, on sélectionne la liste des vdocs indexées par ce descripteur (c'est-à-dire qui le contiennent, ou dans le cas de termes libres ou de termes de thésaurus, les vdocs pour lesquelles le descripteur a été non seulement trouvé, mais reconnu comme un "bon" descripteur par filtrage statistique).

Indexation

Ce mot peut être pris dans deux sens, indexation 1 ou indexation 2.

Indexation 1

Il s'agit de l'indexation dite "full-text". Opération qui consiste à produire à partir d'un ensemble de documents, une liste (dite "inversée") de mots avec les liens vers les documents dans lesquels chaque mot apparaît. Cette opération est faite pour l'ensemble des mots sauf ceux déclarés dans un "anti-dictionnaire", généralement des mots outils, verbe être, etc.

- Dans Lingway KM une indexation full-text est réalisée ; elle est utilisée par la recherche sémantique.

Indexation 2

Opération consistant à affecter des *descripteurs* à des documents, chacun servant d'intitulé à un index thématique.

Voir aussi *Descripteur, Descripteur libre et Descripteur contrôlé.*

Indexation contrôlée

Mode d'indexation des documents par des *descripteurs* choisis dans une *liste d'autorité.*

- Dans Lingway KM il est possible de définir une liste d'autorité. Dans ce cas, un texte comportant un descripteur de la liste d'autorité sera systématiquement indexé par ce descripteur, indépendamment de tout calcul statistique.

Voir aussi *Indexation libre.*

Indexation libre

Mode d'indexation des documents par des *descripteurs* choisis indépendamment d'une *liste d'autorité.*

- Dans Lingway KM les descripteurs libres sont choisis parmi l'ensemble des *termes* et *entités nommées* extraits du texte, par des méthodes statistiques tenant compte de la fréquence du terme dans l'ensemble du corpus et dans le document à indexer.

Indexation mixte

Mode d'indexation des documents combinant *indexation contrôlée* et *indexation libre*.

- Lingway KM permet une indexation mixte.

Indexation thématique

- Dans Lingway KM, l'indexation des documents par descripteurs (définis ou extraits) et la regroupement de ceux-ci par types (thèmes, entités nommées, etc.) permet de constituer différents index thématiques.

[à ne pas confondre avec Catégorisation ou Classification].

L

Lemmatisation

Identification de la forme canonique (ou lemme) d'un mot, à partir des formes (conjuguées, pluriel...) que l'on trouve dans les textes.

- Lingway KM applique une lemmatisation aux documents des applications, et aux requêtes des utilisateurs. Ainsi une requête avec un mot au pluriel ou au singulier a le même résultat.

Lexique

Ensemble des mots de la langue courante, ou des mots d'un domaine spécialisé, dictionnaire spécialisé ou glossaire, lexique de mots rares.

- Relativement à Lingway KM, le *dictionnaire électronique* à la base de l'analyse linguistique des textes comporte une partie lexique (400.000 mots couvrant l'anglais, le français, l'allemand, l'espagnol, le néerlandais et le portugais, reliés à une couche conceptuelle indépendante de la langue).

Liste d'autorité

Liste de *descripteurs* devant être retenus pour indexer un document. Cette liste peut être "à plat" ou structurée par différentes relations (voir *thésaurus*).

M

Méta-donnée

En général, donnée sur une donnée. Dans notre domaine, il s'agit des données structurées sur un document ou un ensemble de documents. Ces données peuvent inclure les descripteurs, mais également tout type d'information sur le document, auteur, publication, date, infos légales, typologie du document, etc. A l'âge de l'Internet, les méta-données sont aux documents ce que les fiches des catalogues sont encore parfois aux livres des bibliothèques.

- Lingway KM produit automatiquement des méta-données de chaque document traité, dans le format standard en s'inspirant des conventions du Dublin Core.

O

Ontologie

Ensemble de concepts, structurés par des relations sémantiques et hiérarchiques.

- Le *dictionnaire électronique* de Lingway KM est constitué de mots, reliés à une ontologie de 150.000 concepts, liés par différentes relations sémantiques. Voir *réseau sémantique*.

P

Plan de classement

Structure hiérarchique permettant le classement et le repérage de documents ou d'ensembles documentaires.

- Il est possible de charger un plan de classement dans Lingway KM.

Voir *Classification*.

Post coordination

Mode d'indexation combinant plusieurs descripteurs élémentaires entre eux. Par exemple un document décrivant un garage sera décrit par les 2 descripteurs « réparation » et « voiture ».

Voir aussi *Pré Coordination*.

Pré coordination

Mode d'indexation par des descripteurs complexes, mots composés ou expressions. Par exemple un garage sera décrit par un descripteur « réparation de voiture ».

Voir aussi *Post Coordination*.

- Lingway KM sait reconnaître des descripteurs pré coordonnés.

Voir *plan de classement.*

R

Réseau sémantique

Ensemble d'objets (de concepts) reliés entre eux par des relations sémantiques.

- Dans Lingway KM, désigne les concepts et le graphe des relations sémantiques dont il existe une vingtaine de types, dont la relation hiérarchique, la proximité sémantique, partie-de, etc.

S

Sémantique

"Qui signifie". Etude du sens des unités linguistiques et de leurs combinaisons.

T

Taxonomie

Réseau sémantique dans lequel la seule relation est la relation hiérarchique (générique-spécifique). Voir *Plan de classement.*

Terme

Mot simple, mot composé ou groupe de mots plus ou moins complexe désignant généralement un objet ou une opération.

- Lingway KM identifie les termes saillants d'un document ou d'un corpus sur une double base linguistique et statistique. Des « patrons » linguistiques décrivent la forme syntaxique possible d'un terme (nom-préposition-nom, nom-adjectif, etc.) et des calculs statistiques déterminent les termes à retenir comme *descripteurs* pour l'*indexation* d'un document ou d'un corpus.

Thème

- Dans Lingway KM, désigne un *descripteur* qui est un *terme* par opposition à une *entité nommée*.

Thésaurus

Liste d'autorité munie d'une structure de type *réseau sémantique* généralement constituée de deux relations principales : la relation hiérarchique (TG d'un descripteur à son générique ou son inverse TS d'un descripteur à un spécifique) et la relation de proximité (TA ou "terme associé"). De plus, les thésaurus recensent souvent des termes non descripteurs qu'ils rattachent à des descripteurs (relation EM ou "employer" d'un terme au descripteur, ou EP "employé pour" du descripteur vers le terme non descripteur.

- Il est possible d'intégrer un thésaurus (répertoire de termes normalisés pour le classement documentaire) dans Lingway KM. Les *descripteurs* apparaissent alors dans la liste des *thèmes* préfixés par « TH ».

V

vdoc

- Dans Lingway KM, version de document. Pour plus de détail, lire « Les vdocs » page 14.

Sources d'information et contact

Démonstrations en ligne

http://demo.lingway.com/lkm

(codes d'accès à partir de la page Démonstrations du site web)

Site web

Brochures, livres blancs et inscription aux événements Lingway :

http://www.lingway.com

Contact

Lingway SAS
Immeuble Paritalie
18, Rue Pasteur
94270 LE KREMLIN BICETRE, France.

+33 (0)1 58 46 12 40

Lingway est un éditeur qui met en œuvre de puissants composants sémantiques multilingues* pour proposer des moteurs de recherche « qui comprennent votre métier », couplés à des outils de catégorisation et d'aide à la lecture.

Moteur généraliste, moteurs de la gamme « clés en main » pour le recrutement, le médical, la propriété intellectuelle, le commerce électronique…, moteurs « sur mesure » construits grâce à Lingway Custom Search, tous les moteurs de Lingway s'appuient sur la plate-forme de développement sémantique Lingway KM.

Société innovante ayant reçu en 2005 le Prix 01 Informatique-Technologia de la Jeune Entreprise High-Tech, Lingway compte aujourd'hui une centaine de clients – dont l'AFNOR, l'APEC, BNP Paribas, le CNRS, l'INPI, Géopost, le Groupe PSA Peugeot-Citroen, le MINEFI, Questel.Orbit, Würth … – ainsi que de nombreux partenaires éditeurs ou intégrateurs, tels que ATOS, Cap Gemini, Ever, Exalead, Openwide, ProfilSoft, Sopra…

* Représentant plus d'une centaine d'années-homme de développement.