

Manuel Utilisateur

1^{ère} Partie : Usage

TABLE DES MATIERES

1.	Qu'est-ce qu'InfoCodex ?	2
2.	Démarrage d'InfoCodex	3
3.	Recherches et Résultats	4
3.1	Masque standard de recherche	4
3.2	Affichage des résultats	8
3.3	Recherche avancée (combinaisons, recherche sur champs d'indexation, recherche booléenne)	11
3.4	Familles de documents dans InfoCodex	13
4.	Créer une Nouvelle Collection	14
4.1	Déclaration standard d'un ensemble de documents	14
4.2	Choisir les sources de données	14
4.3	Exécution des importations (analyse de contenu/ indexation)	18
4.4	Options avancées pour la définition d'un corpus	20
5.	Evaluation of Content	21
5.1	Catégorisation des Documents	21
5.2	Suppression d'enregistrements de document	22
5.3	Création de sous-collections	22
5.4	Comparaison des documents ajoutés avec l'ensemble initial	23
6.	Ajout de Documents	24
7.	Synonymes/ Taxonomie	26
8.	Administration Système (Résumé de la 2^{ème} Partie)	27
8.1	Protection des Données / Concept des Domaines IC	27
8.2	Gestion des Collections	28
8.3	Intervention sur la Catégorisation	29
9.	Quelques Exemples	30

1. Qu'est-ce qu'InfoCodex ?

"InfoCodex Knowledge Manager" est un outil logiciel pour la gestion des connaissances et des documents. Au moyen d'agents de collecte, il identifie le contenu textuel de différents types de documents à travers différentes plateformes (intranet et internet, boîtes aux lettres, etc). Puis il analyse le contenu intrinsèque de ces textes quelle qu'en soit la langue, regroupe les documents par thème et représente l'ensemble de la collection de documents ou corpus sous forme de carte graphique de l'information – tout ceci sans aucune intervention humaine.

Extraire les Connaissances à partir de Différentes Sources

Un instrument, réellement utile pour faciliter le travail d'examen de masses d'informations, doit être capable de collecter les documents de différentes plateformes et de différents formats puis de les regrouper aussi clairement que possible en fonction de leur contenu thématique. Il doit être aussi capable de reconnaître que la traduction anglaise d'un document français a le même contenu thématique que le document original. L'indexation automatique pour une recherche cross-linguale efficace est donc indispensable.

Caractéristiques principales d'InfoCodex

Formats des Documents

Des agents de collecte sont disponibles pour les fichiers PDF, les documents Word, les tableaux Excel, les fichiers PowerPoint, PostScript et EPS, RTF et autres fichiers texte, HTML, XML, divers formats de courriers électroniques incluant leurs pièces jointes, les fichiers PST (archives Outlook), JPG, GIF, TIF, BMP, Lotus Notes, ZIP, GZIP et GZ.

Reconnaissance du Contenu

Elle repose sur une base de données linguistique multilingue, dont les enregistrements sont liés à une taxonomie universelle.

Cette base de données, qui dispose à l'heure actuelle de 2,9 millions de mots et expressions, est construite à partir de travaux renommés tels que WordNet de l'Université de Princeton, EuroVoc, AgroVoc, JuriVoc, CIS et beaucoup d'autres vocabulaires techniques spécialisés. Elle permet de traiter l'anglais, l'allemand, le français, l'italien et l'espagnol (le chinois etc. sont prévus ultérieurement).

Catégorisation

L'analyse est réalisée à l'aide de réseaux de neurones auto-organisant (carte de Kohonen)

Indices de Recherche

Mots et groupes de mots (expressions) pour les recherches classiques en texte intégral, synonymes pour les recherches avancées et les recherches cross-linguales, mesures de similarité pour les recherches par similarité conceptuelle (basée sur les réseaux de neurones).

Quelques exemples typiques d'application

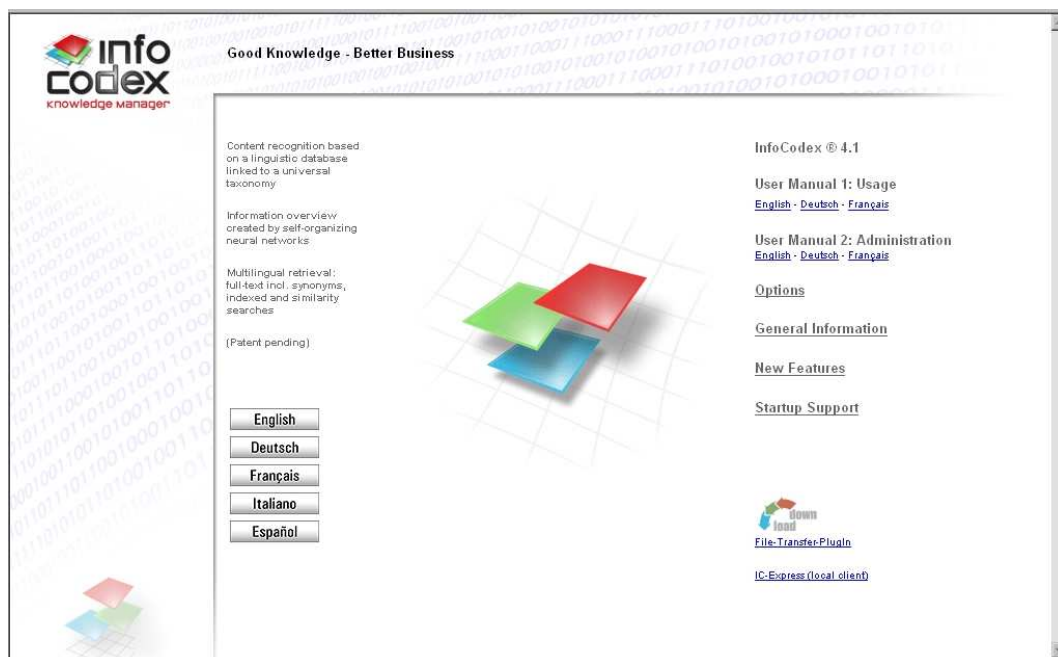
- Moteur de recherche d'entreprise (Enterprise Search Engine), outil de recherche d'informations sur le réseau interne de l'entreprise.
- Recherche d'information : combinaison des résultats de différents moteurs de recherche sur Internet avec les documents des bases de données internes, d'intranet et des courriers électroniques, etc.
- Veille marketing, concurrentielle, technologique et économique. Recherche de brevet.
- Coordination des compétences (par groupes de projet et avec des postes externes)
- Génération automatique de mots-clés (classement par thème des documents)
- Gestion de la réponse (applications orientées processus)

2. Démarrage d'InfoCodex

InfoCodex utilise un navigateur standard (Internet Explorer, Netscape, Mozilla, Firefox) comme interface utilisateur.

Adresse de démarrage : *servername/ic-start.html*

Le langage de l'interface utilisateur (anglais, allemand, français, italien, espagnol) peut être sélectionné sur la première page de l'application qui inclut également la fenêtre d'identification de l'utilisateur.



La connexion de l'utilisateur (login) dépend du mode d'installation choisi :

- L'identifiant utilisateur et son mot de passe sont ceux du réseau informatique sous-jacent
Ceci est le cas quand la gestion des utilisateurs d'InfoCodex est couplée avec LDAP ("point unique d'administration, single point of administration", par exemple ADS de Windows). Lorsque les droits d'accès du système d'exploitation doivent être respectés (File System Security), ce mode d'organisation est obligatoire.
- L'identifiant utilisateur et son mot de passe sont spécifiques à InfoCodex
Dans ce cas l'administration des utilisateurs et l'allocation des droits d'accès sont gérées exclusivement par InfoCodex. Ce mode d'organisation permet la mise en place d'un utilisateur "public" sans mot de passe.
- Connexion automatique
Avec cette configuration de type "single sign-on" aucun identifiant ou mot de passe n'est nécessaire.

Après la connexion, le masque de base d'InfoCodex apparaît (masque de recherche standard), et l'utilisateur peut sélectionner les collections de document et démarrer une recherche.

3. Recherches et Résultats

Dans InfoCodex, l'ensemble global de tous les documents est subdivisé en corpus individuels ou collections de documents qui peuvent être librement combinés. Par exemple :

- Tous les documents du réseau interne de l'entreprise
- Tous les documents d'un site web d'une organisation spécifique
- Les résultats de recherches sur Internet avec différents moteurs de recherche
- Les documents résultant d'une conférence professionnelle
- Les courriels et leurs pièces jointes, combinés avec les documents de répertoires sélectionnés sur un serveur et les documents de sa propre machine cliente

Les différents corpus sont indépendants les uns des autres et classés individuellement, chacun dans sa propre bibliothèque virtuelle. Ils sont tous répertoriés par InfoCodex. Les recherches s'appliquent sur une collection donnée.

La création d'une nouvelle collection est expliquée dans la section 4. Supposons pour l'instant que plusieurs collections de documents existent déjà et regardons comment des documents peuvent être recherchés et identifiés.

Tout d'abord on sélectionne une collection en cliquant sur l'icône correspondante dans la fenêtre gauche de la page, comme si l'on sélectionnait un dossier avec l'Explorateur Windows. Les différentes sources de données de la collection apparaissent alors, comme pour l'affichage des sous-répertoires de l'Explorateur Windows. En cochant les cases correspondantes, il est possible de restreindre la recherche à certaines sources de données. Une requête peut alors être spécifiée dans la fenêtre de recherche.

3.1 Masque de Recherche Standard



Le texte de la requête peut être entré en anglais, allemand, français, italien ou espagnol. L'utilisateur peut alors sélectionner le type de recherche à effectuer :

- Recherche *exacte* La recherche s'applique avec les termes exacts indiqués.
- Recherche par *synonymes* "bicyclette" prendra également en compte «vélo» ainsi que "bicycle", "bike", "Fahrrad", etc. avec recherche cross-linguale.

- Recherche par *similarité* Cette recherche par l'exemple est basée sur la similarité conceptuelle.

Les caractéristiques des différents types de recherche sont détaillés comme suit :

a) Recherche exacte (= recherche texte-entier)

- La recherche "exacte" correspond au mode texte intégral classique en usage dans les moteurs de recherche bien connus tels que Google, Altavista, etc. InfoCodex possède toutefois une option de troncature à droite.
- Une correspondance exacte des mots ou termes est exigée à l'exception des majuscules et minuscules qui ne sont pas discriminées (comme sur Google). Avec "bicyclette", les mots "bicyclettes", "cycle" ou "vélo de course " ne sont pas pris en compte.
- Les nombres et les mots non significatifs tels que "the", "as", "le", "la" etc. sont également pris en compte.

Dans ce mode de recherche la troncature à droite est également possible. Avec "Financ*", des termes comme "Finance", "financier", "financing" etc. sont pris en compte et plus généralement tout ce qui commence par "Financ".

- Tous les termes de la recherche doivent être présents dans les documents, si aucun opérateur OU n'a été spécifié (voir ci-dessous). Si le nombre de documents trouvé est faible, InfoCodex demande alors s'il doit inclure dans les résultats les documents contenant un terme de recherche en moins. Ceux-ci seront alors pondérés plus faiblement que ceux contenant tous les termes.
- Règle pour les opérateurs ET/OU :
Les mots séparés par des blancs ou des virgules doivent tous être présent (condition ET comme par exemple dans Google). Les mots séparés par un ";" (point-virgule) sont traités comme des groupes reliés par OU.

Exemple

Signification

Traffic Car	→	A la fois "Traffic" et "Car" doivent être présents (ainsi que leurs synonymes).
Traffic ; Car	→	Soit "Traffic", soit "Car" doit être présent (ainsi que les synonymes correspondants).
Traffic Car ; Transport	→	Soit ("Traffic" et "Car" à la fois), soit "Transport" doivent être présents (ainsi que les synonymes correspondants).

- N.B. : les mots séparés par des espaces blancs sont également analysés pour vérifier s'ils correspondent à des expressions ou mots composés (exemple "pomme de terre"). Si l'on souhaite éviter cela, les mots doivent être séparés par des virgules (par exemple "pomme, de terre").

b) Recherche par synonymes (=recherche sémantique cross-linguale)

Ceci est donc bien différent des techniques de recherche des moteurs de recherche Internet les plus populaires, tels que Google, Altavista, etc. En effet avec InfoCodex :

- Les synonymes sont pris en compte :
L'entrée "bicycle" recherchera les documents avec "bike", "pushbike", "racing cycle" etc.

- La recherche est *cross-linguale* (anglais, allemand, français, italien, espagnol) :
L'entrée "bicycle" en anglais recherchera les documents avec l'allemand "Fahrrad", le français "bicyclette", l'italien "bicicletta", mais aussi "vélocipède", "push-bike" etc.
- Les *expressions composées (groupes de mots)* sont reconnues :
Par exemple "common sense", "Europäische Union", "office fédéral de la statistique".
- Les règles pour les opérateurs ET/OU sont les mêmes que pour la recherche exacte.

c) Recherche par similarité

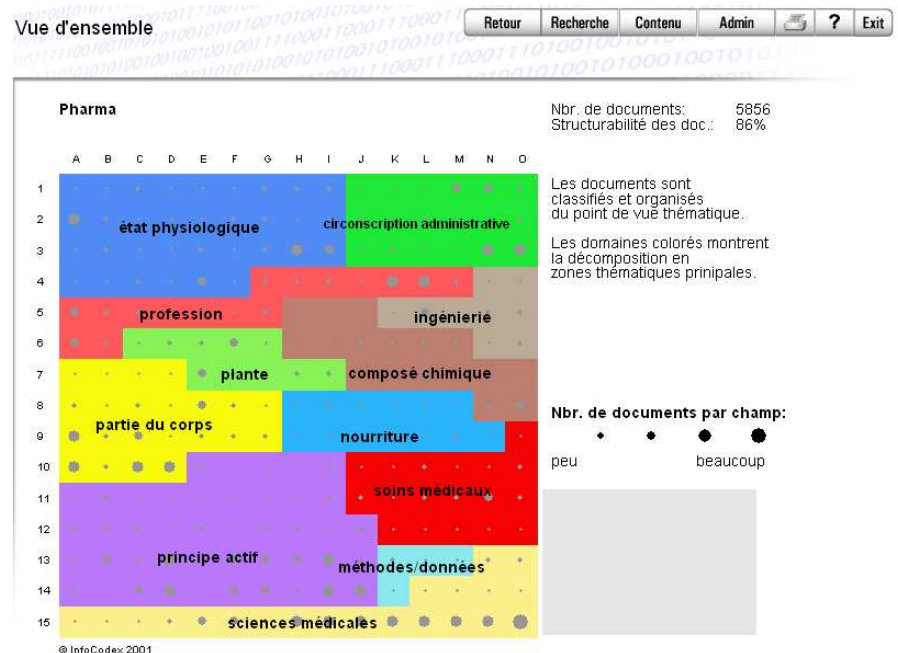
- La recherche par similarité (ou recherche thématique, recherche par l'exemple) signifie
→ *trouver des documents similaire au document exemple*.
- Les documents sont pondérés en fonction de la similarité conceptuelle de leur contenu avec le contenu du texte de requête, dont les mots ne sont pas forcément contenus explicitement dans les documents retrouvés.
- Les documents les plus pertinents reçoivent un ordre de classement supérieur qui leur permet d'apparaître en tête de liste et un code de couleur qui leur permet d'être repérés facilement sur la carte de l'information de la collection.
- La requête, dans ce cas, peut consister en un véritable texte libre et pas seulement en une simple liste de termes. Il peut être inséré dans la fenêtre de recherche, par exemple (via "copier/coller") à partir d'un courriel, d'un fichier PDF, Word, HTML, etc.
- Applications typiques de la recherche par similarité :
 - Recherche de brevet
 - Planification des travaux de recherche : comparaison de nouveaux projets ou nouvelles intentions de recherche avec les travaux existant dans une collection de documents
 - Comparaison de publications étrangères avec ses propres travaux de recherche
 - Traitement des demandes clients → Recherche pour des cas similaires dans le passé
 - Identification des normes et règlements s'appliquant à un problème ou projet particulier
- Les *recherches par similarité pure* avec seulement 1 à 3 termes (sans autres restrictions supplémentaires) n'ont guère d'intérêt. Dans ce cas, InfoCodex pratique automatiquement une recherche par synonyme (avec un opérateur OU entre les termes et leurs hyperonymes/ termes génériques).

Combinaisons et Opérateurs Booléens

Avec la syntaxe spécifique <boole> ... </boole>, toutes les combinaisons booléennes désirées, utilisant ET, OU, SAUF et les parenthèses peuvent être élaborées. Les règles de syntaxe sont affichées en cliquant sur le bouton ?

d) Autres Possibilités et Outils d'aide

- Recherche avancée** Il est possible de combiner les recherches par synonymes, sur termes exacts et par similarité, et de les compléter également par une recherche sur les champs d'indexation (auteur, date du document, type de fichier, etc.) ou encore par une recherche booléenne avec des combinaisons complexes de conditions ET/OU/SAUF. Les détails sont donnés dans la section 3.3.
- Afficher les termes de recherche** Avec cette fonction, les mots de recherche apparaissant dans un ensemble donné de documents sont affichés. Ceci peut être très utile lorsque l'on recherche des mots très particuliers dont on a du mal à se souvenir. Cela sert également à repérer les fautes d'orthographe dans les documents.
- Options utilisateur Règlages** L'utilisateur peut sélectionner divers paramètres en fonction de ses préférences comme l'ordre de tri et l'étendue des listes de résultats de recherche. De plus, il peut choisir le langage pour l'interprétation des termes de recherche (par défaut il s'agit du langage de l'interface utilisateur). Ceci s'avère particulièrement intéressant lorsque qu'un terme a des significations différentes pour chaque langue (par exemple "war" en allemand et "war" en anglais) et dans les situations exceptionnelles où InfoCodex ne peut pas déterminer clairement le langage de la requête à partir du texte de recherche.
- Carte thématique de l'information** C'est la représentation graphique de l'ensemble des documents d'une collection, que l'on peut voir comme une bibliothèque virtuelle où les documents sont classés logiquement par thème. La taille des cercles de chaque case est proportionnelle au nombre de documents qui s'y trouvent.



3.2 Affichage des Résultats

Après l'exécution d'une recherche (par synonymes, exacte ou par similarité) apparaît un écran donnant des informations statistiques de base et proposant différentes options pour l'affichage des résultats :

Sommaire des résultats pour **Recherche synonyme** Pharma (5856 doc.)

Condition		Nbr de doc.
malaria		
-> malaria		50
Nbr.de coups		50

Critères de tri

<input type="checkbox"/>	par champ (neuron)	<input type="checkbox"/>	par date
<input type="checkbox"/>	par taille de document	<input type="checkbox"/>	par nombre d'accès
<input checked="" type="checkbox"/>	par pertinence (c.recherche)	<input type="checkbox"/>	par origine
<input type="checkbox"/>	par portion graphique	<input type="checkbox"/>	par portion des nombres

Nbr.de docum.

<input type="checkbox"/>	premiers 100
<input type="checkbox"/>	premiers 200
<input checked="" type="checkbox"/>	tous, maximum <input type="text" value="1000"/>

Recommandation pour le choix des options :

- Liste le type d'affichage le plus simple ; approprié pour un nombre de documents trouvés inférieur à 100.
- Grappe les résultats sont groupés par thème; approprié pour 100 à 500 documents trouvés.
- Carte de chaleur les documents trouvés sont mis en évidence sur la carte graphique ; approprié lorsqu'il y en a plus de 500.

a) Option Liste → sommaire → document original

Par défaut, les documents sont ordonnés par ordre de pertinence décroissante. Pour chacun, le nom du fichier, le nombre de mots, le titre, l'auteur (s'il y en a un), la date du document, le nombre de fois où il a été sélectionné pour lecture, son classement dans la liste et ses trois mots-clés les plus significatifs sont indiqués.

Simil	Date	Mots	Origine	Describeurs
1	16.03.04	1503	C:\dem...iten/malaria.html	malaria, plasmodium falciparum, "Teufelsinsel"
2	16.03.04	1648	C:\dem...iten/malaria.html	malaria, plasmodium falciparum, "Teufelsinsel"
3	16.03.04	335	C:\dem...amittel/index.htm	amino-benzène-sulfamide, "Managerleistung", trope
4	16.03.04	335	C:\dem...mitel/index.html	amino-benzène-sulfamide, "Managerleistung", trope
5	20.03.04	230	C:\dem...2/20020330_2.html	"state of confusion", décuplement, papier buvard
6	16.03.04	683	C:\dem...kheiten/index.htm	vaginite, "noma", "neurovaskulär"
7	19.03.04	1708	C:\dem...ma\pdf/SCHATT.PDF	profession d'infirmière, "viral haemorrhagic fever", "Weltfl
8	19.03.04	6479	C:\dem...pdf/00430845.PDF	elisa compétitif, clinique, "Feldpostdokument"

En sélectionnant un des documents et en cliquant sur la ligne correspondante apparaît un résumé des caractéristiques du document, une liste des 18 mots-clés les plus significatifs et un résumé du document (générés par InfoCodex pendant la phase d'analyse du contenu).

produit chimique 23 (Neuron P12)

Auteur: ---
 Titre: **Appareils de chauffage indépendants à foyer ouvert utilisant les combustibles gazeux**
 Caractéristique: 15105 mots / français / PDF / 10.06.03
 Nom-document: C:\SNV\Normen\CD_c\9505048_SN EN 13278_fr.pdf
 Classement: 0%

Descripteurs:

appareil	régulateur de pression de gaz	famille
brûleur	organe	nielle
gaz	utilisation	cheminée
pression hydrostatique	combustion	catégorie
allumage	produits de combustion	nutrition
l'essai	écriture portée au débit	régulateur de pression

[Afficher le document](#) [Document suivant](#) [retour](#)
[Afficher le texte extrait](#) [Editer les champs-clés](#)
[Statistique d'utilisation](#) (0 coups)

Abstract:
 Appareils conçus pour l'utilisation exclusive des gaz de la première famille . / Appareils conçus pour l'utilisation exclusive des gaz de la deuxième famille . / Appareils conçus pour l'utilisation des gaz des première et deuxième familles . / Catégories d'appareils conçus pour l'utilisation de gaz des première et deuxième familles . / Evacuation des gaz de cheminée (appareils de type B14, uniquement) . / Brûleur d'allumage ou établissement de la flamme du gaz de démarrage . Appareils avec systèmes non automatiques de contrôle du brûleur . / Appareils avec systèmes automatiques de contrôle du brûleur . / Dispositif d'évacuation des produits de combustion (Appareils de type B14 uniquement) . / La diminution du débit de gaz au brûleur d'allumage peut être obtenue : . / L'appareil étant en fonctionnement, fermer l'alimentation en gaz du brûleur principal. / appareils avec dispositifs d'allumage automatique . appareils avec brûleur d'allumage non permanent simultané .

Le document original peut être affiché en cliquant sur "[Afficher le document](#)" ou directement en cliquant sur son rang dans la liste des résultats de recherche (voir ci-dessus).

Infocodex affiche les documents, chaque fois que cela est possible, en utilisant leur agent d'affichage original (fichiers PDF avec Acrobat, Word avec MS Word, etc.) et ne se restreint pas à l'affichage d'un extrait textuel généré automatiquement.

Toutefois, l'affichage de grands fichiers PDF (ou PowerPoint, Word, Excel) à travers le protocole standard HTTP peut s'avérer particulièrement lent si les fichiers ne sont pas optimisés pour le web. Donc, à la première tentative d'affichage d'un document, InfoCodex met en œuvre le protocole [file:///.....](#) qui est beaucoup plus efficace. Ceci ne peut fonctionner que si le navigateur est paramétré de façon appropriée. Après cette première tentative, l'utilisateur doit indiquer si l'affichage est réussi ou non. Dans la négative, les affichages suivants utiliseront la méthode HTTP (où l'on affiche des copies optimisées pour le web des fichiers PDF de grande taille).

N.B. Paramétrage du navigateur permettant un affichage direct performant :

→ Outils → Options Internet → Sécurité → Sites de confiance → Sites... → et ajouter le serveur InfoCodex à la liste de sites de confiance.

b) Option Grappe

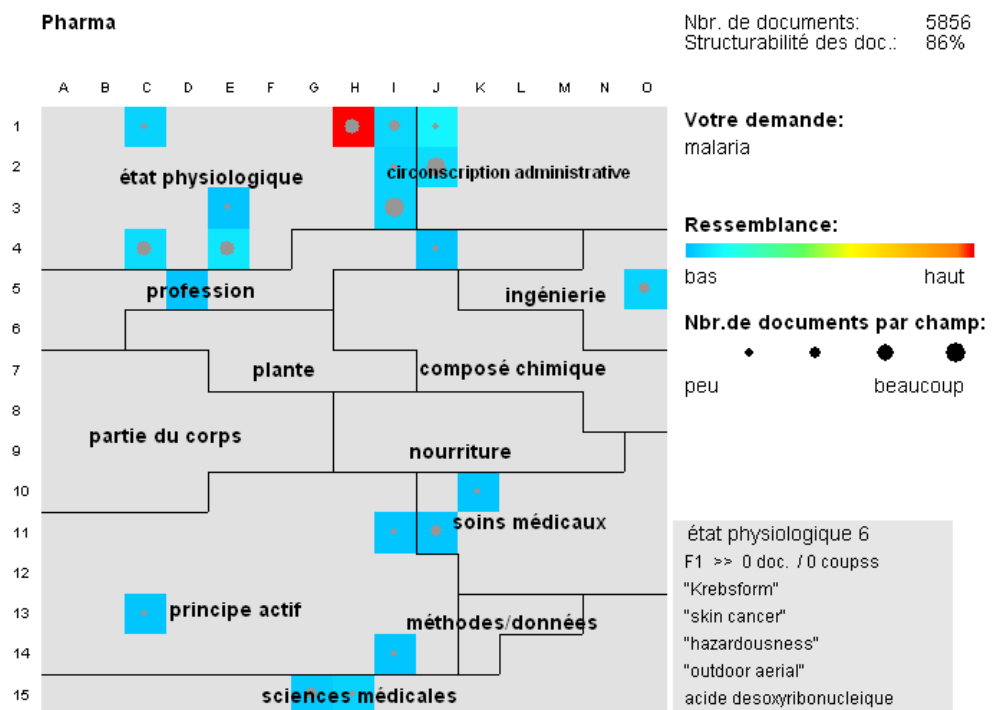
Avec cette option (aussi dénommée clustering), les résultats de recherche sont classés en fonction des thèmes principaux, et l'utilisateur peut afficher une liste de résultats pour chaque thème séparément. Cette option peut faciliter la localisation d'un document spécifique quand le résultat de la recherche en contient plus de 50/100.

Grappes par domaine et pertinence (■=haut, ■=bas)
pour demande "malaria " -> 50 coups

Nbr. de coups	Domaine	Dominante
5	état physiologique	infection, sciences médicales, communication, maladie
+ 19	état physiologique	sciences médicales, maladie infectieuse, maladie, soins médicaux
+ 13	circonscription administrative	nation, sciences médicales, infection, maladie infectieuse
3	soins médicaux	thérapie, sciences médicales, état physiologique
3	principe actif	médicament, douleur, médicament, sciences médicales
3	sciences médicales	maladie, état physiologique, maladie/infirmité
+ 4	Autres	sciences médicales, virus, invertébré, trait psychologique

c) Option Carte de chaleur

Ici la localisation des résultats est affichée directement sur la carte de l'information (ou bibliothèque virtuelle). La couleur des champs (neurones) indique le degré de similarité conceptuelle entre les documents trouvés et la requête. Le bleu (froid) indique une pertinence faible et le rouge (chaud) la pertinence la plus élevée. En déplaçant la souris sur un champ particulier, on affiche les mots-clés les plus significatifs des documents qui y sont contenus (fenêtre en gris, en bas à droite). En cliquant sur le champ choisi on fait apparaître la liste des résultats pour ce champ.



3.3 Recherche avancée (Combinaisons, Recherche par Index, Recherche Booléenne)

Dans une recherche avancée il est possible de combiner les divers types de recherche (par synonymes, sur termes exacts et par similarité) et de compléter également le tout avec une recherche sur index (champs ou attributs structurés d'indexation).

Pharma (5856 doc.)

Recherche par similarité → Trouve des documents *similaires*

Texte de recherche libre ("copy/paste" d'un E-mail, p.ex.)

fichier-texte... remettre

Recherche texte-entier → Les synonymes (ou termes exactes) *doivent* se trouver dans les documents

Rech. synonyme

Rech. exacte

Recherche par indices → Les conditions supplémentaires au méta-données *doivent* être satisfaites

Langue

Type de fichier

Auteur

Titre

Date du doc. jusqu'à

Origine du doc

Nom-fichier masque

Champs-clé

Date d'import. jusqu'à

Recherche standard Recherche Booléenne Démarrer la recherche

Exemples de recherches combinées :

- Utilisation du niveau de pertinence conceptuelle pour améliorer une recherche en texte intégral

On rentre "Tiger" dans la fenêtre de recherche par synonymes. L'utilisateur cependant s'intéresse en fait aux clubs de sport contenant ce mot. En entrant des termes comme "Sport" ou "Compétition" dans la fenêtre de recherche par similarité, le niveau de pertinence et donc le classement dans la liste des documents correspondants sont améliorés.

- Imposer la présence d'un terme dans une recherche par similarité

L'utilisateur peut décider que la présence d'un mot ou d'un de ses synonymes (par exemple "ecology") est indispensable dans les documents résultant d'une recherche par similarité. Ceci réduira encore davantage la liste de documents à consulter. Il suffit alors de compléter le texte entré dans la fenêtre de recherche de similarité par le terme "ecology" dans la partie recherche en texte intégral (soit dans la partie recherche par synonymes, soit dans la partie recherche exacte).

Recherche par Index (partie basse du masque)

Ce type de recherche est basé sur des attributs structurés (méta-données) associés aux documents. Ceci inclut des informations génériques, en général attribuées et stockées de façon automatisée, telles que :

auteur, titre, origine du document et date.

Les méta-données telles que type de fichier, langage (d/e/f/i) et date d'importation dans InfoCodex sont également disponibles.

Une restriction au nom de fichier est également possible ; dans ce cas l'opérateur de troncature en début ou en fin de mot n'est pas nécessaire, ce qui veut dire par exemple que "proj*vermess" est équivalent à "*proj*vermess*".

Enfin pour traiter d'autres caractéristiques de document comme, à titre d'exemple, le numéro client ou le code agence des instructions méta-données peuvent être fournies avec l'aide du champ appelé "Champs clés".

N.B.:

La recherche par index fonctionne comme un filtre pour réduire le nombre de documents résultats d'une requête (comme pour les recherches par synonyme ou sur termes exacts).

Les méta données réellement disponibles dans un corpus donné et pour chaque champ peuvent être affichées et sélectionnées en cliquant sur la flèche à gauche du champ. Par exemple en cliquant sur la flèche de la fenêtre de saisie "Auteur", la liste de tous les auteurs présents dans la collection s'affiche.

Les termes entrés dans la fenêtre de saisie "Titre" doivent être saisis dans leur forme exacte. Seuls les minuscules/majuscules sont ignorées. Mais il n'y a pas utilisation de termes synonymes dans ce cas.

Exemple : "without reservation" est un synonyme pour "unconditional" etc. dans InfoCodex. La saisie de "without reservation" dans la fenêtre "Titre" ne retrouvera que les documents contenant exactement ces termes dans le titre.

Recherche Booléenne (bouton du bas)

C'est une recherche par synonymes avec des conditions logiques complexes. Les opérations suivantes sont possibles :

- enchaînement complexe de ET/OU
- exclusion (" -train" → SAUF train)
- troncature à droite, par exemple Financ* reprend tous les mots commençant par financ
- exclusion avec troncature ("-foot*" → tout mot commençant par foot est exclu...)

Pharma (5856 doc.)

Recherche Booléenne

= Recherche-synonyme utilisant des liaisons complexes
(combinaisons de ET, OU, SAUF et Wildcards)

Entrez les termes de recherche (synonymes sont y compris)
où
-terme --> **SAUF** terme
terme* --> tous commençant par terme

Conditions:

or or or or

AND

or or or or

AND

or or or or

AND

or or or or

AND


or or or or

3.4 Familles de Documents dans InfoCodex

InfoCodex peut rassembler des “documents presque identiques” en familles de documents, par exemple,

- les mêmes documents dans différents formats (PDF, MS Word, HTML, XML etc.)
- des versions légèrement différentes ou corrigées d’un document

Des documents sont membres de la même famille si leurs contenus sont proches, au sens de la mesure de similarité d’InfoCodex (composants d’un vecteur et descripteurs).

La liste résultant de la recherche n’affiche qu’un document par famille. Un symbole spécial indique l’existence de plusieurs documents (par exemple  2).



En cliquant sur le symbole  2 la liste détaillée des documents est affichée :



Benefit

- Les listes de résultats sont plus courtes
- Les efforts de nettoyage des documents dupliqués sont facilités

4. Créer une Nouvelle Collection

4.1 Déclaration Standard d'un Ensemble de Documents

Les collections existantes sont listées dans le masque de recherche standard à la gauche de l'écran et ordonnées par domaine IC (une notion expliquée dans la section 8).

A la fin des listes auxquelles l'utilisateur a accès, apparaît le symbole suivant :



Après avoir cliqué dessus, l'utilisateur doit entrer le nom de la nouvelle collection.

Create New Document Collection

Collection name	<input type="text" value="Plant Protection"/>
Import mode	<input checked="" type="checkbox"/> Update, i.e. for future incremental loads, documents with the same filename as existing entries will be updated (otherwise the documents would be appended to the collection)
Generation of abstracts	<input type="checkbox"/> yes (but it takes some time)
Build document families	<input type="checkbox"/> yes (identify similar documents)
<input type="button" value="Advanced options"/>	
<input type="button" value="OK add documents"/> <input type="button" value="OK save empty collection"/> <input type="button" value="Quit"/>	

Celle-ci est alors sauvegardée en cliquant sur OK. L'utilisateur peut alors spécifier les sources de données pour la collection.

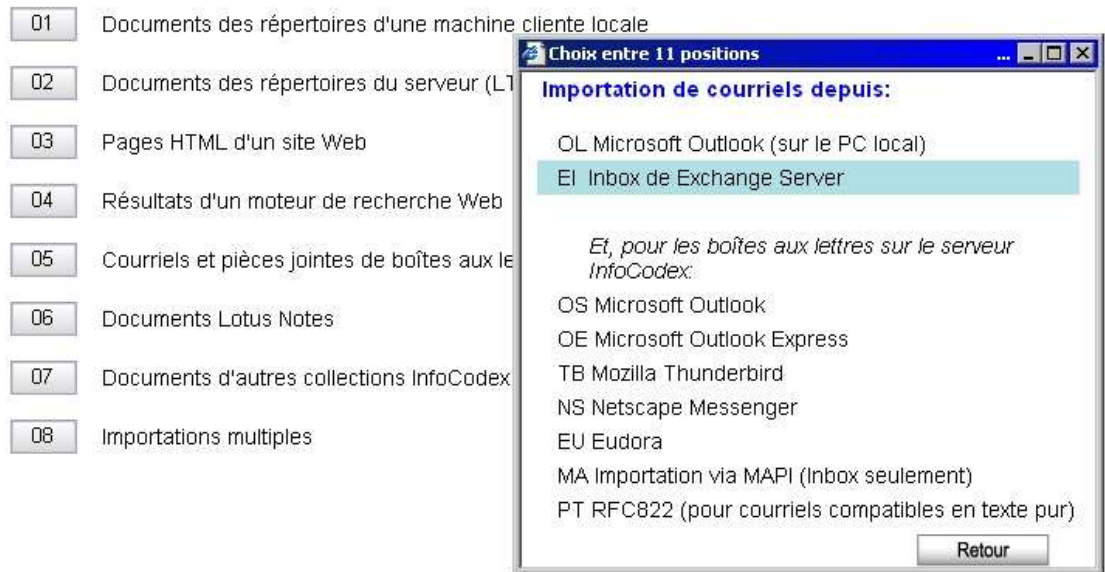
La fonction "Options Avancées" permet de définir des droits d'accès particuliers ou des fonctions de contrôle pour l'analyse de contenu (voir section 4.3).

4.2 Sélection des Sources de Données

Pour une collection spécifique, un maximum de 200 sources de données peut être spécifié. Tous les documents des sources sélectionnées seront inclus, tant qu'InfoCodex sait les analyser.

Les documents originaux ne sont pas copiés, mais seulement lus, analysés et indexés. Les méta-données ainsi que l'adresse des documents sont conservés dans la base de données d'InfoCodex. Le volume des documents n'est pas limité ; il n'a aucune influence sur l'ordre virtuel créé par InfoCodex.

Choisir les sources de données



Choix des Sources de Données

1) Documents des répertoires d'une machine cliente locale

Pour des raisons de sécurité, InfoCodex et le navigateur utilisé n'ont pas d'accès direct aux documents situés sur un ordinateur client local de l'installation.

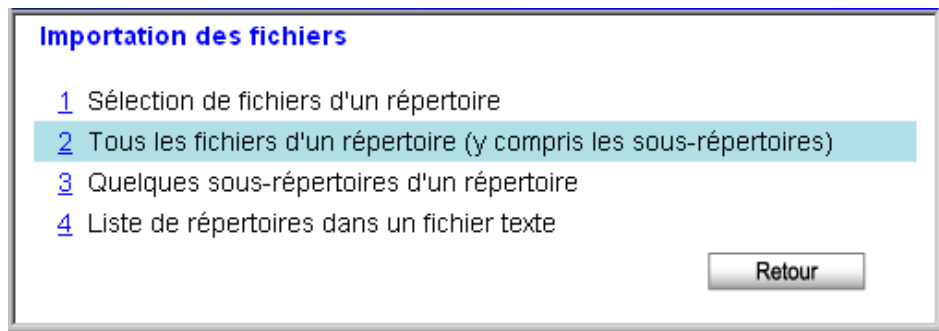
Pour transférer les documents d'une machine cliente vers InfoCodex, l'installation d'une application de transfert de fichiers particulière ("file-transfer plugin") est nécessaire. Elle peut être téléchargée à partir de la page de démarrage d'InfoCodex.



L'installation nécessite les privilèges d'administrateur système sur la machine cliente. Après la mise en place, il faut encore s'assurer que l'utilisateur a les droits d'accès requis pour exécuter les programmes installés.

A l'importation de données d'une machine cliente locale, un répertoire (dossier) est sélectionné et tous les fichiers (y compris les sous-répertoires) sont intégrés dans InfoCodex.

2) Documents des répertoires du serveur (y compris le "Réseau d'entreprise")



Dans la plupart des cas, l'option N°2 est choisie. Par la suite, la sélection des répertoires est faite selon les pratiques de l'Explorateur Windows. Tant que l'utilisateur dispose des droits correspondants, il lui est possible d'accéder à l'ensemble des informations du disque dur et même à la totalité des données du réseau.

3) Sites Web (Internet ou Intranet)

La seule chose nécessaire est d'entrer l'adresse du site désiré (par exemple www.infocodex.com). Infocodex suit alors tous les liens jusqu'à un niveau de hiérarchie spécifié par l'utilisateur, charge les pages pour l'analyse du contenu et finalement extrait les liens additionnels. Non seulement les pages HTML sont téléchargées, mais aussi les fichiers PDF, ZIP, etc. Un délai maximum de 5 minutes est alloué à chaque page.

L'utilisateur peut entrer les données d'identification/mot de passe pour les pages protégées par mot de passe.

4) Résultats d'une recherche faite avec un moteur de recherche Internet (voir section 6)

Après la sélection d'un moteur de recherche (par exemple Google) une fenêtre de saisie apparaît pour entrer le texte de la requête de la même façon qu'on le ferait avec le moteur lui-même.

InfoCodex télécharge alors tous les documents trouvés par le moteur de recherche pour une analyse de contenu. Il est important de noter que beaucoup de moteurs de recherche se limite à un maximum de 1000 documents pour affichage, même si le nombre de documents trouvés est très nettement supérieur.

InfoCodex permet l'accès aux moteurs de recherche protégés par mot de passe (par exemple Web of Science) en proposant une fenêtre appropriée lorsque nécessaire.

5) Courriels et pièces jointes (voir Section 6)

InfoCodex supporte différents formats de messagerie comme Outlook, Outlook Express, Netscape, Thunderbird, Eudora, etc. (voir plus haut). Le contenu des pièces jointes, y compris les fichiers ZIP, est également analysé et chacune d'entre elles est traitée comme un document.

Avant de choisir

OL Microsoft Outlook (sur le PC local d'un utilisateur)
 ou **EI** Inbox de Exchange Server,

il faut s'assurer que Outlook est bien installé sur le serveur. Dans le deuxième cas, l'accès au serveur central Exchange doit être établi et l'utilisateur doit donner un accès partagé à InfoCodex.

L'importation depuis Outlook (**OL**) ne peut pas être exécutée en mode batch, puisque l'utilisateur doit confirmer l'accès à sa boîte à lettres toutes les quelques minutes. En conséquence, il est avantageux d'envoyer les courriels Outlook locaux en tant que fichiers Archive Outlook (fichier **PST**) vers le serveur et de là, les importer via l'option :

2 Documents des répertoires du serveur

6) Documents Lotus Notes

Pour s'interfacer avec Lotus Notes, un client Notes doit être en place et le module InfoCodex **NotesConnect** doit être installé sur le serveur.

A l'import, les documents Lotus Notes sont tout d'abord sélectionnés à travers **NotesConnect** et sauvegardés temporairement comme fichier index dans le répertoire NotesConnect pour traitement.

7) Documents provenant d'autres collections InfoCodex

Lorsque cela est nécessaire, il est possible d'utiliser les documents d'autres collections InfoCodex préexistantes comme sources de données. Dans le principe, c'est comme s'il s'agissait du résultat d'une recherche, avec un moteur de recherche sur Internet.

8) Importations multiples

Avec cette option, un programme d'importation multiple peut être établi pour importer en parallèle, jusqu'à 9 requêtes avec 5 moteurs de recherche différents (soit un maximum de 45 recherches) ainsi que des importations depuis des sites web et des répertoires. Ceci est un agent idéal pour la recherche d'information.

4.3 Exécution des importations (analyse du contenu/ indexation)

Après la sélection des sources de données pour la collection, le traitement automatique peut être démarré par "**OK importer**". De plus, il est possible de spécifier l'ancienneté maximum des documents à importer ou l'utilisation d'un interpréteur OCR pour les fichiers image avec "**Options**" avant l'importation.

<input type="button" value="04"/>	Résultats d'un moteur de recherche Web					
<input type="button" value="05"/>	Courriels et pièces jointes de boîtes aux lettres					
<input type="button" value="06"/>	Documents Lotus Notes					
<input type="button" value="07"/>	Documents d'autres collections InfoCodex					
<input type="button" value="08"/>	Importations multiples					
<table border="0" style="width: 100%; text-align: center;"> <tbody> <tr> <td><input type="button" value="Supprimer"/></td> <td><input type="button" value="Options"/></td> <td><input type="button" value="Séquenceur"/></td> <td><input type="button" value="OK importer"/></td> <td><input type="button" value="Quitter"/></td> </tr> </tbody> </table>		<input type="button" value="Supprimer"/>	<input type="button" value="Options"/>	<input type="button" value="Séquenceur"/>	<input type="button" value="OK importer"/>	<input type="button" value="Quitter"/>
<input type="button" value="Supprimer"/>	<input type="button" value="Options"/>	<input type="button" value="Séquenceur"/>	<input type="button" value="OK importer"/>	<input type="button" value="Quitter"/>		

Les processus suivants sont alors exécutés en séquence :

- Les agents de recherche et de collecte InfoCodex extraient le contenu textuel et les méta-données (auteur, titre, date, etc.) des documents

- Reconnaissance des langues : anglais, allemand, français, italien, espagnol ; analyse lexicale et sémantique
- Analyse de contenu cross-linguale, mise en relation avec la taxonomie
- Construction d'un espace à 100 dimensions basé sur l'espace de contenu de la collection, et projection des documents dans cet espace de contenu
- Classement des documents au moyen de réseaux de neurones auto-organisant (carte de Kohonen) → "rangement des documents dans une bibliothèque virtuelle bien ordonnée"
- Génération de mots-clés pour les documents : création des descripteurs
- Indexation pour une recherche efficace

Tous ces processus sont exécutés en tâches de fond. L'utilisateur n'a pas à en suivre la progression et peut faire un autre travail en parallèle.

Séquenceur de recherche

Au lieu d'une exécution directe avec le choix → "**OK importer**", des instructions d'importation en mode batch peuvent être définies avec l'option

→ "**Séquenceur**"

L'heure de début et la périodicité des importations sont alors entrées dans une boîte de dialogue. Les instructions peuvent être affichées et modifiées par

→ Admin → "**C2 Sources de données**"

Applications typiques du séquenceur de recherche

- Veille concurrentielle
Collecte, de nuit, des dernières informations disponibles sur les activités des concurrents. L'information est disponible le matin, classée et ordonnée de façon claire.
- Mise à jour d'une collection de grande taille
 - Les documents sur le réseau local (intranet) sont chargés une première fois pendant un week-end ("Chargement Initial").
 - Les nouveaux documents sont ajoutés à la collection, par exemple chaque nuit, du lundi au jeudi ("Chargement Incrémental").

4.4 Options Avancées pour la Définition d'une Collection

InfoCodex réalise les opérations décrites dans la section 4.1 automatiquement. Si l'utilisateur souhaite influencer la classification, la fabrication des descripteurs, l'extraction des méta-données ou les paramètres de droits accès, des options peuvent être mises en œuvre avec le bouton "Options Avancées".

Paramétrer la collection

Numéro d'identification: État:

Nom de la collection

allemand:

anglais:

français: obli.

italien:

Répertoire de la collection

Groupes utilisateurs obli.

Paramétrage optionnel de classification

Classification

Table de mots-clés

Catégories prédéfinies

Instructions méta-données

Base de données frontale

Indexer **Indexer** aussi les termes exacts

Options spéciales Génération de résumés Familles de documents

Mode d'importation en cas de noms-fichiers identiques, actualiser l'ancien enreg.

Constitution de la carte

Nombre de colonnes de la carte (nbrs. de col. = nbrs. de lignes)

Nombre de sujets principaux (sous ensembles de la carte)

Description

<i>Groupes Utilisateurs</i>	Restriction des droits d'accès à certains groupes d'utilisateurs particuliers
<i>Table de mots-clés</i>	Une table contenant des mots ou termes particulièrement importants et leurs poids respectifs. Ces mots-clés peuvent influencer à la fois la classification et la fabrication des descripteurs (voir Partie 2, section 5.1)
<i>Catégories prédéfinies</i>	Des tables qui permettent de contrôler ou d'influencer la classification (voir Partie 2, section 5.3)
<i>Instructions Méta-données</i>	Un outil pour la spécification de règles d'extraction des méta-données particulières des documents
<i>Base de données linguistique frontale</i>	En complément à la base linguistique d'InfoCodex (> 2.9 millions d'entrées), il existe la possibilité de créer une base de données individuelle spécialisée de premier plan, incluant sa taxonomie. Elle peut par exemple contenir des abréviations de termes propres à l'entreprise ou encore des termes de spécialistes. Cette base de données frontale a priorité sur la base InfoCodex. Ceci peut avoir une influence considérable sur la classification et la génération de mots-clés (voir Partie 2, section 5.2)
<i>Mode d'Importation</i>	Lorsque l'on met à jour une collection existante, par défaut, les documents sont ajoutés même s'ils ont déjà été importés. En cochant l'option spéciale " <u>Mode d'Importation</u> " les noms de fichiers enregistrés ne seront pas dupliqués et seul leur contenu sera mis à jour.

5. Evaluation de Contenu

Les fonctions d'évaluation sont activées en cliquant sur :



Les possibilités suivantes sont disponibles :

Collection actuelle: Pharma

Documents

10	Catégorisation de documents
11	Ajout de documents
12	Ajouter une collection préparée en format txt
13	Effacer des enregistrements de documents

Analyses

20	Créer une sous-collection
21	Comparer les documents ajoutés avec l'ensemble initial

Informations d'aide

30	Carte thématique
31	Synonymes / taxonomie

La fonction "**10 Catégorisation de documents**" réalise uniquement une comparaison des documents avec la collection active (allocation formelle des documents aux champs de la carte d'information sans que les documents soient effectivement rajoutés à la collection). La collection active reste inchangée.

Au contraire, la fonction "**11 Ajout de documents**" inclut effectivement les documents sélectionnés dans la collection. Cette fonction est décrite en Section 4 et détaillée, plus loin, en Section 6. La fonction "**12 Ajouter une collection préparée en format txt**" se réfère au cas spécifique où une collection externe préparée en format texte doit être incluse. Cette fonction est décrite en détails dans la partie 2 du manuel d'InfoCodex (Administration Système, Section 6).

5.1 Catégorisation des Documents

Avec cette action, les documents de différentes sources à sélectionner sont assignés à des champs individuels de la carte de l'information en fonction de leur contenu. Les documents réunis dans le champ cible auront un contenu similaire, et par conséquent une telle mise en adéquation signifie une catégorisation basée sur la similarité conceptuelle.

Cette fonction est également le fondement des applications de gestion de la réponse : les documents entrés sont catégorisés, et en fonction de la catégorie à laquelle ils sont

individuellement alloués, une action de suivi correspondante peut être mise en œuvre avec l'aide de la matrice de décision fournie.

Exemples d'applications :

- Distribution automatique de courriels qui arrivent à une adresse générique centralisée
- Traitement automatique des demandes ou des plaintes clients, et génération de réponses suggérées sur la base de modèles préparés à l'avance
- Recherche de normes et de règlements sur des sujets spécifiques

Avec la catégorisation, une génération automatique de mots-clés a également lieu avec 20 descripteurs par document. Ils peuvent alors être utilisés pour la mise en archive de documents dans un système de gestion ou de stockage à base de mots-clés.

Séquence :

- a) Appel de la fonction "Catégoriser les documents"
- b) Sélection des sources de données (exactement comme décrit dans la section 4.2)
- c) Sortie des résultats à l'écran ou stockage dans une feuille Excel (ensemble sélectionnable des caractéristiques et descripteurs des documents)

Attention, avec cette utilisation, les documents analysés ne sont pas assimilés dans le corpus existant.

5.2 Suppression d'enregistrements de document

Grâce à cette fonction, des sélections d'enregistrements de documents peuvent être retirées de la collection courante. Toutefois, les documents originaux restent intacts.

- Procédure*
- Démarrage d'une requête pour retirer les documents
 - stockage de la requête
 - sélection du résultat de cette requête et marquage des documents à effectivement retirer (fenêtre popup)
 - puis retirer les entrées sélectionnées

5.3 Création de Sous-collections

Objet Recherche d'un sous-ensemble de documents particulièrement intéressants à l'intérieur d'une collection plus large et génération d'une nouvelle collection réduite (avec sa propre carte thématique plus détaillée)


Procédure Le sous-ensemble est construit par la sélection de thèmes spécifiques issus de la classification et/ou grâce à la liste de résultats d'une recherche sauvegardée.

5.4 Comparaison des documents ajoutés avec l'ensemble initial








Objet Identification précoce de changements (signaux faibles), par exemple pour les domaines suivants :

- veille concurrentielle / étude de marchés
- service de renseignements
- identification de nouvelles technologies

Scénario Une collection de documents est créée pour un sujet d'intérêt spécifique (par exemple par un chargement initial utilisant plusieurs recherches Internet). A des étapes répétées périodiquement, les documents nouvellement apparus à partir des mêmes sources de données sont ajoutés à l'ensemble initial. Une comparaison des documents ajoutés avec l'ensemble initial devrait permettre une identification rapide et sûre des plus importants changements ("Quoi de neuf?").

Procédure  →  Comparer les documents ajoutés avec l'ensemble initial

Dans le masque qui suit, l'utilisateur doit définir quels documents incrémentaux il souhaite sélectionner dans l'optique d'une comparaison avec l'ensemble initial.

Analyse de comparaison       

Institut de recherches : Analyse de comparaison





Les contenus caractéristiques des documents incrémentaux sont comparés avec ceux de l'ensemble à base.

Sélection de l'ensemble de base et des incréments*

Date d'importation	Nbr. de documents	Ensemble de base	Incréments
<input type="text" value="15.08.06"/>	<input type="text" value="100"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="text" value="15.08.06"/>	<input type="text" value="50"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Évaluations possibles

Bulk Content Shift	List the displacements of the thematic characteristics
Carte de chaleur	The fields (neurons) having the largest relative increments are colored in red
Nouveaux sujets	Document list ordered by largest deviations from previous content

Evaluation Au moyen des trois évaluations "Décalage de contenu", "Carte de chaleur" et «Nouveaux sujets», les modifications essentielles sont détectées d'un simple clic sur un bouton.

L'évaluation "Nouveaux sujets", fournit une liste des documents contenant réellement de nouveaux faits par rapport au contenu de l'ensemble initial.

6. Ajout de Documents

Avec la fonction "Ajout de documents", des documents de différentes sources sont intégrés dans la collection courante. Il faut distinguer entre :

- "*Chargement Initial*" pour la génération d'une nouvelle collection (vide)
- "*Chargement Incrémental*" pour intégration dans une collection existante

Dans une action de "Chargement Incrémental" la structure de la carte de l'information (càd le réseau de neurones) reste inchangée. Le traitement des documents, qui ont déjà été importés, dépend du Mode d'Importation fixé au moment de la création de la collection (voir section 4.4) :

- Par défaut : Les documents sont intégrés sans tenir compte du fait qu'ils ont déjà été enregistrés
- Mode d'Importation coché : Les documents qui ont déjà des noms de fichier enregistrés sont mis à jour sans créer un document supplémentaire

Le choix des sources de données est expliqué dans la section 4.2. Voici un complément d'information pour l'importation de données depuis les moteurs de recherche web et les boîtes aux lettres courriels.

Importation à partir des moteurs de recherche Web

Après le choix d'un moteur de recherche, un masque apparaît pour saisir la requête comme dans le moteur lui-même. Des options supplémentaires sont offertes et sont d'un usage particulièrement pratique dans les actions de recherche d'information.

- "Niveaux de hiérarchie"

Si une recherche Google retourne 50 documents et le niveau de hiérarchie est fixé à 0, les 50 pages seront chargées et analysées.

Si le niveau est fixé à 1, les liens de ces 50 pages sont également identifiés et inclus dans les résultats de la recherche.

Avec un niveau établi à 2, les liens avec un niveau de pages supplémentaires sont également suivis.

Ceci est très utile quand on cherche des informations extrêmement récentes (par exemple pour des projets de recherche, de veille concurrentielle, etc.), les informations fournies directement par les moteurs web datant en général de plus de deux semaines
- Bouton "Termes de Recherche Avancée"

Les moteurs de recherche web ne fonctionnent qu'avec les termes exacts entrés dans la fenêtre de saisie. Ils n'utilisent pas de synonymes ou de termes traduits automatiquement dans une autre langue.

Cette option permet donc de sélectionner de façon additionnelle les termes synonymes et les dictionnaires de langue d'InfoCodex. Ceci génère alors des requêtes supplémentaires en parallèle qui utilisent des termes différents bien que synonymes.

Importation de courriels depuis Outlook et Exchange Server

Type de stockage email

Comment importer ?

→ Ajout de documents ; puis

Archive Outlook sur
serveur local de réseau :
fichiers PST

→ Documents des répertoires du serveur
condition : les fichiers PST doivent être sur le serveur et ne doivent pas être ouverts pendant l'importation

Exchange Server
(externe ; Inbox)

→ Courriels et pièces jointes
→ Inbox du serveur Exchange
Condition : l'utilisateur doit partager son Inbox avec InfoCodex

Boîtes aux lettres Outlook
sur **client PC**
(y compris Exchange)

→ Courriels et pièces jointes
→ Microsoft Outlook (sur client PC)
condition : Outlook doit être ouvert et la permission d'accès doit être confirmée toutes les 10 minutes

L'extraction de courriels avec pièces jointes d'un Serveur Exchange externe nécessite, naturellement, davantage de temps que sans les pièces jointes. Ceci est d'autant plus vrai que les pièces jointes sont importantes et représentent une partie significative de l'information des courriels. Sur un serveur Exchange très actif il faut compter environ 1 seconde par courriel.

Règles de bon sens/ Recommandations :

- Les fichiers PST (Archive Outlook sur le serveur réseau) sont importés de façon plus efficace en travail différé de nuit.
- Il vaut mieux également effectuer en différé de nuit les importations de données depuis un serveur Exchange.
- Les importations depuis une machine client locale nécessitent d'être présent devant le PC. Il est souvent plus pratique de les archiver d'abord comme fichier PST puis de les importer depuis le serveur en travail différé.

7. Synonymes/ Taxonomie

Cette fonction permet d'accéder à l'infrastructure linguistique d'InfoCodex. Ceci est particulièrement utile pour observer l'étendue des groupes de synonymes et clarifier des situations apparemment étranges (en particulier, lorsque qu'un document se voit attribuer un descripteur surprenant, qui de premier abord apparaît irrationnel ou illogique).

N.B : Pris individuellement, les mots d'une langue ont souvent plus d'une signification. La traduction dans une autre langue ainsi que le classement dans un groupe de synonymes et le chemin d'accès à la taxonomie sont inévitablement subjectifs. Au final, la classification d'un document via le réseau neuronal n'est pas spécialement problématique. Les êtres humains connaissent le même genre d'incertitudes en lisant et analysant le contenu d'un document.

Synonyme → *afficher des mots similaires*

Entrez le mot dont vous voulez voir les synonymes en 3 langues

Affichage en anglais allemand français italien espagnole

mot français

ou anglais

ou

Hiérarchie de concepts → *système de classification (taxonomie)*

Entrez l'hyperonyme dont vous voulez voir la classification (rien = vue d'ensemble)

terme français

ou anglais

Avec la partie haute du masque (**Synonymes**), on peut voir à quel groupe de synonymes (*nuage sémantique*) appartient un mot donné et son positionnement dans la hiérarchie des concepts.

Dans la partie basse du masque (**Hiérarchie de concepts**) on peut observer l'ensemble du système de classification (c.-à-d la **Taxonomie**), ordonnée de haut en bas jusqu'aux groupes de synonymes eux-mêmes.

Sans instruction particulière, InfoCodex affiche les nœuds absolus supérieurs de l'arbre de la taxonomie.

Si une base linguistique frontale a été installée, ces mêmes structures sont également affichées. De cette manière, l'utilisateur peut vérifier si un terme appartient à la base frontale spécialisée ou à la base standard d'InfoCodex.

Le lien des groupes de synonymes avec l'arbre de la taxonomie, c.-à-d l'allocation d'hypernymes à des groupes de synonymes, forme un ontologie spécifique, mettant en corrélation la signification la plus importante des mots.

8. Administration Système (Résumé de la Partie 2)

Les fonctions du système d'administration d'InfoCodex sont accessibles avec le bouton



Elles sont divisées en deux catégories :

- *Gestion Utilisateurs* Autorisation utilisateur et groupes d'utilisateurs ; protection des données ; interface LDAP à l'administration centralisée des utilisateurs ; création de domaines IC autonomes
- *Administration Système* administration des collections ; interventions sur la catégorisation ; réglages système

Elles sont décrites en détails dans la Partie 2 du manuel. Quelques caractéristiques intéressantes pour l'utilisateur final sont cependant abordées dans la section qui suit.

Ces fonctions ne sont que partiellement visibles pour les utilisateurs non privilégiés. La fonction

"Changement de mot de passe"

par exemple, n'apparaît que sous "Administration utilisateur".

8.1 Protection des Données / Concept des Domaines IC

La protection des données est mise en œuvre à trois niveaux :

- *Groupes Utilisateurs* Allocation de droits d'accès à des groupes sélectionnés au moment de l'établissement d'un projet
- *Sécurité File System* Conformité totale aux droits d'accès dictés par le système d'exploitation sous-jacent, lorsque l'option File System Security est activée
- *Domaines IC* Création de régions autonomes et hermétiques (sur les disques) pour des utilisateurs individuels ou des groupes d'utilisateurs particuliers

Le concept des domaines IC permet une administration complète et flexible des fonctions de sécurité, même lorsque des données hautement confidentielles sont impliquées. Le principe en est le suivant.

La partie d'InfoCodex dans laquelle sont stockées les bases de données des collections peut être divisée en un domaine IC principal et un nombre quelconque de sous-domaines IC ayant chacun en propre un administrateur de privilèges d'accès.

Domaine IC Principal

- Administrateur : "sysadmin"
- Utilisateurs : tous les utilisateurs InfoCodex

Domaine IC "Biotech"

- Administrateur : "Headbio"
- Utilisateurs : Groupe "bio"

Domaine IC "BioMgmt"

- Administrateur : "Headbio"
- Utilisateurs : Groupe "Headbio"

Les domaines IC individuels ne peuvent être créés que par l'administrateur système. Un administrateur de domaine (alloué au moment de la création) a la souveraineté complète sur son domaine. Même l'administrateur système n'a pas accès aux nouveaux domaines, sauf bien entendu, si l'administrateur de domaine lui accorde explicitement les droits correspondants. Dans le cas standard, l'administrateur système n'a que le droit d'annuler le projet.

8.2 Gestion des Collections

Ces fonctions s'appliquent à la définition et au suivi des collections individuelles (voir Section 4.4). Elles ne sont disponibles que pour les utilisateurs ayant les privilèges correspondants.

Fonctions spéciales :

- *Effacer des Collections* Purge entièrement les collections sélectionnées du système
- *Voir le Statut* Consulter le statut du traitement d'une collection (en particulier dans les phases d'importation et d'analyse) ; consulter le rapport de traitement
- *Séquenceur de Recherche* Lorsque l'importation de documents est effectuée avec le séquenceur de recherche (voir Section 4.3), les scripts correspondants peuvent être consultés et modifiés avec → Admin → "C2 Sources de données".

8.3 Intervention sur la Catégorisation

La catégorisation des documents (création de la bibliothèque virtuelle) et l'allocation de descripteurs à chaque document sont largement déterminées par la base de données linguistique et la taxonomie d'InfoCodex. La base contient plus de 2.9 millions de termes en anglais, allemand, français, italien et espagnol. Elle couvre largement tous les domaines de la connaissance.

Il peut toutefois être intéressant ou souhaitable d'introduire, par exemple, des expressions techniques spécialisées supplémentaires, des abréviations de mots spécifiques à l'entreprise ou encore une nomenclature produit propre à l'utilisateur. La catégorisation et l'allocation de mots-clés sont modifiées en conséquence.

L'éventail des possibilités est le suivant :

- *Définition/ Paramétrage des mos-clés* Déclaration d'une liste de mots ou expressions avec des pondérations fortes pour la phase d'analyse de contenu
- *Base de données frontale* Préparation d'une base données spécifiques (avec ses liens vers l'arbre de la taxonomie), contenant des mots, codes ou expressions spécifiques à l'entreprise. Cette base frontale a priorité sur la base standard InfoCodex. Ainsi, durant l'analyse de contenu, un mot est d'abord recherché dans la base frontale et, s'il est identifié, c'est cette signification et importance qui sont utilisées.
- *Catégories Prédéfinies* Ceci est un moyen d'intervenir sur la constitution de la carte d'information ("la bibliothèque virtuelle"). La Catégorisation Ad-hoc est simple et interactive. Elle permet de modifier la catégorisation automatique effectuée au préalable et permet de nombreuses réalisations. D'un autre côté il est également possible de définir à l'avance des catégories fixes et de désactiver les automatismes des réseaux de neurones.

9. Quelques Exemples

a) Recherche d'Information

Exercice (dans le domaine de la R&D) :

Recherche périodique sur le thème "Secondary contamination in raptors after the usage of poison baits against Arvicola terrestris".

Termes de recherche : rodenticide, bromadiolons, raptor, Arvicola terrestris

Procédure :

- Nouvelle Collection → Ajout de documents → Requête sur le web via Google avec → rodenticide, bromadiolone, raptor, "Arvicola terrestris"
- → Bouton "Afficher la progression" → Résultat "Aucun document trouvé"
- → Bouton "Recherche Avancée" → les synonymes des termes de recherche et leur traduction dans les autres langues sont pris en compte
- → Démarrer l'Importation

Ici, plusieurs requêtes Google sont initiées avec les différentes combinaisons de synonymes et des documents sont effectivement trouvés (attention, cet exemple a pu changer avec le temps).

Raffinements :

- la recherche est définie comme ci-dessus
- le séquenceur de recherche est mis en œuvre pour la même recherche toutes les semaines avec l'option, "ajouter seulement les nouveaux documents"

b) Trouver des savoir-faire disponibles

Exercice :

Un scientifique souhaite ouvrir un nouveau domaine de recherche. A la suite de recherches documentaires il dispose d'un papier particulièrement intéressant et veut vérifier s'il y a des connaissances de même nature au sein de son institut qui pourraient servir de base à ses travaux ou au recrutement interne.

Procédure :

- Choisir la collection InfoCodex existante des documents de l'institut "Internal Network"
- Copier le contenu du papier (Word, PDF, HTML ou email),
- puis le coller dans la fenêtre supérieure du masque de recherche
- Eventuellement entrer un synonyme pertinent dans le champs "Recherche par synonymes" (à titre d'exemple biotechnology)
- Démarrer la recherche

Résultats :

InfoCodex retrouve des documents disponibles sur le réseau local de même nature que le papier original. Les mots utilisés dans ces documents ne sont pas nécessairement présents dans le texte de requête, puisque la recherche s'est faite par similarité conceptuelle.