Manuel Utilisateur

2ème Partie : Administration Système

TABLE DES MATIERES

1. Objet de ce Manuel

Cette partie du manuel utilisateur s'adresse aux administrateurs système et aux utilisateurs privilégiés possédant certains droits d'administration des utilisateurs ou du système (au moins pour les sous-domaines). Les sujets suivants sont abordés :

Architecture Système et Concept de Protection des Données

• *Généralité*	Structure des versions "autonome" et "composants" ; Organisation disques (logiciel et base de données linguistique, base de données frontale, interface web) ; Recommandations pour les sauvegardes
• *Concept de protection des données*	Moyens garantissant la protection de l'accès aux données sensibles ; Concept des domaines IC (régions autonomes) ; Sécurité du système de fichiers (File System Security).

Administration Utilisateur

• *Données utilisateurs*	Créer / Supprimer les comptes utilisateurs : allocation des privilèges
• *Groupes d'utilisateurs*	Créer des groupes utilisateurs auxquels certains utilisateurs peuvent être assignés
• *Interface LDAP*	Coordination des données utilisateurs et des groupes d'utilisateurs à travers le système central d'administration des utilisateurs, tel que ADS de Windows (fonction administrateur système)
• *Domaines IC*	Etablissement de régions autonomes pour des groupes particuliers d'utilisateurs (fonction administrateur système)

Administration Système

• *Administration des Collections*	Créer / supprimer des collections de documents ; Procédures d'importation pour les sources de données associées ; Réorganisation/ régénération/ importations en mode différé ; Statut et rapport de traitement
• *Intervention sur la catégorisation*	Moyens disponibles pour influencer la catégorisation (structurer la "bibliothèque virtuelle") et paramétrage des mots-clés
• *Fonctions d'importation et d'exportation*	Importation de données spécifiques utilisateur à partir d'une feuille Excel ; Exportation de données InfoCodex vers une feuille Excel

Paramétrage Système

• *Options de traitement*	Définition des options de traitement Installation dans des environnements particuliers (table de correspondance des espaces disques, etc.)
• *Interfaces spéciales*	Exchange Server ; Lotus Notes Interfaces à des bases de données particulières

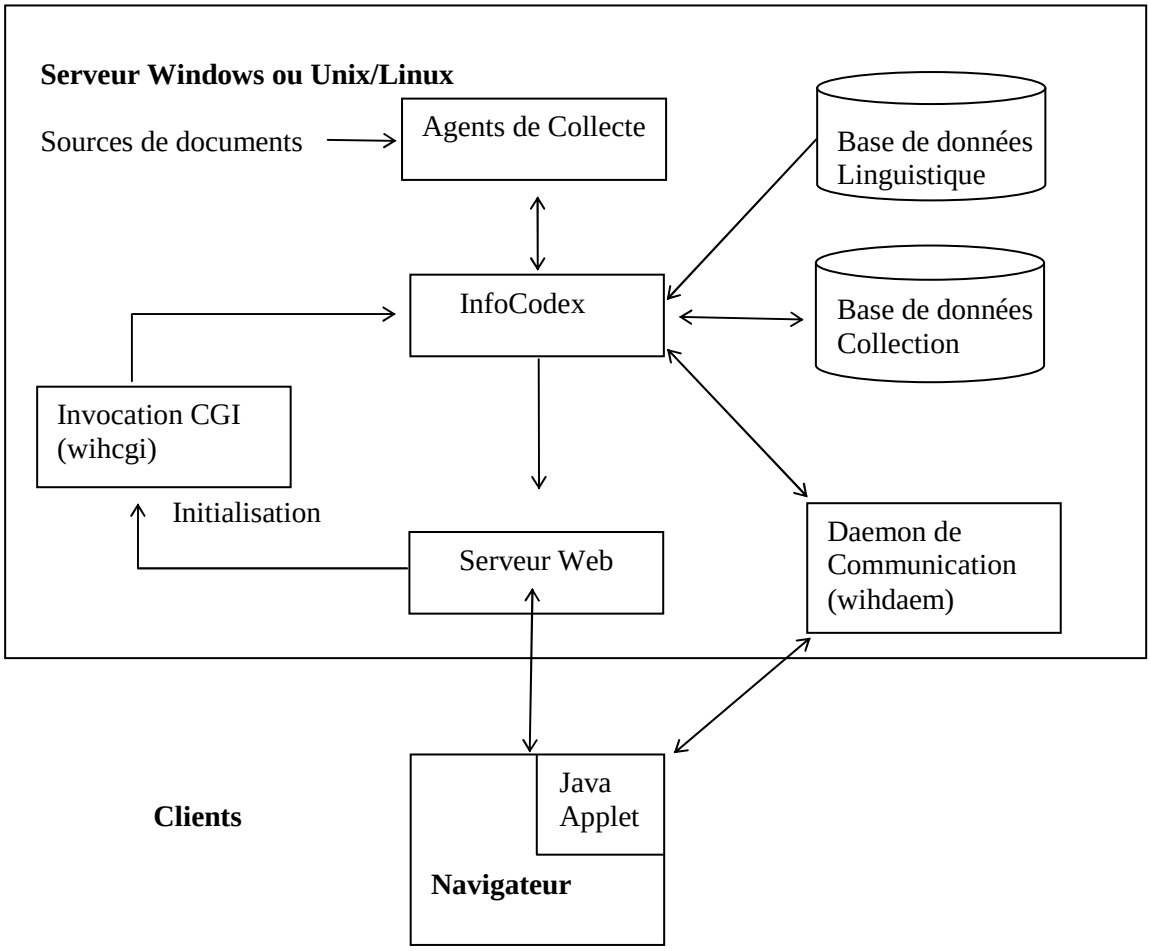
2. Architecture Système et Concept de Protection des Données

2.1 Architecture système

De par sa structure ouverte et sa portabilité, InfoCodex s'intègre de façon relativement aisée dans les environnements existants. Le logiciel InfoCodex est disponible soit en version autonome, soit (pour intégration dans d'autres systèmes) sous forme de composants (API).

Version autonome d'InfoCodex

Basée sur une architecture web, elle s'installe sur une machine serveur en même temps qu'un serveur web standard (Apache ou IIS). Du côté client, le seul logiciel nécessaire est un navigateur standard et l'environnement java (runtime).



Version sous forme de composants d'InfoCodex

InfoCodex peut être intégré dans d'autres applications sous forme de composants (modules API, librairies partagées). Des API en "C" ou Java, spécifiques de l'installation peuvent être aisément mises en œuvre grâce à la structure générique de la version. Des connections à un serveur d'applications J2EE (JSP) ou des modules Perl sont également disponibles

L'interface standard à la version "composants" est basée sur XML. Ceci est décrit dans la Partie 3 du manuel.

2.2 Organisation des disques

Le système InfoCodex est installé dans les trois zones de disque suivantes (répertoires et sous-répertoires du serveur) :

Zone Programme (permanent, aucune sauvegarde nécessaire)

Elle contient le logiciel, la base de données linguistique, les fichiers d'option et quelques fichiers temporaires pour la coordination des traitements.

Les bases de données linguistiques frontales, optionnelles et spécifiques d'un client donné, sont également situées dans cette zone. Elles sont importées à partir de tableaux Excel en complément des données standard. Elles doivent être réinstallées après chaque mise à jour de la base de données linguistique d'InfoCodex.

Zone Données (données utilisateurs et collections de documents)

Elle contient l'administration des utilisateurs et des collections, et dans des sous-répertoires, les bases de données individuelles concernant les collections de documents.

Il est recommandé d'inclure cette zone dans les procédures régulières de sauvegarde.

Interface Web (pas de procédure de sauvegarde à mettre en place)

Cette zone est utilisée pour la communication avec les navigateurs des utilisateurs (pages HTML, etc.).

Droits d'accès pour les répertoires et fichiers respectifs

Zone Programme (wP)	**Zone Données** (wP)	**Interface Web** (wP, r+)
- Logiciel InfoCodex (xP) - Base linguistique (rP) - Bases frontales (wP) (optionnel) - Coord. traitement (wP)	- Admin. utilisateurs (wP) - Gestion des collections (wP) - Bases des collections (wP) (un sous-répertoire par collection)	Un sous-répertoire par langue (e, d, f, i) - "htdocs"/ice (wP, r+) - "htdocs"/icd (wP, r+) - "htdocs"/icf (wP, r+) - "htdocs"/ici (wP, r+)

(Les domaines multiples sont autorisés)

Explications :

w	accès en écriture	P	utilisateur privilégié, qui lance le daemon de communication InfoCodex "wihdaem"
x	exécution et lecture		
r	accès en lecture	+	monde

N.B. : quand des bases de données frontales spécifiques sont utilisées, la base de données linguistique doit aussi avoir un droit d'accès en écriture (wP).

2.3 Concept de protection des données

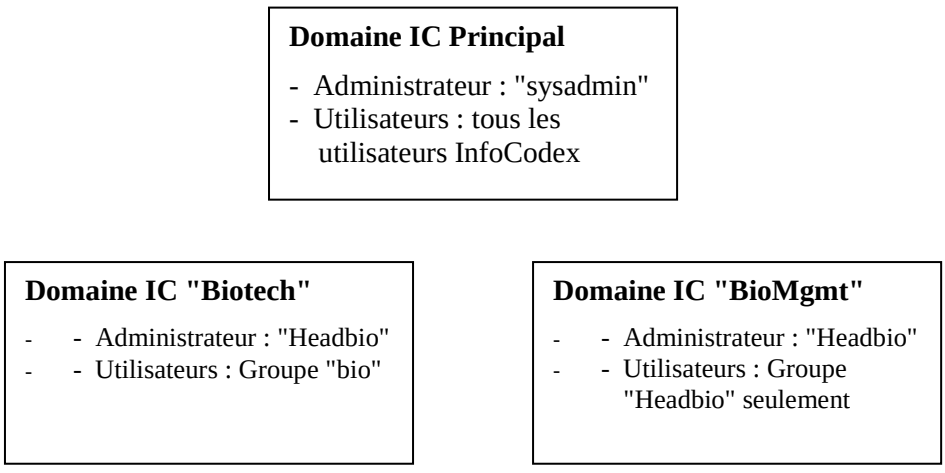
InfoCodex offre un système de protection des données à trois niveaux différents :

- *Groupes utilisateurs* — Allocation de droits d'accès à des groupes sélectionnés au moment de la création d'une collection
- *Sécurité système de fichiers* — Lorsque l'option File System Security est activée, l'accès aux documents individuels est en parfaite cohérence avec les privilèges assignés par le système d'exploitation sous-jacent
- *Domaines IC* — Régions autonomes et protégées pour des utilisateurs ou groupes d'utilisateurs spécifiques

Concept des domaines IC

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

La zone de données peut être partagée en un domaine IC principal et un nombre quelconque de domaines IC indépendants. Chaque domaine possède son propre administrateur de collections et d'utilisateurs, avec attribution de droits d'accès individuels.



Tous les domaines IC ne peuvent être crées que par l'administrateur du domaine principal. Un administrateur de domaine est désigné au moment de la création et reçoit la souveraineté complète sur son domaine. Il peut alors attribuer des permissions d'accès à des utilisateurs individuels ou à des groupes d'utilisateurs. Même l'administrateur système du domaine principal n'a pas accès aux sous-domaines, sauf bien entendu, si l'administrateur de domaine correspondant lui en accorde explicitement les droits.

Ce mode de fonctionnement garantit la confidentialité des données manipulées. Par exemple, les documents du domaine IC "Biotech" ne peuvent être accédées que par les utilisateurs du groupe "bio", tandis que les documents du domaine "BioMgmt" ne peuvent être vus que par les utilisateurs du groupe "Headbio".

Chaque domaine IC possède sa propre administration utilisateurs et sa propre gestion de collections. Tous les utilisateurs et groupes d'utilisateurs, doivent être enregistrés dans le domaine principal, c.-à-d. que l'administrateur de domaines ne peut attribuer des droits d'accès

qu'aux seuls utilisateurs et groupes déjà gérés au niveau du domaine IC principal. Un utilisateur peut être inscrit dans plusieurs domaines IC avec des droits de nature différente pour chaque domaine.

Accès aux documents

L'accès à un document donné n'est accordé que sous trois conditions :

- L'utilisateur appartient à un groupe d'utilisateurs qui a accès aux collections correspondantes et à leurs sources de données (voir Section 4.4 de la Partie 1)
- Lorsque l'option File System Security est activée, l'utilisateur doit avoir un droit d'accès au document au niveau du système de fichiers sous-jacent
- L'utilisateur est explicitement enregistré dans le domaine IC correspondant

Exemple

Soit : La collection "ProX" est dans le domaine IC "DomA" : pour cette collection les droits d'accès ont été donnés aux groupes utilisateurs "Grp1" et "Grp2"

Question : Qui a quels droits d'accès à "ProX"?

Réponse :
- Tous les utilisateurs enregistrés dans le domaine "DomA" qui sont membres de "Grp1" ou "Grp2"
- Le type d'accès pour un utilisateur spécifique (recherche et affichage, ajout de documents, administration des collections, etc.) dépend des droits d'accès génériques attribués à cet utilisateur dans le domaine "DomA".
- Si le système de protection "*File System Security*" est activé, alors l'utilisateur X ne verra que les documents pour lesquels il a obtenu les droits requis au niveau du système de fichiers sous-jacent.

Note :

- Si les documents de la collection "ProX" proviennent de plus d'une source de données, il est possible de restreindre les accès à la première source de données au groupe d'utilisateurs "Grp1", etc. (restrictions par source de données).

- Les droits d'accès à une collection spécifique sont donnés à un ensemble de groupes utilisateurs (et non pas à des utilisateurs individuels). Les caractéristiques des droits d'accès individuels des utilisateurs correspondants sont déterminées au moment de leur enregistrement. Cette simplification ne constitue pas une réelle restriction. En combinant avec le principe des domaines, il est toujours possible d'établir des droits d'accès différenciés dans toutes les configurations imaginables.

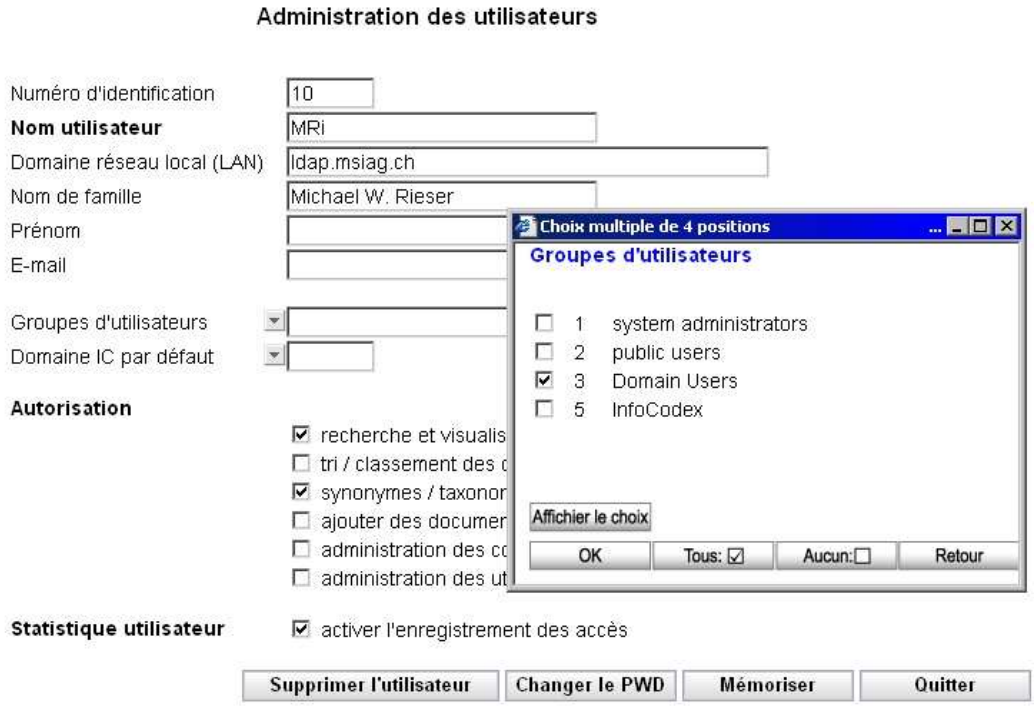
3. Administration Utilisateurs et Droits d'Accès

Cette fonction est activée par le bouton

dans la barre de menu principal d'InfoCodex.

3.1 Données utilisateur et groupes d'utilisateurs

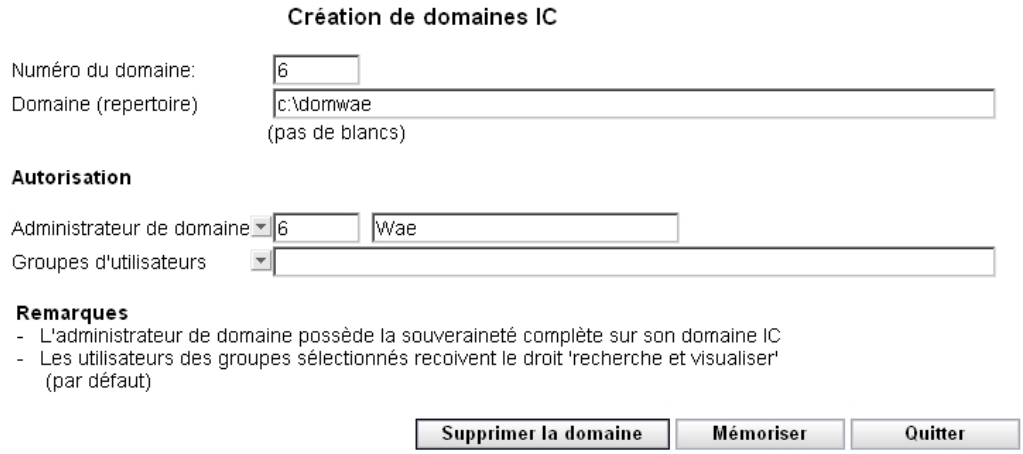
Les droits de l'utilisateur peuvent être définis individuellement pour chaque domaine pour lequel l'utilisateur est enregistré.



3.2 Création de domaines IC (régions autonomes)

Seuls les administrateurs système (utilisateurs avec les privilèges "d'administration utilisateur" sur le domaine IC principal) peuvent créer un nouveau domaine IC.

Un administrateur de domaine est désigné à la création d'un nouveau domaine IC. Ce dernier possède alors la souveraineté complète sur son domaine. Même l'administrateur système initial n'a pas accès au nouveau domaine, sauf si l'administrateur de domaine lui en a donné expressément le pouvoir.
Un administrateur système peut toutefois supprimer n'importe quel domaine IC.

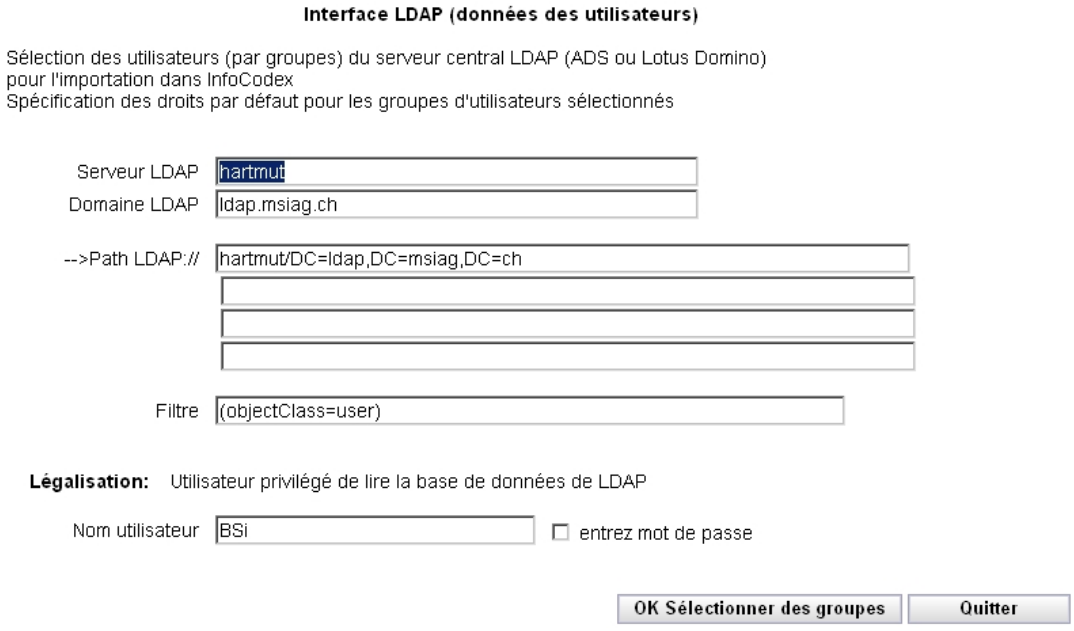


3.3 Interface LDAP à l'administration centralisée des utilisateurs

L'administration utilisateurs d'InfoCodex peut être intégrée à l'administration utilisateurs générale du réseau informatique sous-jacent via le standard *LDAP*.

Sous réserve qu'un serveur LDAP soit utilisé comme point unique d'administration (par exemple"Active Directory Service" ou "Lotus Domino Server"), les données utilisateurs et leurs groupes peuvent être périodiquement importées dans InfoCodex depuis le serveur central LDAP.

Il est recommandé d'introduire un groupe utilisateur spécial, par exemple "Infocodex", intégrant tous les utilisateurs devant utiliser InfoCodex. L'administrateur système peut alors sélectionner tous les utilisateurs appartenant à ce groupe, et le programme interface d'InfoCodex importera tous ces utilisateurs en même temps que les noms des groupes d'appartenance ayant au moins un utilisateur.



Conséquences de l'importation des données utilisateurs via LDAP :

- Les utilisateurs qui ne sont plus présents dans le serveur central LDAP sont supprimés des listes utilisateurs de l'application InfoCodex

- De nouveaux utilisateurs d'InfoCodex et leurs groupes d'appartenance sont insérés dans l'administration utilisateurs d'InfoCodex

- Pour les utilisateurs existants, leurs groupes sont mis à jour. Les droits d'accès de plus haut niveau accordés dans InfoCodex restent en vigueur (par exemple, un administrateur de domaines ne perd pas ses privilèges à l'intérieur de son domaine IC).

- Les nouveaux groupes, ayant au moins un des utilisateurs InfoCodex, sont importés.

Pour que les modifications survenues sur le serveur central LDAP prennent effet dans InfoCodex, le programme interface doit être lancé (ou alors, le lancement périodique du programme d'importation doit être réalisé via le séquenceur de tâches).

3.4 Sécurité du système de fichiers

L'activation de l'option "File System Security" (*FS security, sécurité File System*) implique que les droits d'accès établis dans le système de fichiers du réseau sous-jacent soient respectés. Cette fonction offre une mesure de protection supplémentaire, parallèlement aux mécanismes de sécurité propres à InfoCodex.

La sécurité "File System" garantit que l'utilisateur InfoCodex ne peut voir que les documents auxquels il pourrait avoir accès sur le système de fichier du réseau sous-jacent.

Niveaux de sécurité offerts par InfoCodex (voir Section 7.2)

Niveau de sécurité	Sécurité File System	Les documents protégés apparaissent dans les listes de résultats	Le résumé statistique affiche le nombre total de résultats
0	non activé	oui*	oui
1		non	oui
2		non	non
3	activé	oui *	oui
4		non	oui
5		non	non

*) Les documents protégés sont masqués avec des * dans la liste des résultats de la recherche : ils ne peuvent cependant pas être vus

En ce qui concerne l'importation des documents dans la phase de génération d'une collection ("ajout de documents"), l'option de sécurité FS implique que les documents qui sont lus et analysés sont seulement ceux pour lesquels l'utilisateur qui effectue l'importation a les droits d'accès requis. En conséquence, un utilisateur ne peut pas ajouter à une collection des documents protégés inaccessibles à son compte utilisateur.

L'option sécurité FS est mise en place pour l'ensemble du système informatique par l'administrateur système. Pour des collections individuelles particulières, la sécurité FS peut être désactivée. Les règles suivantes s'appliquent :

- dans la phase d'importation de documents ("Ajout de documents"), la sécurité FS mis en place pour l'ensemble du système par l'administrateur système est respectée.

- pour la recherche et la consultation de documents, c’est la sécurité FS de chaque collection prise indépendamment qui est effective.

L’activation de la sécurité FS a naturellement un impact sur la performance. Sa mise en œuvre pour des collections contenant un très grand nombre de documents doit être soigneusement réfléchie.

4. Gestion des Collections

4.1 Créer/ supprimer une collection

Ces fonctions concernent la création et la mise à jour des collections individuelles. Elles sont décrites dans la Section 4.4, Partie 1 du manuel utilisateur.

N.B. :

- Lorsque des documents sont importés en mode différé (voir Section 4.3, Partie 1), les scripts correspondants peuvent être consultés ou modifiés avec le bouton "**C2 Sources de données**"

- La suppression d'une collection (bouton "**C3 Effacer la collection**") efface toutes les données de la base InfoCodex concernant la collection. Les documents appartenant à la collection, demeurent naturellement inchangés.

4.2 Ajouter une collection préparée en format "txt"

En règle générale, les documents sont chargés dans InfoCodex avec la fonction "Ajout de documents" dans la rubrique"Contenu" du menu principal. Après la sélection des sources de données, les agents de collecte d'InfoCodex rassemblent les documents sélectionnés, les convertissent en simple format de fichier TXT et les stockent, temporairerement, pour être utilisés, plus tard, par les programmes d'extraction de texte (text mining). En parallèle, les agents de collecte établissent une liste "toc.lst" des documents importés, dans laquelle sont enregistrés les noms des document et leurs méta-données.

Alternative

La préparation de fichiers TXT pour importation dans InfoCodex peut évidemment être réalisée de façon indépendante par d'autres moyens. Les documents doivent alors être rassemblés dans un répertoire et accompagnés d'un fichier liste appelé "toc.lst", contenant les noms et les méta-donnés disponibles de tous les fichiers devant être importés.

Le fichier liste "toc.lst" doit avoir la structure d'enregistrement suivante :

```
f2.txt|10.11.01|20|pdf|D:\widas32\gp\GPCIV.PDF|A.Meier|Aggregation by civ
f8.txt|24.06.01|32|pdf|D:\widas32\gp\GPSEX.PDF|A.Meier|Aggregation by sex
______ ________ __ ___ _____________________ _______ ____________________
fname  Date     GP Fmt Nom complet de la source Auteur Titre du document
```

"fname" est le nom du fichier converti en format TXT. Il peut être précédé par son chemin d'accès.
"GP" est une estimation du contenu graphique du document original (en pourcentage).

Le caractère "|" est utilisé comme séparateur. Le premier champ (nom du fichier TXT) est obligatoire.

Les autres champs sont optionnels. Date se réfère à la dernière modification du fichier source.
Pour le format du fichier source (Fmt), on utilise la codification suivante :

word=document Word, pdf=fichier PDF, html=fichier HTML, text=fichier TXT, mail=email, xls=Excel, ppt=Powerpoint, xml=fichier XML, ps=Postscript, jpg=image JPG, etc.

4.3 Réorganiser la carte thématique de l'information

Quand de nouveaux documents sont ajoutés à une collection existante ils sont simplement attribués aux champs existants de la carte thématique dont ils se rapprochent le plus en fonction de leur contenu. Ces nouveaux documents, toutefois, n'ont pas d'influence sur la structure même de la carte thématique de la collection.
Si l'on souhaite le contraire, c.-à-d., que les documents ajoutés depuis la dernière opération de classification/réorganisation, influencent la structure thématique de la collection, il est alors nécessaire de procéder à une réorganisation de la collection. Ceci est à recommander, de toute façon, quand le nombre de documents ajoutés représente 10% à 20% du volume original de la collection, ou si les nouveaux documents relèvent de thèmes notablement différents.

Avec la fonction "**C4 Réorganiser la catégorisation**", les documents ne sont pas réimportés et analysés, mais la catégorisation et l'indexation sont rafraîchies. En conséquence, la durée du traitement n'est pas trop importante.

La fonction "**C5 Régénération de la collection**" (voir Section 6.2) permet une régénération complète de la collection avec une nouvelle importation de l'ensemble des documents mis à jour.

4.4 Suivi de l'état du traitement d'une collection

Cette fonction permet l'affichage de l'état du traitement des phases d'importation et d'analyse en cours. Elle permet également de visualiser le fichier contenant le rapport de traitement.

Les étapes du traitement peuvent être suivies grâce à un code numérique dans le menu "**C1 Paramétrer la collection**". Les chiffres ont les significations suivantes :

0	aucun document importé
1	importation en cours (lancement en interactif)
2	importation en cours (lancement en mode différé)
3	importation complète; analyse de contenu en cours
-1	OK, tous les documents sont chargés et analysés
-99	régénération requise (après des modifications dans la (les) base(s) linguistique(s))

Le code peut être réinitialisé à 0, si nécessaire (voir Section 5.3)

5. Intervention sur la Catégorisation

Pour influencer la catégorisation des documents ("classement dans une bibliothèque thématique existante") et l'allocation de mots-clés, les trois possibilités qui suivent peuvent être mises en œuvre :

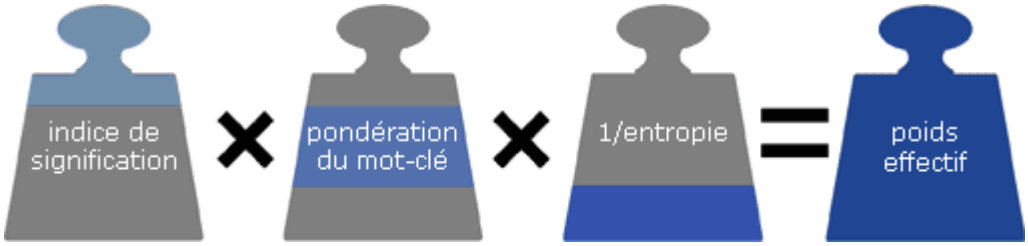
- *Définition/ Paramétrage des mos-clés* — Déclaration d'une liste de mots ou expressions devant avoir une pondération forte durant la phase d'analyse de contenu
- *Base de données frontale* — Préparation d'une base linguistique spécifique (avec ses liens vers l'arbre de la taxonomie), contenant des 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 y est trouvé, 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. Par ailleurs, il est également possible de définir à l'avance des catégories fixes et de désactiver l'automatisme des réseaux de neurones.

5.1 Paramétrage des mos-clés

Principe

Une liste de mots ou expressions (termes) peuvent avoir un attribut d'importance (un facteur de pondération de 2, 4 et 8 fois). Ces termes sont stockés dans une table avec leur facteur de pondération respectif. Une telle table peut être attribuée à une collection particulière au moment de sa création (voir Section 4.4, Partie 1). Ces mots ou expressions ont alors une influence plus forte pendant la phase d'analyse de contenu.

Le poids global effectif du mot est déterminé à partir de la combinaison d'un indice de signification (qui est enregistré dans la base de données linguistique d'InfoCodex), de son paramètre de pondération, et de son entropie pour l'ensemble de la collection considérée.



Explication :

- *Indice de Signification* — Un code, dans la base linguistique, pour le contenu informatif d'un mot ou expression variant de 0 à 4 :

  0 mot de liaison peu significatif, qui est ignoré pour la phase d'analyse de contenu (exemple "the", "in", "le", "la", "dans")

  4 mot rempli de sens et dénué d'ambiguïté .par exemple "World Health Organization"

- *Pondération des mots-clés* — Facteur de pondération assigné via le paramétrage des mots-clés:

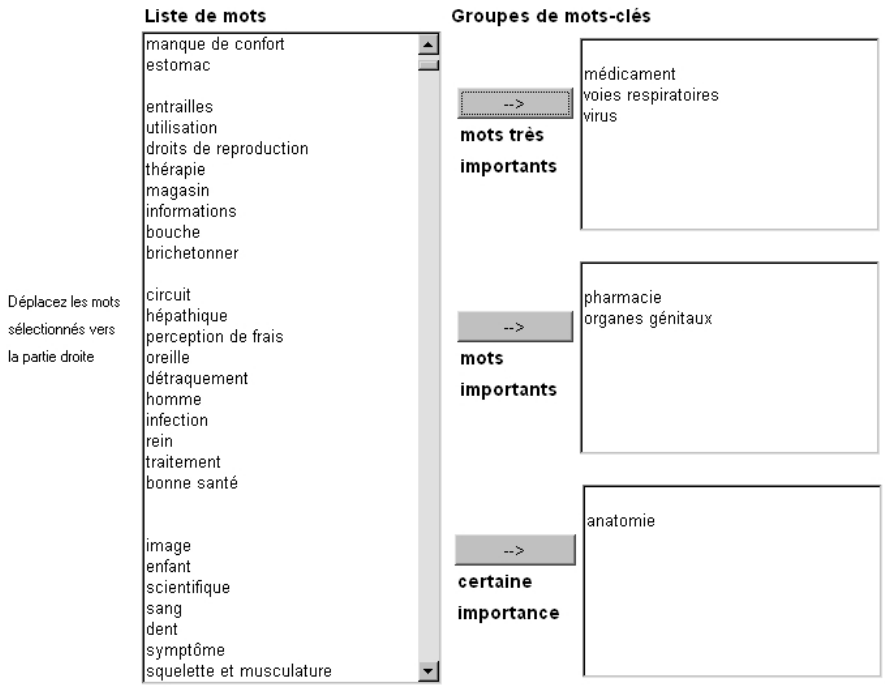
  8 mots très importants
  4 mots importants
  2 mots d'une certaine importance
  1 autres mots (non spécifiés dans le paramétrage)

- *Entropie* — Une mesure de l'incertitude/indétermination d'un mot dans le contexte de la collection de documents considérée.
  Par exemple, quand le mot "Reuters" apparaît dans presque tous les documents, il a une entropie plus importante et contribue très faiblement à la différentiation du contenu des documents. Un tel mot dans ce contexte reçoit, par conséquent, un poids plus faible.

Procédure pour le paramétrage des mots-clés

a) <u>Sélectionner la fonction "Paramétrage Mots-clés" et choisir une liste de mots existants :</u>

soit - les mots les plus pertinents d'une collection existante
ou - des mots provenant de la base de données InfoCodex
ou - des mots que l'utilisateur fournit dans un fichier texte



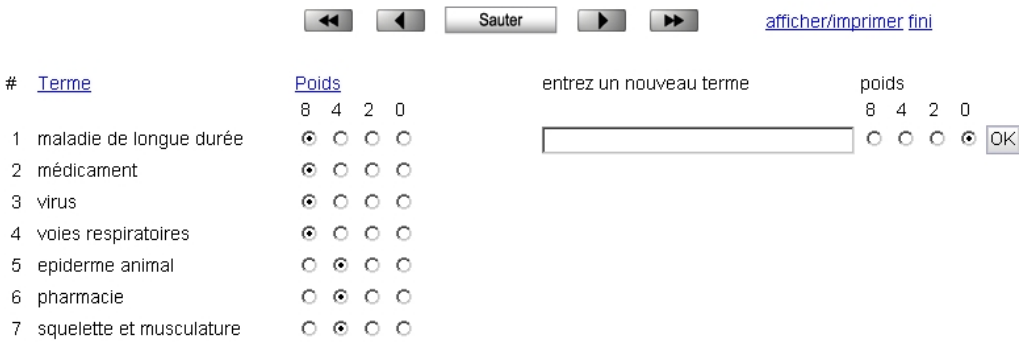
Les mots à pondérer sont tout d'abord sélectionnés avec la souris dans la partie gauche. Ils sont ensuite déplacés vers la partie droite en cliquant sur le bouton qui correspond au niveau de pondération désiré. Après avoir ainsi attribué tous les mots souhaités à leur groupe d'importance, ils peuvent être sauvegardés sous forme d'une table qui contient donc les mots-clés et leur attribut de pondération.

L'utilisateur est alors sollicité pour donner à cette table un nom de fichier qui doit obligatoirement posséder l'extension ".kw". Elle est ensuite rangée dans le répertoire racine du domaine IC concerné.

Toutes les tables ainsi définies sont disponibles pour toutes les collections du domaine IC concerné.

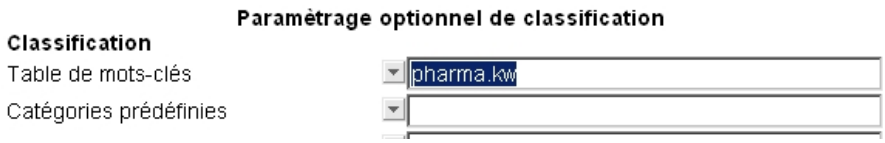
Il est également possible de modifier et d'imprimer une table créée au préalable. La modification du facteur de pondération de chaque terme est réalisée avec des boutons radio. Le choix du poids "0" se traduit par la suppression de ce mot-clé de la table.

La liste peut être triée par terme ou par poids. La flèche simple permet de la faire défiler vers le haut ou vers le bas. La flèche double d'aller directement au début ou à la fin de la liste. Les mots ou expressions ajoutés à la liste doivent naturellement avoir un facteur de pondération supérieur à zéro.



b) Assignation d'une table de mots-clés à une collection

Une table de mots et expressions ainsi spécialement pondérés peut être attribuée à une collection donnée à l'aide de la fonction "Créer/ supprimer une collection".



Pendant l'analyse de contenu, les termes correspondants sont donc pris en considération avec davantage de force. Ceci influence à la fois la catégorisation et l'allocation des descripteurs à chaque document individuel.

Si une table est assignée après la création d'une collection, alors la fonction "Réorganiser la carte thématique de l'information" doit être exécutée.

5.2 Base de données linguistique frontale

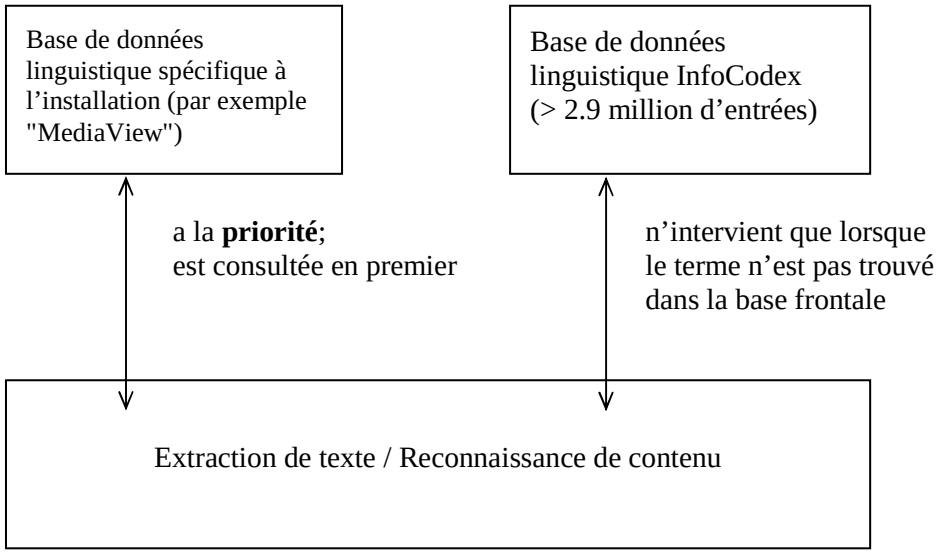
La base de données linguistique d'InfoCodex contient plus de 2.9 millions d'entrées et couvre pratiquement tous les domaines des connaissances. Elle est basée sur des travaux reconnus tels que WordNet de l'Université de Princeton, EuroVoc, Agrovoc, Jurivoc, des taxonomies de l'industrie de la production d'électricité, des télécommunications, de la bancassurance (UBS), ainsi que de nombreuses associations et organismes dépendants de l'Organisation des Nations Unies. Une adaptation à des environnements spécialisés n'est presque jamais nécessaire.

Toutefois, une amélioration peut être obtenue par la mise en œuvre d'une base de données linguistique frontale qui peut avoir un impact significatif sur la qualité des descripteurs des documents.

Illustration : InfoCodex interprète en anglais *SMD* comme "surface mounted device" → component → method → etc.
Supposons que dans un certain environnement d'entreprises, de travail ou géographique, *SMD* soit plutôt utilisé comme abréviation de "Swiss Media Database ".

Remède : Le terme SMD peut être redéfini dans une base frontale et placé dans le même groupe de synonymes que "Swiss Media Database" (avec le chemin d'accès à la taxonomie : name of organization → economic branch → economy/finance).

Principe :



La base de donnée frontale est préparée sous forme d'une feuille Excel et peut être chargée dans InfoCodex avec la fonction correspondante.

Préparation d'une base de données frontale

a) Feuille Excel avec des extensions de l'arbre de la taxonomie (option)

L'arbre de la taxonomie InfoCodex consiste en sept niveaux de hiérarchie et englobe environ 4.100 noeuds. Chaque nœud est identifié par un hyperonyme (= un terme de sens plus général qui recouvre la branche correspondante et celles de sens plus restrictif qui en dépendent ; voir Section 7, Partie 1 du manuel).

Si un utilisateur souhaite introduire des noeuds additionnels dans la taxonomie (par exemple pour une structuration beaucoup plus fine dans un domaine spécialisé), l'arborescence supplémentaire est préparée dans une feuille Excel sur un PC client.

La feuille doit contenir 5 colonnes texte comportant chacune 30 caractères (les colonnes C, D et E peuvent aller jusqu'à 34 caractères):

A	noeuds mère (hyperonyme en anglais)	champ obligatoire
B	noeuds fille (hyperonyme en anglais)	champ obligatoire
C	noeuds fille (hyperonyme en allemand)	optionnel
D	noeuds fille (hyperonyme en français)	optionnel
E	noeuds fille (hyperonyme en italien)	optionnel
F	noeuds fille (hyperonyme en espagnol)	optionnel

Les hyperonymes en anglais doivent être uniques. Les hyperonymes de la colonne A (noeuds mère) doivent être soit déjà présents dans l'arborescence de base d'InfoCodex, soit déclarés dans la colonne B d'une des lignes précédentes.

newnodes.xls

	A	B	C	D	E
1	**Newnodes.xls**				
2	**Déclaration des Noeuds Additionnels dans la Taxonomie**				
3					
4	**Noeuds mère**	**Hypernymes pour les noeuds fille en les 4 langues (le hypernyme en anglais est obligatoire)**			
5	**(anglais)**	**(anglais)**	**(allemand)**	**(français)**	**(italien)**
6					
7	reinsurance	reinsurance contract	Rückversicherungsvertrag	contrat de réassurance	contratto di riassicurazione
8	reinsurance contract	reinsurance premium	Rückversicherungsprämie	tarif de réassurance	premio di riassicurazione
9	reinsurance contract	reinsurance commission	Rückversicherungsprovision	commission de réassurance	commissione di riassicurazione
10	reinsurance	reinsurance law	Rückversicherungsgesetz	droit des réassurances	legge sulle riassicurazioni
11					

Explication

"reinsurance" est un noeud déjà présent dans la taxonomie en langue anglaise d'InfoCodex :

ECONOMY/FINANCE (Niveau 7) → *economic system* (Niveau 6) → *insurance* (Niveau 5) → *reinsurance* (Niveau 4)

Les déclarations de la table Excel introduisent les nouveaux nœuds (soulignés) comme structure supplémentaire de niveau inférieur :

reinsurance (Niveau 4) → *reinsurance contract* (Niveau 3) → *reinsurance premium* (Niveau 2)

→ *reinsurance commission* (Niveau 2)

→ *reinsurance law* (Niveau 3)

b) Feuille Excel avec vocabulaire

La substance de la base frontale doit être saisie dans une table de vocabulaire à 6 colonnes. Les termes sont alloués à des groupes de synonymes et munis de divers attributs.

Vocabulary.xls

	A	B	C	D	E	F
2	Mots/ Expressions pour la Base de Données Frontale					
3						
4	No.	Lan	Type	Sig	Mots/ Expression	Hypernyme
5	Grp.	ge		nif.		(anglais)
6						
7	102	e	1	2	coinsurance	insurance
8	102	e	1	2	co-insurance	
9	102	d	1	2	Mitversicherung	
10	102	f	1	3	assurance additionnelle	
11	102	f	1	2	coassurance	
12	102	i	1	2	coassicurazione	
13						
14	105	e	1	3	underground economy	economic structure
15	105	e	1	3	black economy	
16	105	e	1	3	informal economy	
17	105	e	1	3	shadow economy	
18	105	e	1	3	twilight economy	
19	105	e	1	3	counter-economy	
20	105	e	1	3	grey economy	
21	105	e	1	3	submerged economy	
22	105	e	1	3	unofficial economy	
23	105	d	1	3	Schattenwirtschaft	
24	105	d	1	3	Untergrundwirtschaft	
25	105	d	3	2	schattenwirtschaftlich	
26	105	f	1	3	économie souterraine	
27	105	i	1	3	economia sommersa	

La table de vocabulaire doit être structurée comme suit :

Col.	Format	Description	
A	N5	numéro de groupe [a)]	
B	T1	langue	e = anglais d = allemand f = français i = italien s = espagnol 0 = indépendant (nom, nomenclature)
C	N1	type (qualificatif)	1 = nom 2 = verbe 3 = adjectif 4 = autres (adverbe, pronom)
D	N1	indice de signification [b)]	0 = sans signification ("mot de liaison") 1 = peu significatif, mais peut être utilisé comme terme de recherche dans la recherche par synonymes 3 = important et sans ambiguïté 4 = très important et sans ambiguïté

E	T30	mot / expression par exemple Cours Européenne de Justice
F	T30	hyperonyme anglais [c)]

[a)] Numéro de groupe

Il s'agit d'un numéro arbitraire mais unique pour un groupe de synonymes. Tous les termes ayant la même signification (les synonymes) doivent être regroupés sous le même numéro. Le mot qui apparaît d'abord pour une langue spécifique est le chef du groupe de synonymes correspondant (à moins que le chef de groupe ne soit explicitement désigné par un indice de signification ≥ 5). Le chef de groupe est le mot le plus représentatif d'un groupe de synonymes ; il peut être affiché comme neurone ou étiquette de document dans InfoCodex.

[b)] Indice de signification

Il varie de 0 à 4 en fonction (par exemple) des règles de bon sens spécifiées ci-dessous. Pour désigner de façon explicite le chef de groupe d'un groupe de synonymes, on peut ajouter 5 à l'indice de signification qui prend alors une valeur de 5 à 9. Si l'indice de signification n'est pas indiqué dans la table, InfoCodex génère une valeur par défaut.

[c)] Hyperonyme

Pour relier un groupe de synonymes à l'arbre de la taxonomie, un hyperonyme de langue anglaise doit être spécifié, qui, bien entendu, correspond soit à un noeud de la taxonomie InfoCodex, soit à un noeud défini par l'utilisateur. La saisie de cet hyperonyme est optionnelle. Les hyperonymes valides d'InfoCodex peuvent être exportés sous la forme d'une table Excel.

Règles de bon sens pour l'indice de signification

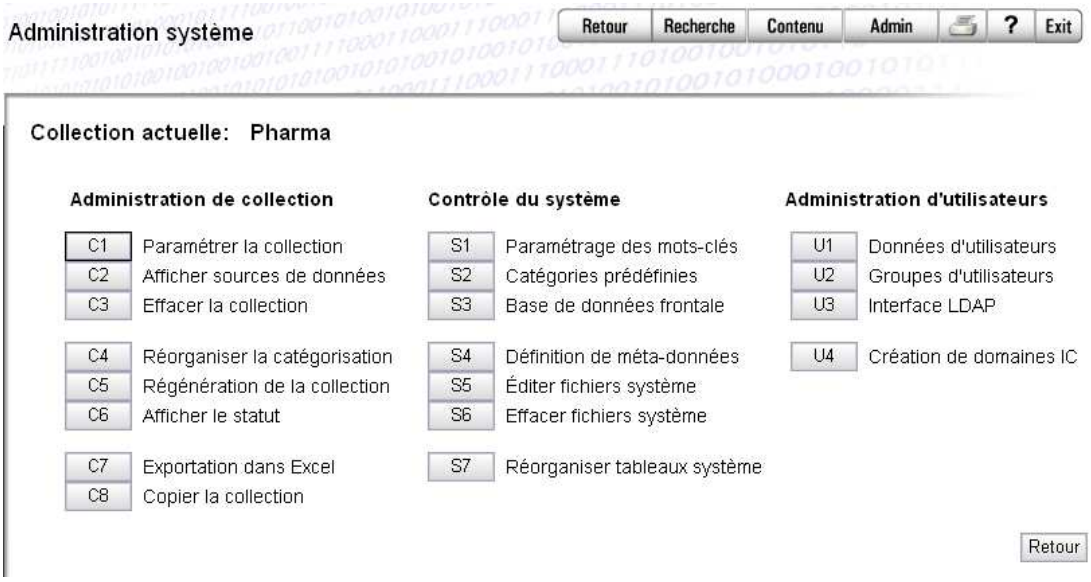
Type	Défaut	1 - 2 lettres ou sans signification	3 - 4 lettres ou plusieurs significations	> 16 lettres ou important	très important
Nom	2 apartment	0 UN, check	1 bank	3 banking institution	4 nuclear power station
Verbe	1 breathe	0 shall	0 ou 1 come sell	2 dehydrogenate	3 internationalize
Adjectif ou Autres	1 commercial	0 otherwise	0 ou 1 well lazy	2 multicellular	3 positively charged

Pour désigner les chefs de groupe qui n'apparaissent pas en première position :
Ajouter 5 à l'indice et donc,

Indice effectif	0	1	2	3	4
Indice des chefs de groupe	5	6	7	8	9

c) Mise en place de la base de données frontale

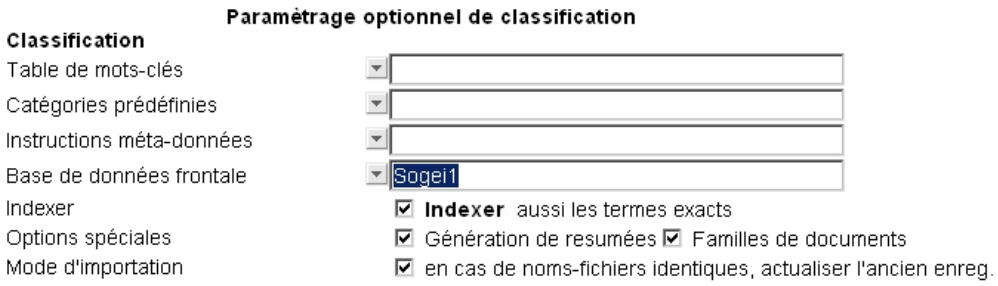
Il faut finalement importer la table Excel dans InfoCodex et l'aligner avec la base linguistique standard :
→ **Admin** → **S3 Base de données frontale**



Avant d'importer la base de données frontale, il faut lui attribuer un nom. Elle doit être rangée dans le répertoire central des programmes InfoCodex (par exemple : `c:\infocodex`). Elle est disponible pour tous les domaines IC définis.

d) Affectation d'une base de données frontale aux collections

Afin d'être utilisée, une base frontale doit être explicitement attribuée au moment de la création d'une collection :
→ **Admin** → **C1 Paramétrer la collection** → puis sélection de la table dans le champ "Base de données linguistique frontale"



Si une base de données frontale est affectée à une collection préexistante, le code de suivi de l'évolution du traitement de la base est basculé sur 99 (c.-à-d. que la collection doit être régénérée, voir Section 6.2).

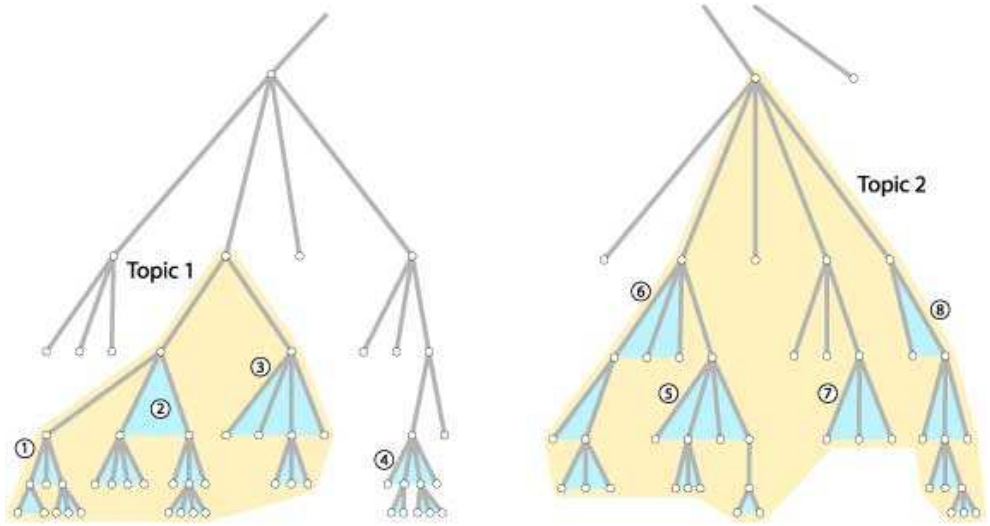
5.3 Catégories et structures prédéfinies

InfoCodex a la capacité d'identifier les thèmes des documents d'une collection sans aucune intervention humaine. Toutefois, un utilisateur peut souhaiter classer les documents selon des thèmes de son choix définis à l'avance, même si cela peut résulter en une baisse de la qualité de la structuration thématique de la collection. En général, ceci est le cas dans les activités d'automatisation de processus pour lesquelles le système doit reproduire des structures d'organisation en place, ou encore pour des tâches exploratoires sur un sujet précis.

InfoCodex offre trois possibilités pour la mise en œuvre de catégorie prédéfinies (spécifications pour l'organisation de la carte thématique de l'information). Il s'agit de :

- Catégorisation Ad-hoc — une recomposition à l'aide de la souris des résultats de la catégorisation automatique
- Catégorisation Analytique — une spécification efficace, bien que relativement complexe des aménagements de la catégorisation
- Catégorisation fixe avec apprentissage — Une définition fixe, et faite au préalable des thèmes principaux et l'apprentissage à la catégorisation à l'aide de collections échantillons

Dans le troisième cas, la catégorisation est dite "*entraînée*" et il ne reste qu'un système de spécification fixe pour l'allocation des documents des nouvelles collections (c.-à-d. les réseaux de neurones sont désactivés pour ce type de classification). L'aperçu, ci-après, de la méthodologie de classification aide à la compréhension des trois procédures.



Combining Agglomeration Clusters into Thematic Topics

InfoCodex est muni d'une taxonomie universelle à sept niveaux. Pour catégoriser de façon optimale les documents d'une collection donnée à l'aide de réseaux de neurones auto-organisants, un espace de contenu à 100 dimensions est taillé sur mesure pour chaque corpus. En faisant cela, 98 composants structurels de l'arborescence (①, ②, ③ etc. comme illustré ci-dessus) sont déterminés comme étant ceux qui reflètent le mieux les termes contenus dans l'ensemble de la collection. Les composants structurels voisins sont regroupés en thèmes principaux (Thème 1, Thème 2, etc. comme ci-dessus). Ces 98

structures, ainsi qu'un composant pour les ratios graphique et numérique du contenu et un composant résiduel technique, constituent les coordonnées de l'espace de contenu à 100 dimensions.

Les 98 composants structurels ont, naturellement, une influence très significative sur la classification par les réseaux de neurones. Il est donc possible d'intervenir sur la catégorisation, en tout cas en partie, en modifiant les caractéristiques de ces composants.

Les composants déterminés de façon automatique par le système pour une collection donnée peuvent être affichés via : **Admin**. → **S2 Catégories prédéfinies** → **Afficher les caractéristiques des composants** (voir figure ci-dessous).

Cette liste montre, en particulier, quels concepts sont présents dans la collection de documents concernée, et avec quel niveau de pertinence.

Composants structurels pour la collection de documents "Pharma"

	Composant	Niveau	Rélévance	Thème principal
1.01	**activité humaine/motif**	6		
	activité humaine/motif	6	984	
1.02	**trait psychologique**	6		
	trait psychologique	6	1085	
1.03	**état psychologique**	6		
	état psychologique	6	822	
1.04	**état physiologique**	6		
	état physiologique	6	1071	état physiologique
	trouble	5	624	état physiologique
	maladie/infirmité	5	1218	état physiologique
	maladie	4	1342	état physiologique
	maladie infectieuse	2	445	état physiologique
	tumeur	3	365	état physiologique
	infection	5	392	état physiologique
	blessure	5	296	état physiologique
	pathologie	5	291	état physiologique
1.09	**ennuis/malheur**	6		
	ennuis/malheur	6	917	
2.01	**groupe sociale**	6		
	organisation	5	381	
2.04	**circonscription administrative**	6		
	nation	5	709	circonscription administrative
	nation européenne	4	607	circonscription administrative
	ville/cité	5	412	circonscription administrative
2.07	**procédure légale**	6		
	procédure légale	6	1428	
3.02	**système économique**	6		
	système économique	6	1020	
3.03	**établissement commercial**	6		
	gestion/direction	5	592	
	lieu d'affaires	5	294	

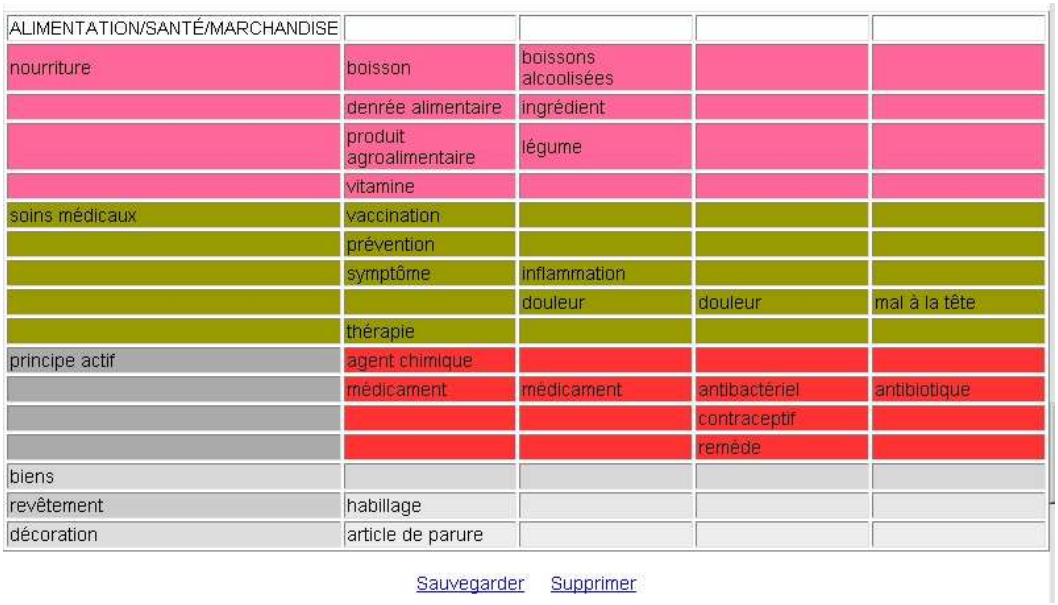
Méthodes pour la définition de catégories prédéfinies

a) Catégorisation Ad-hoc

Cette méthode simple, claire et qui utilise la souris comme mode opératoire est disponible sous → **Admin** → **S2 Catégories prédéfinies** → **Catégorisation Ad-hoc**.

Il faut tout d'abord indiquer un nombre de nœuds maximum à afficher. Le chiffre recommandé se situe entre 300 et 1000. Le système affiche alors les parties de la taxonomie les plus significatives pour la collection traitée. Les niveaux de hiérarchie les plus élevés sont des hyperliens et le détail vers les niveaux inférieurs sont ordonnés sur la droite. Les colorations grises des champs individuels représentent le niveau d'importance

dans la collection : les champs en gris foncé sont davantage mis à contribution que ceux en gris clair. Les surfaces colorées représentent les sections de la taxonomie qui sont utilisées comme thèmes principaux de classification de la collection.



ALIMENTATION/SANTÉ/MARCHANDISE				
nourriture	boisson	boissons alcoolisées		
	denrée alimentaire	ingrédient		
	produit agroalimentaire	légume		
	vitamine			
soins médicaux	vaccination			
	prévention			
	symptôme	inflammation		
		douleur	douleur	mal à la tête
	thérapie			
principe actif	agent chimique			
	médicament	médicament	antibactériel	antibiotique
			contraceptif	
			remède	
biens				
revêtement	habillage			
décoration	article de parure			

Sauvegarder Supprimer

Pour modifier les critères de classification du système, il suffit de sélectionner, supprimer et fusionner les noeuds choisis de la taxonomie pour créer les nouveaux thèmes de classement principaux désirés. Les opérations suivantes sont possibles :

1. *Cliquer sur un noeud dans la zone grise* : la branche de l'arborescence qui commence au noeud sélectionné et qui englobe tous les noeuds des niveaux hiérarchiques inférieurs qui en découlent se voit assigner une couleur et devient par conséquent un des thèmes principaux. Le nom de ce thème peut être modifié.

2. *Cliquer sur un noeud dans la zone colorée* : la branche qui commence au noeud sélectionné est marquée et peut alors
   a) soit être désactivée (c.-à-d devenir une zone grise neutre)
   b) soit être déclarée comme nouveau thème principal
   c) soit être fusionnée avec un autre thème.

Lorsque toutes les modifications souhaitées ont été réalisées, la configuration peut être sauvegardée en tant que modèle de catégorisation prédéfinie en cliquant sur "Mémoriser". L'utilisateur est alors sollicité pour saisir un nom pour le modèle, qui doit obligatoirement avoir l'extension ".pm" (predefined map, carte prédéfinie).

Le modèle est alors disponible pour toutes les collections du domaine IC concerné.

Pour rendre la catégorisation ad-hoc effective, la fonction "**C4 Réorganiser la catégorisation**" doit être exécutée.

b) Catégorisation Analytique

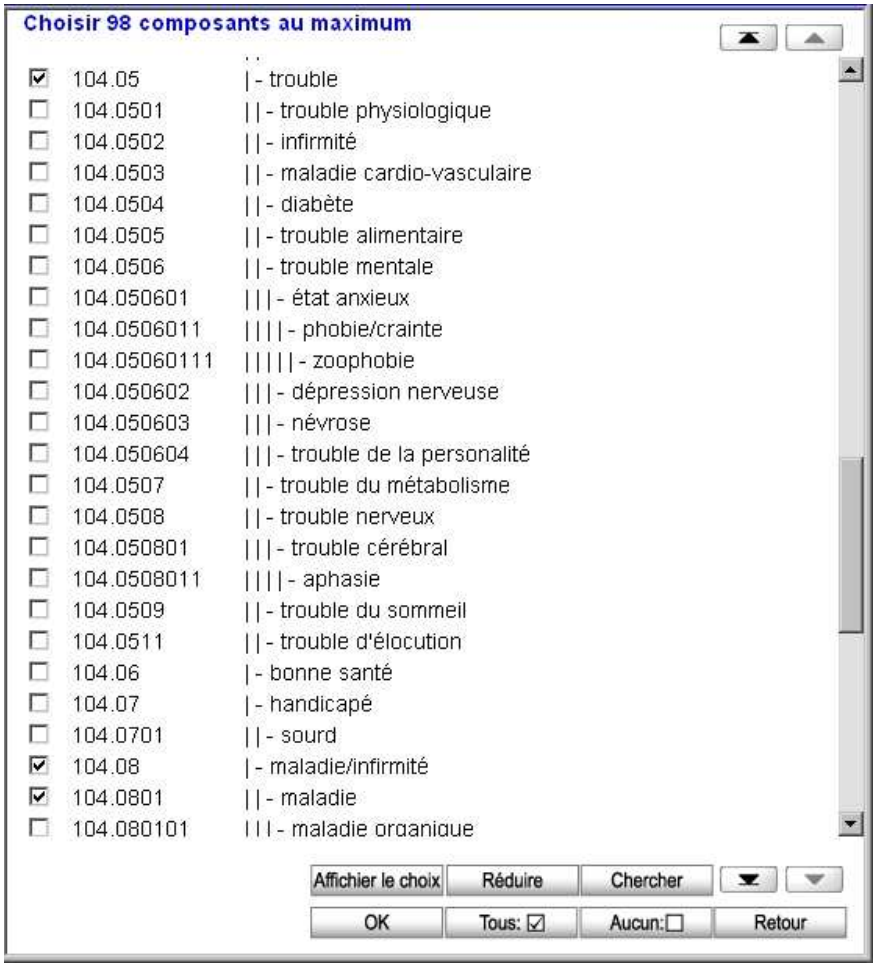
Cette méthode, aux fondements théoriques solides, consiste en l'assignation directe des 98 composants caractéristiques (dimensions de l'espace). Ceci est réalisé à travers le chemin de commandes → **Admin** → **S2 Catégories prédéfinies** → **Catégorisation analytique**.

Etape 1 : Charger une sélection existante de composants

La liste de composants générés automatiquement est utilisée comme point de départ

Etape 2 : Modifier la sélection de composants

Cocher les nœuds de la taxonomie devant devenir des composants caractéristiques (voir figure ci-dessous). En choisissant un noeud, les noeuds des niveaux hiérarchiques inférieurs qui en dépendent sont aussi inclus (par exemple. 104.05 jusqu'à 104.0511 inclus). Si, cependant, un noeud des niveaux inférieurs de la branche est à son tour coché, il sera exclu de la branche mère précédente pour devenir par lui-même un composant caractéristique de l'espace (dans l'exemple ci-dessous, 104.08 englobe tous les nœuds inférieurs sauf 104.0801, qui forme son propre composant).



Si moins de 98 composants sont ainsi sélectionnés, InfoCodex complète automatiquement jusqu'à 98.

Etape 3 : Définir les thèmes principaux

Sélection de 2 à 20 sujets principaux pour la représentation de la carte thématique de l'information.

Etape 4 : Assignation des thèmes principaux

Les 98 composants définis précédemment sont assignés aux thèmes de l'étape 3.

Après l'étape 4, les critères de catégorisation sont sauvegardés dans une table PM et une réorganisation de la carte thématique de l'information doit être effectuée.

c) Catégorisation fixe avec apprentissage

Avec cette variante, la catégorisation doit faire l'objet d'un processus "d'apprentissage" à l'aide d'un échantillon de documents dont le contenu est connu. Les documents de l'échantillon sont assignés aux thèmes particuliers de la carte de l'information.

La catégorisation, dans ce cas, est donc prédéterminée par cette phase d'apprentissage. Elle est fixe pour toutes les importations qui suivent et les capacités d'automatisation des réseaux de neurones sont désactivées.

La procédure principale se déroule comme suit :

Etape 1 : Préparation d'un échantillon pour l'apprentissage

Il faut tout d'abord définir une liste de 2 à 20 sujets principaux correspondant aux thèmes devant apparaître dans la carte thématique de l'information, par exemple. "monopoly", "taxe" et "propriété de l'Etat". Ensuite on prépare une collection d'échantillonnage, dont le contenu est parfaitement connu et maîtrisé, de façon à ce que les documents caractéristiques d'un des thèmes choisis soient regroupés dans un même répertoire. Il y a donc un répertoire par thème, chacun avec au moins 20 documents.

Etape 2 : Création d'un fichier script pour l'apprentissage

Les instructions pour définir la catégorisation prédéfinie à partir d'un échantillon sont indiquées dans un fichier script situé dans le répertoire racine du domaine IC concerné, par exemple

```
"c:\icdata\instruct.txt"
```

où "instruct.txt" est un nom choisi arbitrairement pour le script.

Format du fichier script (fichier TXT)

```
Topic=monopoly
Dir=C:\demo2\sogei\monopoli

Topic=tax
Dir=C:\demo2\sogei\entrate
KW=wine, liquor, cigars

Topic=state property
Dir=C:\demo2\sogei\demanio
KW=castle, national park, railway
KW=*IDB: state property
KW=*UDB: state property
```

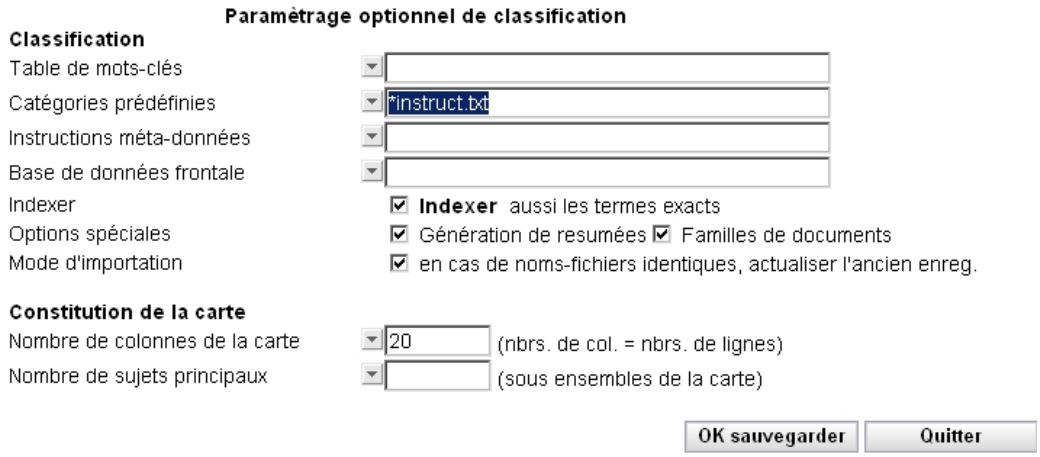
Explication

Topic	Les thèmes doivent être entrés en langue anglaise et correspondre aux hyperonymes d'InfoCodex (les noeuds de l'arborescence dans la taxonomie). Les hyperonymes d'une base de donnée frontale définie par l'utilisateur sont également autorisés.
Dir	Les documents de ce répertoire et de ses sous-répertoires doivent être représentatifs du thème correspondant. Ils sont utilisés pour la phase d'apprentissage à la catégorisation, avec l'aide de mots-clés optionnels.
KW	Mots-clés optionnels KW=wine, liquor, cigars signifie : ajouter les termes "wine, liquor, cigars" à chaque document du répertoire correspondant (seulement pour la phase d'apprentissage) KW=*IDB: state property signifie : ajouter les termes de la base InfoCodex appartenant à l'hyperonyme "state property " à chaque document du répertoire correspondant (seulement pour la phase d'apprentissage) (de la même façon pour la base de données frontale de l'utilisateur avec KW=*UDB: …) La saisie de mots-clés n'est pas obligatoire. C'est seulement un moyen supplémentaire de caractériser les thèmes principaux ciblés.

Etape 3 : Apprentissage à l'aide de l'échantillon

Créer une nouvelle collection (en cliquant sur le symbole correspondant sur la partie gauche de l'écran), entrer le nom de la collection et sélectionner "Options avancées".

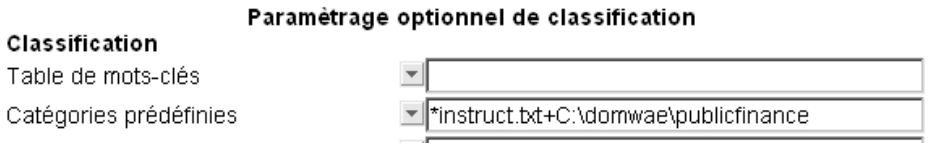
Entrer le nom du fichier script, précédé par un astérisque, dans le champ "Catégories prédéfinies", par exemple *instruct.txt. Il est recommandé d'indiquer un nombre de colonnes pour la carte, par exemple 20. En effet, la dimension de la matrice de cellules (neurones) générée pourrait devenir trop faible si le nombre de documents dans la collection traitée s'avèrait insuffisant.



L'apprentissage commence dès que l'on a cliqué sur le bouton "OK sauvegarder".

<u>Etape 4 :</u> En option, amélioration de la carte générée

Après la génération automatique de la carte, il est possible de la réorganiser avec la technique de catégorisation ad-hoc pour l'aligner sur un concept personnel particulier. Pour rendre cette modification effective, il faut, bien-sûr, enchaîner avec la fonction "**C4 Réorganiser la catégorisation**".



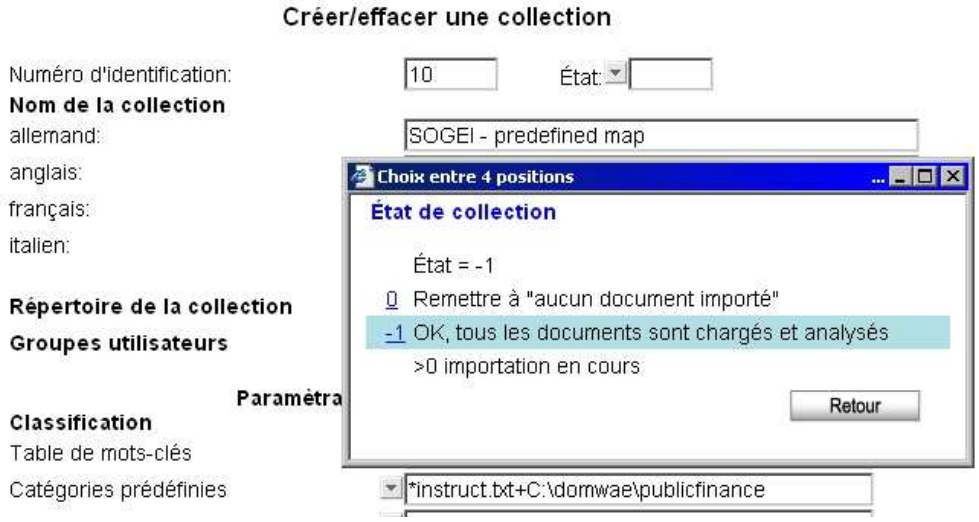
L'entrée dans le champ "Catégories prédéfinies" doit être complètée par le nom de la table PM spécifiée dans la catégorisation ad-hoc.

<u>Etape 5 :</u> Réinitialiser le statut et importer les documents

La carte prédéfinie ou modèle de catégorisation est donc établie. Il est désormais possible de charger des documents en vue de leur classification en fonction de cette carte thématique prédéfinie. Sa structure reste inchangée.

Avant l'importation des documents, le statut de la carte d'apprentissage doit être réinitialisé.

→ **Admin** → **C1 Paramétrer** → flèche vers le bas dans le champ statut → choisir "Pas de documents importés".



Après la réinitialisation, toute collection de documents peut être importée dans la carte thématique formée à partir de l'échantillon.

→ **Contenu** → **Ajouter des documents** → Sélectionner les sources de données.

6. Fonctions Auxiliaires

6.1 Importion/Exportation de feuilles Excel

InfoCodex offre plusieurs possibilités d'échange de données entre le serveur et le tableur Excel d'une machine cliente. Pour que ces transferts soient possibles, il faut avoir téléchargé et installé correctement le logiciel de transfert de fichiers fourni, à partir de la page de d'accueil d'InfoCodex (voir Section 4.2, Partie 1 §1).

Les principales utilisations en sont l'exportation de résultats soit de recherche (voir Section 3.2, Partie 1), soit de catégorisation et de paramétrage de mots-clés (voir Section 5, Partie 1).

D'autres possibilités sont disponibles sous : → **Admin** → **C7 Exportation dans Excel**

On peut mentionner, en particulier, l'exportation des mots d'une collection qui ne sont pas connus dans la base linguistique d'InfoCodex. La liste peut alors être utilisée, d'une part, pour localiser des erreurs d'orthographe dans les documents et d'autre part pour servir de base à la création d'une base frontale.

6.2 Régénération des collections / Mise à jour des statistiques d'utilisation

Ces fonctions sont à utiliser surtout après une mise à jour de la base de données linguistique ou lorsqu'une base de données frontale est assignée à une collection existante. Ces modifications ont en effet pour conséquence que la matière codée extraite des documents est incorrecte et qu'elle doit être reconstruite.

Les moyens mis à disposition permettent cette régénération sans avoir à répéter les actions déjà effectuées d'importation des documents, de construction de tables de mots-clés et de catégorisation.

Statistiques d'utilisation

Il y a un relevé statistique permanent du nombre de clics sur chaque champ (neurone) de la carte thématique de la collection traitée, c.-à-d. de la consultation des documents par les utilisateurs. Ceci peut être utile pour identifier les utilisateurs qui consultent souvent les mêmes groupes d'information et qui donc peuvent s'avérer être des experts du domaine.

Le relevé statistique reste inchangé après une simple réorganisation. Si une régénération est sélectionnée (avec une nouvelle importation de documents), l'utilisateur a la possibilité de conserver le relevé statistique existant ou le remettre à zéro.

6.3 Copier/ Modifier/ Supprimer

Copier (C8 Copier la collection)

Permet de copier une collection existante dans une nouvelle collection ("Clonage")

Modifier les fichiers script (S5 Éditer fichiers système)

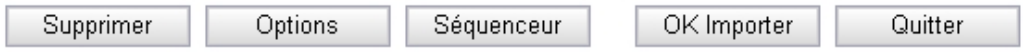
Permet à l'administrateur système de modifier les paramètres du système ("options.ictxt", etc.; voir Section 7).

Supprimer (S6 Effacer fichiers système)

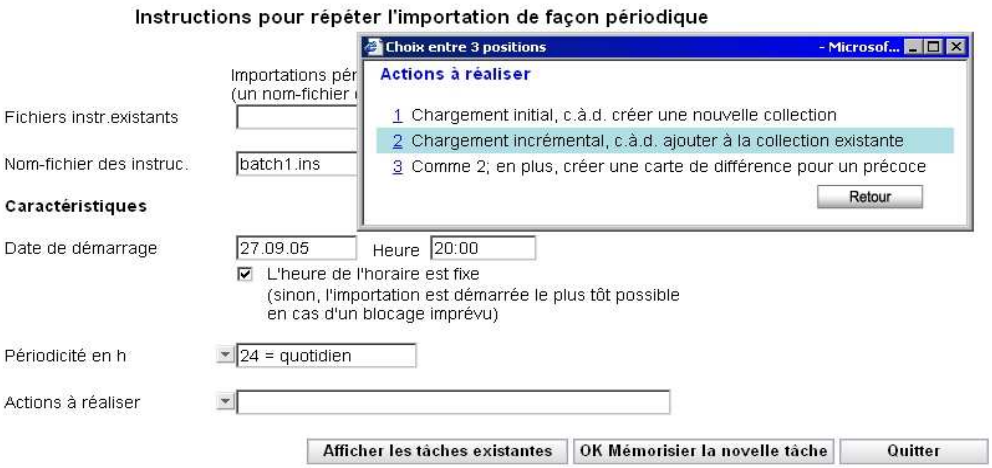
Permet la suppression des tables de mots-clés, des tables PM (catégories prédéfinies), des bases de données frontales, etc.

6.4 Séquenceur de recherche/ Surveillance et contrôle

Lorsque l'importation de documents doit être répétée de façon périodique, l'action peut être automatisée sous la forme d'un programme qui déclenche les tâches correspondantes en mode différé (voir Section 4.3, Partie 1) :



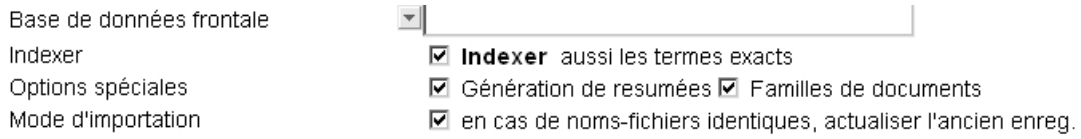
En sélectionnant l'option "**Séquenceur**", une boîte de dialogue apparaît pour spécifier l'heure de départ et la périodicité du traitement.



Dans la fenêtre "Actions à réaliser" on choisit si une nouvelle collection doit être créée ("Chargement initial") ou si les documents doivent être ajoutés à la collection existante ("Chargement Incrémental").

Spécifications pour un "Chargement Incrémental"

Lorsque l'on exécute des importations périodiques différées, on peut, pour les fichiers ayant le même nom que ceux déjà présents dans la collection, soit les ajouter à nouveau séparément, soit les mettre à jour. Cette option doit être initialisée au moment de la création de la collection en cochant la case "Mode d'Importation"



Lorsque la case n'est pas cochée, les documents avec des noms de fichier identiques sont ajoutés à la collection (voir Section 4.4, Partie 1).

Surveillance et contrôle

WimonWin.exe C'est un programme de surveillance, préparé en tant que "Tâche planifiée" (ou "cronjob" sur les plateformes Unix) au moment de l'installation d'InfoCodex.
Il est activé toutes les minutes pour gérer le statut du daemon de communication du serveur "wihdaem" (avec redémarrage en cas de problème). De plus, toutes les 30 minutes il lance tous les travaux d'importation en attente.

ICregen.exe Ce programme est lancé toutes les 30 minutes (par défaut) par le processus décrit ci-dessus et accomplit les travaux en mode différé qui sont en attente.
Il lit le fichier de planification de travaux "importschedule.ictxt", situé dans le répertoire central de programmes InfoCodex et lance les travaux en attente d'exécution. Il incorpore également, dans le fichier "importschedule.ictxt", tous les travaux nouvellement définis par les utilisateurs et stockés dans le fichier d'attente "importqueue.ictxt", et enfin supprime ceux dont les utilisateurs ont demandé l'arrêt.

N.B. : La périodicité d'ICregen est déterminée par une entrée dans le fichier "importnext.ictxt" (le nombre dans la première ligne de ce fichier représente l'intervalle de temps en minutes).

7. Paramétrage Système

7.1 Serveur de proxy ("proxy.ini")

Lorsque le serveur InfoCodex doit avoir accès à Internet ou à un réseau Intranet à travers un serveur de proxy, l'adresse du proxy doit être spécifiée dans le fichier "proxy.ini" situé dans le répertoire de programmes d'InfoCodex (par exemple dans "C:\infocodex"). Les données suivantes doivent être spécifiées :

- nom ou adresse IP du proxy
- port du proxy
- nom utilisateur } si le proxy nécessite une authentification
- mot de passé } (en général non nécessaire)

Format du fichier "proxy.ini"

```
www.extern.proxy
8080
```

Par défaut (absence de fichier "proxy.ini")

S'il n'y a pas de fichier "proxy.ini", InfoCodex utilise le paramétrage du proxy du navigateur installé sur le serveur InfoCodex. S'il y a, à cet endroit, une adresse de proxy, InfoCodex l'utilise. Dans le cas contraire, il assume qu'il n'y a pas de proxy en activité et s'efforce d'établir une connexion directe.

7.2 Options de traitement ("options.ictxt")

Les processus d'arrière plan d'InfoCodex peuvent être contrôlés avec différents paramètres. Ils concernent pour la plupart les agents de collecte, l'importation de documents, l'extraction de texte (text mining), la catégorisation et l'indexation.

Les paramètres sont aussi également stockés dans un fichier central TXT nommé "options.ictxt", dans le répertoire central de programmes d'InfoCodex.

Format du fichier "options.ictxt"

```
SECURITY     = 4
RETARDANT    = 10
CLEAN        = 30000
SHOWMSG      = 1
OL_LocalPrf  =  infocodex-lokal
ExchangePrf  =  infocodex
```

Les différents paramètres sont explicités ci-dessous à l'aide des valeurs par défaut.

a) SECURITY = 0

Spécification des niveaux de sécurité (voir Section 3.4)

b) RETARDANT = 10

L'importation et l'analyse des corpus documentaires sont réalisées par des tâches d'arrière plan qui ne doivent pas être interrompues avant leur bonne fin d'exécution. Certaines versions de Windows peuvent présenter des difficultés de gestion de caches disques lorsque des opérations d'entrée-sortie particulièrement lourdes sont mises en oeuvre. Ceci peut entraîner des arrêts intempestifs des tâches en cours et empêcher leur fin d'exécution.

Dans ce cas de figure, il faut utiliser l'option RETARDANT qui a pour effet de ménager des pauses bien contrôlées aux étapes critiques.

Valeurs recommandées :	0	pour les systèmes (Windows) bien adaptés
	10	pour les systèmes (Windows) avec une fréquence d'horloge très élevée ou multiprocesseurs
	30	pour les situations extrêmes

c) CLEAN = 30000

L'étape de l'extraction de texte (text mining) requiert de grandes ressources mémoire. Par précaution, InfoCodex effectue une opération de regroupement des éléments inutiles (nettoyage mémoire) quand le nombre de mots nouvellement reconnus excède une certaine limite.

Valeurs recommandées :	30000 pour les systèmes avec 256 à 512 Mo de RAM (=défaut)
	15000 pour les systèmes avec 128 Mo de RAM
	50000 pour les systèmes avec plus de 1 Go de RAM

d) NUMBER = 20

Les documents comprenant de grandes tables de données statistiques pourraient conduire à des ensembles énormes de synonymes différents correspondants à tous les nombres rencontrés. Pour prévenir les surcharges qui en résulteraient, les différents nombres à prendre en considération peuvent être plafonnés avec l'option NUMBER.

Exemples :	20	si les seuls 20 premiers nombres d'un document présentent un intérêt (N° client, Code référence, etc.)
	1000	si les 1000 premiers nombres sont intéressants

e) TRACE = 0

Quand les documents sont importés à partir d'un moteur de recherche Internet ou d'un site Internet/Intranet et quand InfoCodex ne détecte pas tous les liens qu'il serait supposé trouver, il est recommandé d'utiliser l'option TRACE.

Exemples :	0	aucune trace des liens
	-2	tous les liens reconnus sont enregistrés dans le fichier de rapport "import.log"

Les options suivantes se rapportent aux boîtes aux lettres Outlook et aux serveurs Exchange.

f) SHOWMSG = 0

Pendant l'importation de courriels, leur contenu est sauvegardé dans un fichier texte pour affichage ultérieur. Les messages importés de Microsoft Outlook sont également sauvegardés sous le format propriétaire MSG. Les fichiers MSG apparaissent sur la machine cliente juste

comme n'importe quel autre courriel, y compris les en-têtes et les pièces jointes, à la condition qu'une version compatible d'Outlook y soit installée.

Valeurs possibles :	0	Afficher les fichiers texte pur
	1	Afficher les messages complet, si disponible (sinon retour aux fichiers texte pur)

g) OL_LocalPrf

Les profils Outlook pour Outlook installés sur le serveur InfoCodex sont utilisés pour importer les messages des fichiers PST (archives Outlook).

Exemple : OL_LocalPrf = infocodex-lokal

h) OL_ExchangePrf

Les profils Outlook pour Outlook installés sur le serveur InfoCodex sont utilisés pour importer les messages du serveur Exchange.

Exemple: OL_ExchangePrf = infocodex

i) ExchangePwd

Mot de passe pour la connexion au serveur Exchange par le profil "OL_ExchangePrf" (si un mot de passe est réellement requis pour "OL_ExchangePrf").

k) OL_CleanPST = 0

Permet de fermer toutes les archives de messages ouvertes (fichiers PST) avant l'importation d'un fichier PST.

Valeurs possibles :	0	Ignore les archives de messages ouvertes
	1F	Ferme toutes les archives de messages ouvertes (sauf celui par défaut)

l) ClientShare = \\infocodexserver\ICExchange$

Lorsqu'une importation de courriels doit être réalisée à partir d'une boîte aux lettres Outlook située du côté client, les courriels importés doivent être copiés temporairement sur le serveur InfoCodex. Ils sont sauvegardés dans des sous-répertoires temporaires, sous des répertoires réseaux spécifiques "ClientShare".

m) ClientDir = InfoCodexProgramDirectory\Exchange

C'est le nom local des répertoires sur le serveur InfoCodex mentionné sous l) ci-dessus.

n) InfoURL

Le nom d'un fichier HTML utilisé pour l'affichage optionnel d'informations spécifiques au client dans le masque standard de recherche InfoCodex.

Exemple : InfoURL = agroscope.html

(les pages d'information qui dépendent de la langue utilisée doivent exister dans "htdocs"/icd,"htdocs"/ice, etc., si leur adresse n'est pas "http:...").

o) translateToEN=0

Option for automatic translation of documents which are not German, English, French, Italian or Spanish.

Possible values:
- 0 do not translate
- 1 use online translation tool (Google translate or Systran) to translate documents into English.
  Primarily supported languages: Russian, Chinese (Mandarin), Arabic, Hindi, Japanese, Portuguese

p) MailServer, MailPort, MailFrom

In order to send news alerts and other notifications, the following three options need to be configured:

MailServer	The host name or IP address of the mail server to use.
MailPort	The port to use (typically 25).
MailFrom	The sender, e.g. "InfoCodex Server <infocodex@example.com>".

InfoCodex will only attempt to send out e-mail if those three options are present.

q) CutUArgs

A space-delimited list of variables to be removed from URLs.
Some websites insert session IDs or other volatile information into hyperlinks, as in "http://www.example.com/page.php?SESSIONID=12345678". To InfoCodex this means that even when the documents are still the same in subsequent imports, the URLs will change every time.
InfoCodex will work around this by removing any variables from a URL that match any of the regex fragments in CutUArgs, e.g.:
CutUArgs = PHPSESSID PHPSESSIONID ASPSESSIONID[A-Z0-9]+ RANDOM

r) AccessLog=0

Store all login attempts (successful and unsuccessful) in a logfile.

Possible values:
- 1 Store all login attempts in %IC%\icaccess.log
- 0 No logging

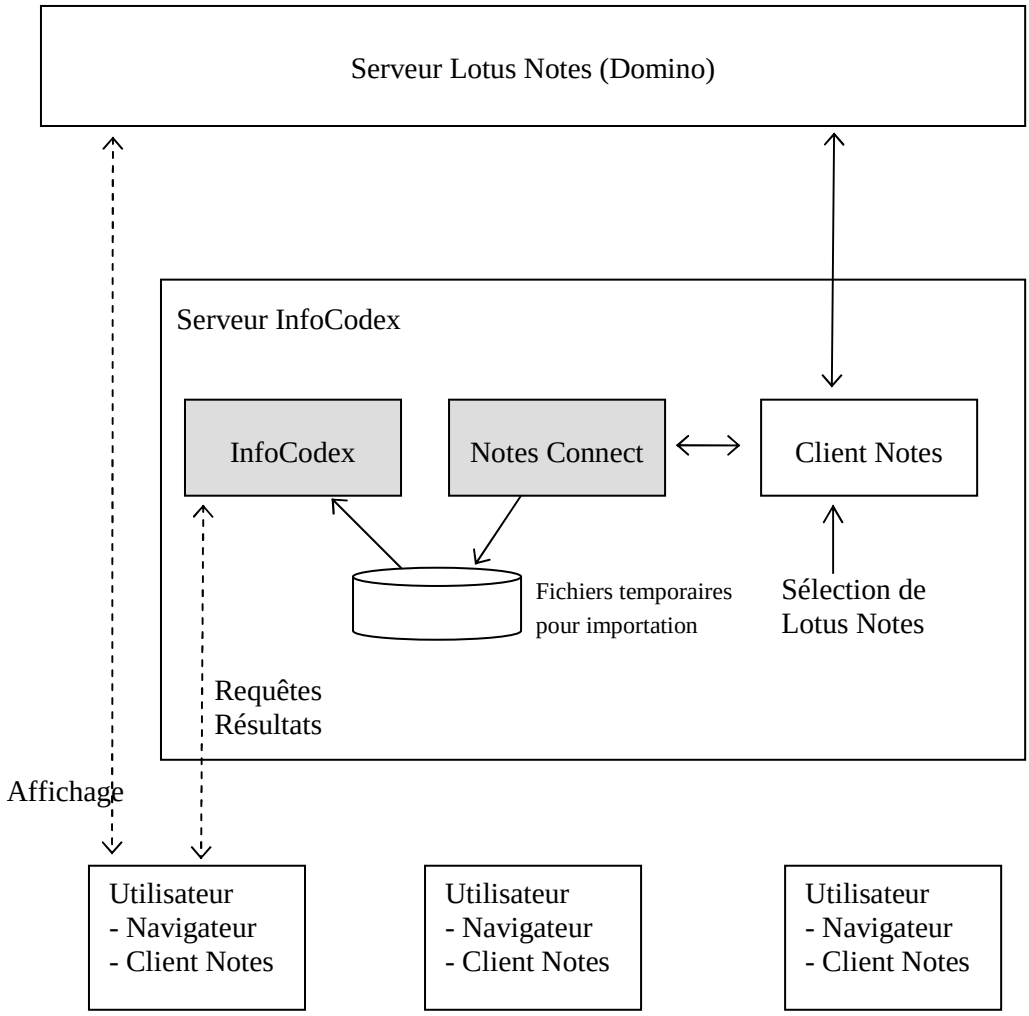
s) TxtFiles

A space-delimited list of file extensions to be considered plaintext (ASCII) files during import.

Possible values:	empty	No effect
	log dat	Files with names ending in .log and .dat will be treated as plaintext files during import

7.3 Interface Lotus Notes

Le module spécial d'InfoCodex "NotesConnect" et le logiciel Notes Client doivent être installés sur le serveur.



Légende :

Modules spécifiques InfoCodex (sur le serveur InfoCodex)

Flux d'informations pendant l'importation

Flux d'informations pendant les recherches (requête, résultats)

Importation de documents Notes

La sélection de documents Notes (et de leurs pièces jointes) dans les diverses bases de données Notes est réalisée à travers le logiciel Notes client (paramétré avec les privilèges nécessaires).

Les documents sélectionnés sont également copiés par une tâche d'arrière plan dans des fichiers temporaires, et supprimés par InfoCodex après la phase d'analyse.

Les documents Notes et leurs pièces jointes sont alors disponibles pour les opérations de recherche d'informations. Les liens vers les bases de données Notes sont conservés dans InfoCodex de telle façon que l'affichage des documents Notes individuels puisse être actionné via Notes Client.

Droits d'accès

Les droits d'accès Notes pour importer des documents Notes peuvent également être reconnus et les accès utilisateurs coordonnés par le serveur LDAP.

7.4 Interfaces Outlook et Exchange Server

Ces deux interfaces nécessitent que Outlook soit installé sur le serveur InfoCodex et que le fichier "options.ictxt" soit correctement défini (voir Section 7.2).

7.5 Restriction à certains types de fichiers ("extensions.ictxt")

La durée d'importation de documents à partir de très grands réseaux informatiques peut être réduite considérablement en spécifiant une liste restreinte d'extensions de fichiers associées aux documents. Cette liste est définie dans un fichier TXT du nom de "extensions.ictxt" dans le répertoire de programmes central d'InfoCodex.

Exemple d'un fichier "extensions.ictxt" :

```
doc,rtf,xls,xlt,ppt,pps,pdf,ps,psc,eps,msg,html,htm,shtml,xml,txt
tif,tiff,jpg,jpeg,jpe,bmp,png,gif,csv,tmp,asc,zip,gzip,gz
pst
```

Activation de iFilters pour des formats de fichier spécifiques

Un iFilter est une dll qui extrait le texte d'un format de fichier spécifique (par exemple, un fichier CAD).
Avec les formats de fichier standard que les agents de recherche InfoCodex manipulent (doc, rtf, pdf etc.) il est possible d'indexer des fichiers de formats arbitraires, à condition que le iFilter correspondant soit installé sur le serveur.

Un iFilter est invoqué par le programme "filtdump" (un composant d'InfoCodex).

S'il y a un iFilter (ou un autre utilitaire) disponible pour un format de fichier particulier, et si ce format doit être indexé (par exemple, les fichiers CAD), alors une ligne doit être ajoutée au fichier texte "extensions.ictxt" comme suit :

```
dwg=filtdump  -b  <in>  >  <out>
```

Il s'agit de la commande qui doit être exécutée pour les fichiers ayant l'extension ".dwg".

(La codification exacte doit être utilisée comme souligné ci-dessous ; filtdump exécutera la dll iFilter correspondante)

Une liste d'extensions multiples de fichiers est acceptable, par exemple

```
dwg,dwf=filtdump  -b  <in>  >  <out>
```

Important :

Si des iFilters spécifiques doivent être utilisés, alors

- soit toutes les extensions standards désirées doivent être listées dans le fichier texte "extensions.ictxt" (doc, xls, pdf, etc.)
- soit le caractère "*" doit être inséré dans "extensions.ictxt", ce qui signifie que toutes les extensions standards doivent être prises en compte.

Exemple 1:

```
*
dwg,dwf=filtdump  -b  <in>  >  <out>
```

Exemple 2:

```
dwg,dwf=filtdump  -b  <in>  >  <out>
```

Dans l'exemple 2, *seuls* les fichiers avec les extensions ".dwg" et ".dwf" seront interprétés

7.6 Table de correspondance des espaces disques ("drives.ictxt")

Dans un environnement réseau Windows, des utilisateurs peuvent allouer des noms particuliers à leurs espaces de stockage sur le réseau. Des correspondances de noms peuvent être déclarées dans un fichier txt "drives.ictxt" dans le répertoire de programmes central d'InfoCodex. Ceci permet aux utilisateurs de voir leur nom d'espace disque habituels comme "R:", "S:" etc. au lieu des noms standard absolus du réseau.

Exemple d'un fichier "drives.ictxt" :

```
S:->\\fal002s\share
R:->\\fal008s\share
Q:->\\fal008s\pool
Z:->\\fal008s\data\info_drive
N:->\\fal008s\private\<user>
P:->\\fal002s\private\<user>
```

La convention "<user>" signifie que le nom de l'utilisateur sera substitué dans le chemin d'accès au lecteur de disque réseau, par exemple

"\\fal008s\private\smith" pour l'utilisateur "smith".

7.7 Paramétrage du Daemon ("icmonitor.ictxt")

Ce fichier contient les paramètres pour les programmes de surveillance "wimonwin" (pour Windows) ou "wimonux" (pour Linux et Unix) respectivement, et le daemon de communication "wihdaem". Il est généré automatiquement à l'installation d'InfoCodex.

Enregistrements spécifiques :

a) Cleanup 03 :00

"Wimonwin.exe" ou "wimonux" respectivement, invoquent un nettoyage journalier (suppression des fichiers auxiliaires restants et redémarrage du daemon de communication). Le paramètre 03:00 définit l'heure du nettoyage.

b) APIDAEM C:\infocodex\pgm\icapidaem.exe

Cet enregistrement doit être présent quand une version API d'InfoCodex est utilisée (par exemple, quand le client de recherche IC-Express est installé).

Cela signifie que le programme daemon nécessaire est lancé et surveillé par "wimonwin.exe" ou "wimonux" respectivement. Le paramètre "C:\infocodex\pgm\icapidaem.exe" indique le nom du programme daemon de l'API.

7.8 Fichier Auxiliaire pour IC-Express ("webserver.ictxt")

Ce fichier doit exister quand le client de recherche IC-Express est utilisé. Il contient le nom du serveur et le nom du répertoire Apache, par exemple

```
LT4
C:\Apache2
```