Nouvelles fonctionnalités du serveur d'investigation Transcriptome : Ajout d'un sous-ensemble du méta-thésaurus UMLS

Solveig Vidal

Université Henri Poincaré

Université Nancy2

Institut National Polytechnique de Lorraine

DESS d'Information Scientifique et Technique

Intelligence Economique

Année 2001-2002

NOUVELLES FONCTIONNALITÉS DU SERVEUR D'INVESTIGATION TRANSCRIPTOME

Ajout d'un sous-ensemble du méta-thésaurus UMLS

Ajout de bases d'arrière-plan

par

Solveig Vidal

Maître de stage : Cecilia Fabry

Stage effectué du 13 mai au 16 août 2002
à l'Institut de l'Information Scientifique et Technique (INIST-CNRS)

Remerciements

Ce stage a été réalisé dans le cadre du DESS Information Scientifique et Technique-Intelligence Economique cohabilité par les trois universités de Nancy (Université Henri Poincaré Nancy 1, Université Nancy 2, Institut National Polytechnique de Lorraine). Il est issu d'une collaboration entre l'Institut National de Recherche et de Sécurité (INRS) et l'Institut de l'Information Scientifique et Technique (INIST). L'objectif du stage consiste à ajouter deux nouvelles fonctionnalités à une application d'une plate-forme documentaire développée par l'INIST mise en place à l'INRS.

Je tiens à remercier :

- Cecilia Fabry pour son soutien durant toute la durée du stage, sa gentillesse.
- Philippe Houdry pour son suivi sur le plan technique, ses précieux conseils et enfin sa relecture du rapport.
- Jacques Ducloy pour la confiance qu'il m'a accordée.
- Alain Zasadzinski pour ses compléments d'information sur l'UMLS et surtout pour nous avoir prêter le CD-Rom UMLS et toute la documentation nécessaire.
- Claire François pour son aide technique en programmation shell.
- Tous les membres du DPS, du centre de documentation et du service informatique de l'INRS pour leur accueil, leur sympathie et leur soutien.

Note : Tous les mots et sigles suivis d'une "*" ont une définition dans le glossaire. Les numérotations entre "[ ]" renvoient à la bibliographie.

Sommaire

Introduction

Les avancées récentes en biologie moléculaire sont à l'origine de l'accroissement exponentiel du nombre d'études portant sur l'analyse des ***génomes****, ***protéomes**** et ***transcriptomes****. La conséquence immédiate est l'augmentation du nombre de publications. Actuellement, le principal défi correspond à l'analyse globale de toutes ces données afin d'en extraire une information biologique pertinente. Ainsi, dans un centre tel que l'Institut National de Recherche et de Sécurité pour la prévention des accidents de travail et des maladies professionnelles (INRS) [10], le centre de documentation utilise des produits tels que le logiciel documentaire « ***AIRS**** Web » et la plate-forme documentaire *Documentation Information LIBrary* (***DILIB****) qui permettent de gérer et exploiter une telle abondance d'informations.

La nouvelle version de la plate-forme DILIB [1], conçue par Jacques Ducloy, responsable du Département Produit et Services de l'Institut National de l'Information Scientifique et Technique (INIST) [2], propose également des fonctionnalités intéressantes pour l'analyse de l'information. Cet aspect a intéressé Bertrand Rihn, chercheur à l'INRS, qui a souhaité utiliser DILIB pour l'exploitation des données bibliographiques liées aux résultats d'une étude des gènes impliqués dans le ***mésothéliome**** humain (cancer de la ***plèvre****). Cette collaboration entre les deux instituts a donné lieu à la création du serveur « Transcriptome », anciennement appelé « Génome » [3,4].

Mon travail a consisté dans un premier temps à une prise en main du serveur en améliorant notamment l'automatisation de certaines étapes lors de sa génération. Cette prise en main a aussi été favorisée par la participation à la relecture de l'article de Bertrand Rihn sur les résultats apportés par le serveur Transcriptome [5]. Dans un second temps, il a fallu intégrer un sous-ensemble du méta-thésaurus ***UMLS**** [6] afin de permettre une navigation à partir des mots-clés ***MeSH**** des corpus Medline. Enfin, le serveur a été enrichi de ***bases d'arrière-plan**** Pascal afin de compléter la couverture documentaire par rapport aux notices Medline de l'existant mais également de suivre l'évolution temporelle des idées et concepts déjà émergents.

1. Présentation de l'Institut de l'Information Scientifique et Technique.

➢ Son rôle et son statut

Unité de service du Centre National de la Recherche Scientifique (CNRS), l'***INIST**** est le premier centre intégré européen d'Information Scientifique et Technique (***IST****).

Fournisseur de copies de documents, producteur de bases de données multilingues et multidisciplinaires recensant l'essentiel de la littérature internationale dans la plupart des domaines de la recherche, l'INIST étend aujourd'hui son offre de services sur internet.

➢ Ses missions

Une mission de service public.

L'INIST a pour principal objectif de servir les différents acteurs de la recherche publique, qu'il s'agisse du CNRS ou d'autres Etablissements Publics à caractère Scientifique et Technique (EPST), ou de l'enseignement supérieur (universités et grandes écoles), afin d'améliorer la collecte, l'analyse et la diffusion de l'information scientifique.

Un accès à l'information pour le milieu socioéconomique.

Les entreprises ont besoin de connaître l'état des recherches dans leur domaine d'activité, afin d'être à même d'adapter au mieux leur propre stratégie de développement. De nombreux laboratoires de recherche privés ont recours quotidiennement aux différents services proposés par l'INIST :

- Services de recherche sur internet (***ARTICLE@INIST****, ***ARTICLESCIENCES****).
- Bases de données (***PASCAL****, ***FRANCIS****).

Développer l'accès à l'information électronique.

L'INIST offre à ses utilisateurs la possibilité d'identifier et de localiser un document, et d'en faciliter l'accès par l'intermédiaire de ses réseaux (service de fourniture de copies de documents primaires). C'est l'un des principaux enjeux lancés aux acteurs de l'Information Scientifique et Technique (IST). C'est dans cette perspective que l'INIST a mis en place en 2001 le portail en IST « ***ConnectSciences**** » qui propose, dans un environnement personnalisé et évolutif, un ensemble de ressources et de services produits par l'INIST et ses partenaires**.**

Développer la veille.

L'INIST étudie et développe de nouveaux outils de veille technologique et documentaire pour le traitement ***bibliométrique**** et ***l'analyse infométrique**** des données issues de diverses sources d'information, et en particulier de ses bases.

Ces applications constituent une aide à l'élaboration de stratégies scientifiques, tant pour les chercheurs que pour les entreprises.

➢ Le Département Produits et Services

Le Département Produits et Services assure la constitution des bases bibliographiques de l'INIST, la fabrication des produits et la mise en place des services et leurs exécutions. Il comprend différents services :

- Des services de production (Fourniture de document, Formation, Traduction)
- Des services scientifiques (Sciences de la vie, Sciences Humaines et Sociales, Sciences Exactes et de l'Ingénieur).
- Des services transversaux (Gestion de Production et Budget, Ingénierie et Partenariat, Cellule de veille).

Ses objectifs consistent à assurer les prestations de production et à entreprendre une mutation technologique, par exemple le déploiement de nouvelles compétences liées aux développements des nouvelles technologies. Dans ce cadre, la boîte à outils DILIB est utilisée tant pour mettre à disposition des résultats de recherche bibliographique (présentation des

résultats de recherches effectuées pour ses clients sous forme de serveurs d'investigations) que dans une optique de mutation technologique (des formations internes à l'utilisation de DILIB).

2. Plateforme Dilib

➢ Définition

Une plate-forme pour l'ingénierie documentaire et l'information scientifique et technique permettant les applications suivantes :

- L'investigation documentaire.
- La construction de Système de Recherche d'Information (***SRI****).
- La mise en place d'outils pour les bibliothèques électroniques.

➢ Historique

C'est le fruit d'un travail collectif qui a connu une première et importante réalisation sous l'ancien nom d'ILIB (Information LIBrary) au sein du Département Recherches et Produits Nouveaux de l'INIST, en coopération avec le Centre de Sociologie de l'Innovation de l'Ecole des Mines de Paris. Cette première application a bénéficié des résultats de nombreux travaux antérieurs :

- La plate-forme de production de l'Association puis Agence Nationale du Logiciel.
- Activités documentaires du ***CIRIL**** et de l'***INALF****.

Ce produit s'est ensuite développé au LORIA et à l'INRIA-Lorraine pour enfin revenir à l'INIST.

➢ Description

Le contenu de cette plate-forme est le suivant :

- Boîte à outils ***SGML/XML****.
- Composants pour construire des Systèmes de Recherche d'Information.

- Modules infométriques.
- Générateurs d’applications infométriques multibases.
- Interfaces Web pour la navigation.

3. l'Unified Medical Language System (UMLS)

L'explosion des connaissances en médecine amène à un besoin d'information de plus en plus pertinent. Or, la multiplication des banques de données dans le domaine biomédical (qui constitue une première difficulté, sachant qu'il faut savoir « où » chercher l'information) entraîne une hétérogénéité des clés d'accès (le vocabulaire étant en effet différent pour chacune d'entre-elles). Face à ces difficultés, la volonté du *National Library of Medecine* (NLM) de construire un outil commun a permis d'engendrer depuis 1986, l'***UMLS**** qui comprend trois sources de connaissances : le ***méta-thésaurus****, le ***réseau sémantique**** et un lexique dit « ***Specialist Lexicon**** » [7, 8]. Nous nous intéresserons uniquement à la première source qui est le méta-thésaurus.

➢ Définition

D.A.B. Lindberg, directeur de la NLM, a proposé, en 1986, la conception et le développement d'un système de langage médical unifié ou "Unified Medical Language System" (UMLS).

Ce système contient l'information sémantique sur les concepts biomédicaux, sur leurs appellations et sur leurs relations entre eux. Il est construit à partir de thésauri, de classifications, de système de codages et il liste tous les termes contrôlés et développés par des organisations telles que ***SNOMED*, AIR93*, DDB00*, DSM****… .

Pour accéder à l'information pertinente, il faut d'une part savoir dans quelle base de données chercher et, d'autre part, être en mesure de formuler la requête (en terme de vocabulaire et de syntaxe). L'UMLS est destiné à pallier ces deux difficultés : le Méta-thésaurus et le Réseau Sémantique permettent de "traduire" la requête de l'utilisateur en "langage UMLS". En plus de ces sources de vocabulaires, il existe des programmes lexicaux s'utilisant avec le méta-thésaurus et le « specialist lexicon » (qui contient tous les variants syntaxiques des termes biomédicaux mais également des verbes qui n'apparaissent pas dans le méta-thésaurus) et permettant de développer des applications avec des bases de données et des systèmes d'indexation personnalisés ou spécialisés. L'UMLS est réactualisé tous les ans, avec

rééditions de la plupart des sources de vocabulaires, ajout de traductions de ces mêmes sources (15 langues différentes).

➢ Le méta-thésaurus : une organisation par concept

Le Meta-thésaurus donne accès à un ensemble uniforme et contrôlé de classifications et vocabulaires biomédicaux dans un format texte ***ASCII**** mais également les liens entre les différents concepts.

Un concept correspond à une définition ou à un sens. Un ensemble de termes peut être lié à un même concept et tous les concepts sont répertoriés dans un même fichier : le MRCON.

L'organisation est la suivante (figure 1) :

Chaque concept a un identifiant unique de concept ou *Concept Unique Identifier* (***CUI** **).

Un concept est subdivisé en plusieurs identifiants uniques communs de chaîne de caractères ***LUI**** (Lexical Unique Identifier) qui correspondent d'une part aux traductions et, d'autre part, aux variations synonymiques.

De même, chaque LUI est lui-même subdivisé en ***SUI**** (*String Unique Identifier*) qui correspondent aux variations pluriel/singulier, orthographiques, capitalisés ou pas.

Chaque concept possède des attributs afin de mieux définir son sens, c'est-à-dire son type sémantique ou les catégories auquel il appartient, sa position hiérarchique et, le plus souvent, sa définition.

L'édition de mai 2002 inclut 871.584 concepts et 2,1 millions de termes dans plus de 95 sources de vocabulaires biomédicales. Certaines sont multilingues.

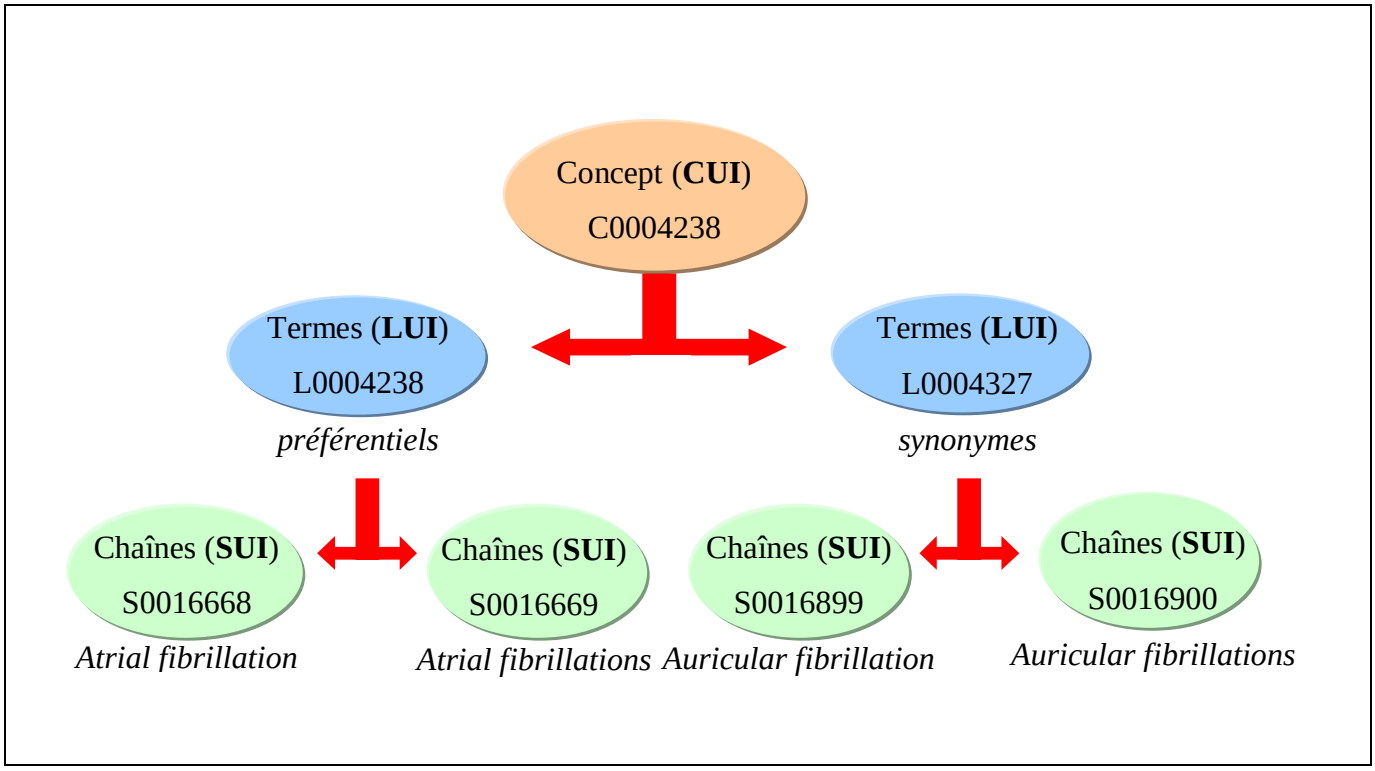


Figure 1 : Le concept et les termes qui lui sont reliés.

Ceci est un exemple simplifié où le multilinguisme n'est pas pris en compte. Chaque traduction aura un LUI différent du LUI originel (anglais) et, pour chaque traduction, il peut y avoir des termes préférentiels et des synonymes. Voir l'extrait du MRCON plus bas.

```
C0000343|ENG|S|L0470427|PF|S0577471|25-Hydroxy ergocalciferol|3|
C0000343|FIN|P|L1507161|PF|S1803070|25-hydroksivitamiini D 2|3|
C0000343|FRE|P|L0177473|PF|S0241954|25-HYDROXYVITAMINE D 2|3|
C0000343|GER|P|L1226360|PF|S1468306|25-Hydroxyvitamin D2|3|
C0000343|GER|S|L1226359|PF|S1468305|25-Hydroxyergocalciferol|3|
C0000343|POR|P|L0319258|PF|S0427563|25-HIDROXIVITAMINA D 2|3|
```

Figure 2 : Extrait du MRCON.

➢ Relations entre différents concepts

Les relations entre les concepts dans le méta-thésaurus proviennent des sources elles-mêmes ou ont été créées par les éditeurs du méta-thésaurus pour lier les concepts.

Neuf types de relations dans le méta-thésaurus.

Dans le méta-thésaurus toutes les relations établies entre les clés CUI sont répertoriées dans un fichier MRREL.

La relation concerne celle établie entre le second concept (désigné sous CUI2 dans le fichier MRREL) **envers** le premier concept. Bien qu'un CUI puisse avoir plusieurs relations avec plusieurs CUI, une entrée MRREL ne concerne qu'une relation entre deux CUI. Ainsi, le fichier MRREL est le plus volumineux des fichiers du méta-thésaurus : il contient plus de neuf millions d'entrées (à comparer au fichier MRCON qui en contient deux millions).

Neuf types de relations :

- **RB** pour « Broader » (générique)
- **RN** pour « Narrower » (spécifique)
- **RO** pour « Other Relations » autres que synonymes, spécifiques ou génériques
- **RL** pour « Like » : les deux concepts sont similaires, synonymes
- **PAR** pour « Parents » : relation parentale ascendante dans la hiérarchie propre à la source de vocabulaire
- **CHD** pour « Child » : relation parentale descendante
- **SIB** pour « Sibling » : relation frère/frère
- **AQ** pour « Allowed Qualifier » : est un qualificateur autorisé dans la source de vocabulaire
- **QB** pour « can be Qualified By » : peut être qualifié par un concept d'une source de vocabulaire du méta-thésaurus.

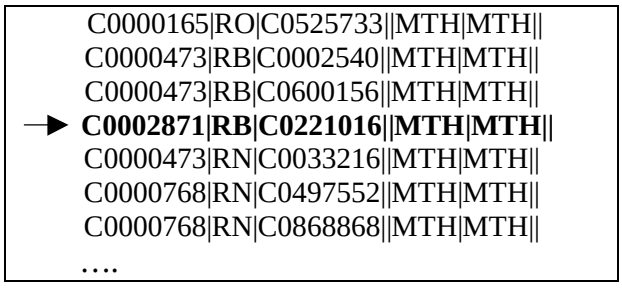


Figure 3 : extrait du MRREL.

C0002871|**RB**|C0221016||MTH|MTH|| : *Red blood cell disorder, NOS* (C0221016) est le **générique** de *Anemia* (C000287).

4. L’application Transcriptome

➢ L’origine du projet

L’amiante et le mésothéliome

De par ses qualités d’isolant et ses propriétés de résistance, l’amiante fut largement utilisé au cours des cinquante dernières années dans l’industrie et le bâtiment. Certaines études ont permis de mettre en évidence le caractère cancérogène des fibres d’amiante. Il a été prouvé que leur inhalation était responsable chaque année en France de nombreuses atteintes pulmonaires (cancer du poumon, fibroses pulmonaires, cancer de la plèvre). Le mésothéliome est le cancer de la plèvre (enveloppe du poumon) résultant en général d’une exposition professionnelle à l’amiante.

L’expression génique

L'ensemble des gènes est défini par le génome. Dans une cellule, un certain nombre de gènes est activé en fonction du type de la cellule et de son environnement. Cet état d'activation des gènes est appelé l'expression des gènes. Les gènes exprimés donnent lieu à la synthèse d'ARN messagers leur correspondant, cette étape s'appelle la transcription. L'ensemble des transcrits (ARN messager) est appelé transcriptome. Ces ARN messagers seront ensuite traduits en protéines qui formeront le protéome.

La technique des puces à ADN

Afin de mieux comprendre la pathogénie du mésothéliome pleural humain (cancer de la plèvre) et de décrire plus précisément ses mécanismes moléculaires, l’équipe de Bertrand Rihn (INRS) [11] a entrepris l’étude des gènes impliqués dans ce cancer [9, 10].

Différentes techniques de biologie moléculaire ont été utilisées à cet effet, en particulier la technique des ***puces à ADN**** (Figure 4). Cette technique récente permet de tester l’état d’activation de milliers de gènes simultanément.

L'expérimentation a consisté à comparer l'expression de 7000 gènes de cultures de cellules saines et de cellules malignes de la plèvre par quantification des ARN messagers.

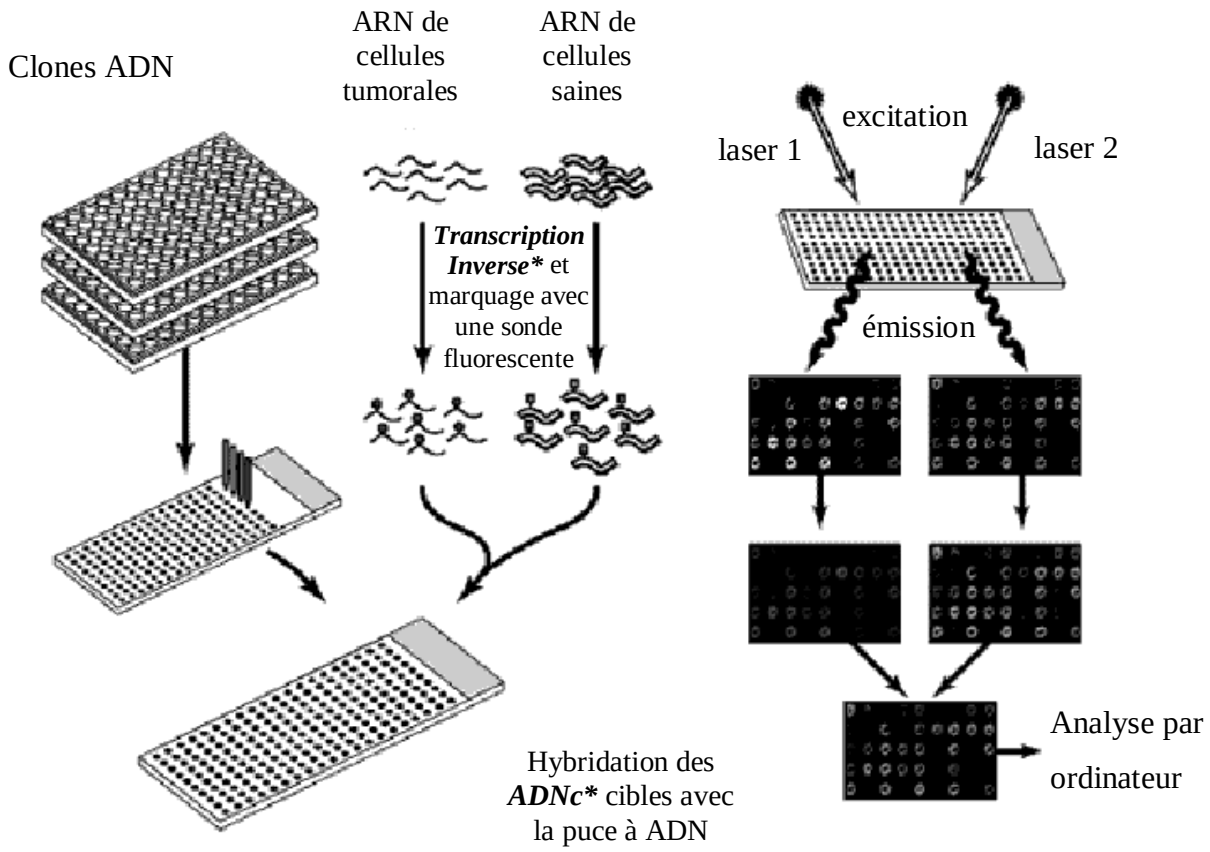


Figure 4 : Principe de la technique des puces à ADN.

Du transcriptome au corpus bibliographique

Les résultats de cette étude ont été rassemblés dans un tableau permettant de visualiser, entre autres, le niveau d'expression de chaque gène dans les cellules saines et cancéreuses de la plèvre. Ces résultats des expériences sur les puces à ***ADN**** permettent également de comprendre certains mécanismes génétiques de la transformation cancéreuse des cellules normales de la plèvre en cellules malignes de mésothéliome. Cette étude a par ailleurs permis d'expliquer la relative résistance du mésothéliome à la chimiothérapie et à la radiothérapie. Afin d'exploiter au mieux les résultats de l'étude, il a été envisagé une analyse bibliométrique réalisée avec la plate-forme DILIB et exploitant des données bibliographiques issues de ***Pubmed**** *(via le numéro d'accession* ***GenBank**** *interrogeable directement dans Pubmed)* en relation avec les gènes mis en évidence par l'expérimentation. L'utilisation d'outils infométriques tels que DILIB permet d'analyser des corpus bibliographiques importants qui ne sont pas exploitables manuellement. Cela est d'autant plus intéressant en génomique compte tenu de l'étendue des bases de données dans ce domaine. L'application mise en place à l'INRS contient deux corpus bibliographiques dans un serveur d'investigation [12], un pour les gènes sur-exprimés (expression différentielle > 1.9), et un pour les gènes sous-exprimés (expression différentielle < -1.6). Cette application de DILIB a été nommée " application Transcriptome ".

L'extraction des corpus s'est faite via les numéros d'accession de GenBank qui sont attribués pour chaque gène. A partir d'un tableau Excel contenant tous les gènes dont l'expression différentielle est suffisamment significative, deux requêtes dans le champ ***Secondary Source Identifier**** (SI) de Medline (qui contient les numéros d'accession GenBank) ont pu être effectuées via les numéros d'accessions GenBank des gènes d'intérêt. Cette méthode, bien qu'indirecte, a permis de générer plus de notices par rapport à une interrogation directe dans GenBank (voir figure 5) (pour plus de détails voir le rapport de Claude Némurat [4]).

Cette extraction s'est faite grâce à un programme ***shell**** « GenereRequetes.sh » (voir figure 6).

Les notices Medline récupérées sont en format texte et non en XML, format indispensable à la prise en charge des notices lors de la navigation dans un serveur d'investigation généré par la plate-forme Dilib.

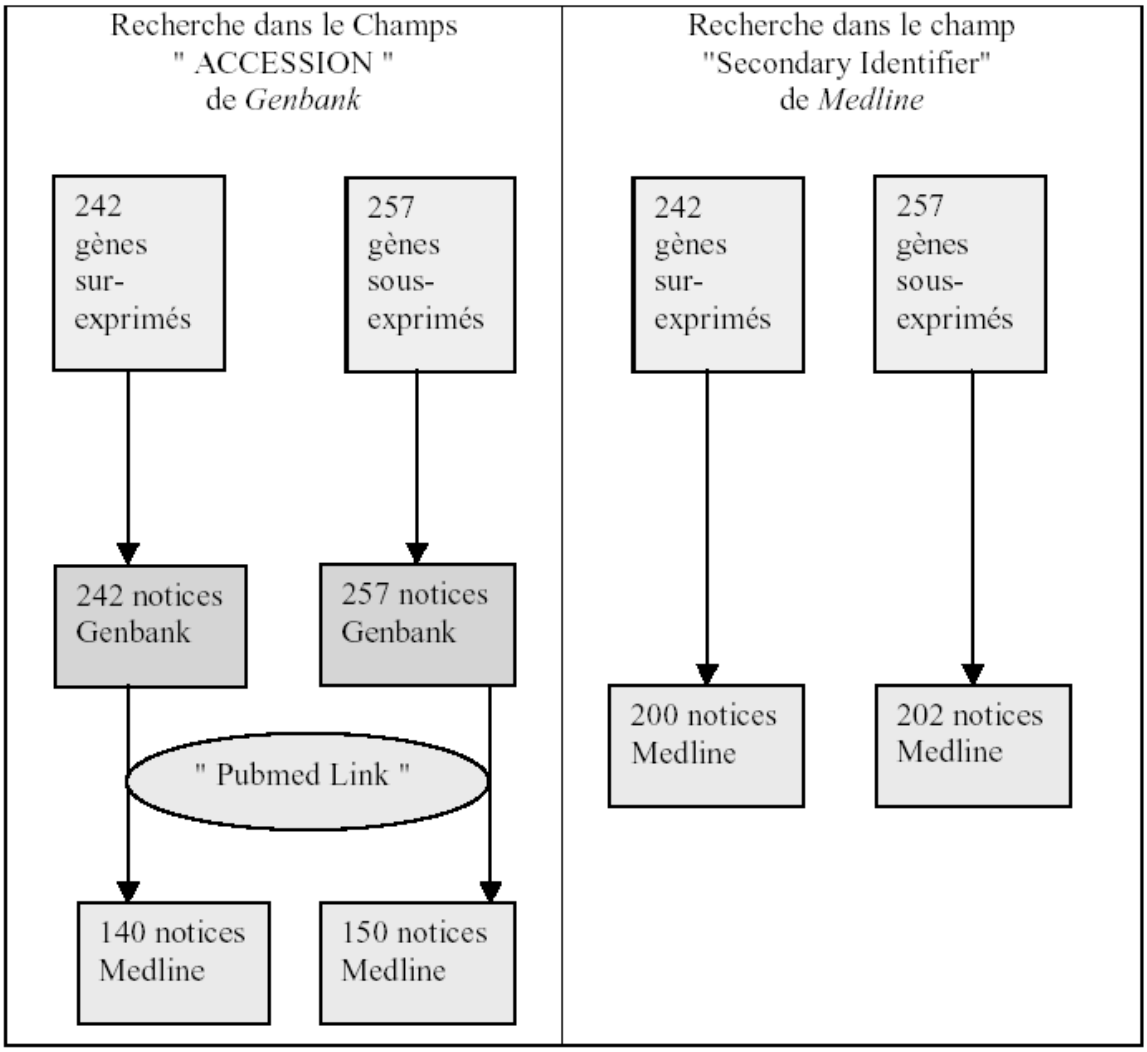


Figure 5 : Comparaison entre les deux méthodes (directe à gauche et indirecte à droite). L'interrogation indirecte via SI de Medline qui contient les numéros d'accession GenBank génère plus de notices. Certains gènes n'ayant pas encore donné lieu à des publications, il est normal d'obtenir 200 ou 202 notices au lieu de 242 et 257.

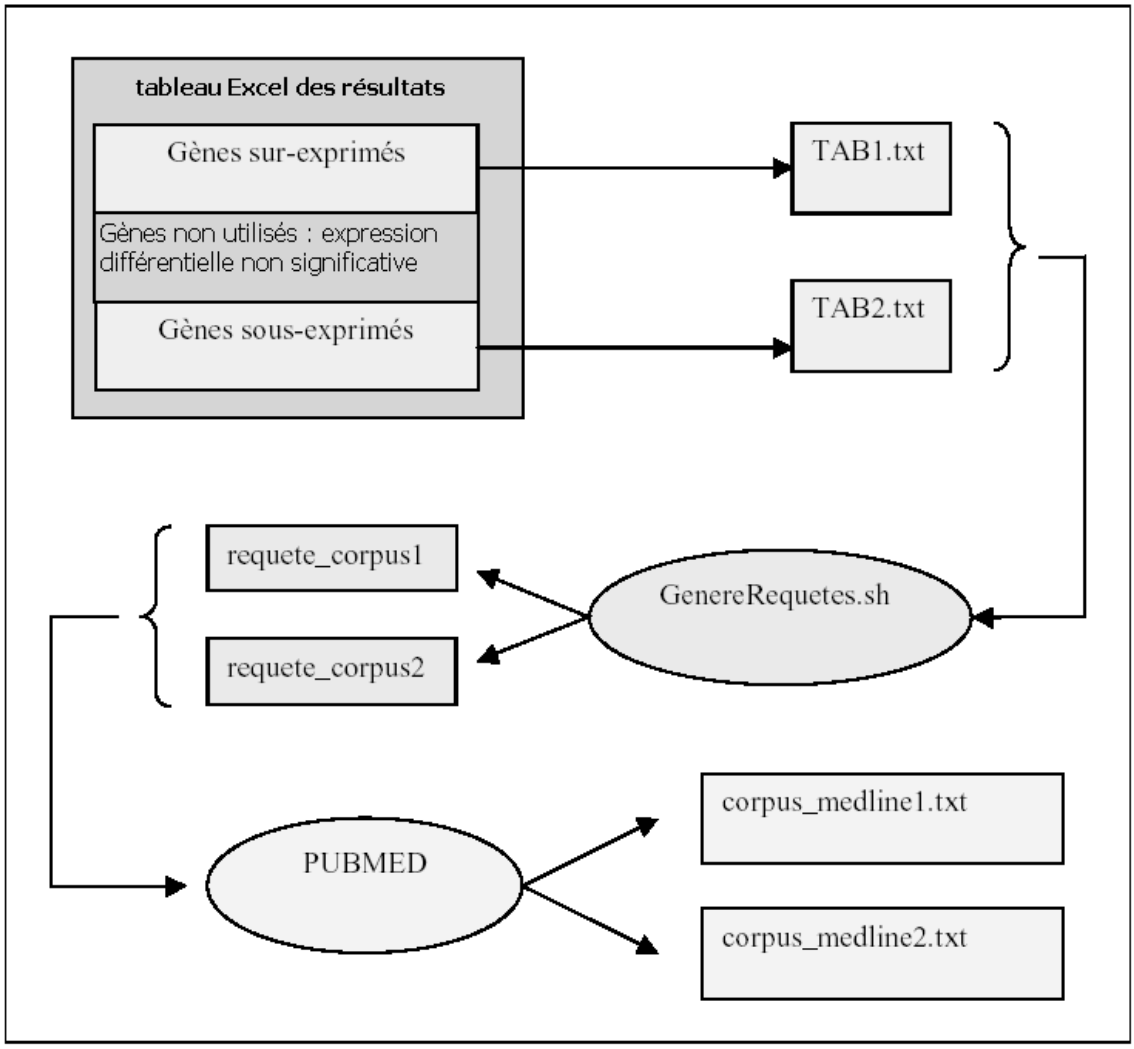


Figure 6: Méthode d'extraction des corpus.

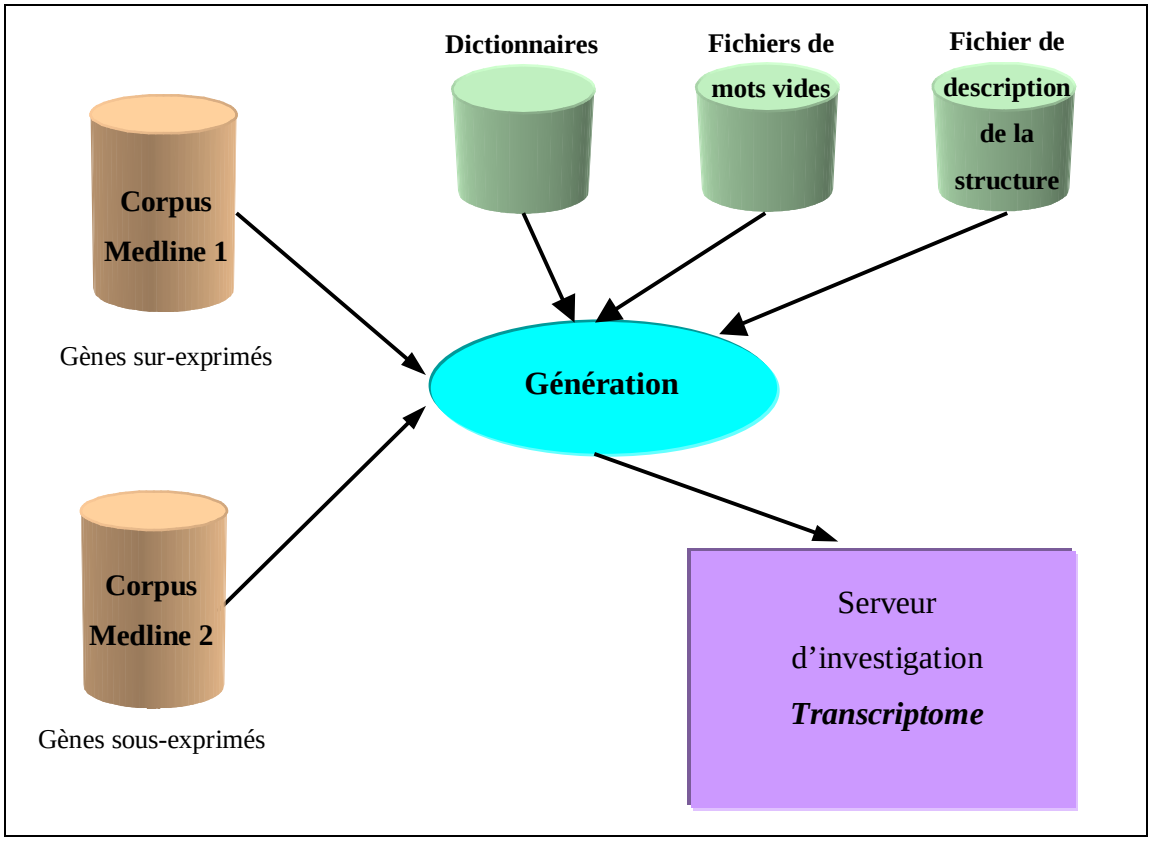


Figure 7: Avant toute génération d'un serveur d'investigation, il est nécessaire de paramétrer la génération :

- Les dictionnaires (anglais et français) contiennent des informations nécessaires à la présentation du serveur sur le Web (titres, sous-titres, intitulés des index…).
- Les fichiers de mots vides permettent de filtrer les termes non significatifs lors des indexations.
- Le fichier de description du serveur (***format ed****)permet de fixer la structure du serveur (par exemple les champs indexés). Les informations de ce fichier sont introduites dans un fichier de ***makefiles**** permettant de générer le serveur en fonction du paramétrage établi dans le fichier de description.

La réalisation comprend plusieurs étapes.

1. Reformatage des notices :

Dans un premier temps, les corpus Medline sont convertis au format XML. Les données sont ensuite stockées dans deux fichiers bibliographiques sous la structure *Hierarchic File organization for Documentation* (***HFD****) dans le cadre de l'usage standard de DILIB. Cette structure permet de stocker en standard jusqu'à un million de références bibliographiques.

2. Indexation des notices :

Ces notices sont par la suite indexées via la génération automatique de fichiers inverses. DILIB utilise de nombreux fichiers inverses afin de récupérer les informations nécessaires à la construction de graphes de navigation. Ces fichiers index d'interrogation permettent d'accéder de façon directe aux objets reliés entre eux par une propriété commune (par exemple tous les mots-clés).

3. Construction des associations :

Cette étape consiste à extraire les associations de termes issus d'un même champ dans les notices. Les fichiers inverses d'association permettent, lors de la navigation, d'accéder aux termes ou auteurs co-occurrents.

4. Construction des clusters :

Un cluster comprend un ensemble d'associations et le nom du cluster correspond au nom de l'association qui a le plus de poids (figure 8). Les clusters représentent des thèmes et sont composés d'un ensemble de termes.

Pour plus de détails dans la construction des clusters, voir le rapport de Claude Némurat [4].

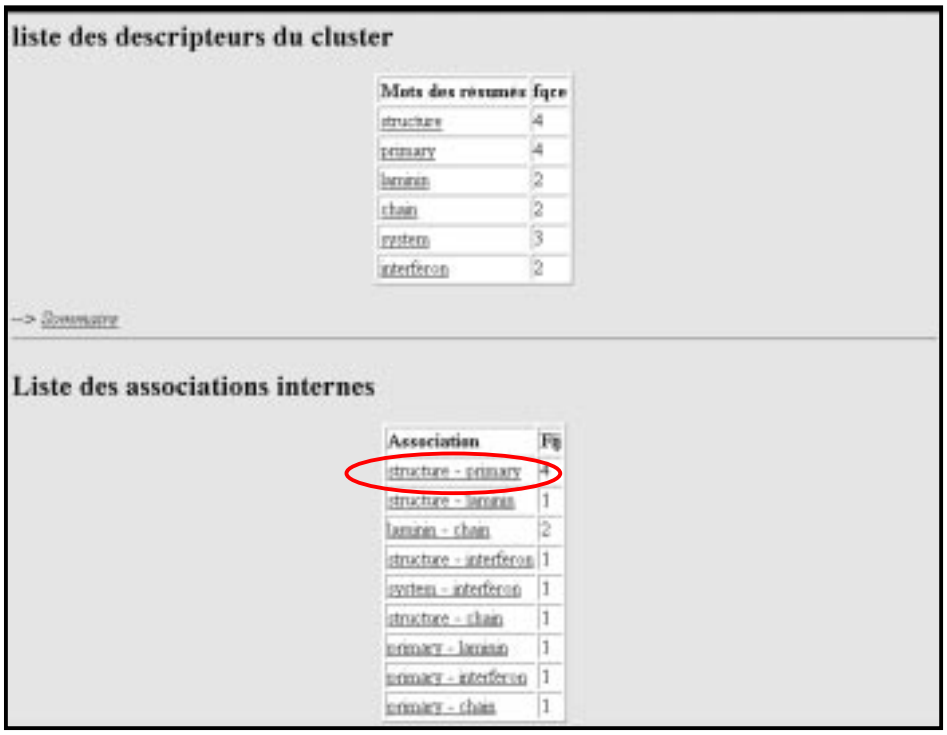


Figure 8 : Exemple du cluster « Structure-Primary »

5. Réalisation d'une navigation sur Internet :

Cette dernière étape permet la visualisation des résultats sur le ***Web****. Tous les fichiers nécessaires à l'exécution des *Common Gateway Interface* (***CGI****) sont créés lors de la génération du serveur d'investigation. Les CGI sont des scripts ou des programmes qui sont exécutés par un serveur ***http**** lorsque le client en donne l'ordre (par un lien par exemple).

Ceux-ci permettent la création dynamique de pages ***HTML****, le stockage ou la prise en compte d'informations (remplissage de formulaires par exemple). DILIB peut alors générer dynamiquement les différentes pages HTML du serveur afin de permettre sa consultation sur le Web.

➢ Les objectifs

Les nouvelles fonctionnalités du serveur Transcriptome concernent deux axes principaux :

- Intégration d'un sous-ensemble du méta-thésaurus UMLS pour améliorer le caractère informatif de la navigation.
- Incorporation de bases d'arrière-plan Pascal avec quatre années d'arriérés pour suivre l'évolution temporelle des concepts émergents.

Cependant, il était également souhaité d'améliorer la gestion des mots-clés afin de solutionner les problèmes survenant lors de l'interrogation de certains mots-clés et préparer l'intégration du méta-thésaurus. Enfin, pour optimiser le confort de la génération du serveur, il a fallu automatiser une étape de personnalisation du serveur, restée manuelle jusqu'à présent.

Intégration d'un sous-ensemble du méta-thésaurus UMLS conçu par le National Library of Medicine.

Bertrand Rihn souhaitait, à partir des descripteurs MeSH des corpus Medline, pouvoir "rebondir" sur des concepts voisins ou nouveaux grâce à une arborescence issue du MeSH. Il était intéressant de savoir, par exemple, ce que le terme "Integrin" pouvait ramener comme termes lui étant génériques ou spécifiques.

Le méta-thésaurus UMLS constituait donc un outil idéal dans la mesure où les termes de MeSH y sont inclus. Nous avons donc choisi de concevoir un fichier thésaurus, qui sera baptisé ThesGenome, en format XML pour qu'il soit intégrer sous Dilib dans le serveur Transcriptome. Ce thésaurus contiendra les *termes MeSH* des corpus Medline de départ (Gene1 pour les sur-exprimés et Gene2 pour les sous-exprimés), leurs *génériques* et les *relations*.

Incorporation de bases d'arrière-plan Pascal thématique.

Le premier avantage d'ajouter des bases d'arrière-plan est l'enrichissement de la navigation dans le serveur. Cela permet également de suivre un concept à travers des domaines différents mais voisins ou complémentaires. De plus, les bases d'arrière-plan de troisième niveau, constituées d'années complètes, permettent de suivre l'évolution temporelle d'un concept.

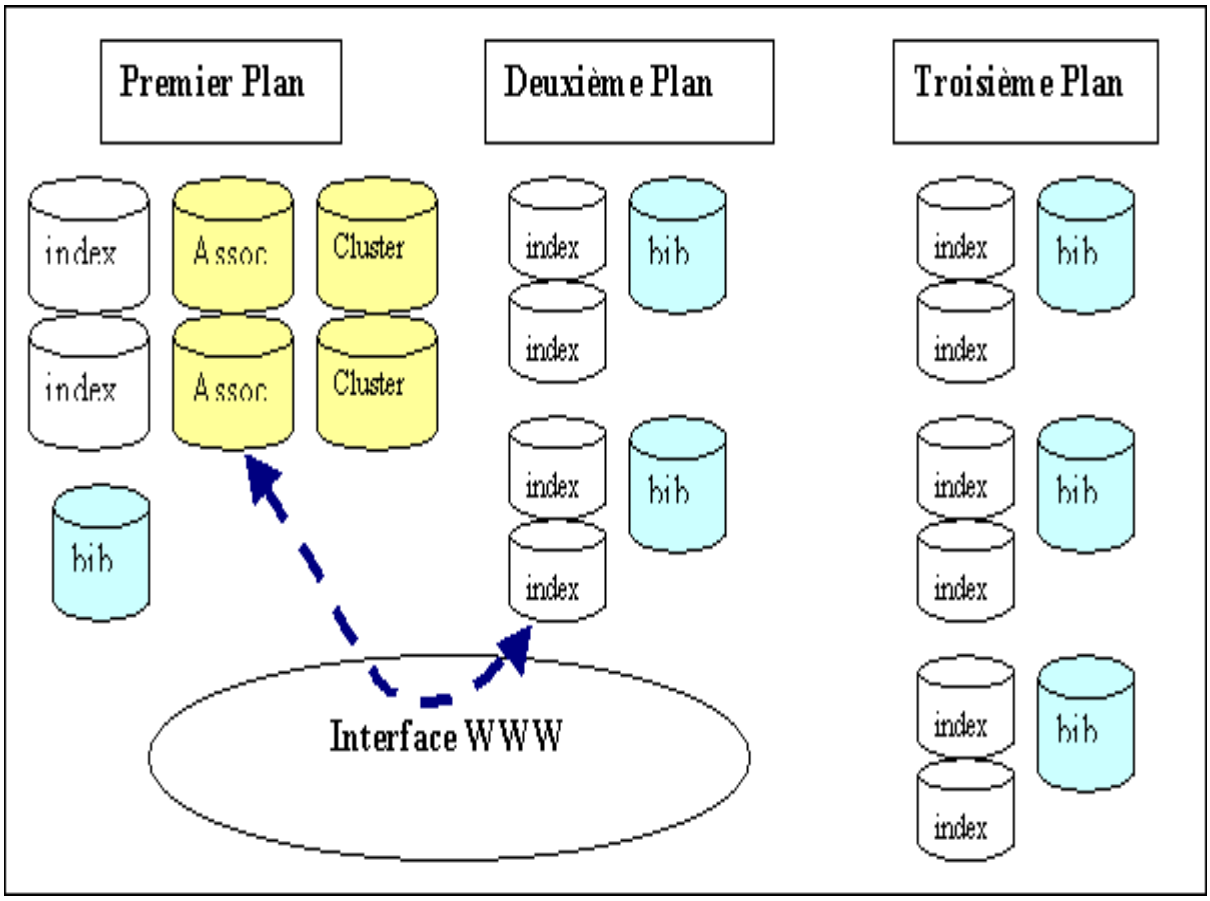


Figure 9 : L'organisation des bases d'arrière-plan dans un serveur d'investigation.

En accord avec Bertrand Rihn, il a été décidé de faire profiter des avantages que procurent des bases d'arrière-plan au serveur Transcriptome. Il a semblé intéressant d'intégrer des bases Pascal qui pourraient ainsi compléter la couverture documentaire procurée par Medline.

➢ Modifications apportées au serveur

Opérations de pré- et postprocessing.

1. Préparation des mots-clés MeSH (préprocessing).

Les termes du MeSH sont contenus dans le champ MH des notices *Medline*. Ce champ MH est utilisé pour constituer l'index "descripteur" (DE) du serveur. Le champ MH des notices *Medline* contenant les termes du MeSH est particulier (Figure 10). Chaque occurrence de ce champ est en fait une suite de termes incluant des termes génériques (*Subheadings*). Si cette

présentation particulière est utile pour faire des recherches sur Medline, elle n'est pas très appropriée à l'établissement d'un serveur d'investigation réalisant des associations et des *clusters* sur les termes spécifiques. De plus pour pouvoir naviguer dans le thésaurus, il nous faut des mots-clés MeSH isolés (sans qualifieur).

```
<DE>
<e>Aorta/drug effects/metabolism/pathology</e>
<e>Blotting, Northern</e>
<e>Cell Division/drug effects</e>
<e>Cell Movement/*drug effects</e>
<e>Cells, Cultured</e>
<e>Comparative Study</e>
<e>Cytokines/metabolism/*pharmacology</e>
<e>*Gene Expression/drug effects</e>
</DE>
```

Figure 10 : Structure du champ des mots-clés du MESH.

L'astérisque désigne le qualifieur ou le concept principal. Par exemple, le terme « Cytokines » est à prendre en compte principalement dans le contexte de la pharmacologie et non du métabolisme.

Afin de pouvoir exploiter isolément tous les mots-clés d'une notice, les contenus du champ MH/DE ont été découpés via différents traitements (langage C et Dilib).

Le résultat est le suivant : les termes MeSH d'une même notice (une fois découpés et isolés) sont tous rangés dans un seul nœud XML « <DE> » et individualisés dans des nœuds-fils <e> comme indiquée ci-dessous :

```
<DE>
    <e> Chromosome Banding(c)</e>
    <e>Chromosomes, Human, Pair 10</e>
    <e>Genetics</e>
    <e>Human</e>
    <e>In Situ Hybridization, Fluorescence</e>
</DE>
```

Figure 11 : Nouvelle présentation des mots-clés MeSH, après scission en uni-termes, dans une même notice.

C’est à partir de ces notices, que la génération du serveur avec des index d’interrogation « propres » est effectuée. De plus l’extraction du vocabulaire pour le thésaurus est maintenant possible.

2. Automatisation de l’adaptation des pages HTML après la génération du serveur.

Les pages d'accueil de ce serveur diffèrent de celles d'un serveur d'investigation standard.

En effet, il a été rajouté, en plus du traditionnel tableau présentant les index accessibles pour la navigation, une table comprenant une liste des auteurs, une liste des affiliations et une liste des titres d'articles selon la demande de Bertrand Rihn (voir annexe 1).

L'ajout de ce tableau lors de chaque génération du serveur, modifiant en partie l’interface de navigation standard, avait jusqu'à présent toujours été effectué à la main sur les pages HTML (cette opération se répétant quatre fois dans la mesure où le serveur comprend deux corpus et que, pour chaque corpus, il fallait deux pages d'accueil pour respecter le bilinguisme anglais-français).

Il était donc souhaité que ce rajout se fasse de manière automatique.

Cette automatisation fut réalisée grâce à un programme shell, qui doit être lancé après chaque génération du serveur (post-processing).

La réalisation et l’intégration de l’extrait du méta-thésaurus UMLS

- Etude préalable de l'UMLS.

Il en ressort que deux fichiers clés seront utilisés : le MRCON qui associe un terme à une clé CUI (Concept Unique Identifier) et le MRREL qui décrit le type de relation entre deux CUI.

- Conversion XML des fichiers clés.

Ces deux fichiers ayant en commun le CUI, il a été décidé de manipuler ces fichiers par cette clé CUI commune pour construire un thésaurus à partir des descripteurs MeSH des deux corpus Medline du serveur « Transcriptome ». Afin de rendre ces fichiers plus manipulables,

ils ont été reformatés en format XML ce qui facilite la fabrication du thésaurus grâce aux commandes Dilib.

- **Le fichier MRCON** :

Cette étape a nécessité trois programmes Lex du fait de la complexité de l'organisation hiérarchique entre les CUI, LUI et SUI (voir Figures 12-17). Entre chaque niveau de clé, il a fallu faire un dédoublonnement (avec la commande ***DilibUniq***) pour éviter les redondances.

De plus, nous avons dû au préalable filtrer le MRCON pour ne garder que les termes anglais, d'une part et éliminer les synonymes et autres variants lexicographiques d'autre part (en effet, nous manipulerons les descripteurs MeSH qui sont très majoritairement des termes préférentiels).

Dans le fichier MRCON est indiqué, pour chaque niveau de clé (LUI et CUI), les propriétés du terme.

Pour un CUI est indiqué la langue et si c'est un terme préférentiel (P) ou un synonyme (S).

Plus en profondeur, pour la clé LUI sont indiqués tous les variants lexicographiques (VO pour Variant orthographique, VC pour distinction dans la casse…par rapport à un terme préférentiel qui est indiqué PF). A ce niveau, pour chaque variant correspond un SUI et une chaîne de caractère.

Ainsi, le premier programme Lex ***UmlsMrcon.1*** sépare les clés des contenus.

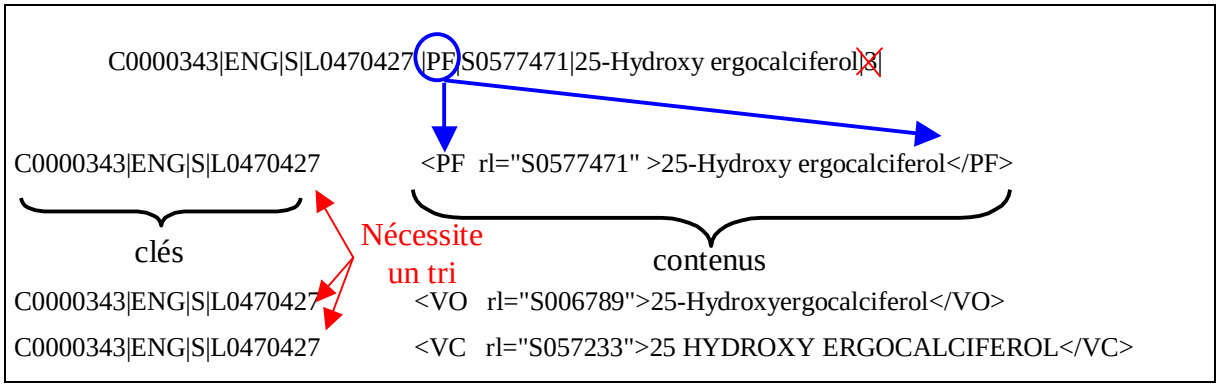


Figure 12 : résultat du programme ***UmlsMrcon.1***

En haut, une entrée du MRCON avant traitement de UmlsMrcon.1. « PF » devient une balise englobant le contenu, qui est d'ailleurs séparé par une tabulation de la clé (à droite).

En bas, sous les flèches bleues, trois mêmes clés au contenues différents (variations lexicographiques).

Le tri permet de ranger sous une racine unique toutes les balises qui correspondent au même LUI, de ce fait on peut trier au niveau supérieur, c'est-à-dire au niveau CUI.

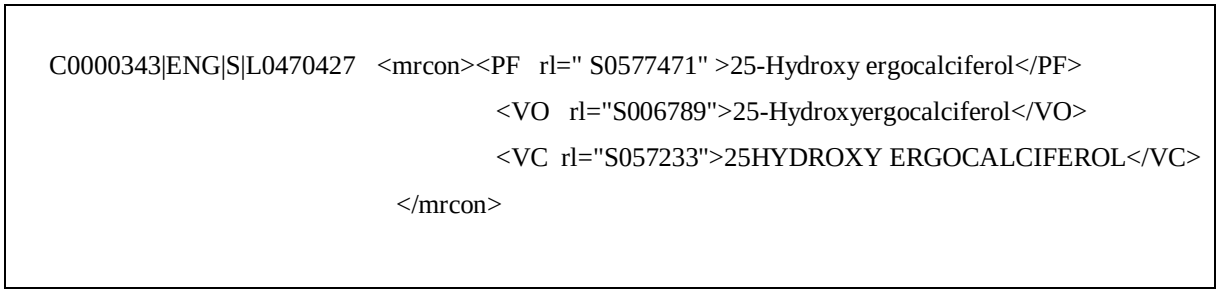


Figure 13: Résultat de la première commande de tri *DilibUniq –r mrcon*.

Le second programme Lex ***UmlsMrcon.2*** prépare le tri entre les LUI préférentiels (P) ou synonymes (S). Ainsi, il y a une séparation au niveau des indications S ou P.

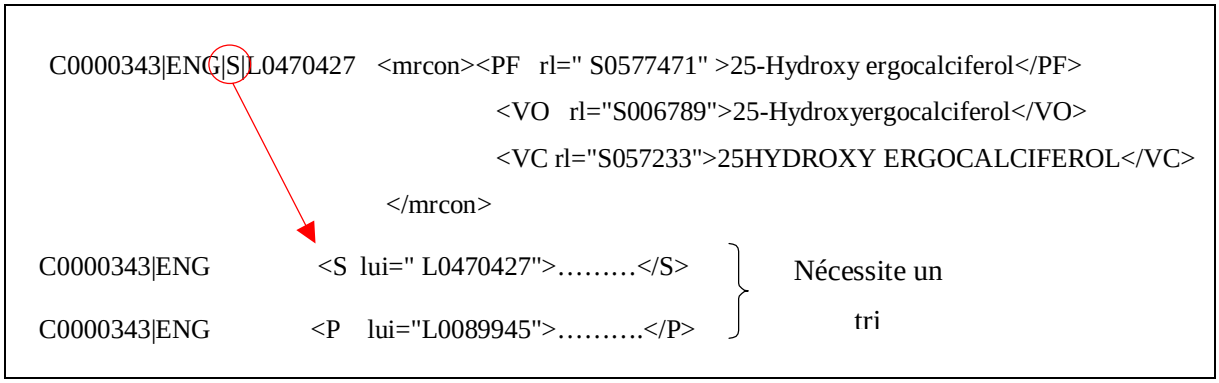


Figure 14 : résultat du programme ***UmlsMrcon.2*** .

En haut, l'entrée MRCON avant UmlsMrcon.2 . les différentes variations lexicographiques d'une même clé LUI sont rassemblés sous la balise <mrcon>. « S » ou « P », après séparation par une tabulation de la clé CUI, deviennent des balises <S> ou <P> pour désigner les termes synonymes ou préférentiels.

```
C0000343|ENG    <mrcon><S  lui="L0470427"><PF   rl=" S0577471" >25-Hydroxy ergocalciferol</PF>
                                    <VO   rl="S006789">25-Hydroxyergocalciferol</VO>
                                    <VC rl="S057233">25 HYDROXY ERGOCALCIFEROL</VC></S>
                          <P   lui="L0089945"><PF……..</P>
                  </mrcon>
```

Figure 15: Résultat du second tri avec *DilibUniq –r mrcon.*

Sous une même balise <mrcon>, tous les variants lexicographiques des termes préférentiels ou synonymes d'un même concept sont rassemblés. Ceci pour une langue.

Le troisième programme Lex ***UmlsMrcon.3*** achève le reformatage en séparant la langue de la clé CUI, la langue (si différente de l'anglais) devient un attribut (exemple lang= « FRE ») d'une balise <lv> :

```
C0000343|ENG    <mrcon><S  lui="L0470427"><PF   rl=" S0577471" >25-Hydroxy ergocalciferol</PF>
                                    <VO   rl="S006789">25-Hydroxyergocalciferol</VO>
                                    <VC   rl="S057233">25 HYDROXY ERGOCALCIFEROL</VC></S>
                          <P   lui="L0089945"><PF……..</P>
                  </mrcon>

C0000343          <mrcon><S>……</S>……<P>……</P>…….<mrcon>
C0000343           <lv     lang="FRE">.....</lv>
```

Figure 16: Résultat du dernier programme Lex UmlsMrcon.3 .

En haut, avant traitement de UmlsMrcon.3 . La balise <lv> apparaît quand il y a une traduction du concept.

La dernière commande de tri ***DilibUniq*** permet de classer tous les variants linguistiques sous une même racine <mrcon>. On obtient donc le fichier final sous la forme suivante :

```
<mrcon>
 <CUI>C0000039</CUI>

 <LV LANG="FRE">
  <P LUI="L0176992">
   <PF id="S0241473" rl="">1,2-DIPALMITOYLPHOSPHATIDYLCHOLINE</PF>
  </P>
 </LV>

 <LV LANG="GER">
  <P LUI="L1226153">
   <PF id="S1468099" rl="">1,2-Dipalmitoylphosphatidylcholin</PF>
  </P>
  <S LUI="L1246976">
   <PF id="S1488922" rl="">Dipalmitoyllecithin</PF>
  </S>
 </LV>

  <P LUI="L0000039">
   <PF id="S0007564" rl="">1,2-Dipalmitoylphosphatidylcholine</PF>
   <VW id="S1357296" rl="">1,2 Dipalmitoylphosphatidylcholine</VW>
 </P>
 <S LUI="L0000035">
   <PF id="S0007560" rl="">1,2-Dihexadecyl-sn-Glycerophosphocholine</PF>
   <VW id="S1357276" rl="">1,2 Dihexadecyl sn Glycerophosphocholine</VW>
 </S>
</mrcon>
```

Pas de balise <LV> quand c'est de l'anglais

Figure 17: Fichier MRCON en format XML et trié.

Les trois programmes Lex et les commandes de tri sont réunis dans un même script shell "***CreateCuiTab.sh***" qui va permettre de créer une table d'équivalence ne contenant que les termes préférentiels et anglais de l'UMLS avec leurs clés CUI respectives. Cette table servira pour remplacer les CUI par leurs termes dans le thésaurus, baptisé ThesGenome.

- **Le fichier MRREL** :

Le reformatage de MRREL en format XML est effectué par un programme Lex baptisé ***UmlsMrrelToXml***. Les modifications sont les suivantes :

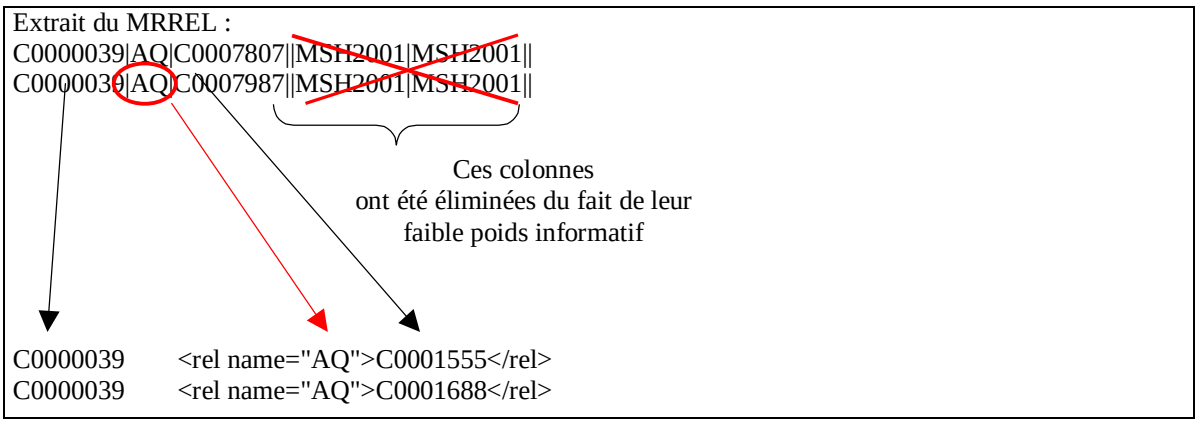


Figure 18 : Résultat du programme UmlsMrrelToXml.

- Création de tables d'équivalence.

Pour construire le thésaurus qui sera inclus dans le serveur (ThesGenome), nous avons dû réaliser plusieurs tables d'équivalence intermédiaires à partir des fichiers UMLS (MRCON et MRREL) mais également à partir des corpus du serveur « Transcriptome ».

- La table **Cui2TP.tab**.

Il était indispensable, une fois le thésaurus ThesGenome construit, de pouvoir intervertir les clés CUI par leurs équivalents mots-clés. La difficulté résidait dans le fait qu'un CUI pointe sur plusieurs chaînes de caractères puisqu'un concept inclut les synonymes, les traductions et tous les variants lexicographiques. Il a donc fallu filtrer le fichier MRCON pour ne garder que les termes anglais (nos descripteurs MeSH de départ étant en anglais).

Généralement, le vocabulaire d'entrée du MeSH correspond à des descripteurs contrôlés ou « preferred descriptors » pour chercher sur PubMed. Ainsi, en plus de la restriction du MRCON sur les termes anglais, nous avons limité ce fichier avec uniquement les termes P et PF du MRCON qui correspondent à ces descripteurs contrôlés du MeSH, en partant du principe que les mots-clés des deux corpus Gene1 et Gene2 sont majoritairement indexés par ces termes (voir figure 19).

Cette table d'équivalence a été réalisée par le programme suivant :

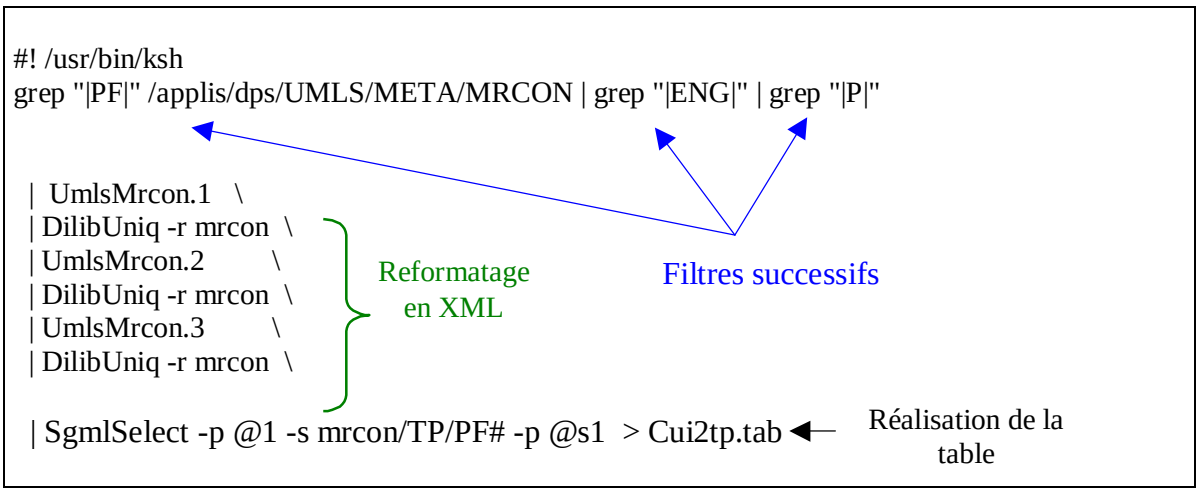


Figure 19 : programme CreateCuiTab.sh .

Le reformatage du fichier MRCON était nécessaire pour pouvoir sélectionner avec une commande simple les clés CUI et les termes correspondants.

Cette table contient 797.360 entrées (un petit extrait est présent dans l'annexe 2).

- Les deux tables d'équivalence à partir des corpus Medline : **tableequiv1.tab** et **tableequiv2.tab**.

Pour que les extractions des génériques MeSH provenant de l'UMLS correspondent au sujet du serveur Transcriptome, il a été préférable de partir directement des descripteurs inclus dans les deux corpus Gene1 et Gene2 qui sont du vocabulaire MeSH (la source MeSH faisant partie de l'UMLS).

Nous voulions les clés CUI correspondant aux descripteurs Medline des deux corpus, afin de rechercher leurs génériques dans le fichier MRREL. L'ensemble de ces clés va constituer le cœur du thésaurus thesGenome qui sera inclus dans le serveur Transcriptome.

Les index mots-clés du serveur correspondant aux descripteurs MeSH ont été utilisés, et les termes en ont été extraits puis triés. Il se trouve qu'aux termes communs aux deux corpus ont été rajoutés des (c) (c'était un vœux de Bertrand Rihn pour améliorer la lisibilité lors de la navigation). Or ces (c) entravent la reconnaissance des mots-clés dans une table d'équivalence CUI/terme telle que « Cui2tp.tab ». Ainsi, un programme Lex a été conçu pour *nettoyer* les deux listes de descripteurs MeSH provenant des corpus Gene1 et Gene2.

On peut noter une différence entre le nombre de mots-clés au départ (536 termes dans Gene1 par exemple) et le nombre de mots-clés dans la table (515 au lieu de 536).

On peut expliquer cette différence par le fait que certains mots-clés sont soit des synonymes MeSH au lieu des « Preferred Descriptors », soit des « SubHeading » ou des « qualifiers » du MeSH ne faisant pas partie de l'UMLS.

- La table des clés de termes génériques **subsetCUI.list**.

Pour obtenir le cœur de ThesGenome, il fallait non seulement les mots-clés des deux corpus Medline du serveur mais également tous leurs génériques du MeSH/UMLS. Ainsi à partir des clés CUI correspondantes aux descripteurs des deux corpus, et du MRREL en format XML, il fallait récupérer tous les génériques. Un programme récursif a donc été conçu pour chercher dans une première boucle les génériques directs des termes de Gene1 et Gene2, puis à partir

de ces génériques, rechercher dans le MRREL leurs génériques (avec dédoublonnage des termes récupérés). Ce programme, réalisé en langage C, s'appelle ThesFind (voir Annexe 3). Le résultat de ce programme se trouve dans une table subsetCUI.list avec 2760 entrées CUI. Ce programme ThesFind est appelé dans un script shell « CreateSubThes.sh » qui va créer à partir de l'ensemble des tables décrites plus haut et du MRREL en format XML, le thésaurus final qui sera inclus dans le serveur.

- Synthèse de ThesGenome.

Les différentes étapes de la construction de thesGenome sont indiquées dans le schéma de la figure 20.

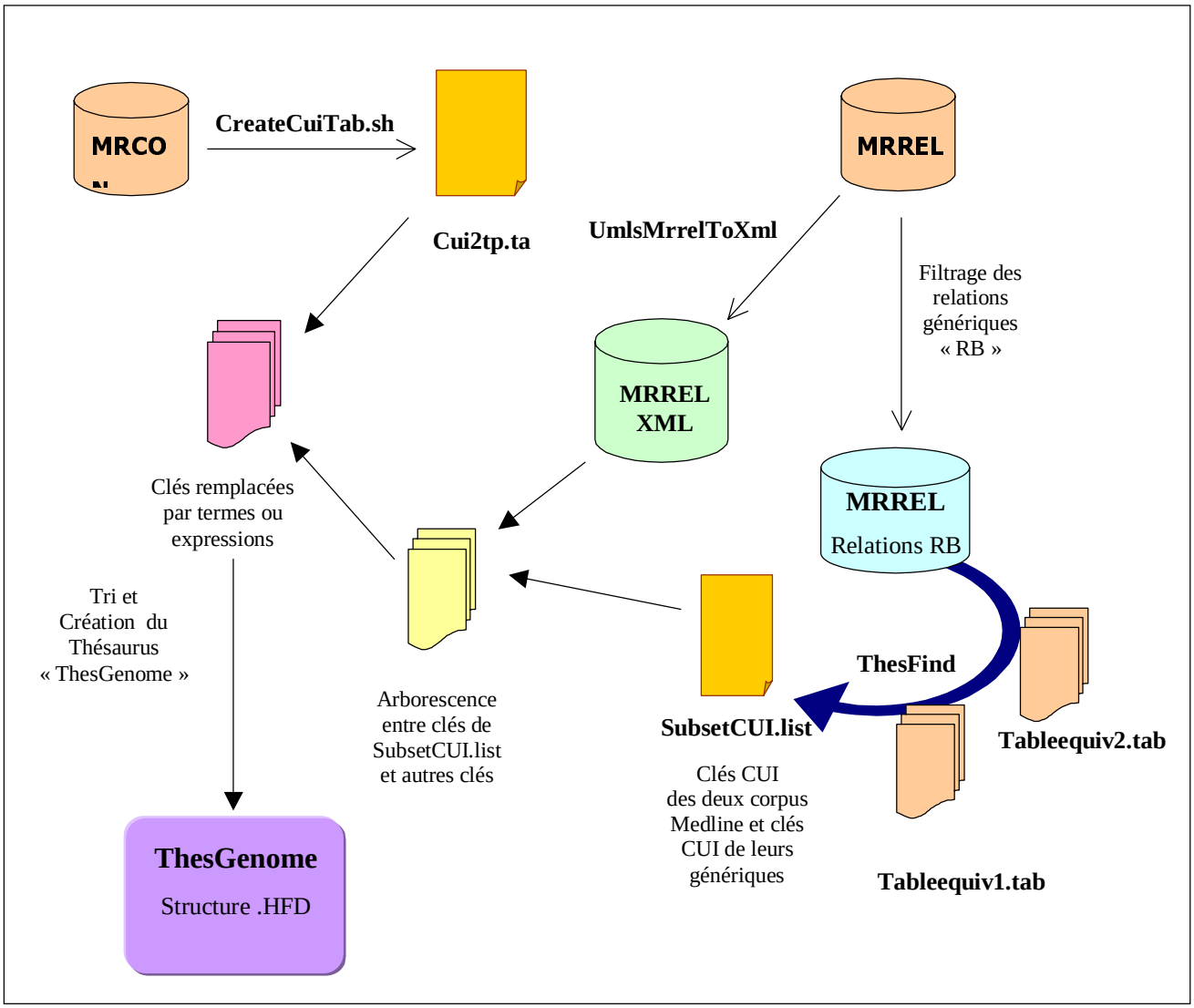


Figure 20: schéma synthétique de la réalisation du thésaurus « thesGenome » qui sera intégrer dans le serveur d'investigation Transcriptome.

Les deux fichiers clés de l'UMLS sont reformatés en XML, le MRCON par trois programmes Lex inclus dans le shell CreateCuiTab.sh, et le fichier MRREL par le programme UmlsMrrelToXml.

Ces deux fichiers étant reconvertis en format XML, il a été possible pour le MRCON de créer une table d'équivalence Cui2tp.tab (équivalence entre clé CUI et terme) et pour le MRREL de pouvoir extraire les relations correspondantes entre les clés CUI d'intérêt, c'est-à-dire les CUI correspondants aux mots-clés MeSH des deux corpus Medline, et tous leurs CUI génériques.

L'extraction de ces relations s'est faite grâce à la commande Dilib ***StrSearchKey*** sur la table SubsetCUI.list. Cette commande a été répétée pour récupérer les relations inverses afin de pouvoir naviguer dans le thésaurus en pouvant suivre les deux types de relations.

- Déclaration du thésaurus ThesGenome dans le paramétrage du serveur d'investigation.

```
<thes code=MH>
    <input env=DEMO_ROOT>Import/Data/ThesGenome
    <rel code=RN inv=RB>
    <rel code=RB inv=RN>
 </thes>
```

Figure 21: extrait de GenomeDesc.ed (fichier descripteur du serveur en format « ed »).

Dans la déclaration, chaque relation est présentée avec son inverse.

L'incorporation de bases d'arrière-plan Pascal

Pour intégrer trois ou quatre années de la base Pascal qui vont constituer les bases d'arrière-plan de troisième niveau, il est nécessaire d'utiliser en amont une base d'arrière-plan de second niveau qui servira d'intermédiaire pour établir une passerelle sémantique entre les bases Medline et les bases Pascal de troisième niveau.

Le serveur d'investigation est donc enrichi au final d'une base Pascal de second plan et de quatre bases Pascal de troisième plan qui concerneront les années 1999, 2000, 2001 et 2002.

Le problème d'équivalence d'indexation s'est posé dès l'extraction du corpus Pascal intermédiaire. Comment établir une passerelle sémantique si les habitudes d'indexation diffèrent selon la base de données ?

De plus, le niveau de spécification est plus précis dans Medline que dans Pascal. Par exemple pour le concept de régulation d'expression génique, dans Medline on a le descripteur « Gene Expression Regulation » alors que, dans Pascal, ce concept est décrit par deux mots-clés « Gene expression » et « Regulation ». Il est même possible qu'un concept décrit en une seule expression dans Medline corresponde à trois descripteurs Pascal.

Nous avons donc décidé d'avoir, dans la mesure du possible, un corpus Pascal clone des deux corpus Medline afin de pouvoir créer une passerelle sémantique entre Medline et Pascal.

- La réalisation du corpus clone Pascal.

Nous avons utilisé Miriad qui est un système d'interrogation des bases Pascal et Francis sur l'intranet de l'INIST. Concernant ce corpus clone, il a semblé judicieux de créer une requête sur Miriad avec les concepts les plus fréquents des deux corpus Medline.

On envisage un corpus clone d'au moins 200 x 2 notices environ (chaque corpus Medline faisant 200 notices), c'est-à-dire environ 500 notices. Ainsi, on obtiendrait un corpus avec une volumétrie et une thématique équivalente avec un maximum de notices identiques à celles présentes dans le serveur d'origine. Les termes Pascal équivalents aux termes Medline les plus fréquents génèrent cependant beaucoup de bruit. La requête a donc été affinée les auteurs des deux corpus Medline (voir annexe 4).

Il est à noter que la plupart des ISSN, donc des périodiques, présents dans les bases Medline du serveur sont absents de Pascal. Ce dernier point fait qu'il a été impossible d'avoir un corpus réellement clone. Nous avons décidé de récupérer un corpus de 1000 notices, avec la vingtaine de notices communes.

Cependant, les auteurs ayant écrit dans d'autres périodiques au cours de leur carrière tout en poursuivant leurs recherches sur le même sujet, nous avons pu observer la présence d'articles très voisins et nous avons considéré que ce corpus était satisfaisant dans la mesure où les

principaux acteurs et concepts du serveur de départ étaient présents avec des proportions voisines à celles des deux corpus Medline. Par exemple, dans le corpus Gene1 (comprenant 200 notices), le concept « Transcription Factor » qui a une fréquence de 8 se retrouve à une fréquence de 41 pour un corpus de 1000 notices.

Nous nous sommes heurtés à certaines difficultés pour constituer ce corpus « pseudo » clone.

- La grande proportion de termes très génériques dans Pascal.

Afin de pouvoir mieux se rendre compte de la validité du corpus Pascal qui servirait de base intermédiaire, il a semblé intéressant de générer de manière provisoire un serveur monobase avec ce corpus. Ceci a permis d'améliorer le confort de l'analyse du corpus obtenu après le lancement de la requête sur Miriad. Ce serveur étant généré, une grande prépondérance de termes très génériques comme « Human » , « Rat », « Experimentation » etc… a été observée. Ce fait a rendu l'analyse de ce corpus plus difficile que supposé au départ. Il était impossible, au premier abord, de se rendre compte de la présence ou non dans cette base Pascal des concepts qui étaient les plus importants dans le serveur Medline de départ ni de savoir, s'ils étaient présents, dans quelle proportion.

Nous avons donc créé une table de mots vides (ou table de rejet, ou anti-dictionnaire) comprenant tous les mots trop génériques (voir annexe 5). Cette table a été inspirée de celle créée l'année dernière pour générer le serveur Transcriptome.

- Le problème de volumétrie.

L'année 2001 de Pascal comprend 450.000 notices et correspond à la volumétrie moyenne d'une année Pascal. Sachant qu'il nous faut quatre années en plus des deux corpus Medline de départ, du thésaurus, et du corpus intermédiaire, il a fallu se résoudre à limiter la volumétrie générale pour des raisons de stockage physique des données

Ainsi, les bases de troisième plan ont été réduites entre 1000 et 1500 notices. Ce qui supposait la constitution de corpus thématiques plutôt que de prendre des années entières.

Il a donc fallu également réduire le corpus de second plan servant de « pseudo » clone aux corpus Medline. De plus, il était nécessaire de réduire le bruit pour avoir une volumétrie proche des deux corpus de départ. Nous avons donc choisi de réduire ce corpus à environ 500

notices pour aussi avoir une volumétrie équivalente aux deux corpus de départ sans perte d'information significative. Ainsi, la commande IndexQuery a permis de réduire le corpus Pascal de 1000 notices à 463 en utilisant l'opérateur booléen « OU » pondéré sur les descripteurs, via l'option –f. Cela a l'avantage d'une part de permettre l'absence de certains mots-clés dans les notices sélectionnées comme avec un « OU » booléen ordinaire (ce que ne permet pas le « ET » booléen) et, d'autre part, de calculer un score avec les termes de sélection cooccurents (exemple une notice avec quatre mots-clés a un score de 4). A partir de là, avec une commande de sélection Dilib, il est possible de filtrer les notices aux scores les plus forts et d'abandonner celles avec les scores bas. Cependant, cette dernière opération n'a pas été réalisée, dans la mesure où la sélection des 10 mots-clés les plus fréquents (voir figure 22) sur le corpus de 1000 notices a entraîné sa réduction à 463, avec les notices d'un score de 1 compris. Afin d'avoir, cette fois, une volumétrie suffisante, nous avons préféré garder les notices d'un score de 1. Cela revient donc à filtrer le corpus de 1000 notices en utilisant un « OU » booléen ordinaire.

```
1.IndexQuery -h corpus.ED.i \
      -f -k "C-Onc gene" -k "Signal transduction" -k "Binding protein" \
      -k "Messenger RNA" -k "Gene organization" -k "Membrane receptor" \
      -k "Transcription factor" -k "Liver" -k "Transcription factor" \
      -k "Multigene family" -k "Tumor" -k "Tissue specificity" \
      -k "DNA binding protein" -k "Heat shock protein" \
      | SgmlSelect -s idx/l/e# -p @s1 | DamHfdSelect -h corpus.bib.hfd \
    > CorpusRestreint.txt
```

La commande IndexQuery va sélectionner toutes les notices qui comportent ces mots-clés.

Figure 22 : Réduction du corpus « pseudo » clone de 1000 à 463 notices.

- L'équivalence entre les descripteurs MeSH et les descripteurs Pascal.

Comme décrit précédemment, pour un même concept, le niveau d'indexation, et à fortiori le terme choisi, diffère complètement quand on passe d'une notice Medline à une notice Pascal, même quand il s'agit du même article.

Adrenal Cortex Neoplasm → Adrenal cortex diseases + Tumor

Antigens, CD → CD antigen

Chromosomes, Human, Pair 1 → A1-Chromosome

Maple Syrup Urine Disease → Leucinosis

Microphthalmos → Microphthalmia

Figure 23: Quelques exemples de disparités d'indexation. A gauche, les termes MeSH, à droite leurs équivalents Pascal.

Nous avons dû chercher manuellement les équivalents Pascal des descripteurs Medline les plus fréquents. L'utilisation de l'outil Folfic sur le vocabulaire INIST a été précieuse, bien que l'on se soit heurté à un défaut de mise à jour. Nous pouvions interroger par un mot MeSH, et rechercher le ou les équivalents Pascal proposés. Cependant, dès qu'un concept générait plus d'un terme ou expression Pascal, Folfic ne donnait aucun résultat. La consultation à la main d'index édités en 1990 a permis de compléter ces lacunes.

Nous avons ainsi établi un tableau d'équivalence entre les termes MeSH présents dans les corpus Medline de départ et les termes Pascal.

Il s'est alors rapidement posé le problème quant à la gestion des descripteurs Medline qui généraient des équivalents multi-descripteurs Pascal pour la réindexation automatique des notices Medline .

Deux tables d'équivalence ont été créées : Med2Pas et Pas2Med. Ces tables ont été éditées dans une structure XML nécessaire à leur utilisation ultérieure via Dilib.

```
Abnormalities, Multiple <list><term>Malformation</term><term>Multiple</term></list>
Acetyl-CoA Carboxylase <term>Acetyl-CoA carboxylase</term>
Actins <term>Actin</term>
Ca(2+)-Calmodulin Dependent Protein Kinase <list><term>Calmodulin</term><term>Calcium
</term><term>Protein kinase</term></list>
```

Figure 24: Extrait de Med2Pas.

Malformation	<if><input>Multiple</input><then>Abnormalities, Multiple</then></if>
Acetyl-CoA carboxylase	<term>Acetyl-CoA Carboxylase</term>
Actin	<term>Actins</term>
Calmodulin	<if><and><input>Calcium</input><input>Protein kinase</input></and><then>Ca
(2+)-Calmodulin Dependent Protein Kinase</then></if>	

Figure 25: Extrait de Pas2Med

Grâce à ces deux tables, nous avons pu réindexer les notices Medline et Pascal via un script shell « Reindex.sh» contenant la nouvelle commande Dilib *IndexGateWay* (qui est en fait un programme en langage C) développée spécifiquement pour ce besoin.

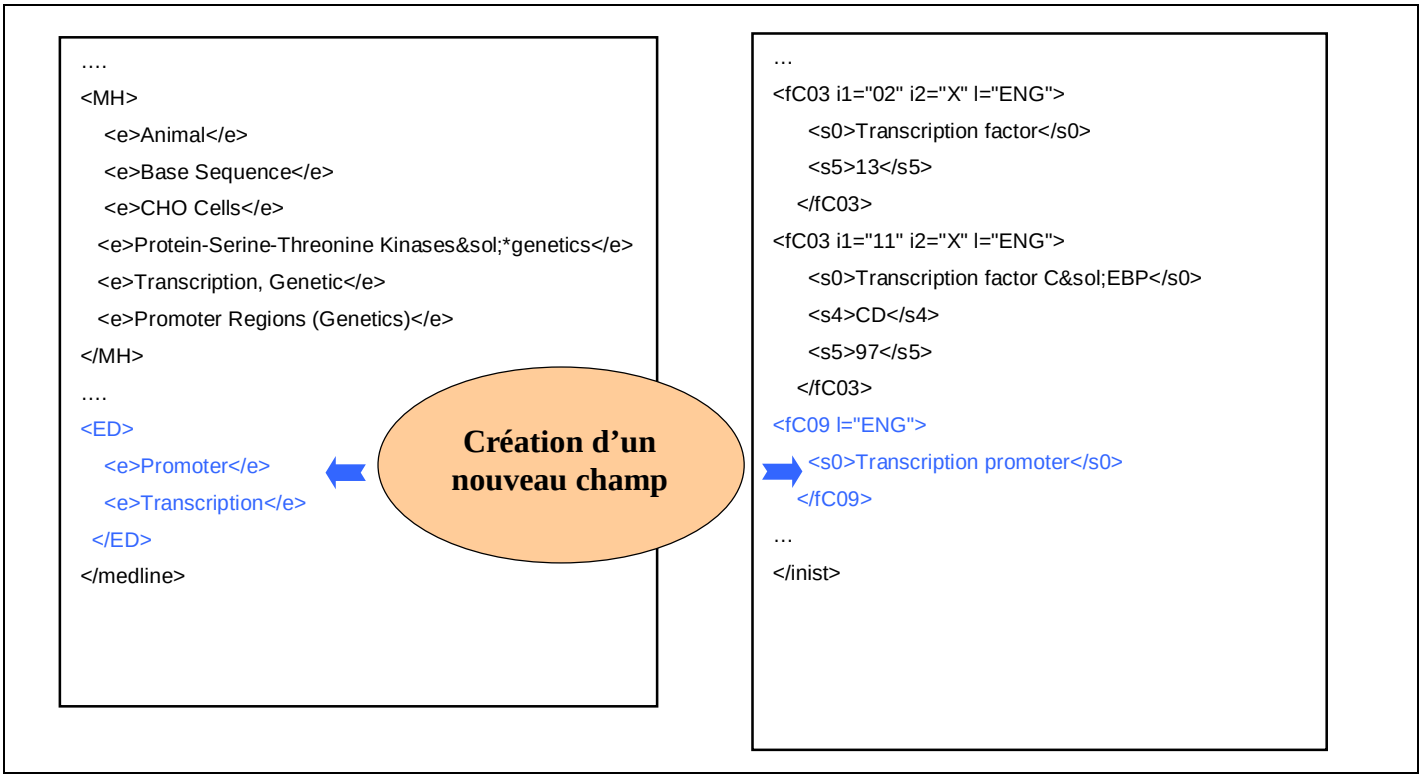


Figure 26 : Résultat du programme ReIndex.sh sur une notice Medline (à gauche) et sur une notice Pascal (à droite).

Les descripteurs générés par la réindexation des corpus (Medline et Pascal) ont été stockés dans un nouveau champ à l'intérieur des notices concernées, pour conserver la possibilité d'utiliser l'indexation initiale (par exemple, pour la navigation dans le serveur), celle-ci devait être conservée intacte.

- La réalisation des bases de troisième plan.

La requête sur Miriad a été la suivante pour chaque année de Pascal :

```
(((((tumor* ou cancer* ou tumeur*) et
(cc=002a07c* ou cc=002a04c* ou cc=002b23* ou cc=002b11* ou cc=002b02r* ou
cc=002b03* ou cc=002a31c* ou cc=002b02a* ou cc=002b30* ou cc=002a08* ou
cc=002b11*)) sauf (monocotyledones ou dicotyledones))) et "appareil
respiratoir*")
```

Figure 27: Requête thématique sur Miriad.

« cc » correspond aux codes de classement internes de la base Pascal.

La requête comprend des codes de classement correspondant à la génétique des eucaryotes, la génétique moléculaire et cellulaire, la biotechnologie et tout se concerne les sciences médicales (mais réduit aux sous-thématiques suivantes : pharmacologie, toxicologie, tumeurs, pneumologie, santé publique). L'ensemble générant beaucoup de notices concernant la génomique végétale (hors de propos dans notre thématique), nous avons utilisé l'opérateur « sauf » pour les exclure.

- La déclaration des bases d'arrière-plan pour l'intégration dans le serveur d'investigation.

```
<background from=pascal>
   <backBase code="pas99">
     <name>Pascal 1999
     <root>/applis/dps/INRS/GenomeAP/Back/Pas1999
    </backBase>
…
```

Figure 28: Extrait de genome.desc.ed (format « ed »).

Le fichier de paramétrage de la structure du serveur genome.desc.ed (voir annexe 6) s'est trouvé enrichi par plusieurs déclarations : d'abord par la déclaration du thésaurus, ensuite par celles d'une part du corpus Pascal pseudo-clone et d'autre part des quatre corpus Pascal de troisième plan.

Conclusion

Les nouvelles fonctionnalités ajoutées au serveur Transcriptome ont nécessité de longues phases d'étude de faisabilité tant sur l'incorporation du thésaurus ThesGenome que sur celle des bases d'arrière-plan.

En effet, une étude préalable de l'UMLS était indispensable afin de savoir comment extraire une arborescence d'une part cohérente avec l'existant et, d'autre part, apportant lors de la navigation des informations significatives et complémentaires (éventuellement sur les concepts émergents). La conception de ce thésaurus final ThesGenome a nécessité l'utilisation de nombreuses tables d'équivalence pour pouvoir se servir de programmes et commandes Dilib, ce qui a permis de simplifier la chaîne de réalisation.

L'ajout des bases documentaires d'arrière-plan supposait une grande maîtrise du sujet (génomique cancéreuse et plus particulièrement du mésothéliome), mais aussi de l'outil Miriad afin d'extraire des corpus Pascal cohérents avec les deux corpus Medline. De plus, il a fallu répondre à deux contraintes techniques fortes qui étaient de réduire la volumétrie générale mais, surtout, de mettre en place une passerelle sémantique entre Medline et Pascal pour naviguer de manière croisée au sein d'un même serveur.

Une évolution envisagée serait de lier des clusters de gènes aux bases de données autres que Pubmed comme LocusLink, Oncolink, Genecard ou toute autre base de données génomique existant sur Internet afin d'apporter une fonctionnalité à visée ***bioinformatique**** au serveur.

Ce stage m'a permis d'approfondir mes connaissances techniques, mais également de travailler en équipe dans le cadre d'un projet. De plus, j'ai eu l'occasion de représenter l'INIST lors du congrès TKE2002 (Terminologie et ingénierie des connaissances), en présentant une conférence en anglais intitulé « Multi-database server and semantic link ».

Glossaire

ADN : Acide Désoxyribonucléique. Enchaînement de nucléotides formant deux brins complémentaires antiparallèles dépositaires de l'information génétique.

ADNc : Acide Désoxyribonucléique Complémentaire. ADN simple brin obtenu par rétro transcription à partir d'ARNm (ARN messager).

AIR93 : AI/RHEUM. Bethesda (MD): National Library of Medicine, Lister Hill Center, 1993. Banque de données sur maladies auto-immunes et arthrites rhumatoïdes.

AIRS : AIRS est un système de recherche documentaire qui permet de classer, d'indexer et de retrouver des documents en fonction de leur contenu. AIRS Web permet un accès aux données sur Internet et Intranet.

ARN messagers : Acide Ribonucléique Messager. Copie complémentaire de l'ADN qui spécifie la séquence d'acide aminés d'une protéine.

ARTICLE@INIST : Catalogue en ligne du fonds INIST contenant 7 millions de références d'articles et monographies.

ARTICLESCIENCES : Moteur de recherche et de commande en ligne de copies d'articles scientifiques et techniques de l'INIST, disponibles en 4 langues avec possibilité de paiement en ligne.

ASCII : American Standard Code for Information Interchange.

Base d'arrière-plan : elle peut être de deux niveaux (de second plan ou de troisième plan). Une base de second plan permet de ré-interroger à partir d'un concept d'un corpus de départ (base de premier plan) : il y a complémentarité au niveau de la couverture quand les deux bases n'ont pas la même origine (Medline←→Pascal) mais elle permet également de créer et d'utiliser des passerelles sémantiques (pour un même concept, le descripteur Medline est différent du descripteur Pascal). Une base de troisième plan correspond en général à une année complète. On rajoute en général plusieurs bases de ce type pour suivre l'évolution temporelle d'un concept.

Bibliométrie : définie en 1969 comme "l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication" (Pritchard). Utilisée pour mesurer les modes d'utilisation et de publication du matériel documentaire.

Bioinformatique : définie comme étant l'étude, à l'aide de techniques informatiques, des composants du monde vivant et plus particulièrement des molécules telles que les protéines, les acides nucléiques et les sucres, et de leurs interactions.

CGI: Common Gateway Interface. Petit programme informatique, écrit en langage de script, qui permet de réaliser des pages dynamiques et d'interagir avec le serveur Web.

CIRIL : Centre Inter-universitaire de Ressources en Informatique de Lorraine.

Cluster : Ensemble de mots ayant des associations fortes, référant à un thème commun.

ConnectSciences : Portail CNRS d'information scientifique et technique en ligne qui met gratuitement à la disposition du public un ensemble de ressources documentaires et de services produits par l'INIST.

CUI : Unique Identifier for Concept. C'est l'identifiant unique d'un concept du méta-thésaurus auquel est rattaché un terme ou une chaîne de caractères.

DDB00 : Diseases Database 2000. London (England): Medical Object Oriented Software Enterprises Ltd., 2000.

DILIB*: Documentation and Information LIBrary*. Plate-forme documentaire développée par l'INIST servant à générer des serveurs d'investigation.

DSM : Diagnostic and Statistical Manual of Mental Disorders. Washington (DC): American Psychiatric Association.

Fichier inverse : Fichier d'indexation facilitant la gestion des relations entre les termes d'un index et les notices qui les contiennent.

Fichier d'association : Fichier permettant la mise en relation de termes co-occurrents d'un même document.

Format "ed" : Format XML simplifié ne contenant pas toutes les balises de fin. C'est un format pivot pour DILIB. Le passage systématique par ce format permet d'utiliser un programme unique "minibibFromEd" pour convertir différent format au format XML. Cela est intéressant car le format "ed" est facile à obtenir à partir de tout les formats de notices bibliographiques.

FRANCIS : Base de données de l'INIST. C'est la principale base de données européennes sur les sciences humaines et sociales avec près de 2.5 millions de références bibliographiques depuis 1972.

GenBank : base de données des séquences nucléotidiques, créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information, Los Alamos, US).

Gène : région d'ADN qui contrôle un caractère héréditaire précis correspondant habituellement à un ARN unique.

Génome humain : Ensemble des gènes (environ 35 000) localisés sur les 23 paires de chromosomes de l'homme. Il comprend la totalité de l'information génétique appartenant à une cellule ou à un organisme.

HFD : *Hierarchical File organisation for Documentation.* Cette structure est utilisée par DILIB pour stocker les fichiers inverses et les fichiers d'association. Cette structure permet de traiter en standard jusqu'à un million de références bibliographiques réparties en 100 répertoires de 100 fichiers contenant chacun 100 références.

HTML : *HyperText Markup Language*. Il s'agit d'un "langage à balises" (format ASCII), contenant des instructions entre les balises (*tags*) qui sont délimitées entre crochets. Ce langage permet de coder une page à l'aide de commandes de mise en forme. Ces dernières sont ensuite interprétées par un navigateur (*browser*) et apparaissent sur l'écran de l'ordinateur.

http : *HyperText Transfert Protocol*. Ce protocole permet à un serveur de communiquer avec un ou plusieurs clients sous la forme de requête et de réponse.

INALF : Institut National de la Langue Française, cet institut à été renommé par la suite ATILF (Analyse et Traitement Informatique de la Langue Française).

Infométrie : terme adopté en 1987 par la F.I.D. (International Federation of Documentation, IFD) pour désigner l'ensemble des activités métriques relatives à l'information, couvrant aussi bien la bibliométrie que la scientométrie .

INIST : INstitut de l'Information Scientifique et Technique.

INRS : Institut National de Recherche et de Sécurité.

IST : Information Scientifique et Technique.

LORIA : Laboratoire Lorrain de Recherches en Informatique et ses Applications.

LUI : Lexical Unique Identifier. C'est l'identifiant unique pour un groupe de variants lxicaux (termes) auxquels une chaîne de caractère correspond à un des variants. Les chaînes qui partagent un LUI ont la même forme de chaîne normalisée dû au programme « Specialist « norm » » distribué avec le méta-thésaurus.

Makefiles : Fichier utilisés lors de la génération d'une application de DILIB permettant de générer le serveur conformément au paramétrage établi dans le fichier de description.

Méta-thésaurus : c'est une base de données d'informations sur les concepts qui apparaissent sur un ou plusieurs vocabulaires et classifications contrôlés dans le monde biomédical. La couverture du méta-thésaurus dépend des couvertures combinées des différentes sources de vocabulaires.

MeSH : *Medical Subject Headings*. Thésaurus des mots clés de la base *Medline faisant* référence dans le domaine biomédical.

Mésothéliome : Tumeur faite d'une prolifération des cellules de la plèvre. Cette tumeur, de la plèvre, toujours maligne, est habituellement en relation avec une exposition à l'amiante.

PASCAL : Base de données multidisciplinaires et multilingue de l'INIST. Cette base de données signale près de 14 millions de références bibliographiques en sciences, technologie et médecine depuis 1973.

PCR : *Polymerase Chain Reaction.* Réaction *in vitro* permettant d'amplifier des régions spécifiques d'ADN grâce à des cycles multiples de polymérisation de l'ADN, chacun suivi d'un traitement thermique bref pour séparer les brins complémentaires.

Plèvre : Enveloppe du poumon lui permettant de glisser lors de la respiration dans la cage thoracique et cible de l'action toxique de l'amiante.

Protéome : Ensemble des protéines synthétisées à partir des ARN messagers.

Puces à ADN : Technique d'hybridation permettant une analyse génomique comparative de l'expression d'un grand nombre de *patterns* de mRNA. Immobilisés sur un support solide (matrice), des oligonucléotides (simples brins) spécifiques de différents gènes ou ADNc connus constituent les sondes dont le rôle est de détecter des cibles marquées complémentaires, présentes dans le mélange complexe à analyser (ARNm extraits de cellules, tissus ou organismes entiers et convertis en ADNc). Les sondes sont soit greffées sur le support, soit synthétisées *in situ* (unité d'hybridation = plot). Les signaux d'hybridation sont détectés selon le type de marquage, radioactivité ou fluorescence, par mesure radiographique ou par fluorescence, et quantifiés.

REL : Relationship. Présent dans le fichier MRREL. C'est la relation qui lie le second concept CUI2 au premier concept CUI1.

Relation interne : Association entre deux mots-clés appartenant au même cluster.

Relation externe : Association entre deux mots-clés appartenant à des clusters différents.

Registry Number : Index des noms de substances et de leur numéro de nomenclature internationale lorsque celui ci existe (RN pour les produits chimiques, EC pour les enzymes…).

Le réseau sémantique : procure une catégorisation de tous les concepts représentés dans le méta-thésaurus UMLS.

Scientométrie : on peut la considérer comme la bibliométrie spécialisée au domaine de l'IST. Toutefois, la scientométrie désigne d'une manière générale l'application de méthodes

statistiques à des données quantitatives (économiques, humaines, bibliographiques) caractéristiques de l'état de la science.

Secondary Source Identifier (SI) : Intitulé du champs Medline contenant les numéros d'accès vers d'autres bases de données (par exemple Genbank, SwissProt …).

Serveur : Ordinateur qui met ses ressources à la disposition d'autres ordinateurs sous la forme de services, qui peuvent être : Espace disque, Information, Base de données, Traitements automatisés.

SGML : *Standard Generalized Markup Language.*

Shell : Langage de script utilisé sous Unix. Ce langage permet d'écrire des petits programmes appelés script shell.

SNOMED: Systematized nomenclature of medicine. College of American Pathologists.

Specialist Lexicon : développé pour donner toutes les informations lexicales et sert de lexique général qui inclut tous les termes biomédicaux.

SRI : Système de Recherche d'Information.

SUI : Unique Identifier for String. L'identifiant unique pour chaque chaîne de caractère présent dans le méta-thésaurus. C'est l'identité d'une chaîne de caractère. Deux chaînes qui varient seulement d'une majuscule pour la première lettre se verront attribués des SUI différents.

Transcription : Synthèse d'ARNm à partir du brin codant d'ADN.

Transcription inverse : Synthèse d'un brin d'ADN complémentaire à partir d'un ARN.

Transcriptome : C'est l'ensemble des transcrits (ARN messagers) issus de la transcription des gènes exprimés.

Unix : Système d'exploitation multi-utilisateur.

Web : Abréviation de World Wide Web, c'est un ensemble site informatique de type multimédia interconnectés pour constituer une hyper-base de données mondiale accessible en utilisant un navigateur (Netscape Communicator, Internet Explorer).

XML : Ext*ensible Markup Language.* Language de balisage établi pour répondre au besoin d'élargir la nature des documents à échanger sur le Web, de faciliter l'interopérabilité entre applications, et de permettre des descriptions plus précises.

Liste des Figures.

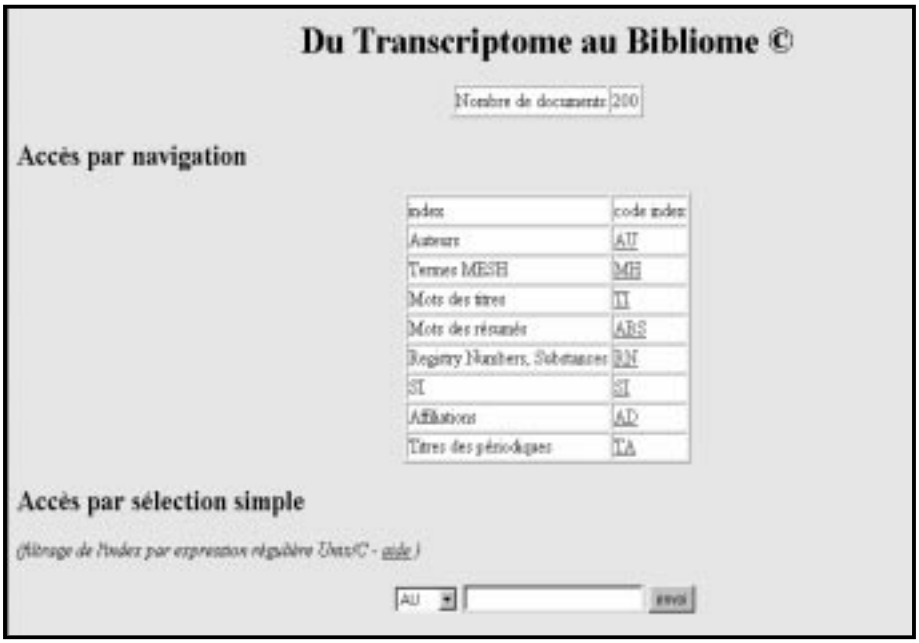
Bibliographie.

1. DUCLOY J. (2001). Plateforme et boîte à outils DILIB. http://www.loria.fr/projet/dilib
2. INIST http://www.inist.fr
3. NEDELCOT A. (2000). Mise à jour et installation de DILIB v0.21, son utilisation dans les applications « IMD » et « Genome ». Rapport de stage de DESS Information scientifique et Technique, Nancy.
4. NEMURAT C. (2001). Développement des applications de DILIB « IMD » et « Transcriptome ». Rapport de stage de DESS Information Scientifique et Technique, Intelligence Economique, Nancy.
5. RIHN B., HOUDRY P., VACHENC S., MOHR S., NEMURAT C., MOULIN D., GRANDJEAN F., DUCLOY J. (2002). " From transcriptomics to bibliomics ". Rédaction en cours.
6. UMLS http://www.nlm.nih.gov/research/umls
7. BORST F., SCHERRER J. -R. (1991)." Les interfaces U.M.L.S. (Unified Medical Language System) ". Informatique et Santé (volume 4), éditions Springer-Verlag (Paris), p. 113-120.
8. LUNIN L. F., HERSH W. R. (1995). "Automated retrieval from multiple disparate information sources: the World Wide Web and the nlm's Sourcerer". Project. J. Am. Soc. Inf. Sci., p. 755-764.
9. RIHN B., MOHR S., McDOWEL S.A., BINET S., LOUBINOUX J., GALATEAU G., LEIKAUF K. and G.D. (2000). « Differential gene expression in mesothelioma ". FEBS Letters, p. 480.
10. MOHR S. (2000). *Etude des transcriptomes de cellules humaines mésothéliales et de mésothéliome*. Rapport de DEA Biologie Moléculaire et Cellulaire, Université Louis Pasteur, Strasbourg.
11. INRS http://www.inrs.fr
12. Adresse Url du serveur Transcriptome : http://portail.inist.fr/dilib/v0.3/DilibBottom/Local/WWW/Veille/Private/Genome/Server/EN.Presentation1.html

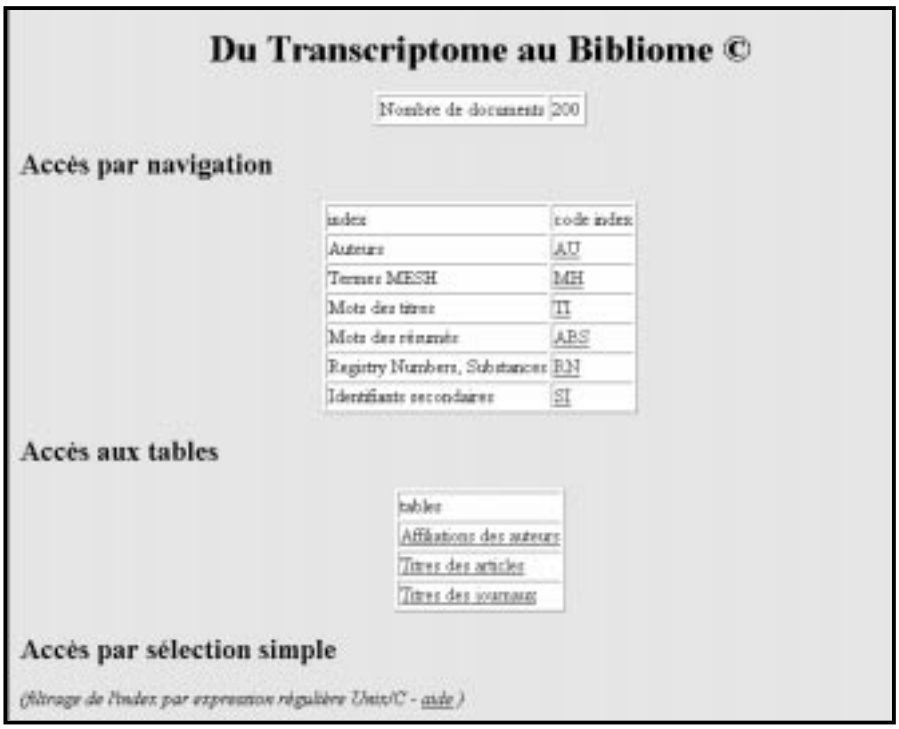
Annexe 1 :

Modification dans la page d'accueil au niveau d'un corpus.

Avant :



Après :



Annexe 2 :

Table Cui2tp.tab (extrait)

C0000005	(131)I-Macroaggregated Albumin
C0000039	1,2-Dipalmitoylphosphatidylcholine
C0000052	1,4-alpha-Glucan Branching Enzyme
C0000074	1-Alkyl-2-Acylphosphatidates
C0000084	1-Carboxyglutamic Acid
C0000096	1-Methyl-3-isobutylxanthine
C0000097	1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine
C0000098	1-Methyl-4-phenylpyridinium
C0000102	1-Naphthylamine
C0000103	1-Naphthylisothiocyanate
C0000107	1-Sarcosine-8-Isoleucine Angiotensin II
C0000119	11-Hydroxycorticosteroids
C0000120	11-Hydroxycorticosteroids, Synthetic
C0000132	15-Ketosteryl Oleate Hydrolase
C0000137	15S RNA
C0000139	16,16-Dimethylprostaglandin E2
C0000151	17 beta-Hydroxy-5 beta-Androstan-3-One
C0000152	17 beta-Hydroxysteroid Dehydrogenases
C0000163	17-Hydroxycorticosteroids
C0000164	17-Hydroxycorticosteroids, Synthetic
C0000165	17-Hydroxysteroid Dehydrogenases
C0000167	17-Ketosteroids
C0000172	18-Hydroxycorticosterone
C0000173	18-Hydroxydesoxycorticosterone
C0000176	19-Iodocholesterol
C0000184	2',3'-Cyclic-Nucleotide Phosphodiesterases
C0000189	2',5'-Oligoadenylate Synthetase
C0000190	2'-CMP
C0000194	2,2'-Dipyridyl
C0000204	2,3-Diketogulonic Acid
C0000215	2,4,5-Trichlorophenoxyacetic Acid
C0000220	2,4-Dichlorophenoxyacetic Acid
C0000232	2,6-Dichloroindophenol
C0000246	2-Acetolactate Mutase
C0000248	2-Acetylaminofluorene
C0000254	2-Amino-5-phosphonovalerate
C0000257	2-Aminoadipic Acid
C0000263	2-Aminopurine

….

Annexe 3 :

Programme ThesFind

```
#include "Html.h"
#include "Server.h"
#include <stdlib.h>
#include "Index.h"
#include "StrSearch.h"
```

Appel des bibliothèques

```
int getopt();
extern char *optarg;

findOtherTerm(thes, table, rel, term)
    Index *thes;
    StrSearchTable *table;
    char *rel;
    char *term;
{
 SgmlNode *inputThesNode;
 if (inputThesNode=IndexReadSgml(thes,term))
   {
    SgmlNode *relNode;
    SgmlNode *itemNode;
    SgmlNode *thesNode;

    thesNode=SgmlCopy(inputThesNode);
    relNode=SgmlGetChildTagAtt(thesNode,"rel","name", rel);

    if ((relNode)&&(itemNode=SgmlFirst(relNode)))
      {
       do
         {
          char *candidate;
          char *newTerm;
          candidate=SgmlLeafGetData(itemNode);
          if (StrSearch(table,candidate ))continue;
          printf("%s\n", candidate);
          newTerm=strdup(candidate);
          StrSearchAdd(table, newTerm, newTerm);
          findOtherTerm(thes, table,rel, newTerm);
         }while((itemNode=SgmlNext(itemNode)));
      }
    SgmlFree(thesNode);
   }
}
```

Fonction findOtherTerm

Traitement récurssif

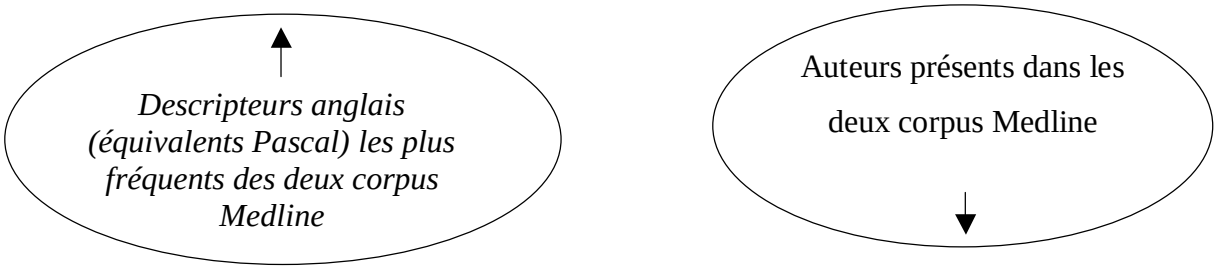
```
int main(argc,argv)
int argc;
  char **argv;
{
  int cod_arg;
  Index *thesHfd;
  char *relName;
  Buffer *bufInput;
  char *newInputTerm;
  StrSearchTable *tabTerm;

  tabTerm=StrSearchTableCreate(100,100);

  while ((cod_arg = getopt(argc,argv,"t:r:"))!=EOF)
    {switch(cod_arg)
      {
      case 't':
        thesHfd=IndexOpenRead(optarg);
        break;
      case 'r':
        relName=optarg;
        break;
      }
    }

  bufInput=BufferCreate(10,10);

  while((BufferGets(bufInput)))
    {
      if (!StrSearch(tabTerm,BufferString(bufInput) ))
        {
          newInputTerm=strdup(BufferString(bufInput));
          printf("%s\n", newInputTerm);
          StrSearchAdd(tabTerm, newInputTerm, newInputTerm );
          findOtherTerm(thesHfd,tabTerm ,relName, newInputTerm );
        }
    }
  exit (0);
                                                    }
```

Programme principal

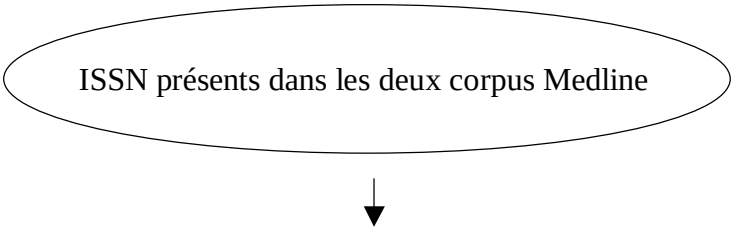
Annexe 4 :

requête Miriad pour obtenir le corpus pseudo clone

(ed="binding protein" ou ed="multigene family" ou ed="membrane protein" ou ed="heat hock protein" ou (ed="gene expression" et ed="regulation") ou ed="repetitive dna" ou ed="repeated sequence" ou ed="dna binding protein" ou ed="brain" ou ed="cell nucleus" ou ed="transcription factor" ou ed="nuclear protein" ou ed="mitochondria" ou …)



ET (au="abu-elheiga l" ou au="adachi y" ou au="aizawa h" ou au="albertsen hm" ou au="albrecht c" ou au="alexander c" ou au="alpan rs" ou au="amatruda jf" ou au="andersson b" ou au="arfin sm" ou au="barillari g" ou au="bartels jl" ou au="baughman g" ou …)



ET (sn="1061-4036" ou sn="0028-0836" ou sn="0008-5472" ou sn="0964-6906" ou sn="1340-2838" ou sn="0378-1119" ou …) .

Annexe 5 :

MyEmptyTable

Human
Rat
Mouse
Genetics
In vitro
Animal
Aminoacid sequence
Aminoacid
Amphibia
Analysis
Antibodies
Monoclonal
Bacteria
Bacterial Proteins
Baculoviridae
Nucleotide sequence
Biosynthesis
Blister
Northern blotting
Southern blotting
Western blotting
Caenorhabditis
Caenorhabditis elegans
Case Report
Catalysis
Cattle
Cell count
Cell extracts
Cell line
Established cell line
Mechanism of action
Cell culture
Centrifugation
Cercopithecus aethiops
Chemistry
Cell proliferation
Enzyme
Enzymatic activity
Chromosome DNA
Molecular interaction
Gene expression
Complementary DNA
Molecular cloning
In vivo
Mutation
Immunoblotting assay
Comparative study
Cell cycle
Biological activity

Annexe 6 :

Fichier paramètre genome.desc.ed (extrait)

NB : un seul des corpus Medline est signalé dans cet extrait

```
<server code=genome>
 <base code=Gene1 from=medline>
   <input>
     <file env=DEMO_ROOT>Temp/corpus_medline1.xml
   </input>
   <index code=ED lexicon=yes>
    <path>medline/ED/e#
    <minibibTag>ED
    <cross>AU
    <cross>ED
    <cross>ETI
    <cross>country
    <cross>date
   </index>
     <backBase base="pas99">
     <backBase base="pas00">
     <backBase base="pas01">
     <backBase base="pas02">
 </base>
 <base code=pas from=pascal>
   <input>
     <file env=DEMO_ROOT>Temp/pascal.xml
   </input>
     <backBase base="pas99">
     <backBase base="pas00">
     <backBase base="pas01">
     <backBase base="pas02">
 </base>
 <background from=pascal>
  <backBase code="pas99">
    <name>Pascal 1999
    <root>/applis/dps/INRS/GenomeAP/Back/Pas1999
   </backBase>
  <backBase code="pas00">
    <name>Pascal 2000
    <root>/applis/dps/INRS/GenomeAP/Back/Pas2000
   </backBase>
  <backBase code="pas01">
    <name>Pascal 2001
    <root>/applis/dps/INRS/GenomeAP/Back/Pas2001
   </backBase>
  <backBase code="pas02">
```

← Chargement du corpus Medline Gene 1

Index mots-clés MesH

Index croisés

Appel bases Pascal d'arrière-plan

← Chargement du corpus Pascal « pseudo » clone (second plan)

Chargement des 4 bases Pascal de troisième plan

```
      <name>Pascal 2002
<root>/applis/dps/INRS/GenomeAP/Back/Pas2002
    </backBase>
  </background>
  <base code=MULTI type=multi>
      <index code=aut>
      </index>
      <index code=ED>
      </index>
      <index code=FD>
      </index>
      <index code=ETI>
      </index>
      <index code=FTI>
      </index>
  </base>
 <thes code=MH>
    <input env=DEMO_ROOT>Import/Data/ThesGenome
    <rel code=RN inv=RB>
    <rel code=RB inv=RN>
 </thes>
</server>
```

Chargement base Multi

(Medline + Pascal pseudo-clone)

Chargement du thésaurus

ThesGenome

Résumé :

Le serveur d'investigation Transcriptome est le résultat d'un travail de coopération entre l'Institut de l'Information Scientifique et Technique (INIST) et l'Institut National de Recherche et de Sécurité (INRS). Ce serveur est construit grâce à la plate-forme documentaire *Documentation and Information Library* (DILIB) et offre des fonctionnalités d'analyses infométriques.
La première partie du stage a consisté à intégrer un sous-ensemble du méta-thésaurus de *l'Unified Medical Language System* (UMLS) cohérent avec la thématique des deux corpus Medline préexistants dans le serveur Transcriptome afin de pouvoir améliorer le caractère informatif de la navigation.
La seconde partie du stage a été consacrée à l'intégration de bases Pascal avec quatre années d'arriérés pour suivre l'évolution temporelle des concepts émergents.

Mots-clés :

Transcriptome, mésothéliome, UMLS, MeSH, base Pascal, XML, DILIB, passerelle sémantique, multibase, infométrie, bibliométrie.