Mise à jour et installation de DILIB v0.21 Son utilisation dans les applications "IMD" et "Génome"

Aude Nedelcot

Université Henri Poincaré Nancy 1
Université Nancy 2
Institut National Polytechnique de Lorraine

D.E.S.S Information Scientifique et Technique
Année Universitaire 1999-2000

Mise à jour et installation de DILIB v0.21
Son utilisation dans les applications "IMD" et "Génome"

par

Aude NEDELCOT

Maîtres de Stage :

Jacques DUCLOY
Institut de l'Information Scientifique et Technique
2, allée du Parc de Brabois
F-54514 Vandœuvre-lès-Nancy

Dr Bertrand RIHN
Françoise GRANDJEAN
Institut National de Recherche et de Sécurité
BP 27
54500 Vandoeuvre-lès-Nancy

Stage effectué du 1er mai au 31 juillet

à

Institut National de Recherche et de Sécurité
Pour la prévention des accidents du travail et des maladies professionnelles

Plan du mémoire

Avertissement:

Tous les mots suivis d'une "*" ont une définition dans le glossaire. Tous les sigles sont définis dans le chapitre "Abréviations".

Remerciements

Je tiens à remercier :

- Jacques DUCLOY pour la confiance qu'il m'a accordée et le temps qu'il m'a consacré.
- Bertrand RIHN pour son encadrement dans le projet et pour sa patiente relecture.
- Françoise GRANDJEAN pour m'avoir permis d'effectuer ce stage.
- Michel SERVAIS pour son aide lors de l'installation de DILIB.
- Stéphane SCHNEIDER pour ses encouragements, son soutien, son aide, et sa disponibilité.
- Alain ZASADZINSKI pour sa précieuse aide lors de la partie purement biomoléculaire de ce travail.
- Philippe HOUDRY et Florian MAZUR pour leurs conseils et leur aide.
- Tous les membres du DPS pour leur accueil, leur sympathie et leur soutien.

Introduction

Dans un centre de recherche comme l'Institut National de Recherche et de Sécurité pour la prévention des accidents du travail et des maladies professionnelles (INRS), la masse de l'information bibliographique est très importante. Afin de produire, exploiter et gérer de façon efficace une telle abondance d'informations documentaires, le système informatique se doit d'être performant. Par ailleurs, le fonds documentaire doit être accessible à tous. C'est pour ces différentes raisons que le centre de documentation utilise des produits tels que "AIRS Web"* et la plate-forme documentaire *Documentation Information LIBrary* (DILIB).

Jacques Ducloy, concepteur de DILIB et responsable du Département Produits et Services de l'INstitut de l'Information Scientifique et Technique (INIST), a permis le développement récent de DILIB en une nouvelle version intéressant l'INRS. DILIB est un outil permettant la gestion et l'exploitation de gros corpus documentaires tels que ceux possédés par l'INRS (environ 1.2 Go). Le centre de documentation de l'INRS utilise DILIB depuis plusieurs années car cet outil offre l'avantage d'évoluer selon les besoins propres de l'institut et permet, grâce à l'Intranet, une consultation aisée et dynamique de toute l'information documentaire.

Par ailleurs, DILIB a suscité l'intérêt de Bertrand Rihn, chercheur à l'INRS, qui a souhaité l'utiliser pour l'exploitation de données bibliographiques issues d'une étude des gènes impliqués dans le mésothéliome humain: il s'agit de l'application "génome" de DILIB

La mise à jour de l'application du centre de documentation et la création de l'application "génome" constituent mes deux axes de travail. Ils seront donc développés dans le présent mémoire, compte rendu du stage que j'ai d'effectué à l'INRS dans le cadre du D.E.S.S. Information Scientifique et Technique.

I. Présentation de l'INRS

1. Missions et moyens

L'INRS a pour mission d'agir pour la prévention des accidents du travail et des maladies professionnelles. C'est une association loi 1901 créée en 1968 par la CNAM (Caisse Nationale d'Assurance Maladie).

L'institut emploie environ 600 personnes réparties sur trois sites : Paris (siège social), Vandoeuvre-lès-Nancy et Neuves-Maisons. Il exerce ses activités autour de quatre thèmes :

- mieux connaître les risques professionnels,
- analyser leurs conséquences pour la santé de l'homme au travail,
- rechercher comment les combattre et les maîtriser,
- faire connaître et enseigner les moyens de leur prévention.

Ces quatre missions se répartissent de la façon suivante :

- études et recherches (essentiellement sur le site de Vandoeuvre-lès-Nancy) : épidémiologie, toxicologie, chimie, nuisances physiques, machines dangereuses, robotique, psychosociologie du travail, multi-nuisances…

- assistance : les experts de l'INRS assurent une assistance technique et médicale répondant aux demandes de conseils et d'études émanant des ministères, des organismes de sécurité sociale, des services de médecine du travail ou des entreprises relevant du régime général de l'assurance maladie. Ils participent également à des groupes français et européens de normalisation. Ils examinent, testent et permettent la certification européenne (norme CE) de substances, machines ou équipements de protections.

- formation : il s'agit de formation continue à la prévention des risques professionnels des salariés des entreprises, de formation à la prévention des risques professionnels dans l'enseignement technique et de la formation à la prévention de risques professionnels dans les écoles d'ingénieurs. La formation est destinée aux ingénieurs, techniciens, médecins du travail, contrôleurs des Caisses Régionales d'Assurance Maladie (CRAM), personnel des entreprises et aux organismes de formation.

- information : l'INRS se doit de sensibiliser aux risques professionnels. La prévention se concrétise par la réalisation de documents destinés à un public varié. Ces documents ont plusieurs formes :
    - brochures (plus de 400 titres),
    - affiches (environ 190 documents),
    - films (environ 120 titres),
    - périodiques (5 documents : un mensuel, trois trimestriels et un qui paraît sept fois par an),
    - CD-ROM (7 titres).

Les partenaires de l'INRS sont en premier lieu les services prévention des Caisses Régionales d'Assurance maladie, les médecins du travail, les partenaires en entreprises et leurs représentants, ainsi que les instances et organismes spécialisés en prévention.

2. Le centre de Vandoeuvre-lès-Nancy

Le centre de Vandoeuvre-lès-Nancy compte 390 personnes, et six départements Etudes et de Recherches. Il est chargé d'élaborer les projets du programme d'Etudes et de Recherches de l'institut et de les soumettre à l'avis des différentes instances qui les examinent avant leur adoption éventuelle par le conseil d'administration.

Les projets de recherche correspondent à des demandes émanant de l'extérieur : CNAM, CRAM, médecine du travail, organisations professionnelles…, ou à des initiatives propres aux chercheurs.

3. La documentation et le réseau informatique

L'INRS Paris possède un important fonds documentaire, constitué d'ouvrages, de brochures, d'articles et de périodiques. Son corpus s'accroît de façon importante chaque année, puisqu'il progresse de plus de 3000 documents par an, dont environ 1000 ouvrages et brochures, le reste étant essentiellement constitué d'articles et de périodiques. Ce service a également constitué deux bases de données : INRS-B (base de données recensant tous les documents retenus par le centre de documentation depuis 1981) et INRS-R. (base internationale recensant les recherches en cours et les projets de recherche). INRS-B contient environ 40000 références (soit 1.2 Go).

L'ensemble du fonds documentaire est consultable par le grand public dans le centre de documentation parisien de l'INRS. Cependant, le centre de Vandoeuvre-lès-Nancy possède son propre centre de documentation créé en 1970.

a. Le centre de documentation de Vandoeuvre-lès-Nancy

Le centre de Vandoeuvre-lès-Nancy, composé de quatres personnes (deux documentalistes et deux secrétaires) met à la disposition des chercheurs ses propres ressources (plus de 160000 références) ainsi que toutes les ressources de l'institut (deux bases de données : NDOC et PDOC représentant plus de 1 Go). Ce service principalement à usage interne, a pour mission de fournir, aux chercheurs du centre, l'information dont ils ont besoin pour la réalisation de leurs programmes d'études et de recherches. Toutefois, il est accessible aux personnes extérieures à l'INRS (étudiants, médecins, …) qui peuvent consulter, sur place, ce fonds documentaire très spécialisé. Celui-ci se compose d'ouvrages, de périodiques, de normes et de notes techniques.

Des ordinateurs, en libre service, permettent l'accès à de nombreux documents électroniques sur des supports tels que les CD-ROM (Medline, Toxline, Cc-Info, …) et la consultation des différentes bases en Intranet ou sur Internet via le réseau scientifique et technique interne.

Au sein même du centre de documentation de Vandoeuvre-lès-Nancy, les chercheurs ont une large place dans la chaîne documentaire. En effet, ils effectuent eux-mêmes les recherches documentaires propres à leur sujet d'études et signalent les références intéressantes au centre de documentation qui se charge de les commander. Chaque laboratoire peut alimenter une base de données qui regroupe les différentes publications disponibles dans son domaine (avec le logiciel AIRS).

Ce logiciel est utilisé par l'institut pour permettre une saisie des références de chaque service. L'ensemble de ces bases est consultable par tous les chercheurs de l'institut *via* l'Intranet, grâce à des outils comme "AIRS Web"* et grâce à la plate-forme documentaire DILIB comme nous le verrons dans la suite de ce mémoire.

La médiathèque de Neuves Maisons est plus spécialement réservée aux stagiaires en formation à l'INRS. Son fonds documentaire est également géré sous AIRS.

b. Le réseau informatique

Le réseau informatique de l'INRS se compose de plusieurs réseaux locaux propres à chaque bâtiment, fédéré par un réseau d'entreprise utilisant le protocole TCP/IP*, le tout ouvert sur l'extérieur avec accès au REseau NAtional de télécommunications pour la Technologie, l'Enseignement et la Recherche (RENATER), via le Centre Inter-universitaire de Ressources en Informatique de Lorraine (CIRIL) situé à Vandoeuvre-lès-Nancy. Ce réseau fonctionne sous Unix ou sous Windows NT.

Sur le site de Vandoeuvre-lès-Nancy, le Centre de Services Informatiques de Lorraine (CSIL) est responsable de la gestion du réseau.

4. Le laboratoire de cancérogenèse

Le laboratoire de cancérogenèse, dirigé par Bertrand Rihn, appartient au département Polluants et Santé qui a pour vocation la recherche en toxicologie dans le domaine de l'évaluation des risques dus à l'exposition en milieu professionnel.

Le groupe dirigé par Bertrand Rihn étudie en particulier l'action mutagène des toxiques industriels (produits chimiques, particules solides) sur des modèles murins transgéniques.

Bertrand Rihn est à l'origine d'une thématique de recherche utilisant les techniques de puces à ADN pour mieux élucider la biologie des cancers professionnels comme le mésothéliome*. Ce sont les résultats de l'étude des gènes du mésothéliome que nous avons exploités dans le cadre de l'application "Génome" créée à partir de DILIB.

II. DILIB et la norme *Standard Generalized Markup Language* (SGML)

L'INRS, comme nous l'avons vu, travaille sur des données bibliographiques au format AIRS. Cependant, ce type de format est spécifique à une base de données particulière. Pour que l'exploitation des données soit accessible à tous, il faut reformater les données et leur donner un format "universel".

Le langage standard international visant à décrire un document sous sa forme logique est SGML.

1. La norme et ses avantages

Selon les bases de données, les notices bibliographiques ont une structure différente. Le *Standard Generalized Markup Language* est une norme qui permet de décrire et de représenter tous les documents structurés. Au départ il était conçu pour faciliter l'échange de gros volumes de documents électroniques scientifiques et techniques dans le cadre du projet *Computer-aided Acquisition and Logistic Support* (CALS) du département de défense des Etats Unis (*Department Of Defence*: DOD). Cette norme s'est avérée d'un usage beaucoup plus général en devenant un langage de balisage généralisé. Il décrit la structure logique de documents, indépendamment de leur format de départ. Le traitement des documents est ainsi indépendant des systèmes utilisés pour les générer ce qui facilite leur exploitation ultérieure.

Le balisage d'un document correspond à l'intégration de marques délimitant les différentes parties de celui-ci pour faciliter leur identification. L'utilisateur pourra, à sa guise, nommer les différentes parties de son document pour en donner une arborescence logique.

Si nous prenons l'exemple d'une notice Medline (Annexe I), celle-ci comporte un certain nombre de champs qui forment sa structure (Figure 1). Nous avons décidé de garder tous ces champs dans l'architecture SGML. Nous obtenons alors un document qui a la structure suivante:

```
<medline>
<UI>97444286</UI>
<AU><e>Choudhury BK</e>
      <e>Kim J</e>
      <e>Kung HF</e>
      <e>Li SS</e></AU>
<TI>Cloning and developmental expression of Xenopus cDNAs
encoding the Enhancer of split groucho and related
proteins.</TI>
....
<DE><e>Amino Acid Sequence</e>
      <e>Animal</e>
      ....
      <e>Xenopus&sol;growth &amp;
development&sol;*genetics</e></DE>
....
<EM>199712</EM>
....
<SI>GENBANK&sol;U18775</SI>
....
<SO>Gene 1997 Aug 11;195(1):41-8</SO>
</medline>
```

Figure 1 : Structure d'une notice Medline.

La norme SGML offre de nombreux avantages, entre autres :

- la codification d'un document se fait en utilisant un jeu de caractères minimum (ISO 646) et une codification spécifique pour les caractères spéciaux (signes mathématiques, langue latine, …) facilitant ainsi sa manipulation et sa portabilité,
- la norme sépare le contenu et la forme des documents. La cohérence de l'architecture des données traitées est donc indépendante du contenu de celles-ci,
- le balisage favorise l'utilisation d'analyseurs lexico-syntaxiques simplifiant l'exploitation des résultats,
- la norme permet une navigation dans le monde de l'Internet (*world wide web* : www). En effet, le langage de balisage utilisé sur le web, *HyperText Markup Language* HTML, n'est autre qu'une application (Définition de Type de Document : DTD) de la norme SGML, aussi, la consultation et l'utilisation de documents en SGML sont-elles favorisées,
- la norme SGML est largement répandue.

DILIB, bénéficiant des atouts de la norme SGML, offre par ailleurs, des outils facilitant le traitement des données au format SGML. C'est ce que nous allons voir dans cette seconde partie.

2. La plate-forme DILIB

a. Généralités

DILIB (*Documentation Information LIBrary*) est une plate-forme d'exploitation de l'information pour l'Ingénierie du Document de l'Information Scientifique et Technique. Il s'agit d'une boîte à outils permettant la conversion au format SGML, l'exploration et l'exploitation de corpus documentaire pour une analyse infométrique poussée.

La première version de la plate-forme SGML pour la documentation, nommée ILIB, a été réalisée dans le cadre d'un projet du Département Recherche et Produits Nouveaux de l'INIST. Elle a ensuite évolué au "Laboratoire Lorrain de Recherches en Informatique et ses Applications" (LORIA) sous le nom de DILIB.

DILIB permet de manipuler et d'exploiter des documents de formats initialement différents grâce à des outils tels que :

- une bibliothèque de fonctions en langage C,
- des commandes de manipulation de documents,
- des données normalisées et des exemples,
- des interfaces avec des logiciels disponibles sur le marché.

Cette plate-forme de gestion de l'information contient également des fonctionnalités permettant de construire des systèmes de recherche d'information :

- des modules d'interface avec Internet,
- des outils d'interrogation (opérateurs booléens*),
- des applications prédéfinies (par exemple : gestion d'une base de bandes dessinées, gestion d'une base de données sur la dégénérescence Wallerienne des neurones) ,
- des outils de construction de système.

Les domaines d'application de DILIB sont vastes, mais il existe trois produits cibles :

- l'investigation documentaire,
- la constitution de systèmes de recherche de l'information,
- la construction de plates-formes d'exploitation de l'information.

L'étude des gènes impliqués dans le mésothéliome (laboratoire de cancérogenèse, Bertrand Rihn) a généré un important fichier de données (2.9 Mb dans sa totalité) candidat à une exploitation par la plate-forme DILIB.

b. Génération d'une application

Il existe plusieurs étapes (Figure 2) dans la construction d'un outil d'exploitation de l'information à l'aide de DILIB. Celles-ci permettront une visualisation des résultats sur Internet :

- acquisition d'un corpus documentaire (par Internet ou dans une base de données),
- transformation des données initiales au format SGML,
- création de fichiers inverses* (Figure 5),
- création de fichiers d'associations* (Figure 6),
- création de clusters*,
- création et génération de pages grâce au langage HTML.

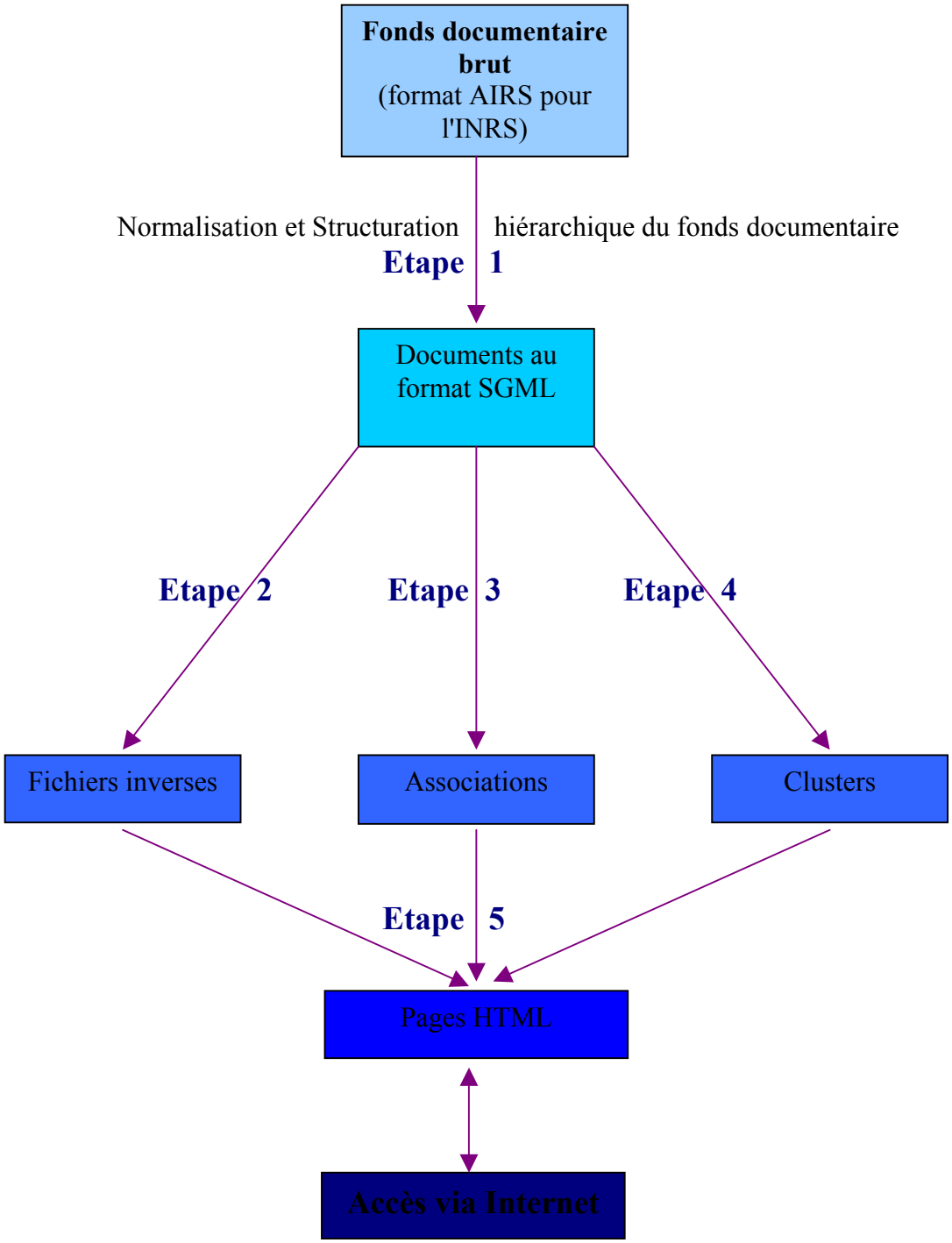


Figure 2 : Les étapes de création d'une application pour la gestion de l'information

→ Utilisation de fonctions ou de modules de DILIB

A chaque étape de la génération d'une application correspond l'utilisation d'un module particulier de DILIB faisant intervenir des plusieurs fonctions.

- Etape 1, normalisation. Cette étape nous permet tout d'abord de convertir le fonds documentaire sur lequel nous voulons travailler au format SGML. L'homogénéisation des données est possible grâce à une fonction spécifique. C'est le cas de la fonction "AirsToSgml" utilisée dans le développement de l'application du centre de documentation de l'INRS pour lequel les notices bibliographiques initiales se trouvent au format AIRS.
  Lors de cette étape, les données seront stockées selon une architecture dite HFD pour *Hierarchic File organization for Documentation*. Cette structure offre un temps de traitement plus rapide du fonds documentaire et un accès plus facile aux données nécessaires.

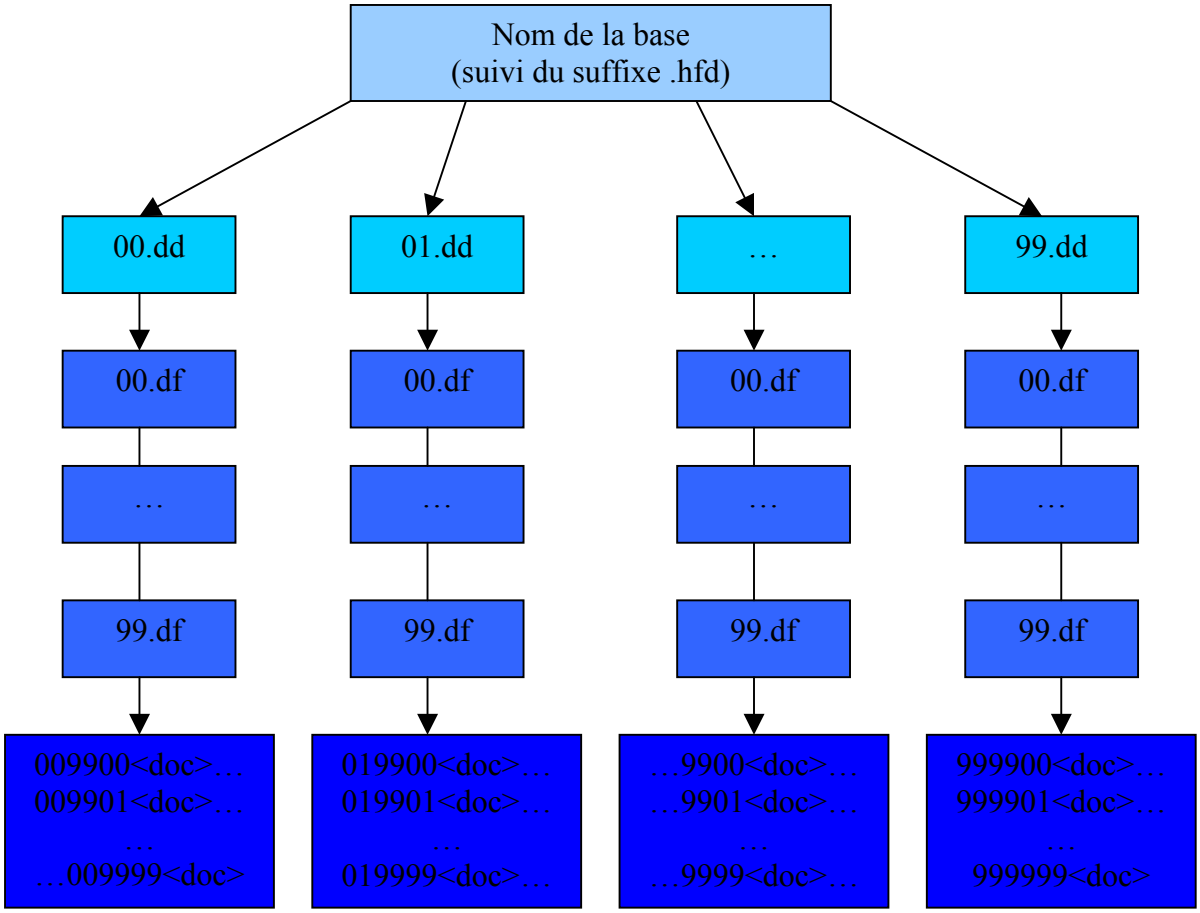


Figure 3 : Structure hiérarchique d'un corpus documentaire

La Figure 3 est un schéma simplifié de la structure HFD nous permettant de mieux appréhender l'organisation d'une application générée par DILIB. Dans cet exemple, la base contient un million de références bibliographiques, réparties en 100 répertoires contenant chacun 100 fichiers contenant respectivement 100 documents.

Cette structure permet la gestion des répertoires de fichiers inverses et des répertoires de fichiers d'associations.

Afin de faciliter l'accès aux données, un fichier ".hcs" (Figure 4), possédant lui aussi la structure SGML, est généré.

```
<hfd>
<struct>
<dir><digit>0123456789</digit><digit>0123456789</digit></dir>
<file><digit>0123456789</digit><digit>0123456789</digit></file>
<key><digit>0123456789</digit><digit>0123456789</digit></key>
</struct>
<name>/users/nedelcot/genome/Server/genome.bib</name>
<state>created</state>
<lastkey>000444</lastkey>
<nrec>445</nrec>
</hfd>
```

Figure 4 : Exemple d'un fichier ".hcs"

- Etape 2, indexation des données. Cette étape consiste en la création des fichiers inverses (ou index). DILIB utilise de nombreux fichiers inverses afin de récupérer les données nécessaires à la construction de graphes de navigation. Les fichiers inverses permettent d'accéder de manière directe aux objets reliés entre eux par une propriété commune (par exemple tous les titres). Il s'agit donc d'un mécanisme d'indexation facilitant la gestion des relations entre les objets.

  Pour chaque terme indexé, sont indiqués le nombre de références dans lequel se trouve ce terme et le numéro de toutes ses références.

```
000022
<idx>
<kw>*Chromosomes, Human, Pair 5</kw>
<lc>*chromosomes, human, pair 5</lc>
<f>4</f>                    (1)
<l>
<e>000082</e>
<e>000137</e>               (2)
<e>000292</e>
<e>000317</e>
</l>
</idx>
```

Figure 5 : Exemple de structure d'un fichier inverse.
(1) : Fréquence du mot-clé (2) : Documents dans lesquels se trouve le terme indexé

- Etape3, création des associations. Cette étape consiste à extraire les associations de termes (deux auteurs ou deux mots-clés) des notices, avec leur fréquence d'apparition et de co-apparition (Figure 6).

Il faut signaler que chaque association est exprimée une seule fois et que les associations non significatives (trop élevées ou trop petites) sont éliminées pour limiter le volume d'information et pour éviter de biaiser les résultats de la génération des clusters qui va suivre.

```
000730
<assoc>
<ti>
<kw>Support, U.S. Gov&amp;apos;t, P.H.S.</kw>
<f>123</f>                                    (1)
</ti>
<tj>
<kw>Species Specificity</kw>
<f>16</f>                                     (2)
</tj>
<fij>6</fij>                                  (3)
</assoc>
```

Figure 6 : Exemple de structure d'un fichier d'association
(1) : fréquence du premier mot-clé (2) : fréquence du second mot clé (3) : fréquence de co-apparition

- Etape 4, création des clusters. Un cluster* représente un thème. Il est composé d'un groupe de mots, de ses relations internes* et externes* avec d'autres clusters. La construction des clusters s'effectue à partir des associations précédemment établies. L'algorithme permettant la génération des clusters consulte la liste des associations de termes classées par ordre de pertinence. Les clusters sont alors générés par incrémentation en regroupant les associations possédant des termes communs. Le cluster regroupe l'ensemble des associations les plus pertinentes. L'ensemble des associations restantes est utilisé de manière additionnelle pour constituer les liens entre les différents clusters.

- Etape5 : Accès Internet : Cette étape permet la visualisation des résultats *via* Internet. Tous les fichiers nécessaires à l'exécution des *Common Gateway Interface* (CGI) ont été générés.
  A partir des pages HTML générées, la consultation des résultats sur Internet est alors possible.

3. DILIB à l'INIST

a. Le plan de mutation technologique

Au cours de mon stage, j'ai été intégrée au DPS (Département Produits et Services), dirigé par Jacques Ducloy, initiateur de ILIB, première version de ce qui est devenu aujourd'hui DILIB. Au sein de ce département, l'utilisation de DILIB est inscrite dans le "plan de mutation technologique". Celui-ci est élaboré à partir d'un plan de mise en place de nouveaux services, accompagné d'un plan de formation.

L'introduction d'une nouvelle technologie telle que DILIB n'est pas toujours aisée. Mais cet outil doit permettre, aux ingénieurs de l'INIST, la création de nouveaux produits, l'amélioration de la fabrication des bases et services existants et le développement de partenariats avec la portabilité des nouveaux moyens technologiques.

Le SIP (Service Ingénierie et Partenariat) est le service pilote de l'implantation de DILIB à l'INIST. Cette boîte à outils devrait permettre, outre la simple gestion d'applications existantes, le développement et la création de nouvelles applications. C'est dans ce service que j'ai appris à manipuler DILIB et à construire des applications utilisant cette plate-forme documentaire. Une nouvelle version de DILIB a été développée par Jacques Ducloy afin d'avoir une boîte à outils standard qui puisse être portable et mise à jour facilement sur des sites extérieurs à l'INIST. Cette version a une architecture claire (Annexe II) qui permet une navigation plus aisée et une adaptation à chaque site d'exploitation. Pour que l'utilisation de DILIB soit totalement possible, il a fallu par ailleurs installer un serveur "Apache" pour l'exploration des futures applications.

b. Le serveur "Apache".

La consultation sur Internet des applications créées avec DILIB nécessite l'utilisation d'un serveur "http"* pour la prise en compte de l'exécution des différentes commandes dynamiques (CGI) de DILIB. Les CGI sont des systèmes ou des programmes qui sont exécutés par un serveur lorsque le client en donne l'ordre (par un lien par exemple). Ils permettent la création dynamique de documents en HTML, le stockage ou la prise en compte d'informations (remplissage de formulaire par exemple).

"Apache" est le serveur "http" le plus utilisé dans le monde de l'Internet de par sa robustesse et sa gratuité. A l'INIST, il existe actuellement plusieurs serveurs "Apache", mais chaque serveur est destiné à une application particulière. Il était donc nécessaire d'installer un nouveau serveur, avec des paramètres spécifiques, pour toutes les applications qui allaient être créées avec DILIB. Les informations concernant le paramétrage (notamment la sécurité d'utilisation) ont été trouvées sur le site d'Apache (http://www.apache.org) ainsi que dans le livre "Apache". Les spécialistes "Apache" de l'INIST ont aidé à donner une touche finale correcte à ces paramétrages. Une fois le serveur paramétré correctement, il ne faut pas oublier de lui donner l'accès à trois répertoires :

- le répertoire contenant les CGI de la DILIB (répertoire "htbin"),
- le répertoire contenant les applets java (répertoire "java"),
- le répertoire contenant les bases de l'application (répertoire "Server").

Pour cela une commande devait être créée : un "ScriptAlias" favorisant l'accessibilité par le serveur aux documents contenus dans ces différents répertoires. Il faut cependant noter que le paramétrage des "Alias" ne s'est pas fait de la même façon dans le serveur de l'INIST et dans le serveur de l'INRS. Les consignes de sécurité n'étaient pas tout à fait les mêmes. Nous avons utilisé des liens symboliques dans le serveur de l'INIST, que nous ne pouvions pas établir au sein du réseau de l'INRS (voir chapitre suivant). La Figure 7 est l'illustration des liens symboliques possibles dans le serveur de l'INIST.

```
ScriptAlias /bin/ "/applis/dps/WWW/site.deploie/cgi/"  (1)
<Directory "/applis/dps/WWW/site.deploie/cgi/">
   AllowOverride None
   Options ExecCGI FollowSymLinks
   <Limit GET POST>
      Order allow,deny
    …….
   </Limit>
 </Directory>
```

Figure 7 : Exemple d'un "ScriptAlias"
(1) : "cgi" est un répertoire contenant un lien symbolique vers le répertoire "htbin" de DILIB)

4. DILIB à l'INRS

Depuis plusieurs années, l'INRS utilise DILIB. Cet outil expérimental est en perpétuelle évolution et il est souvent nécessaire d'effectuer des mises à jours pour que les dernières améliorations de la plate-forme documentaire soient accessibles. L'INRS fait partie des partenaires de l'INIST qui emploient DILIB pour l'exploitation d'applications déjà créées. Contrairement aux ingénieurs de l'INIST, les chercheurs de l'INRS utilisent DILIB comme un outil de consultation et non comme un outil de génération d'applications.

La première application mise en place est celle du centre de documentation, "IMD", détaillée dans le troisième chapitre. Dans le cadre de cette application, DILIB rend accessible et exploitable, toutes les bases ayant comme format originel, le format AIRS. La seconde application est celle destinée à l'exploitation des résultats de l'étude menée par l'équipe de Bertrand Rihn, étudiée dans le quatrième chapitre. L'analyse des résultats portera en particulier, sur les associations et les clusters construits par cette application "Génome".

III. L'application *INRS Multi Data* (IMD) du centre de documentation

Afin d'améliorer l'application IMD, celle-ci a été installée à l'INIST durant mon stage. Un premier travail a consisté à modifier la structure de l'application pour ensuite rectifier les calculs statistiques erronés des résultats.

1. L'application et ses inconvénients

L'application utilisée pour la gestion des bases de données documentaire du centre de documentation évolue d'années en années. Elle se conforme à l'amélioration des capacités et des performances de chaque nouvelle version de DILIB.

Au cours de l'été 1999, l'application a été améliorée et de nombreuses fonctions ont été ajoutées. Ces fonctions n'étaient pas intégrées dans l'architecture même de DILIB. Certaines d'entre elles, comme celles qui permettent l'exploitation de données en multibase, ont donc été améliorées et inclues dans DILIB. Ainsi les ajouts parallèles de fonctions des années précédentes ont été supprimés et assimiler à la dernière version de DILIB.

Tous les documents des bases de données sont au format AIRS. Lors du passage des notices au format SGML, certaines informations étaient ignorées. Par conséquent, les résultats fournis par l'application étaient incorrects et incomplets. Il a été établi que le problème provenait de l'ajout de tabulations dans les documents lors du formatage des notices au format AIRS.

Par ailleurs, afin de simplifier la gestion de l'application, cette dernière était générée par un enchaînement de programmes complexes rendant difficile la maintenance.

2. Les modifications apportées

a. Première modification

La première modification a concerné l'architecture de l'application. En effet, la nouvelle version de DILIB ayant une architecture "standard", toutes les applications devant être créées à partir de cet outil devaient également être construit selon une architecture "standard".

Les répertoires de génération de l'application ont été réorganisés pour obtenir une architecture "standard". Ce changement a permis la suppression de *makefiles* (programme de générations de fichiers) très nombreux et trop compliqués pour assurer une maintenance rapide.

Ces améliorations devraient simplifier les futures opérations de mise à jour de l'application "IMD".

- Nouvelle architecture de l'application IMD :

L'application contient au départ 6 fichiers :IMD.init.sh, IMD.make.sh, IMD.index.mk, IMD.bib.mk, IMD.desc.ed et IMD.dict

Elle contient également 4 répertoires :
- "bin" contient IMD.mk (*makefile* de l'application) ainsi que des fichiers d'initialisation des variables: "def.sh" et "def.csh",
- "Prog" contient tous les programmes de génération des bases : "Nom_pgme.bib.sh",
- "Server" contient les fichiers générés,
- "Text" contient les bases et leurs données.

- Lancement de l'application IMD :

Pour lancer l'application, il faut, dans l'ordre :

- sous IMD : lancer le programme "IMD.init.sh" qui permet d'initialiser les variables d'environnement de DILIB et de les sauvegarder. Cela permet la création des 4 répertoires nécessaires à la construction des bases et la déclaration des variables d'environnement dans le répertoire "bin".

- sous IMD : lancer le *makefile* de l'application, "IMD.make.sh".

- Importation d'une nouvelle base dans l'application IMD :

Pour incorporer une nouvelle base dans l'application, il faut :

- dans le répertoire "Text" : enregistrer le contenu de la base (EXEMPLE.BIB.1).

- dans le répertoire "Prog" : créer le shell de génération de la base (EXEMPLE.bib.sh).

- dans le fichier "liste_bases" : ajouter le nom de la base (EXEMPLE).

- dans le fichier "def.path.input" : déclarer la base.

- Dans le fichier "IMD.index.mk" : compléter le shell avec la suite de ce *script* (Figure 8) qui permet de créer la structure des documents de la base.

```
#-------------------------------------------------------------------------
#
#         Base EXEMPLE
#
#-------------------------------------------------------------------------

#---- AUTE ----

$(SERVER_ROOT)/EXEMPLE.AUTE.i.hcs: $(SERVER_ROOT)/EXEMPLE.bib.hcs     \
                        $(SERVER_PROG)/EXEMPLE.bib.sh      \
                        $(SERVER_PROG)/reject.tab
        IndexBuildUsual -k doc/AUTE/e# -i AUTE  -b $(SERVER_ROOT)/EXEMPLE -r
$(SERVER_PROG)/reject.tab

#---- DESC ----

$(SERVER_ROOT)/EXEMPLE.DESC.i.hcs: $(SERVER_ROOT)/EXEMPLE.bib.hcs     \
                        $(SERVER_PROG)/EXEMPLE.bib.sh      \
                        $(SERVER_PROG)/reject.tab
        IndexBuildUsual -k doc/DESC/e# -i DESC -b $(SERVER_ROOT)/EXEMPLE -r
$(SERVER_PROG)/reject.tab
```

Figure 8 : Structure pour compléter le fichier "IMD.index.mk"

- dans le fichier "IMD.desc.ed" : compléter le shell avec la structure (Figure 9) qui permet de créer la structure complète de la base.

```
<base code=EXEMPLE>
   <input type=command>bidon
   <title>doc/TITR#
    <author>doc/AUTE/e#
      <index code=AUTE>
        <path>doc/AUTE/e#
      </index>
        <index code=DESC>
          <path>doc/DESC/e#
      </index>
 </base>
```

Figure 9 : déclaration de la nouvelle base dans le fichier "IMD.desc.ed"

- dans le fichier "IMD.bib.mk" : compléter le programme suivant (Figure 10) qui permet le lancement du shell de génération de la base.

```
#-------------------------------------------------------------------------
#                    MAKEFILE de generation de la base sur EXEMPLE
#-------------------------------------------------------------------------

$(SERVER_ROOT)/EXEMPLE.bib.hcs:
/users/nedelcot/IMD/Text/EXEMPLE.BIB.1 \
                        $(SERVER_PROG)/EXEMPLE.bib.sh
        $(SERVER_PROG)/EXEMPLE.bib.sh
```

Figure 10 : Shell de lancement du programme de génération de la base

- dans le Répertoire "Server", dans le fichier "FR.dict" : compléter le shell avec une liste identique à celui des autres bases pour que l'affichage html des résultats soit possible.

Dans un premier temps, il est vrai que ces manipulations sont plus longues que celles proposées dans l'ancienne version de l'application, cependant, elles permettent de corriger plus facilement les erreurs introduites dans les bases. Ces manipulations seront améliorées.

b. Seconde modification

La seconde manipulation a consisté à étudier les notices au format AIRS de façon à supprimer la perte d'information dans les données résultats. Nous avions remarqué une nette différence entre le nombre de notices contenues dans certaines bases et le nombre de notices calculé par le programme de calcul statistique des résultats.

L'étude de la structure des notices nous a révélé que le format de certains champs n'était pas standard. Le programme "AirsToSgml" n'était pas tout à fait au point puisqu'il ne permettait pas la restitution de toutes les notices contenues initialement dans les bases de données. La fonction a donc été modifiée, de manière à gérer la longueur des champs et la présence de tabulations produisant des erreurs lors de l'exécution du programme "AirsToSgml".

Les résultats statistiques se sont révélés justes et le format AIRS était alors exploitable par l'application "IMD".

3. DILIB et son installation

L'architecture de la dernière version de DILIB (v0.21) est très différente des précédentes. Il a donc été nécessaire d'installer de façon intégrale la plate-forme documentaire. Pour cela, Jacques Ducloy a créé un site de déchargement de celle-ci facilitant l'installation (Annexe III) de toute l'architecture de DILIB dans sa version v0.21 ou la mise à jour (Annexe IV) pour les installations possédant déjà cette nouvelle architecture.

Ces deux opérations se trouvent aux adresses suivantes :
- http://portail.inist.fr/dilib/Operations/creMaster.fre.html
- http://portail.inist.fr/dilib/Operations/updateMaster.fre.html

DILIB a donc été installé à l'INRS. Les différentes étapes de l'installation sont décrites sur le site se rapportant à la création d'un site maître. Il faut signaler que le répertoire de base de l'application est $DILIB_BOTTOM=/doc/AIRS/Dilib. Par ailleurs, il ne faut pas oublier de modifier les fichiers d'initialisation des paramètres. Il s'agit des fichiers "bottom.def.sh" "init.sh", "init.test.sh", et "Test.def.sh". Dans le fichier "Test.def.sh", par exemple, il faut changer l'adresse des URL "Htbin" pour que les CGI soient localisés dans le fichier "cgi-dilibTest". Ce chemin d'accès au répertoire contenant les CGI doit être correctement déclaré dans les fichiers d'initialisation de DILIB

Un autre point délicat lors de l'installation de la plate-forme DILIB est le suivi effectif de tous les liens symboliques. Lors du déchargement de la plate-forme documentaire les droits (de lecture, d'écriture et d'exécution) des liens symboliques ne sont pas maintenus. Il faut donc vérifier et rétablir ces droits. Les fichiers correctement initialisés, il a fallu corriger le fichier de configuration du serveur "Apache".

4. Apache et son installation

Contrairement aux manipulations effectuées à l'INIST, il ne s'agissait pas de configurer un serveur mais d'actualiser les adresses de répertoires contenus dans les "Alias" et "ScriptAlias" nécessaires à DILIB. Une option a été ajoutée à l'Alias "dilib". En effet, les répertoires et les fichiers contenus dans le répertoire "WWW" de la nouvelle version de DILIB contiennent des liens symboliques. Il fallait donc que ceux-ci soient pris en compte lors de l'exécution des programmes. Nous avons donc ajouté l'option "FollowSymlinks" qui offre cette fonctionnalité.

```
ScriptAlias /cgi-dilibTest/ "doc/AIRS/Dilib/v0.21/Targets/Test/Dilib/htbin/"
<Directory "doc/AIRS/Dilib/v0.21/Targets/Test/Dilib/htbin/"
        AllowOverride None
        Options None
        Order allow,deny
        Allow from all
</Directory>
```

Figure 11 : ScriptAlias permettant l'accès aux données du répertoire "htbin"

L'accès aux données du répertoire "htbin" (Figure 11) peut se faire par le raccourci serveur "/cgi-dilibTest/". Elles peuvent aussi être visualisées à l'URL "http://www.inrs.fr/cgi-dilibTest"

```
Alias /IMD/ "/doc/AIRS/Dilib/Applications/IMD/"
<Directory "/doc/AIRS/Dilib/Applications/IMD/">
        Options Indexes multiviews
        AllowOverride None
        Order allow,deny
        Allow from all
</Directory>
```

Figure 12 : Alias rendant accessibles les documents du répertoire "IMD"

```
Alias /dilib/ "/doc/AIRS/Dilib/v0.21/WWW/"
<Directory "/doc/AIRS/Dilib/v0.21/WWW/"
        Options Indexes Follow Symlinks      (1)
        AllowOverride None
        Order allow,deny
        Allow from all
</Directory>
```

Figure 13 : Alias rendant accessible les documents du répertoire "WWW"
(1): option très importante pour que les liens symboliques fonctionnent

```
Alias /java-dilib/ "doc/AIRS/Dilib/v0.21/Targets/Test/Dilib/java/"
<Directory "doc/AIRS/Dilib/v0.21/Targets/Test/Dilib/java/">
        Options Indexes multiviews
        AllowOverride None
        Order allow,deny
        Allow from all
</Directory>
```

Figure 14 : Alias rendant accessibles les documents du répertoire "java"

Ces modifications (Figures 11 à 14) effectuées, il ne restait qu'à installer la structure des programmes permettant la gestion des données du centre de documentation. Seuls les programmes et la nouvelle structure de l'application IMD ont été installés à l'INRS. Ces quelques répertoires recopiés dans l'arborescence de DILIB, nous avons lancé la génération de l'application. Celle-ci s'est déroulée sans incident, et nous avons pu visualiser les résultats et apprécier la disparition des erreurs qui étaient générées auparavant.

IV. L'application "Génome" de l'INRS

1. L'origine de cette application

De par ses qualités d'isolants 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 montré le caractère cancérogène des fibres d'amiante et il a été prouvé que leur inhalation était responsable chaque année, en France, de nombreuses atteintes pulmonaires. Le mésothéliome est le cancer de l'enveloppe du poumon (plèvre) résultant, en général, d'une exposition professionnelle à l'amiante.

Afin de mieux comprendre la biologie du mésothéliome et de décrire le plus exhaustivement possible ses caractéristiques moléculaires, l'équipe de Bertrand Rihn (INRS) a entrepris l'étude des gènes impliqués dans le mésothéliome humain (cancer de la plèvre). C'est ainsi que le fonctionnement de plus de 7000 gènes ($1/7^{ème}$ du génome humain) a été analysé sur des cellules cancéreuses et normales de la plèvre. La technique des puces à ADN a été utilisée à cet effet. C'est une méthode de biologie moléculaire récente permettant de tester simultanément le fonctionnement de milliers de gènes dans les cellules. Les techniques antérieures n'autorisaient que des études très parcellaires de quelques gènes impliqués dans les cancers.

Les résultats de cette étude ont été rassemblés dans un tableau permettant de visualiser, entre autres, le niveau de fonctionnement de chaque gène dans les cellules normales et cancéreuses de la plèvre.

Le niveau de fonctionnement des gènes est appelé, en biologie moléculaire, expression* des gènes. Le rapport de l'expression d'un gène dans les cellules cancéreuses à celle dans les cellules normales est appelé l'expression différentielle ("Diff Expr" (Tableau I)).

$$\text{"Diff Expr"} = \frac{\text{Expression d'un gène donné dans les cellules cancéreuses de la plèvre}}{\text{Expression d'un gène donné dans les cellules normales de la plèvre}}$$

Ce rapport est >0 si l'expression d'un gène dans la cellule cancéreuse est supérieure à l'expression de ce même gène dans la cellule normale, et <0 dans le cas contraire. Un gène est considéré comme exprimé différentiellement si son " Diff Expr " (expression différentielle) est >+2 ou <-2. Dans notre cas, environ 400 gènes possédaient une expression différentielle >+2 et <-2.

A chaque gène exprimé différentiellement a été associé un article de la base de données "Genbank" grâce au numéro d'accession. Le numéro d'accession est une sorte de numéro d'identité qui est octroyé à chaque gène lors de sa découverte. Actuellement, "Genbank", la plus grandes bases de gènes, compte environ 6 000 000 de gènes individualisés dans 70 000 espèces différentes. Dans notre étude, le nombre de gènes exprimés différentiellement étant important, l'exploitation manuelle de ces données bibliographiques était difficile voire impossible. Ceci a motivé la création d'un outil permettant l'exploitation et la manipulation de ce volumineux corpus documentaire.

2. La méthodologie appliquée

Les résultats de l'étude de l'équipe de Bertrand Rihn sont rassemblés dans le tableau (disponible à l'adresse http://www.inrs.fr/actualités/amiante/genmesoen.htm) où apparaissent les caractéristiques de chaque gène exprimé.

Ces caractéristiques comprennent, entre autres, le niveau d'expression différentielle, le nom du gène et la référence de ce dernier dans la base de gènes Genbank (http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html).

Sur 1/7ème du génome, 200 gènes sont sur-exprimés (Exp Diff >+2) et 200 gènes sont sous-exprimés (Exp Diff <-2) dans les cellules cancéreuses par rapport aux cellules normales de la plèvre.

Tableau I : Les gènes et leur référence Genbank

Diff Expr (1)	*GeneName* (2)	***AccessionNum* (3)**
32,2	plasminogen activator inhibitor, type II (arginine-serpin) {Incyte PD:557012}	**M31551**
19,2	fatty acid binding protein 5 (psoriasis-associated) {Incyte PD:2537805}	**AA972250**
18,6	deiodinase, iodothyronine, type II {Incyte PD:2714959}	**AF007144**
11,9	serum-inducible kinase {Incyte PD:1255087}	**AF059617**
11,3	ESTs {Incyte PD:1397926}	**N35555**
10,6	Human EV12 protein gene {Incyte PD:511738}	**M55267**
9,1	annexin A1 {Incyte PD:79576}	**X05908**
8,1	ornithine decarboxylase 1 {Incyte PD:1930235}	**M81740**
7,3	pentaxin-related gene, rapidly induced by IL-1 beta {Incyte PD:1700077}	**M31166**
7,2	integrin, alpha 6 {Incyte PD:1518328}	**X53586**
6,3	deiodinase, iodothyronine, type II {Incyte PD:2869983}	**AF093774**
6,2	heat shock 105kD {Incyte PD:1922468}	**AB003334**
6,1	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor) {Incyte PD:2803366}	**X16983**
6	integrin, beta-like 1 (with EGF-like repeat domains) {Incyte PD:1258790}	**AF072752**
5,5	pregnancy specific beta-1-glycoprotein 1 {Incyte PD:64457}	**M20881**
5,2	ESTs {Incyte PD:1512478}	**AI744081**
5,1	ESTs, Weakly similar to STE20-like kinase 3 [H.sapiens] {Incyte PD:2793922}	**AA191319**
4,8	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase) {Incyte PD:2506867}	**AI928978**
4,7	ribonuclease L (2',5'-oligoisoadenylate synthetase-dependent) inhibitor {Incyte PD:1674405}	**X76388**
4,5	heat shock 60kD protein 1 (chaperonin) {Incyte PD:3132987}	**M34664**
4,5	glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2) {Incyte PD:661259}	**M22632**
4,3	interferon, gamma-inducible protein 16 {Incyte PD:2508261}	**S75433**

(1): Expression différentielle (2): nom du gène (3) : numéro d'accès dans la base Genbank

Ce tableau est une copie partielle simplifiée du fichier originel (2.9Mb). Pour faciliter la lecture des informations et l'évaluation de l'outil créé, nous avons décidé de ne travailler que sur les 200 gènes les plus exprimés et les 200 gènes les moins exprimés. Nous n'avons donc retenu que les numéros Genbank correspondant à ces gènes (en **gras** dans le Tableau I).

3. De Genbank à Medline

Une notice Genbank est présentée dans l'Annexe V. Nous n'avons gardé que le champ que nous utiliserons dans l'élaboration de cette application.

En effet, les notices Genbank ne nous sont utiles que pour accéder aux références Medline, donc seuls les champs "Accession", "Version" et "Medline" sont intéressants. Les deux premiers nous permettent de garder la trace des notices initiales et d'étendre la recherche à toutes les notices Genbank dont la référence est indiquée dans ces champs. Le second nous donne le numéro de la notice Medline que nous allons télécharger (Figure 15).

```
LOCUS       AW513437     535 bp    mRNA          EST      03-MAR-2000
ACCESSION   AW513437
VERSION     AW513437.1  GI:7151515
MEDLINE   97044478
```

Figure 15 : Champs conservés dans les notices Genbank

Des exemples de notices entières se trouvent en annexe. Toutes les notices sont sous ce format, mais certaines contiennent des références Medline (Annexe V) et d'autres non (Annexe VI). Nous avons tout d'abord décidé de ne travailler que sur les notices Medline pour l'exploitation des données bibliographiques. Dans un premier temps, tous les champs ne nous intéressaient pas. Nous avons conservé les principaux : le titre, les mots-clés (du *MEdical Subject Heading section* ou Mesh*) et les auteurs.

Voici une Notice Medline (Figure 16) contenant uniquement les champs initiaux conservés pour l'élaboration de l'application "Génome". Une notice Medline complète se trouve en Annexe I.

UI - 98225166
AU - Ludwig J
AU - Kerscher S
AU - Brandt U
AU - Pfeiffer K
AU - Getlawi F
AU - Apps DK
AU - Schagger H
TI - Identification and characterization of a novel 9.2-kDa membrane sector-associated protein of vacuolar proton-ATPase from chromaffin granules.
LA - Eng
MH - Amino Acid Sequence
MH - Animal
MH - Base Sequence
MH - Cattle
MH - Chromaffin Granules/*enzymology
MH - H(+)-Transporting ATP Synthase/*metabolism
MH - Human
MH - Membrane Proteins/chemistry/*metabolism
MH - Mice
MH - Molecular Sequence Data
MH - Sequence Homology, Nucleic Acid
MH - Support, Non-U.S. Gov't
SI - GENBANK/Y15285
SI - GENBANK/Y15286
SI - GENBANK/P81103
SI - GENBANK/P81134</u>

Figure 16 : Champs conservés dans les notices Medline

Dans un premier temps, nous avons donc collecté toutes les références Medline se trouvant dans les notices Genbank, en séparant bien les numéros de notices correspondant aux 200 gènes les plus exprimés de ceux correspondant aux 200 gènes les moins exprimés dans les cellules cancéreuses. En effet, il est important de pouvoir comparer les résultats obtenus pour les deux groupes de gènes. Nous avons obtenu une suite de numéros Medline représentée à la Figure 17, pour chacun des groupes de gènes.

M31551 or AA972250 or AF007144 or AF059617 or N35555 or M55267 or X05908 or M81740 or M31166 or
…….
X53586 or AF093774 or AB003334 or X16983 or AF072752 or M20881 or AI744081

Figure 17 : Liste de références Medline

Nous avons ensuite téléchargé de façon manuelle, à partir du site "Pubmed" (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=search&DB=PubMed), les notices Medline correspondant aux numéros de référence trouvés des notices Genbank.

Ce déchargement de notices pourra être optimisé, à l'avenir, de façon automatique, mais auparavant, il nous a semblé qu'il valait mieux nous concentrer sur l'exploitation des données que sur leur téléchargement.

Les notices sont stockées dans deux fichiers distincts : medline.txt (pour les 200 premiers gènes sur-exprimés) et medlineII.txt (pour les 200 derniers gènes sous-exprimés). Un programme en "Lex"* permet de sortir les données contenues dans tous les champs des notices Medline. C'est le début du passage du format "Medline" au format SGML qui sera ensuite utilisé pour travailler sur le corpus documentaire (Figure 18 et 19).

```
%START SKIP AUT MESH ABS
%%
^UI[ ]+\-[ ]+           { printf("<medline>\n<UI>");}
<MESH>^MH[ ]+\-[ ]+      ;
<AUT>^AU[ ]+\-[ ]+       ;
<AUT>\n/[^A][^U][^ ]     {ECHO; BEGIN 0;}
<MESH>\n/[^M][^H][^ ]    {ECHO; BEGIN 0;}
<ABS>[ ]*\n[ ]+          printf(" ");
<ABS>\n/[^A][^B][^ ]     {ECHO; BEGIN 0;}
^AU[ ]+\-[ ]+            { printf("<AU><e>\n"); BEGIN AUT;}
^MH[ ]+\-[ ]+            { printf("<DE><e>\n"); BEGIN MESH;}
^AB[ ]+\-[ ]+            { printf("<AB><s>\n"); BEGIN ABS;}
^[A-Za-z]+               { printf("<%s>",yytext);BEGIN SKIP;}
<SKIP>[ ]*\-[ ]*         BEGIN 0;
%%
main()
{
 yylex();
}
```

Figure 18 : Programme de séparation des champs de la notice Medline

Ce programme est utilisé dans le programme suivant :

```
#!/bin/sh

. $DILIB_CONFIG

SgmlCharSetTr     -f     iso8859
\
 | MacNewLine                \
 | MedLineWww1                 \
 | MedLineWww2                 \
 | MiniBibFromEd
```

Figure 19 : Programme de mise au format SGML

Ce dernier programme (Figure 19) nous donne un format SGML simplifié des notices, c'est à dire une architecture en arbre (Figure 20).

```
<medline>
<UI>97444286</UI>
<AU><e>Choudhury BK</e>
      <e>Kim J</e>
      ....
<TI>Cloning and developmental expression of Xenopus cDNAs encoding
the Enhancer of split groucho and related proteins.</TI>
<LA>Eng</LA>
<DE>  <e>Amino Acid Sequence</e>
      <e>Animal</e>
      ....
      <e>Xenopus&sol;growth &amp;
development&sol;*genetics</e></DE>
<RN>0 (enhancer of split protein)</RN>
...
<AB><s>.... <s></AB>
....
<PMID>0009300818</PMID>

<SI>GENBANK&sol;M20571</SI>
....
</medline>
```

Figure 20 : Arborescence d'une notice Medline

Nous pouvons donc faire ressortir les champs désirés et obtenir une notice au format SGML avec des champs prédéfinis (par exemple le titre aura la balise "tit"). Le programme suivant (Figure 21) permet de nommer de façon personnelle les différents champs que nous voulons utiliser.

```
. $DILIB_CONFIG

USAGE="usage: MedLineBaseUsual -b Base [-f inputFile] [-s] [-m maxAssoc]"

LIST_FILE=""
m_OPTION=1000

while getopts b:f:h:m:s c
do
  case $c in
   b) b_OPTION=$OPTARG;;
   h) LIST_FILE="$LIST_FILE $OPTARG";;
   f) LIST_FILE="$LIST_FILE $OPTARG";;
   s) LIST_FILE="$LIST_FILE -";;
   m) m_OPTION=$OPTARG;;
   \?) echo -$c $OPTARG unknowed option
       echo $USAGE
       exit 2;;
  esac
done

BASE_PATH=`DamGetPath -Arn $b_OPTION`

if test "$LIST_FILE"
then
   rm -rf $BASE_PATH.bib.*
   DamCat $LIST_FILE | SgmlSelect -p @0 | DamHfdBuild -h $BASE_PATH.bib
fi

IndexBuildUsual -k medline/DE/e# -i de -h $BASE_PATH
IndexBuildUsual -k medline/AU/e# -i au -h $BASE_PATH

AssocFastBuildUsual -i de -b $BASE_PATH -m $m_OPTION
AssocFastBuildUsual -i au -b $BASE_PATH -m $m_OPTION

ClusterBuildCos -b $BASE_PATH -i de -m $m_OPTION
ClusterBuildCos -b $BASE_PATH -i au -m $m_OPTION

ClusterBuildFreq -b $BASE_PATH -i de -m $m_OPTION
ClusterBuildFreq -b $BASE_PATH -i au -m $m_OPTION

DamCat $BASE_PATH.bib.hfd                          \
   | SgmlSelect -s medline/TI#  -p @1 -p '<tit>@s1</tit>'      \
   | MiniVocFromTit -b $BASE_PATH -t ti -e
```

Figure 21: Programme de gestion des champs et de leurs fichiers

Ce programme (Figure 21) nous permet de créer les fichiers index pour les champs auteurs, mots-clés, et titre. Le passage au format HTML des données se fait *via* la fonction "MedlineWwwUsual". Les fichiers initiaux de notices Medline ont donc subi différentes modifications jusqu'à la visualisation possible de ceux-ci sur Internet. Ces différents traitements sont résumés dans la suite de commandes décrite dans la Figure 22.

```
MedlineFromWww < medline.txt | MedlineBaseUsual -b Server/genome -s
MedlineWwwUsual -b Server/genome -s
```

Figure 22: Commandes permettant la génération de l'application

La première fonction (MedlineFromWww) nous donne un format SGML simplifié (séparation et balisage des champs) des notices Medline contenu dans le fichier "medline.txt". La seconde fonction (MedlineBaseUsual) permet la création des index, associations et clusters. A cette étape nous pouvons personnaliser les intitulés des différents champs. La troisième fonction (MedlineWwwUsual) construit les fichiers nécessaires à la visualisation des résultats par une navigateur.

4. Les bases, leurs évolutions et le résultat

Nous avons donc obtenu un certain nombre de données telles que des fichiers inverses, sur les auteurs, les titres et les mots-clés. A partir de ceux-ci, nous avons créé des bases qui regroupaient tous les fichiers et permettaient l'exploitation des résultats. Ceux-ci ont été présentés à Bertrand Rihn afin d'avoir un avis sur leur pertinence. Il s'agissait de savoir si, par exemple, les résultats des associations pouvaient mettre en évidence de nouvelles co-apparitions qui n'avaient pas encore été relevées par l'étude bibliographique manuelle.

Il s'est avéré qu'un certain nombre de mots-clés du Mesh (mots-clés Medline) était trop généraliste et non pertinent. Nous avons donc établi la liste de tous les mots contenus dans les fichiers "mots-clés" et "mots du titre". Les mots les plus récurrents ont été supprimés par Alain Zasadzinski, ingénieur documentaliste de l'INIST et spécialiste en microbiologie et biologie moléculaire afin de n'en garder qu'un échantillon pertinent. Le tri de ces mots-clés ayant été validé par Bertrand Rihn, nous avons pu créer un fichier de mots vides que nous avons inclus dans les programmes de génération de l'application "Génome". En effet, il existe une option dans la fonction "IndexBaseUsual" permettant d'intégrer un fichier de mots (non pertinents ou mots vides) qui seront supprimés des résultats. A ce stade de la construction de l'application "Génome", nous avions donc quatre bases de données. Les deux premières correspondent aux 200 premiers gènes sur-exprimés avec et sans le tri des mots-clés (un total de 1758 mots-clés qui seront au nombre de 1131 après le tri). Les deux dernières correspondent aux 200 derniers gènes sous-exprimés avec et sans le tri des mots-clés clés (un total de 1124 mots-clés qui seront au nombre de 854 après le tri). Ces bases et les fonctionnalités (associations, clusters…) qui les accompagnent ont permis à Bertrand Rihn de découvrir, par exemple, des associations de termes induisant des relations jusque là inédites entre certains gènes ou certaines protéines.

Un exemple pertinent de l'analyse lexicale possible d'un terme est présenté dans les Tableau II et III. Le terme "Antigen CD" (pour "Antigens Cell Differentiation") apparaît 10 fois dans le cluster lexical des cellules cancéreuses et 7 fois dans celui des cellules normales (Tableau II). L'analyse bibliométrique sommaire montre que, pour cet exemple du moins, la fonction de différenciation cellulaire des antigènes de surface parait être plus importante dans les cellules cancéreuses.

Après avoir affiné l'analyse et supprimé les termes trop généraux (Tableau III), on s'aperçoit que les cellules cancéreuses de la plèvre ont un rapport avec les antigènes CD 34, 36 et 95, alors que les cellules normales de la plèvre ont un rapport avec les antigènes de surface 4 et 59. L'antigène CD 55 est quant à lui, cité dans les lexiques des deux types cellulaires. Bien sur, cette première analyse sémantique n'a pas, en soi, de valeur biologique.

Par contre, les liens entre les mots-clés et les documents étant conservés, il est possible de revenir aux publications sources et d'analyser plus en détail la signification biologique de cette assertion. Lorsque les filtres sémantiques sont déterminés, DILIB permet au chercheur en biologie, de ne se concentrer que sur les publications qui paraissent les plus importantes dans les volumineux corpus documentaires possibles.

L'étape suivante de ce travail consiste a ne garder que les termes "a haute valeur biologique ajoutée" et de constituer un thésaurus propre a chaque type cellulaire et un thésaurus commun aux 2 types cellulaires.

Il sera possible de définir quelles fonctions cellulaires, quelles voies métaboliques, quels gènes ou protéines cellulaires peuvent être considérés comme spécifiques de chaque état, normal et pathologique des cellules de la plèvre.

Tableau II : Analyse primitive des occurrences " Antigens CD"

Item informatif du MESH	Cellule cancéreuse	Cellule normale
Antigens, CD, genetics	x	
Antigens, CD, genetics, metabolism	x	
Antigens, CD, metabolism		x
Antigens, CD, chemistry, genetics, metabolism	x	
Antigens, CD, genetics, physiology	x	
Antigens, CD, metabolism		x
Antigens, CD34	x	
Antigens, CD34, analysis	x	
Antigens, CD36, genetics	x	
Antigens, CD4, genetics		x
Antigens, CD55, genetics, immunology	x	
Antigens, CD55, metabolism		x
Antigens, CD59, metabolism		x
Antigens, CD95, metabolism	x	
Antigens, Differentiation, chemistry, genetics		x
Antigens, Surface, biosynthesis, genetics, physiology		x
Antigens, Surface, genetics, metabolism	x	

Tableau III : Analyse complémentaire de l'occurrence " Antigens "

Item informatif	Cellule cancéreuse	Cellule normale
Antigens, CD34	x	
Antigens, CD34	x	
Antigens, CD36	x	
Antigens, CD4		x
Antigens, CD55	x	
Antigens, CD55		x
Antigens, CD59		x
Antigens, CD95	x	

En sachant que l'étude ne s'est faite que sur 1/7è du génome humain, quelles perspectives s'offrent aux chercheurs lorsque les 50 000 gènes, soit l'intégralité du génome humain seront évalués? On peut s'attendre, dans ce cas, à ce qu'environ 3 000 gènes soient activés (exprimés) différentiellement dans les deux populations cellulaires : normales et cancéreuses. Ces 3 000 gènes pourront être associés à environ 20 000 références bibliographiques qu'il devrait être possible d'analyser avec les outils de DILIB que nous avons développés.

Une application telle que celle-ci est un outil intéressant dans la mesure où elle permet l'exploitation d'un volumineux corpus de données non réalisable manuellement. Son intérêt est d'autant plus grand qu'il permet d'obtenir des informations supplémentaires sur les relations entre certains gènes.

5. Les améliorations à venir

Grâce à l'application "Génome", des informations nouvelles et ayant une pertinence biologique ont pu être mises en évidence. Or, il ne s'agissait, dans un premier temps, que d'un travail sur les mots-clés et les mots du titre. L'objectif est d'améliorer les performances de cet outil en développant l'application vers l'exploitation du résumé, les mises à disposition de liens entre les mots-clés et les auteurs. La comparaison des résultats portant sur les 200 gènes sur-exprimés et les 200 gènes sous-exprimés pourra constituer une autre source d'informations non négligeable pour le biologiste spécialiste du génome. Les fichiers de mots-clés communs ou propres aux deux catégories de gènes peuvent être mis en parallèle et comparés. L'introduction de thésaurus est une autre extension imaginable de l'application.

Conclusion

La plate-forme de gestion d'information DILIB est un outil encore en plein essor. Ce stage m'a permis de comprendre comment un tel produit était conçu et comment il pouvait être utilisé lors de la création d'outils infométriques permettant la gestion de volumes importants de données. Les connaissances informatiques nécessaires à l'utilisation d'une telle plate-forme ne sont pas négligeables. Il m'a fallu découvrir et apprendre à me servir correctement de langage tels que le "shell" ou le "C-shell" pour développer et corriger l'application "IMD" du centre de documentation de l'INRS avant de débuter la construction de l'application "Génome".

La version de DILIB installée à l'INRS est la dernière version actuellement disponible de la plate-forme documentaire. Son architecture "standard" offre une facilité de mise à jour qui permettra aux personnes du Départements Produits et Service de l'INIST chargées de la maintenance de l'application "IMD" de pouvoir fournir un suivi plus aisé et une amélioration des fonctionnalités de l'application.

L'élaboration de tels produits m'a permis de réaliser l'importance de la gestion de l'information quelque soit le domaine d'application. En effet, si DILIB permet d'accéder de façon aisée à des bases de données diverses et variées, cette plate-forme nous a montré qu'elle était capable de donner des résultats intéressants et prometteurs lors de la manipulation de données telles que celles obtenues en biologie sur l'exploitation du génome humain. Cette gestion des données pourra, par ailleurs, être appliquée à n'importe quel domaine de recherche ce qui ouvre des perspectives intéressantes d'installation et de développement pour DILIB.

BIBLIOGRAPHIE

Apache "Apache". URL http://www.apache.com

DA POZZO P. "Aide mémoire de Unix". Edition Marabout.

DUCLOY J. "Plate-forme et boîte à outils DILIB". URL http://www.loria.fr/projets/dilib/#doc

INRS. "INRS". URL http://www.inrs.fr

LAURIE B. &LAURIE P. "Apache, Installation et mise en œuvre". Edition O'REILLY révisée 1999.

MIDON L. "Installation et amélioration de la plate-forme documentaire DILIB-0.2". Rapport de stage de DESS Information Scientifique et Technique (Nancy1). 1999.

MOHR S. "Etude des transcriptions de cellules humaines mésothéliales et de mésothéliome". DEA de Biologie moléculaire et Cellulaire, Strasbourg, 2000.

PELISSIER C. "UNIX. Utilisation. Administration. Réseau Internet". Edition HERMES.

PUCHET D. "Mémoire de stage". Rapport de stage de DESS Informatique (Nancy1). 1996.

RIHN B., MOHR S.,McDOWEL S.A., BINET S., LOUBINOUX J., GALATEAU G., LEIKAUF K. and G.D. "Differential gene expression in mesothelioma". FEBS Letters , 480 (sous presse) 2000.

Glossaire

- AIRS Web : 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.

- Base de données AIRS : Base de données gérées par le logiciel AIRS qui est un système de gestion de bases de données documentaires.

- Cluster : Classe de mots ayant des associations fortes.

- DTD : C'est l'ensemble des règles, déterminées par une application, qui appliquent SGML au balisage des documents d'un type particulier. Elle définit en cela le vocabulaire et la grammaire du balisage.

- Expression des gènes : Les gènes s'allument ou s'éteignent (sont activés ou inhibés) selon un schéma génétique préétabli. C'est ce programme qui permet, par exemple, à un globule rouge de produire de l'hémoglobine.
  Un sous-ensemble activé du génome est responsable de l'aspect et de la fonction caractéristique des 200 types cellulaires différents qui composent le corps humain. C'est aussi la cause et/ou la conséquence de la transformation maligne des cellules.
  C'est cette même expression des gènes qui gouverne la naissance, le devenir et la mort des cellules.

- Fichiers inverses : Fichier d'indexation facilitant la gestion des relations entre les objets.

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

- Format AIRS : Format des notices contenues dans les bases de données gérées par le logiciel AIRS.

- Génome humain : Ensemble des gènes (environ 50 000) localisés sur les 23 paires de chromosomes de l'homme.

- Html : 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.

- Lex : Langage permettant la génération de programme pour l'analyse lexicale de texte.

- Mesh : Thésaurus de référence dans le domaine biomédicale. C'est le thésaurus de la base Medline.

- 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.

- Navigateur (ou *browser*) : Logiciel de visualisation qui permet de consulter des pages au format HTML sur Internet ou pour un intranet.

- Opérateurs booléens : Opérateurs permettant de combiner ou d'isoler des mots-clés (ET, OU, SAUF).

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

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

- 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.

- Protocole http : Permet à un serveur de communiquer avec un ou plusieurs clients sous la forme de requête et de réponse.

- 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.

Abréviations

CALS : *Computer-aided Acquisition and Logistic Support*
CE : Certification Européenne
CGI : *Common Gateway Interface*
CIRIL : Centre Inter-universitaire de Ressources en Informatique de Lorraine
CNAM : Caisse Nationale d'Assurance Maladie
CRAM : Caisse Régionale d'Assurance Maladie
CSIL : Centre de Services Informatiques de Lorraine

DESS : Diplôme d'Etude Supérieur Spécialisé
DILIB : *Documentation Information LIBrary*
DOD : *Departement Of Defense*
DPS : Département Produits et Services
DTD : Définition de type de document

HFD : *Hierarchical File organisation for Documentation*
HTML : *HyperText Markup Language*
HTTP : *HyperText Transfer Protocol*

INIST : INstitut de l'Information Scientifique et Technique
INRS : Institut National de Recherche et de Sécurité

LORIA : Laboratoire Lorrain de Recherches en Informatique et ses Applications

MESH : *Medical Subject Heading Section*

RENATER : RÉseau NAtional de télécommunications pour la Technologie, l'Enseignement et la Recherche.

SGML : *Standard Generalized Markup Language*
SIP: Service Ingénierie et Partenariats

TCP/IP : *Transfert Control Protocol / Internet Protocol*

URL : *Uniform Resource Locator*

WWW : *World Wide Web*

Les figures et tableaux

ANNEXES

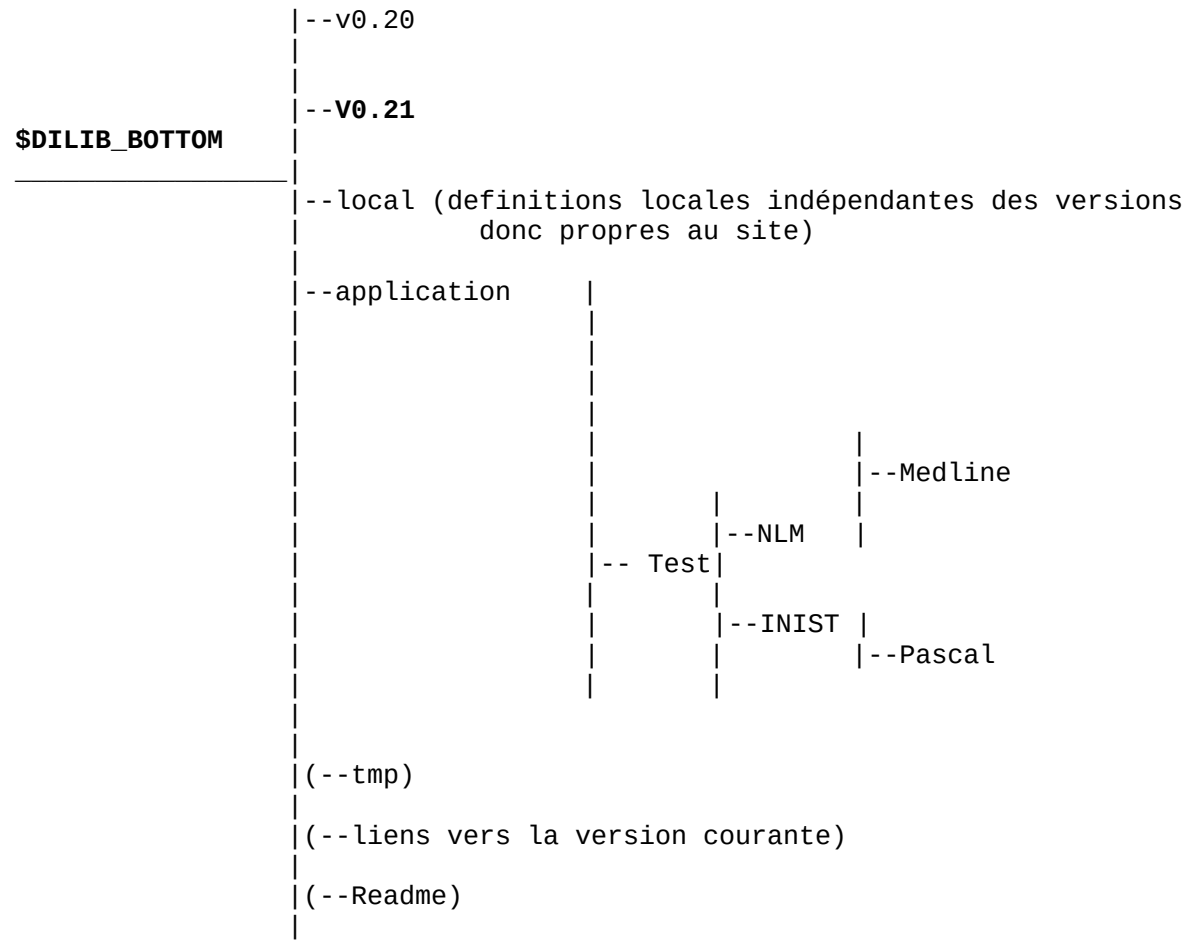
ANNEXE I
Exemples de structure de Notices Medline

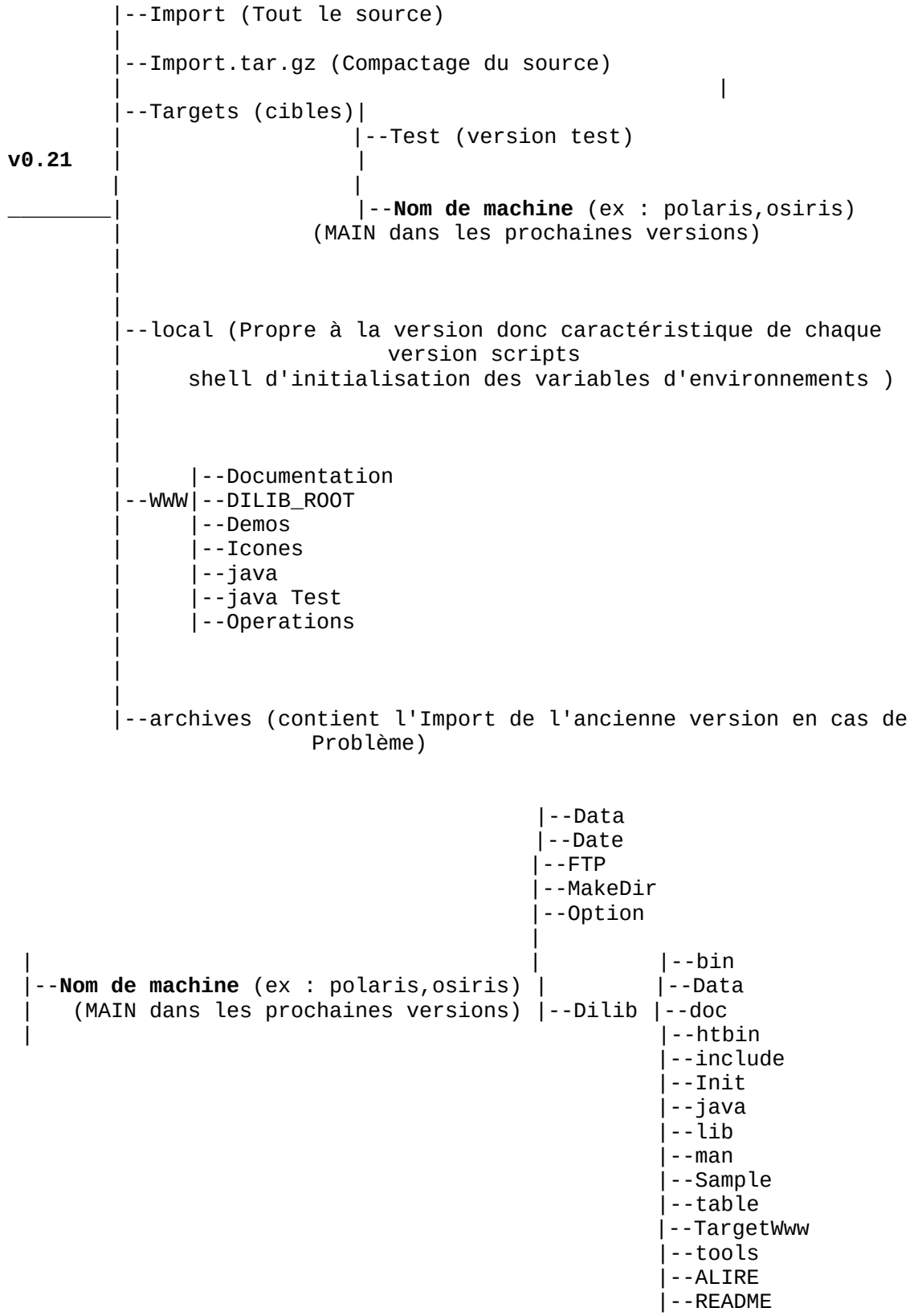
UI - 99115052
AU - Deniziak M
AU - Mirande M
AU - Barciszewski J
TI - Cloning and sequencing of cDNA encoding the rice methionyl-tRNA synthetase.
LA - Eng
MH - Amino Acid Sequence
MH - Cloning, Molecular
MH - DNA, Complementary
MH - Human
MH - Methionine-tRNA Ligase/chemistry/*genetics
MH - Molecular Sequence Data
MH - Rice/*enzymology
MH - Sequence Homology, Amino Acid
MH - Support, Non-U.S. Gov't
RN - EC 6.1.1.10 (Methionine-tRNA Ligase)
RN - 0 (DNA, Complementary)
PT - JOURNAL ARTICLE
DA - 19990324
DP - 1998
IS - 0001-527X
TA - Acta Biochim Pol
PG - 669-76
SB - M
CY - POLAND
IP - 3
VI - 45
JC - 0B4
AA - Author
EM - 199905
AB - Three overlapping clones of cDNA, Mos43, Mos28 and Mos60, coding for methionyl-tRNA synthetase were obtained by screening the Oryza sativa lambda gt11 library. Their nucleotide sequence of 2850 bp was determined. The deduced amino-acid sequence of the isolated clones contains a HLGN and KFSKS motifs, which are conserved for this family of enzymes and have been proposed to be the signature sequences for class I aminoacyl-tRNA synthetases. A comparison of the rice MetRS primary structure with those deposited in EMBL/GenBank points to its high homology to yeast, human and Caenorhabditis elegans MetRSs. Interestingly, a great similarity of its C terminus to endothelial-monocyte-activating polypeptide II (EMAPII) and yeast protein G4p1 was observed.
AD - Institute of Bioorganic Chemistry of the Polish Academy of Sciences, Poznan.
PMID- 0009918493
SI - GENBANK/AF040700
SI - GENBANK/Z73427
SI - GENBANK/V01316
SI - GENBANK/X14629
SI - GENBANK/AB004537
SI - GENBANK/Z98978
SI - GENBANK/X94754
SI - GENBANK/U67567
SI - GENBANK/AE000841
SI - GENBANK/AE001003
SI - GENBANK/U32807
SI - GENBANK/AE001160
SI - GENBANK/X57925

SI - GENBANK/D26185
SI - GENBANK/M64273
SI - GENBANK/AE000557
SI - GENBANK/U76417
SI - GENBANK/D64002
SI - GENBANK/AE000015
SI - GENBANK/Z94752
SI - GENBANK/U39680
SI - GENBANK/Y13943
SI - GENBANK/U31348
SI - GENBANK/U10117
EDAT- 1999/01/26 03:01
MHDA- 1999/01/26 03:01
SO - Acta Biochim Pol 1998;45(3):669-76

ANNEXE II
Architecture DILIB

sur polaris et Osiris : DILIB_BOTTOM =/applis/dilib




|--Import (Tout le source)
|--Import.tar.gz (Compactage du source)
|--Targets (cibles)|
|--Test (version test)
v0.21
|--Nom de machine (ex : polaris,osiris)
(MAIN dans les prochaines versions)
|--local (Propre à la version donc caractéristique de chaque
version scripts
shell d'initialisation des variables d'environnements)
|--Documentation
|--WWW|--DILIB_ROOT
|--Demos
|--Icones
|--java
|--java Test
|--Operations
|--archives (contient l'Import de l'ancienne version en cas de
Problème)
|--Data
|--Date
|--FTP
|--MakeDir
|--Option
|--Nom de machine (ex : polaris,osiris)
(MAIN dans les prochaines versions)
|--Dilib
|--bin
|--Data
|--doc
|--htbin
|--include
|--Init
|--java
|--lib
|--man
|--Sample
|--table
|--TargetWww
|--tools
|--ALIRE
|--README

ANNEXES III
Création d'un site maître

ANNEXES IV
Mise à jour d'un site maître

ANNEXE V
Exemples de structure de Notices Genbank

Une notice peut avoir **une référence Medline**, c'est le cas de la notice suivante:

```
LOCUS       T28987       378 bp    mRNA           EST      06-SEP-1995
   DEFINITION  EST63362 Human White blood cells Homo sapiens cDNA 5' end similar
               to B-myb (GB:X13293) (HT:1470), mRNA sequence.
   ACCESSION   T28987
   VERSION     T28987.1  GI:611085
   KEYWORDS    EST.
   SOURCE      human.
     ORGANISM  Homo sapiens
               Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
               Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
   REFERENCE   1  (bases 1 to 378)
     AUTHORS   Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A.,
               Bult,C.J., Lee,N., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D.,
               White,O., Sutton,G., Blake,J.A., Brandon,R.C., Chiu,M.-W.,
               Clayton,R.A., Cline,R.T., Cotton,M.D., Earle-Hughes,J., Fine,L.D.,
               FitzGerald,L.M., FitzHugh,W.M., Fritchman,J.L., Geoghagen,N.S.M.,
               Glodek,A., Gnehm,C.L., Hanna,M.C., Hedblom,E., HinkleJr,P.S.,
               Kelley,J.M., Klimek,K.M., Kelley,J.C., Liu,L.-I., Marmaros,S.M.,
               Merrick,J.M., Moreno-Palanques,R.F., McDonald,L.A., Nguyen,D.T.,
               Pellegrino,S.M., Phillips,C.A., Ryder,S.E., Scott,J.L.,
     TITLE     Initial Assessment of Human Gene Diversity and Expression Patterns
               Based Upon 83 Million Basepairs of cDNA Sequence
     JOURNAL   Nature 377, 3-174 (1995)
MEDLINE   96026280
   COMMENT     Other_ESTs: THC23534
               Contact: Venter, JC
               The Institute for Genomic Research
               932 Clopper Rd, Gaithersburg, MD 20878
               Tel: 3018699056
               Fax: 3018699423
               Email: tdbinfo@tdb.tigr.org
               For clone availability, additional sequence and expression
               information related to this EST, please contact the TIGR Database
               (tdbinfo@tdb.tigr.org)
               Seq primer: M13 Reverse.
   FEATURES             Location/Qualifiers
      source          1..378
                      /organism="Homo sapiens"
                      /db_xref="ATCC (inhost):104862"
                      /db_xref="taxon:9606"
                      /tissue_type="white blood cells"
                      /clone_lib="Human White blood cells"
                      /note="Organ: blood"
   BASE COUNT       68 a    116 c    114 g     78 t      2 others
   ORIGIN
         1 ccacttcacg acacctgccc ctatgtccag tgcctggaag acggtggcct gcggggggac
        61 cagggaccag cttttcatgc aggagaaagc ccggcagctc ctgggccgcc tgaagcccag
       121 ccacacatct cggaccctca tcttgtcctg aggtgttgag ggtgtcacga gcccattctc
       181 atgtttacag gggttgtggg ggcagagggg gtctgtgaat ctnagagtca ttcaggtgac
       241 ctcctgcagg gagccttctg ccaccagccc ctccccagac tctcaggtgg aggcaacagg
       301 gccatgtgct gccctgttgc cgagcccagn tgtgggcggc tcctggtgct aacaacaaag
       361 tttccacttc caggtctg
```

ANNEXE VI
Exemples de structure de Notices Genbank

Une Notice peut ne contenir **aucune référence Medline**, c'est le cas de la notice qui suit:

```
LOCUS       AW513437      535 bp    mRNA            EST       03-MAR-2000
DEFINITION  xo44h05.x1 NCI_CGAP_Ut1 Homo sapiens cDNA clone IMAGE:2706873 3'
            similar to gb:L06505 60S RIBOSOMAL PROTEIN L12 (HUMAN);, mRNA
            sequence.
ACCESSION   AW513437
VERSION     AW513437.1  GI:7151515
KEYWORDS    EST.
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 535)
  AUTHORS   NCI-CGAP http://www.ncbi.nlm.nih.gov/ncicgap.
  TITLE     National Cancer Institute, Cancer Genome Anatomy Project (CGAP),
            Tumor Gene Index
  JOURNAL   Unpublished
COMMENT     Contact: Robert Strausberg, Ph.D.
            Tel: (301) 496-1550
            Email: Robert_Strausberg@nih.gov
            Tissue Procurement: Christopher Moskaluk, M.D., Ph.D., Michael R.
            Emmert-Buck, M.D., Ph.D.
             cDNA Library Preparation: Life Technologies, Inc.
             cDNA Library Arrayed by: Greg Lennon, Ph.D.
             DNA Sequencing by: Washington University Genome Sequencing Center
             Clone distribution: NCI-CGAP clone distribution information can be
            found through the I.M.A.G.E. Consortium/LLNL at:
            image.llnl.gov/image/html/iresources.shtml

            Seq primer: -40UP from Gibco
            High quality sequence stop: 402.
FEATURES             Location/Qualifiers
     source          1..535
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /clone="IMAGE:2706873"
                     /tissue_type="well-differentiated endometrial
                     adenocarcinoma, 7 pooled tumors"
                     /clone_lib="NCI_CGAP_Ut1"
                     /lab_host="DH10B"
                     /note="Organ: uterus; Vector: pCMV-SPORT6; Site_1: SalI;
                     Site_2: NotI; Cloned unidirectionally.  Primer: Oligo dT.
                     Average insert size 1.75 kb. Life Technologies catalog #:
                     11538-014"
BASE COUNT      111 a    134 c    131 g    158 t      1 others
ORIGIN
        1 gtcaaatgat cctttattga aatgttttcc tttgtgctta actggctggg cattccacag
       61 caccactgtt gatgtcatcg atgatgtcat gaggatggcg gccatcaaca ttacagccca
      121 ctgactgggc agtccccagg atctctttaa tggttccaga gagttctctg gctaaggatc
      181 ggtgccgcat ctgtcgagca atgttgacaa tctcatcaaa agtgatattc ccactgtgtt
      241 taatgttttt ctgtttcttt ctgtctcttg gtggttcctt gagggctttg atgatcaggg
      301 cagaggcaga aggcaccacc tcaatctggg cctgtctgtt ctgaatggtc agtttcactg
      361 taatcctcag gcccttccag tcacccgttg ccttggcaat gtcatcacca actttttttg
      421 gagacagacc caggggggccg atcttggggg ccagagcaga agtggcaccg acttcaccta
      481 ccgtgcacct caagtatacg actttgatct cgttgggtcn aacttcggcg gcatg
```

Résumé

L'Institut National de Recherche et de Sécurité dispose d'un important fonds documentaire qui doit être exploitable et consultable *via* son intranet. Pour que cette gestion soit possible, le service de documentation de Vandoeuvre-lès-Nancy, a choisi d'utiliser la plate-forme documentaire *Documentation Information LIBrary* (DILIB). Celle-ci est constituée d'un ensemble d'outils informatiques permettant la manipulation de bases de données et la visualisation de celles-ci sur une interface web.

DILIB est un outil de gestion d'information expérimental qui évolue et acquiert de nouvelles fonctionnalités d'année en année. Celles-ci peuvent être utilisées dans divers domaines puisqu'elles offrent le moyen de faire de l'infométrie et de la gestion de l'information sur des fonds documentaires très importants.

Un exemple concret de l'utilisation de cette plate-forme est la gestion et l'exploitation des données issues des résultats de recherche de l'équipe du Docteur B. Rihn (INRS) dans le cadre de l'étude de l'expression des gènes impliqués dans le cancer de la plèvre chez l'homme.

Mots-clés

DESS IST, DILIB, expression des gènes, génome, infométrie, Information Scientifique et Technique, ingénierie de l'information, INIST, INRS, format, gestion de bases de données, mésothéliome, plate-forme documentaire, rapport de stage, SGML.