Mining medical data using multiple corpora interaction:the Transcriptomics investigation server experiment

Solveig VIDAL, Jacques DUCLOY and Philippe HOUDRY
INIST CNRS 2 allée du Parc de Brabois
54514 Vandoeuvre-lès Nancy, France

ABSTRACT

Exponential increasing of studies (which deal with **DNA**, **RNA** and proteic sequences analysis), owing to recent improvements in molecular biology, modifies researchers needs. Currently, the principal challenge corresponds to global analysis in order to extract pertinent and useful information in biology from various data.
New version of DILIB platform proposes interesting functionalities for information analysis and visualization. This aspect interested biologist Bertrand Rihn who wished to facilitate exploitation of bibliographic data linked to genes study implicated in human pleural cancer.
This collaboration gave creation to "Transcriptomics investigation server".
The characteristics of this server are as following: automatic extraction of Medline references from experimental results, integration of UMLS metathesaurus extract, creation of semantic link and addition of four corpora from a scientific database (called Pascal) hosted in INIST.
The first section shows basic foundation of an investigation server with DILIB platform introduction.
The second section develops UMLS metathesaurus extract integration in this server.
The third section deals with semantic link building and crossed navigation between bibliographic corpora from different databases (Medline and Pascal). The last section shows how Transcriptomics server has been linked to factual data from bioinformatic database.
The conclusions of this work are finally exposed.

Keywords : Transcriptomics, Investigation Server, DILIB, Mesothelioma, UMLS metathesaurus, Datamining.

INTRODUCTION

As post-genomic era generates massive biological data flow located both in factual and bibliographic databases, the main challenge for researchers is to get pertinent information.
In this paper, we describe a prototype tool which is able to analyse bibliographic data using infometric statistics. This tool, called documentary investigation server, is generated thanks to a Documentation and Information LIBrary (DILIB) platform, developped at INIST (Institut de l'Information Scientifique et Technique – scientific and technic information institute) cooperating with LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications – computer science research laboratory).
A comparative study of expression level from more than 7,000 genes (transcriptomics and expressed tags (**ESTs**)) in cancerous and healthy pleural cells has been launched by a cancer research laboratory, directed by Bertrand Rihn, from INRS (Institut National de Recherche et Securité – national reseach and safety institute). This type of pleural cancer, called mesothelioma, is mainly provocated by fibre amiante exposition. Both INIST and INRS institutes collaborated in order to facilitate exploitation of bibliographic data linked to this study. An investigation server dedicated to this study, called Transcriptomics, was generated and offers functionalities such as infometric analysis using hypertextual browsing.
We will describe first the technical base of an investigation server describing DILIB workbench, and show particularities of Transcriptomics server i.e. new features added in this server. Then we will introduce UMLS metathesaurus and show how an extract of this metathesaurus is integrated in the server .
We will also describe how crossed navigation was allowed between two types of corpora, from Medline [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed] and from Pascal database (provided by INIST) and show how four Pascal corpora were extracted before beeing implemented in Transcriptomics server.
Finally, the fourth section concerns the last evolution of this server i.e. linking Transcriptomics server to factual bioinformatic data.
The conclusions of this work are finally exposed.

1. INVESTIGATION SERVER AND DILIB PLATFORM

1.1. Technical base of an investigation server : DILIB platform

DILIB (Document and Information LIBrary) is a workbench developped by INIST/ CNRS and INRIA. It works under the Unix operating system and is composed of two main kinds of tools : first, functions for converting heterogeneous data into SGML/XML trees and for handling such trees or records, and, second, for building information retrieval systems like investigation servers.
Records are stored in Hierarchical File organization for Documentation (HFD repository). This DILIB platform's method allows the storage of up to 10^6 records in basic configuration of the HFD. Other items such as inverted files, whose records are also coded into SGML/XML form, improve easy access of related records (namely records with similar keys (e.g., keywords or authors)) as it can be handled by DILIB's commands.

1.1.1. Homogenisation of information, SGML/XML.

Sgml/Xml markup languages are used, the most often, to unify description of textual materials and can code three main classes of information : content itself, bibliographical description and all the internal data such as inverted files.
The first set of tools is used to help converting heterogeneous data into standardized format with tagged fields. Two main levels (character set and data structure) are taken into account. As XML derived from SGML standard (ISO-8879 norm) is used in DILIB plateform, character such as french one "à" is represented by the string "à", so

DILIB provides commands which allows such conversions. Concerning data structure level, the approach is to keep a DTD close to original bibliographic format. For instance, each textual field become a tagged field. Records coming from Medline or from formats based on ISO2709 (UNIMARC, USMARC, Inist standard), which are organized with fields and subfields, can then be converted.

1.1.2. Clusterization and information retrieval tools. Another part of DILIB workbench consists in building information retrieval systems such as"Investigation Servers". Basic access mechanisms have been defined and are very close to Unix file system. On this basis, other items have been defined such as inverted files whose records are also coded into SGML form, thus all the internal data such as inverted files can be handled by DILIB. This advantage has been used to develop most of the infometric modules. From an inverted file, an association file can be built first which look like this :

```
<assoc><ti><kw>usmarc</kw><f>223</f></ti>
        <tj><kw>SGML</kw><f>300</f></tj>
        <fij>15</fij><list><e>000035</e>…
…</list></assoc>
```

where “fij“ is the frequency of co-occurences of the keywords “usmarc” and “SGML”. From this association file, a cluster file may be built and is also a set of SGML documents, one per cluster. Each cluster is mainly made up of a group of words with their inter-relations. A cluster constitutes a group of “internal associations”, associations describing relations existing between descriptors inside the cluster, which represents a scientific theme. Other associations may link several clusters and are called “External associations”, on the opposite of “Internal association”.
The size of each cluster is limited to the group with the most significant associations which were used to build it.

1.1.3. WWW interface. One of Dilib's tool allows Web interface for investigation servers corresponds to a strategic choice of DILIB's developpers in order to explore bibliographical information with a more intuitive and especially the most used way in Internet. As a result, data analysis is facilitated. This interface contains classical static HTML pages but also dynamic ones with cgi programs. Infometrics analysis are called with such .cgi programs.
There are two main access from the Home page of an investigation server : "browsing access" with infometrics analysis, vizualization of articles, clusters (of keywords and of authors), and "selection access" which leads via a word's query to articles. Infometrics analysis are based on association files and, as a result, on clusters. Java applets are then automatically built in order to obtain a graphic representation of each co-occurrence of keywords (in association file) or to have an idea of each clusters' size in a corpus.

1.2. Particularities of Transcriptomics server

1.2.1. Automatism of bibliographic data retrieval thanks to experimental results. Compared expression level of approximatively 7,000 genes was conducted in healthy pleural cells versus cancerous cells (or mesothelioma cells) and 4% differentially expressed genes lured Bertrand Rihn's interest. As shown in **figure 1**, 242 genes (over 7,000) were overexpressed in mesothelioma cells than in healthy ones (called here "overexpressed genes") and 257 were underexpressed (named here "underexpressed genes"). As each gene is referenced in GenBank database [www.ncbi.nlm.nih.gov/Genbank/] by its Accession Number, corresponding also to "Second Identifiers" [SI] from Medline database, Medline records were retrieved with queries, which are lists of Accession Numbers, in Pubmed [www.ncbi.nlm.nih.gov/entrez/query.fcgi].
However, only 200 and 202 Medline records were obtained; they corresponded to overexpressed genes and underexpressed genes in mesothelioma cells. These differences were due to genes which do not belong any Medline records.
Two bibliographic corpora corresponding to overexpressed genes and underexpressed genes were loaded in two databases in Transcriptomics server and following Medline fields were kept in theses corpora [AB, AD, AU, ID, MH, RN, SI, TA, TI].
Then, four indexes were achieved : MH (MeSH descriptors), TI (title descriptors), ABS (abstract descriptors), RN (registry number) assigned by the Enzyme Commission to designate a particular enzyme or by the Chemical Abstracts Service (CAS), namely MH, TI, AB and RN for each retrieved Medline records. All MeSH words were retrieved but not kept, as a rejected vocabulary table, constituted of non significant words, was performed. Those words were not relevant or too general, e.g. "software", "human", "animal", "support US government" The same selection was done with trivial words in titles and abstracts. MeSH [MH] and Registry Number [RN] descriptors present in both databases (under- and overexpressed genes) were labeled with "c" (for common words i.e words present in both corpora) making comparisons more easy.
Compared to a classical investigation server, Transcriptomics server is enriched of a third access mode called " Tables access ". Hyperlinks lead to lists of genes, affiliations, articles titles and journal titles corresponding to Medline fields SI, AD, TI and TA.

1.2.2. Thesaurus browsing. A new functionality was added to Transcriptomics server. Indeed, implementation of thematic extract UMLS Metathesaurus was performed in order to allow navigation in a thesaurus and to discover connecting concepts.

1.2.3. Crossed navigation and multibase aspect. It is possible to make correlations between several corpora from heterogeneous format on a same topic with keyword's frequency, co-occurring words (relationships), or with a cluster's size. To obtain such a result, it is possible to add secondary corpora from different origin to first ones already installed in a investigation server. Secondary corpora are usually called "background databases". Transcriptomics server benefited from implementation four Pascal corpora.

2. UMLS AND TRANSCRIPTOMICS SERVER

2.1. UMLS description

Difficulties to retrieve relevant information in the biomedicine arise because there are many databases in this field (records, news) and most of the time, every database has its own vocabulary, which distorts the query even though the concept behind the term is well known.
"The Metathesaurus is a database of information on concepts that appear in one or more of a number of different controlled vocabularies and classifications used in the field of biomedicine"; [definition from the National Library of Medicine [NLM]: http://www.nlm.nih.gov/ research/umls/META2.HTML#s2].

This Metathesaurus is organized as an array of subtrees and is fed by more than 100 vocabulary sources and classifications from different organizations. In fact, UMLS Metathesaurus is organized by concept or meaning and not by term. 2003 last version of this powerful tool contains more than 870,000 concepts i.e. 2.4 millions of terms. Concepts are located in MRCON file and each concept, identified by a unique key named CUI [for Concept Unique Identifier], is a cluster of all synonymic and lexicographic terms variations.
Relationships concern CUI and not terms. More than 10 millions of relationships are contained in the MRREL file.

2.2. Extracting of UMLS extract and its implementation in the server

As Dilib workbench is based on XML language, both MRCON and MRREL files have been reconverted in XML format in order to facilitate CUI identifiers manipulations.
Moreover, as both files are voluminous, they were filtered, only preferred terms were kept on MRCON (synonymics and other lexicographic variations were rejected), and four over eleven genuine types of relations were kept on MRREL (RN for Narrower Relationship, RB for Broader Relationship, PAR for PARent and CHD for ChilD) i.e. only hierachical relationships.
The main goal was to obtain, from the Metathesaurus sources, a well-suited thesaurus close to the Transcriptomics server's topic. The best way, shown in **figure 2**, was to extract MeSH keywords already in records from Transcriptomics server and thanks to these keywords catch all generics terms (from Server's keywords to roots) from MRREL file. It was performed with a recursive program which raises progressively to Metathesaurus' roots and retrieves, at each step, generic words which are finally added to the server. Relationships themselves, between all words, have been extracted from MRREL file in order to build our thesaurus. At last, CUI keys were remplaced by corresponding preferred terms from MRCON file.
This thesaurus is then implemented in Transcriptomics server's file structure.

3. SEMANTIC LINK FOR CROSSED NAVIGATION IN TRANSCRIPTOMICS SERVER

3.1. Semantic link building

As it was wished to incorporate secondary set of records from Pascal database into Transcriptomics server, which contains already two Medline corpora, we had to find a way to allow browsing between both types of corpora (Pascal and Medline). Indeed, indexing procedure differs between Medline and Pascal formats. Finally, the adopted solution is re-indexing Pascal records with MeSH descriptors and re-indexing Medline records with Pascal descriptors. Inverted files were built from added fields and give access to same records by both Mesh and Pascal keywords.
In order to automatize as most as possible this step of re-indexing, two equivalence tables were built. These tables ,shown in **figure 3**, are needed to build semantic bound.
An XML structure was used as it was suited to Dilib technology (XML based).For Medline-to-Pascal table, the easiest case is to build terminology correspondance between monoterms: "Actins" [in MeSH indexing] to "Actin" [Inist Indexing]. In this case we only use the "<term>" markup.A more difficult case is to build the correspondence between one term and a multiterm descriptor. For example, in **Figure 3**, with "Colorectal Neoplasms" [MeSH keyword], "<list>" and "<term>" markups are used in order to point to three Pascal terms corresponding to Medline concept e.g. "Tumor", "Rectum" and "Colon". Each Pascal term is delimited by the "<term>" markup and all Pascal terms are delimited by the "<list>" markup.Reciprocally, for Pascal-to-Medline table, one of the terms from Pascal expression keywords "colon" is used as the key of the table. Each of the other Pascal terms is delimited by an "<input>" markup. As there are several Pascal terms in the second column, a "<and>" boolean operator is introduced to link them.
This markup syntax allows to encode Pascal keywords.The correspondence with Medline keywords is delimited by a "<then>" markup. All data of the right column has to be encapsulated by the final "<if>" markup. These logical markups allow a DILIB program to find out, from a Pascal monoterm (the key in Pascal-to-Medline table) followed by other Pascal monoterms (in right column of Pascal-to-Medline table), the Mesh corresponding terminology (in bold characters in Pascal-to-Medline table).

3.2. Automatized double re-indexing of documents

Thanks to both tables of vocabulary equivalences, an automatic double re-indexing of all records has been performed.
For the records, this new indexing is added in a new field to keep the original indexing intact. Finally, both Pascal and Medline corpora were re-indexed.

3.3. PASCAL corpora extracting

As semantic link between Pascal indexing and Medline indexing has been solved, it was possible to implement four years of Pascal database. Four Pascal corpora corresponding to years 1999, 2000, 2001 and 2002 were extracted using an internal bibliographic platform, called MIRIAD, which can be used to create corpora from boolean queries and also may provide statistical analyses.
A thematic query was performed for each year of Pascal database in order to limit volumetry of corpora to papers which concern eucaryotic genetics, cellular and molecular genetics, biotechnologies, all medical fields and especially pharmacology, toxicology, cancers, tumors, pneumopathologies and public health. Each corpus contains an average of 1000 to 1500 articles. Theses corpora were then added in the file structure of Transcriptomics server.

4. TRANSCRIPTOMICS SERVER LINKED TO FACTUAL DATABASES

One of the main researchers' wishes since the last decade is to have several points of view on information gathered in the Web and/or databases. In fact, it is even more understandable in biomedical fields as two major types of information exist, bibliographic or textual data on one hand and, on the other hand, factual data e.g. DNA sequences, proteic sequences, graphics for **chromosomal mapping** or 3-D pictures of proteic structures. INRS's team is not an exception, most of all, as the conducted study concerns more than 7,000 genes. A trivial way to build straight automatical access to factual databases from bibliographical records linked to studied genes already available in Transcriptomics server has been performed.

4.1. Heterogeneity and hyperlinks

As right from the outset, National Center for Biotechnology Information 's site (NCBI) has been used to extract the first Medline corpora with Accession Numbers (**AccNum**) queries (these numbers are indexed in GenBank factual database [www.ncbi.nlm.nih.gov/Genbank/]), it is natural to start with these AccNum in order to gather more information

on genes themselves which are implicated in cancerous transformation of pleural cells.
The first idea was to access automatically GenBank entries corresponding to genes from Transcriptomics server thanks to AccNum.
GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis. There are more than 22,000,000 records on August 2002. GenBank provides information but this information is often heterogeneous as type of nucleotid sequence differs. Indeed, a gene can be represented by a cluster of partial nucleotide sequences called Expressed Sequence Tags (EST). Thus, AccNum could be an identifier for a **mRNA** (messenger RiboNucleotid Acid which is a product of an entire gene expression), or for an EST's sequence.
In a GenBank entry, gene's name and hyperlinks to Medline references are not present for a EST's sequence but are available for an mRNA 's sequence.
For instance, GenBank entry for AL135243 AccNum, which corresponds to an EST's sequence, doesn't mention the gene name *adducin 3*.
Otherwise, GenBank entry for NM_006670 AccNum, which is a mRNA's sequence, provides the name of *human trophoblast glycoprotein***,** its abbreviated name *TPBG*, and also several Medline references which lack on previous entry.
How could it be possible to have homogeneous information concerning all sequences of Transcriptomics server ?
In fact, we may have more information on an EST but several hyperlinks corresponding to other databases have to be visited and, as a result, an automatic process catching factual data may be impaired.

4.2 A possible solution : URLs constructing to Unigene and LocusLink chosen databases.

LocusLink and UniGene were chosen to complete informative aspect of GenBank and in the same time to homogenize retrieved factual data which are in our interest. Both databases are hosted also in NCBI and as a result are linked each other. Moreover, browsing access to these databases provides complementary informations for a same sequence of nucleotides.UniGene is organized with gene-oriented clusters i.e. each UniGene cluster contains sequences that represent a unique gene. For instance, the EST's sequence AL135243, from previous sub-section 4.1, is located in UniGene cluster Hs324470 which corresponds to gene of Adducin 3. Moreover, UniGene provides information on the tissue in which the gene has been expressed and its map location. There is also a link to several nucleotide databases hosted in NCBI such as OMIM and LocusLink. There are now more than 100,000 EST's entries for human in UniGene.
LocusLink contains descriptive information about genetic **loci** i.e. physical location of a gene. It gives official nomenclature, genes' aliases, sequence homologies but especially many complementary information on gene's final product i.e. the protein with description of its 3-D structure and its functions.
The best way to have homogeneous information, whatever type of sequence corresponding to AccNum, is to provide automatical access to UniGene database first and, thanks to LocusLink's hyperlink already present on a UniGene entry, a researcher may also consult LoculLink's entry.
Straight access to UniGene entry is possible thanks to URL's building with a .cgi program. Indeed, this URL is a .CGI query (formed with AccNum already in Transcriptomics server) to UniGene database and provide as a result UniGene entry corresponding to the query i.e. AccNum.
As all AccNum are already available in Transcriptomics server, it is easy to modify these lists of AccNum by adding hyperlinks to UniGene entries with such URL.
Moreover, URL for UniGene entry contains two parts, one constant string such as "http://www.ncbi.nlm.gov/UniGene/clust.cgi?ACC=" concatenated with a variant string which is AccNum itself (indexed SI field from Medline records). For instance, "http://www.ncbi.nlm.gov/UniGene/clust.cgi?ACC=" + "AL135243" gives a hyperlink such as AL135243 for table's line corresponding to Adducin 3 gene.
This table's modification has been performed automatically with a .LEX program.

CONCLUSION

In the very first steps of transcriptomics server[1], datamining was allowed thanks to results of **microarray** technology which permit gene expression evaluation. Associations between gene expression and litterature have been performed and analysed in order to redefine research area and discover other functions for genes differentially expressed in mesothelioma cells.
Indeed, current researcher's needs seem to correlate all sources of data in order to have a global view of biological phenomenon. Jenssen [2] developped a tool which answers to such a need, as Transcriptomics server, gene expression were linked to litterature, in order to associate gene expression patterns with patient survival in breast cancer. Jenssen's work allows clusterization of genes with a virtual gene network called Pubgene (www.pubgene.uio.no). Our method of gene clusterization is different as we use DILIB functions which allow linkage between clusters of Mesh keywords [MH field] and gene(s) [SI field].
Another evolution of these tools is to get ahead of data clusterization and allow crossing between heterogeneous data in order to unify retrieved data and to broaden useful information area for researcher. It is even more an emergency as differences in databases format and nomenclature problems occur more and more frequently. That's why new axis of research in Information Research System concern data unification or some way to link heterogeneous information together in order to facilitate its further exploitation.
Indeed, a study is currently performed to get closer to user's point of view in proposing a model, called Xmap, for information retrieval from heterogeneous databases involving assistance to navigation, estimation of degree of confidence and structuring of retrieved data in order to facilitate exploitation [3]. It is applied to the retrieval of mapping data on human genome.
Moreover, another study has been performed in 1997 with MedExplore project [4] which tends to show that conciliation between data exploitation and heterogeneity is not a current researcher's need. Indeed, in this project, user can navigate through information with very different aspect such as Medical pictures, bibliographic databases, biomedical data with multilingual access handled by UMLS metathesaurus. This biomedical tool has been also performed with DILIB technology as Transcriptomics server. In fact, all this diversity on information is not yet treated in our tool but it is our ultimate goal.
As array results are purely numeric, the corresponding genes should be linked to databases like Medline to make the array

results more informative. Such work was achieved by Masys et al. [5]. They used, as we did, controlled terminology, e.g. MeSH descriptors and RN keywords, which is known to reflect biomedical papers. Retrieved MeSH descriptors were translated into parent concept identifiers, according to the classification of the UMLS metathesaurus.
As this last aspect of Masys's study was interesting, we improved Transcriptomics server by adding MESH thesaurus extract from UMLS metathesaurus. Moreover, another functionnality such as crossed navigation between different types of bibliographic sources i.e. Medline and Pascal allows user to correlate infometric results and enlarge access to bibliometric information in order to broaden his(her) experimental field and discover other research areas. At last, giving access thanks to AccNum to biological factual data such as UniGene database's content is a first step, for this tool, to handle such type of data, as classical investigation server only contain bibliographical records. It could then be envisaged to extract at once from PubGene (as it is now automatically accessible) biological information such as loci, tissular localisation where gene is expressed, pathologies linked to over- or under-expression, protein function(s) and even chromosome mapping pictures in order, thanks to DILIB functions, to clusterize such data and linked these new clusters to those already present (from MH, RN, SI, TI Medline fields) in Transcriptomics server.
This new step could enhance analysis and exploitation of INRS microarray results.

REFERENCES

[1] B. Rihn, S.Vidal, C. Nemurat, S. Vachenc, S. Mohr , F. Mazur, P. Houdry, F. Grandjean, S. Visvikis, J. Ducloy, "From Transcriptomics to bibliomics", **Medical Science Monitor** (accepted).
[2] N. Boudjlida, M-D. Devignes, M. Smaïl-Tabbone, "Services for a Genomic Open Distributed Environment". Position Paper, **XEWA Workshop**, League City, TX, December 2000.
[3] T.K. Jenssen, W.P. Kuo, T. Stokke, E. Hovig, "Associations between gene expressions in breast cancer and patiant survival", **Human Genetics,** Vol. 111, 2002, pp. 411-420.
[4] E. Nauer, J. Ducloy, J-C. Lamirel, "Using of multiple data source for information filtering:first approaches in the MedExplore project", **5th DELOS Workshop** Budapest 10-12 november 1997.
[5] D.R. Masys, "Linking microarray data to the literature", **Nature Genetics**, Vol. 28, 2001, pp. 9-10.

GLOSSARY

AccNum : Abbreviation of Accession Number (identifier of nucleotide sequence in GenBank database).
Chromosomal mapping : physical cartography of a chromosome.
DNA : sort of nucleotide sequence present in genome. Genes are made of DNA.
EST : For Expressed Sequence Tags. Constitutes peaces of a whole nucleotide sequence (DNA or mRNA) usually used in experiments.
Locus (pl. loci) : localisation of a gene on a chromosome.
Microarray techniques : experimental process to evaluate gene expression.
mRNA : For messenger RiboNucleotide Acid. Result of gene's transcription, it is an intermediary step between the gene located in cell's kernel and protein which has a role in physiology.
Nucleotide's sequence : sequence present in cells or synthetized in laboratory which contains genetic information (DNA, RNA or cDNA).
Protein : result of a gene after two steps in gene's expression (fisrt one : transcription, second one : translation).
Reverse transcription : experimental procedure to obtain from a RNA a DNA.
Sequence homology : comparison between several nucleotide sequences from different species, for instance between a human gene and a mouse, a rat, a fly etc…
Transcription : step procedure to transform a DNA in mRNA. Transcription level is proportionnal to activation of gene expression.
Translation : procedure to transform a mRNA in functional protein.

WEB SITES

INIST home page: http://www.inist.fr/

INRS home page: http://www.inrs.fr/index_fla.html

Transcriptomics server home page:
http://dilib.inist.fr/dps/sdv/GenomeAP/Server/EN.genome.index.html

DILIB home page: http://dilib.inist.fr/

Bioinformatics and Genomics, presentation of Transcriptomics server (in french):
http://www.forumlabo.com/2002/abstracts/2002/26bibliom.htm

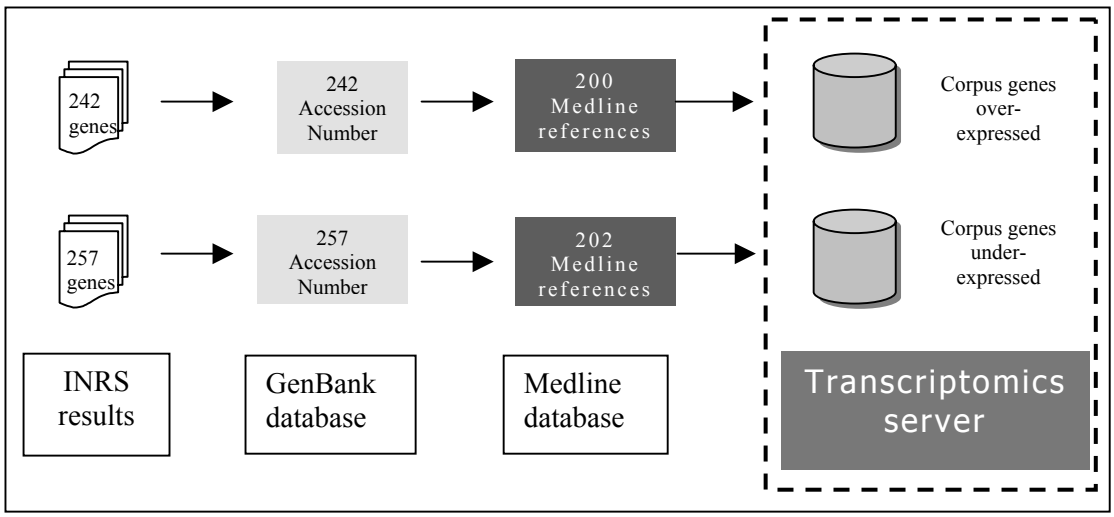

Figure 1 : Automatized processing of bibliographical data retrieval and Transcriptomics server generation. This server contains two Medline sets of records but will be completed by UMLS metathesaurus extract and four sets of Pascal records.

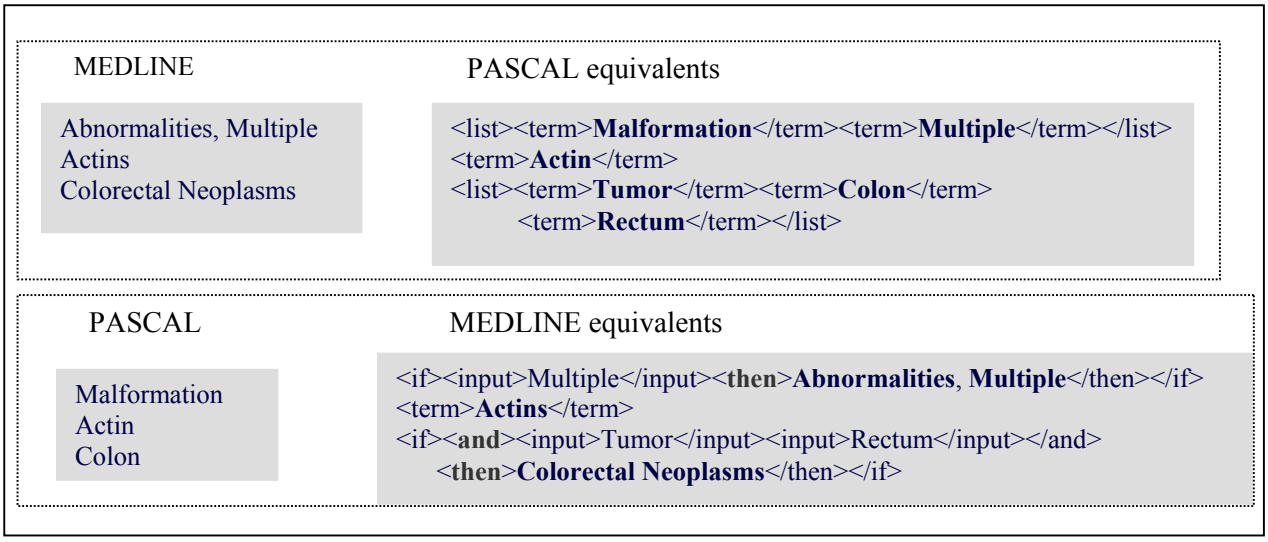

Figure 3 : On top, **Medline-to-Pascal table** with Medline keyword as keys on left column and Pascal equivalents on right column. On bottom, **Pascal-to-Medline table** with Pascal descriptor as key on left column and Medline equivalents on right column. Each table contains tagged fields in XML structure on right column in order to handle multiterms descriptors.

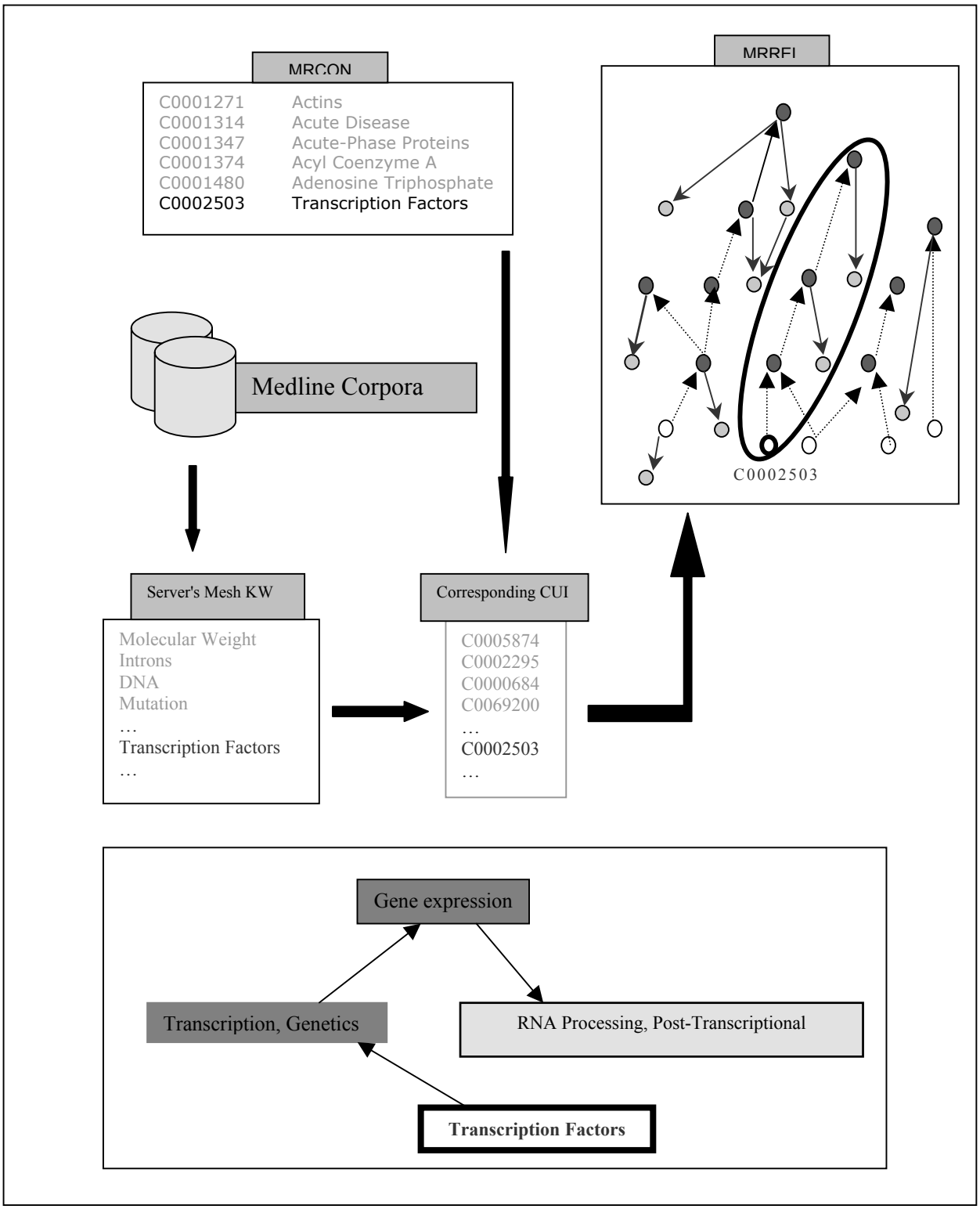

Figure 2 : Thesaurus building processing. On top, both MRCON and MRREL files where dotted arrows in MRREL file represent recursive rise of Dilib program, non-dotted arrows show how other terms (light gray balls) have been caught thanks to generic ones (dark grey balls). White balls represent MeSH terms from Transcriptomics server.