AN INFORMATION COMMUNICATION ASSISTANCE SYSTEM

LOUIS-CLAUDE PAQUIN, PH.D., PROFESSOR
COMMUNICATIONS DEPARTMENT
ASSOCIATE MEMBER OF THE UNESCO-BELL CHAIR
FOR COMMUNICATION AND INTERNATIONAL DEVELOPMENT
UNIVERSITÉ DU QUÉBEC À MONTRÉAL

This paper offers a solution to the problem of information overload caused by the massive flow of documents on Internet. The volume keeps increasing to the point where it is now calculated in gigabytes (one page equals 2 kilobytes). On the other hand, our ability to process and absorb information remains limited, resulting in an exformation phenomenon, that is, a collection of unprocessed information for want of time and skill. Worse still, the increase in the volume of information can lead to focus reduction, a tunnel effect. Of course, many systems ara available to help locate documents through key words. Some, like Lycos, are remarkably effective, considering the wide volume they cover. However, this ease of access to documents obfuscates the difficulty of finding relevant information. How can one be sure that all relevant documents have been scanned? Why are so many irrelevant documents ferreted out?

Analytical processing is unavoidable. If the operation is not carried out before the documents are circulated, it has to be done by users who must keep reformulating their request and scan vast amounts of extraneous material to achieve a result that is not necessarily satisfactory. Information scientists agree on the need to prepare the documents to be circulated by marking their logical structure (parts, sections, subsections, etc.) and conceptual indexing. Both forms of document preparation have until now been carried out semi-manually by professionals, entailing substantial costs and delays.

The Information Communication Assistance System (ICAS), which we are now developing, is a work flow, that is, a set of computer procedures designed to speed up and support human decisions in the low-cost processing of existing documents to allow satisfactory access to their content. ICAS helps perform two operations: i) converting the word processor's typographical codes (character type, paragraphs, indents, tables, etc.) into information on the documents' logical structure; ii) revealing the terminology used in such documents, i.e. all terms - often made up of several words - designating concepts in the reference field.

The identification and marking of the documents' logical structure are achieved by the following modules: a converter of the binary formats specific to word processing into a Sandard Generalized MarkUp Language (SGML) format; a learning module of the Document Type Description (DTD), that is the set of tags and their syntax for a given range of documents, and the rules of conversion of typographical signs into components of the logical structure proceeding from direct handling: elements of the documents are selected on the screen and a label is assigned to them, the rules are automatically defined; a document analyzer which identifies and marks out the logical structure of the document from the learned typographical signs, breaks it into segments and identifies the references to other documents; this analyzer is in fact an expert system whose certainty accretion enables it to solve noisy situations; and, finally, a module validating the results of the analysis to help resolve structural ambiguities and change the rules by direct handling.

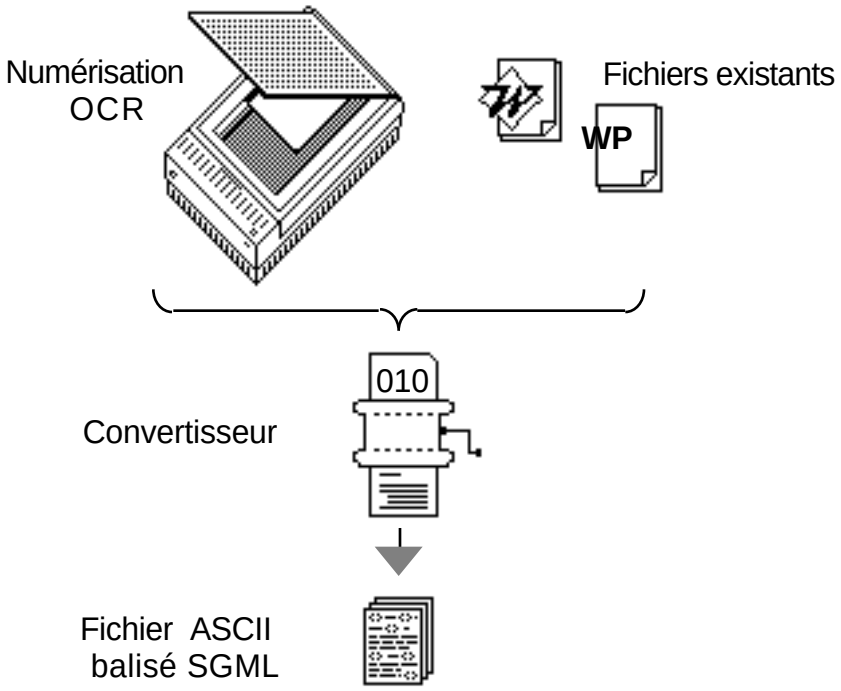
The approach to reveal the terminology can be called "mixed" since it is based on the sheer force of an algorithm for the statistical detection of co-occurrences, tracking the frequency of certain word groups within a given documentary space, to which applies a restriction with morphological categories. The expressions are arranged in ascending order from the shortest to the longest one. A hypertext-type interface makes it possible to navigate from one term to another and forward the selected term to the server's research module.

In short, to allow for the consistent, faster preparation of documents by less skilled users, ICAS uses the techniques of uncertain reasoning, supervised learning, direct-handling graphic interfaces and automated processing of natural languages. This research and development project, estimated at $1.5 million, has been granted a $500,000 subsidy by the Quebec government.

Voici maintenant une description détaillée des modules qui constituent SACI accompagnée, le cas échéant, des hypothèses sous-jacentes :

Convertisseur de formats binaires

Le convertisseur de formats binaires propres aux logiciels de traitement de texte, constitue le point d'alimentation de la chaîne de traitement comme le montre le schéma suivant :



Le convertisseur a pour fonction d'uniformiser tous les formats liés à des produits donnés pour les rendre admissibles à l'analyseur. Le format produit par la conversion des fichiers binaires est de l'ASCII balisé en SGML. Le convertisseur s'applique aux caractères accentués ainsi qu'aux codes typographiques : justification, gras, italique, tabulation, retrait, changement de fonte, de grosseur, etc. De plus, ce livrable effectuera la conversion des tableaux en transformant les codes de la structure matricielle (cellules, rangées et colonnes) en SGML. Il effectuera aussi la conversion du format vectoriel (les diagrammes et les schémas) en image «bitmap» après avoir extrait et balisé les chaînes de caractères présentes afin de permettre le repérage. Ce livrable repose sur les hypothèses suivantes : une conversion en SGML des codes mentionnés plus haut peut être réalisée sans aucune perte d'information; un balisage commun et indépendant des logiciels permet d'uniformiser tout traitement subséquent; les chaînes de caractères qui sont contenues dans la première rangée et la première colonne des tableaux, de même que dans les schémas et les diagrammes constituent un très bon indice des concepts traités.

Module d'apprentissage de la DTD et des règles de transformation

L'utilisation de SGML pour des fins de marquage requiert la description préalable d'une Document Type Description (DTD) qui définit les balises et leur syntaxe pour une classe de documents donnée. Ainsi, par exemple, une DTD prescrira que le document comporte des parties et que les parties comportent des sections, que les sections comportent un titre facultatif et des sous-sections qui, elles, comportent des paragraphes. La conception de DTD est une opération complexe qui demande un entraînement aux langages formels et à l'analyse de texte. Il s'agit d'un obstacle majeur sinon à l'utilisation de SGML, du moins à une optimisation du potentiel de ce langage. La plupart du temps, les DTD sont constituées par modification de DTD déjà existantes. Les DTD sont très difficiles à lire et à comprendre car elles sont composées d'une série de définitions, d'abord de composantes simples, puis de

composantes de plus en plus complexes fabriquées à partir de composantes de complexité moindre.

Il ne s'agit pas dans le projet SACI de concevoir un module pour confectionner n'importe quelle DTD que l'on voudrait appliquer à tout type de structure ou de composante des documents. La seule structure qui nous intéresse est la structure logique du document, c'est-à-dire sa subdivision en parties, en sections, sous-sections, etc. L'envergure du problème se trouve donc réduite à des proportions telles qu'il devient possible de constituer un inventaire des catégories pertinentes et de concevoir un module d'assistance à la conception de DTD pour la structure logique des documents. Ce module comporte les éléments suivants : un interface graphique à manipulation directe pour identifier sur le document les différentes composantes structurelles, un générateur de règles de transformation ainsi qu'un compilateur de définitions des composantes de DTD.

L'interface graphique permettra d'afficher un fichier en ASCII qui a reçu un balisage SGML des codes typographiques. L'utilisateur n'aura qu'à sélectionner une partie du document qui correspond à une composante et à en identifier la catégorie et les attributs à l'aide de menus. Les attributs permettent de déterminer si la composante est obligatoire ou facultative, unique ou répétable. Ainsi, par exemple, une section peut ou ne pas comporter un titre mais celui-ci est toujours unique. Les catégories déjà identifiées sont affichées dans la marge de gauche pour vérification. Cette opération d'identification va des composantes les plus simples aux composantes de plus en plus complexes. À partir de cette catégorisation, deux traitements sont accomplis : la constitution de la DTD et de la base de règles de transformation des codes typographiques en composantes de la structure logique du document. Voici un exemple de règles de transformation : si une chaîne est seule et est suivie d'un paragraphe, alors c'est un intertitre; si une chaîne de caractères est seule, en gras et suivie d'un paragraphe ou d'un intertitre, alors c'est un titre de section.

Trois hypothèses sont à la base de la conception de ce module : la technique d'apprentissage par l'exemple constitue la meilleure stratégie pour faciliter la mise au point des DTD et la construction des règles de transformation des codes typographiques vers la structure logique ; un interface graphique à manipulation directe est le dispositif le plus efficace réaliser un apprentissage par l'exemple ; il est possible de construire un compilateur de DTD à partir d'une syntaxe formelle.

Module de classification des documents

L'ensemble des documents à diffuser constituent un espace documentaire. Cet espace peut être ordonné par un plan de classification dont on peut tirer parti pour accéder à l'information des documents. Pour permettre la classification des documents, ce module sera doté d'un interface graphique à manipulation directe qui permettra l'inscription d'un nouveau document dans l'espace documentaire représenté sous forme d'un arbre. Ce module pourra aussi être utilisé pour réaménager un espace documentaire déjà réalisé. Les documents seront représentés par des icônes et les niveaux intermédiaires de l'arborescence par des icônes d'un autre type. Les opérations d'édition possibles seront l'insertion et l'élimination d'un document, la création et l'élimination d'un niveau intermédiaire, le déplacement de documents ou de niveaux intermédiaires. Ainsi, lorsqu'un sous-groupe de documents, représenté par une portion de l'arborescence, sera sélectionné et déplacé, tous les documents appartenant à ce sous-groupe seront aussi sélectionnés et déplacé de la même façon. Pour faciliter la manipulation de l'arborescence, des opérations de focalisation seront possibles : des *zoom out* qui ont pour effet de cacher une partie inférieure de l'arborescence pour avoir une vue d'ensemble, des *zoom in* qui permettent le contraire, de voir une portion de l'arborescence dans le détail. L'hypothèse à la base de ce module est que les taxonomies sont plus faciles à concevoir et à gérer directement dans leur forme graphique.

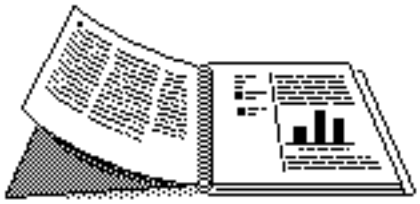
Analyseur de documents

Ce module effectue de façon automatique les trois opérations suivantes sur les documents : la segmentation des composantes de la structure logique, l'identification et le balisage des composantes de la structure logique ainsi que l'identification des références. Ces opérations sont effectuées à partir de la DTD et des règles de transformation des codes typographiques en composantes de la structure logique produites par le module d'apprentissage. Rappelons que ces règles de transformation prennent en prémisse des codes typographiques : si on constate un changement du type de caractères de la grosseur du jeu de caractères ainsi qu'un retrait négatif alors on balise «sous-titre».

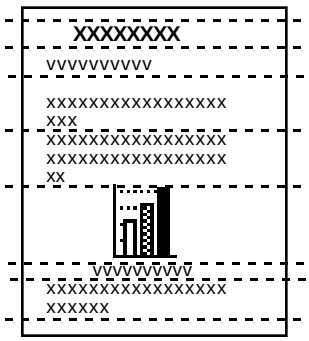
L'hypothèse qui est à la base de ce livrable est que l'utilisation des techniques du raisonnement incertain permet une tolérance dans le déclenchement des règles de transformation tout en fournissant un indice de «certitude» qui permettra d'orienter l'intervention humaine pour les corrections éventuelles. Ainsi, l'analyseur sera un système expert doté d'un cumul de certitude qui lui permettra d'oeuvrer dans des situations «bruitées», c'est-à-dire lorsque les indices typographiques nécessaires au dépistage d'une constituante de la structure ne sont pas tous là ou encore lorsque plusieurs interprétations contradictoires sont plausibles. Un exemple d'une situation «bruitée» serait un segment qui compte une ligne et demie. Il peut s'agir d'un paragraphe si les intertitres ne comptent pas d'identificateurs numériques, utilisent la même fonte que le reste du texte et commencent à la ligne comme les paragraphes. Les système experts par le traitement des connaissances incertaines peuvent poser un diagnostic différencié de ces situations et offrir assistance à une désambiguisation manuelle. Ainsi, pour l'exemple précédent, le système expert déterminera qu'il s'agit d'un titre avec un indice de certitude de 75% et d'un paragraphe avec un indice de 50%. De plus, il est possible dans les situation de conflit d'interprétation comme celle-ci lorsqu'on utilise un système expert de formuler des méta-règles qui viennent trancher en prenant en compte des décisions passées, des probabilités d'occurrence,

etc.

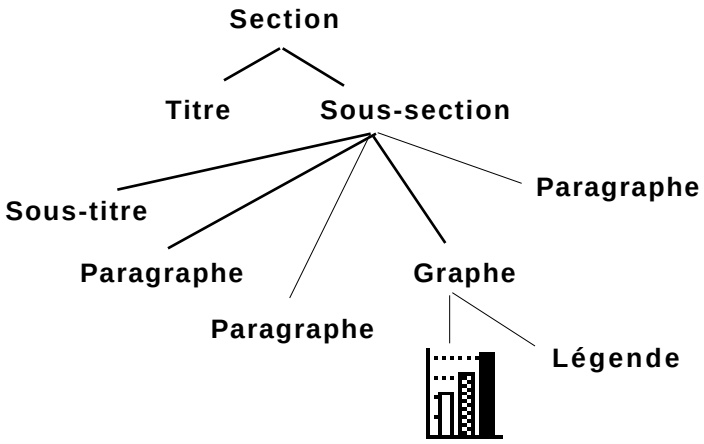
Voici un schéma illustrant les opérations effectuées par l'analyseur :



Segmentation des composantes



Balisage de la structure logique



En plus de segmenter le document en composantes de la structure logique, d'identifier et de baliser ceux-ci, l'analyseur dépiste les références. Ce dépistage est effectué avec des règles qui sont constituées elles aussi par apprentissage. Cet apprentissage est effectué en sélectionnant des expressions indiquant qu'il y a référence. Ces expressions sont transformées en patrons appliqués par l'analyseur au document. Par exemple, le mot «article» ou son abréviation «art.» en italique suivi

d'un nombre, indique la référence à des lois, règlements, contrats ou convention collectives. Tout comme pour les composantes de la structure logique, le cumul des coefficients de certitude permet de détecter la possibilité de référence dans les situations «bruitées». En plus de recevoir un balisage, les références dépistées sont inscrites dans une table qui permettra, lorsque le segment correspondant à la référence sera traité, d'établir un lien.

Module de validation

Ce module interactif a pour fonction de valider les composantes de la structure logique qui ont été identifiées et les balises qui leur a été attribuée lors l'analyse automatique effectuée précédemment. Il a aussi pour fonction d'ajouter d'autres indications permettant l'accès à l'information qui ne peuvent l'être automatiquement. Ces indications sont des liens associatifs avec d'autres segments du même document ou d'autres documents d'un même espace documentaire.

L'interface de ce un module sera conçu pour faciliter l'intervention humaine, autant par une ergonomie adaptée à la tâche que par une aide en contexte. Cet interface comporte une zone d'affichage du segment avec déroulement qui permet de valider la segmentation qui a été effectuée. Si le découpage est inopportun, l'utilisateur peut regrouper plusieurs segments en un seul segment ou encore il peut décomposer le segment en plusieurs segments par sélection de blocs. Dans ce dernier cas, l'analyseur est appliqué aux nouveaux segments pour corriger le balisage. De plus, les références qui ont été dépistées sont mises en relief dans le texte du segment. l'utilisateur peut les invalider ou en ajouter d'autres. Les autres zones de l'interface sont consacrées respectivement à un identificateur séquentiel des segments qui sert à l'adressage des liens qui pourront être établis avec les autres segments, à l'indication du niveau du segment dans la structure du document, au titre du segment et aux identificateurs des segments qui sont liés.

Le titre du segment, lorsqu'il est présent dans le document et qu'il a été dépisté par des indices typographiques est affiché dans la zone réservée à cette fin, sinon la zone est laissée en noir. L'utilisateur peut accepter tel quel le titre dépisté ou encore il peut le changer. Il est important de noter que la plupart du temps les titres qui sont donnés aux subdivisions des documents sont ambigus et incomplets en eux mêmes. Ils sont habituellement attribués avec un souci de concision dans un contexte de lecture linéaire du document sur support papier. Les intertitres attribués aux segments sont cumulés aux sous-titres et aux titres de sections et parfois même aux titres de parties. L'interface sera dotée d'options qui faciliteront la tâche de l'utilisateur en lui permettant de cumuler au titre du segment un ou plusieurs titres des niveaux supérieurs dans la structure. Il est à noter que dans le cas où un titre est changé, le titre original est conservé par souci d'intégrité de l'archive.

La table des matières qui sera offerte lors de l'accès à l'information est donc constituée a posteriori et ne correspondra pas, dans la plupart des cas, à la table des matières dont la version papier du document peut être dotée. En plus de la compositionnalité des titres vue précédemment, les tables des matières de la version papier des documents ne se rend que très rarement au niveau des segments, se contentant d'indiquer les grandes divisions et laissant ensuite le soin au lecteur de parcourir celles-ci afin de trouver les segments.

Une dernière zone contient les identificateurs des autres segments qui ont été liés au segment présentement affiché. Ces liens sont établis un à un par l'utilisateur. Pour ce faire, le module de validation offre les facilités d'un serveur de fichiers pour interroger et naviguer dans l'espace documentaire en cours de constitution. L'utilisateur peut effectuer une requête, soit à partir des mots du texte, soit à partir de la terminologie (cf. livrables suivants) ou encore par la table des matières et, suite à un examen des passages repérés, il peut établir un lien entre certains segments et le segment en cours de vérification.

Les hypothèses qui sont à la base de ce module sont les suivantes : il est impossible de réussir totalement une analyse de texte automatique; il est beaucoup plus facile et rapide d'accomplir une tâche de vérification à l'aide d'un interface à manipulation directe; les titres des composantes d'un document papier doivent être revus pour une diffusion électronique; les liens associatifs ne peuvent être établis automatiquement de façon satisfaisante sur la base de chaînes de caractères identiques, la relation conceptuelle doit être très forte.

Analyseur terminologique

Ce module a pour objectif de pallier aux déficiences de la recherche d'information sur un serveur par les mots du document qui est source de «silence» et de «bruit», sans pour autant recourir à un solution aussi lourde qu'une indexation conceptuelle. Les hypothèses sous-jacentes sont les suivantes :un inventaire de la terminologie permet de restreindre le «bruit» lors de la recherche d'information sur un serveur à partir des mots des documents si les multi-termes y sont répertoriés; une stratégie mixte de repérage des multi-termes, basée à la fois sur un calcul de fréquence d'occurrences et sur une restriction par les catégories morphologiques, est la plus performante et efficace; seul l'utilisateur est en mesure de valider les multi-termes qui sont pertinents.

La solution retenue réside dans un inventaire de la terminologie présente dans un espace documentaire donné. Par terminologie il est entendu l'ensemble des termes qui désignent des concepts dans le domaine de référence. Les termes sont deux types : les uni-termes et les multi-termes. Les uni-termes qui correspondent à un mot sont les moins fréquents et servent à constituer plusieurs multi-termes d'où le «bruit» au repérage d'information. Les multi-termes sont des expressions composées de plusieurs termes dont la signification est différente de leur signification individuelle. L'expression «traitement de texte» est un bon exemple de multi-terme.

La stratégie employée ici qui constitue une approche originale peut être qualifiée de «mixte», en ce qu'elle est basée sur la force brute d'un algorithme de dépistage statistique de cooccurrences auquel on applique une restriction en fonction des catégories morphologiques en présence. La cooccurrence est le phénomène de la répétition fréquente d'un certain groupe de mots dans un espace documentaire donné. L'algorithme de dépistage de cooccurrences identifie pour chacun des mots de cet espace documentaire, pris comme pôle, l'ensemble des expressions où ce mot apparaît avec une fréquence qui dépasse un seuil arbitrairement fixé. Les expressions sont ordonnées des plus courtes aux plus longues. Les résultats d'un algorithme de dépistage de cooccurrences sont très volumineux car les mêmes expressions reviennent pour chacun des mots qui les composent. Ainsi, pour l'exemple précédent, on retrouvera les mêmes expressions sous les mots-pôles «congés», «droit» et «prise».

Tous les multi-termes sont dépistés par un algorithme de dépistage de cooccurrences. Toutefois ce ne sont pas les seules expressions qui sont dépistées. En fait, les multi-termes constituent une faible proportion des expressions dépistées. Ils peuvent être considérés comme des cooccurrences qui on sens complet par eux-mêmes. C'est ainsi qu'une restriction doit être opérée sur les cooccurrences dépistées. La première restriction est appliquée dès le dépistage par le recours à un anti-dictionnaire qui permet de ne pas constituer de liste d'expressions pour les mots qui ne peuvent désigner un concept comme les conjonctions, les articles, et autres mots-outils. Cette restriction a pour effet de réduire considérablement le temps de calcul et le volume des résultats. La deuxième restriction a pour but d'éliminer les mots-pôles qui ne désignent pas des concepts et les contextes d'occurrences non significatifs ou incomplets. Pour ce faire, les résultats sont décrits par un analyseur morphologique. À partir des catégories, tous les mots qui ne sont pas des noms communs ou propres sont écartés en position de pôle. Puis, les expressions dont la formation est incomplète ou inappropriée c'est-à-dire dont la catégorie morphologique du premier mot n'est pas un nom ou un adjectif et dont la catégorie

morphologique n’est pas un nom, un adjectif, un participe ou un infinitif.

L'approche proposée est entièrement automatique donc n'engendre aucun coût de main d’oeuvre. Cette approche serait insatisfaisante si elle avait pour but de constituer un répertoire des termes du domaine, car d’autres facteurs doivent entrer en ligne de compte pour sélectionner des termes bien formés. Elle est toutefois particulièrement bien adaptée à la formulation de requête d’information dans un serveur de documents en ce qu’elle permet à loisir à l’utilisateur de restreindre le «bruit», soit les segments qui ne sont pas pertinents. En effet, choisir un multi-terme court aura pour effet de dépister un plus grand nombre de segments alors que choisir le plus long permettra un plus grande précision à la recherche.

Module de navigation dans la terminologie

Le résultat de l’analyse terminologique précédente consiste en une liste alphabétique de termes, appelés expressions principales ou têtes. Une liste de contextes d'occurrence, appelés expressions subordonnées est adjointe à chacune de ces expressions principales. Le module de navigation dans la terminologie présente un interface qui facilite la manipulation de la liste d’expressions principale et des listes d’expressions subordonnées qui leur sont adjointes. C’est ainsi que la fenêtre-écran est divisée en deux sections, chacune dotée d’une barre de défilement pour localiser l’expression désirée. les expressions principales sont affichées en permanence dans la section de gauche, alors que les expressions subordonnées de l’expression principale sélectionnée sont affichées dans la section de droite.

Pour localiser rapidement une expression principale donnée, l’utilisateur appuie sur les lettres du clavier correspondant aux premières lettres de l’expression. Une fois l’expression principale localisée, celle-ci est affichée en vidéo-inverse et la liste d’expressions subordonnées de cette expression est affichée dans la section de droite. Dans l’expression subordonnée, l’expression principale est remplacée par

deux tirets. En visionnant les différentes expression subordonnées, l'utilisateur peut, avant même de formuler une requête, opérer une restriction qui lui fera d'économiser du temps de validation des segments obtenus suite à une requête au serveur. Lorsqu'une expression subordonnée est sélectionnée, elle s'affiche dans la zone du bas. Pour lancer une requête au serveur afin d'obtenir tous les segments qui contiennent l'expression subordonnée sélectionnée, il suffit de valider la sélection.

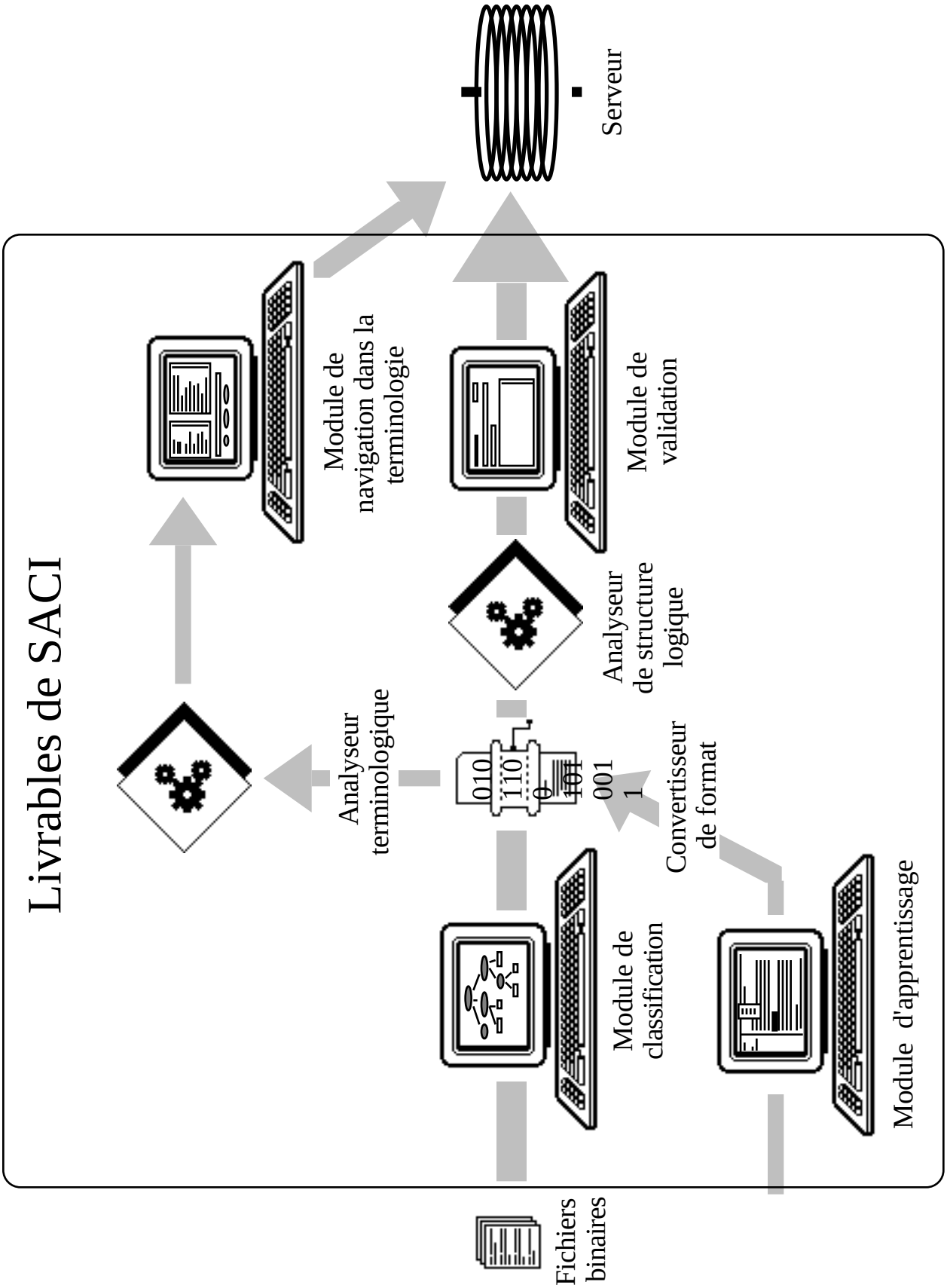
L'interface permet aussi de sélectionner dans la zone du bas qui comporte l'expression subordonnée complète un terme qui deviendra la l'expression principale et dont les expressions subordonnées viendront s'afficher dans la section de droite. Ce dispositif de navigation permet de prendre contact avec l'univers terminologique d'un corpus de document donné. De même, il donne accès à l'information même si l'on ne connaît pas exactement la terminologie. Ce module repose sur l'hypothèse que la langue étant très productive, il est très difficile a priori de prédire les multi-termes des documents qui seront jugés pertinents. Il vaut mieux alors un dispositif qui facilite la navigation au travers une liste de candidats et que cela soit l'utilisateur lui-même, celui qui recherche de l'information, qui reconnaisse les multi-termes qui l'intéressent.

Module d'alimentation du serveur

Le module d'alimentation du serveur intervient à plusieurs niveaux et touche plusieurs types de documents, tous balisés avec SGML. : les segments de document après leur validation; la table des matières des documents après la validation de tous les segments; le plan de classification de l'espace documentaire après le traitement de tous les documents; la terminologie de l'espace documentaire après le traitement de tous les documents. Une fois que les indications ajoutées au segment par l'analyseur sont validées à la satisfaction de l'utilisateur et que des liens ont été ajoutés, les segments sont versés dans le serveur et le titre du segment est versé dans une table des matières. Une fois que tout un document a été traité, la table des matières est

alors versée dans le serveur. Une fois que tous les documents ont été inscrits dans le plan de classification, analysés et validés, le plan de classification est versé dans le serveur. De même, une fois que la terminologie a été extraite, le résultat est versé dans le serveur.

Voici en terminant un schéma qui illustre les livrables de SACI et leurs relations :


Livrables de SACI
Serveur
Module de navigation dans la terminologie
Module de validation
Analyseur de structure logique
Analyseur terminologique
010
110
001
Convertisseur de format
Module de classification
Module d'apprentissage
Fichiers binaires