De l'information primaire à l'information à valeur ajoutée dans le contexte du numérique

Sahbi Sidhom, Abiodun-Charles Robert, Amos David

De l'information primaire à l'information à valeur ajoutée dans le contexte du numérique

Sahbi SIDHOM
MCF & Chercheur
de l'équipe SITE du LORIA

Charles ROBERT
Doctorant
de l'équipe SITE du LORIA

Amos DAVID
Professeur & Responsable
de l'équipe SITE du LORIA

LORIA - Université Nancy2, BP. 239, 54506 Vandoeuvre Cedex

Sahbi.Sidhom@loria.fr **Robert@loria.fr** **Amos.David@loria.fr**

Résumé

Notre objet d'étude porte sur la « *gestion de l'information numérique et la coordination avec l'information à valeur ajoutée* ». Le traitement de l'information quelque soit sa nature et ses origines se trouve au confluent de plusieurs disciplines que sont : l'analyse linguistique automatique, l'informatique, les mathématiques, les réseaux d'information, la socio-économie, l'intelligence économique (IE), etc.

Cette dimension pluridisciplinaire sur l'information numérique occupe de nos jours une place prépondérante dans l'activité des institutions (gouvernementales, scientifiques, socio-économiques et autres), et offre des possibilités de moduler des interactions complexes en matière des besoins informationnels. Ainsi, dans un processus d'IE, l'information associée à une ressource documentaire permet d'en favoriser l'utilisation, l'exploitation et l'annotation par un *agent* humain (veilleur ou décideur), du fait de son exploitation par un *agent* logiciel (plate-forme et outils informatiques).

Dans le cadre de cette étude, nous vous proposons une architecture logicielle pour la mise en œuvre des traitements qui partent de l'information primaire pour aboutir à l'information à valeur ajoutée.

Abstract

Our objective focuses on the "management of digital information and the coordination with value-added information". Information processing involves several disciplines whatever its nature or origin. These disciplines include automatic linguistic analysis, computer science, mathematics, information network, social economy, economic intelligence (EI), etc. This pluridisciplanry dimension of digital information occupies a predominant position within the activities of institutions (governmental, scientific, socio-economic and others), and offers the possibility to modulate the complex interactions as regards information needs. As such, the information associated with documentary resources within the process of EI allows to improve the utilisation and annotation by the human agent (in this case, information watcher and decision-maker), through the use of software agent (computer tools and platforms). Within the framework of this study, we propose a software architecture for the implementation of the processing, starting from primary information to value-added information.

Mots-clés

Document numérique, information multiforme, information à valeur ajoutée, Intelligence Economique, Modélisation de l'utilisateur, Système de recherche d'information, Outils logiciels.

Keywords

digital document, multiform information, value-added information, economic intelligence, user modelling, information retrieval system, software tools.

1. Introduction

Le document numérique est devenu incontournable dans toutes les chaînes de manipulation de l'information. Il a gagné rapidement en importance dans les procédures de traitement documentaire des bibliothèques, des centres d'informations et des institutions économiques, gouvernementales, industrielles ou autres. Dans ce travail, le document numérique est présenté comme un document qui n'est lié à aucun support physique, lequel n'est qu'un véhicule transitoire. En revanche, c'est un document "structuré" totalement ou partiellement, qui contient les éléments essentiels de sa description, de ses accès (identification, autorisation, …) et de son administration [BACHIMONT, 99].

Et sous l'aspect « chaîne de traitement » de documents numériques, la complexité consiste à la fois en une description (externe et interne), une indexation d'accès (représentation et analyse de contenu) et des indications de localisation (disposition, format de données, …). Les données administratives qui permettent la gestion du document (les règles de communication, la confidentialité, les mouvements de prêt, …) sont réparties entre différents systèmes (prêt, règlement intérieur, acquisition, …).

De nos jours, les bibliothèques sont désormais confrontées à la gestion de documents numériques pour de multiples raisons. D'abord, parce qu'elles font partie d'une institution qui créée elle-même des documents de ce type. C'est le cas dans les universités, qui produisent des supports pédagogiques et de recherche pour étudiant, enseignant et chercheur. Ces supports doivent tous être valorisés : mémorisation, accès, disposition de contenus (notice, source). Dans cet article, nous illustrons nos propositions par des applications sur des documents de l'institut national de l'audiovisuel (INA) à Paris, qui a pour mission la conservation des programmes audiovisuels (chaînes radios et TVs) et la constitution d'un savoir audiovisuel pour les futures générations de France. Ce choix s'explique par la diversité des documents : le multimédia.

Dans le contexte du document numérique sur le multimédia, l'objectif recherché par les professionnels de l'information est la constitution d'un fonds documentaire à partir de la description du contenu, l'identification des demandes [BUENO & al., 2001] (politique de sélection, usages, besoins, publics, …) et d'en informer les usagers.

Pour atteindre cet objectif, les professionnels pratiquent sur les documents multimédia un certain nombre d'opérations intellectuelles [ALQUIER, 00]. Pour ceux de l'INA, il s'agit de l'analyse documentaire audio-visuelle selon des méthodes propres, désignées par les *grilles d'analyse chronologique*. Ce qui conduit à représenter le contenu (thèmes, résumés, séquences, …) et la

constitution d'outils terminologiques et thésauriques (descripteurs multi-niveaux) afin de faciliter la recherche ou le filtrage d'informations. À l'INA, l'analyse audiovisuelle comprend deux opérations complémentaires pour la constitution d'une banque de données numériques. Cette dernière est associée dans sa gestion à des bases de données bibliographiques, pour parvenir à :

1 L'élaboration de résumés en vue de présenter à l'utilisateur une version abrégée ou concise sur le contenu du document, lui permettre d'en juger de l'opportunité de la lecture, de la visualisation ou de l'audition de la version source du document. Il s'agit de la production d'un texte bien formé selon des critères méthodologiques bien définis : les *grilles d'analyse chronologique* ;
2 L'indexation, c'est-à-dire le relevé des concepts significatifs porteurs d'information et caractérisant le document analysé. Ce critère fait la qualité de la mémoire documentaire. Il s'agit de compléter le contenu selon des spécifications thésauriques bien définis : les *thésaurus (INA) de l'audiovisuel*.

Dans la démarche de notre étude sur l'analyse des documents multimédia (textes, images fixe ou animée, séquences audio, séquences vidéo), les contraintes de l'*analyse documentaire* préservent la variabilité de la forme des productions textuelles (plusieurs variantes de résumés : chapeau, résumé, séquences du document, résumé producteur), bien que leur élaboration soit en fonction d'une grille d'analyse [SIDHOM, 02]. Ainsi, cette activité nous a conduit à considérer le texte résumé dans sa globalité (bien formée) et dans sa particularité (ses structures) comme cadre d'étude pertinent sur la chaîne de traitements de l'information (*cf.* Figure 1.), en partant du document primaire (numérique) jusqu'au document d'annotation (information à valeur ajoutée).

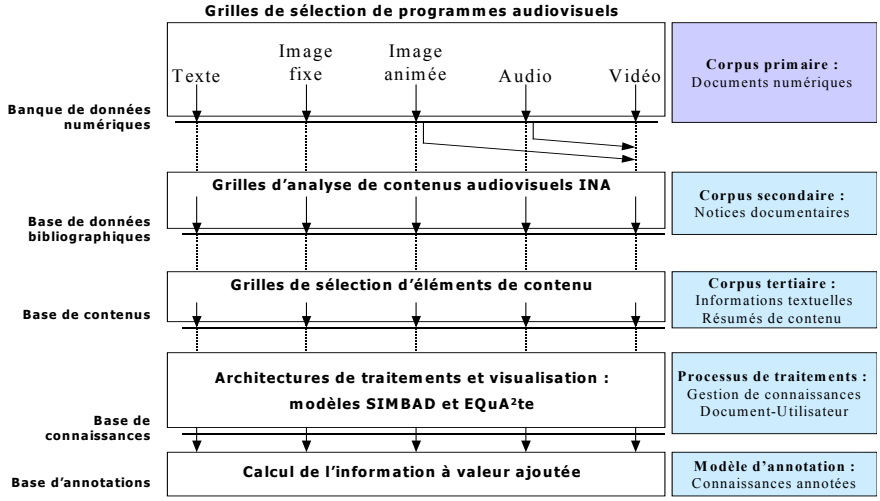


Figure 1. Architecture en couches : du numérique aux informations à valeur ajoutée.

L'enjeu de cette perspective de recherche porte sur :

1 La problématique de l'exploitation de l'information numérique : *Quels savoirs issus de la recherche ayant comme source l'audiovisuel et de quelle manière ?*

Le document numérique s'est organisé au travers différents métiers, compétences et savoirs-faires. Dans un tel environnement, une synthèse professionnelle sur l'analyse de contenu documentaire s'avère nécessaire.

2 La problématique du traitement de l'information secondaire : *Une fois disposé d'un*

texte représentant un contenu audiovisuel, comment l'analyser, l'indexer et visualiser la connaissance ? et comment s'enrichit-il ?

Notre travail de recherche nous a amené à évoluer vers un système complet d'ingénierie linguistique et de la gestion des connaissances qui s'appliquent aux résumés de contenu des documents primaires de l'INA.

3 La problématique du modèle utilisateur dans la recherche d'information: *Quel potentiel, quelle souplesse, le système peut-il répondre de façon optimisée, non seulement à des requêtes utilisateurs, mais spécifiquement à des besoins qui changent dans le temps et dans l'espace ?*

Dans le contexte d'utilisation d'un SRI augmenté de fonctionnalités pour l'aide à la décision, nous privilégions, ici, la modélisation de l'utilisateur : décideur, veilleur.

4 La problématique de l'annotation d'un document : *Comment personnaliser la lecture d'un document et la mettre en valeur ?*

La construction de classes de données (sélection, traitement et calcul informationnels) s'avère nécessaire pour les descriptions structurées des documents dans les banques de données numériques, les bases de données bibliographiques ou les bases de connaissances [RÖSCHEISEN & al., 03]. Une approche pour décrire un modèle d'annotation sur les documents est proposée.

Notre contribution se distingue par une différenciation de la représentation de l'utilisateur et celle des SRI, qui est susceptible d'interagir avec des modèles dédiés à l'IE [MARTINET, 95], [MARTRE, 94] en terme de processus. En vertu de la présentation des problématiques liées aux SRI et sur l'évolution des processus de gestion de documents (primaire, secondaire, annotations), nous présenterons dans ce qui suit les aspects fonctionnel et opérationnel de chacune énoncée et de ses interactions par construction de modèles.

2. Problématique de l'exploitation de l'information numérique

Depuis l'invention de l'écriture dans la société de Mésopotamie et de Chine, les *systèmes d'écriture* ont singulièrement modifié l'histoire de la connaissance et de la transmission des savoirs. Pour l'anthropologue Jack Goody [GOODY, 98], l'*écriture alphabétique*, en particulier, aurait permis aux hommes d'analyser leur propre discours grâce à la forme continue qu'elle donnait au message qu'il soit oral ou écrit. Ainsi, les *systèmes de communication* sont en rapport direct avec ce que l'homme peut faire de son monde à la fois « *interne* » en terme de pensées et « *externe* » en terme d'organisations culturelle, sociale et économique.

Dans le contexte de l'analyse de contenu multimédia et particulièrement la recherche d'un lien commun entre le graphique (l'image), l'audio (la parole) et le texte (l'écrit), il était évident de remonter dans le temps et de rechercher les origines de l'activité langagière, qui est source des connaissances et des savoirs partagés de l'humanité, afin de retrouver le vecteur commun entre les différentes formes d'expression.

Dès lors, nous convenons à émettre un avis sur les différents canaux communicationnels comme sources de connaissances et l'établissement d'un lien commun entre eux :

1 les arts graphiques donnent accès à la lecture et à l'écriture : systèmes de correspondance entre la langue, la graphie et les signes figuratifs ;
2 l'écriture graphique (ou imagée) est la source des écritures syllabiques et alphabétiques ;
3 l'écriture a ses racines dans les arts graphiques ;

4 le système graphique réussit à doubler et à amplifier le système linguistique : la correspondance sémantique entre le mot, le signe et la correspondance phonétique ;
5 les sources de connaissances visuelles ne sont pas indépendantes et infranchissables des sources langagières, en exemple : l'apport de la rhétorique.

Par le biais des mécanismes de correspondance entre ces canaux, l'analyse des documents multimédia [GUARINO, WELTY, 00a-b] subordonnés à la représentation de leur contenu textuel, constitue une nouvelle source de connaissances, que nous considérons une source complémentaire à la nature du document primaire. Dans un tel environnement, une synthèse professionnelle sur l'analyse de contenu « documentaire » s'avère nécessaire. Accommodement, les technologies de l'information et de la communication permettant la manipulation et l'extraction automatiques de connaissances seront amenées à jouer un rôle essentiel dans la société de l'information [RICH, 83].

Nous étalons, dans ce qui suit, un bref historique des évolutions qui ont affecté ces sources de connaissances et les difficultés liées à l'analyse du multimédia [LUSTIÈRE, 99], [SIDHOM, 02].

En 1970, l'ORTF (Office national de la Radio et Télévision Française) charge le cabinet d'organisation SEMA (Documentation et gestion) et ses experts de mettre au point une méthode d'analyse afin de gagner du temps sur la préparation de l'indexation des films d'actualités [DEGEZ, 70]. L'analyse plan par plan, appelée « *analyse chronologique* », se révèle indispensable pour situer le sujet dans son contexte et faciliter la réponse à la demande. La grille d'analyse chronologique mise au point à cette époque est utilisée jusqu'à présent à l'INA et est étendue à l'ensemble de documents multiformes [LUSTIÈRE, 99], (*cf.* Figure 2a.).

On peut y relever des informations élaborées, les niveaux d'analyse avec le concours de cinémathécaires sur un document donné :

1 la position des plans et la description plan par plan,
2 les mouvements de caméra,
3 les personnes identifiées et les lieux de l'action,
4 la distinction entre image et son ou d'autres média.

NATURE DE PRODUCTION	DE	DES	RES	SEQ
PRODUCTION PROPRE CODE 01	Descripteurs thématiques et géographiques, personnes morales, personnes physiques évoquées.	Descripteurs séquences et images. Nom des villes si images réutilisables. Nom des personnes si visibles à l'image.	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Description par séquence du sujet.

Figure 2a. Présentation d'une Grille d'Analyse : « PRODUCTION PROPRE ».

Par construction, on trouve deux types de notices documentaires :

1 la notice d'émission : une notice documentaire d'une émission isolée, d'un sujet de journal télévisé ou d'un reportage traité comme un document,
2 la notice d'ensemble : une notice documentaire de présentation et d'historique d'un ensemble d'émissions.

Chaque notice élaborée dispose, en partie, par la méthode d'analyse chronologique de plusieurs champs résumant le contenu : chapeau (résumé court), résumé (résumé détaillé), séquences (principaux évènements en relation avec des personnages, des lieux, des actions, etc.), descripteurs primaires et secondaires (à partir des thésaurus INA), titres (propre, thématique), le résumé producteur (auteur, maison de production) du document en question s'il est joint (*cf.*

Figure 2b.).

Titre propre :	ECLIPSE : CONSEILS SECURITE ROUTIERE
Titre collection :	19/20
Descripteurs principaux :	FRANCE ; POLITIQUE INTERIEURE ; ECLIPSE ; SOLEIL ; PREVENTION ; CAMPAGNE D'INFORMATION ; SECURITE
Descripteurs secondaires:	LUNETTES ; HOMME POLITIQUE ; FEMME ; MINISTRE ; DEMESSINE MICHELLE ; AUBRY MARTINE ; BUFFET MARIE GEORGE ; GAYSSOT JEAN CLAUDE ; LIEBART BERNARD ; ANIMATION ; CIRCULATION ROUTIERE
Producteurs (aff) :	PRD, PARIS : FRANCE 3 (F3), 1999
Nature de production (aff):	PRODUCTION PROPRE
Résumé :	Présentation des mesures de sécurité recommandées, notamment en matière de circulation routière, lors de l'éclipse solaire du 11 août prochain.
Séquences :	- Sur le perron de l'Observatoire de PARIS, quatre MIN (Michèle DEMESSINE, Martine AUBRY, Marie-George BUFFET et Jean-Claude GAYSSOT), posent avec des lunettes "spéciales éclipse" / conférence de presse. - SYNTHE CARTE montrant l'amplitude de l'éclipse. - DP circulation sur autoroute.

Figure 2b. Notice obtenue par la Grille d'Analyse : « PRODUCTION PROPRE ».

Ainsi, la consultation du document primaire se fera immédiatement après l'interrogation et la lecture des résumés dans la notice (document secondaire), voire même la consultation de la séquence cherchée (écoute, visualisation ou lecture) [BITITICI & al., 97] .

Par extension à l'étude de la problématique posée, nous avons cherché à étudier la stabilité des descriptifs textuels (résumés), tout particulièrement ceux développés à l'INA (sources de INAthèque et INAactualités), afin d'établir par une analyse statistique les composantes grammaticales et syntaxiques de la phrase, comme prévision au processus d'analyse morpho-syntaxique. Ce cadre de recherche fera l'objet de la problématique suivante.

3. Problématique du traitement de l'information secondaire

Les experts et professionnels de l'INA ont eu comme mission la constitution d'un savoir multimédia à transmettre pour les futures générations. Pour atteindre cet objectif, ces derniers pratiquent sur le document numérique un certain nombre d'opérations intellectuelles. Il s'agit de l'analyse documentaire regroupant l'analyse chronologique, le catalogage spécifique, l'enrichissement du contenu, etc. à appliquer sur le multimédia et la constitution d'outils terminologiques et thésauriques.

Pour remédier à la méthode INA en matière de gestion et de recherche d'information, qui sont basées sur la recherche par mots-clés soit spécifiquement à des champs de notice ou en texte intégral, nous avons développé le système SIMBAD (ie. « ***S****ystème d'****I****ndexation du* ***M****ultimédia* ***B****asé sur l'****A****nalyse* ***D****ocumentaire* ») pour l'indexation de documents multimédia à partir des représentations textuelles associées (résumés). Il regroupe principalement le *parseur* morphosyntaxique (PARSE_Fr) associé à des outils dictionnairiques électroniques (DIC_Fr), le noyau d'indexation et de recherche d'information (SRI_Fr) pour la gestion des connaissances autour du syntagme nominal (SN).

3.1. Système d'analyse automatique : SIMBAD

Le côté expérimental de ce travail de recherche nous a amené à évoluer vers un système complet d'ingénierie linguistique, la gestion et le management des connaissances [CHIN, 89]. Pour l'ingénierie linguistique, le modèle théorique mis en application a permis l'identification des

syntagmes SN [SIDHOM, HASSOUN, 03] , tout en mettant en évidence les transitions entre les mots du lexique (prédicats libres) et les SN (prédicats liés) qui pointent sur des objets de la réalité extra-linguistique.

Ce processus de transitions s'effectue à travers l'identification de structures syntaxiques dans le texte du document, en repérant les SN. Ce modèle a été conçu en ayant les objectifs suivants [SIDHOM, 02] :

1 identifier les SN dans le contexte d'analyse : textes de résumés dans les notices ou ceux de requêtes,
2 déterminer la structure SN en mettant en évidence les relations entre ses constituants.
3 permettre l'entreposage des représentations SN et leur relation, puis augmenter les fonctionnalités sur les SN pour la recherche d'information,
4 faciliter le mécanisme de passage de la logique intensionnelle (les mots qui appartiennent au lexique de la langue : prédicats libres) à la logique extensionnelle (les unités à valeur référentielle ou les SN : prédicats liés).

Dans l'approche logico-sémantique du modèle SN, ce dernier est en effet *« l'unité minimale du discours qui permet de désigner un objet»* [LE GUERN, 89]. Nous sommes donc confrontés à deux ordres logiques afin que le SN se définisse comme étant la plus petite unité d'information porteuse d'une valeur référentielle, à savoir [LE GUERN, 91] :

5 La logique intensionnelle qui est « *sans référentiel et sans classe, constituée de relations (entre les mots) et de propriétés (du mot) envisagées indépendamment de quelque objet que ce soit* ». Il ne s'agit que des propriétés, isolés et sans références du mot, en tant qu'unité du lexique hors discours.
6 Alors qu'au niveau de la logique extensionnelle, le mot (prédicat libre) prend ses valeurs sur un univers du discours, là on peut envisager une classe d'objets « *à la possibilité de déterminer des classes, au moins virtuelles ; c'est le basculement de la logique intensionnelle à la logique extensionnelle ; c'est la mise en relation des mots et des choses* ».

Le mécanisme d'analyse automatique des textes résumés de l'INA s'est concrétisé par la conception d'un noyau d'indexation automatique. Le noyau d'indexation se scinde en trois modules (*cf.* Figure 3.1.) :

1 **Module 1** : la conception de différents outils automatiques servant à l'analyse morphologique du langage naturel. Il s'agit des ressources linguistiques composées essentiellement par les dictionnaires électroniques (DIC_Fr) et la grammaire de réécriture des éléments syntagmatiques (GRAM_Fr) et celle du modèle de la phrase résumé (*cf.* Figure 3.1. a) ;
2 **Module 2** : l'implémentation de l'analyseur morphosyntaxique (PARSE_Fr) par la compilation des règles de réécriture, GRAM_Fr, pour l'analyse des corpus textuels (*cf.* Figures 3.1. a,b) ;
3 **Module 3** : l'extraction des SN à partir des arbres syntagmatiques des phrases analysées (*cf.* Figures 3.1.b,c). L'architecture de l'analyseur est fondée sur les automates à transitions augmentées (ATN) et en cascade (CATN) de W. Woods [WOODS, 80-97]. C'est à partir des objets décorés (*cf.* Figure 3.1.b) que le processus de filtrage automatique des SN s'opère.

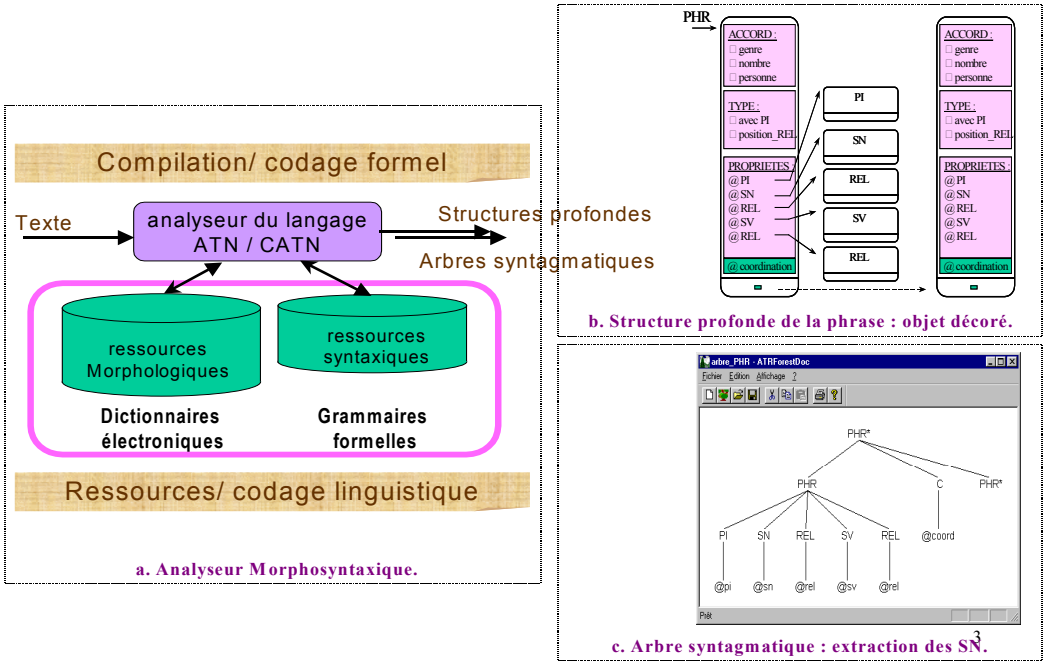
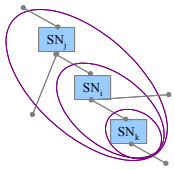


Figure 3.1. Architecture de l'analyseur morphosyntaxique : PARSE_Fr.

3.2. Système de recherche d'information : SRI_Fr

Les syntagmes nominaux ont une organisation naturelle. Dans un sens, ils ont un rapport d'*emboîtement* les uns avec les autres, ce qui permet de les classifier linéairement en des niveaux informationnels distincts :

Logique d'emboîtement	Conditions	Schéma : graphe linéaire
$\forall SN_i, \exists \{ SN_j \wedge SN_k \} /$ $\{ SN_j \supset SN_i \supset SN_k \}$;	*avec :* $SN_i = \{ non\varnothing / saturé / non saturé \}$; $SN_j = \{ \varnothing / saturé / non saturé \}$; $SN_k = \{ \varnothing / non saturé \}$.	

En exemple :

« Elise Lucetti donne le thème de l'émission consacrée aux OGM ».

= « Elise Lucetti donne [le thème de [l'émission consacrée à [les OGM]^{SN_1}] ^{SN_2}] ^{SN_3} ».

Ainsi : $SN_3 \supset SN_2 \supset SN_1$;

avec :

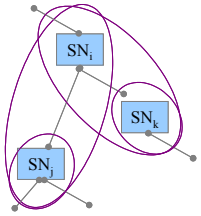
SN_3 = « *le thème de l'émission consacrée aux OGM* », *qui est un SN saturé* ;

SN_2 = « *l'émission consacrée aux OGM* », *qui est un SN non saturé* ;

SN_1 = « *les OGM* », *qui est un SN non saturé* ;

Et dans l'autre, ils ont un rapport de ramification, dans le cas où le syntagme nominal se retrouve

dans un schéma arborescent. Cette dernière propriété permet d'ordonner et de distinguer les classes d'informations (structure d'arbre des classes d'informations) :

Logique d'arborescence	Conditions	Schéma : graphe d'arbre
$\forall SN_i, \exists \{ SN_j \wedge SN_k \} /$ $\{ SN_i \supset SN_j \} \wedge \{ SN_i \supset SN_k \}$;	*avec :* $SN_i = \{$non $\varnothing$ / saturé / non saturé$\}$; $SN_j = \{$ non $\varnothing$ / non saturé$\}$; $SN_k = \{$ non $\varnothing$ / non saturé$\}$.	

En exemple :

« l'annonce contre la DCA allemande du débarquement allié du 6 juin 1944 ».

« [l'annonce contre [la DCA allemande] $^{SN}{}_{1g}$ de [le débarquement allié de [le 6 juin 1944] $^{SN}{}_{1d}$] $^{SN}{}_{2d}$] $^{SN}{}_{max}$ ».

Ainsi:

$\{SN_{max} \supset SN_{2d} \supset SN_{1d}\}$ $\wedge$ $\{SN_{max} \supset SN_{1g}\}$;

avec :

SN_{max} = « *l'annonce contre la DCA allemande du débarquement allié du 6 juin 1944* » ;

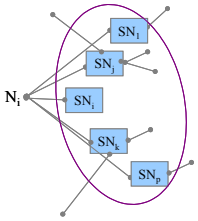
SN_{1g} = « *la DCA allemande*» ;

SN_{1d} = « *le débarquement allié du 6 juin 1944* » ;

SN_{2d} = « *le 6 juin 1944* » ;

Ces caractéristiques logiques permettent de construire une architecture de gestion des connaissances, qui exploite les informations autour du SN, au moyen de la navigation dans des structures d'arbres ou des emboîtements. Par la superposition de ces deux logiques, la navigation entre les SN s'intègre dans une architecture d'un treillis de connaissances.

A ces deux aspects logiques de navigation, le centre nominal (N) dans un SN est un élément appartenant à la logique intensionnelle. N ne peut construire un objet de discours, mais comme trait d'une classe pour accéder à ses éléments. Ce prédicat libre N va contribuer à la description d'une classe d'objets <SN> : nous le situons comme la clé d'accès à cette classe ayant comme attribut commun le trait <N>.

Logique d'appartenance	Conditions	Schéma : classe de SN
$\forall SN_i, \exists !\ N_i$ / $\{ N_i \in SN_i \} \wedge$ $\{ \exists !\ SN_{1..k} / N_i \in \{ SN_{1..k} \} \}$;	*avec :* $SN_i = \{$non $\varnothing$ / saturé / non saturé$\}$; $SN_{1..k} = \{$ saturé / non saturé $\}$.	

En exemple :

« un envoie de transport de troupes américain est attaqué par les escadrilles japonaises ».
= « [un envoie de transport de [des troupes] américain] est attaqué par [les escadrilles japonaises] ».
Ainsi les { $N_i \in SN_i$ } :

envoie $\in$ un envoie de transport de troupes américain $\supset$ des troupes américain
troupe $\in$ des troupes américain
escadrille $\in$ les escadrilles japonaises

Dans ce contexte, l'indexation d'un document (par sa représentation textuelle) ou celle d'une requête passent par le même mécanisme d'analyse dans SIMBAD.

Ainsi, le schéma d'interrogation (*cf.* Figure 3.2.), dans le système de recherche d'information SRI_Fr, consiste à retrouver les SN d'une requête (SN-requête) qui sont présents dans la base de documents (SN-base). Bien entendu, les documents qui répondent le mieux à la requête sont ceux identifiés par des SN saturés, bien moins que par ceux identifiés par les SN non-saturés, et moins par ceux identifiés par les de prédicats intensionnels (N) ou leur synonyme (N_s).

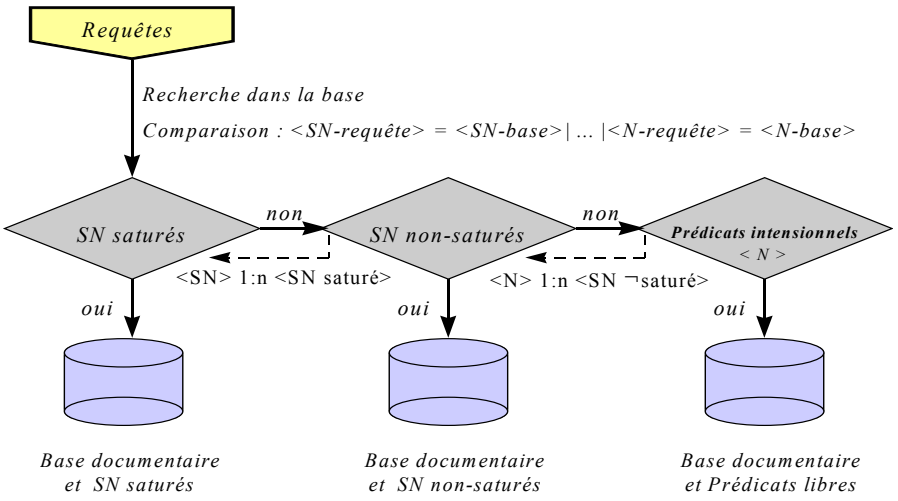


Figure 3.2. Schéma d'interrogation de la base de connaissances : SRI_Fr.

4. Problématique du modèle utilisateur dans la recherche d'information

La modélisation de l'utilisateur se donne comme objectif de pouvoir personnaliser les réponses du SRI. Il s'agit de formaliser la façon de représenter un utilisateur et ses comportements. Cela concerne également la façon d'exploiter les connaissances dont nous disposons à son sujet. Trois catégories de modèle sont proposées [DAVID, 99] :

(a) Le **profil de l'utilisateur** : l'utilisateur est associé à sa requête, qui exprime ses besoins informationnels. Dans ce contexte, son besoin est relativement stable. Le profil est appliqué aux nouvelles informations afin de lui proposer celles qui sont pertinentes.

(b) Le **modèle implicite** de l'utilisateur : son comportement et ses préférences sont déterminés d'une manière implicite. Par exemple, la visualisation d'un document par l'utilisateur envisage l'interprétation comme une adéquation du document par rapport à sa requête.

(c) Le **modèle explicite** de l'utilisateur : son comportement et ses préférences sont également représentés mais selon ses spécifications. Par exemple, si l'utilisateur visualise un document,

cela n'envisage pas l'adéquation du document par rapport à sa requête, cas (b), sauf s'il indique son opinion sur le degré de pertinence du document visualisé.

L'exploitation d'un profil utilisateur (a) est généralement individualisée. Les modèles implicite ou explicite, (b) et (c), peuvent être traités par la méthode de **stéréotype**. Par la technique de stéréotypage, les utilisateurs sont regroupés dans des classes puis une interprétation est appliquée à tous les utilisateurs d'une classe donnée.

La représentation des paramètres cognitifs sur les utilisateurs formalise, par exemple, des paramètres nécessaires pour connaître le niveau de connaissances d'un utilisateur. Ces derniers paramètres permettent une meilleure interprétation de sa requête et nécessite la sauvegarde du modèle de l'utilisateur au travers ses sessions individualisées. Ainsi, la notion de processus cognitif réside dans une représentation qui intègre potentiellement les problématiques de l'utilisateur, en privilégiant les interactions dynamiques et évolutives de ce dernier.

Le modèle cognitif de l'utilisateur est construit sur des phases cognitives identifiées dans un processus d'apprentissage humain. Quatre phases, qui correspondent à des niveaux d'habitudes évocatives, ont été intégrées dans le modèle :

(i) **La phase d'observation** : l'apprenant prend connaissance de son environnement par le processus d'observation ;

(ii) **La phase d'abstraction élémentaire** : l'apprenant désigne les objets observés par des mots, ce qui correspond également à une phase d'acquisition de vocabulaire ;

(iii) **La phase de symbolisation et de raisonnement** : l'apprenant emploie des vocabulaires spécialisés qui relèvent d'un niveau d'abstraction des concepts élevés. Par exemple, quelqu'un à un niveau d'abstraction bas peut dire « je vois un poisson », mais ne peut pas dire « je vois un piscivore » (i.e. un oiseau qui mange des poissons) ;

(iv) **La phase de créativité** : l'apprenant découvre et s'approprie des connaissances qui ne sont pas présentées d'une manière explicite dans le système. C'était le cadre d'expérimentations du prototype BIRDS [DAVID, 99].

Le modèle de l'utilisateur a permis de proposer une architecture SRI qui repose sur l'évolution cognitive de l'utilisateur. L'architecture permet à l'utilisateur d'explorer la base d'informations [FROISSART, 01] pour découvrir son contenu, de formuler des requêtes, d'effectuer des annotations et de lier des activités de recherche à un besoin d'information, pour des objectifs décisionnel ou stratégique.

La contribution de cette étude sur un SRI dans le domaine de l'IE, est la proposition d'une architecture fonctionnelle : EQuA^{2}te (ie. « ***E**xplore, **Qu**ery, **A**nalyse, **A**nnota**te** ») [DAVID, THIERY, 02] et la gestion, l'exploitation de la base de données numériques, la base de connaissance sur les utilisateurs.
Nous explicitons les quatre modules de EQuA^{2}te [THIERY, DAVID, 03] (*cf.* Figure 4.2.) :

1 " **Explorer** " l'entrepôt de données, c'est naviguer dans les données ; Par exemple, l'outil COGNOS propose deux modules d'exploitation des données : –" Explorer "– permet d'explorer les données, –"Reporter "– permet de faire des interrogations et des rapports sophistiqués sur l'entrepôt ;
2 " **Interroger** " l'entrepôt, c'est utiliser des requêtes prédéfinies qui sont proposées par des outils associés (COGNOS, SQL) et de poser des requêtes classiques. Les entrepôts étant pour la plupart relationnels ;

3 " **Analyser** " c'est utiliser des techniques de fouille de données pour extraire de nouvelles connaissances de l'entrepôt. De tels outils comme *Scenario* de COGNOS ou *Enterprise Miner* de SAS ;
4 " **Annoter** ", seulement dans ce module, il s'agit de prendre des décisions et de les entrer dans la base de connaissances du système d'IE.

Pour illustrer, un décideur pouvait être en mesure d'explorer l'historique des recherches d'information dans l'entrepôt de données pour déceler si un cas similaire à son besoin informationnel s'est exprimé. Par le biais de l'architecture EQuA^{2}te, ce décideur aura la facilité d'explorer, d'interroger, d'analyse et d'annoter la base de connaissances sur ses activités, éventuellement, celles des autres utilisateurs. Par extension de ce processus, cette base de connaissances serait partie intégrante de la modélisation de l'utilisateur.

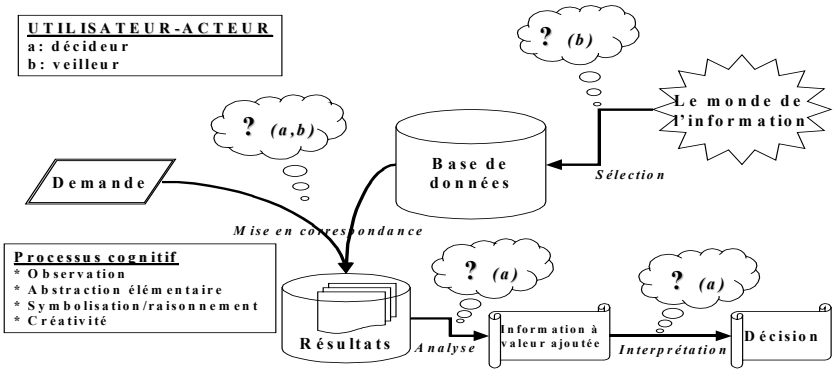


Figure 4.1. Architecture d'un système d'intelligence économique [DAVID, 99].

Le même principe peut s'appliquer au veilleur dans le contexte d'IE [JAKOBIAK, 92-95], qui est souvent confronté à des problèmes comparables à d'autres cas passés. Il peut explorer, interroger, analyser et annoter des solutions antérieures afin d'associer à son problème présent les recherches d'information rencontrées. De manière explicite, nous pouvons considérer que la base d'information d'un système d'IE (*cf.* Figure 4.1.) est l'entrepôt de données [THIERY, DAVID, 03] : des connaissances sur les documents [RICH, 83], des connaissances sur les utilisateurs et sur le domaine 'métiers' [DAVID, 99].

Ainsi, dans le contexte d'exploration de contenus et de recherche d'information, les modules d'EQuA^{2}te et ceux de SIMBAD peuvent rentrer en phase de collaboration entre processus dans le contexte d'IE (*cf.* Figure 4.2.). La mise en correspondance entre les requêtes et les documents sera aussi bien l'analyse par une approche issues des traitements automatiques des langues naturelles, que par une approche fonctionnelle sur la modélisation de l'utilisateur.

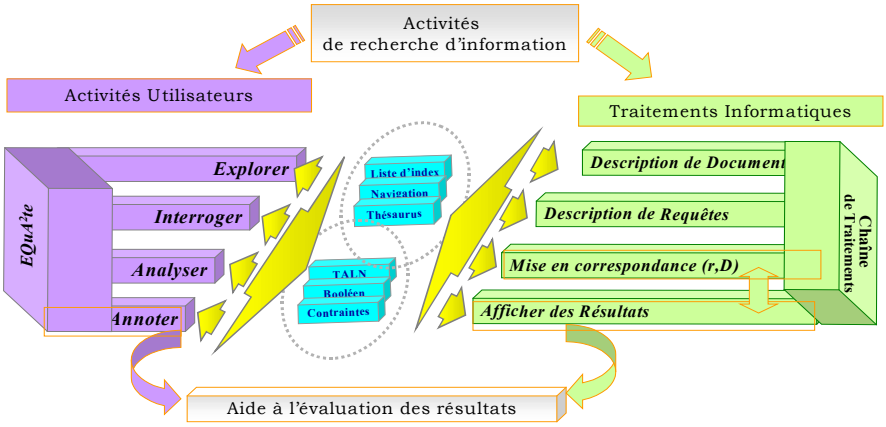


Figure 4.2. Architecture EQuA^{2}te dans le processus de recherche d'information.

5. Problématique du modèle d'annotation d'un document

Le modèle d'annotation AMIE (ie. « ***A**nnotation **M**odel for **I**nformation **E**xchange* ») est la conjonction de paramètres caractéristiques à l'annotation de documents par rapport à un SRI. Les processus qui s'associent dans la communication de l'information au modèle d'annotation se retrouvent au niveau d'autres modèles : EQuA^{2}te et SIMBAD.

Ainsi, Le modèle AMIE disposera parmi ses fonctions la résolution de problèmes liés aux échanges de l'information entre les acteurs d'un système d'IE [SALLES, 00] en terme de besoins informationnels et informations décisionnelles [BOUAKA, DAVID, 03], aussi bien les utilisateurs d'un SRI [DESMONTILS & al., 03] en terme d'informations calculées et information à valeur ajoutée.

En conséquence, les fonctionnalités suivantes sont considérées comme partie intégrante du modèle d'annotation, en cours de réalisation :

1 **sélection** : sélection de critères à utiliser pour décrire les activités communicationnelles ;
2 **collection** : enregistrement et stockage des observations avec estimation des interventions liées à la communication entre les acteurs d'un système (IE ou SRI) ;
3 **explication** : informations additionnelles sur des critères *sélectionnés* (sélection). Un critère est définit comme un ensemble d'attributs et de valeurs ;
4 **présentation** : création de sorties (visualisation) à partir des informations *enregistrées* (collection) ;
5 **enregistrement** : sauvegarde des activités d'un acteur en phase d'annotation ;
6 **communication** : transmission et réception d'informations annotées entre les acteurs d'un système ;
7 **calcul** : manipulation de données et de fonctions numériques pour décrire une information à valeur ajoutée ;
8 **qualification** : attribution d'une qualité (valeur calculée, information complémentaire)

à l'information traitée ;

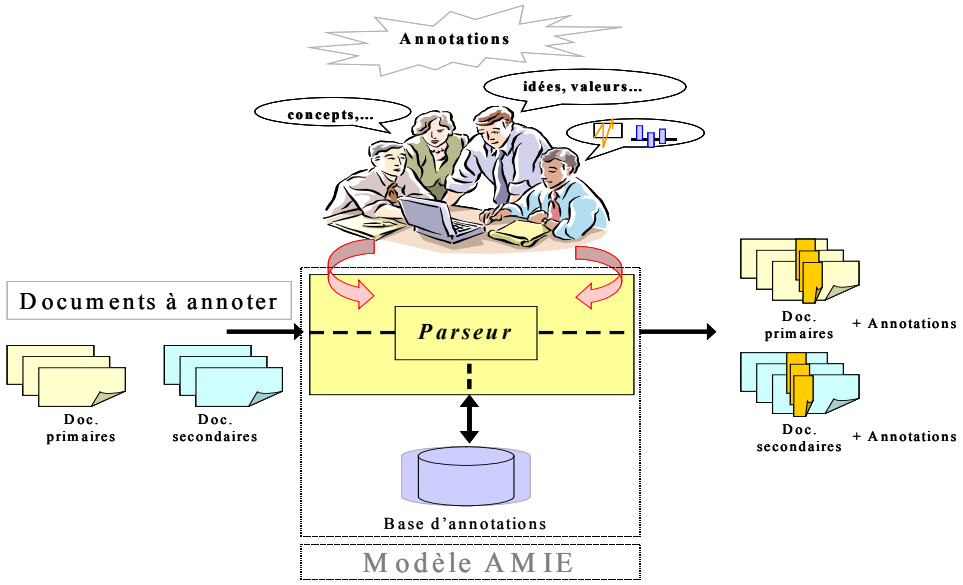


Figure 5. Processus d'annotations : modèle AMIE.

Dans la construction de ce modèle, il est question de définir des classes de données (sélection, traitement, calcul informationnel) et celles de descriptions structurées à partir de données numériques ou informationnelles. La résultante des classes construites n'est autre que de l'information annotée [MATTHEW & al., 96], [HECK & al., 03]. Cette dernière se caractérise par des segments additionnés au contenu du document (primaire, secondaire ou traité par EQuA^{2}te / SIMBAD) et ainsi, il est augmenté avec de la valeur ajoutée [THIERY, DAVID, 02], [DAVID & al., 01] suite au processus d'annotation. Cette voie de recherche permet d'interconnecter AMIE aux autres modèles opérationnels. Ainsi, ce modèle prend place dans la <boîte à outils> avec ses fonctionnalités (*cf.* Figure 6).

L'architecture du modèle AMIE (*cf.* Figure 5.) se retrouve entraîner dans la chaîne de traitements de l'information et vers l'information à valeur ajoutée pour la prise de décision.

6. Conclusion

Nous avons différencié la notion de l'utilisateur à celle du processus de recherche d'information dans le contexte de l'information numérique. Ainsi, les techniques sont dissociées des tâches organisationnels dans cette *architecture logicielle*, afin de contribuer à la résolution de certaines complexités dans la conception de nos modèles.

La gestion de l'information coordonnée avec le processus d'interprétation dans un SRI ont permis d'intégrer la complexité relative au mécanisme de représentations de contenu sur les documents numériques. Dans cette démarche, il est question d'affiner un raisonnement aux "frontières" des modèles et des problématiques (SRI, Modèle utilisateur et IE), en vue de formaliser les éléments d'un dialogue pluridisciplinaire.

Le thème " recherche d'information " nous a fournit un cadre d'analyse sur des problèmes liées au document numérique et l'évolution de sa gestion : modélisation de l'utilisateur (veilleur, décideur), information à valeur ajoutée dans la prise de décisions. Dans ce travail, les modèles,

leur expérimentation et leur synchronisation, offrent la dimension architecturale pour un processus d'IE. Pour nous, le concept d'IE s'affirme, puisqu'il s'agit d'étudier les processus impliqués dans la production des indicateurs interprétables avec prise de décisions sur des informations internes et externes à l'organisation en question.

Par intégration à cette architecture, évolutive dans sa construction et qui formalise un environnement multidisciplinaire, les modèles EQuA^{2}te, SIMBAD et AMIE (*cf.* Figure 6.) vont contribuer dans un avenir proche à la visualisation des sujets de recherche de l'utilisateur (recherche antérieure réactivée, mise à jour ou synchronisée avec d'autres), aussi bien à la représentation d'une carte conceptuelle des informations disponibles dans un entrepôt de données.

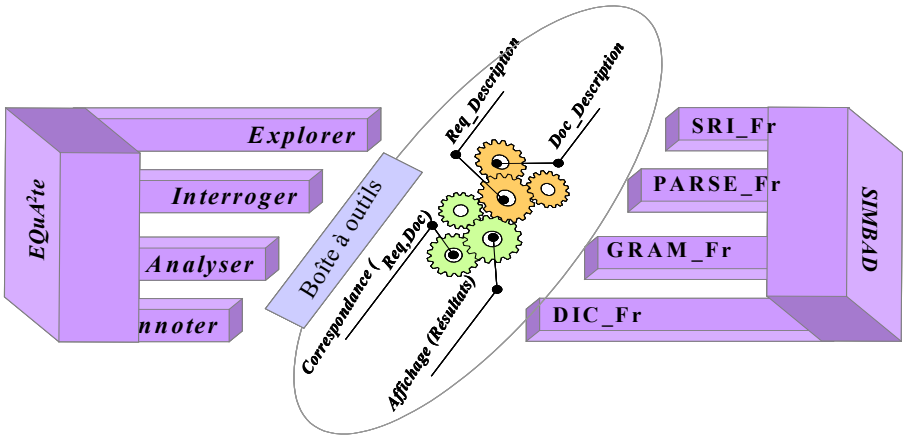


Figure 6. Synchronisation entre les modèles de traitement de l'information.

L'ensemble dispose de possibilités architecturale et fonctionnelle dans un tel système organisationnel, qui se donne comme objectifs d'affiner la recherche d'information, de synchroniser différentes applications dans le Management de l'information et de connaissances et d'intégrer l'utilisateur soit par sa modélisation ou par son influence sur le contenu d'un document (annotation). L'information en question n'a pas cessé de muter, du document primaire au document secondaire au document annoté et enfin au document à valeur ajoutée. L'information s'enrichit par les différents procédés intellectuel, automatique et semi-automatique tout en restant dans l'*environnement du numérique*.

7. Bibliographie

[ALQUIER, 00] : ALQUIER Anne-Marie (2000), « Quelques principes méthodologiques pour la conception de Systèmes d'Information d'Intelligence Economique en fonction des exigences en aide à la décision », Revue d'Intelligence Economique, N° 6-7, Association Française pour le Développement de l'Intelligence Economique, Oct. 2000.

[BACHIMONT, 99] : Bachimont, B. (1999), " Engagement sémantique et engagement

ontologique : conception et réalisation d'ontologies en ingénierie des connaissances.", In J. Charlet, M. Zacklad & G. Kassel (Eds.), Ingénierie des connaissances, Paris : Eyrolles.

[BITITICI & al., 97] : Bititici U.S., Carie A.S., McDewitt L. (1997), "Integrating performance measurement systems : a development guide", International Journal of Operations and Production Management, Vol. 17,N°5, pp. 522-534.

[BOUAKA, DAVID, 03] : Najoua BOUAKA et Amos DAVID. Modèle pour l'Explicitation d'un Problème Décisionnel: Un outil d'aide à la décision dans un contexte d'intelligence économique. *in Conférence "Intelligence Economique : Recherches et Applications"*, Nancy :14-15 avril 2003.

[BUENO & al., 2001] : Bueno David, Conejo Ricardo, Carmona Cristina and David Amos (2001). METIORE: A Publications Reference for the Adaptive Hypermedia Community. Hypermedia : Openness, Structural Awareness, and Adaptivity. (Aarhus, Denmark). 2001.

[CHIN, 89] : CHIN David N., User Models in dialog systems. Berlin : Springer, 1989.

[CHIN, KNOME, 86] : CHIN David N., KNOME : Modeling What the User Knows in UC. In UM 86: First International Workshop on User Modeling, 1986.

[DAVID & al., 01] : David, Amos and Bueno, David and Kislin, Philippe. Case-Based Reasoning, User model and IRS. In The 5th World Multi-Conference on Systemics, Cybernetics and Informatics - SCI'2001. International Institute of Informatics and Systemics (IIIS). (Orlando, USA). 2001.

[DAVID, 99] : David, Amos. Modélisation de l'utilisateur et recherche coopérative d'information dans les systèmes de recherche d'informations multimédia en vue de la personnalisation des réponses. Mémoire HDR, Université Nancy 2, Mai 1999.

[DAVID, THIERY, 02] : David, Amos and Thiery, Odile. Application of "EQuA^{2}te" Architecture in Economic Intelligence. In Information and Communication Technologies applied to Economic Intelligence - ICTEI'2002. (Ibadan, Nigeria). 2002.

[DEGEZ, 70] : Danièle Degez. Conclusions de l'étude expérimentale des méthodes d'analyse et d'indexation SEMA-ORTF. Publications Archives audiovisuelles de la télévision,Novembre 1970.

[DESMONTILS & al., 03] : E. DESMONTILS, C. JACQUIN & L. SIMON. « Dinosys : un outil d'annotation pour l'enseignement à distance sur le Web », *in Colloque "Miage et e-mi@ge" : Méthodes Informatiques Appliquées à la Gestion des Entreprises, Marrakech 15-17 Mars 2004*, (URL visité : 11/2004) http://e-miage.ups-tlse.fr/colloque/papiers/E.DESMONTILS.pdf

[FROISSART, 01] : Froissart C. « De la communication homme-machine à la recherche d'information dans la documentation technique ». Mémoire pour l'Habilitation à diriger des Recherches, Université Jean Monnet – Saint-Etienne, 2001.

[GUARINO, WELTY, 00a] : Guarino N., Welty, C. A Formal Ontology of Properties. in Dieng, R., and Corby, O., eds, Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management. Spring-Verlag LNCS. October, 2000.

[GUARINO, WELTY, 00b] : Guarino N., Welty, C. Ontological Analysis of Taxonomic Relationships. in Laender, A. and Storey, V. eds, Proceedings of ER-2000: The 19th International Conference on Conceptual Modeling. Springer-Verlag LNCS. October, 2000.

[GOODY, 98] : Jack Goody. De l'oral à l'écrit. Propos recueillis par Nicolas Journet, in Sciences Humaines, Mai 1998, N°83, p.38-41.

[HECK & al., 03] : Rachel M. Heck, Sarah M. Luebke, Chad H. Obermark, « A Survey of Web

Annotation Systems ». *Work supported by Grinnell College Noyce Science Summer Research Fund*, (URL visité : 11/2004)

http://www.math.grin.edu/~rebelsky/Blazers/Annotations/Summer1999/Papers/.

[JAKOBIAK, 92] : F. JAKOBIAK. Exemples commentés de veille technologique. Les Edition d'Organisation, 1992.

[JAKOBIAK, 95] : F. JAKOBIAK. L'information scientifique et technique. Presses Universitaires de France, 1995.

[LE GUERN, 89] : Le Guern M. Sur les relations entre terminologie et lexique. in actes du colloque: les terminologies spécialisés - Approches quantitatives et logico-sémantique, et Meta Vol.34, No.3., sept. 89.

[LE GUERN, 91] : Le Guern M., Un analyseur morpho-syntaxique pour l'indexation automatique. Revue de linguistique française : Le Français moderne . n°1, juin 1991.

[LI, 96] : Li S.H. , « Precision and Recall of Ranking Information-Filtering Systems », Journal of Intelligent Information Systems, Vol(7) n°3 ,287-306, 1996.

[LUSTIÈRE, 99] : Colette Lustière. Les différents types de notices dans BASINA: les étapes du traitement documentaire à la vidéothèque d'actualités INA. Rapport technique (INA actualités).1999.

[MARTINET, 95] : B. MARTINET. L'intelligence économique. Les Editions d'Organisation, 1995.

[MARTRE, 94] : MARTRE, Henri, "Intelligence économique et stratégie des entreprises", Rapport du Commissariat Général au Plan, Paris, La Documentation Française, 1994.

[MATTHEW & al., 96] : Matthew A. Schickler, Murray S. Mazer and Charles Brooks. (*1996*). Pan-Browser Support for Annotations and Other Meta-Information on the World Wide Web. *in Fifth International World Wide Web Conference, 6-10 May, 1996 , Paris (France).* (URL visité : 11/2004) http://www5conf.inria.fr/fich_html/papers/P15/Overview.html.

[RICH, 83] : RICH E., Users are individuals: individualizing user models, International journal of Man-Machine Studies, Volume 18, p. 199-214, 1983.

[RÖSCHEISEN & al., 03] : Röscheisen Martin, Christian Mogensen and Terry Winograd. (1995) "Interaction Design for Shared World-Wide Web Annotations", in CHI '95 Proceedings, (URL visité : 11/2004)

http://www.acm.org/sigchi/chi95/Electronic/documnts/shortppr/cmn_bdy2.htm

[SALLES, 00] : Salles, M. (2000), « Problématique de la conception de méthodes pour la définition de Systèmes d'Intelligence Economique », Revue d'Intelligence Economique, N° 6-7, Association Française pour le Développement de l'Intelligence Economique, Octobre 2000.

[SIDHOM, 02] : SIDHOM, Sahbi. " Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherché d'information: de l'écrit vers la gestion des connaissances.", Thèse de Doctorat à l'Université Claude Bernard Lyon1, France, Mars 2002.

[SIDHOM, HASSOUN, 03] : SIDHOM Sahbi, HASSOUN Mohamed. « Morpho-sytactic Parsing for a Text Mining Environment ». In Official Journal « Knowledge Organization » KO. 29(2002) No. 3-4, Edited by Olson, Hope A. – Saranchuk, Georgina R. Zaharia, (c) 2003 Ergon Verlag.

[THIERY, DAVID, 03] : Thiery, Odile et David, Amos. L'architecture EQuA^{2}te et son application à l'intelligence économique. Conférence "Intelligence Economique : Recherches et

Applications" - IERA'2003. (INIST, France). 2003.

[WOODS, 80] : William A. Woods. Cascaded ATN Grammars. in American Journal of Computational Linguistics, January-March 1980, vol.6, n°1.

[WOODS, 97] : William A. Woods. Conceptual Indexing : a better way to organize knowledge. Technical Report SMLI TR-97-61 : SUN Micosystems, Lab. Mountain View Canada, April 1997.