

## **Sémiotique des Médias.**

### **Le genre du documentaire audiovisuel**

---

#### **Cours VI: Indexation, thesauri et ontologies**

---

**Peter Stockinger**

**Séminaire de DESS à l'Institut National des Langues et  
Civilisations Orientales (INaLCO)**

**Paris, 2001 - 2002**

## Sommaire

---

1) INTRODUCTION.....	3
2) DE L'INDEXATION.....	4
3) INDEXATION ET DÉCOUPAGE DE DOCUMENTS.....	7
4) LE THESAURUS.....	10
4.1) FONCTION ET RÔLE D'UN THESAURUS.....	10
4.2) PRINCIPAUX ÉLÉMENTS D'UN THESAURUS.....	12
5) L'ONTOLOGIE.....	21
5.1) TYPES D'ONTOLOGIES.....	21
5.2) CONCEVOIR ET ÉDITER UNE ONTOLOGIE.....	23

## 1) Introduction

---

Dans ce cours nous discuterons l'intérêt de la description thématique par rapport à :

- La problématique de l'indexation de documents (audiovisuels)
- L'outil "thesaurus" qui est très difficilement remplacé par d'autres outils dans la gestion de l'information
- L'outil "ontologie" qui repose sur des théories souvent assez sophistiquées en sciences du langage (sémantique) et en intelligence artificielle (représentation des connaissances, théories des agents, ...) mais qui a, actuellement, "le vent en poupe" ...

Dans un petit chapitre, nous reviendrons, de nouveau, sur la question du découpage - physique ou purement conceptuel - de documents (audiovisuels) dans une optique gestion de l'information .

## 2) De l'indexation

---

"Indexer" un document (audiovisuel) ou un corpus de documents (audiovisuels) veut dire, *proposer une description du document ou du corpus de documents*. D'une façon plus générale, "indexer" veut dire, expliciter une certaine organisation sémantique d'une ou d'un ensemble de ressources d'informations - l'organisation sémantique devant correspondre d'une part aux contraintes structurales (i.e. aux spécificités "internes") de ces ressources et d'autre part aux contextes et objectifs d'application, d'utilisation, d'exploitation.

La description sous forme d'indexation se fait sur la base d'un ensemble de thèmes préalablement définis et organisés en des catégories, structures de classification, etc. Les principaux "outils" pour ce procédé sont :

- Le vocabulaire "structuré"
- Le thesaurus
- L'ontologie
- La base de connaissances.

Le **vocabulaire "structuré"** correspond assez précisément à un dictionnaire spécialisé (un dictionnaire pour un domaine). Il est organisé suivant de grandes thématiques à l'aide desquelles on prétend obtenir une compréhension d'un domaine. A chaque grande thématique correspond un certain nombre d'expressions spécialisées, de termes. Ils - ces termes - sont organisés d'une manière alphabétique. Chaque terme constitue, en règle générale, l'entrée à un article dans lequel on trouve une définition du terme et, parfois, d'autres informations (linguistiques ou référentielles). Souvent, ces dictionnaires spécialisés sont composés, outre du vocabulaire stricto sensu, d'illustrations dans lesquelles sont référencées les différents termes. Citons ici l'exemple très intéressant du vocabulaire de l'architecture de Jean-Marie Pérouse de Montclos (Jean-Marie Pérouse de Montclos : Architecture. Méthode et vocabulaire. Paris, Imprimerie Nationale 1972).

Le **thesaurus** ressemble au vocabulaire mais lui ajoute des dimensions d'explicitation due notamment au fait qu'il positionne les termes (descripteurs) les uns par rapport aux autres à l'aide de relations de classification et d'autres indications d'utilisation d'un terme.

Nous y reviendrons. Citons comme exemples connus les différents thesaurii de l'UNESCO : le Unesco Thesaurus, développé sous la direction de Jean Aitchison (Paris, UNESCO 1977); les différents thesaurii, développé sous la direction de Jean Viet (UNESCO-MSH, Paris), tel que le Thésaurus international du développement culturel (Paris, UNESCO 1980) ou le Thesaurus pour le traitement de l'information en sociologie (Paris, UNESCO 1971)

Une **ontologie** ressemble à un thesaurus. Une ontologie est, selon les spécialistes tels que Gruber ou Guarini, un vocabulaire qui explicite sous forme de définitions une certaine conception d'un domaine. Par exemple, un domaine de connaissances tel que la description de sites web peut être conceptualisé selon des points de vue et des "intérêts" différents : description socio-économique, description technologique, description du point de vue du management d'un site, description sémiotique (structurale). Chaque conception possède sa "propre ontologie" (même si elle n'est pas disjointe des autres) : des "assumptions" à propos de l'objet, des connaissances, des théories, etc. Tout cela s'exprime obligatoirement à travers un *langage* qui, d'une manière la plus rudimentaire, est composé de *termes* (de mots) et de *règles de composition* de termes (une syntaxe). Autrement dit, même si deux conceptions partagent un sous-ensemble de termes, les termes en question acquièrent un sens particulier selon leur appartenance à une conception. Une ontologie est donc supposée rendre compte de ces différences en:

- a) reconstruisant le langage et la conception qui le sous-tend
- b) définissant les termes (le "vocabulaire") dudit langage
- c) classifiant les termes (ou, plutôt, le contenu des termes).

Ceci étant, il est, d'un point de vue empirique, souvent très difficile de tenir compte d'une manière satisfaisante du premier point cité ci-dessus de sorte qu'une ontologie ressemble bien souvent à une sorte de thesaurus.

Enfin, une **base de connaissances** est, traditionnellement, assimilée à des systèmes informatiques utilisant une expertise humaine formalisée pour résoudre certains problèmes dans des domaines circonscrits. On connaît, par exemple, les systèmes expert, les systèmes d'aide à la décision, certains systèmes plus proche au traitement du langage tels que les aides à la traduction, à la rédaction et ainsi de suite.

Il existe différents formalismes et systèmes de représentation pour décrire une certaine connaissance. Les plus populaires sont les graphes conceptuels (John Sowa), les frames et scripts (Schank & Abelson, Wilensky), les réseaux sémantiques, etc. Ils se ressemblent tous - certains reposent sur des théories formelles plus sophistiquées que d'autres, certains autres sont très proches de "langages" tout fait (tel que les différents KRLs - Knowledge Representation Language), etc.

Le point essentiel est que l'on essaie de décrire les connaissances d'un domaine sous forme de petites configurations ou réseaux (sémantiques). Certaines configurations sont considérées plus basiques que d'autres, on en dérive alors des configurations plus spécialisées, etc. En somme, c'est un peu comme si l'on considérait chaque taxème comme un petit réseau sémantique, comme un graphe conceptuel qui, tel quel est autonome d'autres taxèmes, d'autres graphes ou réseaux. Néanmoins cette "autonomie" n'est que partielle dans la mesure où tel ou tel spécifieur d'un taxème (tel ou tel noeud dépendant dans un graphe ou réseau) peut se trouver également dans un autre taxème (dans un autre graphe ou réseau). D'où justement la possibilité de comparer plusieurs taxèmes, de passer d'un taxème à un autre, de sérier des taxèmes en des hiérarchies partielles, etc - bref : de les utiliser dans le cadre de solution de problèmes s'appuyant sur une expertise humaine.

Dans ce cours, nous allons considérer surtout les deux outils - le thesaurus et l'ontologie.

### 3) Indexation et découpage de documents

---

Avant de prendre en considération l'outil "thesaurus", encore un mot au sujet du découpage de documents (audiovisuels) en général et des documentaires (audiovisuels) en particulier.

Le découpage (**physique** ou "**logique**", cf. infra) de documents (audiovisuels) a comme but la constitution d'une bibliothèque de segments (audiovisuels), chaque segment acquiert de ce fait le statut d'un document "à part entière". Le document (audiovisuel) originaire devient, lui, une sorte de réseau de composants documentaires (i.e. de segments audiovisuels, dans notre cas) qui peut être modifié, adapté à des besoins et aux contraintes spécifiques d'un contexte d'application donné. Dans ce sens, on parle du découpage (physique ou logique) en terme d'un processus **d'enrichissement** de documents (audiovisuels), en terme d'un processus de production **d'une valeur ajoutée** d'un document (cf. à ce propos le très intéressant article de Lainé - Cruzel - Guinet). L'enrichissement d'un document par l'outil de son découpage (physique ou logique) permet, par exemple :

- Une **réutilisation** plus aisée de certaines parties (thématiques) d'un document (audiovisuel) dans de "nouvelles" productions.
- Une "**circonscription**" plus fine du sens; i.e. d'une certaine thématique pouvant se manifester, se développer dans différents "endroits", dans différents segments d'un même document (audiovisuel) ou à travers plusieurs documents (audiovisuels).
- Un **accès** plus direct aux informations pertinentes relatives à une question (i.e. un besoin exprimé).
- La **publication personnalisée** de segments (audiovisuels, textuels, ...) appartenant à un ou différents documents (audiovisuels) - défi à la fois conceptuel, technique et socio-économique majeur pour les métiers de l'édition et de la diffusion au sens "traditionnel".

- Une **mise à jour plus aisée** car seulement locale de "complexes documentaires" (notamment : documentation technique) .
- etc.

Rappelons encore que le processus du découpage d'un document en un certain nombre de segments repose sur des critères textuels (début et fin d'un plan, transitions, données sonores, données relatives à la parole, etc.) et surtout thématiques (i.e. relatifs au contenu ). Ce sont surtout ces derniers critères qui constitueront les méta-données (de description, d'indexation) des segments .

Enfin, le découpage d'un document (audiovisuel) en un certain nombre de segments (audiovisuels) peut se réaliser :

- Sous forme de création de **fichiers** physiquement distincts
- Sous forme purement **conceptuelle** (i.e. sous forme d'une description relative à une partie - un "segment" - d'un document audiovisuel).

Dans le premier cas, on découpe "réellement" (en général) une copie du fichier informatique contenant le document (audiovisuel) originale. Ce processus exige, bien sûr, la connaissance et l'utilisation de logiciels appropriés (pour le découpage physique de documents audiovisuels, il faut disposer, par exemple, du logiciel Premiere d'Adobe).

Signalons ici le cas particulier de production de plusieurs fichiers informatiques lors du processus de la numérisation d'un document audiovisuel à support analogique (cassette VHS, ...) ou du processus de l'acquisition d'une vidéo numérique (cassette DV). En règle générale, les cartes d'acquisition (analogique/numérique et/ou numérique/numérique) d'un certain niveau sont fournies avec un logiciel d'acquisition permettant, entre autre, l'acquisition sélective de telle ou telle partie d'un document audiovisuel ("batch capture", en anglais). Nous y reviendrons encore plus tard dans le cours sur l'acquisition de documents audiovisuels (cours VII).

Dans le cas du découpage purement "conceptuel" d'un document audiovisuel en un certain nombre de segments, le document-source n'est pas touché.

On crée seulement différents vues ("schémas") sur les parties, les segments d'un document. Ces vues constituent des méta-données qui sont ensuite stockées soit sous forme de fichiers ("méta-fichiers") qui "pointent" vers le segment en question soit sous forme d'enregistrements dans une base de données. Ainsi on peut avoir  $n$ -interprétations plus ou moins divergentes, plus ou moins proches ou contradictoires d'un même document sans qu'il faille pour autant toucher au fichier original lui-même.

L'intérêt de cette deuxième approche est qu'elle permet de créer ses propres "archives" autour d'une bibliothèque centrale de fichiers de segments ou de documents entiers.

Bien évidemment, les deux procédures peuvent co-exister ce qui est le cas, par exemple, dans le projet Opales où découpage physique et constitution de vues sur un segment, un document) concourent dans la constitution des vidéothèques spécialisées pour la recherche et l'enseignement.

## 4) Le thesaurus

---

### 4.1) fonction et rôle d'un thesaurus

---

Voici une citation d'une grande spécialiste dans le domaine de l'élaboration et de la gestion de thesauri montrant l'importance centrale accordée à cet outil en gestion de l'information en général et de bibliothèques, vidéothèques, médiathèques, .... en particulier.

"Le thesaurus est un outil fondamental en documentation : c'est un élément essentiel de la chaîne qui assure la liaison entre le document et son utilisateur. Que l'on parte d'un document à indexer, ou d'une demande d'information auprès d'une bibliothèque (ou d'un centre d'information) on aboutit au thesaurus, outil qui permet de transcrire en un langage documentaire les mots du langage naturel." (Marie-Thérèse LAUREILHE, Le thesaurus. Lyon, ENSB 1977, p.1)

Selon le même spécialiste, le thesaurus "est un vocabulaire de termes d'indexation contrôlés, structurés de sorte qu'il mette en évidence les relations à priori entre les concepts " (p.7); il est "une liste d'autorité organisée de descripteurs et non-descripteurs obéissant à des règles terminologiques propres et reliés entre eux par des relations hiérarchiques ou sémantiques" (Marie-Thérèse LAUREILHE, Le thesaurus. Lyon, ENSB 1977, p.1)

Dans ce sens, un thesaurus :

- Est la pièce centrale pour la **normalisation** des termes-thèmes utilisés lors de la description - indexation d'un document ou d'un corpus de documents
- Constitue le **guide** par excellence dans le processus de l'indexation (guide de définition, d'explication, de coordination, etc)
- Produit l'**index thématique** d'un corpus de ressources d'information - index thématique qui est le résultat même du processus de la description-indexation d'un tel corpus et qui constitue un sous-ensemble des thèmes composant le thesaurus.

L'*index thématique* représente ou, plutôt, recouvre la structure sémantique ou encore le schéma thématique propre à un document ou à un corpus de ressources d'information. Le

thesaurus, lui, est supposé contenir tous les schémas thématiques d'un certain type de ressources d'information (et pas seulement de tel ou tel corpus, tel ou tel fonds, telle ou telle collection, ...).

Un *schéma thématique* correspond à ce que l'on appelle aussi un *scénario* (thématique), un *modèle* (thématique, conceptuel, sémiotique, ...) ou encore, en "anglais", un "*template*". En d'autre terme, un schéma thématique tel qu'un index thématique est une vue sur une ressource d'information ou encore un corpus de ressources d'information.

L'index thématique, par ailleurs, peut être rapproché à ce que l'on appelle également des *méta-données* - des méta-données qui (comme le terme le suggère déjà) interprètent les données (i.e. les documents, corpus de documents, voire les parties ou segments de documents textuels, audiovisuels ou autres).

En effet, un schéma thématique ou encore un index thématique relatif à une ressource d'information est le résultat d'un processus de description ("sémiotique", lato sensu) qui peut être "codé" (cf. ccours I et cours 10 de l'année passée) dans un certain standard informatique de traitement et de gestion de l'information dont, notamment, le standard XML ou, plus précisément les standards plus appropriés -

- pour la gestion et le traitement de tel ou tel type d'information (cf. le standard MPEG 7 pour le codage des thèmes audiovisuels; le standard TEI pour le codage de thèmes relatifs à la littérature linguistique, sociologique, etc.);
- voir pour la gestion dans tel ou tel domaine ou contexte professionnel (cf. le cas du Dublin Core qui est un des standards en gestion de ressources d'information proche à l'optique documentaliste ...).

Toujours est-il que le thesaurus constitue en quelque sorte l'outil central et pratiquement indispensable pour toutes ces différents contextes d'exploitation. Comme déjà dit ci-dessus, le thesaurus est essentiellement :

- un **outil de normalisation** (de termes-thème),
- un **outil de production d'index thématiques** (normalisés) et
- un **guide** (d'une description "normalisée").

En d'autres termes, la mauvaise presse de cet outil ne corrobore pas avec son statut et ses fonctions dans tout projet de gestion d'information et de connaissances.

## 6.2) principaux éléments d'un thesaurus

---

Un thesaurus existe pour des raisons pragmatiques, c'est-à-dire que son organisation, sa composition de thèmes sont évidemment déterminés par des intérêts d'exploitation pratique.

Conception, réalisation et maintenance d'un thesaurus constituent l'objet de plusieurs normes et standards dont notamment la norme ISO .... et la norme française AFNOR Z 47.100. Ces normes définissent :

- d'une part la composition et les constituants principaux d'un thesaurus ainsi que
- d'autre part sa planification, gestion et exploitation.

Un thesaurus s'organise très typiquement autour des éléments et procédures suivants qui seront brièvement présentés ci-après :

- |  |
|--|
| <ul style="list-style-type: none"><li>d) des descripteurs (non-descripteurs)</li><li>e) un vocabulaire normalisé</li><li>f) des relations principales qui organisent le vocabulaire</li><li>g) des parties constitutives du thesaurus</li><li>h) des principes de construction selon tel ou tel type spécifique de thesaurus</li></ul> |
|--|

### a) les descripteurs

Le descripteur est un mot ou un groupe de mots choisi (parmi, en général, un ensemble de mots "candidats") pour représenter, exprimer une notion (un thème, un "lieu de savoir") dans le thesaurus.

La classe des non-descripteurs recouvrent, par contre, les expressions qui ne sont pas employées comme descripteurs mais qui auraient pu l'être qui, autrement dit, servent également à exprimer un thème, une notion.

On en tient compte, typiquement, à l'aide de la relation "EM" ("employer") ou "EP" (employé pour). Prenons l'exemple suivant tiré de l'ouvrage déjà cité de Marie-Thérèse LAUREILHE (i.e. *Le thesaurus*. Lyon, ENSB 1977, p.5). Dans celui-ci, le descripteur est "ordinateur" et les non-descripteurs sont "computer" et "calculateur" :

ORDINATEUR  
EP Computer, calculateur

Computer  
EM ORDINATEUR

Il est clair que l'utilisation de tel ou tel terme, tel ou tel syntagme linguistique comme descripteur d'un thème dépend entièrement du choix de l'auteur d'un thesaurus. Si "ordinateur" est le descripteur dans le thesaurus A, il peut être un non-descripteur dans le thesaurus B au détriment du terme "calculateur", par exemple (cf. à ce propos le statut des termes tels que "logiciel"/"software", "e-mail"/"mel", etc).

Ceci étant voici quelques précisions importantes. Le descripteur :

- Fait partie du **vocabulaire normalisé** ou "contrôlé" d'un thesaurus;
- Joue un **rôle équivalent** à celui du **taxème** dans la description thématique se positionnant en amont d'un thesaurus (d'une version corrigée, augmentée, ... d'un thesaurus)

La première précision stipule bien que l'élaboration (linguistique, sémiologique) d'un terme ou syntagme représentant un thème ou une notion est soumise à des contraintes propres à la rédaction d'un thesaurus (cf. ci-après). La deuxième précision montre que le

descripteur joue un rôle particulier dans le thesaurus, que ce rôle est de déterminer le **niveau de pertinence** (de généralité, de spécialisation, d'orientation, ...) du contenu d'un thesaurus.

C'est dans ce sens que le descripteur est comparable à un taxème et que la "conversion" ou encore le "codage" d'une description thématique en un thesaurus se fait sur la base **taxème → descripteur**.

Pour prendre notre exemple info-touristique (cf. le cours précédent), les "grands" thèmes :

1. Quête
2. Sensation (forte)
3. Communauté
4. Environnement
5. Artefact
6. Lieu (social)
7. Temps (social)

servant à la localisation, contextualisation et classification des taxèmes (à la rencontre des gens, témoignages, ... dans la catégorie "quête") peuvent être utilisés directement à la classification des taxèmes-descripteurs dans un thesaurus info-touristique. Sous chaque "grand" thème, on trouvera, ensuite, par ordre alphabétique, les différents taxèmes-descripteurs.

Conçu dans cette optique, l'organisation thématique du thesaurus reflète assez fidèlement une certaine vision du tourisme : c'est une quête visant des sensations non-quotidiennes, un regard sur une autre communauté (que celle du touriste, ...), etc.

Ceci dit, la "conversion" taxème → descripteur est souvent une simplification (de l'organisation structurale, interne d'un taxème). Nous y reviendrons.

### b) un vocabulaire normalisé

Tout thesaurus repose un vocabulaire "contrôlé", "normalisé". Ceci est nécessaire pour des raisons à la fois *technique* et *de gestion* (d'un thesaurus, voire d'un index thématique) . Souvent, le prix à payer est celui d'une certaine artificialité du langage (du vocabulaire) et une mauvaise compréhension de tel ou tel terme (syntagme linguistique, icônique, ...) par l'utilisateur (même professionnel). D'où l'importance des **définitions** et des **exemplifications** dont on est souvent assez avare, dans le domaine des thesaurii.

La "normalisation d'un vocabulaire (ou autre systèmes d'expression de thèmes), soulève plusieurs problèmes importants comme, par exemple :

- Le choix d'un système d'"écriture";
- La confusion - souvent implicite - entre d'une part termes et syntagmes linguistiques et d'autre part concepts, notions ou encore thèmes ;
- Le contrôle, plus spécifiquement, des synonymes ou quasi-synonymes et aussi des homonymes dans les termes ou syntagmes utilisés;
- Le besoin d'éviter la polysémie dans les termes (i.e. le fait qu'un terme puisse exprimer plusieurs "notions", plusieurs "lieux de savoir" ce qui est normal, en contexte de la langue naturelle);
- La question du traitement des noms propres (référent ? terme ? lieu de savoir ?, ...);
- La construction explicite des termes composés avec d'une part un élément central ("entête") et d'autre part un élément distinctif ("modificateur");
- La construction de systèmes d'expression multilingue;
- Les "passerelles" entre des systèmes d'expression linguistique et des systèmes d'expression non-linguistique ou hybride (linguistico-icônique, par exemple, ...)

Nous n'avons pas réellement problématisé cette question du vocabulaire dans le cadre de l'analyse thématique. Mais il y existe des conventions d'écriture (par exemple, les lettres majuscules, les crochets ([ ]) ou encore les barres obliques (/ /) réservées à l'expression des thèmes, etc). Il va de soi, que dans le cadre de la construction d'un thesaurus, ces questions deviennent importantes ....

Enfin, il faut noter ici encore une distinction qui est souvent mal perçue. Il s'agit de celle entre **terme** (d'un thesaurus) ou terme faisant partie d'un vocabulaire contrôlé et **mot clé**. Le mot-clé est une chaîne de caractères dans un texte ou ressource textuelle : il appartient donc au "langage naturel" et non au vocabulaire du méta-langage de description d'un thesaurus ...

Même s'il peut y avoir homonymie entre mot-clé et terme dans un thesaurus, il n'y a pas forcément synonymie entre ces deux entités et, de toute façon, même s'il y a synonymie entre ces deux entités, elles appartiennent à deux catégories logiques différentes : le mot-clé appartient à ce que l'on appelle le **langage-objet** et le terme faisant partie du vocabulaire d'un thesaurus fait partie à ce que l'on appelle le **méta-langage**.

### c) les relations principales

Comme connu, il existe un canon de relations "lexico-sémantiques" pour structurer, organiser les descripteurs . Voici les plus souvent utilisés :

#### Relation général/spécifique

**TG** (= terme dit générique, i.e. général)

**TS** (=terme spécifique)

#### Relation dite associative

**TA** (terme associatif)

#### Relation de restriction (= "point de vue)

**NA** (= note d'application)

**Exemple :** (adapté d'après Marie-Thérèse LAUREILHE, Le thesaurus. Lyon, ENSB 1977, p. 5)

MEMOIRE

NA : au sens de partie d'un ordinateur conservant les informations

TG : Ordinateur

TS : Mémoire vive, mémoire morte, mémoire virtuelle; disque

TA : écran, clavier, unité centrale, ...

#### **Attention:**

Pas de distinction entre relations taxinomiques et relations méréonymiques.

Pas de définition au sens propre

**Note :** dans certains thesauri - des types de relations plus fines, plus élaborées

Rôle, fonction

Lieu

**Note:**

Il faut être conscient qu'il s'agit d'un canon de relations qui peut être remplacé ou modifié (spécialisé, raffiné, ...). Tout dépend si on dispose d'une bonne typologie de relations incluant, outre les relations méronymiques, les relations narratives et rhétoriques, voire encore les relations modales, etc.

On aura compris que les spécifieurs d'un taxème forment les thèmes (plus généraux, plus spécialisés, associés, ...) autour d'un descripteur. Les relations (taxinomiques, méreologiques, narratives, rhétoriques, ...) qu'ils entretiennent avec leur taxème (voire entre eux ...) sont souvent (mais pas nécessairement) réduites au canon cité.

#### d) les parties constitutives d'un thesaurus

Typiquement, on distingue entre les parties suivantes qui entrent dans la composition d'un thesaurus :

- la classification-définition d'un descripteur (entrée vedette + principales relations)
- la liste alphabétique : des descripteurs dans leur environnement de classification
- la liste de la hiérarchie thématique (où on regroupe les descripteurs en quelques grands champs sémantiques ou thématiques; cf. le vocabulaire contrôlé de l'architecture)
- une table alphabétique qui comporte descripteurs, non-descripteurs, (grands) thèmes avec renvoi vers la liste alphabétique
- des listes particulières (noms des lieux, noms de collectivités, personnes, sigles, codes, ...)
- l'index permuté.

Voici encore un exemple d'un index permuté" (Marie-Thérèse LAUREILHE, Le thesaurus. Lyon, ENSB 1977, p. 9) - le terme "groupe" intervenant dans plusieurs descripteurs d'un thesaurus de sociologie (dont l'auteur est Jean Viet) :

	TERME	
Analyse de	GROUPE	
Appartenance au	GROUPE	
Attraction du	GROUPE	
Cohésion du	GROUPE	
Comportement du	GROUPE	
	GROUPE	social
	GROUPE	spiritual
Composition du	GROUPE	
	GROUPE	marginal

#### e) différents types de thesauri

On distingue entre différents types de thesauri qui se basent sur de modes et de "logiques" de confection particuliers. Voici un échantillon des principaux types :

- **thesaurus structuré** ou explicite (le standard des thesauri comportant une liste alphabétique de descripteurs dans leur environnement sémantique respectif complété, souvent, par une table thématique organisant les descripteurs en différents champs sémantiques)
- **thesaurus à représentation graphique** (sous forme de "schémas fléchés" ou "graphes"; sous forme d'arborescence ou encore sous forme de représentation circulaire et concentrique)
- **thesaurus à facette** (i.e. intégrant des "points de vue" particuliers, cf. ci-après)
- **macro-thesaurus** (utilisé pour différents domaines de connaissances et engageant souvent plusieurs grandes organisations de production et de gestion d'information - se rapproche le plus à la notion actuelle du standard ...)
- **thesaurus sectoriel ou spécialisé** (i.e. thesaurus conçu et élaboré pour tel ou tel domaine, voire sous-domaine de connaissances)

- **thesaurus multilingue** (thesaurus proposant son vocabulaire dans différentes langues)

Le type de thesaurus (peut-être le plus intéressant) est le thesaurus à facette qui introduit la notion du "point de vue" dans la conception et l'élaboration d'un vocabulaire pour indexer des ressources d'information. Voici un exemple très simplifié d'un thesaurus à facette (dans : Marie-Thérèse LAUREILHE, Le thesaurus. Lyon, ENSB 1977, p. 19)

FUELS

Subdivide by source

Coal  
Fuel oil  
Gasoline  
Kerosene  
Natural gas  
Petroleum

Subdivide by phase

Solid fuels  
Coal  
Peat  
...  
Liquide fuels  
Fuel oil  
Gasoline  
...

Subdivide by application

Illuminating fuels  
Aviation fuels  
Jet fuels  
Propellants  
...

etc.

En nous référant à la description thématique d'un corpus de documents (audiovisuels), nous pouvons en fait distinguer deux catégories de facettes :

- **a) l'axe thématique** (et correspondant, grosso modo, à la Note d'application, i.e., en anglais, au "scope")
- **b) le type/la catégorie de spécificateurs** (cf., dans le cours V, le '**modèle analytique**')

C'est la notion de "modèle analytique" qui devient central dans les tentatives de mise en place de normes, standards, etc. en gestion de l'information. Voici l'exemple de la norme Afnor Z 47 100 qui propose un modèle analytique en six facettes pour l'organisation des thesauri techniques, technologiques, industriels, etc. :

- **processus** (action voulue, provoquée par un agent humain ou anthropomorphe)
- **phénomène** (action naturelle échappant au contrôle de l'homme)
- **propriétés** (physiques, chimiques, ...)
- **matériaux** (terre, bois, ...)
- **outils ou équipement** (objets fabriqués, ...)
- **sciences, technologie, domaine** (domaines de connaissances, théories, ...)

## 5) L'ontologie

---

Gruber (GRU 1993) définit une ontologie comme une « spécification explicite d'une conceptualisation » d'un domaine d'expertise ou de référence.

Dans une perspective de représentation de connaissance traditionnelle, la conceptualisation est un ensemble de termes (souvent improprement appelés « concepts ») qui « nomment » des pièces de connaissances et qui, par leur formes d'expression suggèrent plutôt une certaine compréhension qui l'explicitent, i.e. qualifient ou la définissent.

Ce qui manque, ce sont les définitions pour chacun de ces termes constituant le vocabulaire. La production de telles définitions est la tâche d'une ontologie. Dans ce sens, Guarino précise : « the main purpose of an ontology is to specify the intended meaning of a vocabulary, i.e. its underlying conceptualization » (GUA 1997 : 3).

D'où aussi la nécessité d'une analyse thématique "en amont" de toute tentative de définitions des termes constituant une ontologie, reposant sur une théorie du contenu (i.e. du contenu d'un signe textuel au sens large du terme et d'un signe audiovisuel, en particulier).

Or, comme nous le savons déjà, s'il est relativement aisé à expliciter sous forme d'une ontologie la conceptualisation d'un domaine technique ou scientifique, il n'en est pas de même en ce qui concerne des domaines de référence sociaux, culturels ou historiques ou encore en ce qui concerne des domaines de référence « pratiques » recouvrant par exemple, les savoirs et savoirs-faire pratiques.

### 5.1) types d'ontologies

A la base, une ontologie ressemble (n'en déplaisent aux défenseurs de ce type de recherche-application) à un thesaurus (à facette).

- Une ontologie est composée d'un vocabulaire de "termes".
- Chaque terme est défini dans son environnement relationnel ( i.e. dans une configuration de relations qui lui est propre); cette définition pouvant être représentée dans un langage (semi-)formel.

- Un terme peut posséder une définition dite en "langage naturel", voire une "définition dite encyclopédique".
- Les termes peuvent être organisés - à la manière des "grands thèmes" dans un thesaurus - dans des champs ou catégories thématiques.

Le vocabulaire d'une ontologie est souvent l'objet de distinction en des sous-types plus spécifiques. Les deux sous-types les plus connus sont :

- Celui représentant un concept ou un thème
- Celui représentant une relation conceptuelle ou thématique .

Thesaurus et ontologie peuvent se distinguer surtout par rapport aux deux aspects suivants :

- Usage systématique et "réglementée", dans les ontologies, de la définition d'un terme (concept ou relation)
- Usage fréquent , dans les ontologies, de variétés importantes de types de relations (mais pas toujours contrôlés d'un point de vue théorique).

On distingue entre différents types d'ontologies. Voici un échantillon de quelques types:

Les **top-level ontologies** : comprennent les définitions de termes très généraux dont on assume qu'ils peuvent se retrouver, indistinctement, dans une grande diversité de domaines de référence plus spécifique. L'un des projets les plus connus de la spécification d'une « top-level ontology » est le projet Ontolingua (Stuttgart).

Les **ontologies de domaines** : analogue aux thesauri spécialisé, il commence à exister toute une panoplie d'ontologies de domaines réunissant les connaissances variées d'un domaine particulier.

Les **ontologies de tâches** : ce sont des ontologies "procédurales" qui proposent un vocabulaire pour la compréhension (et le partage) de tâches particulières (diagnostic, production, vérification, ...). Un des points de départ ici est le KADS (Knowledge Acquisition Development System), développé au début des années 90.

Les **ontologies d'applications** (concrètes) : Ce sont des ontologies que l'on utilise pour mieux gérer, par exemple, un atelier logiciel (les cycles de développement d'un logiciel), etc.

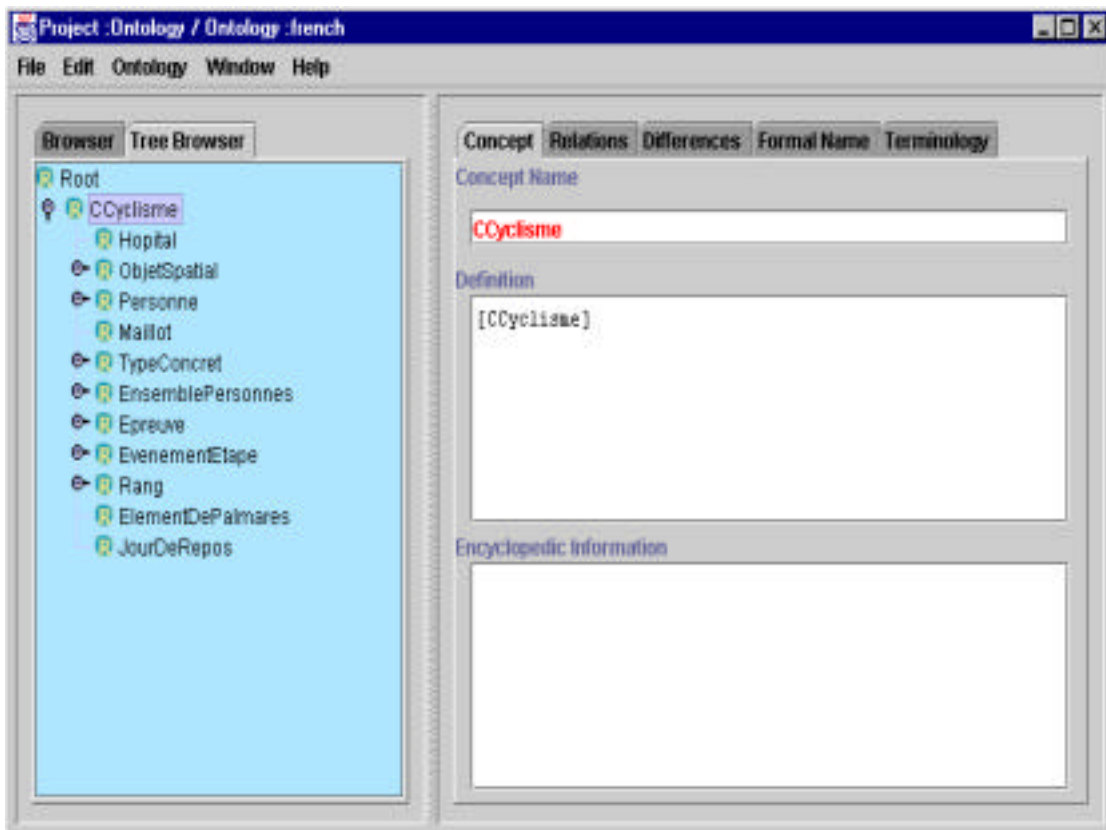
Un cas particulier constituent les « **meta level ontologies** » qui proposent des formats de représentation et de définition aux ontologies que nous venons de citer ci-dessus. Parmi

ces méta-ontologies, citons la « frame ontology » (GRU 1993) ou encore la théorie des graphes conceptuels (SOW 1984).

### 5.2) concevoir et éditer une ontologie

Nous présentons ci-après à l'aide de quelques images de page-écran un éditeur d'ontologie développé par l'INA et utilisé dans le cadre du projet Opales. Ensuite, nous terminerons en indiquant comment passer d'une description thématique à une ontologie

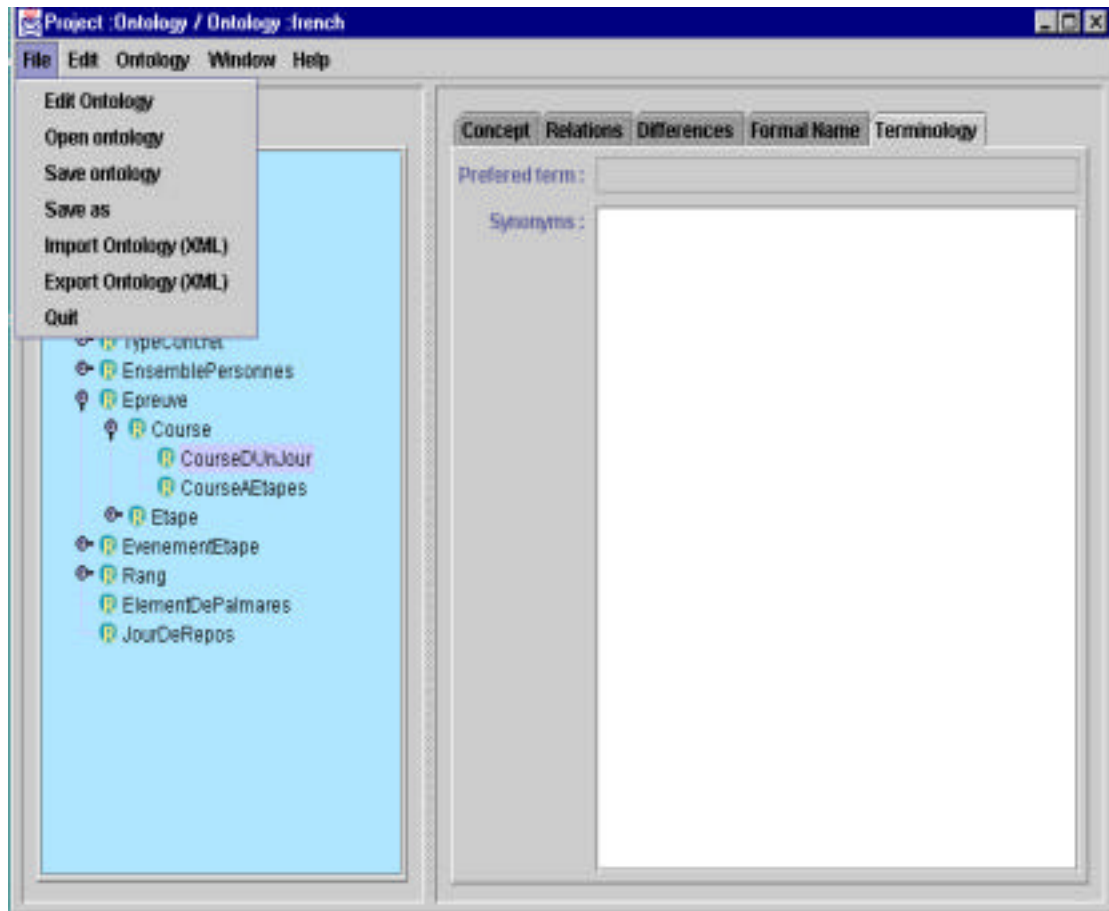
Voici l'interface de l'éditeur en question. L'ontologie ici concerne le domaine du cyclisme.



(figure 1)

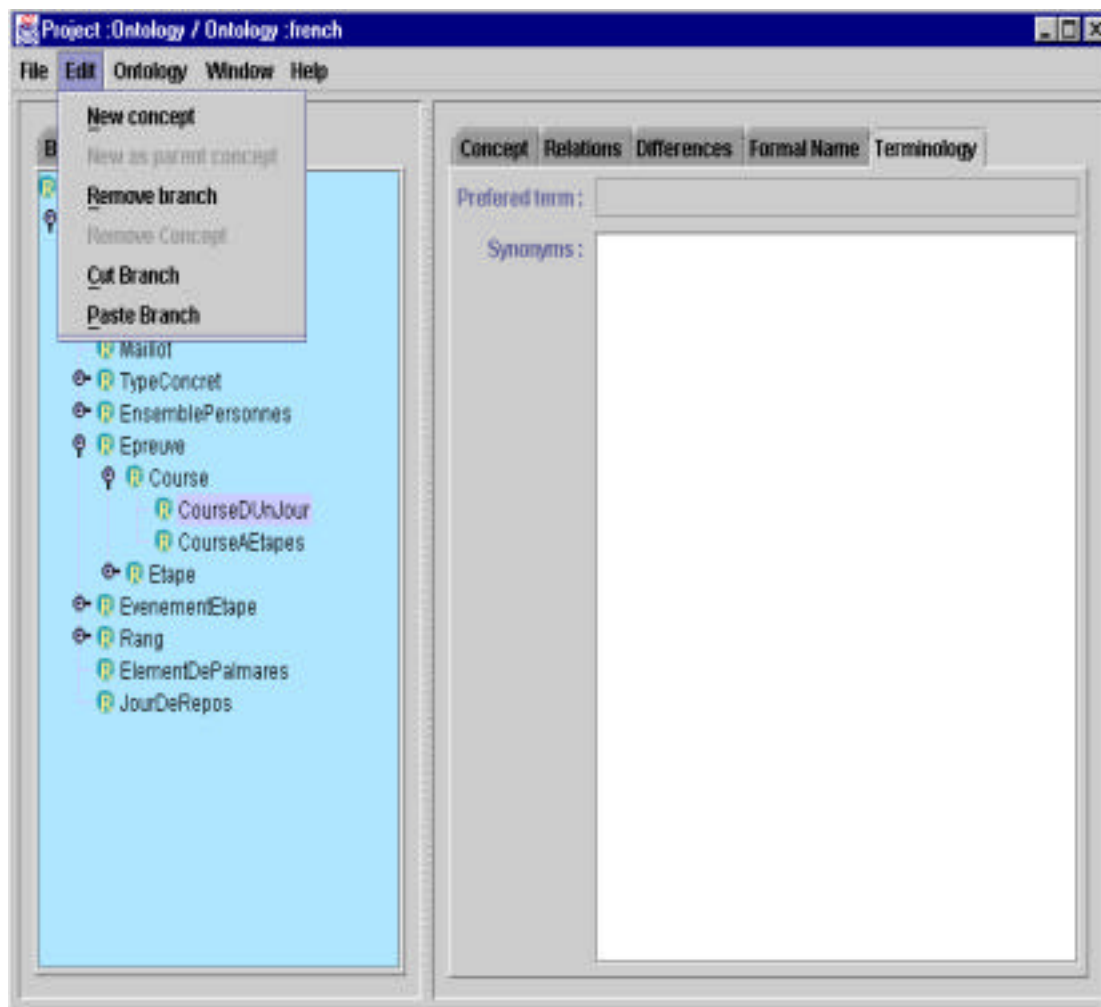
Dans la figure 1, on voit les différents "concepts" (thèmes) qui constitue l'ontologie en question ainsi qu'une série d'options que nous verrons ci-après.

Pour commencer, il faut ouvrir la rubrique "file" dans la barre de menu (figure 2). Là, on a le choix soit d'éditer une nouvelle ontologie, soit d'ouvrir une ontologie déjà existante (ne fonctionne pas dans cette version), soit d'importer une ontologie déjà existante en XML (remplace dans cette version l'option "open"); soit d'exporter une ontologie en format XML



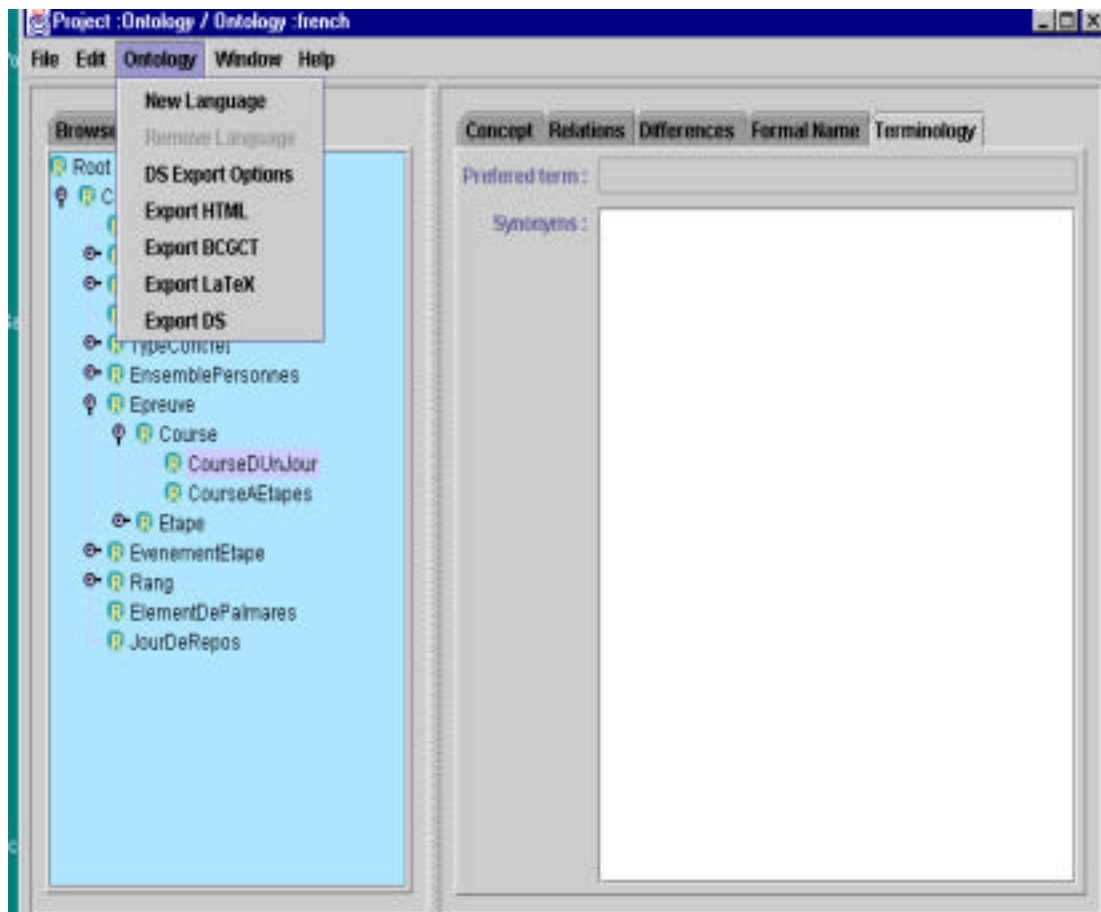
(figure 2)

La rubrique "edit" (figure 3) propose un ensemble d'options importantes pour l'édition d'une nouvelle ontologie : éditer un nouveau concept, couper/coller des "branches" dans l'arbre d'une ontologie, etc ....



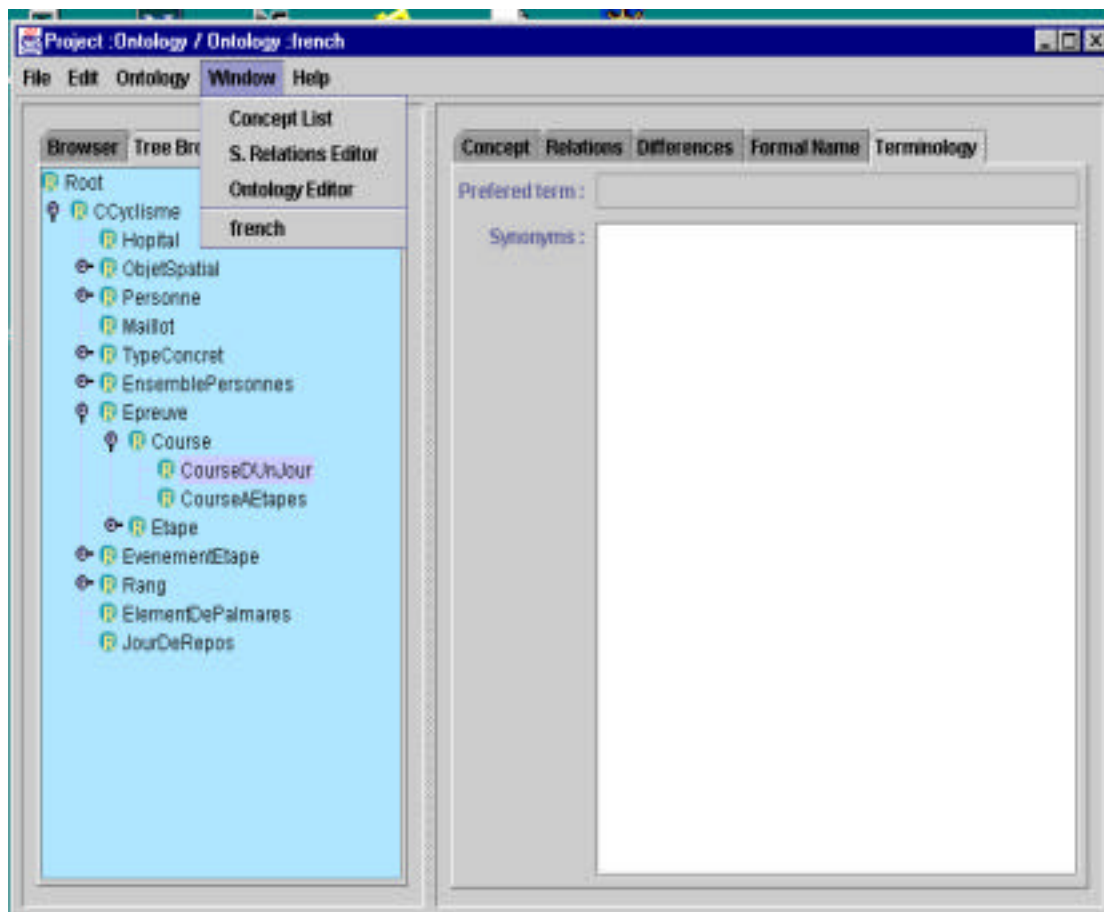
(figure 3)

La rubrique "ontologie" (figure 4) réunit un ensemble d'options concernant d'une part l'exportation d'une ontologie (en format HTML, LaTeX, ...) et d'autre le choix d'une langue (français, anglais, ...) pour rédiger le vocabulaire d'une ontologie



(figure 4)

Dans la rubrique "Window", on peut choisir entre différentes interfaces de travail : interface de travail pour éditer les concepts (thèmes); interface de travail pour éditer les relations conceptuelles (relations thématiques tels que relations taxinomiques, méréologiques, narratives, ...)

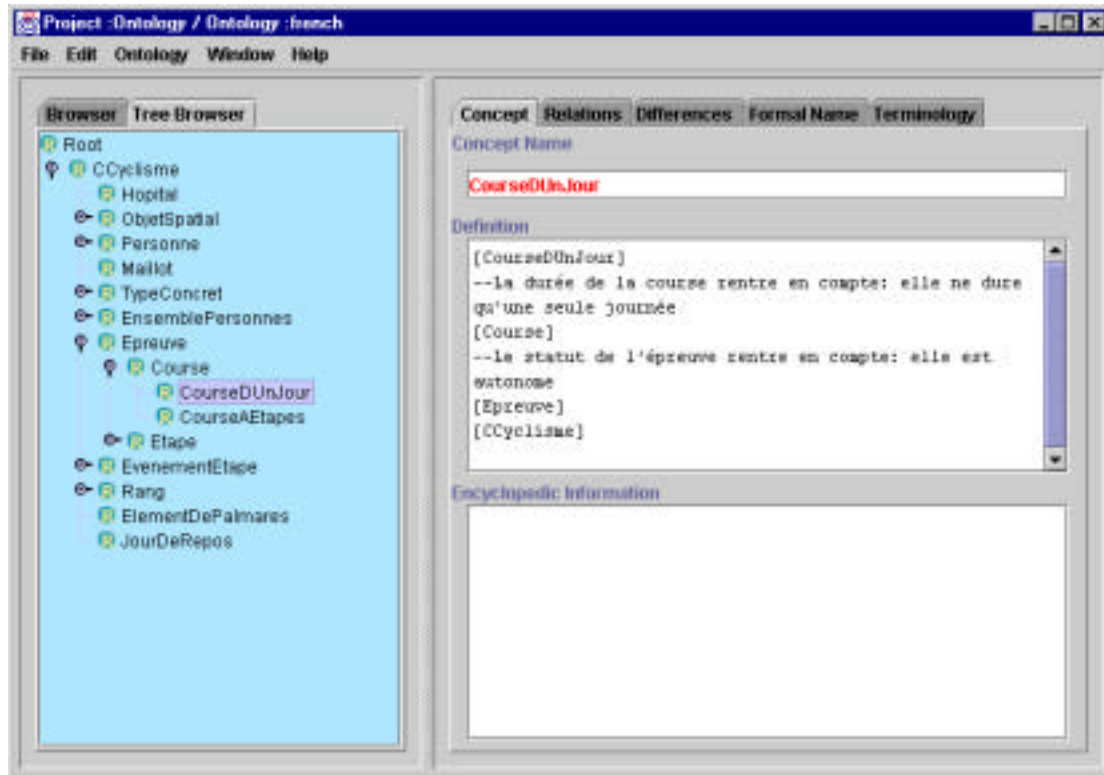


(figure 5)

Dans les figures 1 à 5, nous avons pu nous rendre compte qu'une ontologie de concepts et de relations entre concepts s'articule en fait comme une taxinomie. Autrement dit, la relation taxinomique est donnée telle que, elle est présupposée aussi bien dans la production et définition des termes exprimant des concepts (thèmes) que dans celles des termes exprimant les rapports entre concepts (thèmes).

La figure 1 nous montre également comment et surtout où introduire et nommer (exprimer par un terme, un syntagme linguistique) un concept (un thème). Dans la figure 6, ci-après, nous voyons comment procéder à une définition explicite d'un terme. Une telle

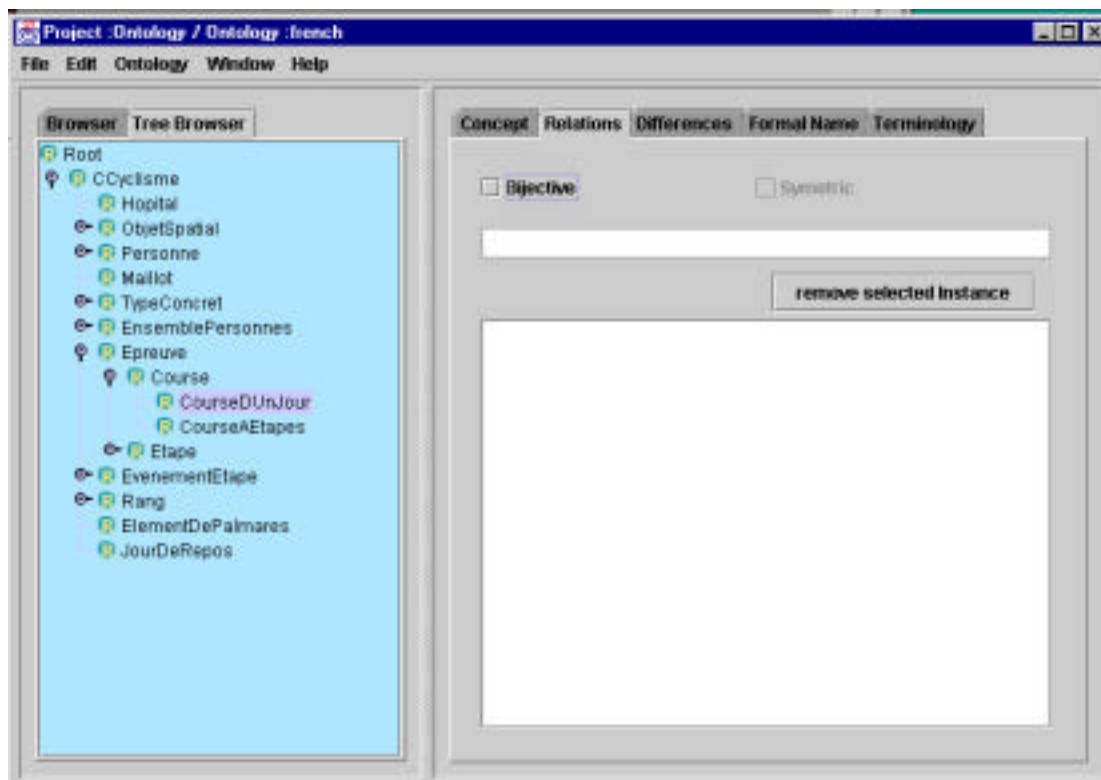
définition explicite repose sur le contexte dans lequel s'inscrit un terme dans une hiérarchie de termes et des commentaires ("libres").



(figure 6)

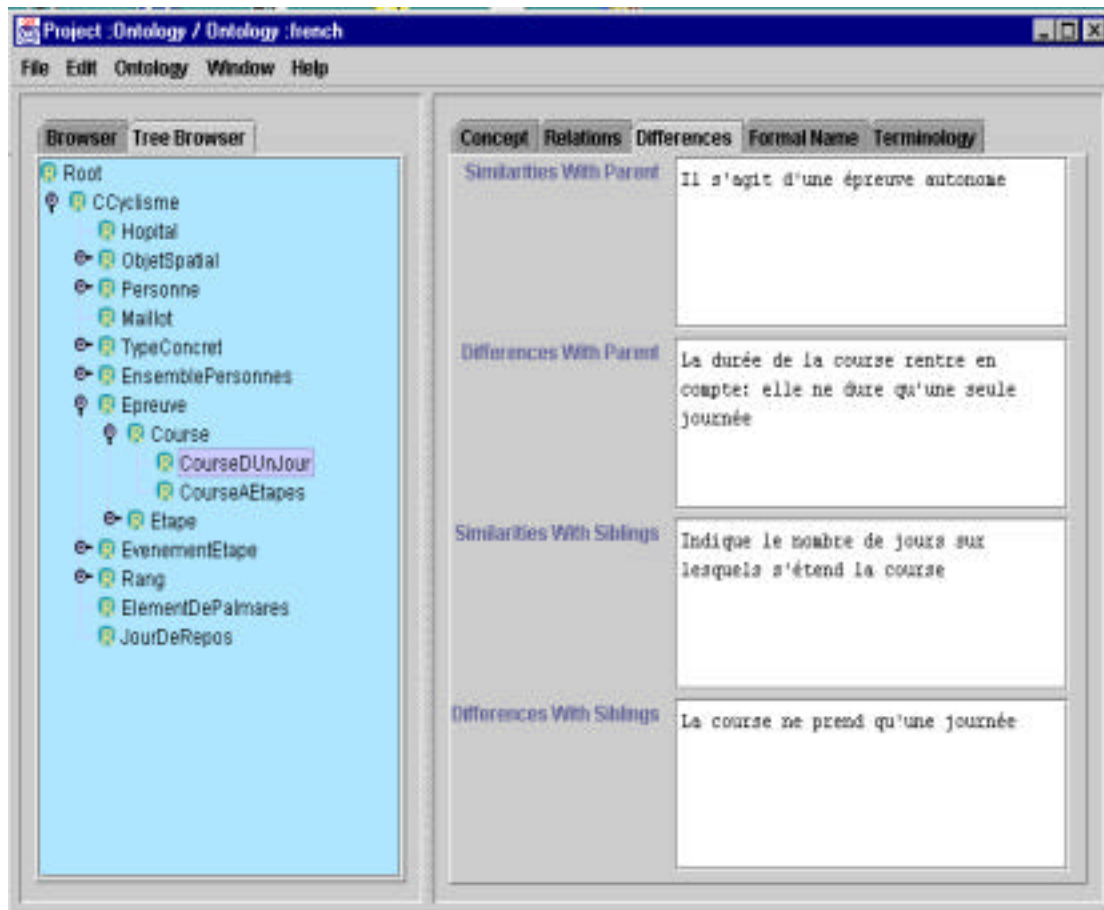
Dans le champ "encyclopedic information", on introduit toute sorte d'informations permettant d'obtenir une compréhension plus riche, plus "dense" du concept.

Dans la figure 7, on voit qu'il faut (que l'on peut) "attacher" à un concept (thème) des relations quelconques le liant avec d'autres concepts (thèmes) dans l'ontologie. C'est une option très intéressante dans la mesure où elle permet justement de définir les taxèmes en tenant compte de leurs spécifieurs (taxinomiques, attributifs, partitifs, actantiels, ....; narratifs, etc.). Cette énorme richesse offerte au concepteur d'une ontologie est certainement l'un des avantages d'une ontologie par rapport à un thesaurus (mais cette liberté connaît aussi ses pièges ...)



(figure 7)

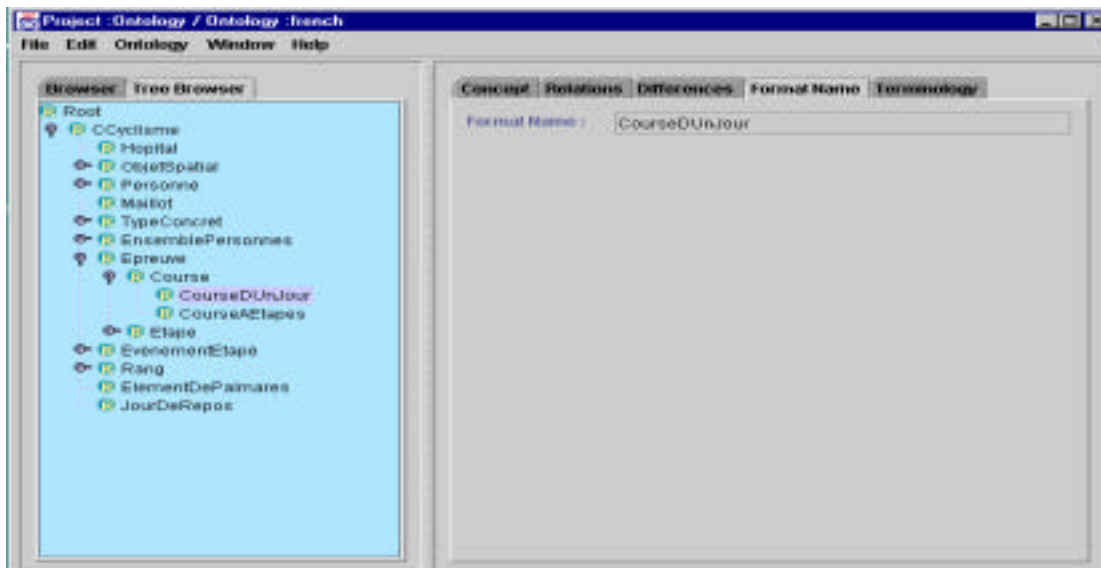
Dans la figure 8, on voit une option particulière qui concerne l'explicitation des différences entre un concept (thème) donné, son concept-père et ses concepts associés (dans une hiérarchie de concepts donnée). C'est une option spécifique qui veut rendre compte d'un principe de base de la sémantique structurale en particulier (A.J. Greimas, B. Pottier) et de la linguistique structurale en général à savoir que tout "sens" est différentiel (ou, comme l'a formulé avec plus de précision A.J. Greimas, "relationnel").



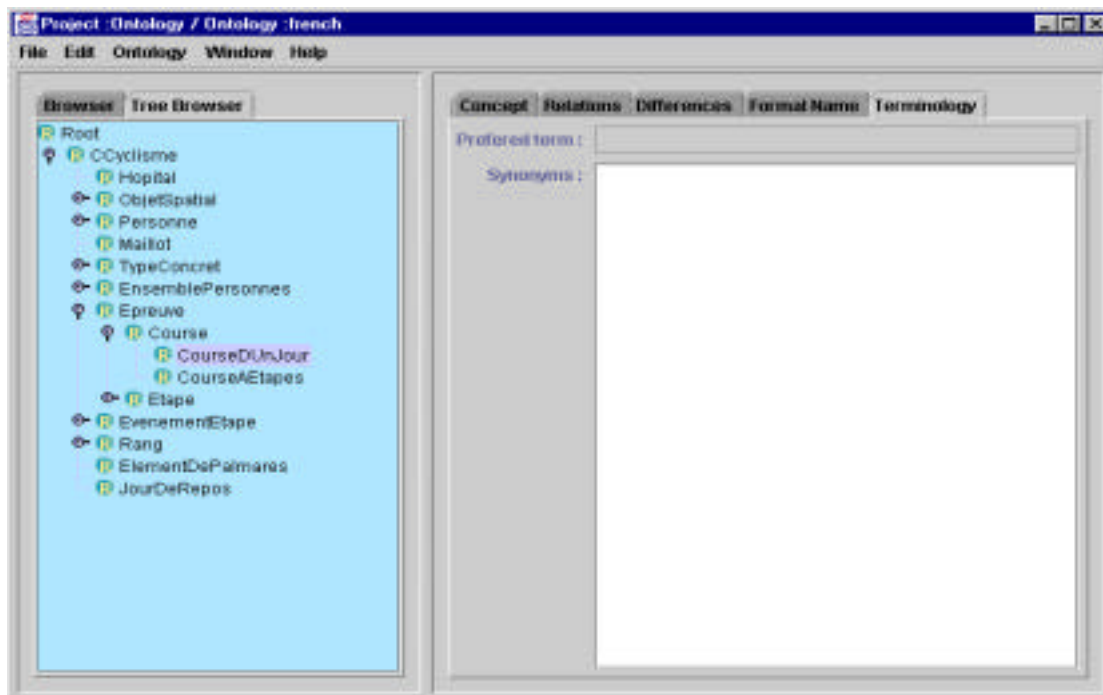
(figure 8)

Les options "Formal Name" et "Terminology" (figure 9 et 10) permettent de distinguer entre un système de représentation "technique" interne des concepts et relations conceptuelles et leur terminologie (distinction qui peut être intéressante dans un contexte multilingue).

Par ailleurs, comme la figure 10 la montre, l'option "Terminologie" permet de spécifier un terme (syntagme) préféré et aussi une liste de synonymes (ou, plutôt, quasi-synonymes).



(figure 9)



(figure 10)

L'élaboration d'une ontologie présuppose obligatoirement une description d'un domaine ou, plutôt, d'un corpus de documents (audiovisuels, dans notre cas). La description thématique est en quelque sorte le "*modèle de description*" ou encore la *conceptualisation* d'un domaine (selon Gruber ou Guarino, cf. le début du chapitre) indispensable pour définir un vocabulaire qui l'exprime.

Les taxèmes d'une description thématique peuvent être classifiés en des "grands" thèmes, comme nous l'avons déjà vu dans le cours IV à propos de la description de notre corpus info-touristique. Certains taxèmes peuvent être plus spécialisés que d'autres et donc constituer des thèmes plus spécialisés. Certains autres taxèmes ne possèdent pas de relations de spécialisation les uns par rapport aux autres et n'entretiennent entre eux que des rapports "associatifs" dans ce sens ils font tous partie d'un même "modèle de description, d'une même conceptualisation. C'est à une ontologie de préciser ces rapports "associatifs" dans la mesure où tous les thèmes faisant partie d'une conceptualisation doivent trouver leur place dans une grande hiérarchie de concepts (ou thèmes; cf. ci-dessus les figures 1 à 5).

Autrement, au lieu d'introduire directement les taxèmes dans une ontologie, il convient:

- d'expliciter les différents groupes de thèmes qui forment des petites taxinomies, de petits groupes taxinomiques (même si certains thèmes restent isolés, seuls);
- de réunir les différents groupes taxinomiques et les différents thèmes isolés en des catégories plus générales
- de les éditer sous forme d'un "arbre" hiérarchique général qui englobe tous les catégories, groupes taxinomiques et thèmes isolés (mais attention toutefois : il ne faut jamais oublier que des taxinomies dépassant trois à quatre niveaux hiérarchiques sont très peu compréhensibles et semblent être très artificiels)
- de réunir les différentes relations thématiques utilisées dans la description thématiques et de les déclarer de la même façon dans l'éditeur de relations
- enfin, de déclarer, de définir les taxèmes sous forme de "scènes" à l'aide des concepts (thèmes) et relations conceptuelles (thématiques) préalablement déclarés et définis dans l'ontologie.