

Construction de modèles conceptuels à partir de textes

Nathalie Aussenac-Gilles

IRIT – Toulouse (F)

aussenac@irit.fr

- ❑ **Domaine** : Représenter des connaissances à partir de textes en IC
- ❑ **Structures de données** : Ressources terminologiques et ontologiques
- ❑ **Normes** : Langages pour la représentation d'ontologies
- ❑ **Logiciels** pour la construction d'ontologies
- ❑ **Applications « phare »** : Web Sémantique, gestion documentaire
- ❑ **Discussion** : Au delà de la représentation des connaissances, vers des modèles plus dynamiques

Ingénierie des connaissances

- ❑ **L'IC** intervient pour
 - Définir une aide à l'utilisateur (méthode, logiciel, organisation ...)
 - Modéliser des connaissances de natures différentes
 - Rendre ces connaissances accessibles dans un système informatique
- ❑ **La recherche en IC** produit
 - Des méthodes et des techniques de recueil, d'analyse et de structuration des connaissances
 - Des plates-formes de modélisation
 - Des **représentations des connaissances** opérationnelles ou non, en fonction du contexte

Représenter des connaissances à partir de textes en IC

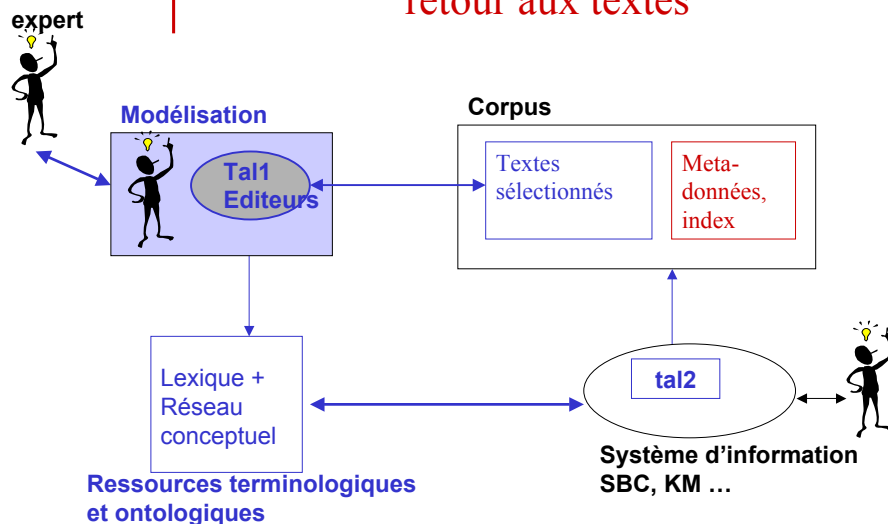
❑ Modèles conceptuels

- Supports à l'explicitation des connaissances
- Conceptualisent les entités du domaine et leurs interactions

❑ Processus

- Observation de l'usage des connaissances
- Normalisation (structuration / construction)
- Formalisation : représentation des connaissances au sens IA

Des textes aux modèles pour un meilleur retour aux textes

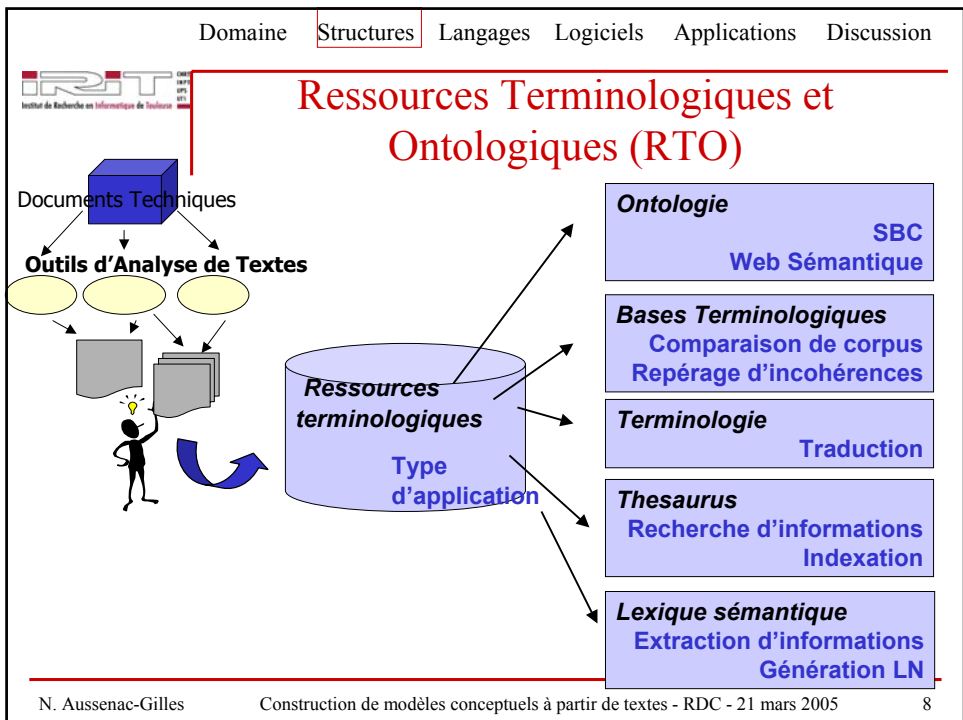
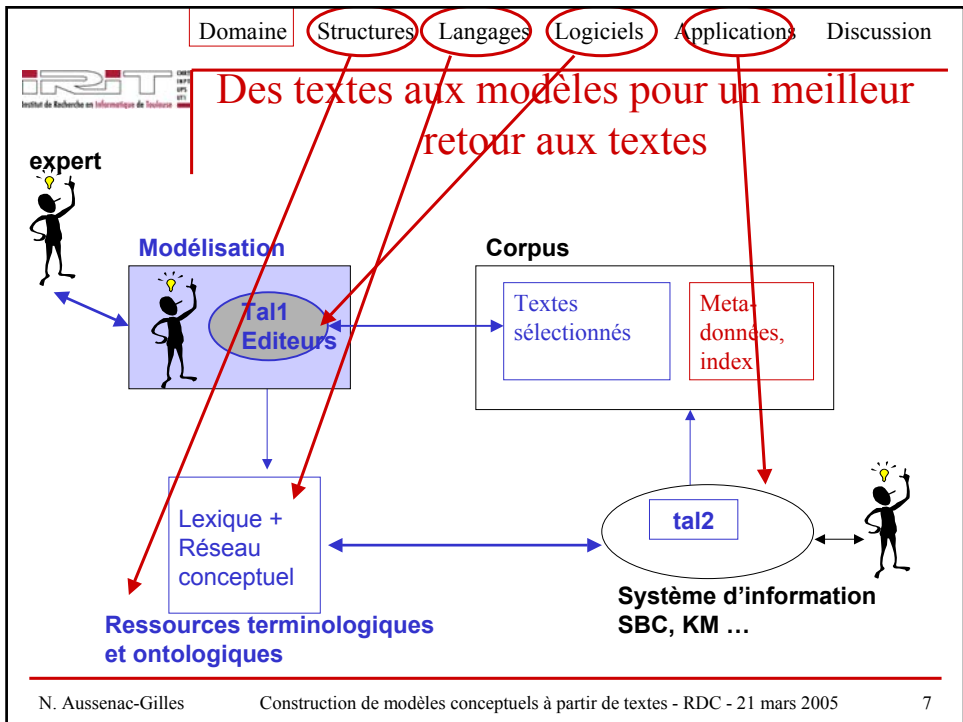


Nos hypothèses

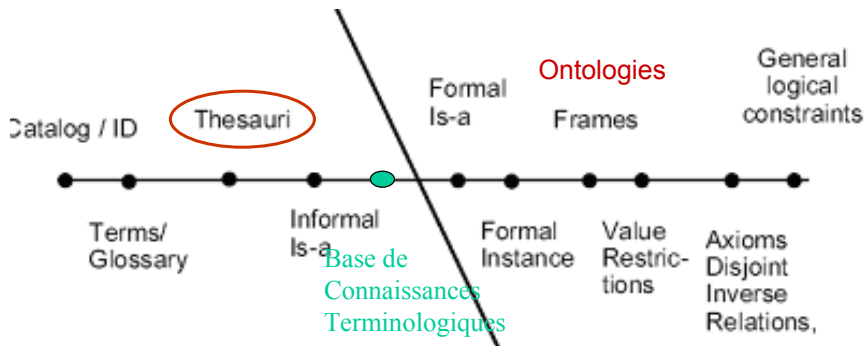
- ❑ Focalisation sur un domaine spécialisé
- ❑ Modélisation = **interprétation**, choix, construction
 - **Semantique de corpus** : occurrences des termes comme indices de concepts
 - Ni termes ni concepts ne pre-existent
- ❑ L'analyste au cœur du processus
 - prend en compte les besoins des utilisateurs
 - applique des critères ontologiques de structuration
- ❑ Double statut des RTO
 - résultats du processus d'acquisition
 - ressources pour le TAL ou le système d'information

Pourquoi recourir aux textes ? Quels textes ?

- ❑ Intérêt des textes
 - **Traces de connaissances** en usage, partagées et stabilisées
 - Améliorer la lisibilité et la maintenance des modèles
 - Complémentaires de l'expertise humaine
 - Automatisation de l'analyse : gain de temps, réduction de coûts
- ❑ Nature et contenu des corpus
 - Le corpus est construit
 - En fonction de l'application visée et des caractéristiques des documents (contenu, genre textuel, date, auteurs, format, etc.)



Les RTO sur l'axe de la formalisation

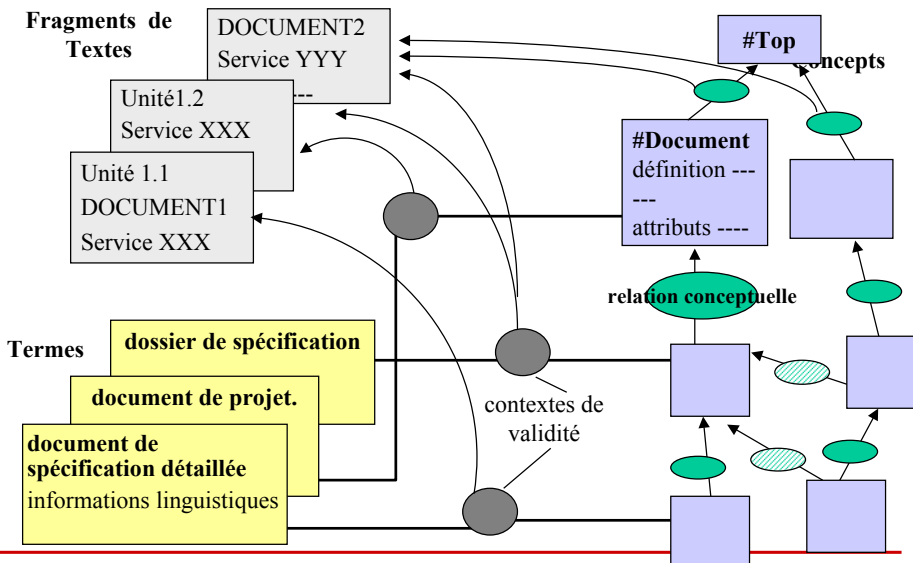


RTO : quels modèles construisons-nous ?

- ❑ Composants
 - Composant terminologique : termes et informations associées, patrons de recherche ...
 - Réseau conceptuel : concepts, relations, rôles, ...
 - Corpus : occurrences des termes, relations lexicales, ...

- ❑ Pourquoi une composante terminologique ?
 - Utile pour le TAL
 - Traçabilité et maintenance : reconstruire l'interprétation

Le modèle des données d'une BCT



Thesaurus

hémopéritoine

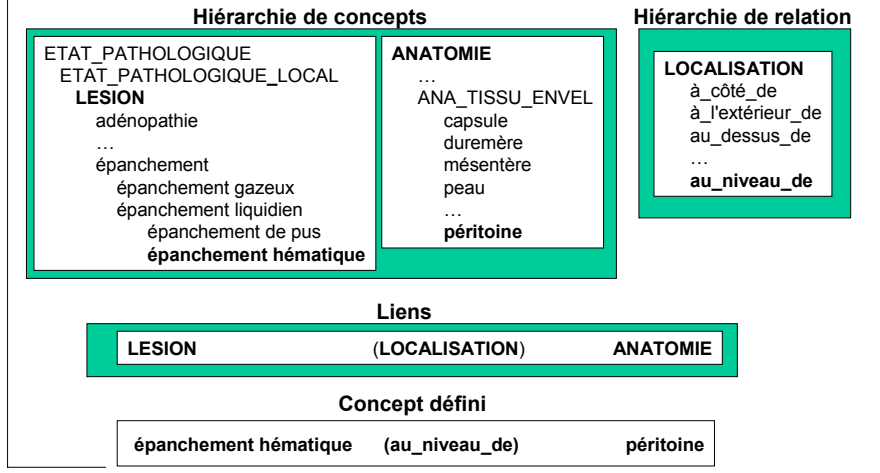


A	B
212	
213	3- APPAREIL DIGESTIF
214	
215	Intestin mésentère
216	Fistule du grêle ou du colon
217	Infarctus mésentérique et colite ischémique aiguë
218	Grêle court
219	Colite pseudomembraneuse
220	Syndrome pseudo occlusif aigu [POCA]
221	Colite et gastro-entérite infectieuse
222	
223	Péritoine
224	Hémopéritoine non traumatique
225	Péritonite primaire (sans perforation d'organe)
226	Péritonite appendiculaire
227	Péritonite secondaire à une perforation intestinale non traumatique
228	Péritonite post-opératoire par désunion d'anastomose
229	
230	Paroi
231	Cellulite infectieuse
232	Infection de paroi consécutive à un acte opératoire

Thésaurus SRLF et de la SFAR

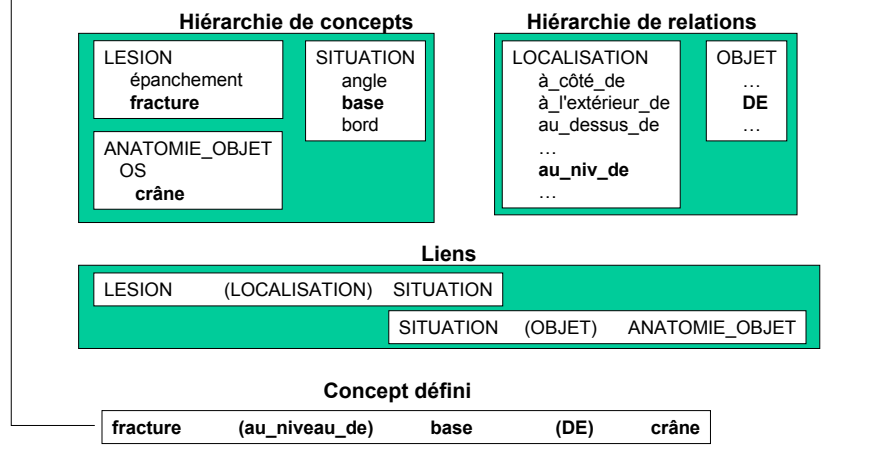
Ontologie formelle

hémopéritoine : « épanchement hématique localisé au niveau du péritoine »



Ontologie formelle

fracture à la base du crâne



Ontologies : définition

Ontologie INGÉNIERIE DES CONNAISSANCES. *Spécification normalisée représentant les classes des objets reconnus comme existant dans le domaine. Construire une ontologie, c'est aussi décider d'une manière d'être et d'exister des objets.*

□ Motivations

- réutiliser certains composants dans les systèmes informatiques
- assurer la communication entre systèmes informatiques

□ Contenu

- Classes génériques, relations et règles

Fondements des ontologies : principes

□ Pertinence pour une application, une tâche donnée

- Les concepts, relations, axiomes résultent d'une **décision** et relèvent d'un **point de vue** porté par l'analyste
- **Engagement ontologique** : seules les propriétés pertinentes sont représentées ; les concepts sont des constructions artificielles

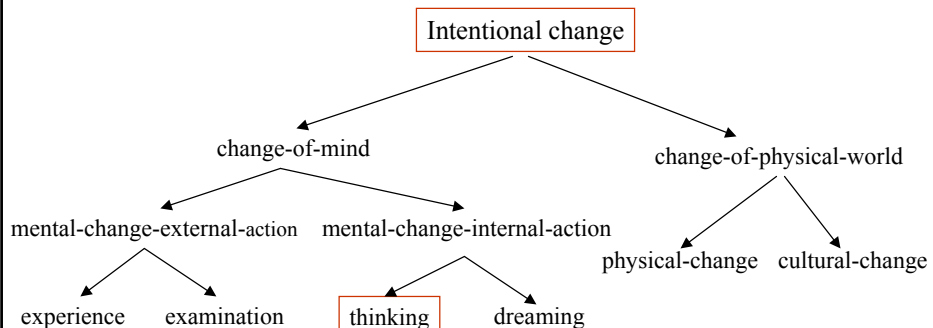
□ Normalisation selon des principes ontologiques

- Principes différentiels (arbre de Porphyre, B. Bachimont),
- Propriétés formelles (méta-propriétés définies par Guarino et Welty)

Principes de normalisation

- ❑ Critères de normalisation
 - Point commun entre 1 concept et son père
 - Différence entre 1 concept et son père
 - Points communs entre 1 concept et ses frères
 - Différences entre un concept et ses frères
- ❑ Justifier la signification d'un concept par les relations
 - Concept ou instance
 - Différenciation des concepts
 - Unicité de définition
 - Homogénéité de point de vue
 - Cohérence des descriptions

Ménélas : différenciation des concepts



Thinking : « On peut penser que c'est dû à la stenose »

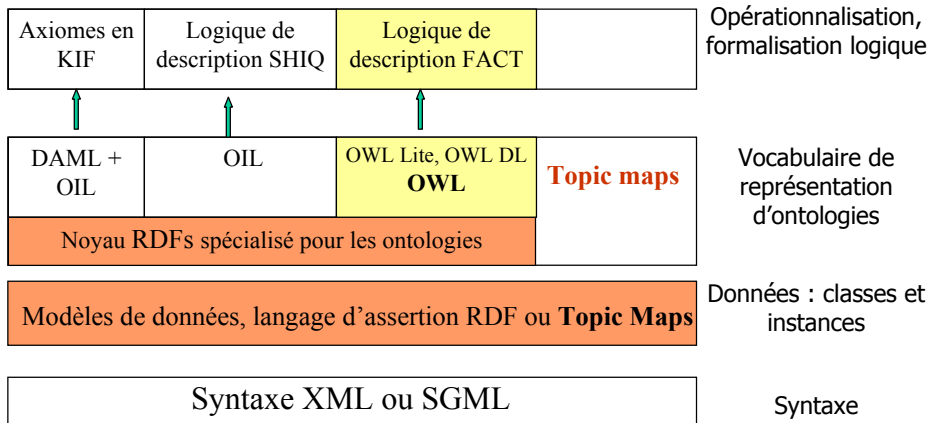
Contraintes sur la représentation des ontologies

- ❑ 2 rôles symétriques des ontologies
 - Définir / fournir une **sémantique formelle** pour l'information permettant son exploitation par un ordinateur
 - Définir / fournir une sémantique d'un domaine du monde réel fondée sur un **consensus** et permettant de lier le contenu exploitable par la machine avec sa **signification pour les humains**
- ❑ Exigences sur les langages
 - Lisibilité ; continuité avec le langage naturel
 - Portabilité
 - Inférence

Représentation des connaissances d'une ontologie : historique

- ❑ Référence historique : Réseaux sémantiques (Brachman, Levêques)
- ❑ Logique du 1er ordre : CycL, KIF
- ❑ Frames : Frame Logic, Ontolingua
- ❑ Logiques de description
- ❑ Graphes conceptuels

Représentation des connaissances : vers des architectures en couches

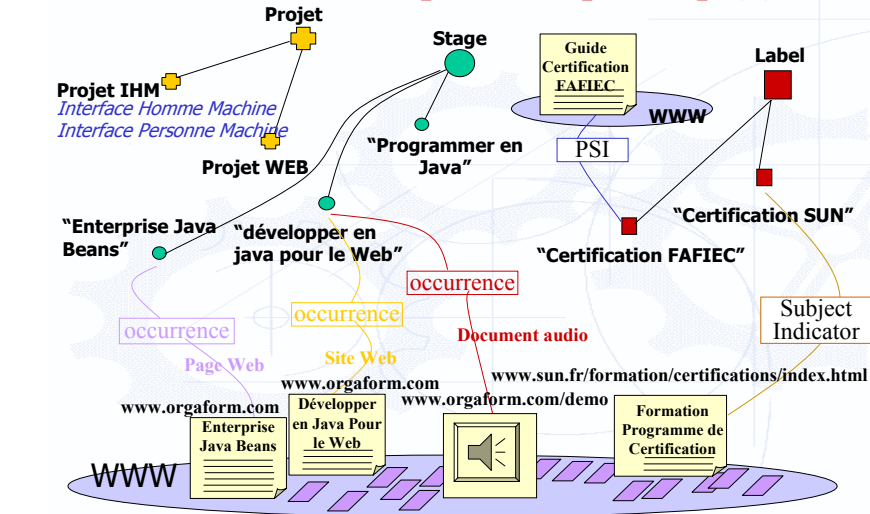


Les Topic Maps

- ❑ Paradigme remonte à 1993
 - Développé dans un cadre documentaire
 - Donne naissance à la norme XTM 1.0 en février 2001
- ❑ Contient les sujets (*topics*) dont les ressources « parlent », et les relations (*associations*) entre ces sujets
 - Les associations peuvent être elles-même des *topics*
 - Le *scope*, une notion de contexte, de portée du *topic*
- ❑ Un formalisme pour : index, table des matières, thésaurus, glossaire... réseaux sémantiques
- ❑ Pas de hiérarchies

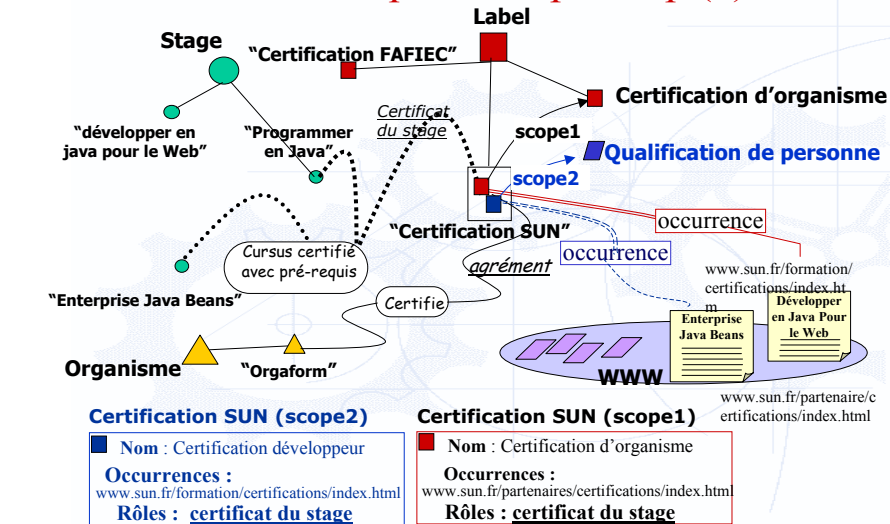
<http://www.topicmaps.org/>

Exemple de Topic map (1)



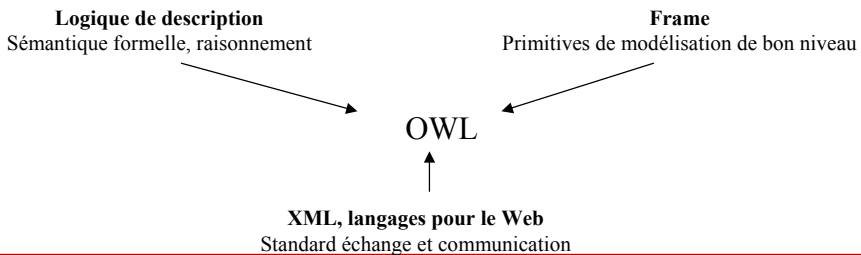
D'après J. Caussanel, M. Zacklad *et al.*

Exemple de Topic map (2)

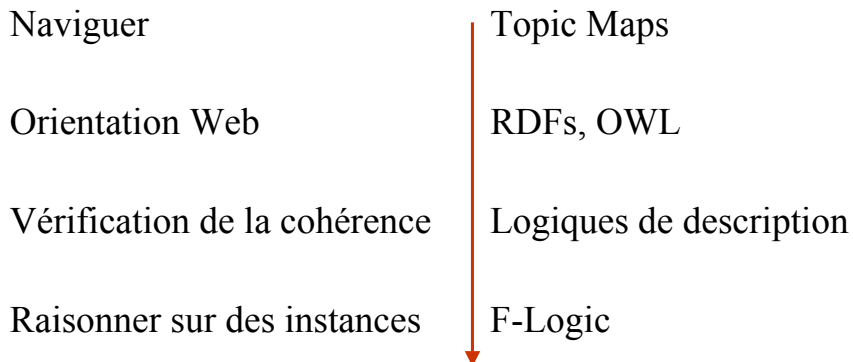


Représentation des connaissances : de RDF à OWL

- ❑ XML et RDF : RDFs
 - Une ontologie ne peut pas être spécifiée dans une DTD
 - RDFs peut convenir pour décrire des ontologies
 - XOL et OIL : logiques de descriptions pour ontologies basés sur XML
- ❑ Norme définie par le W3C : ontologie pour le Web Sémantique



Bilan sur les représentations



capacité inférentielle croissante

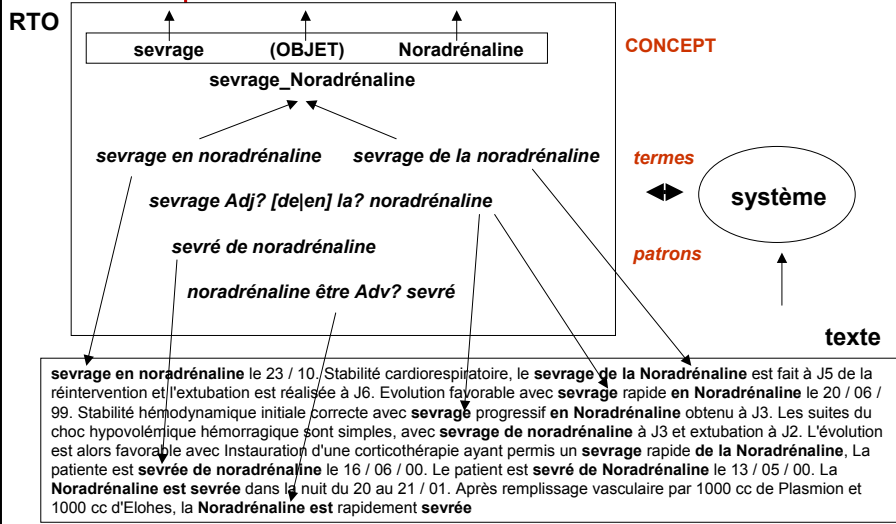
Logiciels pour la construction de RTO

- ❑ Outils d'analyse de textes
- ❑ Editeurs
 - OilEd, OntoEd, DOE
- ❑ Plates-formes = éditeurs + logiciels d'analyse + principes de structuration (méthode)
 - Protégé-2000
 - KAON
 - TERMINAE

Outils de TAL : extraire et structurer

- ❑ Typologie fonctionnelle :
 - Extraire des candidats termes : Termino, Lexter, Ana, Syntex, Acabit
 - Extraire des relations candidates : Prométhée, Caméléon, Asium
- ❑ Autres typologies
 - Méthodes linguistique / méthodes statistiques
 - Construction de RTO / mise à jour de RTO
 - Phase d'amorçage / phase d'enrichissement
 - Apprentissage automatique / supervisé

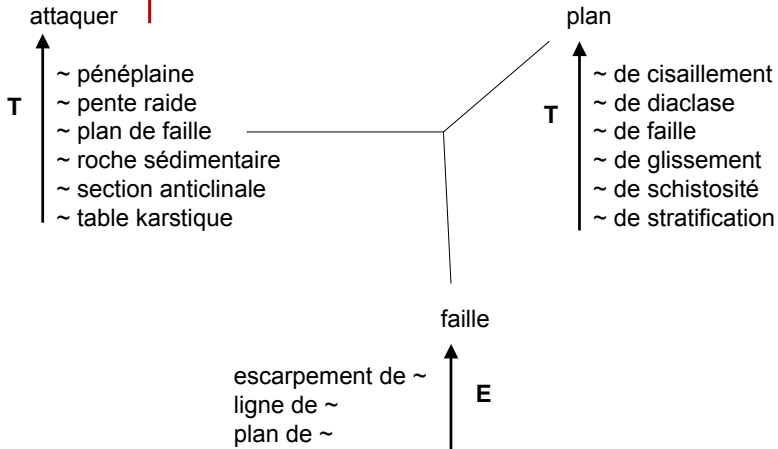
Traitement de l'information textuelle



Syntex (Bourigault, 2000)

- ❑ Extracteur de termes et mise en relation
- ❑ A partir d'un corpus étiqueté, produit :
 - un corpus analysé syntaxiquement
 - un réseau de syntagmes
 - ❑ syntagmes verbaux, nominaux, adjectivaux
 - ❑ réseau structuré par les relations Tête et Expansion
- ❑ Interface de consultation OntoTerm
- ❑ Principes de base
 - Analyse syntaxique et analyse distributionnelle

Syntax : réseau de syntagmes



Syntax : exploitation des contextes

Contextes pour le terme : <i>murmure vésiculaire</i>	Termes pour le contexte : (patient présenter , OBJ)
(abolir , OBJ)	<i>amyotrophie</i>
(abolir à gauche , OBJ)	<i>détresse</i>
(abolition , DE)	<i>douleur</i>
(diminuer , OBJ)	<i>douleur thoracique</i>
(diminuer à gauche , OBJ)	<i>dyspnée</i>
(diminution , DE)	<i>fièvre</i>
(percevoir , OBJ)	<i>fracture</i>
	<i>hématome</i>
productivité = 7	<i>Syndrome</i> productivité = 9

Caméléon : repérage et modélisation de relations (Séguéla, 2000)

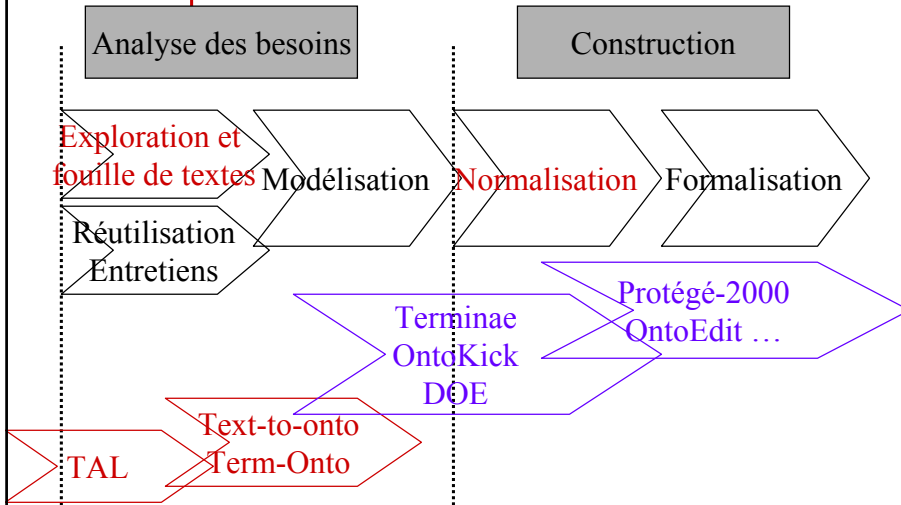
- ❑ La notion de marqueur :
 - Eléments lexico-syntaxiques permettant de repérer une relation conceptuelle
 - ❑ Det N1 est Det N2 (qui, adj, p.passé, p.présent)
 - ❑ Tous les N2 sauf det N1
 - ❑ Det N1 comme det N2
- ❑ Hypothèses :
 - Une même relation peut s'exprimer par différents marqueurs
 - Les relations peuvent dépendre du corpus
 - Les marqueurs peuvent dépendre du corpus
- ❑ Corpus étiqueté en entrée

Etapas d'utilisation

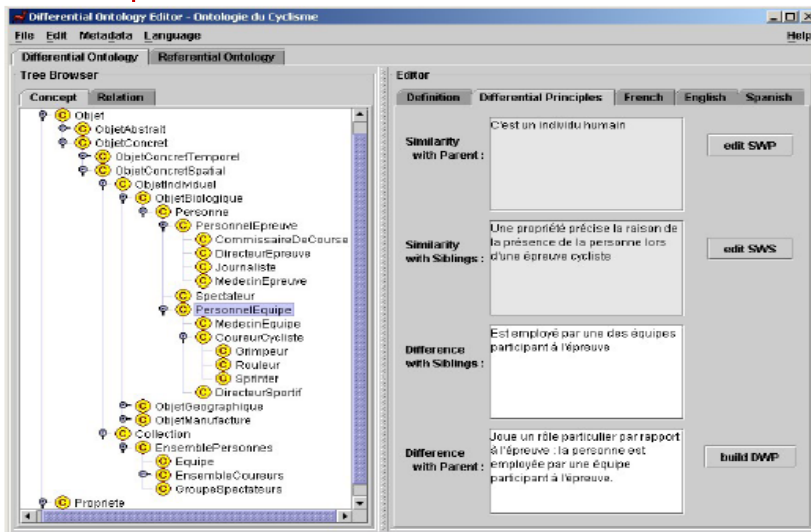
- ❑ Mise au point des marqueurs pour le corpus
 - Evaluation et adaptation de marqueurs génériques
 - Identification de patrons et relations spécifiques
- ❑ Enrichissement du modèle conceptuel
 - Evaluer les phrases où se trouvent les marqueurs :
 - ❑ présence ou non d'une relation sémantique
 - ❑ Identification des concepts en relation
 - Intégrer ces relations ou non dans le modèle



Logiciels supports



DOE : Differential Ontology Editor



Plates-formes pour la construction d'ontologies

□ Architecture

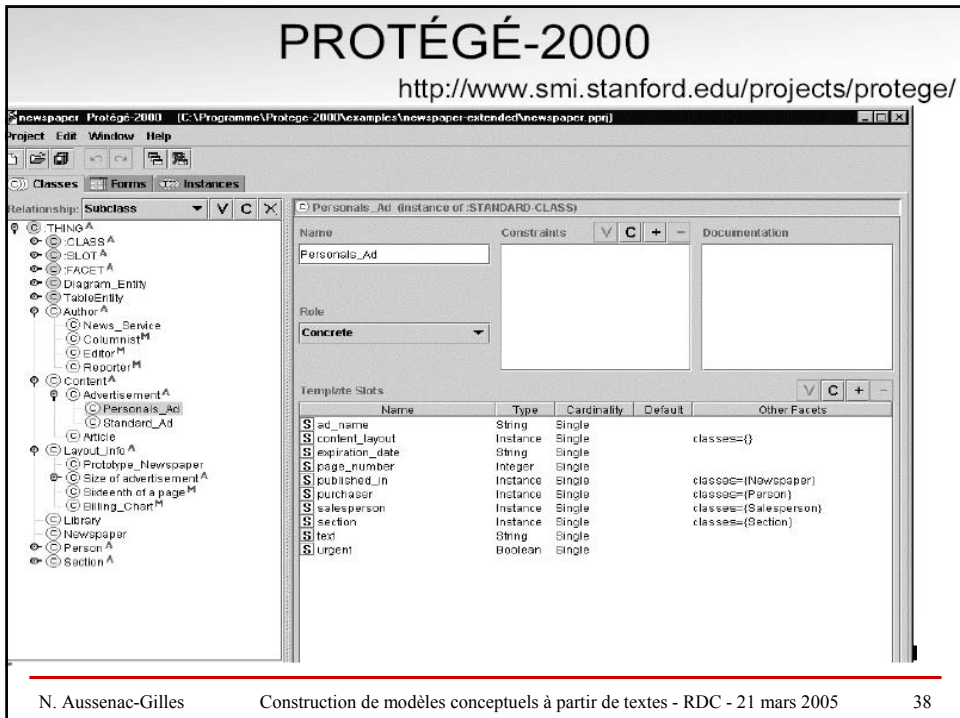
- Editeur + plug-in : Protégé-2000 (Stanford, Musen)
- Serveur et modules complémentaires : WebODE (UPM, Gomez-Pérez)
- Suite d'outils complémentaires : KAON = OntoQuick + Text-to-Onto + OntoEdit (karlsruhe, DFKI, Maedche et Staab)

□ Exportation en OWL

□ Couverture du processus

PROTÉGÉ-2000

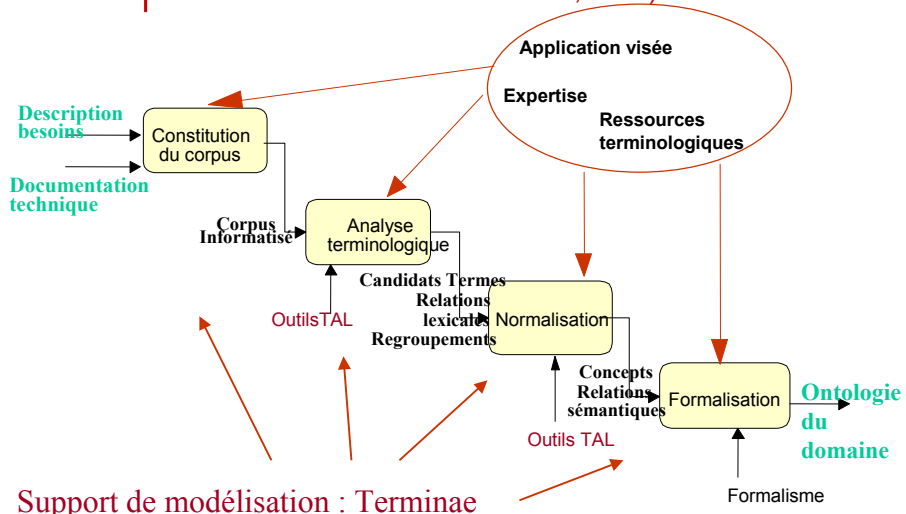
<http://www.smi.stanford.edu/projects/protege/>



The screenshot shows the Protégé-2000 interface. On the left is a class hierarchy tree with 'THING' at the root, followed by 'CLASS', 'SLOT', 'FACET', 'Diagram_Entity', 'TableEntity', 'Author', 'News_Service', 'Columnist', 'Editor', 'Reporter', 'Content', 'Advertisement', 'Personals_Ad', 'Standard_Ad', 'Article', 'Layout_info', 'Prototype_Newspaper', 'Size_of_advertisement', 'Sideenth_of_a_page', 'Billing_Chart', 'Library', 'Newspaper', 'Person', and 'Section'. The right pane shows the 'Personals_Ad' class details, including its name, role (Concrete), and a table of template slots.

Name	Type	Cardinality	Default	Other Facets
S ed_name	String	Single		
S content_layout	Instance	Single		classes={}
S expiration_date	String	Single		
S page_number	Integer	Single		
S published_in	Instance	Single		classes={Newspaper}
S purchaser	Instance	Single		classes={Person}
S salesperson	Instance	Single		classes={Salesperson}
S section	Instance	Single		classes={Section}
S text	String	Single		
S urgent	Boolean	Single		

TERMINAE : méthode (Biébow, Szulman, Aussenac-Gilles, 2000)



Support de modélisation : Terminae

Terminae, plate-forme de modélisation

❑ Fonctionnalités

- Etude linguistique
 - ❑ travail sur les résultats des outils Lexter ou Syntex
 - ❑ étude des relations Linguae
- Création de fiches terminologiques et conceptuelles
- Création d'ontologies (logique de description)

❑ Originalité

- Traçabilité des textes vers les modèles
- Intégration de résultats et outils d'analyse de textes
- Construction de terminologies ou d'ontologies

Applications utilisant des ontologies

- Le Web Sémantique
 - Annotation de pages web
 - Recherche d'informations hétérogènes (PICSEL)
 - Les web services
- Indexation de ressources
 - Textes, documentation d'entreprise, bases documentaires
 - Vidéo (INA)
- Veille technologique,
- Recherche d'information (Web ou intranet)

Web Sémantique : infrastructure

- **Infrastructures standards** : *protocoles, langages, mécanismes, ... pour*
 - exprimer ces descriptions
 - les échanger et intégrer des informations hétérogènes
 - traduire entre différents formalismes
 - raisonner à partir de ces descriptions, sécuriser ces échanges ...
- **Couche "sémantique"** permettant
 - la description des contenus (*métadonnées*)
 - la structuration des ressources (*liens hypertextes étiquetés*) ...
 - ... avec un certain niveau de formalisation
- **Vocabulaires** partagés pour différentes communautés
 - **ontologies** ... comme conceptualisations partagées, *au cœur du WS*
 - Permettent d'exprimer différents types de *métadonnées*

Scénario (1) : l'organisation de la visite d'une ville

Un outil dédié qui exploite et combine les ressources pour une tâche donnée

- Descriptions de ressources touristiques précises : *métadonnées*
 - Les différents musées, monuments, ...
 - Réutilisation d'une *ontologie* des objets culturels et historiques
- Connaissances générales sur le tourisme : *ontologie*
 - Les différents moyens de transport sont ...
- Connaissances sur les préférences de l'utilisateur : *profil attaché à sa page Web personnelle comme un ensemble de métadonnées* ...
 - Art Baroque, Art nouveau, Le Gréco, ...
- Connaissances sur les itinéraires ...

Implications du scénario (1)

L'outil dédié

- peut utiliser les mêmes langages de représentation des ontologies et de méta-données : *RDF, OWL* ...
- nécessite en plus : *l'intégration de sources de données hétérogènes et la médiation de requêtes globales vers des structures locales*
 - musées ayant chacun leur format de métadonnées
 - ...
- nécessite des capacités de raisonnement : *formalisation*

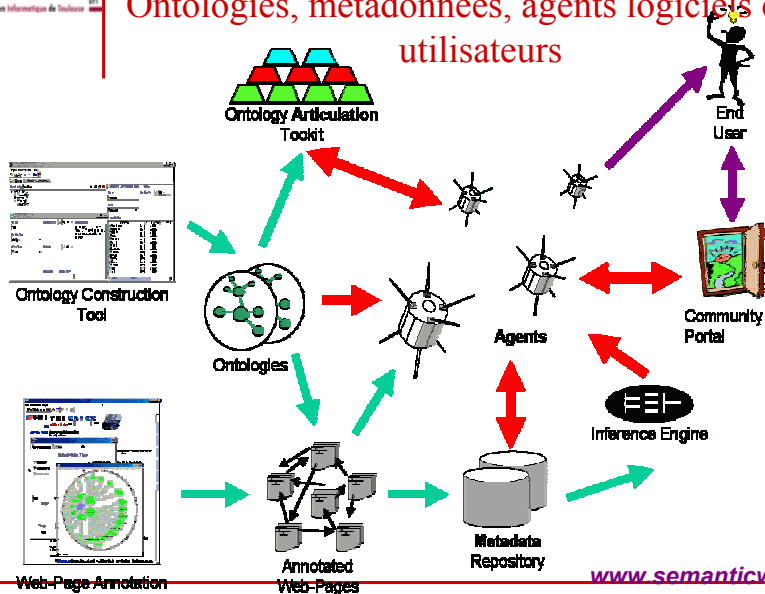
Scénario (2) : les services Web .. sémantiques

Un outil qui fait appel, exploite et combine des services

- Réservation d'un musée, d'un guide, d'un bateau
- Commande de brochures ...

- Accès aux services Web proposés par le portail du musée, de la compagnie de bateau
- Accès aux services Web de réservation des guides de la ville ...

Ontologies, métadonnées, agents logiciels et ... utilisateurs



Discussion : limites et points critiques

- ❑ Limites des ontologies
 - Coût de construction des RTO
 - Qualité, validation des RTO
 - Coût de l'annotation de pages et d'indexation avec une ontologie
 - Adéquation modèles/applications/utilisateurs
- ❑ Limites des approches à partir de textes
 - Comment décider de la description d'un concept ?
 - Consensus ? Portabilité ? Réutilisation ?
- ❑ Maintenance conjointe des modèles et des ressources
- ❑ Passage à l'échelle du Web, pertinence de ressources « générales »
- ❑ QUI va construire les ressources ? Annoter les pages Web ?

Recherches en cours

- ❑ Automatisation :
 - apprentissage d'ontologies (TAL et extraction d'information)
 - réutilisation et adaptation semi-automatique,
 - indexation ou annotation automatisée ...
- ❑ Fusion, comparaison, alignement d'ontologies
- ❑ Nécessaire inter-disciplinarité
 - Question du sens : linguistique, terminologie, IA, IC, recherche d'information ...
 - Question des usages : sciences de l'information, sociologie, ergonomie, IC ...

Questions d'actualité

- ❑ est-ce que cela a du sens de figer des représentations ?
sont-elles vraiment des connaissances pour un système ?
ou des ressources utiles à des applications ?
- ❑ comment rendre compte de la dynamique du langage,
des connaissances, des usages, des corpus de documents
dans lesquels on cherche des informations ?
- ❑ vers des processus dynamiques de reconstruction
régulière des réseaux terminologiques annotant des
documents

Éléments de bibliographie

- ❑ Uschold M. M., Gruninger M. Ontologies : principles, methods and applications. *Knowledge Engineering Review*. 1996
- ❑ Gomez-Pérez A., Fernandez-Lopez M, Corcho O., *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Verlag, 2004.
- ❑ Bachimont B. *Art et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle*. Mémoire d'habilitation à diriger des recherches de l'université de technologie de Compiègne. Janvier 2004.
- ❑ Charlet J., *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*. Mémoire d'habilitation à diriger des recherches en Informatique de l'université de Pierre et Marie Curie. Décembre 2002.
- ❑ Charlet J. Laublet P. Reynaud C., *Web sémantique*, rapport final de l'action spécifique 32 du CNRS/STIC. Déc. 2003. <http://www.ensib.fr/rtp-doc/>
- ❑ D. Bourigault, M.-C. L'Homme & C. Jacquemin (eds), *Recent Advances in Computational Terminology*, John Benjamins. 2000.
- ❑ Maedche A., *Ontology learning for the Semantic Web*. Kluwer Academic Publisher. 2002.
- ❑ Guarino N., Welty C., A formal Ontology of Properties. In *Proc. of the 12th International Conference on Knowledge Engineering and Knowledge Management*. Juan-Les-Pins (F). Oct 2000. R Dieng and O. Corby (Eds). LNAI Vol 1937. Springer Verlag. 2000. 97-112.
- ❑ M. Pierrel, M. Slodzian *Revue d'Intelligence Artificielle (RIA)*, Numéro spécial « Terminologie ». J (Ed.). Hermès : Paris. Vol. 16. N°1/ 2004