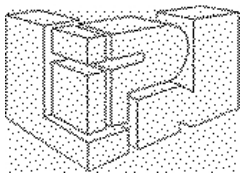


Méthode et outil de construction d'ontologies à partir de textes

Sylvie Szulman

LIPN
Université Paris 13

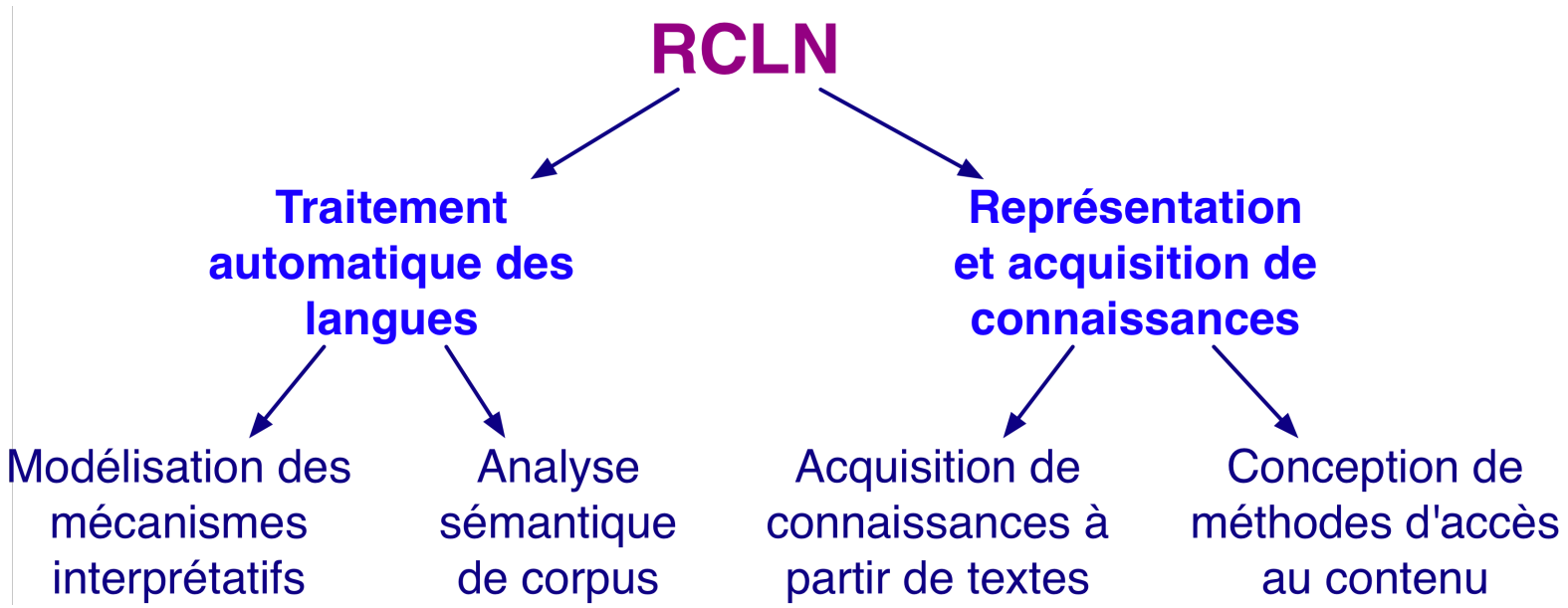
16 novembre 2011

Plan

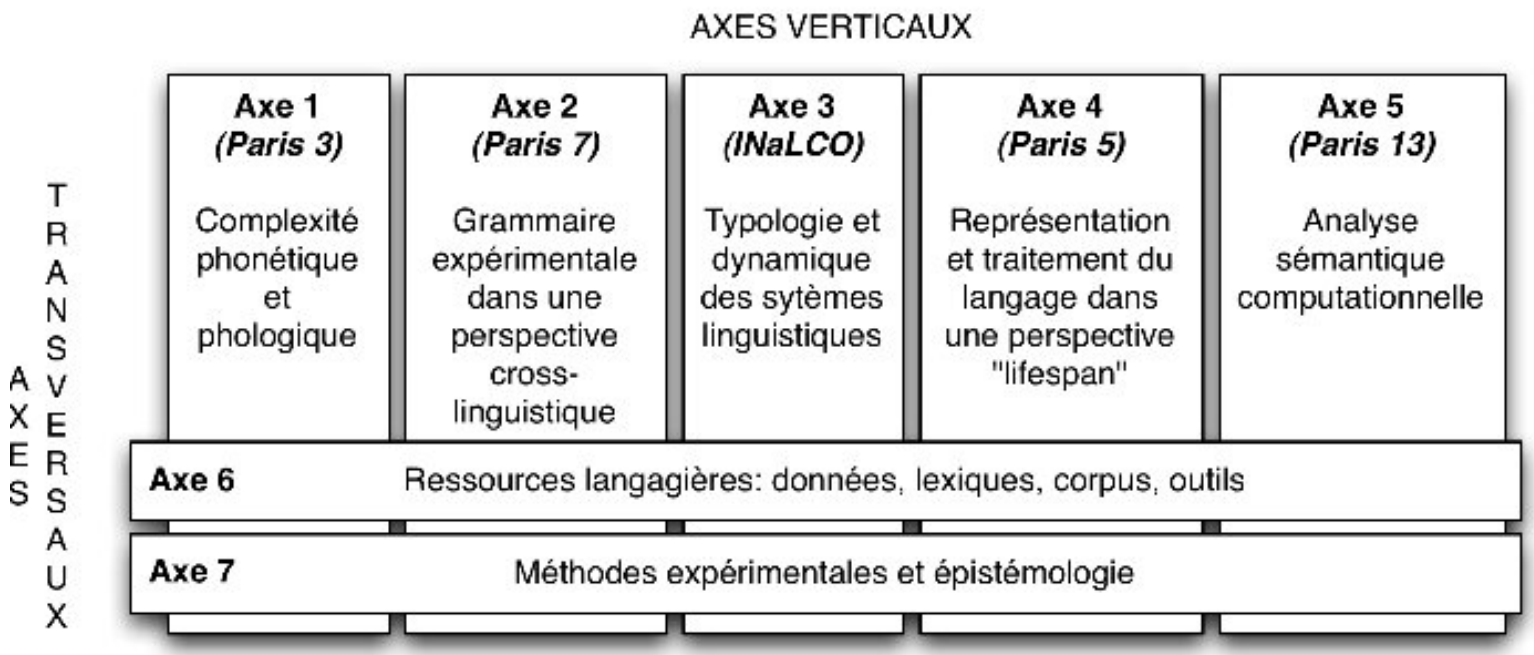
- Le LIPN -équipe RCLN
  - Equipe RCLN
  - LabEx EFL
  - Thématiques de Recherche
- Projets récents
  - Le projet Dafoe4App
    - Présentation générale
    - Objectifs
    - Réalisations
  - Le projet OntoRule
    - Présentation générale
    - Architecture du projet
    - Acquisition des modèles
    - Exploitation et recherche dans l'index
- De la terminologie à l'ontologie : le réseau de termino-concepts

LIPN = Laboratoire d'Informatique de Paris Nord (P13)
dirigé par Christophe Fouqueré
UMR7030
> 100 membres
5 équipes dont l'équipe RCLN

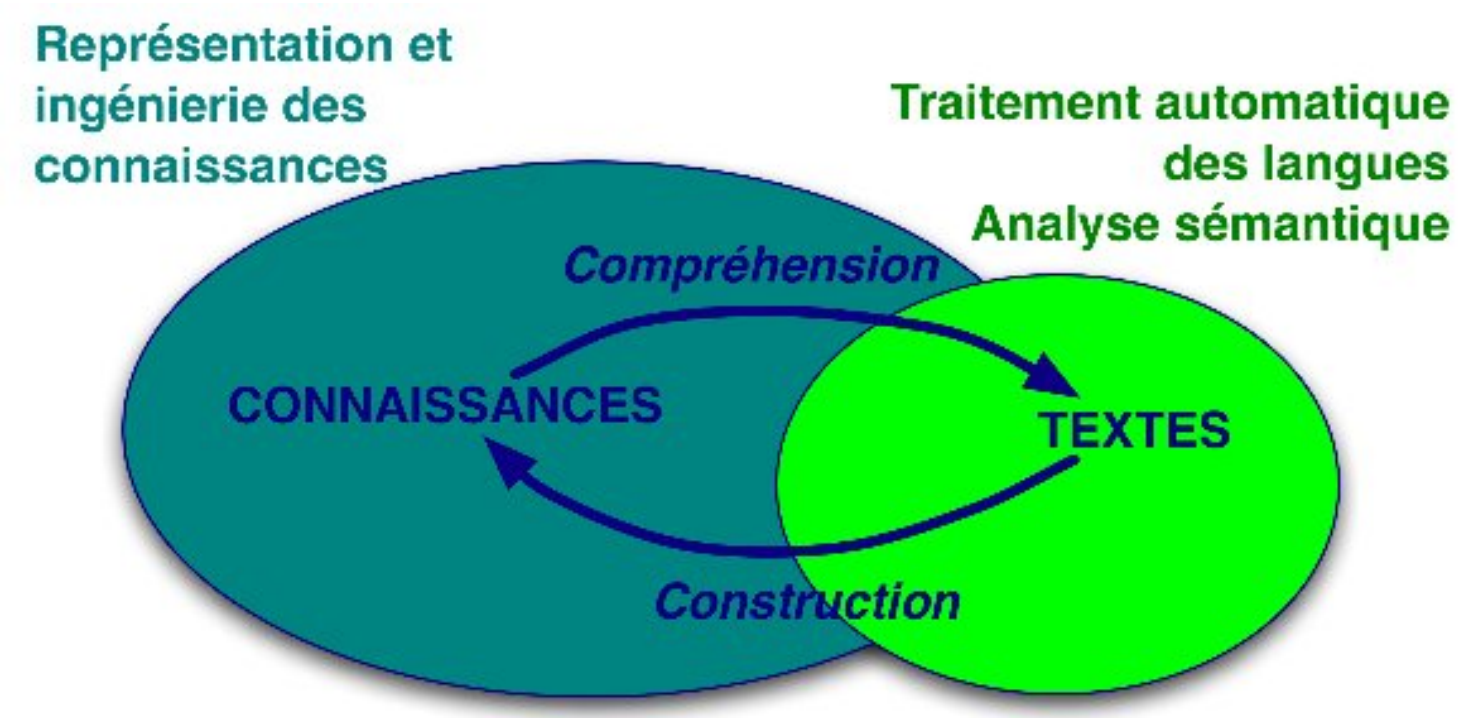
dirigée par Adeline Nazarenko



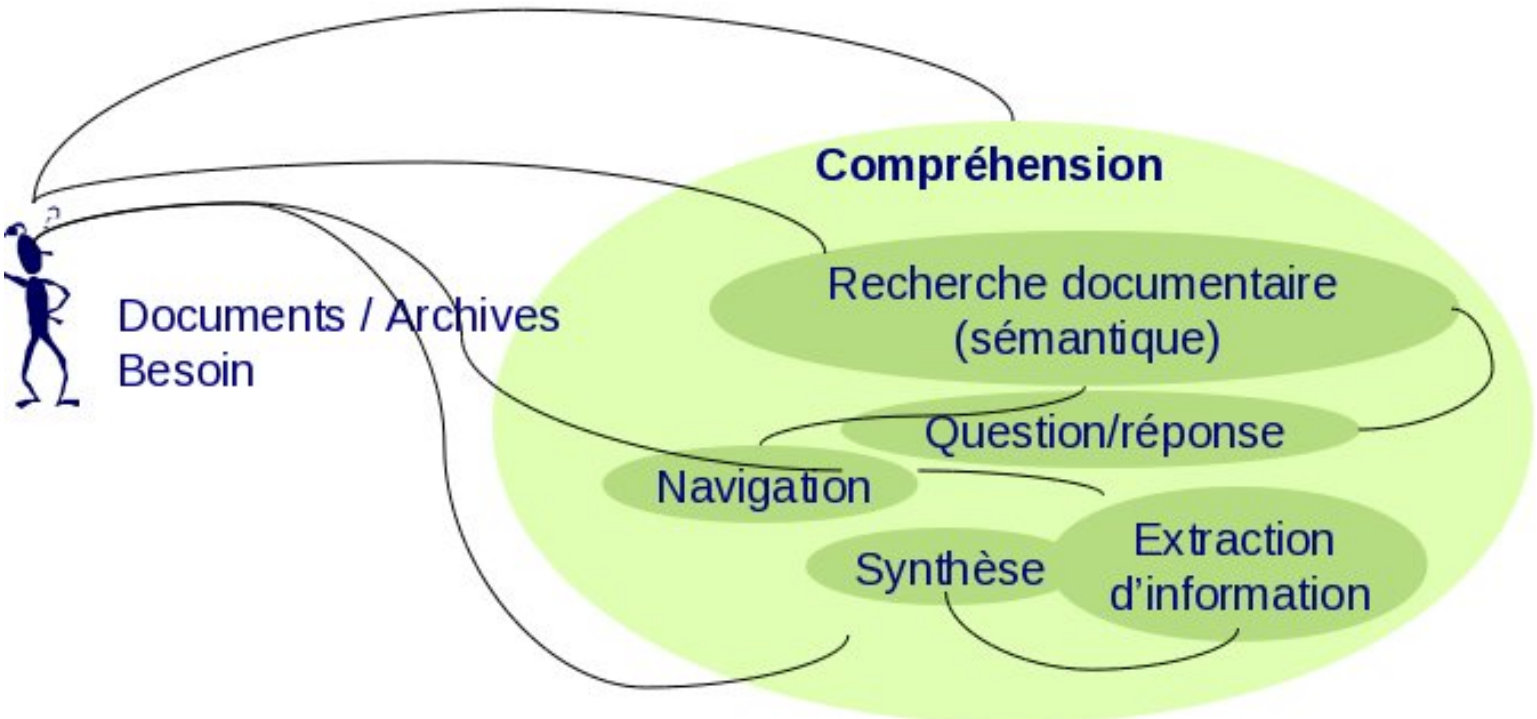
Fondements empiriques de la linguistique



Entre Textes et Connaissances

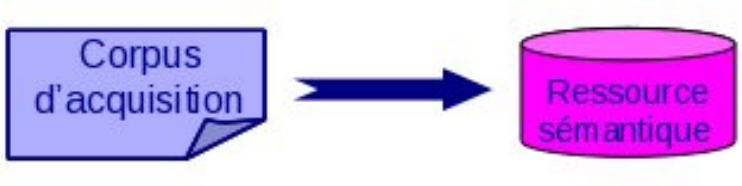


Compréhension



- Construction de connaissances à partir de textes
- Modèles et outils d'accès au contenu

Construction de connaissances à partir de textes



- Résultats
  - Outils de construction de terminologies : YaTea
  - Méthodes et outils d'acquisition de thesaurus, ontologies . . .
  - Méthodes et outils d'évaluation des méthodes d'acquisition
- Méthodes
  - Exploitation d'outils de TAL
  - Techniques d'apprentissage
  - Méthodes d'Ingénierie des connaissances : interaction, formalisation

Plan

Le LIPN -équipe RCLN
Equipe RCLN
LabEx EFL
Thématiques de Recherche
Projets récents
Le projet Dafoe4App
Présentation générale
Objectifs
Réalisations
Le projet OntoRule
Présentation générale
Architecture du projet
Acquisition des modèles
Exploitation et recherche dans l'index
De la terminologie à l'ontologie : le réseau de termino-concepts

Présentation générale

Colloque ANR contenus interactions et robotique
ANR-06-TLOG-010 (2007-2009)
Dafoe4App = Differential and formal ontology editor for Applications

- Responsable : J. Charlet (INSERM)
- Partenaires : INSERM UMR_S 872 equipe 20 (Paris 6/Paris 5), ENST/GET (Paris), IRIT (Toulouse 3), LIPN (Paris 13), LISI (Poitier), Mondeca (Paris), Supelec (Saclay), UTC (Compiègne)

Objectifs -1

- Créer une plateforme technique pour la construction de ressources sémantiques (thésaurus, ontologie) à partir de différentes sources (textes, base de données)
- Manipuler des éléments de connaissances tirés de ces sources afin de préparer une ontologie
- Conserver une trace entre l'ontologie construite et ces éléments de connaissances
- Fournir en résultat du processus une ressource à laquelle est associée une composante terminologique

Objectifs -2

- Pouvoir passer à l'échelle pour créer des ressources importantes à partir de corpus sans se préoccuper de leur taille.
- Fournir un éditeur d'ontologie qui prend en charge toute la question de la sémantique de ces ontologies, à travers des questions épistémologiques liées aux concepts formels de haut niveau (la top ontology ou ontologie catégoriale).

Réalisations

- Séparation du coeur de la plateforme et des greffons
- Réalisation d'un cahier des charges avec prise en compte de l'ensemble des scénarios de l'ensemble des participants
- Spécification précise du modèle de données

4 espaces de travail

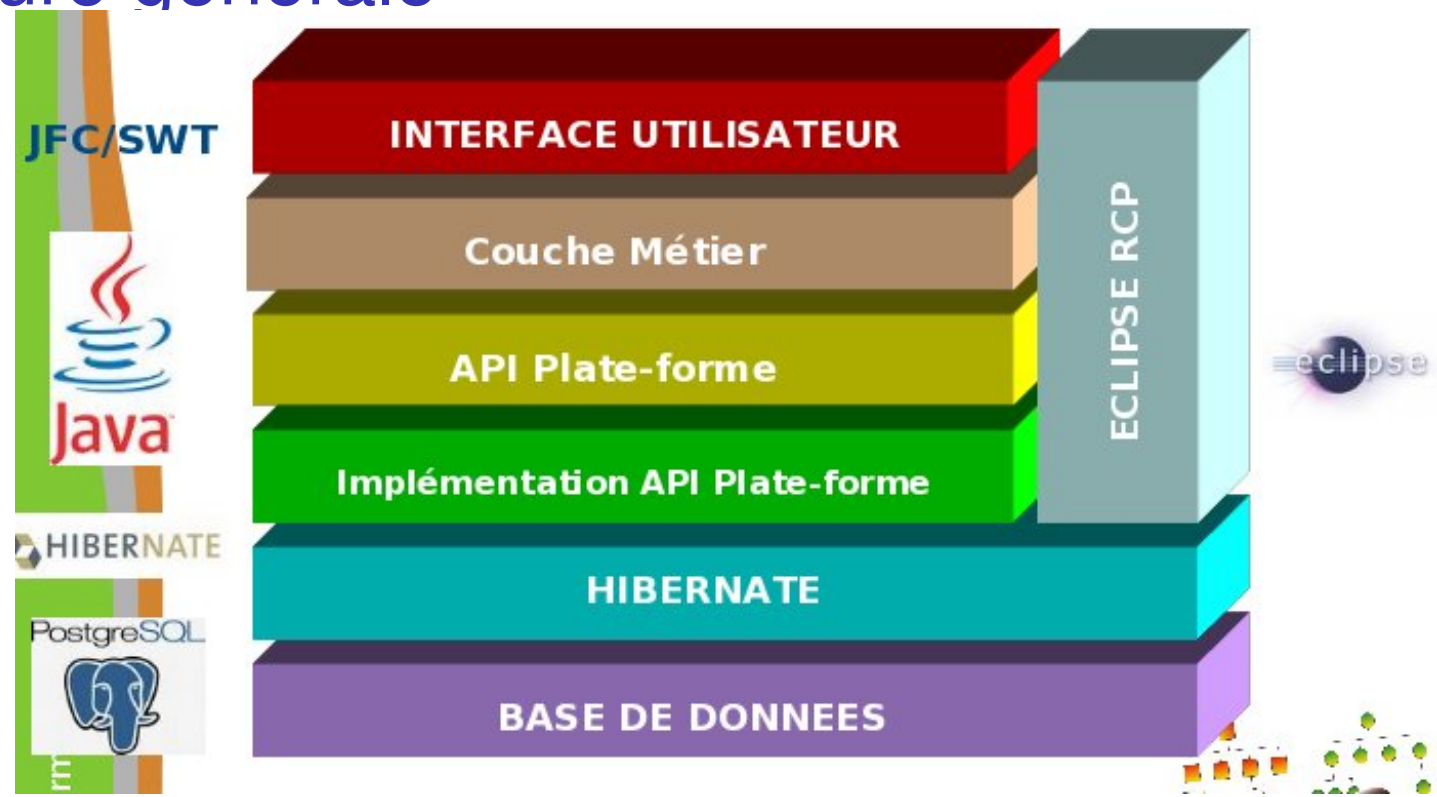
- A chaque niveau du modèle de données, correspond un espace de travail identifié par une couleur et une icône.

Niveau 3
Niveau 2
Niveau 1
Niveau 0

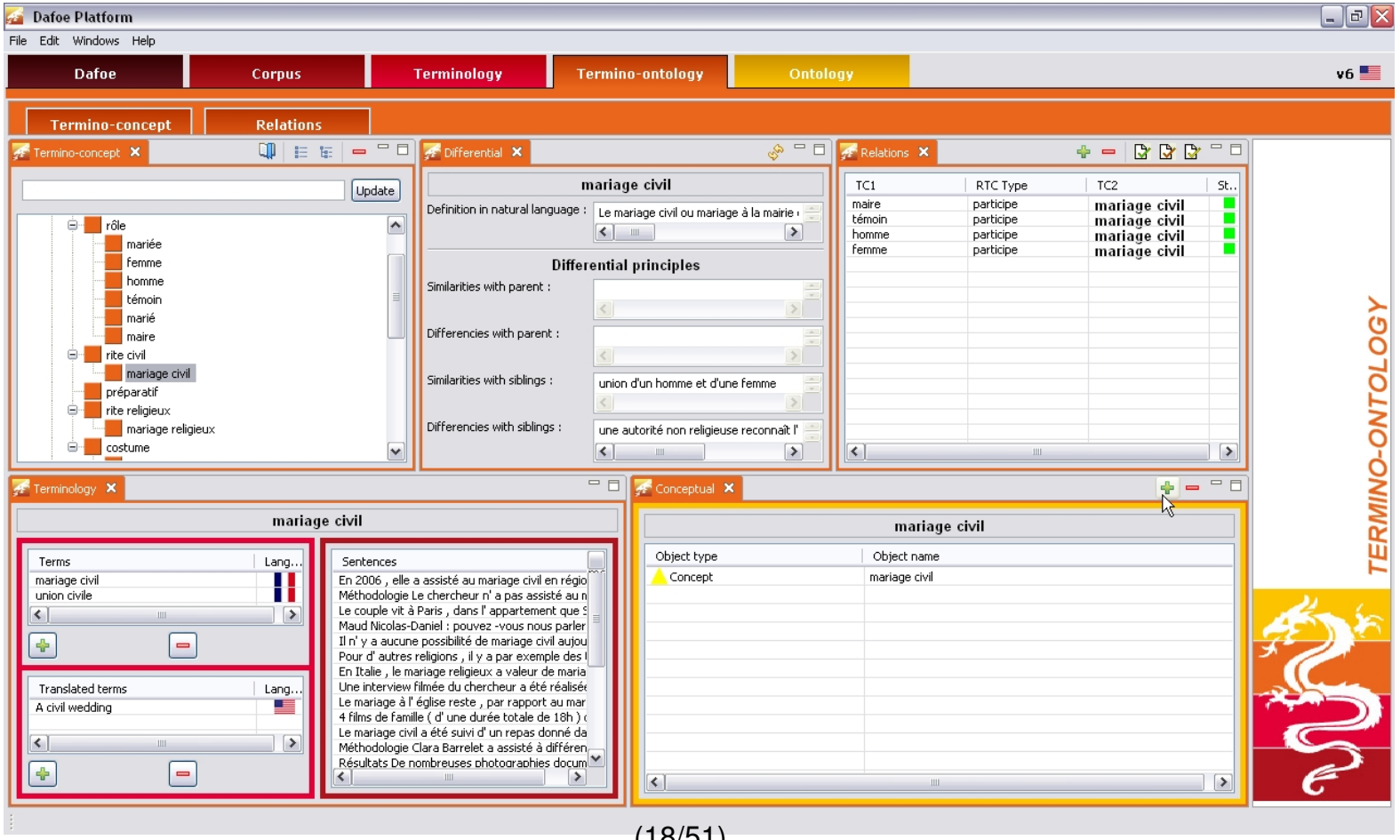
Développement

- Une société de développement
- Réaliser comme une application eclipse en Java (Open Source)
- Un prototype a été livré

Architecture générale



Exemple


Dafoe Platform
File Edit Windows Help
Dafoe
Corpus
Terminology
Termino-ontology
Ontology
v6
Termino-concept
Relations
Termino-concept
Update
rôle
mariée
femme
homme
témoin
marié
maire
rite civil
mariage civil
préparatif
rite religieux
mariage religieux
costume
Differential
mariage civil
Definition in natural language :
Le mariage civil ou mariage à la mairie
Differential principles
Similarities with parent :
Differencies with parent :
Similarities with siblings :
union d'un homme et d'une femme
Differencies with siblings :
une autorité non religieuse reconnaît l'
Relations
TC1
RTC Type
TC2
maire participe mariage civil
témoin participe mariage civil
homme participe mariage civil
femme participe mariage civil
Terminology
mariage civil
Terms
Lang...
mariage civil
union civile
Translated terms
A civil wedding
Sentences
Conceptual
mariage civil
Object type
Object name
Concept
mariage civil
TERMINO-ONTOLOGY

Objectifs

OntoRule (http ://ontorule-project.eu/)

ONTORULE
Ontologiesmeet Business Rules

(Janvier 2009 - Décembre 2011)

Les outils décisionnels utilisent des moteurs d'inférence puissants mais la création et la maintenance de la base de règles doivent être améliorées.

Plateforme d'acquisition, de gestion, de maintenance et d'exécution de règles métiers.

Fournir une méthodologie et des outils logiciels

- utilisant les standards (SBVR, PRR, RIF, OWL)
- permettant de gérer et maintenir l'ensemble de règles et leurs ontologies

Partenaires

- IBM France (coordinateur)
- Ontoprise
- Free University of Bolzano
- Vienna University of Technology
- PNA
- Université Paris 13
- Fundación CTIC
- Audi
- ArcelorMittal

Les objectifs détaillés du projet pour le LIPN (WP1-WP2)

- Acquérir les règles et l'ontologie à partir de textes
- Travail assisté, non automatique
- reposant sur
  - annotation sémantique des textes
  - navigation dans l'ontologie/ les textes /les règles
- pour
  - Acquérir le "business model"
  - Expliquer les décisions en se référant aux textes plutôt qu'aux règles formalisées,
  - Maintenir le modèle lorsque la

Les objectifs détaillés du projet pour le LIPN (WP1-WP2)

- Acquérir les règles et l'ontologie à partir de textes
- Travail assisté, non automatique
- reposant sur
  - annotation sémantique des textes
  - navigation dans l'ontologie/ les textes /les règles
- pour
  - Acquérir le "business model"
  - Expliquer les décisions en se référant aux textes plutôt qu'aux règles formalisées,
  - Maintenir le modèle lorsque la

Les objectifs détaillés du projet pour le LIPN (WP1-WP2)

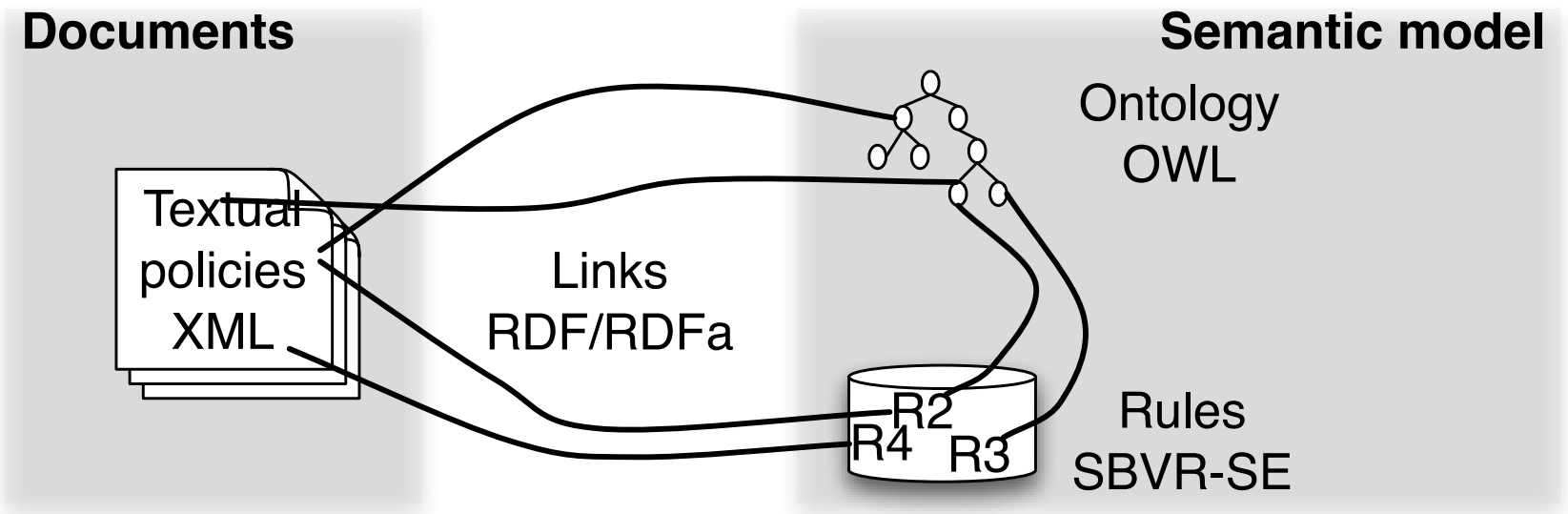
- Acquérir les règles et l'ontologie à partir de textes
- Travail assisté, non automatique
- reposant sur
  - annotation sémantique des textes
  - navigation dans l'ontologie/ les textes /les règles
- pour
  - Acquérir le "business model"
  - Expliquer les décisions en se référant aux textes plutôt qu'aux règles formalisées,
  - Maintenir le modèle lorsque la

2 cas d'utilisation :

- AAdvantage (cas d'utilisation jouet)
  Le texte décrit les règles d'attribution des bonus lors des voyages avec la compagnie American Airlines
- Audi
  Le texte est une directive européenne décrivant les règles à observer par les constructeurs automobile relatives à la conformité des ceintures de sécurité.

Modèle des règles métiers documentés

- Progressivement construit lors de la phase d'acquisition
- Construction d'une structure de données qui lie le document source, l'ontologie et les règles appelée "index"



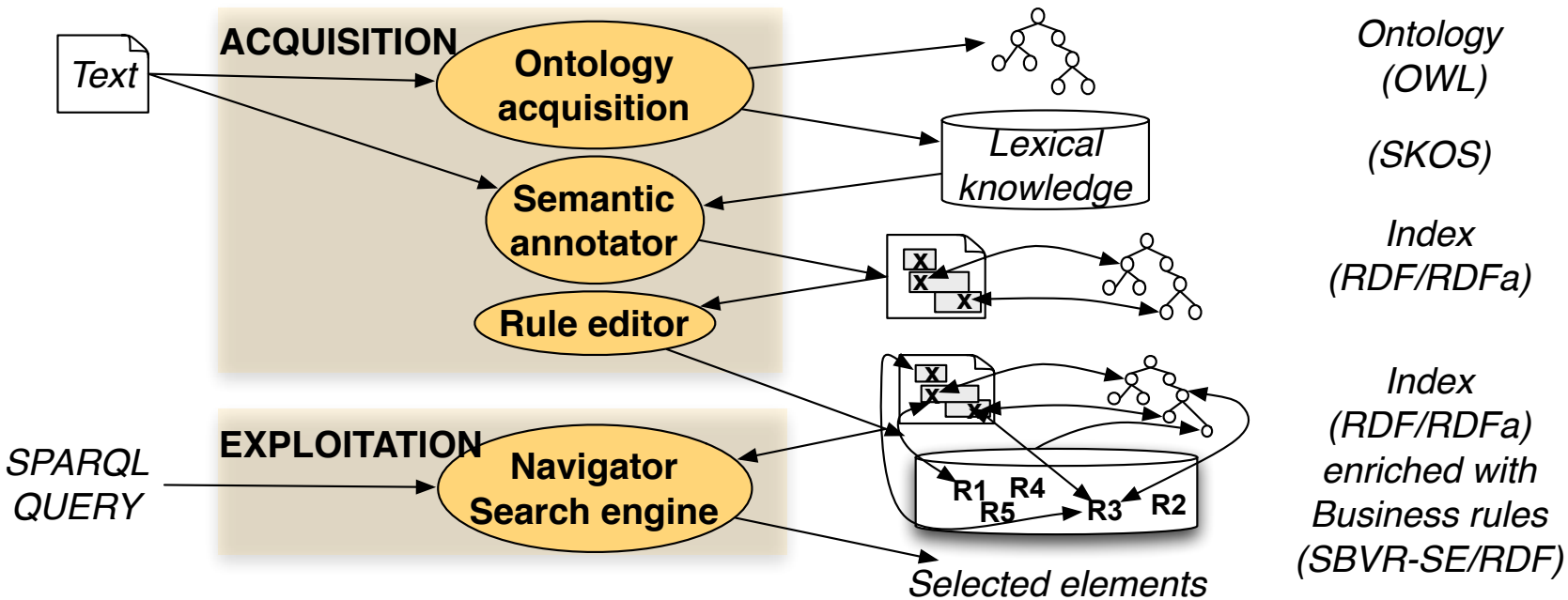
Les modèles -1

- modèle pour le document
  - un ensemble d'unités textuelles
  - une unité textuelle est repérée par l'offset de son premier caractère
  - une annotation d'une unité textuelle dans le document est exprimée à l'aide de RDFa

Les modèles -2

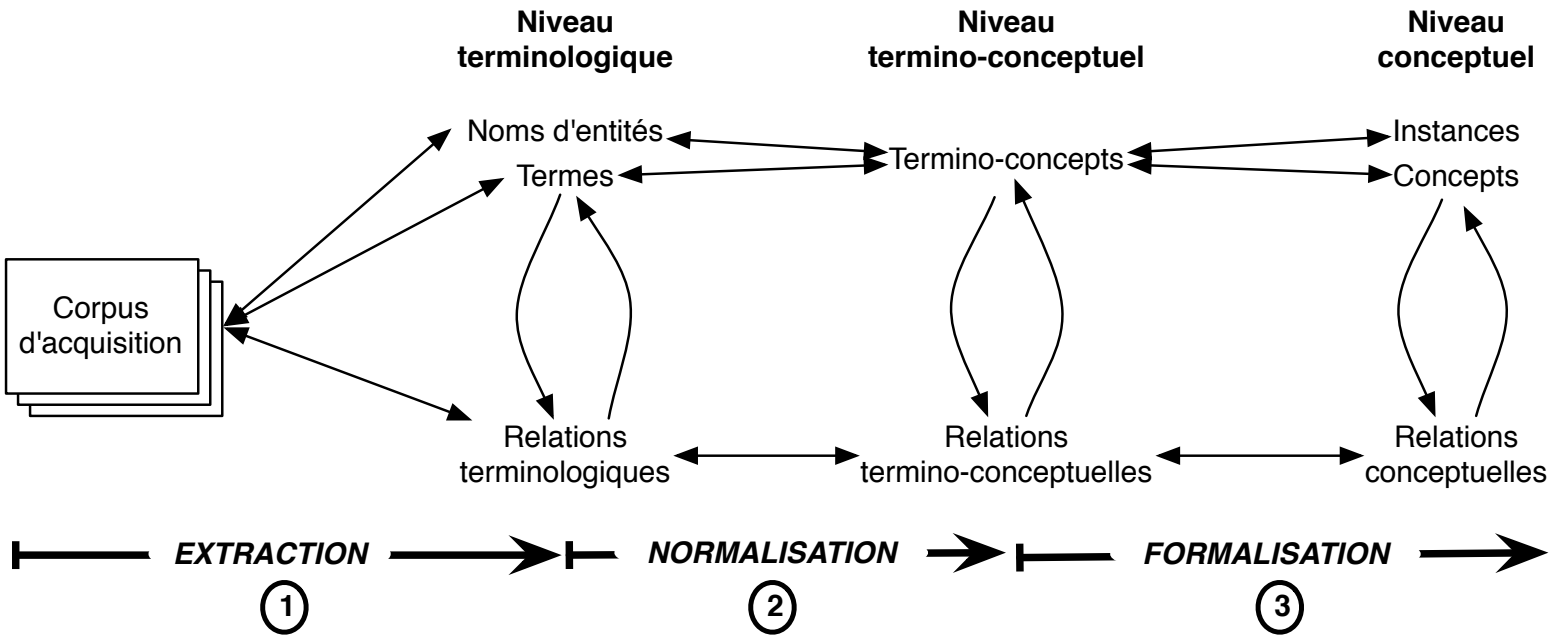
- une ontologie exprimée en OWL
- un modèle pour les règles
  - définition de règles *candidates* (expression dans le texte d'une règle)
  - réecriture de la règle candidate éventuellement en plusieurs règles dans un langage naturel en évitant les phénomènes de langue comme les ellipses, les métonymies . . .
  - classification de la règle : contrainte statique, règle effective, réglementaire. . .
  - Ecriture en SBVR (Semantic of Business Vocabulary and Rule) -Anglais simplifié

Le processus d'acquisition



Création de l'ontologie de domaine

Méthode TERMINAE



Analyse linguistique : filtrage

- Extraction automatique des unités textuelles : ex *airline participant*
- Filtrage manuel ou semi-automatique des unités textuelles
- Regroupement d'unités textuelles : ex regroupement des candidats termes
  e.g. *airline participant* regroupé avec *Airline participant* et *participant*

Normalisation

- Plan linguistique
- constitué de 2 phases.
  - Analyse terminologique :
    - Identifier les variantes de natures diverses : graphique (NY vs. New York), morphologique (member vs. members) ou bien morpho-syntaxique (program member vs. member of the program)
    - Identifier des unités synonymes (qui décrivent un même sens ou renvoient à une même entité du monde) et les regrouper sous une forme canonique
  - Construction du réseau termino-conceptuel

Niveau Termino-conceptuel/1

Un termino-concept est un terme désambiguïsé dont le sens est défini par son usage dans le corpus

Un termino-concept est décrit par :

- un terme vedette (terme préféré)
- un ensemble de termes synonymes
- une définition en langage naturel
- un ensemble de liens avec d'autres termino-concepts.

ex participant
2 utilisations soit membre, soit organisation
Création de 2 termino-concepts ***Participant*** et ***Member***
Les 2 termino-concepts sont liés au terme *participant*

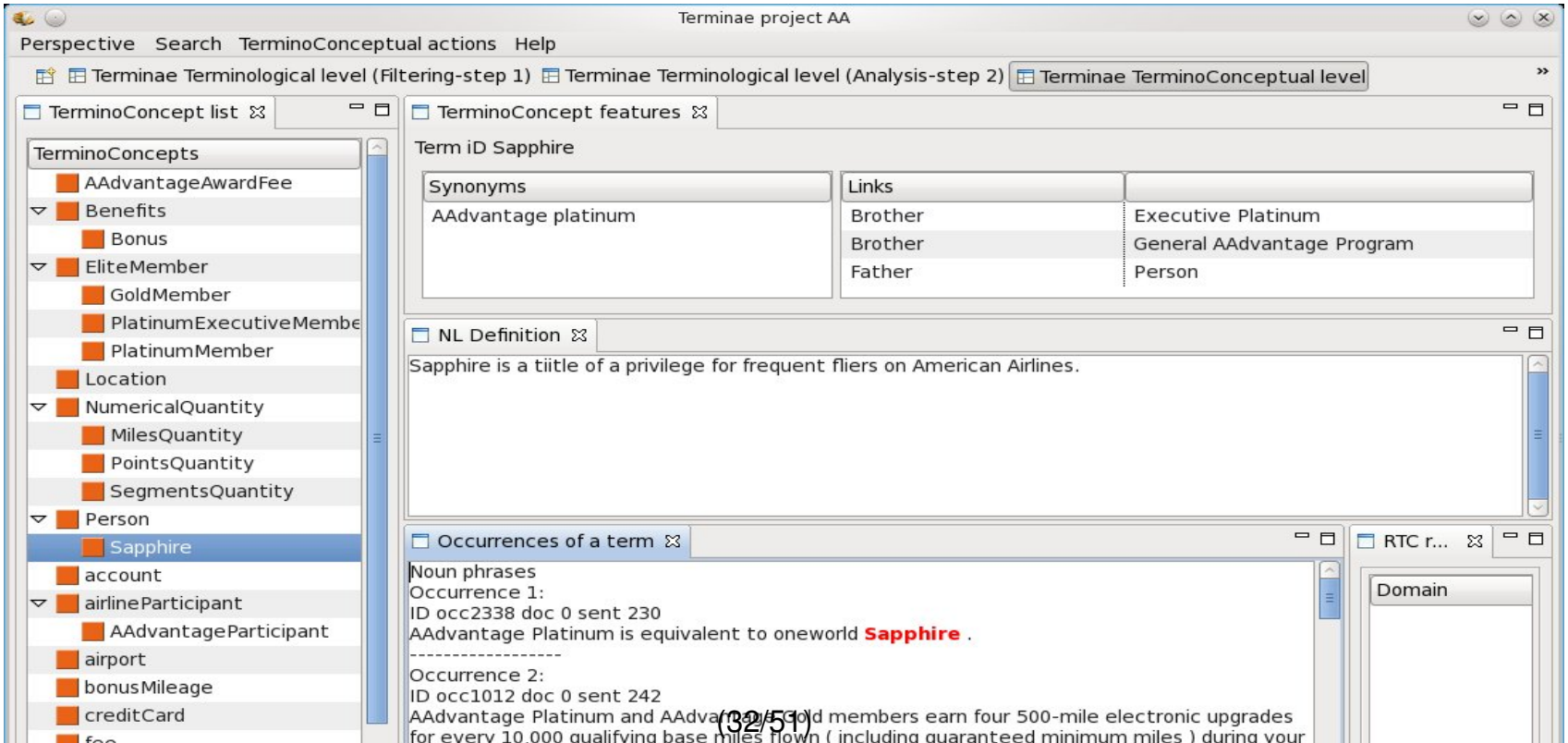
Formalisation

- Formalisation dans l'ontologie e.g. Creation de 2 concepts :
  - **Participant** fils du concept **Company**
  - **Member** fils du **Client**

Un outil pour l'expert

L'outil Terminae

(http ://www-lipn.univ-paris13.fr/ szulman/TerminaeWorkbench/)



Annotation des textes

- Les connaissances linguistiques sont sauvegardées en SKOS

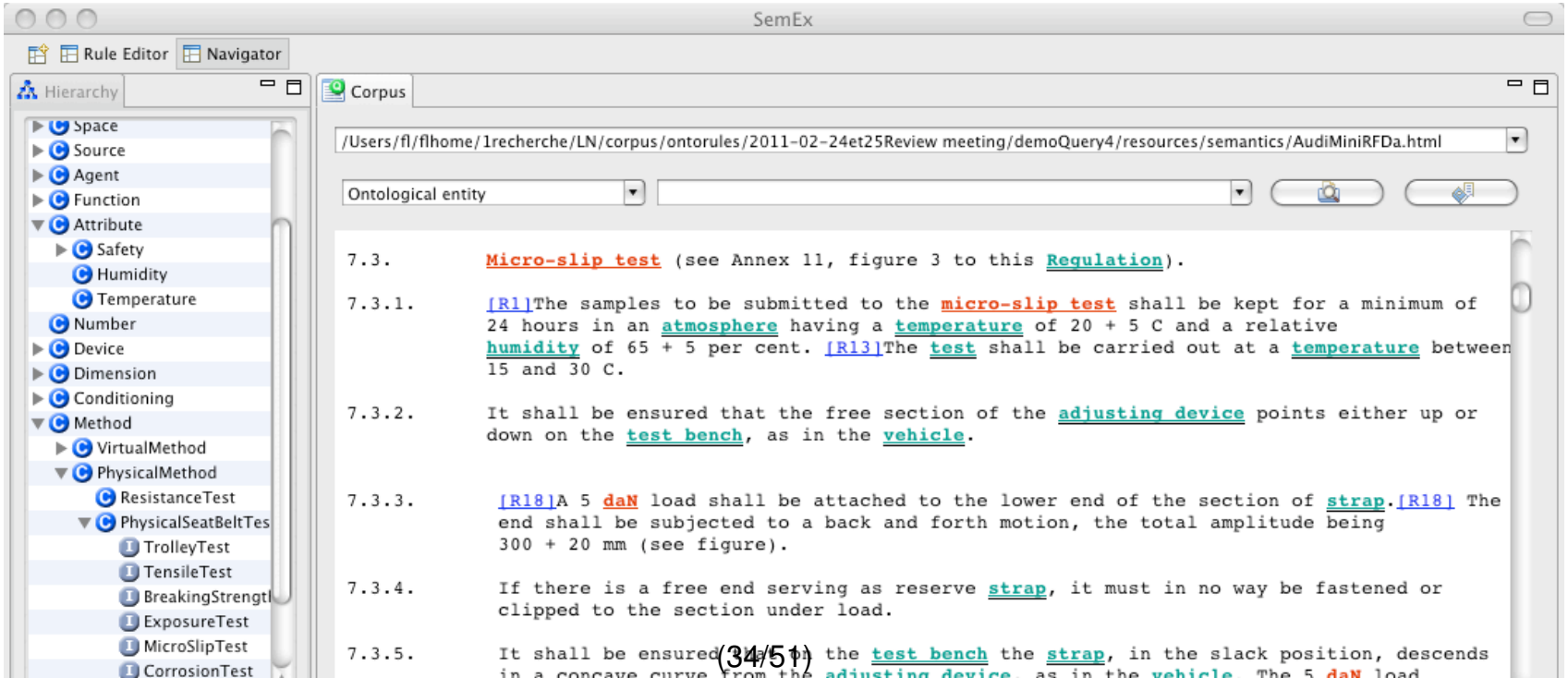
```
<skos:Concept rdf:about="http://ontorule#AdjustingDevice">
    <skos:prefLabel>adjusting device</skos:prefLabel>
    <skos:definition />
    <skos:altLabel>adjusting devices</skos:altLabel>
    <skos:altLabel>belt adjustment devices</skos:altLabel>
    <skos:example>index Corpus = 0 DOC = 0  Sentence = 48  Start_position = 5  End_posi
Text =Belt adjusting device .</skos:example>
    <skos:example>index Corpus = 0 DOC = 0  Sentence = 1004  Start_position = 119
End_position =135  Text =When the belt is being worn , it shall either adjust automatic
    <skos:example>index Corpus = 0 DOC = 0  Sentence = 383  Start_position = 5
End_position =21  Text =Belt adjusting device .</skos:example>
```

- peut être utilisées pour annoter de nouvelles versions ou d'autres textes sur le même domaine.

Acquisition des règles métiers

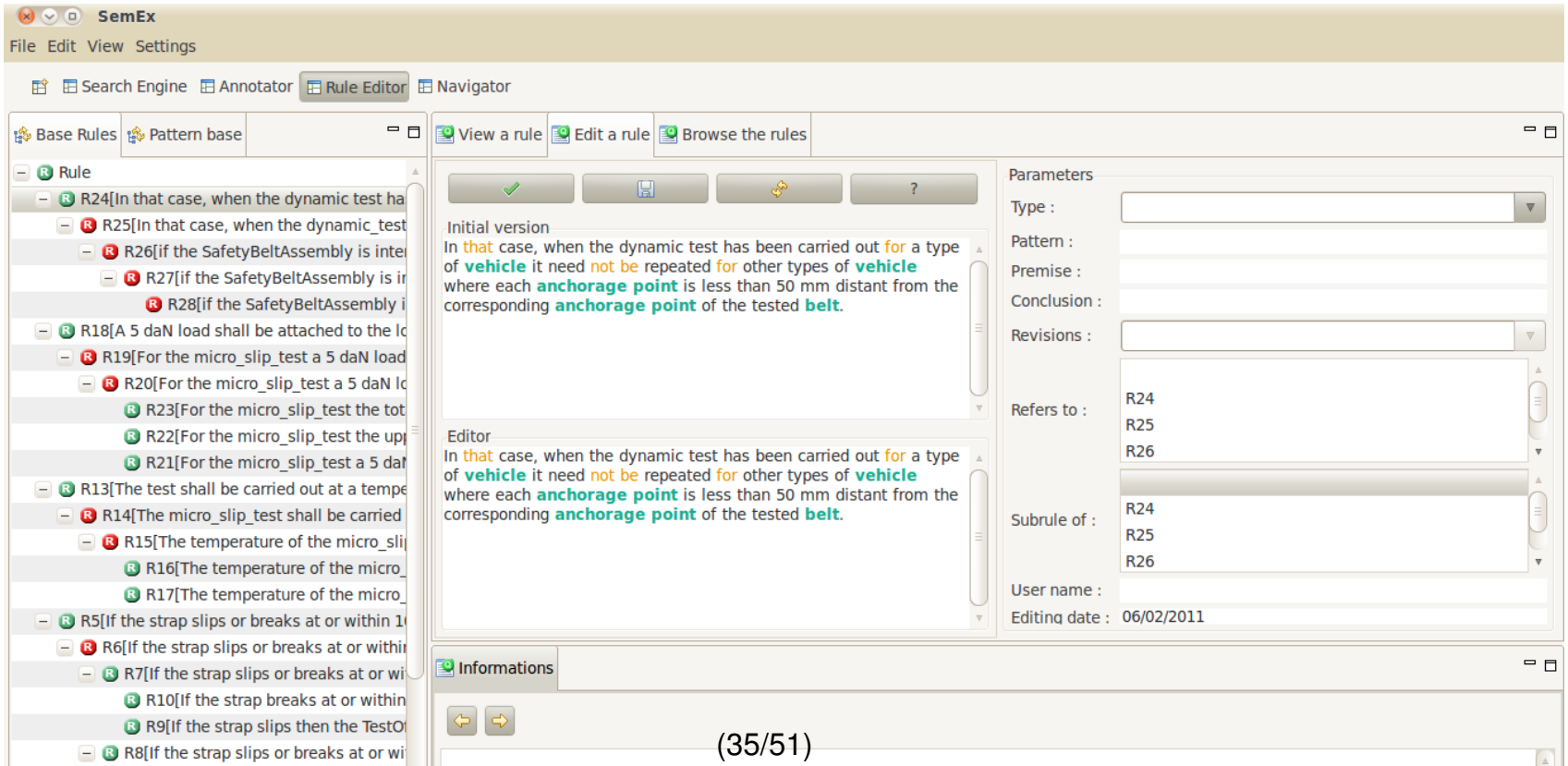
- Strategie : analyse et exploration du texte
- SemEx : un outil pour eliciter et explorer les règles métiers

exploration sémantique



Le projet OntoRule

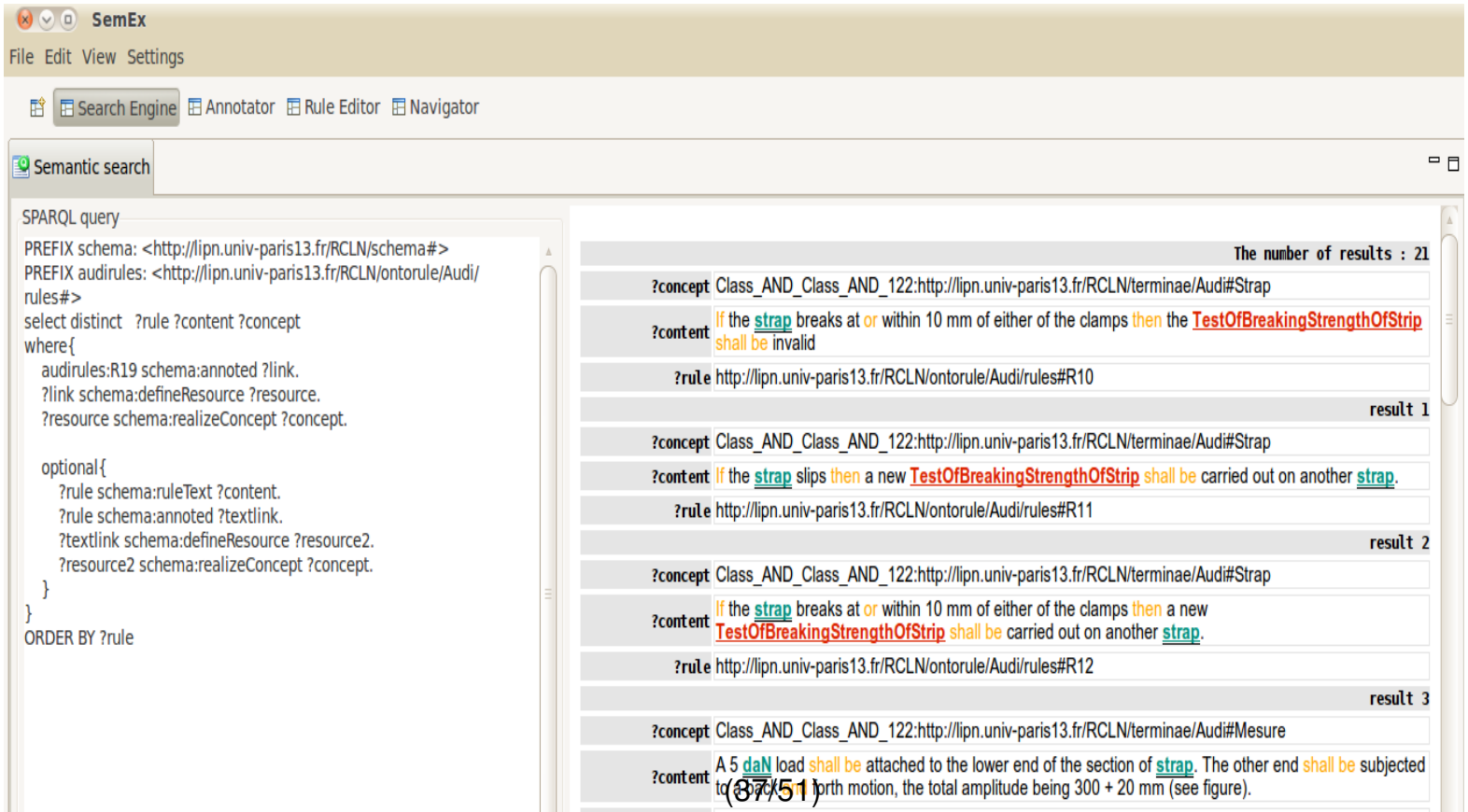
- les marqueurs linguistiques sont soulignés
- selection d'une règle et copie dans l'éditeur de règle



Recherche sémantique

- Recherche classique sur le texte
- Recherche dans l'ontologie
- Feuilleter la base de règles
- Naviguer d'une ressource à une autre
  - des concepts aux phrases qui utilisent les termes qui leur sont associés
  - des concepts aux règles qui utilisent les termes qui leur sont associés
  - entre les phrases et les règles qui sont associés aux même concepts

Questions Sparql sur tout l'index



Plan

- Le LIPN -équipe RCLN
  - Equipe RCLN
  - LabEx EFL
  - Thématiques de Recherche
- Projets récents
  - Le projet Dafoe4App
    - Présentation générale
    - Objectifs
    - Réalisations
  - Le projet OntoRule
    - Présentation générale
    - Architecture du projet
    - Acquisition des modèles
    - Exploitation et recherche dans l'index

Travail de réflexion mené avec Adeline Nazarenko à la suite des projets

- Terminae ancienne version
  définition du concept terminologique
- Dafoe4App
  définition du termino-concept
- OntoRule
  Utilisation du réseau de termino-concepts

Ressources sémantiques

- Terminologie
- Thesaurus
- Ontologie
- Ressource Termino-ontologique

Terminologie

- Terminologie : une terminologie est une construction linguistique qui fournit une liste de termes. Pour chaque terme, il y a un ensemble de caractéristiques linguistiques comme (éléments morphologiques, l'équivalent dans d'autres langages . . .)

Thesaurus

- un thesaurus définit un langage formel constitué de descripteurs, organisés hiérarchiquement liens BT/NT pour broader than/ narrower than). Les descripteurs sont des éléments linguistiques qui ont un sens specifique et stable dans un domaine donné.
- un thesaurus est généralement utilisé pour classer des documents et pour faire de la recherche d'information
- exemple : le MeSH (Medical Subject Headings) est un thesaurus servant à indexer les articles de PubMed.

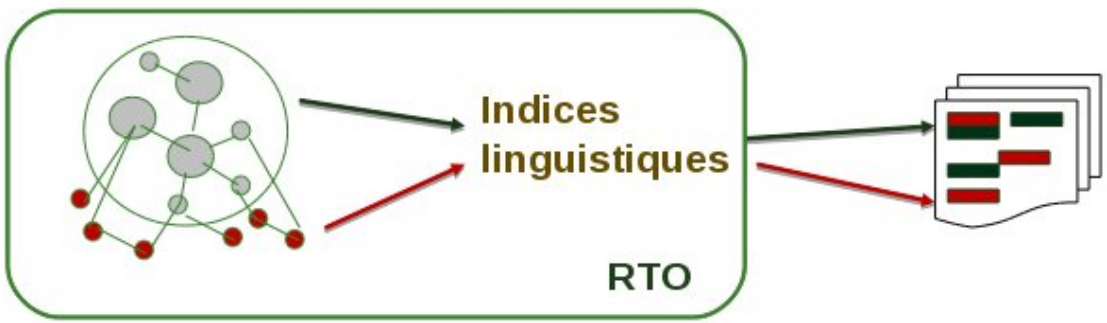
Ontologies

Une ontologie est une spécification normalisée représentant les classes des objets reconnus comme existant dans un domaine (Charlet 2002).

Définitions

Deux significations dans l'état de l'art

- Ontologie construite à partir de ressources linguistiques
- Ontologie enrichie, complétée d'informations linguistiques



emprunté à N. Aussenac-Gilles

Vocabulaire : mots, termes, concepts

Element	Ensemble
mots relations syntaxiques	Textes
Termes - relations terminologiques	Terminologies
Descripteurs - relations sémantiques	Thesaurus
Concepts - Roles propriétés	Ontologies

Un termino-concept peut s'apparenter à un descripteur dans un thesaurus mais en plus du terme qui le désigne, sa signification est exprimée par les occurrences dans le corpus de l'ensemble des termes qui lui correspondent.

- Un termino-concept est caractérisé par
  * sa pertinence pour le domaine à modéliser
  * il est non-ambigü
  * il n'a pas de synonyme au niveau termino-conceptuel

- Plusieurs termes peuvent correspondre au même termino-concept. Ils sont considérés comme synonymes dans le contexte de la modélisation.
- Un terme peut correspondre à plusieurs termino-concepts si il est ambigü et les différentes significations sont pertinentes dans le domaine.
- Certains termes n'ont pas d'équivalent au niveau termino-conceptuel car considérés comme non pertinents pour le domaine à modéliser.

Les termino-concepts sont organisés sous forme de réseau sémantique dont les relations sont :

- soit génériques/spécifiques non transitives, sans héritage de propriétés mais sans cycle
- soit associatives qui peuvent être symétriques ou inverses
- si deux termino-concepts sont liés par une relation associative, ils ne peuvent être liés par une relation spécifique/générique

- Un réseau termino-conceptuel est construit à partir d'un réseau de termes (terminologie)
- Un réseau termino-conceptuel sert à construire une ontologie.
  Chaque termino-concept peut être formalisé par :
  - un concept (classe)
  - une instance
  - une propriété

- Pour la classification comme un thesaurus
- Comme une terminologie normative pour le modèle d'un domaine
- Si le réseau de termino-concepts est construit à partir d'une terminologie et si il sert de base à construction d'une ontologie, il permet de réaliser une RTO sous la forme d'un couplage faible entre les unités lexicales et les éléments conceptuels créés dans une ontologie.e

Pour finir

terme <> concept ! ! !