Construction d'une ontologie à partir d'un thésaurus de l'astronomie: vers une représentation plus sémantique des données textuelles

Nathalie Hernandez, Josiane Mothe

IRIT, 118 route de Narbonne, 31040 Toulouse cedex
{hernande,mothe}@irit.fr

Contexte général

Système d'information

- Système de recherche d'information
- Système d exploration

- Contenu des documents
- Meta-données

→ Représentation des données textuelles

Problématiques

- Vocabulaire (besoin utilisateur, corpus)
- Approches de la littérature : mots clés de la requête recherchés dans les documents [Salton 1971] [Spärck Jones 2003]

$\rightarrow$perte du contexte et du vocabulaire

$\Rightarrow$Maintenir ces liens pour s'adapter au vocabulaire de l'utilisateur, de la tâche, de la collection

Problématiques

- Indexation
  - Indexation intégrant la variabilité des mots utilisés dans les collections vs indexation par mots clés

  ⇒ Meilleure représentation pour un meilleur appariement
- Accès
  - Vues globales et spécifiques sur la connaissance modélisée et sur le contenu de la collection

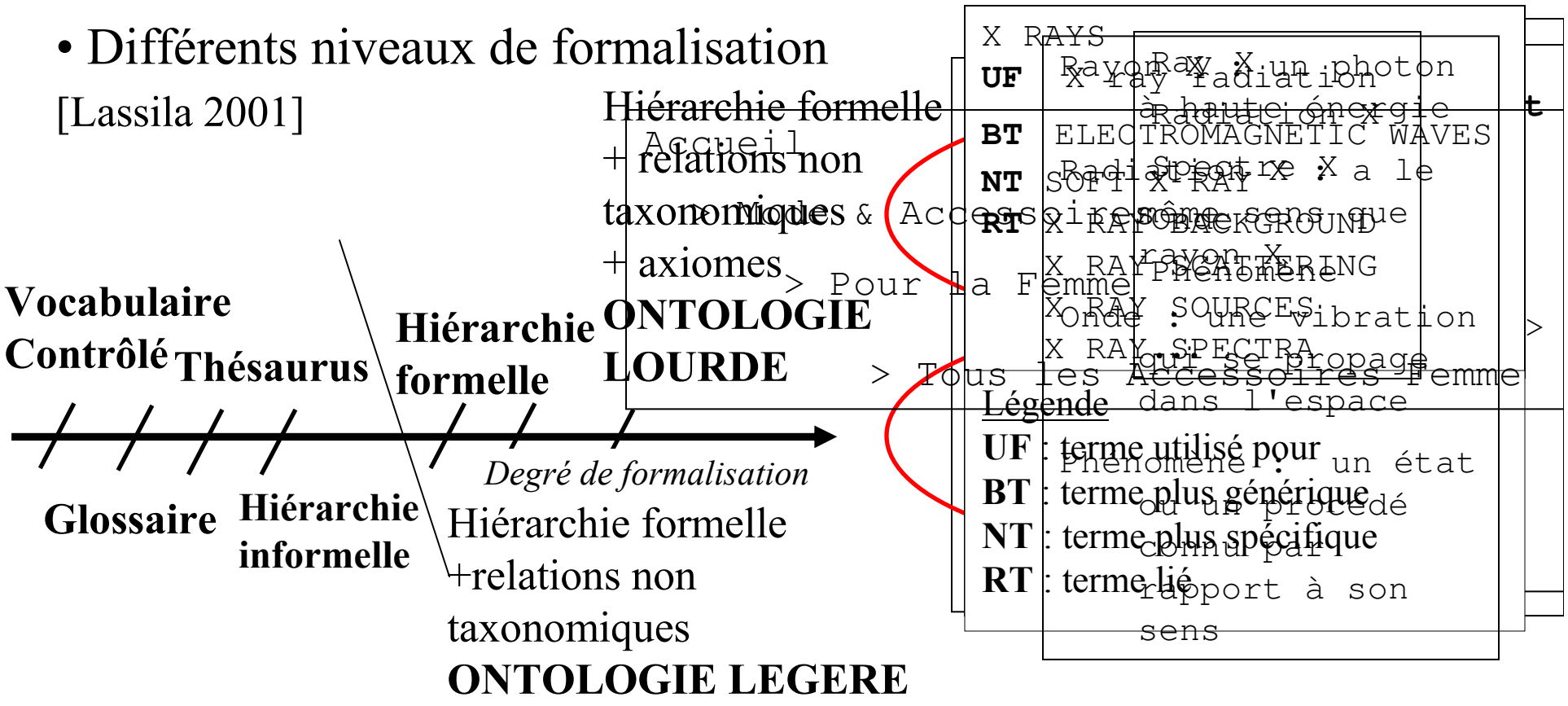
  ⇒ Exploration du corpus à partir du contexte

⇒ Connaissances à représenter et intégrer dans le système

Différentes représentations de la connaissance

- Différents niveaux de formalisation

[Lassila 2001]



Plan

- Thesaurus vs ontologie
- Présentation générale de la méthode
- Description des 3 étapes principales
- Utilisation d'ontologies en Recherche d'Information
- Conclusion

Thesaurus

- Thesauri = ressources lexicales
  - Collection de termes organisés hiérarchiquement
  - Relations entre les termes
- Thesaurus IAU utilisés par des documentalistes pour
  - Indexer manuellement des documents
  - Formuler manuellement des requêtes

Principaux inconvénients

- Créés dans les années 1990

  → Ne contiennent pas la connaissance actuelle

- Normes sur leur contenu (ISO 2788 - ANSI Z39), MAIS aucun format uniforme (ascii, html, bases de données)

  →Outils pouvant les utiliser limités (visualisation, annotation, …)

  →Utilisation limitée par les systèmes d'information (phase d'adaptation)

Principaux inconvénients

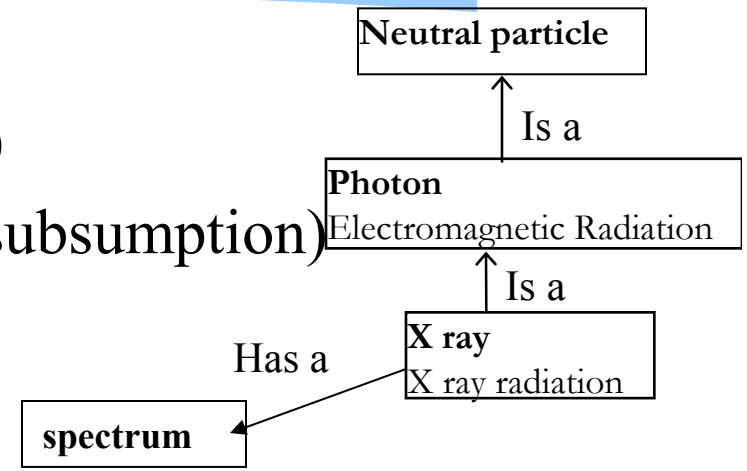
- Faible degré de formalisation dans la représentation de la connaissance
  - Pas de niveau d'abstraction (concepts, concepts génériques)
  - Pas de distinction entre un concept et sa lexicalisation
  - Relations entre termes ambigus ("est lié à")

  $\Rightarrow$représentation de domaines de connaissance en terme de terminologie et de catégories d'indexation et non pas en terme de sens

  $\rightarrow$ difficile à utiliser par des applications automatiques (eg indexation)

Potentiel des ontologies légères

- Ontologies légères =
  - concepts définis par des labels (termes)
  - concepts organisés hiérarchiquement (subsumption)
  - Relations associatives entre termes



- Référence pour la communication entre machines et humains
- Indexation sémantique de données hétérogènes

Elaboration d'ontologies

- Approches existantes
  - A partir de "rien" [Uschold 1996] [Guarino1998a] [Fernandez 1997]
  - A partir de textes [Maedche 2000] [Velardi 2001]
  - A partir de thésaurus (mais sans mise à jour de la connaissance) [Soergel 2004] [SKOS schéma w3c] [Hahn 2004]
- Notre méthode :
  - Tirer le bénéfice des termes du thésaurus
  - Extraire la connaissance implicite des thésaurus
  - Mettre à jour cette connaissance à partir de documents textuels

Méthode pour transformer un thesaurus

- Principales étapes à partir de la méthodologie Terminae [Aussenac 2000]
  - Spécification des besoin: indexation de documents (termes du domaine, concepts, relations entre concepts)
  - Choix d'un corpus de référence du domaine: A&A 1995, 2002
  - Analyse linguistique du domaine: Analyse syntaxique du corpus (Syntex) + termes et relations extraits du thesaurus
  - Normalisation (concepts et relations)
  - Formalisation : OWL-Lite [w3c]

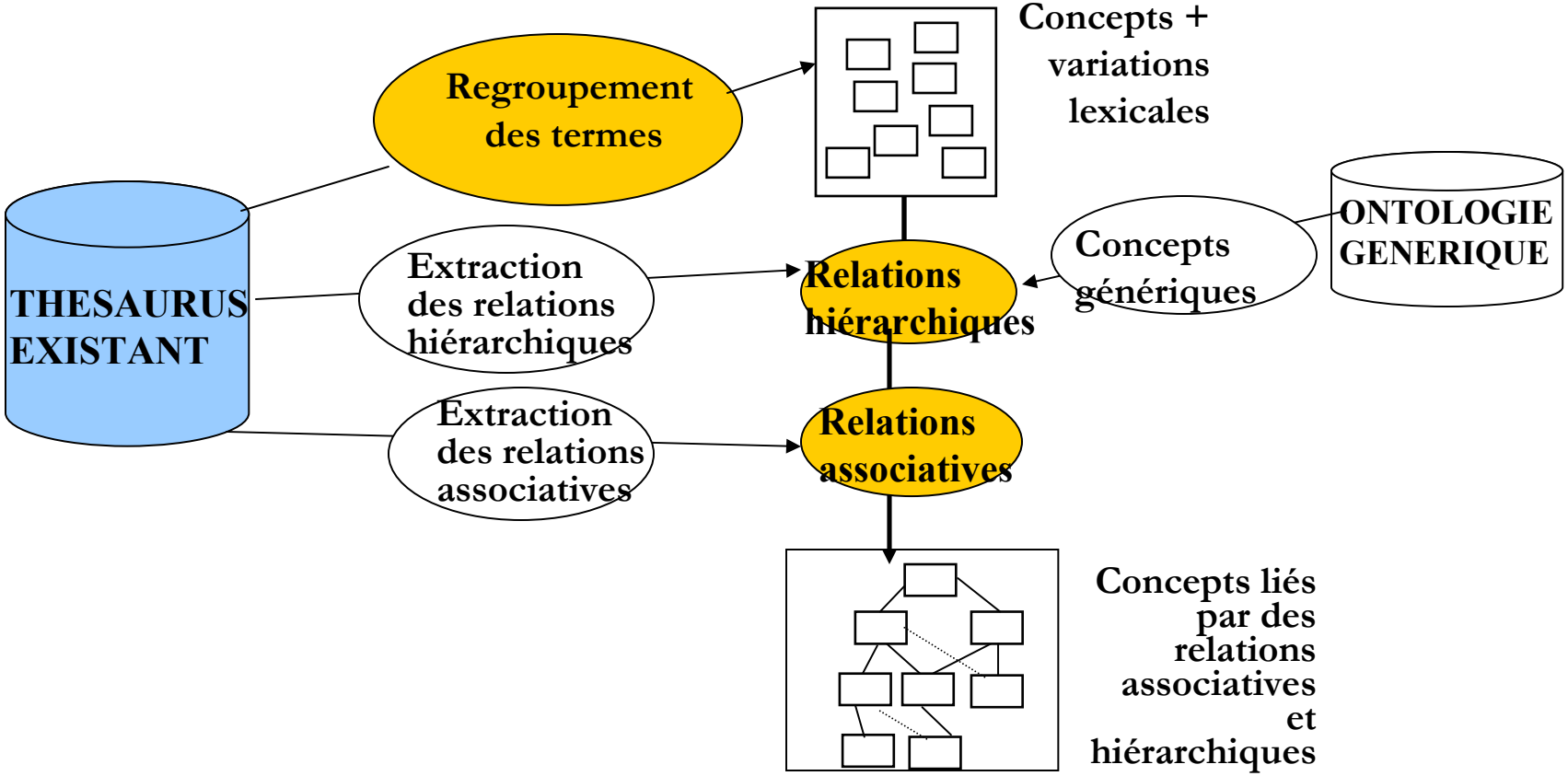
3 étapes semi-automatiques

- Extraction des concepts de l'ontologie et de sa structure (relations entre concepts) à partir du thésaurus
- Capture de nouvelles relations entre concepts non présentes dans le thésaurus
- Mise à jour de l'ontologie à partir de nouveaux termes et relations

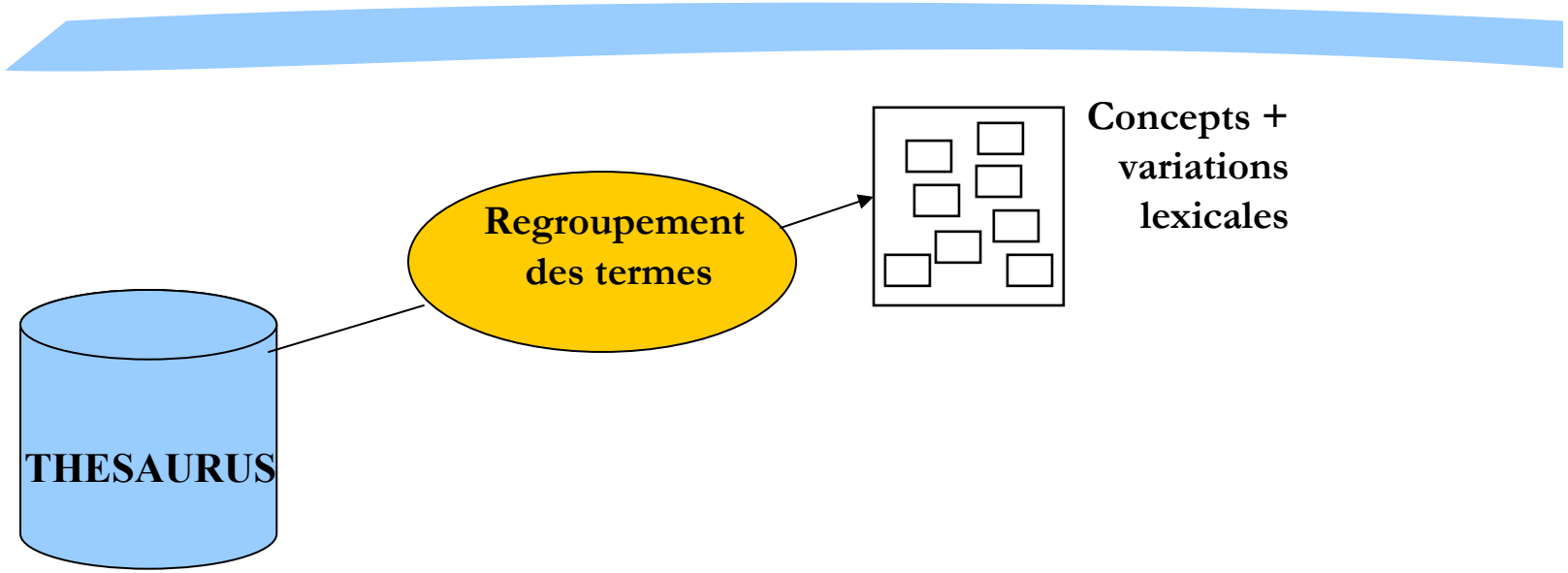
3 étapes semi-automatiques

- Extraction des concepts de l'ontologie et de sa structure (relations entre concepts) à partir du thésaurus
- Capture de nouvelles relations entre concepts non présentes dans le thésaurus
- Mise à jour de l'ontologie à partir de nouveaux termes et relations

Etape 1: extraction des concepts et relations du thésaurus



Etape 1: Extraction des concepts



Extraction des concepts

- Conceptualisation du lexique
  - Relations du thesaurus
    - Term1 **USE** term2
    - Term3 **USED FOR** term2

  →Regroupements à partir de la fermeture transitive de ces relations

  - *Exemple :*

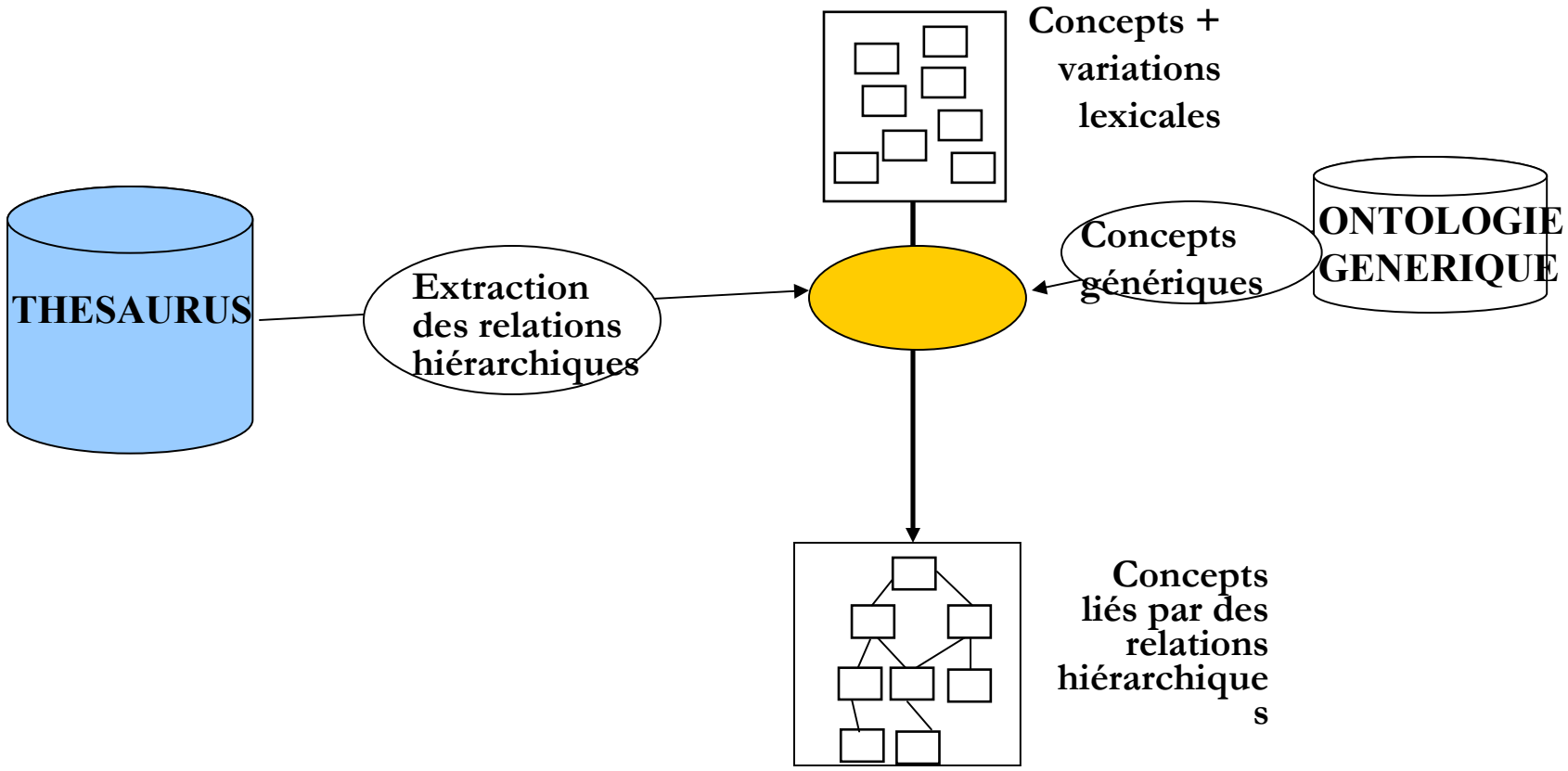
    ELLIPSOIDAL VARIABLE STARS **USE** photometric binary stars

    ellipsoidal binary stars **USED FOR** ELLIPSOIDAL VARIABLE STARS

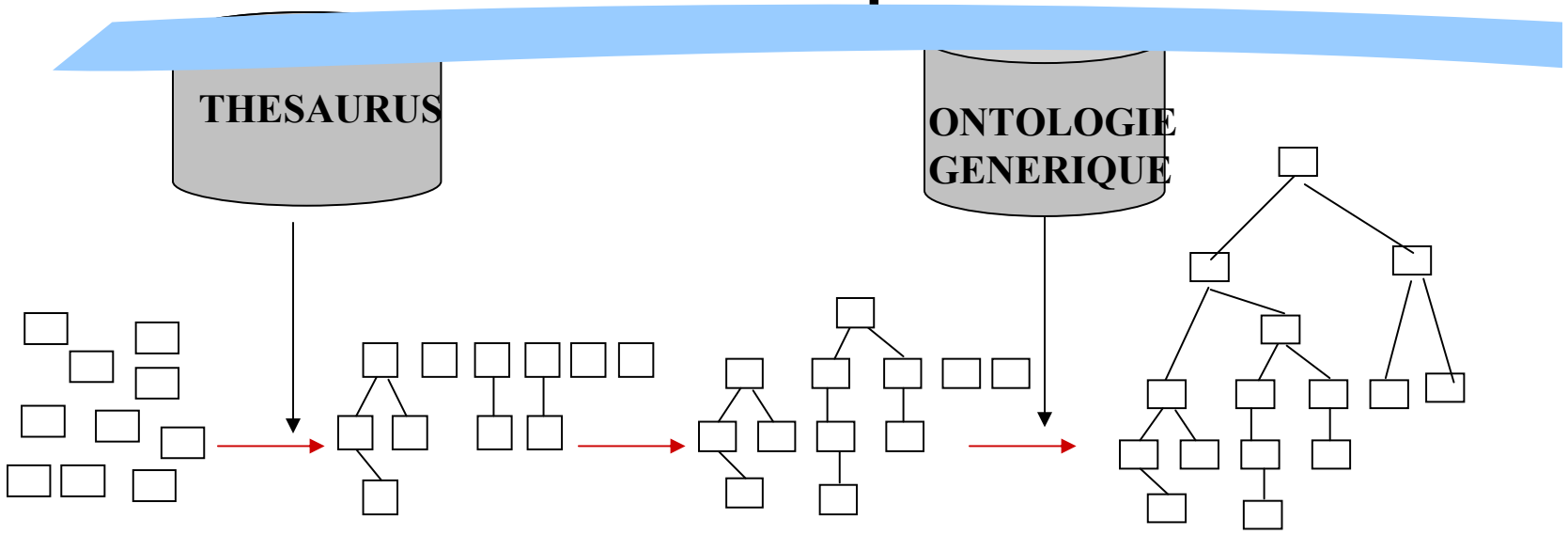
  ⇒Concept : ELLIPSOIDAL VARIABLE STARS

    labels : photometric binary stars, ellipsoidal binary stars, ellipsoidal variable stars

Extraction des relations hiérarchiques



Extraction des relations hiérarchiques



Extraction des relations taxonomiques explicites

Nouveaux niveaux obtenus à partir de la tête et de l'expansion des labels

Nouveaux niveaux obtenus à partir des concepts génériques

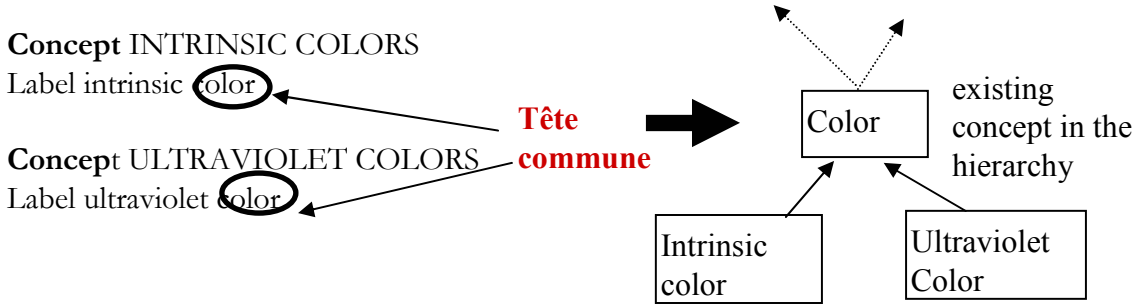
Extraction des relations explicitées dans le thésaurus

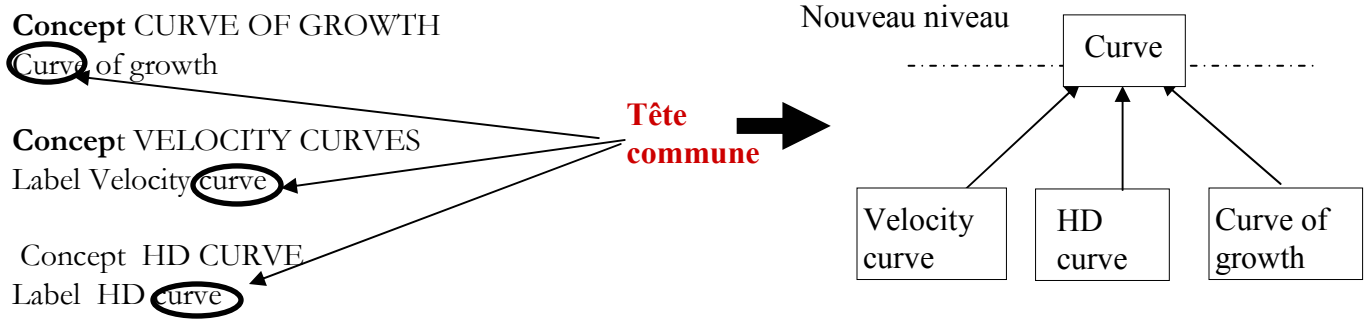
- Relations du thesaurus
  - Term1 **Broader Term** Term2
  - Term3 **Narrower Term** Term4

  →Création de liens hiérarchiques entre concepts dont les labels sont liés par ces relations

Nouveaux niveaux hiérarchiques

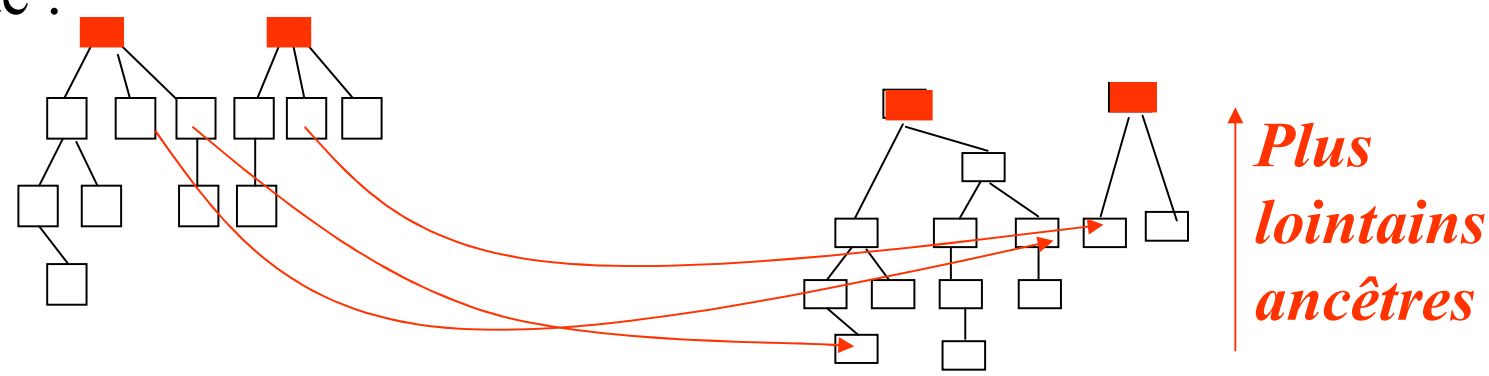
- A partir des labels des concepts





Extraction des concepts génériques

- Concepts génériques = structuration de l'ontologie
- Extraction des concepts génériques à partir d'ontologies génériques (WordNet)
- Procédé :



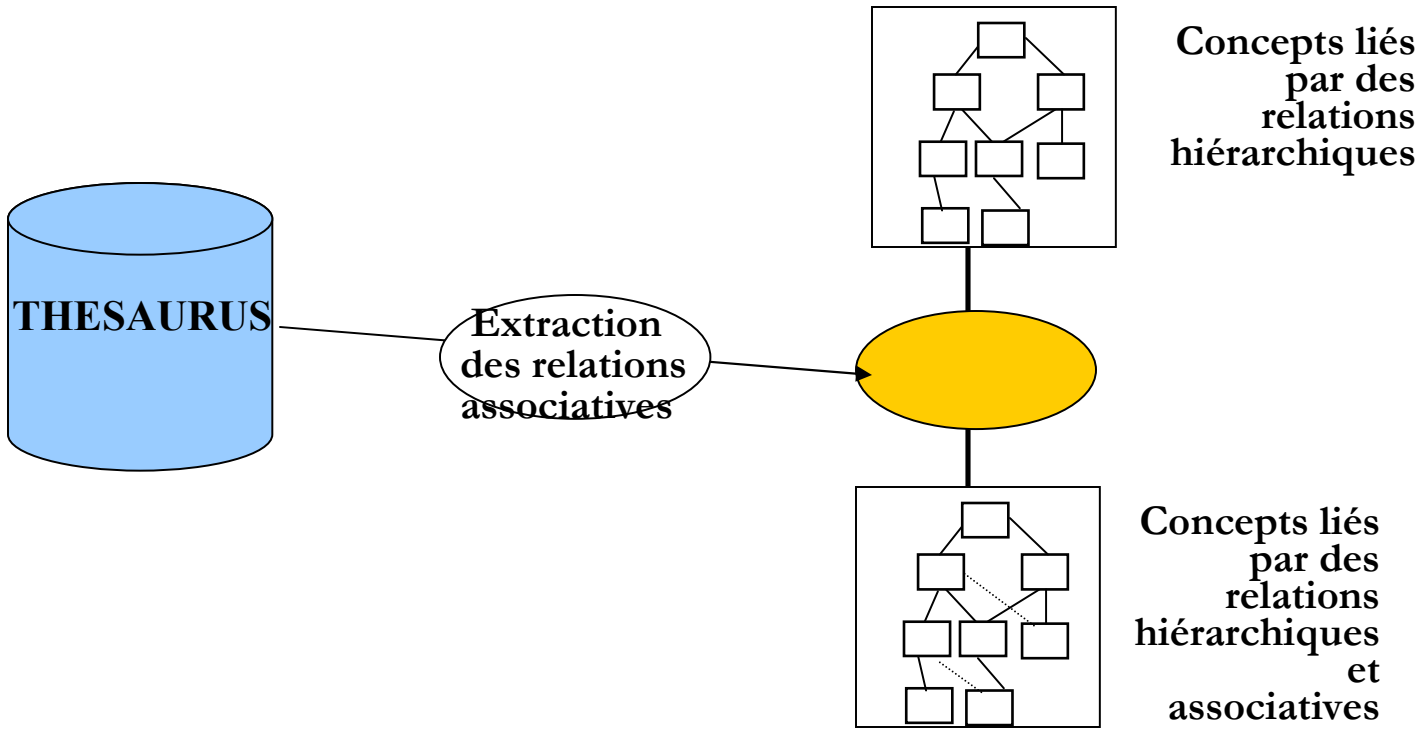
Ontologie extraite du thésaurus

Mise en correspondance

Ontologie générique

Exemple : *phenomenon, natural object*

Etape 1: extraction des relations associatives



Extraction des relations associatives

- Relations explicitées dans le thésaurus
  - Term1 **RELATED TO** term2

  →Extraction de ces relations entre concepts dont les labels sont liés dans le thésaurus

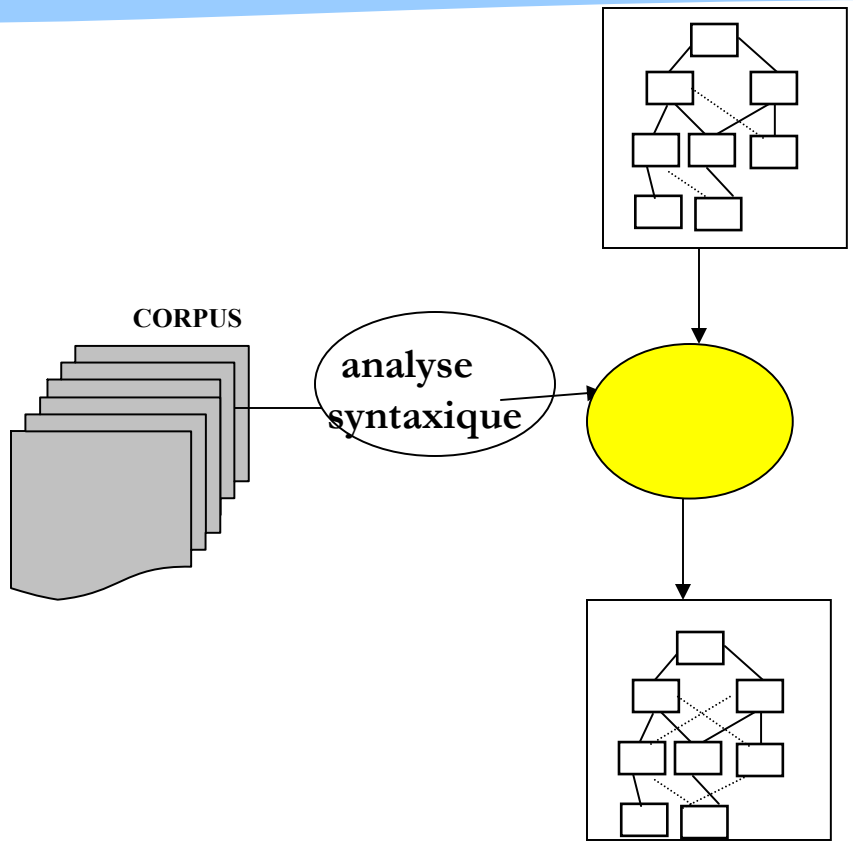
  →relations vagues et ambigus
- Désambiguïsation de ces relations à partir des concepts génériques

*Exemple : coronagraph **RT** solar corona →*
*coronagraph **OBSERVES** solar corona*

3 étapes semi-automatiques

- Extraction des concepts de l'ontologie et de sa structure (relations entre concepts) à partir du thésaurus
- Capture de nouvelles relations entre concepts non présentes dans le thésaurus
- Mise à jour de l'ontologie à partir de nouveaux termes et relations

Etape 2 : Capture de nouvelles relations



Ontologie à l'étape 1

Ontologie enrichie

Capture des relations

- Analyse syntaxique du contexte des labels de concepts dans le corpus de référence
- Si un label apparaît dans le contexte d'un label d'un autre concept

⇒Création d'une nouvelle relation entre les deux concepts

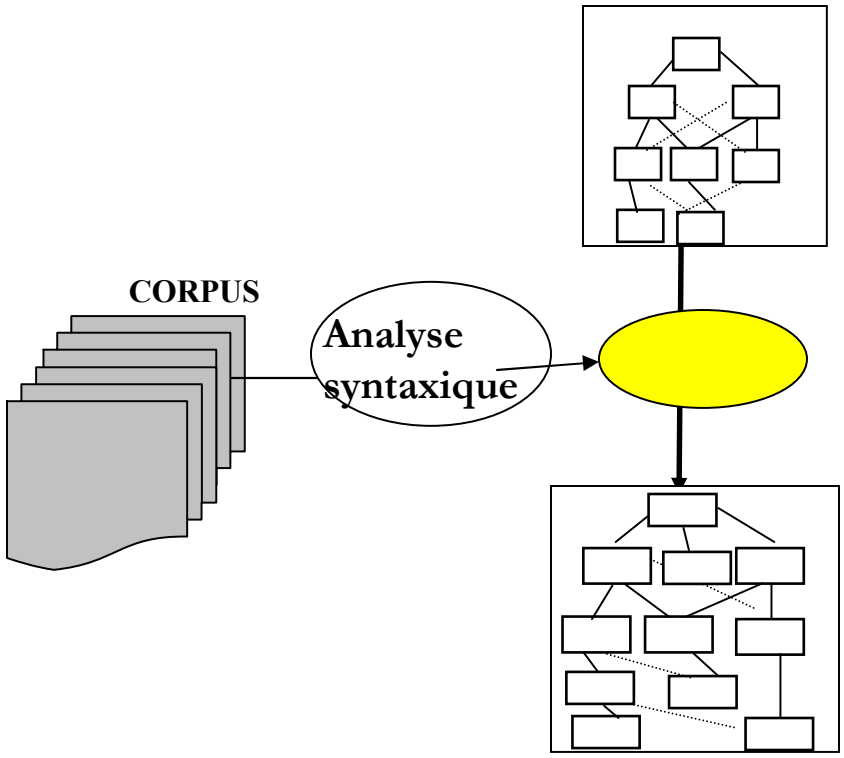
Exemple : « intensity » trouvé dans le contexte de « radial velocity » (the intensity of radial velocity)

⇒ Création de la relation « is a property of » entre les concepts « radial velocity » et « intensity »

3 étapes semi-automatiques

- Extraction des concepts de l'ontologie et de sa structure (relations entre concepts) à partir du thésaurus
- Capture de nouvelles relations entre concepts non présentes dans le thésaurus
- Mise à jour de l'ontologie à partir de nouveaux termes et relations

Etape 3 : mise à jour de l'ontologie



Ontologie de l'étape 2

Ontologie enrichie

Extraction de nouveaux termes

- Deux fonctions d'extraction

Termes généraux	Termes spécifiques
high resolution	Yarkovsky force
globular cluster	Relativistic gravity
binary system	Suprathermal electron
soft X ray	Halpha knot
orbital period	Penumbral wave
stellar population	Mean free path
power law	Integral magnitude
absorptance line	Mixing layer
line emission	stellar population
active region	

Intégration des termes dans l'ontologie

- 2 approches :
  - Nouveaux sous-concepts de concepts existants
  - Nouvelles relations entre concepts

Intégration des termes dans l'ontologie

- Nouveaux concepts sous-concepts de concepts existants

  → tête d'un terme= label d'un concept existant

Exemple : nouveau terme "soft X Ray"

concept existant "X Ray"

⇒ creation du concept "soft X Ray" sous-concept de "X Ray"

Intégration des termes dans l'ontologie

- Nouvelles relations associatives entre concepts existants

  → tête et expansion d'un terme = labels de concepts existants

Exemple :

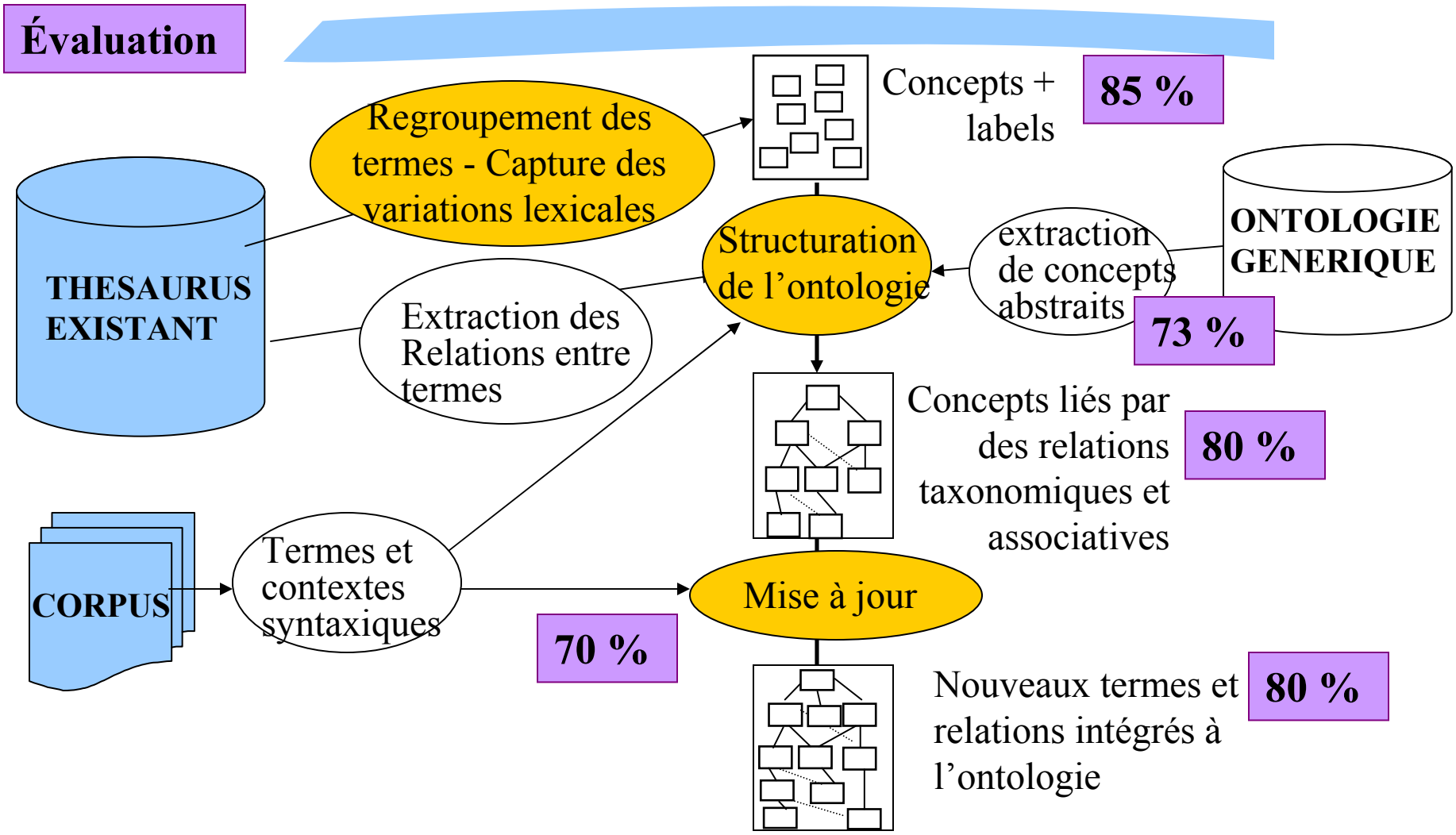
Nouveau terme : star mass

Concepts existants :

- star (natural_object)
- mass (property)

⇒ creation de la relation « has property » entre les concepts star et mass

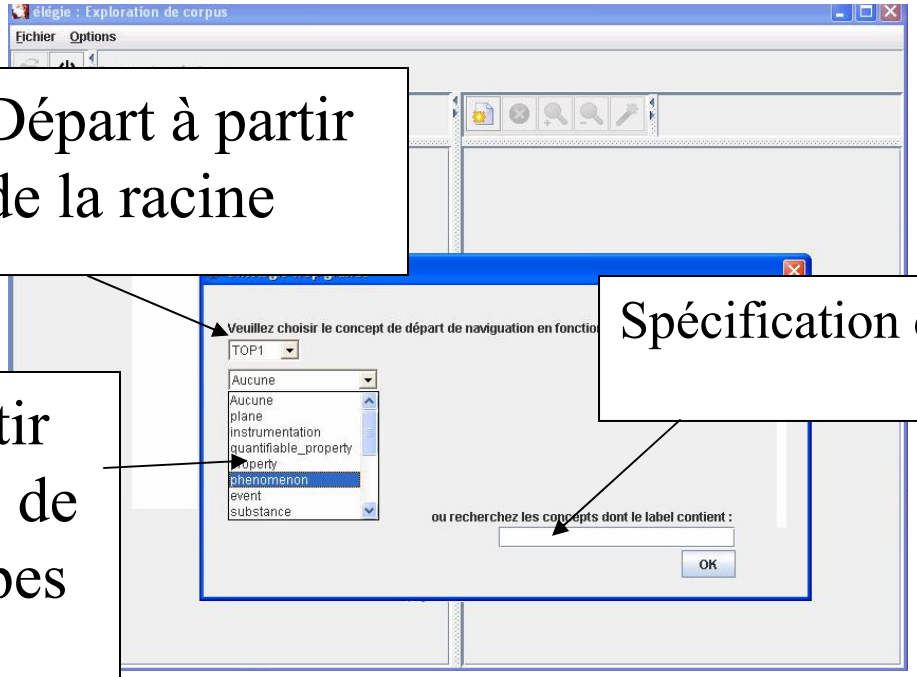
Evaluation


Évaluation
Regroupement des termes - Capture des variations lexicales
Concepts + labels
85 %
THESAURUS EXISTANT
Structuration de l'ontologie
extraction de concepts abstraits
ONTOLOGIE GENERIQUE
73 %
Extraction des Relations entre termes
Concepts liés par des relations taxonomiques et associatives
80 %
CORPUS
Termes et contextes syntaxiques
Mise à jour
70 %
Nouveaux termes et relations intégrés à l'ontologie
80 %

- Résultats :
  - Ontologie de l'astronomie évaluée à partir de 10% du thésaurus
  - Listes des nouveaux termes
  - Interface permettant de naviguer dans l'ontologie en OWL

Exploration d'ontologies volumineuses

Départ de la navigation



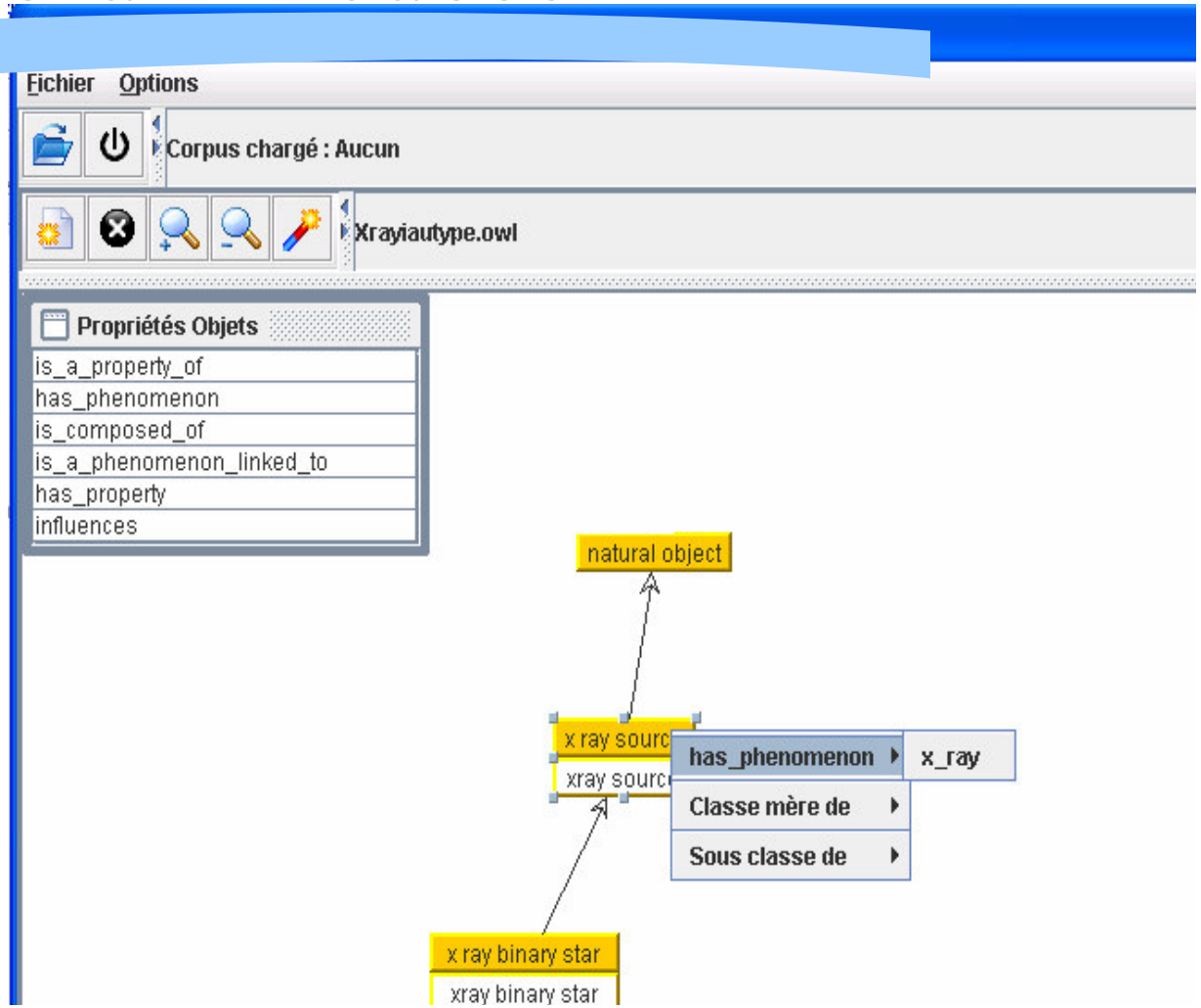
Départ à partir de la racine

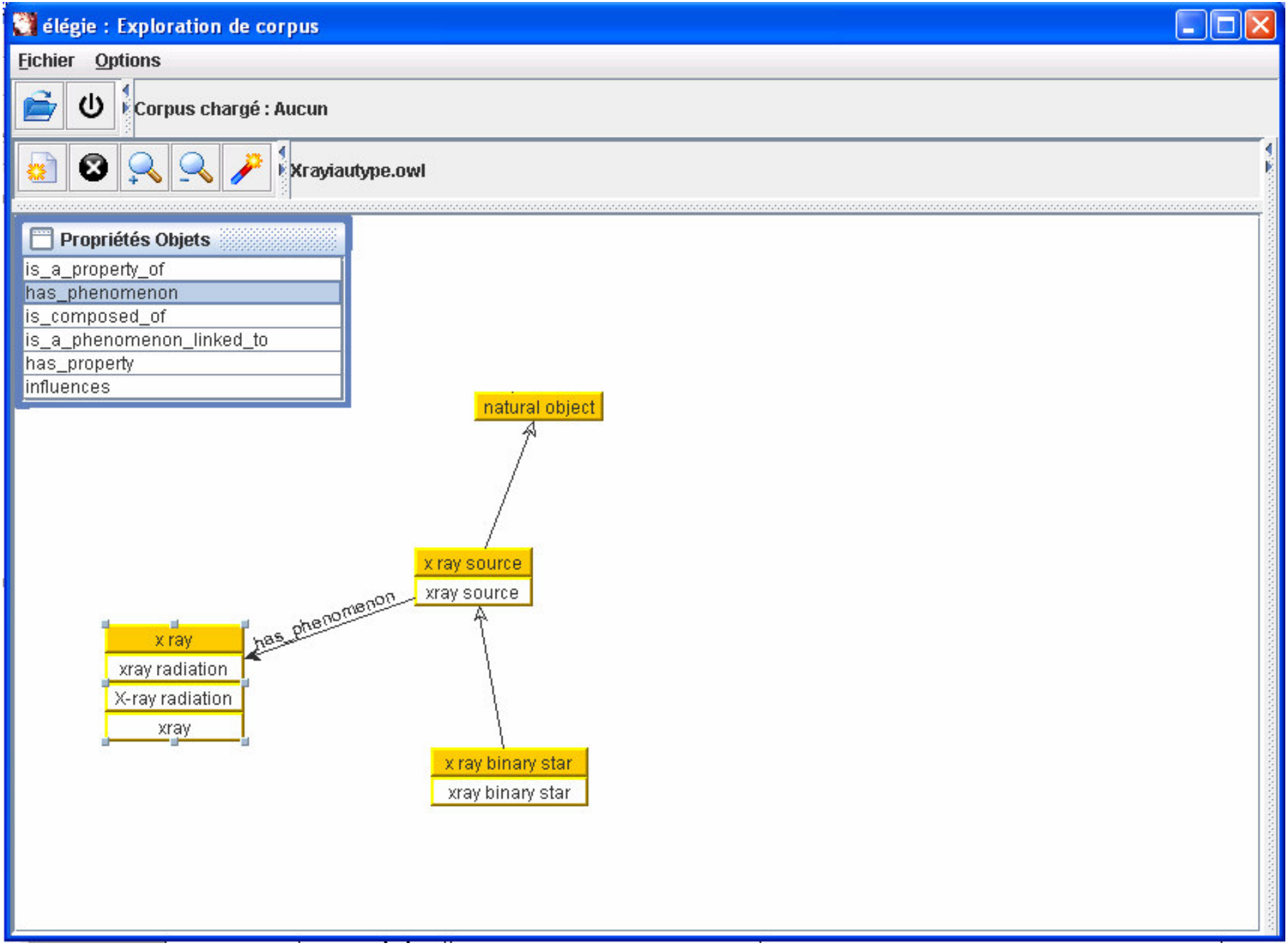
Spécification d'un label

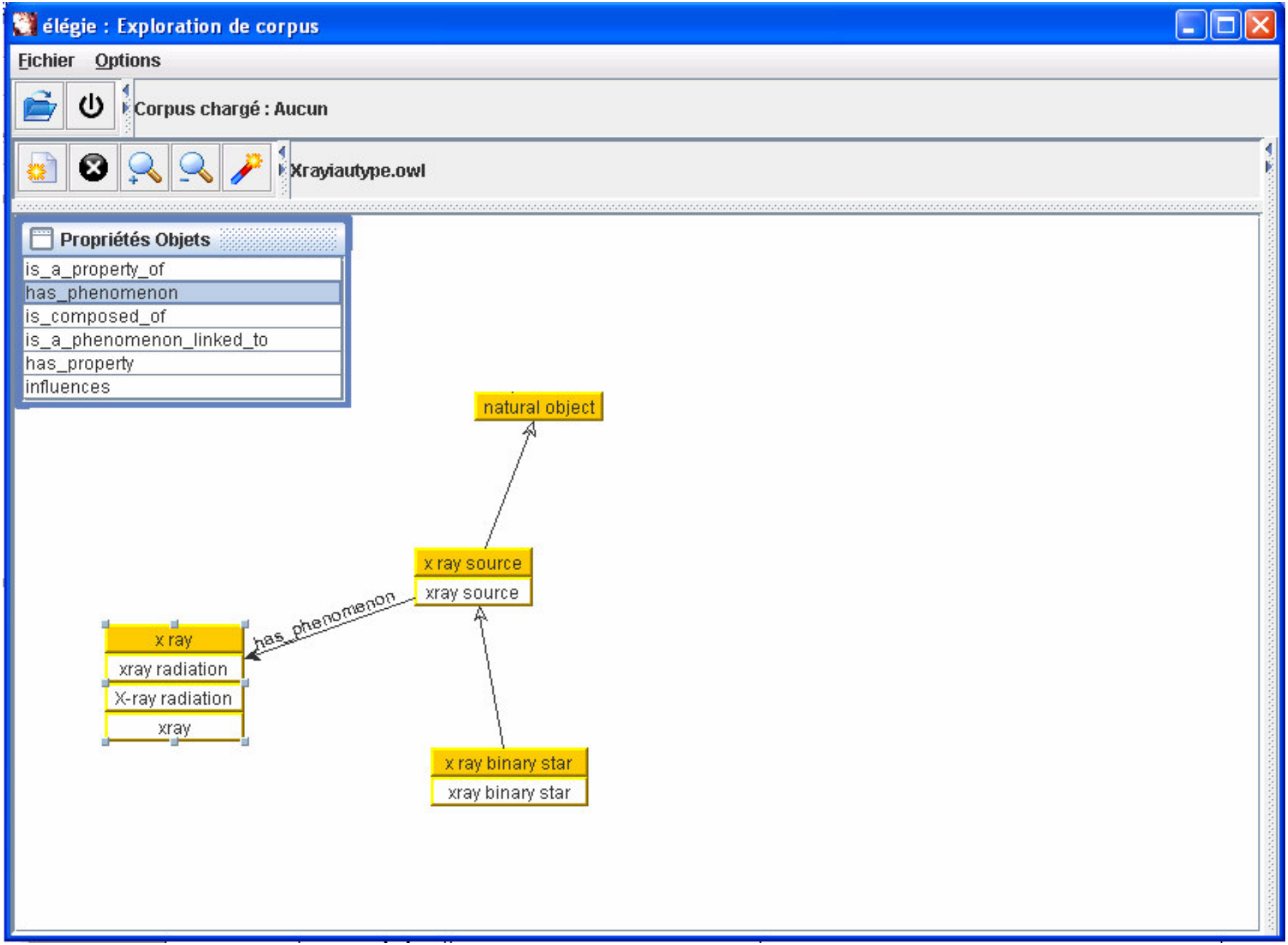
Départ à partir des concepts de niveau 1 (types abstraits)

Exploration d'ontologies volumineuses

- Visualisation du contexte d'un concept



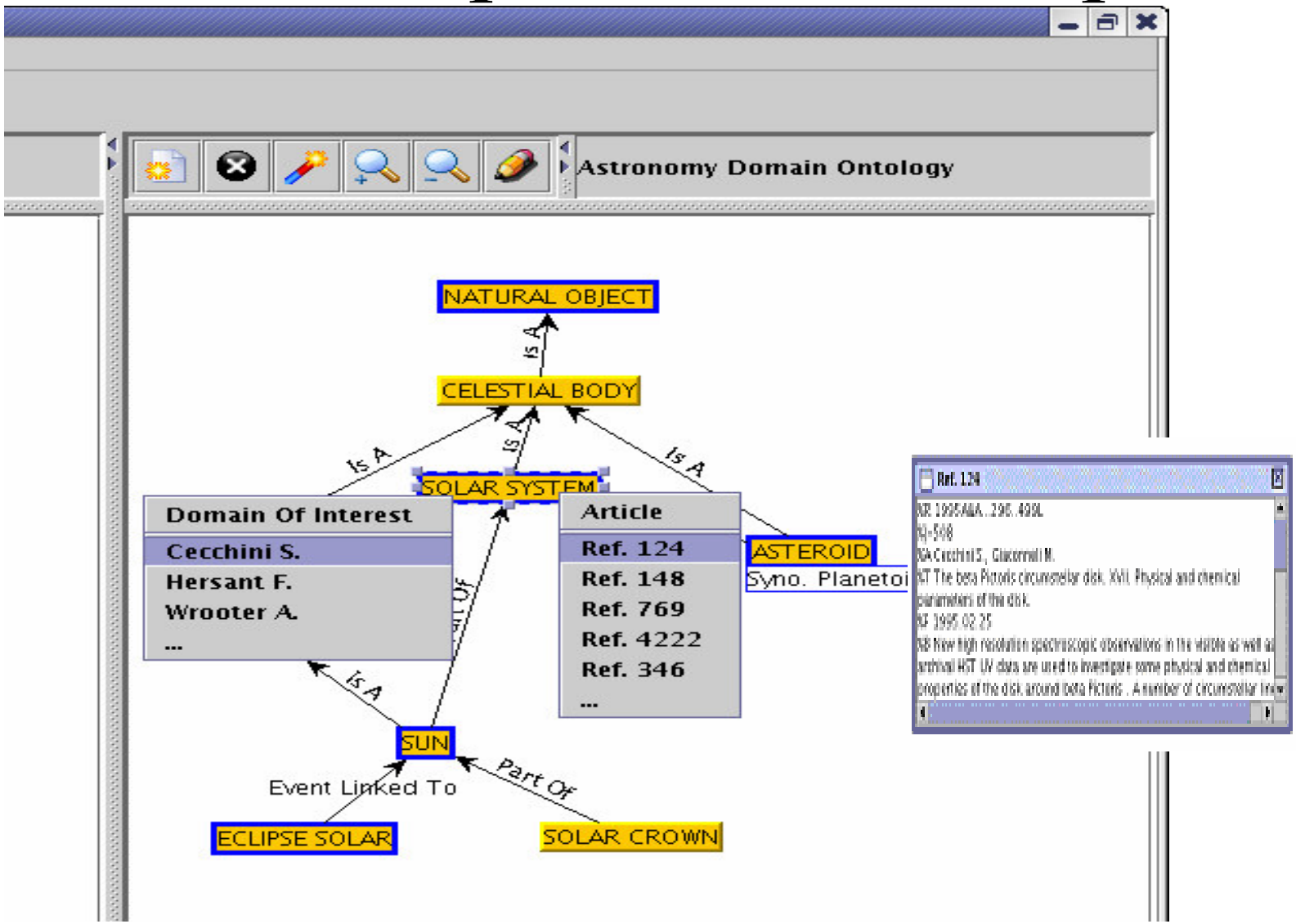

élégie : Exploration de corpus
Fichier
Options
Corpus chargé : Aucun
Xrayiautype.owl
Propriétés Objets
is_a_property_of
has_phenomenon
is_composed_of
is_a_phenomenon_linked_to
has_property
influences
natural object
x ray source
xray source
has_phenomenon
x ray
xray radiation
X-ray radiation
xray
x ray binary star
xray binary star


élégie : Exploration de corpus
Fichier
Options
Corpus chargé : Aucun
Xrayiautype.owl
Propriétés Objets
is_a_property_of
has_phenomenon
is_composed_of
is_a_phenomenon_linked_to
has_property
influences
natural object
x ray source
xray source
has_phenomenon
x ray
xray radiation
X-ray radiation
xray
x ray binary star
xray binary star

Indexation sémantique à partir de l'ontologie de l'astronomie

- Indexation sémantique [Haav 2001]
  - Non plus uniquement à partir de considérations statistiques
  - A partir des objets du monde référencés
- Descripteurs des granules documentaires choisis à partir des concepts de l'ontologie de l'astronomie
  - Détection des concepts dans les granules
    - Identification des labels référençant des concepts
    - Désambiguïsation des labels
  - Pondération des concepts à partir de leur représentativité des granules
    - Poids statistique
    - Poids sémantique : liens avec les autres concepts

Exploration du corpus



Conclusion

- Méthode pour le transformation d'un thésaurus en une ontologie légère
- Évaluation dans le cadre de l'astronomie résultat efficace (+80% des propositions validées)
- Utilisation de l'ontologie créée pour l'indexation et l'exploration de corpus
- Application aux types d'objets

Désambiguïsation des labels retrouvés dans les documents

« ***Polarization*** varies noticeably with emergent ***photon*** energy below 40keV, being up to 30% and down to -10% for different angles of view; these variations cover the range of observed ***magnitudes***.»

Label ambigu :

polarization concerning wave

polarization concerning charge separation

Labels non ambigus

→Nombre d'arcs dans l'ontologie séparant les deux concepts candidats aux concepts identifiés dans la phrase

→Choix du concept ayant le nombre d'arcs minimum

Désambiguïsation des relations RT