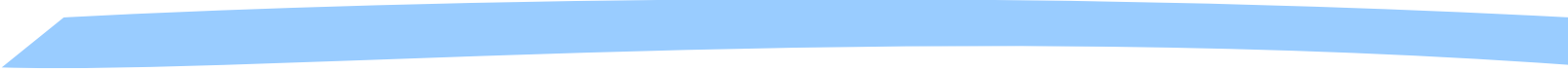


Building and updating ontologies from thesauri



Application to the
astronomical field

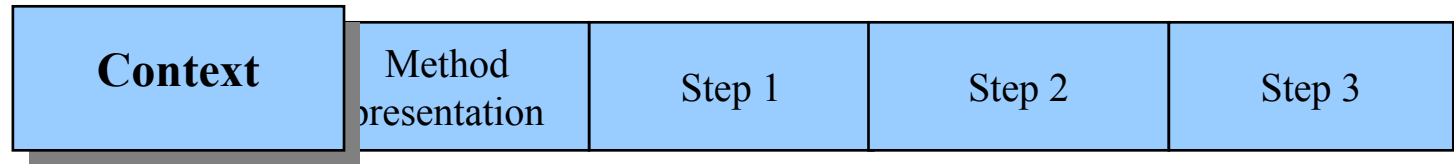
N. Hernandez, J. Mothe

IRIT

Overview



- Context (thesaurus /ontology)
- General presentation of the method
- Description of the three steps of the method
- Conclusion



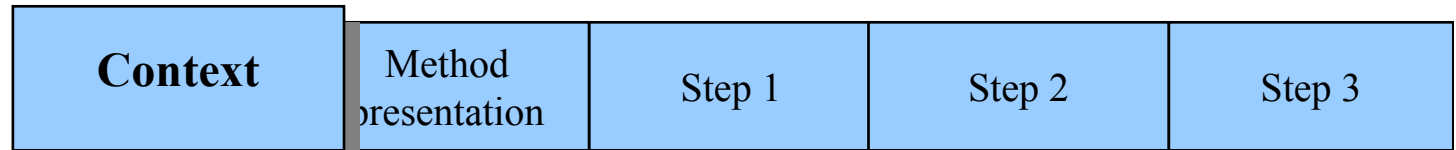
General context

- Indexing system
- Exploration system

→ Domain knowledge

- Meta-data
- Document content

→ Model based on ontologies



Thesauri

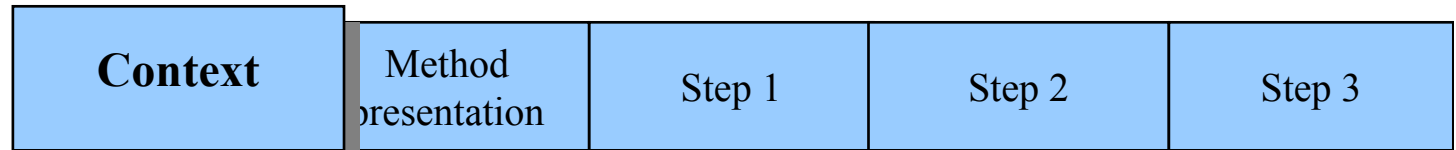
- Thesauri = lexical resources
 - Collection of terms organised hierarchically
 - Relations between terms
- Many existing thesauri developed in order to help librarians
 - Manually indexing document resources
 - Manually formulating queries

- Astronomical thesaurus
IAU by the International
Astronomical Union
in created in 1995

Sample of IAU

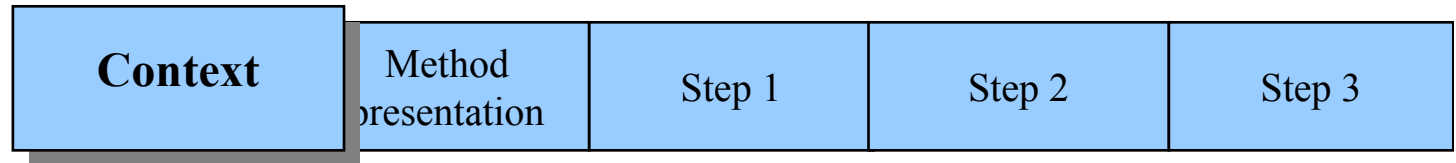
```
Ae STARS
  UF herbig stars
  BT A STARS
    EARLY TYPE EMISSION STARS
  RT Be STARS

AERONOMY
  RT ATMOSPHERES
  ...
```



Main drawbacks

- Built 10 years ago
 - Do not contain recent knowledge
- Norms on their content (ISO 2788 - ANSI Z39), BUT no uniform format (ascii, html, data-base)
 - Limited tools that can use them (visualisation, annotation, ...)
 - Limited use in Information Retrieval Systems (adaptation phase)



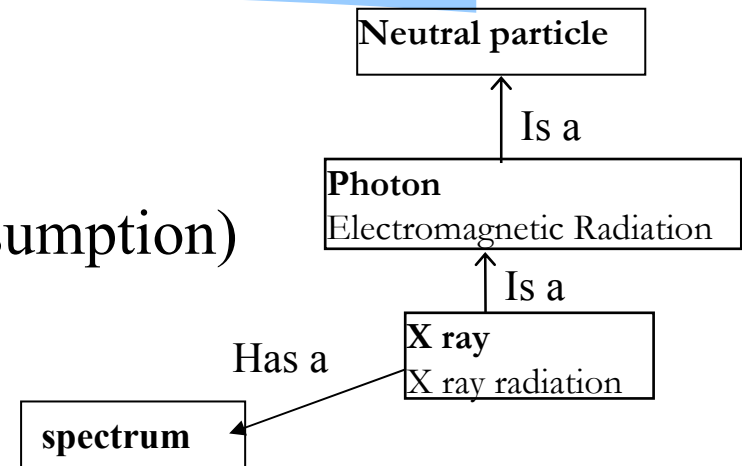
Main drawbacks

- Low degree of formalisation for knowledge representation
 - No conceptual abstraction level
 - No distinction between a concept and its lexicalisation
 - Ambiguous relations between terms (“is related to”)
 - ⇒ Domain representation in terms of terminology and indexing categories and not in terms of meaning
 - difficult to use in automatic application (eg indexing)

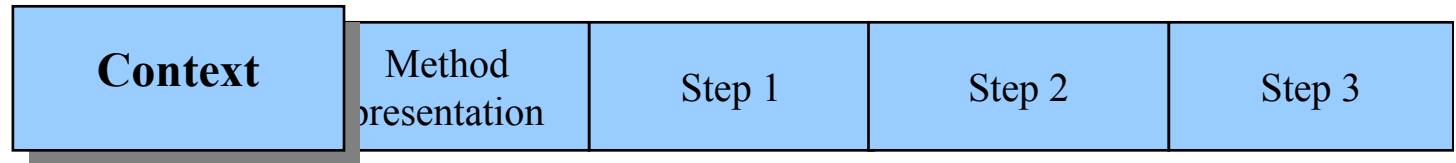


Potential of lightweight ontologies

- Lightweight ontology =
 - concepts defined by labels (terms)
 - concepts organised hierarchically (subsumption)
 - associative relations between concepts



- Reference for communication between machines and between machines and humans
- Semantic indexing for heterogeneous data



Ontology elaboration

- Existing approaches for ontology elaboration
 - From scratch [Uschold 1996] [Guarino1998a] [Fernandez 1997]
 - From texts [Maedche 2000] [Velardi 2001]
 - From thesauri (but no knowledge update) [Soergel 2004] [SKOS schéma w3c] [Hahn 2004]
- Our method :
 - Take advantages of terms stated in thesauri
 - Extract implicit knowledge from thesauri
 - Update ontology knowledge from text analysis



Method for updating thesauri

- Main stages according to the methodology

Terminae [Aussenac 2000]

- Needs specification: indexing language for IR (domain terms, concepts, relations between concepts)
- Reference domain corpus choice: A&A 1995, 2002

– Linguistic analysis of domain: Syntactical analysis of corpus (Syntex) + terms and relations extracted from thesaurus

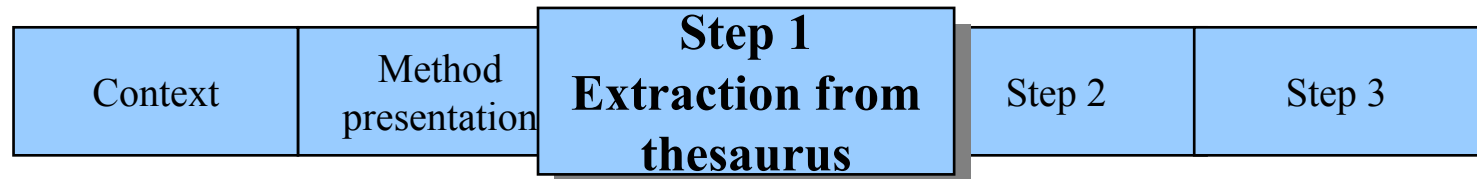
– Normalisation (concepts and relations)

- Formalisation : OWL-Lite [w3c] « **magnetic connection** between **black holes** and **disks** are observed »



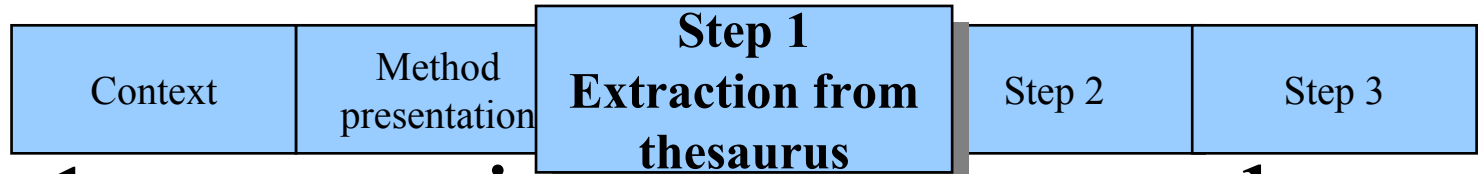
Method for updating thesauri

- 3 semi-automatic steps :
 - Extraction of ontology concepts and structure (relations between concepts) from thesaurus
 - Capture of new relations between concepts not stated in the thesaurus (from texts)
 - Ontology update with new terms and concepts

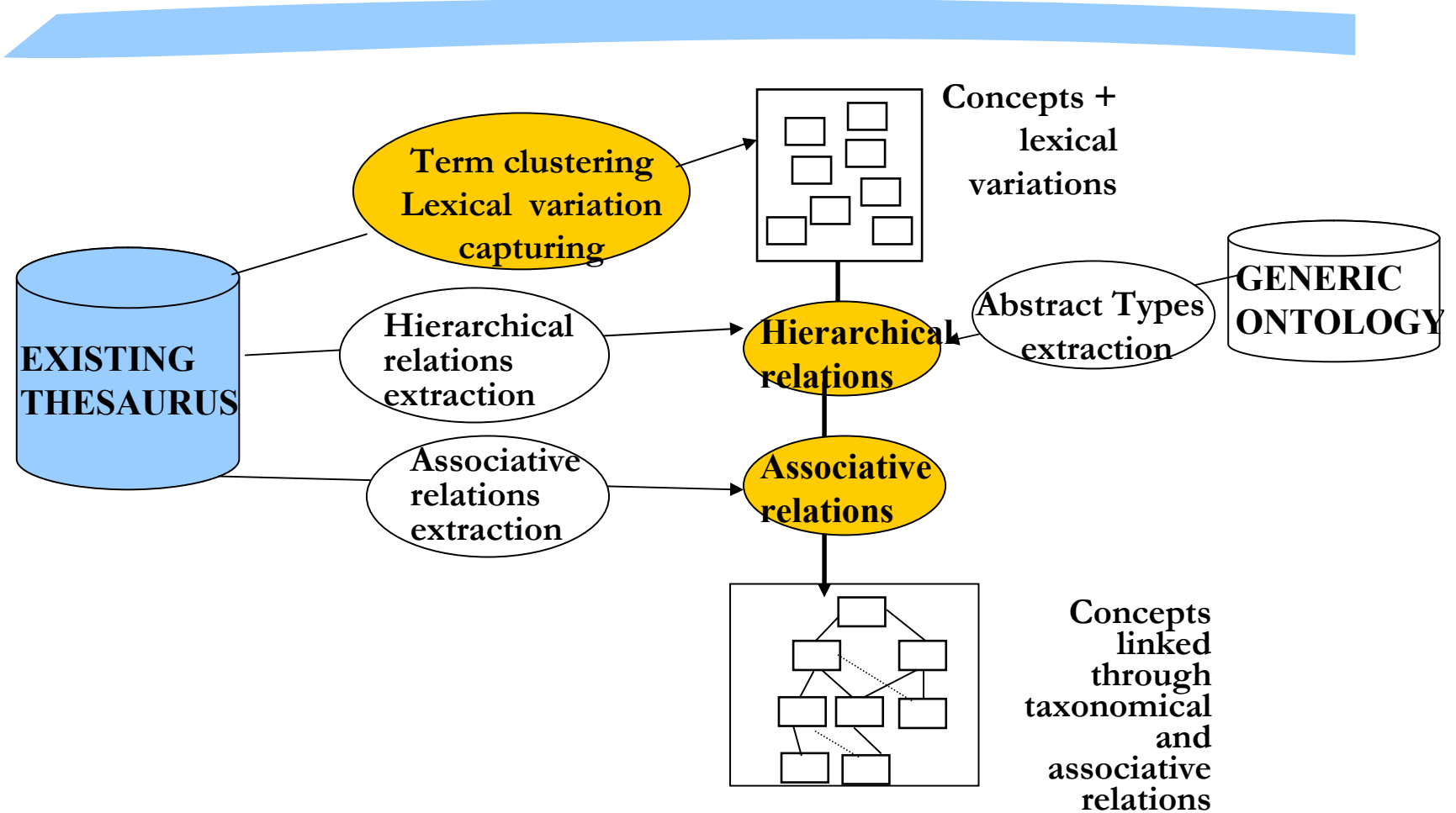


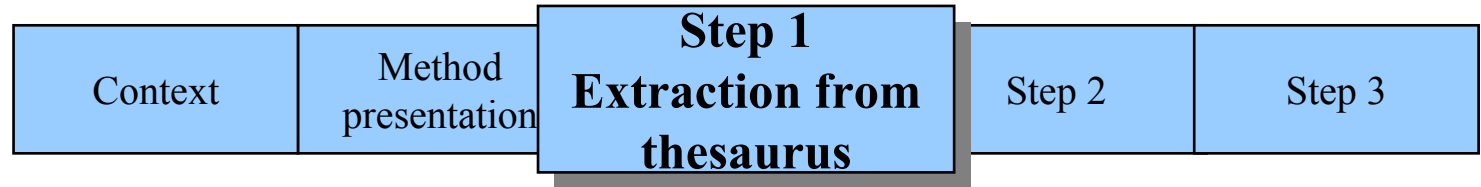
Method for updating thesauri

- 3 semi-automatic steps :
 - Extraction of ontology concepts and structure (relations between concepts) from thesaurus
 - Capture of new relations between concepts not stated in the thesaurus (from texts)
 - Ontology update with new terms and concepts

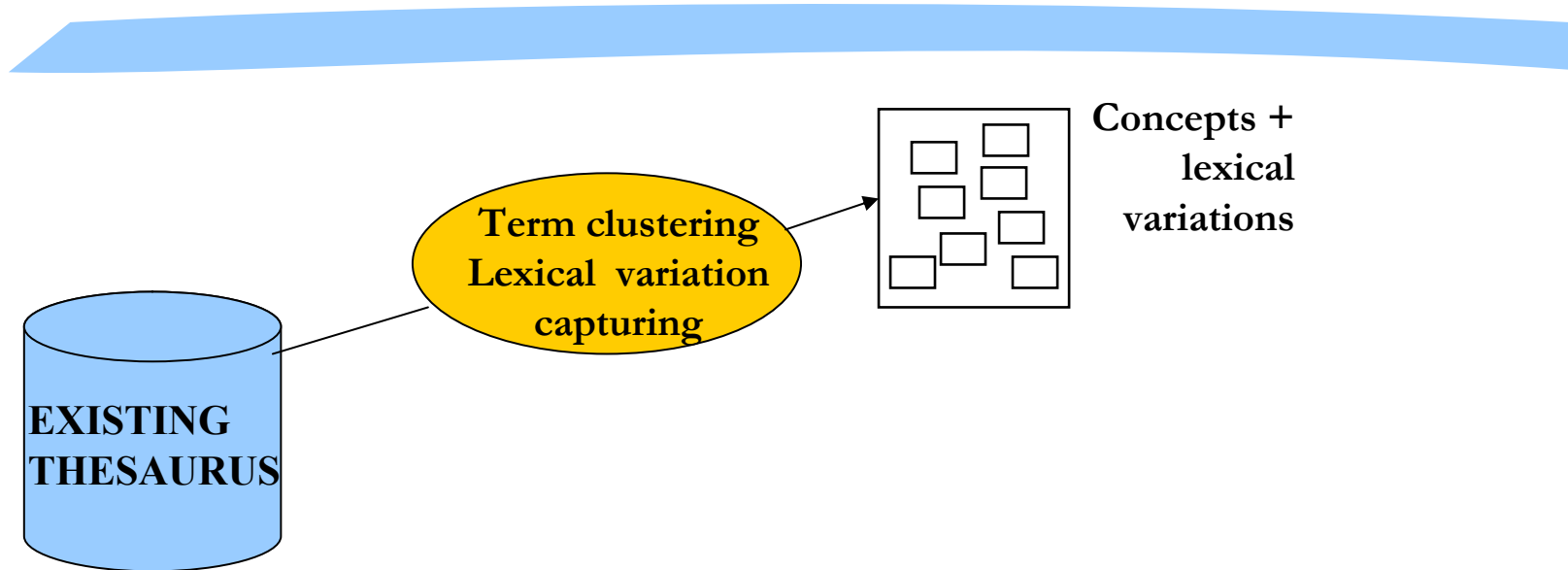


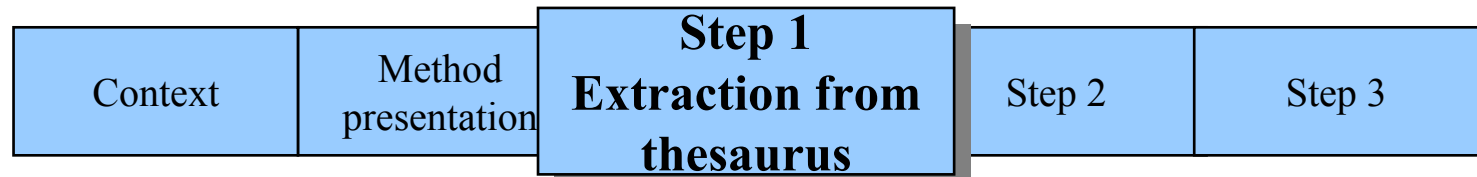
Step 1: extracting concepts and structure from thesaurus





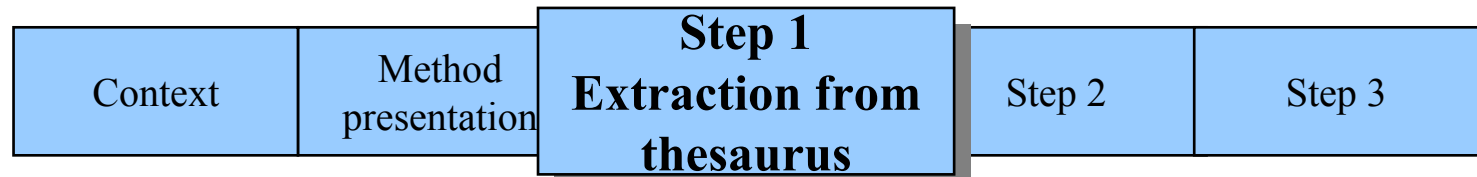
Step 1: Concept extraction





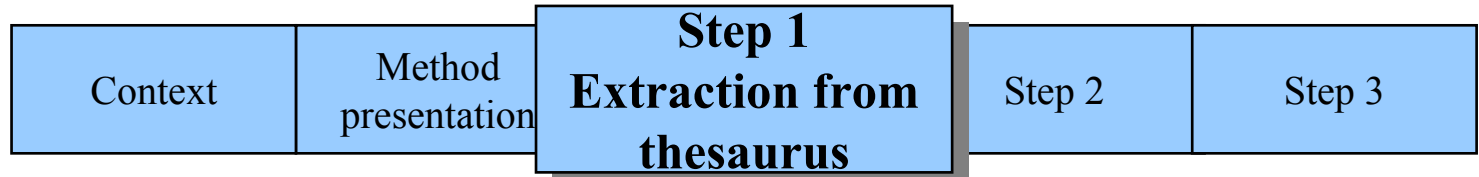
Concept extraction

- Lexis conceptualisation
 - Thesaurus relations
 - Term1 **USE** term2
 - Term3 **USED FOR** term2
 - Clustering according to the transitive closure of these relations
 - *Example* :
 - 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

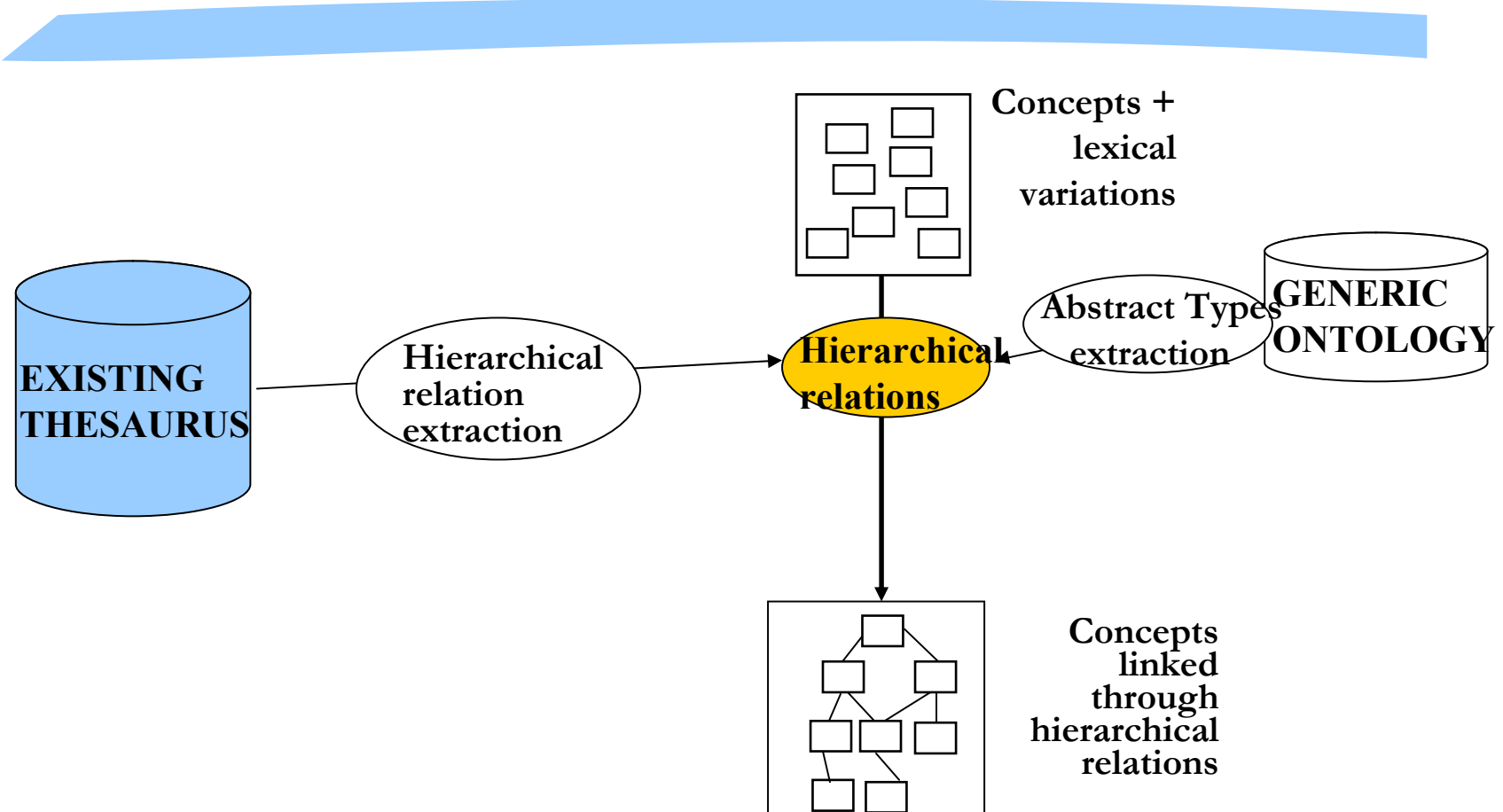


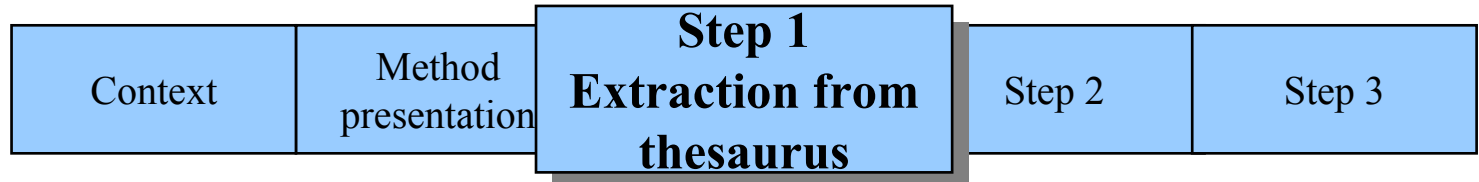
Concept extraction

- Extraction of lexical variants
 - Thesaurus terms in plural
 - extraction of singular form
 - Case verification
 - Example : B dwarf stars, ab variable stars, E layer
e process

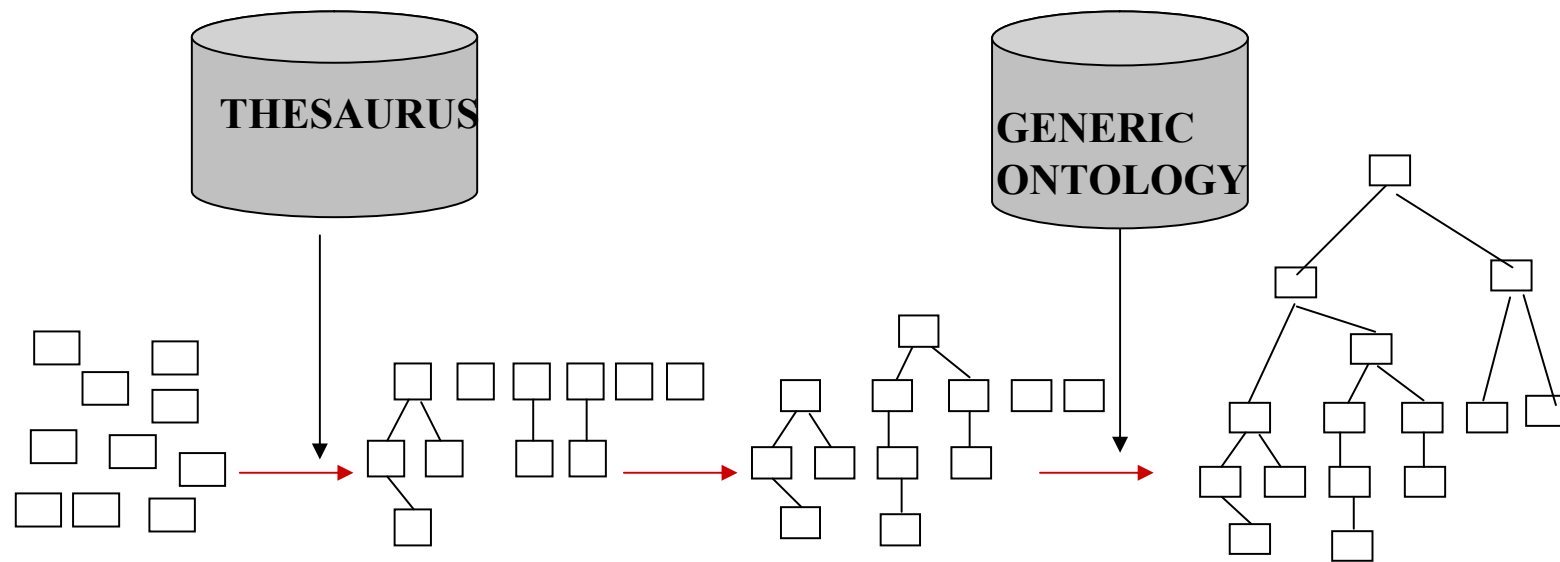


Step 1: Extraction of hierarchical relations

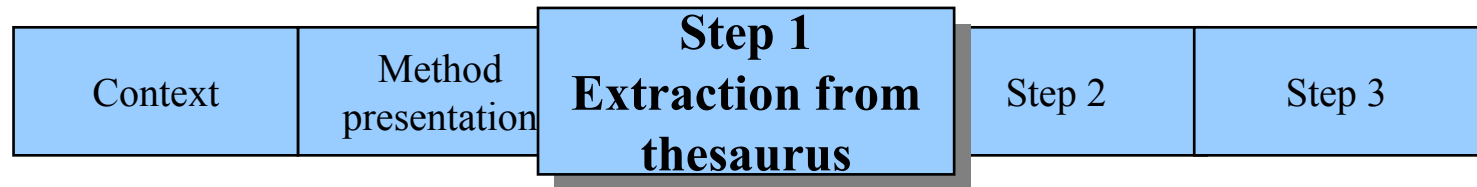




Hierarchical relation extraction

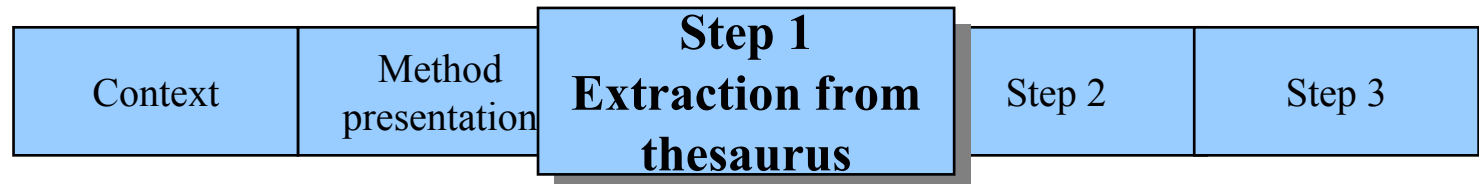


<p>Extraction of explicit taxonomical relations of the thesaurus</p>	<p>New hierarchical level from head and expansion of labels</p>	<p>New hierarchical level obtained from abstract types</p>
---	--	---



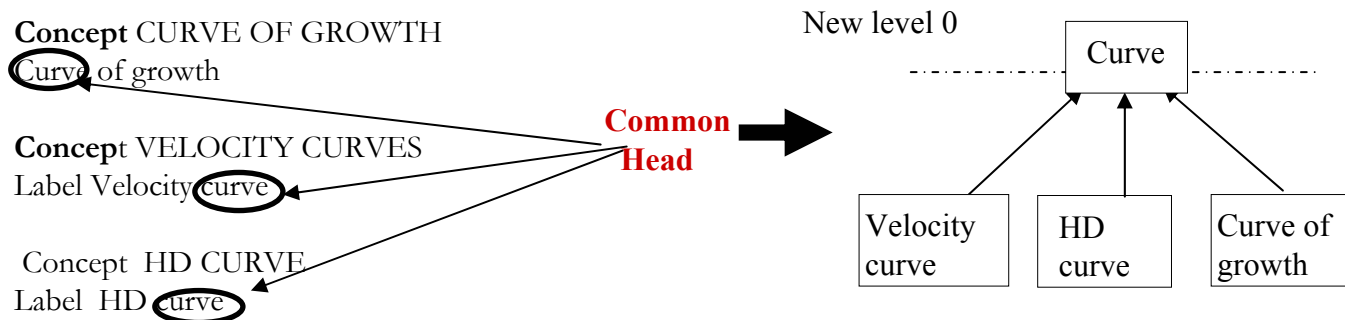
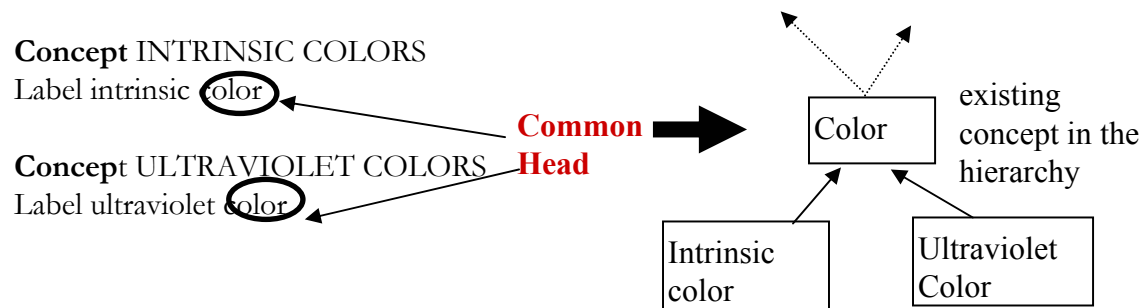
Extraction of explicit taxonomical thesaurus relations

- Thesaurus relations
 - Term1 **Broader Term** Term2
 - Term3 **Narrower Term** Term4
 - Creation of taxonomical relations between concepts whose labels have BT or NT relations in the thesaurus
 - ⇒ For IAU : 1132 top concepts

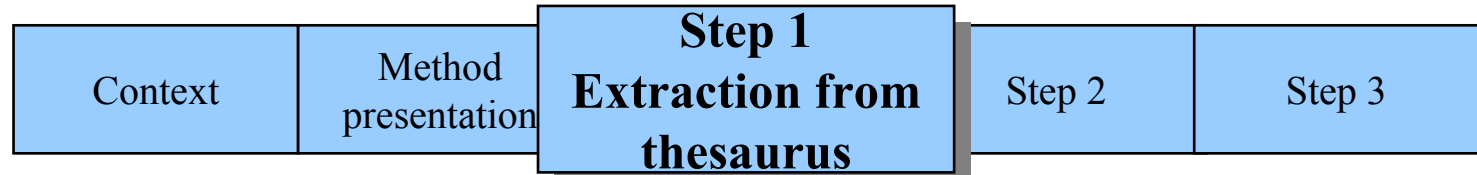


New Taxonomical relations

- New generic level extraction according to label head

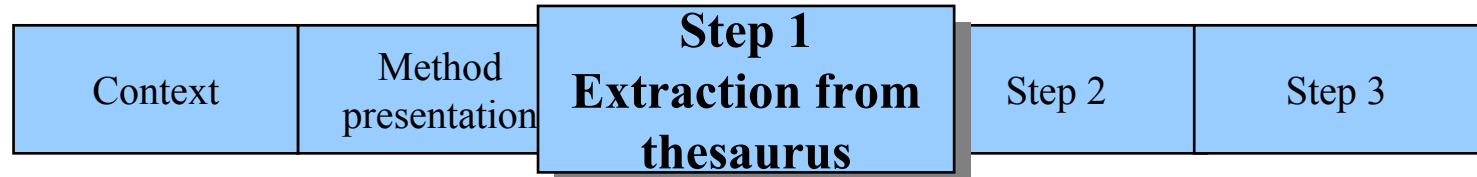


⇒ For IAU : 682 top concepts



Extraction of abstract types

- Abstract types = generic concepts structuring the ontology
- Extraction of the abstract types from a generic ontology (WordNet)
- Process :
 - Automatic mapping ontology's top concepts to concepts of WordNet (62% of the top concepts)
 - Extraction of most generic concepts of the mapped concepts in WordNet (19 concepts)



Extraction of abstract types

- Validation of 14 abstract types

Property : a basic or essential attribute shared by all members of a class

Phenomenon : any state or process known through the senses rather than by intuition or reasoning

Event : *something that happens at a given time*

Science : a particular branch of scientific knowledge

Instrumentation : an artifact (or system of artifacts) that is instrumental in accomplishing some end

Substance : that which has mass and occupies space

Relation : an abstraction belonging to or characteristic of two entities or parts together

Location : a point or extent in space

Angle : the space between two lines or planes that intersect; the inclination of one line to another

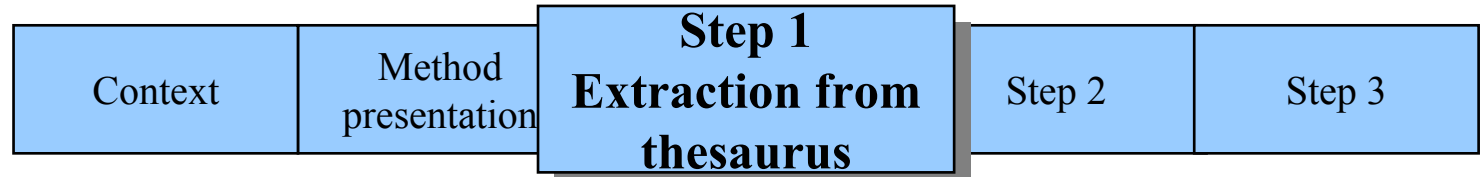
Plane : an unbounded two-dimensional shape

Region : the extended spatial location of something;

Object : a tangible and visible entity

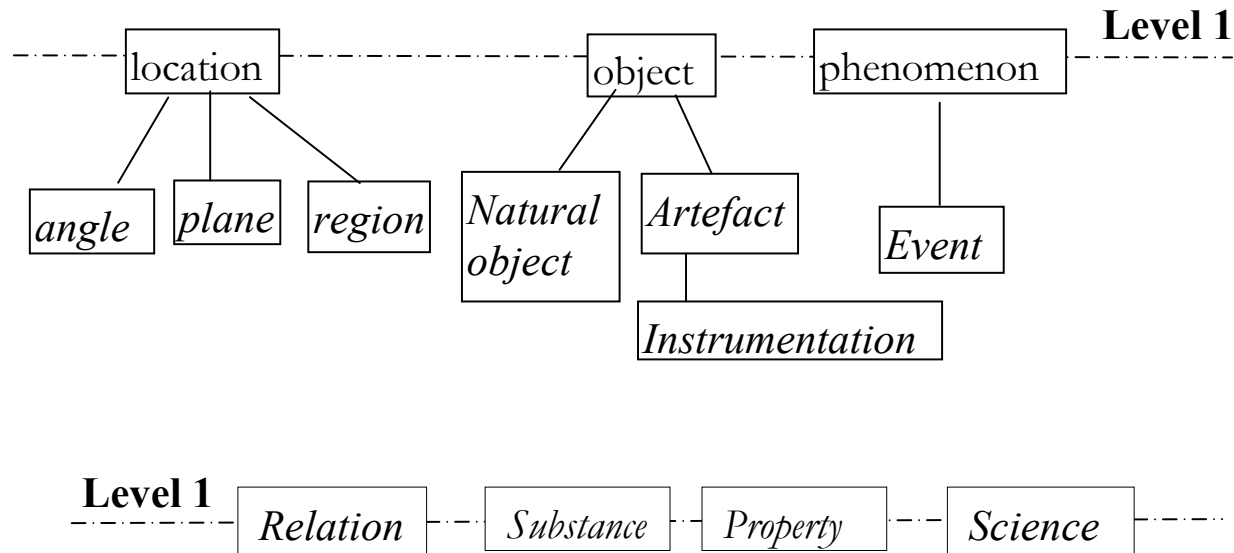
Natural object : an object occurring naturally; not made by man

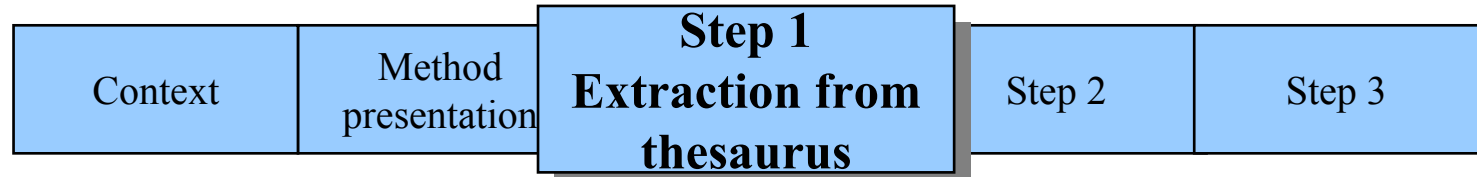
Artefact : a man-made object taken as a whole



Extraction of abstract types

- Organisation



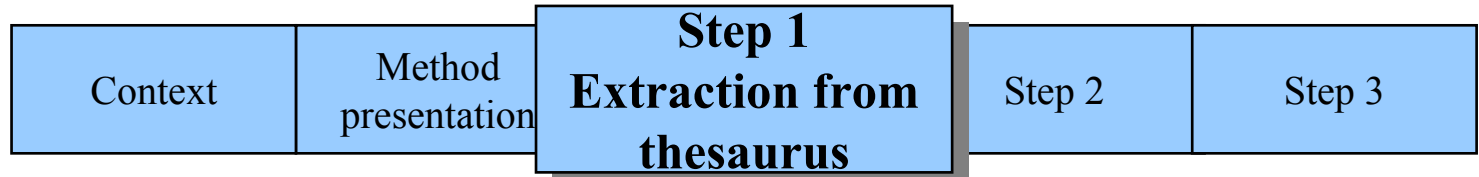


Extraction of abstract types

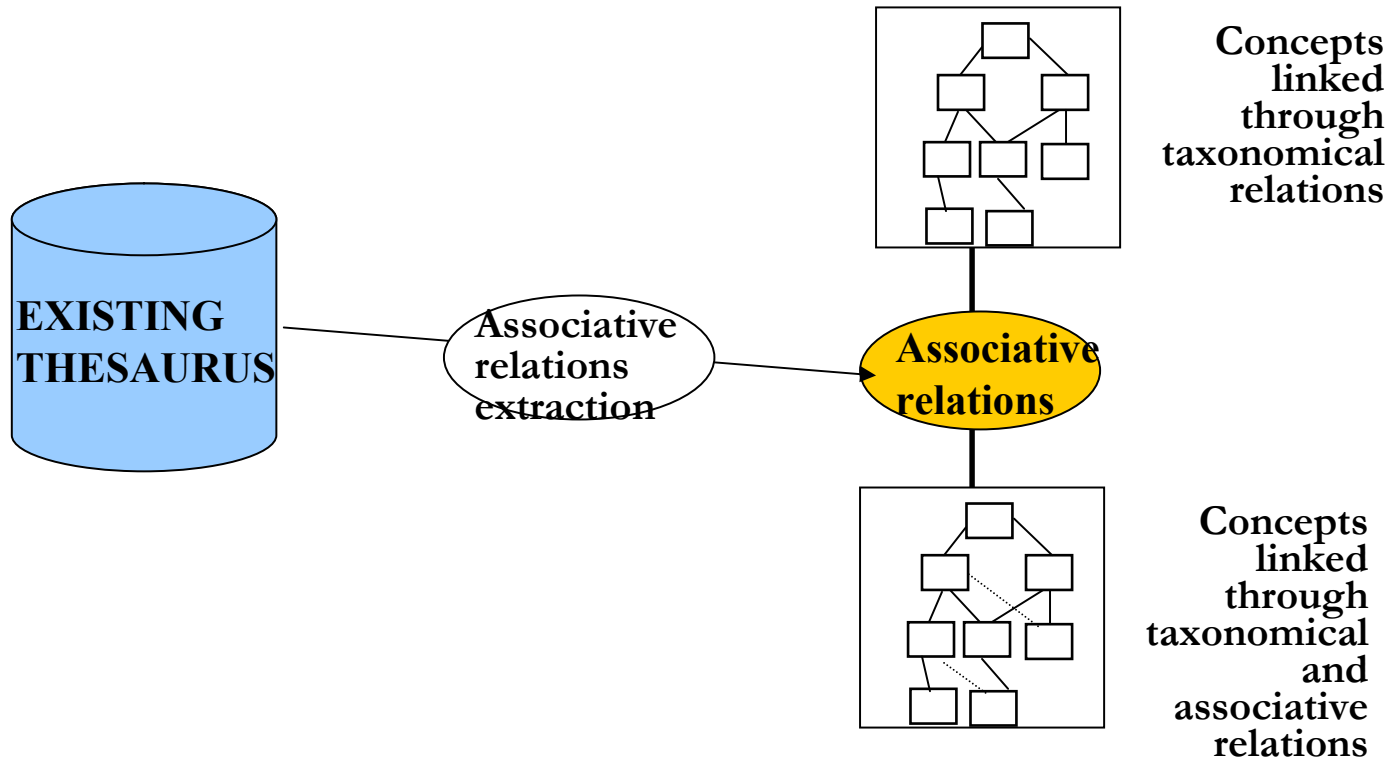
- Validation of the abstract types associated to the ontology's top concepts

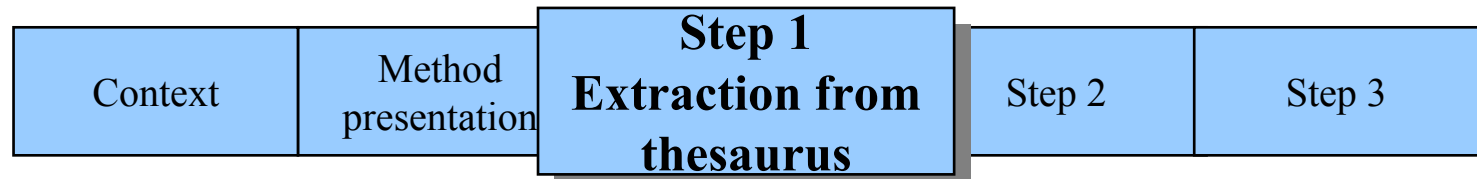
Abstract types	Number of concepts evaluated	Concept for which the abstract type is correct
PROPERTY	53	75%
PHENOMENON	68	67%
EVENT	14	42%
SCIENCE	30	93%
INSTRUMENTATION	13	100%
SUBSTANCE	4	100%
RELATION	19	100%
ANGLE	5	100%
PLANE	4	100%
REGION	15	100%
NATURAL_OBJECT	10	100%
ARTEFACT	34	85%

→ efficient approach



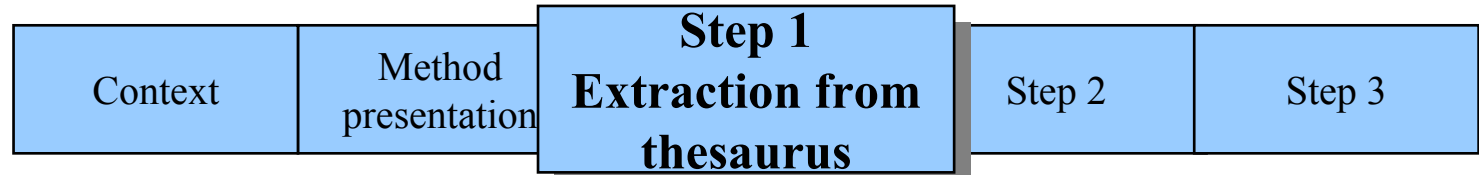
Step 1: extraction of associative relations





Extraction of associative relations

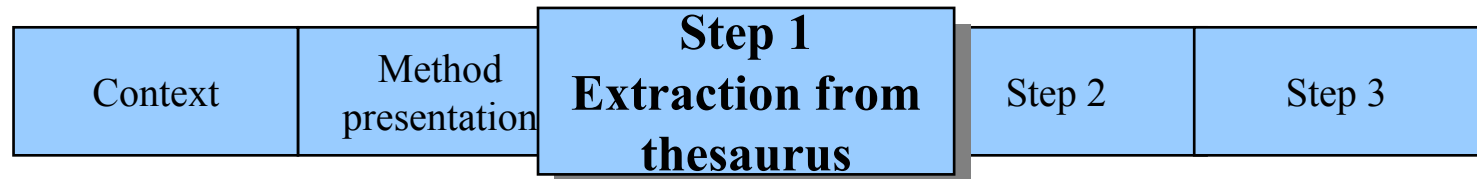
- Explicit relations in thesaurus
 - Term1 **RELATED TO** term2
 - Extraction of associative relations between concepts whose labels are related
 - But relations vague and ambiguous
- Disambiguation of relations according to abstract types



Extraction of associative relations

- Manual definition of relations between types

	Property	Event	Science	Natural object	Instrumentation
Property	Influences Is influenced by Determined by Determines Exclude Has part Is part	Is a property of induces	Is studied by	Is a property of	Is made by Is observed by Is a property of
Instrumentation	Makes Observes Has property	Observes Measures	Is Used to studied	Is observed by	Is ou has part exclude



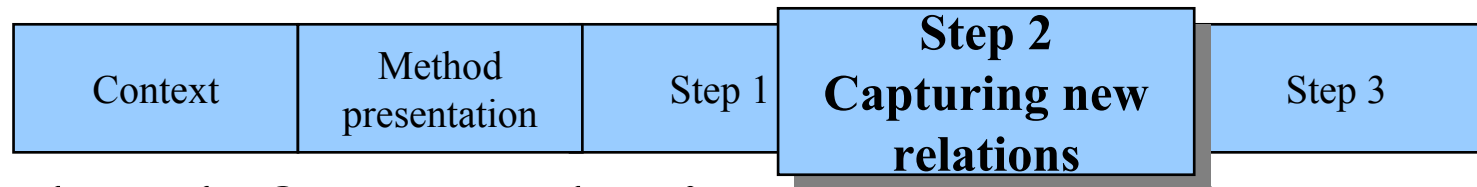
Extraction of associative relations

- Application of the relations between types to the disambiguation of the relation « is related to »

- Validation

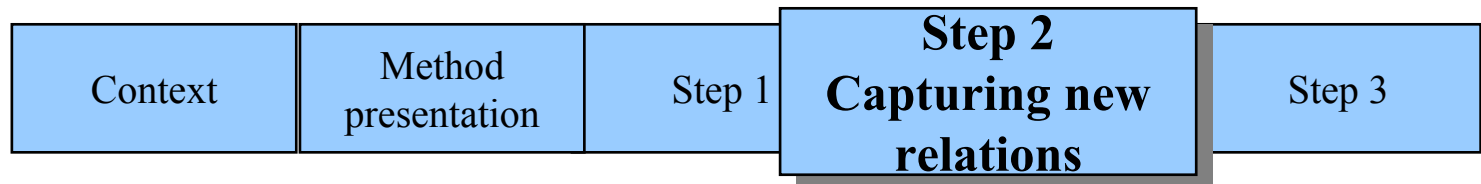
	Number of vague relations evaluated	Number of wrongly desambiguated relations
Concepts linked to the abstract type « property »	34	5
Concepts linked to the abstract type « instrumentation »	15	3

→efficient approach

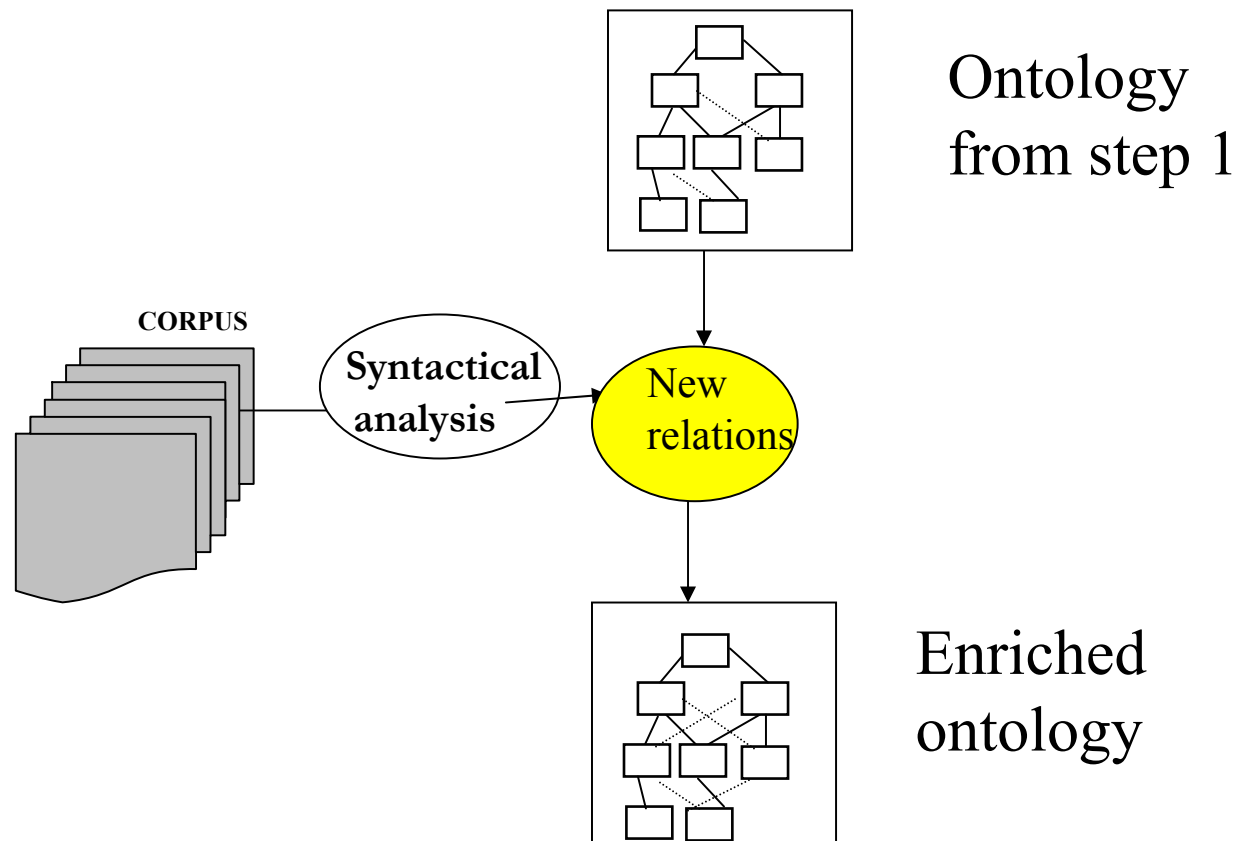


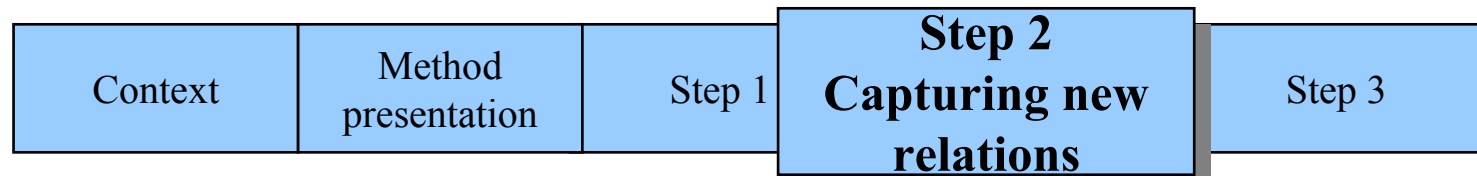
Method for updating thesauri

- 3 semi-automatic steps :
 - Extraction of ontology concepts and structure (relations between concepts) from thesaurus
 - Capture of new relations between concepts not stated in the thesaurus (from texts)
 - Ontology update with new terms and concepts



Step 2 : Capturing new relations





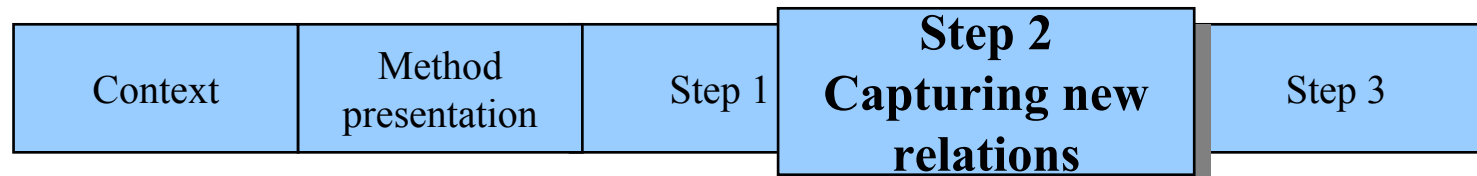
Capturing new relations

- Syntactical analysis of labels' context in reference corpus
- If a concept label occurs in the context of a label of another concept

⇒ Creation of new associative relations between the two concepts

Example : « **intensity** » found in the context of « **radial velocity** » (the intensity of radial velocity)

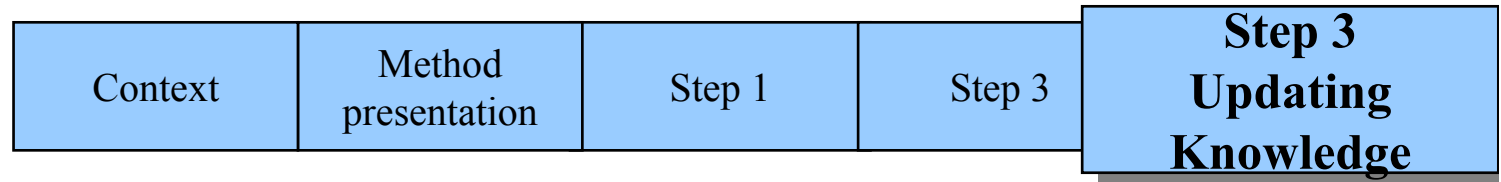
⇒ Creation of the relation « **is a property of** » between the concept « **radial velocity** » and « **intensity** »



Capturing new relations

- Validation

	Number of proposed relations	Number of relations incorrect	Number of incorrect labels
Concepts linked to the abstract type « property »	47	3	2
Concepts linked to the abstract type « instrumentation »	27	2	8

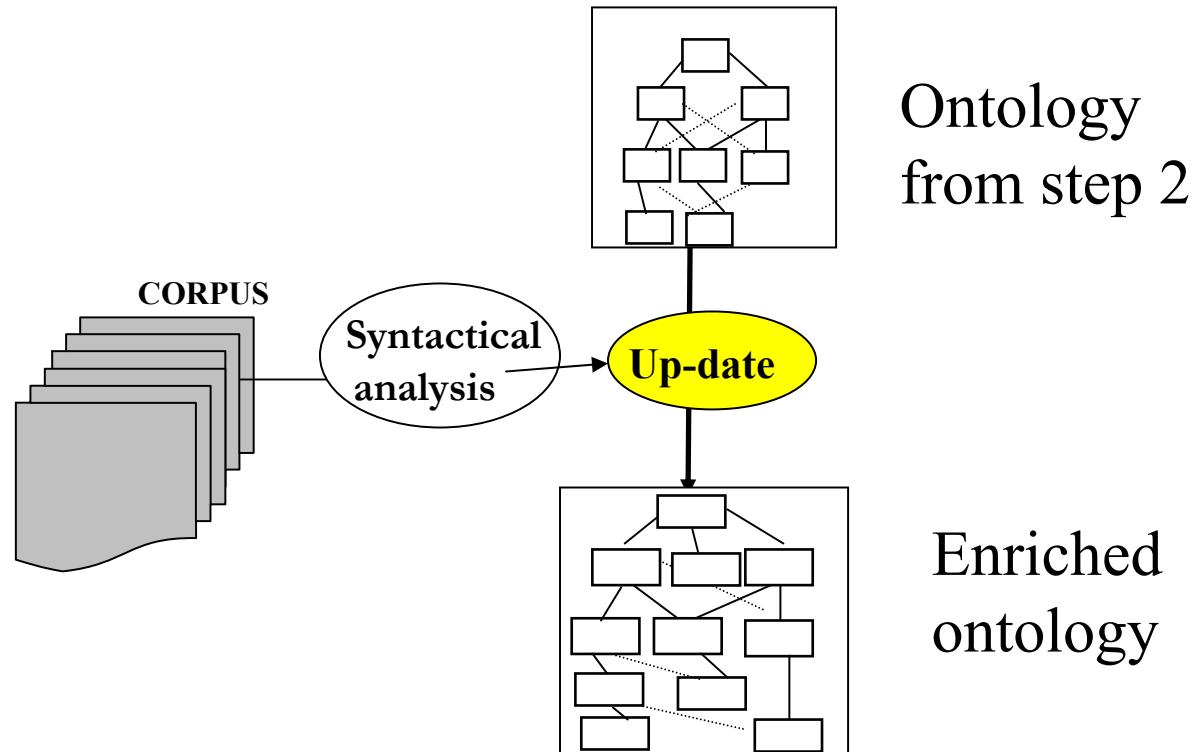


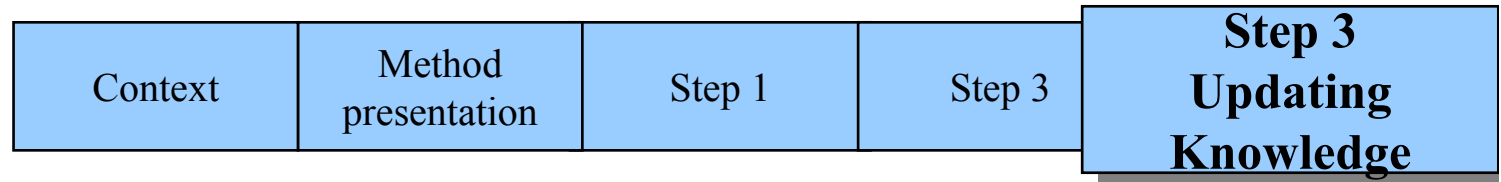
Method for updating thesauri

- 3 semi-automatic steps :
 - Extraction of ontology concepts and structure (relations between concepts) from thesaurus
 - Capture of new relations between concepts not stated in the thesaurus (from texts)
 - Ontology update with new terms and concepts

Context	Method presentation	Step 1	Step 3	Step 3 Updating Knowledge
---------	---------------------	--------	--------	--

Step 3 : updating ontology with new concepts





Extraction of new terms

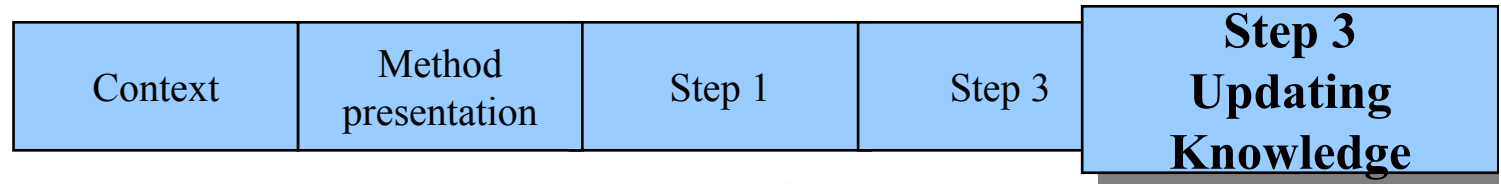
- Two extracting functions of new terms in reference corpus

General terms

high resolution
 globular cluster
 binary system
 soft X ray
 orbital period
 stellar population
 power law
 absorptance line
 line emission
 active region

Specific terms

Yarkovsky force
 Relativistic gravity
 Suprathermal
 electron
 Halpha knot
 Penumbral wave
 Mean free path
 Integral magnitude
 Mixing layer
 stellar population



Terms integration in the ontology

- 2 approaches :
 - New concepts sub-concepts of existing ones
 - New relations between existing concepts
- Others :
 - New labels (synonyms)
 - ...

Context	Method presentation	Step 1	Step 3	Step 3 Updating Knowledge
---------	---------------------	--------	--------	--

Terms integration in the ontology

- New concepts as sub-concepts of existing concepts
→ head of term = label of existing concept

Example : new term “soft **X Ray**”

existing concept with label “**X Ray**”

⇒ creation of the concept “soft X Ray” sub-concept of “X Ray”

Context	Method presentation	Step 1	Step 3	Step 3 Updating Knowledge
---------	---------------------	--------	--------	--

Terms integration in the ontology

- New associative relation between two existing concepts
 - head and expansion of term = labels of existing concepts

Example :

New term : **star mass**

Existing concepts :

- **star** (natural_object)
- **mass** (property)

⇒ creation of the relation « **has property** » between the concepts star and mass

Conclusion



- Method for transformation of a thesaurus into a lightweight ontology
 - Extraction of concepts, labels, relations
 - Update of knowledge using text analysis
- Encouraging results of the method on samples of the thesaurus
- Extend the method to the whole thesaurus

Indexing Process

- Semantic indexing
 - 2 phases
 - Concept detection in documents
 - Concepts weighting
- Indexing with concepts and not with terms often ambiguous
- Help manual indexing using domain knowledge