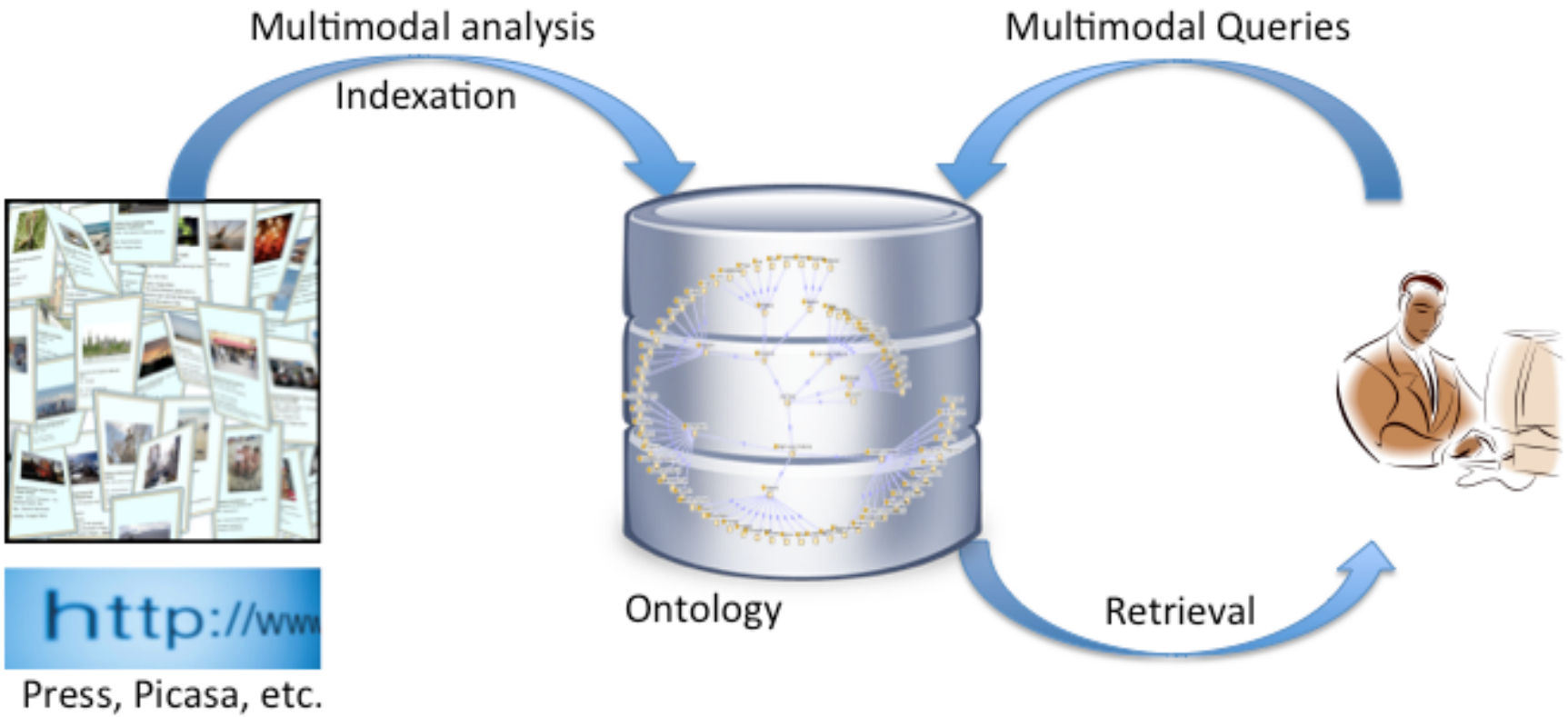
Ontology-driven content extraction using interlingual annotation of texts in the OMNIA project

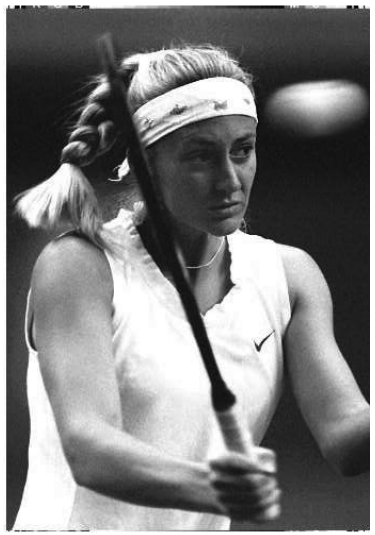
A. Falaise / D. Rouquet / D. Schwab / C. Boitet / H. Blanchon

Context : the OMNIA project



Belga-News dataset : 500.000 images and texts

1012—Australian Open champion Mary Pierce of France volleys the ball during the second-round match against Kyoko Nagatsuka of Japan in the Toray Pan Pacific Open women's tennis tournament in Tokyo 02 Feburary. Pierce defeated Nagatsuka 6-4, 6-0.



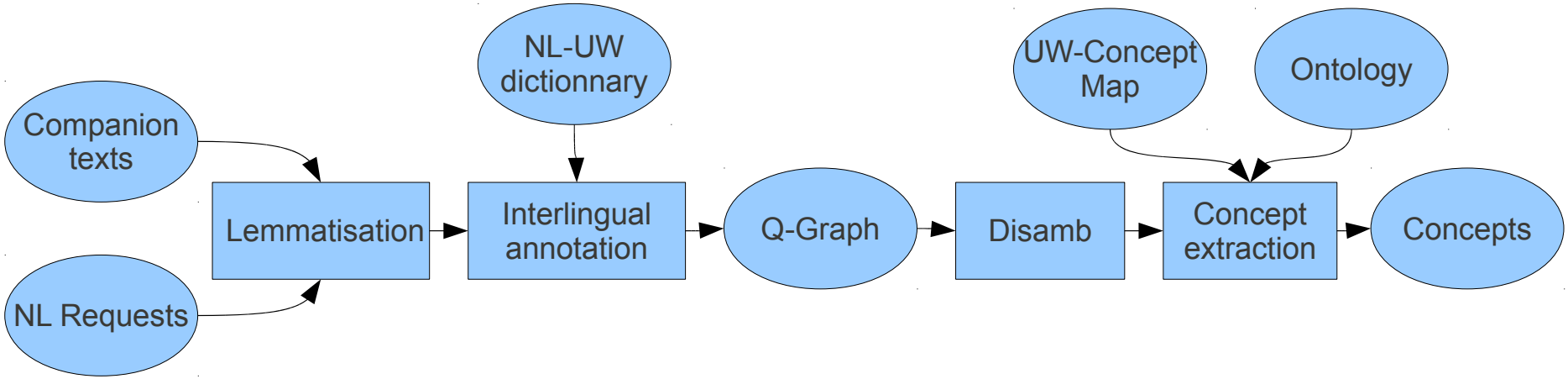
1050730—AWA05 - 20020924 - BAGHDAD, IRAQ : Iraqi women sit under a portrait of Iraqi President Saddam Hussein in a waiting room in Baghdad's al-Mansur hospital 24 September 2002. Saddam Hussein is doggedly pursuing the development of weapons of mass destruction and will do his best to hide them from UN inspectors, the British government claimed in a 55-page dossier made public just hours before a special House of Commons debate on Iraq. Iraqi Culture Minister Hamad Yussef Hammadi called the British allegations "baseless." EPA PHOTO AFPI AWAD AWAD

Typical requests

I would like a photo of war, very violent, with a lot of red and fire in the background

I need a black and white picture of an active woman tennis player during Roland Garros 2010

Architecture of the content extraction



Specificities of our approach :

- **Generic** :
  - Language-independent (using interlingual annotations)
  - Domain-independent (taking an ontology as parameter)
- **Modular** : service-oriented implementation
- **Scalable** : indexation of 500K images, on-the-fly processing of requests
- **Ambiguity management** : keeps NL ambiguities in dedicated structures

Interlingual annotation

Main ideas :

- **Use interlingual lexemes (UWs) as lexical annotations** :
  - not directly conceptual symbols (too ambitious, and fixed)
  - **UWs** (Universal Words) are the lexemes of UNL, an *anglo-semantic pivot* usable for many other NLP applications
- **Don't use full UNL hypergraphs** (representing NL utterances) because
  - no analyzer (or UNL enconverter) produces more that 30% complete structures on arbitrary, spontaneous sentences)
  - if a level of annotation is not good enough, theory & practice prove that it is counterproductive to use it for IR.
    - F-measure decreases
    - computing time increases

Data structure for annotated texts

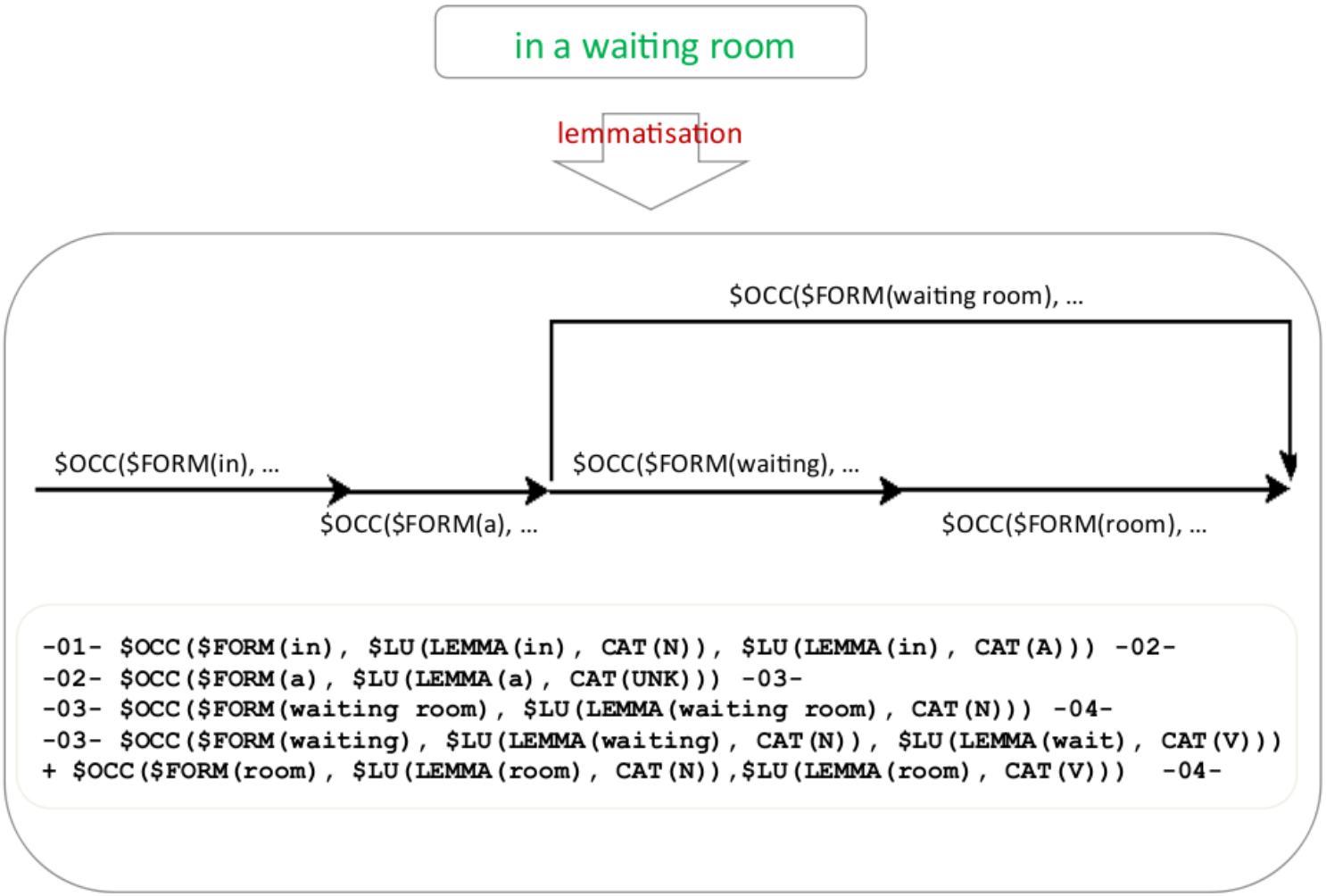
Q-language

- structures manipulated: chain graphs of labelled trees (Q-graphs)
- transformed by sequences of sets of rewriting rules (Q-systems)
  - addition step: all rules are applied, results are *added* to the Q-graph
  - cleaning act: if and when addition terminates, the maximum well-formed Q-graph not containing any used arc is produced.
- developed for MT by Alain Colmerauer in 1967
  - used to implement the TAUM-meteo system (1976), then to run it operationnally (1977-1992) — 30M words/year in E↔ F.
  - then replaced by the similar but more typed and less non-deterministic language, GramR (METEO system, J. Chandioux, 1992-2002)

Interest

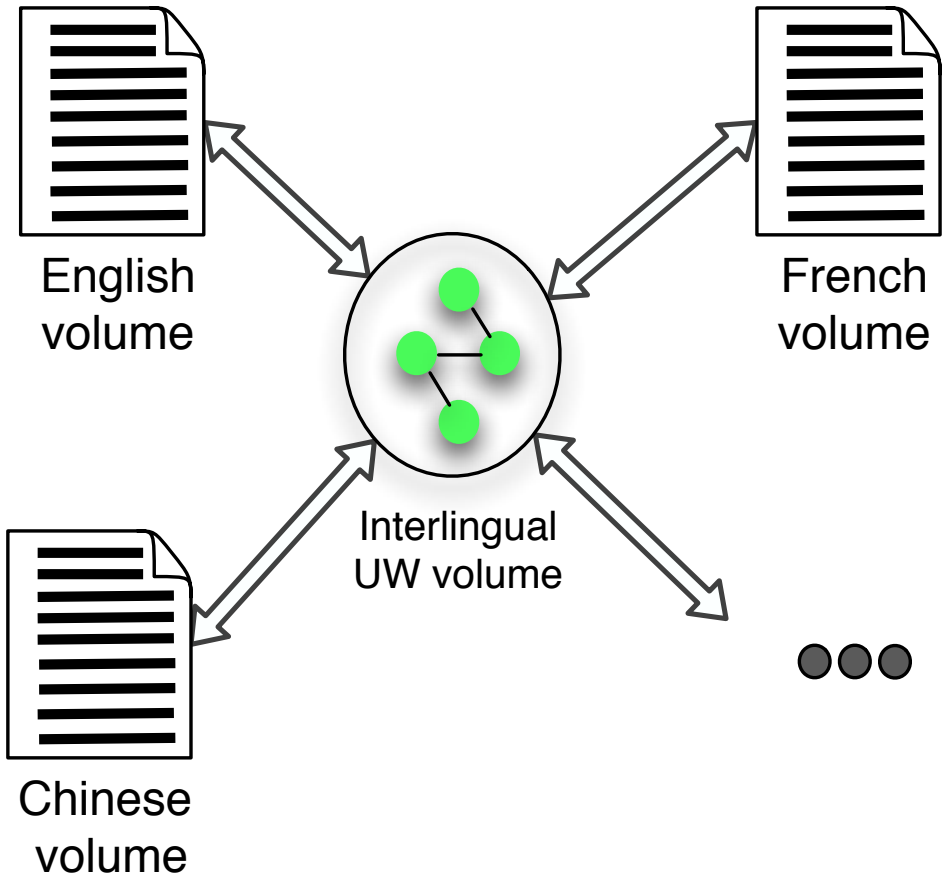
- A Q-graph can represent all ambiguities in a text
- It is easy to write powerful graph rewriting systems as Q-systems
- The Q-language is very easy for linguists to learn and use

Transformation of a lemmatized text into a Q-graph



- Lemmatization is done with a language-dependent piece of lingware (NooJ, DELAF...)

{UWs} as an interlingual lexical pivot



Interlingual lexemes for annotation

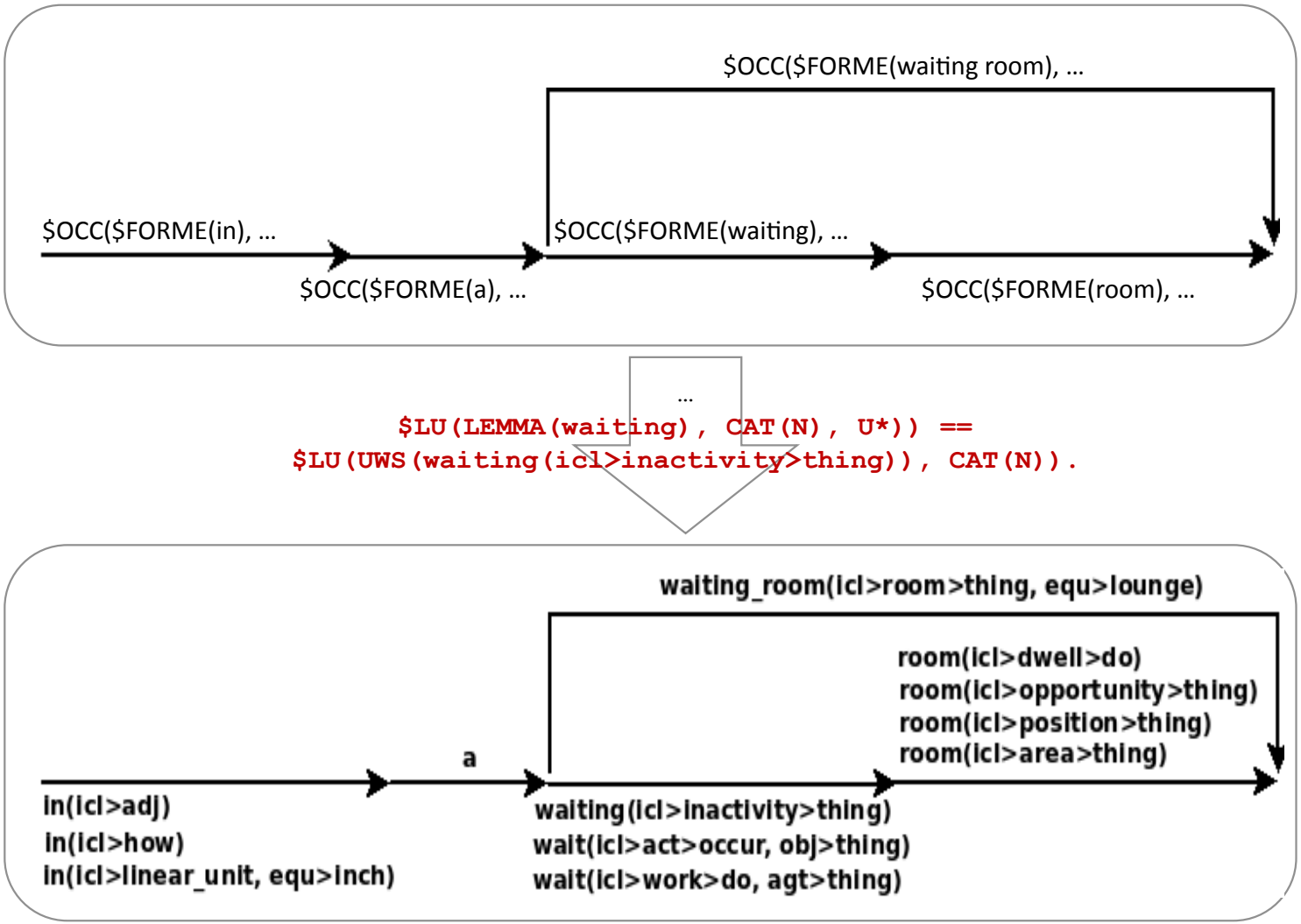
Universal Network Language (UNL)

- an international project started in 1996 by the United Nations University
- an abstract pivot language (anglo-semantic hypergraphs)

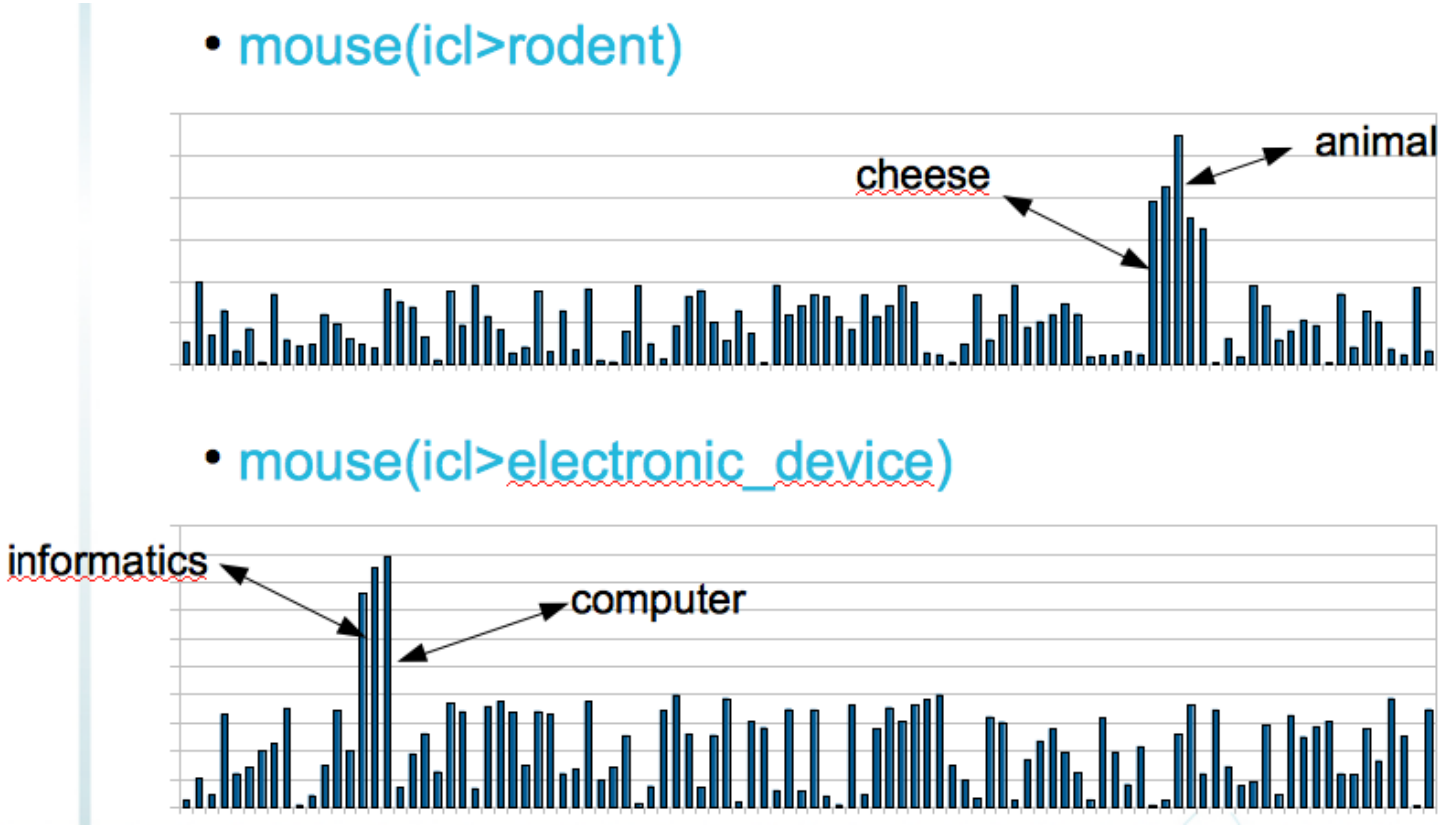
Universal Words (UWs)

- UW : Headword "(" semantic_restrictions ")" ; (ANTLR syntax)
- UWs represent acceptions without ambiguities
- examples :
  - book(icl>do, agt>human, obj>thing)
  - book(icl>thing).
  - ikebana(icl>flower_arrangement).
- 200,000 UW++ built from WordNet synsets

A Q-system is used to add UWs to the lemmas

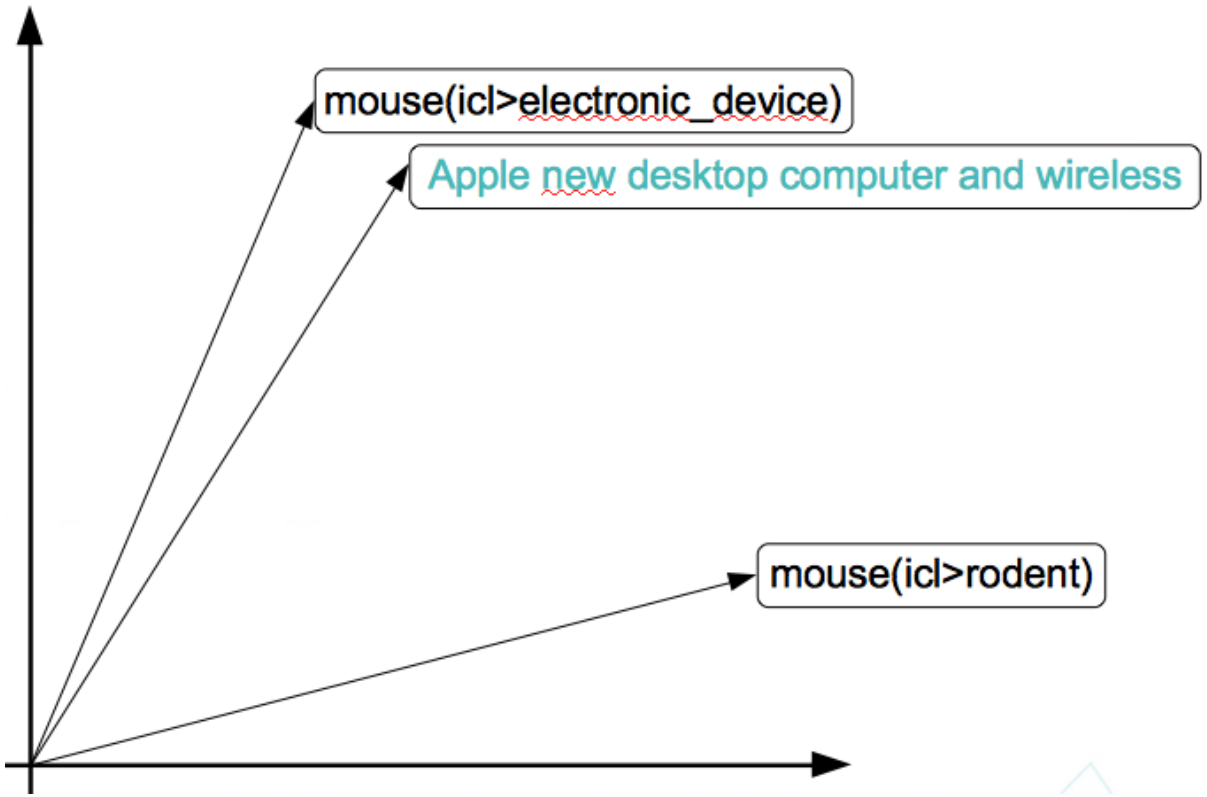


Conceptual vectors



Vectors (which components correspond to *emerging "ideas"*) are associated to UWs and (by a composition rule) to sentences.

Word Sense Disambiguation (WSD) based on Conceptual Vectors



The smaller the angular distance between a word sense and its context, the more likely it is that word sense is the correct one in this context.

Ontology-driven content extraction

An ontology is our "domain parameter"

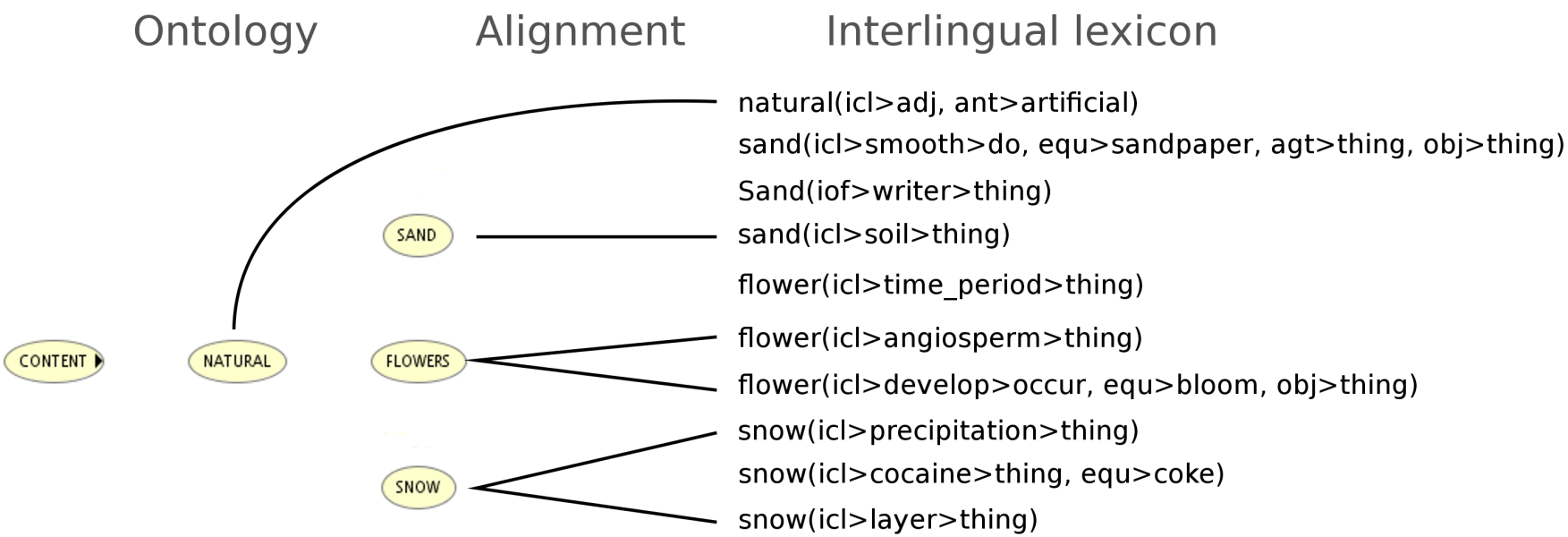
. . . it formalizes the "things we want to extract" :

- Ontologies give an axiomatic description of a domain, based on formal logics (soundly usable by software agents)
- Ontological structures are close to the organisation of ideas as semantic networks in human minds
- Semantic Web normative initiatives (W3C) offer many shared tools for editing, querying, merging, etc.

The OMNIA ontology

- 732 concepts (animals, politics, religion, army, sports, monuments, transports, games, entertainment, emotions, etc.)
- Instances are the images to classify

Alignment between the ontology and the UWs

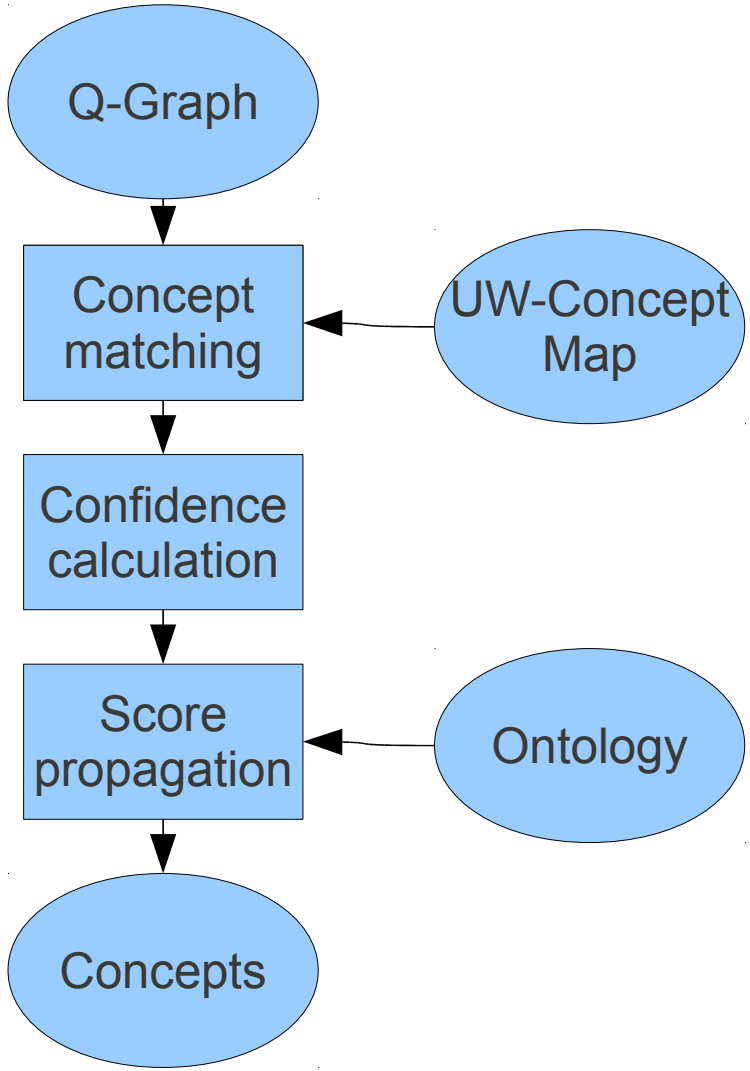


- **Representation** using ontology alignment formats
- **Computation** using the Ontology Alignment API[1]
  1. base alignment with a bilingual dictionary (UWs - Ontology symbols)
  2. disambiguation adapting NLP techniques or structural ontology matching (to score the alignments)

[1] http://alignapi.gforge.inria.fr

Generic content extractor

A 3 steps process :



Score computation

The likelihood score of a concept is equal to :

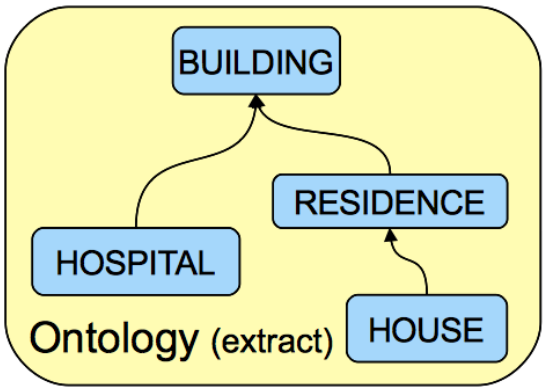
(score of the UW) $\times$ (score of the UW–concept alignment)

Interpretation in fuzzy sets theory

A concept score is the "membership degree" of an image in the class corresponding to the concept in the ontology

Score propagation

Membership degrees of an image "i" are propagated through the ontology hierarchy according to given fuzzy operators (for each superclass, we compute a lower bound of the membership degree).
For instance, with the following ontology and Gödel fuzzy operators :



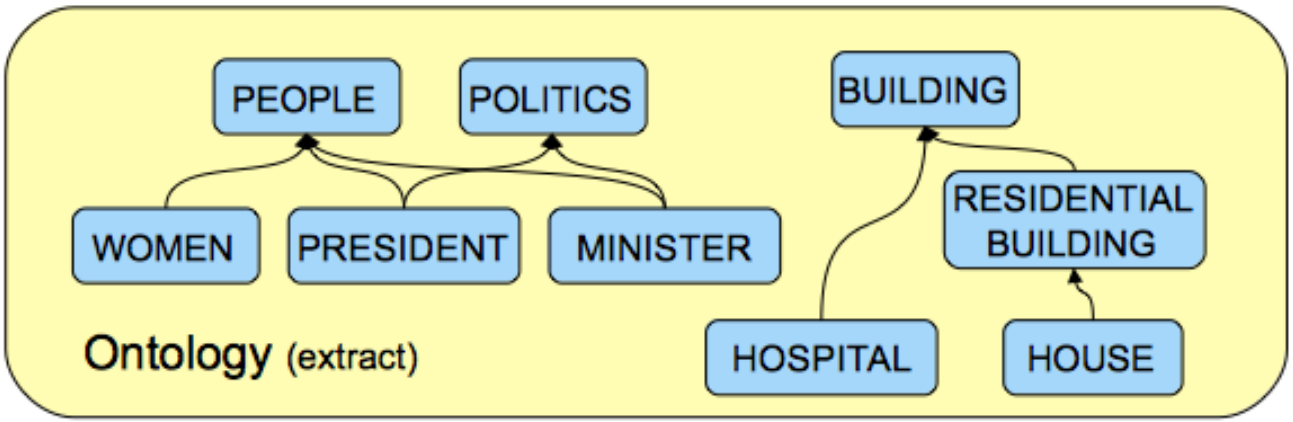
if $M_{HOUSE}(i) = 0,5$ and $M_{HOSPITAL}(i) = 0.7$
we know that $M_{RESIDENCE}(i) \geq M_{HOUSE}(i) = 0,5$ and
$M_{BUILDING}(i) \geq max(M_{HOSPITAL}(i), M_{RESIDENCE}(i)) = 0,7$ because
$HOUSE \subseteq RESIDENCE$ and $(HOSPITAL \cup RESIDENCE) \subseteq BUILDING$

First experiments

Output example

Concept	Score
RESIDENTIAL BUILDING	0.0002
PRESIDENT	0.0004
BUILDING	0.0004
HOUSE	0.0002
HOSPITAL	0.0002
POLITICS	0.0044
MINISTER	0.0040
WOMAN	0.0002
PEOPLE	0.0146

AWA05 - 20020924 - BAGHDAD, IRAQ : Iraqi women sit under a portrait of Iraqi President Saddam Hussein in a waiting room in Baghdad's al-Mansur hospital 24 September 2002. Saddam Hussein is doggedly pursuing the development of weapons of mass destruction and will do his best to hide them from UN inspectors, the British government claimed in a 55-page dossier made public just hours before a special House of Commons debate on Iraq. Iraqi Culture Minister Hamad Yussef Hammadi called the British allegations "baseless." EPA PHOTO AFPI AWAD AWAD



First experiments in the OMNIA project

Dataset

- The OMNIA ontology (732 concepts)
- 1000 images+texts from Belga News

Automatic evaluation

- Average computation time: 8s with disambiguation, 3s without
- 23% of the images received no class
- In average, images belonged to 6 classes

Manual evaluation on 30 images/texts

- 127 concepts were identified
- **Visual relevance** (the concept is in the image) :
  99 concepts (80%)
- **Textual relevance** (the concept is in the text) :
  124 concepts (89%)

Conclusion and future work

We presented a textual content extraction method that is :

- **language-independent**: annotations by *interlingual lexemes*
- **domain-independent**: the ontology is a parameter
- **modular**: service-oriented implementation
- **scalable** "out of the box"

We are still working on :

- optimization
- exploitation of linguistic features (negation, intensification, etc.)
- integration with the other OMNIA multimodal components
- multilinguality (testbed in preparation by collaborative postedition of MT of a sample)

Thanks for your attention