Ontologies and multilinguality

Dra. Guadalupe Aguado de Cea
lupe@fi.upm.es
http://www.oeg-upm.net
Ontological Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
Campus de Montegancedo sn,
28660 Boadilla del Monte, Madrid, Spain

Outline

- Definition and purpose of multilinguality
- Localization vs. internationalization
- From monolingual to multilingual systems
- NLP systems including multilinguality
- Multilinguality in KB systems
- Multilinguality in ontologies
  - Information
  - Realization
  - Modelling
- A new proposal: Linguistic Information Repository- LIR

Multilinguality. What for?

- Multilinguality is required in different NLP applications
  - Question answering systems
  - Multilingual information retrieval
  - Multilingual speech processing
  - Machine translation
- Knowledge sharing → ontologies
- Reusing Ontologies → Semantic Web

How can we provide multilinguality? Localization vs. internationalization

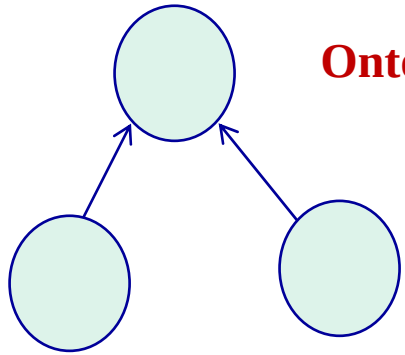
- ***Localization*** involves taking a product and making it linguistically and culturally appropriate to the target locale (country/region and language) where it will be used and sold (LISA)
- In **economics**: adapting a product to *a non-native environment*.
- In **software y *web design***: adapting contents, language, and design to the target language and culture
- In **ontologies:** *Ontology Localization involves the process of adapting an ontology to a particular language and culture*.

Internationalization

- ***Internationalization*** *is the process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for re-design. Internationalization takes place at the level of program design and document development (LISA).*
- Important aspects:
  - Separating text from the source code – > prevents translators from changing the source code
  - Internationalization is also applied to *online help, documentation and web sites*
  - Technical writers have to take internationalization into account: "*writing for a global audience*", "*web site globalization*"

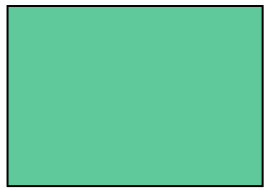
From software localization to ontology localization

SW Localization

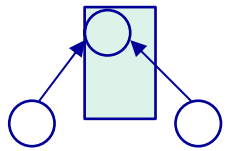


Ontology localization at different levels

SW Internationalization



Metamodel



Ontology model

From software localization to ontology localization: Similarities

- Internationalization:
  - **Lexical content : characters and symbols handled by the computer (ASCII encoding, UNICODE, etc.)**
  - **Grammatical content: characters, syntactic structures and symbols used in certain ontology languages (RDF(S), OWL)**
  - **Representation paradigm layer: frames, semantic networks, DL, (Ontologies)**

- Localization:
  - **Lexical-terminological content: terms or words used to name ontology elements.**
  - **Conceptual content: conceptualization decisions: granularity, expresiveness, perspective, etc., mainly in domain ontologies.**
  - **Pragmatic content: final result of the model (GUI, etc.)**

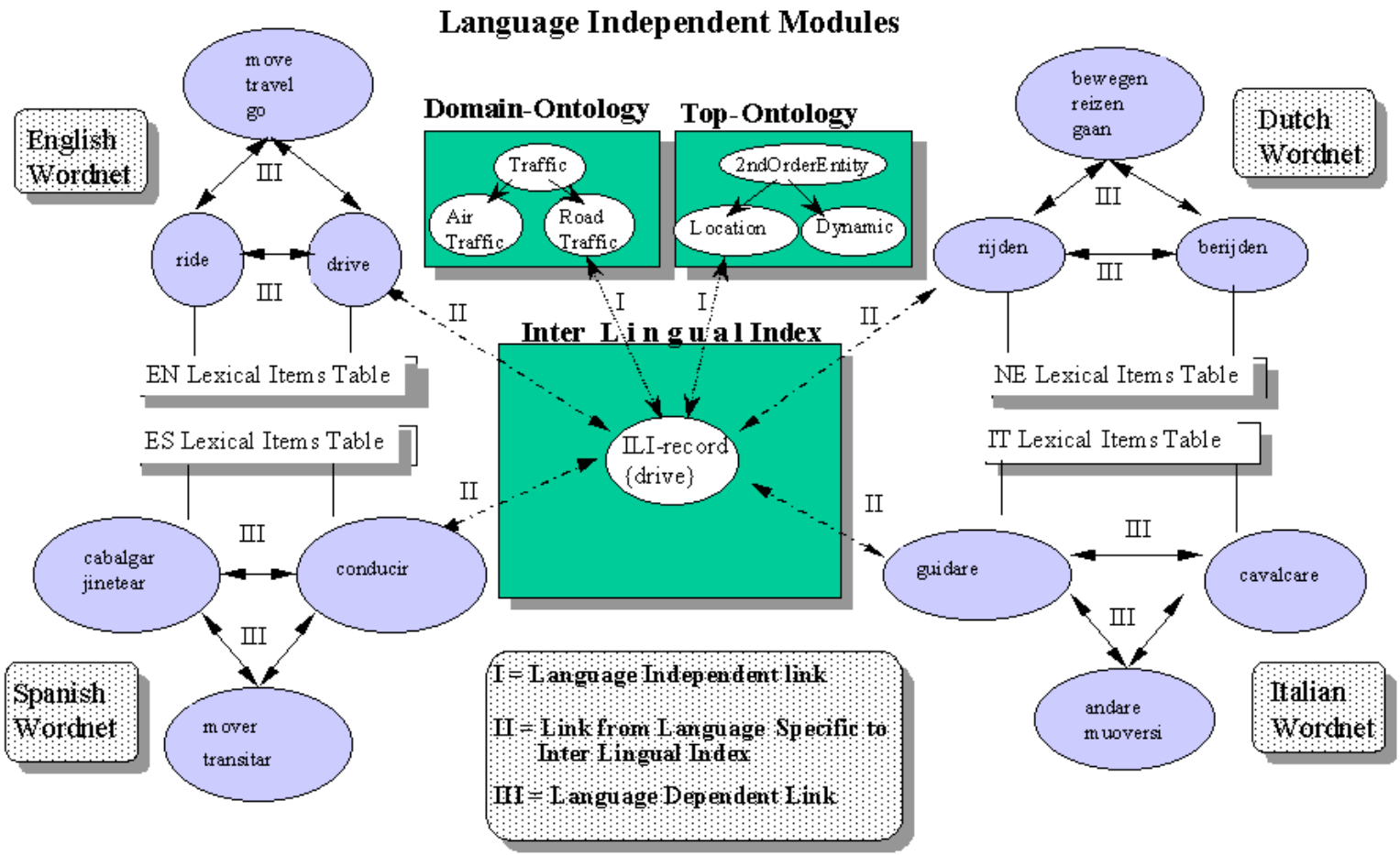
From monolingual to multilingual systems

- Few multilingual ontologies
  - http://olp.dfki.de/ontoselect/
  - 1652 ontologies
  - 149 with language information
  - 130 in English, 10 in Spanish
  - 5: en-es, 4: en-es-fr
- Scarce information available about how to represent multilinguality
- Recent interest in international research groups:
- LISA (*Localization Industry Standards Association*)
- OSCAR (*Open Standards for Container/Content Allowing Re-use*)
- OASIS (*Organization for the Advancement of Structured Information Standards*)
- W3C
- ISO *International Standards Organization*

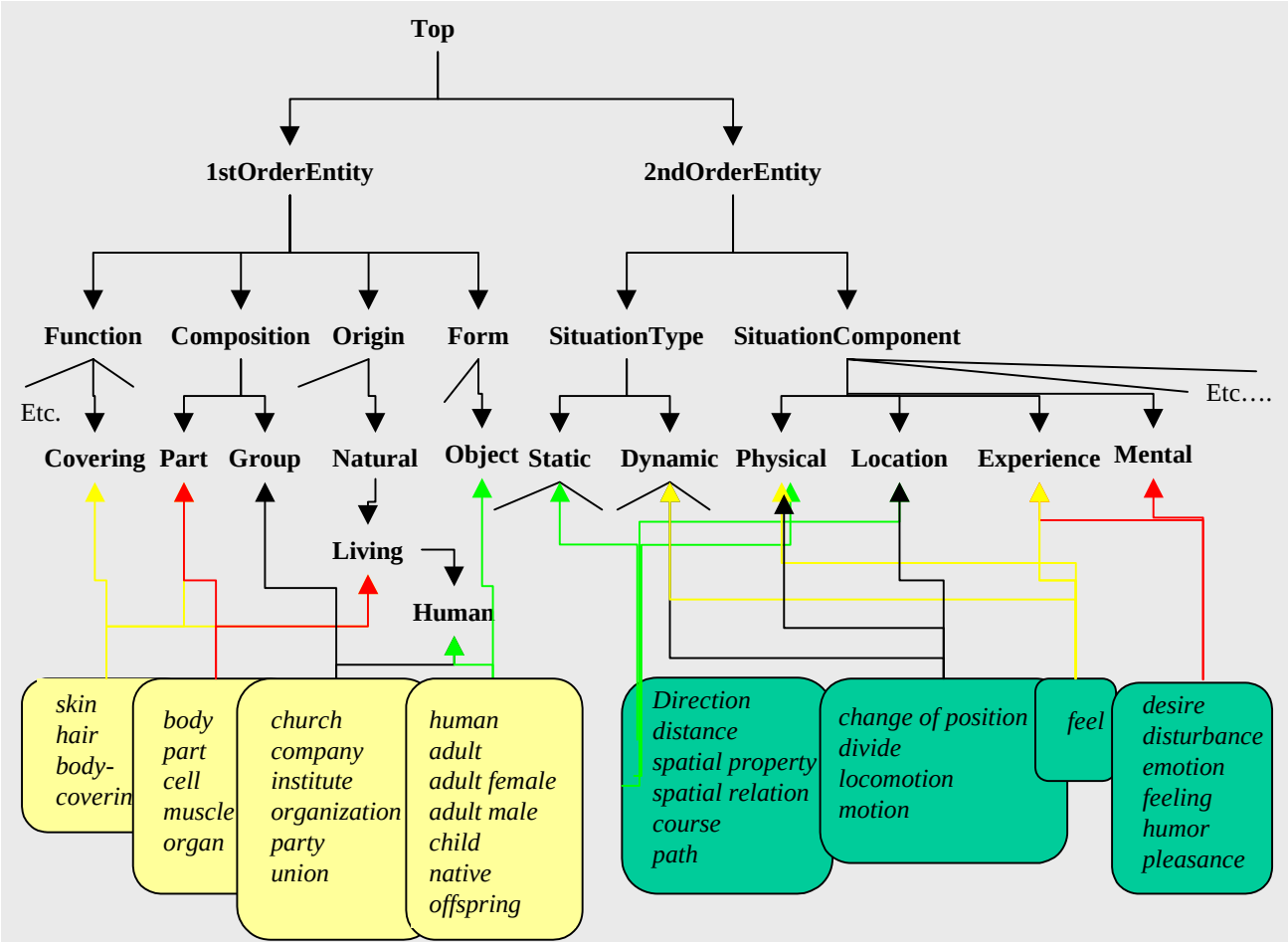
NLP systems including multilinguality: EWN

- **EuroWordNet**
  - Based in Wordnet, http://wordnet.princeton.edu/perl/webwn
- **Objectives**
  - Building a multilingual database with wordnets for several languages: Dutch, Italian, Spanish, English, German, French, Estonian and Czeck
  - Building ***wordnets:*** monolingual autonomous ontologies, connected by an ILI
  - Based on Wordnet *synsets:* set of synonymous word meanings and basic semantic relations
  - Using existing national resources to build networks independently
  - Maintain the language specific structures and compare concept relations
- **Participants: 8 universities (UNED, UPC), 3 Business firms.**
- **Funded by the EU.**

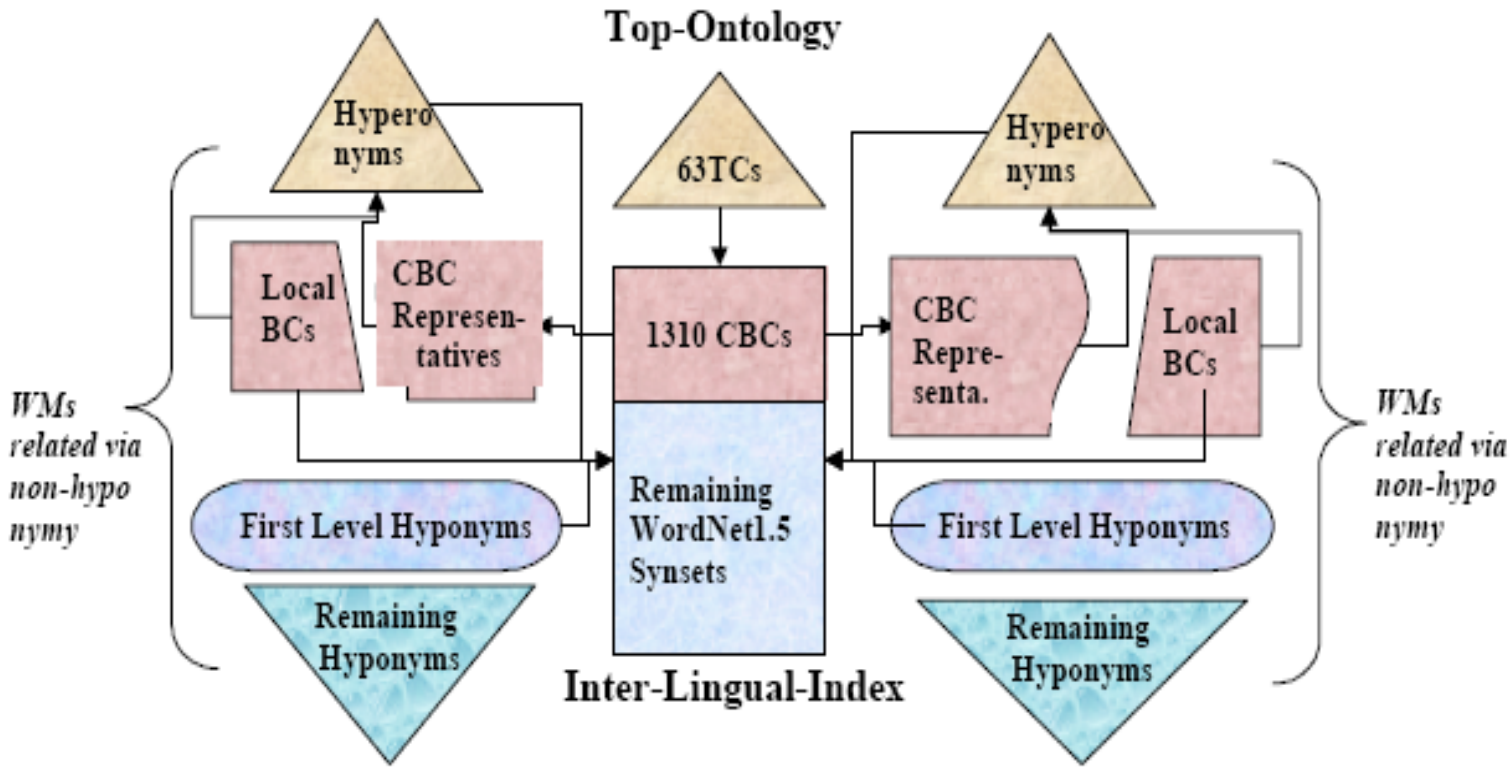
EuroWordNet. Architecture



EuroWordNet: Top Ontology



Mapping two EWN *wordnets* to the Wordnet ILI



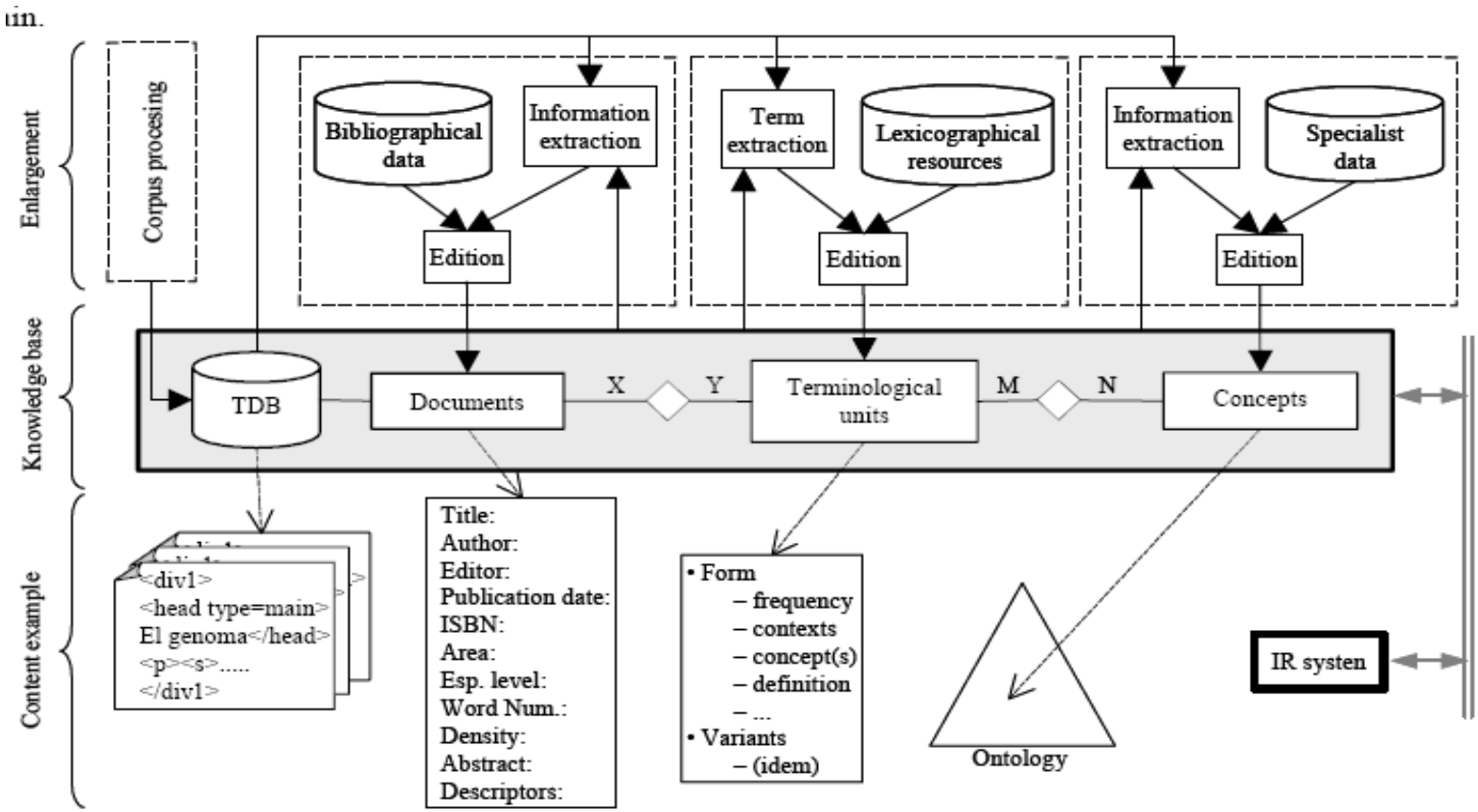
(Vossen, 2002)

TC: Top concepts
CBC: Common Base Concepts
BC: Base Concepts

NLP systems including multilinguality: Genoma-KB

- **Ontological Module:** (MikroKosmos) 21 basic concepts: ALL, OBJECT (physical, mental, social), EVENT (physical, mental, social), PROPERTY (attribute, relation), etc.
  - Feliu (2004) describes certain relations
    - Similarity, Hyponymy, Sequential relations (place and time)
    - Causality, Instrument, Meronymy, Association
- **Terminological Module:**
  - Multilinguality, POS, context, sources, lemma, administrative information
- **Corpus Module:** multilingual texts (En-Cat-Sp)
- **Entities Module:**
  - Bibliography: complete references of terms and texts
  - Factual module: research centres, people, institutions, etc.

Knowledge base architecture: GENOMA-KB

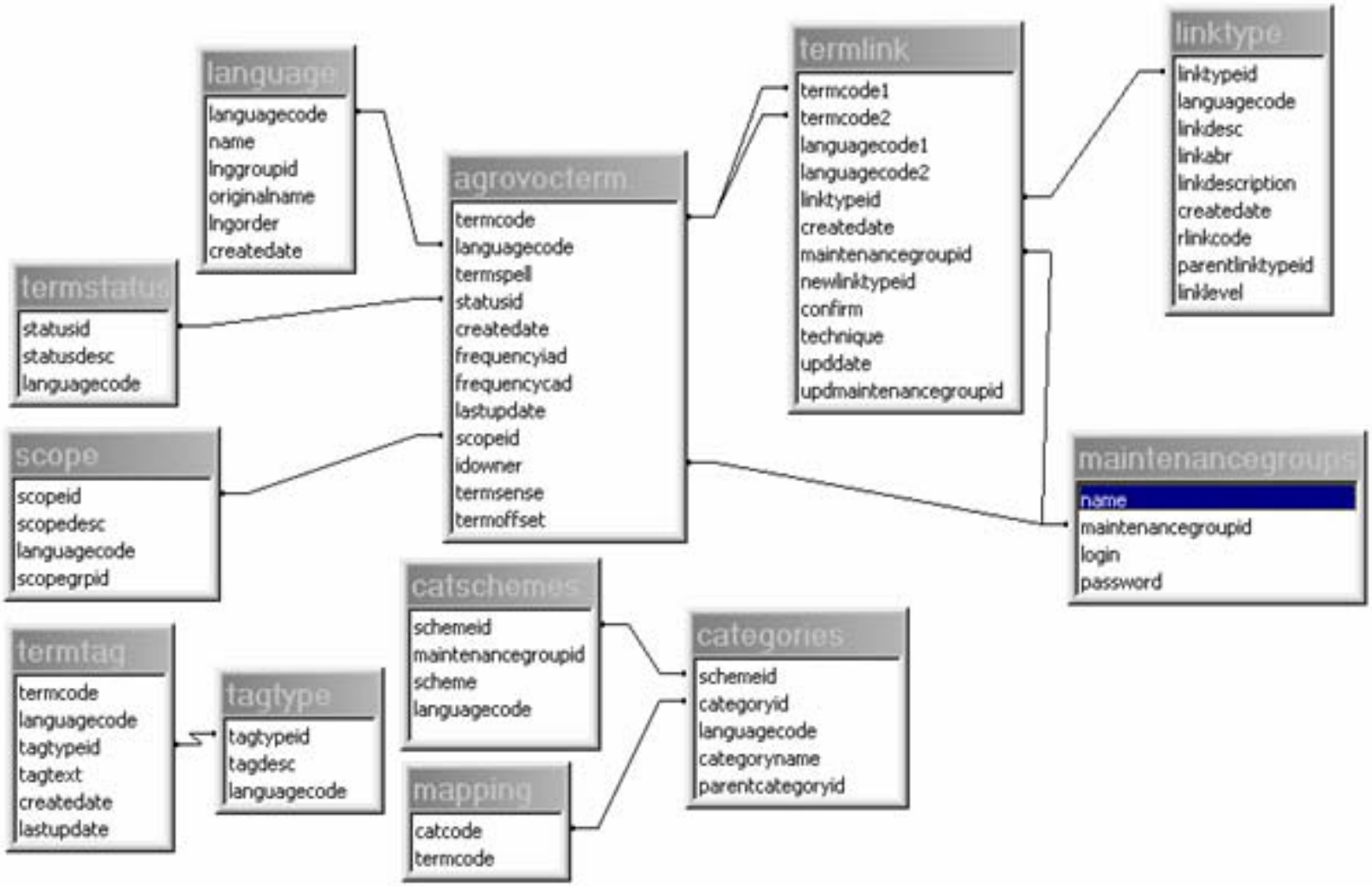


(Feliu, Vivaldi y Cabré, 2002)

NLP systems including multilinguality: AGROVOC

- AGROVOC Thesaurus developed by the FAO (*Food and Agriculture Organization*) and the EU in 1980/1982.
- Initially, 3 languages. Now, 10 languages. They want to include more.
- It is defined as "*a multilingual structured and controlled vocabulary*".
- Used to index and retrieve data about fisheries and food
- It shows the amount of terms in real time (41,580 terms in Spanish)
  - URL http://www.fao.org/aims/ag_figures.jsp

AGROVOC: representation of multilingual information

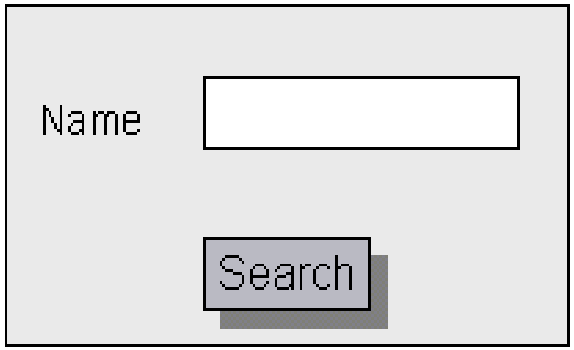


Multilinguality in KB systems (1)

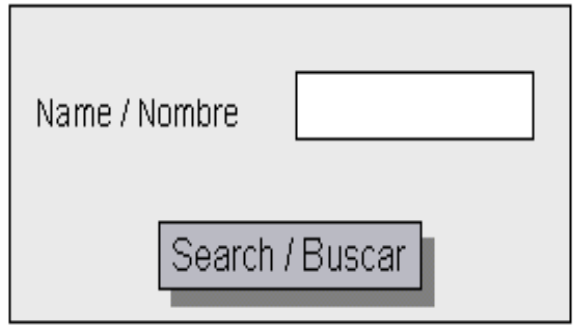
- Multilinguality can be included at three levels:
  1. Interface
     - (a) Messages
     - (b) Contents
  2. Data
  3. Knowledge representation

- *Aguado de Cea,G., Montiel Ponsoda, E., Ramos Gargantilla, J.A. "Multilingualidad en una aplicación basada en el conocimiento", Procesamiento del lenguaje natural, nº 38, Abril 2007*

Multilinguality in KB systems
1. Interface (a) Message visualization



1. Monolingual messages

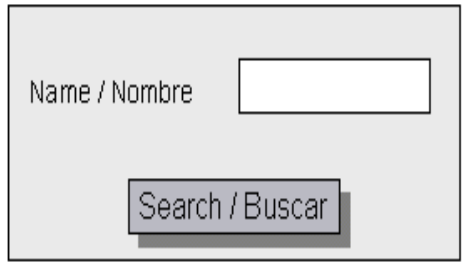


2. Simultaneous bilingual messages

3. Non-simultaneous multilingual messages

Advantages and disadvantages of option (a)

- With simultaneous visualization, adding more languages requires modifying the visualization code.



- With non simultaneous visualization, there is no need to modify the whole code, but it requires:
  - Increasing the number of interfaces
  - Modifying the interface as for the selection options.

(b) Multilingual contents visualization

- When the KB is multilingual
  - The application invokes the KB once
  - The interface shows the contents in the selected language
- When the KB is monolingual
  - The application invokes the KB
  - It uses a translation system (multilingual resource: dictionary, glossary, lexicon, DB, etc)
  - The interface shows the translation
- Similar interface in both cases to message visualization.

Advantages and disadvantages of option (b)

- When the KB is multilingual
  - Time employed in obtaining contents = reply time (RT) from the KB
    - Reason: multilinguality has been provided in design time
    - Disambiguation: in design time
- When the KB is monolingual
  - Time employed in obtaining contents = RT from the KB

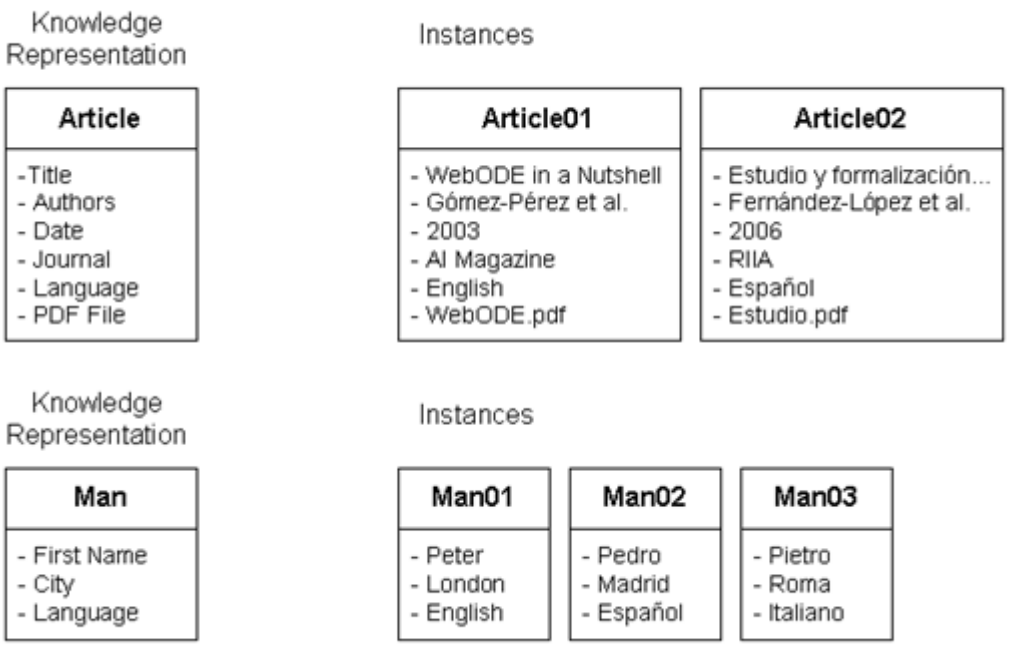
  + RT from the multilingual resource
  - Translation is carried out during execution time
  - Disambiguation: it may increase the reply time (RT)

Multilinguality in KB systems (2)

- Multilinguality can be included at three levels:
  1. Interface
     - (a) Messages
     - (b) Contents
  2. Data
  3. Knowledge representation

- *Aguado de Cea,G., Montiel Ponsoda, E., Ramos Gargantilla, J.A. "Multilingualidad en una aplicación basada en el conocimiento", Procesamiento del lenguaje natural, nº 38, Abril 2007*

Multilingual data in KB systems



- Information about individuals is multilingual
- Multilinguality will be dealt with as another aspect of the domain to be modelled

Multilingual data in a KB system that includes `Language`

Multilinguality in KB systems (3)
Knowledge representation

- **Data**: instances or individuals, lower level in Knowledge Representation (KR) (Mickey, Minnie, Pluto, Madroño…)
- **Model**: intermediate level in KR. It represents the organization of data structure. (Ontology of fiction animals and ontology of real animals)
- **Metamodel**: upper level in KR. It represents the structure model. (Ontology composed of concepts, relations, ...)

- ***Mapping***: Relation between elements of different set of resources: two ontologies, one ontology and one DB, etc.

Multilinguality in ontologies
Levels of representation

realization

information

modelling

Metadata:
Information about whether the KR is in one or several languages

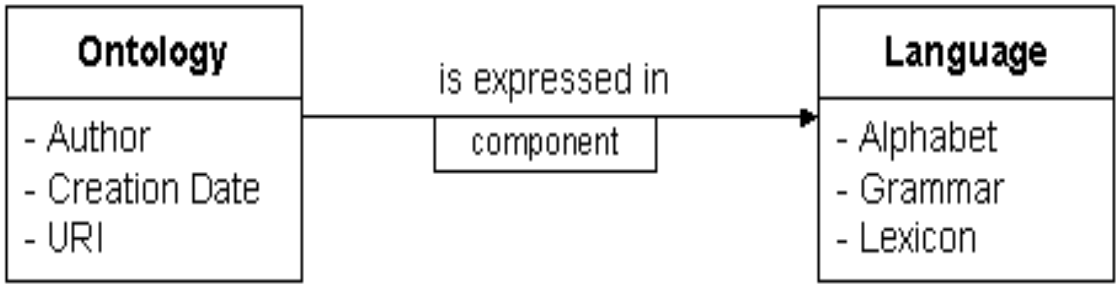
Ontology:
Linguistic realizations that show multilinguality in the KR

Metamodel:
Representing components that provide KR with multilinguality

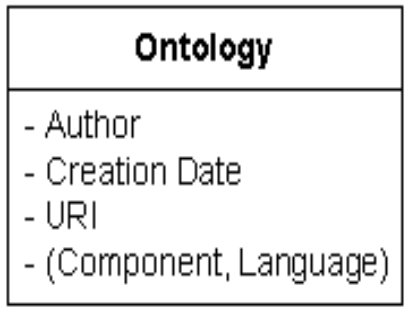
Multilinguality in ontologies

1. Information. Example

Standard: OMV (Ontology Metadata Vocabulary)

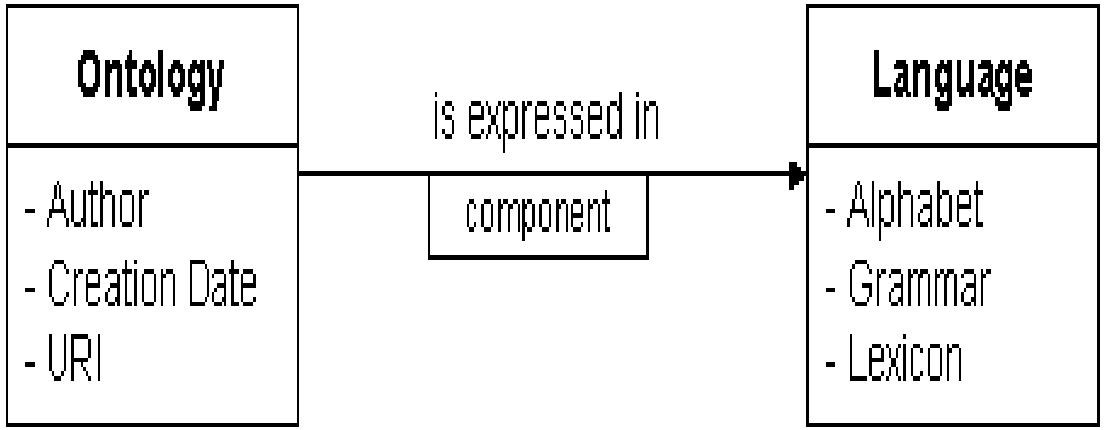


Option 1. Multilinguality by relation



Option 2. Multilinguality modifying the concept `ontology`

Advantages and disadvantages. Option 1



➢**Advantage: it is possible to include a certain amount of linguistic information**

➢**Disadvantage:**

- **Difficulty to instantiate the `language` concept with all information**
- **Few systems have relations with associated semantic information**

Advantages and disadvantages. Option 2

Ontology
- Author - Creation Date - URI - (Component, Language)

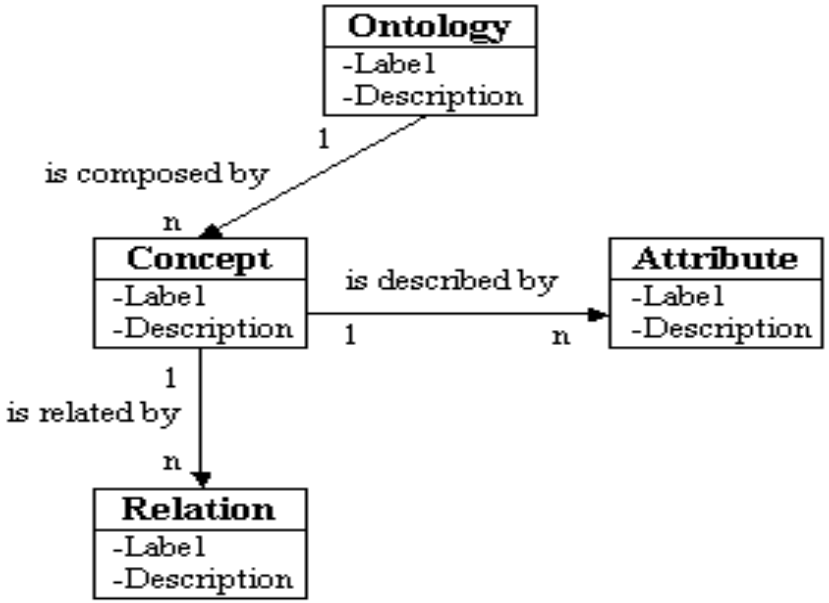
- **Advantage: It is easier to implement**
- **Disadvantage: A lot of linguistic information is lost**

Multilinguality in ontologies
2. Realization

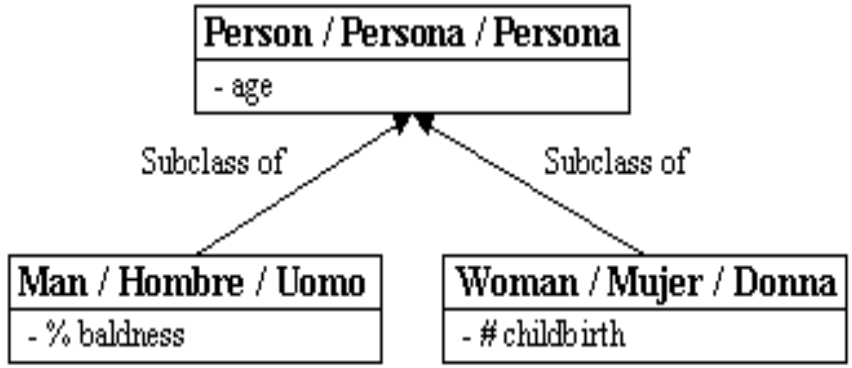
- Closely related to modelling
- It is the instantiation of the model
- Two possibilities:
  - Linguistic information inside the ontology
  - Linguistic information outside the ontology
    - Relational DB
    - Terminological DB
    - Multilingual lexicon
    - Multilingual thesaurus

2. Realization

Linguistic information inside the ontology (1)

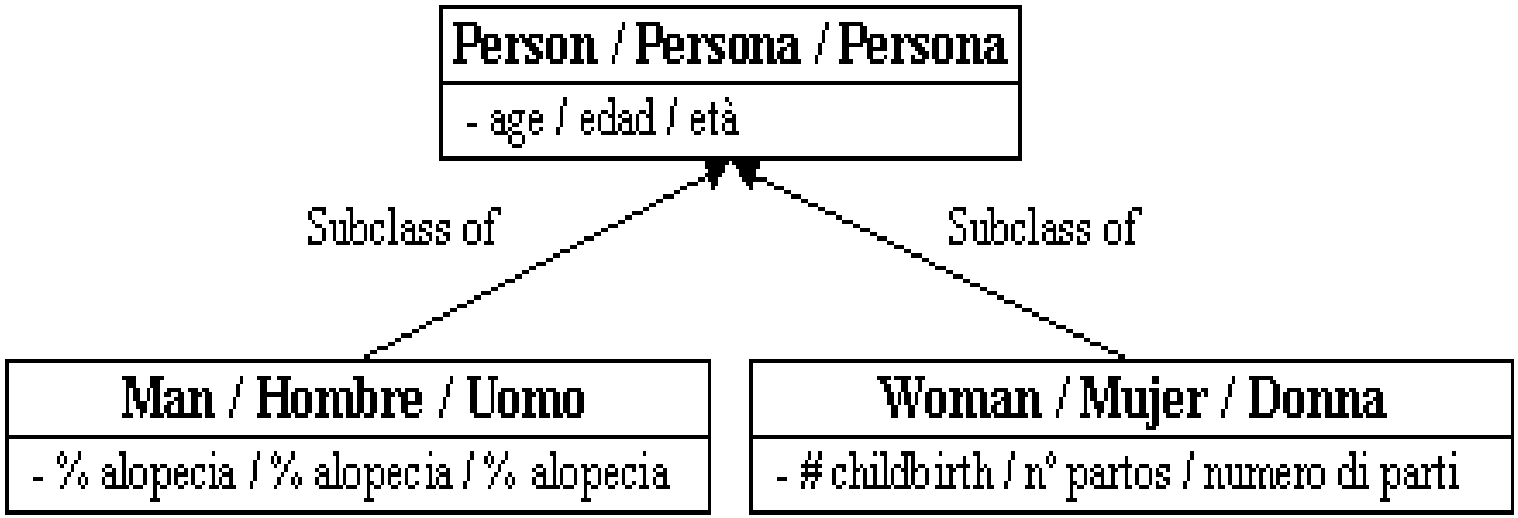


Ontology metamodel



Multilinguality in the ontology: concepts, not attributes

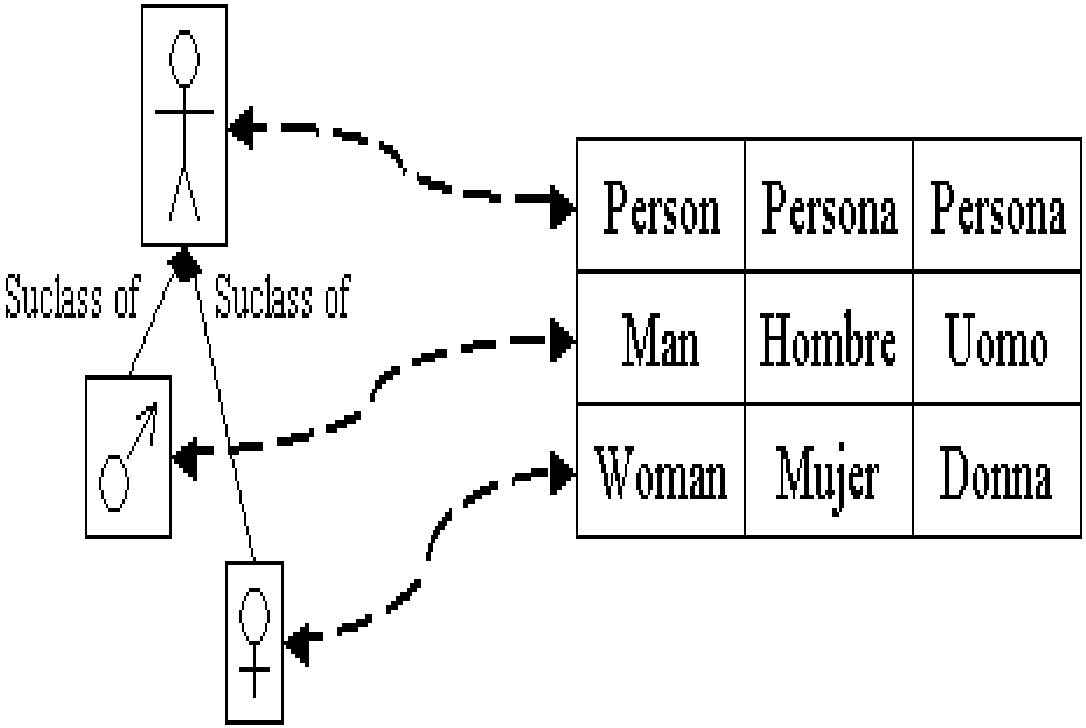
2. Realization. Linguistic information inside the ontology (2)



Same ontology metamodel

Multilinguality in attributes

2. Realization.
Linguistic information outside the ontology



Ontology metamodel, 'alingual' ontology model, linguistic resource model Genoma KB

Multilinguality in ontologies
3. Modelling

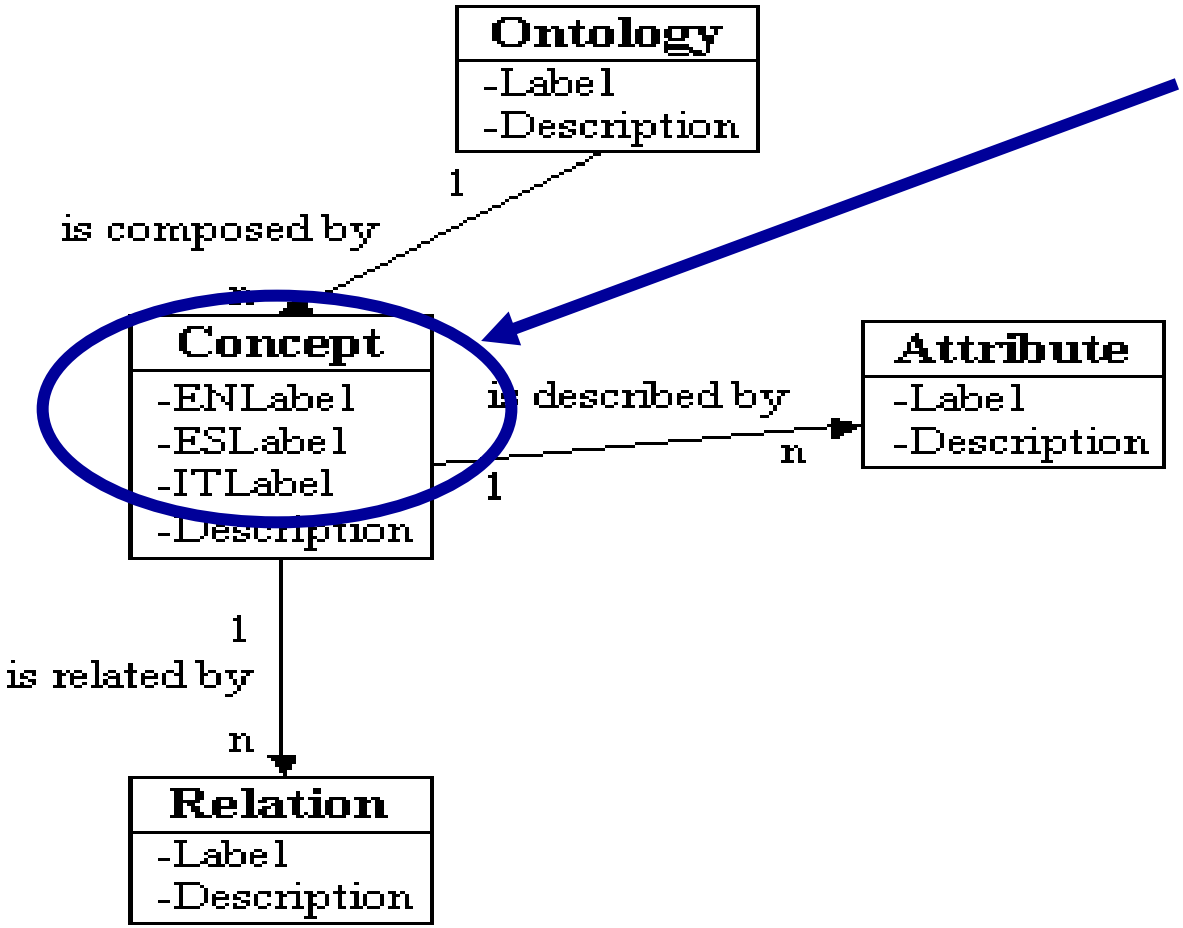
Three ways:

A. Including multilingual information in the ontology metamodel

B. Combining the ontology meta-model with a mapping model

C. Combining a linguistic information model and an ontology metamodel

3. Modelling:
A. Including multilingual information in the *ontology meta-model* (1)



- Multilinguality is limited to concepts
- By means of `rdfs:label` and `rdfs:comment` properties to define labels and descriptions in NL for classes.
- Localization at the terminological layer.
- It is the most widespread modelling modality.

A. Including multilingual information in the *ontology meta-model*

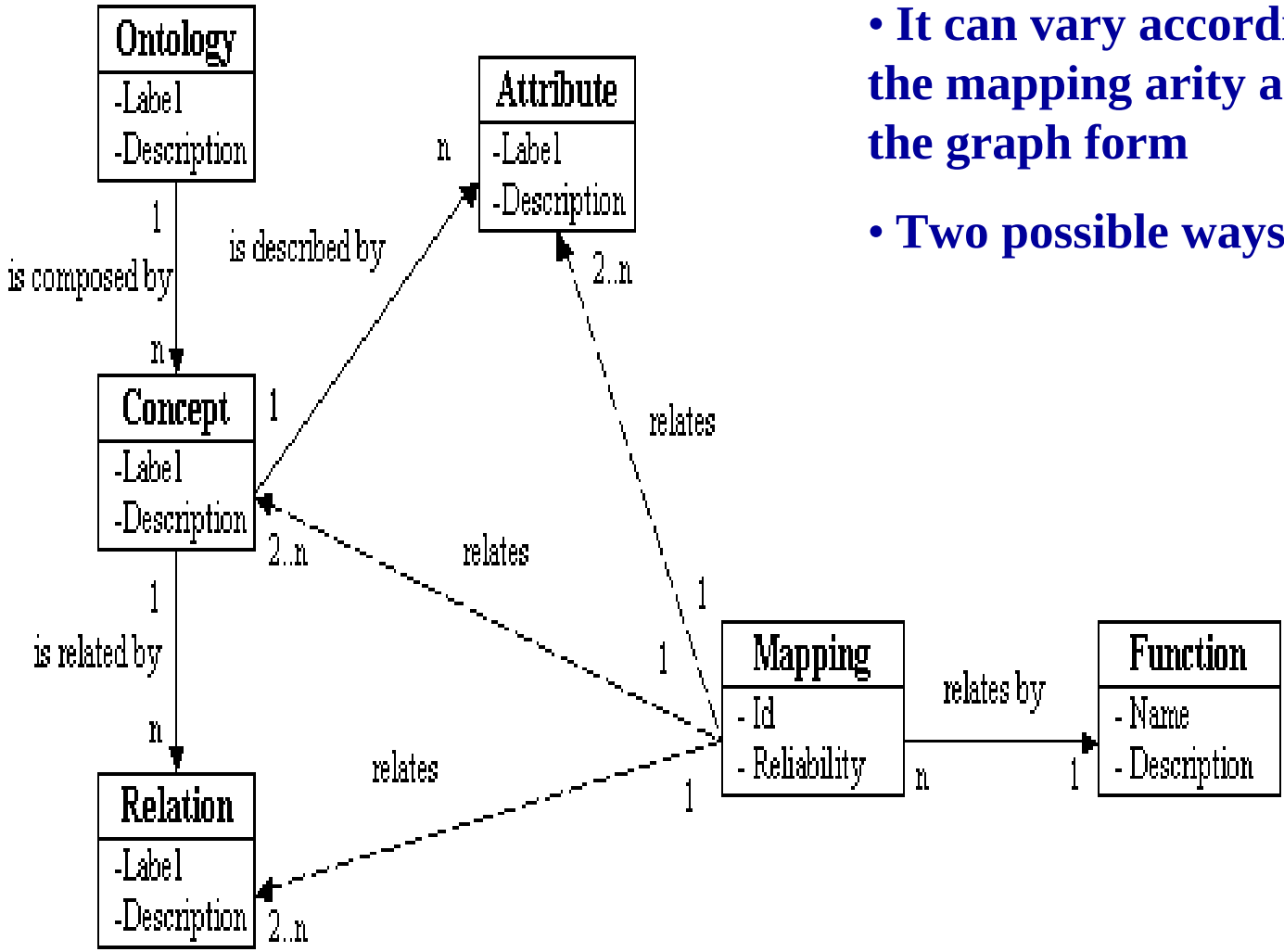
➢ Advantages:

- Increase of other languages is easily done by including just labels in the ontology.
- Suitable for highly specialized domain ontologies: knowledge shareable among different linguistic community experts.

➢ Disadvantages:

- Linguistic information included in the ontology is limited.
- Full synonym relation is assumed among labels in the different languages, but it is not true.

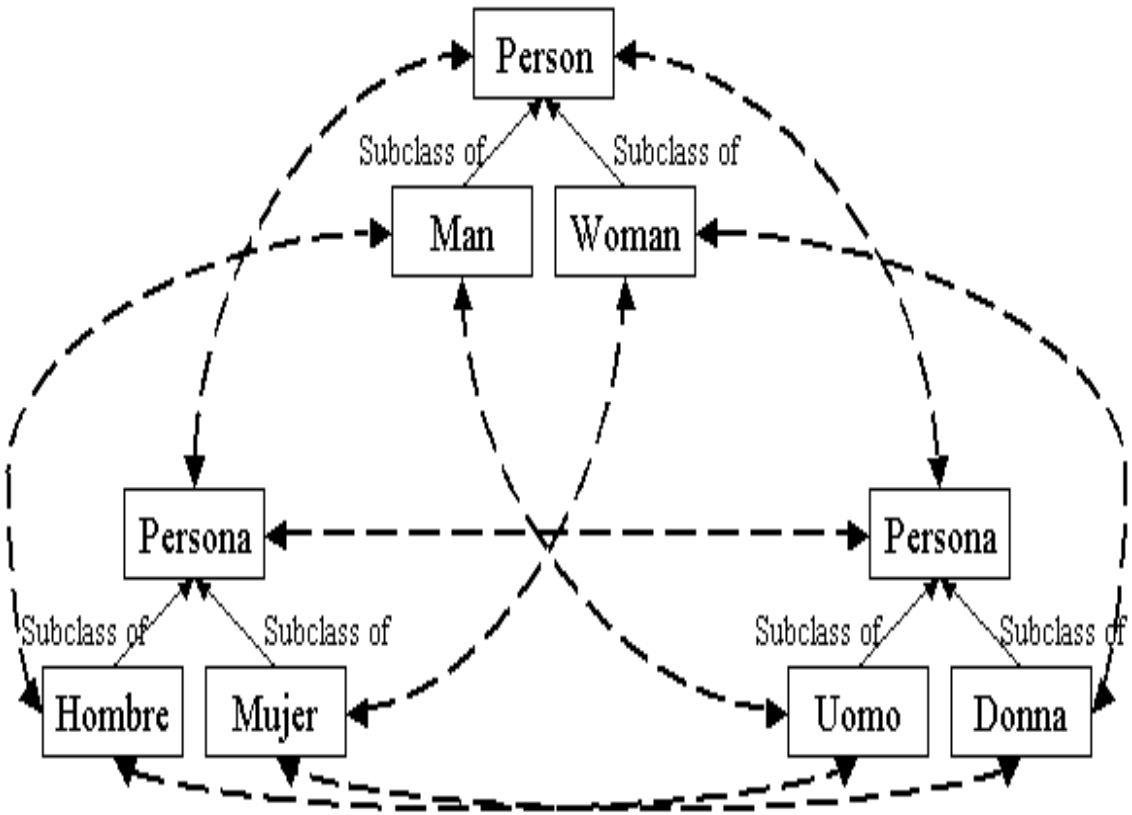
B. Ontology metamodel and a *mapping model:* Example



- **It can vary according to the mapping arity and the graph form**
- **Two possible ways**

B. Combining the *ontology meta-model* with a *mapping model*

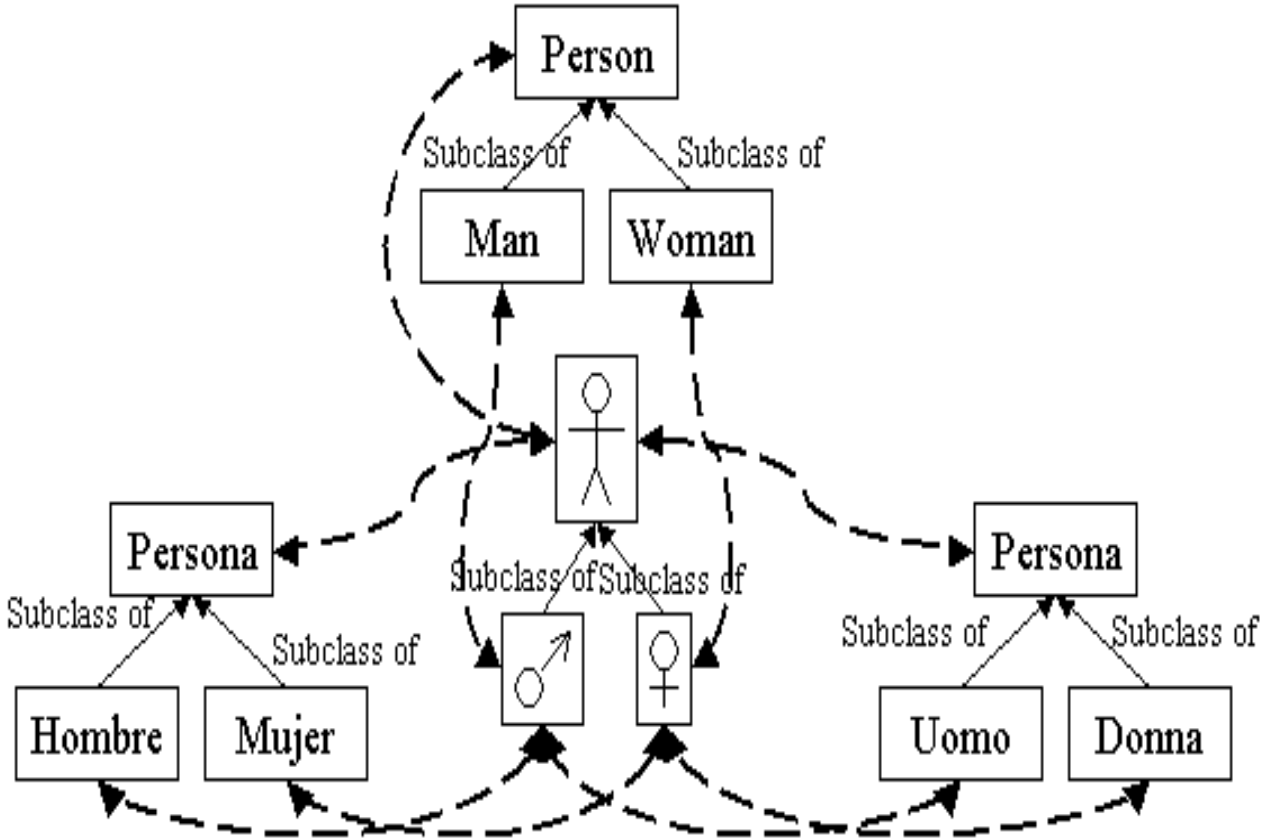
1. Binary mappings in an orthogonal graph.



- Localization at the conceptual layer.
- Each monolingual ontology structure knowledge is mapped to the rest of ontologies in a pair-wise way
- Less intuitive from the abstraction viewpoint

B. Combining the *ontology meta-model* with a *mapping model*

2. Binary mappings in a radial graph



- Localization at the conceptual layer.
- Monolingual ontologies mapped to each other through an interlingua: set of common concepts to establish equivalences
- EWN

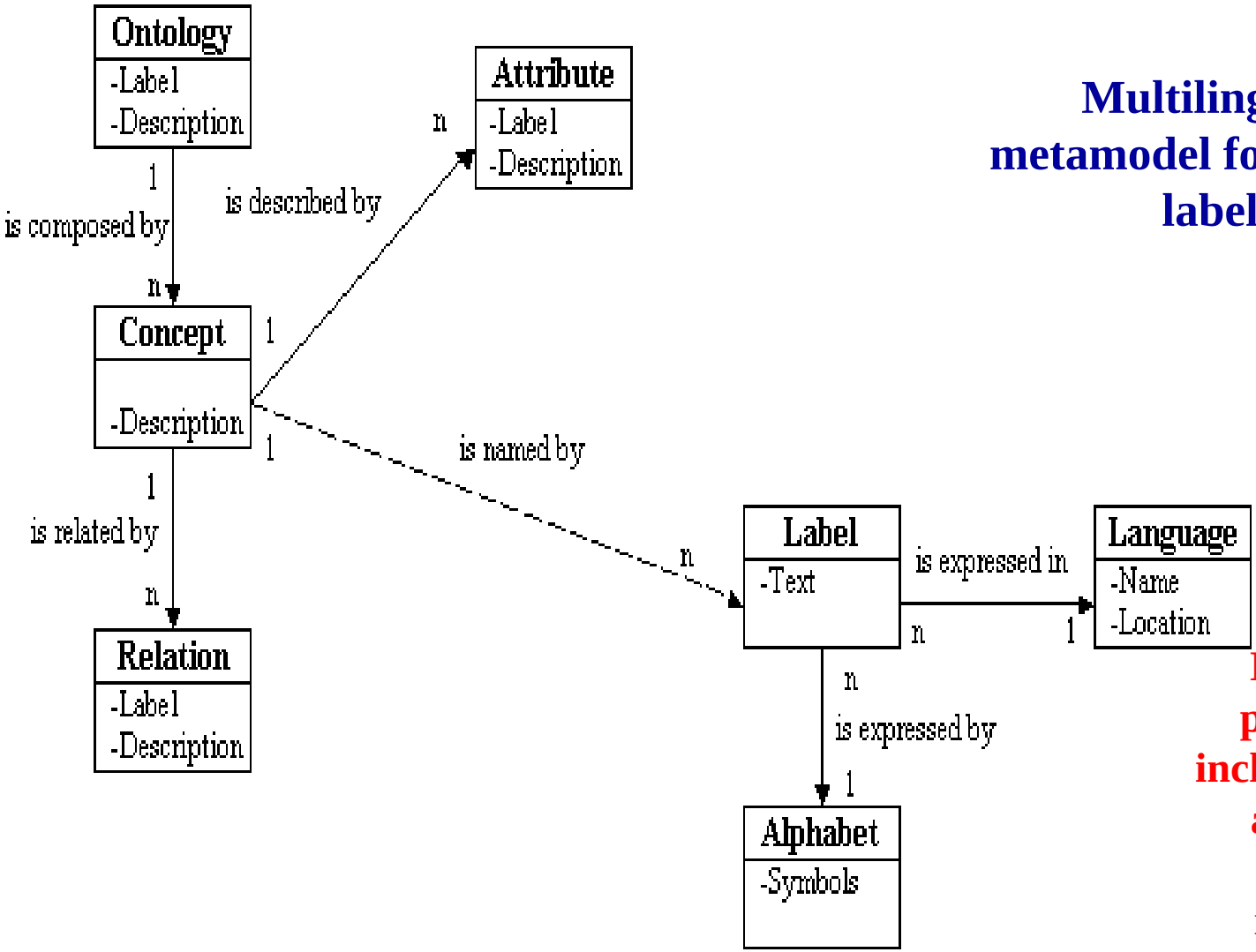
Advantages and disadvantages of combining the *ontology meta-model* with a *mapping model*

- **Advantages:**
  - Conceptualizations are maintained in each language
  - Suitable for ontologies highly dependent of a certain culture: the judiciary field.
- **Disadvantages:**
  - Huge effort needed to conceptualize the same domain in different natural languages.
  - Three types of expertise are required:
    - domain expertise,
    - linguistic expertise
    - ontology engineering expertise.

C. Associating the *ontology meta-model* to a *multilingual linguistic model*

- Localization at the terminological and conceptual layers
- Elements of the ontology are linked to multilingual data stored outside the ontology.
- Different ways for representing and organizing the linguistic information: DB (as in GENOMA-KB or OncoTerm), an ontology, etc.
- The ontology conceptualization layer can undergo modifications to meet localization needs, as the creation of language specific ontology modules,

C. *Ontology meta-model* linked to a *multilingual linguistic model* : Example



Multilinguality metamodel for concept labels

Increases the possibilities of including linguistic and ontology component information.

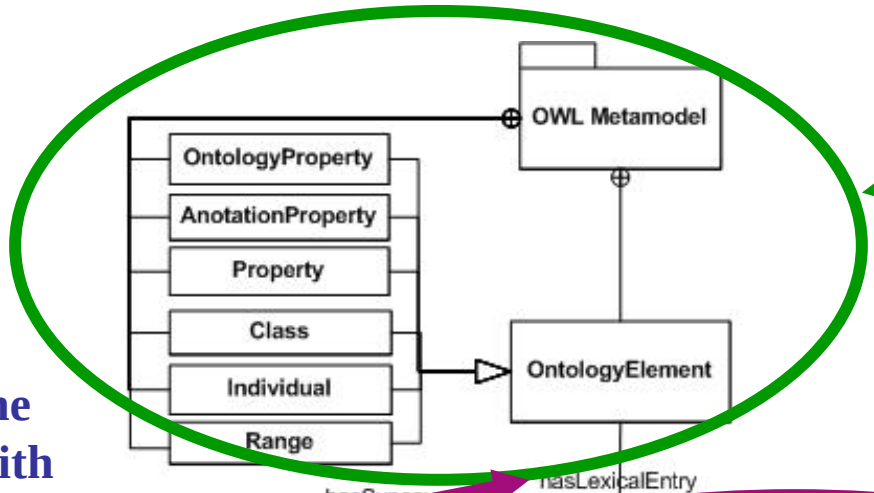
Advantages and disadvantages

- **Advantages**:
  - Including as much linguistic information as wished is possible
  - Linguistic elements within one language or across languages can be linked.
  - Nuances or differences between languages can be reported and formalized at the terminological layer
  - Relevant information as, e.g., the provenance of the linguistic elements, can also be included.
  - Ontology development expertise is not necessary for linguists and domain experts to access the terminological layer in a distributed environment.
- **Disadvantages**:
  - Some language specificities could be lost, unless captured in language specific ontology modules, i.e., in the conceptual layer, or in the linguistic model, i.e., at the terminological layer.

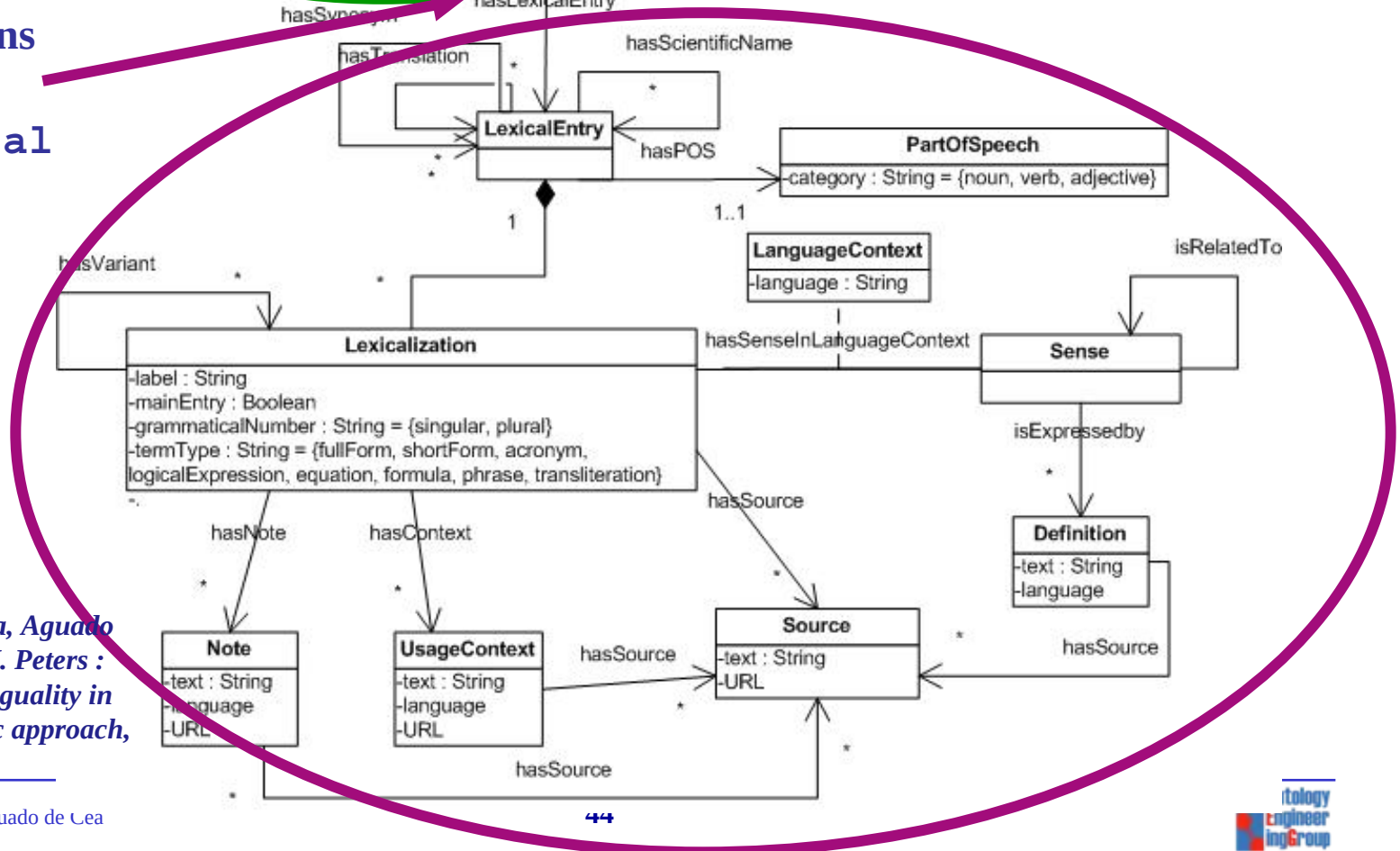
A new proposal
Linguistic Information Repository - LIR

- It is modelled as an ontology.
- The linguistic information captured in the LIR is organized around the `LexicalEntry` class.
- A lexical entry is a ternary relation: `Lexicalization`, `Sense` and `LanguageContext`.
- `Note` is linked to the `Lexicalization`, but it could be linked to any other class in the model to include supplemental information.
- By linking `Note` to the `Sense` or `Definition` classes, possible differences or nuances among senses in different languages can be made explicit.

Ontology metamodel in OWL



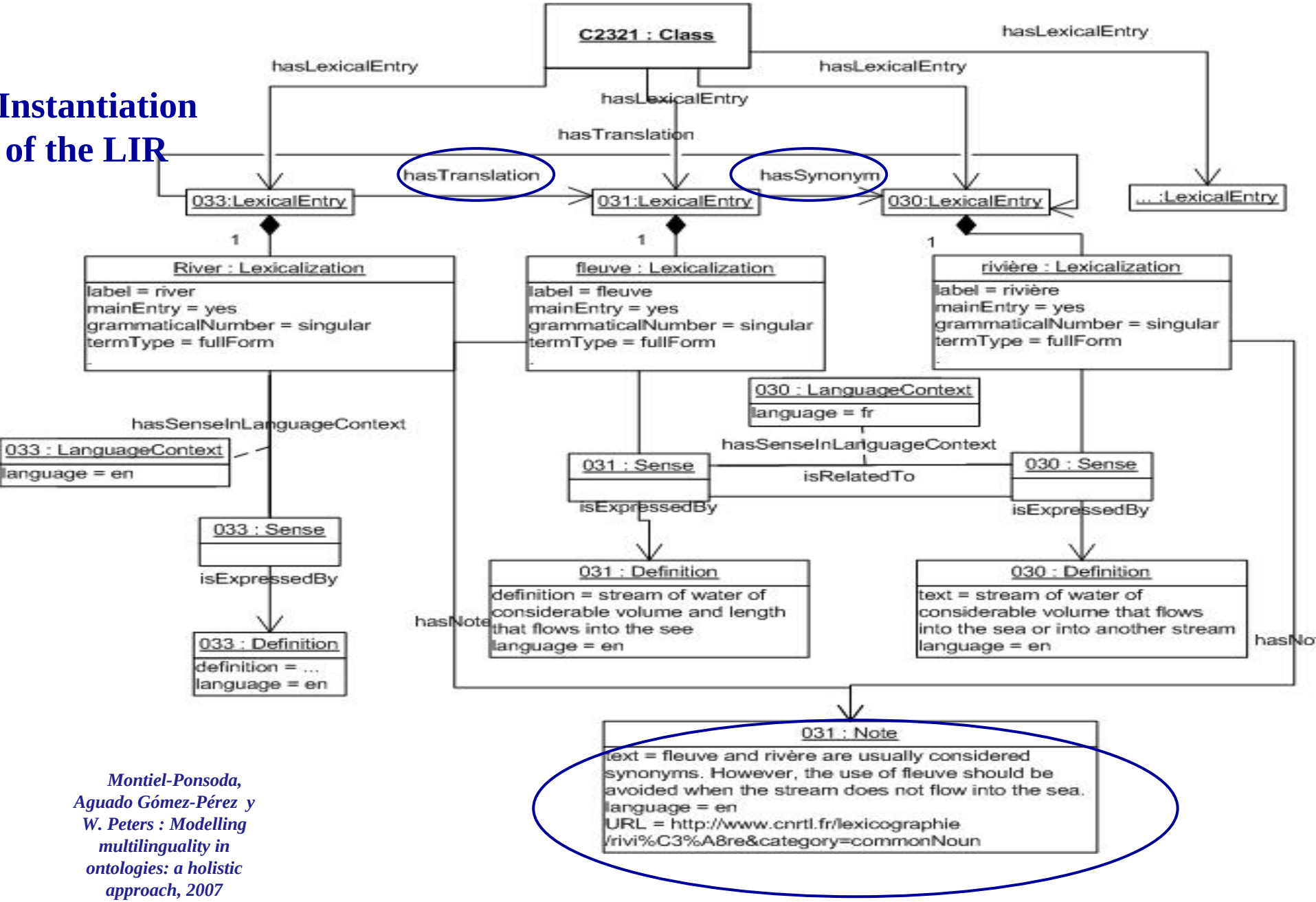
Linking the Ontology with LIR by means of the "hasLexicalEntry" relation



LIR

Montiel-Ponsoda, Aguado Gómez-Pérez y W. Peters : Modelling multilinguality in ontologies: a holistic approach, 2007

Instantiation of the LIR



Montiel-Ponsoda, Aguado Gómez-Pérez y W. Peters : Modelling multilinguality in ontologies: a holistic approach, 2007

Advantages of the LIR proposal

- Preserves the independence between the ontology and the linguistic layer.
- Links multilingual information with all ontology elements.
- By adopting linguistic standards for describing linguistic features helps to maintain language specificities
- Allows localization at the terminological and conceptual level.
- Facilitates interoperability and extensibility if more information is needed.
- Solves conceptualization mismatches
- Access to multilingual resources is possible thanks to certain tools: LabelTranslator.