

Ontologías y multilingualidad

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

Indice

- Multilingualidad: definición y finalidad
- Localización vs. internacionalización
- De los sistemas monolingües a los sistemas multilingües
- Sistemas de PLN con multilingualidad
- La multilingualidad en los SRC
- La multilingualidad en las ontologías
 - Información
 - Realización
 - Modelización
- Una nueva propuesta : Linguistic Information Repository- LIR

Multilingualidad ¿para qué?

- Necesidad de multilingualidad en los sistemas de PLN
 - Sistemas de búsqueda de respuestas
 - Búsqueda de información multilingüe
 - Recuperación de información
 - Traducción automática
- **Compartición** de conocimientos → ontologías
- **Reutilización** de conocimientos
- Ontologías → Web semántica

¿Cómo conseguir la multilingualidad?

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)
- En **economía**: “adaptar un producto a un entorno distinto del original (*a non-native environment*).
- En **software y diseño web**: adaptar el contenido, la lengua y el diseño a la cultura y la lengua de llegada
- En **ontologías**: **Ontology Localization** involves the process of adapting an ontology to a particular language and culture.

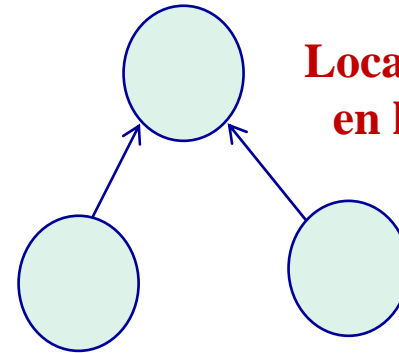
Internacionalización

- **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).
- Es importante:
 - Separar el texto del código fuente – > evita que los traductores cambien el código fuente
 - No se limita al software: *online help, documentation and web sites* pasan por este proceso
 - Para los escritores técnicos: “*writing for a global audience*”, “*web site globalization*”

De la localización de SW a la localización de ontologías

Localización de SW

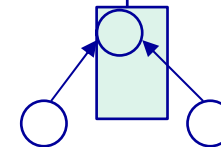
Internacionalización de SW



**Localización de ontologías
en los diferentes niveles**



Metamodelo



**Modelo de
ontología**

Semejanzas entre ambos procesos

- **Internacionalización:**
 - Contenido **léxico**: caracteres y símbolos que maneja el ordenador (ASCII encoding, UNICODE, etc.)
 - Contenido **gramatical**: caracteres, estructuras sintácticas y símbolos utilizados en determinados lenguajes de ontologías (RDF(S), OWL)
 - Paradigma de representación del conocimiento: marcos, redes semánticas, LD, (ontologías)
- **Localización:**
 - Contenido **léxico-terminológico**: términos o palabras que sirven para denominar los elementos de la ontología
 - Contenido **conceptual**: en cuanto a decisiones de conceptualización, como la granularidad, expresividad, perspectiva, etc. Especialmente en ontologías de dominio.
 - Contenido **pragmático**: resultado final del modelo (GUI, etc.)

De la monolingualidad a la multilingüidad

- Pocas ontologías multilingües
 - <http://olp.dfki.de/ontoselect/>
 - 1652 ontologías
 - 149 con algún tipo de información lingüística
 - 130 en inglés, 10 en español
 - 5: en-es, 4: en-es-fr
- Poca información disponible sobre representación de multilingüidad
- Reciente interés en los grupos de investigación internacionales:
- 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*

Sistemas que incorporan multilingüalidad y ontologías: EWN

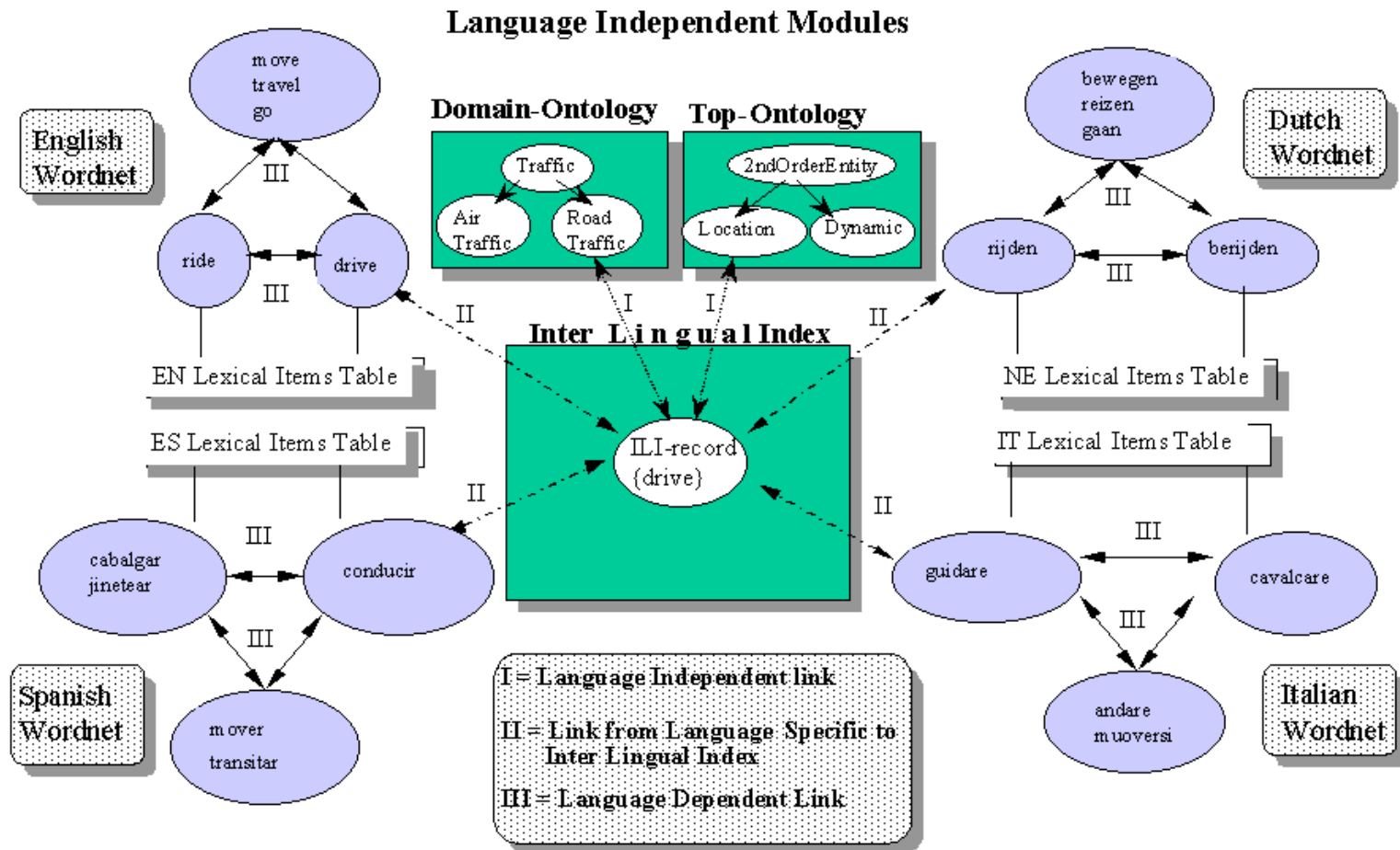
➤ EuroWordNet

- Basado en Wordnet, <http://wordnet.princeton.edu/perl/webwn>

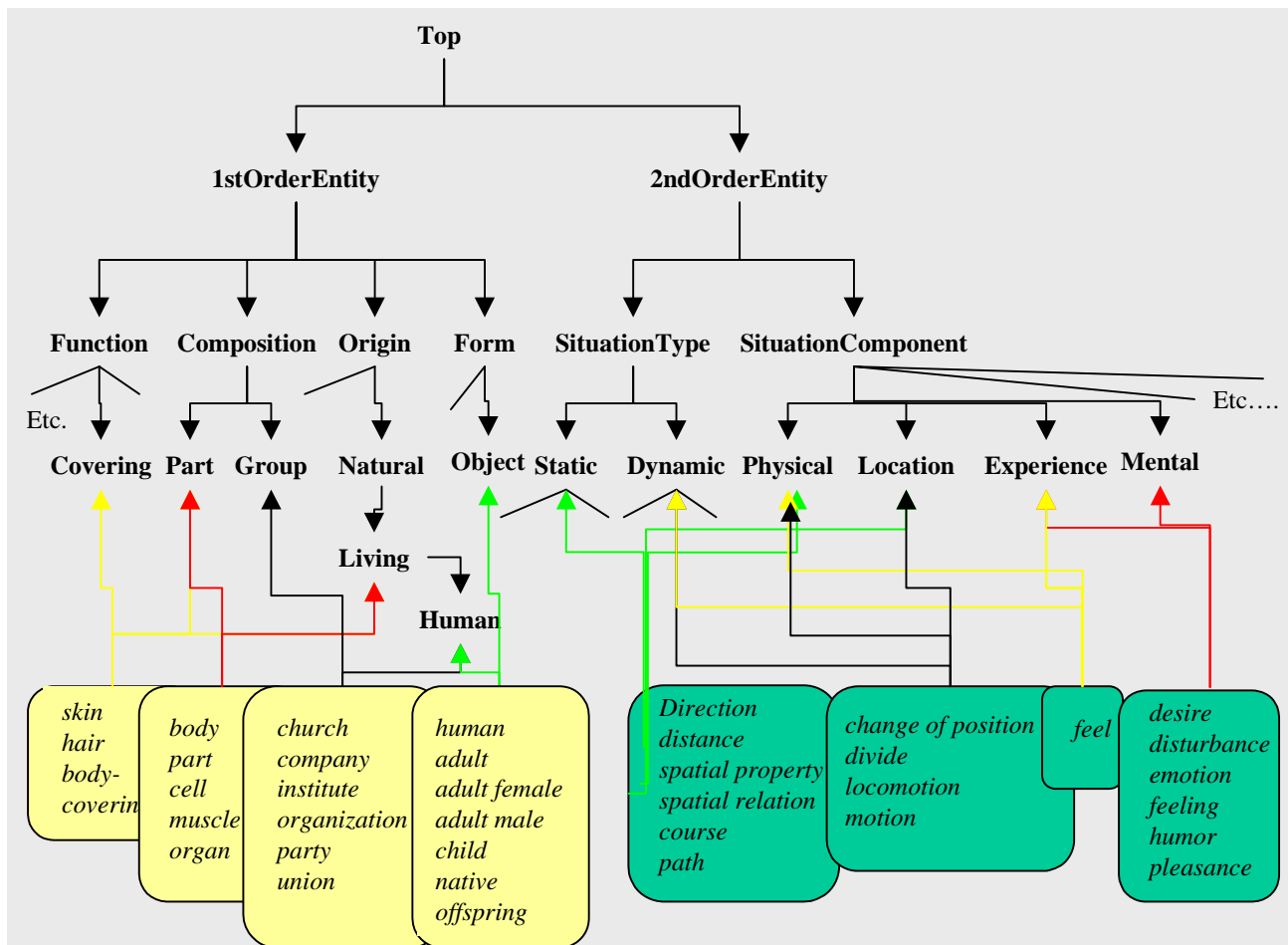
➤ Objetivos:

- Crear una BD léxica **multilingüe** para holandés, italiano, español, inglés, alemán, francés, estonio y checo
 - Mantener las relaciones específicas de las lenguas en sus redes
 - Lograr la máxima compatibilidad entre los distintos recursos
 - Construir las redes de forma independiente, reutilizando recursos propios de cada lengua
 - Redes de palabras: **wordnets** -> ontologías autónomas monolingües, conectadas mediante un ILI
 - **Synsets** (sinónimos)
-
- **Instituciones: 8 universidades (UNED, UPC), 3 empresas.**
 - **Financiado por la U.E.**

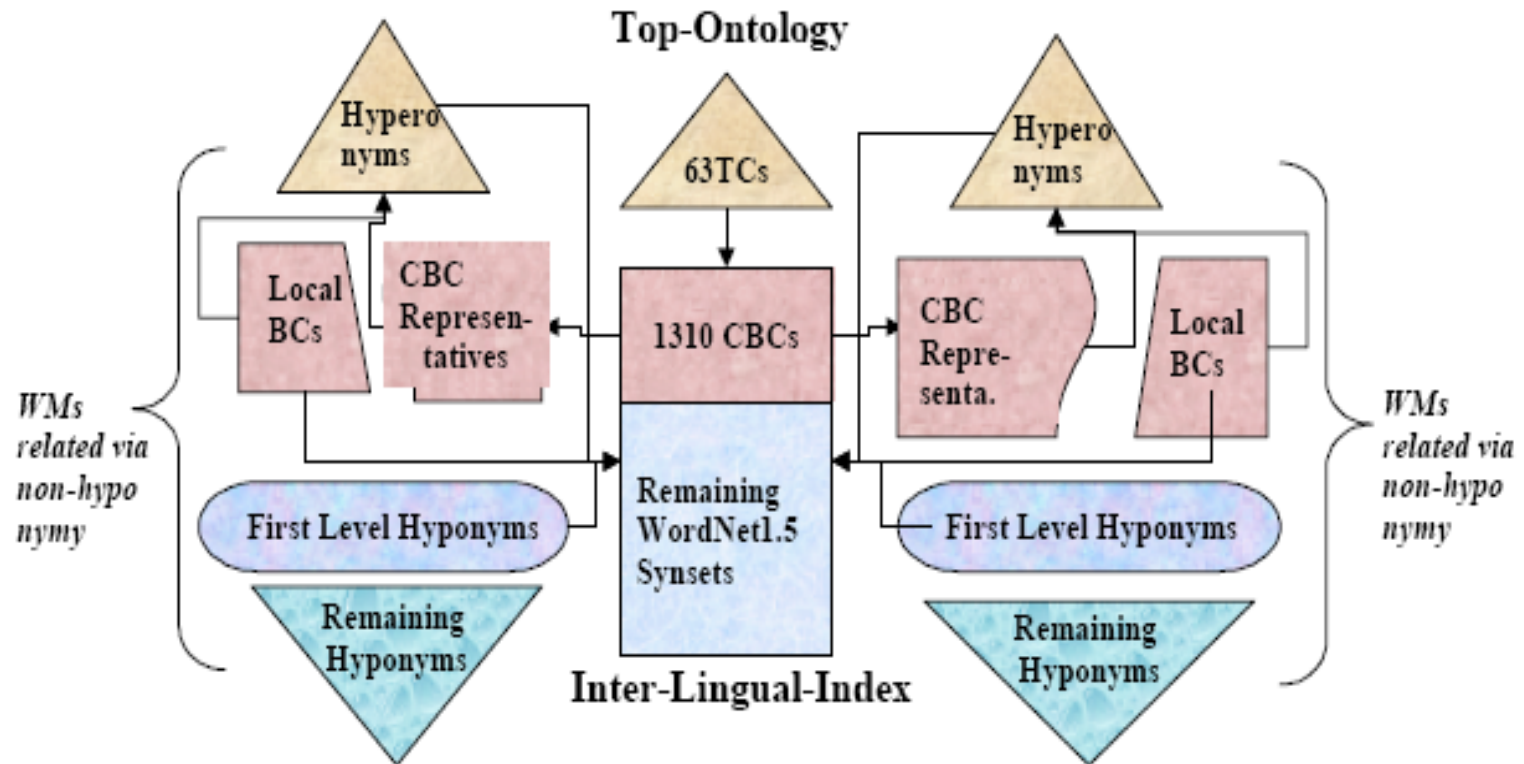
EuroWordNet. Arquitectura



EuroWordNet: Top Ontology



Esquema general de dos *wordnets* mapeados al ILI Eurowordnet



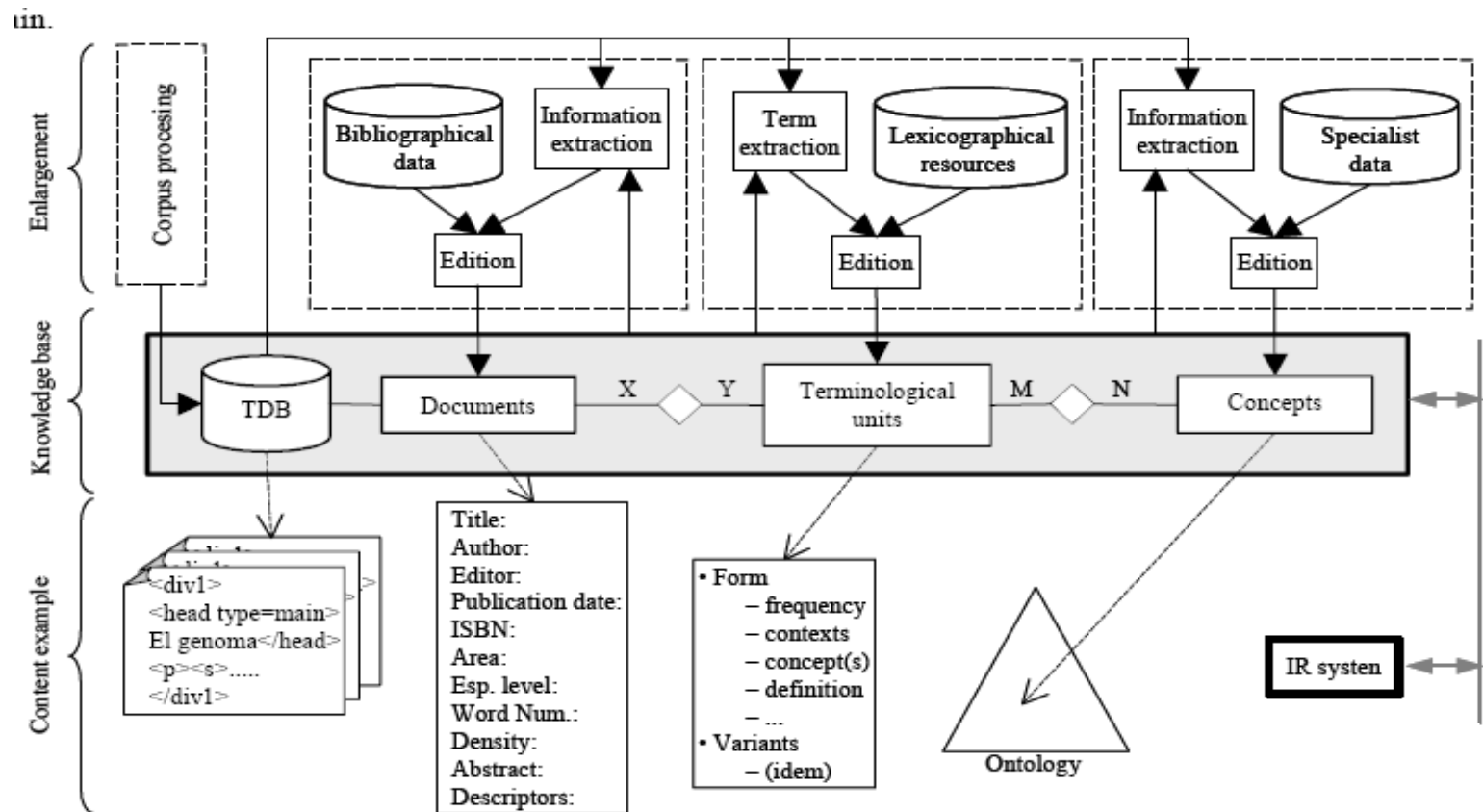
(Vossen, 2002)

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

Sistemas que incorporan multilingüalidad y ontologías: Genoma-KB

- **Módulo ontológico:** (MikroKosmos)
 - 21 conceptos básicos ALL, OBJECT physical, mental, social), EVENT (physical, mental, social), PROPERTY (attribute, relation), etc.
 - 100 conceptos propuestos por expertos del dominio
 - Las relaciones descritas (Feliu 2004) son:
 - Similaridad, Hiponimia, Secuencialidad (lugar y tiempo)
 - Causalidad, Instrumentalidad, Meronimia, Asociación
- **Módulo terminológico**
 - Multilingüalidad, POS, contexto, fuentes, lema, información administrativa
- **Módulo de corpus:** textos multilingües
- **Módulo de entidades:**
 - Módulo bibliográfico: refer. completas de textos y términos
 - Módulo factográfico: centros de invest. , personas, instituciones,

Arquitectura de la base de conocimiento GENOMA-KB

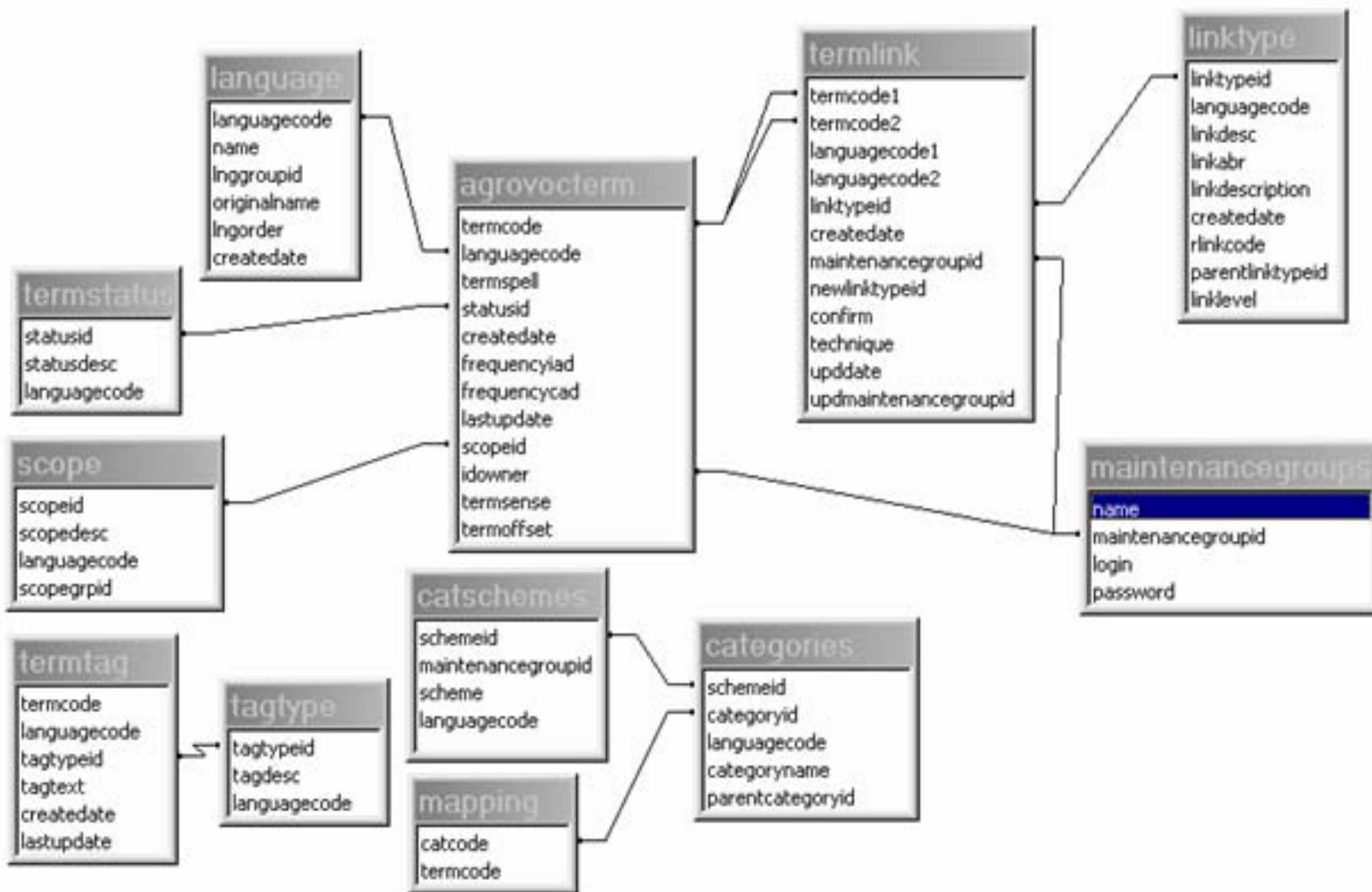


(Feliu, Vivaldi y Cabré, 2002)

Sistemas que incorporan multilingüalidad y ontologías: AGROVOC

- AGROVOC Thesaurus desarrollado por la FAO (*Food and Agriculture Organization*) y la UE en 1980/1982.
- 3 lenguas iniciales. Actualmente 17. Se incorporarán algunas más.
- Se define como “*a multilingual structured and controlled vocabulary*”.
- Utilizado para indexar y recuperar datos sobre pesca y alimentación.
- Basado en modelización UML (Unified Modelling Language)
- Muestra el nº de términos en tiempo real (41,580 términos en español)
 - URL http://www.fao.org/aims/ag_figures.jsp

AGROVOC: representación de la información multilingüe



Multilingualidad en los SBC (1)

- La multilingualidad puede darse en tres niveles:

1. Interfaz

a) Mensajes

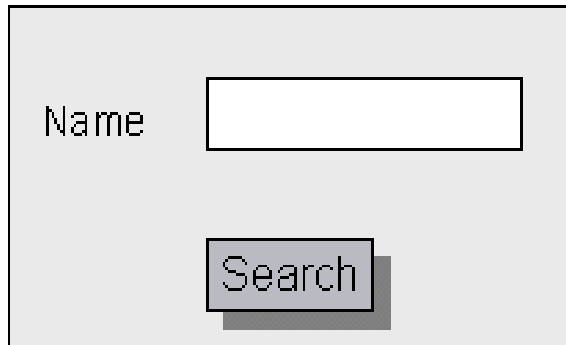
b) Contenido

2. Datos

3. Representación de conocimiento

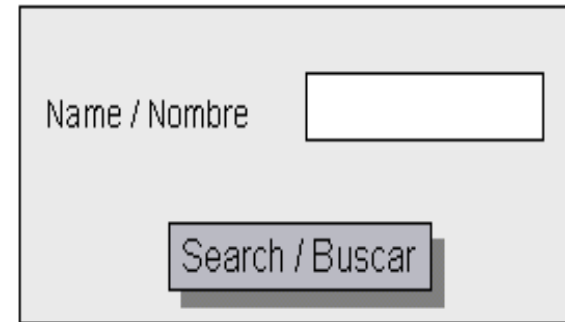
- *Aguado de Cea, G., Montiel Ponsoda, E., Ramos Gargantilla, J.A. “Multilingualidad en una aplicación basada en el conocimiento”, Procesamiento del lenguaje natural, n° 38, Abril 2007*

1. Interfaz: (a) visualización de mensajes



A diagram of a monolingual user interface. It consists of a light gray rectangular box. Inside, the word "Name" is positioned to the left of a white rectangular input field. Below the input field is a gray rectangular button with the word "Search" in white text.

1. Mensajes monolingües



A diagram of a simultaneous multilingual user interface. It consists of a light gray rectangular box. Inside, the text "Name / Nombre" is positioned to the left of a white rectangular input field. Below the input field is a gray rectangular button with the text "Search / Buscar" in white text.

2. Mensajes multilingües simultáneos



A diagram of a non-simultaneous multilingual user interface. It consists of a light gray rectangular box. Inside, the word "Name" is positioned to the left of a white rectangular input field. Below the input field are two small flag icons: the United Kingdom flag (Union Jack) and the Spanish flag. To the right of the flags is a gray rectangular button with the word "Search" in white text.

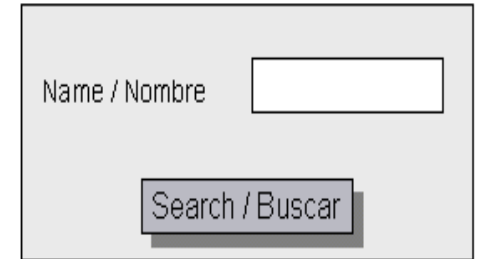


A diagram of a non-simultaneous multilingual user interface. It consists of a light gray rectangular box. Inside, the word "Nombre" is positioned to the left of a white rectangular input field. Below the input field are two small flag icons: the United Kingdom flag (Union Jack) and the Spanish flag. To the right of the flags is a gray rectangular button with the word "Buscar" in white text.

3. Mensajes multilingües no simultáneos

Ventajas y desventajas de la opción (a).

- En la **visualización simultánea**, la incorporación de otras lenguas requiere **modificar** código de visualización.



A screenshot of a web interface for simultaneous visualization. It features a single text input field labeled "Name / Nombre" and a single button labeled "Search / Buscar".

- En la **visualización no simultánea**, no implica modificar todo el código, sino ampliar el nº de interfaces y modificar las opciones de selección.



A screenshot of a web interface for non-simultaneous visualization. It features a text input field labeled "Name", two language selection buttons (UK flag and Spanish flag), and a button labeled "Search".



A screenshot of a web interface for non-simultaneous visualization. It features a text input field labeled "Nombre", two language selection buttons (UK flag and Spanish flag), and a button labeled "Buscar".

1. (b). Visualización del **contenido** de forma multilingüe

- Cuando la BC es **multilingüe**
 - La aplicación consulta a la BC
 - La interfaz muestra el contenido en el idioma seleccionado
- Cuando la BC es **monolingüe**
 - La aplicación consulta a la BC
 - Se utiliza un sistema de traducción (**recurso multilingüe**)
 - La interfaz muestra la traducción
- Interfaz similar en ambos casos a visualización de mensajes, **PERO** el tiempo de respuesta varía si la BC es multilingüe

Ventajas y desventajas de la opción (b)

➤ BC es **multilingüe**

- Tiempo de obtención de contenidos = tiempo de respuesta (TR) de la BC
- Razón: se ha conferido multilingualidad a la BC en **tiempo de diseño**
- Desambiguación: **en tiempo de diseño**

➤ BC es **monolingüe**

- Tiempo de obtención de contenidos = TR de BC + TR del recurso multilingüe
- La traducción se realiza en **tiempo de ejecución**
- Desambiguación: puede alargar el TR

Multilingualidad en los SBC (2)

- La multilingualidad puede darse en tres niveles:

1. Interfaz

- a) Mensajes
- b) Contenido

2. Datos

3. Representación de conocimiento

Multilingualidad en los datos de los SBC

Knowledge Representation

Article
- Title - Authors - Date - Journal - Language - PDF File

Instances

Article01
- WebODE in a Nutshell - Gómez-Pérez et al. - 2003 - AI Magazine - English - WebODE.pdf

Article02
- Estudio y formalización... - Fernández-López et al. - 2006 - RIIA - Español - Estudio.pdf

Knowledge Representation

Man
- First Name - City - Language

Instances

Man01
- Peter - London - English

Man02
- Pedro - Madrid - Español

Man03
- Pietro - Roma - Italiano

- La información sobre los individuos es multilingüe

- La multilingualidad se tratará como otro carácter más del dominio que se va a modelar

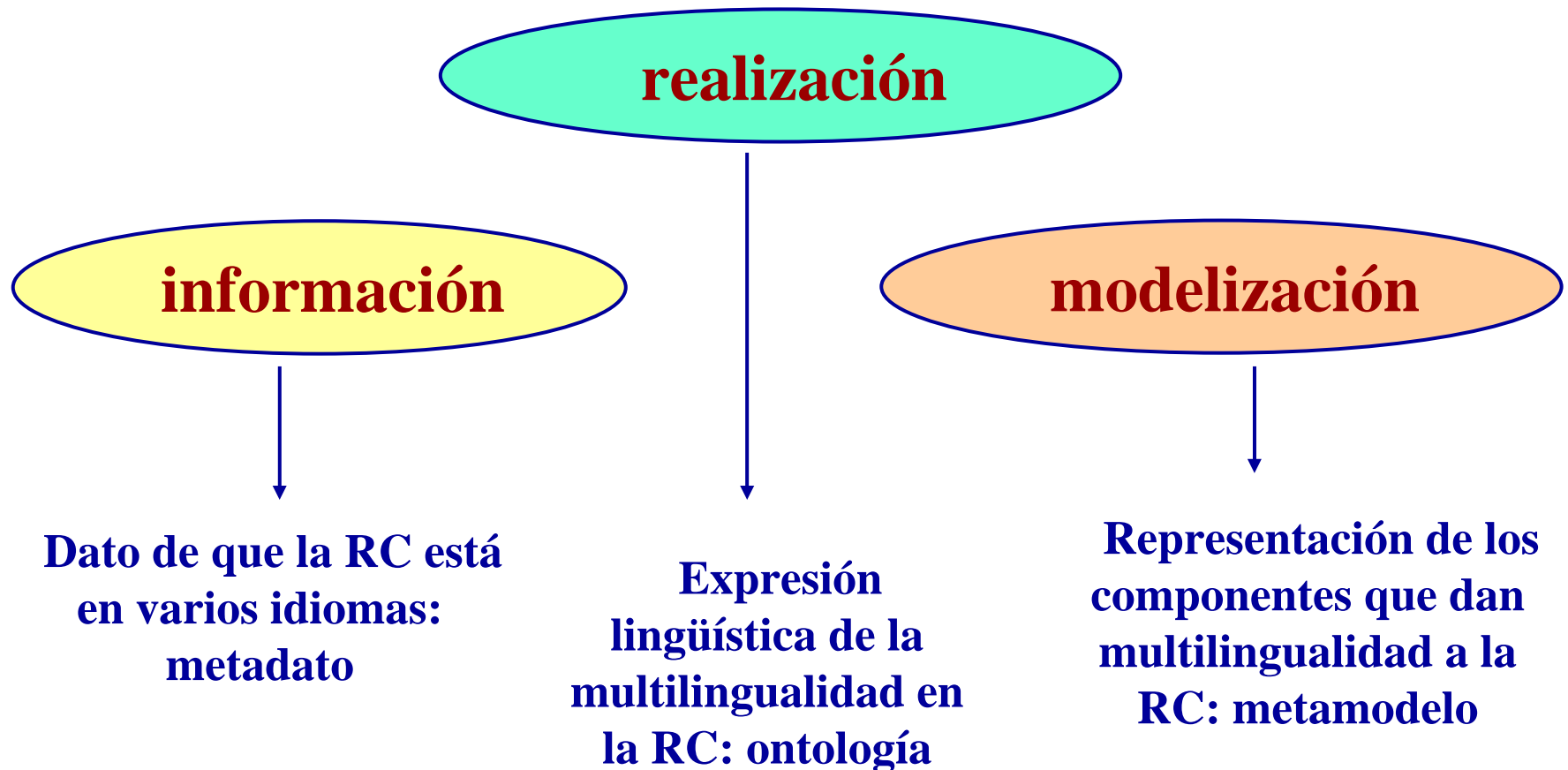
Datos multilingües en una RC monolingüe que considera la característica Language

Multilingüidad en los SBC

3. Representación del conocimiento

- **Datos:** instancias o individuos, nivel inferior de la RC (Mickey, Minnie, Pluto, Madroño...)
- **Modelo:** nivel intermedio y representa la estructura de los datos. (Ontología de Animales de ficción y Ontología de Animales Reales...)
- **Metamodelo:** nivel superior y representa la estructura del modelo. (Ontología que se compone de conceptos, relaciones...)
- **Mapping:** Relación entre elementos de conjuntos diferentes: dos ontologías, una ontología y una BD, etc.

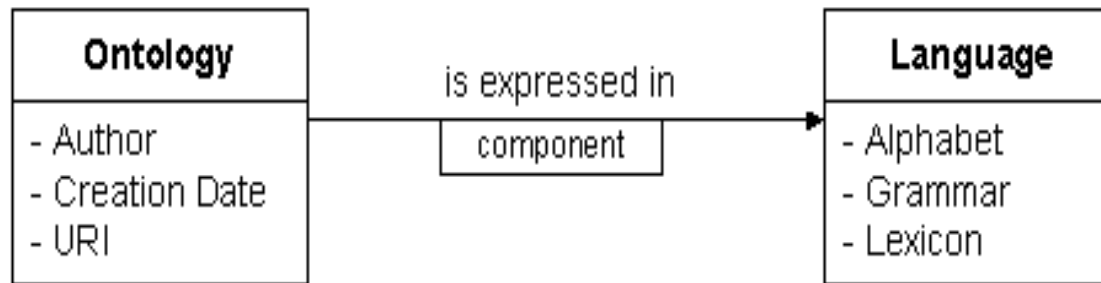
Multilingüalidad en Representación del conocimiento: ontologías



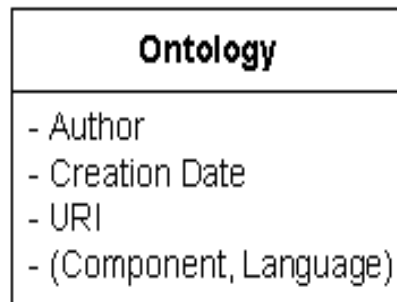
Multilingualidad en las ontologías

1. Información. Ejemplo

Estándar: OMV (Ontology Metadata Vocabulary)

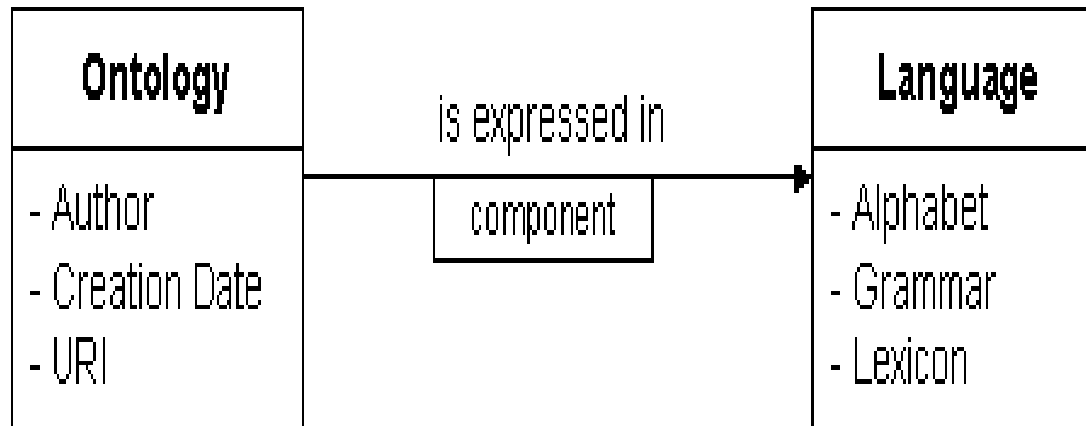


Opción 1. Multilingualidad mediante relación



Opción 2. Multilingualidad modificando los metadatos del concepto ontology

Ventajas y desventajas. Opción 1



➤ Desventaja:

- Dificultad de instanciar el concepto **language** con toda la información
- Pocos sistemas tienen relaciones con información semántica asociada

➤ Ventaja: Riqueza de información lingüística

Ventajas y desventajas. Opción 2

Ontology
<ul style="list-style-type: none">- Author- Creation Date- URI- (Component, Language)

➤ **Desventaja:** se pierde información lingüística

➤ **Ventaja:** es más sencilla, más fácil de implementar

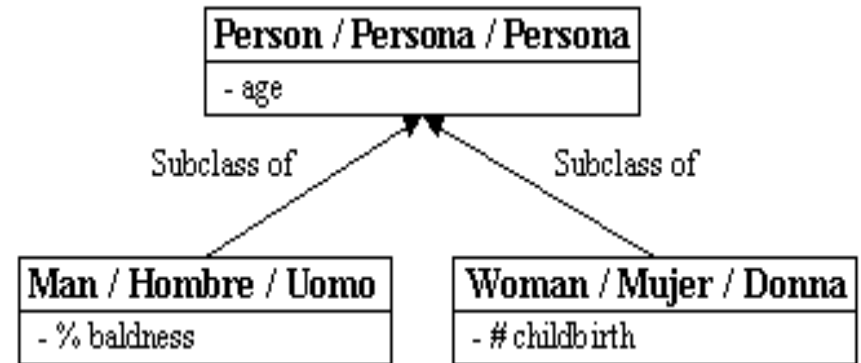
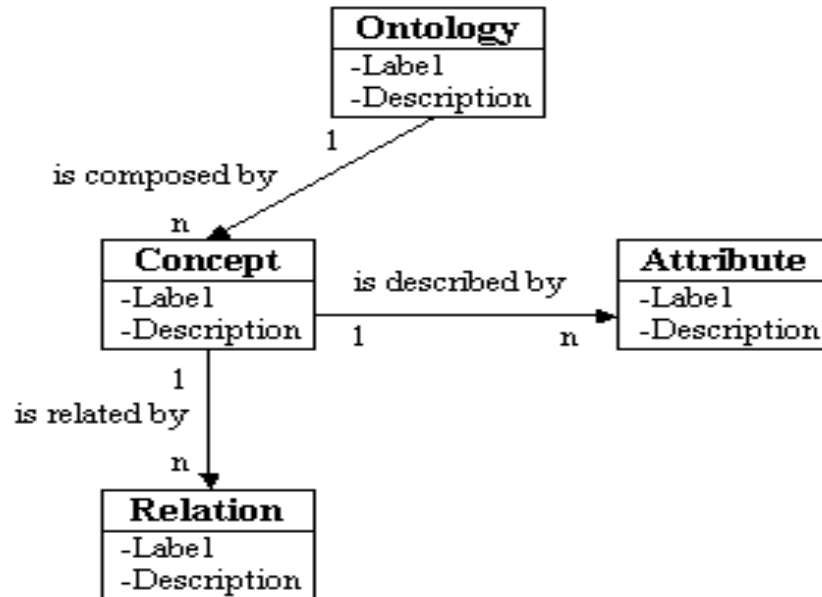
Multilingüidad en ontologías

2. Realización

- Estrechamente ligada a modelización
- La realización es la instanciación del modelo
- Pueden darse dos opciones:
 - Información lingüística **dentro** de la ontología
 - Información lingüística **fuera** de la ontología
 - BD relacional
 - Base terminológica
 - Lexicón multilingüe
 - Tesauro multilingüe

2. Realización.

Información lingüística **dentro** de la ontología (1)

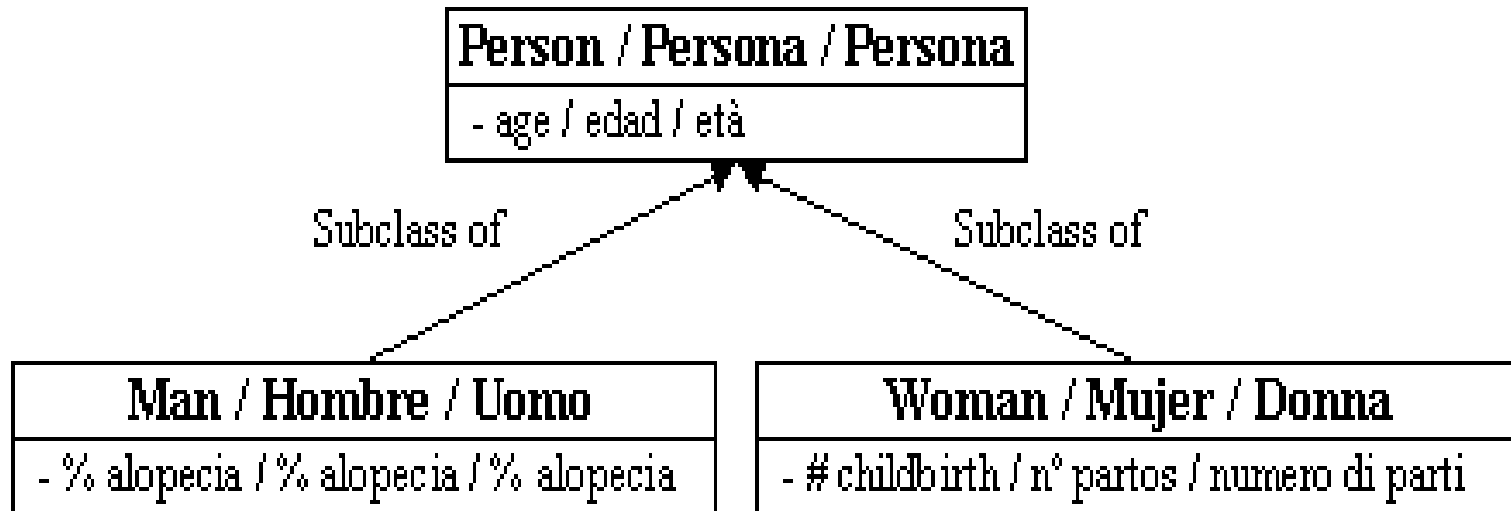


**Multilingualidad en la ontología:
conceptos, no atributos**

**Metamodelo de
ontología**

2. Realización.

Información lingüística **dentro** de la ontología (2)

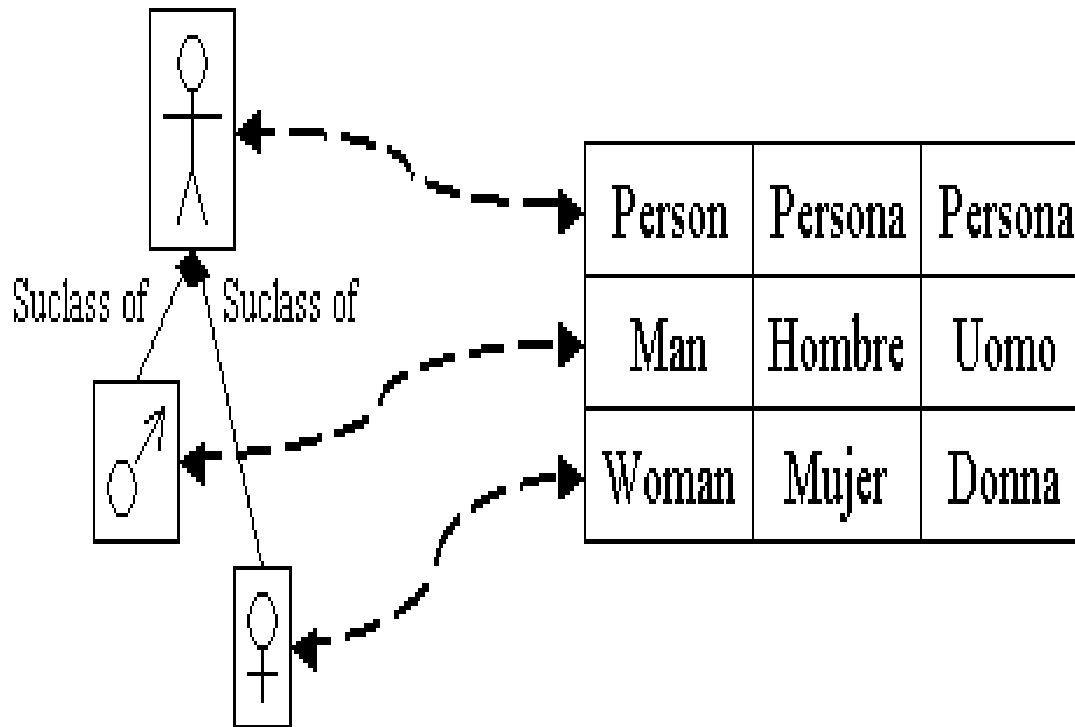


Mismo metamodelo de ontología

Multilingualidad en atributos

2. Realización.

Información lingüística **fuera** de la ontología (1)



**Metamodelo de
multilingualidad con
metamodelo de ontología
“alingüe” y modelo de
recurso lingüístico**

Genoma KB

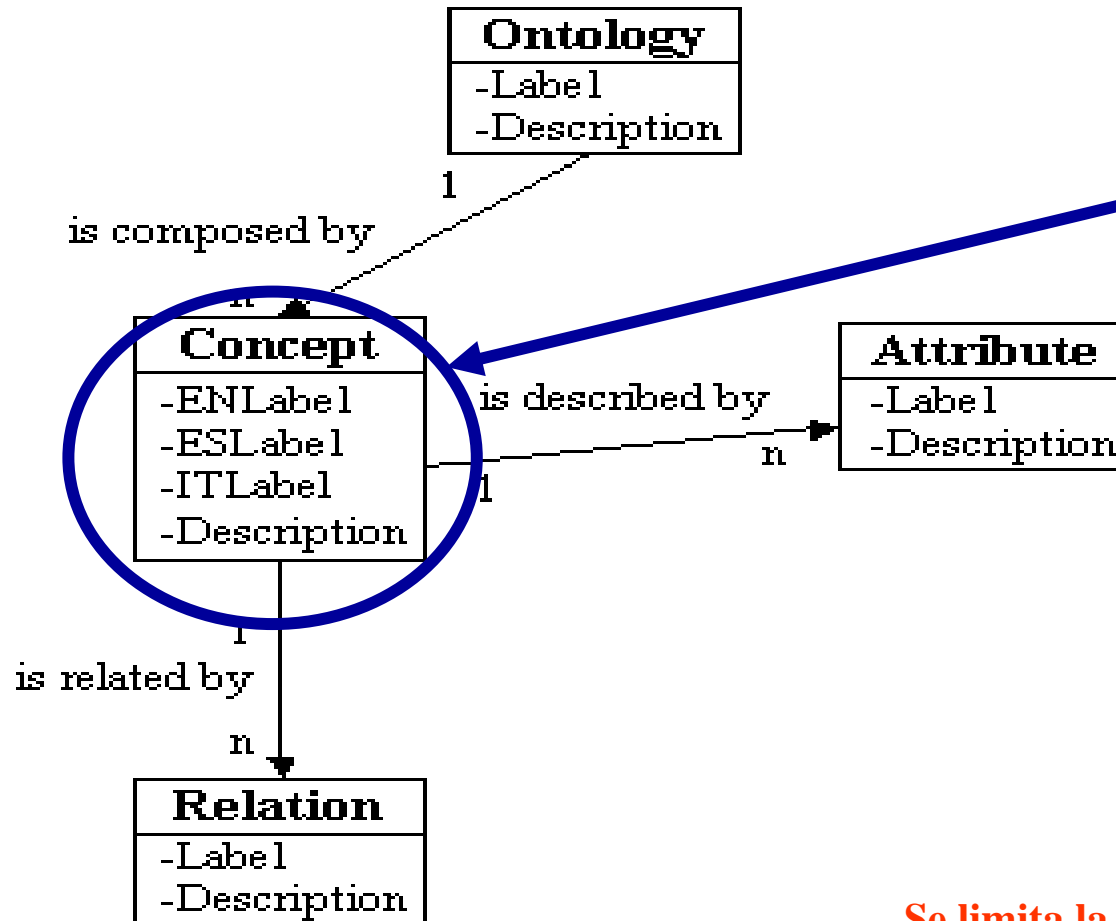
3. Modelización

Tres opciones:

- A. **Ampliar** el metamodelo de ontologías con información lingüística
- B. **Agregar** un modelo de información lingüística y relacionarlo con metamodelo de ontologías
- C. Utilizar *mappings* para relacionar ontologías monolingües

3. Modelización.

Ampliación de metamodelo con info lingüística Opción A. (1)



- Multilingualidad en conceptos mediante las propiedades para definir etiquetas y descripciones
Rdfs:label
Rdfs:comment
- Localización a nivel terminológico

Se limita la info lingüística a un dato

Ventajas y desventajas

Ampliación de metamodelo con info lingüística

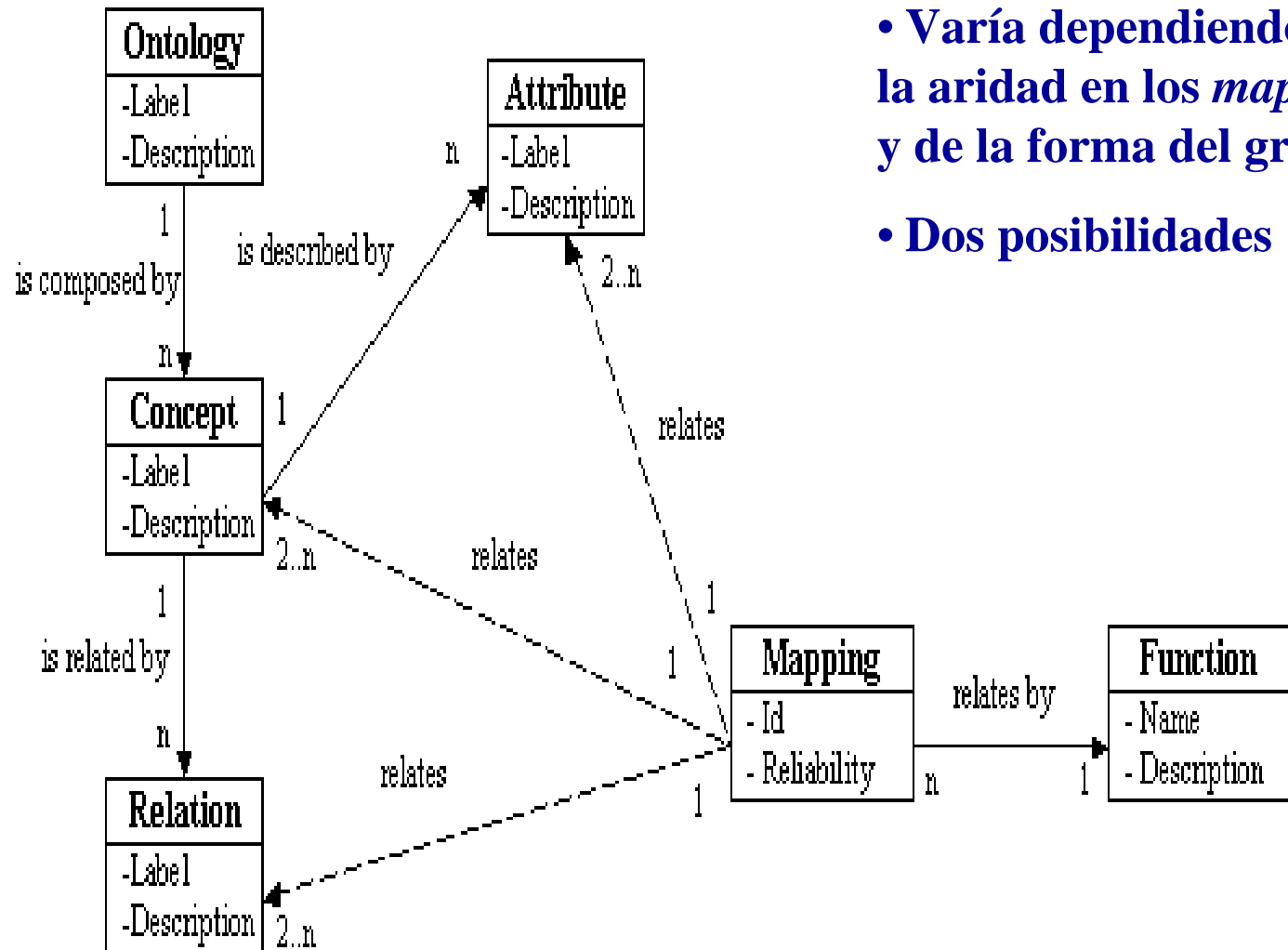
➤ Ventajas

- La ampliación a otras lenguas es fácil
- Adecuado para dominios muy especializados:
conocimiento compartido por comunidades lingüísticas de usuarios

➤ Desventajas

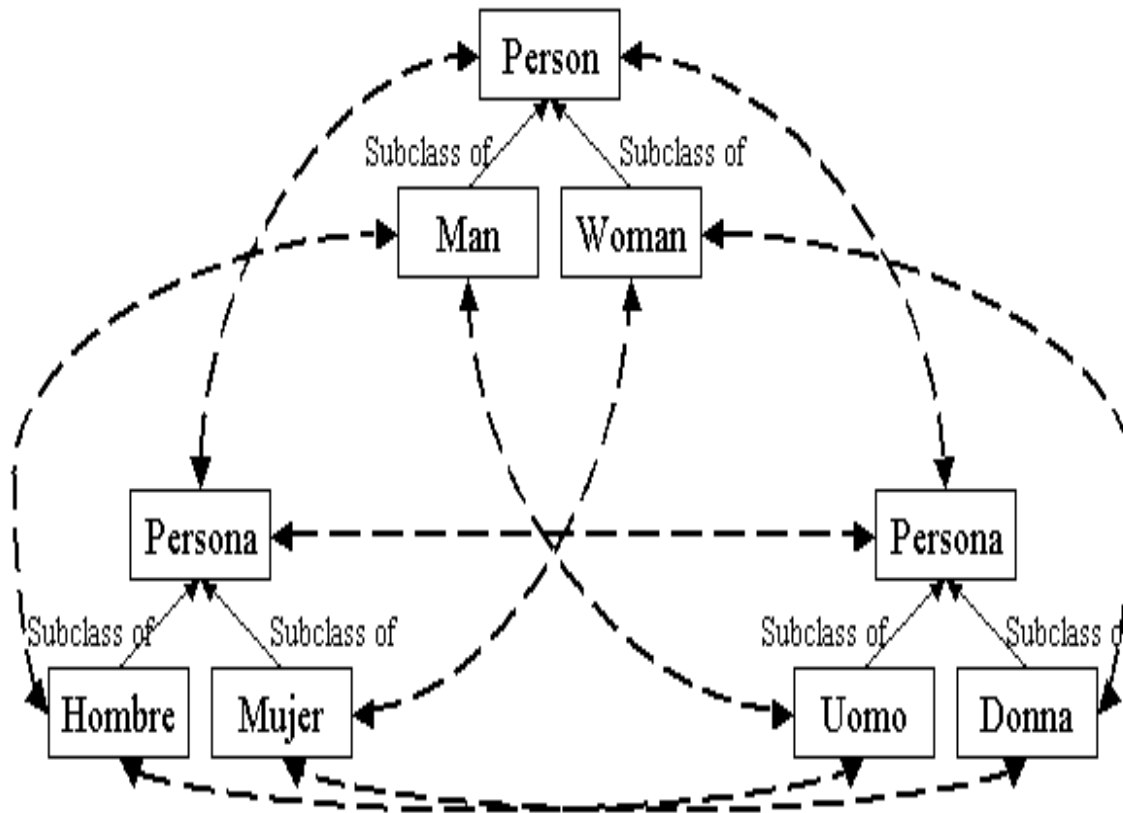
- Información lingüística limitada a etiquetas para las clases
- Se asume una total sinonimia aunque no sea cierta

B. Metamodelo de Ontología y modelo de *mapping*: Ejemplo



- Varía dependiendo de la aridad en los *mappings* y de la forma del grafo
- Dos posibilidades

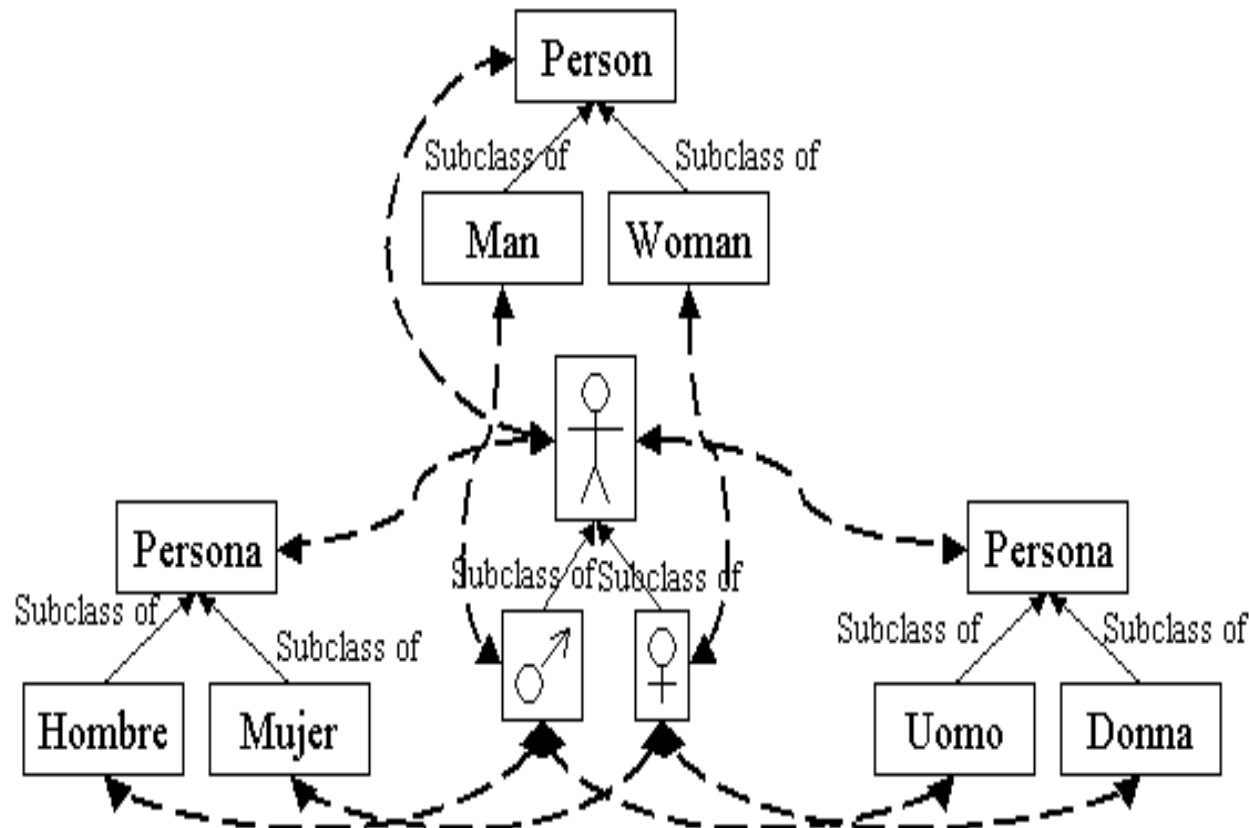
B. Metamodelo de ontología y modelo de *mappings* (1)



- Localización a nivel conceptual
- Menos intuitivo desde el punto de vista de abstracción

***Mappings* binarios en grafo ortogonal**

B. Metamodelo de ontología y modelo de *mappings* (2)



- Localización a nivel conceptual
- Los mappings se relaizan a través de un modelo “interlingüe” (EWN)

***Mappings* binarios en grafo radial**

Ventajas y desventajas de la opción B

Metamodelo de ontología y modelo de *mappings*

➤ Ventajas

- Se mantienen las conceptualizaciones en cada lengua
- Adecuado para dominios muy dependientes de la lengua: el ámbito judicial

➤ Desventajas

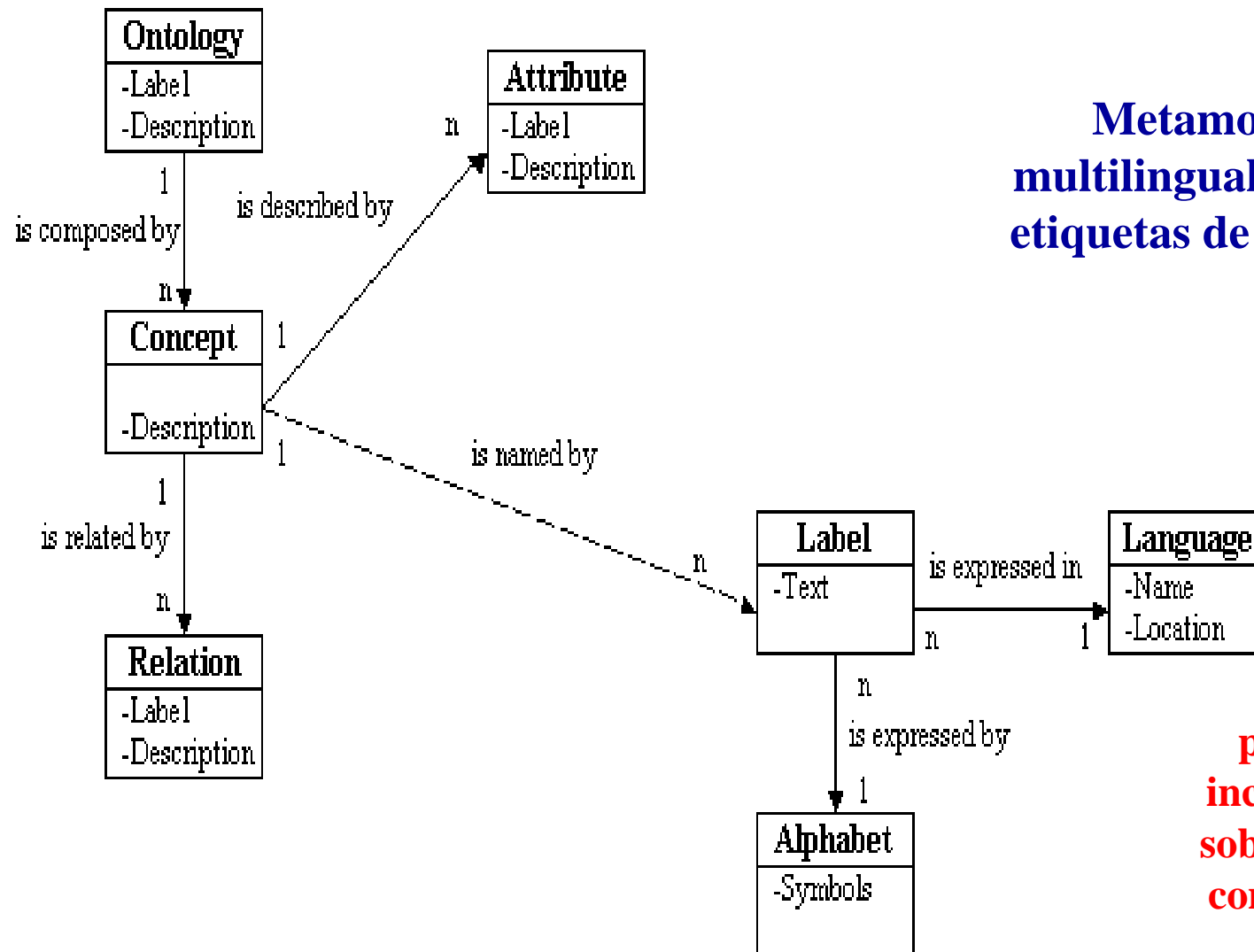
- Se requiere mucho esfuerzo para modelizar el mismo dominio en lenguas distintas
- Se requiere dominio lingüístico, del campo de conocimiento y ontológico

C. Modelo de información lingüística y relacionarlo con metamodelo de ontologías

- Localización al nivel terminológico y conceptual
- Los elementos de la ontología se enlazan con los datos multilingües almacenados fuera de la ontología
- Diferentes formas de organizar y representar la información lingüística: BD (Genoma KB, Oncoterm), una ontología, etc.
- La conceptualización permite modificaciones para satisfacer las necesidades de localización: creación de módulos

Metamodelo de ontología y modelo de recurso lingüístico

Opción C. Ejemplo.



**Metamodelo de
multilingüidad para
etiquetas de conceptos**

**Aumenta las
posibilidades de
incluir información
sobre la lengua y los
componentes de las
ontologías.**

Ventajas y desventajas

- **Ventajas**

- Se puede incluir toda la información lingüística que se desee
- Se enlazan los elementos lingüísticos dentro de una lengua y entre varias lenguas
- Las diferencias y especificidades se pueden formalizar al nivel terminológico
- Se preserva información relevante: fuentes, etc.
- No es necesario el conocimiento del experto en ontologías para acceder al nivel terminológico en un entorno distribuido

- **Desventajas**

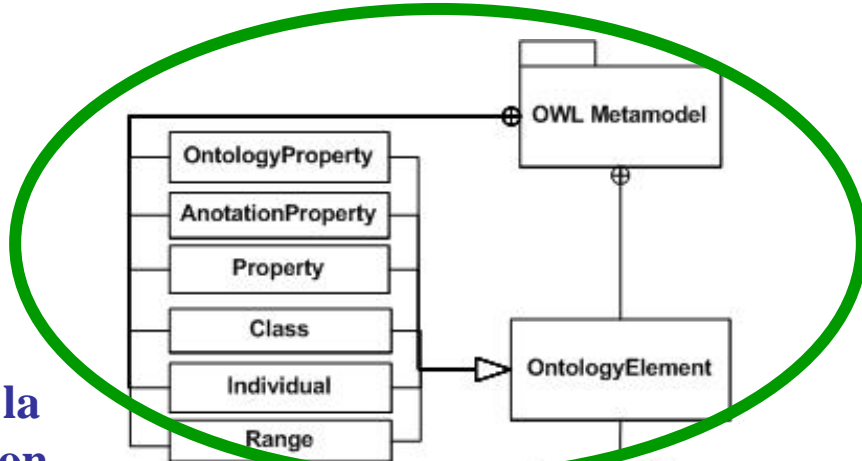
- Se pueden perder especificidades propias de una lengua, salvo que se reflejen en los módulos de la ontología específicos de la lengua

Una nueva propuesta: LIR

Linguistic Information repository

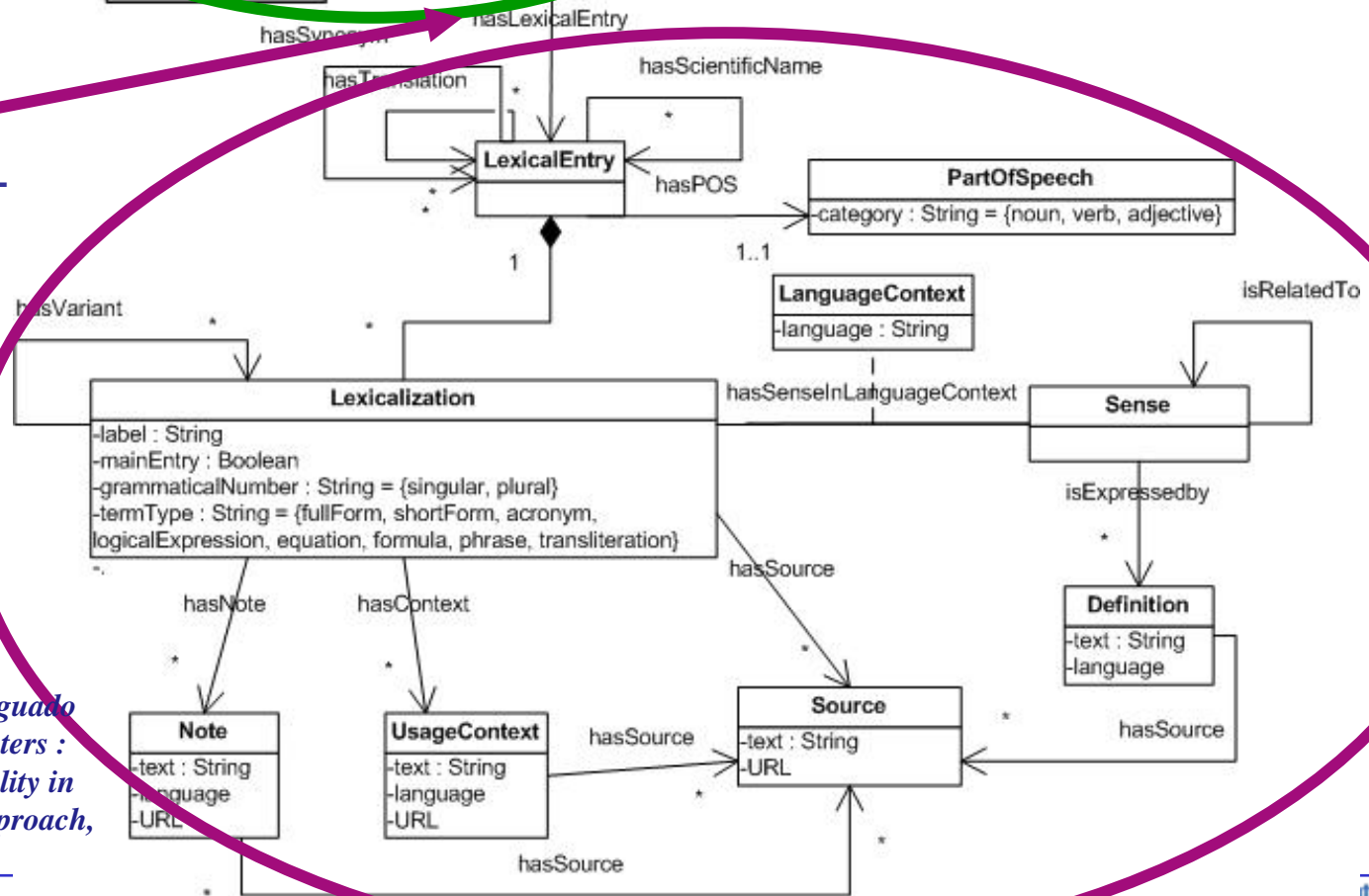
- Modelizado como una ontología
- Se enlaza con el metamodelo mediante la clase `LexicalEntry`
- Una *lexical entry* es una relación ternaria: `Lexicalization`, `Sense` y `LanguageContext`.
- `Note` se enlaza con `Lexicalization`, pero también se puede enlazar con cualquier otra clase del modelo para incluir información suplementaria
- Al enlazar `Note` con las clases `Sense` o `Definition`, se manifiestan las posibles diferencias en las lenguas

Metamodelo
de la ontología
en OWL



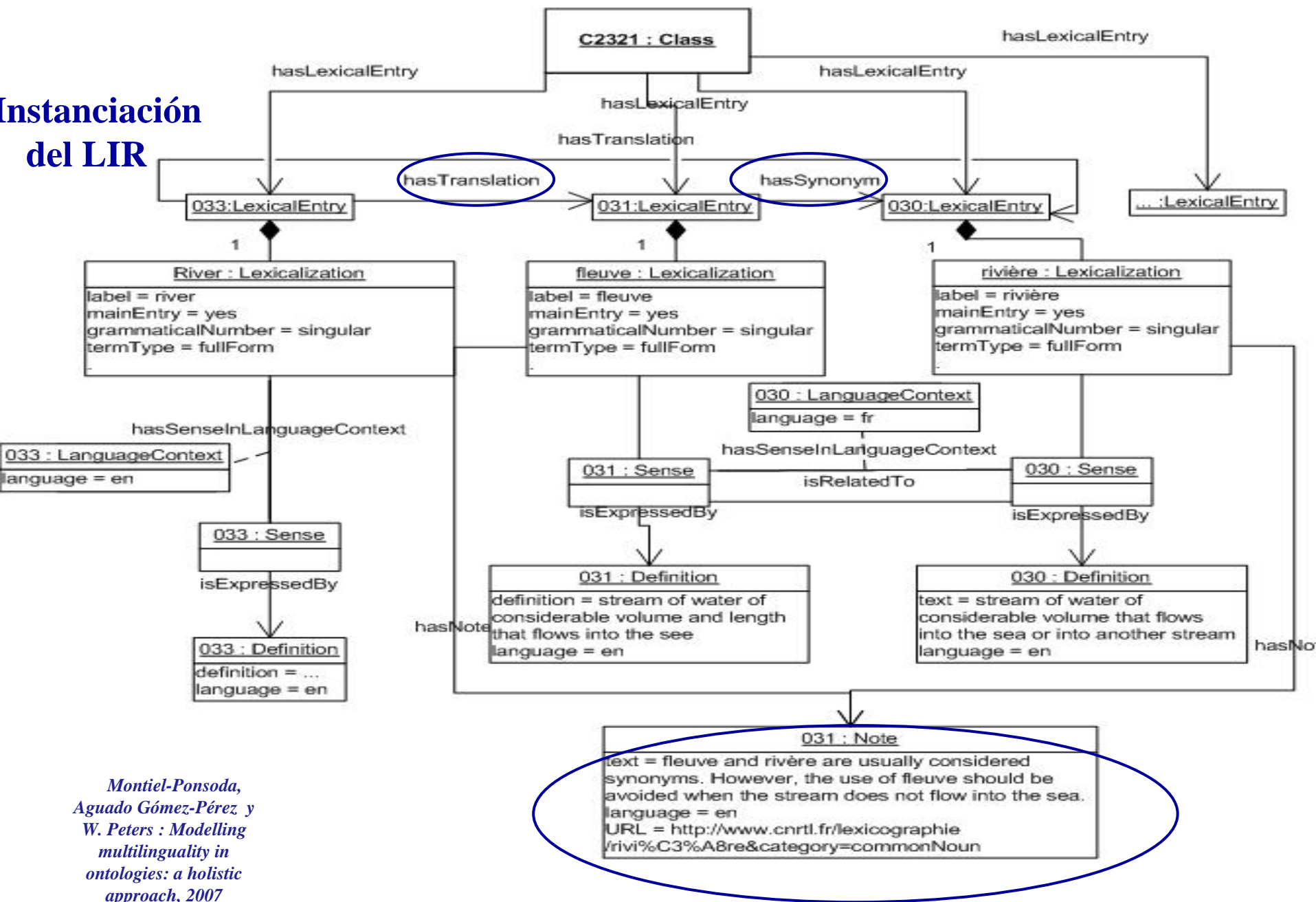
Enlace de la
Ontología con
LIR mediante
la relación:
“hasLexical
Entry”

LIR



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

Instanciación del LIR



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

Ventajas de esta propuesta (LIR)

- Se preserva la **independencia** entre la ontología y la capa multilingüe.
- Permite **conectar** la información multilingüe con cada elemento de la ontología.
- La adopción de estándares en la descripción lingüística ayuda a **mantener las especificidades** de la lengua y **garantiza** la localización del significado en la ontología.
- Facilita la **interoperabilidad** y la **extensibilidad** si se requiere más información.
- El **acceso** a los recursos multilingües externos es posible gracias a herramientas como **LabelTranslator**.