UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en microelectrónica y sistemas distribuidos

Semantic Mining

Jose Aguilar

TEPUY Research+Development

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en microelectrónica y sistemas distribuidos

Index

1. Introduction
2. Web Semantic Mining
3. Ontology Mining
4. Graph Mining
5. Linked Data

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en microelectrónica y sistemas distribuidos

Introduction

Semantic Mining

Data Mining is a mature area in Computer Science, whose **main objective is the extraction of knowledge.**

Data Mining has required to be enriched in recent years, due to the need to analyze **semantic content.**

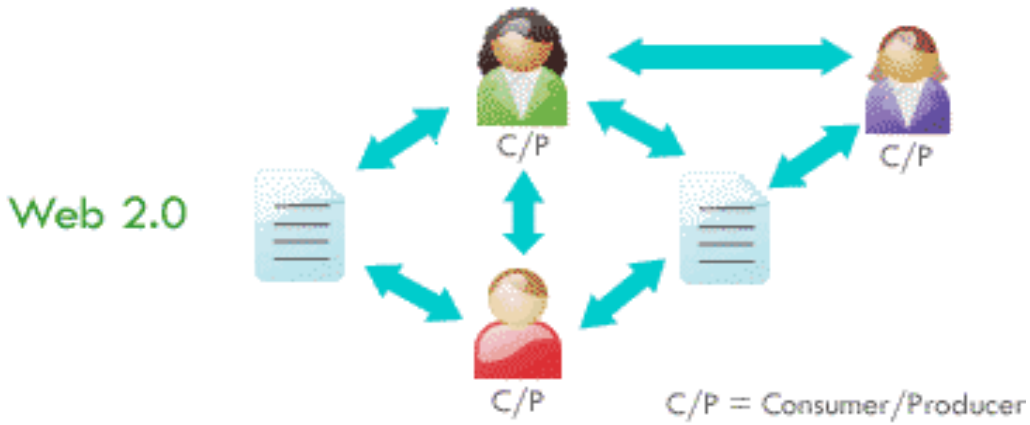
Some kinds of Mining

- Data Mining
  - Spatial Data
  - Spatial-temporal
  - moving targets
  - multimedia data
  - data streams
- Process/Service Mining
- Mining of domains: health, air traffic control, foods, energy
- Text mining
  Web Mining
- …

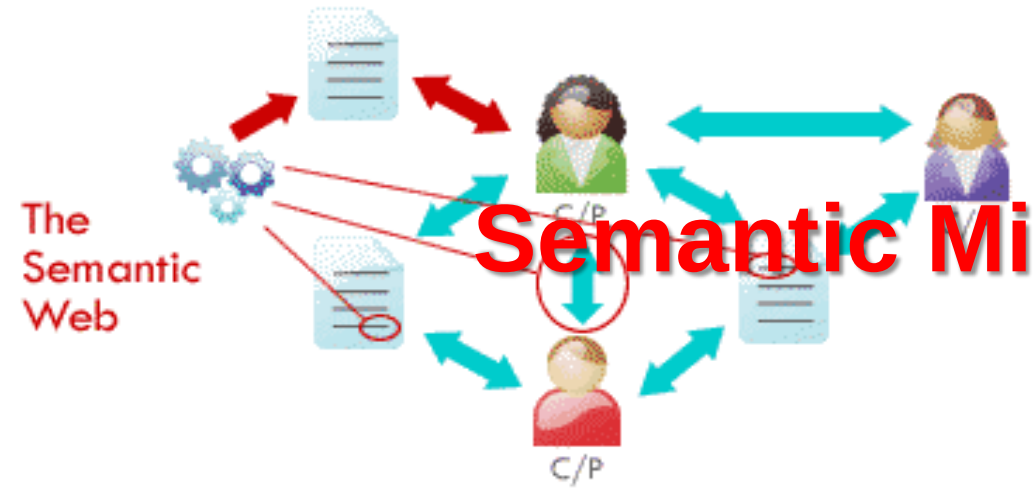
Semantic Mining

WEB *.0 -> WEB 3.0

- Static pages
- HTML.



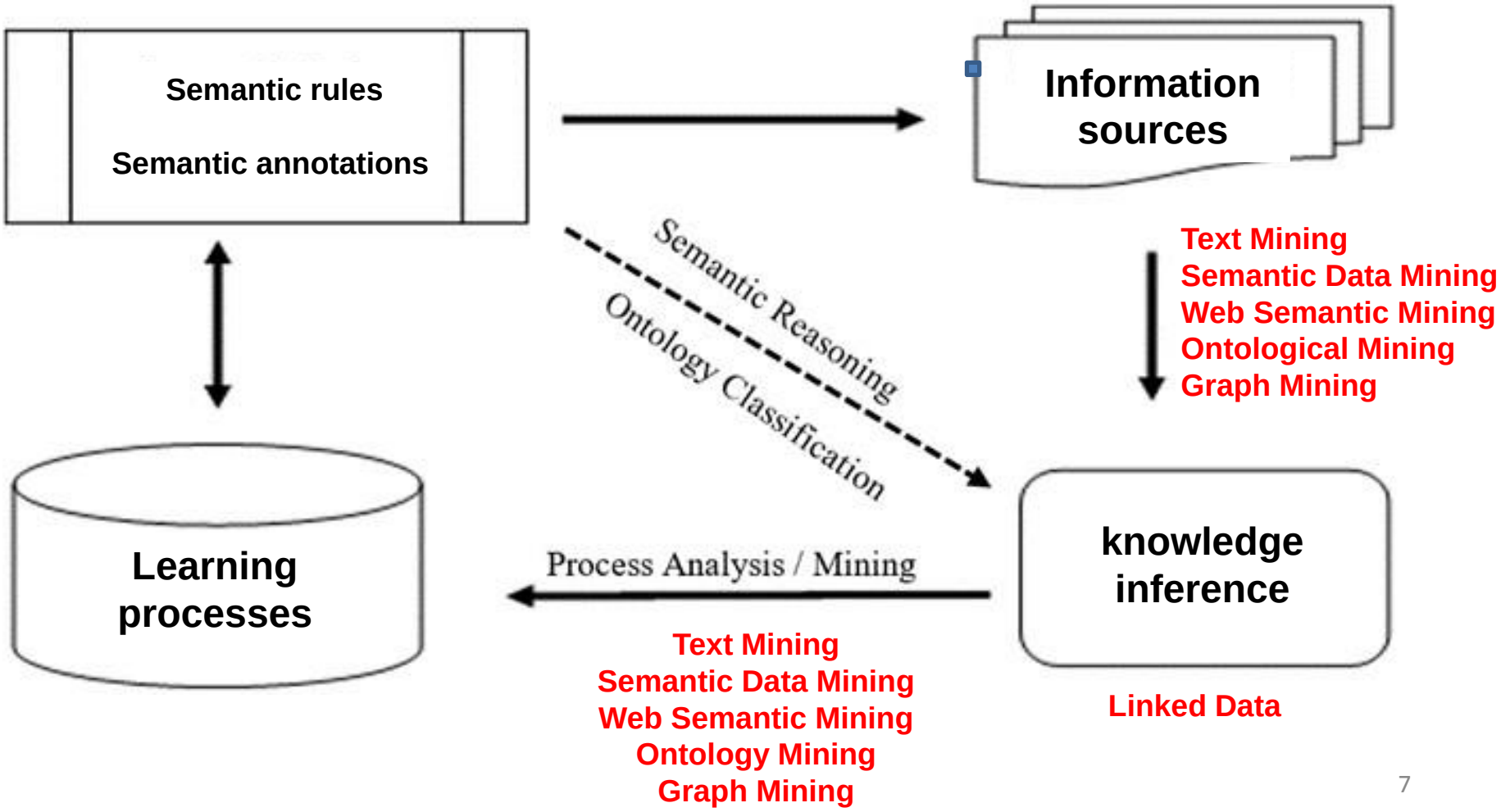
3 basic principles
- The web as a platform
- Take advantage of Collective Intelligence
- Enriching experiences of the user



Semantic Mining

Web 3.0 is based on
- A more "intelligent" Internet
- Users carry out searches close to natural the
- Information has associated semantics
- The website deduces information through rules associated with the meaning of the content

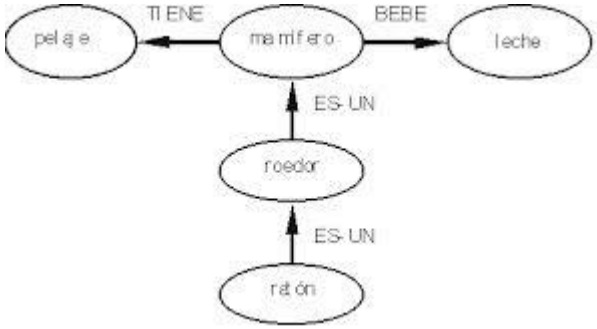
General idea



Semantic Mining

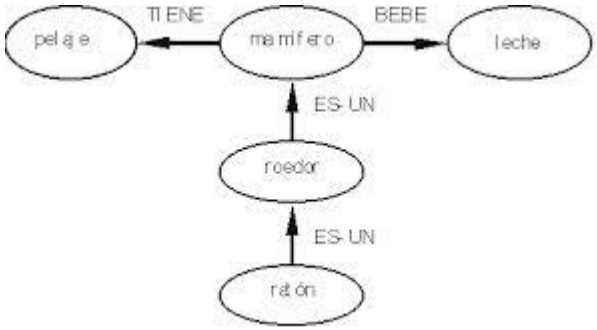
	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874215	BOSA	$ 150.000
6	Liliana Cruz	Chef Panadera	3201478951-7451825	SUBA	$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	$ 800.000
8	Camila Davalos	Cajera	3214875961-7584621	TEUSQUILLO	$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	$ 600.000
10	Pamela Carrazco	Mesera	3157485912-2485796	CANDELARIA	$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	$ 600.000
12	Jairo Arevalo	Parqueadero	3002157459-2861459	BOSA	$ 489.500
13				TOTAL	$ 8.199.500

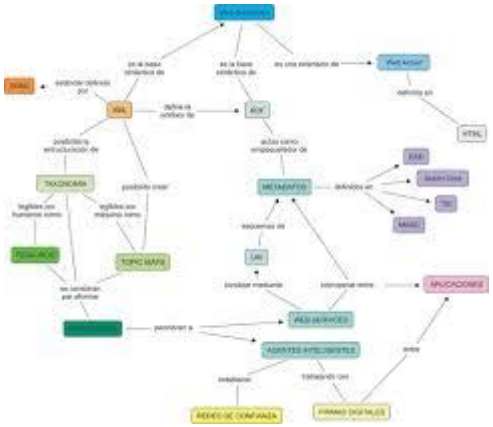
Mining



Semantic Mining

	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874215	BOSA	$ 150.000
6	Liliana Cruz	Chef Panadera	3201478951-7451825	SUBA	$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	$ 800.000
8	Camila Davalos	Cajera	3214875961-7584621	TEUSQUILLO	$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	$ 600.000
10	Pamela Carrazco	Mesera	3157485912-2485796	CANDELARIA	$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	$ 600.000
12	Jairo Arevalo	Parqueadero	3002157459-2861459	BOSA	$ 489.500
13				TOTAL	$ 8.199.500





- Websites,
- Unstructured content on the web,
- Structured content on the web,
- Labeled Graphs,
- Ontologies,
- Data Table, among others

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Metodologies

MIDANO

MASINA

MEDAWEDE

ApEm

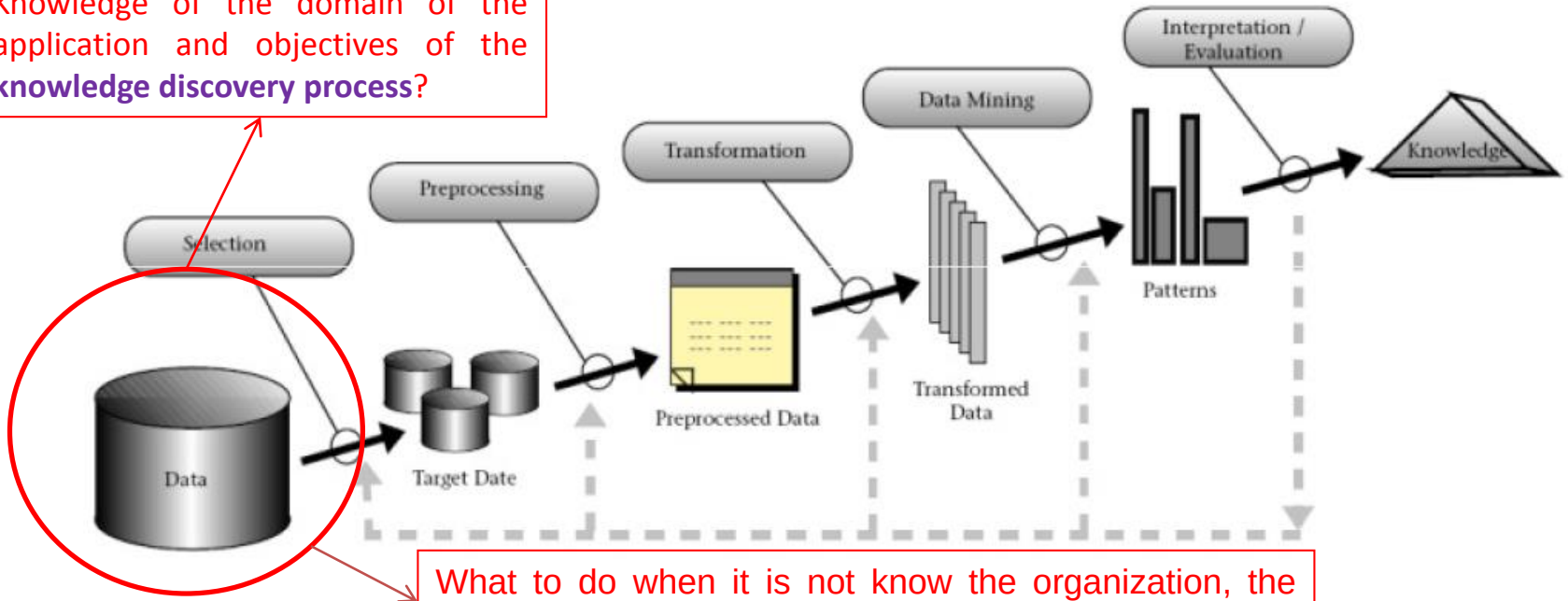
MIDANO

"Methodology for the development of data mining applications based on organizational analysis"

Extended to be used in data analysis

MIDANO

Knowledge of the domain of the application and objectives of the **knowledge discovery process**?

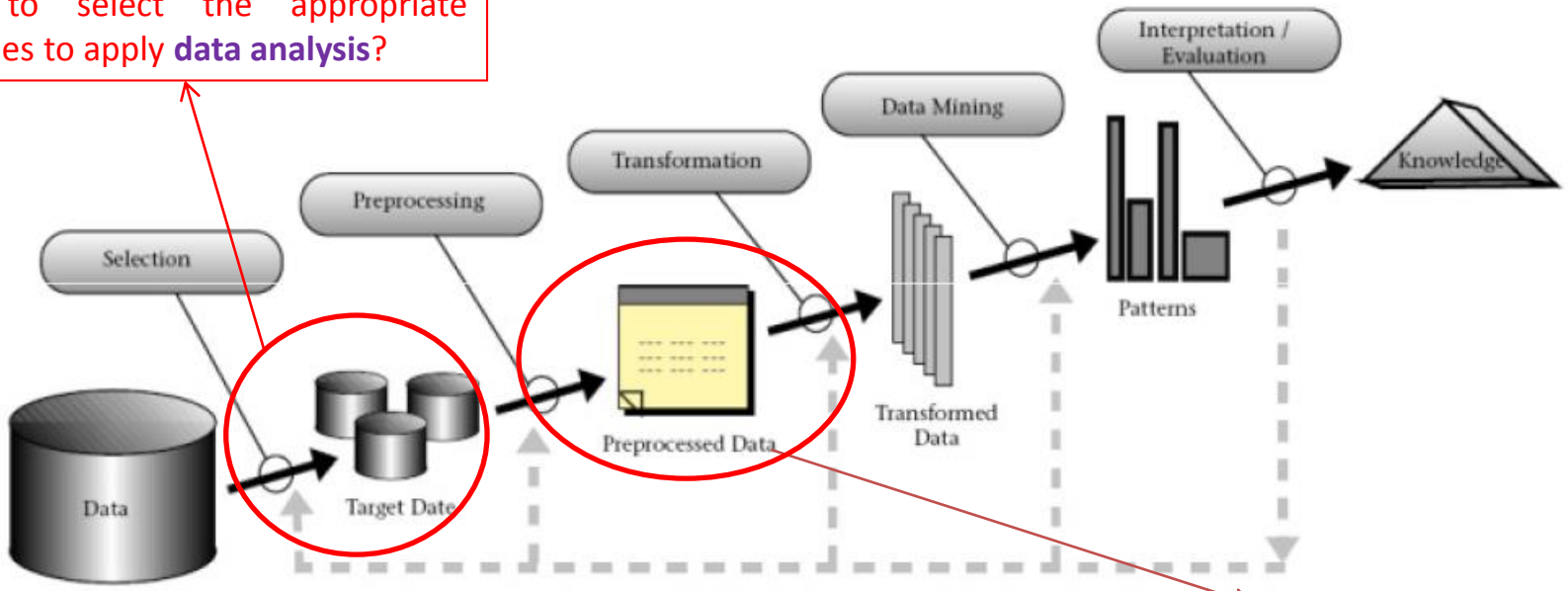


What to do when it is not know the organization, the problem, or the processes to be studied?

"Methodology for the development of data mining applications based on organizational analysis"

MIDANO

How to select the appropriate variables to apply **data analysis**?



How to perform **data processing?**

"Methodology for the development of data mining applications based on organizational analysis"

MIDANO-AofD

Three phases

Phase 1

Identification of the sources of knowledge extraction in an organization

Phase 2

preparation and processing of data

Phase 3

Development of autonomic cycle of data analysis tasks

Business intelligence **Data sciences** **Data analysis**

Phase 1: Identification of the sources of knowledge extraction in an organization

Future scenarios should be aimed at achieving them

Statistical metrics, knowledge models, …

Description of the future scenario

Strategic objectives to be achieved	Associated actor (s)	Associated variables	AofD activities to be carried out	New features	Results to be obtained (achievement indicators)

Future scenario : < xxx >

The set of future scenarios defines an organizational strategic technology plan

Phase 1: Identification of the sources of knowledge extraction in an organization

For each autonomous cycle

Strategic objective to be achieved : < … >

	Name	General sources of data required	Indicators generated	Expected effects on the strategic objective
AofD Observation tasks				
AofD Analysis tasks				
AofD Decision Making Tasks				

Statistical metrics, knowledge models, ... that produce

Used in the future as an AC quality metric

Relationships between the tasks of the AofD

	AofD1 Task	AofD2 task	AofD3 task
AofD1 task			
AofD2 Task			
AofD3 Task			

Phase 2: preparation and processing of data

- Definition of the data model
- Data treatment

- Conceptual View
- Operational View

With these views, the multidimensional data model

Phase 2: preparation and processing of data (data science)

Name	Name of the fact table
Keys to dimension tables	
Objectives Variables	Variables that describe or are associated with extracted knowledge (predictions, etc.)
Other variables	Variables required by the AofD task, for example, derived from dimension processing operations or OLAP

Name	Name of the dimension table
Keys of the dimension	
Attributes of the dimension	Attributes that describe the theme associated with that dimension

Variable	Extraction	Transformation	Load
Name of the variable			

ETL table

CCA table

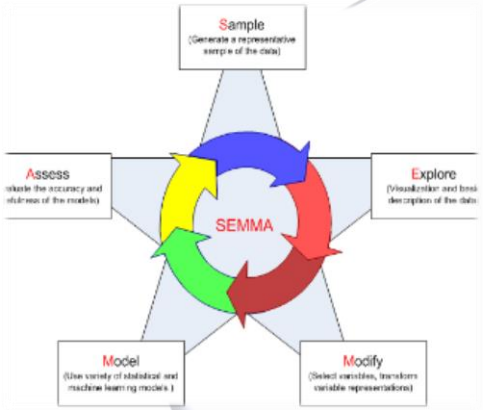
Variable	Collection	Curation	Analysis
Name of the variable			

Phase 3: Development of autonomic cycle of data analysis tasks

- Specification of the decision-making system
- Technological specification of the autonomic cycle
- Development of the autonomic cycle
- Validation

Phase 3: Development of autonomic cycle of data analysis tasks

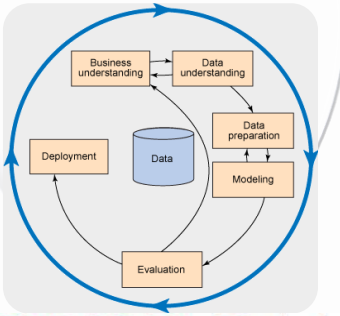
Development of AofD tasks



SEMMA

Any DM methodology can be used for this phase of AofD tasks,

CRISP-DM

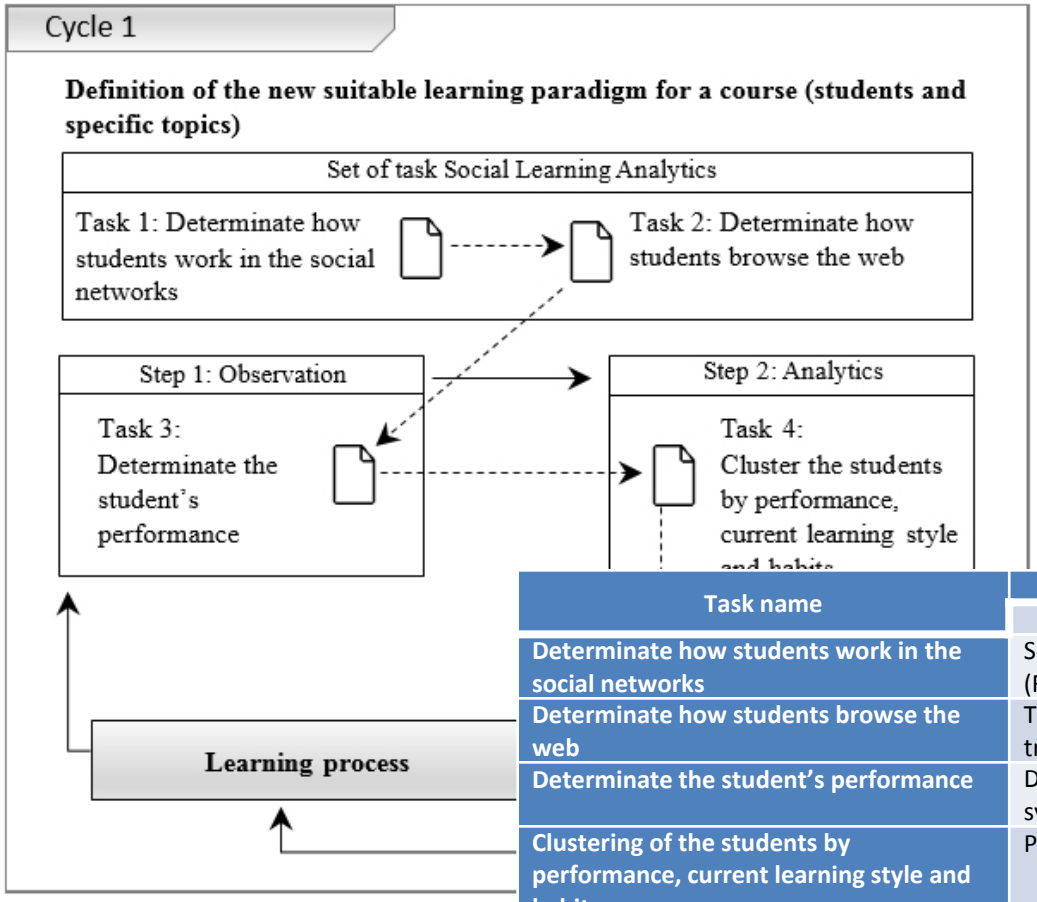


CATALYST

Specification of the Autonomic Cycles of Learning Analytic Tasks for a Smart Classroom

- ACOLAT 1: Definition of the current learning paradigm.
- ACOLAT 2: Determination of the educational resource for a given student.
- ACOLAT 3: Identification of students with special needs.
- ACOLAT 4: Avoid Student Desertion.
- …

Specification of the Autonomic Cycles of Learning Analytic Tasks for a Smart Classroom



Task name	Description of the task		
	Data source	Data analytic techniques	Role
Determinate how students work in the social networks	Social networks (Facebook, Twitter)	Social analytics network techniques	Observation
Determinate how students browse the web	The Internet, navigation trace	Web mining	Observation
Determinate the student's performance	Database of the academic system	Data mining	Observation
Clustering of the students by performance, current learning style and habits	Previous results	Data mining	Analysis
Determinate the new learning style suitable for the course/group	Previous results	Data mining	Decision

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

cemisid
centro de estudios en microelectrónica y sistemas distribuidos

Specification of the Autonomic Cycles

- Intelligent Autonomic System for Oil Processes

- To improve the quality of services in the communications networks

- Autonomous cycle for smart cities

Web Semantic Mining

Web Semantic Mining

It is the integration of two areas of knowledge:

- **Semantic Web**
- **Web Mining**

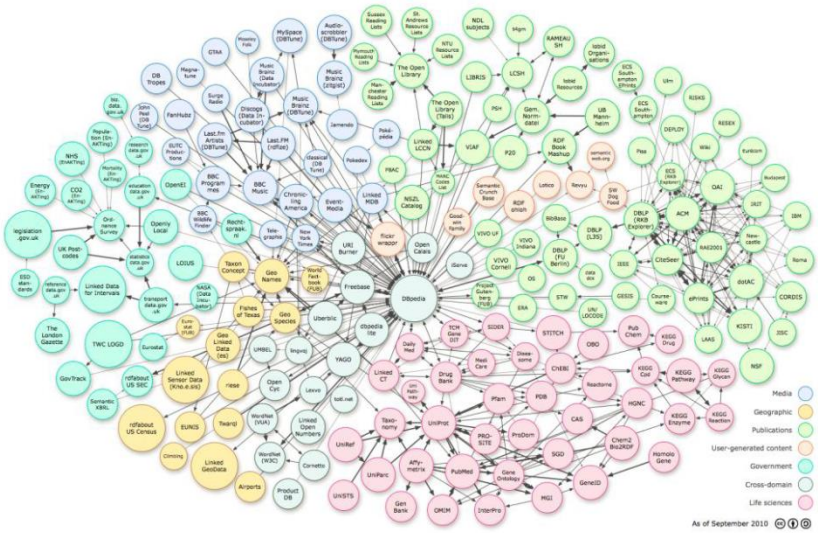
The **Semantic Web** is used to give meaning to the data found on the Web.

Web Mining is used to extract patterns of behavior on the Web.

Web Semantic Mining

Change of paradigm from data mining to knowledge mining

Semantic Web Mining: **Mining of knowledge in the web** encoded in domain ontologies, etc.



Types of semantic resources

- Domain Ontologies
- Ontologies in the semantic web

Web Mining

There are several types :

- The content of the web

Search results
Content of the website

Text Mining

- The structure of the web

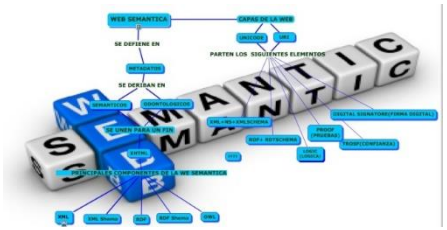
Links

Graph Mining

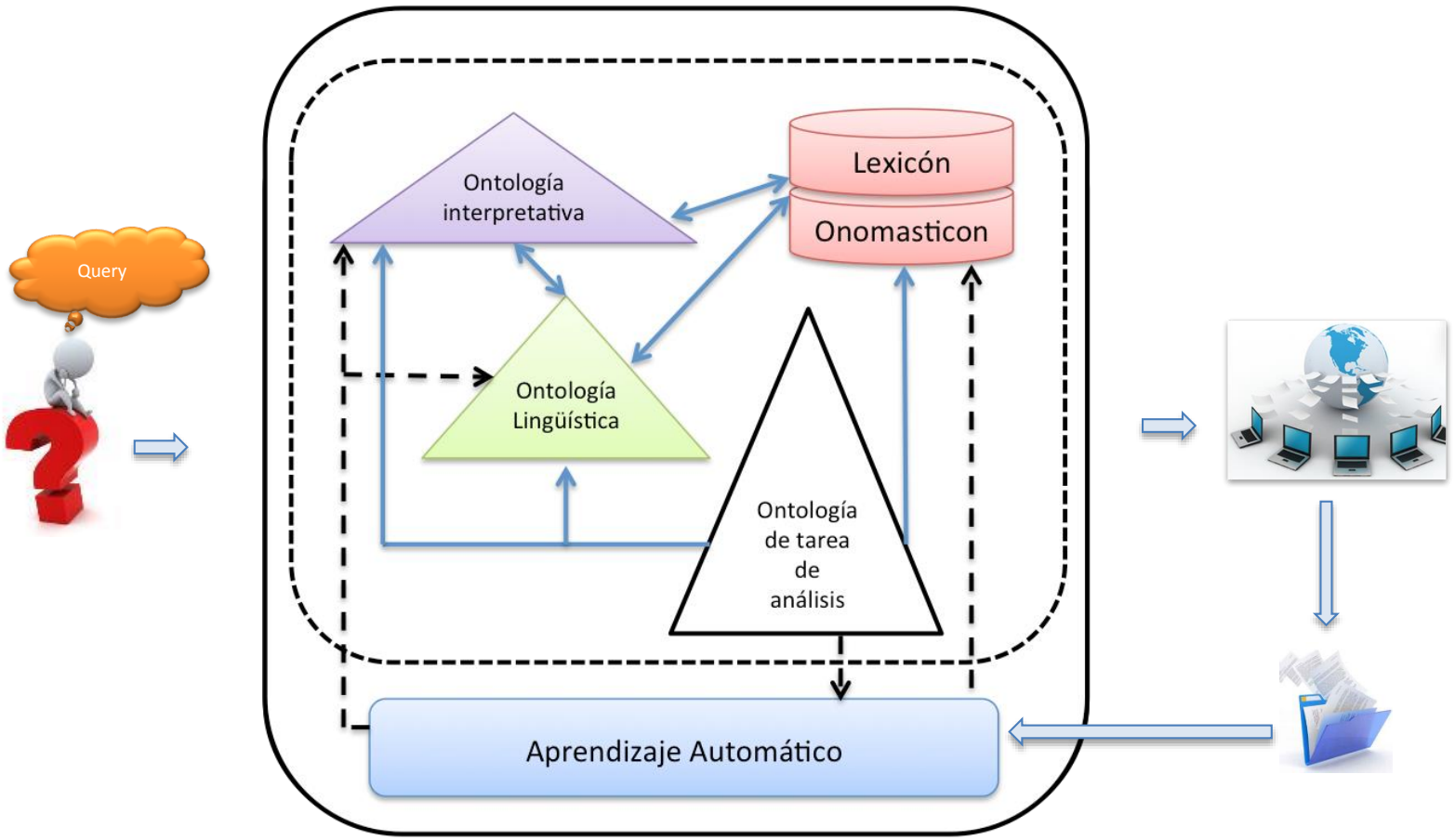
- The use made of the web.

General patterns of use
Personal access patt

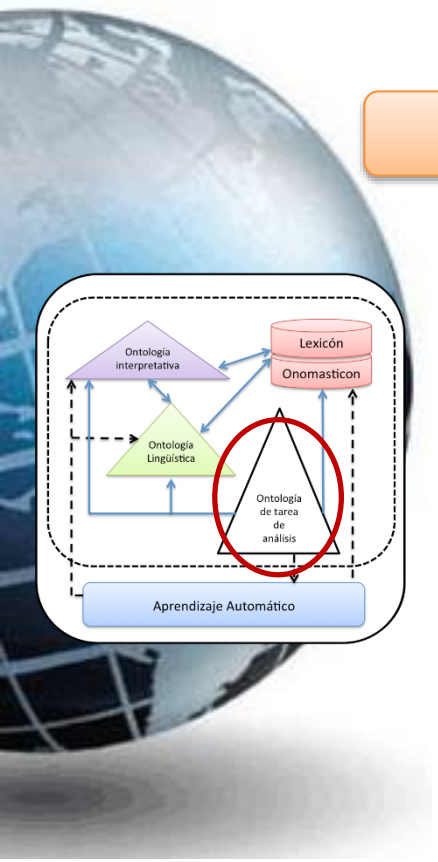
User Mining



Dynamic Semantic Ontological Framework



Task Ontology

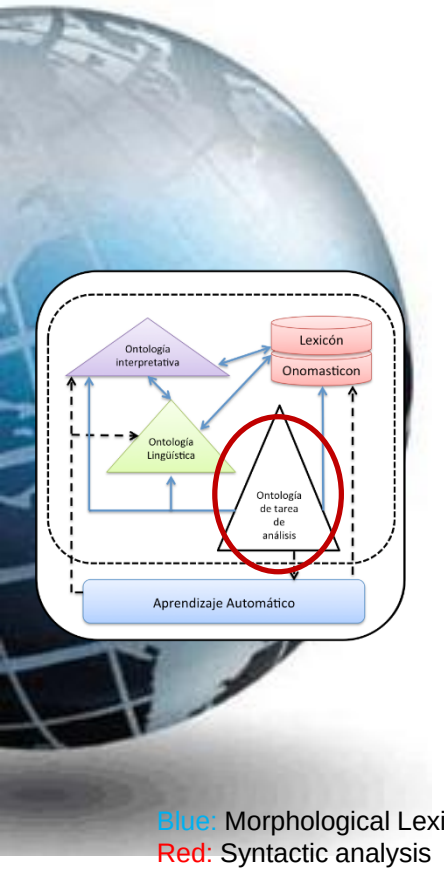


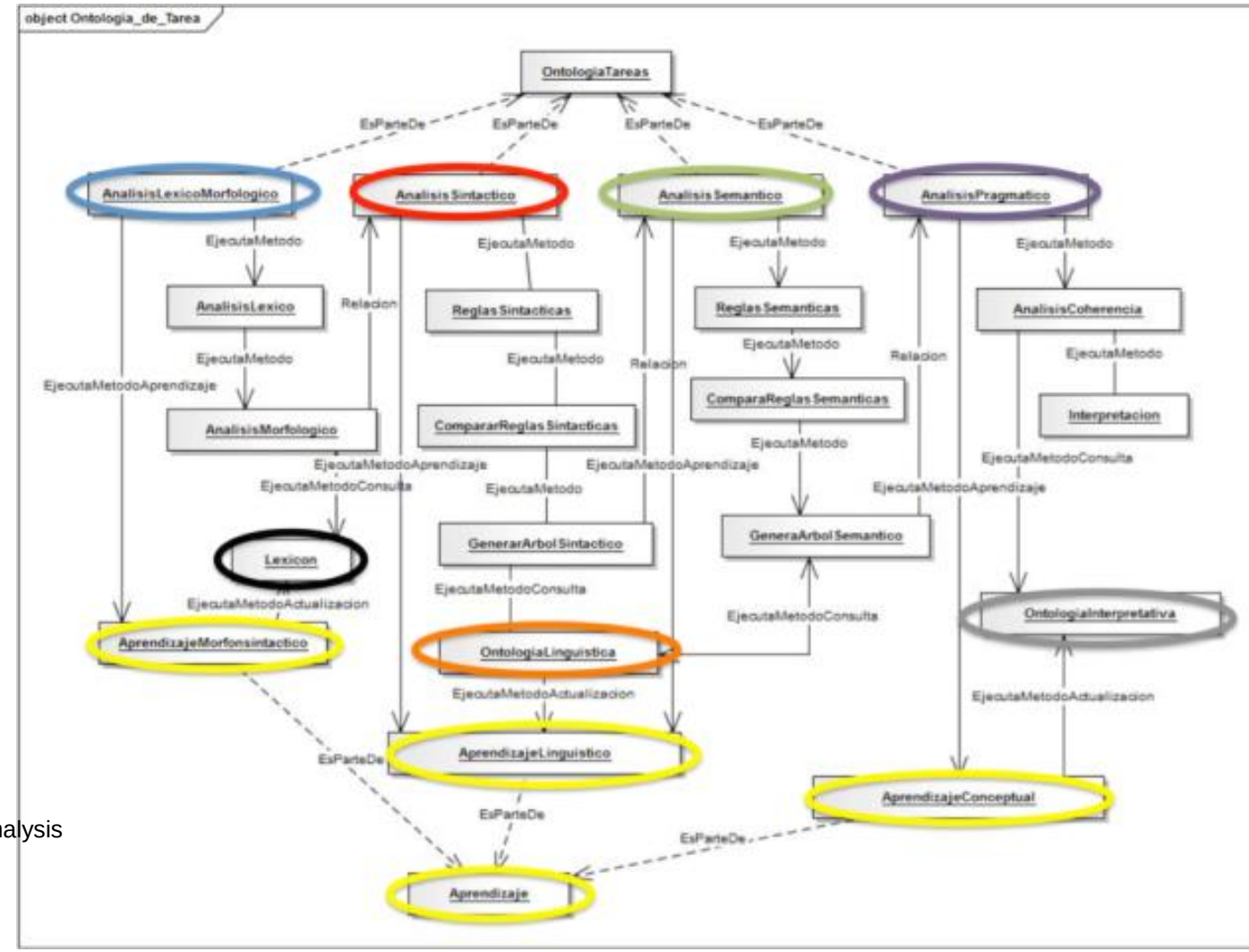
Task Ontology

Natural language processing
- Morphological Lexicon Analysis
- Syntactic analysis
- Semantic Analysis
- Pragmatic Analysis

Integration of:
- Onomasticon
- Lexicon
- Linguistics ontology
- Interpretive ontology
- Learning

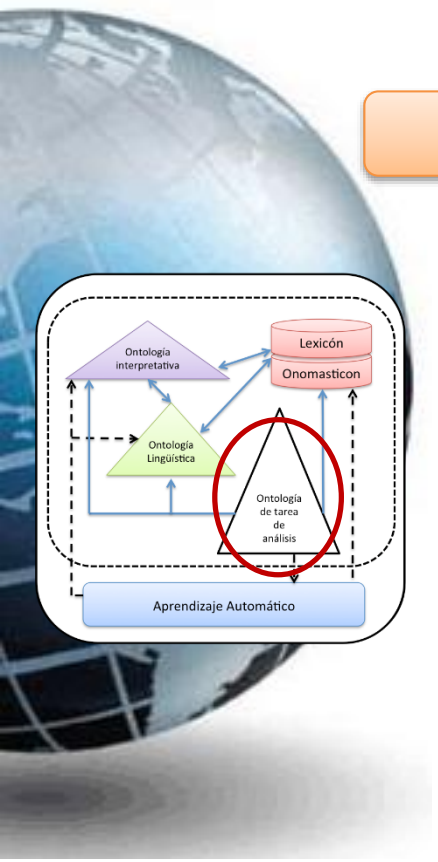
Task Ontology





Blue: Morphological Lexicon Analysis
Red: Syntactic analysis
Green: Semantic analysis
Purple: Pragmatic analysis
Black: Lexicon
Orange: Linguistic Ontology
Gray: interpretive ontology
Yellow: Learning

Task Ontology

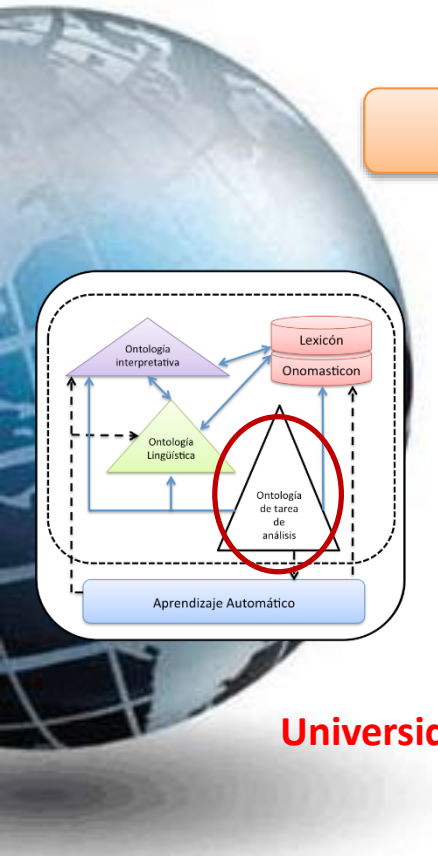


Task Ontology

Receive the query

For example: Universidad de Los Andes

Task Ontology



Task Ontology

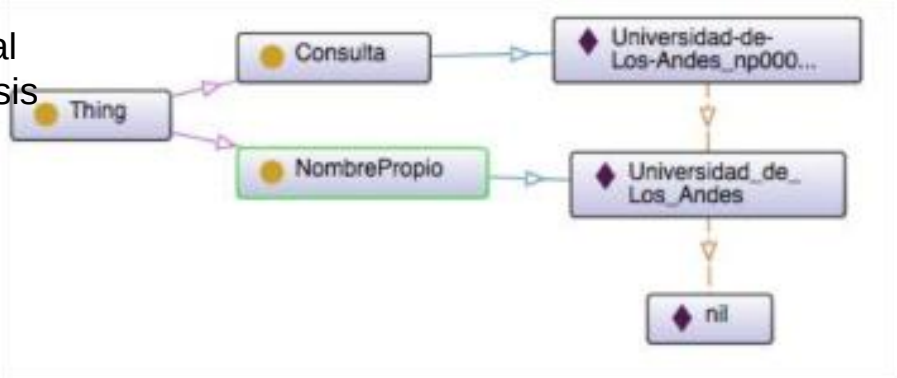
Lexicón

Onomasticon

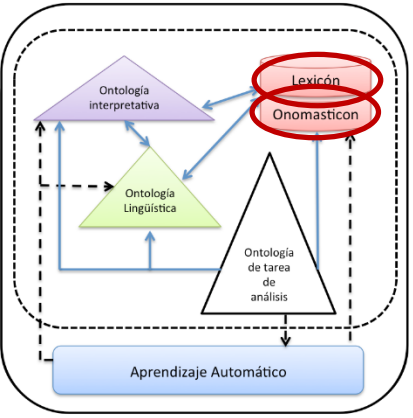
Morphological Lexicon Analysis

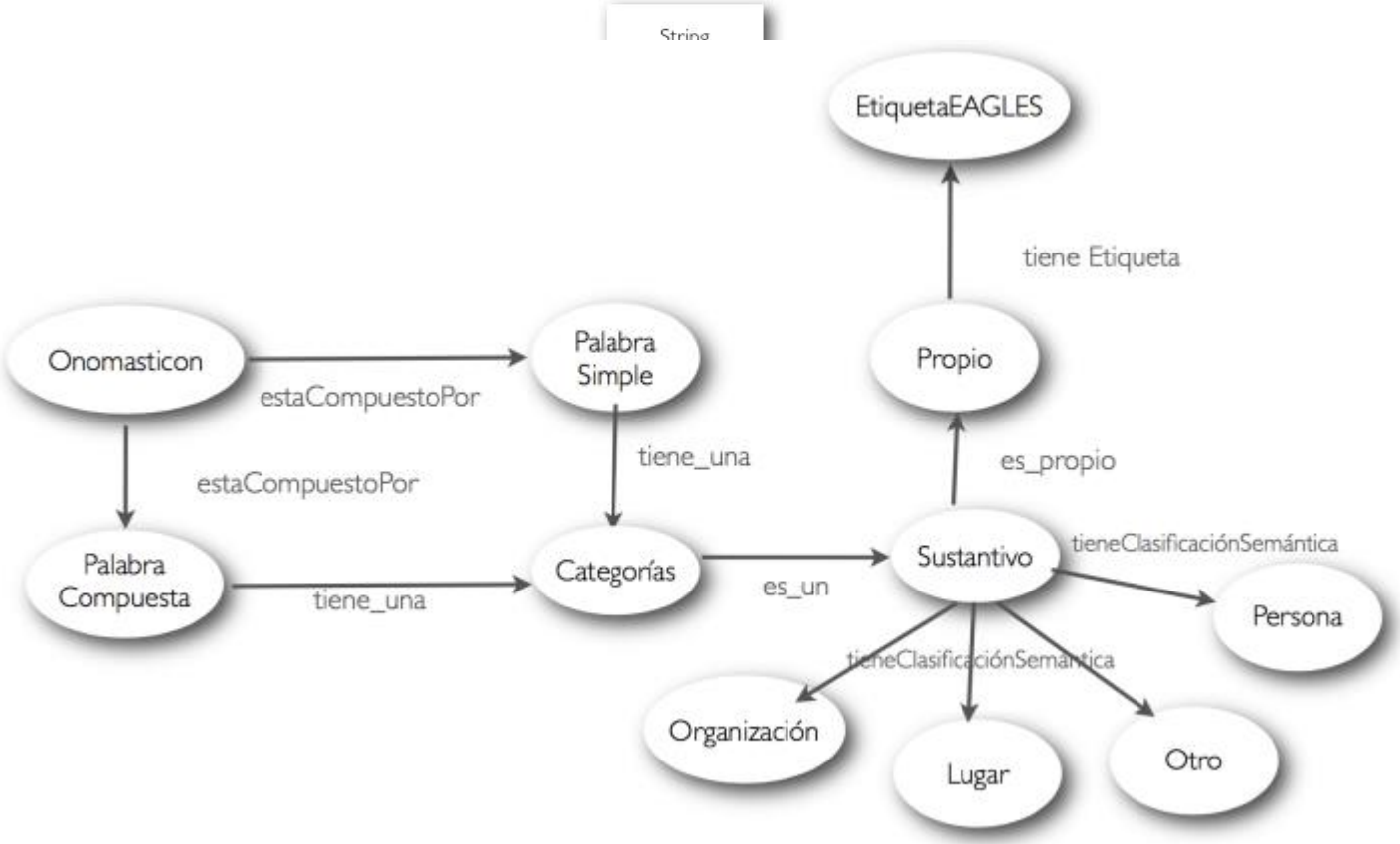
Universidad de Los Andes

Morphological Lexicon Analysis

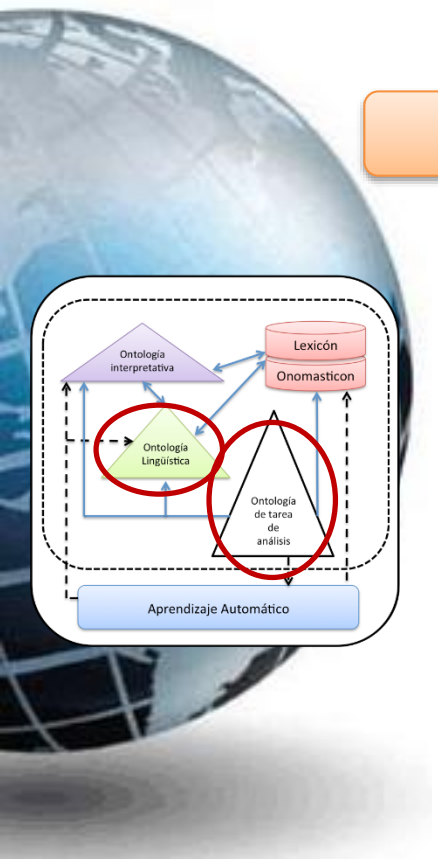


Lexicon and Onomasticon





Task Ontology



Task Ontology

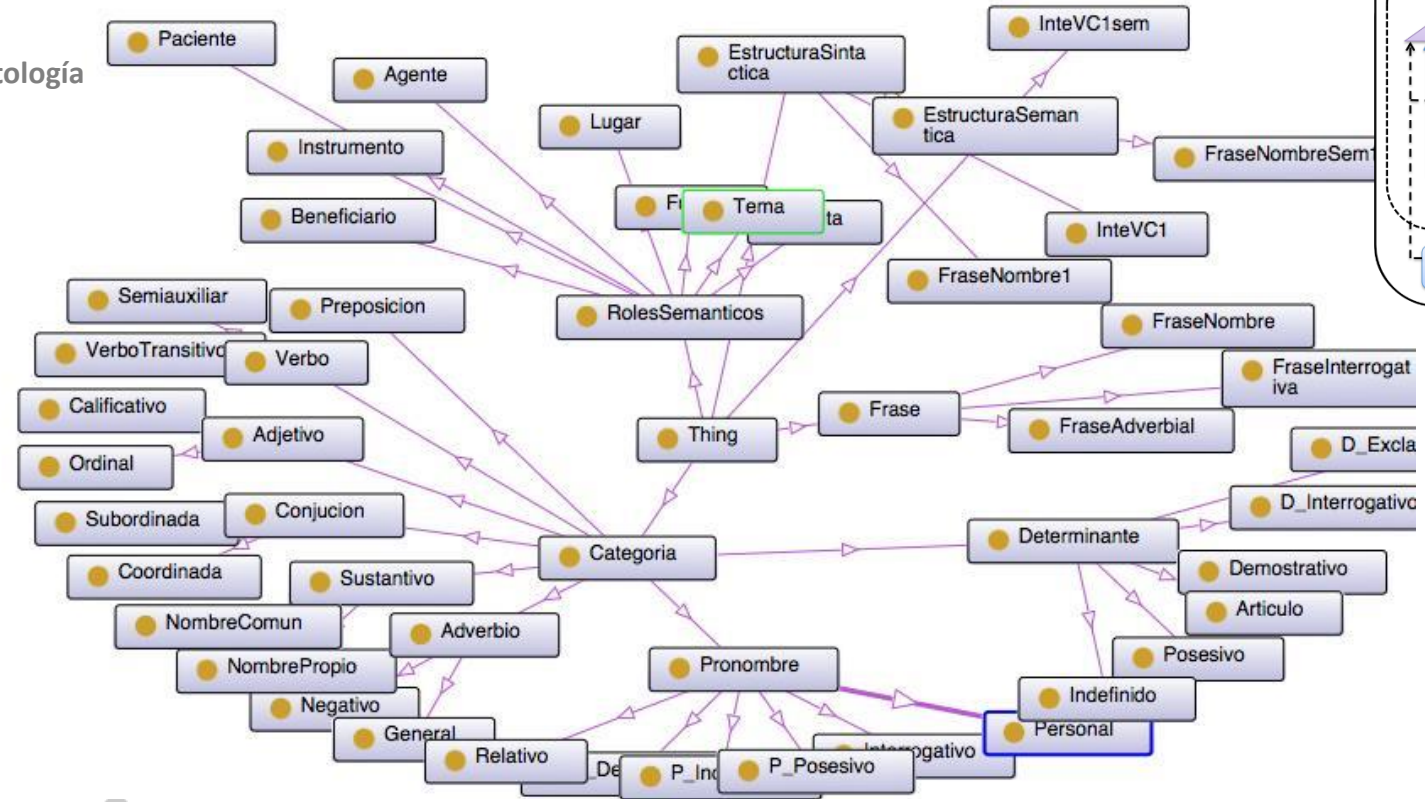
Syntactic Analysis

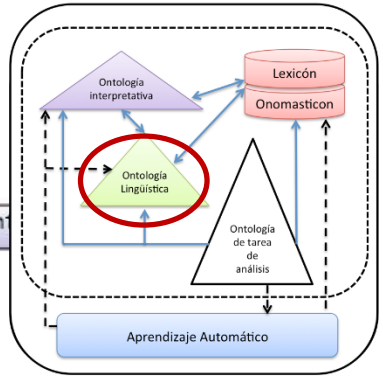
Semantic Analysis

Ontología Lingüística

Linguistics Ontology

Ontología





Rules

Articulo(?a), Consulta(?c), Interrogativo(?i), NombreComun(?nc), NombrePropio(?np), Preposicion(?p), Semiauxiliar(?v), hasNext(?a, ?nc), hasNext(?c, ?i), hasNext(?i, ?v), hasNext(?nc, ?p), hasNext(?p, ?np), hasNext(?v, ?a) -> InteVC1(?c)

InteVC1(?y), Tema(?x) -> InteVC1sem(?x)

InteVC1(?x) -> FraseInterrogativa(?x)

Agente(?x), FraseNombre1(?y) -> FraseNombreSem1(?x)

Agente(?x), InteVC1(?y) -> InteVC1sem(?x)

FraseNombre1(?x) -> FraseNombre(?x)

FraseNombre1(?x), NombrePropio(?y) -> Agente(?y)

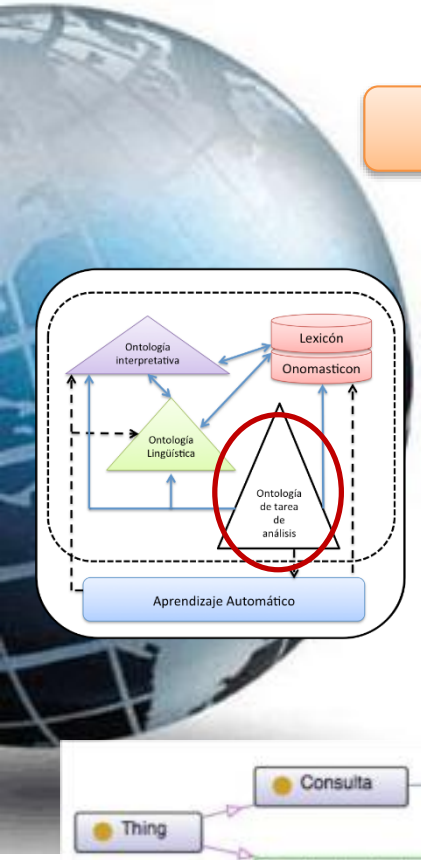
InteVC1(?s), NombrePropio(?np), Preposicion(?p), hasNext(?p, ?np) -> Tema(?np)

Consulta(?x), NombrePropio(?y), hasNext(?x, ?y), hasNext(?y, nil) -> FraseNombre(?x)

Articulo(?a), InteVC1(?x), NombreComun(?nc), hasNext(?a, ?nc) -> Agente(?nc)

SWRL rules

Task Ontology

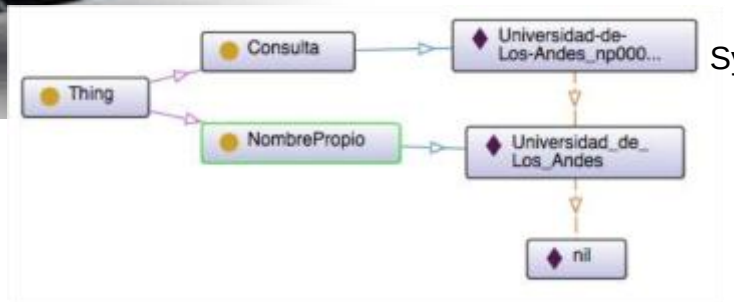


Task Ontology

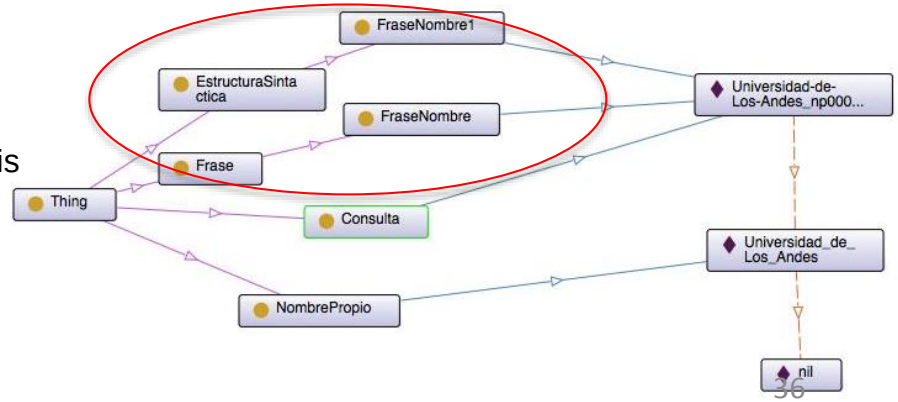
Ontología Lingüística

Syntactic Analysis

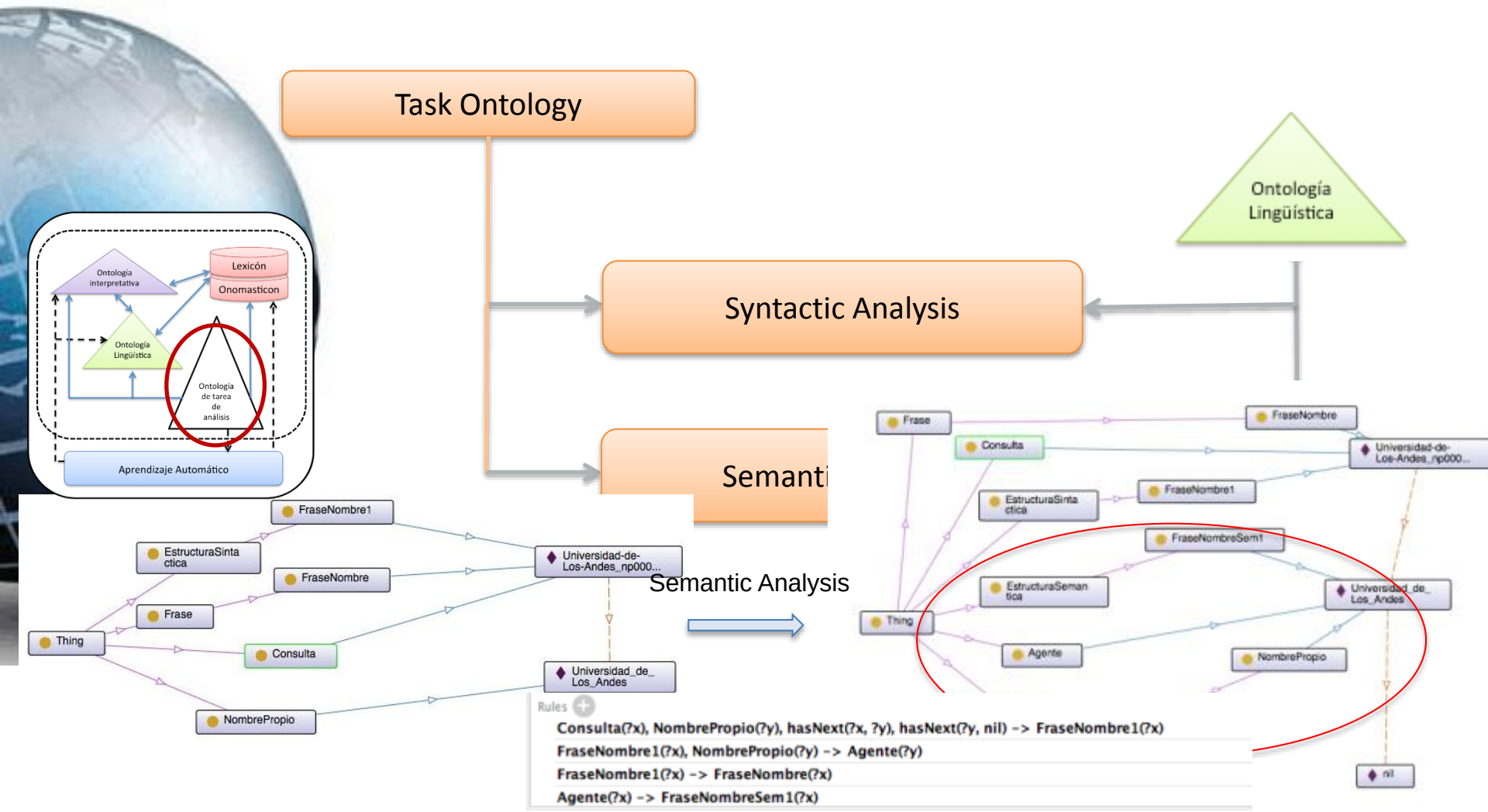
Consulta(?x), NombrePropio(?y), hasNext(?x, ?y), hasNext(?y, nil) -> FraseNombre1(?x)



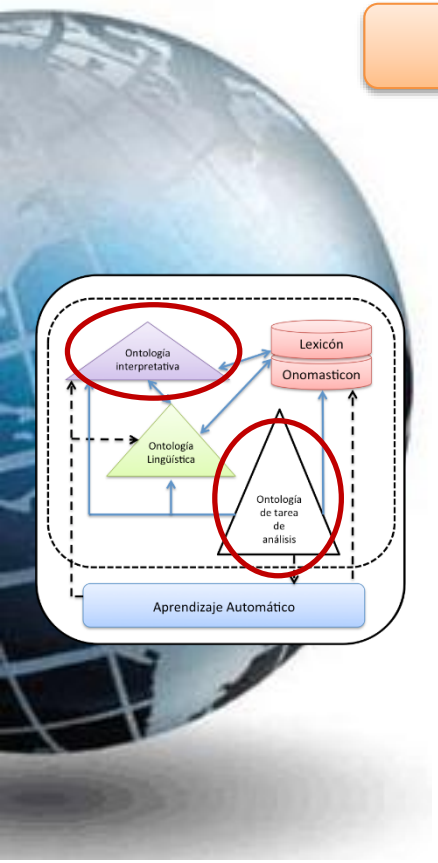
Syntactic Analysis

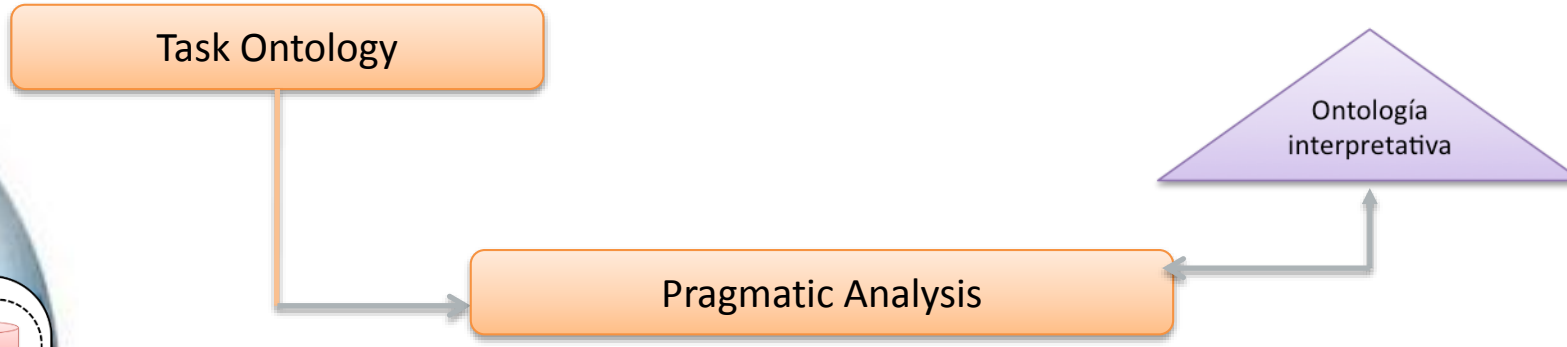


Task Ontology

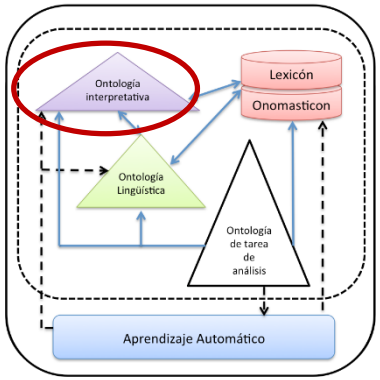


Task Ontology



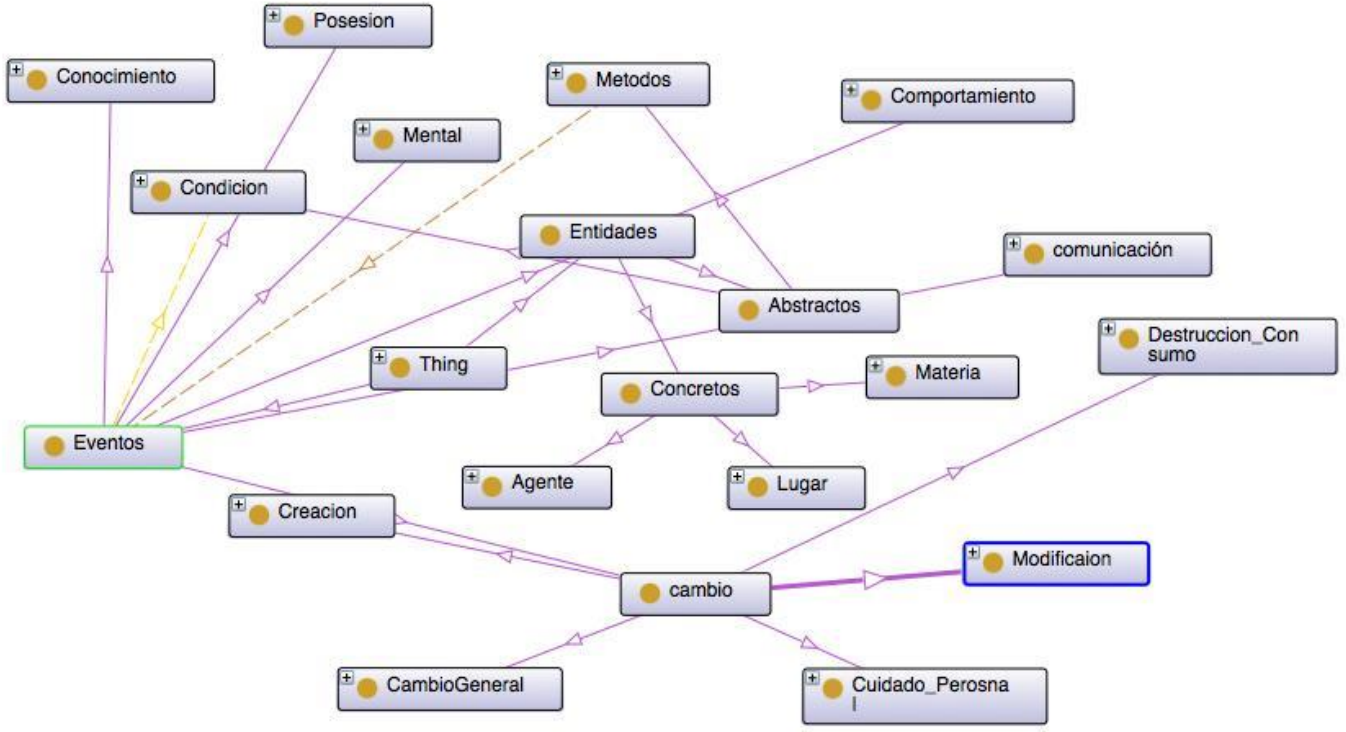


Interpretive Ontology



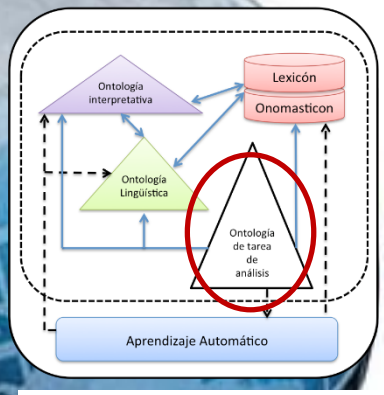
MODS ENTIDADES	Definición [10]
Abstractos	Los abstractos pueden ser números, conjuntos, definiciones
Concretos	Los concretos son objetos físicos o que se pueden definir en algo específico (por ejemplo, el planeta Venus, ese árbol,

MODS EVENTOS
Comunicación • General (Decir, hablar) • Valoración (Criticar, felicitar) • Mandato (Suplicar, ordenar)
Comportamiento



Cobrar

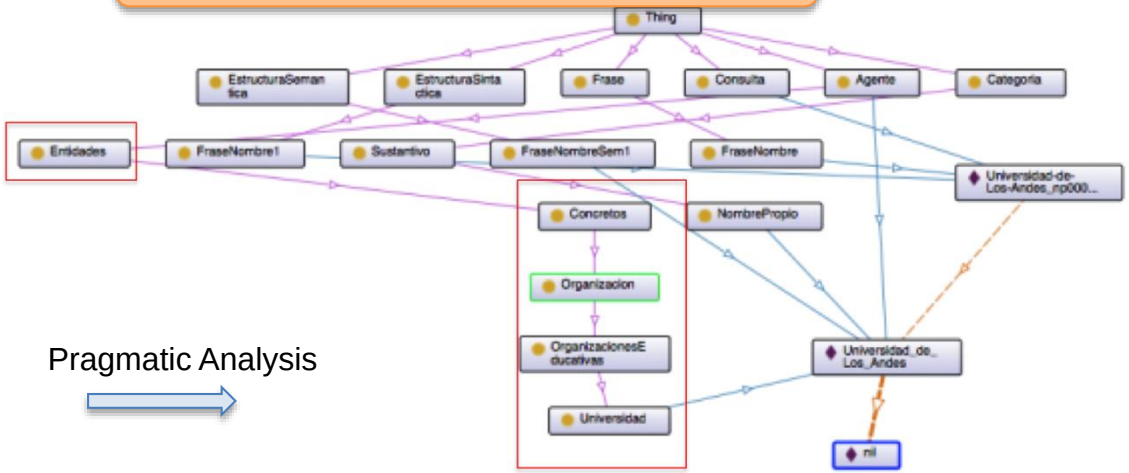
Task Ontology

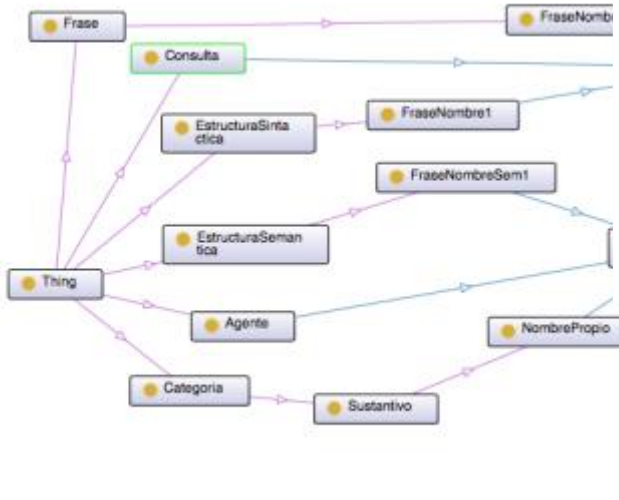


Task Ontology

Ontología interpretativa

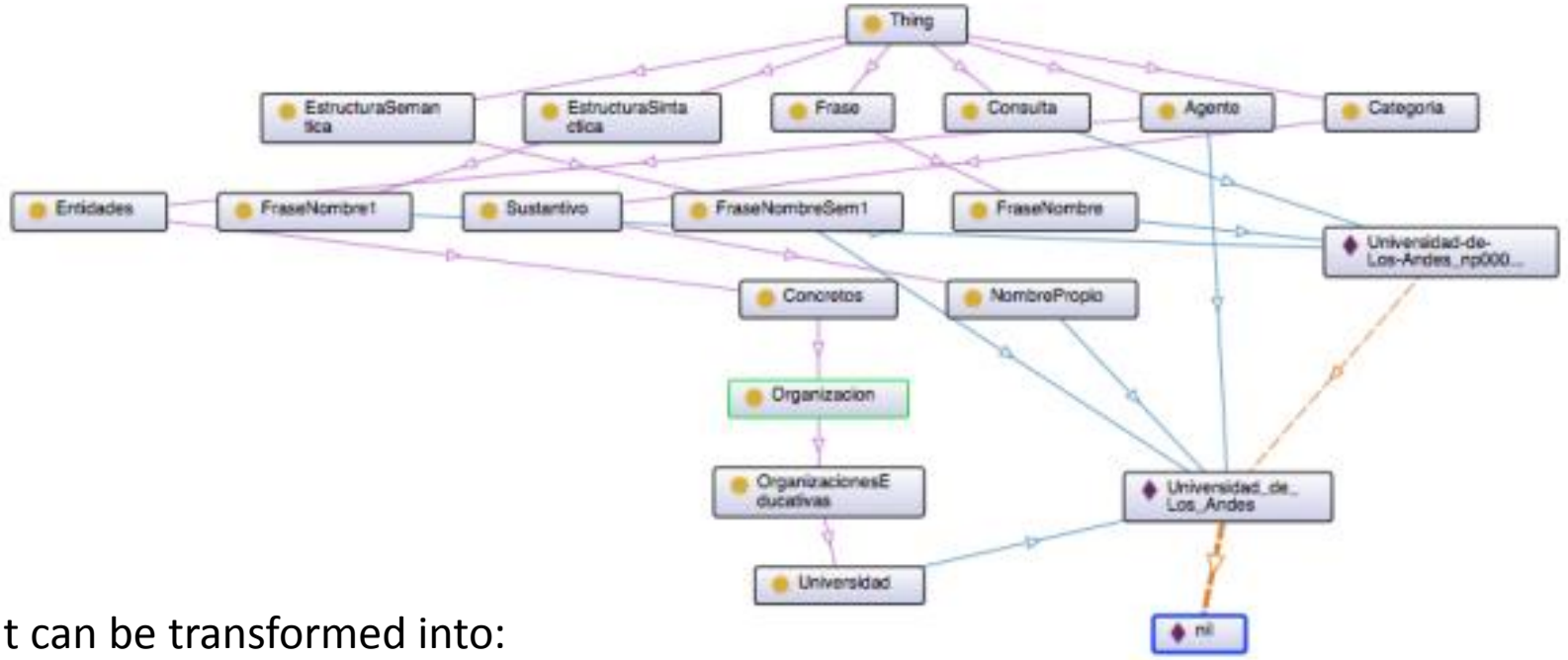
Pragmatic Analysis





Pragmatic Analysis

Interpreted query



It can be transformed into:

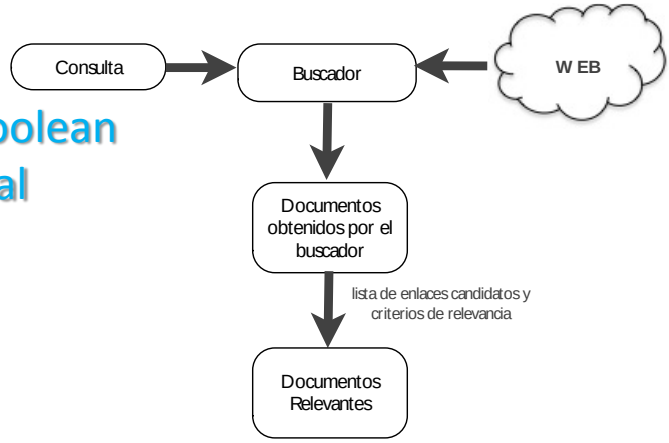
Boolena	Information Recovery Systems
SPARQL SPARQL-DL SeRQL	Ontologies
SQL	For relational database

SPARQL query

```
SELECT ?Clases
    WHERE
    { ?subject j.0:nombre   "Universidad de Los Andes"^^xsd:string .
      ?subject rdf:type  ?Clases}
```

Query with MODS:
Extraction of Relevant Documents from the Web

Relevant documents according to the degree of similarity between the user's query and the recovered documents



To determine this similarity, different models can be used : **Boolean expressions, a vector model, those based on fuzzy logic, neural networks or Bayesian networks, etc.**

Vector model: This model represents the query and the documents as vectors. Thus, a vocabulary of size t will define a t-dimensional space such that:

- a document d_j is represented by a vector $d_j = (w_{1j}, \ldots, w_{tj})$
- a query q is represented as a vector $q = (w_{1q}, \ldots, w_{tq})$

Query with MODS:
Extraction of Relevant Documents from the Web

Weighting of the terms of the documents and of the query

Frequency analysis, number of occurrences of the terms found in the query and in the recovered documents,

TF-IDF weights, the importance of a term to discriminate the document and/or collection of documents.

$$IDF(term) = \log(N/DF)$$

$$TF - IDF = TF(term) * IDF(term)$$

Where: N = number of documents in the collection, DF = number of documents in which the term appears, TF = Frequency of appearance of the term in the document

Frequency and weights determine the similarity

Query with MODS:
Extraction of Relevant Documents from the Web

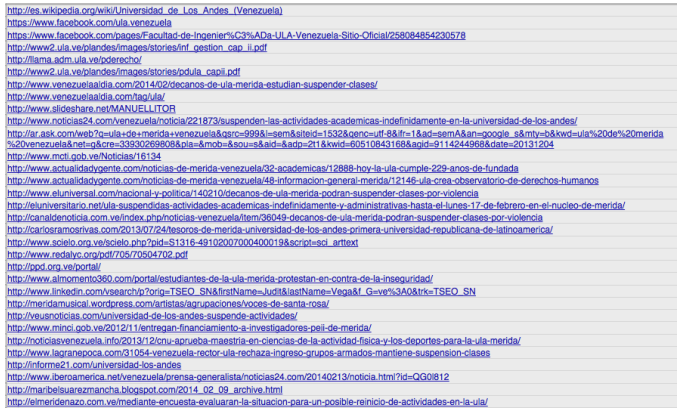
Suppose a user performs the following query in natural language in the google search engine:

“Universidad de Los Andes de Mérida”

MODS gets the following Boolean query:

(“Universidad de los Andes” and Mérida and Venezuela) or (ULA Mérida and Venezuela) or (“Universidad de los Andes” and “Núcleo Mérida” and Mérida and Venezuela) or (ULA and “Núcleo Mérida” and Mérida and Venezuela),

This query is made in the Google search engine, and MODS gets the set of links to documents


http://es.wikipedia.org/wiki/Universidad_de_Los_Andes_(Venezuela)
https://www.facebook.com/ula.venezuela
https://www.facebook.com/pages/Facultad-de-Ingenier%C3%ADa-ULA-Venezuela-Sitio-Oficial/258084854230578
http://www2.ula.ve/plandes/images/stories/inf_gestion_cap_ii.pdf
http://fiama.adm.ula.ve/derecho/
http://www2.ula.ve/plandes/images/stories/pdula_capii.pdf
http://www.venezuelaaldia.com/2014/02/decanos-de-ula-merida-estudian-suspender-clases/
http://www.venezuelaaldia.com/tag/ula/
http://www.slideshare.net/MANUELLITOR
http://www.noticias24.com/venezuela/noticia/221873/suspenden-las-actividades-academicas-indefinidamente-en-la-universidad-de-los-andes/
http://ar.ask.com/web?q=ula+de+merida+venezuela&qsrc=999&l=sem&siteid=1532&qenc=utf-8&lfr=1&ad=semA&an=google_s&mtv=b&kwd=ula%20de%20merida%20venezuela&net=g&cre=33930269805&pla=&mob=&sou=s&aid=&adp=21&xwid=60510843168&agid=911424496&date=20131204
http://www.mcti.gob.ve/Noticias/16134
http://www.actualidadygente.com/noticias-de-merida-venezuela/32-academicas/12888-hoy-la-ula-cumple-229-anos-de-fundada
http://www.actualidadygente.com/noticias-de-merida-venezuela/48-informacion-general-merida/12146-ula-crea-observatorio-de-derechos-humanos
http://www.eluniversal.com/nacional-y-politica/140210/decanos-de-ula-merida-podran-suspender-clases-por-violencia
http://eluniversitario.net/ula-suspendidas-actividades-academicas-indefinidamente-y-administrativas-hasta-el-lunes-17-de-febrero-en-el-nucleo-de-merida/
http://canaldenoticia.com.ve/index.php/noticias-venezuela/item/36049-decanos-de-ula-merida-podran-suspender-clases-por-violencia
http://cariosramosrivas.com/2013/07/24/tesoros-de-merida-universidad-de-los-andes-primera-universidad-republicana-de-latinoamerica/
http://www.scielo.org.ve/scielo.php?pid=S1316-49102007000400019&script=sci_arttext
http://www.redalyc.org/pdf/705/70504702.pdf
http://ppd.org.ve/portal/
http://www.almomento360.com/portal/estudiantes-de-la-ula-merida-protestan-en-contra-de-la-inseguridad/
http://www.linkedin.com/vsearch/p?orig=TSEO_SN&firstName=Judit&lastName=Vega&f_G=ve%3A0&trk=TSEO_SN
http://meridamusical.wordpress.com/artistas/agrupaciones/voces-de-santa-rosa/
http://veunoticias.com/universidad-de-los-andes-suspende-actividades/
http://www.minci.gob.ve/2012/11/entregan-financiamiento-a-investigadores-peii-de-merida/
http://noticiasvenezuela.info/2013/12/cnu-aprueba-maestria-en-ciencias-de-la-actividad-fisica-y-los-deportes-para-la-ula-merida/
http://www.lagranepoca.com/31054-venezuela-rector-ula-rechaza-ingreso-grupos-armados-mantiene-suspension-clases
http://informe21.com/universidad-los-andes
http://www.iberoamerica.net/venezuela/prensa-generalista/noticias24.com/20140213/noticia.html?id=QG0I812
http://maribelsuarezmancha.blogspot.com/2014_02_09_archive.html
http://elmeridenazo.com.ve/mediante-encuesta-evaluaran-la-situacion-para-un-posible-reinicio-de-actividades-en-la-ula/

Query with MODS:
Extraction of Relevant Documents from the Web

Accuracy: proportion of documents retrieved that are relevant.

$$Accuracy = \frac{Recovered\ relevant\ documents}{Recovered\ documents}$$

	Yahoo	Google
Accuracy	*0,2716*	*0,3699*

Our system gives all documents as relevant (100% accuracy).

Random Queries	Yahoo	Google
Query 1	*0,2456*	*0,311*
Query 2	*0,299*	*0,323*
Query 3	*0,253*	*0,312*
Query 4	*0,2612*	*0,3419*

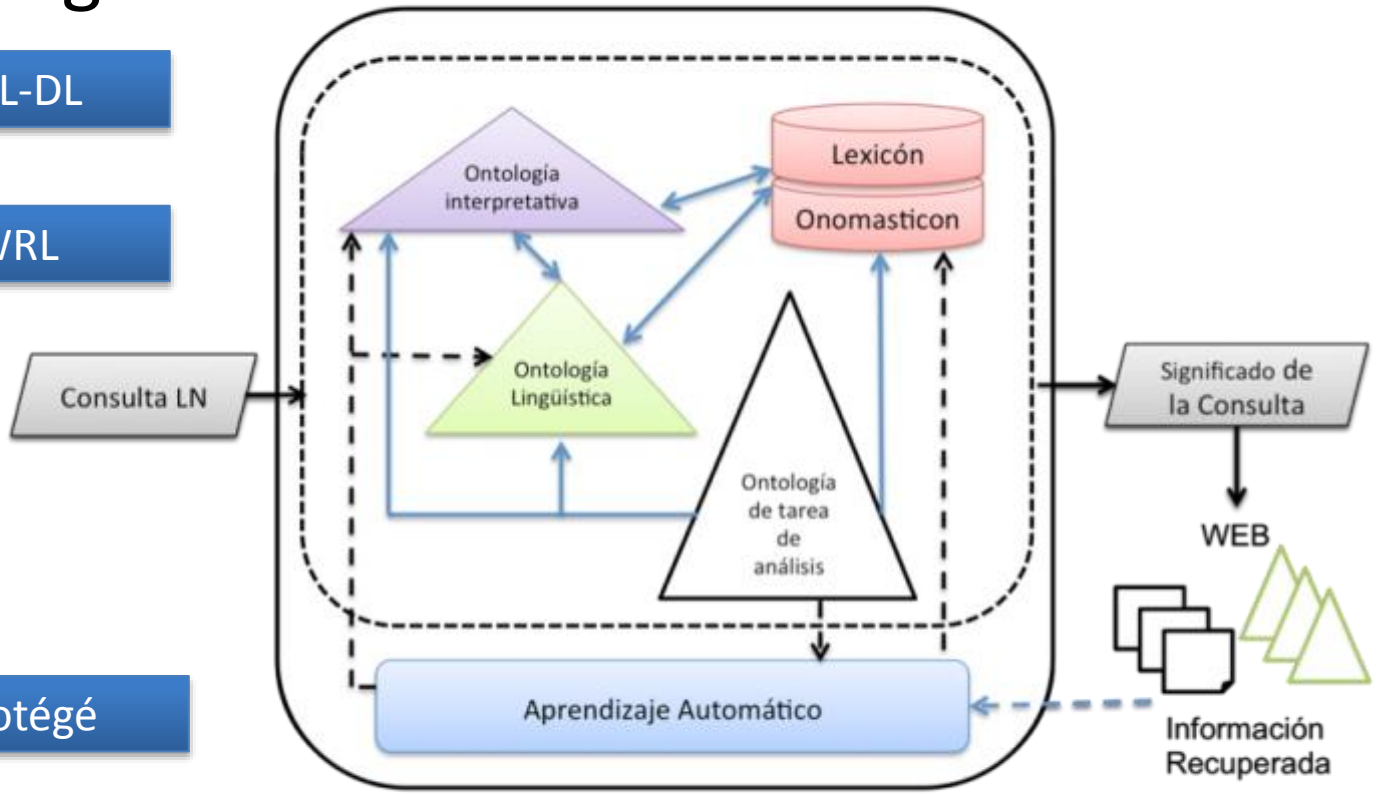
Tools: MODS Design

Ontologies

OWL-DL

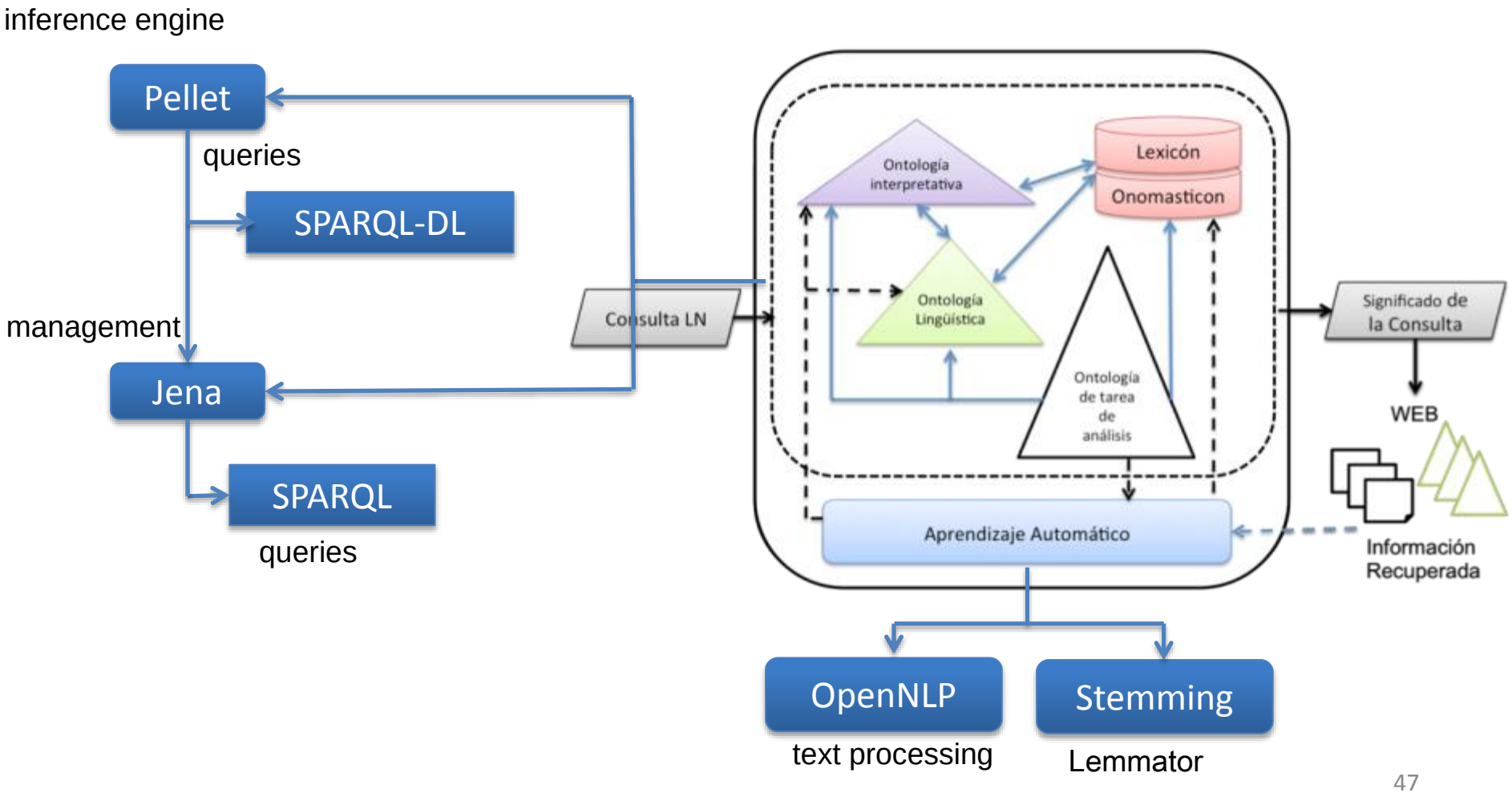
SWRL

Protégé

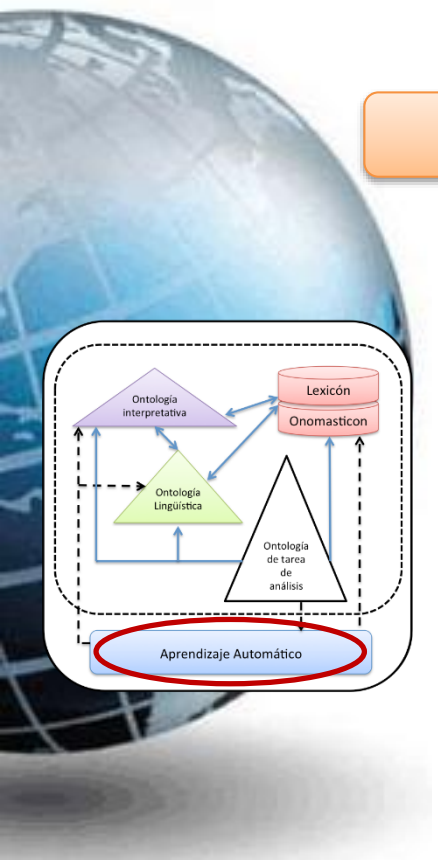


r

Tools: Implementation of MODS



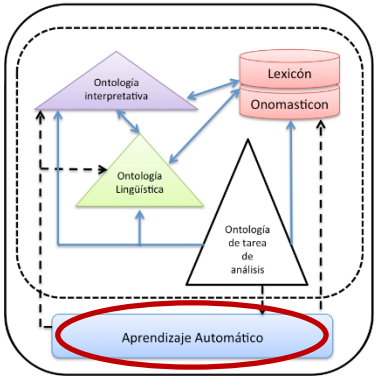
Task Ontology

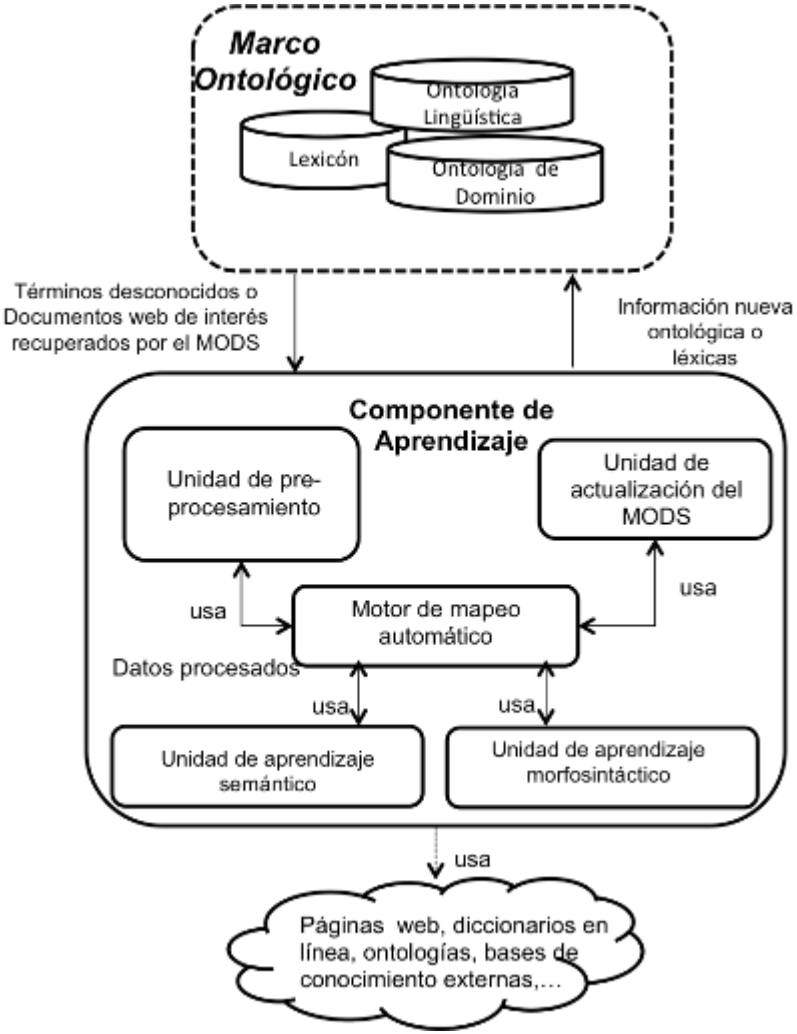


Task Ontology

Invokes Learning

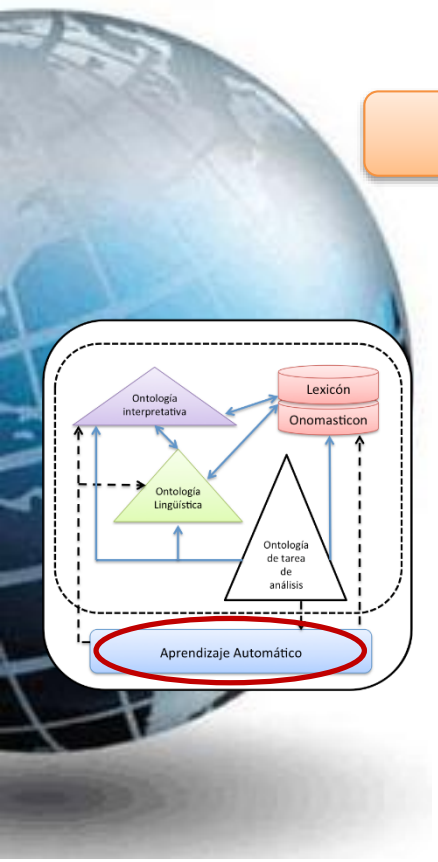
Learning





- ✓ **Unknown terms**
  - ✓ Automatic mapping engine
  - ✓ Morphosyntactic learning unit
- ✓ **Recovered documents**
  - ✓ Automatic mapping engine
  - ✓ Semantic learning unit

Task Ontology



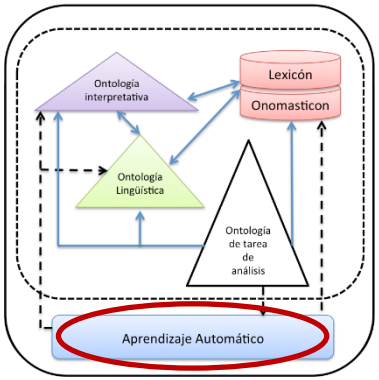
Task Ontology

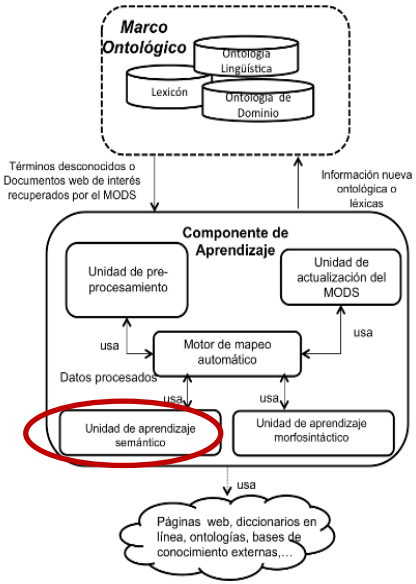
Invokes Learning

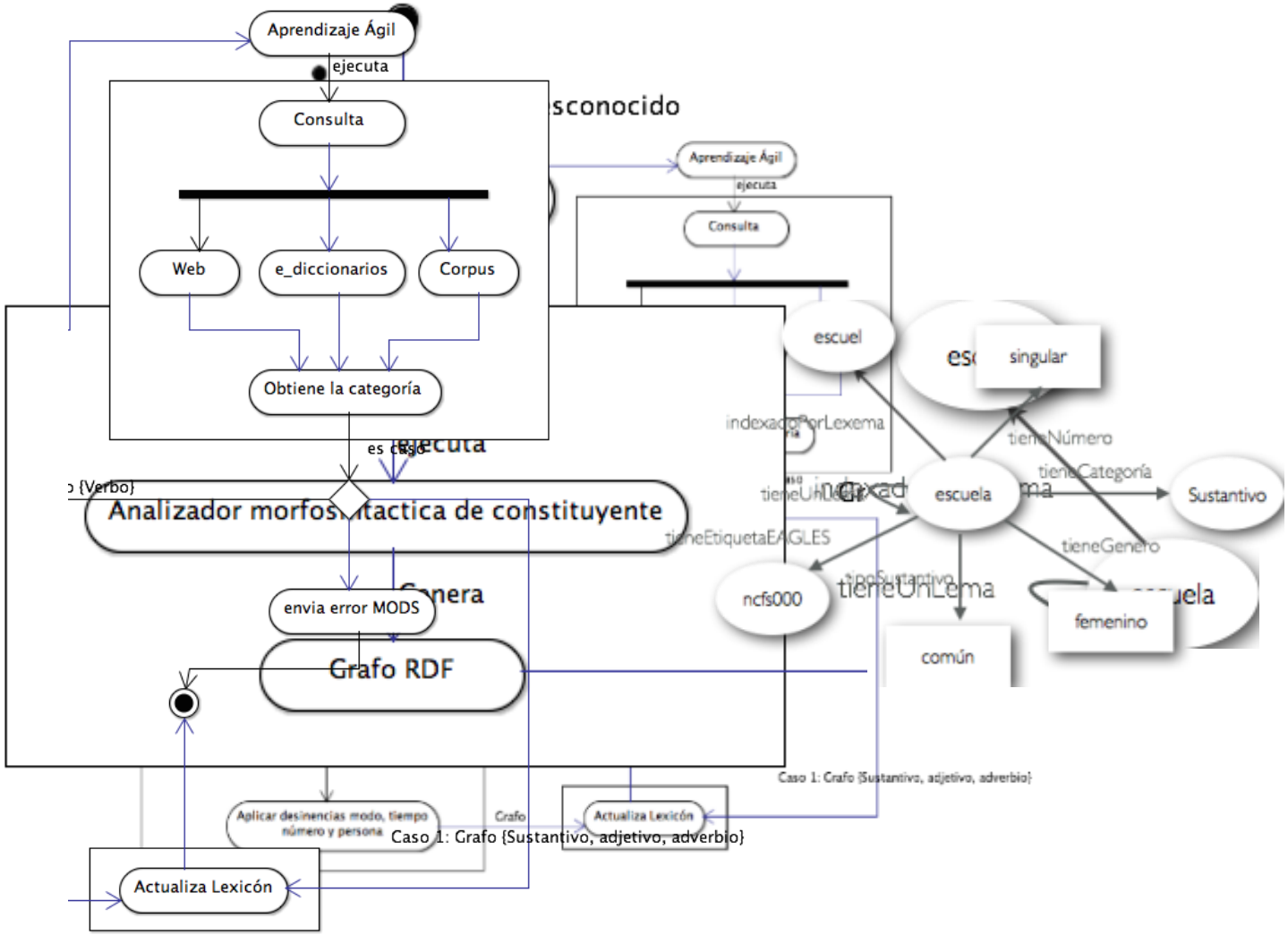
Invokes Morphosyntactic Learning

Unknown term

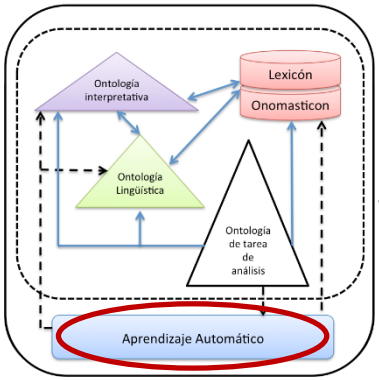
Morphosyntactic Learning

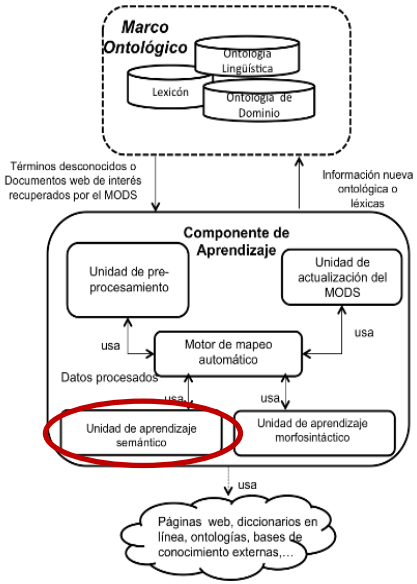






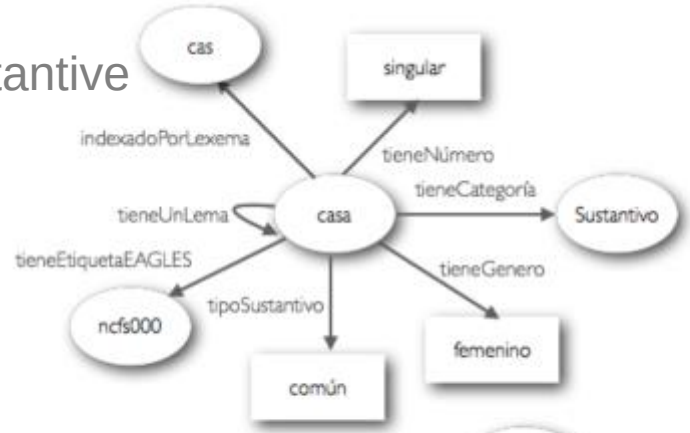
Morphosyntactic Learning

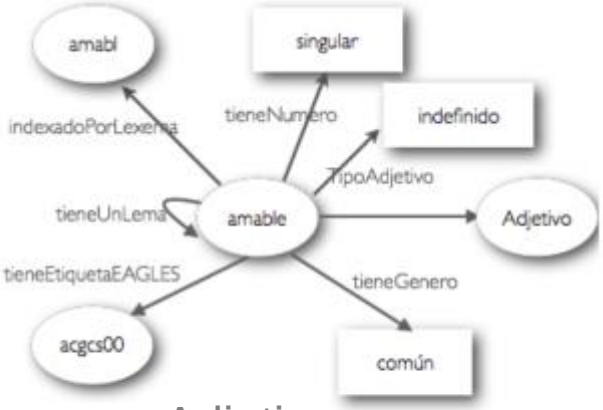




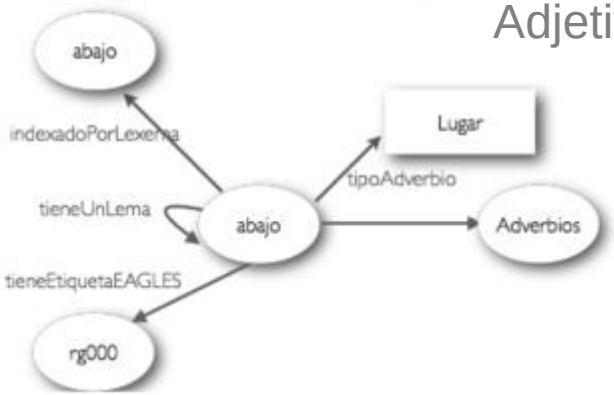
Substantive, Adjective or Adverb Learning

Sustantive



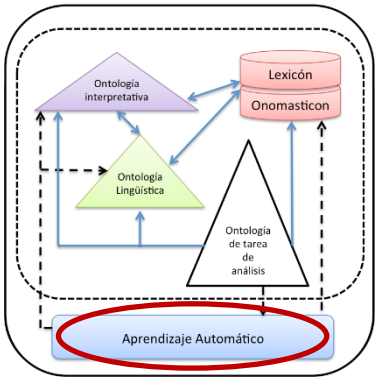


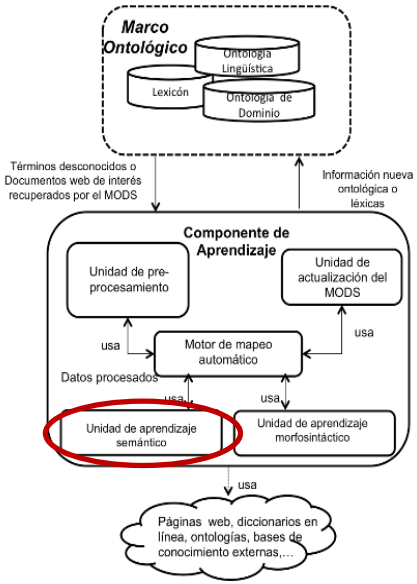
Adjetive

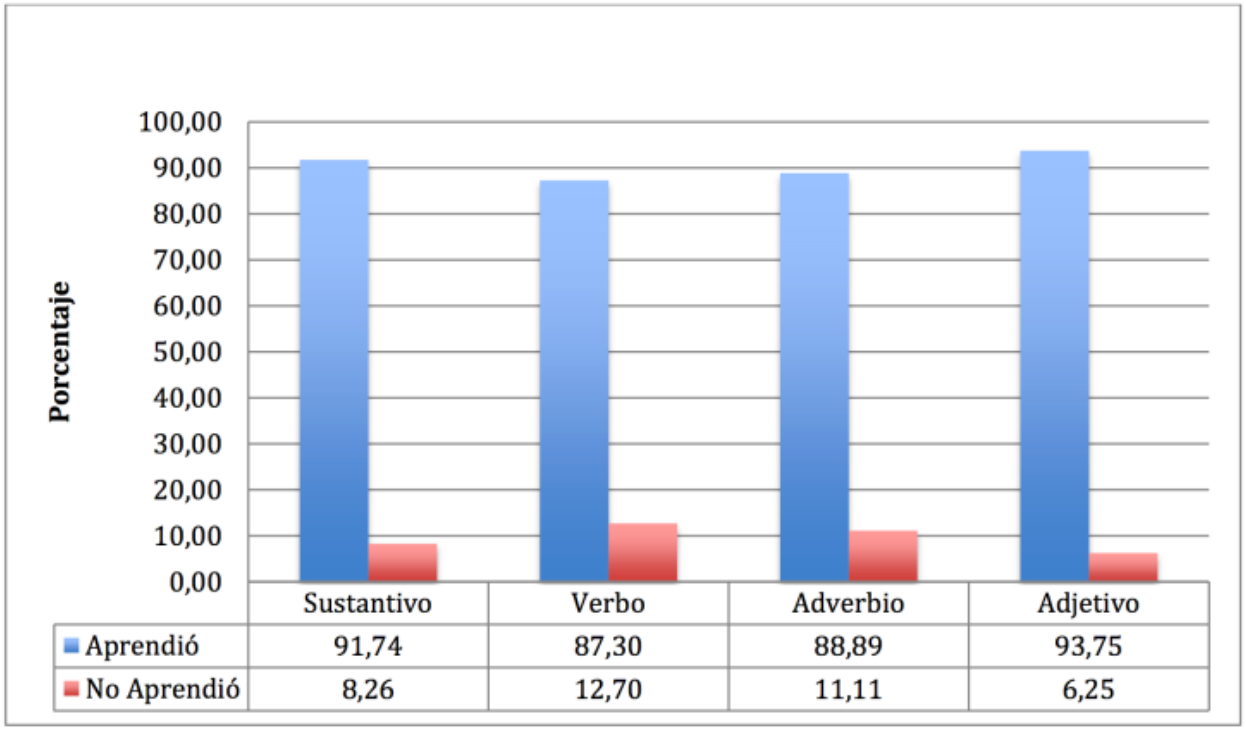


Adverb

Morphosyntactic Learning



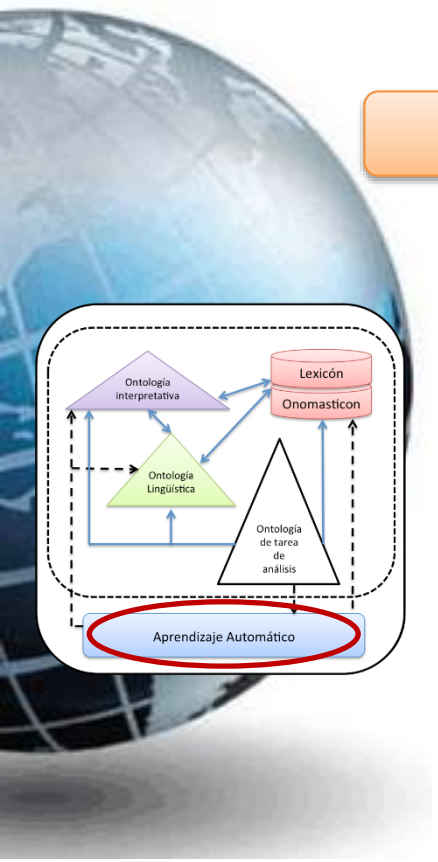




	Sustantivo	Verbo	Adverbio	Adjetivo
Aprendió	91,74	87,30	88,89	93,75
No Aprendió	8,26	12,70	11,11	6,25

random queries with unknown terms

Task Ontology



Task Ontology

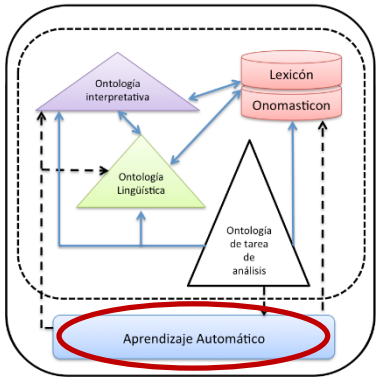
Invokes Learning

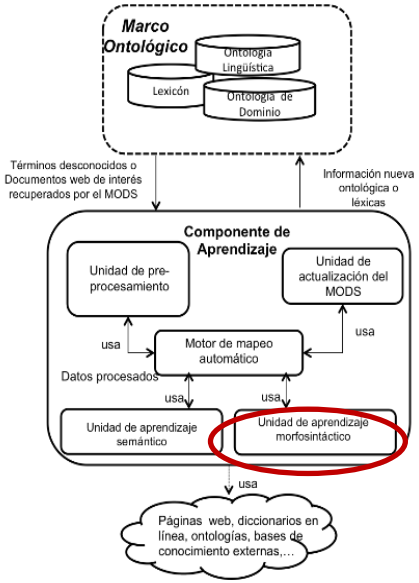
Invokes Morphosyntactic Learning

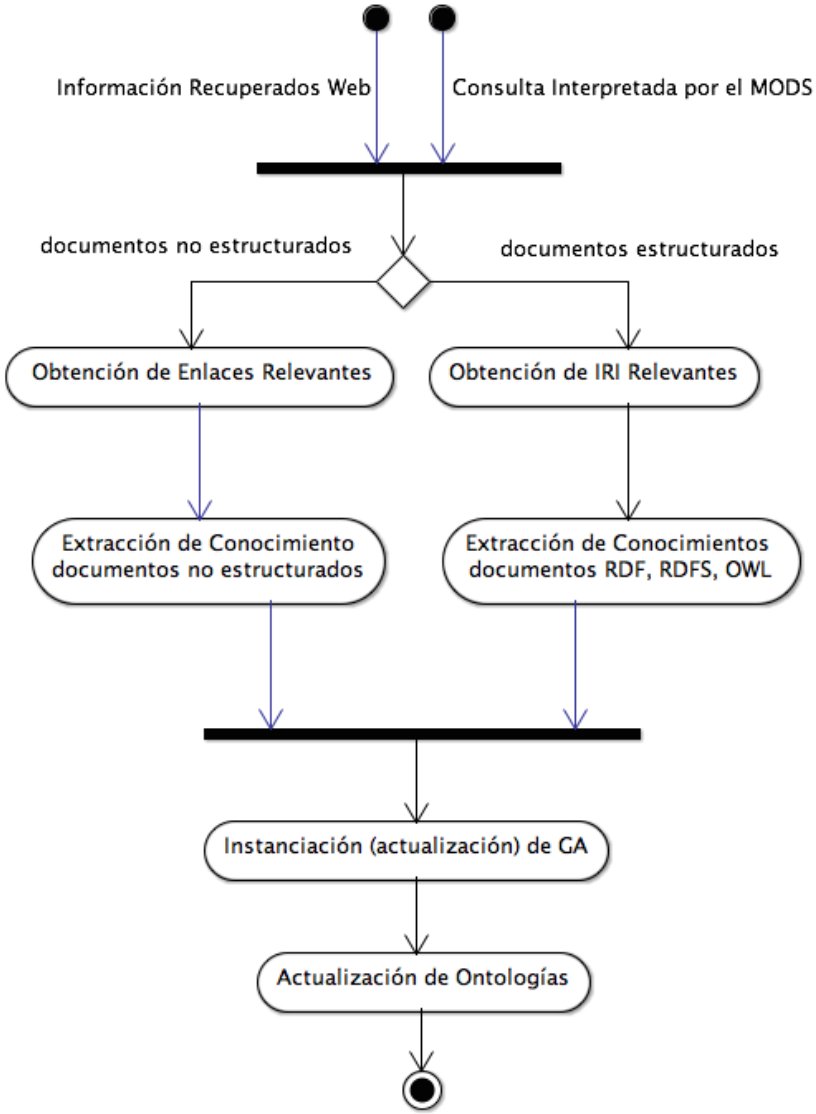
Invokes Semantic Learning

Recovered documents

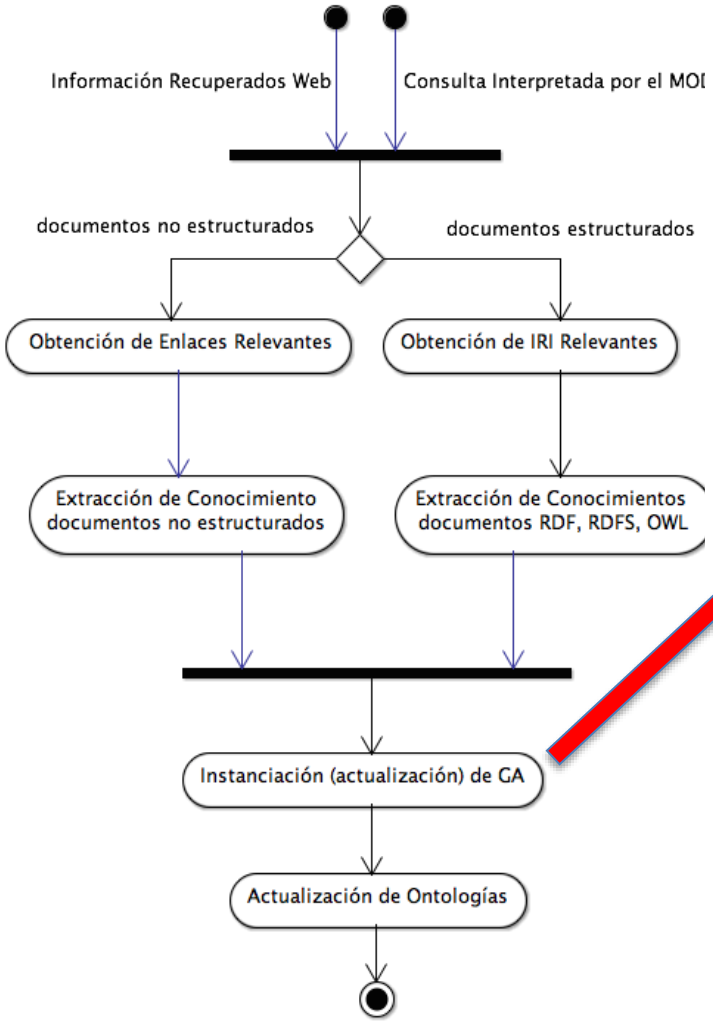
Semantic Learning

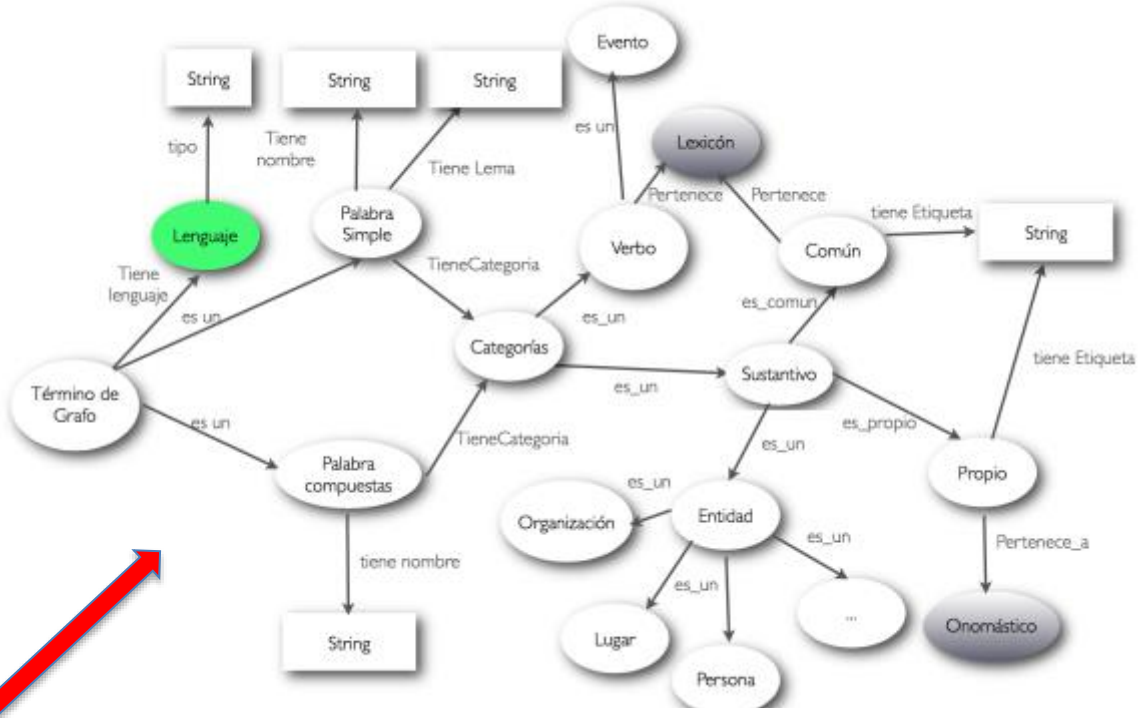


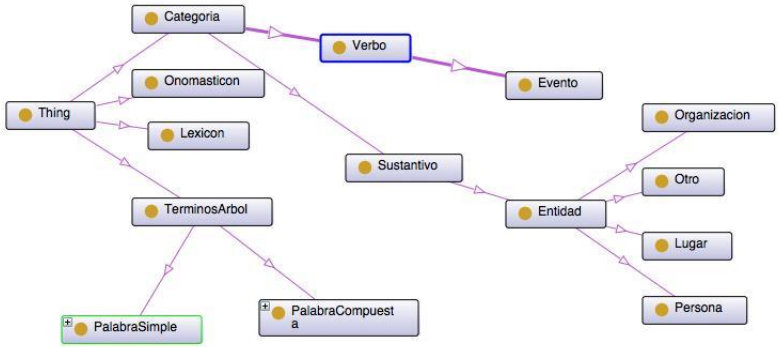




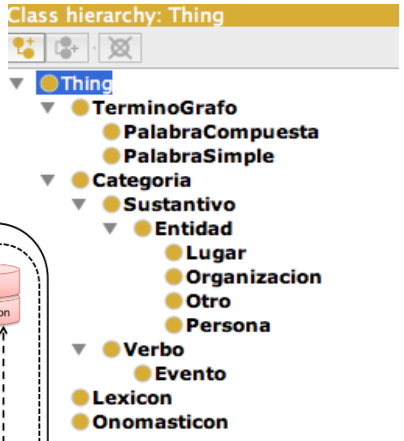
Learning Graph

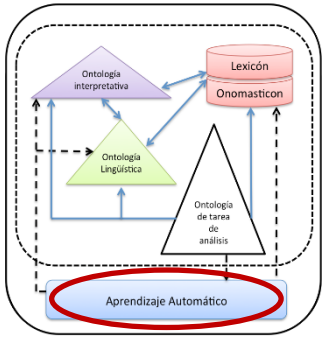






Learning Graph





The graph has **the entities and relationships relevant**, which is calculated

$$w_j = tf_j * fi_j$$

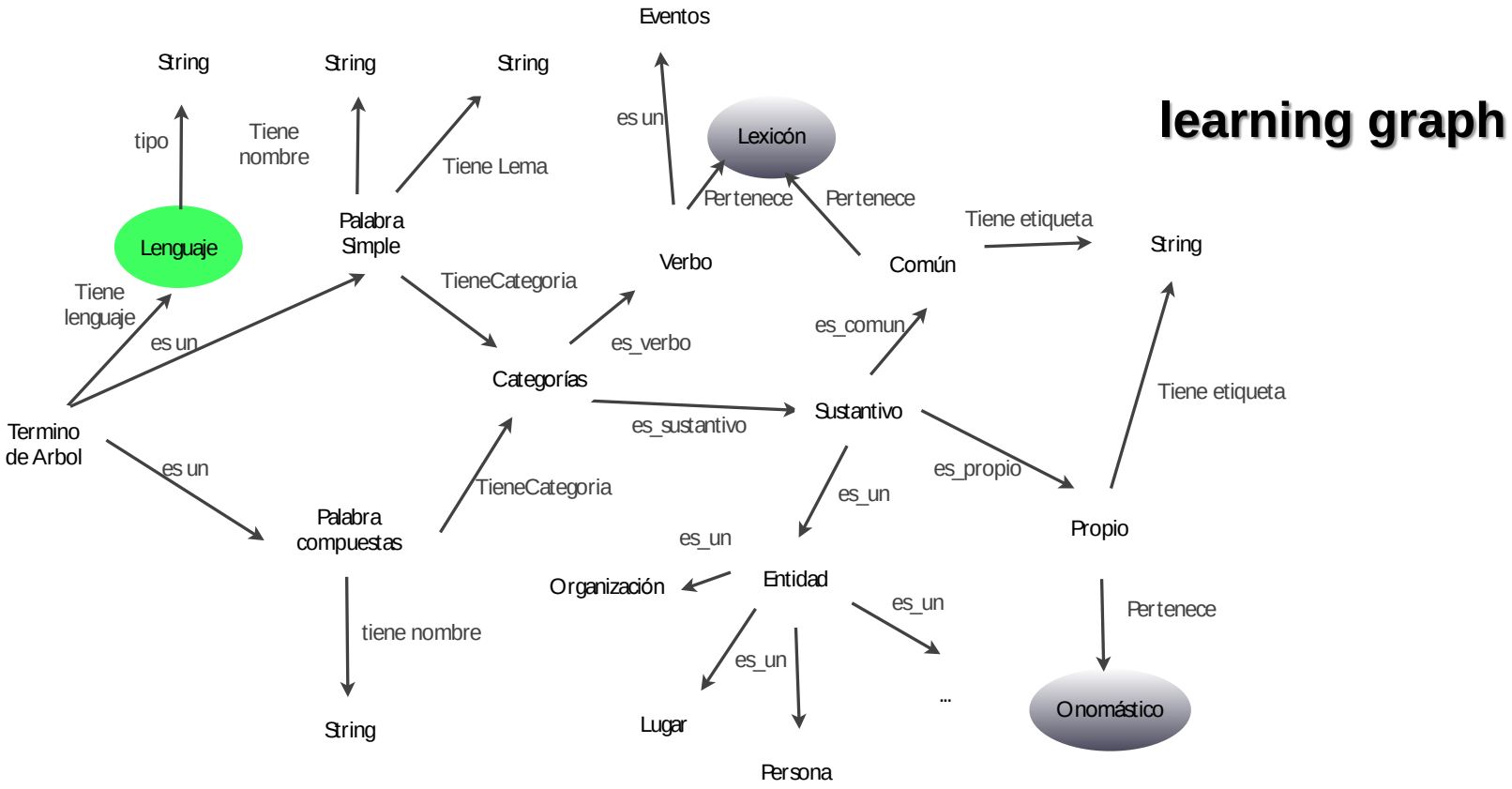
tf_i *average* TF-IDF *in the processed texts*

$$fi_j = \log\left(\frac{processed\ texts}{texts\ with\ the\ term}\right)$$

It will be relevant if its weight is greater than or equal to the average of the weights.

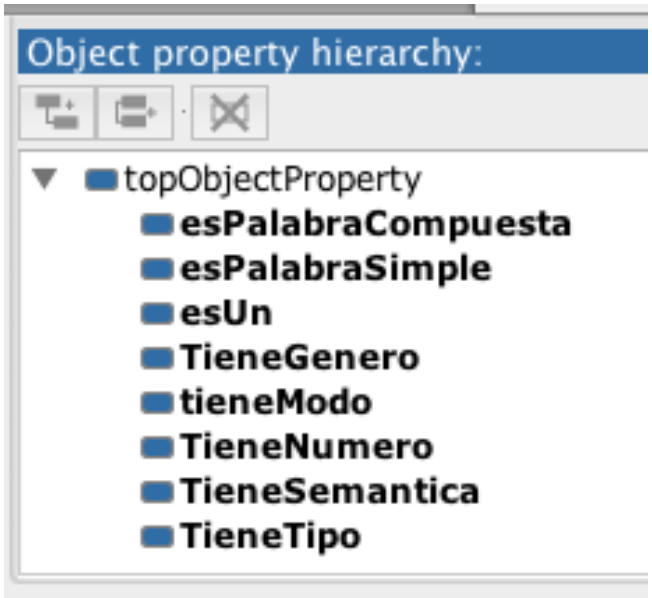
Semantic Mining based on the Learning Graph

The learning graph is composed of a set of basic axioms to infer new knowledge.



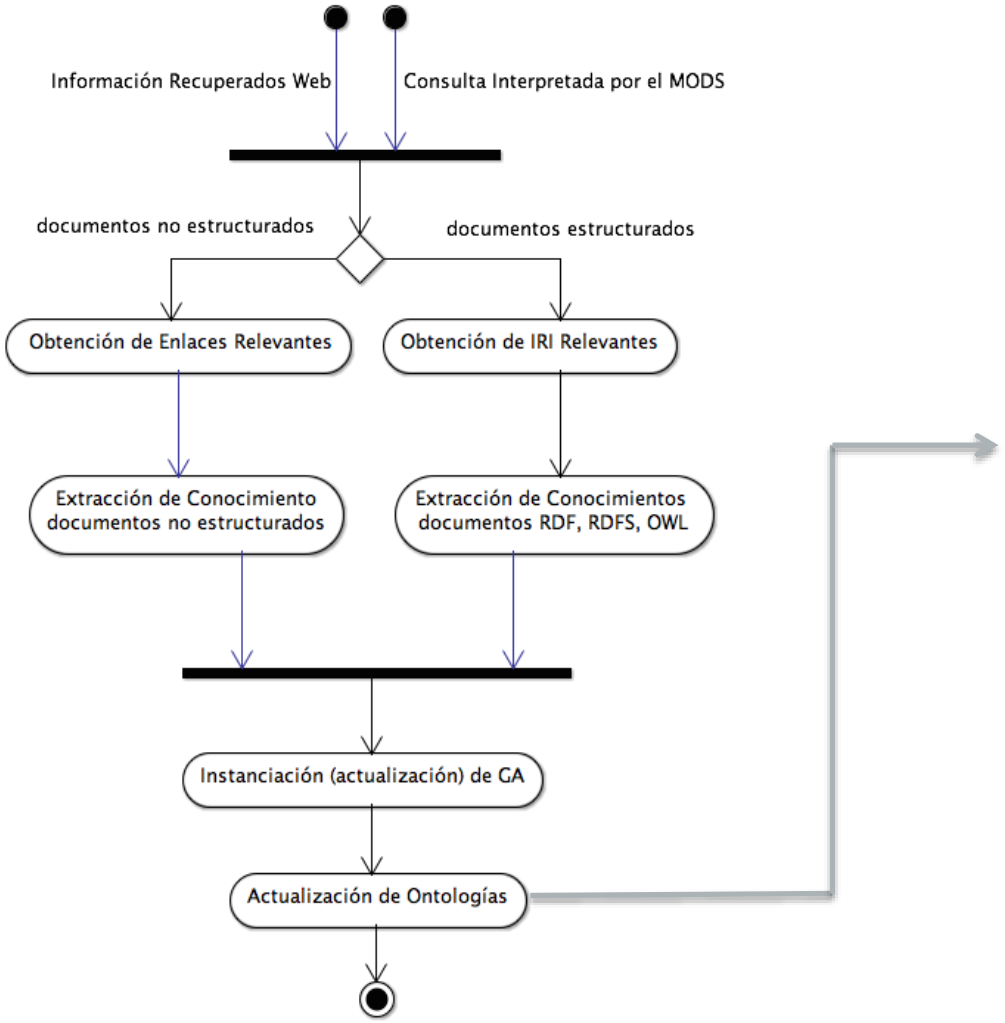
Semantic Mining based on the Learning Graph

The classes defined in the learning graph

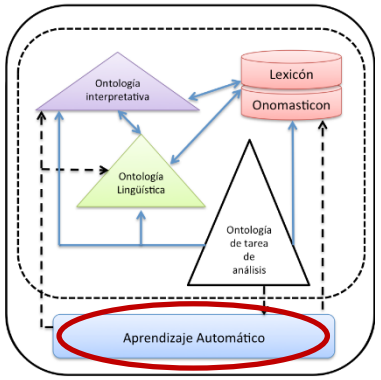


The relationships defined in the learning graph

Update of Ontologies



- Lexicon
- Onomasticon
- Interpretive



Semantic Learning

("Universidad de los A

Doc	Enlace	Similitud
73	http://erevistas.saber.ula.ve/index.php/visiongere	0,97

Entidades Relevantes

nombre	peso
ula	183.88309208431204
jurado	54.66794629533601
artículo	35.8351893845611
créditos	32.30378644724401
miembros	32.30378644724401
reglamento	24.849066497880003
Correo	22.364159848092005
Electrónico	22.364159848092005
Nombres	22.364159848092005
Doctoral	21.50111363073666
profesores	21.50111363073666
trabajo	21.50111363073666
tutor	20.873633484694086
aspirante	20.79441541679836
José	19.879253198304003
actividades	19.879253198304003
co	19.879253198304003
lapso	19.879253198304003
doctorado	19.775021196025975
grado	19.709354161508603
investigación	17.577796618689757
Carlos	17.394346548516

Relaciones Relevantes

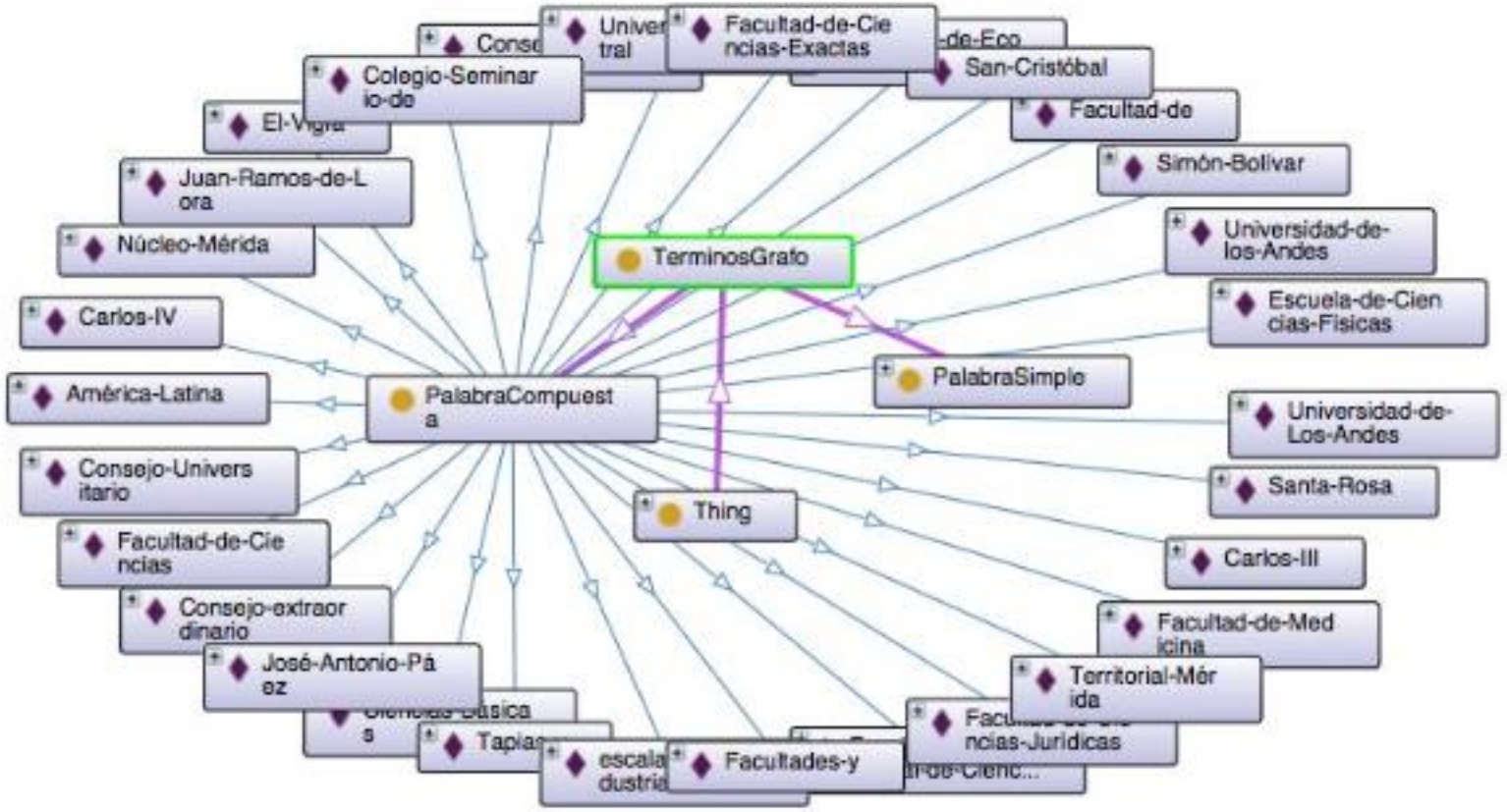
Relación	Peso
VE	176.42837213494803
SERÁ	37.62694885378915
SER	29.112181583517703
PODRÁ	24.849066497880003
PRESENTAR	21.50111363073666
PARTE	19.879253198304003
DEBERÁ	19.709354161508603
PODRÁN	14.909439898728003
TENDRÁ	14.909439898728003
DEBERÁN	14.33407575382444
ESTARÁ	12.424533248940001
SERÁN	12.424533248940001
ELABORAR	9.939626599152001
TENDRÁN	9.939626599152001
DEBE	9.704060527839234
ESTAR	7.4547199493640015
HABER	7.4547199493640015
NOMBRAR	7.4547199493640015
APROBADO	7.16703787691222
CUMPLIR	7.16703787691222
DEBEN	7.16703787691222
POSEER	6.931471805599453

Web Mining based on the Learning Graph

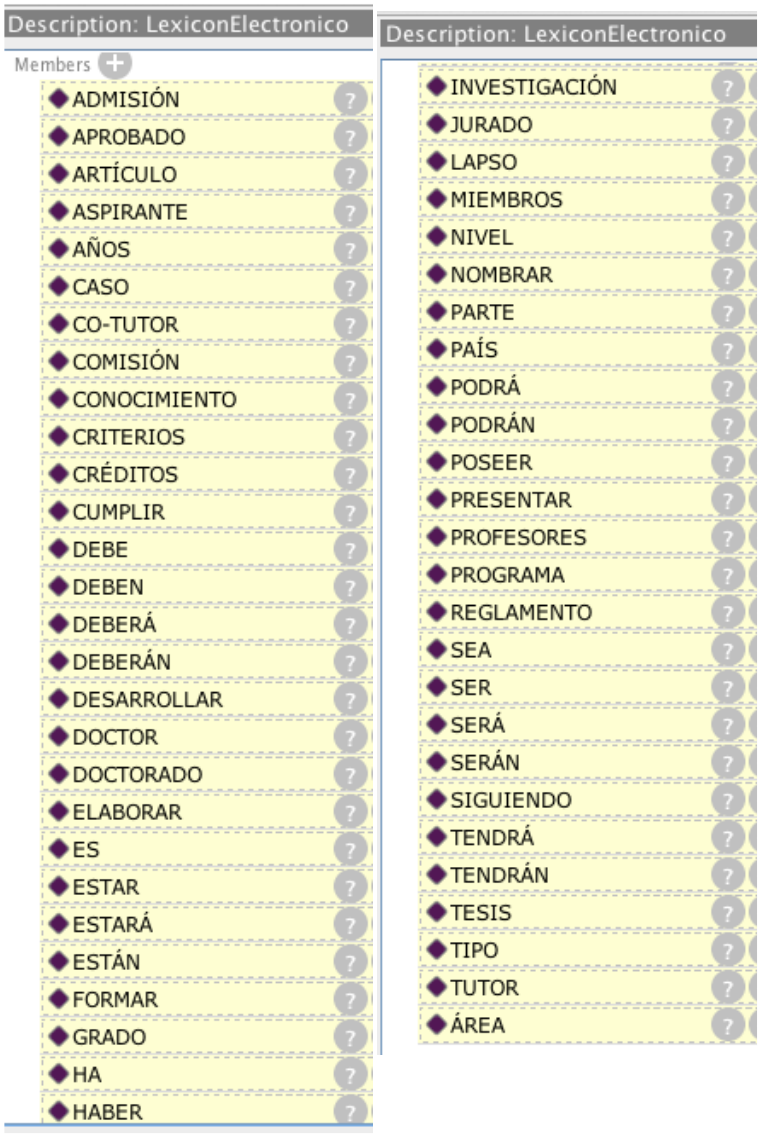
- Automatic creation of the **terminological base** of a domain
- **is_a** Relationship detection (superclass-subclass),
- Creation of **electronic lexicons** for nouns and verbs.

Semantic mining

- ✓ Build a **Terminological Base** in a specialized domain.



Semantic Mining based on the Learning Graph



Electronic lexicon of nouns and verbs

Semantic Mining based on the Learning Graph

Example: websites analyze

Procesamiento de texto no estructurado

Introducir el Texto

La Universidad de Los Andes es una universidad pública y autónoma ubicada en los andes venezolanos con su sede principal y rectorado en la ciudad de Mérida; fundada por el clero como casa de estudios el 29 de marzo de 1785, elevada luego a seminario y finalmente reconocida como Universidad el 21 de septiembre de 1810 bajo decreto expedido por la Junta Gubernativa de la provincia de la Corona de España.
Es una de las principales universidades de Venezuela por la cantidad de estudiantes que alberga, por su nivel académico y por sus aportes en investigación que han contribuido al estudio y desarrollo de las ciencias. La universidad tiene como propósito fortalecer la formación integral iniciada en los ciclos de educación primaria y secundaria, además de formar equipos profesionales y técnicos necesarios para el desarrollo y progreso de Venezuela.
La universidad está conformada por 11 facultades repartidas en el Núcleo Mérida (ubicado en la ciudad de Mérida), 3 núcleos autónomos localizados en las ciudades de San Cristóbal, Trujillo y El Vigía, dos extensiones universitarias con estudios de pregrado, postgrado y actualización profesional en Tovar y en Valera, extensiones de actualización profesional en las ciudades de Barinas, Guanare, Barquisimeto, Maracaibo, Caracas, entre otras, y diversas instalaciones universitarias dentro del territorio nacional como estaciones experimentales, haciendas de producción agricolas, reservas naturales para el desarrollo de la fauna y flora y laboratorios de investigación.

Separar Sentencia

Unstructured text

Reconocimiento de Entidades y Relaciones Candidatas

Entidades Candidatas

Universidad de Los Andes	universidad	andes	sede
ciudad	Mérida	universidades	Venezuela
cantidad	estudiantes	nivel	aportes
investigación	estudio	desarrollo	ciencias
propósito	formación	ciclos	educación
equipos	progreso	facultades	Núcleo
núcleos	ciudades	San Cristóbal	Trujillo
El Vigía	extensiones	estudios	pregrado
actualización	Tovar	Valera	Barinas
Guanare	Barquisimeto	Maracaibo	Caracas
instalaciones	territorio	estaciones	haciendas
producción	reservas	fauna	flora
laboratorios			

Relaciones Candidatas

es	Es	alberga	han	
contribuido	tiene	fortalecer	formar	
está				

Actualizar el Arbol

Universidad de Los Andes
Desarrollado por Tatiana Rodríguez y Jose Aguilar

Entities and Relations Candidates

Semantic Mining based on the Learning Graph

Example: websites analyze

	Total
Candidate Entities	251
Candidate Relations	121

	Frecuencia	Peso	Porcentaje de Relevancia
Consejo Directivo	38	36,83	96,92
tesis	25	28,78	75,73
jurado	21	25,76	67,80
programa	21	25,76	67,80
tutor	21	25,76	67,80
Aspirante	18	23,29	61,28
doctorado	17	22,42	58,99
investigación	17	22,42	58,99
Plan de Formación	17	22,42	58,99
miembro	15	20,59	54,19
profesor	15	20,59	54,19
examen	14	19,64	51,68
créditos	12	17,64	46,42
caso	11	16,58	43,64
Área	10	15,49	40,76
estudiante	10	15,49	40,76
Estudios	10	15,49	40,76
grado	10	15,49	40,76
lapso	10	15,49	40,76
Programa de doctorado	10	15,49	40,76
artículo	9	14,35	37,77
comisión	9	14,35	37,77
conocimiento	9	14,35	37,77
doctoral	9	14,35	37,77
Facultad de Ingeniería	9	14,35	37,77
informe	9	14,35	37,77
postgrado	9	14,35	37,77
publicación	9	14,35	37,77
Año	8	13,17	34,65
candidatura	8	13,17	34,65
consejo	8	13,17	34,65
doctor	8	13,17	34,65
grupo	8	13,17	34,65
mes	8	13,17	34,65
reglamento	8	13,17	34,65
investigador	7	11,93	31,39
nivel	7	11,93	31,39
tipo	7	11,93	31,39
Admisión	6	10,63	27,96
criterios	6	10,63	27,96
curso	6	10,63	27,96
país	6	10,63	27,96
Programa de doctorado en ciencias aplicadas	6	10,63	27,96
actividades	5	9,25	24,34

Relevant entities with weight criteria> = 10.06

build text summaries

Web Mining based on the Learning Graph

Example: Ontologies Construction

In the analyzed texts of the Doctorate website, we find the following sentences that have the verb 'to be‘

The **doctoral student** is totally immersed in the dynamics of the research group to which his tutor belongs and follows the guidelines previously established by him in the training plan.

Any **qualified researcher** who is a member of a consolidated research group at the Universidad de Los Andes is, potentially, **a tutor** of the program.

If the **Training Plan** is not accepted by the Admission Committee, the applicant and his/her tutor can modify it and submit it once again to the Commission for consideration, within a period of one month.

Web Mining based on the Learning Graph

Example 2: Ontologies Construction

For each candidate sentence, the morphosyntactic analysis determines.

- **In sentence 1** cannot be established a relationship among the entities.
- **In sentence 2** is established the relationship that researcher is a tutor

- **In sentence 3** cannot be established a relationship the verb is: "is accepted“

Ontology Mining

Ontology Mining

To explore techniques that can extract additional knowledge from a set of ontologies, to achieve a wider domain of knowledge.

- Extraction of knowledge patterns,
- Build or enrich ontologies.
- Establish relationships between ontologies
- ...

Ontology Mining

- **Extraction of Rules**
- **Integration of Ontologies**
  - **Linked Ontologies**
  - **Merge of Ontologies**
  - **Ontology Alignment**

Our works

Alignment of ontologies

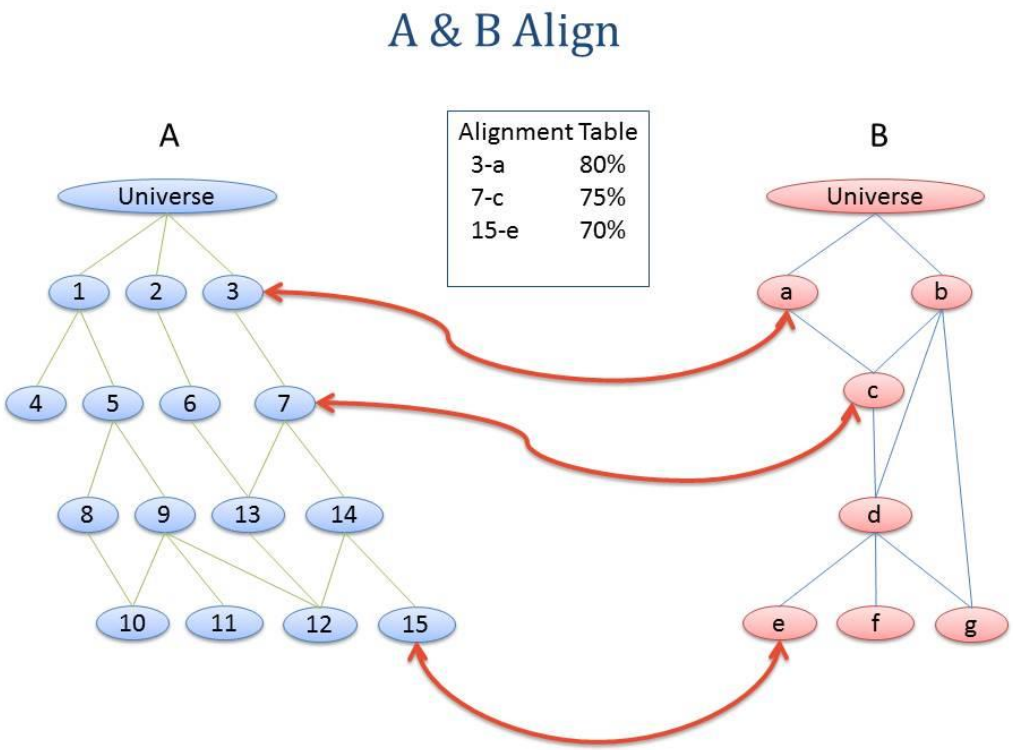
Identify concepts of an ontology that are similar in the other ontologies

Semantic distance between each pair of concepts in different ontologies

Methods and tools for the alignment of ontologies

Alignment of ontologies

ontologies alignment techniques finds the correspondences between the concepts of the ontologies.



Alignment of ontologies

Ontology alignment techniques

Distance of Levenshtein

- **Based on linguistic matching**
  - Similarity based on terms.
  - Semantic similarity:
    - Similarity between properties of the classes
    - Similarity between super-classes

- **Based on graph matching**

Example of Method to Calculate Near Concepts

Assume that CA is a concept in the ontology A and PA its predecessor. CB is a concept CB in the ontology B (PB is its predecessor).

Four cases to calculate the similarity:

Case A: The CA concept coincides with CB in B, and the predecessors PA and PB

Case B: PA matches PB, but there is no match between CA and CB.

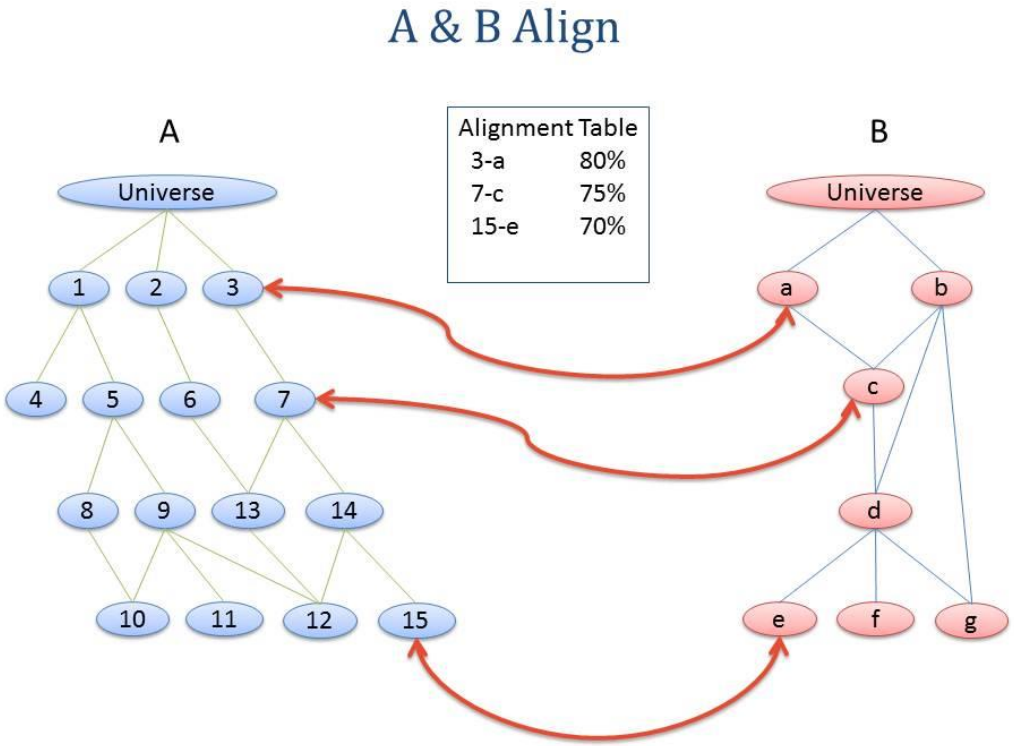
Case C: CA matches CB, but there is no match between PA and PB.

Case D: CA does not coincide with the CB and PA does not coincide with PB.

Ontology Alignment Recommendation System using ABC algorithm

The problem is to determine which alignment technique should be used at a given context.

We propose an **ABC** based technique, which **automatically select** the proper alignment **technique**



ABC Algorithm: the main steps

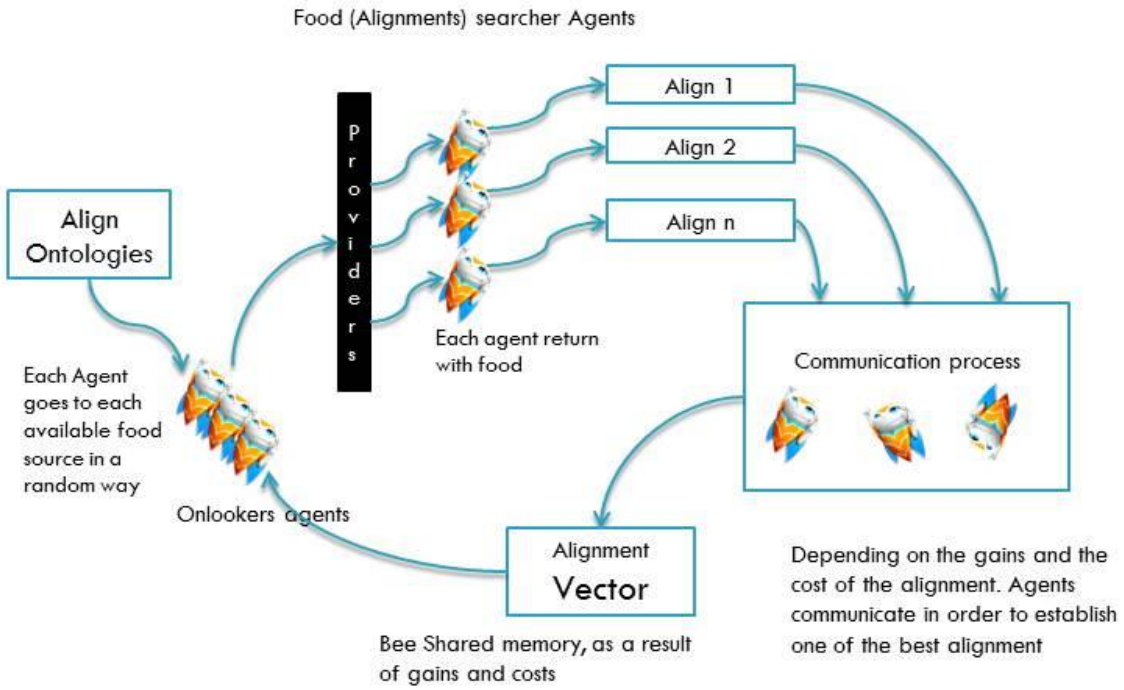
The algorithm based on the Colonies of Bees, called Artificial Bee Colony (**ABC**), motivated by the intelligent behavior of bees.

1. Send the scouts bees to find food sources
2. *REPEAT*
   a. *Send the employed bees to identified food sources and determine their amounts of nectar.*
   b. *Calculate the probability value of the sources (quality) with which the onlookers bees will prefer sources.*
   c. *Send onlookers bees to food sources using a stochastic selection process based on the amount of nectar in each source.*
   d. *Stop the process of exploitation of sources exhausted by bees.*
   e. *Send scouts to the search area to discover new food sources randomly.*
   f. *Save the best food source found so far.*
3. *UNTIL (the requirements are met)*

Emergent Alignment by using our ABC Approach

The **problem** of the ontology **alignment** is to be able to decide which of the techniques of semantic alignment must be used.

For it, it is used the **ABC** algorithm in order to let it **choose automatically** the **technique** to perform the alignment.



Emergent Alignment by using our ABC Approach

The gain $G(S_i)$ is calculated as follows, in the equation

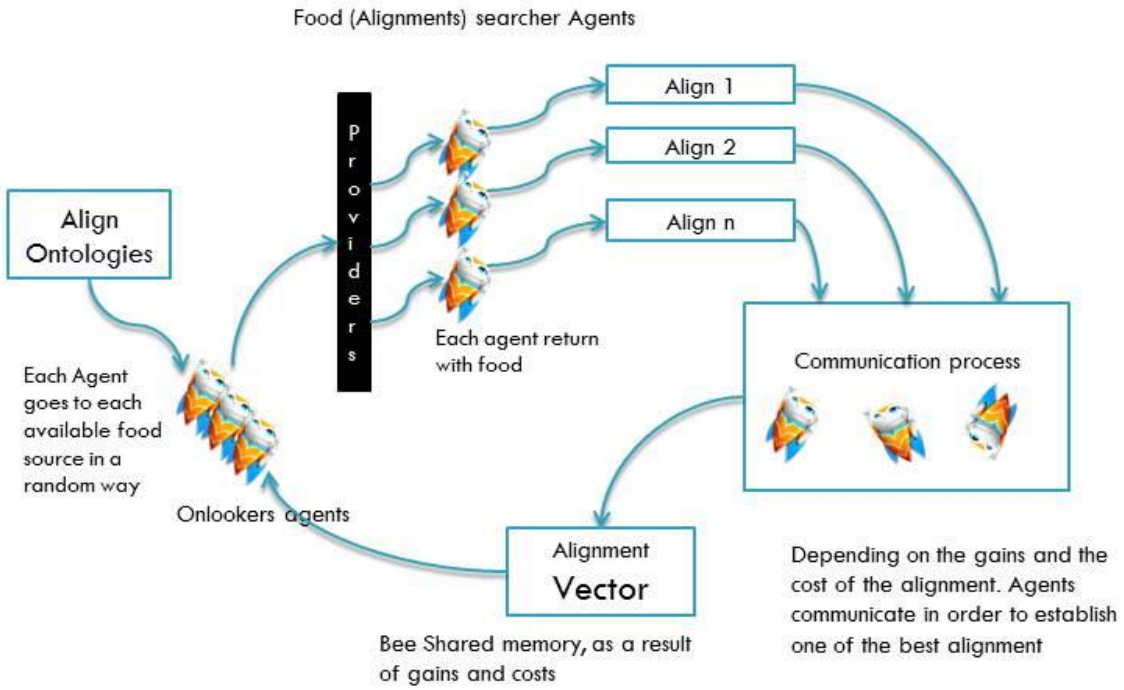
$$G(S_i) = \frac{S_a(S_i)}{CA(S_i)} \times P_c$$

- **S_i:** Service that can be utilizes to resolve a request (In our case, the alignment techniques). That is**, each alignment technique** is a source of nectar.
- **$G(S_i)$:**): Profit, that is obtained by the use of the service Sj (one alignment technique), defined by the equation, which determine the **quality of nectar** (alignment technique).
- **$Sa(S_i)$:** Satisfaction of the Bee, when the service Sj is performed. It is also relate with the quality of nectar; in our case is the **number of aligned nodes** of the ontologies
- **$CA(S_i)$:** Cost, it is represented in this work as **the execution time** of the service Sj to return a result (also affects the quality of nectar).
- **Pc:** Probability of preserving the food source. **Pseudo-random** value, with a **normal distribution** within the range of 0 and 1, which changes the value of $G(S_i)$

Emergent Alignment by using our ABC Approach

The algorithm is iterative, and it is done for finite iterations to make **several suggestions**,

The bees arrive at different services (sources of nectar); **they suggest various services**.



Used Alignment Techniques

a) **Class Structure:** based in finding similitude taking into consideration the **graph structure** from the classes.

b) **Distance edited name:** semantic distance required to make **the classes names** equals.

c) **Distance edited subclass name:** semantic distance required to make **the sub-classes names** equals, from the parent classes

d) **Name and properties:** **similitude** between classes names and properties.

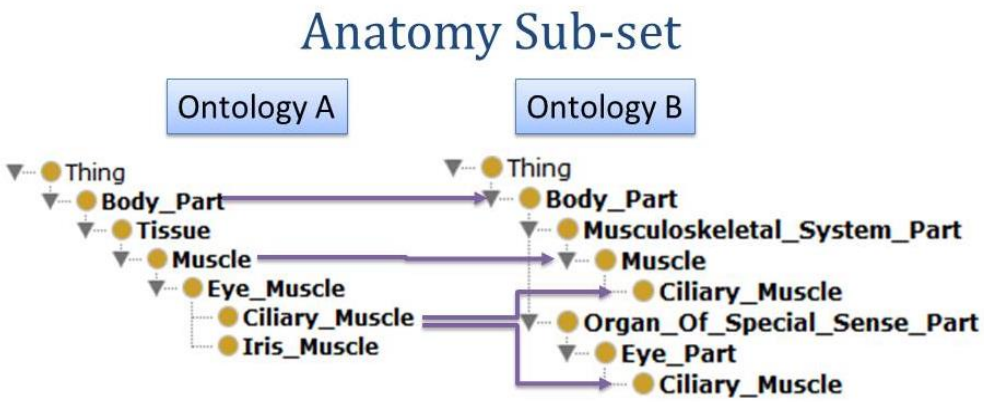
e) **Same names:** semantic similitude if **names from classes** are equals.

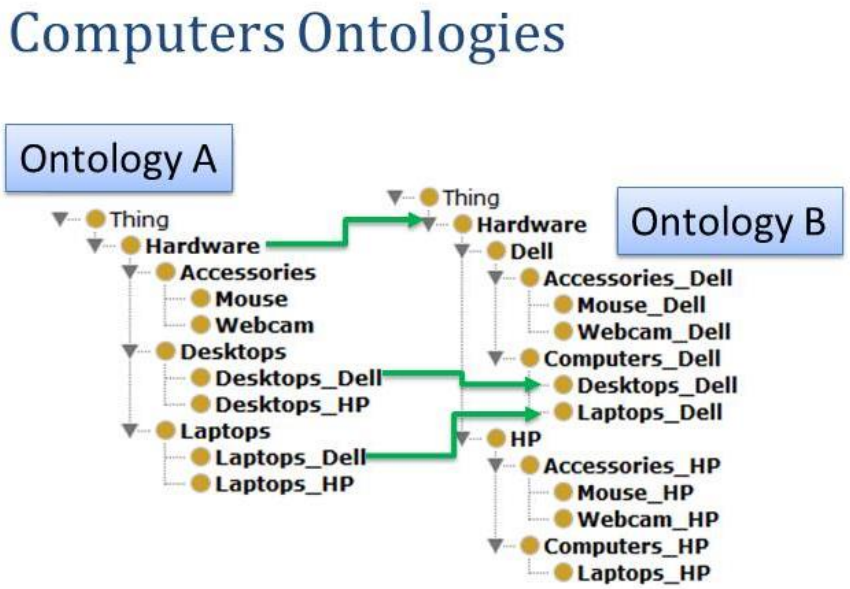
f) **Distance SMOA name** (A String Metric for Ontology Alignment): similitude between entities as **parts in common minus their differences**.

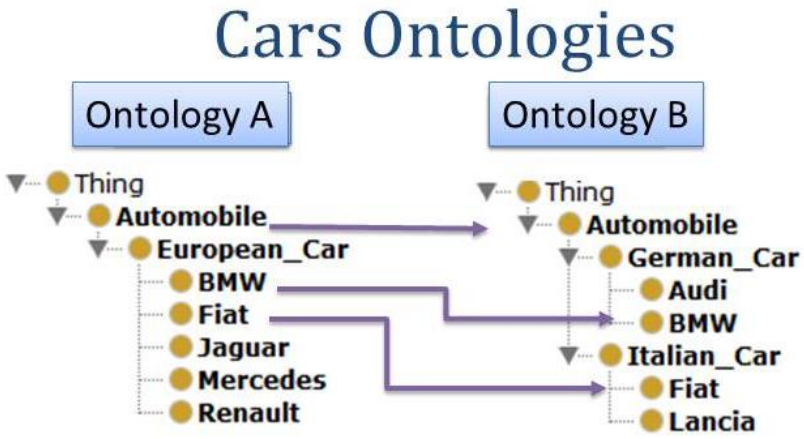
g) **String Distance:** takes classes, properties and instances **as simples strings** to compare them.

h) **Sub structures distance:** based in finding similitude taking into consideration the **graph structure** from the **sub-classes**.

Experiments







Results

Set	Execution Time (sec)	# Aligned nodes	#Times that each Align Technique is chosen
Cars	1:00	0	a) 0
	0:50	3	b) 6
	0:51	3	c) 3
	0:49	3	d) 2
	0:41	**3**	**e) 7**
	0:40	**3**	**f) 7**
	0:59	3	g) 3
	1:12	3	h) 2
Anatomy Sub set	0:50	0	a) 0
	0:59	3	b) 0
	0:53	3	c) 0
	0:47	**5**	**d) 12**
	0:45	3	e) 0
	0:35	5	f) 8
	0:59	3	g) 0
	0:53	**5**	**h) 10**
Computers	1:02	0	a) 0
	1:06	4	b) 0
	0:56	7	c) 0
	0:49	**9**	**d) 12**
	0:49	4	e) 0

In this cases, all techniques from b) to h) align **the same numbers of nodes**, as the **times are closed**, it is chosen any of this techniques

In this two cases the number of aligned nodes are different, it is chosen any of the techniques with the bigger **quantity of aligned nodes**, **choosing the one with less time**.

Linking Ontologies

It is the process to find relationships between entities that belong to different ontologies.

Weak link of Ontologies: it is a **correspondence between identical concepts.**

Strong Link of Ontologies: It is carried out in a **semi-automatic wa**y, with the help of a global knowledge expert that is linking, which can define new concepts, as well as **links that relate concepts of different ontologies**, thus creating a **Meta-Ontology**.

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Definitions:

A **meta-ontology** provides generic terms in the form of meta-concepts

A **meta-concept** represents a generic class, which has inheritable properties.

Similarity of Properties $$Sim\_P(C, C') = \frac{|P \cap P'|}{|P \cup P'|}$$

lexical similarity

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

		Concepts of the Source Ontology			
		C_1	C_2	...	C_N
Aligned Concepts	C'_1	PC_1,... PC_N	...	...	...
	C'_2	...			
	...				
	C'_N	...			...
New acquired concepts	CN_1	...			...
	CN_2	...			...
	...				
	CN_N	...			...

Collective Learning of Properties Matrix (MACP)

Common Property	Concepts
PC_1	C_1 , C_2 , $C_{3....}$, C_N
PC_2	C_1 , C_2 , $C_{3....}$, C_N
...	...
PC_N	C_1 , C_2 , $C_{3....}$, C_N

Table of Common Properties among Concepts (TCPC)

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Definitions

A Context X=(C, P, M, R, S) is a combination of a set of ontologies, where C is a set of concepts, P is a set of properties, M is a sub-set of the Cartesian product CxC, R a set of relationships of incidence between properties of concepts, and S a set of relationships between parent-child concepts.

A Category is a collection of concepts that have one or more properties in common. In a context X, a category Cat1 will be defined as (C, P), where C is a set of Concepts and P is a set of Properties

The incidence relation R can be represented as: P→M or P→CxC.

P

R: S ----------> O

The Scope of a category is the set of domain concepts

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Definitions

A Cat1 category is a **Sub-Category** of Cat2 (Cat1 $\subseteq$ Cat2) if C1 $\subseteq$ C2 and P1 $\subseteq$ P2

A **generic** sub-category **Sub-Cat-G** is one with a range (the object of R) greater than a threshold.

A **specific** sub-category **Sub-Cat-E** is one with scope (the subject of R) equal to a threshold

A **list of Ordered Sub-Categories** (LSO) for a context X is defined as:

$$LSO=\{Sub\text{-}Cat_1,\ Sub\text{-}Cat_2, \ldots, Sub\text{-}Cat_N \mid \forall\ Sub\text{-}Cat_i \subseteq Cat\text{-}O\ y\ Sub\text{-}Cat_{i+1} \subseteq Sub\text{-}Cat_i\ \}$$

sub-categories are sorted from the most general (greater scope), to the most specific (lower scope).

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Macro-Algorithm for the Generation of Meta-concepts

Inputs: A context X=(C,P,M,R,S)

Process:

1. The possible sub-categories are defined.
2. The LSO is created for the context X.
3. The sub-categories are classified as Sub-CAT-G and Sub-Cat-E.
4. The Sub-CAT-G are established as candidates for a meta-concept.
5. To structure the meta-ontology for the context, based on the Sub-Cat-G, the relationships of "Sub-Class" between these sub-categories are established.

A Cat1 is "Sub-Class" of a Cat2, if Cat1 is "Sub-Category" of Cat2.

⇩

The scope of the parent class is greater than the scope of the child class.

Output: The Meta-Ontology for the context X = (C, P, M, R, S))

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Quality Metrics of a Meta-Ontology

Robustness: A meta-ontology MO is Robust (R) with respect to an ontology O, if each meta-concept in MO represents at least one concept (or perhaps several) in the ontology O.

$$R(MO, O) = \frac{|MC\_R|}{|MC|}$$

Where: MC_R Set of Meta-Concepts that meeting the criterion of robustness, MC: Set of all Meta-Concepts

Completeness: A meta-ontology MO is Complete (C) with respect to an ontology O, if each concept in O is represented by at least one meta-concept in MO.

$$C(MO, O) = \frac{|C\_MC\_C|}{|C|}$$

Where: C_MC_C : Set of concepts of O that are defined for some meta-concept, C: Set of all concepts of O

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

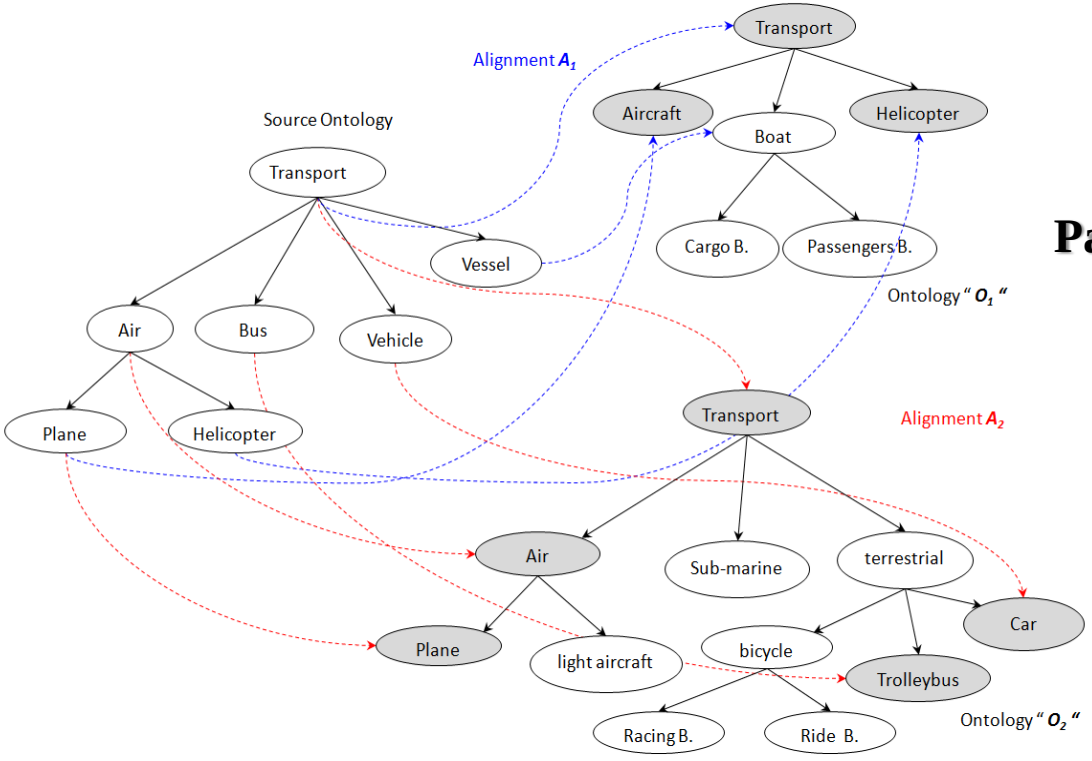
Quality Metrics of a Meta-Ontology

Precision: A meta-ontology MO is Precise (P) with respect to an ontology O, if each concept is associated maximum to a meta-concept (or in any case none) in MO.

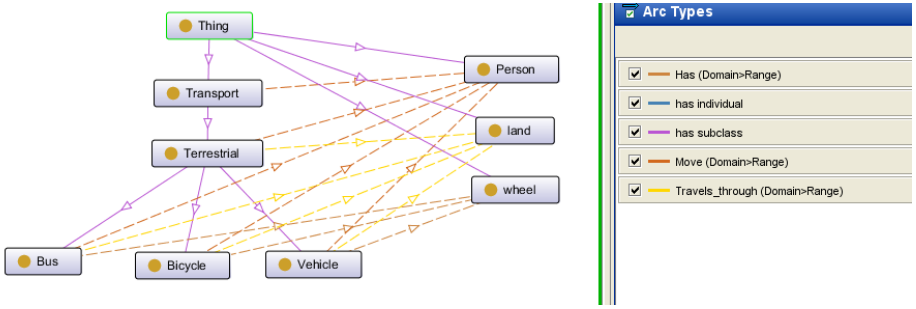
$$P(MO, O) = \frac{|C\_MC\_P|}{|C|}$$

Where: C_MC_P: Set of O concepts that are defined for only one meta-concept

Generation of Meta-Ontologies from Multiple Alignments between Ontologies



Participating ontologies in the multiple combination process



Part of the resulting Ontology after of the combination process

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

	Concepts of the Source Ontology			
Aligned Concepts	Concept	Transport	Bus	Vehicle
	Transport(O1)	Move_Person		
	Transport(O2)	Move_Person		
	Trolleybus (O2)		Move_Person, Travels_through_Land, Has_Wheel	
	Car (O2)			Move_Person, Travels_through_Land , Has_Wheel
New acquired concepts	Terrestrial	Move_Person	Move_Person, Travels_through_Land	Move_Person, Travels_through_Land
	Bicycle		Move_Person, Travels_through_Land, Has_Wheel	Move_Person Travels_through_Land , Has_Wheel

MACP

Common Property	Concepts
Move_Person	Transport, Bus, Vehicle, Terrestrial, Bicycle
Travels_through_Land	Bus, Vehicle, Terrestrial, Bicycle
Has_Wheel	Bus, Vehicle, Bicycle

TCPC

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Set of ontologies: O1, O2 and O_source.

Set of concepts C:
C={ Transport (Tra),Terrestrial (Ter), Bicycle (Bic), Bus (Bus), Vehicle (Veh), Land (Lan), Wheel (Whe),Person(Per)}

Set of properties P:
P={Move_Person(Mov_Per),Travels_through_Land(Tra_Lan), Has_Wheel (Has_Whe) }

Set M of cartesian product cxc:
M={<Tra,Per>,<Ter, Per>, <Bic, Per>, <Bus, Per>,<Veh, Per>, <Ter, Lan>, <Bic, Lan>, <Bus, Lan>,<Veh, Lan>, <Bic, Whe>, <Bus, Whe>,<Veh, Whe>}

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Set R of incidences between properties and objects:

R={ Mov_Per → <Tra, Per>, Mov_Per → <Ter, Per >, Mov_Per → <Bic, Per >, Mov_Per → <Bus, Per >, Mov_Per → <Veh, Per >, Tra_Lan → <Ter, Lan>, Tra_Lan → <Bic, Lan>, Tra_Lan → <Bus, Lan>, Tra_Lan → <Veh, Lan>, Has_Whe → <Bic, Whe>, Has_Whe → <Bus, Whe>, Has_Whe → <Veh, Whe>}

Concepts that belong to the domain and range:

Domain={Tra, Ter, Bic, Bus, Veh}

Range={Lan, Whe, Per}

Set S of parent-child relationships:

S={Tra←Ter, Ter ← Bic, Ter ←Bus, Ter ← Veh }

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Macro-algorithm:

1. Cat-O=({ Tra, Ter, Bic, Bus, Veh, Lan, Whe, Per},
{Mov_Per, Tra_Lan, Has_Whe})

2. Some sub-categories are:
Sub-Cat1=({Tra,Ter,Bic,Bus,Veh,Per,Lan,Whe},
{Mov_Per,Tra_Lan,Has_Whe}) Scope= 5
Sub-Cat2=({Ter,Bic,Bus,Veh,Lan,Whe}, {Tra_Tie,Has_Whe})
Scope= 4
Sub-Cat3=({Bic, Bus, Veh, Whe}, {Has_Whe})
Scope= 3

3. LSO={Sub-Cat1, Sub-Cat2, Sub-Cat3}

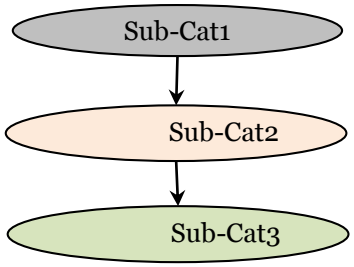
4. **Sub-CAT-G:** Sub-Cat1, Sub-Cat2, Sub-Cat3. In this case, **Sub-CAT-E**=empty.

Generation of Meta-Ontologies from Multiple Alignments between Ontologies

Macro-algorithm:

5. **Sub-CAT-G's candidates** for meta-concepts: Sub-Cat1, Sub-Cat2 and Sub-Cat3.

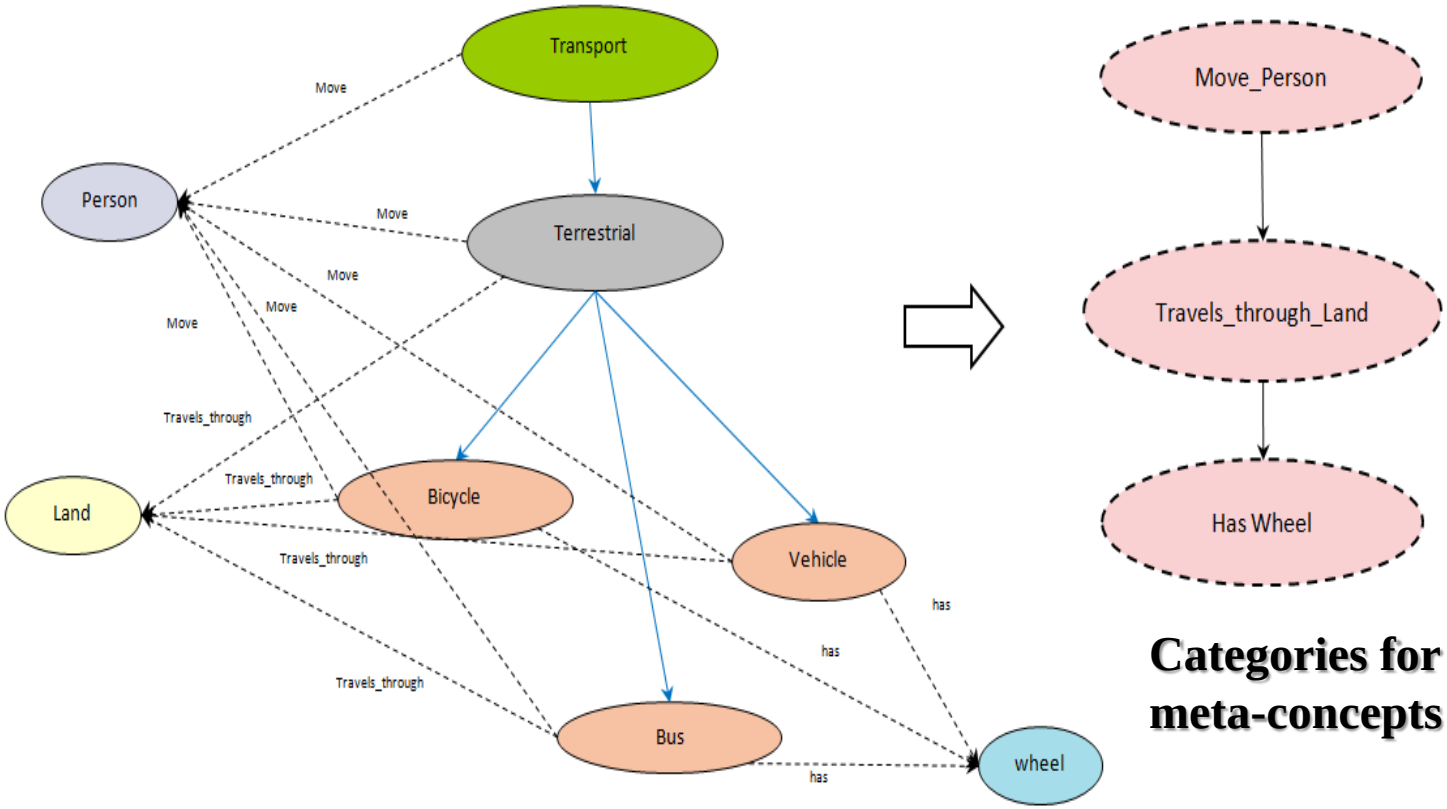
6. **Relationships of "Sub-Class"** between sub-categories: Sub-Cat3 ← Sub-Cat2 ← Sub-Cat1



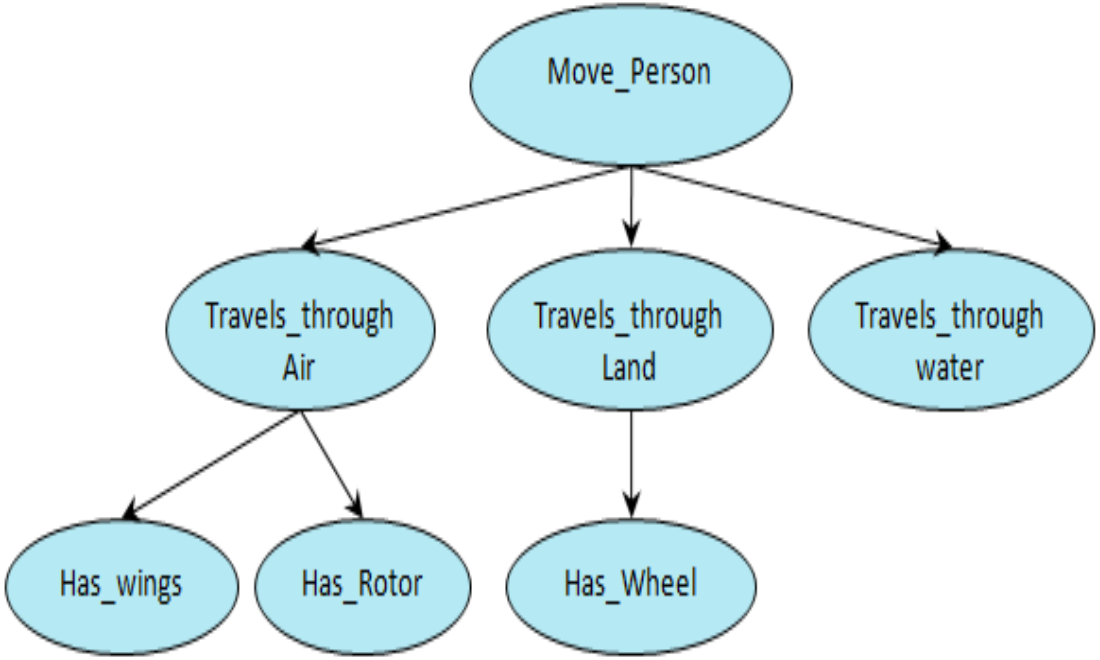
Candidate Sub-Categories for Meta-Concepts

The name to identify the meta-concepts of a meta-ontology, will be given by the properties and the range that define each subcategory

Generation of Meta-Ontologies from Multiple Alignments between Ontologies



Generation of Meta-Ontologies from Multiple Alignments between Ontologies



Resulting meta-ontology for the transport domain

	O_Source	O1	O2	General
Robustness	1	0.7	0.85	0.83
Completeness	1	1	1	1
Precision	1	1	1	1

Getting a value of 1 in robustness is very difficult, and would correspond to a perfect alignment between all the ontologies involved in the integration process

Merge of ontologies

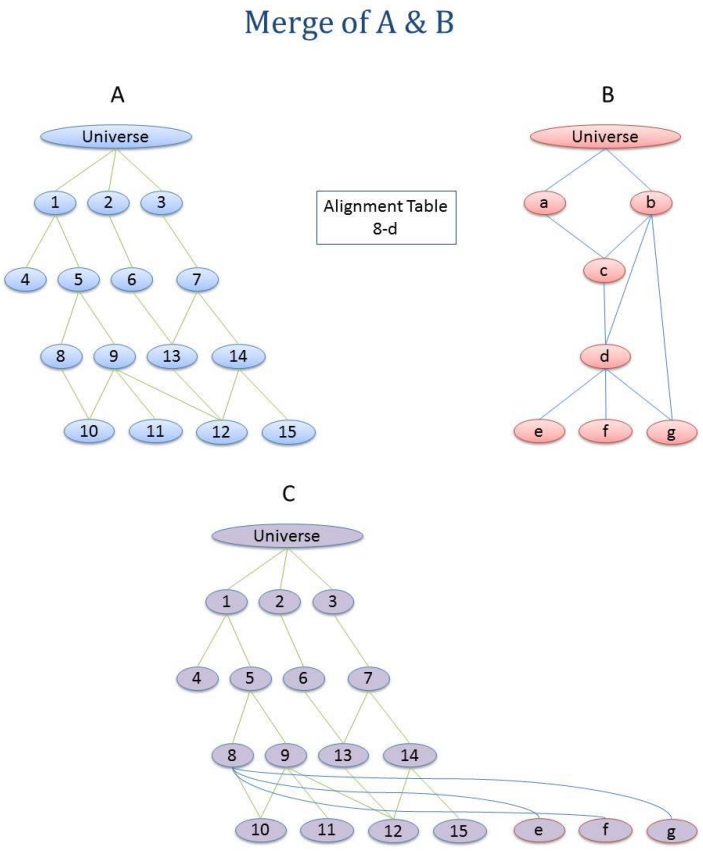
It is the process where several ontologies within the same domain come together to standardize knowledge, make knowledge grow or have total knowledge locally.

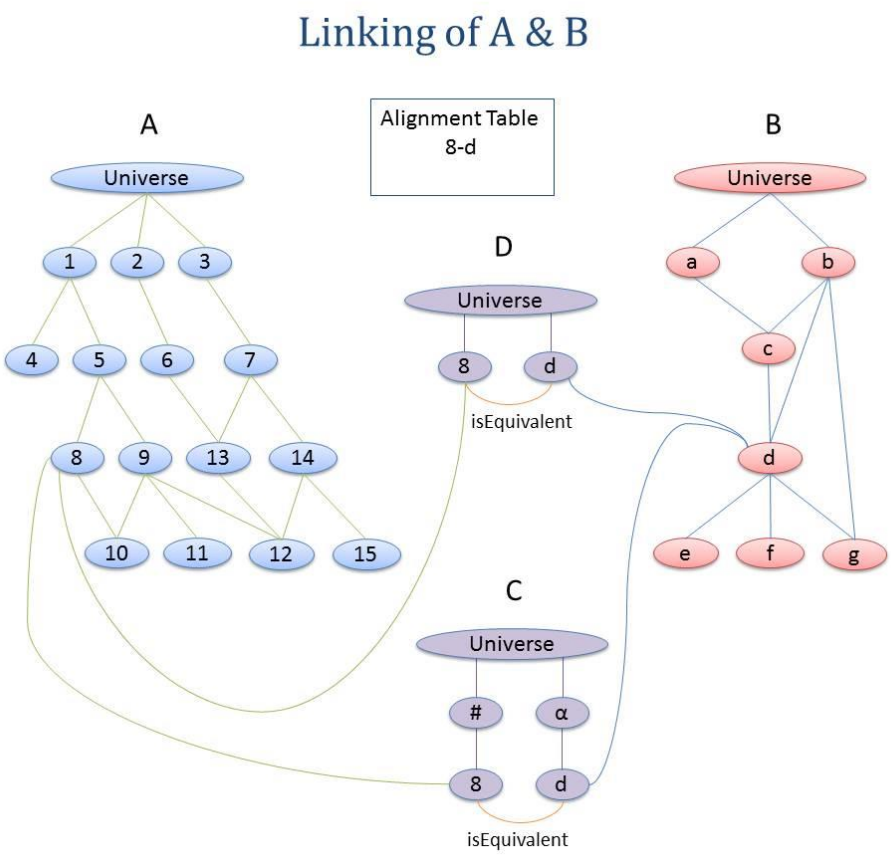
The ontologies handle the same knowledge, but with different representations, or have partial representations of that knowledge.

Weak mixture of Ontologies: take an ontology A, copy it as a result C, and enrich it with the other B, comparing all the concepts of ontology C (which are the same as A at this time) with those of ontology B, enriching the concepts of C with their similar concepts of B, **leaving out part of the knowledge of B**.

Strong mixture of Ontologies: It is a weak mixture, but incorporating the knowledge left out of B,

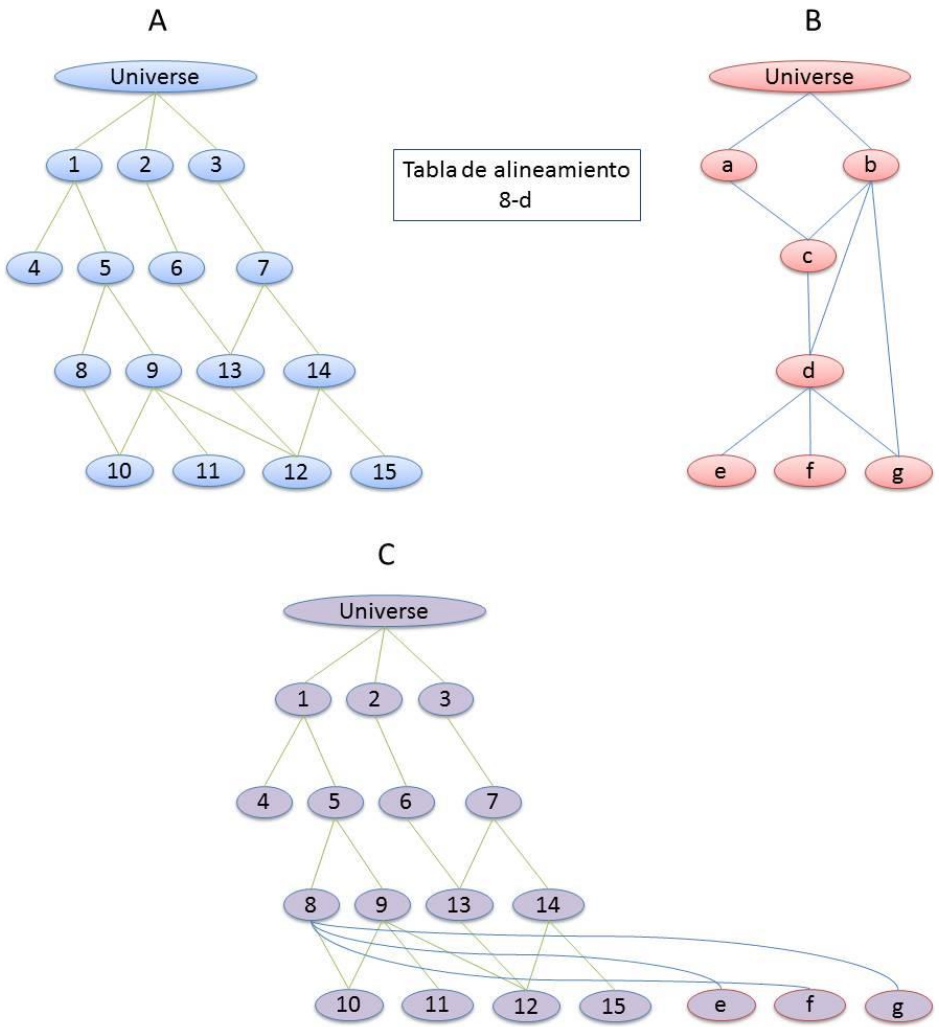
Ontology Mining (OM)





Ontology Merge Algorithms

The problem of the traditional mixture of ontologies (which we have called weak mixture), is that it leaves knowledge without being incorporated in the resulting ontology.

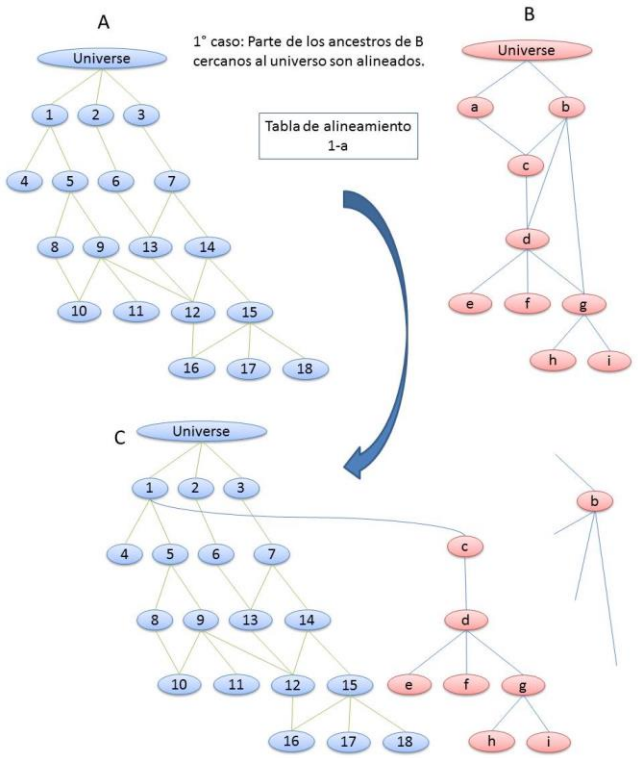


Ontology Merge Algorithms

Our Strong Blend proposal is made in two parts,

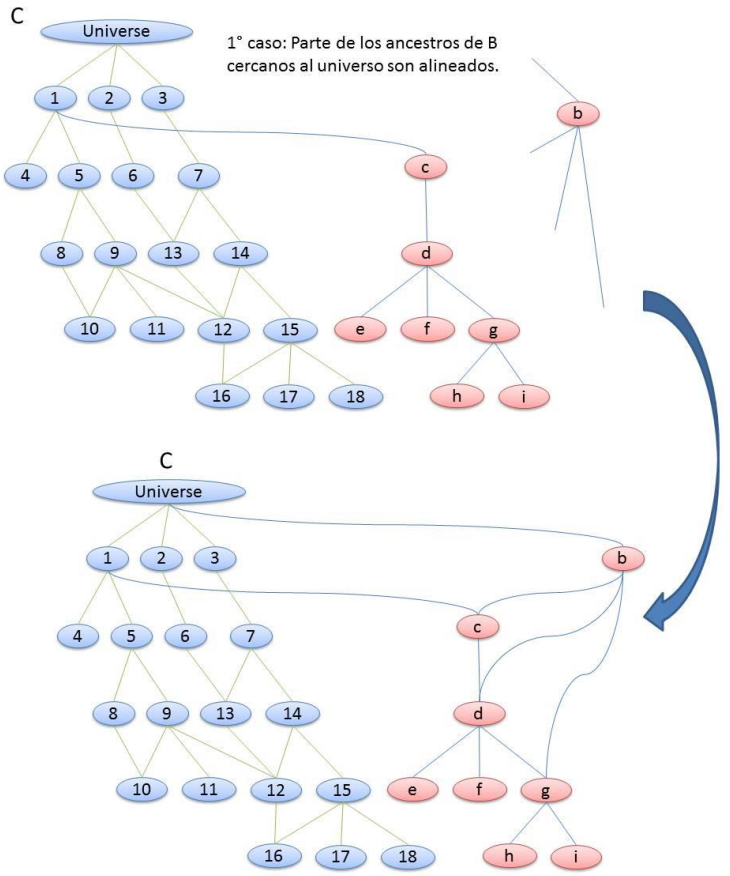
- The weak mixture is made,
- The concepts and relationships left outside are incorporated.

First Part: Our system performs the weak mixture of two consistent ontologies A, B in an ontology C



Ontology Merge Algorithms

Second Part: if the ontology to which the knowledge is being extracted to be added to the first has still knowledge without being added, the following cases are analyzed:



Case 1:

The concepts of B were partially aligned and unaligned nodes are not copied into the result ontology. Only would suffice with:

- add nodes not aligned to C
- copy the relationships that were not copied or aligned.

Ontology Merge Algorithms

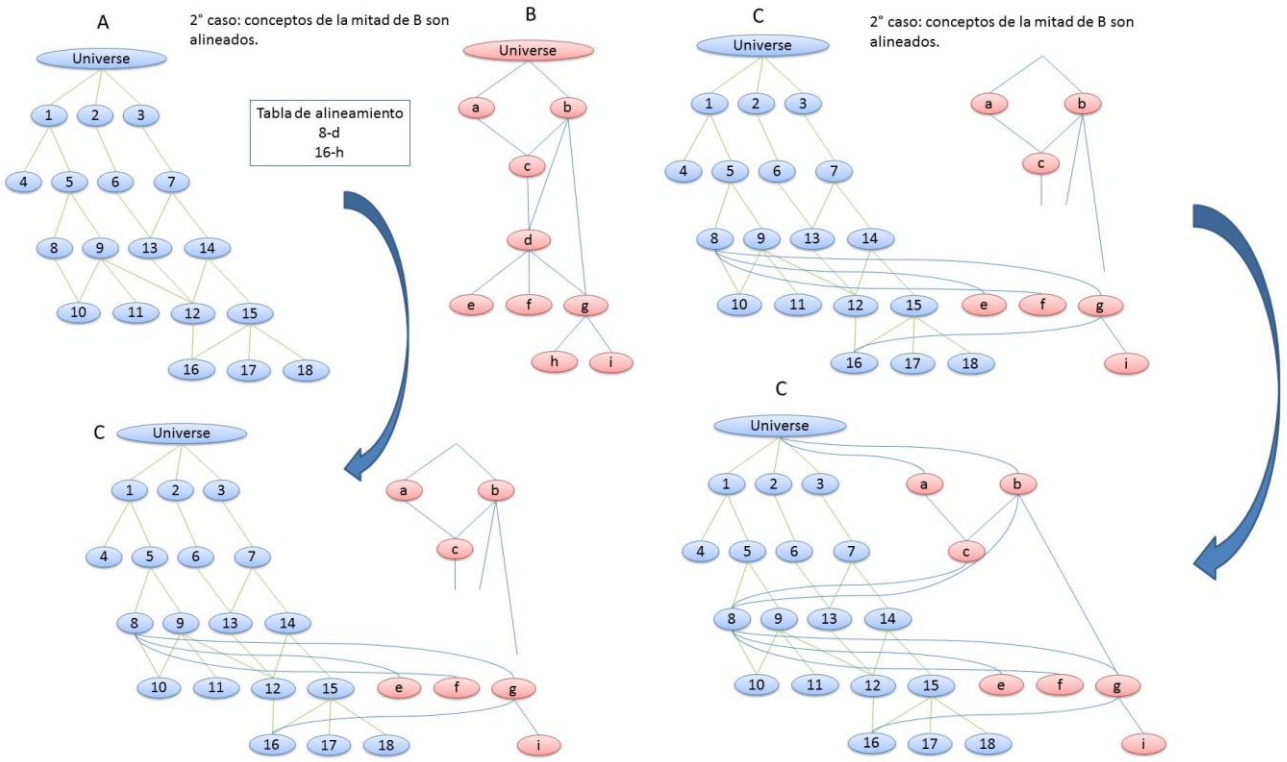
Case 2

- Certain intermediate nodes were aligned
- It leaves the ancestors without being copied

The strong mixture must:

- add these concepts to the universe of C
- Add the relationships where they participate

(see **b-g ,c-8 and b-8**)



Ontology Merge Algorithms

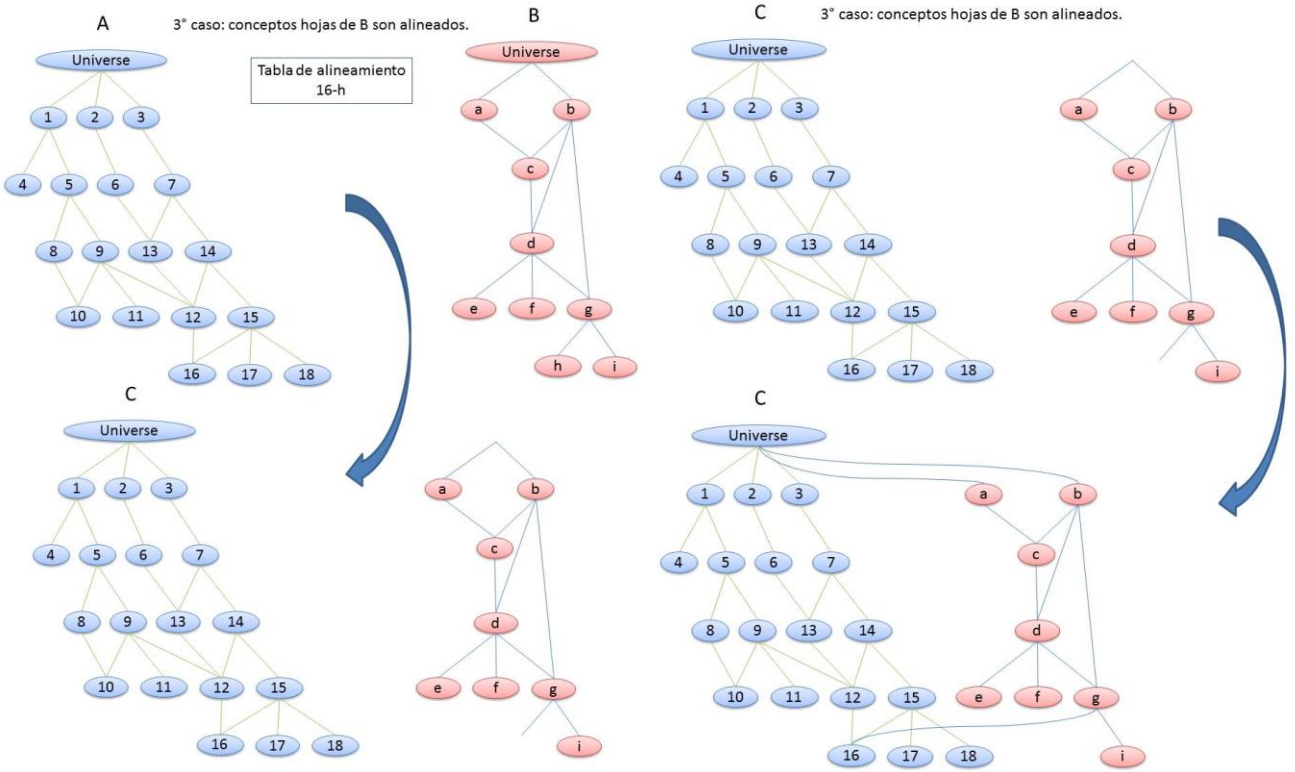
Case 3:

- Only the leaf nodes are aligned
- A large knowledge set of B. is left out of C.

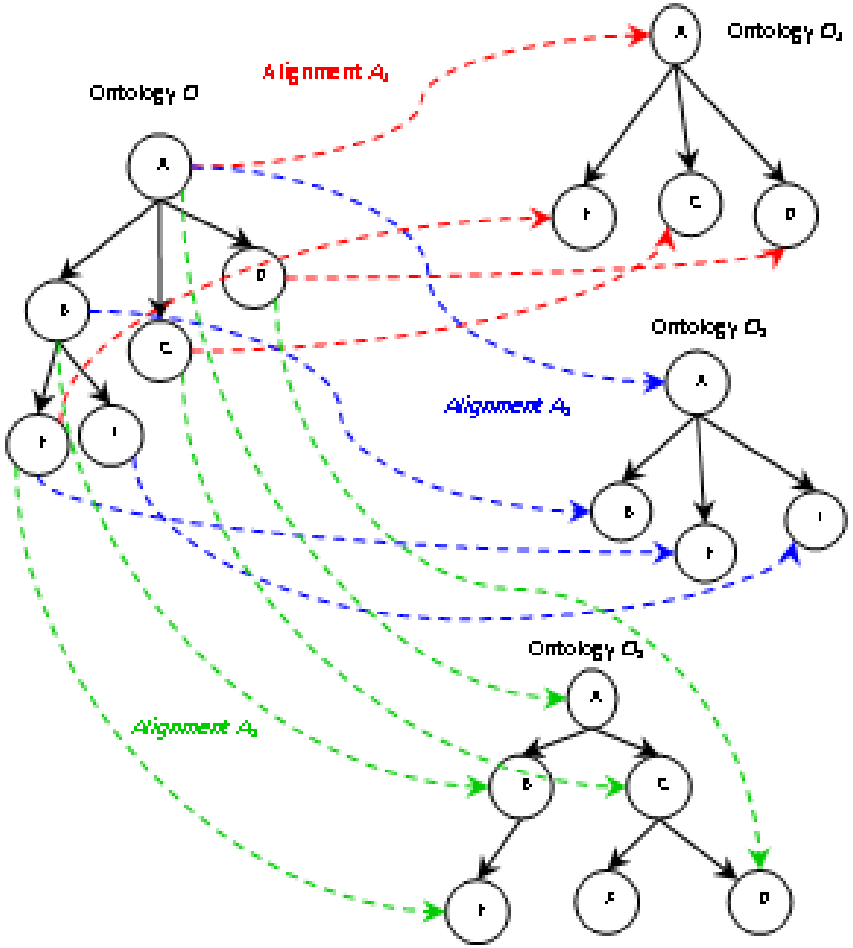
The strong mixture must

- add all concepts not copied to C
- look for the relationships that these concepts already had with others in B
- copy also those relationships with the concepts that were copied or with which they were aligned

(g-16)



Combination of multiple ontologies



Combination of multiple ontologies

DETERMINATION OF THE SPACE OF SOLUTIONS

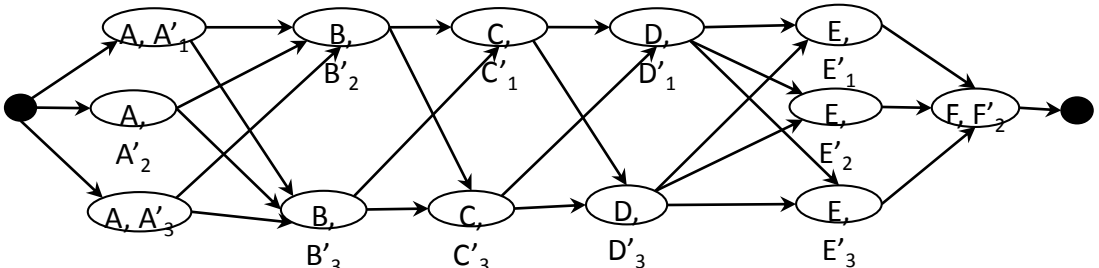
The alignments ***A*$_1$, *A*$_2$ *y A*$_3$** already defined can be described as follows:

$$\mathbf{A_1 = \{(A, A'_1), (C, C'_1), (D, D'_1), (E, E'_1)\}}$$
$$\mathbf{A_2 = \{(A, A'_2), (B, B'_2), (E, E'_2), (F, F'_2)\}}$$
$$\mathbf{A_3 = \{(A, A'_3), (B, B'_3), (C, C'_3), (D, D'_3), (E, E'_3)\}}$$

Among each pair of concepts must be defined, by the existing alignments, a similar measure:

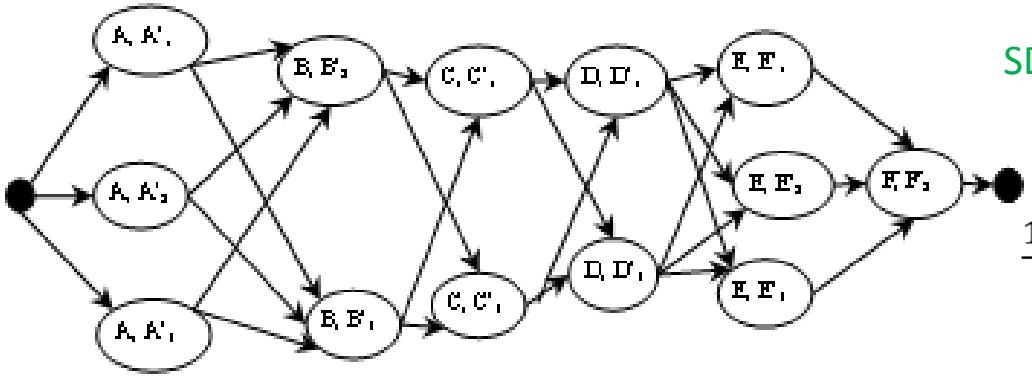
MC(C,C')

This similarity measure must be ranging from 0 to 1, which refers to the **grade of similarity with the aligned concepts (p.e lexical similarity)**.



For **M** concepts in the source ontology and **N** alignments, we have a maximum of **N^M** possible solutions for this problem.

Combination of multiple ontologies



SD: Similarity of Siblings of C and C

$$SS(C,C') = PC_S \times Sim(C,C') + \frac{1-PC_H}{n}\sum_{i=1}^{n}\max(Sim(S_i,S_1'),\ldots,Sim$$

SA : Similarity of Ancestors of C and C'

$$SA(C,C') = PC_A \times Sim(C,C') + \frac{2(1-PC_A)}{n(n+1)}$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(n+1-i)Sim(Anc_i(C),Anc_j(C'))$$

SS: Similarity of Descendants of C and C'

$$SD(C,C') = PC_D \times Sim(C,C') + \frac{1-PC_D}{n}\sum_{i=1}^{n}\max(Sim(H_iH_1'),\ldots,Sim(H_i,H_n'$$

)

Similarity measure

$$MS(C,C') = \frac{SA(C,C') + SD(C,C') + SS(C,C')}{3}$$

Combination of multiple ontologies

Probability of Transition

In order to build the solution, each ant must choose the next element of the solution from the *"r"* location. To do this, it uses a function of probability to select the element *"s":*

$$P_{(r,s)}^{k} = \frac{\gamma_{(r,s)}{}^{\alpha} \cdot \eta_{(r,s)}{}^{\beta}}{\sum_{u \in J_r^k} \gamma_{(r,u)}{}^{\alpha} \cdot \eta_{(r,u)}{}^{\beta}} \quad \text{Si } s \in J_r^k$$

Where:

γ_{rs} : is the amount of pheromone.

η_{rs} : is the heuristic information (Similarity Measure (MS)).

J_r^k : is the node not visited yet by the k ant from r.

α y β: define the importance of the memory information (pheromone) and heuristic information.

Combination of multiple ontologies

Pheromone Updating

While an ant is in the construction process of the solution, each selected edge must update the pheromone, delivering an amount of pheromone
:

$$\gamma_{(r,s)} = \gamma_{(r,s)} + \Delta\gamma_{(r,s)}$$

Where:

$\Delta\gamma_{(r,s)}$: It is the increment of the pheromone, corresponding to sum of amounts of pheromone leaving by ants in the edge (r, s);

$$\Delta\gamma_{(r,s)} = \sum_{k=1}^{M} \Delta\gamma_{(r,s)}^{K}$$

Where:

$\Delta\gamma_{(r,s)}^{K}$: It is the amount of pheromone leaving by k ant in the edge (r,s) which is directly related to the “Quality of the Solution” found by k ant.

$$\Delta\gamma_{(r,s)}^{K} = f(GE(C, C_s))$$

Where:

C_s : selected alignment by k ant as the best solution for C

$GE(C, C_s')$: Grade of enrichment of the alignment.

Combination of multiple ontologies

"Degree of enrichment" (GE)

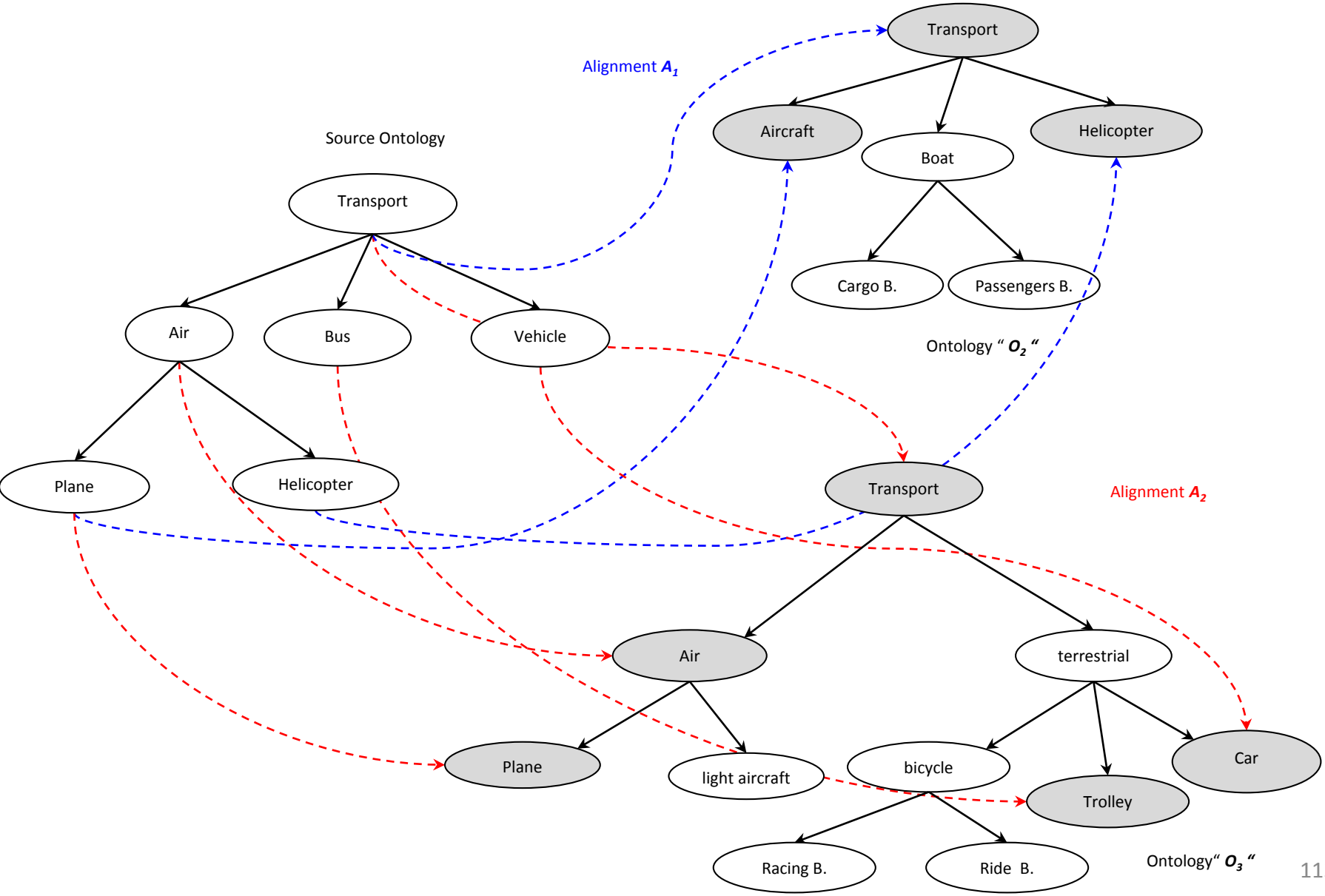
amount of new concepts obtained by the source ontology after selecting an alignment for a concept.

GE of the ontology after selecting the alignment of a concept C with C 'of the new concepts that can be added to the correspondents of the ontology:

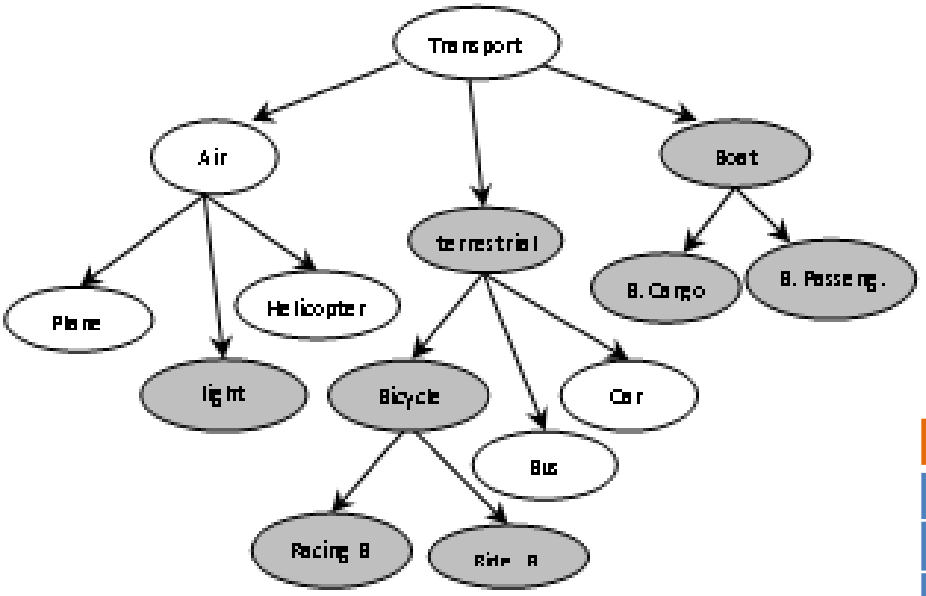
- Children of concepts C 'not aligned (New hyponyms) and their descendants
- The siblings of the concepts C 'not aligned with the immediate aligned ancestor (father) (New Cohyponyms) and their descendants.
- Concepts ancestors of C 'not aligned (New hyperonyms).

$$GE(C, C') = CHildren\_Non\_Aligned\,(C') + Siblings\_Non\_Aligned(C') + Ancestors\_Non\_Aligned(C')$$

Combination of multiple ontologies



Combination of multiple ontologies



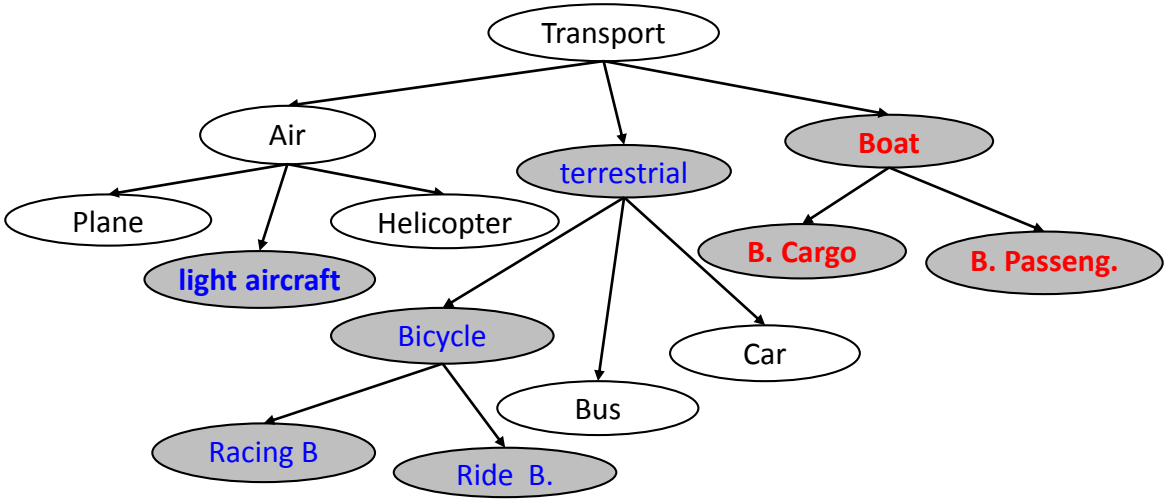
Quality metric

- Coverage (or completeness)
- Compactness
- Redundancy

Ontology O	Ontology O1	MS(C,C´)
Transport	Transport	1
Air	-	-
Bus	-	-
Vehicle	-	-
Plane	Aircraft	0.8
Helicopter	Helicopter	1
Ontology O	**Ontology O2**	**MS(C,C´)**
Transport	Transport	1
Air	Air	1
Bus	Trolley	0.5
Vehicle	Car	0.8
Plane	Plane	1
Helicopter	-	-

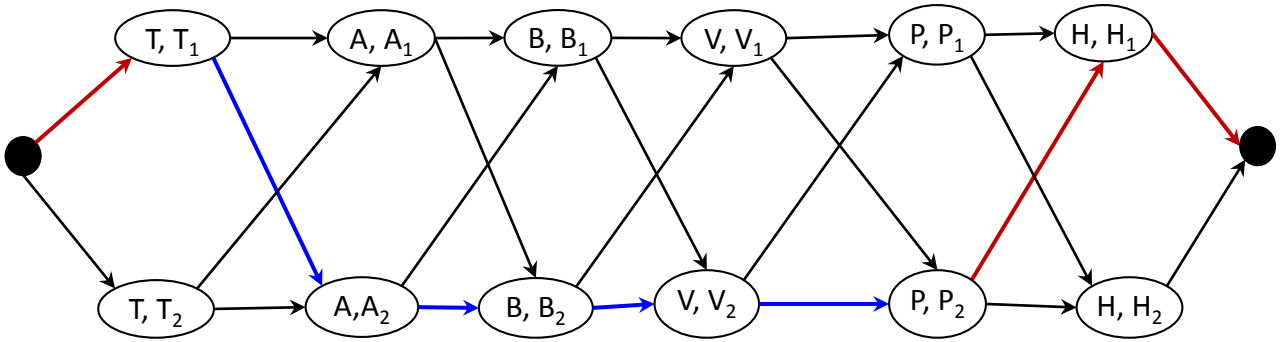
Combination of multiple ontologies

A possible resultant ontology after combination process, with a GE of 8, where new concepts acquired by ontology are highlighted.



This is a route graph for the solution obtained

Concepts
Transport (T)
Air(A)
Bus (B)
Vehicle (V)
Plane (P)
Helicopter (H)



Graph Mining

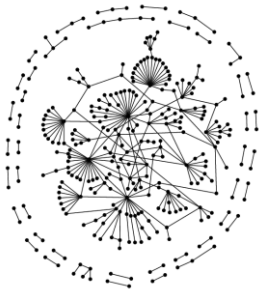
Graph Mining

Objective: Develop algorithms to analyze graphs in order to extract knowledge.

- Search for patterns in them
- Search for groups of similar graphs (clustering)
- Construction of prediction models for graphs (classification)

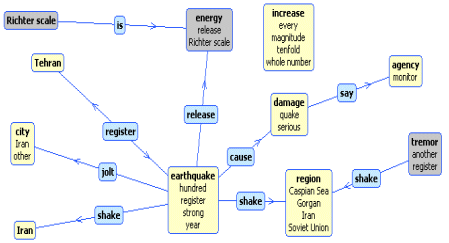
- Applications
  - discovery structural motive
  - protein recognition
  - reverse engineering in VLSI
  - Much more …

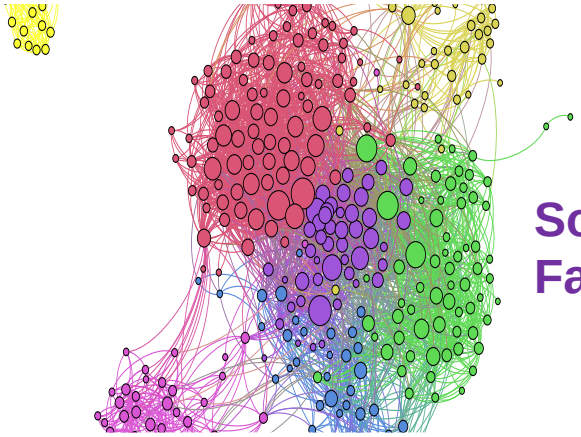
Graph Mining



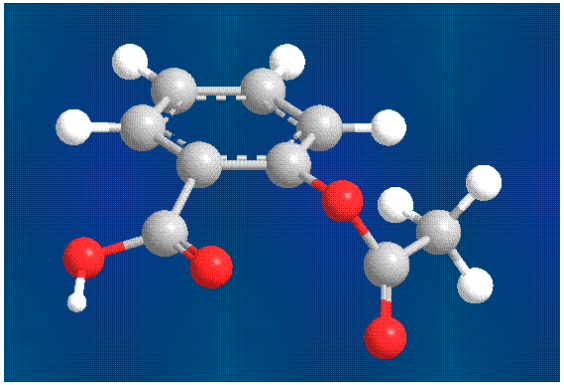
Interactions between yeast proteins

Semantic network



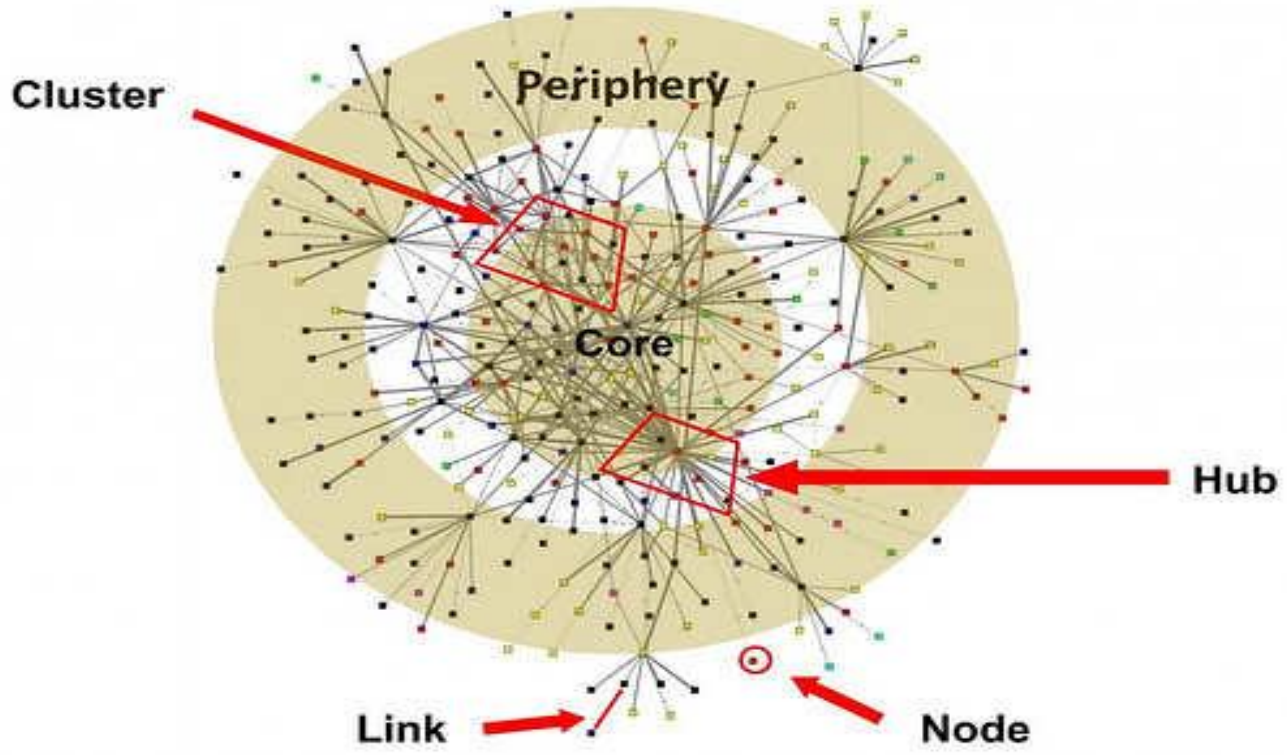


Someone's Facebook



Aspirin

Graph Mining



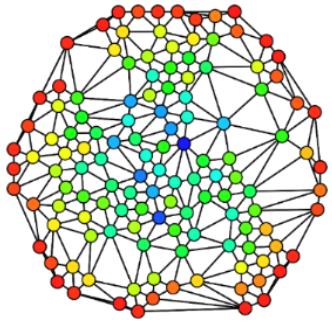
Similarity measures based on graph patterns

- Similarity measures based on characteristics
  - Each graph is represented as a vector of characteristics
  - Distance vector
- Similarity measure based on Structure
  - Maximum common subgraph
  - Grafo edita distancia: insertion, deletion, and re-labeling

Network metrics

Each network metric answers the following questions:

- **Question: Who is more central?**
  - **NETWORK METRICS: centrality**
    - a) degree centrality.
      - 1) Indegree
      - 2) Outdegree
    - b) Closeness centrality.
    - c) Betweenness centrality.



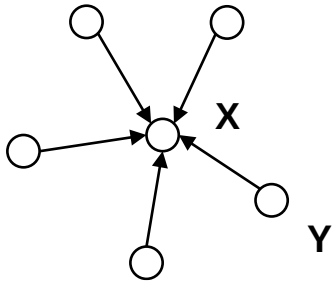
- **Question: Everything is connected?**
  - **NET METRICS: connected components**
    - Strongly connected components:
    - Weakly connected components
  - **NET METRICS: giant component**

- **Question: How far are things?**
  - **NET METRICS: shorter routes**

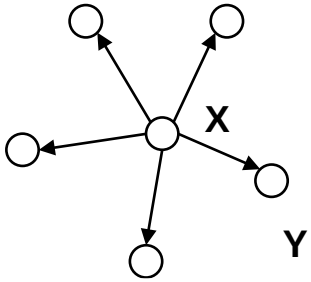
- **Question: How dense are they?**
  - **NET METRICS: density of the graph**

Centrality

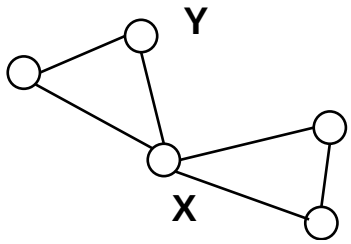
Possible measurements of a vertex in a graph, which determines its relative importance within it



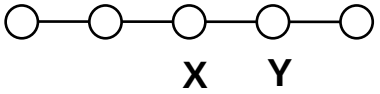
indegree



outdegree

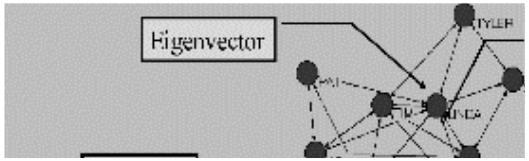


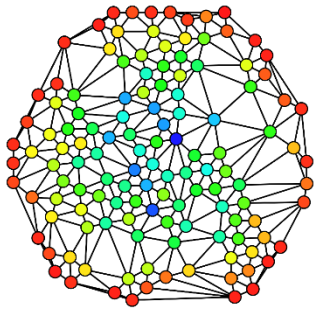
Betweenness



Closeness

eigenvector centrality





The color (from red = 0 to blue = maximum) of each node indicates intermediation centrality.

Community Metrics

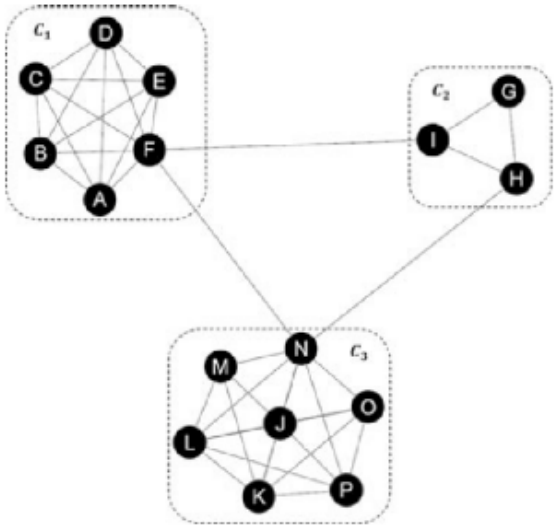
Designed to divide the network into modules, clusters, communities.

A network with high modularity means that:

- It is very dense between nodes of the same community
- But with dispersed connections between nodes of different communities

- ❑ Mutuality: Each member knows all the members
- ❑ Frequency: Each member knows at least k members of the group
- ❑ Closeness: The members are separated by a maximum of n jumps

Communities Metrics



- 3 communities
- Each communities has its graph
- The density among the communities is low

Global metrics:

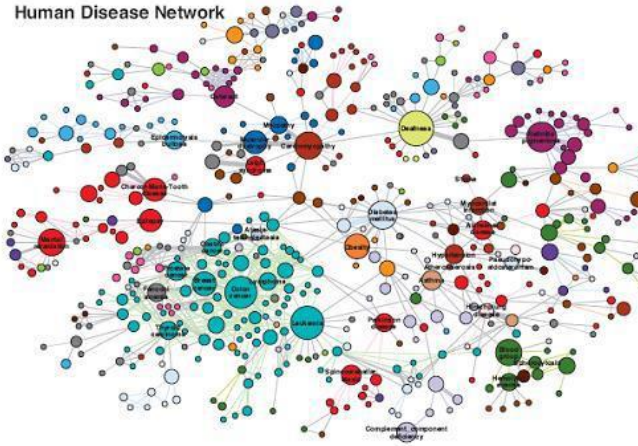
- **average distance**
- **average grade**
- **diameter and radius**
- **...**

Clusters of Signaling Pathways Networks

A **signaling pathway** is the set of reactions involved in the reaction of a cell to an external stimulus.

The clusters do not give much information per se, but by identifying the biological functions that identify each cluster, families can be defined.

The activation of the receptor caused by binding to a ligand is directly associated with **the response of the cell.**

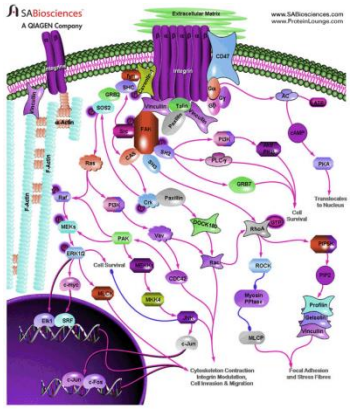


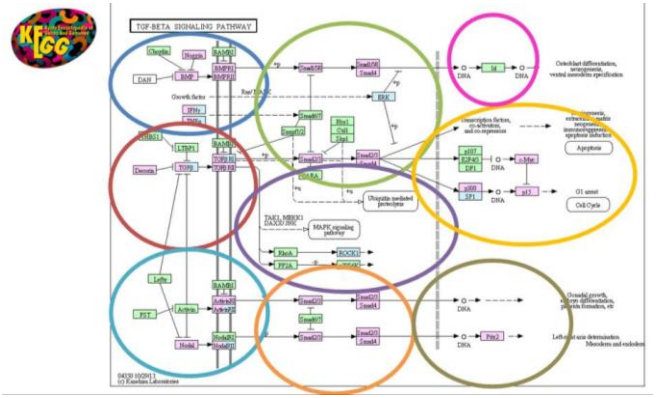
SeMiC Macro-Algorithm

Macro algorithm that allows to detect the clusters within a network signaling pathway, and enrich them with GO.

1. Receive a **signaling pathway network** as input
2. Take it to a **network format** (proteins will be treated as nodes and reactions as relationships)
3. Calculate the **modularity** for each node in the network
4. Perform the **hierarchical cluster**,
5. Calculate the **centroids** of each cluster, using network centrality techniques.
6. **Enrich each centroid** semantically with GO

SeMiC Macro-Algorithm

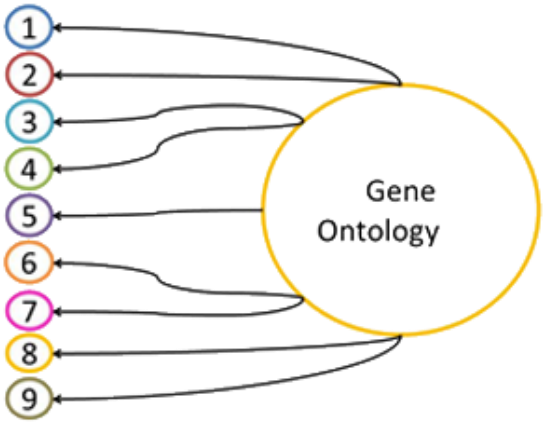




Example the encyclopedia of Genes and Genomes TGF-β

The clusters are calculated (step 4). The centroids are then removed (step 5)

take the network, which can be received, for example, in OWL format, to a traditional network format to be analyzed (step 2): (NET, DOT and CSV). Then the modularity of the nodes is calculated (step 3).

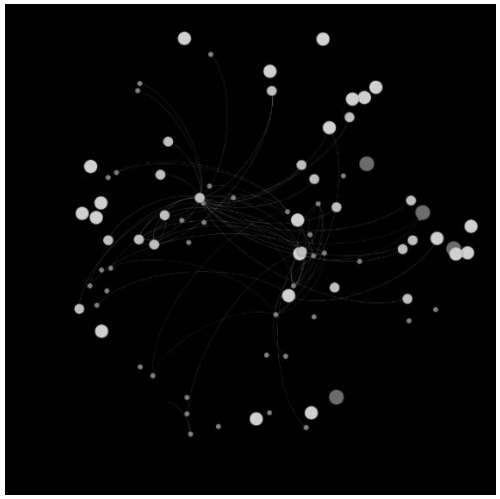


The centroids pass to a semantic enrichment (step 6) using Gene Ontology (GO)

tool used: Gephi

SeMiC Macro-Algorithm

Alzheimer Network view, with the main central nodes goo.gl/JFyu26



Label	Degree	Closeness Centrality	Id of the cluster
_:A2034	4	6,215	0
_:A2095	4	6,145	8
_:A1967	4	5,761	3
_:A1943	4	5,717	3
_:A1943	4	5,269	3

Alzheimer nodes, with semantic content extracted from GO

Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class
Methyl-CpG-binding domain protein 5 MBD5 ortholog	METHYL-CPG-BINDING DOMAIN PROTEIN 5 (PTHR16112:SF18)	-
Tyrosine-protein kinase JAK2 JAK2 ortholog	TYROSINE-PROTEIN KINASE JAK2 (PTHR24418:SF179)	non-receptor tyrosine protein kinase non-receptor tyrosine protein kinase
Hepatocyte nuclear factor 4-alpha HNF4A ortholog	HEPATOCYTE NUCLEAR FACTOR 4-ALPHA (PTHR24083:SF41)	nuclear hormone receptor receptor nucleic acid binding
Leptin receptor gene-related protein LEPROT ortholog	LEPTIN RECEPTOR GENE-RELATED PROTEIN (PTHR12050:SF3)	cytokine receptor
Appetite-regulating hormone GHRL ortholog	APPETITE-REGULATING HORMONE (PTHR14122:SF1)	-

Linked Data

Current Web

Microformats

XFN (**X**HTML **F**riends **N**etwork)

XFN quick reference

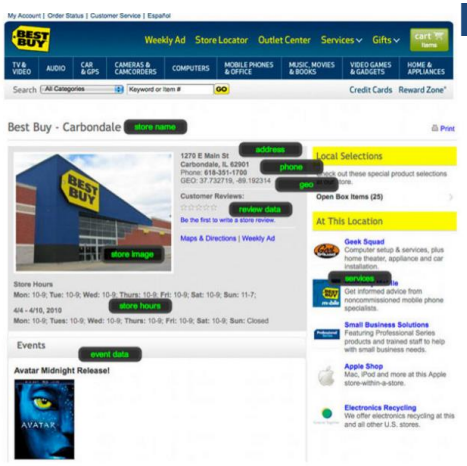
relationship category	*XFN values*
friendship (at most one):	friend acquaintance contact
physical:	met
professional:	co-worker colleague
geographical (at most one):	co-resident neighbor
family (at most one):	child parent sibling spouse kin
romantic:	muse crush date sweetheart
identity:	me

hCard

hcalendar

Social Data Analytics
Social Network Analytics
Linked Data

RDFa



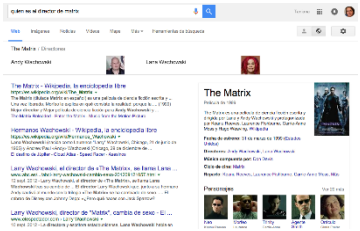
Best Buy employees entered information into the blogs every day, using online forms that output RDFa. Myers told us that the use of RDFa makes "human input from our store employees more visible on the Web."

Best Buy is using Good Relations, a Semantic Web vocabulary for e-commerce that describes product, price, and company data.

FOAF

```
<?xml version="1.0" standalone="yes"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">

  <foaf:Person>
    <foaf:name>
                Taniana Josefina Rodrígiuez de Paredes
    </foaf:name>
    <foaf:mbox rdf:resourse=mailto:taniana@ula.ve/>
    <foaf:knows>
                <foaf:Person>
                    <foaf:name> Jose Aguilar </foaf:name>
                    <foaf:mbox rdf:resourse=mailto:aguilar@ula.ve/>
                </foaf:Person>
    </foaf:Knows>
  </foaf:Person>
</rdf:RDF>
```

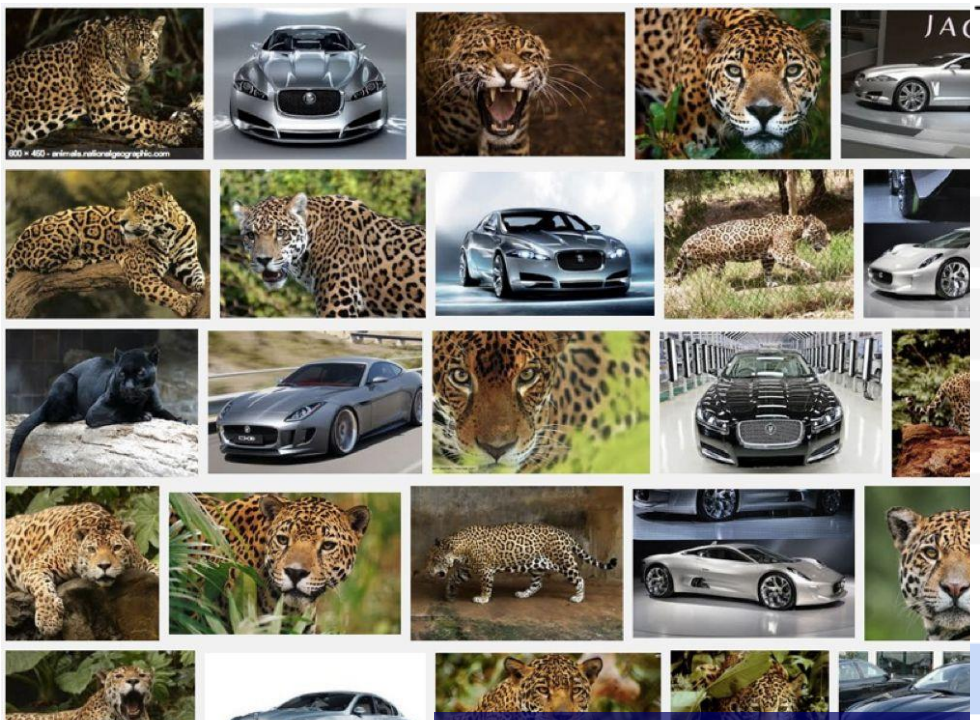
SEO



Knowledge Graph

Why Linked Data?

Problem in recovery of the information



Text: "**Pluto**"

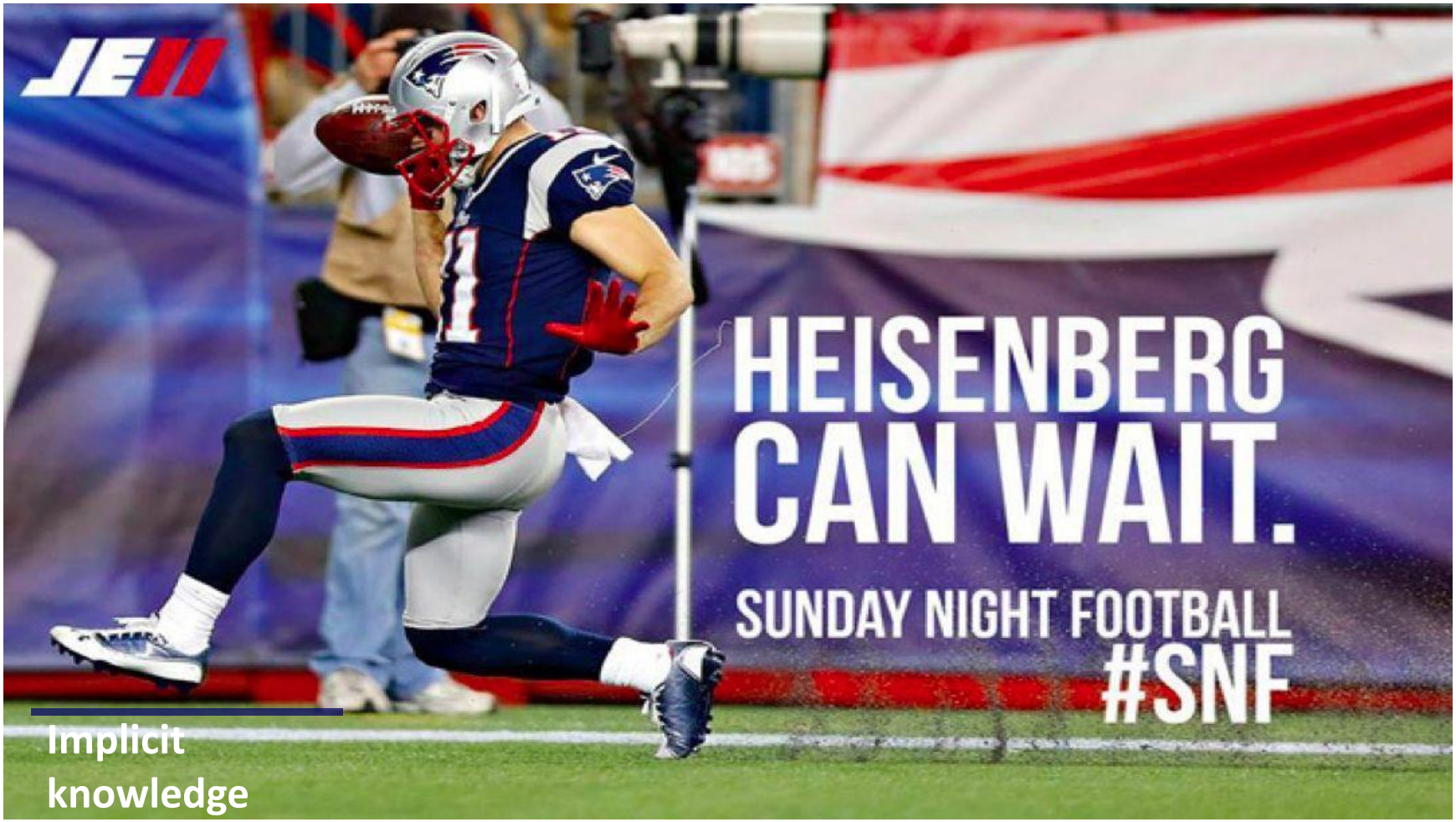
Entity Mapping
Disambiguation

Pluto	a Disney cartoon character
Pluto	a Roman god
Pluto	a song by Björk
HMS Pluto	a ship
...	
Pluto	a dwarf planet

- **Ambiguity of natural language**
  Different words / expressions for the same concept (Synonyms, metaphors)

Why Linked Data?

Problem in recovery of the information



Implicit knowledge

Linked Data

Method of publishing data so that they can be interconnected

It is based on Web technologies, such as HTTP, FOAF, OWL, RDF and URI, but instead of using them for web pages, they are extended to share information in a way that can be read automatically by computers.

- DBpedia - Wikipedia; 3.4 million concepts described by a billion triples (1000 million),
- Bibliografía DBLP - scientific articles, with information of 800,000 articles, 400,000 authors and approximately 15 million triples
- riese - statistical data of 500 million Europeans

Why Linked Data?

- Many ontologies **with similar information in some of their parts**:

  For example, Names, CI, Address, Phone number

- These common parts **could interconnect**, and gather all the data from multiple ontologies in a giant collection of data, to be consulted.

That should lead to creating a swarm of ontologies in the world, and each ontology would be a node of the giant graph.

Why Linked Data?

- The ontologies would be "linked" through the common parts (Name, Address, etc.)
- Users know the ontologies that they need to consult
- Queries are made about individual ontologies
- The common part of an ontology connects with the similar parts of the other
- From that "linked", a subset of that global shared ontology is extracted locally

Knowledge representation

- How do I represent the following fact:
  "*Pluto has been discovered in 1930*"?

Pluto : Planet
discovered = 1930

UML instance

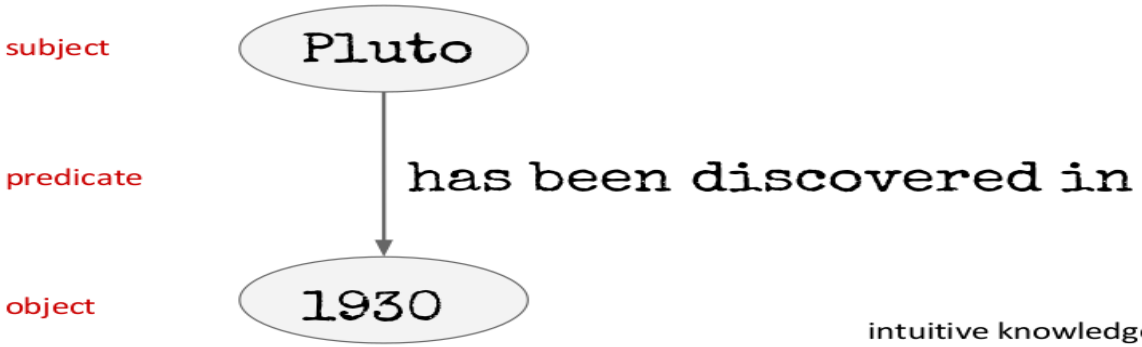
```
<a href="http://en.wikipedia.org/wiki/Pluto">
 Pluto
</a> has been discovered in 1930.
```

HTML

```
<planet name = "Pluto" discovered="1930" />
```

XML

- How do I represent the following fact:
  "*Pluto has been discovered in 1930*" in an intuitive way?



intuitive knowledge representation with a **directed graph**

Knowledge representation

- **RDF Statements (RDF-Triple):**

Subject	+	Property	+	Object / Value
URI		**URI**		**URI / Literal**

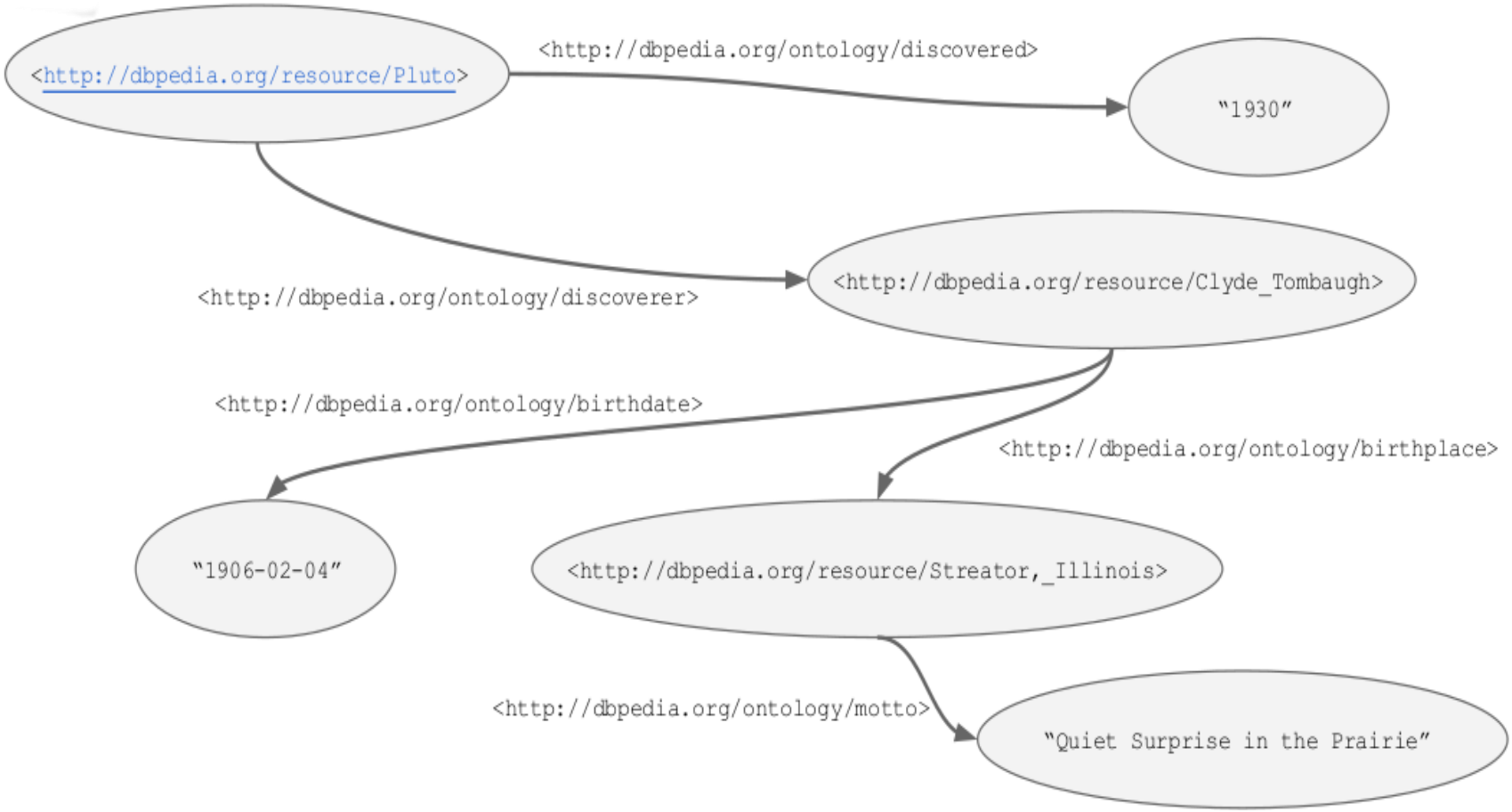
RDF Building Blocks

N-Triples Serialization

```
<http://dbpedia.org/resource/Pluto>    <http://dbpedia.org/ontology/discovered>    "1930" .
```

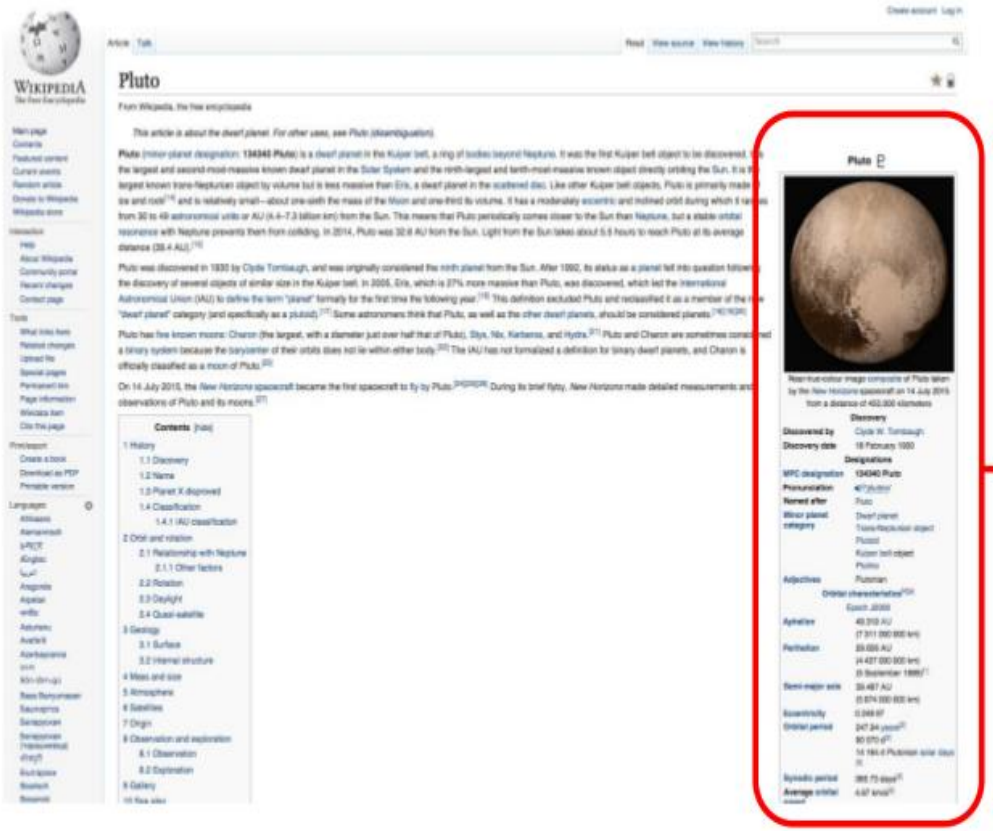
graph representation

Knowledge representation



URI: the key element

http://en.wikipedia.org/wiki/Pluto



http://dbpedia.org/resource/Pluto

Knowledge representation

<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discovered> "1930" .
<http://dbpedia.org/resource/Pluto> <http://dbpedia.org/ontology/discoverer> <http://dbpedia.org/resource/Clyde_Tombaugh> .
<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/CelestialBody> .
<http://dbpedia.org/resource/Pluto> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/place> .
... ...

<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthdate> "1906-02-04" .
<http://dbpedia.org/resource/Clyde_Tombaugh> <http://dbpedia.org/ontology/birthplace> <http://dbpedia.org/resource/Streator,_Illinois> .
... ...

<http://dbpedia.org/resource/Streator,_Illinois> <http://dbpedia.org/ontology/motto> "Quiet Surprise in the Prairie" .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#lat> "41.120834"^^xsd:float .
<http://dbpedia.org/resource/Streator,_Illinois> <http://www.w3.org/2003/01/geo/wgs84_pos#long> "-88.835281"^^xsd:float .
... ...

Subject Property Object

RDF Triples

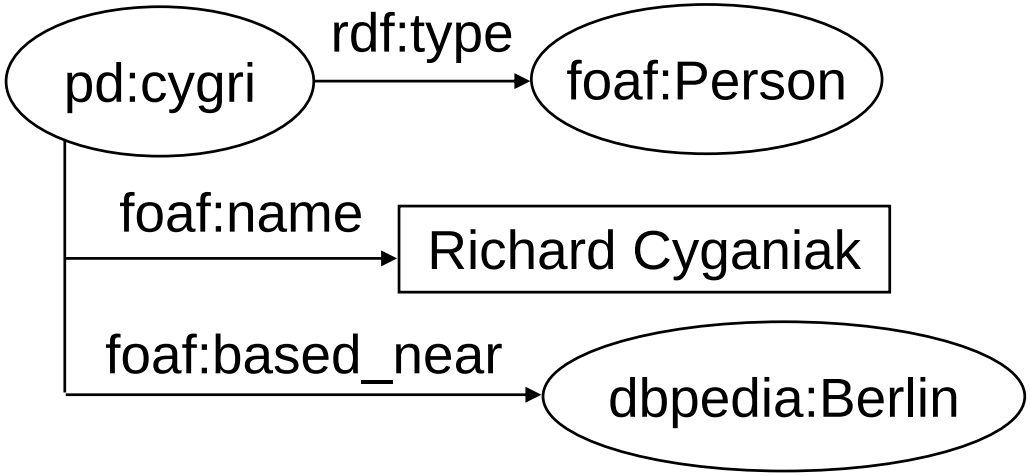
Individuals (Entities)

Literals / Values

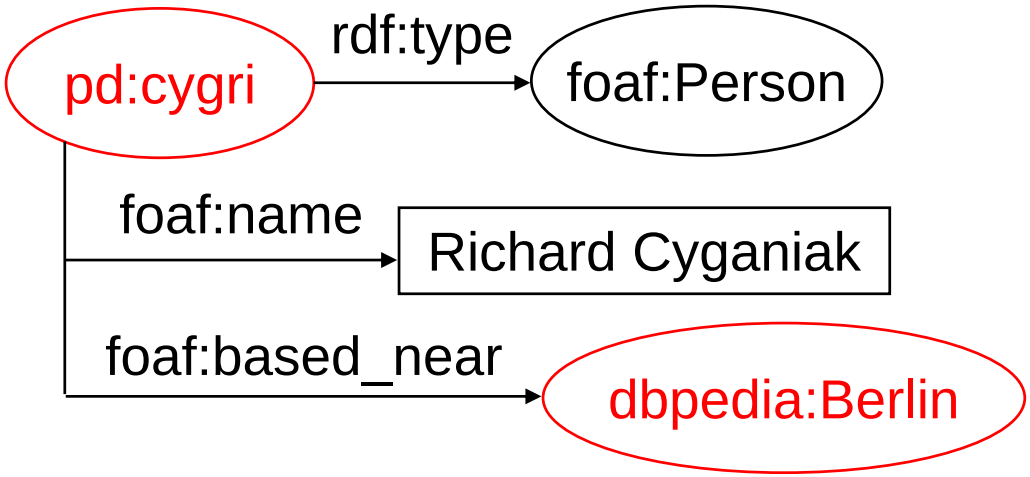
Classes

Vocabularies / Ontologies

Model RDF


pd:cygri
rdf:type
foaf:Person
foaf:name
Richard Cyganiak
foaf:based_near
dbpedia:Berlin

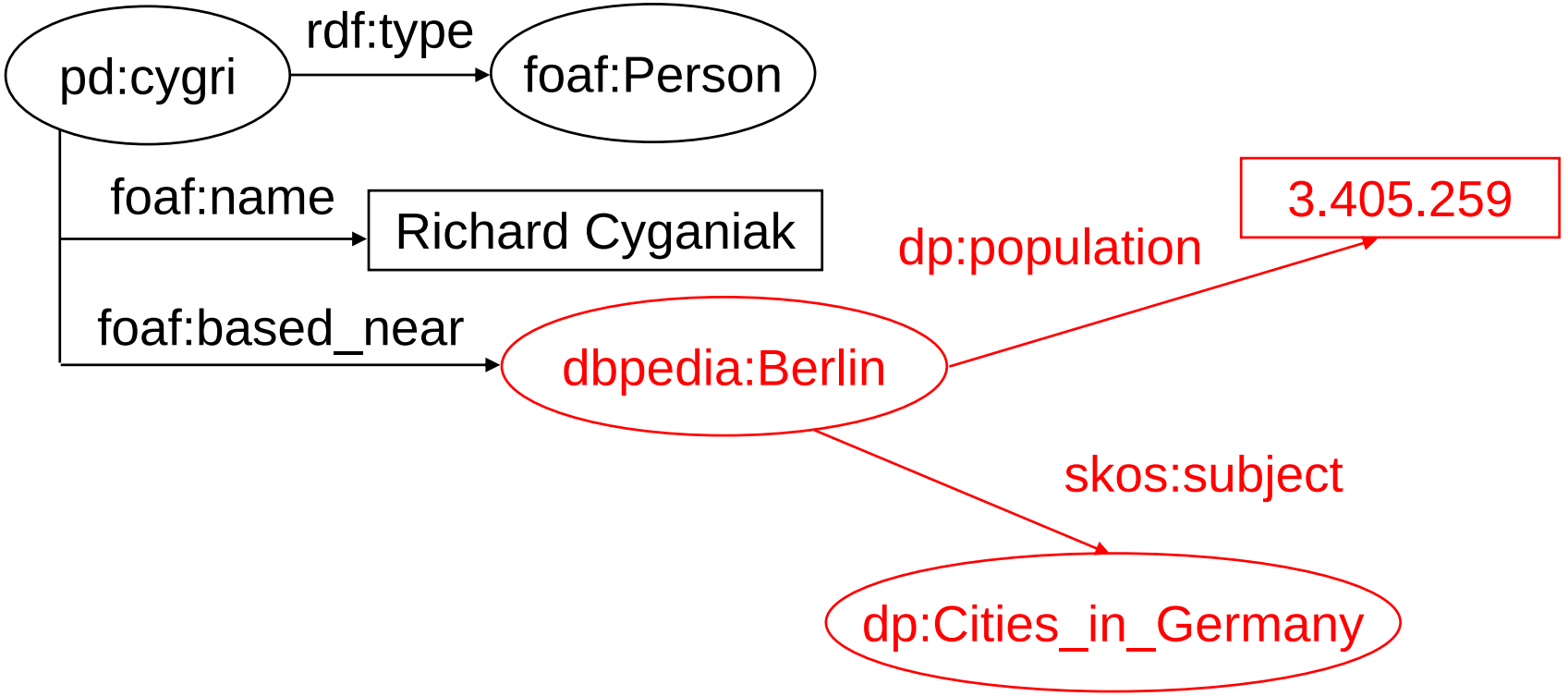
Data are identified with URIs



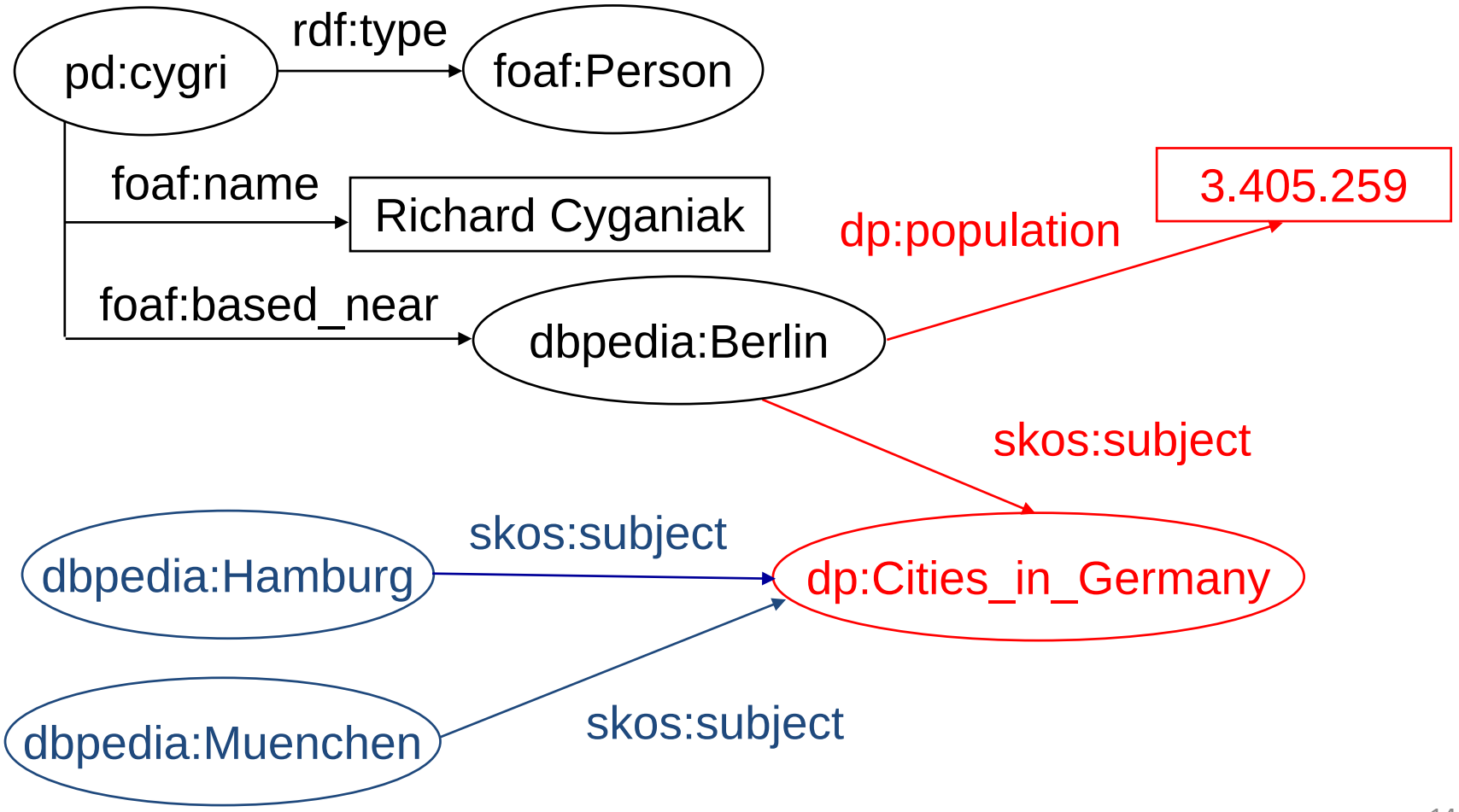
pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri

dbpedia:Berlin = http://dbpedia.org/resource/Berlin

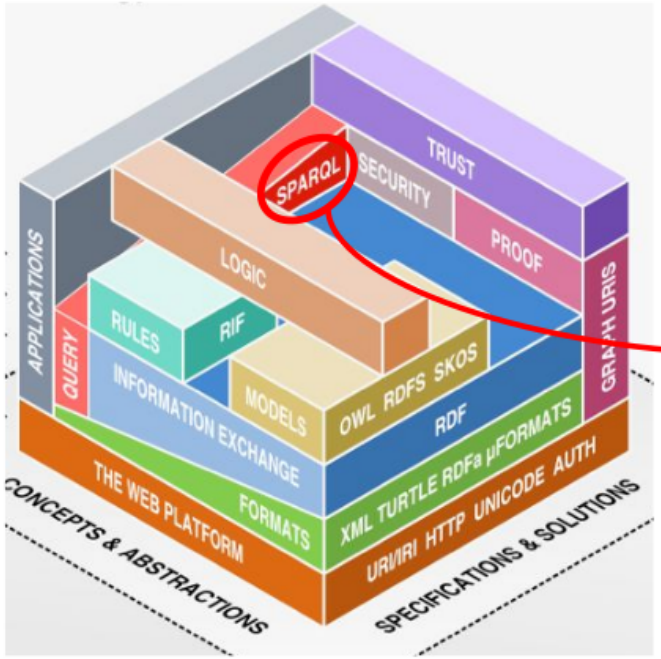
Data are identified with URIs



Data are identified with URIs



Knowledge representation (SPARQL)



search all authors and the titles of their notable works:

specifies namespaces

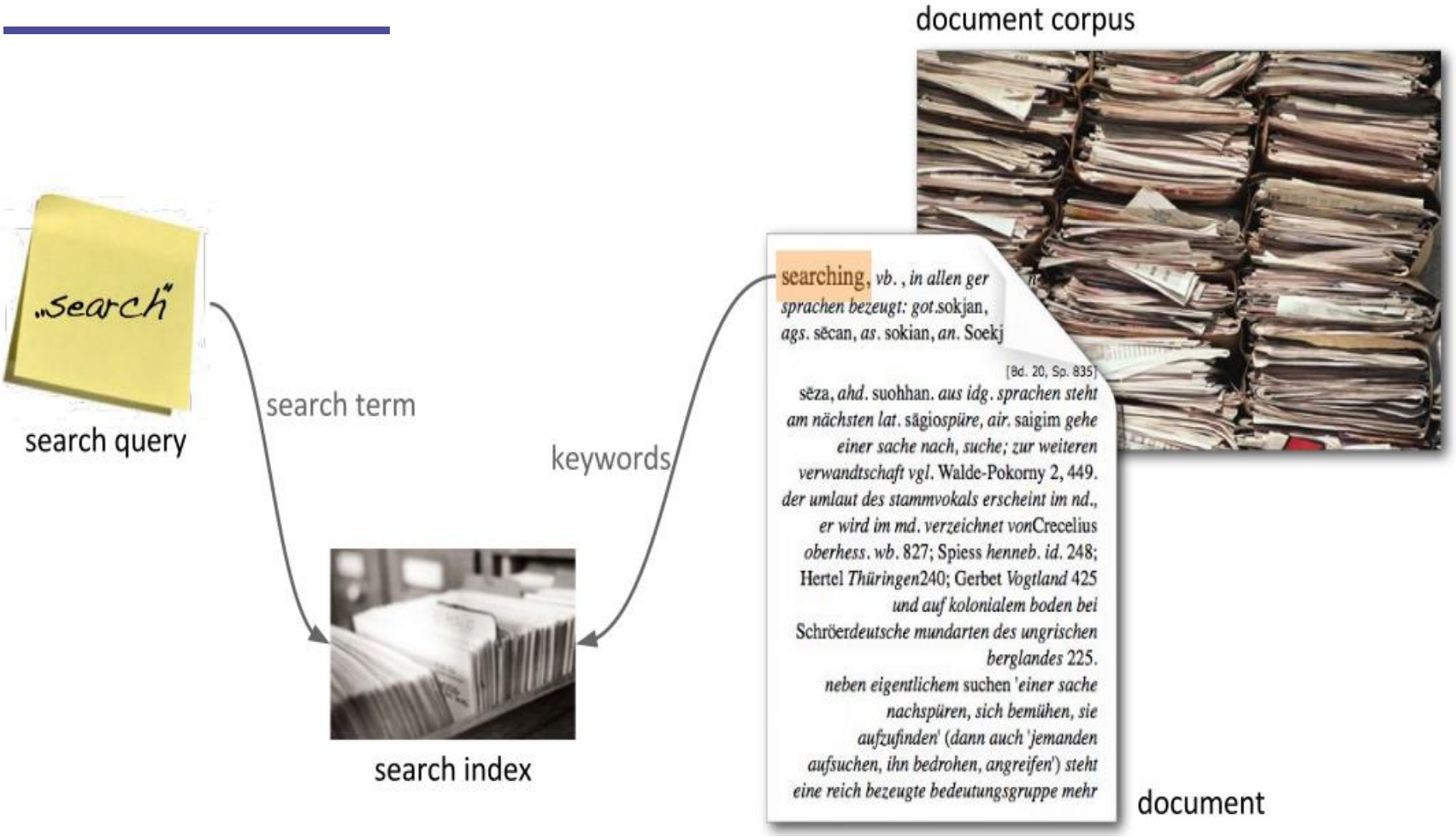
```
PREFIX :     <http://dbpedia.org/resource/>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo:  <http://dbpedia.org/ontology/>

SELECT ?author_name ?title

FROM <http://dbpedia.org/>

WHERE {
     ?author rdf:type dbo:Writer .
     ?author rdfs:label ?author_name .
     ?author dbo:notableWork ?work .
     ?work rdfs:label ?title .
}
```

specifies output variables

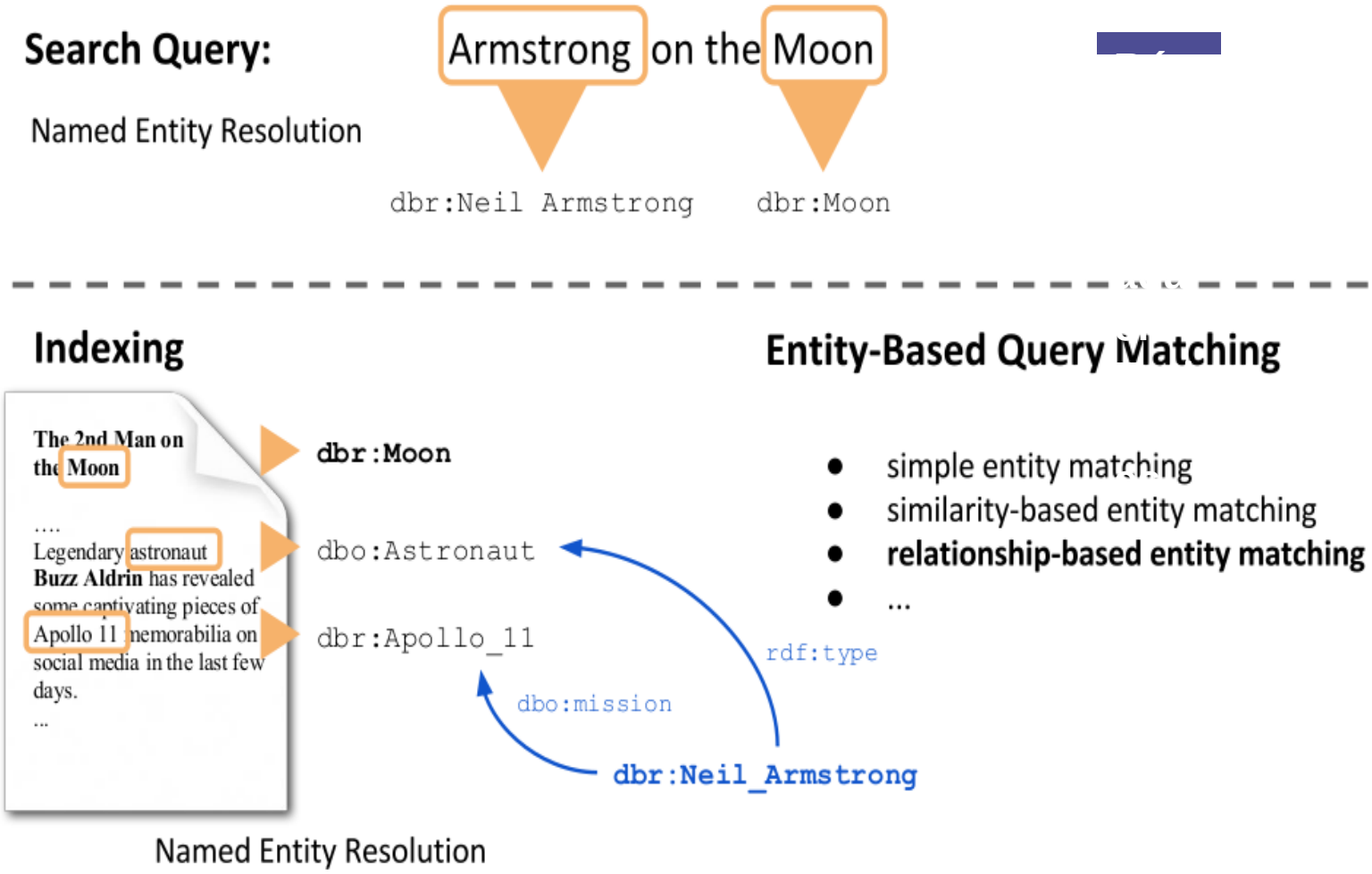
specifies graph to be queried

specifies graph pattern to be matched

Semantic Search



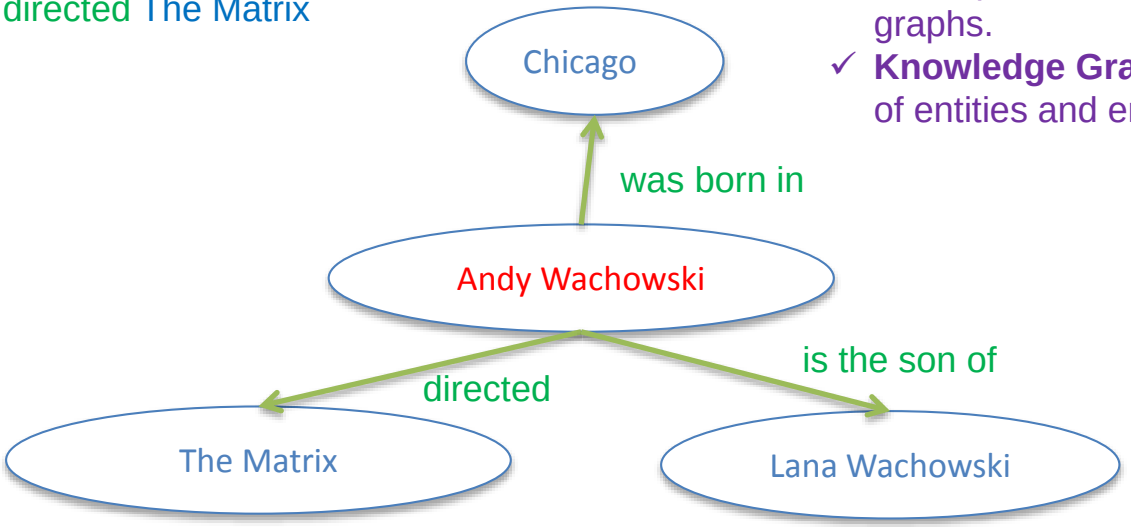
Semantic Search



SEO (Search Engine Optimization)

Andy Wachowski was born in Chicago
Andy Wachowski is the son of Lana Wachowski
Andy Wachowski directed The Matrix

- ✓ The triplets are represented in graphs.
- ✓ **Knowledge Graph** -> is a database of entities and entities among them

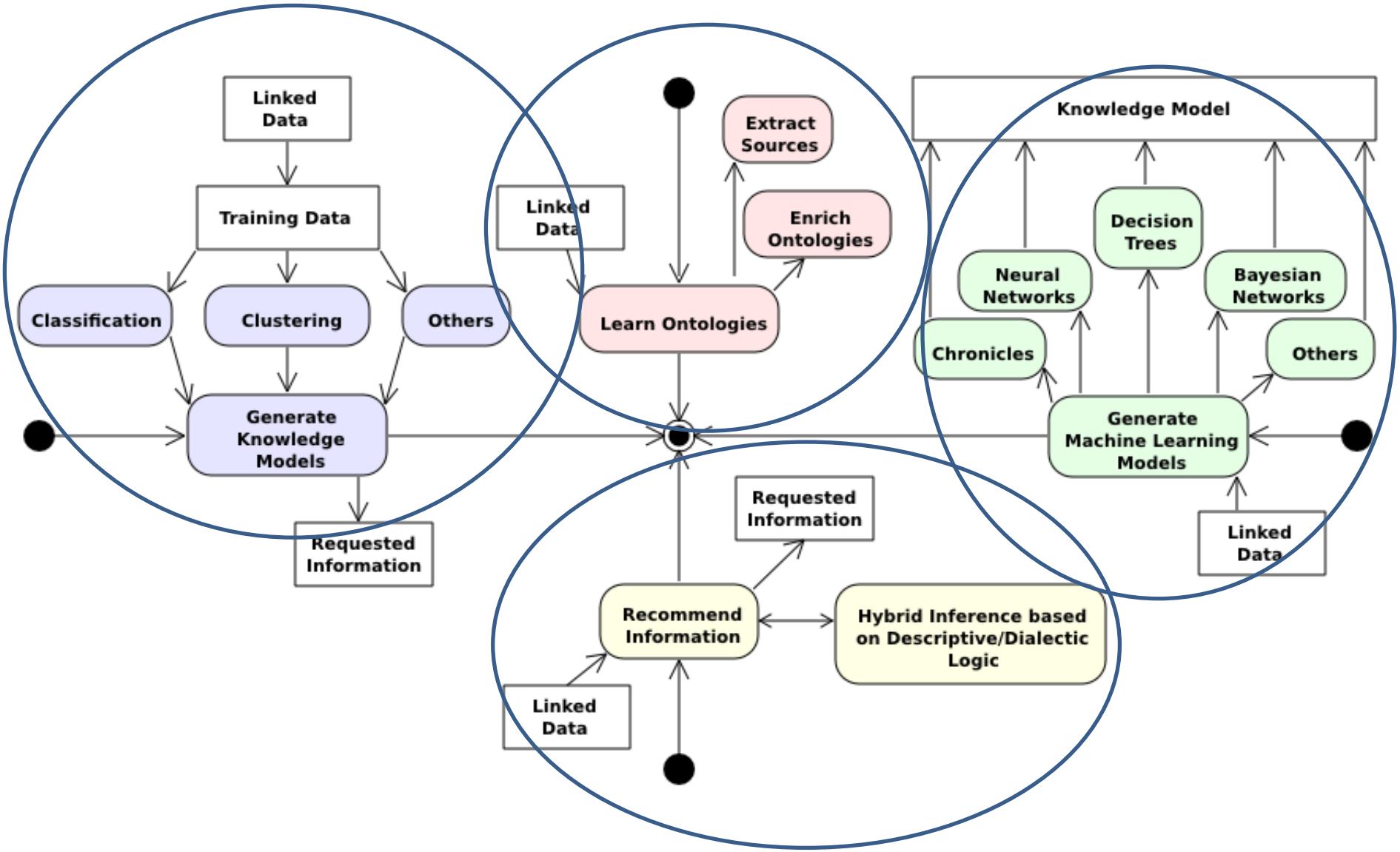


✓Semantic SEO aims to help search engines understand exactly what your pages are about.

✓ To do this, follow the next steps

- ✓Determine the entities corresponding to the page.
- ✓Disambiguate them directly
- ✓Disambiguate them indirectly.

Challenges from linked data



Challenges from linked data

Natural language phenomena

- Vagueness,
- Contingent statements about the future, presupposition failure,
- Counterfactual reasoning,
- Fictional discourse.

Hybrid Inference

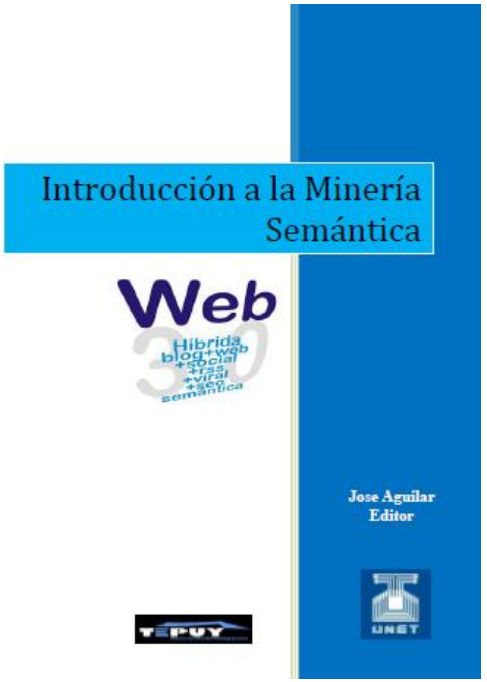
Discovery

DL or RM3 Engine

SEO

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

TEPUY Research+Development



"If you are looking for different results, then do not always do the same"

A. Einstein

www.ing.ula.ve/~aguilar
aguilar@ula.ve

cemisid
centro de estudios en microelectrónica y sistemas distribuidos