

Integración semántica de los recursos de información en una memoria corporativa

Erik Alarcón Zamora

Enero 2014. México, D.F.

Asesores:

Dra. Reyna Carolina Medina Ramírez

Dr. Héctor Pérez Urbina

Contenido I

- 1 Contexto y Motivación
 - Memoria Corporativa
 - Tecnologías Semánticas
 - Integración Semántica
- 2 Descripción del Problema
 - Pregunta Investigación
 - Objetivos
 - Metodología
 - Hipótesis
 - Aportaciones
- 3 Estado del Arte
 - Integración de la Información
 - Búsqueda y Recuperación de la Información
 - Gestión de una Memoria Corporativa
- 4 Integración Semántica de una Memoria Corporativa
 - Marco de Referencia
 - Arquitectura de la Integración Semántica
 - Casos de Uso
 - Representación el Conocimiento
 - Enriquecer el conocimiento en el modelo semántico
 - Buscar y recuperar la información en el modelo semántico
- 5 Prototipo (Aplicación)

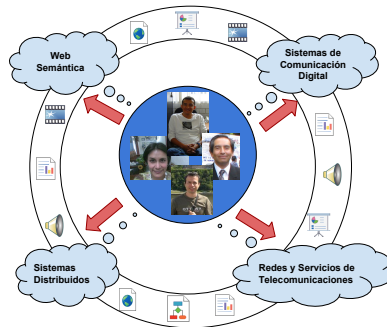
Contenido II

6 Referencias

Memoria Corporativa I

Definición

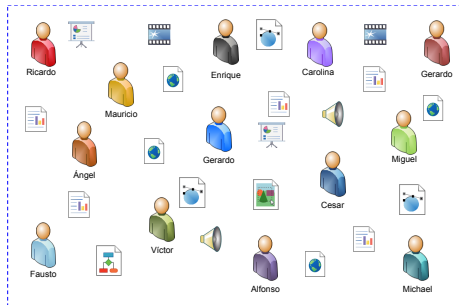
La representación explícita, tácita, consistente y persistente del conocimiento de una organización. [Gandon, 2002]



Memoria Corporativa II

Recurso de Información

Un elemento que representa y encapsula una parte del conocimiento de una organización (investigaciones, colaboraciones, proyectos, cursos, temas de interés, objetos e ideas).



Memoria Corporativa III

Diversidad en formato



pdf, doc, odp, html, txt, xls, wav, png, mp3, mp4, mpeg, mov, ppt, mov

Diversidad en contenido



p2p, middleware, estado global, replicación, concurrencia, sincronización

Diversidad en estructura



estructurados

semi-estructurados

sin estructura

Homonimia

radio ---> Química, Comunicación, Anatomía o Geometría

Sinonimia

herramienta = aparato = instrumento = mecanismo = artificio

Tecnologías Semánticas

Definición

Un conjunto de metodologías, lenguajes, aplicaciones, herramientas y estándares para suministrar u obtener el significado de las palabras, información y las relaciones entre éstos. [Alfred et al., 2010]



Resource Description Framework (RDF) I

Definición

Marco genérico para describir el conocimiento e información explícita de los recursos mediante sus características y relaciones. [Bouzid et al., 2012]

Recurso

Puede ser cualquier cosa (persona, documento, entidad del mundo real o concepto abstracto) que tiene un identificador único de recursos (URI).

Declaración

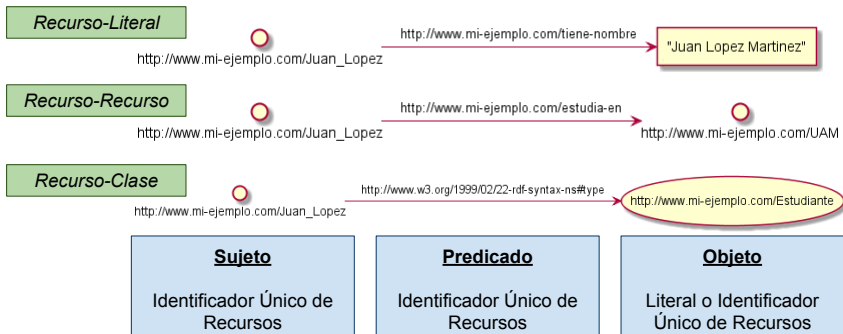
Una afirmación de un hecho explícito sobre un recurso.

- Juan se llama “Juan López Martínez”.
- Juan estudia en la UAM.
- Juan es un estudiante.

Resource Description Framework (RDF) II

Tripleta RDF

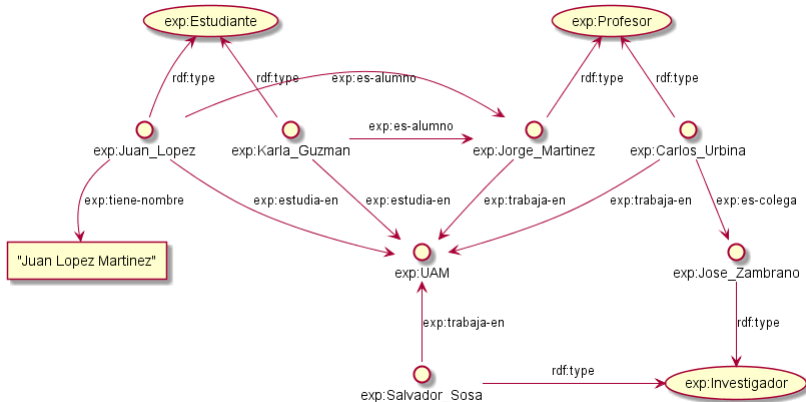
La forma básica para representa una descripción en un modelo semántico.



Resource Description Framework (RDF) III

@prefix exp: <http://www.mi-ejemplo.com/>

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>



Existen distintas sintaxis de serialización: N3, turtle, RDF/XML, N-triples.

SPARQL I

Definición

Lenguaje de consulta y protocolo de acceso a RDF, para la búsqueda y recuperación de la información en un grafo RDF.

###Lista de prefijos

PREFIX exp: <http://www.mi-ejemplo.com/>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

Variables a recuperar

SELECT ?x

WHERE {

Patrones tripletas

?x exp:estudia-en exp:UAM.

?x rdf:type exp:Estudiante. }

Ontología I

Definición

Una definición formal, explícita y compartida de los conceptos, así como las relaciones de un determinado dominio. [Gruber, 1993]

Componentes

- **Componente Asertivo (ABox)** está constituido por descripciones que afirman que los individuos son instancias de una clase o propiedad.
- **Componente Terminológico (TBox)** describe las clases y propiedades relevantes, así como las reglas de inferencia que permiten aprovechar la manera en que las instancias se relacionan entre sí.

Ontología II

Reglas de inferencia o Axiomas

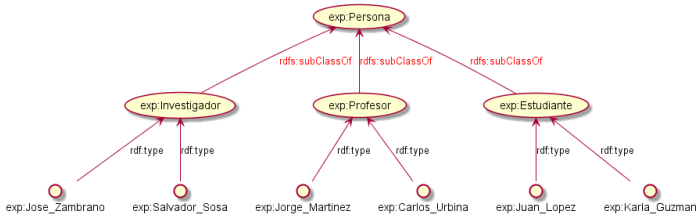
Expresiones para enriquecer un grafo RDF con conocimiento implícito.

-Algo que es evidente para una persona, no lo es para una maquina-.

@prefix exp: <http://www.mi-ejemplo.com/>

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

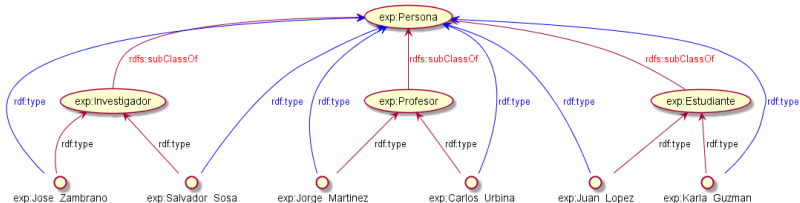
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>



Inferencia

Razonador

Un programa que deduce declaraciones a partir de los axiomas y declaraciones explícitas en la ontología.



Integración Semántica

Definición

La búsqueda y recuperación significativa de información existente en los recursos de información para responder una consulta dada por un usuario.

Etapas

Representar el *conocimiento* de los *recursos de información* en un *modelo semántico*.

Buscar y recuperar información existente en la memoria corporativa mediante la interrogación del modelo semántico.

Pregunta Investigación

*¿Las **tecnologías semánticas** son viables para solucionar la **integración semántica** de los **recursos de información** de una **memoria corporativa**?*



Objetivos

Objetivo Principal

Contribuir a la integración semántica de los recursos de información en una memoria corporativa, mediante el uso de las tecnologías semánticas.

Objetivos Particulares

- 1 Desarrollar una **marco de referencia** para la *integración semántica* de los *recursos de información* existentes en una *memoria corporativa*.
- 2 Implementar un **modelo semántico** que representa el *conocimiento explícito e implícito* de los *recursos de información*.
- 3 Implementar un **prototipo de interfaz gráfica de usuario** que permita a los usuarios una interacción amigable para la integración semántica de los recursos de información.
- 4 Evaluar los **resultados devueltos** y **tiempos de procesamiento** en la *integración semántica* para el dominio de redes y telecomunicaciones.

Metodología

Marco de Referencia

- 1 Identificar los *casos de uso* para encontrar los principales *recursos de información* existentes en la memoria, así como los criterios de búsqueda asociados a éstos.
- 2 Evaluar las herramientas semánticas para la edición de descripciones semánticas, edición de reglas de inferencia, gestión de modelos semánticos.
- 3 Recopilar los recursos de información de acuerdo a los casos de uso.

Modelo Semántico

- 4 Describir el conocimiento explícito de los *recursos de información* recopilados en un modelo semántico (ontología).
- 5 Identificar y construir las reglas de inferencia, con base en el diagrama de clases.

Metodología II

Marco de Referencia

- ⑥ Identificar las preguntas en lenguaje natural y transformarlas a *consultas SPARQL*.
- ⑦ Emplear un razonador para hacer explícito el conocimiento implícito.
- ⑧ Buscar y recuperar información en la memoria corporativa, interrogando el modelo semántico (ontología).

Prototipo de interfaz gráfica de usuario

- ⑨ Diseñar un prototipo para interacción (búsqueda y navegación) amigable y transparente de los usuarios de la memoria con la ontología.
- ⑩ Implementar el prototipo y realizar pruebas del mismo.

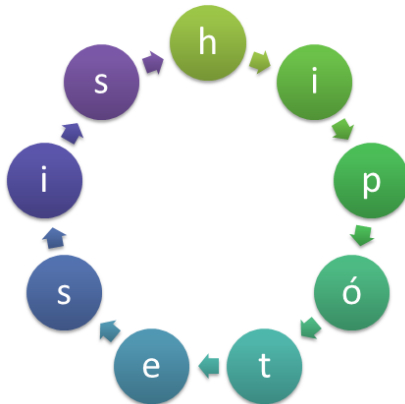
Metodología III

Evaluación

- 11 Evaluar la calidad de los resultados con y sin inferencia, comparando los *recursos relevantes recuperados* para una pregunta con los recursos que realmente responden la pregunta.
- 12 Evaluar los *tiempos promedios* que toma la herramienta electa de gestión de los modelos semánticos, para consultar los modelos con/sin inferencia.

Hipótesis

El uso de las *tecnologías semánticas* es adecuado para lograr la *integración semántica de recursos de información* en una *memoria corporativa*.



Aportaciones

- 1 Un *marco de referencia* para lograr la integración semántica de recursos de información.
- 2 Un modelo semántico que representa el conocimiento de una memoria corporativa, el cual tiene tres ramas principales (Personas, Recursos Digitales y Conceptos del Redes y Telecomunicaciones).
- 3 Un prototipo (interfaz gráfica de usuario) para la interacción amigable (búsqueda y consulta de información) de los usuarios al modelo semántico.
- 4 Los resultados de nuestra evaluación experimental.
- 5 Un par de scripts para la generación automática y controlada de descripciones (conocimiento explícito) de los *recursos de información*, con el fin de poblar la base de conocimiento.

Estado del Arte

Ejes claves

- 1 Integración de la información a partir del uso de tecnologías semánticas.
- 2 Búsqueda, recuperación y publicación de la información desde una ontología.
- 3 Gestión de una memoria corporativa.

Integración de la Información I

[Moner et al., 2006]

Una *arquitectura dual* que representa consistentemente la información clínica de cualquier persona, para que los profesionales de la salud accedan al historial clínico de los pacientes.

[Zhai et al., 2008]

Una arquitectura basada en ontologías y el *lenguaje de marcado extensible*, para la *integración semántica* en *sistemas de información de energía eléctrica* (sistemas heterogéneos).

[Yang and Steele, 2011]

Un *marco de integración semántica* para la integración de la información en el dominio del alojamiento en-línea (información en constante cambio), con el fin de reunir y compartir esta información.

Búsqueda y Recuperación de la Información I

[Cao et al., 2011]

Un sistema para la recomendación y el acceso a la información turística semántica que emplea una ontología para definir conceptos y propiedades asociadas a la información y conocimiento sobre el turismo.

[Ha et al., 2011]

Una propuesta para representar y recuperar información sobre los contenidos de los *manuals de mantenimiento electrónicos*, basado en una ontología, así como diseñar un *sistema de visualización* de la información en una ontología.

Búsqueda y Recuperación de la Información II

[Suganyakala and Rajalaxmi, 2013]

Un *marco de trabajo* para la recuperación de información basado en ontología (Película), así como el diseño de una *interfaz de consulta* para que un usuario pueda escribir una consulta en lenguaje natural.

[Salam, 2013]

Un método basado en ontologías para recuperar la *información contenida en un documento* a partir de los conceptos asociados que están en una ontología, mediante el uso de un *motor de búsqueda e indexación semántico*.

Gestión de una Memoria Corporativa

[Xin and Guangleng, 2001]

Un enfoque basado en las **ontologías**, para describir los *objetos de justificación del diseño*, así como acceder de manera uniforme a los *recursos de información*.

[Chakhmoune et al., 2011]

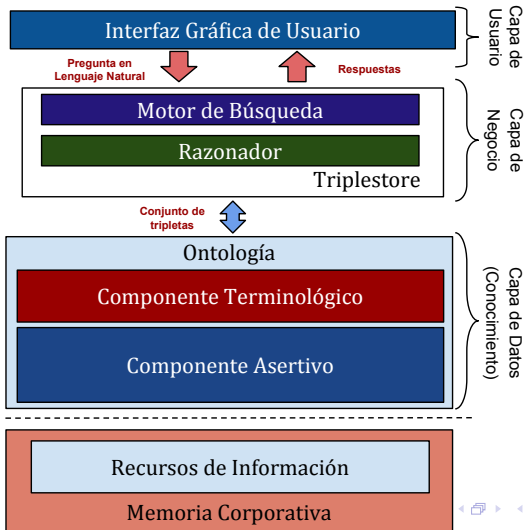
Un entorno de colaboración para la gestión de memorias corporativas, cuya función es construir una *memoria corporativa* a partir de *lluvia de ideas* y un *mecanismo de toma de decisiones consensuadas*.

Marco de Referencia

Etapas

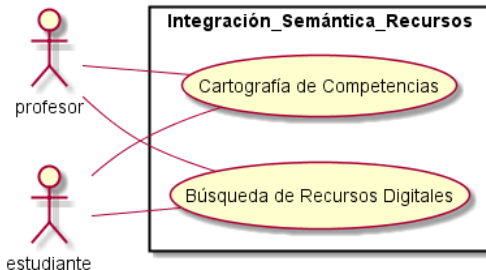
- 1 ***Representación del conocimiento explícito de los recursos*** consiste en recuperar los recursos de información de la memoria corporativa y representar las características y/o relaciones de los mismos con el estándar RDF.
- 2 ***Enriquecimiento del conocimiento en el modelo*** consiste en introducir axiomas, para enriquecer el modelo semántico con conocimiento implícito del dominio.
- 3 ***Búsqueda y recuperación de la información en el modelo*** consisten en identificar las principales consultas de los usuarios e interrogar la ontología para responder estas consultas.

Arquitectura de la Integración Semántica



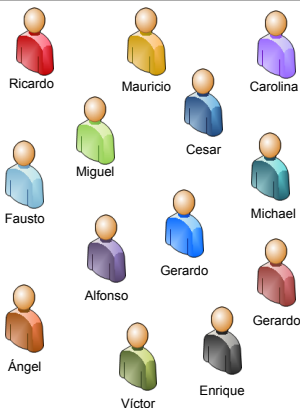
Casos de Uso

- **Cartografía de Competencias** consiste en la búsqueda y recuperación de información significativa de las personas a partir de las características personales y profesionales de las mismas.
- **Búsqueda de Recursos Digitales** consiste en la búsqueda y recuperación de información significativa de los documentos y archivos multimedia a partir del contenido de los mismos.

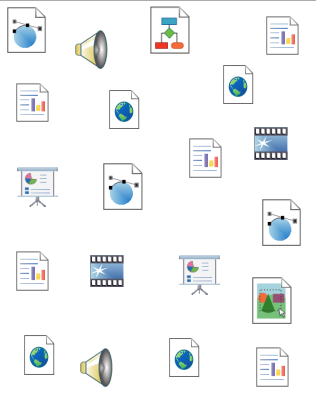


Identificar los principales recursos de información

Cartografía de Competencias

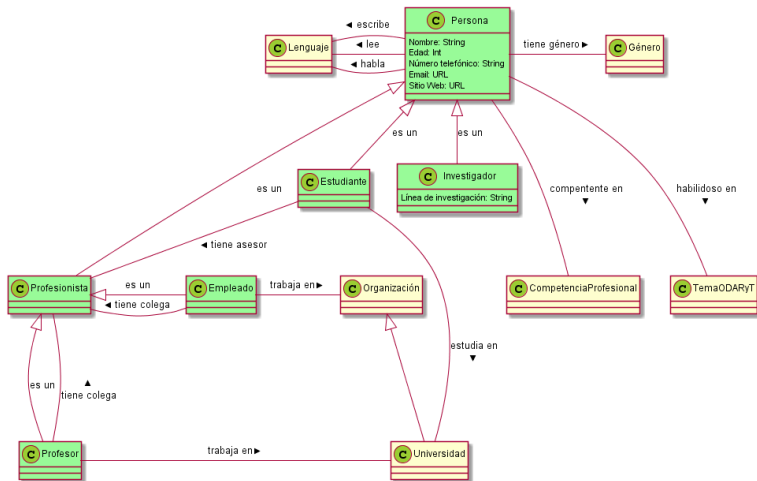


Búsqueda Recursos Digitales



Memoria Corporativa del área de RyT

Adquirir y expresar el conocimiento de los recursos de información



Representar el conocimiento e información mediante el estándar RDF I

Actividades en la representación del conocimiento

- 1 Asignar un identificador único de recursos para cada recurso de información en la memoria corporativa.
- 2 Asignar los identificadores únicos de recursos a las propiedades.
- 3 Reconocer los valores de las propiedades: otro recurso o literal.
- 4 Generar las tripletas RDF asociadas a las descripciones de los recursos de información.

```
@prefix sirp: <http://arte.izt.uam.mx/ontologies/personRyT.owl#>  
@prefix sird: <http://arte.izt.uam.mx/ontologies/digiResourceRyT.owl#>  
@prefix redes: <http://mcyti.izt.uam.mx/arios/odaryt.owl#>
```

Representar el conocimiento e información mediante el estándar RDF II

@prefix sirp: <http://arte.izt.uam.mx/ontologies/personRyT.owl#> .

@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

@prefix redes: <http://mcyti.izt.uam.mx/arios/odaryt.owl#> .

sirp:RicardoMarcelinJimenez

a sirp:Teacher ;
sirp:has-name "Ricardo Marcelin Jiménez"^^xsd:string;
sirp:has-email "calu@xanum.uam.mx"^^xsd:anyURI;
sirp:has-webSite "http://cbi.izt.uam.mx/electrica/profs/ricardo_marcelin.html"^^xsd:anyURI;
sirp:has-gender sirp:Male;
sirp:worksIn sirp:UAM;
sirp:researchesOn "El almacenamiento distribuido, las redes inalámbricas de sensores y la simulación de eventos discretos."^^xsd:string;
sirp:expertiseIn redes:Distributed_Systems, redes:Distributed_Storage, redes:MDS_Codes, redes:Performance_evaluation, redes:Semantic_Annotations, redes:Image_compression, redes:Routing_Protocols, redes:Distributed_Algorithms, redes:Wireless_Sensor_Networks, redes:N_and_ST;
sirp:competentIn sirp:Article_Reviewing_Skills, sirp:Thesis_Supervision_Skills, sirp:Oral_And_Written_Communication_Skills, sirp:Area_Expert, sirp:Analysis_Skills, sirp:Decision_Making_Skills, sirp:Research_Skills, sirp:Problem_Solving_Skills, sirp:Synthesis_Skills, sirp:Abstraction_Skills, sirp:Counseling_Skills_for_Social_Service, sirp:IT_And_Communication_Skills;
sirp:has-colleague sirp:MiguelLopez, sirp:CarolinaMedinaRamirez;
sirp:reads sirp:Spanish, sirp:English;
sirp:writes sirp:Spanish, sirp:English;
sirp:speaks sirp:Spanish, sirp:English.



Representar el conocimiento e información mediante el estándar RDF III



```
@prefix sird: <http://arte.izt.uam.mx/ontologies/digiResourceRyT.owl#> .
```

```
@prefix sirp: <http://arte.izt.uam.mx/ontologies/personRyT.owl#> .
```

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

```
@prefix redes: <http://mcyti.izt.uam.mx/arios/odaryt.owl#> .
```

```
sird:What-isLinkedData-2012-flv
```

```
  a      sird:Video ;
```

```
  sird:has-title "What is Linked Data?"^^xsd:string ;
```

```
  sird:has-author "Manu Sporny"^^xsd:string ;
```

```
  sird:has-filePath "http://www.youtube.com/watch?v=4x_xzT5eF5Q"^^xsd:anyURI ;
```

```
  sird:has-fileExtension sird:flv ;
```

```
  sird:has-languageSource sirp:English ;
```

```
  sird:has-topic redes:Linked_data, redes:Data, redes:Web, redes:Semantic,  
                redes:Knowledge_graph ;
```

```
  sird:has-synopsis "A short non-technical introduction to Linked Data, Google's  
                  Knowledge Graph, and Facebook's Open Graph Protocol"  
                  ^^xsd:string ;
```

```
  sird:has-yearOfCreation "2012"^^xsd:int .
```

Enriquecer el conocimiento en el modelo semántico

Axiomatización

Para cada *caso de uso* debe encontrarse el respectivo conjunto de axiomas (TBox).

Lenguajes

Especificaciones para describir clases, propiedades e individuos.

- *RDF Schema* **RDF(S)**
- *Web Ontology Language* **OWL**

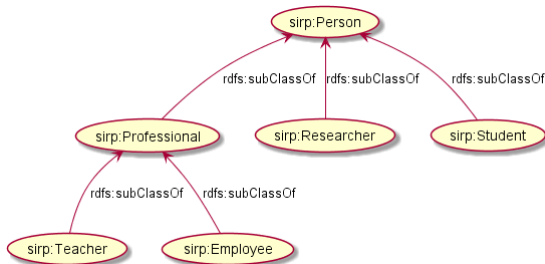
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>

@prefix owl: <http://www.w3.org/2002/07/owl#>

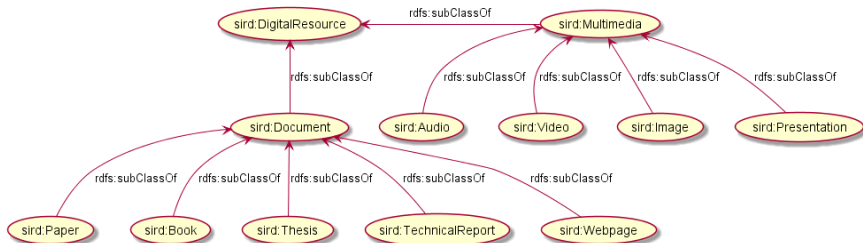
Herencia de Clases I

Subclase (rdfs:subClassOf)

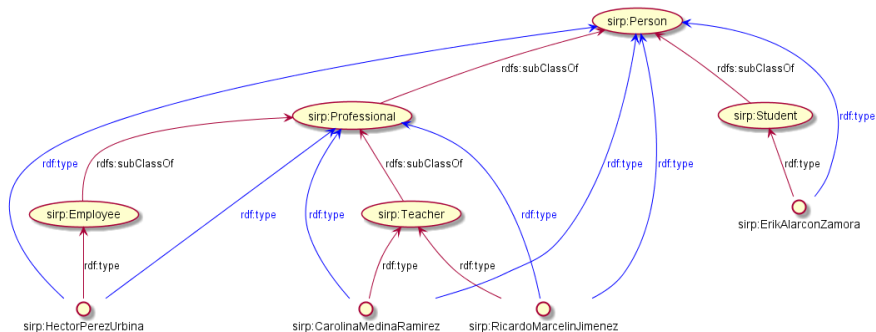
Afirma que una *clase A* se subsume por una *clase B*, es decir, la clase A es un caso particular de la *clase B*. En este caso, las instancias de la clase A son instancias de la clase B.



Herencia de Clases II



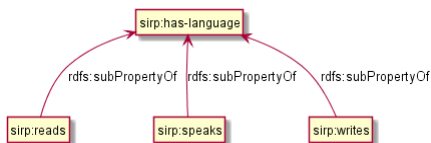
Herencia de Clases III



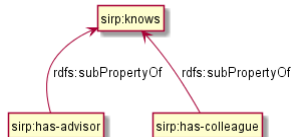
Herencia de Propiedades I

Subpropiedad (rdfs:subPropertyOf)

Afirma que todos los recursos que se relacionan por la *propiedad X*, también se relacionan por la *propiedad Y*.

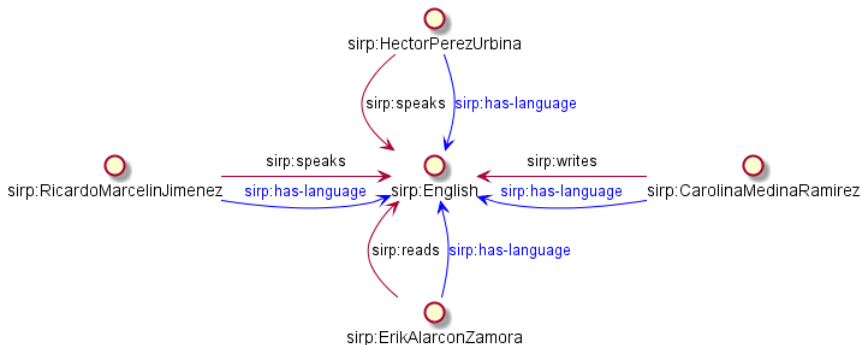


(a) Jerarquía de propiedades para las habilidades lingüísticas



(b) Jerarquía de propiedades para las relaciones profesionales entre personas

Herencia de Propiedades II



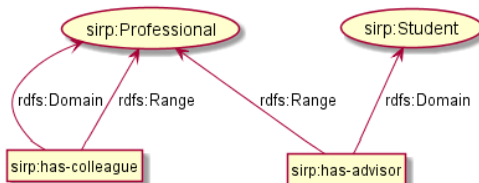
Dominio y Rango en las Propiedades I

Dominio (rdfs:domain)

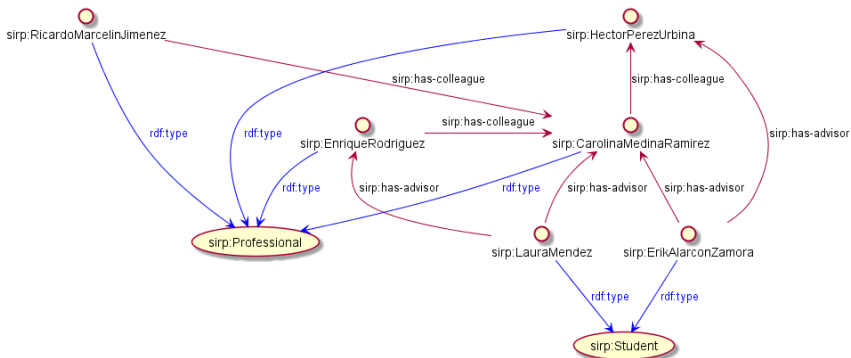
Especifica qué clase se aplica a una propiedad.

Rango (rdfs:range)

Especifica los valores (clase o tipo de literal) que puede asumir una propiedad.



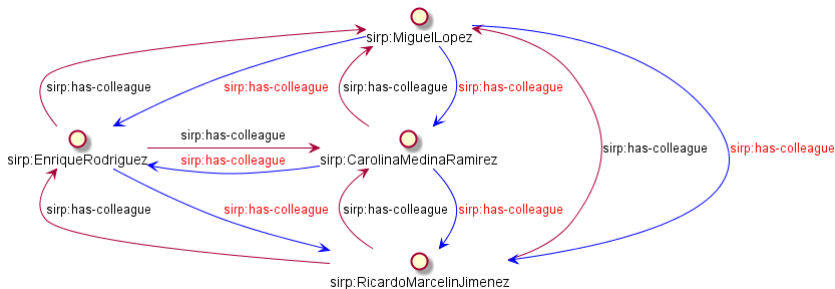
Dominio y Rango en las Propiedades II



Características en las propiedades

Propiedad simétrica (owl:SymmetricProperty)

Afirma que la *propiedad X* es su propia propiedad inversa, es decir, si la *propiedad X* relaciona al *individuo A* con el *individuo B*, entonces, esta propiedad debe relacionar al *individuo B* con el *individuo A*.



Buscar y recuperar la información en el modelo semántico

Objetivo

La búsqueda y recuperación de la información para responder las preguntas o necesidades informativas de los usuarios del área de Redes y Telecomunicaciones (RyT).

Actividades

- 1 Identificar las preguntas en lenguaje natural.
- 2 Transformar las preguntas a una consultas SPARQL.
- 3 Ejecutar las consultas mediante un motor de búsqueda SPARQL.

Transformar las preguntas a una consultas SPARQL I

¿Cuáles son los nombres y sitios Web de las personas que conocen a Carolina Medina Ramírez?



```
PREFIX sirp: <http://arte.izt.uam.mx/ontologies/personRyT.owl#>

SELECT ?name ?ws
WHERE
{
    {?x sirp:has-colleague sirp:CarolinaMedinaRamirez.} UNION
    {sirp:CarolinaMedinaRamirez sirp:has-colleague ?x.} UNION
    {?x sirp:knows sirp:CarolinaMedinaRamirez.} UNION
    {sirp:CarolinaMedinaRamirez sirp:knows ?x.} UNION
    {?x sirp:has-advisor sirp:CarolinaMedinaRamirez.}
    ?x sirp:has-name ?name;
        sirp:has-webSite ?ws.
}
```

Uso de inferencia I

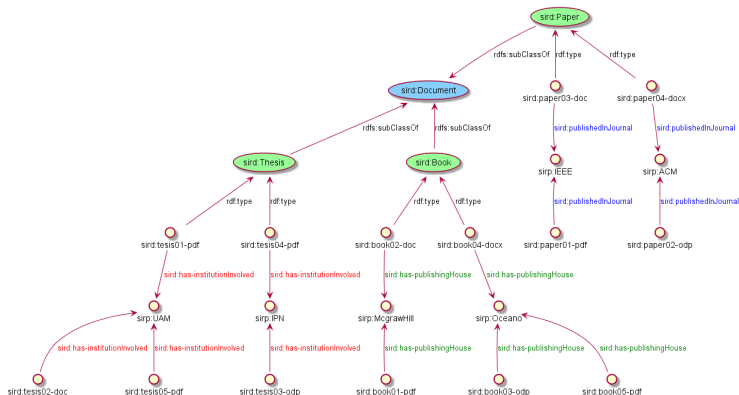


Figura : Grafo RDF sin inferencia

Uso de inferencia II

```
PREFIX sird: <http://arte.izt.uam.mx/ontologies/digiResourceRyT.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
SELECT ?x
WHERE
{
    {?x rdf:type sird:Paper.} UNION
    {?x rdf:type sird:Book.} UNION
    {?x rdf:type sird:TechnicalReport.} UNION
    {?x rdf:type sird:Thesis.} UNION
    {?x rdf:type sird:Webpage.} UNION
    {?x rdf:type sird:Document.}
}
```

(a) Consulta sin inferencia

?x
<i>sird:tesis01-pdf</i>
<i>sird:tesis04-pdf</i>
<i>sird:book02-doc</i>
<i>sird:book04-docx</i>
<i>sird:paper01-pdf</i>
<i>sird:paper02-odp</i>

(b) Resultados de la consulta

Uso de inferencia III

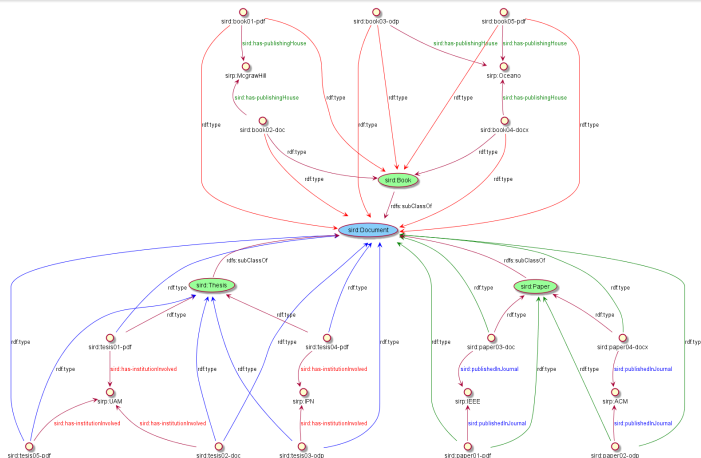


Figura : Grafo RDF con inferencia

Uso de inferencia IV

```
PREFIX sird: <http://arte.izt.uam.mx/ontologies/digiResourceRyT.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
SELECT ?x
WHERE
{
  ?x rdf:type sird:Document.
}
```

(a) Consulta con inferencia

<i>?x</i>
<i>sird:tesis01-pdf</i>
<i>sird:tesis02-doc</i>
<i>sird:tesis03-odp</i>
<i>sird:tesis04-pdf</i>
<i>sird:tesis05-pdf</i>
<i>sird:book01-pdf</i>
<i>sird:book02-doc</i>
<i>sird:book03-odp</i>
<i>sird:book04-docx</i>
<i>sird:book05-pdf</i>
<i>sird:paper01-pdf</i>
<i>sird:paper02-odp</i>
<i>sird:paper03-doc</i>
<i>sird:paper04-docx</i>

(b) Resultados de la consulta

Prototipo (Aplicación) I

Dificultad

La búsqueda y recuperación en este modelo no son sencillas, porque se requiere que un usuario tenga un grado especializado de conocimientos en el uso de las tecnologías semánticas y los vocabularios en las ontologías.

Referencias I

- [Alfred et al., 2010] Alfred, S., Arpah, A., Lim, L. H. S., and Sarinder, K. K. S. (2010). Semantic technology: An efficient approach to monogenean information retrieval. In *Computer and Network Technology (ICCNT), 2010 Second International Conference on*, pages 591–594.
- [Bouzid et al., 2012] Bouzid, S., Cauvet, C., and Pinaton, J. (2012). A survey of semantic web standards to representing knowledge in problem solving situations. In *Information Retrieval Knowledge Management (CAMP), 2012 International Conference on*, pages 121–125.
- [Cao et al., 2011] Cao, T.-D., Phan, T.-H., and Nguyen, A.-D. (2011). An ontology based approach to data representation and information search in smart tourist guide system. In *Knowledge and Systems Engineering (KSE), 2011 Third International Conference on*, pages 171–175.
- [Chakhmoune et al., 2011] Chakhmoune, R., Behja, H., and Marzak, A. (2011). Building corporate memories in collaborative way using ontologies: Case study of a ssii. In *Next Generation Networks and Services (NGNS), 2011 3rd International Conference on*, pages 23–28.

Referencias II

- [Gandon, 2002] Gandon, Fabien, L. (2002).
Ontology Engineering: a Survey and a Return on Experience.
Technical Report RR-4396, INRIA.
- [Gruber, 1993] Gruber, T. R. (1993).
A translation approach to portable ontology specifications.
Knowl. Acquis., 5(2):199–220.
- [Ha et al., 2011] Ha, I., Oh, K.-J., and Jo, G.-S. (2011).
Ontology-driven visualization system for semantic search.
In Information Science and Applications (ICISA), 2011 International Conference on, pages 1–6.
- [Moner et al., 2006] Moner, D., Maldonado, J., Bosca, D., Fernandez, J., Angulo, C., Crespo, P., Vivancos, P., and Robles, M. (2006).
Archetype-based semantic integration and standardization of clinical data.
In Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE, pages 5141–5144.
- [Salam, 2013] Salam, F. (2013).
New semantic indexing and search system based on ontology.
In Emerging Intelligent Data and Web Technologies (EIDWT), 2013 Fourth International Conference on, pages 313–318.

Referencias III

- [Suganyakala and Rajalaxmi, 2013] Suganyakala, R. and Rajalaxmi, R. (2013).
Movie related information retrieval using ontology based semantic search.
In Information Communication and Embedded Systems (ICICES), 2013 International Conference on, pages 421–424.
- [Xin and Guangleng, 2001] Xin, W. and Guangleng, X. (2001).
Design rationale as part of corporate technical memory.
In Systems, Man, and Cybernetics, 2001 IEEE International Conference on, volume 3, pages 1904–1908 vol.3.
- [Yang and Steele, 2011] Yang, K. and Steele, R. (2011).
A semantic integration solution for online accommodation information integration.
In Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on, pages 1105–1110.
- [Zhai et al., 2008] Zhai, J., Li, J., and Wang, Q. (2008).
Using ontology and xml for semantic integration of electricity information systems.
In Electric Utility Deregulation and Restructuring and Power Technologies, 2008. DRPT 2008. Third International Conference on, pages 2197–2201.