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

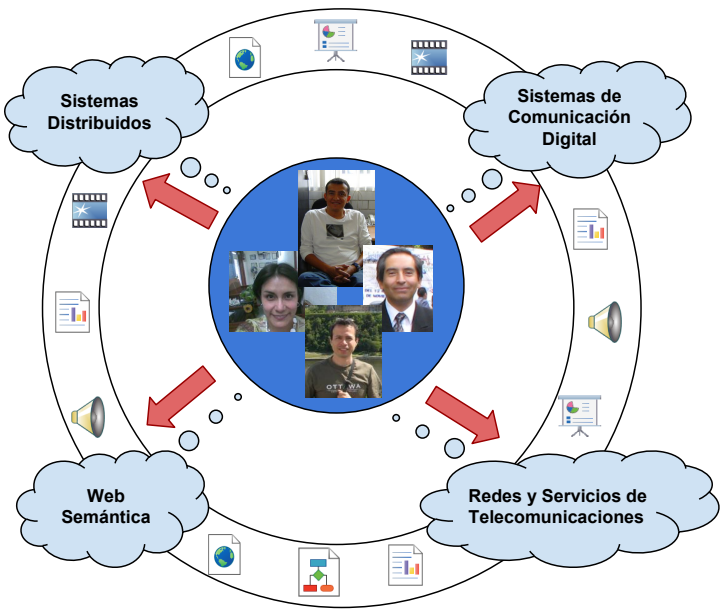
Integración Semántica de los Recursos de Información en una Memoria Corporativa

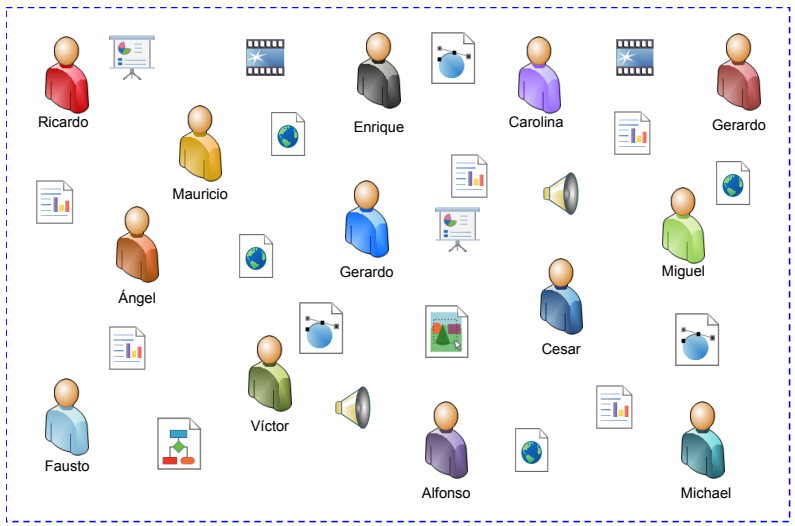
Búsqueda y recuperación de información significativa existente en los recursos de información.

Memoria Corporativa

Definición

Una memoria corporativa (MC) es una representación explícita, tácita, consistente y persistente del conocimiento de una organización.





Heterogeneidad y significado de la información

Diversidad en formato

pdf, doc, odp, html, txt, xsl, 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, Telecomunicaciones, Anatomía, Geometría

Sinonimia

herramienta ≡ aparato ≡ instrumento ≡ mecanismo ≡ artilugio

Tecnologías Semánticas

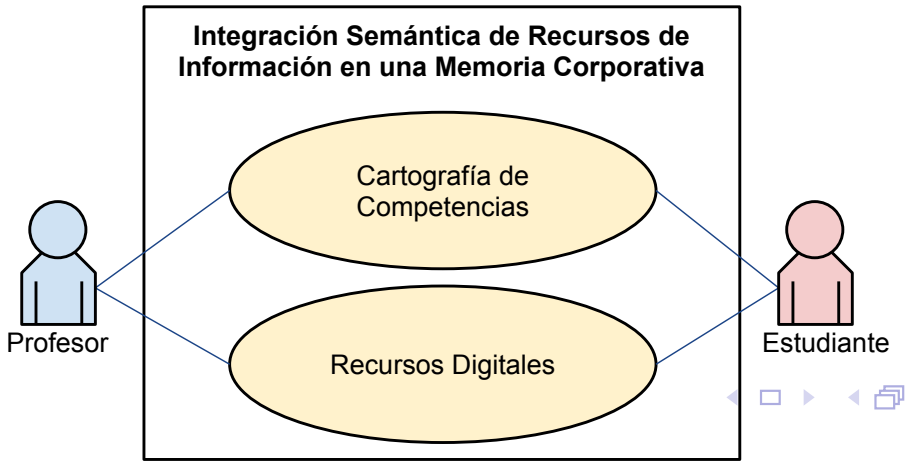
Definición

Las tecnologías semánticas (TS) son un conjunto de metodologías, lenguajes, aplicaciones, herramientas y estándares para suministrar u obtener el significado de las palabras, información y las relaciones entre éstos.

Integración Semántica de los Recursos de Información en una Memoria Corporativa

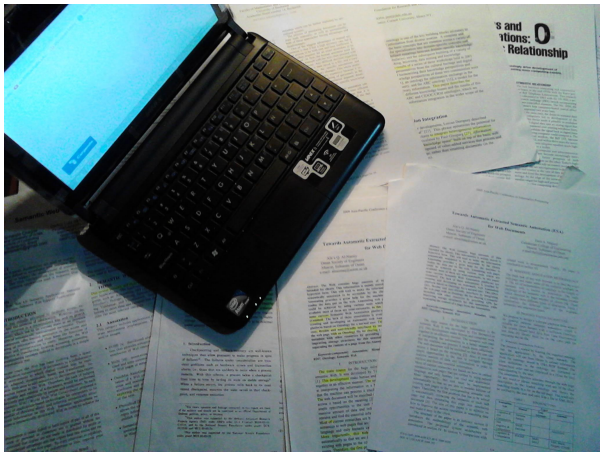
Casos de Uso

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



Estado del Arte

1. Representación del conocimiento mediante modelos semánticos.
2. Búsqueda, recuperación y/o publicación de información empleando tecnologías semánticas.
3. Gestión semántica de una memoria corporativa.



Comparativa

Autor	Dominio	Modelo	Tecnologías Semánticas	Representación del conocimiento	Búsqueda y recuperación de información	Motor de búsqueda e inferencia
Moner et al.	Salud	Orientado a objetos y Arquetipos	No	Sí	No	No
K. Yang y R. Steele	Alojamiento en-línea	Ontología	Sí	Sí	No	No
Jun Zhai et al.	Electricidad	Ontología	Sí	Sí	No	No
Tuan-Dung et al.	Turismo	Ontología	Sí	No	Sí	No
Ha Inay et al.	Mantenimiento de aeronaves	Ontología	Sí	No	Sí	No
Suganyakala y Rajalaxmi	Películas	Ontología	Sí	No	Sí	No
Salam	Urología	Ontología	Sí	No	Sí	No
Xin y Guangleng	Justificación del diseño (Ing. Soft.)	Ontología	Sí	Sí	Sí	No
Chakhmoune et al.	Memoria Documental	Ontología	Sí	Sí	No	No

Problema

Pregunta investigación

¿El *conocimiento implícito* es un factor importante para la obtención de *recursos de información pertinentes* a una consulta?

Hipótesis

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

Objetivo General

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

Objetivos Particulares

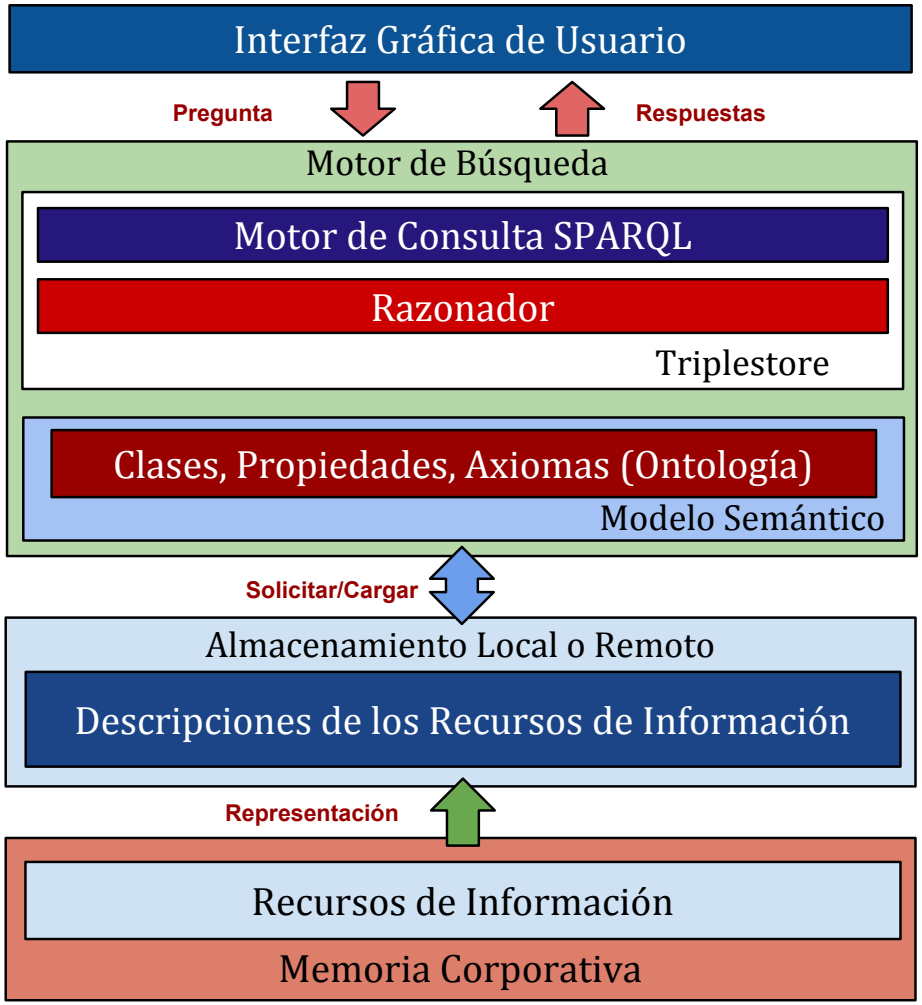
1. Proponer un **marco de referencia** para la *integración semántica* de los *recursos de información* existentes en una memoria corporativa.
2. Proponer un **modelo semántico** que represente el *conocimiento explícito e implícito* existente en los *recursos de información*.
3. Implementar un **prototipo** que permita a los usuarios buscar y recuperar *recursos de información* existentes en una memoria corporativa, así como visualizar las caracterizaciones de estos recursos.

Integración Semántica mediante tecnologías semánticas

Metodología

1. Representar las características y/o relaciones de los *recursos de información*, para construir un modelo semántico.
2. Introducir *reglas de inferencia* en el modelo, para obtener conocimiento implícito.
3. Buscar y recuperar conocimiento en el modelo semántico para responder un conjunto consultas que lleven a *recursos de información*.

Arquitectura



Representación del conocimiento

Resource Description Framework

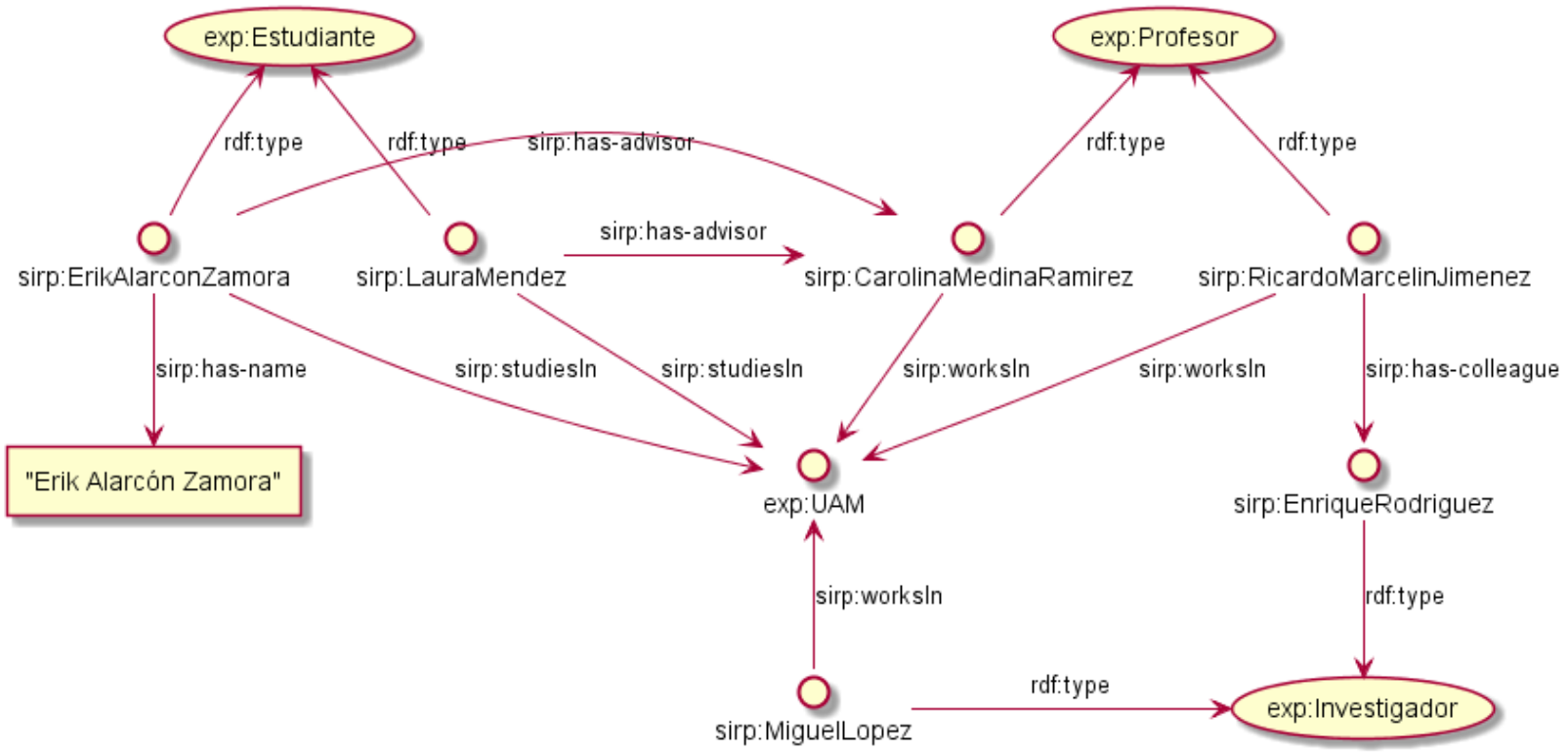
Es un marco genérico para describir el conocimiento e información explícita de los recursos mediante sus características y relaciones.

```
@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.
```

Grafo RDF

```
@prefix sirp: <http://arte.izt.uam.mx/ontologies/personRyT.owl#>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

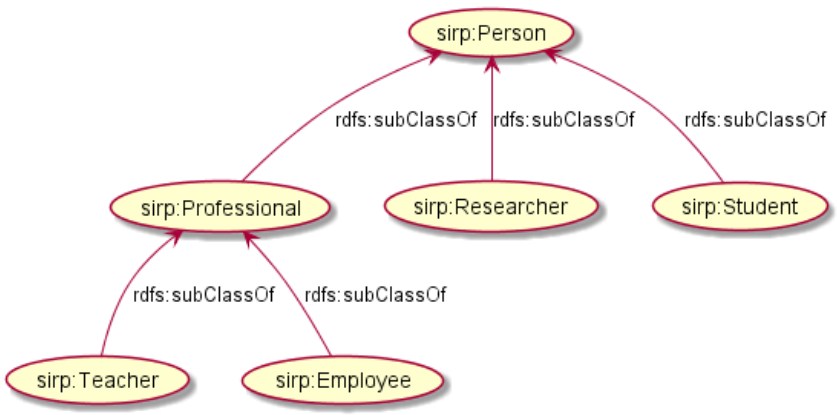


Existen distintas sintaxis de serialización: N3, turtle, RDF/XML, N-triples.

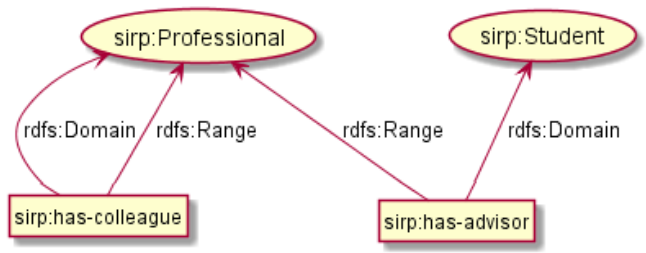
Introducir reglas de inferencia

Reglas de inferencia o Axiomas

Son expresiones para enriquecer un grafo RDF con conocimiento implícito de las clases y relaciones.



(c) Jerarquía de Clases

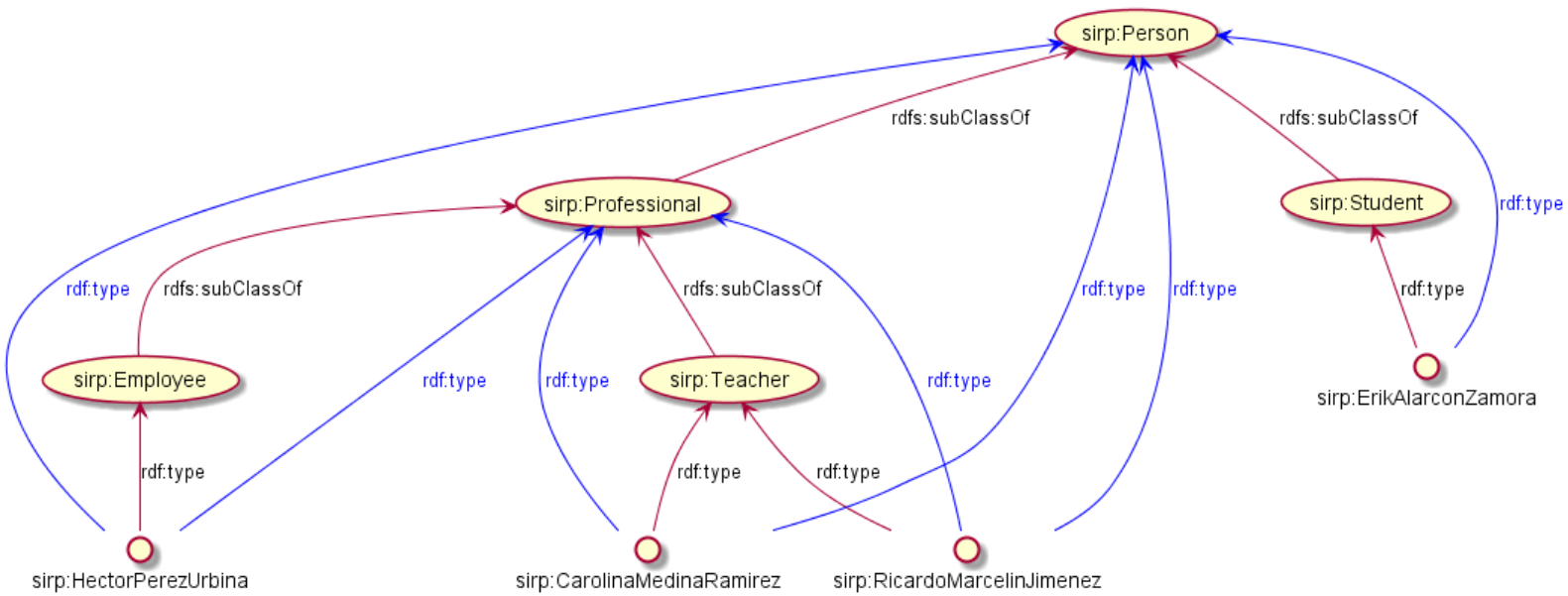


(d) Dominio y rango

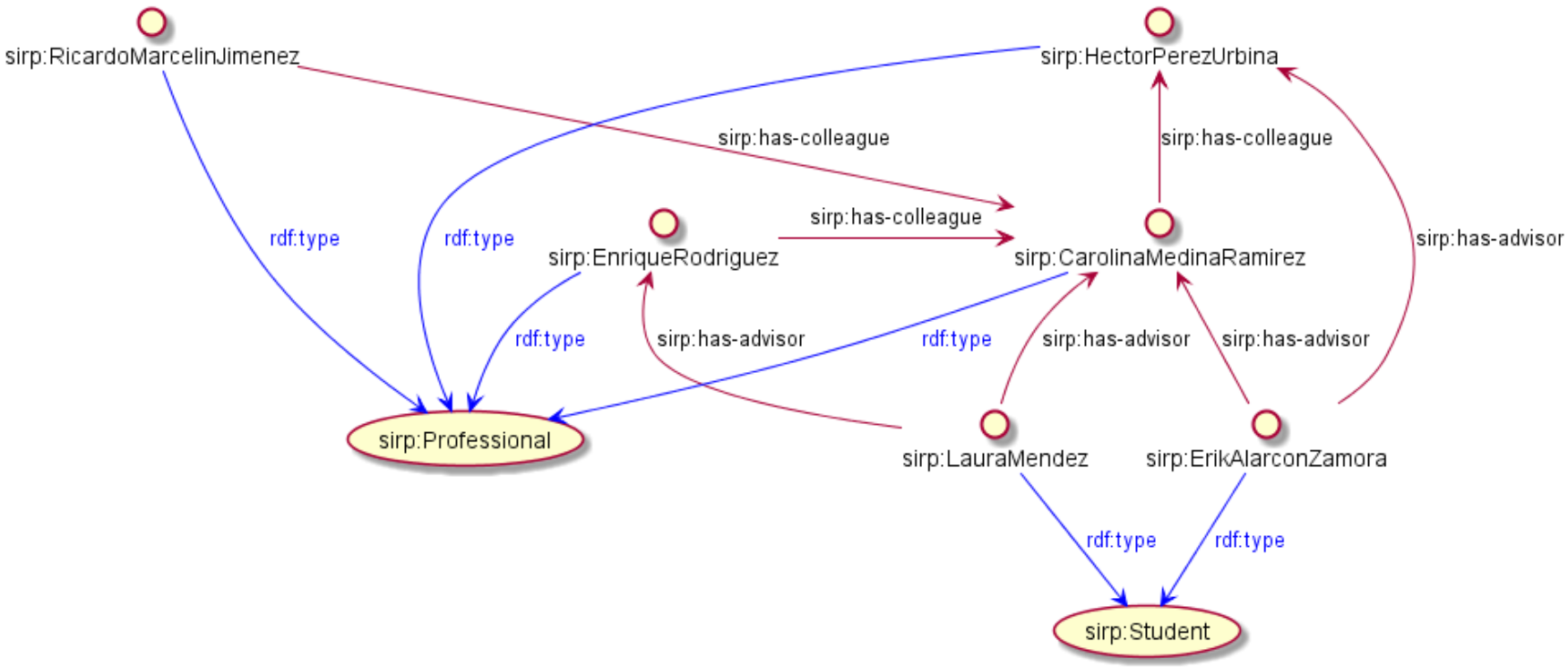
Emplear un programa para inferir conocimiento

Razonador

Es un programa que deduce conocimiento a partir de los axiomas y declaraciones explícitas en un modelos semántico.



Emplear un programa para inferir conocimiento



Buscar y recuperar la información en el modelo semántico

SPARQL

Es un lenguaje de consulta para la recuperación de información en un grafo RDF.

¿Cuáles son los nombres y sitios Web de los recursos de información que son Personas?

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

```
SELECT  ?name ?ws
    WHERE
     {        {?x  rdf:type sirp:Persona.} UNION
              {?x  rdf:type sirp:Profesionista.} UNION
              {?x  rdf:type sirp:Estudiante.} UNION
              {?x  rdf:type sirp:Profesor.} UNION
              {?x  rdf:type sirp:Empleado.}
              ?x   sirp:has-name ?name;
                   sirp:has-webSite ?ws.
     }
```

Consulta sin inferencia

```
SELECT  ?name ?ws
    WHERE
     {
              ?x   rdf:type sirp:Persona;
                   sirp:has-name ?name;
                   sirp:has-webSite ?ws.
     }
```

Consulta con inferencia

Herramientas para la Integración Semántica

Descriptor Semántico de Recursos

Herramienta para crear y almacenar tripletas RDF, en varias sintaxis de serialización, a partir de la información explícita de los recursos de información. ***OntoMat Annotizer***, ***MnM***, ***GATE*** y ***Aktive Media***.

Editor de Ontologías

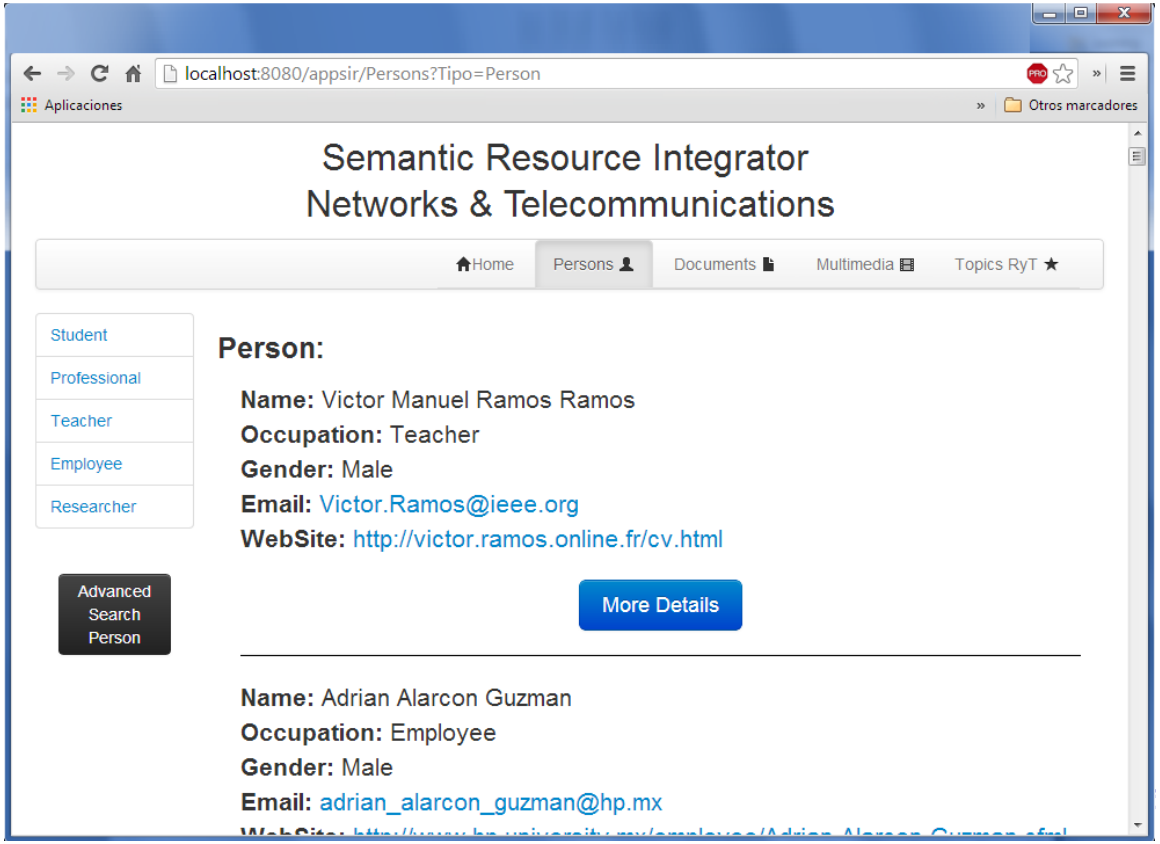
Herramienta que proporciona una serie de interfaces amigables para la construcción y mantenimiento de ontologías. ***Protégé***, ***pOWL***, ***TopBraid Composer*** y ***SWOOP***.

Triplestore

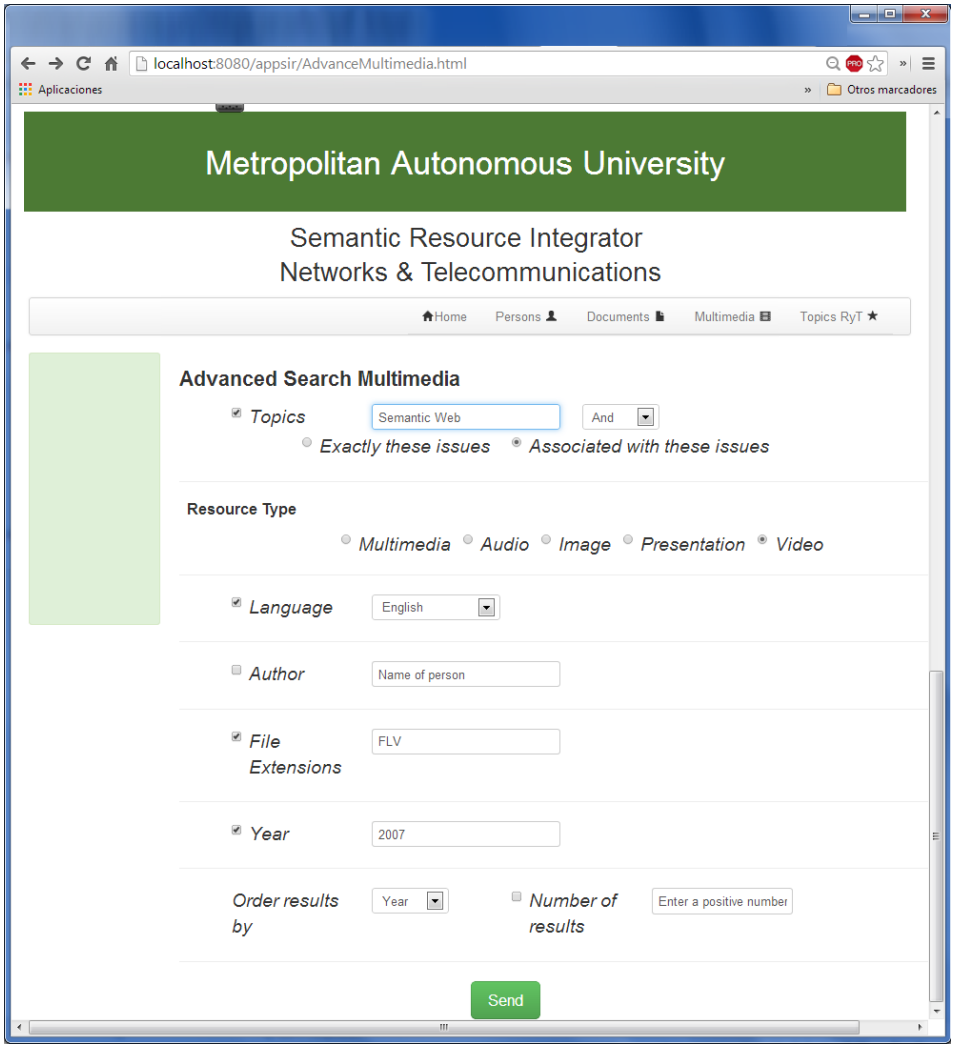
Programa para el almacenamiento e indexación de tripletas RDF, con el fin de permitir la consulta eficiente de información sobre estas tripletas. ***Jena***, ***Stardog***, ***4store*** y ***Sesame***.

Construir un Prototipo (Aplicación)

El prototipo es una aplicación Web que permite a los usuarios estructurar sus preguntas. Éstas a través del uso de un modelo semántico recuperan recuperan los *recursos de información*, así como las características de los mismos.



Construir un Prototipo (Aplicación)



Evaluar la integración semántica

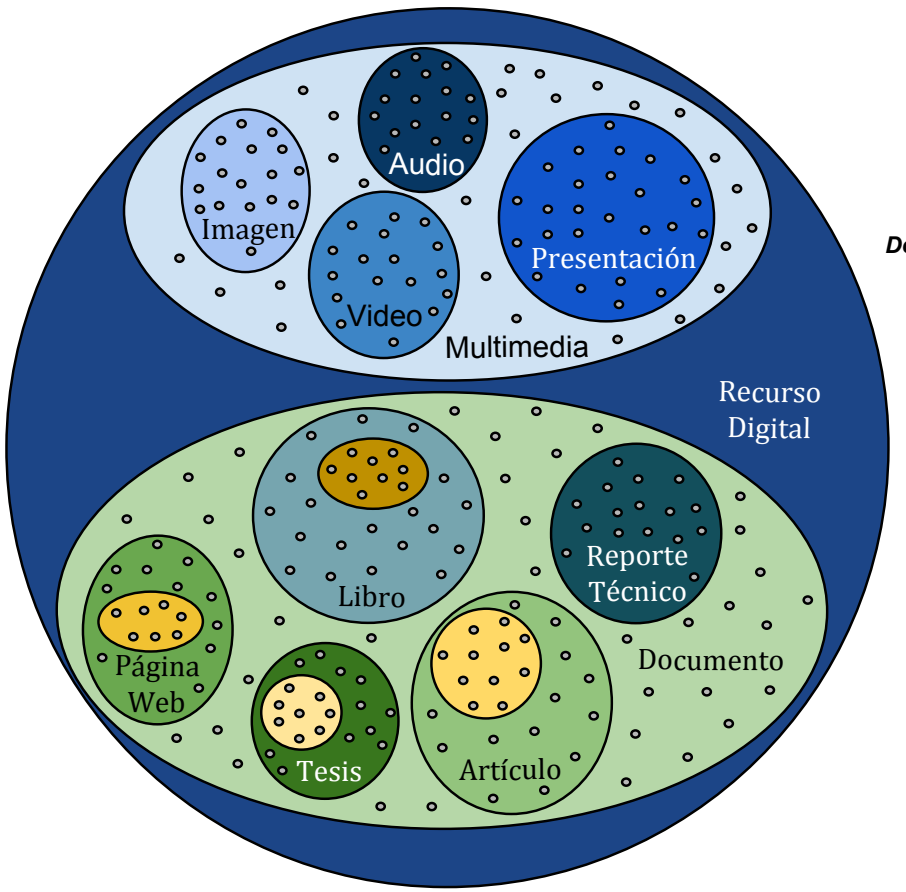
Evaluar la calidad de los resultados

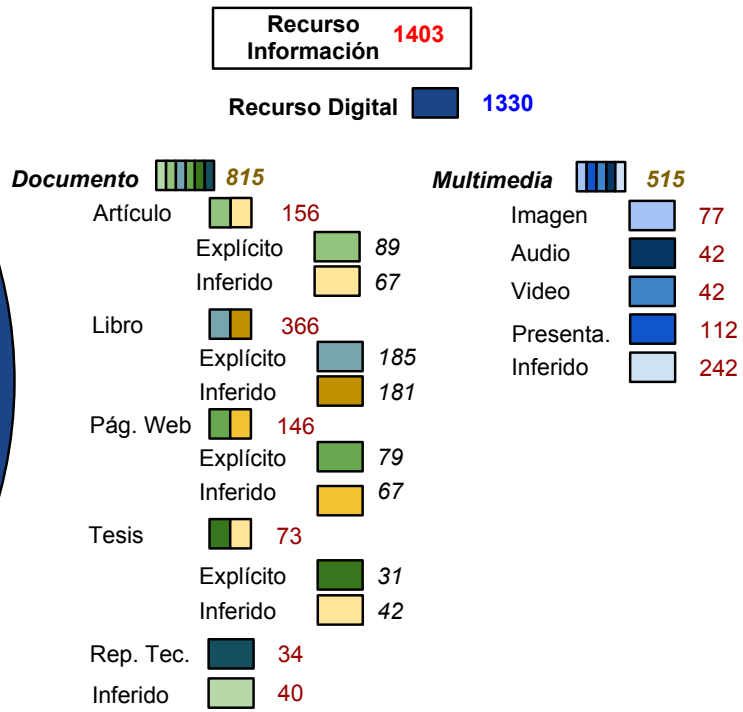
Esta evaluación consiste en comparar los *recursos relevantes recuperados* por Jena (con/sin inferencia) para una consulta dada, con los resultados que de antemano se sabe responden a esta consulta (total de recursos relevantes).

Medir los tiempos promedio de procesamiento de Jena

Esta evaluación consiste en comparar los tiempos de consulta para un modelo con inferencia y otro que no emplea ésta; estos tiempos se toman desde la ejecución de la consulta hasta la presentación de los resultados.

Diagrama de Venn asociado a los recursos digitales





Preguntas en lenguaje natural

Id. Consulta	Pregunta	Total de Recursos Relevantes
Q2.1	¿Cuáles son los títulos, rutas, extensión, idioma de todos los recursos digitales de RyT?	1330
Q2.2	¿Cuáles libros tratan sobre algunos temas de Sistemas Distribuidos?	103
Q2.3	¿Qué recursos fueron publicados por la UAM?	18
Q2.4	¿Qué documentos son para dar un curso de Sistemas P2P?	31
Q2.5	¿Qué recursos multimedia son mayores al año 2009?	119
Q2.6	¿Cuáles documentos tratan sobre Ontologías?	30
Q2.7	¿Qué recursos fueron publicados en una Revista científica?	156
Q2.8	¿Qué recursos tienen en su contenido las palabras "linked data¿	159
Q2.9	¿Cuáles documentos en inglés y mayores al año 2000 son de autoría de Erik Alarcón Zamora?	2
Q2.10	¿Cuáles la tesis de Samuel Hernández Maza?	4

Calidad de los recursos recuperados

Q2.2.- ¿Cuáles libros tratan sobre algunos temas de Sistemas Distribuidos?

Id. Consulta	Recursos relevantes recuperados sin inferencia	Recursos relevantes recuperados con inferencia	Total recursos relevantes
Q2.1	1330	1330	1330
Q2.2	0	103	103
Q2.3	18	18	18
Q2.4	15	31	31
Q2.5	66	119	119
Q2.6	15	30	30
Q2.7	156	156	156
Q2.8	159	159	159
Q2.9	0	2	2
Q2.10	3	4	4

Exhaustividad y Precisión

Q2.2.- ¿Cuáles libros tratan sobre algunos temas de Sistemas Distribuidos?

Id. Consulta	Sin inferencia		Con inferencia	
	Exhaustividad	Precisión	Exhaustividad	Precisión
Q2.1	1	1	1	1
Q2.2	0	-	1	1
Q2.3	1	1	1	1
Q2.4	0.484	1	1	1
Q2.5	0.555	1	1	1
Q2.6	0.5	1	1	1
Q2.7	1	1	1	1
Q2.8	1	1	1	1
Q2.9	0	-	1	1
Q2.10	0.75	1	1	1

Tiempos de Procesamiento

Q2.2.- ¿Cuáles libros tratan sobre algunos temas de Sistemas Distribuidos?

Id. Consulta	Tiempo promedio (milisegundos)	
	Modelo sin inferencia	Modelo con inferencia
Q2.1	24	3520
Q2.2	9	4016
Q2.3	12	3520
Q2.4	16	3472
Q2.5	42	3451
Q2.6	14	3392
Q2.7	13	3431
Q2.8	32	3312
Q2.9	34	3570
Q2.10	11	3398

Aportaciones

Un **marco de referencia** para lograr la integración semántica de recursos de información en una memoria corporativa.

Un **modelo semántico** que representa el conocimiento de una memoria corporativa, el cual es flexible, extensible y reutilizable.

Un **prototipo** para la búsqueda y recuperación de recursos e información.

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.

Conclusiones

- El uso de las tecnologías semánticas contribuye a la integración semántica de los recursos de información en una memoria corporativa.
- El conocimiento implícito es determinante para la obtención de *recursos de información pertinentes*.
- La inferencia no es gratis, tiene costo en tiempo.
- Un marco de referencia, modelo semántico, prototipo y scripts son las contribuciones para la integración semántica de recursos de información en una memoria corporativa.

Recomendaciones

- Introducir nuevos *casos de uso* para modelar mayor conocimiento.
- Mejorar la seguridad del prototipo y agregar un recuadro para búsquedas por *palabras clave*.
- Construir un módulo (aplicación) para generar *tripletas RDF* a partir de las descripciones de los *recursos de información*.
  - Generación guiada por los usuarios.
  - Generación automatizada.
- Comparar los tiempos de procesamiento y calidad de los recursos con otros triplestores: Stardog y Sesame.

Integración semántica de los recursos de información en una memoria corporativa

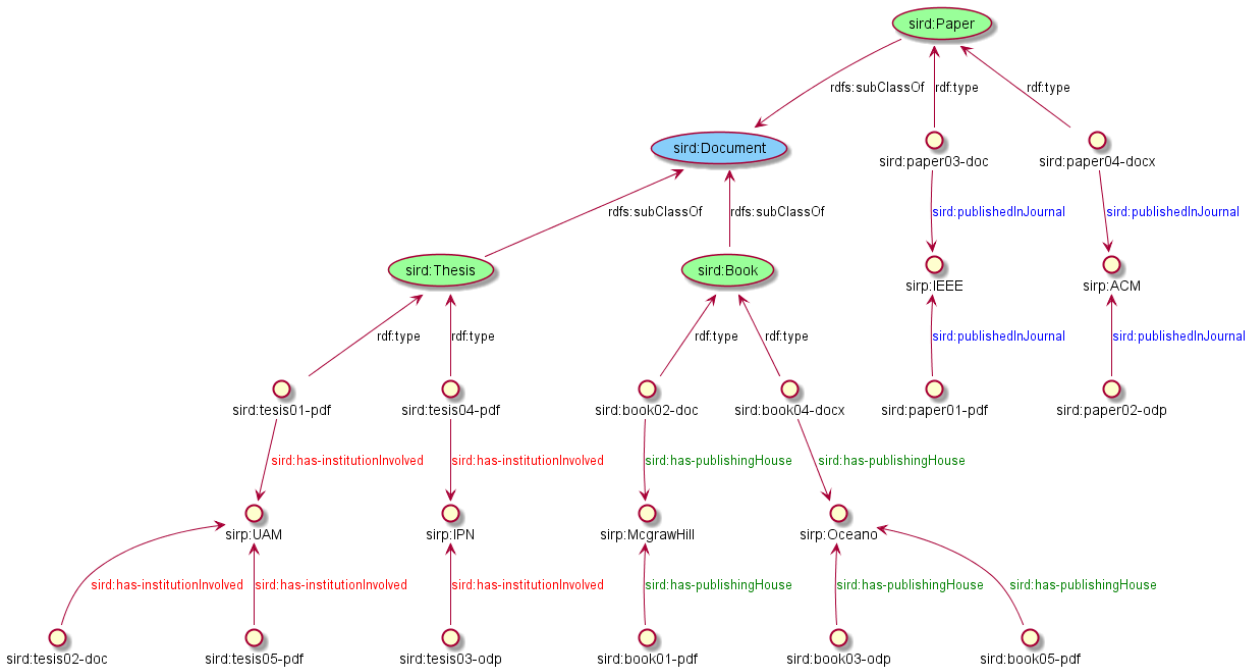
Erik Alarcón Zamora

Enero 2014. México, D.F.

Asesores:
Dra. Reyna Carolina Medina Ramírez
Dr. Héctor Pérez Urbina

Búsqueda sin inferencia

Grafo RDF sin inferencia



Búsqueda sin inferencia

```
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.}
    }
```

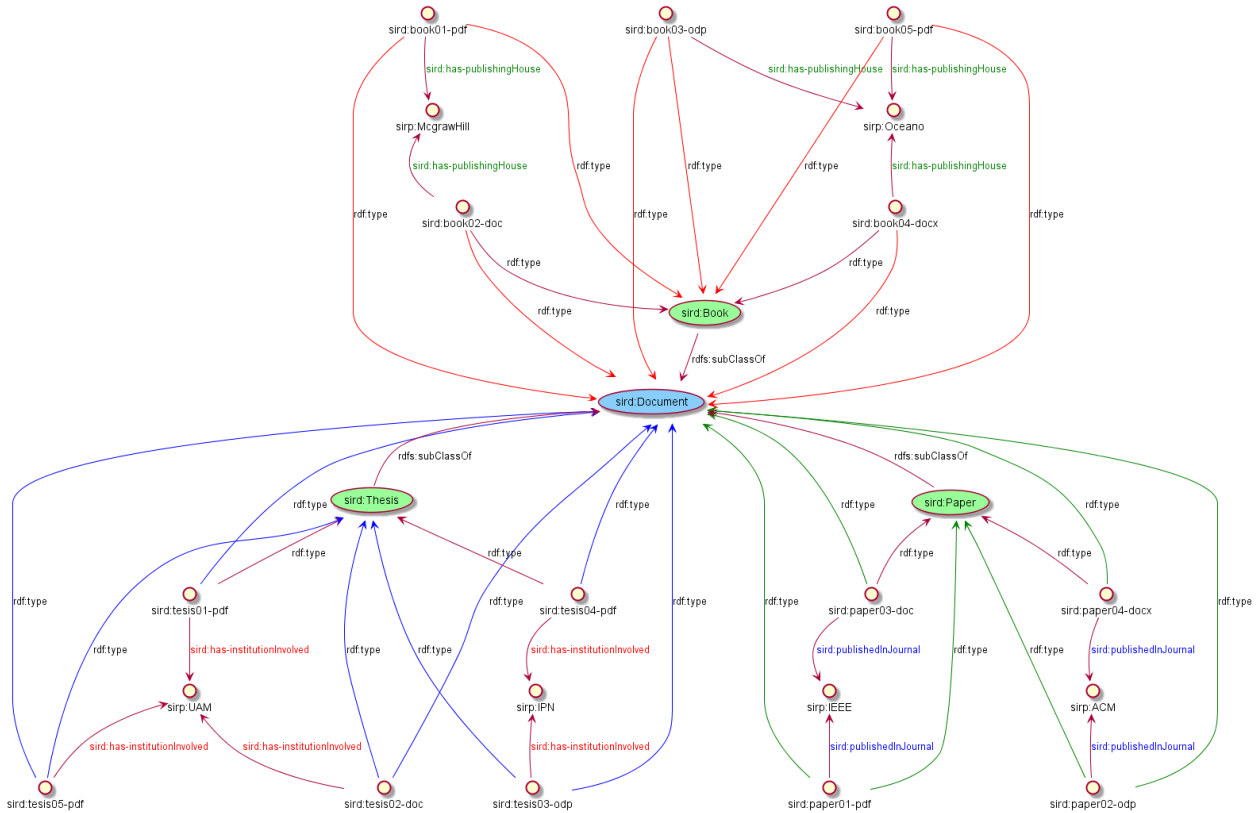
(e) Consulta sin inferencia

?x
sird:tesis01-pdf
sird:tesis04-pdf
sird:book02-doc
sird:book04-docx
sird:paper01-pdf
sird:paper02-odp

(f) Resultados de la consulta

Búsqueda con inferencia

Grafo RDF con inferencia



Búsqueda con inferencia

```
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. }
```

(g) Consulta con inferencia

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

(h) Resultados de la consulta