TRABAJO PRACTICO FINAL- PRIMERA PARTE	SISTEMAS DE RECUPERACION DE INFORMACION EN LA WEB

Elaborado por: Jaime Guzmán L.

Revisado y corregido por: Jaime Guzmán L.

Fecha de entrega y sustentación práctica: abril 18 a las 8 am en salón de clase

---

OBJETIVO

Aplicar los conocimientos adquiridos sobre las tecnologías Linked Data, el lenguaje de ontología OWL, la realización de inferencias sobre la base de conocimiento desarrollada en OWL y los sistemas de recomendación, aplicadas en la solución de un problema relacionado con la recuperación de información en la Web.

Para ello se proponen dos problemas a solucionar: el primero de carácter investigativo y el segundo de carácter práctico.

PROBLEMA 1- Investigación de datos abiertos enlazados portal Colombia (Valor 1.0)

Dado los dos documentos entregados sobre los datos enlazados en Colombia (articles-9407_Guia_Apertura y Resumen_Ejecutivo_Datos_Abiertos) se pretende realizar una breve investigación sobre las políticas, avances y problemas relacionados con los datos abiertos en Colombia. Para ello

- Primero buscar 5 artículos que hablen sobre este tema ya sea en revista de tipo A en el índice de Colciencias o artículos de las bases de datos Scopus o JCR. Para esto utilizar el servicio de Bases de datos que se tienen en línea por la biblioteca de la universidad.
- Segundo, investigar sobre el estado la situación actual de los datos enlazados y como se pueden aplicar para mejorar y apoyar el tema de los datos abiertos en Colombia. Para ello obtener y analizar 5 artículos que traten sobre una revisión de su estado actual y su futuro para posteriormente proponer soluciones en el contexto Colombiano.
- Por último, generar un artículo corto de 3 páginas con la investigación realizada en los 2 numerales anteriores (El documento debe incluir: título, resumen, introducción, desarrollo del tema, conclusiones y bibliografía).

Nota: se entrega el documento How to do a systematic literature review and meta-analysis.pdf como guía de cómo hacer la revisión de la literatura.

Material a entregar:

1. Archivos de los 5 artículos consultados para la consulta. Adicionalmente incluir una hoja donde se indique de que base de datos se obtuvo cada documento.
2. Archivo con al artículo de 3 páginas (Usar formato de la revista dyna).

PROBLEMA 2- Implementación de sistema de datos abiertos enlazados asociados al portal Colombia (Valor 4.0)

Se busca que el estudiante proponga un problema a solucionar en el cual se pueda aplicar las técnicas vistas en clase. El sistema solución deberá cumplir básicamente los requerimientos que se describen a continuación.

DESCRIPCIÓN DETALLADA

El trabajo consiste de cuatro partes principales. En la primera parte se deberán plantear el problema de recomendación a resolver, en la segunda parte se plantea la implementación de la base de conocimiento que

soportará el sistema de recomendación a resolver, la tercera parte se plantea el desarrollo del sistema de recomendación con su respectiva interfaz, en la cuarta parte se plantea la implementación de algunas pruebas de validación del sistema a desarrollar.

Parte 1. Especificación del problema a solucionar (Valor 20%).
Descripción clara y específica de un problema que busque recomendar sobre algo y haga uso de las tecnologías de datos enlazados. Este problema deberá tener en cuenta que la información asociada al problema está distribuida. Para ello primero se deberá seleccionar diferentes dataset del portal de datos abiertos colombiano (http://www.datos.gov.co/frm/buscador/frmBuscador.aspx) con el fin de que usted plantee su caso de estudio con información proveniente de estos portales la cual será implementada en las bases de conocimiento que desarrollarán. Luego, haciendo uso de las técnicas de la Web semántica y de los sistemas de recomendación vistos en clase implemente dicho sistema de recomendación el cual deberá ser alimentado por 3 sitios que publican *Linked Data* implementados por usted para este trabajo.

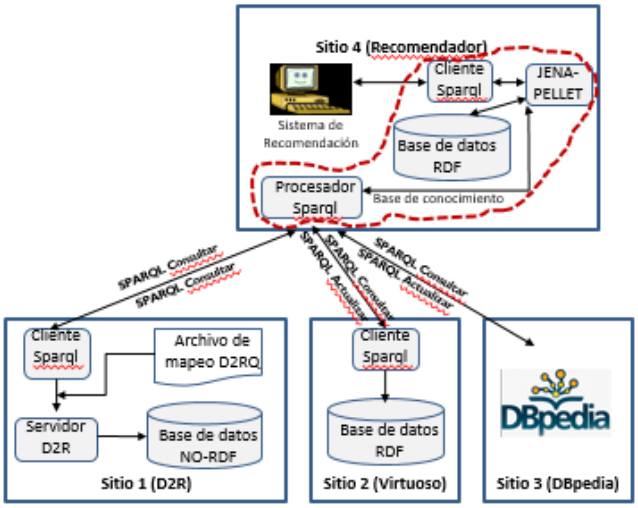
Producto a entregar en esta actividad: Documento con la definición del problema junto con una revisión del marco teórico y estado del arte que soporte el problema. En este punto se deberá justificar el porque es útil el uso de las técnicas de la Web semántica y los dataset que contienen la información de los portales de datos elegidos para la solución de su problema.
NOTA: Este documento deberá cumplir con lo solicitado en el anexo 1.

Parte 2. Diseño detallado de la arquitectura del sistema (valor 30%)
En este punto se detallará los componentes del sistema y sus interacciones, el modelo de datos utilizado y el diagrama de proceso del sistema.

Producto a entregar en esta actividad:
1. Documento con la Arquitectura del sistema con los elementos solicitados. El informe deberá ser lo suficientemente detallado como para entender claramente el alcance del sistema que pretende solucionar el problema.
2. Detallar en el informe Mediante una sección particular “Aportes al Diseño del Proyecto” que elementos del diseño aportó cada miembro del grupo. Se indicará cual fue su papel en esta etapa según las actividades propuestas en el cronograma de la propuesta de investigación.

Parte 3. Implementación de la base de conocimiento que soportará el sistema de recomendación (valor 40%).
La información (relacionada con la temática escogida) que soporta la base de conocimiento del problema estará distribuida en tres sitios de publicación de datos enlazados en la Web. Cada sitio hace uso de una plataforma diferente de publicación (uno utilizando D2R server, otro utilizando Virtuoso y el tercero la base RDF de DBpedia) y de un vocabulario independiente y diferente (que hace uso de los vocabularios básicos vistos en clase) lo que nos da como resultado tres vocabularios distintos que aunque tratan de la temática a resolver están hechas con expresiones diferentes y hacen uso de los vocabularios básicos FOAF, Skos y Dublin core. Así mismo, para las dos primeras plataformas de publicación estos vocabularios implementarán la información obtenida desde los dataset del portal datos abiertos colombiano. Es importante aclarar que estos 3 sitios deben tener acceso desde la Web

Para integrar esta información se desarrollar un nuevo repositorio que tiene la siguiente arquitectura básica mostrada en la figura 1. Este cuarto sitio de publicación de Linked Data deberá recoger e integrar **en tiempo real** la información procedente de las tres bases haciendo uso de una ontología hecha en OWL que permita razonar e integrar en un solo repositorio temporal los tres vocabularios de dichas fuentes de información para conformar la información asociada a los objetos a recomendar. Este sitio de publicación se deberá implementar haciendo uso de JENA y PELLET para llevar a cabo una serie de consultas sobre la base de conocimiento desarrollada haciendo uso del Sparql y el empleo de técnicas de inferencia para realizar consultas complejas que utilizará el sistema de recomendación de información. A continuación se describe en detalle cada uno de los aspectos de este repositorio.



Sobre la ontología OWL (que integra los tres vocabularios de las tres Bases de conocimiento) es importante aclarar que esta debe cumplir con las siguientes características:

- La ontología deberá incluir los siguientes elementos del OWL:
  - owl: inverseOf
  - owl:SymmetricProperty
  - owl:TransitiveProperty
  - owl:equivalentClass
  - owl:sameAs
  - owl:FunctionalProperty
  - owl:InverseFunctionalProperty
- Crear al interior de dicha ontología dos clase diferentes para cada uno de los siguientes elementos:
  - Restriction y:owl:someValuesFrom
  - Restriction y:owl:allValuesFrom
  - Restriction y:owl:maxCardinality
- Emplear en su ontología adicionalmente en dos casos los siguientes elementos:
  - owl:complementOf
  - owl:disjointWith

Adicionalmente se deberán crear instancias inferidas de elementos y relaciones para lo cual deberán configurar su archivo OWL finalmente para esto. Las inferencias se realizarán haciendo uso de JENA y PROTEGE.

Productos a entregar en esta actividad: La ontología OWL abstracta que integra los tres vocabularios. Los dos vocabularios abstractos generados para las dos bases de conocimiento a implementar en el proyecto. Códigos SQL que permita la creación y el llenado de la base de datos del sistema D2R al igual que los archivos RDF (y/o OWL) y D2RQL (vocabulario y contenido) que permite su manejo como SPARQL. Código configurado que permita llenar la base de Virtuoso (vocabulario y contenido). El archivo RDF del sitio 3 que contenga la ontología y sus instancias. Adicionalmente se deberá presentar en el momento de la sustentación el portal web de las 2 bases de conocimiento QUE SE IMPLEMENTAN DESDE CERO trabajando correctamente en la Web. Por último se deberá entregar el código fuente relacionado con el funcionamiento de la base de conocimiento asociada al sistema de recomendación Web (que incluye JENA y PROTEGE) que permite su llenado y actualización. Adicionalmente, se debe al momento de la sustentación mostrar en funcionamiento la base de conocimiento de este sistema de recomendación Web. Se debe entregar también la documentación de esta base de conocimiento donde se describa de manera detallada sus algoritmos de

funcionamiento y actualización al igual que sus componentes (utilizar ilustraciones y una definición clara de sus algoritmos). Por último se debe entregar un documento con el código de cada una de las consultas SPARQL asociadas al sistema JENA y PELLET que se requieren al igual que su explicación respectiva para que sirve dentro del sistema recomendación. Estas consultas incluyen las realizadas a las bases de conocimiento distribuidas al igual que las realizadas a la base de conocimiento para llevar a cabo tanto la tarea de actualización de la base de conocimiento del sistema de recomendación como de la información consumida por el mecanismo de recomendación.

En la implementación deberán participar cada uno de los miembros del equipo. En este punto se deberá hacer uso de Git y GitHub para compartir los proyectos con el profesor y el monitor del curso. Ver Anexo2-Taller1_GitHub.pdf.

Parte 4. Sustentación de la propuesta y su estado intermedio de solución (valor 10%).

El día de la sustentación se deberá preparar un breve informe oral donde se exponen los anteriores temas de manera resumida teniendo cada equipo un tiempo de sustentación de 8 minutos distribuidos así:

- Exposición de la propuesta: 2 minutos
- Exposición del diseño: 2 minutos
- Exposición de la implementación (base de conocimiento): 2 minutos
- Preguntas: 2 minutos.

Para dicha sustentación deberán hacer uso de una presentación en video beam y las ideas deberán ser lo suficientemente claras y concretas.
El tiempo para estar listos entre exposición y exposición es de 2 minutos por lo cual deberán tener dispuestos sus equipos para conectarse al video beam.

Productos a entregar en esta actividad:

- Diapositivas en PDF enviadas el día anterior a la exposición.

Nota 1: En este punto se valorará la claridad de las ideas y el buen manejo del tiempo de la exposición. Todos los miembros del equipo deberán intervenir en dicha exposición.
Nota 2: Durante la sustentación de este informe se deberá mostrar en GitHub el código implementado mostrando lo que ha hecho cada uno de los integrantes.

ENTREGABLES EN GENERAL

(1) Un CD con los productos descritos al final de cada parte del trabajo.

BIBLIOGRAFIA

- El marco de trabajo JENA: http://jena.sourceforge.net/index.html
- Razonador Pellet: http://clarkparsia.com/pellet
- OWL: http://www.w3.org/TR/owl-ref/
- Capítulos 5 y 8 del libro Semantic Web Programming. John Heber et al. Ed. WILEYPublishing, inc. 2009.
- Capítulos 9 y 10 del libro Semantic Web for Working Ontologist. Dean Allemang y Jim Hendler. Ed. Morgan Kaufman. 2008.

ANEXO1
ESPECIFICACION DEL PROBLEMA Y SOLUCION A PROPONER

Definir un problema y plantear su solución. Para ello implementar un ensayo que contenga los siguientes elementos:

1. PLANTEAMIENTO DEL PROBLEMA

Definir un problema es caracterizarlo, definirlo, enmarcarlo teóricamente. Generalmente un problema se formula a través de un interrogante, pero también existe la opción de presentarlo de manera descriptiva. Lo importante es que a través del trabajo investigativo, principalmente en los resultados, se dé respuesta a la(s) pregunta(s) problema.

2. OBJETIVOS

2.1. OBJETIVO GENERAL
2.2. OBJETIVOS ESPECIFICOS

"Defina, el propósito general del proyecto en términos de su contribución o coherencia con el problema planteado o su contribución a la competitividad de la empresa, sector o cadena productiva.
Formular un solo objetivo general y defina los objetivos específicos necesarios para alcanzar el objetivo general en función de la(s) alternativa(s) tecnológica(s) identificada(s) para resolver el problema planteado.
Recuerde que no debe confundir objetivos con actividades o procedimientos metodológicos."

3. MARCO TEÓRICO

Destaca la estrecha relación que existe entre teoría, práctica, proceso de investigación, realidad, entorno, y revela las teorías y evidencias empíricas relacionadas con la investigación (ESTADO DEL ARTE DE LA INVESTIGACIÓN)

4. ANTECEDENTES

Revise la bibliografía más reciente para evidenciar si el problema ya tiene alguna respuesta o parcial. Analice todos los trabajos anteriores que estén relacionados con el problema que desea resolver. En este punto se deberá realizar una exploración de las bases digitales de la universidad: IEEE y Scopus (ver figuras 1). Realizar una tabla comparativa de los trabajos existentes donde se verifique la existencia del problema propuesto en su trabajo.

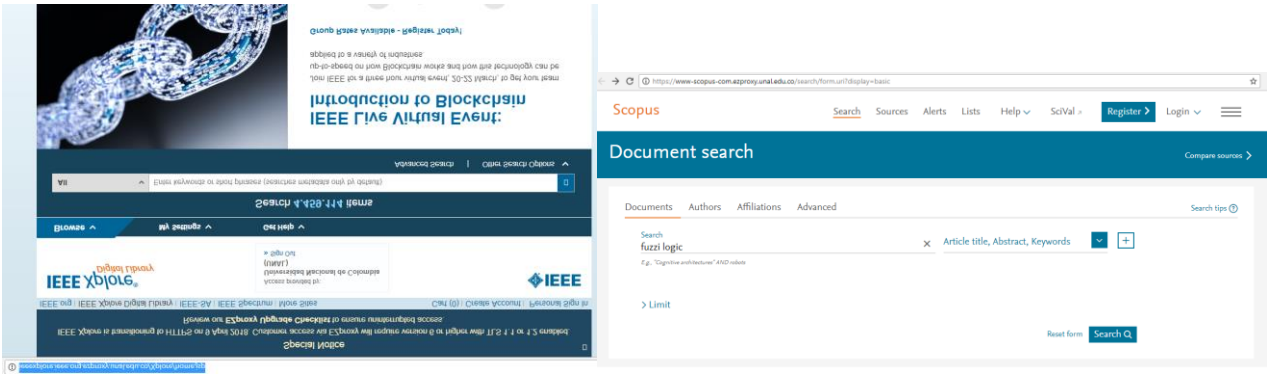


Figura 1. Bases de datos IEEE y Scopus

Para su revisión seguir un protocolo como el detallado en el anexo1- WHAT IS A SYSTEMATIC LITERATURE REVIEW AND HOW DO I DO ONE?

Con lo anterior identificarán los antecedentes que soportan el problema a resolver. En este punto se deberán anexar en formato digital los 4 artículos (2 de IEE y 2 de Scopus) que consideren más relevantes en formato PDF justificando cada él porque fueron seleccionados.

5. PLAN DE TEMAS Y CRONOGRAMA

Es un plan de trabajo o un plan de actividades, que muestra en un orden lógico y secuencial la duración del proceso investigativo, en una forma gráfica o de tabla. En este aspecto se para cada actividad se incluirá la descripción de la misma, su justificación frente a los objetivos y el personal que atenderá la actividad.

6. BIBLIOGRAFÍA Y FUENTES DE INFORMACIÓN

Describa las obras y demás materiales de carácter informativo de consulta para la elaboración de la propuesta de trabajo.

PRODUCTOS A ENTREGAR:

1. Documento con la propuesta de trabajo que contiene los numerales anteriormente descritos.(5 páginas a espacio sencillo)
2. Los 4 artículos más relevantes que soportan su investigación

Nota: En la propuesta de investigación se debe distribuir el trabajo a realizarse entre los miembros del grupo con el fin de en cada etapa asociada a los objetivos de la propuesta solución cada uno de los miembros del equipo aporte a la misma. Esto se debe ver reflejado en la distribución de las actividades y las personas que las van hacer.

ANEXO 2
TALLER DE GIT Y GITHUB

Se anexan en formato digital dos documentos asociados a este tema.

ANEXO 3
LISTA DE GRUPOS

Se tomarán los mismos grupos de las exposiciones definidos al inicio del curso con las modificaciones informadas al profesor.