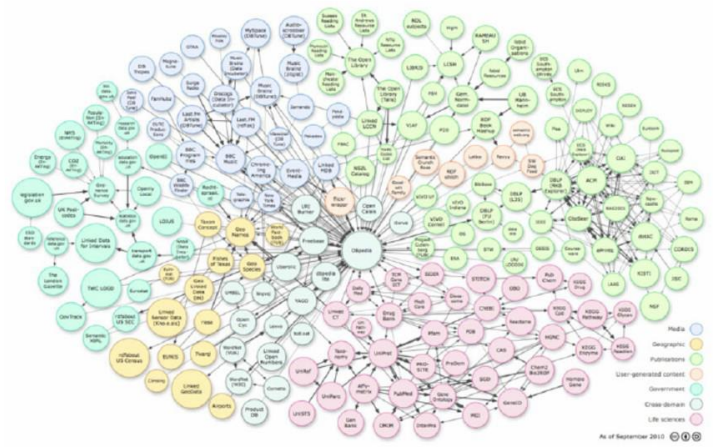
FUNDAMENTOS DE LA WEB SEMANTICA – 2023

UNIDAD 4: Modelos semánticos – Grafos de Conocimiento – RDF – Linked Data



https://eva.fing.edu.uy/course/view.php?id=536&section=5

Regina Motz

InCo- Fing - Universidad de la República

Objetivos de la Unidad

Entender las limitaciones de XML como modelo de datos semántico. Comprender lo que es un grafo de conocimiento y poder modelar una realidad utilizando RDFS.

Relación de esta Unidad con lo que estudiamos en la Unidad anterior:

En la unidad anterior repasamos los Modelos de Datos.
Estudiamos XML Schema como modelo para validar tipos de documentos. Comparamos el Modelo Relacional y el Modelo Orientado a Objetos según el nivel de abstracción que ofrecen. Diferenciamos el nivel de abstracción que tienen los modelos con su capacidad de expresión. Recordemos que la capacidad de expresión esta en relación con cuanta semántica del mundo real es capaz de representar el modelo. En esta unidad estudiaremos el modelo de Grafos de Conocimientos como uno de los modelos de mayor capacidad de expresión que XML.
Nos interesa entender lo que es la Web Semántica, también llamada Web de Datos o Web 3.0. En la unidad anterior terminamos con un dibujo en capas de los elementos necesarios para la catalogación de libros en las bibliotecas en correspondencia con los elementos que hacen a la Web Semántica. En esta unidad profundizaremos en lo que significa la capa RDF de ese stack de tecnologías de la Web Semántica, identificando a RDF como un tipo de grafo de conocimientos.

Contenido de este documento:

¿De dónde viene el nombre Modelos Semánticos?

En la conferencia SIGMOD de 1978 aparece el siguiente trabajo:

The semantic data model: a modelling mechanism for data base applications

Autores: Michael Hammer, MIT Laboratory for Computer Science, Cambridge, MA
Dennis McLeod, MIT Laboratory for Computer Science, Cambridge, MA

Publicado en:

Proceeding SIGMOD '78 Proceedings of the 1978 ACM SIGMOD International Conference on Management of Data. Pages 26-36

Su resumen dice:

Modelos de datos convencionales no son satisfactorios para el modelado de sistemas de aplicación de base de datos. Las características que se ofrecen son de nivel muy bajo para la representación de la semántica de una base de datos que se expresa directamente en el esquema. El modelo de datos semántico (SDM) ha sido diseñado como un mecanismo de modelado de aplicaciones natural que puede captar y expresar la estructura de un entorno de aplicación. Las características del SDM corresponden a las principales estructuras intencionales que ocurren naturalmente en las aplicaciones actuales de base de datos. Además, las facilidades para expresar información derivada (redundante) son una parte esencial del SDM, la información derivada es tan prominente en un esquema SDM como son los datos primitivos. El SDM está diseñado para mejorar la eficacia y la facilidad de uso de las bases de datos informatizadas. Puede servir como una especificación formal y mecanismo de documentación de la base de datos, puede soportar una variedad de poderosas instalaciones de la interfaz de usuario, y puede ser utilizado como una herramienta en el proceso de diseño de la base de datos.

1. ¿Qué limitaciones tiene XML como Modelo Semántico?

En la unidad anterior vimos lo que significa que un documento XML es válido con respecto a un documento XML esquema. Lo que el XML Schema controla es la ESTRUCTURA del documento XML. En la actividad grupal se pedía crear instancias válidas para el documento XML Schema que representa la estructura acordada entre el emisor y el receptor de la información.

La pregunta que nos inquieta ahora es: ¿cuál es la SEMÁNTICA que podemos interpretar de un documento XML?

Antes de continuar realiza la ACTIVIDAD INDIVIDUAL 1 del EVA: ***¿Qué nos dice un documento XML?***

Si realizaste la ACTIVIDAD INDIVUDUAL 1 en las conclusiones encuentras las dos grandes limitaciones de XML.

Por supuesto, XML es a pesar de estas limitaciones uno de los pilares de la Web Semantica. XML tiene muchas características que lo hacen fundamental para su uso en la Web. Entre ellas podemos destacar que facilita el intercambio de información entre fuentes tecnológicamente heterogéneas y es fácil de entender tanto por humanos como por más maquinas. Su aplicación fundamental es la representación estructural de los datos y la fácil distribución de los mismos. Otras cualidades destacables de XML es que existen complementos como XQuery, Xpath, XSLT y XML Schema (XSD) para consultar, navegar, transformar y validar los datos representados en un documento XML.

Un ejemplo para mostrar este potencial podría ser el de dos instituciones médicas A y B con sistemas tecnológicamente diferentes que intercambian información de sus afiliados cuyos datos son almacenados en bases de datos heterogéneas. En este caso, la institución A puede solicitar los datos de los nuevos afiliados a partir de una fecha determinada a un servicio expuesto por la institución B el cual recupera de su base local los registros de sus afiliados, los serializa en un documento XML y los devuelve a la institución A. Luego la institución A puede validar la consistencia del documento usando un archivo de definición de esquemas XML Schema, y en caso de ser correcto procesarlo usando XQuery y XPath para luego persistirlos en su base local. Adicionalmente la institución B puede utilizar XSLT para transformar el documento XML en una página web o mostrar a un administrador.

Por todo esto XML juega un papel fundamental en la web de datos abiertos y la integración de sistemas.

Sin embargo, XML presenta carencias cuando lo que queremos modelar son relaciones conceptuales y no solo la estructura de documentos. Entonces, podemos decir que XML es necesario pero no suficiente! Y es allí donde aparecen construcciones más ricas semánticamente como los Grafos de Conocimiento (Knowledge Graphs), y en especial RDF.

"In short, XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean"

De: XML and the Second-Generation Web. Jon Bosak, Tim Bray. Scientific American, Mayo 1999.

2. Grafos de Conocimiento (Knowledge Graphs)

Como ya analizamos una de las ventajas de XML es la de validar la estructura de documentos donde la estructura lógica de documentos XML es la de ARBOL, mientras que la estructura de la web es de GRAFO. Por consiguiente, para modelar la web deberíamos utilizar un modelo que represente GRAFOS.

2.1 El Grafo de Conocimientos de Google

El termino ***Grafo de Conocimiento*** se hizo popular al ser presentado por Google en el año 2012. Para resolver una consulta, Google no solo presenta el resultado que más se acerca a un término de búsqueda, sino que también establece conexiones más amplias entre los datos. Google, por lo tanto, recopila y analiza cantidades masivas de datos sobre personas, lugares, cosas y hechos y desarrolla formas de presentar los hallazgos de manera accesible en el llamado *Panel de Conocimiento*. Usando "El Grafo de Conocimiento", es que Google consigue ampliar el resultado de una consulta y presentarlo en el panel, o tablero, de conocimiento que se muestra arriba a la derecha de los resultados. Vean en la Figura 1 el tablero que presenta Google al buscar: "Facultad de Ingenieria, UdelaR".

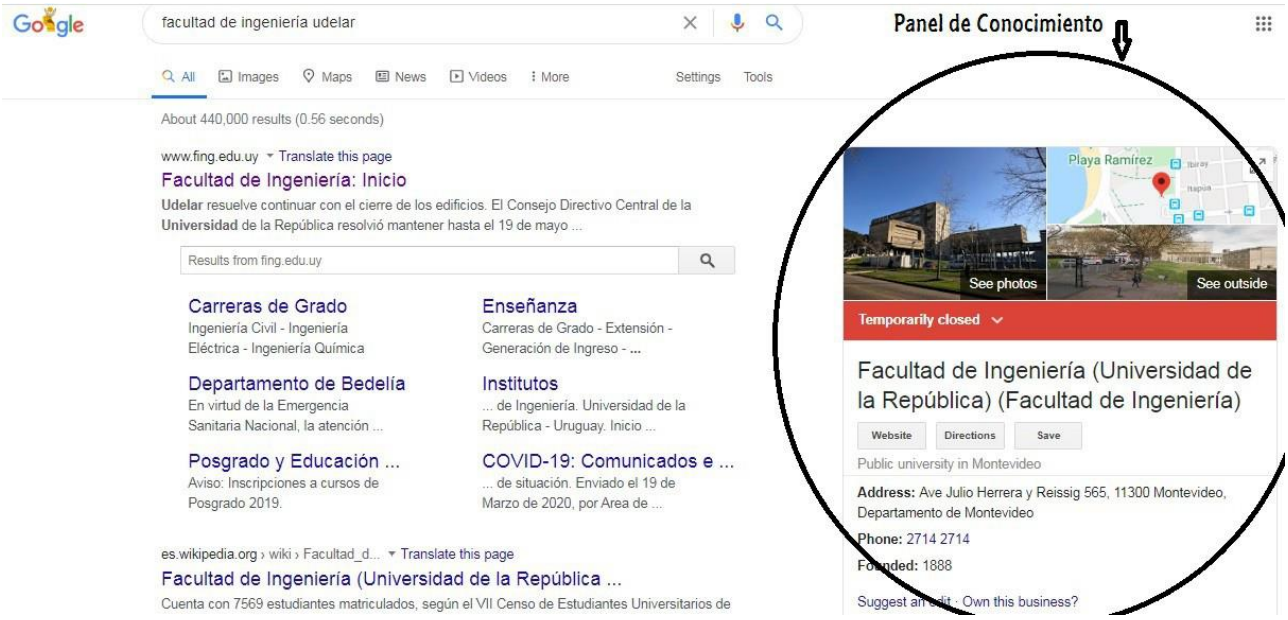


Figura 1. Ejemplo de Panel de Conocimiento de Google.

El panel de conocimiento de Google es el resultado de agrupar hechos, personas y lugares, en relación a una consulta para crear resultados de búsqueda interconectados. Un modelo sencillo que represente las relaciones que conectan a distintos recursos de la web es ***un grafo dirigido con aristas etiquetadas.***

La Figura 2 nos muestra un grafo respecto al panel de conocimiento de la busqueda "Facultad de Ingeniería, UdelaR", correspondiente a la Figura 1.

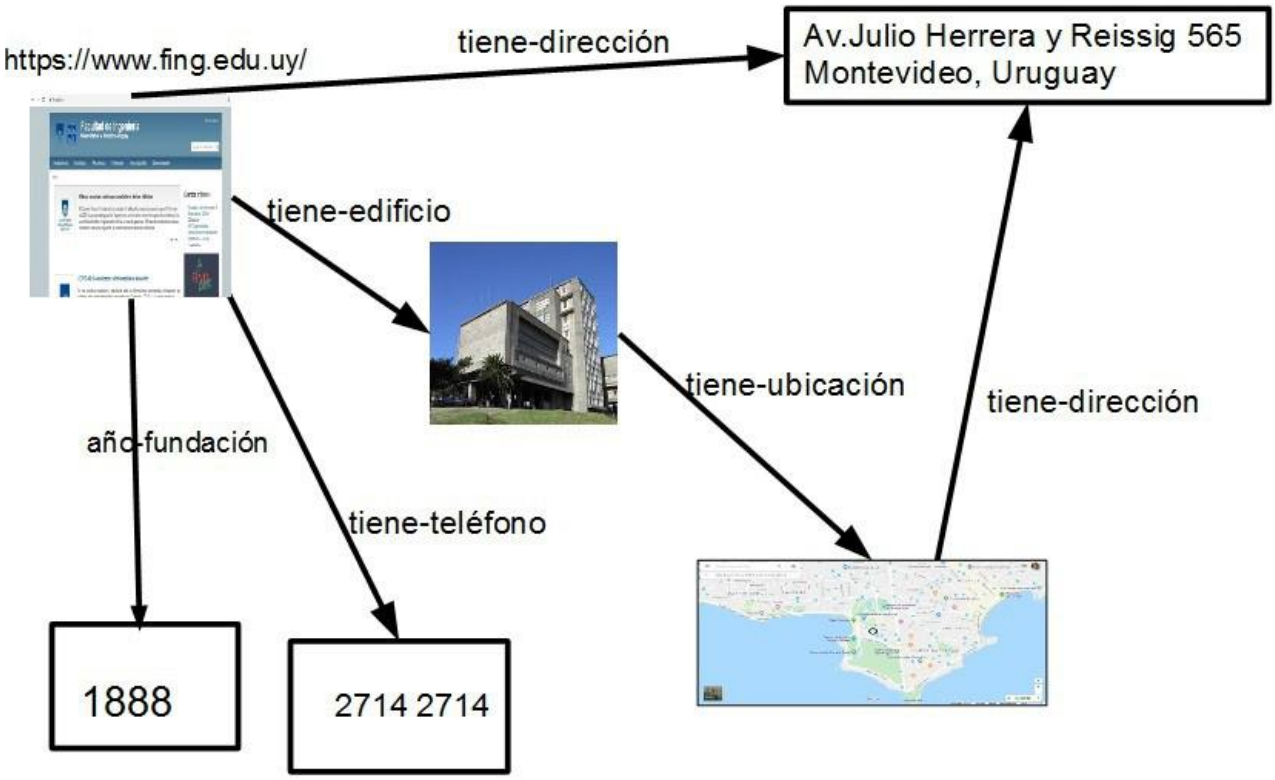


Figura 2. Ejemplo de grafo de la Facultad de Ingenieria UdelaR.

En https://www.youtube.com/watch?v=mmQl6VGvX-c puedes ver el video de presentación de Google cuando lanzaron el Panel de Conocimiento explicando el uso del "Grafo de Conocimiento" en el 2012.
Actualmente el grafo representa recursos mucho más interactivos que al comienzo. Si se busca por ejemplo un negocio se puede obtener la franja horaria con indicaciones de las horas más ocupadas.

Es interesante buscar distintos recursos en la web, como grupos musicales, artistas, alojamientos y observar que también se muestran en el panel las redes sociales donde tiene presencia y recomendaciones relacionadas.

Podrías responder estas preguntas?:

¿Por qué trabajar con un grafo de conocimientos y no trabajar con bases de datos relacionales?

¿Qué tiene un grafo de conocimientos que lo hace diferente a un grafo de datos?

En esta unidad del curso trataremos de contestar estas preguntas.

La Figura 2 muestra un ejemplo de grafo en relación a lo que vemos en el Panel de Conocimiento de Google, lo que nos está mostrando es un **grafo de datos**. Un grafo de datos es un grafo donde los nodos son datos y las aristas representan relaciones entre los datos.

2.2 Características de un grafo de datos

a. Open World Assumption

El grafo de datos utiliza el criterio de Open World Assumption (OWA), esto significa que la ausencia de un dato o una relación en el grafo, no nos permite asumir que ese dato o esa relación no existe en el mundo real.

El criterio contrario, cuando se asume que la ausencia de un dato significa que el dato no existe en el mundo real se llama Closed World Assumption (CWA) y es el criterio que utilizan las bases de datos relacionales. Si por ejemplo en una tabla relacional de CiudadesTuristicas no está el dato "Montevideo" se asume que "Montevideo" no es una CiudadTuristica, mientras que si en un grafo de datos no existe relación entre el dato "Montevideo" y el dato "CiudadTuristica" no se puede decir que Montevideo no sea una CiudadTuristica.

Una consecuencia de adoptar CWA es que la adición de una arista al grafo de datos podría contradecir lo que anteriormente se suponía que era falso (debido a la falta de información), mientras que con OWA, una declaracion que se demuestra como falsa continúa siendo falsa con la adición de aristas. Esto hace que la opción de OWA sea la más adecuada para modelar datos incompletos.

b. El grafo de datos puede prescindir de un esquema

En un grafo de datos se puede trabajar sin la definición de un esquema (observando la Figura 2 vemos que solo conecta datos sin definir ningún esquema).

Modelar los datos como un grafo prescindiendo del esquema ofrece una mayor flexibilidad para integrar nuevas fuentes de datos comparado con el modelo relacional estándar, donde un esquema debe ser definido de antemano y seguido en cada paso. Sin embargo, también se puede modelar datos en grafos usando esquemas para describir una estructura y/o semántica de alto nivel que el grafo sigue o debería seguir.

En general se trabaja con alguno de los siguientes tres tipos de esquemas de grafos: esquema semántico, esquema de validación y/o esquema emergente.

Un *esquema semántico* permite definir el significado de términos de alto nivel (también conocido como vocabulario o terminología) utilizado en el grafo, lo que facilita el razonamiento sobre los grafos utilizando esos términos.

A pesar de que la opción de OWA es la más adecuada para datos incompletos, en algunos escenarios es posible que deseemos garantizar que nuestro grafo de datos, o partes específicas del mismo, estén en algún sentido "completos". Por ejemplo, podemos asegurarnos de que todas las representaciones de "eventos" tengan al menos un nombre, un lugar, una fecha de inicio y una fecha de finalización, de modo que las aplicaciones que usan los datos, pueden garantizar que tengan la información mínima requerida. Podemos definir tales restricciones en un *esquema de validación* y validar el grafo de datos con respecto al esquema resultante, esto sería al estilo de un XML-Schema. Así, mientras que los esquemas semánticos permiten inferir nuevos datos desde los datos existentes en el grafo, un esquema de validación permite validar que los datos existentes en el grafo son los requeridos. Una forma estándar de definir un esquema de validación para grafos es usando *shapes* [5].

Tanto los esquemas semánticos como los de validación requieren que un experto en el dominio explique y especifique las definiciones y restricciones que deben cumplir los datos del grafo. Sin embargo, ocurre a veces que un grafo de datos tiene estructuras latentes que se pueden extraer automáticamente como un *esquema emergente*. Una forma de definir un esquema emergente es el de los *quotient graphs*, que particionan grupos de nodos en el grafo de datos de acuerdo con alguna relación de equivalencia, conservando algunas propiedades estructurales del grafo.

En esta unidad nos limitaremos a estudiar el esquema semántico al estudiar RDF. Un buen trabajo sobre los distintos esquemas de grafos es el artículo de Čebirić *et. al* [6].

c. El grafo de datos representa fácilmente ciclos:

Mientras que otros modelos de datos estructurados como los árboles (XML, JSON, etc.) serían muy similares en flexibilidad, los grafos no requieren organizar los datos jerárquicamente. El grafo permite representar y consultar ciclos entre relaciones de los datos directamente, mientras que en los modelos de datos relacionales es necesario realizar muchas veces varias operaciones de join para trabajar sobre ciclos de datos.

d. Analíticas de grafos de datos:

La analítica es el proceso de descubrir, interpretar y comunicar patrones significativos inherentes a (generalmente grandes) colecciones de datos.
La analítica de grafos es entonces la aplicación de procesos de análisis sobre (típicamente grandes) grafos de datos. Un grafo de datos permite tener analíticas de sus datos aplicando técnicas de la teoría de grafos y del análisis de redes, obteniendo así propiedades interesantes de los datos del grafo como ser:

Centralidad (identificar los nodos más importantes -centrales- en el grafo)

Detección de Comunidades (detectar sub-grafos en los cuales los nodos estan más densamente conectados) y *Similaridad de nodos* (detectar conjunto de nodos que son similares por los tipos de aristas que conectan con sus vecinos).

¿Bases de datos relacional o grafo de datos?

Para decidir si es mejor usar un modelo de datos relacional o un grafo de datos debemos indicar **cuál es el objetivo de cada uno de los modelos de datos.**

Si nuestra aplicación necesita las funcionalidades cubiertas por el objetivo del modelo relacional este será más adecuado, si al contrario, nuestra aplicación está enfocada a resolver objetivos priorizados por el modelo de grafo de datos, entonces un grafo de datos será más adecuado.

Mientras en el modelo de datos relacional el objetivo es priorizar las transacciones ACID (sigla de: Atomicity, Consistency, Isolation and Durability, o Atomicidad, Consistencia, Aislamiento y Durabilidad en español)**, en los grafos de datos el objetivo es priorizar las propiedades FAIR de los datos.**

FAIR se refiere a los términos en inglés de:

Findable, Accessible, Interoperable, Reusable

En 2016, los principios rectores FAIR para la gestión y administración de datos científicos se publicaron en la revista Scientific Data [1], accesible en la sección de materiales complementarios de esta unidad: FAIR-data.pdf y también en el enlace: https://www.nature.com/articles/sdata201618.pdf. En ese artículo los autores proporcionan pautas para mejorar la capacidad de búsqueda, accesibilidad, interoperabilidad y reutilización de recursos digitales. Los principios enfatizan la capacidad de acción de la máquina (es decir, la capacidad de los sistemas computacionales para encontrar, acceder, interoperar y reutilizar datos con ninguna o mínima intervención humana) porque los humanos dependen cada vez más del soporte computacional para manejar los datos como resultado del aumento en el volumen, complejidad y velocidad de creación de datos. En la Figura 3 presentamos el resumen de los principios FAIR.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Figura 3. Principios FAIR de [1]

Un ejemplo de un modelo de datos estandarizado basado en un grafo de datos dirigido con aristas etiquetadas que utiliza metadatos es RDF (Resource Description Framework). RDF es recomendado por el W3C y tambien considerado un tipo de grafo de conocimientos, lo estudiaremos en la Parte II de esta unidad.

2.3 Diferencias de un grafo de conocimientos de un grafo de datos

Según M. Nickel, K. Murphy, V. Tresp and E. Gabrilovich en su artículo *A Review of Relational Machine Learning for Knowledge Graph* (2016 [2]) definen un Grafo de Conocimientos como:

[...] un grafo de conocimientos es un grafo estructurado que representa **bases de conocimientos** (KBs) que almacenan información factual en forma de relación entre entidades".

Esta definición de grafo de conocimientos se puede visualizar como un grafo de datos destinado a acumular y transmitir conocimientos del mundo real, cuyos nodos representan entidades de interés y cuyas aristas representan las relaciones entre estas entidades.

El grafo de datos se ajusta a un modelo de datos basado en un grafo, que puede ser un grafo dirigido con aristas etiquetadas, un grafo de propiedades u otro tipo de grafo. Por conocimiento, se hace referencia a algo que se conoce. Esos conocimientos pueden acumularse a partir de fuentes externas o extraerse del grafo de conocimiento en sí mismo.

El conocimiento puede estar compuesto por simples declaraciones, como "Montevideo es la capital de Uruguay", o declaraciones cuantificadas, como "Todas las capitales son ciudades". Declaraciones simples pueden acumularse como aristas en el grafo de datos. Para que el grafo de conocimientos pueda acumular declaraciones cuantificadas ("Todas las capitales son ciudades"), entonces es necesaria una base de conocimientos.

Al hablar de una base de conocimientos se hace referencia a que se utilizan métodos deductivos para derivar y acumular más conocimientos (por ejemplo, a partir de que "Montevideo es una capital" y que "Todas las capitales son ciudades" entonces se deriva que "Montevideo es una ciudad"). Esta es la gran diferencia entre un grafo de datos y un grafo de conocimientos.

El grafo de conocimientos tiene además del grafo de datos un mecanismo que le permite hacer inferencias. Las inferencias pueden ser inductivas utilizando redes neuronales o deductivas, utilizando bases de conocimiento.

El grafo de conocimientos tiene además del grafo de datos
un mecanismo que le permite hacer inferencias.

La utilización de mecanismos de inferencias a partir de un conjunto de hechos o afirmaciones nos permite hacer explícito conocimiento implícito. Este mecanismo es fundamental para lograr capturar los componentes que hacen a la *semántica* de los datos en la Web Semántica o Web de Datos.

En la Web Semántica es necesario además tener mecanismos para poder desambiguar las entidades, "Montevideo es una ciudad" pero también "Montevideo es un departamento". O para entender el significado del concepto "Jaguar" (con sus posibilidades de ser comprendido como auto o como animal) es necesario primero comunicarlo utilizando símbolos de un lenguaje. Luego que lo conseguimos comunicar decimos que "se entiende" si se **interpreta correctamente** la información del contenido del recurso o del mensaje comunicado.

Para **interpretar correctamente** un concepto es importante: tener elementos del contexto, tener experiencia con el uso del concepto y comprender la pragmática (intención del que envía el mensaje o crea el contenido). Estos elementos se describen en el llamado Triángulo Semiótico de Ogden *et. al* [3]. La Figura 4 nos muestra el triángulo semiótico para el concepto "jaguar".

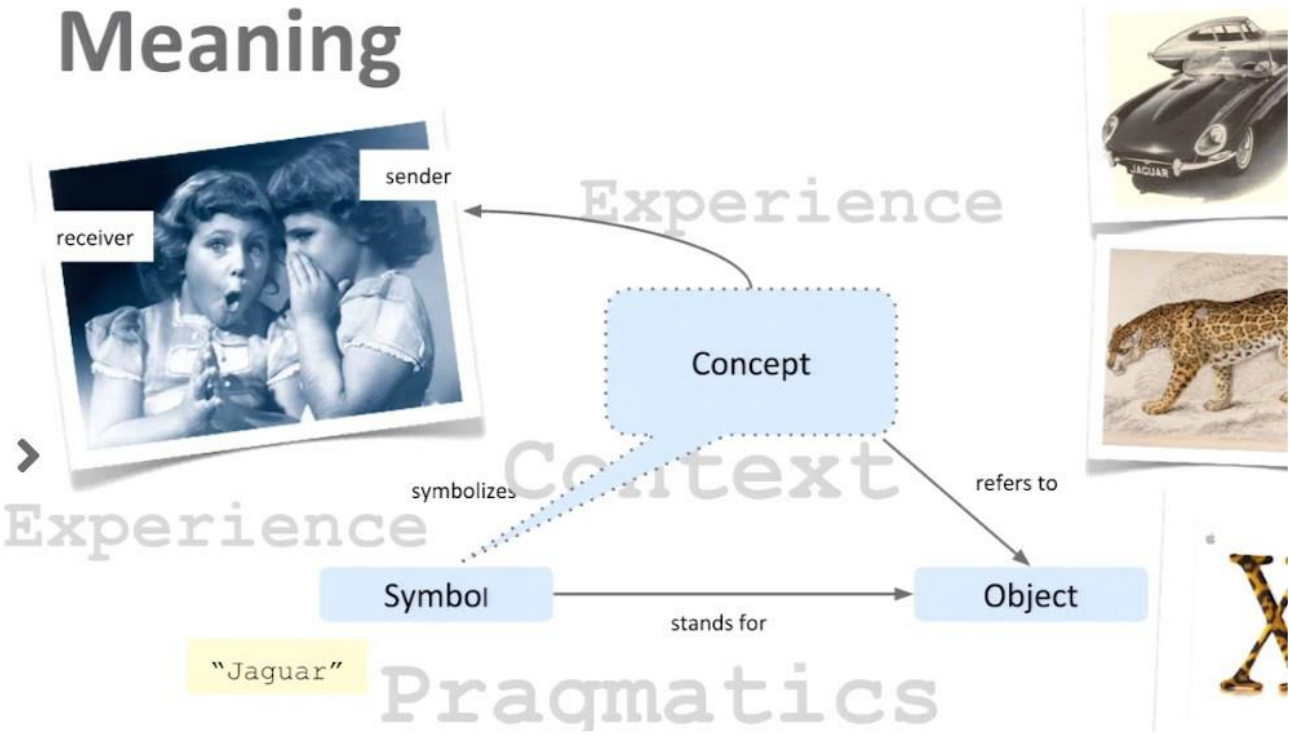


Figura 4. Ejemplo de Triangulo Semiótico.
(Tomado de presentacion de Dr. Harald Sack – FIZ Karlsruhe.)

La tercera generación de la web, la Web 3.0 es la Web Semántica, también llamada la Web de Datos, no solo datos leídos, decodificados, por programas sino "entendidos", **interpretados correctamente** por programas.

Para lograr la interpretación correcta de los datos de la web por programas necesitamos encontrar, según el triángulo semiótico, su pragmática, su contexto y sus experiencias de uso.
Estos temas junto con el análisis de su sintaxis son temas de estudio ampliamente tratados por el área de Procesamiento de Lenguaje Natural (NLP de su sigla en inglés).

Otra forma, complementaria al NLP, consiste en expresar de manera explícita la semántica de los datos expresando explícitamente con metadatos la pragmática, contexto y experiencias de uso de los datos. Este enfoque corresponde a la aplicación de técnicas de la Web Semántica utilizando grafos de conocimientos, en particular grafos RDF.

2.4 Grafos de Conocimiento en el mundo real

Existen grafos de conocimientos abiertos, como: Dbpedia, Wikidata, Freebase, YAGO. DBpedia y Wikidata complementan a la Wikipedia. DBpedia construye un grafo de conocimiento extrayendo datos de las publicaciones de Wikipedia. Por otro lado, Wikidata construye su grafo de conocimiento directamente con aportes de las personas y es la Wikipedia quien utiliza a la Wikidata para complementar sus datos (pueden observar en una pagina de la Wikipedia la opción de seguir navegando a través de las etiquetas de herramientas con la opción de Wikidata).

Existen también varios grafos de conocimientos empresariales, algunos casos que puedes leer sobre el uso que hacen de los grafos de conocimiento son:

AirBnb https://medium.com/airbnb-engineering/scaling-knowledge-access-and-retrieval-atairbnb-665b6ba21e95

Zalando https://jobs.zalando.com/en/tech/blog/semantic-web-technologies/

Thomson Reuters https://neo4j.com/blog/intelligent-recommendation-engine-financial-analysts/

MATERIAL RECOMENDADO:

***Industry-scale Knowledge Graphs: Lessons and Challenges:* Five diverse technology companies**

Natasha Noy, Google; Yuqing Gao, Microsoft; Anshu Jain, IBM Watson; Anant Narayanan, Facebook; Alan Patterson, eBay; Jamie Taylor, Google.

Este articulo [4] recopila el panel realizado en la conferencia *International Semantic Web Conference in Asilomar, California, in October 2018 (*http://iswc2018.semanticweb.org/panelenterprise-scale-knowledge-graphs/*)*

REFERENCIAS

[1] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich.(2016). *A review of relational machine learning for knowledge graphs*. Proc. of the IEEE. (Encuentran el artículo en el archivo definition-KG.pdf en el EVA del curso)

[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data, 3,160018 .
https://doi.org/10.1038/sdata.2016.18 (Encuentran el artículo en el archivo FAIR.pdf en el EVA del curso)

[3] Ogden, C. K., Richards, I. A., Malinowski, B., Crookshank, F. G., & Postgate, J. P. (1923). *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism*. London: K. Paul, Trench, Trubner & Co.

[4] Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson,A.&Taylor, J. (2019). *Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies.* acmqueue, Volume 17, issue 2. (Encuentran el artículo en el archivo acmqueue.pdf en el EVA del curso)

[5] Holger Knublauch and Dimitris Kontokostas. 2017. Shapes Constraint Language (SHACL), W3C Recommendation 20 July 2017. W3C Recommendation. World Wide Web Consortium.
https://www.w3.org/TR/2017/REC-shacl-20170720/

[6] Šejla Čebirić, Francois Goasdoue, Haridimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika. 2019. Summarizing semantic graphs: a survey. The Very Large Data Base Journal 28, 3 (2019), 295–327.

---------- Fin Unidad4-Parte I ----------