Informatio
22 (1), 2017, pp. 3-12 ISSN: 2301-1378

Explorando el "Anáforas": minería de datos a través del QDA – Mine Lite y estudio de los modelos semánticos aplicables

Diana Comesaña[1]

Resumen
Se considera la exploración de la prensa digitalizada por la Facultad de Información y Comunicación (Universidad de la República, Uruguay), disponible a través del portal Anaphoras, ampliando el concepto de minería de datos al estudio de estos documentos. El propósito es agregar semántica a estos documentos, de acuerdo con el concepto de informática, para lograr la recuperación automática de los datos. Como caso especial consideramos los eventos climáticos extremos y las huellas que la prensa recoge a través de sus efectos detectables. Para el desarrollo del modelo semántico se emplean la clase eventos meteorológicos de las ontologías Falcons y la ontología Usos Especiales del meta-modelo desarrollado en la tesis "Modelo Conceptual de Información Geográfica para IDE - Uruguay" (Comesaña, 2015). Las fuentes terminológicas utilizadas fueron: el glosario publicado en la página del Instituto Uruguayo de Meteorología (IN.U.MET) y el glosario de la Organización Meteorológica Mundial (OMM).

Palabras clave: Minería de Datos, Recuperación automática, modelos semánticos.

Exploring the "Anáforas": Data Mining Through QDA - Mina Lite and study of applicable semantic models

Abstract
The exploration of the digitized press is considered by the Faculty of Information and Communication (University of the Republic, Uruguay), available through the portal Anaphoras, extending the concept of data mining to the study of these documents. The purpose is to add semantics to these documents, according to the concept of computing, to achieve automatic retrieval of data. As a special case we consider the extreme climatic events and the traces that the press collects through its detectable effects. For the development of the semantic model, the meteorological-events class of the Falcons ontologies and the ontology Usos_Especiales of the meta-model developed in the thesis "Conceptual Model of Geographical Information for IDE - Uruguay" (Comesaña, 2015) are used. The terminological sources used were: the glossary published on the page of the Uruguay Meteorological Institute (IN.U.MET) and the glossary of the World Meteorological Organization (WMO)

Key words: Data Mining, Automatic Recovery, Semantic Models.

[1] Departamento de Tratamiento y Transferencia de la Información, Instituto de Información y Comunicación. Facultad de Información y Comunicación, Universidad de la República. Correo electrónico: diana.comesana@fic.edu.uy

1. Introducción

Se llama Minería de Datos o Knowledge Discovery of Data (KDD) al:

> proceso asistido por computadora que analiza una enorme cantidad de datos y extrae el conocimiento exacto de los mismos. El conocimiento extraído permite predecir el comportamiento presente y futuro. Esto permite tomar decisiones positivas. … El conocimiento se extrae de los datos históricos mediante la aplicación de reconocimiento de patrones, técnicas estadísticas y matemáticas, lo que da lugar al conocimiento en forma de hechos, tendencias, asociaciones, patrones, anomalías y excepciones. (Amala Jayanthi, Swathi, Tharakai, 2016)

Si se habla de Minería de Datos Semántica, hay que incorporar el conocimiento del dominio a estudio en el proceso de búsqueda y análisis de los datos; así el primer paso es lograr la representación y construcción del conocimiento por modelos comprensibles para los ordenadores y tales que éstos puedan realizar inferencias a partir de ellos. Como indican Amala Jayanthi, Swanthi y Tharakai (2016) es un desarrollo interdisciplinario emergente entre la gestión del conocimiento y la informática; se trata de un proceso de análisis del descubrimiento de información a partir de un gran volumen de datos de cualquier tipo.

Generalmente el término "minería de datos" hace referencia a la búsqueda de los mismos en grandes volúmenes de datos contenidos en bases de datos, pero ¿dónde encontrar mayor cantidad de datos que nos muestren cómo una sociedad ha visto los acontecimientos de su tiempo, sino en la prensa escrita?

La minería de datos es un proceso que comprende tanto la formulación de preguntas sobre estos datos y la creación de un modelo para responderlas, como la implementación del modelo en un entorno de trabajo. Debe definirse el problema, preparar los datos para su explotación, explorarlos, generar modelos semánticos, implementarlos y actualizarlos.

Dice Pierre Vilar (1980) "comprender es imposible sin conocer. La historia debe enseñarnos, en primer lugar, a leer un periódico". Se considera, entonces, la exploración de la prensa digitalizada por la Facultad de Información y Comunicación (FIC – Universidad de la República, Uruguay), disponible a través del portal Anáforas (http://anaforas.fic.edu.uy/jspui/)

El proyecto Anáforas consiste en el rescate digitalizado de documentos del pasado. Este proyecto corresponde al Seminario de Fundamentos Lingüísticos de la Comunicación de la Facultad de Información y Comunicación (FIC), donde constituye una de las unidades curriculares del Ciclo de Graduación de la Licenciatura en Comunicación. El equipo que lleva adelante este proyecto está dirigido por la docente Lisa Block de Behar, es coordinado por Arturo Rodríguez Peixoto y está compuesto por Rodrigo Echániz, Maximiliano Basile, Mariana Noguera, Ileane Sequeira e Ignacio Saraiva.

El portal web del proyecto Anáforas cuenta con tres colecciones de documentos digitalizados: Biblioteca Digital de Autores Uruguayos, Publicaciones Periódicas del Uruguay y Figuras. "Publicaciones Periódicas del Uruguay" tiene el propósito de hacer accesibles, colecciones de revistas y diversos periódicos del Uruguay, abarcando desde el siglo XIX hasta las primeras ocho décadas del siglo XX. La información que contienen estos documentos para la investigación de las más variadas disciplinas es riquísima, y la componen un sin fin de datos escondidos. Para poder encontrarlos y

aprovecharlos necesitamos realizar una verdadera labor de minería.

Se puede ampliar el concepto de minería de datos a este caso particular, sin pretender realizar un proceso de minería de datos tal como se concibe normalmente, pues nuestra información no se halla estructurada en bases de datos: son textos que, según su antigüedad y estado de conservación de los documentos originales, algunos han sido digitalizados, en condiciones ideales con un programa OCR (del inglés Optical Character Recognition), y otros como una imagen. Se tiene el propósito de agregar semántica a los mismos concebida según el concepto informático de la misma, para lograr la recuperación automática de los datos contenidos.

2. Objetivos

El propósito de este trabajo es determinar un método ágil y automático de análisis de los textos digitalizados que permita luego agregar semántica, según la acepción informática de ésta, para lograr la recuperación automática de los datos e interoperabilidad de los mismos.

Como caso particular se consideran los eventos meteorológicos extremos: tormentas, precipitaciones intensas, sequías. Todos estos eventos dejan huella en la vida del hombre y sus efectos son recogidos por la prensa: inundaciones, caminos y rutas cortadas, voladuras de techos, casas derrumbadas, pérdidas agro-económicas y tantos otros desastres que marcan la vida de una comunidad.

3. El modelo semántico del dominio

Como explican Dejing Dou, Hao Wang y Haishan Liu (2015), las ontologías constituyen un modelo exitoso para el conocimiento de un dominio, "son una especificación formal y explícita de una conceptualización" (Gruber, 1993).

Nos muestran lo que existe en un dominio del conocimiento y definen un vocabulario común incluyendo la interpretación de los conceptos básicos de dicho dominio y las relaciones entre ellos, restricciones y reglas sobre esos conceptos, de una forma estándar. Actúan como modelos de referencia comunes entre sistemas diferentes que utilizan conceptos similares facilitando la interoperabilidad.

Una ontología posibilita que los programas informáticos “razonen semánticamente” sobre determinados conceptos. Definen una teoría lógica que les permite realizar inferencias, pues se basan en la lógica descriptiva y contienen: Clases (conceptos generales en el dominio), Instancias (instancias particulares del concepto), Relaciones entre instancias/clases, Propiedades de clases e instancias, Funciones y procesos que involucran a clases e instancias y Restricciones y reglas de inferencia.

Se desarrollan en el lenguaje OWL (Lenguaje de Ontologías Web) que proporciona un mejor mecanismo de interpretabilidad de contenido Web, al permitir expresar mejor las relaciones y establecer restricciones y reglas de inferencia.

Procesar datos que están en la web implica depurarlos para eliminar “ruido”, normalizar su presentación y descripción y desarrollar los mecanismos de extracción pertinentes, en estos pasos suelen existir brechas semánticas. La semántica de los datos es necesaria para comprender las relaciones de los mismos y las ontologías han demostrado ser beneficiosas en este proceso.

La IDE - Uruguay (Infraestructura de Datos Espaciales Uruguaya) “es un órgano

desconcentrado de Presidencia de la República, con autonomía técnica, que tiene por finalidad ordenar la producción y facilitar la disponibilidad, el acceso y uso de productos y servicios de información geográfica del territorio nacional, como apoyo a los procesos de toma de decisiones para el desarrollo sostenible" (IDE-Uy, 2014)

En la tesis Modelo conceptual de información geográfica para la IDE – Uruguay (Comesaña, 2015) se propone meta - modelo conceptual para la información geográfica conformado por once ontologías que forman una red de ontologías para cubrir todo el dominio de la información geográfica.

Este meta-modelo es tal, que un concepto en una ontología es una instancia de otro concepto, llamado meta-concepto, que a su vez es instancia de otro meta-meta-concepto, tal como nos muestran Rohrer, Motz y Severi en Reasoning for ALCQ extended with a exible meta-modelling hierarchy (2014), permitiendo crear un número indeterminado de niveles de meta-conceptos.

Entre las ontologías de la tesis anteriormente mencionada, la que recibe el nombre de "Usos_Especiales" (super-clase), tiene como una de sus clases "Información_meteorológica", mencionada, pero no desarrollada. Las ontologías se crean para reutilizarse, adaptarse y compartirse, por lo que se parte de las ontologías Falcons search, especialmente de aquella que presenta la clase meteorological_event, sub-clase de weather_events (http://umbel.org/reference-concept/?uri=ImmediateWeatherProcess) y se desarrolla y adapta a las necesidades del proyecto y del meta-modelo mencionado, empleando como fuente terminológica, para la selección del vocabulario, el glosario publicado en la página del Instituto Uruguayo de Meteorología (IN.U.MET., http://www.meteorologia.com.uy/biblioteca/glosario) y el glosario de la World Meteorological Organization (W.M.O., http://wmo.multicorpora.net/MultiTransWeb/Web.mvc).

Los efectos ocasionados por los eventos meteorológicos suficientemente notorios como para ser recogidos en los periódicos de época, se transformarán en instancias del modelo desarrollado.

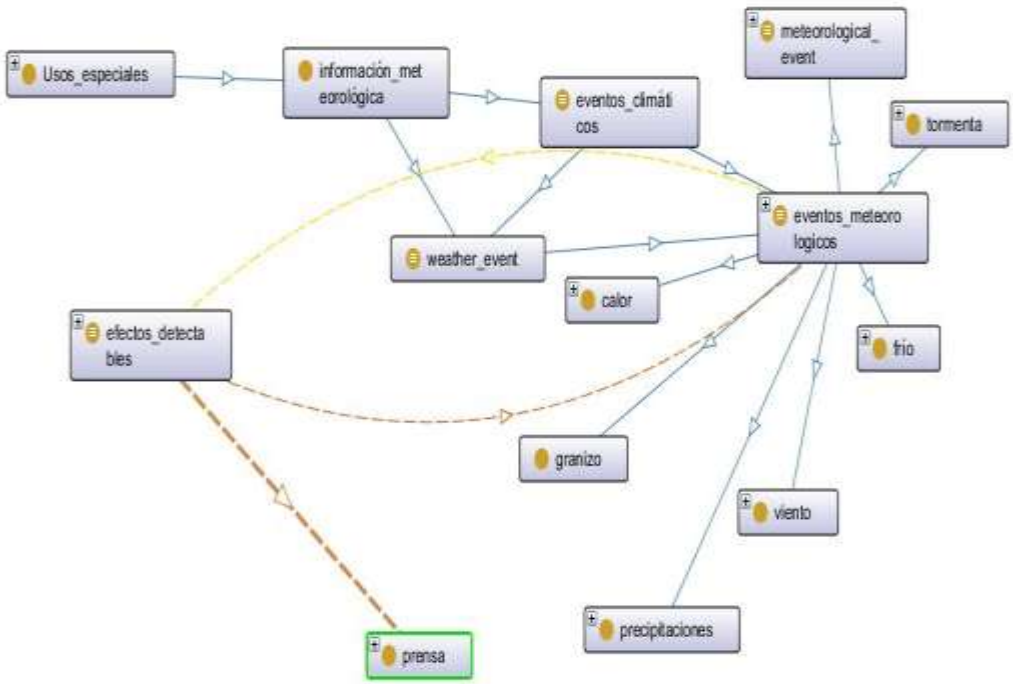

Figura 1: Ontógrafo de la subclase eventos_meteorológicos, clase información meteorológica, ontología Usos_especiales, del meta-modelo Objetos_geográficos

El ontógrafo de la figura 1, muestra las relaciones entre los eventos meteorológicos, sus efectos detectables y la prensa que se analiza.

3. La minería en el "Anáforas"

Toda minería comienza con un pre-tratamiento de los datos, que comprende la limpieza, integración, selección y transformación de los mismos, para su mejor explotación. Este proceso ya ha sido llevado a cabo por parte del equipo del "Anáforas", por lo que la tarea a desarrollar es aquella que lleve a encontrar y recuperar automáticamente la información deseada.

Como el método de digitalización ha seguido dos caminos, según el estado físico de los documentos, la digitalización como imagen para los documentos con mayor estado de deterioro y el empleo de OCR para los demás. La exploración debe seguir dos caminos, según el formato de digitalización empleado.

El siguiente diagrama de flujo nos muestra los caminos a emprender, según el tipo de digitalización llevado a cabo:

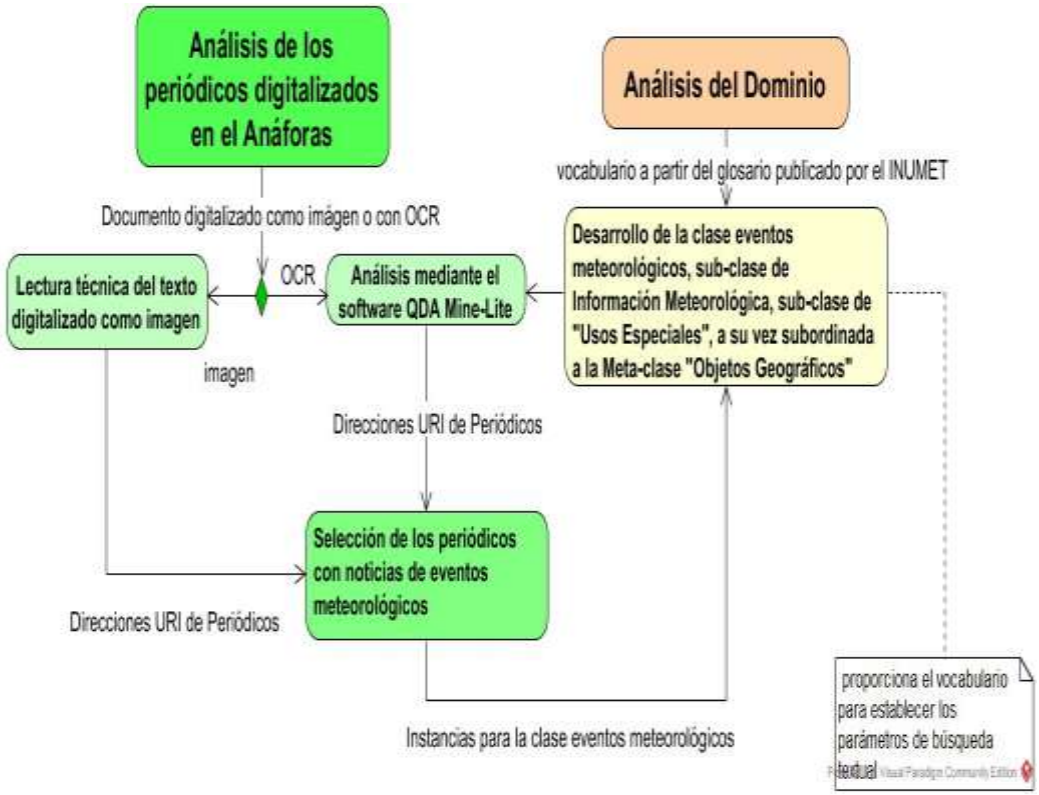

Figura 2: Diagrama de flujo del trabajo

El camino de la lectura técnica es ampliamente conocido, por lo que este documento se centra en el uso del software de análisis de datos cualitativos QDA – Miner Lite. El QDA – Mine Lite es la versión gratuita del QDA Miner. Esta es una herramienta de análisis cualitativo y mixto de datos. Es un software destinado a codificar textos y fotografías en datos que tiene la capacidad de codificar y recuperar texto, almacenándolos en una base de datos interna y realizar otras tareas que corresponden al análisis estadístico.

Permite importar documentos de texto en formatos RTF, HTML, PDF, datos almacenados en Excel, MS Access, CSV, archivos de texto delimitados por tabuladores. Importa datos desde los softwares de codificación cualitativa como el Altas.ti, HyperResearch, Etnograph, y desde herramientas de transcripción como Transana y Transcriber, así como desde archivos de Sistemas de Información de Referencia (.RIS).

Con él se puede realizar una codificación intuitiva con códigos organizados en estructura de árbol, añadir notas de comentarios a los segmentos codificados, a cada archivo estudiado (casos) y a todo el proyecto. Incluye herramientas booleanas de búsqueda rápida de texto para recuperar y codificar segmentos de texto y puede recuperar por la de codificación con operadores booleanos (y, o, no) y de proximidad (incluye, incluido, cerca, antes, después).

Estas cualidades y su interfaz amigable lo transforman en una herramienta muy útil para estudiar la prensa digitalizada con las características de OCR y publicada en el repositorio a estudiar. En nuestro caso, solamente nos interesa un uso de éste tipo.

4. Selección del tema

¿Por qué se tomó como tema los eventos climáticos extremos ocurridos en Uruguay? Recordemos un poco lo ocurrido durante el año 2016. El 15 abril de este año, todo el país se vio conmovido por un tornado ocurrido en la ciudad de Dolores que puso en la conciencia de los uruguayos la vulnerabilidad ante estos fenómenos. Según el diario “El Observador”, en su versión digital, el Sistema Nacional de Emergencia (SINAE) contabilizó 34 tornados entre 1968 y 2011 y en la escala Fujita, utilizada para medir tornados, que va del nivel 0 al 5, el tornado en Dolores se ubicó en el nivel 3, catalogado como "severo", según un informe del Departamento de Ciencias de la Atmósfera de la Facultad de Ciencias.

Y, ¿qué pasó antes de 1968? ¿qué eventos climáticos extremos se produjeron antes de los registros que maneja el SINAE? La prensa de época puede proporcionar “pistas” de lo ocurrido en esos años. Los fenómenos podrán ubicarse por los efectos percibidos y documentados en ella.

El desarrollo de la subclase eventos_meteorológicos, clase información meteorológica, ontología Usos_especiales, del meta-modelo Objetos_geográficos nos proporcionará los parámetros textuales para la búsqueda con el QDA Mine-Lite: temporales, bajas temperaturas, inundaciones, cortes de rutas y caminos, voladuras de techo, incendios forestales, pérdidas agro-pecuarias y otras similares.

5. Obtención de las instancias para la clase eventos meteorológicos, de la ontología "Usos Especiales"

Tomemos, para realizar el prototipo, el periódico bimensual "LaTrinidad" de Uruguay, publicado entre los años 1878 – 1882. Está digitalizado con OCR, por lo que puede aplicarse el análisis a través del QDA – Mine-Lite.

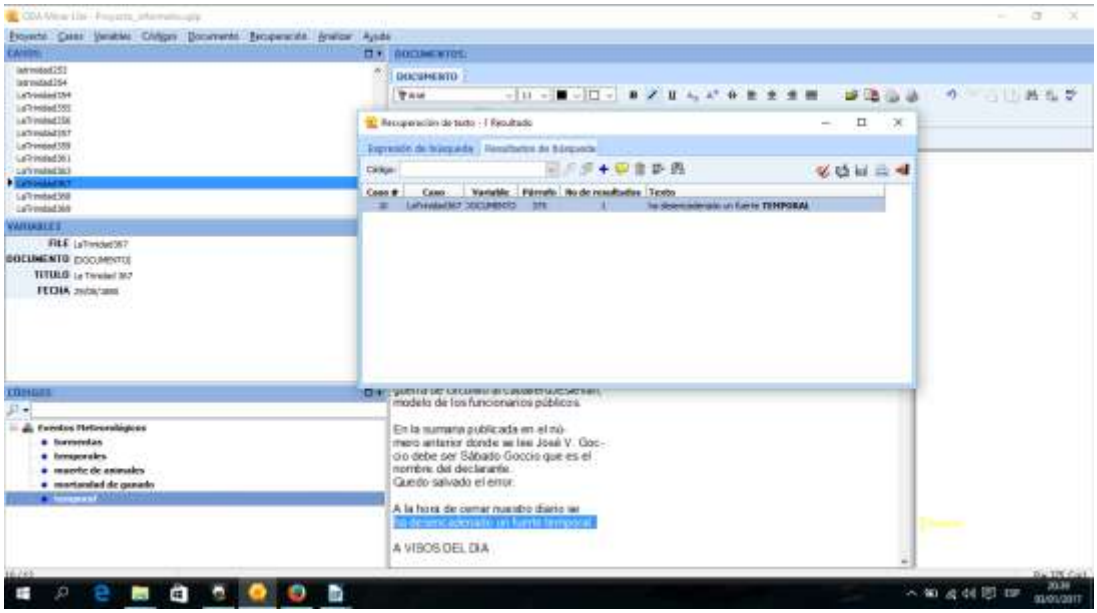

Figura 3: Búsqueda de texto con el QDA Mine-Lite

La figura 3 muestra la captura de pantalla donde la herramienta detecta las instancias para el término "temporal e identifica en que periódico se encuentra. Aplicamos, en este ejemplo, como parámetros de búsqueda: temporal, muerte de animales y mortandad de ganado, estos dos últimos por ser efectos detectables del evento meteorológico temporal. Hay que destacar, que en esta convulsa época de los inicios del país, los efectos tienen que ser de gran magnitud para que los recoja la prensa. Dado que el estado de los originales no es muy bueno, su digitalización con OCR, no es óptima, por lo que de una lectura técnica obtuvimos dos instancias más: en La Trinidad Nº 254, del 24 de julio de 1879, aparece la muerte de animales asociada a un temporal y en La Trinidad Nº 361, del 8 de agosto de 1880, la mortandad de ganado debida a condiciones climáticas.

Una vez obtenidas estas instancias, las trasladamos a nuestra ontología y una consulta SPARQL, proporciona el listado de los periódicos en que se detectan efectos de los eventos climáticos (figura 4):

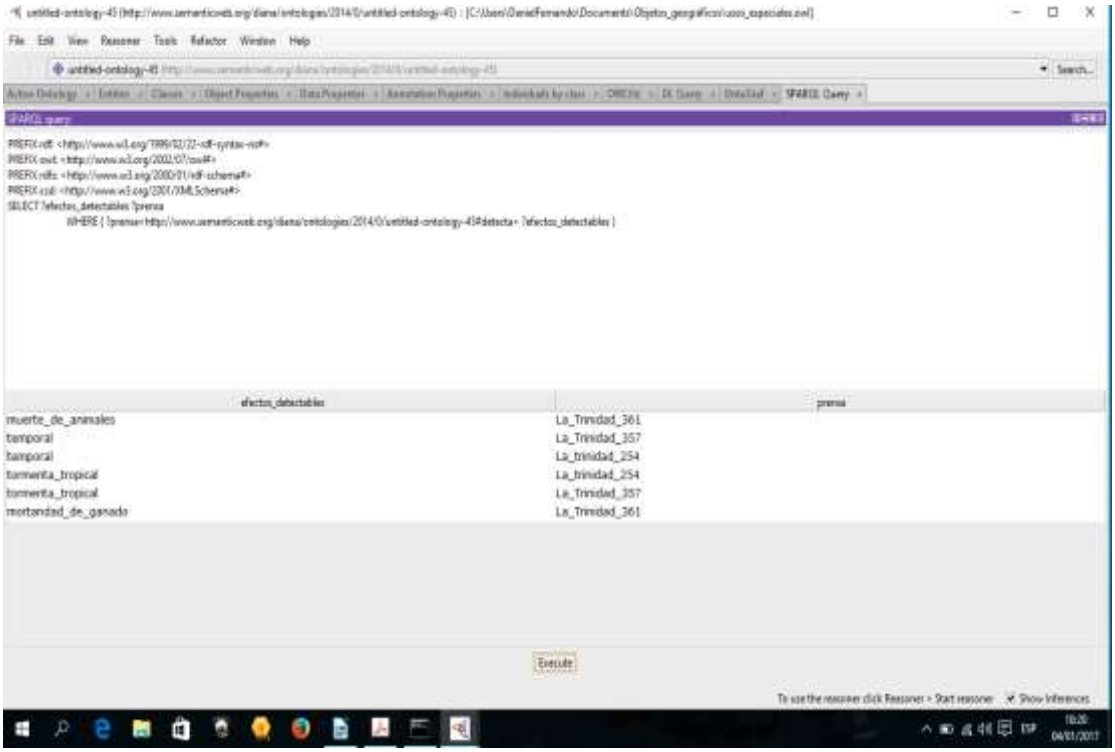

Figura 4: Resultado de una consulta SPARQL al modelo.

4. Reflexiones finales

El procedimiento para realizar nuestra "minería de datos" en la prensa digitalizada no presenta mayores dudas y dependerá exclusivamente del tipo y calidad de la digitalización. El uso de un sofware de análisis cualitativo de datos como el QDA-Mine Lite nos exige primero tener una noción de los términos que buscamos, y a través de los artículos encontrados hallaremos otros relacionados.

Los resultados de las pruebas, se comprueba que, como dice Chantell LaPan (2013) los investigadores que utilicen estas herramientas, deben hacerlo para la investigación puramente exploratoria, teniendo cuidado de no asumir que la computadora puede hacer toda la codificación para ellos, es decir es necesario complementar con una lectura técnica.

Sin embargo, avanzando en el proyecto se comprueba solamente un 15% de casos no detectados, por lo que, ante la existencia de tiempos acotados para la investigación, el método es aceptable.

Respecto a la ventaja de emplear el desarrollo de una ontología, para la representación del conocimiento, está dada por el hecho de que al estar éstas basadas en la lógica descriptiva, constituyen un modelo abstracto que permite que los programas informáticos "razonen semánticamente" sobre los conceptos representados y realicen inferencias, lo que contribuye a la interoperabilidad de los datos

Referencias

AMALA JAYANTHI. M. SWATHI, S. THARAKAI, R. (2016). Data Mining– A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 6
Recuperado de: https://www.ijarcsse.com/docs/papers/Volume_6/4_April2016/V6I4-0294.pdf

CALDÓN, E.; URIBE, G.; M. LÓPEZ, D.; de OLIVEIRA, M.; KRUG WIVES, L. (2014). *Mecanismos de Anotación semántica de Contenidos en Plataformas de Redes Sociales*. [versión electrónica] Recuperado de:
http://seer.ufrgs.br/cadernosdeinformatica/article/view/v5n1p89-99

COMESAÑA, D. (2015.). *Modelo conceptual de información geográfica para la IDE – Uruguay.* Tesis de Maestría. Universidad de la República (PRODIC – FIC). [versión electrónica]. Recuperado de:
https://www.colibri.udelar.edu.uy/simple-search?query=comesa%C3%B1[a]

DOU, D., WANG, H., LIU, H. (2015). *Semantic Data Mining: A Survey of Ontology-based Approaches. Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015).* [versión electrónica]. Recuperado de: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7045128

EL OBSERVADOR (2016). *Tornado en Dolores fue "severo", según informe de Facultad de Ciencias.* Diario El Observador N°899022 [versión electrónica]. Recuperado de:
http://www.elobservador.com.uy/tornado-dolores-fue-severo-segun-informe-facultad-ciencias-n899022

FACULTAD DE CIENCIAS. Departamento de Ciencias de la Atmósfera. (2016). *Análisis del evento del 15 de abril del 2016: Tornado en la Ciudad de Dolores.* [versión electrónica]. Recuperado de:
http://www.fisica.edu.uy/~barreiro/files/Tornado_vfinal.pdf

GRUBER T. (1993). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Technical 1993*. [versión electrónica]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/downloaddoi=10.1.1.89.5775&rep=rep1&type=pdf

IDE_Uy (2014). *¿Qué es la IDE?*. Recuperado de: http://ide.uy/institucional_lista

LAMARCA LAPUENTE, M. J. (2013). *Hipertexto, el nuevo concepto de documento en la cultura de la imagen.* Tesis Doctoral. Universidad Complutense de Madrid, 2013. [versión electrónica]. Recuperado de: http://www.hipertexto.info/

LAPAN, CH. (2013). Review of QDA Miner. *Social Science Computer Review* 31(6). Recuperado de: http://ssc.sagepub.com/content/31/6/774.full.pdf

ROHRER, E., MOTZ, R., SEVERI, P. (2014). *Reasoning for ALCQM Extended with a*

Flexible Meta-Modelling Hierarchy. [versión electrónica]. JIST2014:47-62 Recuperado de: http://link.springer.com/chapter/10.1007%2F978-3-319-15615-6_4

VILAR, P. (1980). *Iniciación al vocabulario histórico*. Barcelona: Grijalbo.