Construcción de una red de ontologías sobre eventos meteorológicos a partir de periódicos históricos

Developing an ontology network about meteorological events from historical newspapers

Luis Manuel VILCHES-BLÁZQUEZ[1] 0000-0001-5799-469X

Diana COMESAÑA[2] 0000-0002-2623-7343

Lorena de Jesús ARRIETA MORENO[3] 0000-0003-3573-7427

Resumen

En la actualidad, pocos niegan el valor del contenido del periódico para comprender cuestiones relacionadas con la política, cultura y sociedad. De esta manera, la digitalización de los archivos de periódicos ha permitido rescatar artículos históricos y culturales relevantes. Sin embargo, aún no se ha sacado a la luz una infinidad de "datos menores" que se encuentran ocultos en estos periódicos. En este artículo se afrontan los desafíos para acceder y tratar los fondos de las hemerotecas nacionales de Colombia, Ecuador, México y Uruguay, que recogen noticias sobre eventos meteorológicos entre los siglos XIX-XX. Sobre estos periódicos se conforma un *corpus* de noticias que mediante lecturas técnicas y la aplicación de un proceso de bibliominería, utilizando diversas herramientas, permite iniciar la construcción de una red de ontologías. Esta red se compone de diferentes módulos (técnico, general y noticias), que son construidos utilizando diferentes enfoques (*top-down y bottom-up*) y metodologías (Methontology y NeOn), con el fin de proveer un entendimiento común y compartido de los eventos meteorológicos históricos en Latinoamérica. Por tanto, este trabajo supone un acercamiento de los fondos de las hemerotecas/bibliotecas a la *Web* Semántica.

Palabras clave: Bibliominería. Meteorología. Ontología. Prensa.

Abstract

Currently, only a few people would deny the value of newspaper content for understanding issues associated with politics, culture, and society. Therefore, the digitalization of the newspapers' archives allows retrieving outstanding historical and cultural articles. However, there is a large amount of "minor data" hidden within these newspapers. This paper addresses the challenge for accessing and dealing with the resources of National newspaper and periodicals libraries from Colombia, Ecuador, México and Uruguay, which

[1] Instituto Politécnico Nacional, Unidad Profesional Adolfo López Mateos, Centro de Investigación en Computación. Av. Juan de Dios Bátiz, s/n., Nueva Industrial Vallejo, 07738, Ciudad de México, México. Correspondencia para/*Correspondence to*: L. M. VILCHES-BLÁZQUEZ. *E-mail*: <lmvilches@cic.ipn.mx>.

[2] Universidad de la República, Facultad de Información y Comunicación, Departamento de Tratamiento y Transferencia de la Información. Montevideo, Uruguay.

[3] Pontificia Universidad Javeriana, Facultad de Ingeniería, Departamento de Ingeniería de Sistemas. Bogotá, DC, Colombia.

Este trabajo está basado en la tesis de maestría de L. J. ARRIETA MORENO, titulada "Construcción de una red de ontologías colaborativa sobre eventos asociados con el cambio climático". Pontificia Universidad Javeriana, 2017.

Recibido el 1 de septiembre del 2018, presentado el 19 de julio del 2019 y aprobado el 18 de septiembre del 2019.

Como citar este artículo/*How to cite this article*

Vilches-Blázquez, L. M.; Comesaña, D.; Arrieta Moreno, L. J. Construcción de una red de ontologías sobre eventos meteorológicos a partir de periódicos históricos. *Transinformação*, v. 32, e180077, 2020. http://dx.doi.org/10.1590/1678-9865202032e180077

collect newspapers where news about meteorological events from the XIX-XX centuries were posted. A news corpus is developed on these newspapers, which through technical readings and a bibliomining process using different tools allow building an ontology network. This network is composed of different modules (technical, general and news), which are built using different approaches (top-down and bottom-up) and methodologies (Methontology and NeOn), for providing a common and sharing understanding of the historical meteorological events in Latin America. Hence, this work entails an approach to take the newspaper and periodicals libraries to Semantic Web.

Keywords: *Bibliomining. Meteorology. Ontology. Newspaper.*

Introducción

Los periódicos se encuentran entre las fuentes más valiosas para los estudiosos interesados en investigar la opinión pública y su configuración en el transcurso del tiempo. No obstante, entre algunos investigadores hubo cierto escepticismo sobre la utilización de periódicos históricos como fuente de información, debido a su dudosa precisión y naturaleza efímera (Bingham, 2010). Sin embargo, en la actualidad, pocos niegan el valor del contenido del periódico para comprender cuestiones relacionadas con política, cultura y sociedad, así como para conocer cómo los desarrollos e ideas se percibieron y se extendieron por diferentes países (Bingham, 2010; Neudecker; Antonacopoulos, 2016). Por todo ello, los periódicos facilitan un nuevo punto de acceso a la información que se almacena en recursos históricos (Smits, 2014), proporcionando acceso a una amplia gama de fuentes de información.

Uno de los desarrollos más útiles para los historiadores modernos es la digitalización de los archivos de periódicos (Bingham, 2010), ya que permite rescatar artículos históricos y culturales relevantes y centrar el esfuerzo de análisis y descripción en ciertas publicaciones. Además, existen infinidad de "datos menores" ocultos en estos periódicos, por lo que se convierte en un desafío acceder a ellos puesto que pueden suministrar valiosa información no recogida en otros documentos.

La creciente digitalización proporciona "nuevos" desafíos técnicos, ya que frecuentemente no se aplican tecnologías de reconocimiento de texto sobre estos recursos de información (Neudecker; Antonacopoulos, 2016). Esto obstaculiza la posibilidad de hacerlos más accesibles a otras instituciones y dificulta la aplicación de tecnologías para la extracción de eventos en formas narrativas o reconstrucción y análisis de patrones (Wijfjes, 2017) que aporten mayor claridad al contexto histórico recogido en los periódicos. No obstante, refleja que el acceso a la información ha cambiado y que existen nuevos modelos de producción, distribución y consumo, pues la información digital presenta características distintas a las tradicionales (Rodríguez García, 2016).

Hoy resulta común hablar sobre la *World Wide Web*, Internet, bibliotecas digitales, repositorios digitales, XML, *Web* 2.0, *Web* Semántica, entre otros tópicos. Técnicamente, muestran cómo las tecnologías de la información protagonizan el escenario con relación al acceso a la información en la biblioteca (Rodríguez García, 2016). En este contexto de la *Web* y las bibliotecas digitales, Chowdhury y Chowdhury (2007) aclaran que las ontologías juegan un papel significativo porque sus mecanismos permiten el análisis del significado de los recursos, favoreciendo su desempeño en el proceso de acceso a la información al organizar y reunir la información heterogénea contenida en los recursos digitales.

El término ontología ha sido objeto de estudio en diferentes áreas de investigación y en varios dominios de conocimiento (Almeida, 2013). En este sentido, existen diferentes aproximaciones para definir este concepto, como puede comprobarse en Almeida (2013). En el contexto de este trabajo, se toma como referencia la adopción de la definición de Gruber (1993), donde afirma que una ontología es una especificación formal y explícita de una conceptualización compartida, entendiendo por conceptualización un modelo abstracto de la realidad en el cual se identifican los conceptos relevantes de un área; por explícita se entiende que todos sus componentes deben estar definidos explícitamente; formal se refiere al hecho de que la ontología debe ser entendible por las máquinas; y compartida refleja el hecho de que debe capturar el conocimiento consensuado por un grupo o comunidad de expertos (Studer; Benjamins; Senfel, 1998). Asimismo, aunque existen diferentes tipos de ontologías (ver:

Gómez-Pérez; Fernandez-López; Corcho, 2003 para más detalle), cuando el dominio del conocimiento involucra varias aristas, es común encontrar redes de ontologías que son una colección de ontologías relacionadas entre sí por diferentes tipos de relaciones (Suárez-Figueroa *et al*., 2012). Así, según Chowdhury; Chowdhury (2007), las ontologías capturan el conocimiento consensual y pueden ser reutilizadas y compartidas a través de aplicaciones software y por grupos de personas, facilitando la organización y el procesamiento de la información digital, otorgándole significado para permitir su uso, acceso y recuperación en el desarrollo de la *Web* Semántica (Berners-Lee; Hendler; Lassila, 2001).

En este artículo se afrontan los desafíos para acceder y tratar fondos hemerográficos que recogen noticias sobre eventos meteorológicos en la sociedad Latinoamericana entre los siglos XIX-XX. Para ello, se seleccionan periódicos históricos digitalizados, procedentes de las hemerotecas nacionales de Colombia, Ecuador, México y Uruguay. La selección de este periodo de tiempo se justifica por la consolidación de la prensa escrita (inicios del siglo XIX) y por la ausencia de registros climatológicos oficiales[4] (primera mitad del siglo XX) en los países que conforman este estudio. Estos periódicos permiten conformar un *corpus* de noticias relacionadas con eventos meteorológicos que se utilizan como base para la creación de una red de ontologías de eventos meteorológicos históricos en Latinoamérica.

Existen diversos trabajos que han propuesto ontologías relacionadas con fenómenos meteorológicos (Bally, 2004; Atemezing *et al*., 2013; Bart *et al*., 2017) y emergencias y desastres meteorológicos (Zhang *et al*., 2016; Zhong *et al*., 2017). Otros trabajos han tratado periódicos y documentos digitalizados desde un enfoque semántico (Castells *et al*., 2004; Ahonen; Hyvönen, 2009; Ambinder; Marcondes, 2013; Monteiro; Jacyntho, 2016). Sin embargo, las propuestas existentes no tratan noticias históricas sobre eventos meteorológicos ni se presentan redes de ontologías sobre dichos eventos en el periodo de tiempo considerado en el contexto latinoamericano.

La principal contribución de este artículo se centra en la construcción de una red de ontologías de eventos meteorológicos históricos en Latinoamérica a partir de la conformación de un *corpus* de noticias. Este *corpus* se trata mediante un proceso de bibliominería para iniciar la conformación de la red de ontologías. Esta red, escrita en el lenguaje de descripción de ontologías OWL (*Ontology Web Language*), se compone de diferentes módulos que recogen diversos tipos de conocimiento (espacial, temporal, meteorológico, *etc*.), con el fin de proveer un entendimiento común y compartido de los eventos meteorológicos históricos en Latinoamérica. Estos módulos también utilizan diversos estándares y glosarios internacionales asociados con los dominios de conocimiento considerados para proporcionar mayor universalidad.

Este artículo se organiza de la siguiente manera: en la segunda sección se describen las características de las fuentes de información, en la tercera sección se presenta una breve descripción de las herramientas y la metodología definida, mientras que en la cuarta sección se muestran los resultados obtenidos y, finalmente, en la quinta sección se enuncian las principales conclusiones.

Características de las fuentes de información

Las hemerotecas nacionales, adscritas a las Bibliotecas Nacionales, tienen como objetivo custodiar, organizar, preservar y difundir su acervo, conformado por publicaciones periódicas nacionales y, en ocasiones, extranjeras. Bajo este contexto, a continuación, se caracterizan los fondos hemerográficos presentes en las hemerotecas nacionales de Colombia, Ecuador, México y Uruguay. Particularmente, esta caracterización aborda la prensa que se encuentra digitalizada del periodo considerado (siglos XIX-XX); por tanto, no se someten a revisión los fondos que aún se encuentran en formato de papel.

[4] En 1951 se fundó la Organización Meteorológica Mundial, por tanto, en este trabajo se asume que, a partir de este momento, resultaría factible hallar datos cuantificados y análisis de los fenómenos meteorológicos.

Partiendo de este contexto, un análisis pormenorizado de los periódicos de los diferentes fondos hemerográficos consultados permite identificar que los periódicos presentan una total ausencia de estructura hasta la segunda mitad del siglo XIX. Además, entre estos periódicos se detecta una ausencia de secciones y noticias sin título, encontrando, en el mejor de los casos, noticias separadas, unas de otras, por guiones.

Durante el siglo XIX los periódicos adquieren estructura, apareciendo secciones estables que distinguían su contenido conforme al género de los lectores. No obstante, resulta poco frecuente hallar noticias de la vida cotidiana, especialmente, relacionadas con eventos meteorológicos, ya que se asumían como acontecimientos inevitables.

El análisis realizado permite identificar una serie de problemas adicionales en los fondos: (a) la prensa analizada presenta carencias de continuidad en el tiempo, debido a las características socio-político-económicas de estos países durante el periodo de tiempo analizado; (b) las colecciones suelen estar incompletas y en su mayoría tienen importantes problemas de conservación física, debido a la inestabilidad propia del soporte y a sus condiciones de preservación; (c) la digitalización muestra ciertas limitaciones de calidad, producto de los problemas de conservación y/o del desconocimiento o falta de recursos humanos o tecnológicos. Estas situaciones derivan en un gran número de periódicos digitalizados como imágenes y, cuando se utiliza un software de *Optical Character Recognition* ([OCR] Reconocimiento Óptico de Caracteres), los resultados presentan "ruido". Un amplio análisis de las características de las fuentes de estudio consideradas en este trabajo se puede encontrar en Comesaña, Vilches-Blázquez (2019).

Herramientas y Procedimientos

La necesidad de conocer los eventos meteorológicos históricos en Latinoamérica conduce al estudio de recursos hemerográficos por medio de la utilización de herramientas orientadas al proceso de bibliominería, aplicando procesamiento de lenguaje natural, enriquecimiento semántico y construcción de redes de ontologías. El objetivo de este trabajo se centra en construir una red de ontologías de eventos meteorológicos históricos en Latinoamérica mediante lecturas técnicas y la utilización de QDA MinerLite (Provalis Research, Montreal, 2012), los softwares libres R (Vienna, 2017), y Protégé (Stanford, 2017) y las *Application Programming Interfaces* (API) *MonkeyLearn* y Google API. Los detalles asociados con cada una de estas herramientas se describen a lo largo de la sección de Resultados.

Con respecto a la metodología empleada, esta se compone de tres actividades orientadas a la conformación del *corpus* de noticias, su tratamiento y la construcción de la red ontológica. La Figura 1 recoge las tareas asociadas a cada actividad, que son presentadas a continuación, de forma general y de forma detallada, en la sección de Resultados.

El *corpus* de noticias se conforma aplicando un vocabulario primario que permite una mejor identificación de las noticias pertinentes para este trabajo. Para obtener este vocabulario se realiza un Análisis de Dominio (AdD) basado en la propuesta de (Hjørland; Albrechtsen, 1995). Este paradigma plantea estudiar los dominios del conocimiento como comunidades discursivas, tomando en consideración el contexto psico-social y sociolingüístico de la sociología del conocimiento y de la ciencia. En este análisis se combinan los enfoques de estudios terminológicos y estudios de usuarios empíricos propuestos por Hjørland (2002).

Con la terminología obtenida se abordan los fondos hemerográficos mediante lecturas técnicas o un procesamiento de análisis cualitativo. Estas técnicas permiten conformar el *corpus* de noticias relacionadas con eventos meteorológicos que será utilizado para su tratamiento posterior.

El tratamiento del *corpus* de noticias se encara mediante un proceso de bibliominería, entendida como el uso de técnicas (inteligencia artificial, aprendizaje automático, estadística, *etc.*) que permiten sondear los datos

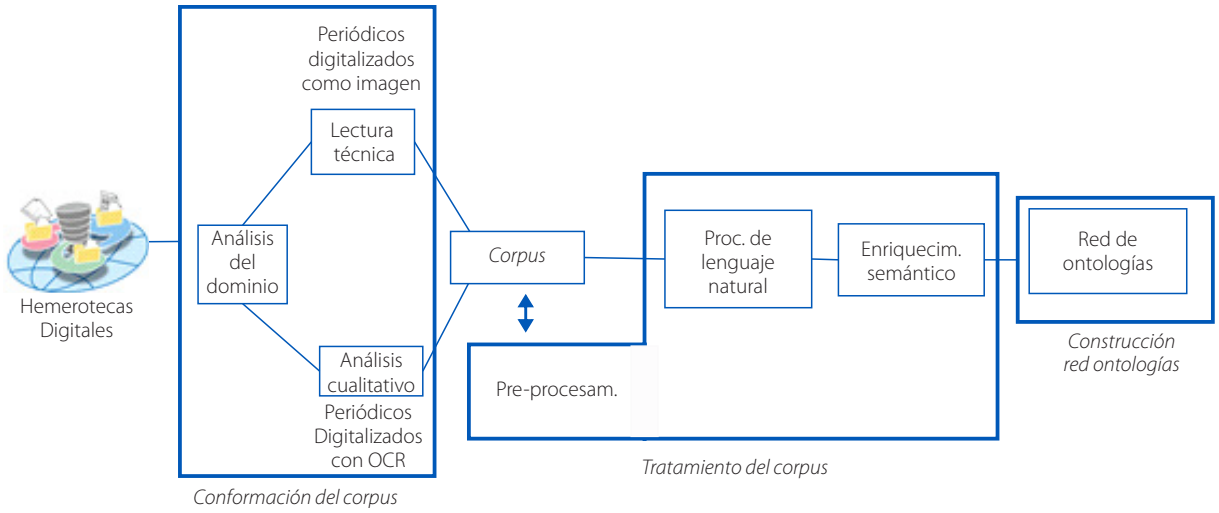

Figura 1. Flujo de actividades.

Nota: OCR: Optical Character Recognition.
Fuente: Elaborada por los autores (2018).

generados por los fondos digitalizados de las bibliotecas (Nicholson, 2004). Con este marco se realiza el tratamiento automático del *corpus*, combinando técnicas de procesamiento de lenguaje natural y enriquecimiento semántico para generar resultados que sirven de insumo al proceso de representación del conocimiento presente en el mismo.

Según Senso Leiva-Mederos, Domínguez-Velasco (2011), en la representación del conocimiento es cada vez más común el uso de las ontologías como herramientas que permiten relacionar los conceptos que aparecen dentro de un área de conocimiento. Estas especificaciones formales cuentan con varios mecanismos y fases para ser construidas. Así, para la construcción de la red de ontologías se utilizan las metodologías Methontology (Gómez-Pérez , Fernández-López, Corcho, 2003) y NeOn (Suárez-Figueroa, 2010; Suárez-Figueroa *et al.*, 2012) que permiten desarrollar diversos módulos (teórico, general y noticias), construidos conforme a diferentes enfoques (top-down y bottom-up) (Kaufmann *et al.*, 2014), que son integrados posteriormente de forma manual para conformar dicha red.

Resultados

En los fondos de las hemerotecas de Colombia, Ecuador, México y Uruguay se analizan más de 22.500 documentos digitalizados. Con un primer análisis de estos documentos se identifica que el 52% de los fondos hemerográficos contienen temas culturales, un 28% está asociado a temas políticos y un 20% presenta temas generales, donde se encuentran las noticias (sobre eventos meteorológicos) consideradas en este trabajo. Sobre estos fondos se realiza un análisis manual donde se localizan 2737 noticias relacionadas con eventos meteorológicos comprendidas entre los años 1810-2000 y se detecta que más del 60% tiene una mala digitalización y que un volumen importante de estos periódicos no presenta digitalización mediante OCR.

Esta caracterización inicial permite establecer las actividades para conformar el *corpus* de noticias sobre eventos meteorológicos del conjunto de fondos hemerográficos considerado. Para ello, se utiliza el vocabulario construido en el AdD, generado a partir de entrevistas a expertos en meteorología y complementado con el estudio del glosario de la Organización Meteorológica Mundial y de los glosarios de los Institutos Nacionales Meteorológicos de cada uno de los países intervinientes. El conocimiento obtenido permite conformar el vocabulario base para la recuperación de las noticias objeto de estudio.

La recuperación se aborda de dos formas distintas, dependiendo de las características de los periódicos (Figura 1). Así, por un lado, se procede a la realización de lecturas técnicas en aquellos fondos que están digitalizados como imagen, con el objeto de recabar información para la descripción temática y catalogación de las noticias y, por otro lado, se aplica un análisis cualitativo automático a los periódicos digitalizados con OCR mediante la herramienta QDA MinerLite (Provalis Research, Montreal, 2012), y la utilización del vocabulario obtenido en el AdD. El empleo de QDA MinerLite, permite ejecutar los módulos SimStat, para el análisis estadístico de datos y WordStat, para aplicar análisis cuantitativo de contenido y minería de texto, ya que ambos se encuentran integrados en la mencionada herramienta.

El proceso seguido permite conformar un *corpus* de más de 2500 noticias relacionadas con eventos meteorológicos contenidas en los fondos hemerográficos considerados.

Con el *corpus* de noticias conformado y teniendo en cuenta que gran parte de los periódicos no fueron digitalizados utilizando OCR, se realiza un proceso de transformación automática de las imágenes de los periódicos a texto, utilizando la herramienta Tesseract (San Francisco, 2017). Esta transformación permite realizar un tratamiento automático de las noticias y aplicar técnicas que generen insumos para la construcción de la red de ontologías.

Dicha transformación produce diversos problemas en el texto resultante, tales como: palabras sin terminar, caracteres "extraños", píxeles de la imagen representados con cadenas de números, espacios consecutivos, *etc.* Ante tales problemas, se realiza un preprocesamiento de los textos resultantes donde se lleva a cabo la supresión de signos de puntuación, acentos, caracteres "extraños" (por ejemplo: "@", "\", "-", *etc.*), secuencias de números, stopwords (por ejemplo: "de", "al", "con", *etc.*) y espacios adicionales entre palabras, mediante el software R (Vienna, 2017), con el objetivo de eliminar el ruido en los textos de las noticias.

Tras la aplicación de estas tareas se consigue tener un *corpus* con mayor número de noticias procesables y con textos depurados para realizar las tareas, asociados con el procesamiento de lenguaje natural. El Cuadro 1, recoge un ejemplo del resultado anterior y posterior a la aplicación del proceso de limpieza.

Procesamiento de lenguaje natural

Tras la realización del preprocesamiento, se llevan a cabo actividades encaminadas a emplear técnicas de procesamiento de lenguaje natural sobre los textos de las noticias consideradas. Así, en primer lugar, se procede a la tokenización del *corpus* de noticias sobre eventos meteorológicos utilizando el software R. La tokenización (Manning; Raghavan; Schütze, 2008) consiste en realizar un proceso automático para separar las palabras. Esto permite que al tokenizar el texto recogido en una noticia, por ejemplo:

"Gran congestión por lluvia. Un fenomenal trancón interrumpió el tráfico automotor"

Se obtenga la siguiente división de palabras:

Gran congestión por lluvia. Un fenomenal trancón interrumpió el tráfico automotor

Cuadro 1. Ejemplo del resultado del preprocesamiento.

Antes	Después
un recorrido especial\nro hubo demoras a l resta-'nlecer el servicio debido a la\ ndificultad de acceso a los si-\ntios de trabajo, por las inun-\n\ndaciones que c ausaron con-\nestion vehicular y por la di-\ncultad de manip las li-\n\nneas baj o el aguacero.\n\nA las 11 de la noche s e\nlogro normalizar el servicio.\n\nReco mendaciones\n\nLa Empresa de Energia\nha ce las siguientes reco-\nmendaciones.	Recorrido especial hubo demoras restalecer servicio debido dificultad acceso tios trabajo inundaciones causaron conestion vehicular dicultad manip líneas bajo a guaceroA noche selogro normalizar serviciorecomendaciones Empresa energiahace siguientes recomendaciones.

Fuente: Elaborada por los autores (2018).

http://dx.doi.org/10.1590/1678-9865202032e180077

Seguidamente, se construye una matriz de términos que permite identificar la frecuencia de repetición de cada palabra en el *corpus* mencionado. Con esta frecuencia se genera un ranking de términos que se encuentran en el *corpus* de noticias. En el ranking realizado sobre el conjunto del *corpus* destacan por su frecuencia palabras como: "personas", "calle", "ciudad", "aguas", "emergencia", "bomberos", "inundaciones", "deslizamientos", "casas", *etc*.

A continuación, se generan nubes de palabras mediante la utilización de R. Estas nubes, que no fueron limitadas por números mínimos de presencia, permiten construir una visualización de la frecuencia de palabras del *corpus*, donde los términos más presentes de las noticias reflejan mayor tamaño y viceversa, como se aprecia en el ejemplo de la Figura 2 (a).

Tras obtener diferentes formas de visualizar la frecuencia de las palabras presentes en las noticias, se procede al *Named Entity Recognition* ([NER] Reconocimiento de Entidades Nombradas), es decir, a localizar y categorizar nombres relevantes y sustantivos propios en un texto. Por ejemplo, en las noticias, los nombres de personas, organizaciones y ubicaciones suelen ser importantes (Mohit, 2014). En este contexto, se realizan las actividades de procesamiento de lenguaje natural dirigidas a la detección de *Part-of-Speech* (POS), reconocimiento de entidades y correlación.

El POS conlleva determinar el tipo de palabras que conforman el texto (sustantivos, verbos, adjetivos, *etc.*). Para ello se utiliza la librería RDRPOSTagger, basada en el software *Stanford Log* lineal *Part-of-Speech*, que permite obtener la categoría gramatical (sustantivo, verbo, adjetivo, *etc.*) e ir conformando el vocabulario sobre eventos meteorológicos. Para esto se toman las entidades obtenidas por la librería y se aplica un filtro por la categoría sustantivo (Noun), ya que los eventos meteorológicos se encuadran en ella.

En paralelo, se realiza un reconocimiento de entidades utilizando las API de *Google Cloud Natural Language* y *MonkeyLearn* para obtener los siguientes tipos de entidades: Lugar, Persona y Organización. Adicionalmente, estas API permiten obtener descripciones del recurso en Wikipedia.

Asimismo, utilizando el software R, se buscan correlaciones entre los términos extraídos, es decir, para cada término se busca qué términos aparecen cuando el término está presente en la noticia. Con esta información se

Figura 2. Nube de palabras y relaciones entre términos del *corpus*.

Fuente: Elaborada por los autores (2018).

puede obtener un grafo de asociaciones entre las palabras presentes en el *corpus* de noticias. Así, por ejemplo, se identifica que la palabra "inundaciones" se asocia con "precipitaciones", "evacuación", "cosechas", *etc*., como se muestra en la Figura 2 (b).

Enriquecimiento semántico

Tras identificar los términos asociados con eventos meteorológicos, se procede a descubrir la presencia de dichos términos en la ontología de DBpedia con el objetivo de recuperar descripciones semánticas de los términos originales recopilados. Se utiliza como ontología de referencia la propuesta por DBpedia debido a que es una ontología de dominios cruzados, creada manualmente, basada en la información más utilizada de Wikipedia. Esta ontología, en su versión 3.7, cubre más de 685 clases y contiene más de 4.2 millones de instancias, asociadas con lugares, personas, organizaciones, *etc*.

Para proceder con el enriquecimiento semántico se utiliza la librería SPARQL en el software R. Esta librería permite enviar una consulta semántica (SPARQL) (SPARQL Working Group, 2013) a la ontología DBpedia para descubrir la presencia de los términos recopilados y, en el caso de encontrar elementos similares, retorna una URI (*Uniform Resource Identifier*)[5] con la descripción del concepto. En múltiples ocasiones, el término consultado muestra relaciones con otros términos a través de *rdfs:seeAlso*, como el caso de Borrasca.

Para enriquecer semánticamente los términos descubiertos en las noticias del *corpus* se desarrolla un proceso iterativo que permite navegar por los conceptos de DBpedia y sus términos relacionados y construir un modelo conceptual que sirve de insumo para la construcción de la red de ontologías.

Construcción de la red de ontologías

En esta sección se describe cómo se construyen los diferentes módulos (teórico, general y noticias) que conforman la red de ontologías de eventos meteorológicos históricos, utilizando las metodologías "*Methontology*" (Gómez-Pérez; Fernández-López; Corcho, 2003) y NeOn (Suárez-Figueroa *et al.*, 2012; Suárez-Figueroa, 2010), adoptando diferentes enfoques (*top-down* y *bottom-up*) y usando el software Protégé (Stanford, 2017).

Módulo de conocimiento técnico

Este módulo adopta la metodología "*Methontology"* y sigue un enfoque *top-down*, donde toma como referencia modelos y conocimiento de alto nivel sobre el dominio (Kaufmann *et al*., 2014). Conforme a la mencionada metodología y al vocabulario obtenido durante el AdD, se construye un glosario que recoge el nombre del término, sinónimos, descripción y tipología (concepto, relación o atributo). El Cuadro 2 muestra un extracto del glosario conformado para la creación de este módulo de la red.

A partir del glosario y de la jerarquía que implícitamente conlleva, se construye un árbol terminológico utilizando Protégé (Stanford, 2017), cuyo elemento principal es "Fenómeno meteorológico". Además, se establecen las clases "Efectos detectados" y "Registro documental", junto con las relaciones "produce" y "documentado en" y sus inversas "producido por" y "documenta", que permiten relacionar al "Evento meteorológico" con sus resultados en el medio y con la noticia que los registra.

Módulo de conocimiento general

Este módulo se construye utilizando la metodología NeOn, que presenta diversos escenarios que se pueden combinar de maneras diferentes y flexibles. En este caso se utilizan los escenarios 1, 3 y 4, centrados en

[5] Se define como una cadena de caracteres utilizada para identificar recursos en la Web Semántica.

Cuadro 2. Extracto del glosario del módulo de conocimiento técnico.

Nombre	Sinónimo	Descripción	Tipología
Agua nieve	Cellisca	Tipo de precipitación en la que el agua presenta dos estados teniéndose una mezcla de agua congelada y agua líquida.	Concepto
Produce		Un evento meteorológico produce cierto efecto detectable.	Relación
Velocidad del viento		Aire en movimiento, medido en km/h.	Atributo

Fuente: Elaborada por los autores (2018).

la especificación e implementación, reutilización y la reutilización y reingeniería de recursos ontológicos, respectivamente. Una descripción detallada de estos escenarios se recoge en Suárez-Figueroa (2010) y Suárez-Figueroa *et al.* (2012).

La primera tarea que se aborda en el marco de estos escenarios es la búsqueda de ontologías a reutilizar. Para ello se consideran los estándares y/o recomendaciones existentes en los diferentes dominios abordados (geoespacial, temporal, *provenance*, personas y organizaciones). La selección de ontologías conforme con los estándares de los diferentes dominios considerados permite construir un módulo que adopta las buenas prácticas de las diversas áreas de conocimiento.

Así, en el contexto del escenario 3, se reutiliza la ontología asociada con el estándar GeoSPARQL (Open Geospatial Consortium, 2012), propuesto por el *Open Geospatial Consortium* (OGC), que desarrolla un vocabulario diseñado para representar datos geoespaciales en la *Web* Semántica. En esta ontología se definen las clases "*Feature*" y "*Geometry*" utilizadas para modelar cualquier objeto que tenga una representación espacial. Además, resulta relevante para este módulo la relación "*hasGeometry*" que permite la conexión entre las clases mencionadas.

Con respecto al escenario 4, se procede a la reutilización y reingeniería de diversas ontologías, entendiéndose como el proceso de recuperar y transformar un modelo conceptual existente e implementado en una ontología en un nuevo modelo, más correcto y completo, conforme a las necesidades del contexto (Suárez-Figueroa, 2010 y Suárez-Figueroa *et al.*, 2012). Las ontologías sobre las que se lleva a cabo este escenario son *Event, Time Ontology*, FOAF (*Friend of a Friend*) y PROV-O. Sobre estas se realizan diversos tipos de modificaciones, tales como: creación o restructuración de jerarquías, incremento del nivel de detalle, redefinición de clases y unificación de nombrado.

(1) La ontología *Event* se centra en la noción de evento, visto como la forma en que los agentes cognitivos clasifican las regiones arbitrarias de tiempo/espacio. En este trabajo, al presentar esta ontología un nivel de abstracción superior se incorpora como subclase de *Event* el concepto "Evento meteorológico". Además, dependiendo de las características de velocidad, intensidad, dirección, *etc*., el "Evento meteorológico" puede tener un determinado "Efecto" (consecuencia).

(2) *Time ontology in OWL* es un vocabulario del W3C que proporciona conceptos para describir las propiedades temporales en el mundo. Sobre esta ontología se realiza una especialización en la clase *MonthOfYear*, adicionando las instancias correspondientes a los meses del año, ya que en el *corpus* aparecen eventos asociados a meses específicos.

(3) La ontología FOAF (*Friend of a Friend*) vincula personas e información a través de la *Web*. Por tanto, esta ontología se utiliza para identificar información de los afectados y organizaciones involucradas en la emergencia de un determinado evento meteorológico.

(4) La ontología PROV-O proporciona elementos para representar e intercambiar información de procedencia generada en diferentes contextos. En este módulo se utiliza para modelar la procedencia de las noticias del *corpus*.

Módulo de conocimiento extraído de las noticias

Este módulo de la red se construye utilizando la metodología NeOn, concretamente los escenarios 1 y 2 relacionados con la especificación e implementación y la reutilización y reingeniería de recursos no ontológicos (Suárez-Figueroa *et al.*, 2012; Suárez-Figueroa, 2010). Además, se sigue una aproximación bottom-up (Kaufmann *et al*., 2014), por lo que se va de lo particular (*corpus* de noticias) a lo general (conocimiento sobre meteorología). A partir del *corpus* se realiza su tratamiento para extraer el conocimiento relevante. La aplicación de técnicas de procesamiento de lenguaje natural y el enriquecimiento semántico realizado, descritos con anterioridad, permiten construir una taxonomía de eventos meteorológicos presentes en el *corpus*.

Partiendo de esta taxonomía se construye este módulo de la red de ontologías con los términos extraídos y el asesoramiento de expertos del dominio. Para ello, se aplica el escenario 1 de la mencionada metodología, que recoge las actividades centrales (conceptualización, formalización e implementación) que deben realizarse en cualquier desarrollo de ontología.

Este módulo tiene relaciones definidas dentro de sus elementos. A modo de ejemplo, la clase "Viento" está asociada con una zona de presión y tiene una "dirección" y "velocidad" determinada. Al ser un "Evento meteorológico" provoca un "Efecto" en "Tierra" o "Mar". Además, tiene asociadas unas subclases que obedecen a la tipología mencionada en las noticias ("Borrasca", "Brisa", "Ciclón", "Tormenta", "Temporal", *etc.*).

Construcción de la red de ontologías

Tras el desarrollo de los diferentes módulos (técnico, general y noticias) se conforma la red de ontologías sobre eventos meteorológicos históricos. Para ello, es necesario definir correspondencias semánticas y realizar procesos de reingeniería entre los diferentes módulos, que se llevan a cabo de forma manual por medio de los siguientes pasos:

(1) *Especificación de correspondencias de taxonomía*. Se definen las equivalencias semánticas entre los elementos que componen las diferentes ontologías desarrolladas en los módulos. Por ejemplo, se establecen equivalencias entre la clase *"Time"* de la ontología *Event* con "*TemporalEntity"*, que corresponde a un intervalo de tiempo o instante en la ontología *Time* y se establece la equivalencia de la relación *"time"* (*Event*) con *"hastime"* (*Time*).

(2) *Establecimiento de relaciones*. Se establecen relaciones entre los diferentes conceptos de los distintos módulos generados con el objetivo de conectarlos y conformar la red de ontologías. Por ejemplo, un evento meteorológico se produce en un lugar determinado y tiene asociada una geometría. Para definir esto, la clase "*Evento meteorológico*" se relaciona con la clase *"Geometry"* a través de la relación "*hasGeometry"*, ambos elementos asociados a la ontología GeoSPARQL, presente en el módulo de conocimiento general.

(3) *Reingeniería*. Al conectar los diferentes módulos de la red se producen procesos de reingeniería de conocimiento que permiten construir la red definitiva. Por ejemplo, se realiza la reestructuración de taxonomías, tomando la clase "*Event"* y ubicándola dentro de la clase "*Feature"*, debido a que el "Evento meteorológico" tiene una ubicación geográfica y la clase *"Feature"* es una superclase de todo objeto que presente un componente geográfico.

La Figura 3 muestra una visión de alto nivel de la red de ontologías conformada por los diferentes módulos desarrollados, disponible en lenguaje OWL para su consulta y explotación.

Esta red de ontologías fue diseñada en el contexto del proyecto de investigación "Compartiendo la historia escondida del cambio climático en Latinoamérica a través de las Tecnologías de la Información y las Comunicaciones", para

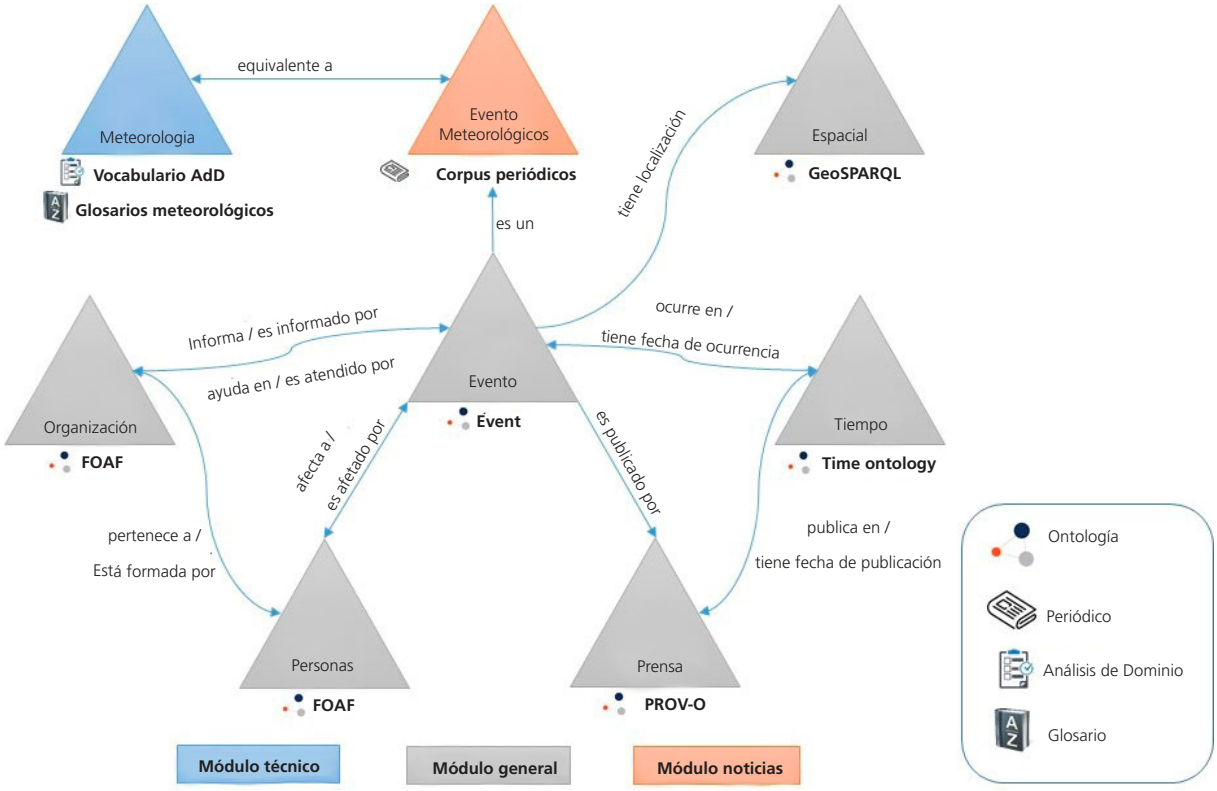

Figura 3. Red de ontologías sobre eventos meteorológicos históricos.

Fuente: Elaborada por los autores (2018).

ser utilizada en un proceso de integración de *corpus* de noticias recopilados de las hemerotecas nacionales de Colombia, Ecuador, México y Uruguay. Esto permite reunir, organizar y consultar la información de las noticias que conforman el mencionado *corpus* de una manera integrada e interoperable semánticamente. Adicionalmente, esta red de ontologías puede ser la base para la implementación de aplicaciones relacionadas con eventos meteorológicos, así como para integrar *corpus* de noticias de otros países o de otras temáticas que tengan relación con este tipo de eventos.

Conclusión

Este trabajo presenta un proceso de construcción de una red de ontologías de eventos meteorológicos históricos en Latinoamérica utilizando como base un *corpus* de noticias, comprendidas entre los siglos XIX-XX, proveniente de las hemerotecas nacionales de Colombia, Ecuador, México y Uruguay.

La conformación de este *corpus* permitió identificar y abordar diversos problemas asociados con las características de las fuentes y con aquellos producidos por una deficiente digitalización de los fondos hemerográficos. Sobre este *corpus* se realizó un proceso de bibliominería, aplicando técnicas de procesamiento de lenguaje natural y enriquecimiento semántico que permitió generar insumos para el proceso de construcción de la red de ontologías. Esta red se compone de diferentes módulos (técnico, general y noticias), donde se utilizan diversos estándares asociados con los dominios de conocimiento considerados.

En definitiva, la red construida permite reunir y organizar la información heterogénea sobre eventos meteorológicos históricos contenida en las hemerotecas, otorgándole significado para permitir su uso, acceso y recuperación en

el desarrollo de la *Web* Semántica. Por tanto, este trabajo supone un acercamiento de los fondos de las hemerotecas/bibliotecas a la *Web* Semántica. Esta relación resulta esencial para organizar las bibliotecas, y sus recursos, como una red de datos (Balaji, *et al*., 2018) y es producto del proceso de evolución, acorde con la aparición de nuevos paradigmas, tendencias e influencias de orden tecnológico.

Agradecimientos

Este trabajo fue desarrollado en el marco del proyecto "Compartiendo la historia escondida del cambio climático en Latinoamérica a través de las Tecnologías de la Información y las Comunicaciones", auspiciado por el Instituto Panamericano de Geografía e Historia, PAT 2017, y "Caracterización de la calidad de los datos generados por iniciativas de crowdsourcing meteorológico", SIP 20195556.

Colaboradores

L. M. Vilches-Blázquez y D. Comesaña colaboraron en la concepción, la recolección y el análisis de datos y en la redacción y revisión del artículo. L. J. Arrieta Moreno colaboró en el análisis de los datos y la revisión del artículo.

Referencias

Ahonen, E.; Hyvönen, E. Publishing historical texts on the Semantic Web: a case study. *In*: International Conference on Semantic Computing, 2009, Berkeley, California. *Proceedings* […]. Berkeley: IEEE Xplore, 2009. Doi: http://dx.doi.org/10.1109/ICSC.2009.9

Almeida, M. B. D. Revisiting ontologies: a necessary clarification. *Journal of the American Society for Information Science and Technology*, v. 64, n. 8, p. 1683-1693, 2013.

Ambinder, D. M. M.; Marcondes, C. H. Novas experiências para apresentação, acesso e leitura de artigos científicos digitais na web. *Transinformação*, v. 25, n. 3, p. 195-201, 2013.

Atemezing, G. *et al*. Transforming meteorological data into Linked Data. *Semantic Web*, v. 4, n. 3, p. 285-290, 2013. Doi: http://dx.doi.org/10.3233/SW-120089

Balaji, B. P. *et al*. An integrative review of Web 3.0 in academic libraries. *Library Hi Tech*, v. 35, n. 4, p. 13-17, 2018. Doi: http://dx.doi.org/10.1108/LHTN-12-2017-0092

Bally, J. *et al*. Developing an ontology for the meteorological forecasting process: Decision support in an uncertain and complex world. *In*: IFIP TC8/WG8.6 Working Conference on IT Innovation for Adaptability and Competitiveness, 7, 2004, Leixlip, Ireland. *Proceedings* […]. New York: Springer, 2004.

Bart, A. *et al*. Ontological description of meteorological and climate data collections. *In*: Internacional Conference on Data Analytics and Management in Data Intensive Domains, Russian Conference on Digital Libraries, 19, 2017, Moscow, Russia. *Proceedings* […]. Moscow: CEUR, 2017. v. 2022, p. 266-272.

Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Scientific American*, v. 284, n. 5, p. 28-37, 2001. Doi: http://dx.doi.org/10.1038/scientificamerican0501-34

Bingham, A. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, v.21, n.2, p.225-231, 2010. Doi: http://dx.doi.org/10.1093/tcbh/hwq007

Castells, P. *et al*. Neptuno: semantic web technologies for a digital newspaper archive. *In*: European Semantic Web Symposium, 2004, Heraklion, Greece; Bussler C. J.; Davies J.; Fensel D.; Studer R. (eds.), The Semantic Web: research and Applications. Proceedings […]. New York: Springer, 2004. Doi: http://dx.doi.org/10.1007/978-3-540-25956-5_31

Chowdhury, G. G.; Chowdhury, S. *Organization information*: from the shelf to the web. London: Facet Publishing, 2007. Doi: http://dx.doi.org/10.1080/14649055.2007.10766175

Comesaña, D.; Vilches-Blázquez, L. M. Un estudio de la prensa latinoamericana entre los siglos XIX y XX con un enfoque en eventos meteorológicos. *Revista de Historia de América*, n. 156, p. 29-59, 2019.

Gómez-Pérez, A.; Fernández-López, M.; Corcho, O. *Ontological engineering*: with examples from the areas of knowledge management, e-Commerce and the semantic web. London: Springer-Verlag, 2003.

Gruber, T. Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, v. 43, p. 907-228, 1993. Doi: http://dx.doi.org/10.1006/ijhc.1995.1081

Hjørland, B.; Albrechtsen, H. Toward a new horizon in information science: domain-analysis. *Journal of the Association for Information Science and Technology*, v. 46, n. 6, p. 400-425, 1995. Doi: http://dx.doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hjørland, B. Domain analysis in information science: eleven approaches – traditional as well as innovative. *Journal of Documentation*, v. 58, n. 4, p. 422-462, 2002. Doi: http://dx.doi.org/10.1108/00220410210431136

Kaufmann, M. *et al.* Combining bottom-up and top-down generation of interactive knowledge maps for enterprise search. *Knowledge Science, Engineering and Management*, v. 8793, p. 186-197, 2014. Doi: http://dx.doi.org/10.1007/978-3-319-12096-6_17

Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to information retrieval*. Cambridge: Cambridge University Press, 2008. Doi: http://dx.doi.org/10.1017/CBO9780511809071

Mohit, B. Named entity recognition. *In*: Zitouni, I. (ed.). *Natural language processing of semitic languages*: theory and applications of natural language processing. New York: Springer, 2014. Doi: http://dx.doi.org/10.1007/978-3-642-45 358-8

Monteiro, L. L. P.; Jacyntho, M. D. A. Use of Linked Data principles for semantic management of scanned documents. *Transinformação*, v. 28, n. 2, p. 241-251, 2016. Doi: http://dx.doi.org/10.1590/2318-08892016000200010

Neudecker, C.; Antonacopoulos, A. Making Europe's historical newspapers searchable. *In*: IAPR Workshop on Document Analysis Systems, 12, 2016, Santorini, Greece. *Proceedings* […]. Santorini: IEEE Xplore, 2016. p. 405-410. Doi: http://dx.doi.org/10.1109/DAS.2016.83

Nicholson, S. The bibliomining process: data warehousing and data mining for library decision-making. *Transinformação*, v. 16, n. 3, p. 253-261, 2004. Doi: http://dx.doi.org/10.1590/S0103-37862004000300005

Open Geospatial Consortium. *GeoSPARQL*: a geographic query language for RDF data. Wayland (MA): OGC, 2012. Available from: http://portal.opengeospatial.org/files/?artifact_id=44722. Acess: July 19 2019.

Protégé Version 5.2.0. Stanford: Protégé, 2017. Available from: https://protege.stanford.edu/. Cited: July 19 2019.

R. The R Project for Statistical Computing. Version R 3.3.3. Vienna: R, 2017. Available from:- https://www.r-project.org/. Cited: July 19 2019.

Rodríguez García, A. A. Las nuevas pautas para el acceso a la información. *Investigación Bibliotecológica*, v. 30, n. 69 p. 121-141, 2016. Doi: http://dx.doi.org/10.1016/j.ibbai.2016.04.015

Senso, J. A.; Leiva-Mederos, A. A.; Domínguez-Velasco, S. E. Modelo para la evaluación de ontologías: aplicación en Onto-Satcol. *Revista Española de Documentación Científica*, v. 34, n. 3, p. 334-356, 2011. Doi: http://dx.doi.org/10.3989/redc.2011.3.788

Smits, T. Problems and possibilities of digital newspaper and periodical archives. *Tijdschrift voor Tijdschriftstudies*, v. 36 p. 139-146, 2014.

SPARQL Working Group. *SPARQL 1.1 Overview*. Cambridge: W3C, 2013. Available from: https://www.w3.org/TR/sparql11-overview/. Acess: July 19 2019.

Studer, R.; Benjamins, R.; Fensel, D. Knowledge engineering: principles and methods. *Data & Knowledge Engineering*. v. 25, n. 1-2, p. 161-197, 1998. Doi: http://dx.doi.org/10.1016/S0169-023X(97)00056-6

Suárez-Figueroa, M. C. *NeOn methodology for building ontology networks*: specification, scheduling and reuse. Thesis (Doctoral), Universidad Politécnica de Madrid, Facultad de Informática, Madrid, 2010. Available from: http://oa.upm.es/3879/. Access 19 jul. 2019.

Suárez-Figueroa, M. C. *et al*. (ed.) Ontology engineering in a networked world. New York: Springer, 2012. Doi: http://dx.doi.org/10.1007/978-3-642-24794-1

Tesseract. Tesseract Open Source OCR Engine. San Francisco: Tesseract, 2017. Available from: https://github.com/tesseract-ocr/tesseract. Cited: July 19 2019.

Wijfjes, H. Digital humanities and historical newspaper research. Tijdschrift voor Mediageschiedenis. v. 20, n. 1, p. 4-24, 2017.

Zhang, F. *et al*. Ontology-based representation of meteorological disaster system and its application in emergency management: illustration with a simulation case study of comprehensive risk assessment. *Kybernetes*, v. 45, n. 5, p. 798-814, 2016. Doi: http://dx.doi.org/10.1108/K-10-2014-0205

Zhong, S. *et al*. A geo-ontology-based approach to decision-making in emergency management of meteorological disasters. *Natural Hazards*, v. 89, n. 2, p. 531-554, 2017. Doi: http://dx.doi.org/10.1007/s1106