UNA APLICACIÓN DE LAS TECNOLOGÍAS DE LA WEB SEMÁNTICA AL ESTUDIO DEL PERIODISMO DE LA EDAD MODERNA

USING SEMANTIC WEB TECHNOLOGIES TO STUDY EARLY MODERN JOURNALISM

Francisco Baena Sánchez
Universidad de Sevilla
frbaena@us.es

Antonia M. Chávez-González
Universidad de Sevilla
tchavez@us.es

Resumen

En el presente trabajo explicamos el diseño de la ontología *Early Modern News*, fruto del trabajo interdisciplinar entre las áreas de la Historia del Periodismo y la Ingeniería del Conocimiento. Dicha ontología conceptualiza de manera formalizada el dominio relativo al periodismo de la Edad Moderna. Se trata de una herramienta tecnológica básica para la posterior creación de un portal semántico que promoverá interesantes avances en el estudio de la primera prensa de la Historia. Con ello, el humanista no sólo podrá acceder a la información visible en las ediciones digitalizadas de los documentos, sino también al conocimiento que las técnicas de razonamiento aplicadas revelarán.

Palabras clave: Ontologías. Web Semántica. Colaboración interdisciplinar. Historia del Periodismo. Edad Moderna.

Abstract

This article is devoted to the description and design of *Early Modern News* Ontology as a result of the interdisciplinary efforts in the two fields of research such as the Journalism History and the Knowledge Engineering. The ontology is a formal conceptualization of the domain of the Journalism on the 17th century. The tool is a basic technology for the subsequent implementation of a semantic portal which will produce interesting advances on the study of the first journalism insights. On this way, humanist researchers will get access not only to visible information on digital editions of the documents, but to the revealed knowledge by applying reasoning techniques.

Keywords: Ontologies. Semantic Web. Interdisciplinarity. History of Journalism. Modern Age.

1. EL NACIMIENTO DEL PERIODISMO EN EUROPA[1]

La prensa actual no difiere tanto de la que se imprimía hace cuatrocientos años. Las bases históricas del periodismo contemporáneo aparecen en la Edad Moderna, un período en el que ya se observan elementos inherentes a la actividad periodística tales como la circulación de las noticias, la dependencia de las fuentes, la periodicidad, la relación con el poder político, el diseño y el estilo, géneros periodísticos tan genuinos como la noticia y el reportaje, la libertad de imprenta, la cultura profesional e incluso el sensacionalismo, que revela ya esa necesidad que tienen los medios de alcanzar la viabilidad económica adaptándose a los gustos e intereses de un público en constante crecimiento, pese al bajísimo nivel de alfabetización. En palabras de Guillamet (2012: 263), "el período que transcurre entre los siglos XVI y XVIII es algo más que el inicio de las publicaciones periódicas, es la epifanía del propio sistema de medios que hoy conocemos, con todos sus factores y condicionantes"[2].

[1] Este trabajo ha sido financiado por el Proyecto del Ministerio de Economía y Competitividad TIN2013-41086-P, cofinanciado con Fondos FEDER; asimismo, se ha realizado en el marco del proyecto *Biblioteca Digital Siglo de Oro 5* (BIDISO 5), con referencia: FFI2015-65779-P, financiado por el Ministerio de Economía y Competitividad del Gobierno de España y el Fondo Europeo de Desarrollo Regional (FEDER) desde el 1-01-2016 hasta el 31-12-2019. El presente estudio forma parte también de la investigación desarrollada por IBEMNEWS (*Iberian Early Modern News*), un grupo que asocia a académicos de varias universidades españolas y que pretende estudiar la primera etapa del periodismo español, en primer lugar, mediante la recuperación del valioso patrimonio documental que representan los fondos de gacetas seriadas o periódicas publicadas en la Península Ibérica en la primera mitad del siglo XVII; y, en segundo lugar, mediante la aplicación de diferentes tecnologías y procedimientos propios de las Humanidades Digitales, tales como la digitalización, la creación de una ontología o la codificación XML-TEI de los textos acopiados.

[2] Las primeras publicaciones que tienen periodicidad –en este caso, semanal– reciben el nombre de *coranto* o *gaceta*. Las más antiguas que se conservan aparecen en tierras alemanas –Estrasburgo (1605), Wolfenbüttel (1609)– y holandesas –Ámsterdam (1618)–. Este modelo centroeuropeo se exporta a Inglaterra, Francia y España antes de 1620.

Uno de esos condicionantes es precisamente la necesaria connivencia entre política y mercado, que, según Conboy (2004: 46), explica la eclosión de la actividad periodística como una "combination of profit, politics and curiosity":

> *Todas las características del periódico-tipo destinado a cristalizar en Europa pueden entenderse como una necesidad de su existencia en el mercado: la periodicidad, la conjunción recurrente entre información y opinión, entre instrucción y entretenimiento, entre moldes cultos y motivos populares; incluso la impresión decepcionante que acarrea al lector, que desde los primeros tiempos lamenta la falta de veracidad implícita en la naturaleza del novedoso discurso periodístico* (Espejo, 2012: 124).

Nuestro conocimiento acerca de los orígenes históricos del Periodismo se ha visto ampliado y mejorado en las últimas décadas, gracias a una serie de estudios notables que explican la aparición de la prensa en Europa, tanto en clave nacional (para el caso español, Guillamet, 2003; Díaz Noci y Hoyo, 2003; Espejo, 2008: 243-267) como en clave paneuropea (Pettegree, 2014; Ettinghausen, 2015), sin olvidar las redes sociales y comunicativas en las que ésta se insertaba (Koopmans, 2005). El primer periodismo europeo es, por tanto, un fenómeno transnacional que, pese a surgir en pleno auge de los nacionalismos modernos, es capaz de sortear las fronteras políticas, religiosas e idiomáticas, que tan férreamente, comenzaban a dibujarse por entonces (Raymond, 2012: 177-206).

De acuerdo con la interpretación más habitual, el periodismo nace en Europa a comienzos de la Edad Moderna vinculado a una serie de factores históricos determinantes, entre los que sobresalen la emergencia de la burguesía, la consolidación de los Estados absolutistas o la implantación de la imprenta. Según esta última línea de explicación, por ejemplo, desde finales del siglo XV se publican impresos con una función claramente informativa por toda Europa[3]. En cualquier caso, más allá de esos factores explicativos o de ciertas singularidades nacionales, la eclosión y consolidación del *auténtico* periodismo europeo no se produce hasta 1618, una fecha-hito marcada por la fiebre informativa que desata el estallido de la Guerra de los Treinta Años (Espejo, 2012: 103-126)[4].

[3] Estos formatos pioneros, antecedentes de los periódicos, reciben una denominación diferente según el estado europeo: *news pamphlets* en Inglaterra, *neue Zeitungen* en Alemania, *ocasionnels* en Francia y *relaciones* en España, Italia y Portugal.

[4] Las noticias circulaban desde Italia y Europa central hasta Holanda y Alemania, potencias económicas del momento, y desde allí, en una verdadera explosión informativa, hacia el resto de Europa.

RELACION
CERTISSIMA
DE LA FELICISSIMA
Victoria que ha tenido don Gonçalo de Cordoua en veinte y nueue de Agosto deste año, en los Estados de Flandes, contra hereges.

GAZETA
ROMANA, Y RELACION
general, de auisos de todos los Reynos y Prouincias del mundo.

CON LICENCIA.

Figura 1. Primera plana de una relación y una gaceta, ambas publicadas en Sevilla durante el primer tercio del siglo XVII.

Las relaciones de sucesos y las gacetas constituyen los dos formatos principales del periodismo impreso de la Edad Moderna (véase Figura 1). Ambos representan el primer periodismo de la Historia, ese periodismo que columpiaba sus intereses entre las noticias serias –la alta política, la guerra, las ceremonias cortesanas y religiosas– y las populares, esos casos espantosos tratados en muchas ocasiones con ribetes morbosos que sorprenden por su proximidad con el sensacionalismo de nuestros días. Asimismo, relaciones y gacetas estaban sometidas a los controles institucionales -licencias, privilegios, censura- que impedían casi sin excepción que se difundiese información poco grata a los ojos de las autoridades. Hasta aquí las similitudes, porque tanto las condiciones de producción y difusión, como la modalidad del discurso, eran distintas para uno y otro formato (Espejo, 2012: 103-126). Las relaciones se publicaban ocasionalmente e informaban habitualmente de una sola noticia en profundidad, siguiendo el canon del discurso historiográfico y literario. En cambio, las gacetas eran de periodicidad semanal y recopilaban varias noticias breves, de uno o dos párrafos de extensión, empleando una nueva modalidad de discurso, despojado de recursos literarios, al que podríamos denominar ya *discurso periodístico*[5].

[5] A efectos de esta investigación, ambos formatos –relaciones de sucesos y gacetas– serán considerados e incluidos en la ontología.

2. HACIA UNA HISTORIA DIGITAL DEL PERIODISMO DE LA EDAD MODERNA

2.1. Antecedentes

La digitalización del patrimonio hemerográfico conservado en las bibliotecas de todo el mundo ha puesto a disposición de la comunidad científica, en las últimas décadas, un acervo de miles de impresos informativos, en cada una de las lenguas europeas de la Edad Moderna. En español, el proyecto *Biblioteca Digital Siglo de Oro* (BIDISO), impulsado por el Seminario Interdisciplinar para el Estudio de la Literatura Áurea Española (SIELAE) de la Universidade da Coruña, ofrece en su *Catálogo y Biblioteca Digital de Relaciones de Sucesos*[6] los registros de más de 6.000 ediciones, de las que más de 2.000 cuentan con reproducción digital (véase Figura 2). Fuera de España, *The Iberian Book Project* (IB), auspiciado por el Centre for the History of the Media de la University College Dublin, permite consultar registros fiables y enlaces a las copias digitalizadas a través del Universal Short Title Catalogue[7].

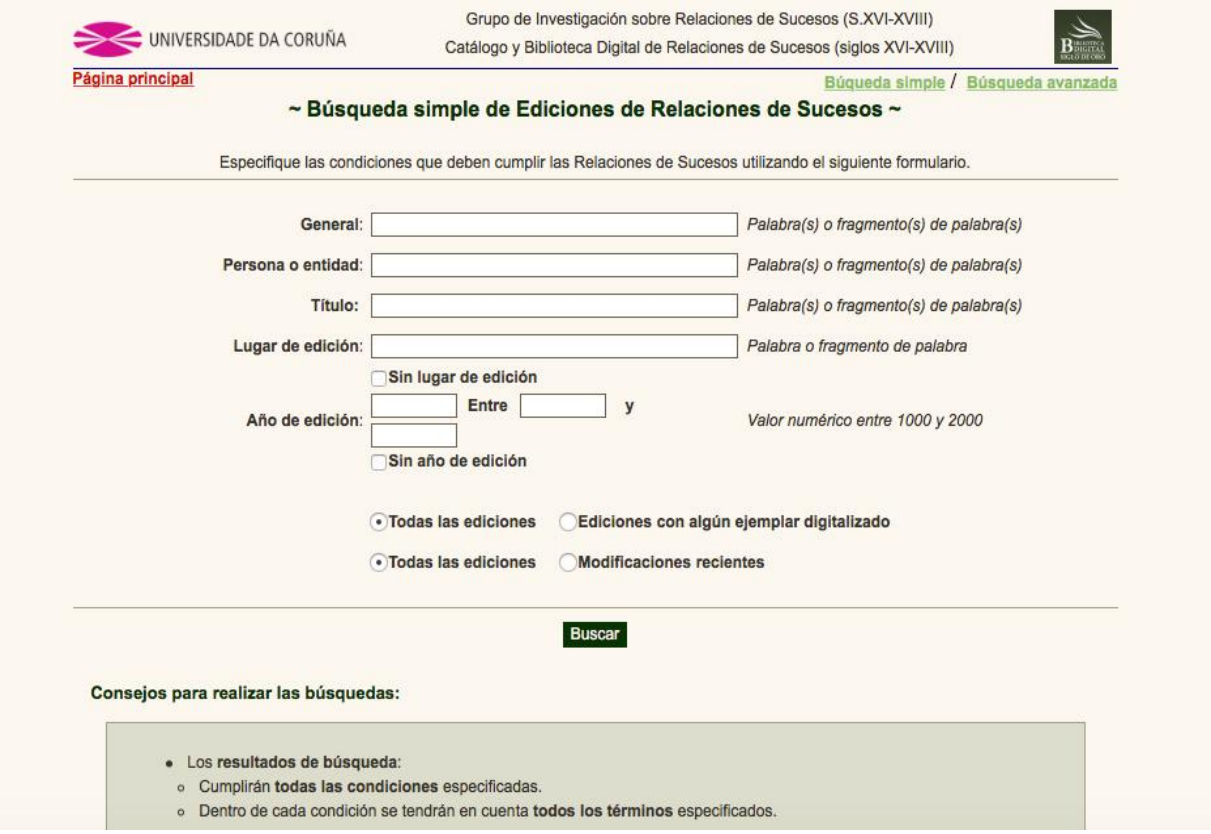

Figura 2. Página de consulta del Catálogo y Biblioteca Digital de Relaciones de Sucesos, fruto de la actividad desarrollada por el proyecto BIDISO.

A pesar de que es probable que sea más lo perdido que lo conservado, contamos ya con ejemplos suficientes para abordar la revisión de la historia de los orígenes del periodismo

[6] Accesible desde http://www.bidiso.es/RelacionesSucesosBusqueda/.
[7] Accesible desde http://ustc.ac.uk/index.php.

en la Península Ibérica. Sin embargo, el investigador de este campo se encuentra con dificultades metodológicas específicas. Hasta ahora, por ejemplo, relaciones y gacetas se encuentran parcialmente catalogadas en bases de datos bibliográficas, donde no cuentan con una descripción precisa en función de su condición de impresos informativos, por lo que resulta necesario proceder a una nueva catalogación que atienda a criterios propiamente periodísticos. A este déficit se suman otras dos dificultades: la primera, el acceso a numerosos impresos conservados obliga a trabajar con grandes series de documentos; la segunda, que, según nos consta, no se ha conservado todo lo que se imprimió durante el periodo. Por tanto, todo ello nos advierte del riesgo de extraer conclusiones erróneas a partir de la observación del corpus disponible.

En el campo académico de la Historia del Periodismo, los especialistas de los diferentes estados europeos han centrado sus esfuerzos, a lo largo de las últimas décadas, en el conocimiento del periodismo de los siglos XVI y XVII (Popkin, 2005: 1-27), hasta entonces poco conocido debido sobre todo a la insuficiencia de los recursos bibliográficos disponibles. Aunque existen algunas aportaciones recientes en el campo de la reedición, el esfuerzo de la investigación contemporánea se ha centrado en la recuperación hemerográfica a través de bibliotecas y hemerotecas digitales.

Para ello, necesitamos poner todas las herramientas de las Humanidades Digitales al alcance y al servicio de la investigación histórica, y en concreto del área a que nos referimos, que se ha denominado *Digital History* (Cohen y Rosenzweig, 2005), así como contemplar aspectos de preservación digital de documentos antiguos (Higgs, 1998; Deegan y Tanner, 2006):

> *Projects that collect and present historical materials online assume a special responsibility for the long-term survival and availability of those materials. Online historians must therefore think prospectively, creatively, and strategically about issues of digital preservation and access* (Cohen y Rosenzweig, 2005).

La *Digital History* es un ejemplo de cómo las Humanidades Digitales han ampliado el alcance y el potencial de las Humanidades tradicionales, planteando renovados desafíos y oportunidades. Precisamente, junto a su carácter heterogéneo e interdisciplinar, el denominador común que vertebra este *new collective singular* reside en el énfasis que se pone sobre hacer, conectar, interpretar y colaborar (Burdick *et al*., 2012). Como afirma Ramsay (2011): “Digital Humanities is about building things. […] If you are not making anything, you are not a digital humanist”; o, tal y como apostillan Cohen y Rosenzweig (2005) a propósito de la *Digital History*: “to preserve, of course, you must first create”.

Los historiadores del primer periodismo español deben afrontar, ante todo, la catalogación del ingente material, en buena parte, aún desconocido o poco sistematizado y

explotado, que puede desembocar no sólo en catálogos de tipo censal (Domínguez Guzmán, 1992). Además, urge la construcción de corpus marcados digitalmente y susceptibles de ser analizados de forma sistemática, en forma de *corpus-driven analysis* (y no meramente *corpus-based analysis*), tal como ha puesto de manifiesto Tognini-Bonelli (2001: 177-178):

> *The primary goal of Corpus-driven Linguistics (CDL) is to make exhaustive and explicit connections between the occurrence and distribution of language items in text, and the meanings created by the text. […] The essential methodology of CDL is to exercise the researcher's intuition in the presence of as much relevant data as can be assembled.*

Resulta, por tanto, indispensable favorecer el intercambio de protocolos técnicos estándar, y en ese aspecto nos centramos en este trabajo, tanto para establecer la tipología de impresos y manuscritos a estudiar, como para catalogarlos de forma unitaria y, sobre todo, acceder de forma semejante y en condiciones óptimas de explotación al contenido de esos productos informativos, es decir, crear herramientas de interoperabilidad a nivel semántico[8].

Por lo que respecta a la constitución de corpus, necesitamos no sólo una mera imagen facsimilar de los documentos –importante sin ningún género de dudas– sino también transcripciones en forma de documentos localizables, "un gran y estructurado conjunto de textos que puedan ser almacenados y analizados electrónicamente" (Tognini-Bonelli, 2001: 65), un *machine-readable set of texts*. Si quisiéramos llevar a cabo algún tipo de análisis cuantitativo del discurso, disponemos del magnífico ejemplo de Haffemayer (2002) en Francia, quien ha trabajado sobre la base de frecuencias y concordancias de palabras y frases. Existen también otros modelos con los que relacionarse, tales como el *Zurich English Newspaper Corpus* (ZEN) y el *Florence Early English Newspapers* (FEEN).

La constitución de un corpus de estas características no es un trabajo fácil. Partiendo de un facsímil en formato de imagen o PDF, la transcripción del texto original puede ser difícilmente confiada en exclusiva a un programa de OCR (Optical Character Recognition), con éxito variable con este tipo de documentos, por lo que necesariamente debe intervenirse manualmente. Necesitamos desarrollar estrategias flexibles de conversión de textos, pero correctamente diseñadas, desde el principio. Esto incluye marcar digitalmente el texto incorporando *tags* del lenguaje XML (eXtensible Markup Language) y produciendo metadatos, por no mencionar la necesidad de traducir todos los textos a una única lengua, si queremos compararlos con otros corpus y aplicarles diferentes softwares de análisis, como *WMatrix* o *CopyCatch Gold*.

[8] A los problemas historiográficos derivados de la existencia de colecciones fragmentarias, deficientemente catalogadas, incompletas, se une la ausencia de estrategias unificadas que empleen los estándares y protocolos más extendidos en la comunidad internacional: OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting), ALTO, una extensión de METS, que mantiene el aspecto facsimilar del documento, y a la vez, indica y coordina los elementos de la página, y SKOS (Simple Knowledge Organization System).

Actualmente, el de la TEI (Text Encoding Initiative) es uno de los esquemas de marcado de textos más longevos e influyentes en el campo de las Humanidades Digitales, dado que, según Burnard (2014): "the TEI will continue to evolve, both in taking on new areas of encoding, and in modifying what has already been proposed to keep up with the changing digital landscape". Una de las razones por las que el marcado en TEI se ha convertido en la práctica en el esquema más aceptado para la edición digital de textos reside en su flexibilidad y sus posibilidades de personalización (Pierazzo, 2014). Prueba de ello, en nuestro campo de conocimiento, es el esquema *ad hoc* que Fernández Travieso (2013) viene proponiendo en los últimos años para marcar los textos periodísticos de la Edad Moderna a partir del vocabulario de etiquetas que proporciona la TEI.

En este nuevo escenario, donde el marcado de los documentos se realiza a nivel *semántico* (empleando etiquetas o metainformación que expresen el significado de los elementos en lugar de su formato), las ontologías se han revelado como una herramienta capaz de facilitar la gestión documental, en general, y el proceso de representación y recuperación de la información, en particular (Arano, 2005; Pedraza-Jiménez *et al*., 2007: 569-578). Con este fin, el proyecto de la Web Semántica rescata para la Inteligencia Artificial el concepto formalizado de Ontología. La principal recomendación del *World Wide Web Consortium* (W3C) para la construcción de ontologías es el lenguaje OWL (Web Ontology Language)[9].

En el campo de la Inteligencia Artificial, una de las definiciones clásicas de ontología se debe a Studer *et al*. (1998: 161-197), quienes, fusionando las anteriores de Gruber (1993: 199-220) y Borst (1997), la describen como "a formal, explicit specification of a shared conceptualisation". Sin embargo, la ontología no es un concepto exclusivo de la Inteligencia Artificial, debido precisamente a su capacidad para compartir el conocimiento que representa, lo que ha propiciado su popularización en otros campos como el de la gestión de los recursos y herramientas digitales (Arano, 2005).

De acuerdo con Lorente Casafont (2005), las ontologías de ámbitos especializados, como la que se propone en este trabajo, constituyen un recurso fiable y efectivo para la recuperación de la información, gracias a la granularidad de la representación conceptual, a la estabilidad de estos conceptos en la comunidad científica y a la formalidad de la gran mayoría de documentos. No obstante, esta misma autora advierte de dos problemas de fondo: la paradoja lingüística y el dinamismo de los conceptos. En la misma línea, Codina y Pedraza-Jiménez (2011: 555-563), aun reconociendo que se trata de una tecnología prometedora y de enorme potencial para los sistemas de información, consideran las ontologías una solución inmadura aún y detectan tres carencias si se las compara con los tesauros, a saber,

[9] Actualmente, el lenguaje OWL va por su segunda versión estable, el OWL2.

imprecisión, insuficiencia e indeterminación. Esta afirmación contrasta con los éxitos cosechados por esta tecnología en el ámbito del Periodismo Digital. Sirva de ejemplo cómo la BBC ha adoptado importantes innovaciones en sus portales de información y en la gestión de sus contenidos[10].

Que sepamos, no se ha creado hasta la fecha una ontología sobre nuestro dominio como tal, el Periodismo de la Edad Moderna, si bien es verdad que existen propuestas relativamente emparentadas. En el ámbito de la Historia Moderna se ha puesto en marcha recientemente la iniciativa Network Ontologies in the Early Modern World, de carácter interdisciplinar, que aspira a compartir los recursos y cruzar la información que han generado varios proyectos de Humanidades Digitales ya consolidados en este campo, con la finalidad de construir una *practical ontology* que responda a la pregunta: "What was there in the early modern world?"[11]. En el ámbito del periodismo se destaca, por ejemplo, el proyecto NEWS (*News Engine Web Services*), uno de los primeros en aplicar las tecnologías de la Web Semántica al dominio de la información, que, en su caso, ha servido para desarrollar una colección de ontologías que describen el contenido de las noticias tomando como referencia los estándares de la International Press Telecommunication Council (ITPC)[12].

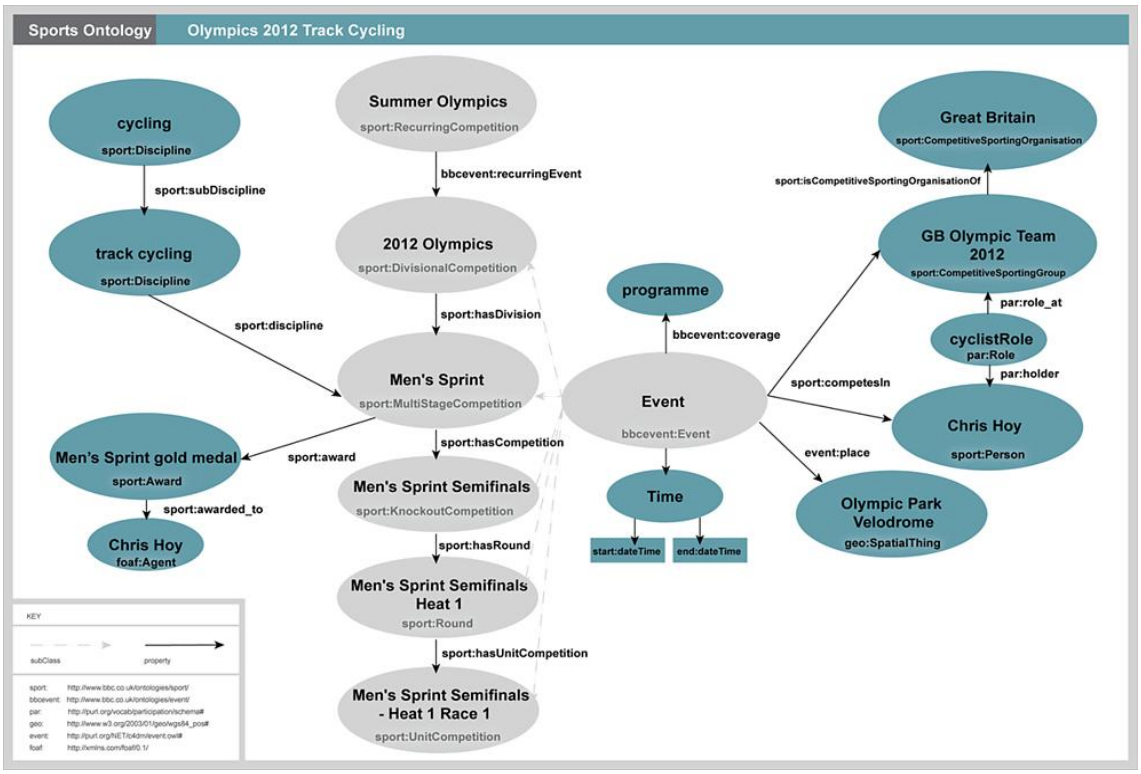

Figura 3. Grafo extraído de la *Sports Ontology*, obra de la BBC británica, que representa la información sobre la modalidad olímpica de ciclismo en pista.

[10] Accesible desde http://www.bbc.co.uk/ontologies.

[11] Los cuatro proyectos implicados en la iniciativa son: *Mapping the Republic of Letters* (http://republicofletters.stanford.edu/), *Itinera* (http://itinera.pitt.edu); *Six Degrees of Francis Bacon* (http://sixdegreesoffrancisbacon.com/) y *Manner of Belonging* (https://projects.iq.harvard.edu/johnson).

[12] Accesible desde http://webtlab.it.uc3m.es/results/NEWS/ontologies.html.

De los 48 casos de aplicación práctica de la Web Semántica recogidos por el W3C en la actualidad, dos se han desarrollado en el área de los medios de comunicación[13]; ambos tienen en común, además, que han sido diseñados en el seno de entes públicos de radio y televisión, la británica BBC[14] y la noruega NRK[15] (véase Figura 3). Bajo la óptica de la Semántica Documental, Codina y Pedraza-Jiménez (2015: 569-578) sostienen que, en caso de implantarse en un medio de comunicación, "una ontología podría servir para proporcionar un sistema de búsqueda inteligente, capaz de realizar inferencias y actuar como si fuera una verdadera inteligencia artificial".

2.2. Motivación y objetivos

En consonancia con la idea de *reinterpretar el mundo* y *construir cosas* referida anteriormente, en las próximas líneas describimos el diseño y la implementación de la ontología *Early Modern News* (EMNO, en adelante), de dominio especializado o restringido, aplicada al estudio del periodismo de la Edad Moderna, fruto del diálogo y la colaboración entre investigadores adscritos a diferentes áreas: la Historia de la Comunicación, por un lado, y las Ciencias de la Computación y la Inteligencia Artificial, por otro.

Movidos por el desafío de avanzar en la investigación sobre la historia del periodismo, recientemente hemos dado un paso más allá de la simple digitalización o de la codificación XML de textos[16] y nos hemos embarcado en una nueva travesía digital consistente en la aplicación de las tecnologías de la Web Semántica a nuestro campo de conocimiento. Con EMNO ponemos a disposición de la comunidad científica una herramienta que, en principio, nos va a permitir describir toda la información (explícita e implícita) contenida en nuestros impresos informativos para que luego pueda ser leída y tratada de manera automatizada por un procesador informático. Es decir, vamos a poder representar el periodismo impreso de la Edad Moderna y recuperar información mediante búsquedas inteligentes gracias a su capacidad para realizar inferencias.

La posibilidad de realizar estas búsquedas semánticas e implícitas -esto es, complejas- nos permitiría hallar propiedades de los impresos a priori ocultas, como, por ejemplo, su pertenencia a una serie, si es copia o continuación de otro impreso, cuántas noticias tiene o cuál es la fuente de información de cada una de ellas. Consideramos, por tanto, que las

[13] Accesible desde http://www.w3.org/2001/sw/sweo/public/UseCases/.
[14] Accesible desde https://www.w3.org/2001/sw/sweo/public/UseCases/BBC/.
[15] Accesible desde http://www.w3.org/2001/sw/sweo/public/UseCases/NRK/.
[16] Durante la última década, en el marco del proyecto BIDISO, referido al inicio del trabajo, hemos desarrollado, en una primera fase, tareas de localización, catalogación y digitalización de impresos informativos de la Edad Moderna; actualmente, sin abandonar el objetivo inicial, hemos comenzado una segunda fase orientada a la transcripción textual de los documentos acopiados, así como a la codificación XML de los mismos mediante el vocabulario de marcado TEI, que se ha instaurado como uno de los estándares más extendidos y más completos a la hora de compartir información digital en el campo de las Humanidades.

ontologías constituyen la herramienta que mejor responde a nuestras necesidades, puesto que no sólo facilitan la accesibilidad a la información, sino que también permiten relacionar todos los datos entre sí; detectar incongruencias o anomalías y normalizarlas; y, lo que es más importante, razonar sobre esa base de datos, es decir, convertir la información en conocimiento.

El objetivo del presente trabajo consiste en describir el proceso de modelización semántica de un dominio que podríamos definir como el origen de la prensa, esto es, el primer periodismo de la Historia. Y lo hemos hecho tomando como muestra los impresos informativos publicados en España en los últimos años del siglo XVI y la primera mitad del siglo XVII[17]. Dicho objetivo plantea desafíos para los dos campos de conocimiento a los que pertenecen los autores. Por un lado, para los investigadores en periodismo representa un desafío plasmar los conceptos y nociones que utilizan y definen para analizar estas manifestaciones editoriales en un objeto rigurosamente formal como es una ontología especificada en OWL. Por otro lado, los investigadores en Ingeniería del Conocimiento se enfrentan a la representación de conceptos y propiedades que se basan en nociones abstractas o vagas para estos, pero muy claras para el historiador del periodismo, tales como evento, noticia, autoría, etc.

Además de este objetivo general que sirve de marco a la investigación, diseñar y desarrollar EMNO, ha derivado en una serie de tareas u objetivos específicos, que podemos agrupar de acuerdo con su plazo de ejecución:

- A corto plazo:
  - Desarrollar una herramienta o asistente que permita poblar la ontología de una forma intuitiva y amigable, de manera que se convierta en un proceso mecánico que pueda realizar un introductor de datos, no el ingeniero ontológico ni el historiador del periodismo.
  - Poblar la ontología progresivamente con los impresos informativos que conforman nuestro corpus, teniendo en cuenta que a día de hoy disponemos de más de 1.500 ediciones registradas, de las cuales cerca de la mitad cuentan con copia digitalizada.
- A medio plazo
  - Traducir la ontología al inglés.
  - Solicitar reconocimiento y almacenamiento a W3C para formar parte de los Semantic Web Standards.
- A largo plazo

[17] En particular, con la ontología aspiramos a formalizar la evolución del discurso periodístico desde las relaciones de sucesos ocasionales a las gacetas seriadas y periódicas o semiperiódicas.

- Alojar la ontología en un servidor, de manera que podamos enlazarla con bibliotecas digitales, que contengan copia de los impresos o con sitios web como Wikipedia, lo que ofrecería una visión más completa de nuestro campo de conocimiento.
- Crear una página web que permita visualizar la ontología y realizar búsquedas semánticas.

Los objetivos, así descritos, son ambiciosos. No obstante, para probar la fiabilidad y la efectividad de las primeras versiones de una ontología de ámbito especializado como la nuestra, se plantea poblarla en una fase inicial a partir de un corpus restringido, integrado por una serie de relaciones de sucesos y gacetas que publicaron en Sevilla, los impresores Juan Serrano de Vargas, Simón Fajardo y Juan de Cabrera en el primer tercio del siglo XVII[18]. En principio, la creación de una ontología sobre nuestro campo de conocimiento se revelaría como una herramienta útil no sólo para facilitar la accesibilidad y la representación del periodismo de la Edad Moderna, sino también para razonar sobre esa base de datos, permitiéndonos trabajar con hipótesis novedosas en el marco del análisis del discurso periodístico de este período.

3. INGENIERÍA ONTOLÓGICA

En este epígrafe pasamos a explicar la metodología de construcción de ontologías que hemos utilizado para el diseño de la que nos ocupa. A continuación, describimos su aplicación al caso, los pasos que hemos seguido y las decisiones que hemos tomado en cuanto a diseño, compromisos ontológicos, etcétera.

3.1. Descripción de la metodología

Para la edición, diseño y desarrollo de la ontología que nos ocupa hemos utilizado el editor Protégé 4.3[19]. Se trata de un editor de código libre, muy utilizado en la construcción de sistemas basados en conocimiento, en áreas diversas como la biología o la medicina. Se siguió la metodología simple extraída de Noy y McGuinness (2005), según la cual los pasos que se han seguido para dar forma a nuestra ontología consisten en:

[18] Juan Serrano de Vargas, Simón Fajardo y Juan de Cabrera forman parte de la importante saga de impresores sevillanos que, al unísono con otros como los barceloneses Mathevad y Liberós, introducen en España los primeros formatos de la información periódica de actualidad, por las mismas fechas que estos se ensayan en el resto de Europa, desmintiendo el lugar común del supuesto *retraso* del periodismo ibérico con respecto al centroeuropeo.
[19] Accesible desde http://protege.stanford.edu/.

- Determinar el dominio y el alcance, el uso que se le va a dar, qué tipo de preguntas deberá responder, quién la usará y la mantendrá.
- Considerar la reutilización de ontologías existentes.
- Enumerar una lista integral de los términos más importantes del campo de conocimiento que pretendemos representar y de sus propiedades.
- Definir las clases y la jerarquía de clases.
- Determinar las propiedades de las clases junto con el tipo/número de valores/lista de permitidos que esas propiedades van a tomar.
- Por último, crear las instancias y completar los valores de sus propiedades.

Cada paso puede dar lugar a una revisión de los anteriores, siempre consensuada con los expertos en el dominio de conocimiento. Se trata, por tanto, de un proceso de diseño iterativo.

El dominio que se pretende representar se ha descrito ampliamente en los primeros epígrafes de este trabajo. Se trata de describir las características que definen a los impresos informativos que se publicaban en España durante la Edad Moderna, características que se refieren a la autoría, al diseño y a la estructura de las publicaciones, al tema y al contenido de las noticias o al estilo y al lenguaje genuinamente periodísticos que ya son evidentes en esta primera prensa de la Historia (véase Figura 4).

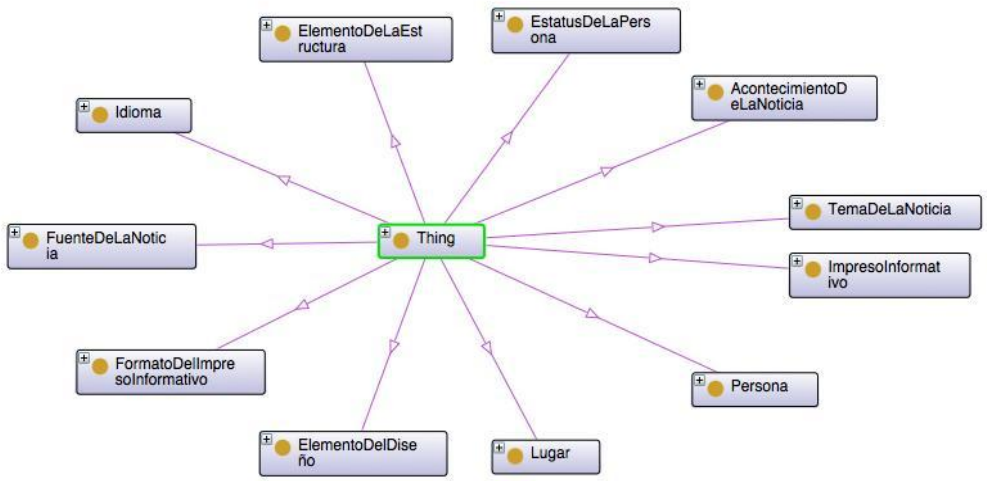

Figura 4. Grafo que representa visualmente las principales clases de la ontología *Early Modern News*.

La idea es catalogar cada detalle *relevante* de los documentos existentes y poder disponer del conocimiento extraído para, mediante técnicas de razonamiento/consulta sobre la ontología, detectar características imperceptibles para el historiador del periodismo que

realiza su estudio a partir del análisis de las digitalizaciones de los documentos originales. Esta utilidad trasciende ayudándonos a confirmar hipótesis y alcanzar objetivos, puesto que nos permite superar la opacidad de los documentos digitalizados, realizar búsquedas más complejas, así como procesar de forma automática la información contenida en nuestro corpus de documentos.

La información extraída de cada impreso se referirá, por ejemplo, a elementos de su diseño (alineación del texto, estilo tipográfico, tamaño de la página o tipo de primera plana) y de su estructura, diferenciando entre los elementos propios de la edición (título, imagen, pie de imprenta, licencia, privilegio, etc.) y de la información, esto es, las noticias. La ontología también permite representar la información contenida en los impresos que hace referencia a personas y lugares; en el primer caso, pueden ser personas implicadas en el proceso de producción del impreso (autor, editor, impresor, promotor, traductor), protagonistas de las noticias o testigos de los acontecimientos; en el segundo caso, los lugares (una ciudad, una región o un estado) indican dónde se ha publicado el impreso, dónde tiene lugar la noticia o de dónde procede la información. De las noticias en sí mismas, además de sus protagonistas o del lugar donde suceden, también podemos representar la información relativa a la fuente o al tema. De manera natural, al describir el dominio de conocimiento que nuestra ontología va a representar formalmente, es inevitable introducir lo que va a ser un primer nivel de conceptos y propiedades de la jerarquía ontológica inicial.

En cuanto al uso pretendido de la ontología, ya se ha comentado más arriba que sería deseable, por ejemplo, poder agrupar aquellas publicaciones que, siendo diferentes, informen sobre la misma noticia o se refieran al mismo personaje o acontecimiento, a fin de poder detectar características que indiquen una posible seriación en los impresos.

Respecto al alcance, el nivel de detalle a la hora de representar la información se decidió que fuera inicialmente muy exhaustivo. En cada concepto a representar no es difícil considerar la instanciación de todas las propiedades útiles si lo que se pretende es la catalogación o registro de toda la información que facilita un impreso en su estructura y contenido. Más tarde, conforme fue evolucionando el diseño de la ontología, fuimos limitando el alcance en algunas cuestiones que detallamos en el apartado dedicado a la toma de decisiones. Estos compromisos de restricción son usuales en Ingeniería Ontológica.

3.2. Conceptos y propiedades

Siguiendo la metodología mencionada, se comenzó realizando un listado acerca de nuestro dominio, el periodismo de la Edad Moderna, que incluyese los conceptos relacionados, así como los elementos relativos a la autoría, el diseño, la estructura, el contenido y estilo de los impresos informativos. Estos términos se incluyeron sin entrar en

consideraciones sobre si se trataba de propiedades o elementos conceptuales. Así, en ese listado aparecieron palabras tales como imagen, editor, impreso, lugar de publicación, fuente de la noticia, fecha de publicación, acontecimiento, privilegio, licencia, título, etc. Los expertos tuvieron que definir y aclarar dichos conceptos a los ingenieros ontológicos; descartar ambigüedades y acordar significados comunes. Por ejemplo, al incluir el papel de las personas relacionadas con cada impreso informativo, inicialmente, algunos de los roles seleccionados fueron impresor, autor, editor, promotor, corresponsal, traductor y homenajeado. El ingeniero ontológico necesitó asimilar las definiciones y las propiedades diferenciadoras de dichos roles: ¿En qué consiste el papel del promotor? ¿Puede ser una misma persona el promotor y el impresor de un documento? ¿Puede haber más de un promotor o más de un corresponsal? ¿A qué nos referimos con homenajeado?

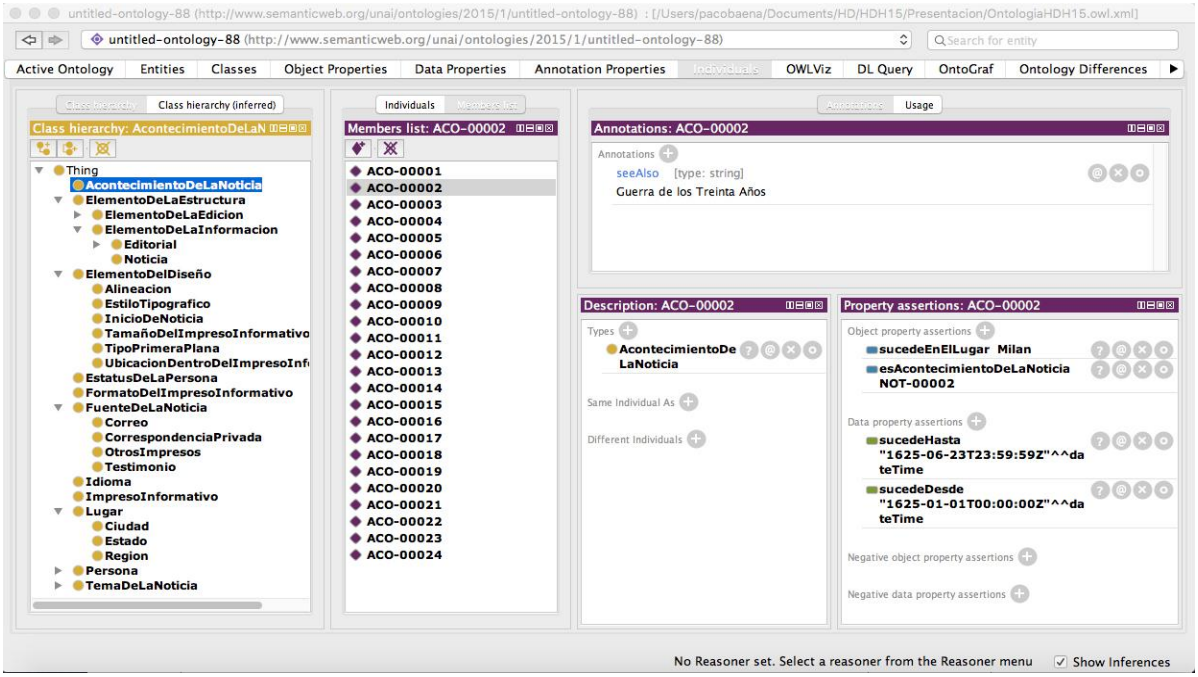

Figura 5. Jerarquía de las clases en la ontología *Early Modern News*.

Una vez establecidos los significados de cada término, se inicia el proceso de jerarquización de conceptos como árbol de clases. Elegimos de la lista de términos elaborada los que pueden representarse mediante clases, aquellos de los que habla la ontología expresando propiedades o características. Obtenemos la jerarquía de clases que se describe en la Figura 5. En ese proceso decidimos qué clases serán disjuntas. Un ejemplo sencillo: dentro de la clase *Imagen*, las subclases *Escudo* e *Icono* serán disjuntas, no permitiremos que una imagen sea a la vez ambas cosas (véase Figura 6).

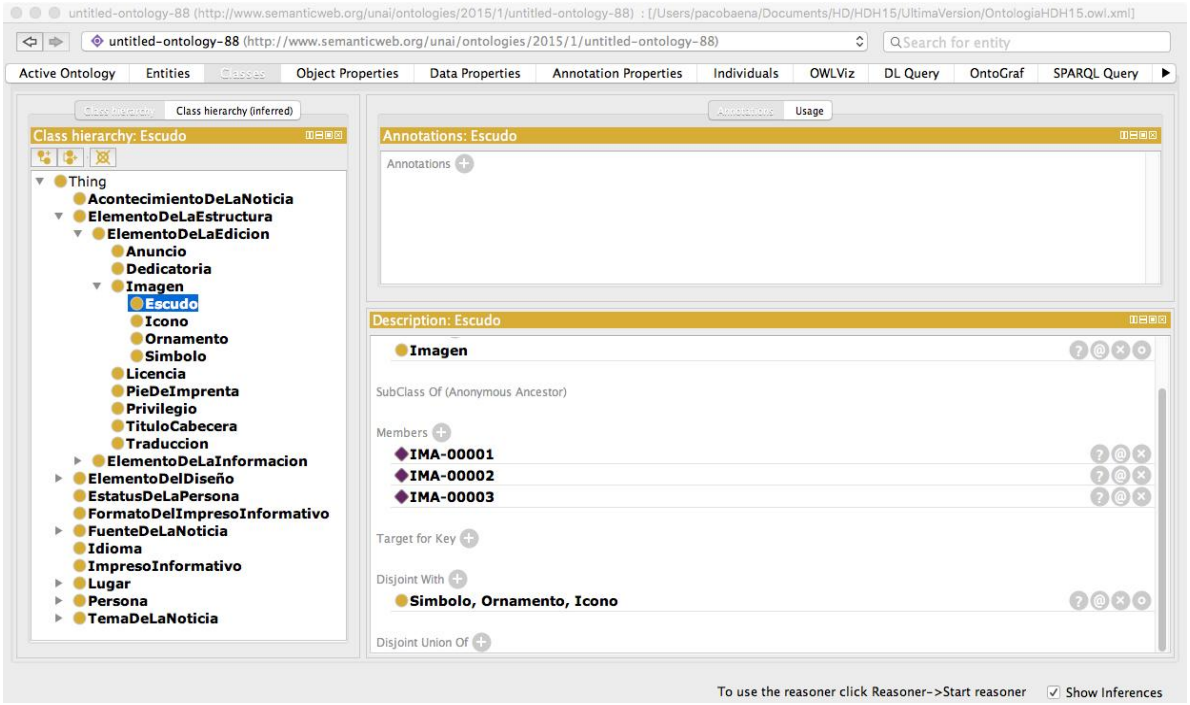

Figura 6. Ejemplo de clases disjuntas.

Habrá que anotar cada uno de los términos para que quede registrada la definición pretendida de cada uno, los comentarios pertinentes y los posibles enlaces o etiquetas que completen la información. Por ejemplo, la clase *Persona* contiene los roles de las personas con respecto a un impreso informativo, una noticia o un testimonio. Entre otros, podemos encontrar los roles *Autor*, *Corresponsal* y *Editor*. El usuario que introduzca los datos relativos a un impreso concreto podría dudar entre asignar a una persona la autoría de un impreso o bien la edición del mismo. Más peligroso es el caso en el que el usuario no sufra dudas, el caso en que entienda erróneamente un término. Por ejemplo, el significado actual de la palabra *corresponsal* difiere del pretendido al nombrar en la ontología la clase *Corresponsal*. En este caso, una persona podrá figurar como corresponsal de la correspondencia privada que puede ser la fuente de una noticia. En estos casos, el usuario que se dedica a poblar la ontología, es decir, a introducir los individuos de la misma, agradecerá estas anotaciones aclaratorias. Esta es una tarea de documentación de la propia ontología que debe ir indicándose a lo largo del desarrollo de la misma y completarse al final, comprobando que encajen la realidad representada y el formalismo obtenido.

A continuación, pasamos a establecer las propiedades de las clases. Hay que indicar que, aunque decimos “a continuación”, el proceso no es lineal: al establecer las clases, ya hemos avanzado necesariamente sobre sus propiedades. Cuando tomamos la decisión de incluir la clase *TemaDeLaNoticia*, es porque hemos sopesado la conveniencia de incluir la propiedad *tieneElTema* para el dominio de la clase *Noticia* y también hemos debatido la utilidad de incluir su inversa *esTemaDeLaNoticia*. Más aún, desde que incluimos la clase

Noticia, ya se inició inevitablemente el proceso de elaboración de la lista de propiedades de las noticias que se consideran necesarias a nuestros propósitos, entre las que están las relacionadas con el tema central de la noticia en cuestión. Lo que hacemos es formalizar las definiciones de esas propiedades, clases dominio, rangos, inversas y demás detalles en la ventana correspondiente de Protégé. En esta fase del desarrollo de la ontología hay que decidir si una propiedad va a ser tipo dato (*DataProperty* en Protégé) o tipo objeto (*ObjectProperty*). Podemos crear una clase que contenga el conjunto de valores limitado que una propiedad puede tomar y, de esta forma, dicha propiedad se convierte en propiedad tipo objeto; o bien asignamos a los valores un tipo genérico como *Integer* o *String* y permitimos introducir cada valor de la propiedad (propiedad tipo dato) de manera no predeterminada. Tras varias revisiones, siempre consensuando el aspecto formal con el real, establecemos la jerarquía de propiedades, que puede observarse en las Figuras 7 y 8. De manera simultánea, vamos depurando la nomenclatura y la normalizamos de manera que refleje la estructura jerárquica y las agrupaciones.

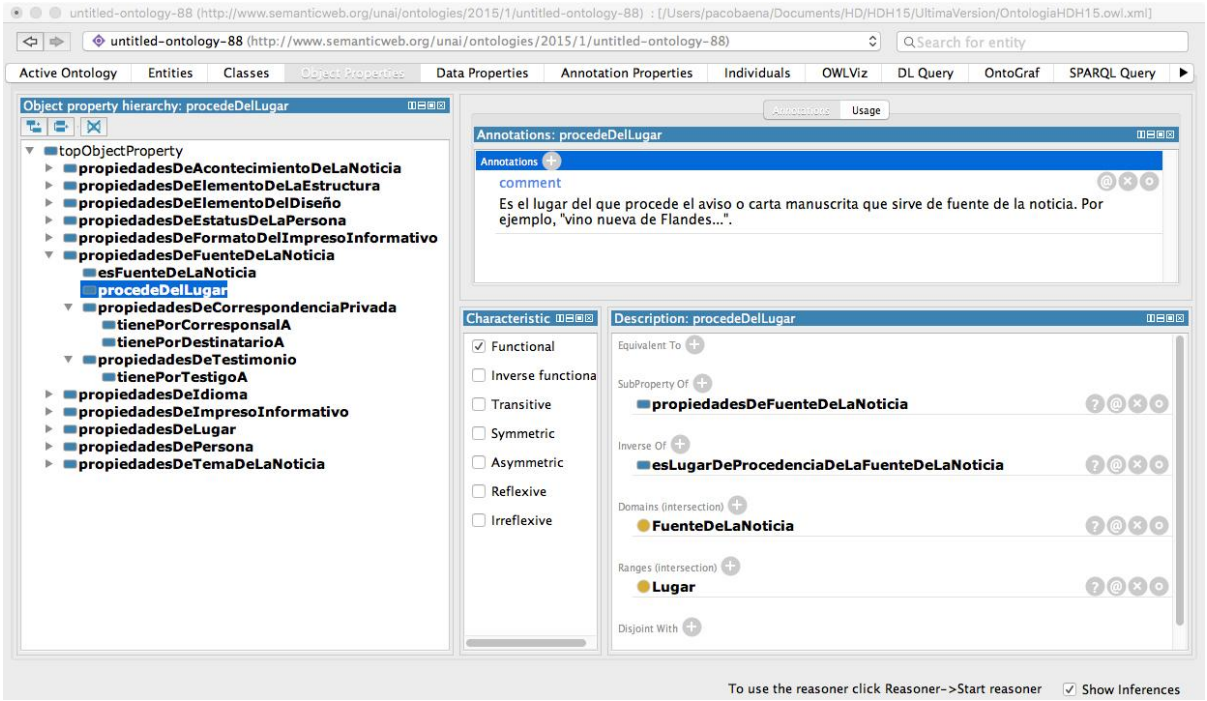

Figura 7. Despliegue parcial de la jerarquía de las propiedades tipo objeto.

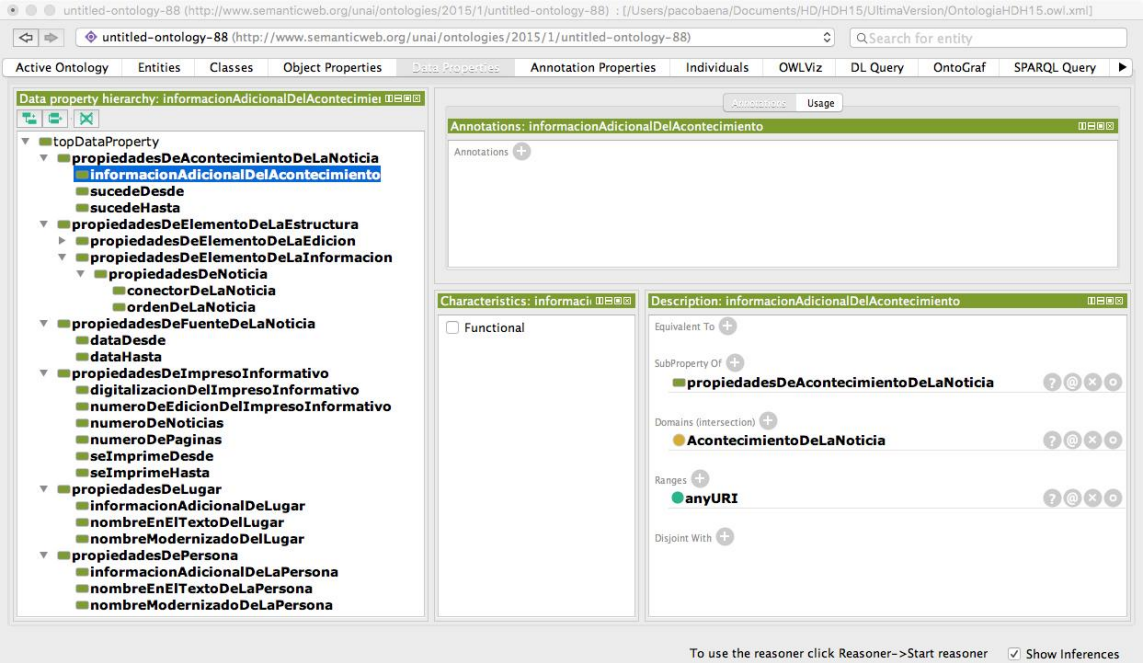

Figura 8. Jerarquía de las propiedades tipo dato en la ontología *Early Modern News*.

3.3. Toma de decisiones

Hemos comentado más arriba que cada paso lleva consigo la toma de decisiones. Algunas de ellas se resuelven con el objeto de homogeneizar el conjunto de clases y propiedades, así como la filosofía que hay detrás de la ontología (pragmatismo, exhaustividad, etc.). En el caso de la exhaustividad, relacionado con el alcance de la ontología, a lo largo del diseño y tras instanciar los primeros documentos, se decidió que no fueran de interés, entre otros, los datos sobre el privilegio o la licencia de un impreso. La ontología responde así a la cuestión de si estos elementos figuran o no en el impreso, pero no alcanza el detalle de quién o cuándo se otorgan, ni tampoco registra el texto concreto de dichos elementos.

En cuanto al pragmatismo, hay decisiones que tomar en relación con el tipo de propiedad que se desea elegir. Elegimos una propiedad que tiene como dominio una clase concreta y, a continuación, decidimos si va a ser *DataProperty* o bien *ObjectProperty*.

Otras decisiones de mayor relevancia se refieren a la posible reutilización de ontologías ya existentes para representar una parte del conocimiento que abarca la que nos ocupa[20]. En este caso, la toma de decisiones conlleva un estudio más pormenorizado de las opciones y de sus ventajas e inconvenientes. Estos aspectos se ejemplifican en los próximos epígrafes.

[20] Desde este sitio se ofrece un listado de ontologías para reutilizar, ordenadas por palabras clave: http://www.daml.org/ontologies/keyword.html.

3.3.1. Propiedades tipo Dato o tipo Objeto

Clasificamos los elementos del diseño de un impreso informativo mediante seis clases como son la alineación, el estilo tipográfico, el inicio de la noticia, el tamaño del impreso informativo, el tipo de primera plana y la ubicación de ciertos elementos como una imagen o una dedicatoria dentro del impreso. Cada una de ellas contiene como individuos los valores admitidos por las propiedades correspondientes. Por ejemplo, las propiedades *tieneAlineacion* y *tieneEstiloTipografico* son de tipo *ObjectProperty* y creando las clases *Alineacion* y *EstiloTipografico*, hemos limitado los valores de esas dos propiedades a los contenidos en esas clases. Así, el título o cabecera de un impreso podrá tener alineación centrada, izquierda, derecha o justificada y ningún otro valor podrá asignarse, y el estilo tipográfico del título podrá ser cursiva, mayúsculas, regular, subrayado o versalitas. Otra opción hubiera sido definir las propiedades con tipo *DataProperty*, de manera que, al completar los datos de las propiedades de los títulos en cuestión, habría que introducir los valores como *String* (por ejemplo), no seleccionando en el editor el valor con un solo clic del ratón. También Protégé nos permite crear una *DataProperty* con un conjunto de valores restringidos, pero, en ese caso, al no disponer de una clase que los agrupe, las opciones a elegir entre los valores posibles quedan ocultas al usuario, no tienen la visibilidad que adquieren al estar en una clase definida como rango de la propiedad.

En este caso es aconsejable elegir propiedades de tipo objeto, ya que el número de valores posibles es limitado y se facilita al introductor de datos la elección del valor correcto de manera directa e impidiendo posibles errores tipográficos o el uso de diferentes grafías para un mismo valor. Esto último se hace patente al definir la clase *Lugar*, cuyas subclases son *Ciudad*, *Estado* y *Region*. Dichas subclases son los rangos de algunas propiedades de tipo *ObjectProperty* como *procedeDelLugar* o *sucedeEnElLugar*. En el caso de la clase *Lugar*, los individuos no están restringidos a unos pocos como ocurre con los cuatro únicos individuos de la clase *Alineacion*. Cada vez que procedamos a incluir un nuevo impreso, es posible que los valores de las propiedades con rango *Lugar* para dicho impreso aparezcan como individuos en la clase correspondiente o, por el contrario, que haya que añadir alguno. Al estar en una clase, bastará añadir ese nuevo lugar a la lista de individuos y estará disponible desde ese momento para ser utilizado, sin errores ni ambigüedad en la grafía y de manera más rápida y sencilla.

3.3.2. Ontologías del tiempo

Como ya hemos señalado, uno de los aspectos en los que la toma de decisiones es crucial en el desarrollo de una ontología se refiere a la posible reutilización de otra ya

existente, que pueda adaptarse o integrarse como parte de la que está en construcción. En primer lugar, intentamos localizar ontologías similares que tratan de representar el mismo dominio o bien subdominios que nos interesen. Después analizamos el grado de adaptabilidad al dominio de nuestra ontología y evaluamos la conveniencia o no de abordar los cambios para lograr la adaptación. Hay casos en los que una ontología cumple nuestras expectativas y la sumergimos en su totalidad, aun cuando eso nos obligue a ceder en algún aspecto menor de nuestra fase inicial. Otras veces, al analizar ontologías de dominio similar al que pretendemos, lo que logramos es extraer ideas para mejorar la representación de nuestro modelo ontológico. En el peor de los casos, decidiremos seguir con el diseño de nuestra ontología descartando la integración total o parcial de otras ontologías ya creadas. Esto último es lo ocurrido con la ontología que estamos diseñando.

En este sentido, una de las decisiones más pensadas y conflictivas ha sido la relativa al tiempo. ¿Cómo debíamos representar las fechas en nuestra ontología? Inicialmente nos topamos con un problema derivado de la singularidad que caracteriza a la prensa de la Edad Moderna en relación con el tiempo. Dicha singularidad estriba fundamentalmente en el hecho de que podemos encontrar hasta tres fechas diferentes en cada impreso informativo: siguiendo un orden cronológico serían, primero, la fecha en que sucede la noticia; segundo, la fecha en que se escribe el aviso que informa de la noticia, es decir, la fecha de la fuente de la noticia; y tercero, la fecha en que se publica finalmente el impreso. Además, la complejidad se ve incrementada por el hecho de que cada una de estas fechas puede ser puntual (un día de un mes de un año concreto) o abarcar un intervalo de tiempo.

Ante el problema descrito arriba se nos planteaban dos posibles soluciones:

- Importar una ontología de tiempo ya existente.
- Crear propiedades de datos del tipo dataTime.

La idea inicial era registrar los intervalos como valores de una cierta propiedad y estudiamos la ontología OWL-Time[21]. Esta ontología (dentro de la estandarización de W3C) abarca los conceptos temporales y su uso pretendido es la descripción, en el aspecto temporal, de páginas y servicios Web. Al estudiarla observamos que excede las necesidades del campo de conocimiento que intentamos describir. No necesitamos unidades de tiempo ni calcular el número de días o meses de un período o intervalo propio. Nuestra necesidad es poner inicio y fin a un intervalo de tiempo del que tal vez sólo conocemos el mes o el año, pero no una fecha y hora concretas. Finalmente, descartamos esta opción, pues la mayor parte de las clases no se utilizan y no contempla la posibilidad de intervalos.

[21] Accesible desde https://www.w3.org/TR/2006/WD-owl-time-20060927/.

Por consiguiente, nos decantamos por la segunda solución, la de considerar el tipo de dato *dataTime*[22] o *dataTimeStamp*[23] como restricción del anterior. Es cierto que esta vía requiere un formato específico al introducir la fecha, pero permite consultas y funciona bien. Además, nos permite introducir intervalos de tiempo, así como ordenar fechas con mucha exactitud por día, mes, año e incluso por hora. El formato para este tipo de dato es *AAAA-MM-DDThh:mm:ssZ*, es decir, año-mes-día, letra T, seguido de horas-minutos-segundos y la letra Z de la zona horaria no obligatoria para *dataTime*. Dentro del formato, cada campo es obligatorio, así, si de una noticia sabemos que sucedió en la fecha 10 de marzo de 1618, tendremos que poner en la descripción del individuo correspondiente, en el apartado *dataProperty assertion*, el valor 1618-03-10T00:00:00Z y en *Type* seleccionar *xsd:dataTimeStamp* o bien *xsd:dataTime*.

Una vez resuelto el problema inicial, pasamos a definir las propiedades de tipo dato para cada una de las clases relacionadas con la representación del tiempo en los impresos informativos de la Edad Moderna, a saber: *sucedeDesde* y *sucedeHasta* para representar la fecha en que sucede el *AcontecimientoDeLaNoticia*; *dataDesde* y *dataHasta* para representar la fecha en que se escribe el aviso que informa de la noticia, es decir, vinculada al dominio *FuenteDeLaNoticia*; y, por último, para representar la fecha en que se publica finalmente el *ImpresoInformativo* utilizamos las propiedades *seImprimeDesde* y *seImprimeHasta*. En este último caso, si de un impreso informativo sólo sabemos que se publicó a lo largo del año 1621, habrá que acordar poner en las propiedades *seImprimeDesde* y *seImprimeHasta* los valores 1621-01-01T00:00:00Z y 1621-12-31T23:59:59Z, respectivamente.

La opción elegida no sólo resulta eficiente y fiable para representar las diferentes fechas relacionadas con los impresos informativos de nuestro corpus, sino que también permite mostrar cómo circulaba la información en la Europa de la Edad Moderna, cuánto tiempo transcurría desde que sucedía la noticia hasta que se publicaba en forma de relación o gaceta, esto es, cuál era su grado de inmediatez y actualidad. Incluso nos permite ser más precisos a la hora de representar la fecha de publicación de los impresos informativos, puesto que podemos detectar fácilmente datos implícitos en los mismos. Por ejemplo, si un impreso informativo se publica en el año 1625, según su pie de imprenta, e informa de una noticia sucedida el 23 de mayo de 1625, podemos deducir rápidamente que dicho impreso nunca se pudo publicar antes de que sucediese la noticia, esto es, antes del 23 de mayo de 1625. Y esto lo podemos representar en la ontología acotando la fecha de publicación de dicho impreso mediante un intervalo de tiempo, gracias a las propiedades descritas anteriormente: *seImprimeDesde* [1625-05-23T00:00:00Z] y *seImprimeHasta* [1625-12-31T23:59:59Z].

[22] Accesible desde http://www.datypic.com/sc/xsd11/t-xsd_dateTime.html.
[23] Accesible desde http://www.datypic.com/sc/xsd11/t-xsd_dateTimeStamp.html.

Además, se incluye una regla que de forma explícita formalice el requisito de que la fecha de publicación sea siempre posterior a la fecha del acontecimiento.

3.4. Consultas

Decíamos en los primeros apartados de este trabajo que con EMNO ponemos a disposición de la comunidad científica una herramienta que permita recuperar la información (explícita e implícita) contenida en nuestros impresos informativos, mediante búsquedas inteligentes, gracias a su capacidad para realizar inferencias. Para realizar las búsquedas utilizamos *SPARQL Query*, un *plugin* instalado en Protégé que nos permite razonar sobre la base de datos disponible en nuestra ontología.

A continuación, exponemos varios ejemplos de consultas semánticas que, como se verá, nos permiten hallar propiedades, a priori ocultas, de las relaciones y gacetas con que hemos poblado inicialmente la ontología.

La primera consulta que hemos realizado para poner a prueba la fiabilidad de la herramienta sería esta: ¿Qué formato tienen los impresos informativos publicados durante la década de 1620? Para formular la pregunta en SPARQL (lenguaje de consulta sobre ontologías pobladas) debemos combinar las clases *ImpresoInformativo* y *FormatoDelImpresoInformativo* con las propiedades tipo dato *seImprimeDesde* y *seImprimeHasta* (véase Figura 9). Así estamos solicitando que el razonador seleccione primero sólo aquellos impresos informativos que, según su fecha de publicación, se imprimieron entre los años 1620 y 1629, para ofrecer luego una relación de todos ellos donde se indique su formato, que a su vez puede adoptar una de estas cuatro propiedades tipo objeto: *RelacionDeSucesos*, *RelacionSeriada*, *GacetaSemiperiodica* y *Gaceta*. El resultado de la consulta aparece en la parte inferior de la pantalla de Protégé.

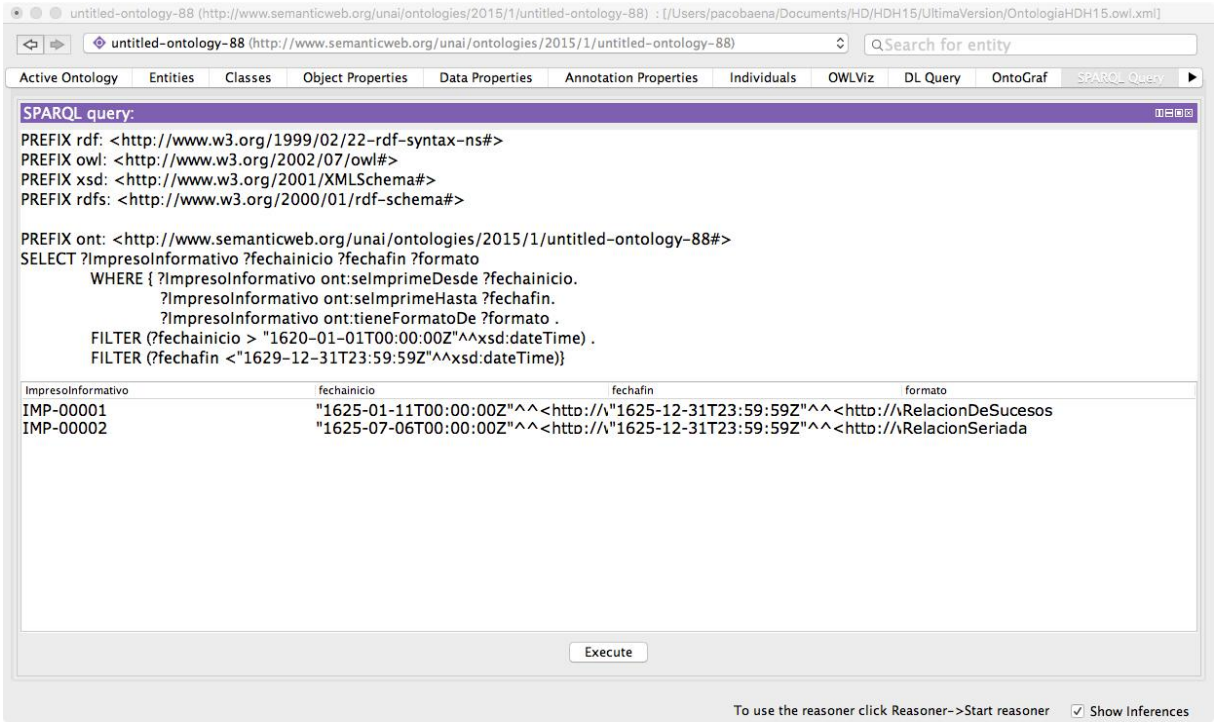

Figura 9. Consulta acerca del formato que presentan los impresos informativos, realizada mediante el plugin SPARQL Query en Protégé.

La consulta descripta arriba nos permite también comprobar que hemos tomado la decisión correcta acerca de cómo representar el tiempo en nuestra ontología. No sólo hemos obtenido el resultado esperado, sino que la respuesta a la consulta muestra los impresos ordenados por fecha, obteniendo únicamente aquellos que se imprimieron en un intervalo concreto de tiempo. Por tanto, podemos confiar en que el formato de fecha y el tipo *dataTime* satisfacen las necesidades del alcance determinado para la ontología en relación con los conceptos temporales.

La segunda consulta que hemos realizado utilizando *SPARQL Query* ha sido esta: ¿Qué noticias tienen como protagonista a un rey? Para formular la pregunta en lenguaje SPARQL, en este caso, debemos combinar las clases *Noticia* y *Persona* con la propiedad tipo objeto *tieneComoProtagonistaA* (véase Figura 10). Así obtenemos una relación de todas aquellas noticias en las que aparece mencionado un monarca.

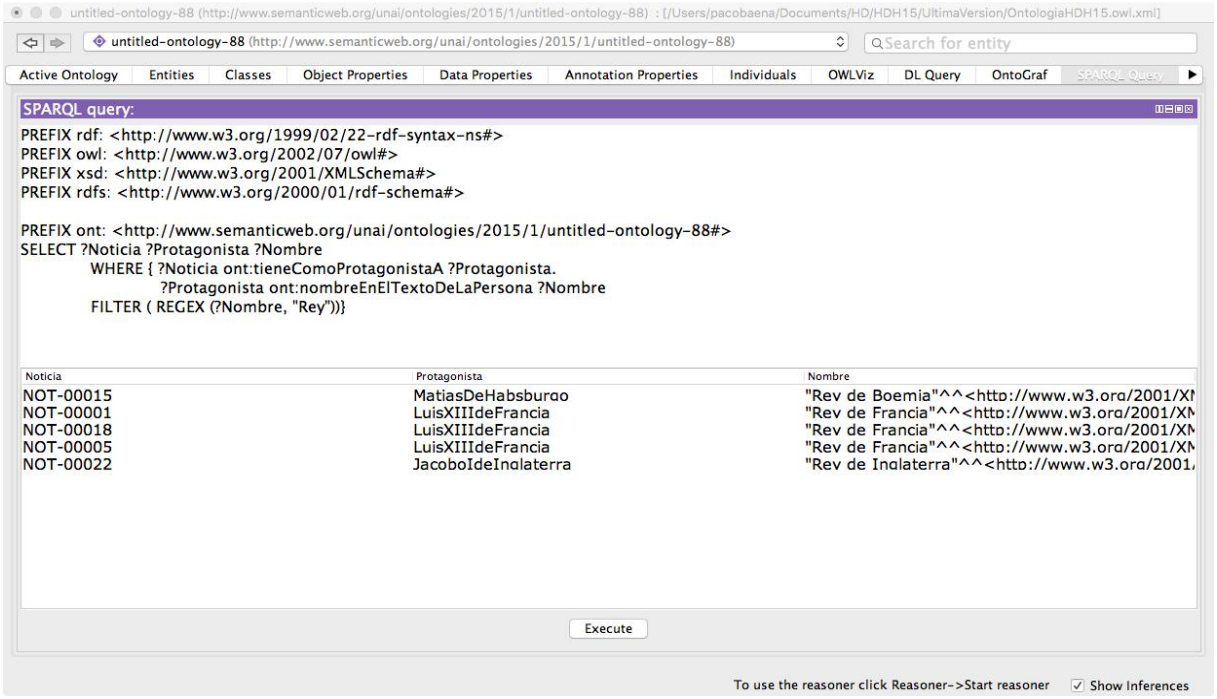

Figura 10. Consulta acerca de las noticias protagonizadas por algún rey, realizada mediante el plugin SPARQL Query en Protégé.

4. RESULTADOS

El diseño de la ontología *Early Modern News* es una primera aproximación a un objetivo mucho más ambicioso: la creación de un portal semántico que permita al usuario acceder a los documentos y a sus digitalizaciones, pero no se limite sólo a eso. Dicho portal deberá facilitar al investigador el conocimiento acumulado y filtrado en los impresos informativos de la Edad Moderna, agrupándolos por fecha de publicación, por formato o por impresor, clasificando las noticias que contienen de acuerdo con su temática o el estilo en el que están escritas, e incluso descubriendo relaciones entre los impresos referidas a su estructura que no estaban visibles en la simple digitalización.

Todo ello es posible gracias a la aplicación de tecnologías de Web Semántica, junto con las herramientas de razonamiento de que disponemos. La creación de la ontología nos ofrece una primera formalización del conocimiento que nos interesa para hacer posible su tratamiento y procesamiento por máquinas de manera automática.

Este trabajo describe la aplicación de una metodología sencilla a la construcción de una ontología cuyo dominio de conocimiento gira en torno a los elementos y características que definen los impresos informativos de la Edad Moderna en España. Se ha conseguido una formalización del conocimiento como se pretendía, obteniendo así la ontología que responde a los objetivos iniciales en cuanto a dominio y alcance de la representación. Se han incluido algunos documentos para la realización de consultas y pruebas de consistencia iniciales. No

obstante, nos encontramos en un estadio inicial en la implementación de nuestra ontología y aún quedan algunas tareas por hacer.

El siguiente paso que tenemos que dar es naturalmente el poblado de la ontología, es decir, la introducción de individuos reales. Hasta ahora habíamos insertado datos relativos a un corpus reducido de impresos informativos con la intención de realizar pruebas de consistencia.

El proceso de poblado se inicia eligiendo una clase de la jerarquía, en nuestro caso, la clase *ImpresoInformativo*, e insertando uno concreto como individuo. A partir de ahí, tendremos que ir describiendo los valores de todas las propiedades tipo dato o tipo objeto que atañen a ese impreso. Hay que tener en cuenta que hay 26 propiedades de tipo *ObjectProperty* y 6 de tipo *DataProperty* para registrar en cada impreso. A la hora de establecer un valor concreto para una propiedad, tipo objeto, por ejemplo, fijado un impreso, debemos antes consultar (en la ventana de Protégé correspondiente) qué rango tenemos que considerar para esa propiedad y elegir el valor en un listado completo de todos los individuos de la ontología no agrupado por clases. Esto significa que, si añadimos un nuevo impreso *IMP-0000a* dentro de la clase *ImpresoInformativo* y seleccionamos en el menú *objectProperty assertion* la propiedad *tienePorPromotorA*, deberemos elegir de la lista de individuos el nombre del promotor y dichos nombres no están agrupados dentro de la clase Promotor, sino en un listado plano de individuos: acontecimientos, imágenes, fuentes, noticias, etc. La situación es también engorrosa si la propiedad elegida es tipo dato, ya que, al seleccionarla, disponemos de una ventana en blanco que no nos orienta acerca del tipo al que corresponde el dato a ingresar, con lo cual podríamos dudar de escribir *uno* o *1*. Esto nos lleva a consultar la descripción de la propiedad cada vez que tenemos que consignarla.

Otro inconveniente en el poblado de la ontología es la ramificación que constantemente se produce con cada dato que introducimos. Si el valor de una *DataProperty* para un individuo A es un individuo B, tendremos que dejar pendiente la cumplimentación de los datos de B como individuo de su clase y continuar con los datos de A. Y no olvidar volver sobre B. Esto, para ontologías sencillas, ya se convierte en un bucle que lleva al grabador de datos a continuas verificaciones.

Estas dificultades podrían resolverse con el desarrollo de un *plugin* para el editor Protégé que se encargue de mostrar un formulario completo que debe ser cumplimentado cada vez que se marque un individuo. Esta necesidad debería cubrirse para facilitar el poblado de la ontología y es el siguiente objetivo de esta primera fase del proyecto que nos ocupa.

Otra de las tareas que queda pendiente es la elaboración de un manual que describa nuestra ontología, su jerarquía, propiedades y definiciones de términos. Este procedimiento debe completarse conforme se va desarrollando la ontología y, de hecho, las anotaciones de

los conceptos y propiedades aparecen ya en el modelo diseñado con Protégé; pero hay que recopilarlas y estructurarlas en un formato independiente del editor utilizado.

Por último, restaría alojar la ontología en un servidor, de manera que mediante su enlace con bibliotecas digitales que contienen copia de los impresos o con sitios web como Wikipedia, se ofrezca al usuario una visión completa de nuestro campo de conocimiento. Este objetivo lleva aparejada la creación de un portal semántico que, finalmente, explote todas las posibilidades que estas tecnologías brindan al investigador, de manera que, respecto al estudio de los inicios del periodismo actual, se puedan confirmar hipótesis y realizar los prometedores avances que hemos vislumbrado.

5. CONCLUSIONES

Este trabajo representa una contribución que no sólo tiene un marcado y evidente carácter interdisciplinar (Periodismo, Historia, Ingeniería del Conocimiento, Inteligencia Artificial), sino que también refleja la innovación metodológica a través de las Humanidades Digitales, rompiendo las barreras entre las distintas ramas del árbol de la ciencia y creando un *machine-readable set of texts* (Tognini-Bonelli, 2001) en el marco de la *digital history* (Cohen y Rosenzweig, 2005).

Tras un primer poblado experimental con un corpus reducido de impresos informativos, EMNO se ha revelado como una herramienta fiable y efectiva aplicada a un dominio especializado como el nuestro, el periodismo impreso de la Edad Moderna, dado que facilita tanto la gestión documental como el proceso de representación y recuperación de la información (Lorente Casafont, 2005; Arano, 2005; Pedraza-Jiménez *et al*., 2007: 569-578). Además, nos ha proporcionado un sistema de búsqueda inteligente, capaz de realizar inferencias y actuar como si fuera una verdadera inteligencia artificial, pese a encontrarse aún en un estadio embrionario, permitiéndonos trabajar con hipótesis novedosas en el marco del análisis del discurso periodístico de este período.

Por ejemplo, la posibilidad de realizar estas búsquedas complejas nos ha permitido hallar propiedades de los impresos a priori ocultas, como su pertenencia a una serie, si es copia o continuación de otro impreso, cuántas noticias tiene o cuál es la fuente de información de cada una de ellas. Por consiguiente, hemos confirmado la hipótesis de partida según la cual las ontologías constituyen la herramienta que mejor responde a nuestras necesidades, puesto que nos ha permitido relacionar todos los datos entre sí, aplicando incluso restricciones a las propiedades de nuestros documentos (por ejemplo, que un impreso no puede tener a la vez dos formatos).

Adicionalmente, hemos podido detectar incongruencias o anomalías y normalizarlas: en nuestros impresos informativos, solemos encontrar diferentes denominaciones para

referirse, por ejemplo, a una misma ciudad, como Amberes, que a veces aparece escrita en español y otras en su idioma original; o una persona, como un monarca, al que a veces una noticia se refiere por su nombre, Luis XIII, y otra por su cargo, el rey de Francia. En definitiva, hemos podido razonar sobre esa base de datos que conforma la prensa de la Edad Moderna, convirtiendo la información en conocimiento.

Por otra parte, la percepción del nacimiento del discurso periodístico como una *preterinternet* (una red implícita que facilita la difusión de noticias que no conservan o respetan un formato predefinido, contenido sintáctico, sujeta a revisiones no controladas y que se alimenta de fuentes oficiales, objetivas, intencionadas, sesgadas, etc.) ha obligado al investigador en Representación del Conocimiento a pensar en las ontologías como elementos útiles para la aproximación al fenómeno en vez de una herramienta de estandarización y semantización clásica. El uso de conceptos que no tienen un perfil formalmente definido nos ha permitido comprender el proceso como una iteración adaptativa que persigue englobar el conocimiento experto sobre el tema en una(s) ontología(s) que aproximen dicho conocimiento. La naturaleza de ambos desafíos nos ha llevado a una revisión de los métodos clásicos de construcción de ontologías. Así, el propio proceso de creación de la ontología nos ha permitido desvelar, gracias a la realización intensiva de entrevistas entre ambos grupos de investigación (propias de la Ingeniería del Conocimiento), rasgos en común en el diseño de la portada de estos impresos, así como regularidades en la redacción del texto, que nos permiten avanzar en nuestro conocimiento de las series periodísticas de la primera prensa española.

REFERENCIAS BIBLIOGRÁFICAS

ARANO, S. (2005). “Los tesauros y las ontologías en la Biblioteconomía y la Documentación”. *Hipertext.net. Anuario Académico sobre Documentación Digital y Comunicación Interactiva*, 3. Recuperado de https://www.upf.edu/hipertextnet/numero-3/tesauros.html el 02/03/2017.

BORST, W. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Ph.D. Dissertation. Enschede: University of Twente. Recuperado de http://eprints.eemcs.utwente.nl/17377/01/t0000004.pdf el 02/03/2017.

BURDICK, A., DRUCKER, J., LUNENFELD, P., PRESNER, T. y SCHNAPP, J. (2012). *Digital Humanities*. Cambridge: The MIT Press.

BURNARD, L. (2014). *What is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. Marseille: OpenEdition Press. Recuperado de http://books.openedition.org/oep/426 el 02/03/2017.

CODINA, L. y PEDRAZA-JIMÉNEZ, R. (2011). "Tesauros y ontologías en sistemas de información documental". *El profesional de la información*, 20. 5, 555-563.

____ (2015). *Taxonomías y ontologías: qué son y cómo se aplican a medios de comunicación*. Recuperado de http://www.lluiscodina.com/?p=2289 el 02/03/2017.

COHEN, D.J. y ROSENZWEIG, R. (2005). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. Recuperado de http://chnm.gmu.edu/digitalhistory el 02/03/2017.

CONBOY, M. (2004). *Journalism: A Critical History*. London: Sage.

DEEGAN, M. y TANNER, S. (eds.) (2006). *Digital Preservation*. London: Nealschuman.

DÍAZ NOCI, J. y HOYO, M. (2003). *El nacimiento del periodismo vasco. Gacetas donostiarras de los siglos XVII y XVIII*. San Sebastián: Sociedad de Estudios Vascos.

DOMÍNGUEZ GUZMÁN, A. (1992). *La imprenta en Sevilla en el siglo XVII. 1601-1650* (*Catálogo y análisis de su producción*). Sevilla: Universidad de Sevilla.

ESPEJO, C. (2008). "El impresor sevillano Juan Gómez de Blas y los orígenes de la prensa periódica: la Gazeta Nueva de Sevilla (1661-1667)". *Zer. Revista de Estudios de la Comunicación*, 13. 25, 243-267.

____ (2012). "Un marco de interpretación para el periodismo europeo en la primera Edad Moderna". En *La aparición del periodismo en Europa. Comunicación y propaganda en el Barroco,* R. Chartier y C. Espejo Cala (eds.), 103-126. Madrid: Marcial Pons.

ETTINGHAUSEN, H. (2015). "How the Press Began: The Pre-Periodical Printed News in Early Modern Europe.". *Janus: Estudios sobre el Siglo de Oro*, 3. A Coruña: SIELAE. Recuperado de http://www.janusdigital.es/anexo.htm;jsessionid=55C6E6E10D1F457F600C289374C0B178?id=7 el 02/03/2017.

FERNÁNDEZ TRAVIESO, C. (2013). *Estudio de Codificación XML/TEI para Relaciones de Sucesos Españolas*. A Coruña: SIELAE. Recuperado de http://www.bidiso.es/sielae/upload/estaticas/file/FTXMLTEIISBN2pr.pdf *el* 02/03/2017.

GRUBER, T.R. (1993). "A Translation Approach to Portable Ontologies". *Knowledge Acquisition*, 5.2, 199-220.

GUILLAMET, J. (2003). *Els origins de la premsa a Catalunya. Catàleg de periòdics antics* (*1641-1833*). Barcelona: Ajuntament de Barcelona.

____ (2012). "Las bases históricas del periodismo: una mirada actual sobre la prensa del Barroco". En *La aparición del periodismo en Europa. Comunicación y propaganda en el Barroco,* R. Chartier y C. Espejo Cala (eds.), 263-276. Madrid: Marcial Pons.

HAFFEMAYER, S. (2002). *L'information dans la France du XVIIe siècle. La Gazette de Renaudot de 1647 à 1663*. Paris: Honoré Champion.

HIGGS, E. (ed.) (1998). *History and Electronic Artefacts*. Oxford: Oxford University Press.

KOOPMANS, J.W. (ed.) (2005). *News and Politics in Early Modern Europe* (*1500-1800*), 13. Leuven: Peeters Publishers.

LORENTE CASAFONT, M. (2005). "Ontología sobre economía y recuperación de información". *Hipertext.net. Anuario Académico sobre Documentación Digital y Comunicación Interactiva*, 3. Recuperado de https://www.upf.edu/hipertextnet/numero-3/ontologia_ri.html el 02/03/2017.

NOY, N.F. y MCGUINNESS, D.L. (2005). *Desarrollo de Ontologías-101: Guía para crear tu primera ontología*. Stanford: Stanford University.

PEDRAZA-JIMÉNEZ, R., CODINA, L. y ROVIRA, C. (2007). "Web semántica y ontologías en el procesamiento de la información documental". *El profesional de la información*, 16. 6, 569-578.

PETTEGREE, A. (2014). *The Invention of the News: How the World came to know About Itself*. Connecticut: Yale University Press.

PIERAZZO, E. (2014). "Digital Documentary Editions and the Others". *Scholarly Editing: The Annual of the Association for Documentary Editing*, 35. Recuperado de http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html el 02/03/2017.

POPKIN, J. (2005). "New Perspectives on the Early Modern European Press". En *News and Politics in Early Modern Europe* (*1500-1800*), J. W. Koopmans (ed.), 1-27. Leuven: Peeters Publishers.

RAMSAY, S. (2011). *Who's in and Who's out*. Recuperado de http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out/ el 02/03/2017.

RAYMOND, J. (2012). "El rostro europeo del periodismo inglés". En *La aparición del periodismo en Europa. Comunicación y propaganda en el Barroco,* R. Chartier y C. Espejo Cala (eds.), 177-206. Madrid: Marcial Pons.

STUDER, S., BENJAMINS, R. y FENSEL, D. (1998). "Knowledge Engineering: Principles and Methods". *Data and Knowledge Engineering*, 25, 161-197.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam-Philadelphia: John Benjamins Publishing Company.