JORNADAS POSICIONAMIENTO es
UNIVERSIDAD DE ALICANTE
DEL 5 AL 7 DE MAYO
El posicionamiento en Internet de instituciones culturales, científicas y educativas

Posicionamiento de contenidos textuales.

5 de mayo. Sala de Juntas de la Facultad de Derecho.

«Desarrollo de entornos semánticos y técnicas de mercado para la agregación de valor a los contenidos», de Joaquín Rodríguez y José Antonio Millán (Residencia de Estudiantes).

1. Planteamiento global: razones para el desarrollo de las tecnologías.

Posicionarse hoy en el vasto mar de la web, destacar entre el aluvión de materiales y sitios en inglés y ocupar el lugar que nuestros contenidos merecen, no es una tarea sencilla. Requiere diseñar una estrategia en la que se integren y complementen, al menos, cuatro factores generales: los económicos y jurídicos, el plan de explotación y la gestión de los derechos de la propiedad intelectual; la detección y análisis del público objetivo, de los potenciales usuarios, a los que va destinado nuestro esfuerzo; la difusión de nuestros logros y nuestros trabajos y la organización de seminarios y debates en torno a ellos; por último, *last but not least*, el desarrollo, instalación y uso de las tecnologías que nos permitan añadir a nuestros contenidos un valor –una riqueza, estimación y cotización, por tanto- que no alcanzarían en sí mismos si se consideraran aisladamente, si se manejara u ofreciera cada documento separadamente, sin resaltar de alguna manera toda su polisemia, el nudo de relaciones semánticas del que forma seguramente parte. Ofrecer a un usuario la posibilidad no sólo de recuperar un documento o una sucesión de documentos individuales clausurados en su significación sino, yendo más allá, brindarle la red de conexiones semánticas de la que cada documento forma parte como un nudo o un enlace, es un avance tan importante y elocuente respecto al estado de aluvión y amorfismo de la información en la web actual que debería contribuir a que nuestra web estuviera posicionada entre las más significativas de la red y, en consecuencia, entre las más visitadas y consultadas.

Nuestra estrategia podría denominarse, por eso, *añadir significado, añadir valor*, y esa adición o suma de esencia y de valoración debe ser el resultado de una táctica global y algo compleja que abarque, de más a menos, la creación de un entorno semántico propio de las humanidades mediante la concepción de la ontología correspondiente que describa las clases, instancias y relaciones en las que quepa cada hecho; el marcado asistido de cada uno de los documentos de nuestros archivos y de los hechos que contengan y, finalmente, la implantación e instalación de un buscador con morfologías lingüísticas específicas capaz de realizar discriminaciones y sugerir relaciones inusitadas para los buscadores comerciales al uso.

2. Onto-H y la semántica de las humanidades.

La Residencia de Estudiantes posee y custodia un gran archivo de la memoria, el denominado *Archivo de la Edad de Plata,* origen del actual proyecto de la *Web de la Edad de Plata*. Durante mucho tiempo, como en todas las bibliotecas y archivos, el esfuerzo se centró en la filmación e incluso digitalización de gran parte de los documentos y en su inclusión y grabación en bases de datos en CDs o en línea a través de la web (www.archivovirtual.org). El tipo de búsqueda que permitían las tecnologías hasta hace poco tiempo se basaba en interrogaciones simples mediante palabras específicas que encontraban su posible ocurrencia en los documentos analizados. Si bien esa clase de recuperación podía llegar a satisfacer a usuarios no especializados, resultaba manifiestamente insuficiente para profesionales con intereses específicos en las relaciones potenciales que pudieran preexistir entre artistas, escritores, obras, allegados y amigos, escenarios y

situaciones, etc. Es decir, había que dar un paso hacia la web semántica, hacia la construcción de un dominio especifico de las humanidades mediante el desarrollo de las tecnologías web específicas. De esa forma pretendíamos dar respuesta a dos imperativos entrelazados: asegurar la máxima y mejor difusión del legado histórico que se custodia mediante el posicionamiento más riguroso y significativo posible.

En resumidas cuentas, las tareas a abordar fueron:

- La construcción de una ontología provisional (ninguna descripción de la realidad puede ser sino provisional) de las humanidades a partir de la contribución de profesionales de diversos ámbitos y el uso de multitud de fuentes;
- Utilizar esa ontología para la anotación semántica de los contenidos mediante una herramienta de edición capaz de asistir y realizar sugerencias pertinentes a los anotadores;
- Publicar los contenidos en el sitio web correspondiente proporcionando funcionalidades específicas para la navegación semántica;
- Integrarla en un buscador.

También, y como consecuencia derivada del trabajo realizado, elaborar una metodología de trabajo que pueda servir a otras instituciones culturales del ámbito de las humanidades para explotar sus contenidos de la misma forma en la web semántica, colaborar, en suma, a que se ubiquen en el lugar que les corresponde en la red.

2.1. El dominio específico de las humanidades y el contexto histórico de la edad de plata.

La materia histórica concreta que ocupa buena parte del trabajo de la Residencia de Estudiantes es la de la *Edad de Plata*, la del movimiento artístico e intelectual que se concretó, sobre todo, entre los años 1868 y 1936 en torno a ella y a la Institución Libre de Enseñanza y al conjunto de figuras señeras que lo habitaron y dirigieron. Es fácil comprender que para extraer el significado más pertinente de cualquiera de los hechos sucedidos no basta con fijarse en el acontecimiento aislado o en el individuo concreto, sino que en este periodo se comprueba fácilmente hasta qué punto es importante el contexto histórico y la tupida e intrincada red de relaciones entre los actores y los sucesos –un pintor, como Dalí, dentro de un movimiento, como el surrealismo, realiza una serie de dibujos, como los de *los putrefactos*, en colaboración con algunos geniales amigos, como García Lorca, en un lugar específico, la Residencia de Estudiantes, en un periodo temporal perfectamente delimitado, el de 1925 a 1926-. ¿Cómo podría un buscador tradicional, basado en el uso de palabras clave, alcanzar el significado pleno y las interrelaciones de todos estos conceptos? Existen, sin duda, soluciones intermedias basadas en las diversas estructuras que ofrecen las bases de datos relacionales y en el uso de tesauros especializados que ofrecen conjuntos de sinónimos, antónimos, epónimos, etc., insuficientes si se pretende ahondar en la búsqueda de relaciones y, por tanto, de significados. Las ontologías vienen a solventar ese problema a costa de un no pequeño esfuerzo que requiere la descripción de un dominio de la realidad específico –el de las humanidades, en el caso que nos ocupa- en donde puedan observarse y recuperarse las relaciones que se establecen entre los conceptos que lo forman.

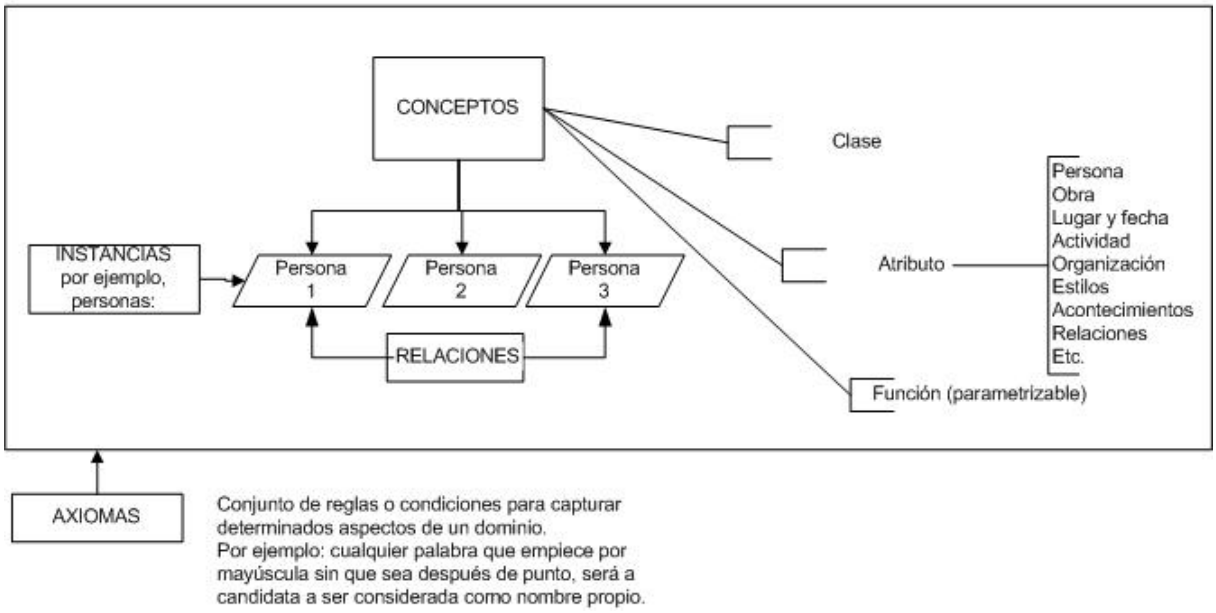
2.2. onto-h o la ontología de las humanidades.

Describir un dominio específico de la realidad de manera que múltiples interlocutores compartan los mismos significados o los comprenda de la misma manera, es el propósito de una *ontología*. La reminiscencia filosófica no es del todo gratuita porque se trata de detallar y pormenorizar el ser de algo –el dominio de las humanidades, en este caso- mediante la especificación explícita y formal de una conceptualización compartida. Podríamos descomponer la última frase y analizar el significado de sus términos para cobrar conciencia plena del alcance de una ontología: debe ser explícita, porque tanto los conceptos usados como las reglas de uso deben ser claras y unívocas; debe ser formal, porque nuestras

máquinas deberán ser capaces de entenderla; debe ser compartida, porque una ontología no puede construirse a partir de la experiencia aislada de una sola persona; y, por último, debe ser el resultado de la concepción y modelización compartidas de la realidad.

De esa manera cabe comenzar a desarrollar la ontología específica basándose en el desarrollo de conceptos que, a su vez, se organicen en todas las taxonomías y conjuntos de atributos que esa realidad requiera. Esquemáticamente cabría dibujar de la siguiente manera el conjunto de nociones que intervienen en la construcción de la ontología:

FORMA DE LA ONTOLOGÍA

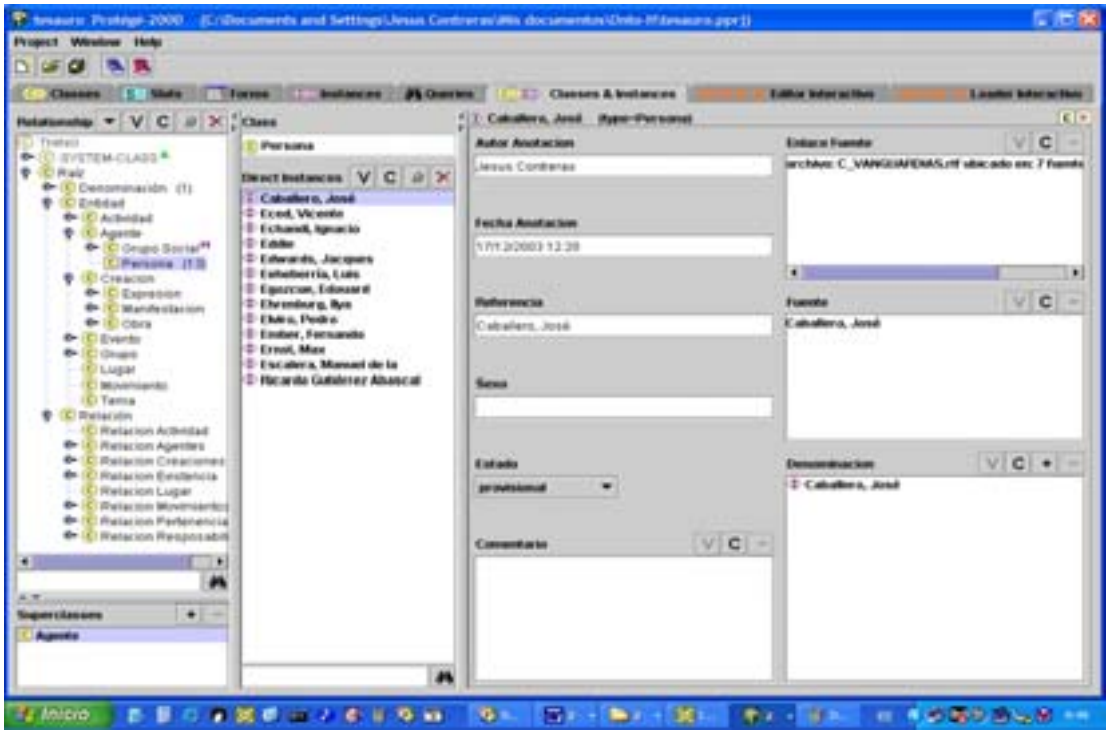


Para la de las humanidades comenzamos a trabajar con el cuadro siguiente basándonos para su construcción en el planteamiento de las denominadas preguntas sobre la idoneidad (*Competency Questions Methodology*[1]) o capacidad para responder al conjunto de cuestiones que sobre ese ámbito del conocimiento debería poder responder una ontología –las preguntas serían, claro, infinitas y por eso mismo la ontología será siempre una aproximación-.

Ejemplos de conceptos	*Competency question*
Persona	• ¿Quién escribió *Donde habite el olvido*? • ¿Quiénes fueron los autores que escribieron en *La Gaceta Literaria?* • ¿Qué pintores estuvieron becados en la *Academia de Roma* durante los años 20?
Obra	• ¿Dónde publicó su primer libro Rafael Alberti? • ¿Qué publicaciones surgieron del I Congreso de Escritores? • ¿En qué obras y qué autores aparece mencionado Pepín Bello?
Lugares y fechas	• ¿Dónde tuvo su librería León Sánchez Cuesta? • ¿Qué poetas nacieron en Madrid entre los años 1890 y 1900? • ¿Cuándo abandonaron la Residencia Lorca y Dalí?
Actividades	• ¿Qué cargos o puestos ocupó Benjamín Jarnés? • ¿Cuándo dio su primer concierto Fernando Remacha y Villar?
Organizaciones	• ¿Quién dirigió la Institución Libre de Enseñanza? • ¿Qué personajes relevantes vivieron en la Residencia de Estudiantes? • ¿Quiénes formaron el grupo de las Misiones Pedagógicas?
Estilos	• ¿Qué pintores formaron parte del realismo mágico madrileño? • ¿Qué obras caerían dentro del ámbito del ultraísmo? • ¿Puede quedar Dalí agrupado en algún movimiento?
Acontecimientos	• ¿En qué representaciones de La Barraca participó García Lorca? • ¿En qué congresos participaron los liberales orteguianos?
Relaciones	• ¿Qué tipo de relación mantuvo Pepín Bello con García Lorca? • ¿Qué relación tuvo Tomás Segovia con México? • ¿Qué clase de vínculo sostuvo Antonio Espina con el grupo SIC?

El planteamiento específico de las preguntas formuladas debería poder ser substituido por interrogantes genéricos del tipo "¿Dónde publicó su primer libro X?" o, también, "¿En qué obras aparece mencionado X?" o, incluso, "¿Qué relaciones mantuvieron X con X?". Esa clase de preguntas son del tipo que un buscador al uso no puede responder porque no puede entenderlas y porque no dispone de la trama semántica que enlaza los documentos que han sido previamente marcados y perfilados guardando y resaltando toda su polisemia. Muy sintéticamente, por tanto, el conjunto de los posibles conceptos[2] (estudios, profesión, institución, organización académica, obra literaria o musical, exposiciones, etc.) se expande a través de diversos tipos de atributos que cualifican tanto su alcance y significado intrínseco como las relaciones que pueda mantener con otros conceptos de primera clase o primer orden (por ejemplo, la relación que pudiera haber mantenido un autor teatral con un teatro madrileño en los años 20 sería del tipo "trabajó_en", donde las fechas de inicio y finalización de su vínculo no afectarían a la esencia del personaje ni de la institución).

El modelo de nuestra ontología de las humanidades tiene dos dimensiones, la técnica y la intelectual: para avanzar en el primer frente, comenzamos a trabajar con el editor de ontologías Protégé, un software libre desarrollado por la Universidad de Standford[3] bajo licencia Mozilla[4]. A partir del código fuente del software, realizamos todas las modificaciones necesarias para adaptarlo a nuestros propósitos; para avanzar en la descripción misma de la ontología, utilizamos algunas ontologías de carácter general capaces de modelar realidades complejas como las de personas, organizaciones, acontecimientos, etc., del tipo SUO[5], *Generalized Upper Model*[6], *WordNet*[7] o *CyCorp*[8]. Dentro del dominio específico de las humanidades encontramos precedentes aprovechables en IFLA[9] y MARC[10].



Pantalla de trabajo de la Ontología de las Humanidades

El árbol de la ontología visto a lo largo de la relación de herencia es el siguiente:

- Entidad
  - Evento
    - Exposición
  - Grupo
    - Grupo Social
      - Empresa
      - Institución
      - Organización Académica
  - Lugar
  - Movimiento
  - Relación
    - Relación Actividad
    - Relación Agentes
      - Relación Grupos
      - Relación Persona Grupo
      - Relación Personas

        - Relación Familiar
        - Relación Profesional
        - Relación Sentimental
    - Relación Creaciones
        - Relación Expresiones
        - Relación Manifestaciones
        - Relación Obras
            - Relación Obras Parte De
    - Relación Existencia
        - Relación Existencia Persona
    - Relación Lugar
    - Relación Movimientos
        - Relación Agente Movimiento
        - Relación Movimiento
        - Relación Obra Movimiento
    - Relación Pertenencia
        - Relación Persona Grupo
    - Relación Resposabilidad
        - Relación Creación
        - Relación Produccion
        - Relación Realizacion
- Tema

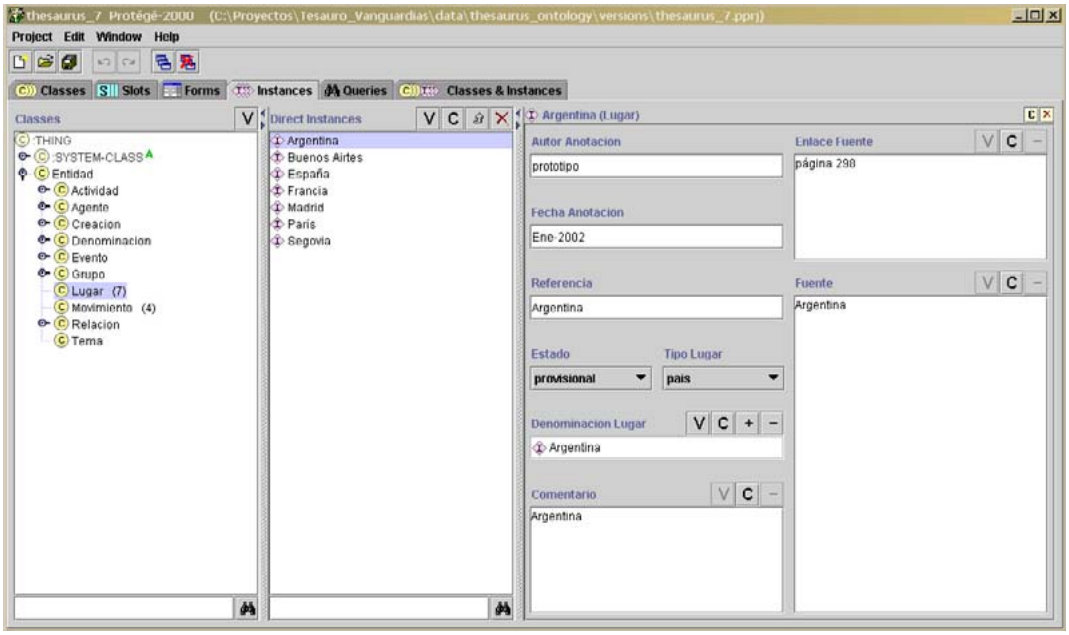
3. Las técnicas de marcado y edición asistida de documentos del Archivo de la Edad de Plata.

Los contenidos de los millones de documentos que se albergan en la Residencia de Estudiantes pueden haber sido ya estructurados –siguiendo las técnicas catalográficas habituales-, semi-estructurados o, simplemente, no tener estructura todavía de ningún tipo. La ontología construida recibirá todos esos contenidos y deberá ser capaz, una vez concluido el proceso de marcado y anotación, de devolver documentos semánticamente enriquecidos. El proceso de anotación, uno de los puntos fundamentales de este trabajo, puede satisfacerse de diversas formas que van desde lo puramente manual hasta el uso de herramientas de edición asistida e, idealmente, hasta la anotación completamente automatizada, métodos de trabajo cuya aplicación parece depender estrechamente del grado de estructuración previa de los contenidos (parece claro que un grado más preciso de estructuración permite que los automatismos del marcado funcionen con mayor exactitud[11]).

Sin lugar a dudas, es en el proceso de anotación de los contenidos donde cualquier institución cultural encontrará la máxima dificultad y los mayores escollos. El editor semiautomático que debe asistir a los profesionales de la anotación en estas tareas, es un *plug-in* que pertenece a la herramienta de Protégé. Tal como se muestra en la figura siguiente, el editor carga el texto que debe anotarse y permite realizar todas las operaciones de edición necesarias sobre la ontología y las instancias, además de otorgar al operario la posibilidad siempre abierta de crear sus propias instancias mediante la sencilla técnica de “arrastrar y soltar”.

Pantalla del editor asistido

Tal como se muestra en la pantalla siguiente, si una instancia no existe y es necesario crearla, basta con arrastrarla al concepto de la ontología al que corresponda para que se genere.



Pantalla de creación de la instancia

O R G A N I Z A N :

Universitat d'Alacant
Universidad de Alicante
Instituto Cervantes
CASA DE AMÉRICA
FUNDACIÓN BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES
CINDOC Centro de Información y Documentación Científica
BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES www.cervantesvirtual.com

En el proceso de anotación no se modifica o altera el contenido textual original sino que genera un vínculo entre la cadena de texto a partir de la que crea la instancia y la instancia misma de manera que a partir de esa "inferencia" puntual se invierten los términos y la cadena de texto se convierte en una ocurrencia o caso concreto de la instancia general. El anotador humano tiene control pleno sobre la herramienta de manera que en función del texto que analice y anote deberá decidir, si un nombre, por ejemplo, es una ocurrencia más de la misma instancia o conviene calificarla como una instancia nueva. El editor asistido, en previsión de los posibles errores que puedan cometerse en el proceso de marcado, dota al anotador humano de herramientas semiautomáticas que realizan sugerencias de adscripción o son capaces, incluso, de generar instancias concatenadas dentro del árbol.

Para el usuario de la Web de la Edad de Plata la ontología de las humanidades no será otra cosa que un recurso omnipresente pero invisible, disponible, en realidad, mediante el uso de un buscador que lanzará consultas a la ontología.

4. Conclusiones: vías de desarrollo y posicionamiento futuros.

En nuestro trabajo inmediato y con el horizonte siempre presente de ocupar la posición que la Residencia de Estudiantes merece dentro de las instituciones culturales del ámbito hispanohablante, abordaremos varias tareas concatenadas:

- depurar nuestra ontología después de que en el periodo de pruebas se hayan producido una serie de contingencias que deben mejorarse;
- comenzar un proceso de anotación masivo de contenidos mediante el uso de la herramienta de edición asistida;
- seguir desarrollando la herramienta de explotación que nos permitirá vincular definitivamente nuestro buscador con nuestra ontología para ofrecer a nuestros usuarios los mejores resultados posibles;
- realizar una campaña de comunicación a nuestros potenciales usuarios, por último, en la que sepamos transmitirles que hemos realizado una ingente tarea de enriquecimiento semántico de nuestros materiales y, por tanto, de adición de valor.

---

[1] M. Uschold y M. Gruninger. 1996. "Ontologies: principles, methods and applications, en *Knowledge Engineering Review*, 11 (2): 93-155.
[2] Para una visión completa de la ontología debe consultarse Juan Manuel Dodero, Jesús Contreras, Richard Benjamins, "Test Case Ontology Specification Cultural Tour". D9.2, *Esperonto Project*, www.esperonto.net.
[3] http://protege.stanford.edu/
[4] http://www.mozilla.org/MPL/
[5] SUO Standard Upper Ontology http://suo.ieee.org/
[6] Generalized Upper Model: http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html
[7] WordNet: http://www.cogsci.princeton.edu/~wn/
[8] CyCorp: http://www.cyc.com/
[9] Federation of Library Associations and Institutions: http://www.ifla.org
[10] MARC http://www.loc.gov/marc/
[11] Contreras et al. D31: Annotation Tools and Services, Esperonto Project: www.esperonto.net