¿QUÉ PUEDE HACER LA WEB SEMÁNTICA POR LA BIOLOGÍA?

Mikel Egaña Aranguren - http://www.mikeleganaaranguren.com

INTRODUCCIÓN

En realidad, la pregunta habría que formularla al revés: ¿Qué puede hacer la Biología por la web semántica? La web semántica es una visión de lo que debería ser la web (o internet, o World Wide Web, ...) pero todavía no lo es. Todavía no lo es por que implantar una nueva tecnología, tan distinta de lo que manejamos hoy en día, requiere de grandes esfuerzos por parte de las comunidades de pioneros que primero adoptan esa tecnología, con todos sus riesgos. La Bioinformática (la parte de la Biología que estudia como manejar los datos Biológicos de una manera eficiente) es una de las pocas disciplinas que se ha atrevido a dar ese paso, beneficiándose de la tecnología de la web semántica y a la vez facilitando una ejército de "cobayas" a los impulsores de la misma, que de otra manera no hubiesen aplicado esa tecnología a ningún problema real. Si la web semántica llega a implantarse algún día, será en gran parte gracias a la Biología, sobretodo a la Biología Molecular y al Bioinformática. Este artículo intenta dilucidar los porqués de ese maridaje, mostrando a la vez los atractivos de una futura web semántica, no sólo aplicada a la Biología.

LA WEB SEMÁNTICA

Tim Berners Lee, un físico que trabajaba en el prestigioso CERN (Laboratorio Europeo de Física de Partículas de Ginebra), publicó a finales de los 80 el principio básico de la web: para facilitar a los científicos compartir artículos y leerlos visitando otros artículos a la vez, se inventó el ya famoso *enlace*. El mecanismo básico era sencillo y robusto: ciertos estándares como HTTP y HTML permitían incluir fácilmente enlaces a otros documentos dentro del documento que el científico estaba leyendo. Así, un científico podía visitar todos los artículos relacionados, navegando de enlace en enlace (y por tanto de documento a documento) por la red (la red existía antes que la web, aunque mucha gente piense lo contrario). La idea caló tan hondo que ese mismo sistema se usa hoy en día para todo tipo de documentos (páginas web) enlazados en una maraña que no para de crecer y hacerse más compleja. Una vez establecido el marco tecnológico básico, el fenómeno no ha parado de crecer, con buscadores como google, blogs (bitácoras), foros, catálogos fotográficos, enciclopedias colaborativas, y todo tipo de espacios cada vez más interactivos y sociales. Tim Berners Lee fundó el W3C (World Wide Web Consortium: http: //www.w3.org/), una fundación que establece recomendaciones oficiales de como deberían ser los protocolos en la red, para mantener una estructura los más abierta y eficiente posible, ya que precisamente el invento de Berners Lee se impuso con tal facilidad debido a que era una estándar completamente abierto y no propietario.

Pero semejante *mare magnum* de información es más inútil de lo que parece: por ejemplo, ¿De qué nos sirve tener acceso a cientos de páginas web de aerolíneas de bajo coste, si no tenemos tiempo de comparar los precios? Es decir, hay un flujo inmenso de información, pero la información útil, que de ahora en adelante denominaremos "conocimiento"[1], sigue siendo mínima; a nosotros lo que nos importa son los mejores vuelos según nuestras preferencias, no todos los vuelos de todas las páginas web. Lo ideal sería que un buscador, con un interfaz sencillo, comparáse por nosotros no sólo los precios, sino todo tipo de parámetros sobre vuelos de bajo coste, y nos devolviese una lista de vuelos, evitándonos el tener que visitar cada página: nosotros introduciríamos nuestras preferencias y el buscador haría el resto. Es más, lo ideal sería que el buscador fuese totalmente general y pudiese dar ese tipo de servicios sin importar el conocimiento que estemos buscando, sea vuelos de bajo coste, vinos, servicios médicos, libros, información sobre un tema, citas con otras personas, etc. Incluso sería deseable que combinase conocimiento de diferentes recursos: podría cuadrar, autómaticamente, los vuelos que deberíamos comprar con las fechas de la gira de nuestro grupo preferido, sólo en una ciudades concretas y a un precio asequible.

Para ser justos, ya existen soluciones de este tipo, pero son soluciones completamente *ad hoc* y un programador con experiencia las tiene que producir para cada caso particular. Los protocolos que posibilitan la programación de esas soluciones generales (como nuestro hipotético buscador "sabelotodo") están en fase muy experimental y no se usan tan masivamente como HTML y HTTP, aunque ya han sido publicadso por el W3C. Como veremos más adelante, sí se usan en la disciplina que nos ocupa, la Biología.

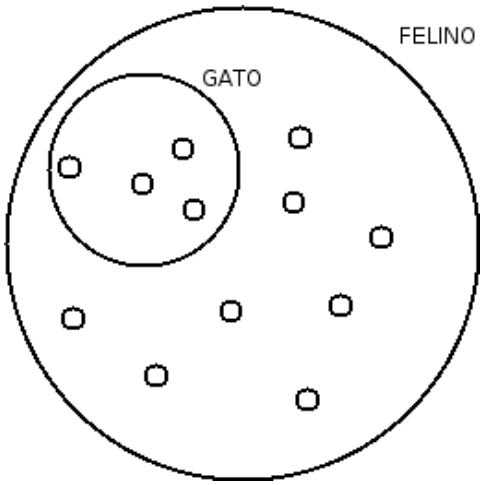
Para que el "buscador de conocimiento" sea posible, los protocolos actuales como HTML no bastan, ya que carecen de algo crucial: lo que a partir de aqui denominaremos "contenido semantico".

El contenido semántico de algo, en ciencias de la información, no es su "significado", como lo entendemos normalmente. La palabra "gato", por ejemplo, tiene un significado concreto, que podemos buscar en el diccionario. Sin embargo, cuando escribimos "gato" (la especie *Felis silvestris*) en la pantalla de un ordenador, para el sistema no es más que una cadena de caracteres. Para darle significado deberíamos codificar el concepto de una manera que el ordenador pueda "entenderlo", o por lo menos ser capaz de gestionar el concepto como tal.

Para que eso sea posible se puede codificar de muchas maneras; por ejemplo, podemos definir "gato" como el conjunto de individuos que forma parte de otro conjunto de individuos llamado "felino". El ordenador puede manejar los conjuntos, debido a que son constructos matemáticos bien definidos, y así gestionar el concepto. Para el ordenador, son los conjuntos lo que importa, y nosotros añadimos las etiquetas "gato" y "felino" para que esos conjuntos sean entendibles para los humanos:

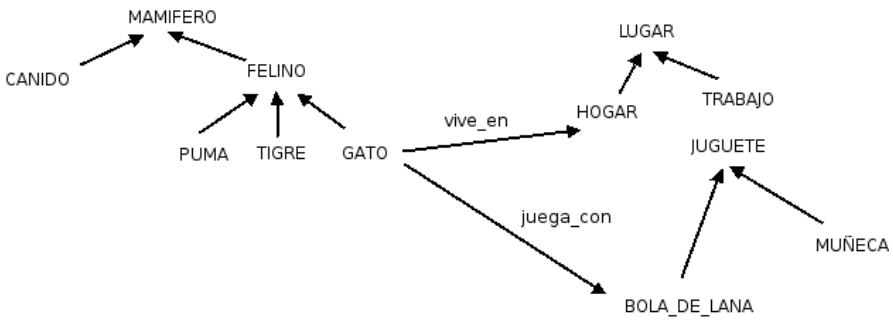
---

[1] La palabra "conocimiento" se usa aquí de una manera muy laxa; no nos referimos al conocimiento como el proceso cognitivo que cualquier humano lleva a cabo por ejemplo al leer este artículo. Definimos "conocimiento" como un grupo de conceptos y sus relaciones que conforman un modelo útil para llevar a cabo alguna función o describir algún dominio.



Los modelos del tipo que acabamos de describir se llaman "ontologías". En filosofía, ontología es la disciplina que estudia lo que es y existe. En informática, una ontología es un modelo matemático que describe un dominio de la realidad (para ser más exactos, el consenso al que han llegado una serie de personas sobre cómo quieren describir un dominio). En ese modelo, "gato" es un clase de individuos cuya superclase es la clase "felino", que también es un conjunto de individuos (la clase "gato" es una subclase o subconjunto de la clase "felino": todos los gatos son felinos, pero no todos los felinos son gatos). Es importante subrayar que los nombres de las clases son completamente triviales para el ordenador: esas mismas clases se podrían llamar "X" e "Y" y serían semánticamente equivalentes, para el ordenador lo que importa es la estructura.

Las ontologías son mucho más complejas que el ejemplo del gato y el felino, pueden definirse propiedades de las clases (juega con algo, vive en algún sitio, ...) y toda una plétora de constructos lógicos de toda clase:



Existen también una serie de programas, llamados razonadores, que son capazes de "analizar" una ontología y contestar "preguntas" que se les haga: ¿Todos los gatos juegan con bolas de lana?¿Cuál es el animal que juega con bolas de lana?¿Si algo juega con bolas de lana, es necesariamente un gato? Etc. (Estas "preguntas" se hacen a través de un interfaz gráfico con una

sintaxis concreta). Los razonadores también pueden deducir cosas a partir de una ontología, o combinar varias ontologías y hacer deducciones todavía más interesantes.

De modo que las ontologías posibilitan que los ordenadores accedan al contenido semántico, y los razonadores pueden gestionar ese contenido y hacer deducciones. Eso las hace muy buenas candidatas para construir la web semántica: volviendo al ejemplo de los vuelos, si hay una ontología que describe vuelos de bajos coste, y una ontología que describe los conciertos de la gira de mi grupo favorito, en teoría un razonador debería ser capaz de deducir para nosotros qué vuelos nos convienen.

Pero todavía falta un elemento clave en esta utopía llamada web semántica: hay que describir el contenido para que sea gestionable a través de las ontologías (o cualquier otra tecnología semántica). Eso se puede conseguir mediante los "metadatos", es decir, datos sobre los datos; es como poner etiquetas a las palabras. Si en nuestra página web sobre gatos, ponemos una "etiqueta" en el término "gato" que diga que es un GATO (el gato de la ontología), el resto lo hará la web semántica por nosotros. A ese proceso se le llama "anotar" la información, y a las "etiquetas", "anotaciones".

Nadie sabe a ciencia cierta si la web semántica llegará algún día. Pero está claro que en el camino está produciendo mucha tecnología muy útil, que ya se está usando en Biologia Molecular. El W3C ya ha propuesto un estándar para crear ontologías en la web, Web Ontology Language (OWL: http://www.w3.org/2004/OWL/) y existen muy buenos programas para crear ontologías (por ejemplo Protégé: http://protege.stanford.edu/) y razonadores (por ejemplo FaCT++: http://owl.man.ac.uk/factplusplus/).

Uno de los problemas de la web semántica es que es de muy difícil implementación, ya que necesita de mucho trabajo por parte de los usuarios: crear una página web es relativamente sencillo, pero crear una buena ontología no. Sin embargo, en dominios como la Biología Molecular, dónde hay usuarios dispuestos a crear ontologías y a anotar términos, un embrión de lo que podría llegar a ser la web semántica está en pleno desarrollo.

BIOLOGIA MOLECULAR: INFORMACION VS. CONOCIMIENTO

La Biología es una ciencia basada en el conocimiento y en la descripción más que en la pura abstracción. A diferencia de la física, por ejemplo, dónde se intenta buscar una ecuación que describa muchos procesos, en Biología Molecular describimos los procesos y eso forma el corpus de la disciplina: por ejemplo, lo más probable es que nunca llegemos a tener una "teoría general unficada" que describa de una manera universal el plegamiento de las proteínas, pero el saber como se pliegan todas y cada una de las proteínas es muy útil y nos permite avanzar en la Biología Molecular.

Eso hace que la recolección y gestión de la información de un modo eficiente y lo más automático posible sea de vital importancia en Biología Molecular. Sobretodo después de la revolución biotecnológica iniciada en los 80, ya que se producen cada vez más y mas datos

(sobretodo secuencias), pero eso no se traduce en "conocimiento" (como lo definiamos al principio): ¿De que nos sirve tener acceso simultáneo a miles de secuencias? Lo que necesitamos saber es en que procesos toman parte, cuáles son sus roles en esos procesos, en qué parte de la célula se localizan, con que otras secuencias interactuan y cómo, etc.

El mismo problema puede encontrarse en la literatura científica: hoy en día se publica como nunca, pero ese volumen de información no es manejable como conocimiento práctico: lo que necesitamos saber son los modelos que describen las publicaciones: por ejemplo, si una veintena de publicaciones tienen secuencias que se relacionan, lo que importa es el modelo de esa relación, no el proceso experimental con el que llegaron a esas conclusiones (probablemente también habrá científicos que estén interesados en el procedimiento experimental, y se encontrarán en una situación similar).

Enfrentados a este problema, los biólogos empezaron a dar pasos hacía una solución, y empezaron a usar ontologías. Muchas de esas ontologías se pueden encontrar en el proyecto Open Biomedical Ontologies (http://obo.sourceforge.net/). La más famosa es Gene Ontology (GO: http://geneontology.org/): GO desribe las propiedades de genes (la localización celular, la función molecular y el proceso biológico). GO provee un sistema para integrar *de facto* diferentes bases de datos que tengan entrada anotadas contra términos de GO. Por ejemplo, podemos buscar genes en esas bases de datos a través de los términos de GO, como "binding", obteniendo todos los genes de esa función molecular. También se puede usar GO para gestionar la información: si tengo un gen anotado a un termino de GO, puedo usar las relaciones que tiene ese término con otro términos para acceder a más información sobre el gen.

GO tuvo un éxito impresionante gracias a su simplicidad: sólo tiene dos tipos de relaciones ("is_a" y "part_of", es decir "es_un" y "part_de") y el formalismo semántico asociado a las ontología es virtualmente inexistente, de modo que los biológos se sienten muy cómodos con la ontología por que es muy intuitiva. Pero desde el punto de vista informático, GO es cuando menos "mejorable", por ejemplo portándola a OWL (un esfuerzo que ya está en marcha).

Otras ontologías Biológicas ya están siendo implementedas en OWL, como BioPAX, que describe "pathways" metabólicos (http://www.biopax.org/), o CCO, que describe el ciclo celular (Cell Cycle Ontology: http://www.cellcycleontology.org). Por ejemplo se espera que CCO sea capaz de generar nuevas hipótesis sobre el ciclo celular, a parte de integrar datos de diferentes bases de datos y contestar a las "preguntas" que tengan mucho biólogos sobre el ciclo celular.

También hay otros proyectos basados en OWL: PhosPhabase (http://www.bioinf.manchester.ac.uk/phosphabase/index.html), por ejemplo, hace uso de OWL y un razonador para clasificar automáticamente familias de fosfatasas, con nuevos resultados bastante interesantes.

Todos estos proyectos y ontologías no son un exponente de la web semántica, pero son una demostración de que la tecnología es útil e implementable. ¿Dará la web semántica el siguiente paso, de la Biología Molecular a los demás usuarios? No lo sabemos. Quizás la web semántica

nunca se implante y se limite a ciertas disciplinas como la Biología, pero en cualquier caso habrá contribuido a una mejor gestión de la información en la Biología Molecular, que es algo muy positivo en sí mismo.

PARA SABER MÁS

La wikipedia (http://es.wikipedia.org) ofrece entradas sobre todos los temas tratados en este artículo.

Un artículo que describe las diferencias entre informáticos y biólogos en cuanto a los formalismos para construir ontologías se puede encontrar en: http://www.biomedcentral.com/1471-2105/8/57.

El artículo original que describe la web semántica, escrito por Tim Berners Lee, se puede encontrar en el volumen de Mayo del 2001 de la revista Scientific American (Investigación y Ciencia).

El W3C también tiene página en español: http://www.w3c.es/.