8ª práctica de laboratorio de SIW
“Información semántica”

Motivación

“Piano piano si arriva lontano”

La mejor forma de valorar las bondades de las tecnologías semánticas es viendo su aplicación práctica en casos reales. Para ello iremos viendo pequeños ejemplos así como herramientas que nos servirán para introducir conceptos como los microdatos en HTML5 o JSON-LD. En consecuencia, esta sesión práctica estará muy guiada y tan solo al final se planteará un ejercicio sencillo para su entrega. Se recomienda seguir los pasos en orden y al ritmo que marque el profesor:

Texto < Hipertexto < Microdatos

Visita el siguiente documento. No trates de traducirlo ni de decodificarlo, esta es la “visión” que tiene una máquina de un **texto plano**; es posible que haya un sentido subyacente pero no puede extraerse de manera inmediata. Con acceso a más texto plano podrían extraerse cadenas significativas, pero seguiría sin haber una semántica accesible para la máquina.

Visita ahora el siguiente hipertexto. No visualices el código HTML, céntrate tan sólo en los aspectos más obvios: los enlaces y las cadenas con énfasis. Esa sería la “visión” de una máquina de un **hipertexto básico**. Sigue sin disponer del significado de ese texto pero sí “sabe” que las siguientes cadenas son importantes y van, además, asociadas a una URL:

- `bsfb pg jñufsftu (http://danigayo.info/research)`
- `tpdjbm nfejb sftfbsdi (http://danigayo.info/research/#socialmediaresearch)`
- `J ibwf qvcmjtife (http://danigayo.info/publications)`
- `JFFF Jñufsñfu Dpnqvujñh (http://danigayo.info/publications/detail.php?id=...)`
- `JFFF Nvmujnfejb (http://danigayo.info/publications/detail.php?id=...)`

- tqfdjbm jttvf pg Jñufsñfu Sftfbsdi pñ uif qsfejdujwf qpxfs pg tpdjbm nfejb (http://www.emeraldinsight.com/toc/intr/23/5)
- dibqufs pñ Qpmjujdbm Pqjñjpñ (http://danigayo.info/publications/detail.php?id=...)

Con independencia de lo que signifique ese texto o las URLs un buscador "sabría" que debería asociarlo como metadatos a las URLs indicadas ya que tal vez pueda ser útil para resolver consultas; después de todo, ya hay al menos un sitio donde se hace referencia a esas URLs con esos términos.

Visita ahora el siguiente documento; se trata en apariencia del mismo hipertexto anterior pero hay diferencias. Antes de profundizar en el código fuente visita esta página web. Allí encontrarás dos herramientas: *Rich Results Test* y *Schema Markup Validator*. Aunque hay un cierto grado de solapamiento entre ambas lo cierto es que persiguen objetivos diferentes así que deberías probar la siguiente URL en las dos:

https://danigayo.info/ejemplos-SIW/02-microdata.html

Comprobarás que la primera indica que no hay ningún "rich result" en la misma (lo cual es cierto, no hay ninguno de estos tipos de contenido) mientras que la segunda sí es capaz de extraer información "accionable" para una máquina gracias a los microdatos incrustados en el HTML:

- Ha descubierto que en ese texto se mencionan 4 publicaciones periódicas:
  - Dpnnvñjdbujpñt pg uif BDN
  - JFFF Jñufsñfu Dpnqvujñh
  - JFFF Nvmujnfejb
  - Jñufsñfu Sftfbsdi
- Ha descubierto un capítulo de libro:
  - titulado: Qpmjujdbm Pqjñjpñ
  - en el libro: Uxjuufs: B Ejhjubm Tpdjptdpqf
  - con ISBN: 9781107500075
  - publicado por la editorial: Dbncsjehf Vñjwfstjuz Qsftt
- Ha descubierto además una persona:
  - Ebñjfm Hbzp-Bwfmmp
  - que trabaja de bttpdjbuf qspgfttps
  - en un college o universidad llamada Vñjwfstjuz pg Pwjfep
  - dentro del departamento Efqbsunfñu pg Dpnqvufs Tdjfñdf

Si observas el código fuente podrás ver cómo toda esa información se incrustó gracias a unas propiedades HTML que tal vez no conozcas:

- `itemscope`
- `itemid`
- `itemtype`
- `itemprop`

Toda la información necesaria para trabajar con esta extensión de HTML la tienes en el documento *HTML Microdata*[1]. Sin embargo, para comenzar te basta saber lo siguiente:

- La primera propiedad, **`itemscope`**, indica que todo el código HTML contenido dentro del elemento que la usa (habitualmente un `div` o un `span`) hace referencia a un único ítem, esto es, un objeto físico o lógico (p.ej., una persona, una película, una organización, una canción, …)
- La propiedad **`itemid`** es opcional pero, de aparecer, siempre va asociada a un `itemscope`. Dicha propiedad recibe una URI válida para identificar al ítem en cuestión. En el ejemplo que estamos usando aparecen una URL de una página personal y dos ISBNs usando URNs[2].
- La propiedad **`itemprop`** va asociada a un elemento HTML anidado dentro de un elemento con `itemscope` (habitualmente un `div` o un `span`) e indica que el código HTML que encierra se refiere a una propiedad del `itemscope` que lo contiene. Obsérvese en el ejemplo que se puede usar en combinación con `itemscope` para definir un ítem dentro de otro ítem.
- Por último, **`itemtype`** permite indicar (con una URL absoluta válida) el tipo (o clase si se quiere) al que pertenece un `itemscope` dado.

Surge entonces una cuestión no trivial: ¿de dónde salen los tipos para `itemtype` y las propiedades para `itemprop`?

El ejemplo vuelve a darnos pistas importantes. Como se puede apreciar todos los tipos usados en el ejemplo (`Book`, `Chapter`, `Periodical`, `Person`, `CollegeOrUniversity`, y `Organization`) han sido definidos en `https://schema.org`[3].

---

[1] No, ahora mismo no todos los estándares web los lleva el W3C, algunos los lleva el WHATWG. Más info aquí: `https://github.com/w3c/whatwg-coord`

[2] Puedes encontrar aquí la lista de espacios de nombres en URNs y aquí información sobre IRIs, URIs, URLs, URNs, en qué se parecen y en qué se diferencian.

[3] Proyecto colaborativo con apoyo de empresas de búsqueda en la Web para crear y mantener vocabularios semánticos para la estructuración de información.

Si visitamos la URL de alguno de esos tipos, como `Person` veremos que aparecen las propiedades de dicho tipo, incluyendo `affiliation` y `jobTitle`. A su vez, al ser `Person` un subtipo/subclase de `Thing` hereda sus propiedades, como `name`.

JSON-LD

Visita el siguiente documento, se trata de la versión "descifrada" de los ejemplos que has estado usando hasta ahora. Para una máquina el texto (más allá de los metadatos) sigue siendo igual de indescifrable, pero a nosotros nos resulta más manejable.

Al igual que en el último ejemplo, en este caso se han usado microdatos HTML5 para etiquetar; eso supone una ventaja (no se usa ninguna tecnología externa) pero también un inconveniente. Por ejemplo, por la forma en que se ha redactado el texto ha sido posible dar el título del capítulo y del libro pero no es viable indicar el título de los artículos publicados en revistas.

Obviamente podría añadirse el título de las publicaciones, pero si nos ceñimos a microdatos tendría que hacerse en el propio texto y podría resultar muy verboso. ¿Habría alguna opción para añadir metadatos visibles para la máquina, pero invisibles para los lectores?

La opción más razonable es JSON-LD (JSON for Linking Data), aquí tienes una versión de los metadatos anteriores serializados en dicho formato. Si se procesan con el *Schema Markup Validator* se obtendría la misma información, junto con un ítem de tipo no especificado asociado a la URL `http://danigayo.info/ … microdata.html`; la razón es que ese archivo JSON-LD se obtuvo originalmente con un traductor[4].

Si cargamos ese código JSON-LD en el *JSON-LD Playground* para visualizarlo[5] nos encontramos con que la información está ahí pero no se están aprovechando todas las posibilidades que ofrece JSON-LD ni los datos enlazados. No solo es que falten los títulos de las publicaciones—que no aparecían en la versión con microdatos, es que no se están vinculando las publicaciones con el autor ni esa persona con la organización con la que está afiliada.

Aquí hay una versión mucho más razonable. Aquí no solo se aprovechan características de JSON-LD sino que se hace uso de identificadores habituales en el mundo académico como ISBN, ISSN y ORCID. Además de validar ese archivo con el *Schema Markup Validator* vamos a visualizar el grafo correspondiente con el *JSON-LD Playground*.

[4] El traductor era *RDF Translator* y aunque la web está online el servicio ya no funciona.
[5] También puedes visualizar RDF con *RDFShape* o con *Sketch*.

Ejemplos para explorar por tu cuenta

Visita algunas de las siguientes URLs y trata de validar los metadatos incrustados usando tanto las herramientas de *Google* como el *OpenLink Structured Data Sniffer*[6].

- Libros:
  - https://www.amazon.com/Ideas-That-Created-Future-Computer/dp/0262045303
  - https://rebiun.baratz.es/rebiun/doc?q=modern+information+retrieval&start=11&rows=1&sort=score%20desc&fq=msstored_mlt155&fv=*&fo=and&redo_advanced=false
- Negocios:
  - https://tierra-astur.com/
- Noticias:
  - https://www.lavozdeasturias.es/noticia/asturias/2022/10/17/hinton-pudieramos-imitar-precision-cerebro-crearamos-sistema-artificial-sentimientos-humanos/000316660023486644849 31.htm
  - https://www.rtve.es/noticias/20221021/reino-unido-johnson-apoyos-diputados-conservadores-volver-downing-street/2406675.shtml
  - https://www.uniovi.es/-/memoria-arboles
- Obras de arte:
  - https://www.museodelprado.es/coleccion/obra-de-arte/los-borrachos-o-el-triunfo-de-baco/4a23d5e2-9fd4-496b-806b-0f8ba913b3d8?searchid=ca4a203d-2d5a-d1d8-6bd5-94b9296bec45
- Productos:
  - https://www.lidl.es/es/disfraz-de-vampiro-infantil/p30856
  - https://www.todocoleccion.net/lotes/show?id_lote=324530218
  - https://www.ebay.es/itm/155177269281
  - https://www.idealista.com/inmueble/2426833/
  - https://es.wallapop.com/item/canon-eos-1100d-840502191

[6] Esta extensión para Chrome (y otros navegadores) extrae todos los metadatos disponibles en una página web. Es capaz de extraer metadatos en diversos formatos como los ya mencionados microdata y JSON-LD pero también RDFa o "POSH". RDFa es similar a microdata en el sentido de que los metadatos están incorporados dentro de HTML en vez de ofrecerse como un script adjunto. "POSH" es terminología propia de la herramienta y se refiere a *"plain old semantic HTML"*, es decir, a la inclusión de información semántica usando fundamentalmente la etiqueta `meta`. El acrónimo "POSH" es poco adecuado puesto que las iniciativas de Twitter y Facebook *Twitter Cards* y *Open Graph* que no tienen nada de *"old"* usan esa etiqueta para que las páginas web incluyan información semántica que facilite su uso dentro de cada plataforma.

Entregable

1. Lee la guía de Google para datos estructurados.
2. Visita la lista de esquemas disponibles en *Schema.org*.
3. Selecciona **[en realidad os lo asignaremos aleatoriamente]** un tipo de contenido a etiquetar con datos estructurados: libro, negocio, noticia, obra de arte catalogada o producto para venta/alquiler online.
4. Localiza en la web dos o tres ejemplos típicos con contenidos suficientes como para preparar un caso ilustrativo ficticio.
5. Prepara un etiquetado con datos estructurados en JSON-LD para ese caso tan exhaustivo como sea posible. Verifica con las herramientas vistas en clase que los datos validan correctamente.
6. **Opcional:** Etiquetar el contenido con *Twitter Cards* y/u *Open Graph*.
7. **Reflexión:** Utiliza la demo de *Dandelion* para procesar el texto de tu caso. ¿Qué diferencias hay entre los datos estructurados que se ofrecen con JSON-LD a las máquinas—en realidad a Google y sus algoritmos—y el tipo de información obtenida mediante entity linking?
8. **Reto:** ¿Es viable etiquetar en todo o en parte el contenido de tu caso usando el vocabulario de *Schema.org* para anotar la información detectada mediante *entity linking*?

Se entregará en el campus virtual un archivo comprimido que contendrá:

a. Enlaces a los textos utilizados como punto de partida para construir el caso ilustrativo.
b. El texto plano del caso.
c. Un archivo JSON-LD con los datos estructurados desarrollados en el punto 5.
d. Opcionalmente: archivos para los datos de *Twitter Cards* y/u *Open Graph*.
e. Un documento con las respuestas razonadas a los puntos 6 y 7.
f. Un archivo JSON-LD que trate de responder al reto del punto 7.

Se valorará positivamente que el material correspondiente a los apartados c) y d) al menos sea un único archivo HTML que incorpore tanto el contenido como los metadatos y pase las pruebas de los validadores.

Referencias bibliográficas

- HTML, Living Standard — Microdata
- JSON for Linking Data