Práctica 7 - Información Semántica

Hugo Pérez Fernández. UO250708@uniovi.es

Sistemas de Información para la Web. Grado de Ingeniería Informática. EII.
universidad de Oviedo. Campus de los Catalanes. Oviedo

1 Introducción

2 Asignación de entidades y creación de microdatos.

Se trata de seleccionar un extracto de una noticia y realizar la asignación de entidades a este en forma de microdatos mediante las herramientas Dandelion, Schema.org y Prueba de datos Estructurados de Google.

2.1 Definicion de entidades mediante Dandelion.

Se ha seleccionado el extracto (Anexo.4.1) de la noticia sobre el ransomware que ha afectado a empresas de España (Noticia aquí). Y, usando la herramienta Dandelion, se han obtenido las siguientes entidades:

- Organisations:
  - Cadena SER.
  - Prisa Radio.
  - Everis.
  - Radio Madrid.
- Concepts:
  - Ransomware.
  - Virus Informático.
  - Internet.
  - Computadora personal.
  - Prensa escrita.
  - Medio de Comunicación.
  - Empresa.

2.2 Refinamiento de entidades con Schema.org

Una vez que tenemos un primer aproximamiento a las entidades que aparecen en el extracto (Anexo.4.1) mediante el uso de la herramienta Dandelion, ahora pasaremos a completar la lista generada mediante nuestro conocimiento del dominio de la noticia y el apoyo de Schema.org, de manera que localicemos mas items o entidades y sus propiedades.

Tras realizar esto generamos el siguiente html con microdatos:

```
<div itemscope itemtype="http://schema.org/NewsArticle">
    <h1 itemprop="headline">Everis y Prisa Radio sufren un grave ciberataque que
        secuestra sus sistemas</h1>
    <span itemprop="datePublished" content="2019-11-04">2019-11-04</span>
    <span itemprop="articleBody">
        Varias empresas españolas han sufrido hoy un serio ataque de
        <span itemscope itemtype="http://schema.org/SoftwareApplication">
        <span itemprop="name">Ransomware</span>
        </span>que recuerda al vivido a mediados de 2017 con
        <span itemscope itemtype="http://schema.org/SoftwareApplication">
        <span itemprop="name">Wannacry</span>
        </span>. Los primeros ataques confirmados de forma oficial
        los han sufrido
        <span itemscope itemtype="http://schema.org/Organization">
        <span itemprop="name">La Cadena SER</span>
        </span> y otras emisoras de <span itemscope itemtype="http://schema.org/Organization">
        <span itemprop="name">Prisa Radio</span>
        </span>, pero también varias consultoras tecnológicas, de las cuales <span itemscope itemtype="http://schema.org/Organization">
        <span itemprop="name">Everis</span>
        </span> ha confirmado oficialmente estar afectada. "Estamos sufriendo
        un ataque masivo de virus a la red de <span itemscope itemtype="http://schema.org/Organization">
        <span itemprop="name">Everis</span>
        </span>. Por favor, mantengan los <span itemscope itemtype="http://schema.org/Product">
        <span itemprop="name">PC</span>
        <img itemprop="image" src="">
        <span itemprop="description"></span>
        </span> apagados". Es el
        mensaje interno que ha remitido <span itemscope itemtype="http://schema.org/Organization">
```

```
30          <span itemprop="name">Everis</span> a sus
                empleados, según ha podido
31          confirmar este diario
32          y han publicado también varios medios
                especializados. La compañía
33          confirma que ha enviado a sus
34          trabajadores a casa hasta que puedan solventar la
                 incidencia.
35
36          <span itemscope itemtype="http://schema.org/
                Organization">
37          <span itemprop="name">Prisa Radio</span> ha
                sufrido esta
38          madrugada un ataque de <span itemscope itemtype="
                http://schema.org/SoftwareApplication">
39          <span itemprop="name">Virus</span>
40          </span> que ha tenido una afectación grave y
41          generalizada de todos nuestros <span itemscope
                itemtype="http://schema.org/Thing">
42          <span itemprop="name">sistemas informáticos</span
                >
43          </span>. Los técnicos especializados en este tipo
44          de situaciones aconsejan encarecidamente la
                desconexión total de
45          todos los sistemas con el fin
46          de evitar la propagación del <span itemscope
                itemtype="http://schema.org/
                SoftwareApplication">
47          <span itemprop="name">Virus</span>. Hablamos, por
                 tanto, de
48          una situación de extrema emergencia,
49          ha comunicado esta mañana la empresa en un
                mensaje interno
50          al que ha tenido acceso este diario.
51
52          A partir de ese momento, se irá chequeando puesto
                 a puesto
53          -siguiendo las instrucciones precisas
54          que os enviará el departamento de Sistemas- para
                autorizar
55          en los casos en que haya garantía
56          absoluta la puesta en marcha de cada equipo. Por
                el momento,
57          y hasta nuevo aviso, la emisión de
```

```
58          <span itemscope itemtype="http://schema.org/
                Organization">
59          <span itemprop="name">La Cadena SER</span>
60          </span> queda centralizada en <span itemscope
                itemtype="http://schema.org/Organization">
61          <span itemprop="name">Radio Madrid</span>
62          </span>; quedan anuladas todas las emisiones
                locales y
63          regionales, salvo aquellas que hayan sido
                autorizadas
64          expresamente por la <span itemscope itemtype="
                http://schema.org/Organization">
65          <span itemprop="name">Dirección de Antena</span>
66          </span>.
67          En este momento, la seguridad es la máxima de la
                compañía
68          por encima de cualquier otro compromiso.
69          Cualquier acción individual no autorizada puede
                poner en
70          peligro el trabajo de rescate que se está
71          llevando a cabo, cierra el mensaje.
72          </span>
73  </div>
```

2.3 Cuestiones sobre el uso de microdatos.

Ya realizado el proceso de los apartados anteriores queda plantearse las siguientes preguntas:

¿Deberían tener valor para itemid todos los items? ¿Qué valor debería asignarse?

Si, todos los itemid deberian tener un itemid de manera que se pudiera identificar cada uno de ellos unequivocamente, además esto favorecería la propiedad de la web Semántica por la que toda entidad debe ser enlazable.

¿Si hubiera varios candidatos cuál escogerías? ¿Por qué?.

En el caso de que hubiera varios candidatos para el identificador de un item elegiria un valor alfanumerico, como realiza WikiData por ejemplo, de este modo evitariamos colisiones y problemas de idiomas.

¿Cuál crees que es la solución de compromiso que podría resultar menos polémica en la mayoría de casos?.

¿Qué inconvenientes concretos te ha supuesto la obligación de incrustarlos metadatos allí donde te "forzaba" la estructura del texto?.

El mayor problema que ha supesto incrustar los metadatos en la estructura del texto es que, por el modo en el que se hace cualquiera de ellos debe aparecer en el texto de manera que algunas propiedades obligatorias de las entidades

quedan sin rellenar porque no aparecen en el texto, y esto al revisarlo mediante la herramienta de Google da errores, que si quisieramos solucionar la unica manera posible para ello seria modificar el texto que se fuera a mostrar para que apareciesen todos los datos de la propiedades de cada entidad.

3 JSON-LD

4 Anexos

4.1 Extracto de la noticia seleccionada.

Varias empresas españolas han sufrido hoy un serio ataque de 'ransomware' que recuerda al vivido a mediados de 2017 con Wannacry. Los primeros ataques confirmados de forma oficial los han sufrido la Cadena SER y otras emisoras de Prisa Radio, pero también varias consultoras tecnológicas, de las cuales Everis ha confirmado oficialmente estar afectada. "Estamos sufriendo un ataque masivo de virus a la red de Everis. Por favor, mantengan los PCs apagados". Es el mensaje interno que ha remitido Everis a sus empleados, según ha podido confirmar este diario y han publicado también varios medios especializados. La compañía confirma que ha enviado a sus trabajadores a casa hasta que puedan solventar la incidencia.

"PRISA Radio ha sufrido esta madrugada un ataque de virus que ha tenido una afectación grave y generalizada de todos nuestros sistemas informáticos. Los técnicos especializados en este tipo de situaciones aconsejan encarecidamente la desconexión total de todos los sistemas con el fin de evitar la propagación del virus. Hablamos, por tanto, de una situación de extrema emergencia", ha comunicado esta mañana la empresa en un mensaje interno al que ha tenido acceso este diario.

"A partir de ese momento, se irá chequeando puesto a puesto -siguiendo las instrucciones precisas que os enviará el departamento de Sistemas- para autorizar en los casos en que haya garantía absoluta la puesta en marcha de cada equipo. Por el momento, y hasta nuevo aviso, la emisión de Cadena SER queda centralizada en Radio Madrid; quedan anuladas todas las emisiones locales y regionales, salvo aquellas que hayan sido autorizadas expresamente por la Dirección de Antena. En este momento, la seguridad es la máxima de la compañía por encima de cualquier otro compromiso. Cualquier acción individual no autorizada puede poner en peligro el trabajo de rescate que se está llevando a cabo", cierra el mensaje.