

# Práctica 8 - Información Semántica

Hugo Pérez Fernández. U0250708@uniovi.es

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

## 1 Introducción

A continuación se realizarán los ejercicios de la practica 8 de la asignatura de Sistemas de Información para la Web, estos se dividirán en dos conjuntos, en el primero se trabajara sobre una noticia (Anexo.4.1) para marcarla con microdata y se resolveran unas cuestiones sobre dicha tecnología, y en el segundo se realizaran una serie de modificaciones en un JSONLD dado para mejorar la información que provea.

## 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 Definición 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:

```
1 <div itemscope itemtype="http://schema.org/NewsArticle">
2   <h1 itemprop="headline">Everis y Prisa Radio sufren
      un grave ciberataque que
3     secuestra sus sistemas</h1>
4   <span itemprop="datePublished" content="2019-11-04">
      2019-11-04</span>
5   <span itemprop="articleBody">
6     Varias empresas españolas han sufrido hoy un
      serio ataque de
7     <span itemscope itemtype="http://schema.org/
      SoftwareApplication">
8       <span itemprop="name">Ransomware</span>
9     </span>que recuerda al vivido a mediados de 2017
      con
10    <span itemscope itemtype="http://schema.org/
      SoftwareApplication">
11      <span itemprop="name">Wannacry</span>
12    </span>. Los primeros ataques confirmados de
      forma oficial
13    los han sufrido
14    <span itemscope itemtype="http://schema.org/
      Organization">
15      <span itemprop="name">La Cadena SER</span>
16    </span> y otras emisoras de <span itemscope
      itemtype="http://schema.org/Organization">
17      <span itemprop="name">Prisa Radio</span>
18    </span>, pero también varias consultoras
19    tecnológicas, de las cuales <span itemscope
      itemtype="http://schema.org/Organization">
20      <span itemprop="name">Everis</span>
21    </span> ha confirmado oficialmente estar afectada
      . Estamos sufriendo
22    un ataque masivo de virus a la red de <span
      itemscope itemtype="http://schema.org/
      Organization">
23      <span itemprop="name">Everis</span>
24    </span>. Por favor, mantengan los <span itemscope
      itemtype="http://schema.org/Product">
```

25 <span itemprop="name">PC</span>  
 26 <img itemprop="image" src="">  
 27 <span itemprop="description"></span>  
 28 </span> apagados. Es el  
 29 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>

```

---

**Listing 1.1.** Archivo HTML con Microdatos embebidos.

Al validar el documento html anterior mediante la herramienta para pruebas de datos estructurados de Google genera el siguiente resultado:

Como se ve en la Fig.1 al validar el html generado este presenta errores y advertencias en las diferentes entidades, esto se debe a decisiones tomadas ante los inconvenientes que presentan los microdatos, que se explicarán a continuación.

La entidad NewsArticle presenta los siguientes problemas:

– Errores:

- Falta el atributo Author obligatrio: Este es debido a que el extracto de la noticia no contiene esa información.
- Falta el atributo image obligatrio: Este es debido a que el extracto de la noticia no contiene esa información.
- Falta el atributo publisher obligatrio: Este es debido a que el extracto de la noticia no contiene esa información.

– Advertencias:



- Falta el atributo description recomendado: Este es debido a que el extracto de la noticia no contiene esa información.
- Falta el atributo image recomendado: Este es debido a que el extracto de la noticia no contiene esa información.
- Falta el atributo sku recomendado: Este es debido a que el extracto de la noticia no contiene esa información.
- Falta algun atributo de tipo identificador global recomendado: Este es debido a que el extracto de la noticia no contiene esa información.

## 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?*

Depende, todos los itemid deberían tener un itemid de manera que se pudiera identificar cada uno de ellos unequivocamente, pero no en todos los casos, puesto que sobrecargaría los textos. 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 elegiría un valor alfanumerico, como realiza WikiData por ejemplo, de este modo evitáramos colisiones y problemas de idiomas.

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

Para decidir si hay que añadir un itemid o no la solución sería si ese item tiene un identificador, ya que quizás no lo tenga, y si mejora la información que nos da.

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

El mayor problema que ha supuesto 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 quisiéramos solucionar la única manera posible para ello sería modificar el texto que se fuera a mostrar para que apareciesen todos los datos de las propiedades de cada entidad.

## 3 JSON-LD

A continuación se presentará la última versión del jsonld que se facilita en la práctica (Anexo.4.2), el cual se ha modificado de la siguiente forma:

- Se han añadido a las entidades Periodical para que el id ahora sea la web oficial de las publicaciones.

- Se ha añadido al contexto del jsonld la url de DOI "<https://doi.org/>", y se han modificado los items de tipo Chapter para que el id apunte a un *"digital object identifier"*.
- Se ha asociado al id de la entidad Person de Daniel Gayo su web oficial. Y de igual modo con las entidades Universidad de Oviedo y Departamento de informatica con sus id's y las web's oficiales.
- Se han añadido a los items de tipo Periodical el atributo hasPart para señalar el articulo del que se habla de toda la publicacion.

---

```
1 <div>
2   <p>
3     Varias empresas españolas han sufrido hoy un
4     serio ataque de 'ransomware' que recuerda al
5     vivido a mediados de 2017 con Wannacry. Los
6     primeros ataques confirmados de forma oficial
7     los han sufrido la Cadena SER y otras emisoras de
8     Prisa Radio, pero también varias consultoras
9     tecnológicas, de las cuales Everis ha confirmado
10    oficialmente estar afectada. Estamos sufriendo
11    un ataque masivo de virus a la red de Everis. Por
12    favor, mantengan los PCs apagados. Es el
13    mensaje interno que ha remitido Everis a sus
14    empleados, según ha podido confirmar este
15    diario
16    y han publicado también varios medios
17    especializados. La compañía confirma que ha
18    enviado a sus
19    trabajadores a casa hasta que puedan solventar la
20    incidencia.
21  </p>
22  <p>
23    PRISA Radio ha sufrido esta madrugada un ataque
24    de virus que ha tenido una afectación grave y
25    generalizada de todos nuestros sistemas
26    informáticos. Los técnicos especializados en
27    este tipo
28    de situaciones aconsejan encarecidamente la
29    desconexión total de todos los sistemas con el
30    fin
31    de evitar la propagación del virus. Hablamos, por
32    tanto, de una situación de extrema emergencia
33    ,
34    ha comunicado esta mañana la empresa en un
35    mensaje interno al que ha tenido acceso este
36    diario.
37  </p>
```

```

19      <p>
20          A partir de ese momento, se irá chequeando puesto
              a puesto –siguiendo las instrucciones
              precisas
21          que os enviará el departamento de Sistemas– para
              autorizar en los casos en que haya garantía
22          absoluta la puesta en marcha de cada equipo. Por
              el momento, y hasta nuevo aviso, la emisión de
23          Cadena SER queda centralizada en Radio Madrid;
              quedan anuladas todas las emisiones locales y
24          regionales, salvo aquellas que hayan sido
              autorizadas expresamente por la Dirección de
              Antena.
25          En este momento, la seguridad es la máxima de la
              compañía por encima de cualquier otro
              compromiso.
26          Cualquier acción individual no autorizada puede
              poner en peligro el trabajo de rescate que se
              está
27          llevando a cabo, cierra el mensaje.
28      </p>
29  </div>
30  <script type="application/ld+json">
31  {
32      "@context": {
33          "cc": "http://creativecommons.org/ns#",
34          "ctag": "http://commontag.org/ns#",
35          "dc": "http://purl.org/dc/elements/1.1/",
36          "dcl": "http://purl.org/dc/terms/",
37          "dcl1": "http://purl.org/dc/elements/1.1/",
38          "dcat": "http://www.w3.org/ns/dcat#",
39          "determs": "http://purl.org/dc/terms/",
40          "doi": "https://doi.org/",
41          "foaf": "http://xmlns.com/foaf/0.1/",
42          "gr": "http://purl.org/goodrelations/v1#",
43          "grddl": "http://www.w3.org/2003/g/data-view#",
44          "hcalendar": "http://microformats.org/profile/
              hcalendar#",
45          "hcard": "http://microformats.org/profile/hcard#",
46          "ical": "http://www.w3.org/2002/12/cal/icaltzd#",
47          "ma": "http://www.w3.org/ns/ma-ont#",
48          "md": "http://www.w3.org/ns/md#",
49          "og": "http://ogp.me/ns#",
50          "org": "http://www.w3.org/ns/org#",
51          "owl": "http://www.w3.org/2002/07/owl#",

```



```

52     "prov": "http://www.w3.org/ns/prov#",
53     "qb": "http://purl.org/linked-data/cube#",
54     "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
55     ,
56     "rdfa": "http://www.w3.org/ns/rdfa#",
57     "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
58     "rev": "http://purl.org/stuff/rev#",
59     "rif": "http://www.w3.org/2007/rif#",
60     "rr": "http://www.w3.org/ns/r2rml#",
61     "schema": "http://schema.org/",
62     "sd": "http://www.w3.org/ns/sparql-service-
        description#",
63     "sioc": "http://rdfs.org/sioc/ns#",
64     "skos": "http://www.w3.org/2004/02/skos/core#",
65     "skosxl": "http://www.w3.org/2008/05/skos-xl#",
66     "v": "http://rdf.data-vocabulary.org/#",
67     "vcard": "http://www.w3.org/2006/vcard/ns#",
68     "void": "http://rdfs.org/ns/void#",
69     "wdr": "http://www.w3.org/2007/05/powder#",
70     "wdrs": "http://www.w3.org/2007/05/powder-s#",
71     "wdsr": "http://www.w3.org/2007/05/powder-s#",
72     "xhv": "http://www.w3.org/1999/xhtml/vocab#",
73     "xsd": "http://www.w3.org/2001/XMLSchema#"
74 },
75 "@graph": [
76     {
77         "@id": "_:N87d20637af2445ec83dea37c29d9085d",
78         "@type": "schema:Chapter",
79         "schema:isPartOf": {
80             "@id": "doi:10.1017/CBO9781316182635"
81         },
82         "schema:name": "Political Opinion"
83     },
84     {
85         "@id": "https://ieeexplore.ieee.org/xpl/
            RecentIssue.jsp?punumber=93",
86         "@type": "schema:Periodical",
87         "schema:name": "IEEE Multimedia",
88         "hasPart": "doi:10.1109/MMUL.2015.47"
89     },
90     {
91         "@id": "https://www.emerald.com/insight/
            publication/issn/1066-2243",
92         "@type": "schema:Periodical",
93         "schema:name": "Internet Research",

```

```

93         "hasPart": "issn:1066-2243"
94     },
95     {
96         "@id": "https://www.computer.org/csdl/magazine/ic",
97         "@type": "schema:Periodical",
98         "schema:name": "IEEE Internet Computing",
99         "hasPart": "doi:10.1109/MIC.2012.137"
100     },
101     {
102         "@id": "https://cacm.acm.org",
103         "@type": "schema:Periodical",
104         "schema:name": "Communications of the ACM",
105         "hasPart": "doi:10.1145/2001269.2001297 |"
106     },
107     {
108         "@id": "_:Nb6734e117ddf48a4a3f95b95e5239145",
109         "@type": "schema:Organization",
110         "schema:name": "Department of Computer Science"
111     },
112     {
113         "@id": "http://danigayo.info",
114         "@type": "schema:Person",
115         "schema:affiliation": {
116             "@id": "_:Nd46c7cbe934541d68dc746e686f6354c"
117         },
118         "schema:jobTitle": "associate professor",
119         "schema:name": "Daniel Gayo-Avello"
120     },
121     {
122         "@id": "urn:isbn:9781107500075",
123         "@type": "schema:Book",
124         "schema:name": "\"Twitter: A Digital Socioscope\"",
125         "schema:publisher": "Cambridge University Press"
126     },
127     {
128         "@id": "http://www.uniovi.es",
129         "@type": "schema:CollegeOrUniversity",
130         "schema:department": {
131             "@id": "http://www.di.uniovi.es"
132         },
133         "schema:name": "University of Oviedo"
134     }
135 ]

```

```
136     }  
137 </script>
```

---

**Listing 1.2.** Archivo HTML con JSONLD embebido.

## 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.

### 4.2 JSONLD original

---

```
1 {  
2   "@context": {  
3     "cc": "http://creativecommons.org/ns#",  
4     "ctag": "http://commontag.org/ns#",  
5     "dc": "http://purl.org/dc/elements/1.1/",  
6     "dc1": "http://purl.org/dc/terms/",
```

```

7      "dc11": "http://purl.org/dc/elements/1.1/",
8      "dcat": "http://www.w3.org/ns/dcat#",
9      "dcterms": "http://purl.org/dc/terms/",
10     "foaf": "http://xmlns.com/foaf/0.1/",
11     "gr": "http://purl.org/goodrelations/v1#",
12     "grddl": "http://www.w3.org/2003/g/data-view#",
13     "hcalendar": "http://microformats.org/profile/
        hcalendar#",
14     "hcard": "http://microformats.org/profile/hcard#",
15     "ical": "http://www.w3.org/2002/12/cal/icaltzd#",
16     "ma": "http://www.w3.org/ns/ma-ont#",
17     "md": "http://www.w3.org/ns/md#",
18     "og": "http://ogp.me/ns#",
19     "org": "http://www.w3.org/ns/org#",
20     "owl": "http://www.w3.org/2002/07/owl#",
21     "prov": "http://www.w3.org/ns/prov#",
22     "qb": "http://purl.org/linked-data/cube#",
23     "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
24     "rdfa": "http://www.w3.org/ns/rdfa#",
25     "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
26     "rev": "http://purl.org/stuff/rev#",
27     "rif": "http://www.w3.org/2007/rif#",
28     "rr": "http://www.w3.org/ns/r2rml#",
29     "schema": "http://schema.org/",
30     "sd": "http://www.w3.org/ns/sparql-service-
        description#",
31     "sioc": "http://rdfs.org/sioc/ns#",
32     "skos": "http://www.w3.org/2004/02/skos/core#",
33     "skosxl": "http://www.w3.org/2008/05/skos-xl#",
34     "v": "http://rdf.data-vocabulary.org/#",
35     "vcard": "http://www.w3.org/2006/vcard/ns#",
36     "void": "http://rdfs.org/ns/void#",
37     "wdr": "http://www.w3.org/2007/05/powder#",
38     "wdrs": "http://www.w3.org/2007/05/powder-s#",
39     "wdsr": "http://www.w3.org/2007/05/powder-s#",
40     "xhv": "http://www.w3.org/1999/xhtml/vocab#",
41     "xsd": "http://www.w3.org/2001/XMLSchema#"
42 },
43 "@graph": [
44     {
45         "@id": "_:N87d20637af2445ec83dea37c29d9085d",
46         "@type": "schema:Chapter",
47         "schema:isPartOf": {
48             "@id": "urn:isbn:9781107500075"
49         }
50     }
51 ]

```

```

50     "schema:name": "Political Opinion"
51 },
52 {
53     "@id": "_:N237308aadd894e9cb88f3f9bd3ca7f85",
54     "@type": "schema:Periodical",
55     "schema:name": "IEEE Multimedia"
56 },
57 {
58     "@id": "_:N0f982a4b306847a78b9496ab615d9502",
59     "@type": "schema:Periodical",
60     "schema:name": "Internet Research"
61 },
62 {
63     "@id": "_:N57a8d3cded6d4061af9d3cb1e7acd39f",
64     "@type": "schema:Periodical",
65     "schema:name": "IEEE Internet Computing"
66 },
67 {
68     "@id": "_:N8381ebd04bda429fbb2c1d900d8b0e2c",
69     "@type": "schema:Periodical",
70     "schema:name": "Communications of the ACM"
71 },
72 {
73     "@id": "_:Nb6734e117ddf48a4a3f95b95e5239145",
74     "@type": "http://schema.org/Organization",
75     "http://schema.org/name": "Department of Computer
        Science"
76 },
77 {
78     "@id": "_:N8013fd042aca458b9f5373abc2c9bbda",
79     "@type": "schema:Person",
80     "schema:affiliation": {
81         "@id": "_:Nd46c7cbe934541d68dc746e686f6354c"
82     },
83     "schema:jobTitle": "associate professor",
84     "schema:name": "Daniel Gayo-Avello"
85 },
86 {
87     "@id": "urn:isbn:9781107500075",
88     "@type": "schema:Book",
89     "schema:name": "\"Twitter: A Digital Socioscope\"",
90     "schema:publisher": "Cambridge University Press"
91 },
92 {
93     "@id": "_:Nd46c7cbe934541d68dc746e686f6354c",

```

```
94      "@type": "http://schema.org/CollegeOrUniversity",
95      "http://schema.org/department": {
96        "@id": "_:Nb6734e117ddf48a4a3f95b95e5239145"
97      },
98      "http://schema.org/name": "University of Oviedo"
99    }
100  ]
101 }
```

---

**Listing 1.3.** JSON Original facilitado en la practica.