

[Publicado en *Letras de Deusto* 100:11-52, julio-septiembre 2003.

*Nota: la paginación en esta versión electrónica no coincide con la de la edición impresa.]*

## Contenidos y metacontenidos en la edición digital

Joseba Abaitua, Guillermo Barrutieta, Josuka Díaz, Inés Jacob, Fernando Quintana  
Grupo DELi  
www.deli.deusto.es  
Universidad de Deusto

### Resumen

Los lenguajes de anotación basados en SGML/XML han propiciado un cambio revolucionario en los métodos de edición tradicional y constituyen el fundamento de la edición digital moderna. El artículo define los conceptos de metadato y metacontenido y reseña las principales iniciativas actuales de gestión de metacontenidos. En el campo filológico se introduce TEI; en el área de Internet se habla de DCMI, RDF, así como de la *web* semántica; en el terreno de la traducción se citan TMX y XLIFF; y, por último, en el ámbito de la recopilación y sindicación de metacontenidos, se presentan OCS y OAI. De esta manera, el artículo ofrece un repaso amplio y representativo de las principales líneas de actuación en la gestión de contenidos y metacontenidos en el marco de la edición digital.

### 1 Introducción

En el desarrollo de las técnicas de edición digital modernas se recogen aportaciones procedentes de disciplinas como la filología, la lingüística, la informática, o la biblioteconomía y todas ellas confluyen en el uso del metalenguaje SGML/XML. La evolución de los procesadores de texto y de otras herramientas de maquetación digital, que representan el área más visible de convergencia entre la informática y el trabajo editorial, ofrecía hasta fechas recientes resultados decepcionantes (con la única excepción del sistema de edición basado en TeX/LaTeX, de Knuth, 1984). Sowa (2000) lo expresa de esta manera:

*Por desgracia, la mayoría de los procesadores de texto sólo abarcan un subconjunto pequeño de la sintaxis y han producido lo que St. Laurent (1999) llama el desastre de WYSIWYG: "El texto plano, por muy tosco que pueda parecer, es mucho más manejable que el resultado de un procesador de texto o maquetador cualquiera". En la práctica, el eslogan "What you see is what you get" significa en realidad WYSIAYG: "What you see is all you get." El texto está tan sobrecargado con códigos de formato que no queda sitio para la semántica o la pragmática. El formato Rich Text Format (RTF) es paradójicamente la forma de representación semántica de texto más pobre que existe. El formato es un aspecto del texto que lo hace parecer bonito, pero que falla en lo cuestión fundamental de lo que significa..*

Por fortuna, a finales de la década de los ochenta se diseñó SGML con unos postulados innovadores en el terreno del tratamiento documental. Para empezar, SGML expresa de manera clara la separación entre formato y contenidos, lo que permite tratar el problema del significado de manera independiente. La gestión de contenidos es precisamente el aspecto que más se beneficia de las tecnologías de la información. Veremos en este artículo las importantes consecuencias que ello conlleva en el marco general de la edición y publicación de textos electrónicos; sobre todo a partir de la explosión de contenidos propiciada por la *web*.

El segundo apartado de este artículo aborda una serie de conceptos básicos que son necesarios para comprender mejor las implicaciones de los nuevos enfoques de la edición digital. Se reseñará XML, sucesor directo de SGML, como metalenguaje estándar y fundamento de toda la actividad actual. Junto a XML se explicarán las nociones de metalenguaje y metadato. En el tercer apartado se introducirá el principal hito en el uso de metacontenidos en el terreno filológico, la *Text Encoding Initiative* (TEI). Pero TEI no agota el abanico de actividades y por ello en el apartado cuarto se presentan otras iniciativas de metadatos que han surgido en el área de Internet: DCMI, RDF y la *web* semántica. El apartado quinto se ocupa de dos estándares para el intercambio de información en el campo de las traducciones: TMX y

XLIFF. Por último, en el apartado sexto se describen iniciativas para la recopilación y sindicación de metacontenidos, OCS y OAI, que representan un área de actividad que tendrá mucho que aportar en los próximos años.

De esta manera, el artículo ofrece un repaso amplio y representativo de las principales líneas de actuación en la gestión de contenidos en el marco de la edición digital.

## 2 Conceptos preliminares

Antes de abordar las que consideramos principales líneas de acción en el panorama actual de la edición digital, vamos a repasar algunos conceptos básicos que ayudarán a comprender mejor la dimensión e implicaciones de los nuevos enfoques. Vamos a presentar XML, el metalenguaje estándar y fundamento de toda la actividad actual. Junto a XML expondremos también las nociones de metalenguaje y metadato.

### Lenguajes y metalenguajes

Siempre que estudiemos el lenguaje, en cualquiera de sus modos (oral, escrito o electrónico), estaremos elevando el discurso a un plano que por definición denominaremos *metalenguaje*. La lingüística, por lo tanto, como ciencia del lenguaje humano se sitúa por vocación propia en el plano de los metalenguajes. Algo similar sucede con la filología, ciencia que estudia los resultados o productos del lenguaje humano; por ello, los estudios que se realicen sobre los textos (estén en el soporte que estén) se plantearán en un plano más elevado de *metatexto*. En la bibliografía especializada y también en este artículo, los términos técnicos de “metalenguaje” o “metatexto” se reemplazarán con frecuencia por los más comunes de “texto” y “lenguaje”.

El desarrollo de tecnologías basadas en metalenguajes tiene especial relevancia para el tratamiento computacional del lenguaje, así como para otros aspectos básicos del procesamiento de textos. Los avances han venido marcados por la evolución de los lenguajes lógicos, formales o *simbólicos*. Los orígenes del simbolismo se remontan a la Grecia clásica y suelen relacionarse con la aportación del matemático griego Euclides, del siglo IV a.C. En su obra *Los elementos*, este autor desarrolló lo que hoy en día se conoce como geometría euclídea, un estudio formal de conceptos geométricos elementales (puntos, líneas, superficies, polígonos y otras figuras). El hecho relevante es que Euclides fue capaz de expresar estos conceptos mediante un lenguaje de estructura lógica. Sobre la base de un conjunto de definiciones iniciales, nociones comunes y axiomas o postulados, se deducían el resto de proposiciones. De esta manera, sentó los fundamentos de una disciplina matemática, la geometría, que en esencia se ha mantenido inalterada hasta finales del siglo XIX.

La geometría euclídea se formuló en una lengua natural, el griego, que sirvió de esta manera de metalenguaje. Euclides como Aristóteles fue discípulo de Platón y ambos formaron parte de una sociedad que empezaba a exteriorizar las consecuencias de una nueva tecnología: la escritura alfabética. La idea de que la generalización de la escritura alfabética fue la causa de importantes cambios en el desarrollo cognitivo y cultural de la Grecia clásica es una hipótesis avanzada por Havelock (1963) y luego reforzada por Ong (1982). De acuerdo con esta hipótesis, las aportaciones de Aristóteles en todos los campos, pero muy especialmente en lógica, así como las de Euclides en matemáticas, serían consecuencia directa de las transformaciones en la manera de aprender y de pensar propiciadas por la escritura alfabética (Ong, 1982):

*Writing makes it possible to separate logic (thought structure of discourse) from rhetoric (socially effective discourse). The invention of logic, it seems, is tied not to any kind of writing system but to the completely vocalic phonetic alphabet and the intensive analytic activity which such an alphabet demands of its inventors and subsequently encourages in all sorts of noetic fields.*

En la evolución de las tecnologías aplicadas al pensamiento humano destaca la invención de Ramón Llull de un artilugio lógico en pleno siglo XIII. En aquellos años de la baja edad media, la Europa cristiana vivía una gran efervescencia proselitista frente al Islam. En este contexto, el autor del *Ars magna* tuvo la audacia de diseñar una máquina de pensar en la que las proposiciones teológicas adquirirían formas geométricas que, manipuladas mediante poleas y manillas, probaban mecánicamente su veracidad o falsedad.

La obra de Lull tuvo un eco moderado en el desarrollo de la filosofía. Leibniz, tres siglos después, rescataba este legado con un nuevo método que trataba de expresar con precisión los razonamientos de la filosofía, el *calculus ratiocinator*. Suponía un nuevo intento de construir un sistema de pensamiento lógico inspirado en los lenguajes matemáticos y alejado de la ambigüedad del lenguaje natural.

Pero no será hasta el siglo XIX cuando George Boole y Gottlob Frege por fin consiguen establecer las bases de los lenguajes formales actuales. En su obra *Laws of thought* de 1854, Boole formuló un álgebra lógica similar en rigor y exactitud al álgebra de las matemáticas. Mientras, la *Begriffsschrift* (o *Ideografía*, 1879) de Frege, con el sugerente subtítulo de *un lenguaje formalizado del pensamiento puro a base del lenguaje aritmético* (Ferrater 1991), marca indudablemente tanto el comienzo de la lógica moderna (mostrando además que la matemática se reduce a ésta), como el uso de los lenguajes simbólicos o formales. A pesar de no tener relación directa con el concepto de metalenguaje, no podemos dejar de mencionar el trabajo fundamental de Kurt Gödel, que en 1931 dio la vuelta a la lógica demostrando, primero, que ésta se reduce a la aritmética, y después, que era incapaz de servir al sueño de Lull, Leibniz, Frege, Hilbert y tantos otros de alcanzar *todas* las verdades universales.

Todos los formalismos gramaticales que se utilizan en la actualidad para expresar *generativamente*, es decir, con precisión, la sintaxis o la semántica de las lenguas naturales (HPSG, LFG, UG, DRT, etc.) son deudoras del álgebra de Boole y de los desarrollos posteriores de Frege, Russell, Whitehead, Tarski, Quine, Carnap, Montague y Chomsky, por citar sólo algunos de los sucesores.

Avanzando sobre el resultado de Gödel, en la década de 1930 Alonzo Church, Stephen Kleene, Emil Post y, sobre todo, Alan M. Turing llevaron a cabo teorizaciones sobre la computación que dieron origen al concepto de cálculo automático o informática. La construcción de las primeras computadoras, fundamentada por cierto en la lógica booleana, comienza en la década de 1940 y tiene a John von Neumann como principal teórico, aunque el propio Turing llevó a la práctica sus ideas participando en proyectos paralelos, que sólo ahora (Davis 2002) se empiezan a reconocer como de gran valía (sobre todo, en comparación con el tradicional mérito atribuido a Von Neumann a este respecto).

La evidente importancia que tiene la informática en la evolución de los lenguajes formales no se hizo sin embargo aparente hasta años después, a finales de la década de 1950 (coincidiendo prácticamente, en una de esas sorprendentes casualidades científicas, con el trabajo seminal de Noam Chomsky sobre la gramática generativa), cuando aparecen los primeros *lenguajes de programación*. Aparte del enorme desarrollo informático que han posibilitado desde entonces, los lenguajes de programación son importantes por constituir el primer ejemplo pleno de lenguajes simbólicos artificiales, y su conocimiento ha permitido el desarrollo y procesamiento de otros metalenguajes, no destinados a la programación específicamente, sino en general a la representación de información, y en particular información lingüística, con lo que llegamos al final de nuestro largo camino hacia el concepto de edición digital.

En lo concerniente a este aspecto, el hito más significativo ha sido el diseño de SGML en 1986 y su adaptación inmediata a Internet en la forma de HTML y XML. Estos metalenguajes, inspirados en los principios de los lenguajes de programación, han propiciado el desarrollo de la *web* como red mundial de conocimiento y son la clave del concepto de edición digital sobre el que nos vamos a extender en los apartados siguientes.

## SGML/XML

En este artículo se presupone cierta familiaridad con los metalenguajes SGML/XML y dedicaremos por ello sólo unas breves líneas a repasar sus aspectos más destacados. XML, definido por el consorcio W3C en 1996, es una variante ligeramente simplificada de SGML con el objetivo expreso de ser utilizado en Internet. SGML había sido definido una década antes (como estándar ISO8879) con un alcance que apenas había trascendido a un círculo de usuarios muy especializado. Antes de que se definiera XML, con mucho la aplicación más exitosa de SGML había sido HTML, lenguaje de hipertexto creado por Tim Berners-Lee en 1991. HTML como es bien sabido es el factor desencadenante de la gestación y rápida popularización de la *web*. Pero HTML es una versión muy limitada, que apenas permite aprovechar unas pocas de las muchas posibilidades que ofrece SGML.

Para explicar el interés que desde el principio despertó SGML en el mundo de la edición digital, y que XML recoge casi en su integridad, hay que fijarse en su capacidad para hacer explícitos los contenidos mediante la aplicación de lenguajes de metacontenidos a cualquier objeto documental. A diferencia de HTML, que se compone de un conjunto predefinido y finito de etiquetas, XML permite el diseño abierto de lenguajes de etiquetas (*tags*) para satisfacer las necesidades de cualquier aplicación. Las etiquetas suelen representar metadatos o metacontenidos, como explicaremos más abajo.

La consecuencia más palpable de este cambio es que, así como los documentos definidos mediante HTML están concebidos para ser leídos e interpretados por humanos, la utilización de XML atiende preferentemente a la necesidad de que los sistemas que gestionan información puedan comunicarse entre sí, como paso previo a generar la versión que se mostrará finalmente al usuario humano. Esto quiere decir que XML permite de hecho que las máquinas procesen datos de manera colaborativa y autónoma, sin necesidad de intervención humana directa. Obviamente el fin último no es prescindir de las personas sino, todo lo contrario, aportar soluciones a la sobrecarga informativa que padecemos los humanos.

Por medio de XML es posible definir los documentos con el grado de exhaustividad que se requiera (sea la edición de una obra literaria, un texto educativo –figura 1–, un documento administrativo bilingüe, un boletín con datos bursátiles –figura 3–, o una cartelera cinematográfica –figura 2). Por eso es muy importante tener muy claros los objetivos y estructuras de estas aplicaciones. Una de las características principales de un documento XML es que permite organizar jerárquicamente todas las unidades informativas de un documento mediante estructuras lógicas. En la terminología de XML, estas unidades se denominan entidades (*entities*) y no son sino datos (contenidos) dispuestos para ser interpretados informáticamente. XML posee mecanismos que permiten revisar la estructura lógica de los documentos con el propósito de que las máquinas que se interconecten entre sí para operar con estos datos lo puedan hacer de manera fluida.

```
<RST>
<RST-S>
  <PREPARATION>
    <S>
      ¿Qué es gestión del conocimiento?
    </S>
  </PREPARATION>
</RST-S>
<RST-N>
  <S>
    Conocimiento, en el contexto de los negocios, es la
    memoria de la organización, lo que la gente sabe colectiva
    e individualmente
  </S>
  <S>
    Gestión es el uso juicioso de recursos para alcanzar un fin
  </S>
  <S>
    Gestión del conocimiento es la combinación de esos dos
    conceptos, GC = gestión + conocimiento
  </S>
</RST-N>
</RST>
```

Figura 1. Texto educativo

```
<?xml version="1.0" standalone="no"?>
<?xmlstylesheet type="text/css" href="Movies.css"?>
<!DOCTYPE movies SYSTEM "Movies.dtd">

<movies>
  <movie type="comedy" rating="PG-13" review="5" year="1987">
    <title>Raising Arizona</title>
    <writer>Ethan Coen</writer>
    <writer>Joel Coen</writer>
    <producer>Ethan Coen</producer>
    <director>Joel Coen</director>
    <actor>Nicolas Cage</actor>
    <actor>Holly Hunter</actor>
    <actor>John Goodman</actor>
    <comments>A classic one-of-a-kind screwball love story.</comments>
  </movie>
```

```

<movie type="comedy" rating="R" review="5" year="1988">
  <title>Midnight Run</title>
  <writer>George Gallo</writer>
  <producer>Martin Brest</producer>
  <director>Martin Brest</director>
  <actor>Robert De Niro</actor>
  <actor>Charles Grodin</actor>
  <comments>The quintessential road comedy.</comments>
</movie>
<movie type="mystery" rating="R" review="5" year="1995">
  <title>The Usual Suspects</title>
  <writer>Christopher McQuarrie</writer>
  <producer>Bryan Singer</producer>
  <producer>Michael McDonnell</producer>
  <director>Bryan Singer</director>
  <actor>Stephen Baldwin</actor>
  <actor>Gabriel Byrne</actor>
  <actor>Benicio Del Toro</actor>
  <actor>Chazz Palminteri</actor>
  <actor>Kevin Pollak</actor>
  <actor>Kevin Spacey</actor>
  <comments>A crime mystery with incredibly intricate plot twists.</comments>
</movie>
</movies>

```

Figura 2. Cartelera cinematográfica

```

<?xml:stylesheet type="text/xsl" href="stock-sorter.xsl"?>
<portfolio xmlns="x-schema:portfolio-schema.xml">
  <description>Technology Stock Index</description>
  <date>1998-10-13T15:56:00</date>
  <stock>
    <symbol>ACXM</symbol>
    <name>axiom corp</name>
    <price>18.875</price>
    <change>-1.250</change>
    <percent>-6.21</percent>
    <volume>0.23</volume>
  </stock>
  <stock>
    <symbol>AFFX</symbol>
    <name>affymetrix inc</name>
    <price>20.313</price>
    <change>-1.438</change>
    <percent>-6.61</percent>
    <volume>0.08</volume>
  </stock>
  <stock>
    <symbol>YHOO</symbol>
    <name>yahoo! inc</name>
    <price>109.938</price>
    <change>-4.500</change>
    <percent>-3.93</percent>
    <volume>6.48</volume>
  </stock>
</portfolio>

```

Figura 2. Cartera bursátil

El acceso a los documentos XML se realiza mediante un *procesador* que revisa la estructura de los documentos e interpreta los contenidos de acuerdo con una gramática que se denomina “DTD” (*declaración de tipo de documento*). La DTD define de manera formal la estructura lógica de los documentos. Como veremos más adelante, estas especificaciones suelen ser compartidas por comunidades de usuarios, instituciones o grupos editoriales con el fin de que sus sistemas de gestión de contenidos puedan trabajar en cadena.

En la figura 3 se muestra un ejemplo de DTD que está tomada de Barrutieta y otros (2002). Se trata de una adaptación de la teoría de estructura retórica del discurso (RTS) de Mann y Thompson (1988) a la generación de documentos educativos. De acuerdo con esta teoría, los textos se conciben con una estructura retórica que se puede definir mediante constituyentes (*constituyentes RST*). La aportación de Barrutieta es representar estos constituyentes como elementos XML. Para ello se define una DTD

particular, a través de la cual se determina la estructura retórica de los textos educativos. Analizando la DTD de la figura 3 podemos comprobar la existencia de un elemento que se llama “materia” (SUBJECT), que contiene uno o más cursos (COURSE+), que a su vez se desarrollan en una o más lecciones (LESSON+), compuestas por explicaciones (EXPLANATION+). Las explicaciones se elaboran en elementos nucleares (RST-N) y satélites (RST-S). Los satélites forman un conjunto muy amplio de opciones, entre las que aparecen *antecedente, preparación, motivación, justificación, antítesis, ejercicio, ejemplo, concesión*, etc. Todos ellos son constituyentes posibles, pero no imprescindibles, en una explicación. Lo interesante del experimento de Barrutieta es que los textos que se generan atienden a las necesidades de los usuarios, determinadas por sus perfiles formativos particulares. Los perfiles tienen en cuenta aspectos como si los alumnos son principiantes, están motivados, el número de veces que han accedido a los materiales, el progreso que van realizando, su disponibilidad de tiempo, etc.

Estos ejemplos de DTD ilustran distintas maneras de definir la estructura de un documento o una familia de documentos en XML. En el trabajo de Barrutieta se describe la estructura retórica de materiales formativos en un contexto de enseñanza universitaria a distancia; pero, como se ha dicho, pueden existir otros modelos de DTD, tantos como tipos de documentos se deseen elaborar (p.ej., la cartelera cinematográfica de la figura 2 mediante la DTD de la figura 4).

```

<!ELEMENT SUBJECT (ADMIN,COURSE+ )>
<!ELEMENT ADMIN
(SUBJECTNAME,DEGREE,MOTIVATION,TIMEDISTRIBUTION,LANGUAGE,PROFESSORS,GOALS,
THEORETICALCONTENT,PRACTICALCONTENT,MATERIAL,METHODOLOGY, EVALUATION,REFERENCES )>
<!ELEMENT SUBJECTNAME (#PCDATA)>
<!ELEMENT DEGREE (S+ )>
<!ELEMENT MOTIVATION (S+ )>
<!ELEMENT TIMEDISTRIBUTION (CREDITS,HOURS,THEORY,EXERCISES,LAB,YEAR,SEMESTER,HOURSAWEEK
)>
<!ELEMENT COURSE (INTRO,LESSON+,CONCLUSION )>
<!ATTLIST COURSE
  LANG (ES|EN|EU) #REQUIRED>
<!ELEMENT INTRO (COURSE_TITLE,S+ )>
<!ELEMENT COURSE_TITLE (S+ )>
<!ELEMENT LESSON (TITLE , EXPLANATION+)>
<!ATTLIST LESSON
  DAY (1|2|3|4|5|6|7|8|9|10) #REQUIRED>
<!ELEMENT CONCLUSION (S+)>
<!ELEMENT TITLE (S+ )>
<!ELEMENT EXPLANATION (RST+ )>
<!ELEMENT RST (RST-S|RST-N)*>
<!ELEMENT RST-N (S|
  RST|
  CONTRAST|
  JOINT|
  SEQUENCE|
  LIST)*>
<!ELEMENT RST-S (EVIDENCE|
  BACKGROUND|
  ELABORATION|
  ELABORATION-LINK|
  ELABORATION-IMAGE|
  PREPARATION|
  ANTITHESIS|
  CIRCUMSTANCE|
  CONDITION|
  ENABLEMENT|
  EVALUATE|
  INTERPRETATION|
  JUSTIFY|
  MOTIVATE|
  NON-VOLITIONAL-CAUSE|
  NON-VOLITIONAL-RESULT|
  OTHERWISE|
  PURPOSE|
  RESTATEMENT|
  SOLUTIONHOOD|
  SUMMARY|
  VOLITIONAL-CAUSE|
  VOLITIONAL-RESULT|
  EXAMPLE|
  EXERCISE|
  CONCESSION)*>

```

```
<!ELEMENT CONTRAST (S+)>
```

Figura 3. DTD en la que se define la estructura retórica de las unidades de aprendizaje en un entorno educativo

```
<!ELEMENT movies (movie)+>
<!ELEMENT movie (title, writer+, producer+, director+, actor*, comments?)>
<!ATTLIST movie
  type (drama | comedy | adventure | sci-fi | mystery | horror | romance |
    documentary) "drama"
  rating (G | PG | PG-13 | R | X) "PG"
  review (1 | 2 | 3 | 4 | 5) "3"
  year CDATA #IMPLIED>
<!ELEMENT title (#PCDATA)>
<!ELEMENT writer (#PCDATA)>
<!ELEMENT producer (#PCDATA)>
<!ELEMENT director (#PCDATA)>
<!ELEMENT actor (#PCDATA)>
<!ELEMENT comments (#PCDATA)>
```

Figura 4. DTD de una cartelera cinematográfica

En apenas un lustro de existencia, XML se ha convertido en la opción más generalizada para el tratamiento de la documentación digital. Existen centenares de aplicaciones en XML, que se plasman en otras tantas DTD disponibles (Robin Cover, 2003). Asimismo, XML se ha convertido en el estándar para el intercambio y publicación de datos en todos los ámbitos, pero especialmente en aquellos que se realizan a través de Internet. La principal ventaja de XML es que aporta un mecanismo sencillo y eficaz para facilitar el tratamiento de los contenidos. XML es la opción lógica de metalenguaje para *metacontenidos* en este momento.

## Metacontenidos vs. metadatos

Preferimos la palabra *metacontenido* a *metadato* porque es más sugerente desde la perspectiva de la edición digital. Hay pequeñas diferencias de matiz en el uso, aunque el término más habitual en la bibliografía especializada es *metadato*. *Metacontenido* sirve para enfatizar las implicaciones que la aplicación de los metalenguajes tiene en la gestión de contenidos. Nos sitúa en un territorio más próximo a la filología, como es la producción literaria y documental.

Podemos definir un documento como un objeto comunicativo que está constituido por contenidos organizados de un determinado modo. Los mismos contenidos pueden dar lugar a distintos documentos, si se cambia ligeramente su organización. Hay otros aspectos, además del orden, como el idioma, el estilo, fecha, número de acceso, etc., que pueden propiciar múltiples versiones de unos mismos contenidos. Por eso resulta más rentable que la gestión se realice directamente sobre los contenidos (por medio de metacontenidos), en lugar de sobre los documentos que los contienen (Bustelo Ruesta, 2003).

Pero esta forma de entender los contenidos es nueva. Hasta fechas recientes, no era posible concebir los contenidos de manera autónoma. Se veían como parte de un proceso comunicativo más completo, que se vuelca en un texto o documento. En el texto, sea manuscrito o impreso, los contenidos se organizan de manera estática y definitiva. La introducción del soporte electrónico ha propiciado un proceso de emancipación del contenido respecto al continente. Con la superación de una de las propiedades más características de los textos, su organización secuencial o lineal, el medio electrónico posibilita el tránsito del texto al hipertexto (Landow, 1992). Sin disposición lineal fija, los contenidos pueden estructurarse en múltiples rutas e itinerarios y configurar lecturas diferentes, en función de las preferencias o necesidades de los destinatarios. En el soporte electrónico además los contenidos se liberan de la condición estática del papel, de forma que pueden actualizarse, adaptarse o personalizarse dinámicamente, según el momento, lugar o perfil del lector. Veremos más adelante cómo los metacontenidos ayudarán a realizar estas operaciones.

El concepto de *metadato* proviene del mundo de la biblioteconomía y es muy utilizado en los sistemas de catalogación y en los servicios de préstamo interbibliotecario. Hasta fechas recientes la mayoría de las propuestas significativas de utilización de metadatos (MARC, AACR, Z39.50, ISO2709, METS, IMLS, etc.) han partido de este sector. Como reacción a los graves problemas de sobrecarga informativa y documental que ha creado la *web*, se ha originado en el mundo de Internet una pléyade de iniciativas que

recurren al empleo de metadatos como principal estrategia para solucionar la avalancha informativa imperante. En el momento presente la apuesta más conocida es la *web semántica* de Tim Berners-Lee (2000), a la que dedicaremos un apartado en las páginas siguientes.

El objetivo de los metadatos es poner orden en los datos y hacer más explícito su significado. El problema con los datos, igual que con los contenidos, o cualquier otro signo en general, es que no tienen significado por sí mismos. Lo adquieren en función de la interpretación que les damos las personas en situaciones concretas. Pongamos un ejemplo. En un ascensor, la comunicación entre personas y máquinas se realiza a través de un panel de números. El significado de cada número se relaciona con el desplazamiento que hará el ascensor a las plantas del edificio identificadas con esos números. Como sabemos, el significado que adquieren los números del ascensor no es intrínseco, ni universal, ni autónomo, sino que depende totalmente del contexto funcional; que en este ejemplo viene dado por el funcionamiento del ascensor y el conocimiento que de él tienen las personas que lo utilizan. El metadato “número de planta” ayudaría a hacer explícito el significado, pero no sería suficiente para hacerlo autónomo. Para lograrlo, habría que definir el metadato en relación con toda la funcionalidad del ascensor, y al propio ascensor en el contexto funcional del uso diario que de él hacen las personas.

Si extrapolamos el ejemplo a otras situaciones algo más complejas (recetas, manuales técnicos, textos científicos, enciclopedias, normativas, etc.), es fácil entender que la interpretación de datos y contenidos en su conjunto es un problema complejísimo. Hasta hace poco, en este proceso de interpretación sólo interveníamos los humanos (incluidos los expertos, especialistas y profesionales de cada rama) y por eso la percepción del problema era pequeña. Pero desde hace unas décadas también intervienen los ordenadores. Es más, desde 1992 prácticamente todos los conocimientos y sistemas conceptuales creados por la mente humana están siendo digitalizados y publicados en Internet. Obviamente, la simple digitalización de los datos no sirve de mucho. Hace falta diseñar programas que procesen los datos, que sean capaces de relacionar unos datos con otros y de asignarles significados. Tradicionalmente, para resolver el sentido de las palabras y de los singos en general se utilizaban glosarios, diccionarios, directorios, catálogos, nomenclaturas, jerarquías, ontologías, etc. Pero ahora todo esto no es suficiente. Sowa (2001) explica así el problema de la asignación de significado:

*Las ontologías contienen categorías, los diccionarios significados de palabras, la terminologías términos, los directorios direcciones, los catálogos índices y las bases de datos números, cadenas de caracteres y objetos binarios. Todas estas listas, jerarquías y redes son colecciones de signos estrechamente interconectadas. Pero las conexiones primarias no se dan en los bits y bytes que los codifican, sino en las cabezas de las personas que los interpretan. El objetivo de las distintas propuestas de metadatos es conseguir hacer explícitas esas conexiones marcando los datos con más signos.*

*Esos metasignos tendrán nuevas interconexiones, que pueden etiquetarse mediante metametasignos. Pero los datos sin significado no pueden adquirirlo al ser etiquetados con metadata sin significado. La fuente última de significado es el mundo físico y los agentes que usan los signos para representar entidades en el mundo y sus intenciones para con ellos.*

El gran reto que presenta el tratamiento a gran escala de la información es enseñar a las máquinas a comprender el significado de los contenidos que depositamos en ellas. Volveremos más adelante a considerar este problema al hablar de la *web semántica*. Vamos ahora a reseñar la evolución de los lenguajes de etiquetado y su incidencia en la edición digital.

### **3 Text Encoding Initiative (TEI)**

El principal hito en el uso de metacontenidos en el terreno más filológico viene dado por la publicación de las recomendaciones del consorcio internacional *Text Encoding Initiative* (TEI) en 1991, 1993 (1995). TEI se fundó el año 1987 en EEUU, por iniciativa de una asociación profesional de gran raigambre en el mundo de las humanidades, la *Association for Computers and the Humanities* (ACH). La propuesta fue inicialmente auspiciada por el organismo institucional encargado de sufragar la investigación en humanidades en EEUU, la *National Endowment for the Humanities* (NEH). En la asamblea constituyente, que tuvo lugar en el Vassar College, se sumaron a la iniciativa otras dos influyentes asociaciones, la *Association for Computational Linguistics* (ACL), que reunía por entonces a varios miles de especialistas en tecnologías lingüísticas a ambas orillas del Atlántico, y la *Association for Literary and Linguistic Computing* (ALLC), réplica europea de la ACH norteamericana.

El objetivo fundacional de TEI era establecer un conjunto de directrices comunes para la anotación electrónica de textos, de manera que se facilitara el intercambio y reutilización de recursos entre distintos centros. Con este fin se eligió la norma SGML. Esta elección fue providencial y particularmente clarividente si tenemos en cuenta que SGML acababa de ser presentado como estándar ISO y carecía de trayectoria. TEI emprendió la marcha a través de una serie de comités encargados de elaborar recomendaciones con objetivos que abarcaban desde cuestiones de catalogación hasta procedimientos de análisis textual. Las propuestas SGML se adaptaron a XML tan pronto como fue promovido a estándar por el consorcio W3C. De hecho varios miembros de TEI formaron parte del comité W3C encargado de la definición de XML.

TEI ha tenido un enorme eco en el ámbito de la filología (sobre todo en los trabajos de lexicografía, creación de corpora y en edición crítica), así como en la lingüística computacional (particularmente en la anotación de recursos y en la importación y exportación de datos). Con los años, las directrices TEI se han ampliado y especializado de diferentes maneras:

- EAGLES (*Expert Advisory Groups on Language Engineering Standards*), un proyecto auspiciado por la Comisión Europea, ha añadido criterios para la codificación de aspectos gramaticales que abarcan desde rasgos fonéticos hasta cuestiones de pragmática y discurso.
- Otro proyecto comunitario, PAROLE, se ha centrado en la creación de recursos léxicos. Uno de los mayores logros de PAROLE ha sido la propuesta de un conjunto homologado de etiquetas morfosintácticas para varias lenguas europeas (inglés, danés, neerlandés, francés, italiano, catalán, español, entre otras).
- Es destacable también MULTTEXT (*Multilingual Text Tools and Corpora*) que ha desarrollado programas modulares para la segmentación y etiquetado de corpus en varias lenguas europeas. Algunas de sus herramientas (como el segmentador MtSeg) han sido reutilizadas en otros proyectos (p. ej. en CRATER y CREA).
- CES (*Corpus Encoding Standards*) ha ampliado la cobertura de las anotaciones y ha abarcado un número mayor de lenguas. Entre sus principales logros está haber recopilado y anotado corpora multilingües con lenguas de Europa oriental.
- Más recientemente hay que citar ISLE (*International Standard for Language Engineering*) que propone estándares para la descripción de metadatos aplicados a recursos lingüísticos multimedia y multimodales.
- OLAC (*Open Language Archives Community*) es otra propuesta similar a ISLE que describe los recursos lingüísticos aplicando el modelo de metadatos DC (que se describe más abajo).

En la actualidad más de un centenar de proyectos de corpus y de edición electrónica utilizan las directrices de TEI (<http://www.tei-c.org/Applications/>), o alguno de sus derivados, y puede considerarse el principal modelo de anotación para tareas filológicas. En el caso del español, las iniciativas más conocidas son los corpora de referencia CORDE (Corpus de referencia diacrónico del español) y CREA (Corpus de referencia del español actual), de la Real Academia de la Lengua, así como las ediciones anotadas del Centro Virtual Cervantes (ambas representadas en este monográfico). Pero el abanico de proyectos en español no se agota con los dos proyectos mencionados. Para ampliar estos datos véase Marcos Marín, 1994; Martín de Santa Olalla, 1999; Pérez, 2002; y Abaitua, 2002. (En la sección bibliográfica se recogen además datos sobre corpora en euskara.)

## La cabecera de TEI

TEI posee como principal virtud un diseño de cabecera <teiHeader> con gran capacidad para incluir información documental, con prestaciones análogas a los sistemas de catalogación bibliográfica (Caplan, 2001; Wittenburg y Broader, 2002).

Una cabecera TEI consta de cuatro partes:

- la descripción del archivo <fileDesc>, que contiene una descripción bibliográfica completa que permita su citación
- una descripción sobre la codificación <encodingDesc>, en el que se precisan las incidencias en el momento de la transcripción
- perfil del texto <profileDesc> que resuelve aspectos contextuales (fecha de creación, categorización, etc.)

- historial de revisiones <revisionDesc> que permite llevar un registro de los cambios realizados sobre la versión electrónica

Por este motivo se puede decir que TEI aporta una metalenguaje sumamente adecuado para describir documentalmente un corpus textual. Un ejemplo de cabecera de un documento en formato TEI se observa en la siguiente figura.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE TEI.2 SYSTEM "teixlite.dtd">
<TEI.2>
<teiHeader>
<fileDesc>
<titleStmt>
<title>Premios fin de carrera</title>
<author>deli</author>
<funder>Proyecto P11999-72. Departamento de Educación, Universidades e Investigación del Gobierno Vasco</funder>
<principal>Proyecto XML-Bi; grupo DELi</principal>
</titleStmt>
<publicationStmt>
<publisher>Universidad de Deusto</publisher>
<availability><p>Documento de uso interno para personal de la Universidad de Deusto</p></availability>
</publicationStmt>
<notesStmt>
<note type="validated" resp="Itzulpen Bulegoa">Yes</note>
</notesStmt>
<sourceDesc><p>Texto de uso interno de la Universidad de Deusto</p></sourceDesc>
</fileDesc>
<encodingDesc>
<projectDesc><p>Proyecto XML-Bi: Diseño de procedimientos para la gestión de flujo documental multilingüe sobre estándares de marcación XML/TEI y TMX. Implementación experimental para la Universidad de Deusto, año 2002</p></projectDesc>
<classDecl>
<taxonomy id="DELi_SareBi_Tipologia_4">
<bibl>http://www.deli.deusto.es/AboutUs/Projects/XML-Bi/Admin/XML-Bi_I2_an1_tipologia4.txt</bibl>
</taxonomy>
</classDecl>
</encodingDesc>
<profileDesc>
<creation><date>2002/09/10 13:34:37.538 GMT+2</date></creation>
<langUsage>
<language id="es">español</language>
</langUsage>
<textClass>
<classCode scheme="DELi_SareBi_Tipologia_4">11303</classCode>
</textClass>
</profileDesc>
</teiHeader>
<text>
<front>
<docDate>2002/05/17</docDate>
<docAuthor>El Rector</docAuthor>
<head type="place">Bilbao</head>
<head type="DepSup">1100</head>
<head type="DepInf">1100</head>
</front>
```

Figura 5. Cabecera de un documento en el formato TEI

La información ofrecida en la cabecera <teiHeader> se complementa con la aportada en la sección <front> del propio texto, donde se recogen datos explícitamente contenidos en el documento, normalmente en la cabecera de la versión impresa. El ejemplo está tomado del sistema SARE-Bi, un gestor de documentos multilingües diseñado por el grupo DELi para la Universidad de Deusto (Díaz y otros, 2003). Los datos recogidos en el <front> son la fecha original del documento <docDate>, el nombre de la persona que lo firma <docAuthor>, el lugar en el que se firma y la filiación. Los códigos 1400 y 1406 de los atributos DepSup y DepInf identifican el centro al que pertenece el documento (se trata en este ejemplo del Departamento de Filología Inglesa de la Facultad de Filosofía y Letras). El sistema maneja 93 códigos de centros (facultades, institutos, departamentos, órganos de gobierno, servicios, etc.).

Un aspecto destacable de la cabecera TEI de SARE-Bi es la taxonomía de categorías documentales. En la versión actual de la taxonomía, que es la cuarta, se han contemplado tres niveles de categorización atendiendo a los criterios de función comunicativa, género y tema. Estos criterios están basados en la propuesta de clasificación tipológica de Trosborg (1997). En la actualidad existen 282 categorías, divididas de manera jerárquica entre tres funciones comunicativas (*reglamentar, informar e inquirir*), 25 géneros y 256 temas. El sistema demuestra las cualidades de TEI para la gestión documental. En el apartado siguiente vamos a mostrar cómo, además de los metadatos de la cabecera, TEI ofrece amplias posibilidades de hacer explícitos los contenidos de un texto.

## Consultas en corpus etiquetados mediante TEI

Existen varios niveles de anotación lingüística que pueden aplicarse en el aprovechamiento de los textos:

- *Anotaciones estructurales* que identifican los elementos que configuran la disposición del texto: epígrafes, párrafos, etc.
- *Anotaciones morfosintácticas* que asignan a cada unidad léxica un código que identifica su categoría morfosintáctica (su *part of speech*, POS), así como otras propiedades morfológicas generalmente asociadas a la flexión (género, número, persona, caso, tiempo, etc.)
- *Lematización*, proceso mediante el cual las formas flexionadas del corpus se emparejan con sus lexemas respectivos, es decir con la forma de citación tal como aparece en los diccionarios.
- *Análisis sintáctico* de las categorías sintagmáticas intraoracionales: grupos verbales y nominales, cláusulas subordinadas, etc.
- *Anotaciones orientadas a la tarea*: como pueden ser las unidades de traducción, o el etiquetado de referencia (numeración, citas, etc.).
- *Códigos de correspondencia*: etiquetas que hacen explícita la correspondencia entre unidades de traducción y que se asignan en el proceso de alineación de textos originales y traducidos.

Todos estos niveles pueden ser abordados mediante el conjunto de etiquetas propuesto por TEI, o por las posteriores ampliaciones de TEI en PAROLE, EAGLES, CES, MULTTEXT, u otras. El etiquetado basado en XML permite ampliaciones para cubrir cualquier necesidad, como se ha visto en la aplicación desarrollada por Barrutieta (2002).

El ejemplo siguiente ilustra la manera en que la tecnología XML permite llevar a cabo en los textos etiquetados operaciones similares a las que se realizan mediante expresiones regulares en las consultas a las bases de datos. Sobre una colección de obras de Shakespeare es posible realizar una búsqueda de elementos estructurales definidos como intervenciones (SPEECH) producidas por el personaje (SPEAKER) Otelo que contengan la palabra *black*. La consulta se realiza mediante Xpath y tiene la forma que se ve en la siguiente figura.

```
document(*)//SPEECH[LINE &='black' and SPEAKER &='OTHELLO']
```

Figura 6. Consulta con Xpath en corpus con obras de Shakespeare

Los tres resultados obtenidos en <http://exist-db.org/> se ofrecen en la figura 7.

```
<SPEECH exist:id ="22583" exist:source ="db/shakespeare/plays/othello.xml" >
<SPEAKER > OTHELLO </ SPEAKER >
<LINE > This fellow's of exceeding honesty, </ LINE >
<LINE > And knows all qualities, with a learned spirit, </ LINE >
<LINE > Of human dealings. If I do prove her haggard, </ LINE >
<LINE > Though that her jesses were my dear heartstrings, </ LINE >
<LINE > I'd whistle her off and let her down the wind, </ LINE >
<LINE > To pray at fortune. Haply, for I am black </ LINE >
<LINE > And have not those soft parts of conversation </ LINE >
<LINE > That chamberers have, or for I am declined </ LINE >
<LINE > Into the vale of years,—yet that's not much— </ LINE >
<LINE > She's gone. I am abused; and my relief </ LINE >
<LINE > Must be to loathe her. O curse of marriage, </ LINE >
<LINE > That we can call these delicate creatures ours, </ LINE >
<LINE > And not their appetites! I had rather be a toad, </ LINE >
<LINE > And live upon the vapour of a dungeon, </ LINE >
```

<pre> &lt;LINE &gt; Than keep a corner in the thing I love &lt;/ LINE &gt; &lt;LINE &gt; For others' uses. Yet, 'tis the plague of great ones; &lt;/ LINE &gt; &lt;LINE &gt; Prerogativd are they less than the base; &lt;/ LINE &gt; &lt;LINE &gt; 'Tis destiny unshunnable, like death: &lt;/ LINE &gt; &lt;LINE &gt; Even then this forked plague is fated to us &lt;/ LINE &gt; &lt;LINE &gt; When we do quicken. Desdemona comes: &lt;/ LINE &gt; &lt;STAGEDIR &gt; Re-enter DESDEMONA and EMILIA &lt;/ STAGEDIR &gt; &lt;LINE &gt; If she be false, O, then heaven mocks itself! &lt;/ LINE &gt; &lt;LINE &gt; I'll not believe't. &lt;/ LINE &gt; &lt;/ SPEECH &gt; </pre>
<pre> &lt;SPEECH exist:id ="22625" exist:source ="/db/shakespeare/plays/othello.xml" &gt; &lt;SPEAKER &gt; OTHELLO &lt;/ SPEAKER &gt; &lt;LINE &gt; By the world, &lt;/ LINE &gt; &lt;LINE &gt; I think my wife be honest and think she is not; &lt;/ LINE &gt; &lt;LINE &gt; I think that thou art just and think thou art not. &lt;/ LINE &gt; &lt;LINE &gt; I'll have some proof. Her name, that was as fresh &lt;/ LINE &gt; &lt;LINE &gt; As Dian's visage, is now begrimed and black &lt;/ LINE &gt; &lt;LINE &gt; As mine own face. If there be cords, or knives, &lt;/ LINE &gt; &lt;LINE &gt; Poison, or fire, or suffocating streams, &lt;/ LINE &gt; &lt;LINE &gt; I'll not endure it. Would I were satisfied! &lt;/ LINE &gt; &lt;/ SPEECH &gt; </pre>
<pre> &lt;SPEECH exist:id ="22643" exist:source ="/db/shakespeare/plays/othello.xml" &gt; &lt;SPEAKER &gt; OTHELLO &lt;/ SPEAKER &gt; &lt;LINE &gt; O, that the slave had forty thousand lives! &lt;/ LINE &gt; &lt;LINE &gt; One is too poor, too weak for my revenge. &lt;/ LINE &gt; &lt;LINE &gt; Now do I see 'tis true. Look here, Iago; &lt;/ LINE &gt; &lt;LINE &gt; All my fond love thus do I blow to heaven. &lt;/ LINE &gt; &lt;LINE &gt; 'Tis gone. &lt;/ LINE &gt; &lt;LINE &gt; Arise, black vengeance, from thy hollow cell! &lt;/ LINE &gt; &lt;LINE &gt; Yield up, O love, thy crown and hearted throne &lt;/ LINE &gt; &lt;LINE &gt; To tyrannous hate! Swell, bosom, with thy fraught, &lt;/ LINE &gt; &lt;LINE &gt; For 'tis of aspics' tongues! &lt;/ LINE &gt; &lt;/ SPEECH &gt; </pre>

Figura 7. Resultado de búsqueda con Xpath en corpus con obras de Shakespeare

Los corpora etiquetados ofrecen múltiples posibilidades de consulta. Añadimos al ejemplo anterior uno similar en que la búsqueda de pasajes en el texto original se resuelve con otros pasajes asociados que contienen correspondencias en algunas de las versiones traducidas. La consulta (figura 8) busca en las intervenciones del personaje Yago usos de la agrupación léxica *green-eyed*. Se obtiene (figura 9) sólo un pasaje.

<pre>document(*)//SPEECH[LG &amp;='green-eyed' and SPEAKER &amp;='IAGO']</pre>
--

Figura 8. Consulta con Xpath en corpus con obras de Shakespeare bilingües

<pre> &lt;SPEECH exist:id ="en_22557" source ="/othello_en.xml" &gt; &lt;SPEAKER &gt; IAGO &lt;/ SPEAKER &gt; &lt;LG &gt; O, beware, my lord, of jealousy; It is the green-eyed monster which doth mock The meat it feeds on; that cuckold lives in bliss Who, certain of his fate, loves not his wronger; But, O, what damned minutes tells he o'er Who dotes, yet doubts, suspects, yet strongly loves! &lt;/ LG &gt; &lt;/ SPEECH &gt; </pre>	<pre> &lt;SPEECH exist:id ="es_1985_22557" source ="/othello_es_1985.xml" &gt; &lt;SPEAKER &gt; IAGO &lt;/ SPEAKER &gt; &lt;LG &gt; ¡Guardaos de los celos, mi buen señor! Pues es monstruo de obscenos ojos que se goza con la carne que lo nutre. Feliz es el engañado que acepta su destino y desprecia a quien le robó su honra. Infortunio tiene quien ama, sin embargo, y además sospecha; quien sospechando, ardiente ama. &lt;/ LG &gt; &lt;/ SPEECH &gt; </pre>
<pre> &lt;SPEECH exist:id ="es_2000_22557" source ="/othello_es_2000.xml" &gt; &lt;SPEAKER &gt; IAGO &lt;/ SPEAKER &gt; &lt;LG &gt; ¡Oh, señor, nunca caigáis en el infierno de los celos! Es un monstruo de ojos verdes que hace sufrir a quien lo alimenta. El cornudo vive feliz si a sabiendas desprecia a quien le engaña; pero, en qué infierno cuenta el tiempo el que siente el amor y la duda, sin dejar de amar. &lt;/ LG &gt; &lt;/ SPEECH &gt; </pre>	<pre> &lt;SPEECH exist:id ="es_1929_22557" source ="/othello_es_1929.xml" &gt; &lt;SPEAKER &gt; IAGO &lt;/ SPEAKER &gt; &lt;LG &gt; ¡Oh, mi señor, cuidado con los celos! Es el monstruo de ojos verdes, que se divierte con la vianda que le nutre. Vive feliz el cornudo que, cierto de su destino, detesta a su ofensor; pero ¡oh, qué condenados minutos cuenta el que idolatra y, no obstante, duda; quien sospecha, y, sin embargo, ama profundamente! &lt;/ LG &gt; &lt;/ SPEECH &gt; </pre>

<pre>&lt;SPEECH exist:id="es_1985_1991" source ="/othello_es_1991.xml" &gt; &lt;SPEAKER &gt; IAGO &lt;/ SPEAKER &gt; &lt;LG &gt; Señor, cuidado con los celos. Son un monstruo de ojos verdes que se burla del pan que le alimenta. Feliz el cornudo que, sabiéndose engañado, no quiere a su ofensora; mas, ¡qué horas de angustia le aguardan al que duda y adora, idolatra y recela! &lt;/ LG &gt; &lt;/ SPEECH &gt;</pre>	<pre>&lt;SPEECH exist:id="es_1881_22557" source ="/othello_es_1881.xml" &gt; &lt;SPEAKER &gt; IAGO &lt;/ SPEAKER &gt; &lt;LG &gt; Señor, temed mucho a los celos, pálido monstruo, burlador del alma que le da abrigo. Feliz el engañado que descubre el engaño y consigue aborrecer a la engañadora, pero ¡ay del infeliz que aún la ama, y duda, y vive entre amor y recelo! &lt;/ LG &gt; &lt;/ SPEECH &gt;</pre>
---	--

Figura 9. Resultado de búsqueda con Xpath en corpus con obras de Shakespeare bilingües

Las variantes del ejemplo se corresponden con las traducciones al español de Jaime Navarra, 2000; Ediciones B; Ángel Luis Pujante, 1991, Austral; Manuel Ángel Conejero, 1985, Cátedra; Luis Astrana Marín, 1929, Aguilar; y Marcelino Menéndez Pelayo, 1881, Arte y Letras.

#### 4 Otras propuesta de metadatos

Se ha presentado TEI por constituir el modelo de referencia de la utilización de metadatos para la edición digital en el terreno filológico. Pero debemos citar otras propuestas de igual o mayor relieve, como son DCMI y RDF, en el plano general de gestión de información en Internet; así como TMX y XLIFF, en el más concreto de las tecnologías de traducción.

#### Dublin Core Metadata Initiative (DCMI)

El origen del *Dublin Core Metadata Initiative* (DCMI) se remonta al segundo congreso internacional de *World Wide Web* celebrado en Chicago en octubre de 1994. Según se expone en el portal de la propia organización, la idea se fraguó en las conversaciones de pasillo de cinco congresistas particularmente ilustres: Yury Rubinsky, representante de la empresa SoftQuad, especializada en sistemas de autor para SGML; Stuart Weibel, Eric Miller, y Terry Noreault, de *Online Computer Library Centre* (OCLC), organización responsable del servicio de préstamo interbibliotecario, introducido en 1979, en el que participan 43.559 bibliotecas de 86 países en todo el mundo; y Joseph Hardin, del *National Center for Supercomputing Applications* (NCSA), centro particularmente influyente en los comienzos de la *web* por la creación en 1992 de Mosaic, el primer navegador gráfico para la *web*, y antecesor de Netscape. Rubinsky presidía en aquel evento los paneles sobre el futuro de HTML y las herramientas de autor para la Web. Stuart Weibel y Eric Miller, además de presentar una ponencia sobre la publicación de ediciones críticas en la *web*, tuvieron unas particiones muy destacadas en los debates sobre la situación de los servicios de préstamos interbibliotecarios por *web*. Noreault, por su parte, era entonces el director de la oficina de investigación del OCLC y Hardin era codirector del NCSA. Las conversaciones giraron en torno a la dificultad de encontrar recursos en Internet (cuando solo había unos 500.000 objetos publicados).

El resultado de estos encuentros llevó a NCSA y a OCLC a organizar un seminario conjunto en Dublin, Ohio (EEUU), en marzo de 1995. En este acto, que se llamó simplemente *OCLC/NCSA Metadata Workshop*, más de cincuenta personas se pusieron de acuerdo en que un conjunto nuclear de elementos semánticos para describir los recursos de la *web* harían mucho más fácil su búsqueda y recuperación. Llamaron al resultado el "Dublin Core metadata", en honor al lugar en el que se desarrolló el seminario.

El conjunto de metadatos nuclear (*Dublin Core Metadata Element Set*), que se consideró esencial para posibilitar la recuperación de documentos electrónicos, lo componen estos quince elementos, según la relación ofrecida por Weibel (1997). Como es fácil comprobar, representan una alternativa simplificada al modelo de cabecera que propone TEI:

- *Title* (Título). El nombre dado al documento electrónico por el autor o editor.

- *Author or Creator* (Autor). Personas u organizaciones responsables del contenido intelectual del documento. (p. ej., autores en el caso de documentos escritos; artistas, fotógrafos o ilustradores, en el caso de recursos visuales).
- *Subject and Keywords* (Asunto y palabras clave). Indican el tema del documento electrónico y pueden definirse a partir de sistemas de clasificación internacionales (CDD, CDU, LCSH), de tesauros, o en última instancia mediante una palabra o conjunto de palabras seleccionadas por el autor.
- *Description* (Descripción). Describe el contenido, bien mediante un resumen del documento o mediante otras explicaciones en el caso de recursos visuales.
- *Publisher* (Editor). Entidad responsable de la presentación del documento en la forma actual. Normalmente serán editoriales, universidades u otras entidades corporativas.
- *Other Contributors* (Otros Colaboradores). Otras personas que hayan contribuido en la realización de la obra (editores, traductores, ilustradores, etc.).
- *Date* (Fecha). La fecha en la que el documento fue publicado en la forma actual.
- *Resource Type* (Tipo de recurso). Hace relación a la categoría tipológica del recurso (género, función, etc.); p.ej., página web, informe, texto de ayuda, registro de datos, hoja de cálculo, etc.
- *Format* (Formato). Naturaleza física del documento electrónico: Postscript, PDF, HTML, XML, etc.
- *Resource Identifier* (Identificador). Serie o número usado para identificar el documento (URL, ISBN etc.).
- *Source* (Fuente). El documento (impreso o electrónico) a partir del cual se ha generado el recurso electrónico.
- *Language* (Idioma). Idioma en el que se expresa el contenido intelectual del documento.
- *Relation* (Relación). Establece la relación del documento con otros documentos (impresos o electrónicos), p.ej., las imágenes que se incluyen en el documento, los capítulos de un libro, o volúmenes de una colección.
- *Coverage* (Cobertura). Ámbito espacial o duración temporal que caracteriza al documento.
- *Rights Management* (Propiedad intelectual). Información sobre los derechos de propiedad intelectual o copyright.

El año siguiente a la celebración del *Dublin Metadata Workshop*, tuvo lugar un segundo seminario, el *Warwick Workshop*, organizado con el propósito de dar continuidad a los debates en torno a la definición de los metadatos. Se concretaron mejor los metadatos nucleares, conocidos como *Dublin Core* (DC), y se dejaron listos para ofrecer interoperatividad a los editores de documentos de la *web*, a los sistemas de catalogación, indizadores y otros sistemas de recuperación de información. El resultado del seminario de Warwick se conoce como *Warwick Framework*.

Es interesante comprobar que los metadatos DC son en realidad *metametadatos*, ya que pueden expresarse en diversos esquemas de metadatos. Andy Powell (2001), del centro UKOLN de la Universidad de Bath, ha diseñado un programa que permite generar cabeceras DC para cualquier documento publicado en Internet y cuyos resultados pueden obtenerse en los formatos HTML, RDF, TEI y otros.

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.title" content="UKOLN: DC-dot Dublin Core metadata editor" />
<meta name="DC.creator" content="Andy Powell" />
<meta name="DC.subject" content="Dublin Core; DC; generator; editor; Warwick Framework; SOIF; TEI; USMARC; XML; GILS; ROADS; RDF; IMS" />
<meta name="DC.description" content="A CGI based Dublin Core metadata generator" />
<meta name="DC.publisher" content="UKOLN, University of Bath" />
<meta name="DC.date" scheme="W3CDTF" content="2001-12-11" />
<meta name="DC.type" content="Text" />
<meta name="DC.format" content="text/html" />
<meta name="DC.identifier" content="http://www.ukoln.ac.uk/metadata/dcdot/" />
<meta name="DC.rights" content="http://www.ukoln.ac.uk/metadata/dcdot/COPYING" />
```

Figura 10. Cabecera DC en HTML del CD-dot generado por el mismo CD-dot de Andy Powell

```
<?xml version="1.0" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://dublincore.org/documents/overview/">
  <dc:title>Overview of documentation for DCMI Metadata Terms</dc:title>
```

```
<dc:description>An overview of official documentation of all DCMI metadata terms.</dc:description>
<dc:date>2003-05-27</dc:date>
<dc:format>text/html</dc:format>
<dc:language>en</dc:language>
<dc:publisher>Dublin Core Metadata Initiative</dc:publisher>
</rdf:Description>
</rdf:RDF>
```

Figura 11. Documento DC en RDF de la propia DCMI

## La web semántica y RDF

La *web* semántica es una propuesta realizada por el propio inventor de HTML y fundador de la W3C, Tim Berners-Lee, en septiembre de 1998, y trata de aprovechar la utilidad de los metadatos para organizar mejor el contenido de la *web*. La idea retoma la mayoría de las cuestiones que se plantearon en 1995 en el Dublin Workshop, pero tiene su principal antecedente en el diseño de RDF (*Resource Description Framework*), cuyo primer borrador lo presentó el W3C en agosto de 1997. El objetivo de RDF fue solucionar el acceso y gestión de contenidos en la web mediante un lenguaje de metacontenidos. Su desarrollo es paralelo a XML y entre ambos se producen algunas interferencias. RDF puede expresarse en XML, aunque no necesariamente. Por otro lado RDF suele adoptar el esquema de datos DC. Es muy utilizado en la actualidad para el intercambio de información y alguna de sus aplicaciones, como RSS para titulares de noticias, tienen una gran difusión.

La idea de web semántica se ha planteado con el objetivo de facilitar la interpretación y potenciar la interoperatividad entre distintas variantes de RDF. RDF es importante para la descripción de los objetos y los tipos de objetos que se encuentran en la red (a los que se suele llamar “recursos”). Por encima de este nivel se necesita, según Berners-Lee (1998), un estrato ontológico (*ontology layer*), porque para describir las relaciones entre los recursos y sus tipos (como puede ser “esta es una propiedad transitiva”) hacen falta ontologías, aunque dichas ontologías no sirvan para decidir cómo debe usarse una relación ontológica de forma computacional. Aparentemente, en muchos sectores se ha detectado la necesidad de contar con ontologías estandarizadas y éste es un trabajo que ha emprendido la W3C dentro de la iniciativa de la web semántica. Sin embargo, como explica Sowa (2000), las ontologías no resuelven completamente el problema:

*Muchas ontologías para objetos de web ignoran los objetos físicos, los procesos, las personas y sus intenciones. Un ejemplo típico es SHOE (Simple HTML Ontology Extensions), en la que se consideran sólo cuatro categorías básicas: cadena, número, fecha y verdad (Heflin et al. 1999). Esas cuatro categorías, necesarias para describir la sintaxis de los datos en la web, no pueden por ellas describir la semántica. Las cadenas son secuencias de caracteres que representan afirmaciones sobre el mundo; los números sirven para contar y medir las cosas; las fechas son unidades de tiempo sujetas a la rotación de la tierra; y la verdad es un término del metalenguaje para hacer correspondencias entre las afirmaciones y el mundo. Esas categorías sólo pueden definirse en relación con el mundo, las personas en el mundo y las lenguas que las personas utilizan para hablar sobre el mundo. Sin esas definiciones, las categorías son etiquetas carentes de significado que no aportan ninguna interpretación a los datos sobre los que se aplican.*

Al debatir la *Resource Description Framework* (RDF), Bray (1998) presentó una visión bastante clarividente sobre las categorías que los metadatos de la web deberían representar:

*Parece improbable que un “tipo de propiedad” (PropertyType) por sí mismo pueda ser muy útil. Sería de esperar que estos tipos vengan en paquetes; por ejemplo, un conjunto de propiedades como autor, título, fecha, etc. Luego un conjunto más elaborado de OCLC y otro alternativo de la Biblioteca del Congreso. Estos paquetes [de metadatos] se llaman vocabularios. Es fácil imaginar vocabularios de propiedades que describan libros, vídeos, ingredientes de pizza, vinos de primera calidad, fondos de inversión inmobiliaria y todo tipo de vida salvaje de la web. Pero la cuestión es: ¿Cómo se relacionan los paquetes entre sí? ¿Cómo se relaciona la propiedad fecha de OCLC con el año de cosecha de un paquete de productos enológicos? ¿Pueden los paquetes heredar las definiciones de otros paquetes? Si dos paquetes compiten entre sí, ¿hay algún tipo de filtro que permita traducir o redefinir las propiedades de uno en las del otro? Un*

*lector humano puede saber que la fecha de cosecha es comparable con la fecha en una ficha OCLC, pero si no hay una definición formal, el ordenador por sí solo no puede saberlo.*

Parecerá una ironía, pero las redes informáticas que hacen que los datos se transmitan con mayor facilidad hacen también que la posibilidad de compartirlos sea más compleja. Continuando con su exposición, Bray toca estas otras cuestiones:

*Nadie debería creer que todo el mundo usará el mismo vocabulario (tampoco deberían usarlo), pero con RDF llegaremos a tener un enorme mercado de vocabularios. Cualquiera los puede diseñar, anunciar y vender. Los mejor promocionados sobrevivirán y prosperarán. Posiblemente, la mayoría de los nichos de información estarán dominados por unos pocos vocabularios, como ya sucede con los catálogos de biblioteca.*

*En la actualidad ya existen miles, si no millones de vocabularios en competencia. Las tablas y campos de cualquier base de datos, o las listas de los catálogos de productos de cualquier negocio en el mundo son vocabularios incompatibles entre sí. Cuando se distribuían en papel estos catálogos, cualquiera, fuera ingeniero o contratista, podía leerlos y comparar sus descripciones. Pero pequeñas variaciones en la terminología de los catálogos informatizados harán imposible que un sistema informático pueda comparar los componentes de diferentes proveedores.*

Con la estandarización de las anotaciones, XML y RDF realizan un primer avance en la resolución del problema, pero es un paso insuficiente si además de compartir los datos no hay alguna manera de comparar, relacionar o traducir los vocabularios. Phipps (2000) ya advirtió de que la estandarización de los vocabularios crearía más dificultades “al ocultar las complejidades que se esconden bajo la apariencia de acuerdos”:

*No será viable un negocio electrónico que no pueda compartir datos en XML, pero incluso con él, los análisis de sistemas que se necesitan para obtener la traducción de los datos suponen un significativo problema. No debemos asumir que XML es la panacea o que la estandarización de vocabularios conllevará automáticamente la interoperatividad. XML aporta un medio de expresar nuestra percepción del significado de los datos, pero todavía tendremos que discernir las realidades y diferencias entre los significados cuando se intercambian los datos.*

Según Sowa (2000), más importante que estandarizar los vocabularios es desarrollar métodos para definir y traducir vocabularios divergentes. Para poseer una semántica y pragmática consistentes, esos métodos deberían relacionar los términos de los vocabularios con las cosas a las que se refieren y con la gente que los usa para comunicar información sobre esas cosas. Los métodos de la lógica y las ontologías pueden usarse para definir, relacionar y traducir los signos de un vocabulario a otro. Esa es la razón del interés en diseñar lenguajes para ontologías en la web y el principal objetivo de la web semántica.

Para Berners-Lee (1998) el término “semántica” significa específicamente “tratable computacionalmente”, que es una interpretación distinta a la que se usa en lingüística. En esta nueva acepción, la semántica debe transmitir lo que las máquinas deben hacer con los datos. El *test* semántico será el que permita comprobar si la máquina ha sabido tratar correctamente un dato dado o no. Con todo, pese a lo que pueda parecer, esta semántica, al igual que todo buen sistema simbólico, debe ser declarativa y no procedimental; es decir, debe servir para expresar lo que un dato “significa” y no lo que se quiere hacer con él.

Un ejemplo de lenguaje declarativo para ontologías en la web es OWL (*Web Ontology Language*). En la figura 12 se muestra la forma de definir hechos.

```
<fact> ::= <individual>
<individual> ::= Individual( [<individualID>] {<annotation>}
                           {type(<type>)} {<propertyValue>} )
<propertyValue> ::= value( <individualvaluedPropertyID> <individualID> )
                   | value( <individualvaluedPropertyID> <individual> )
                   | value( <datavaluedPropertyID> <dataLiteral> )
```

Figura 12. Esquema para la definición de hechos en OWL

En la definición de *hecho (fact)* que se muestra en la figura 12, se estipula información sobre un tipo particular de individuo, consignando las clases a las que el tipo pertenece, además de sus propiedades y valores. A ese individuo se le asignará un identificador (individualID) que lo denotará y que se utilizará para hacer referencias sobre él.

El lenguaje OWL permite todas las operaciones que son propias de una ontología, como crear descripciones (figura 13) sobre clases de individuos, con restricciones y relaciones con otras descripciones:

```
<description> ::= <classID>
| <restriction>
| unionOf( {<description>} )
| intersectionOf( {<description>} )
| complementOf( <description> )
| oneOf({<individualID>})
```

Figura 13. Esquema para la definición de descripciones en OWL

La efectividad que la generalización de las ontologías tendrá en el desarrollo de la web es todavía una incógnita, pero a la vista de sus ventajas y el ímpetu con que se está extendiendo la idea, podemos presagiar que marcarán un nuevo hito en la gestión de la información en un futuro próximo.

## 5 Metadatos para traducción

Vamos a exponer en esta sección un aspecto de la utilización de metadatos que es de especial interés para los autores: la utilización de metadatos en el marco de las tecnologías de traducción.

### Gestión de traducciones en TMX y XLIFF

TMX (*Translation Memory eXchange*) es un formato diseñado por la asociación LISA para conseguir uno de los desiderata de XML, la interoperatividad entre sistemas diferentes. El tipo de sistemas para los que TMX se ha diseñado son los denominados gestores de memorias de traducción (DéjàVu, WordFast, Transit, TranslatorWB), una familia de aplicaciones que agiliza y potencia sobremanera el trabajo del traductor humano y que en la última década ha arraigado con fuerza en el sector de las traducciones, sobre todo en grandes corporaciones públicas y privadas (Fernández García, 2003).

```
<?xml version="1.0" ?>
<!DOCTYPE tmx SYSTEM "tmx11.dtd">
<tmx version="version 1.1">

<header
  creationtool="www.deli.deusto.es"
  creationtoolversion="1.0"
  segtype="paragraph"
  o-tmf="Deli-Bilingual-TMX"
  adminlang="EN-US"
  srclang="ES"
  datatype="PlainText"
  creationdate="20030702T121700Z"
  creationid="Deli"
>
</header>
<body>

<tu>
<tuv lang="ES">
<seg>ORDEN 17 de mayo de 2002, del Recotorado de la Universidad de Deusto por la que se promulga la creación de los Premios Extraordinarios Fin de Carrera, en la Universidad de Deusto.</seg>
</tuv>
<tuv lang="EU">
<seg>Deustuko Unibertsitateko Errektoregoaren AGINDUA, 2002ko maiatzaren 17koa, Deustuko Unibertsitatearen Karrera Amaierako Sariak sortzeko emana.</seg>
</tuv>
</tu>
```

```

<tu>
<tuv lang="ES">
<seg></seg>
</tuv>
<tuv lang="EU">
<seg></seg>
</tuv>
</tu>

<tu>
<tuv lang="ES">
<seg>De acuerdo con lo aprobado por el Consejo de Dirección de esta Universidad celebrado el 16 de mayo de 2002,
promulgo la creación de los Premios Extraordinarios Fin de Carrera, así como la Normativa correspondiente, según el
anexo que se adjunta.</seg>
</tuv>
<tuv lang="EU">
<seg>Unibertsitate honetako Zuzendaritza Kontseiluak 2002ko maiatzaren 16an onetsitakoaren arabera, Karrera
Amaierako Sarriak sortu direla aldarrikatzen dut, baita haien araudia ere, erankinean ageri denez.</seg>
</tuv>
</tu>

<tu>
<tuv lang="ES">
<seg></seg>
</tuv>
<tuv lang="EU">
<seg></seg>
</tuv>
</tu>

<tu>
<tuv lang="ES">
<seg>El Rector</seg>
</tuv>
<tuv lang="EU">
<seg>Errektorea</seg>
</tuv>
</tu>

<tu>
<tuv lang="ES">
<seg></seg>
</tuv>
<tuv lang="EU">
<seg></seg>
</tuv>
</tu>

<tu>
<tuv lang="ES">
<seg>Bilbao, 17 de mayo de 2002</seg>
</tuv>
<tuv lang="EU">
<seg>Bilbao, 2002ko maiatzaren 17a</seg>
</tuv>
</tu>
</body>
</tmx>

```

Figura 14. Ejemplo de segmentos de traducción en TMX

Lo que muestra el ejemplo de la figura 14 es un documento bilingüe. Las memorias de traducción expresadas en TMX se convierten en grandes colecciones de elementos <tu> (unidades de traducción), que equivalen en la práctica a diccionarios de frases hechas. En lugar de almacenar palabras sueltas, se almacenan oraciones, párrafos e incluso documentos enteros. Este tipo de tecnologías son muy exigentes desde el punto de vista de consumo de memoria informática, pero, como es bien sabido, en la actualidad las posibilidades de almacenamiento son inmensas y los precios de los dispositivos de almacenamiento han disminuido considerablemente; luego las exigencias de memoria no suponen ningún problema.

La gestión de traducciones mediante metadatos y su distribución en comunidades de usuarios es una actividad todavía incipiente pero que tendrá una gran expansión en los próximos años. Podemos poner un

ejemplo (figura 15) que ilustra un problema muy frecuente en la traducción, la proliferación de variantes, que una gestión adecuada de traducciones habría evitado.

<p>“Cuando exista una diferencia de altura entre las rasantes del perímetro del edificio tal que permita iluminar o dar acceso independiente a locales de semisótano, éstos contabilizarán en el cómputo de la superficie construida en una proporción igual a la relación entre la superficie de su fachada sobre rasante respecto a la superficie de su cerramiento perimetral, se halle enterrado o no, computado desde el plano horizontal definido por la rasante de menos cota.”</p>	<p>“Eraikinaren perimetroko sestren artean, erdisotoko lokalak argitu edo aparteko sarbidea emateko bestekoa den altuerako tarte bat dagoenean, hauek, eraikitako azaleraren konputuan kontabilizatuko dira, sestraren gaineko euren fatxadaren azaleraren pareko proportzioan, perimetrozko euren zarraketaren azalerari dagokionean, hau lurarren azpian hala gainean dagoela, kotarik txikiena duen sestrak definitutako plano horizontaletik konputatua.”</p>	<p>“Eraikinaren perimetroko sestren arteko altuera diferentzia dagoenean, eta diferentzia horrek erdisotoko lokalak argitu edo horiei sarbide independentea ematen badie, horiek azalera eraikian kontabilizatuko dira, sestra gaineko fatxadaren azaleraren eta perimetroko itxiduraren azaleraren arteko erlazioaren proportzio berean, itxidura hori lurperaturik egon ala ez, eta kota txikieneko sestrak definitutako plano horizontaletik konputaturik.”</p>
<p>Texto original</p>	<p>Variante 1</p>	<p>Variante 2</p>
	<p>“Eraikinaren perimetroaren sestren arteko altueran, erdisotoko lokalak argitzea edo sarrera bananduak ahalbideratzeko aldea dagoenean, hauek eraikitako azaleraren zenbatekoan sartuko dira, bere sestraren gaineko fatxadaren azalera bere itxitura perimetralaren azalarekiko proportzio batean, lurperaturik egon edo ez, eta kota txikiak definitutako plano horizontaletik zenbatua.”</p>	<p>“Eraikinaren perimetroko lerrokaduren artean dagoen altuera-diferentziak erdisotoak argitzeko edo bertako lokaletara sartzeko modua ematen badu, erdisoto horiek eraikitako azaleraren konputoan sartuko dira, eta sartu ere sestra gainean duten fatxadaren azaleraren eta perimetroaren -zorupekoa zein ez-azaleraren arteko proportzio berean sartuko da.”</p>
	<p>Variante 3</p>	<p>Variante 4</p>

Figura 15. Variantes en una traducción

La existencia de variantes de traducción, como hemos visto en el caso de Shakespeare, desvela un hecho en sí mismo enriquecedor, como es que varias manos intervengan para ensayar versiones de un mismo texto de partida. De esta forma, lo que no se hace generalmente con el texto original (que por considerarse canónico queda fijado a perpetuidad), si se hace con sus traducciones. Éstas se corrigen y adaptan según los gustos de cada época, las preferencias estilísticas del editor, o simplemente por motivos económicos.

Pero lo que se considera positivo en el terreno literario, puede cambiar de signo en otros ámbitos, como obviamente en el jurídico y administrativo, pero también en el científico, el técnico, el divulgativo o incluso el informativo. El texto del ejemplo ilustra un problema de la traducción administrativa en euskara que dificulta la implantación de un canon (de intertextualidad) en los documentos redactados en esa lengua. Los sistemas de gestión de memorias de traducción representan una solución tecnológica muy útil para comenzar a paliar este problema. Y TMX es el estándar que los nutre.

TMX sin embargo no es suficiente, ya que carece por ejemplo de mecanismos adecuados para realizar el control de variantes (figura 15). Las variantes forman parte de lo que se denomina el *ciclo de vida* de la documentación. Este es un concepto que abarca todas las fases y facetas que un documento adquiere a lo largo de su existencia, desde que se concibe hasta que se publica, e incluso más allá, cuando se cataloga, archiva y recupera para su consulta o reutilización. Para tratar estas cuestiones se ha definido un formato inspirado en TMX con nuevas prestaciones: XLIFF (figura 17).

```

<header>
<phase-group>
<phase phase-name="complete"
process-name="publication">
contact-email="decanato@fil.deusto.es"
date="2003/03/12 11:30:23.243 GMT+2"/>
<phase phase-name="validated"
process-name="publication">
contact-email="sara@irakaslego.deusto.es"
date="2003/03/13 10:22:19.118 GMT+2"/>
</phase-group>
</header>
...
<trans-unit id="3">

```

```

<source xml:lang="es">Que este Departamento procurará que la investigadora encuentre un entorno de
trabajo adecuado y que pondrá a su disposición los medios necesarios para garantizar una labor formativa óptima.
</source>
<target xml:lang="eu">Sail honek ahalegina egingo duela ikertzaileak behar dituen lan-giro eta baliabideak
izan ditzan bere prestakuntza aurrera eramateko. </target>
<alt-trans>
<source xml:lang="es" phase-name="complete">Que este Departamento procurará que el
investigador encuentre el clima y los medios que precise para llevar adelante su labor formativa como investigador.
</source>
<target xml:lang="eu" phase-name="complete">Sail hau saiatuko da ikertzaileak bere prestakuntza
lana burutzeko behar dituen giro eta baliabideak ematen.
</target>
</alt-trans>
<alt-trans>
<source xml:lang="eu" phase-name="validated">Que este Departamento procurará que la
investigadora encuentre un entorno de trabajo adecuado y que pondrá a su disposición los medios necesarios para
garantizar una labor formativa óptima. </source>
<target xml:lang="eu" phase-name="validated">Sail honek ahalegina egingo duela ikertzaileak
behar dituen lan-giro eta baliabideak izan ditzan bere prestakuntza aurrera eramateko. </target>
</alt-trans>
</trans-unit>

```

Figura 16. Documento en el formato XLIFF

En los próximos años veremos una expansión de recursos traductológicos que utilizan los formatos TMX y XLIFF. Queda por resolver la interacción de estos recursos con otros formatos, RDF, TEI, etc.; pero esta es una cuestión secundaria de fácil resolución, como se ha demostrado en el sistema SARE-Bi (Díaz y otros, 2003).

## 6 Recolección y sindicación de metacontenidos

Una de las principales ventajas que ofrece Internet es el acceso universal a los contenidos. Pero esta ventaja conlleva el importante inconveniente de la sobrecarga informativa. Hemos visto cómo la aplicación de metadatos puede ayudar a paliar el problema. También hemos hablado del efecto beneficioso de poder intercambiar y compartir datos entre agentes productores y consumidores. Vamos a analizar ahora de qué forma se puede optimizar este intercambio.

### Enfoques alternativos: federación, recolección y acumulación

En los EEUU, el organismo encargado de auspiciar la investigación científica, la *National Science Foundation* (NFS), acaba de promover la creación de una comunidad nacional de bibliotecas digitales SMETE (*National SMETE Digital Library*, NSDL). Uno de los especialistas implicados en la gestión de la idea, Warnik (2003), explica que durante la fase de estudio se han considerado tres modos de enfocar la cooperación: *federación*, *recolección (harvesting)* y *acumulación (gathering)*. La diferencia entre ellos radica en dos factores: por un lado, en el grado de implicación que cada biblioteca participante deberá asumir; y, por otro, en los beneficios que la integración reportarán al “descubrimiento” (*discovery*) de la información (que es directamente proporcional al grado de implicación).

El enfoque con mayor carga es el de federación, ya que las organizaciones que deciden participar deben ponerse de acuerdo en los protocolos y estándares de operatividad para construir sobre ellos todos los sistemas que componen la federación. A través de la federación se deben realizar por adelantado importantes esfuerzos para llegar a acuerdos en materia de organización, contenidos y tecnología.

En la comunidad bibliotecaria se encuentran excelentes ejemplos de federación, como son los formatos Z39.50 y MARC, o las reglas angloamericanas de catalogación AACR2. Las actividades federadas suelen tener costes normalmente elevados y para definir, implantar y mantener sus estándares se precisan importantes recursos. En consecuencia, el número de asociados a este tipo de federaciones suele ser reducido.

Al otro lado del espectro, el enfoque con menor carga es el de acumulación (*gathering*). No requiere ningún tipo de cooperación, ya que únicamente depende de motores de búsqueda, al estilo de Google, con los que basta para propiciar el descubrimiento de información, además de algunas aplicaciones especiales,

como el servicio PrePRINT (<http://www.osti.gov/preprint>). Este enfoque es una réplica de la situación actual de la *web*. La ventaja es que no precisa de grandes inversiones.

En mitad del espectro se sitúa el enfoque de recolección (*harvesting*), que mitiga algunos de los inconvenientes tanto de la federación como de la acumulación. La recolección rebaja las exigencias de participación a la vez que facilita el acceso a las colecciones de las bibliotecas digitales. Es el enfoque más adecuado, según Warnik, para muchas tareas del descubrimiento de información.

## Sindicación de metacontenidos

El enfoque de recolección tiene importantes coincidencias con una práctica ya establecida en Internet que se conoce como *sindicación*. La sindicación de *webs* ha alcanzado gran popularidad a medida que más sitios se referencian unos a otros; y no sólo por medio de enlaces, sino compartiendo gran parte de sus contenidos. La idea se remonta al servicio en XML que ofrecía Netscape *Rich Site Summary* (RSS) (<http://www.oasis-open.org/cover/rss.html>). RSS se desarrolló a comienzos de 1999 para llenar el portal My Netscape con entrada de noticias externas o canales (*channels*). Desde entonces, RSS ha adquirido vida propia y ahora miles de sitios web usan RSS para dar actualidad y atraer visitas.

```
<?xml version="1.0" ?>
<!DOCTYPE rss PUBLIC "-//Netscape Communications//DTD RSS 0.91//EN" "http://www.scripting.com/dtd/rss-0_91.dtd">
<!-- http://my.netscape.com/publish/formats/rss-0.91.dtd-->
<rss version="0.91">

<channel>
<title>The XML Cover Pages</title>

<description>A comprehensive online reference work for SGML/XML applications and related standards. The reference collection features news, bibliography, events calendar, and extensive documentation on the application of open, interoperable (meta) markup language standards, including SGML, XML, XML Schema, XSL, XSLT, XPath, XLink, XHTML, DOM, XPointer, HyTime, DSSSL, CSS, SPDL, CGM, ISO-HTML, etc.</description>
<link>http://xml.coverpages.org/covernews.xml</link>
<language>en-us</language>
<managingEditor>robin@isogen.com (Robin Cover)</managingEditor>
<image>
<title>The XML Cover Pages</title>
<url>http://xml.coverpages.org/images/rclogo88.gif</url>
<link>http://xml.coverpages.org/</link>
<width>88</width>
<height>23</height>
</image>

<item><title>Microsoft, IBM, and VeriSign Promote WS-Security Specifications for Web Services</title><link>http://xml.coverpages.org/ni2002-04-11-b.html</link><description>An announcement from Microsoft, IBM, and VeriSign promotes a WS-Security specification which defines a standard set of Simple Object Access Protocol (SOAP) extensions, or message headers. Six companion security specifications are also presented. WS-Security is designed to help organizations build secure, broadly interoperable Web services applications.</description></item>

<item><title>W3C XML Core Working Group Publishes New Working Drafts for Namespaces in XML</title><link>http://xml.coverpages.org/ni2002-04-11-a.html</link><description>The XML Core Working Group has issued two working draft specifications on XML namespaces. 'Namespaces in XML 1.1' is the first draft of a new 1.1 revision of the 'Namespaces in XML' specification which will incorporate several errata to the 1.0 specification, and will make one substantive change applicable to XML version 1.1 instances: the provision of a mechanism to 'undeclare' prefixes.</description></item>

<item><title>University of Hong Kong E-Commerce Center Opens Test Site for OASIS ebXML V2 Registry Implementation</title><link>http://xml.coverpages.org/ni2002-04-10-d.html</link><description>Researchers at the University of Hong Kong's Center for E-Commerce Infrastructure Development (CECID) and Department of Computer Science Information Systems have released a publicly-accessible (beta version) test site which implements the OASIS ebXML Version 2 Registry. This ebXML Registry is one of four ebXML architectural components in the Center's Project Phoenix.</description></item>

<item><title>Siebel Systems Announces XML-Based Universal Application Network</title><link>http://xml.coverpages.org/ni2002-04-10-c.html</link><description>Siebel Systems recently announced a 'Universal Application Network' as a standards-based architecture for end-to-end business process management. The Universal Application Network uses XML/XSLT, and is comprised of three major components: a comprehensive business process library, a state-of-the-art business process design tool, and a market-leading integration server.</description></item>
```

```
</channel>
</rss>
```

Figura 17. Ejemplo de RSS tomado de XML Cover Pages

El estándar RSS 1.0 es una aplicación de RDF, el marco para describir e intercambiar metadatos que la W3C definió en 1997. Como XML, RDF es ampliable y permite añadir nuevos tipos de entidades. Es, como hemos visto, una forma de dar significado a los recursos para permitir el procesamiento automático en la web. RSS es sin duda de algo que ha crecido orgánicamente y que es ampliamente aceptado como estándar. Durante bastante tiempo no fue reconocido por ningún comité de normalización. Pero incluso así llegó a ser muy popular, siendo utilizado de manera muy creativa. Ahora se considera que ha alcanzado su límite y hay demanda para sindicaciones de portales que RSS no puede aportar. Por ello están surgiendo ampliaciones que se solapan con la actividad en torno a la web semántica, por ejemplo OCS (*Open Content Syndication*). Entre las cosas que ofrece OCS está el formato de directorio OCS, diseñado para hacer listados de canales que puedan usarse desde distintos portales, con software de titulares basado en el cliente y otras aplicaciones similares.

## Ejemplos de recolección de metadatos

Según Godby (2003), el sentido especializado del término *recolectar* aplicado a la idea de biblioteca digital es el de hacerse con metadatos de distintos depósitos (*repository*), para fusionarlos (*fusing*) y revelarlos (*disclosing*) en servicios sindicados (*union services*). Los mecanismos de recolección se están convirtiendo en algo habitual y mejor conocido en el mundo bibliotecario. Por este motivo se vive en la actualidad un clima de expectación respecto a los resultados que se producirán con la fusión de los metadatos recopilados.

Como parte del trabajo realizado en el banco de pruebas D-Lib en la Universidad de Illinois (UIUC, 2003), se han recogido en formato XML cerca de 65.000 artículos periodísticos proporcionados por varias editoriales científicas. Los artículos han sido anotados con metadatos documentales, en formato DCQ (*Qualified Dublin Core semantics*) y RDF. La experiencia en la actualidad se está ampliando a otros proyectos de biblioteca digital.

El aspecto más innovador de esta experiencia es que, pese a que requiere un tratamiento "elemento a elemento" de los metadatos, con costes elevados, el resultado final es una colección de objetos de metadatos muy valiosa. La ventaja es que tanto el formato como la estructura puede manipularse y redefinirse fácilmente (p.ej. por medio de XSLT). El mismo archivo de metadatos de base se puede adaptar al usuario final (transformado en XHTML), o ser indizado directamente con un sistema versado en RDF y DCQ, o compartido con otros sistemas por medio del protocolo de recolección OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*).

Los metadatos para los elementos de los artículos de la D-Lib se generan todos usando hojas de estilo XSL. Para ello se utilizaron algunas herramientas del procesador XML, particularmente para las transformaciones, de forma que fueran conformes con las fuentes de datos SQL. Los metadatos de otras colecciones fueron generados a mano y ordenados primero con Microsoft Access o SQL antes de convertirse a XML.

```
<header>
  <identifier>oai:arXiv.org:cs/0112017</identifier>
  <timestamp>2002-02-28</timestamp>
  <setSpec>cs</setSpec>
  <setSpec>math</setSpec>
</header>
<metadata>
  <oai_dc:dc
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
      http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
    <dc:title>Using Structural Metadata to Localize Experience of Digital
```

```

Content</dc:title>
<dc:creator>Dushay, Naomi</dc:creator>
<dc:subject>Digital Libraries</dc:subject>
<dc:description>With the increasing technical sophistication of both
information consumers and providers, there is increasing demand for
more meaningful experiences of digital information. We present a
framework that separates digital object experience, or rendering,
from digital object storage and manipulation, so the
rendering can be tailored to particular communities of users.
</dc:description>
<dc:description>Comment: 23 pages including 2 appendices,
8 figures</dc:description>
<dc:date>2001-12-14</dc:date>
<dc:type>e-print</dc:type>
<dc:identifier>http://arXiv.org/abs/cs/0112017</dc:identifier>
</oai_dc:dc>
</metadata>
<about>
<provenance
xmlns="http://www.openarchives.org/OAI/2.0/provenance"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/provenance
http://www.openarchives.org/OAI/2.0/provenance.xsd">
<originDescription harvestDate="2002-02-02T14:10:02Z" altered="true">

<baseURL>http://the.oa.org</baseURL>
<identifier>oai:r2.org:klik001</identifier>
<datestamp>2002-01-01</datestamp>
<metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc</metadataNamespace>
</originDescription>
</provenance>
</about>

```

Figura 18. Ejemplo de metadatos en el protocolo de recolección OAI-PMH

## 7 Consideraciones finales

El artículo ha ofrecido un repaso que hemos pretendido fuera amplio y representativo de las principales líneas de actuación en la gestión de contenidos en el marco de la edición digital. Tras la decepcionante tendencia que en el área del procesamiento y maquetación de textos había implantado la industria editorial a finales de la década de los ochenta; por fin, en los albores del siglo XXI hemos podido presenciar un cambio radical de panorama.

XML se ha impuesto como estándar de edición digital. Esto ha permitido que el mundo editorial se beneficie de todos los avances de la filología, la lingüística, la biblioteconomía y la informática. Estas cuatro disciplinas convergen en el interés común de gestionar con eficacia toda la información que la sociedad actual genera cada día. Hemos visto las importantes consecuencias que se derivan de ello; sobre todo a partir de la explosión de contenidos provocada por la *web*.

Hemos hablado de TEI, DCMI, RDF, *web* semántica, OWL, TMX, XLIFF, OCS y OAI. Lo importante para los años venideros no es añadir más siglas a esta ya densa y bien condimentada sopa de letras, sino conseguir que los contenidos fluyan libre y raudamente por las redes de comunicación. Por eso, en los próximos años vamos a asistir a un incremento considerable en la demanda de nuevos protocolos de intercambio para contenidos y metacontenidos, así como una trepidante actividad en torno a la idea de *web* semántica. La sociedad de la información no ha hecho más que comenzar. La fiesta continúa.

## 8 Obras citadas

Joseba Abaitua. 2002. Tratamiento de corpora bilingües, *Tratamiento del lenguaje natural*. M. A. Martí y J. Llisterri. Edicions Universitat de Barcelona, 61-90.

Guillermo Barrutieta, Joseba Abaitua y Josuka Díaz. 2002. Cascading XSL filters for content selection in multilingual document generation. *Second Workshop on NLP and XML*. Taipei, 7-12.

- Tim Berners-Lee. 1998. Semantic Web Road map. World Wide Web Consortium (W3C).  
<http://www.w3.org/DesignIssues/Semantic.html>
- Tim Bray. 1998. RDF and Metadata. <http://www.xml.com/xml/pub/98/06/rdf.html>
- Carlota Bustelo Ruesta, 2003. Gestión documental y gestión de contenidos en las empresas: estado del arte 2002 y perspectivas para 2003. En *El Profesional de la Información*, vol. 12.2, 118-120.  
[http://www.inforarea.es/Documentos/IWE\\_estado\\_arte.pdf](http://www.inforarea.es/Documentos/IWE_estado_arte.pdf)
- Priscilla Caplan. 2001. International Metadata Initiatives: Lessons in Bibliographic Control. *Bicentennial Conference on Bibliographic Control for the New Millennium*.  
[http://lcweb.loc.gov/catdir/bibcontrol/caplan\\_paper.html](http://lcweb.loc.gov/catdir/bibcontrol/caplan_paper.html)
- Robin Cover. 2003. Extensible Markup Language (XML). <http://www.oasis-open.org/cover/xml.html>
- Martin Davis. 2002. *La computadora universal. De Leibniz a Turing*. Debate.
- Josuka Díaz, Joseba Abaitua, Garikoitz Araolaza, Inés Jacob y Fernando Quintana. 2003. El sistema SARE-Bi de catalogación y recuperación de documentos multilingües. *Actas de las II Jornadas de Tratamiento y Recuperación de la Información*. Universidad Carlos III, Madrid.
- Juan Rafael Fernández García. 2003. La traducción en el mundo del software libre.  
<http://es.tldp.org/Articulos/0000otras/doc-traduccion-libre/>
- José Ferrater Mora. 1991. *Diccionario de filosofía*. Círculo de Lectores.
- Jean Godby. 2003. OCLC Metadata Switch, [http://iu.berkeley.edu/rdhyee/discuss/msgReader\\$806](http://iu.berkeley.edu/rdhyee/discuss/msgReader$806)
- Eric A. Havelock. 1963. *Preface to Plato*. Harvard University Press.
- Donald E. Knuth. 1984. *The TeX book*. Addison-Wesley.
- George P. Landow. 1992. *Hypertext: The Convergence of Contemporary Critical Theory and Technology*. Johns Hopkins Press
- William C. Mann y Sandra A. Thompson. 1988. Rhetorical Structure Theory: A theory of text organization. Tech. Rep. RS-87-190. Information Sciences Institute. Los Angeles, CA.
- M. Chantal Pérez Hernández. 2002. Explotación de los corpórea textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de Lingüística Española* vol. 18.  
<http://elies.rediris.es/elies18/>
- Francisco A. Marcos Marín. 1994. *Informática y Humanidades*. Gredos.
- Aurora Martín de Santa Olalla Sánchez. 1999. Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española. *Estudios de Lingüística Española* vol. 3. <http://elies.rediris.es/elies3/>
- Walter J. Ong. 1982. *Orality and Literacy*. Routledge.
- Simon Phipps. 2000. Meaning, not Markup. *XML Journal*, vol. 1.1: 66.
- Andy Powell. 2001. A CGI based Dublin Core metadata generator. UKOLN, University of Bath.  
<http://www.ukoln.ac.uk/metadata/dcdot>
- Simon St. Laurent. 1999. *XML: A Primer*. M & T Books.
- John F. Sowa . 2000. Ontology, metadata, and semiotics. En B. Ganter y G. W. Mineau (comp.) *Conceptual Structures: Logical, Linguistic, and Computational Issues*, Springer-Verlag, Berlin, 55-81.  
<http://users.bestweb.net/~sowa/peirce/ontometa.htm>

Anna Trosborg. 1997. Text typology: register, genre and text types. *Text typology and translation*. 3-23. John Benjamins.

Walter L. Warnick. 2003. Using XML and the Open Archive Initiative to Harvest and Reuse Content. <http://www.osti.gov/speeches/asist.html>

Stuart Weibel. 1995. Metadata: the foundations of resources description. *D-Lib Magazine*. <http://www.dlib.org/dlib/July95/07weibel.html>

Peter Wittenburg, Daan Broeder. 2002. Metadata Overview and the Semantic Web. *International Workshop on Resources and Tools in Field Linguistics*. <http://www.mpi.nl/lrec/papers/lrec-pap-04-MD-overview-daan3.pdf>

## **9 Estándares citados**

Corpus Encoding Standard (CES): <http://www.cs.vassar.edu/CES/>

DARPA Agent Markup Language (DAML): <http://www.daml.org/>

Dublin Core Metadata Initiative (DCMI): <http://dublincore.org/>

Expert Advisory Group on Language Engineering Standards (EAGLES): [www.ilc.pi.cnr.it/EAGLES/home.html](http://www.ilc.pi.cnr.it/EAGLES/home.html)

International Standards for Language Engineering (ISLE): <http://www.mpi.nl/world/ISLE/index.html>

ISLE Meta Data Initiative (IMDI): <http://www.mpi.nl/IMDI/>

Localization Industry Standards Association: <http://www.lisa.org/>

Multilingual Text Tools and Corpora (MULTEXT): <http://www.lpl.univ-aix.fr/projects/multext/>

Open Archives Forum: <http://www.oaforum.org/>

Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH): <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

Open Content Syndication (OCS): <http://internetalchemy.org/ocs/>

Open Language Archives Community (OLAC): <http://www.language-archives.org/>

Resource Description Framework (RDF): <http://www.w3.org/RDF/>

Semantic Web: <http://www.semanticweb.org>

Text Encoding Initiative (TEI): <http://www.tei-c.org>

Translation Memory eXchange (TMX): <http://www.lisa.org/tmx>

UIUC Digital Library Testbed Metadata: <http://dli.grainger.uiuc.edu/publications/metadatasestudy/>

Web Ontology Language (OWL): <http://www.w3.org/TR/owl-ref/>

XML Localisation Interchange File Format (XLIFF): <http://www.opentag.com/xliff.htm>

## **10 Corpora con textos en euskara**

Basque Manuscripts (from late XVIth century). López Martín Collection:  
[http://www.ukans.edu/carrie/ms\\_room/martin\\_coll/Basque.htm](http://www.ukans.edu/carrie/ms_room/martin_coll/Basque.htm)

Euskal testuen gordailua: <http://www.vc.ehu.es/gordailua/>

EuskaraCorpusa.net: <http://www.euskaracorpora.net/>

LEGE-Bi corpus eleaniztuna: <http://www.deli.deusto.es/Resources/LEGE-Bi>

Shakespeare euskaraz: <http://www.susa-literatura.com/emailuak/shakespeare/>

Teatro testuak. Itzulpenak: <http://www.teatroa.com/itzulpen.html>