

XML: un estándar crucial para la "plataforma global" Internet

Autor:

Enrique Bertrand López de Roda.
Director de Tecnología.
Software AG España.

Resumen del trabajo:

En muy pocos meses XML, el nuevo estándar Internet definido por el W3C, se ha convertido en el lenguaje de elección para almacenar, publicar e intercambiar documentos electrónicos a través de la Web. En esta comunicación examinamos las poderosas razones que explican esta acogida entusiasta y el valor que supone XML para el futuro de la administración y los negocios electrónicos sobre "la plataforma global" Internet.

Biografía profesional:

Director de Tecnología en Software AG España, empresa en la que trabaja desde 1986. Es licenciado en Ciencias Químicas por la Universidad de Valencia. Empezó su trayectoria profesional en ERIA y en sus casi 20 años de experiencia ha participado como ingeniero de software, jefe de proyecto y consultor en el diseño y construcción de numerosos sistemas informáticos. En la actualidad su actividad principal se centra la concepción y diseño de soluciones Internet/Intranet, con particular énfasis en las arquitecturas basadas en el estándar XML. Ha sido durante varios años Profesor Asociado del Departamento de Ingeniería Informática de la Universidad Autónoma de Madrid.

XML: un estándar crucial para la "plataforma global" Internet

Introducción

En los últimos meses hemos asistido a la consolidación y expansión de un nuevo estándar Internet: XML, acrónimo de "Extensible Markup Language" o Lenguaje de Etiquetas Extensible. Prácticamente ningún ámbito de los nuevos sistemas de información Internet/Intranet ha quedado al margen de este fenómeno. El propósito del presente documento es examinar que hay detrás de las siglas; los porqués de semejante difusión acelerada, incluso para lo que estamos acostumbrados a ver en el mundo Internet; y sobre todo cuál es el valor de XML en un contexto de administraciones y negocios electrónicos a través de la Red.

Antecedentes

Concebida en sus orígenes como un medio para facilitar el acceso sencillo a documentación remota, principalmente en medios científicos y académicos, la Web ha pasado a ser en muy pocos años un medio de intercambio de información universal, empleado en terrenos que van mucho más allá de sus planteamientos iniciales.

Utilizamos o queremos utilizar la Web y sus protocolos básicos para facilitar las transacciones comerciales entre empresas y clientes (o en sus típicas siglas inglesas, transacciones B2C, "business-to-consumer"), las transacciones documentales y jurídicas entre organismos de la Administración y ciudadanos (A2C, "administration-to-citizen"), pero también para vehicular transacciones entre empresas (B2B, "business-to-business") o entre administraciones (A2A, "administration-to-administration"). Por no hablar de las intranets empresariales como medio para gestionar y difundir el conocimiento a lo largo de la compañía (C2E, "company-to-employees").

En todos los casos nos estamos refiriendo a usos de la Web con demandas y requisitos muy diferentes. Personalización y dinamismo cuando hablamos de intercambios con usuarios finales. Automatización e integración directa con proveedores cuando hablamos de intercambios entre empresas u organismos públicos. Estructuración adecuada y facilidades de búsqueda en el caso de las intranets (ver Fig. 1).

Una situación que obligó al W3C a plantearse el futuro de su estándar más conocido, el HTML. El "HyperText Markup

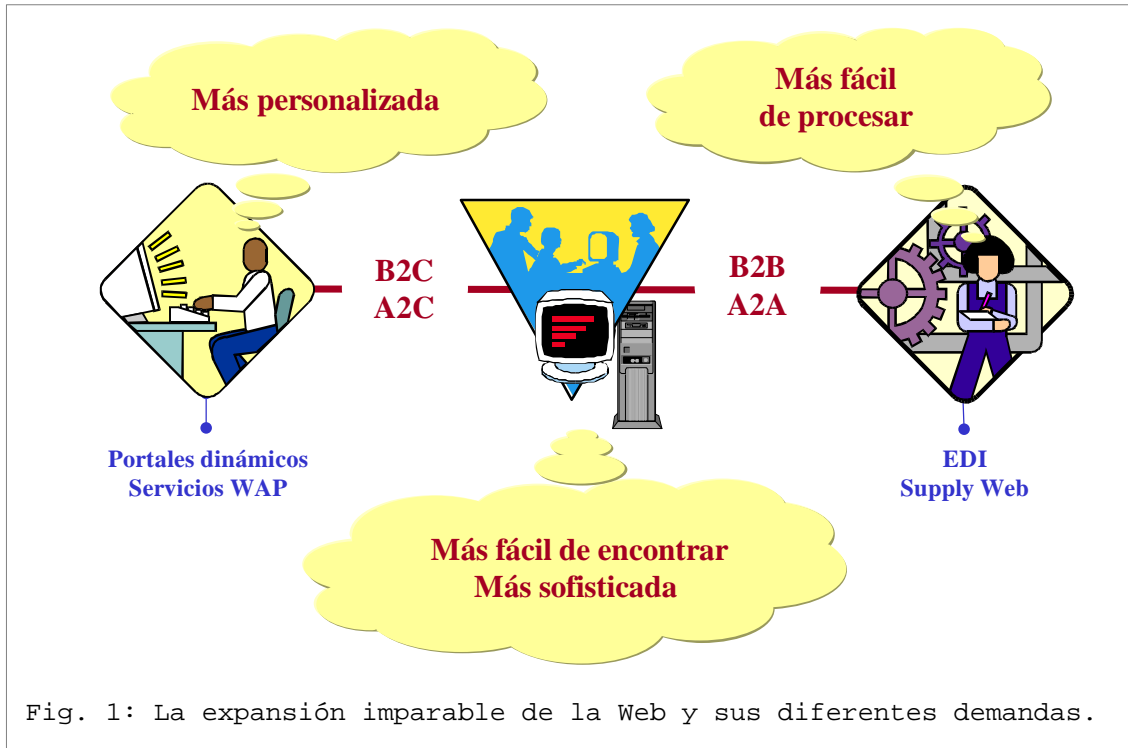


Fig. 1: La expansión imparable de la Web y sus diferentes demandas.

Language" es el lenguaje en el que están publicados la mayor parte de documentos que circulan por la Red. Se trata en lo esencial de un estándar pura y simplemente de visualización, que indica a la herramienta básica de la Web, el navegador o "browser", la forma de mostrar los textos, imágenes, iconos, etc. que contiene un documento. Un lenguaje muy sencillo, cuya facilidad de aprendizaje ha sido, sin duda, una de las razones principales del éxito de la Web. Pero a la vez un lenguaje que no fue concebido para intercambiar información transaccional ni para facilitar búsquedas sofisticadas y contextuales. Un lenguaje que no puede satisfacer las demandas planteadas por los nuevos usos de la Red.

Con objeto de resolver esta situación, surgida como consecuencia de su propio éxito, el World Wide Web Consortium o W3C (el organismo que controla los estándares de la Web) constituyó un grupo de trabajo en 1996 que concluyó sus actividades en 1998 con la publicación de un nuevo estándar: el XML.

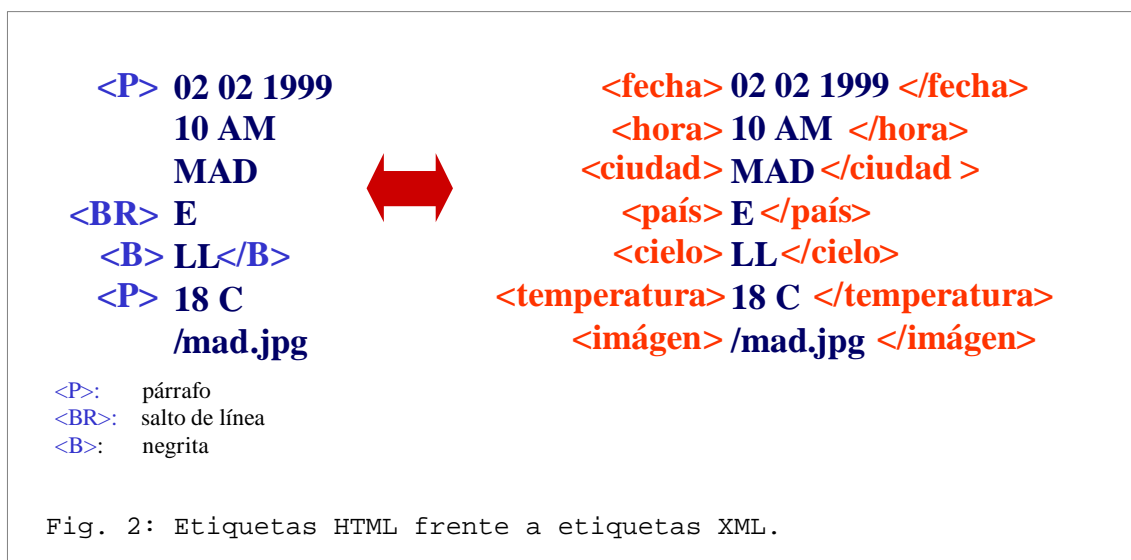
XML: Extensible Markup Language

El desafío del grupo de trabajo radicaba en la necesidad de definir un nuevo lenguaje que fuese mucho más allá que HTML en términos de flexibilidad, sin perder su simplicidad intrínseca. Recuerdese que lo simple nunca ha sido más relevante que en Internet: los estándares que no encuentran el

punto de equilibrio adecuado entre potencia/flexibilidad y simplicidad desaparecen sin más.

Tomaron la decisión de mantener el concepto base de HTML, la etiqueta o "markup", pero cambiando su naturaleza. En lugar de etiquetar aspectos visuales (salto de línea, párrafo o negrita) se trata ahora de etiquetar los contenidos y estructura de un documento o un conjunto de datos. Nace entonces el XML, el "Extensible Markup Language".

En este nuevo lenguaje el concepto de etiqueta se generaliza o "extiende" respecto al utilizado en HTML. De un conjunto restringido de etiquetas (<P>,
, <A>, etc.) en HTML, pasamos en XML a un conjunto de etiquetas definibles por el usuario y específicas de cada tipo de documento. Así podemos crear un juego de etiquetas para una factura, un historial clínico, un curriculum vitae o un parte meteorológico (ver Fig. 2). Tenemos en realidad un lenguaje con el que podemos construir vocabularios específicos (lenguajes) para describir datos (es decir, un metalenguaje).



Tampoco en su trabajo partieron de cero. Su propuesta inicial se apoyó en una versión drásticamente simplificada y adaptada a Internet de un lenguaje de etiquetas ya existente, muy utilizado en el mundo editorial: el SGML o "Standard Generalized Markup Language". Hablamos, pues, de un estándar nuevo y viejo. Nuevo porque está concebido para un mundo dominado por las redes de información y viejo porque asume el legado y la experiencia del más sofisticado lenguaje de etiquetas que haya existido, el SGML.

Consideremos un ejemplo de documento definido en formato XML, que incluye los datos básicos de una compañía, tal y como queremos que figuren en un hipotético registro de empresas:

```

<?xml version='1.0' encoding='ISO-8859-1' ?>
<EMPRESA CODIGO="A78093192">
  <NOMBRE> Software AG España, S.A. </NOMBRE>
  <VENTAS UNIDAD="M PTA"> 12100 </VENTAS>
  <DIRECCION>
    <CALLE> Rda. Luna, 4 </CALLE>
    <CIUDAD> Tres Cantos </CIUDAD>
    <PROVINCIA> Madrid</PROVINCIA>
  </DIRECCION>
  <PRODUCTO EXCLUSIVIDAD="Si">
    <MARCA> Bolero </MARCA>
    <CATEGORIA> Entorno de desarrollo </CATEGORIA>
  </PRODUCTO>
  <PRODUCTO EXCLUSIVIDAD="Si">
    <MARCA> Tamino </MARCA>
    <CATEGORIA> Gestor de base de datos XML </CATEGORIA>
  </PRODUCTO>
</EMPRESA>

```

En la primera línea encontramos un encabezamiento normalizado que identifica este documento como del tipo XML y que indica la versión del estándar utilizado y el juego de caracteres con el que está codificado el documento. A partir de este punto se suceden una serie de etiquetas que contienen la información de la empresa.

Las etiquetas deben satisfacer una serie de reglas relativamente simples. Entre las más importantes:

- Las etiquetas siempre aparecen por pares <xxx> </xxx>, delimitando un contenido determinado

```

<CATEGORIA> Entorno de desarrollo </CATEGORIA>

```

- Una etiqueta pueda descomponerse en otras siempre que las incluya completamente (descomposición jerárquica estricta) y puede repetirse tantas veces como queramos

```

<DIRECCION>
  <CALLE> Rda. Luna, 4 </CALLE>
  <CIUDAD> Tres Cantos </CIUDAD>
  <PROVINCIA> Madrid</PROVINCIA>
</DIRECCION>

```

- Una etiqueta puede incorporar atributos que aclaran ciertos aspectos de la naturaleza del contenido

```

<VENTAS UNIDAD="M PTA"> 12100 </VENTAS>

```

Como vemos, desde el punto de vista sintáctico es muy similar a HTML, pero con reglas de estructuración más precisas.

El punto clave radica en el juego de etiquetas utilizado: ¿quién y por qué ha decidido utilizar <VENTAS> en lugar de <FACTURACION>? ¿quién y por qué ha decidido descomponer <DIRECCION> en <CALLE>, <CIUDAD> y <PROVINCIA>?

Justamente éste es el concepto recogido en el término "Extensible". En XML es posible definir la estructura de etiquetas de un documento de forma enteramente libre, siempre que se respeten las reglas sintácticas. Esta situación es perfectamente asumible cuando manejamos documentos de consumo interno que no deben ser intercambiados con otras instituciones o personas. En caso contrario, sin embargo, deberíamos ponernos de acuerdo con nuestro interlocutor respecto al vocabulario utilizado o, al menos, comunicarle formalmente cual hemos empleado.

Este es el papel de los DTD o "Document Type Definition", definir y formalizar las etiquetas que se pueden utilizar en un documento. Es decir determinar las relaciones de dependencia entre ellas, el carácter obligatorio u opcional de las mismas, la repetibilidad, la posibilidad de añadir atributos, etc.

En XML tenemos, en consecuencia, la libertad como diseñadores para movernos entre dos extremos:

- Documentos que utilizan un vocabulario enteramente libre y sintácticamente correcto (documentos bien formados).
- Documentos que utilizan un vocabulario formalizado que ha sido previamente acordado y/o publicado en su correspondiente DTD. En este caso se puede verificar la validez estructural del documento respecto a su DTD de referencia (documentos validos)

Evidentemente, en el mundo de las transacciones electrónicas a través de la Web sólo tiene sentido esta última opción. Por eso una de las áreas de más actividad en el mundo XML es, como luego comentaremos, la definición de DTDs a escala corporativa, estatal o internacional.

Otro aspecto crítico de XML tiene que ver, paradójicamente, con el origen del problema: la visualización del documento. Si observamos nuestro ejemplo, no hay nada en el mismo que nos indique como se mostrará esta información en un navegador. ¡Precisamente para alcanzar este objetivo se desarrollo el estándar! Se trata de un aspecto irrelevante, por ejemplo, en una interacción entre máquinas en un contexto B2B o A2A. Pero en una interacción con consumidores, en la que se emplea el navegador, no podemos decir lo mismo.

En el caso de XML la solución descansa en un sub-estándar conocido como XSLT, "Extensible Stylesheet Language Transformation". XSLT se apoya en el concepto, clásico en el mundo de los procesadores de texto, de hoja de estilo. Mediante hojas de estilo definidas con el lenguaje XSLT podemos transformar automáticamente nuestro documento XML en muchos otros formatos de salida como, lógicamente, HTML,



Fig. 3: Independencia entre contenido y presentación vía XSLT.

pero también texto plano o páginas WML ("Wireless Markup Language") que nos permiten ver los datos utilizando un teléfono móvil que soporte el protocolo WAP (ver Fig. 3).

Alcanzamos de esta forma uno de los propósitos de origen de XML: independizar de forma sencilla el contenido de un documento de su presentación. Asunto no tan fácil de resolver con el lenguaje HTML.

Después de lo dicho, las razones del éxito extremadamente rápido de XML son fáciles de entender y resumir:

1. XML define un estándar flexible, abierto y universal para describir todo tipo de documentos, estructuras de datos o mensajes.
2. XML es un lenguaje fácil de comprender y utilizar. Tanto las personas como las máquinas pueden interpretar un documento XML sin excesivo esfuerzo (véanse los ejemplos que figuran en este documento). Pero al mismo tiempo, es un lenguaje poderoso que permite representar estructuras jerárquicas de elevada complejidad y controlar vía DTD su validez formal.
3. XML es un estándar Internet desde muchos puntos de vista. Jurídicamente, está controlado por el organismo responsable de la expansión de la Web, el W3C. Es un estándar que aparece con la consolidación del fenómeno Internet y que asume otros estándares perfectamente establecidos (URL/URI, Unicode, HTTP, etc.).

No es extraño entonces que la mayor parte de empresas, organismos, instituciones internacionales, compañías de software apoyen e impulsen decididamente el uso de XML como el lenguaje de elección para almacenar, publicar e intercambiar documentos electrónicos a través de la Red.

Papel de XML en las nuevas aplicaciones Internet/Intranet

XML no es, desde luego, la panacea universal ni la solución mágica ("silver bullet" para los americanos) que resolverá todos nuestros problemas de desarrollo de software. Pero sí que simplifica notablemente la construcción de aplicaciones Internet/Intranet desde todas las perspectivas: usuarios finales (B2C, A2C y C2E), empresas y organismos públicos (B2B y A2A). Veamos algunos ejemplos significativos de las posibilidades que ofrece XML en estos diferentes ámbitos.

En el terreno del B2C y del A2C ("business-to-consumer" y "administration-to-citizen"), XML permite manejar de forma razonable los problemas de personalización. A partir de un contenido único definido en XML es posible generar múltiples niveles de presentación que tienen en cuenta las preferencias de los usuarios o el tipo de dispositivo desde el que están accediendo: navegador estándar, navegar limitado, teléfono móvil, etc.

La solución se basa en la aplicación dinámica de las hojas de estilo XSLT, considerando la información disponible del

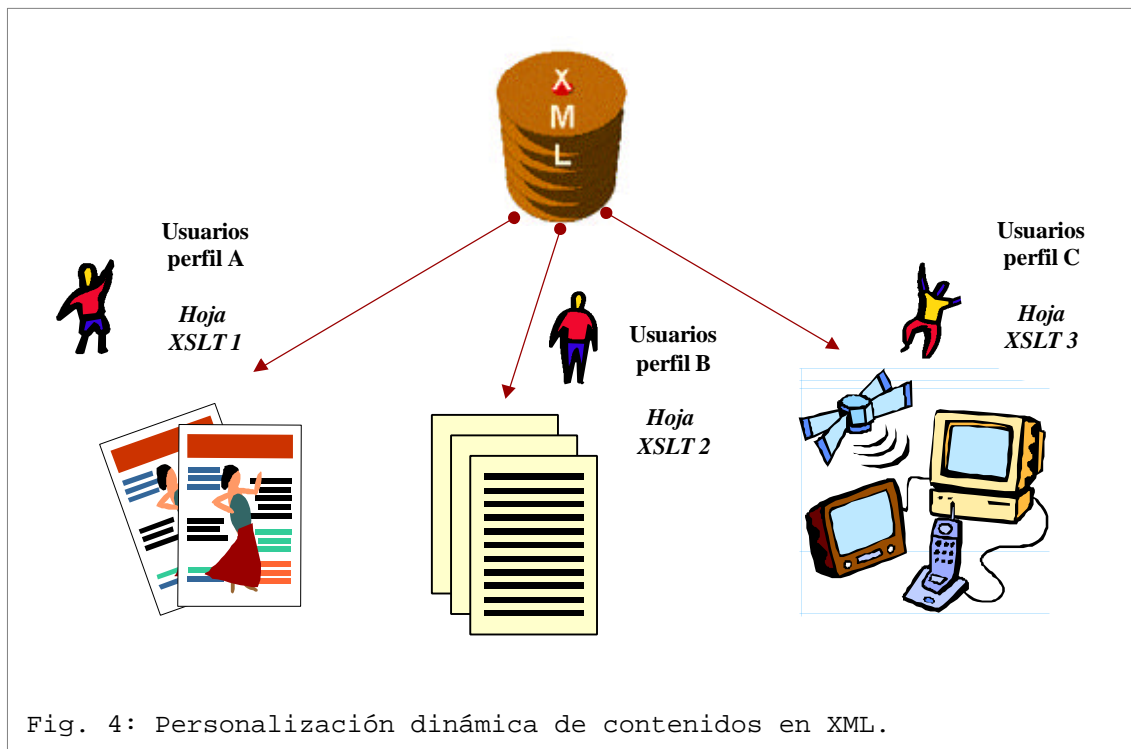


Fig. 4: Personalización dinámica de contenidos en XML.

usuario (si éste se identifica) y los datos respecto al medio de acceso que figuran en las cabeceras de una solicitud HTTP. En función de estos parámetros, se transforma el contenido con distintas hojas XSLT, dando como resultado documentos de salida con diferentes niveles de detalle, niveles gráficos, etc. (ver Fig. 4).

En aplicaciones intranet y C2E ("company-to-employees") el problema más severo es cómo acceder a información que se encuentra repartida y soportada en numerosos productos, cada uno con sus propios formatos: herramientas ofimáticas, herramientas de trabajo en grupo, servidores de aplicaciones, paquetes ERP, etc. XML promete una forma más racional de acceder a toda esta masa de datos en la medida en que la mayor parte de fabricantes de software han anunciado (o ya han implementado) interfaces y APIs XML para sus productos. Estas interfaces permiten extraer la información de estos sistemas, no en el formato propietario de la misma, sino en un formato XML neutro y perfectamente definido.

Pero quizás el terreno donde se vislumbra un crecimiento exponencial en la utilización del lenguaje XML es en los sistemas B2B y A2A ("business-to-business" y "administration-to-administration"), sobre todo para aplicaciones tipo EDI ("electronic data interchange").

El EDI tradicional, apoyado en estándares como EDIFACT o X12 y el uso de redes de valor añadido (VAN), no se ha extendido más allá de las grandes empresas comerciales e instituciones públicas debido a la relativa inflexibilidad de los formatos y a los costes económicos de su explotación. De esta forma, una enorme cantidad de pequeñas empresas y organismos siguen procesando los documentos que reciben de otras empresas de forma enteramente manual.

Una situación difícil de mantener en una economía tan dinámica y global como la que se otea en el horizonte del siglo XXI y en la que de nuevo XML tiene algo que aportar. Podemos plantear una nueva arquitectura para sistemas EDI en la que el formato de los mensajes se define en XML y se utiliza Internet y el protocolo HTTP como mecanismo de intercambio.

Ganamos con este cambio en los dos terrenos antes mencionados:

- Mayor flexibilidad y extensibilidad de los formatos XML frente a los formatos EDIFACT. En estos últimos los contenidos vienen delimitados posicionalmente, mientras que en XML vienen delimitados por las etiquetas. Podemos cambiar la posición de un dato o añadir nuevos datos sin necesidad de modificar las aplicaciones existentes.

- Menores costes de implementación y explotación al utilizar la infraestructura de Internet y software de tipo genérico más asequible que las típicas soluciones EDI.

Evidentemente, uno de los requisitos clave para la viabilidad de las aplicaciones XML/EDI es que los actores en el intercambio se pongan de acuerdo en el vocabulario (DTD) a utilizar (ver Fig. 5). En estos momentos numerosas iniciativas tanto horizontales como verticales están impulsando la consecución de este tipo de acuerdos.

Entre las iniciativas horizontales más destacadas está ebXML, liderada por Oasis y el grupo de estandarización de EDIFACT (UN/CEFACT), para facilitar la transición desde el EDI tradicional al nuevo EDI basado en XML. En esta visión el EDI tal y como lo conocemos no desaparece (lo que no sería realista dadas las inversiones actuales) sino que se adapta y coexiste "pacíficamente" con el nuevo EDI. Los mensajes EDI se reformulan en XML y se pueden seguir utilizando las actuales VAN o la red Internet.

Las iniciativas verticales por su propia naturaleza son mucho más dispares, pero prácticamente no hay ningún segmento de la economía y la sociedad (industrial, financiero, sanitario, turístico, educativo, gubernamental, etc.) al margen del proceso de definición de DTDs específicas para sus documentos principales.

Como ejemplo, en el caso de la Administración Pública espa-

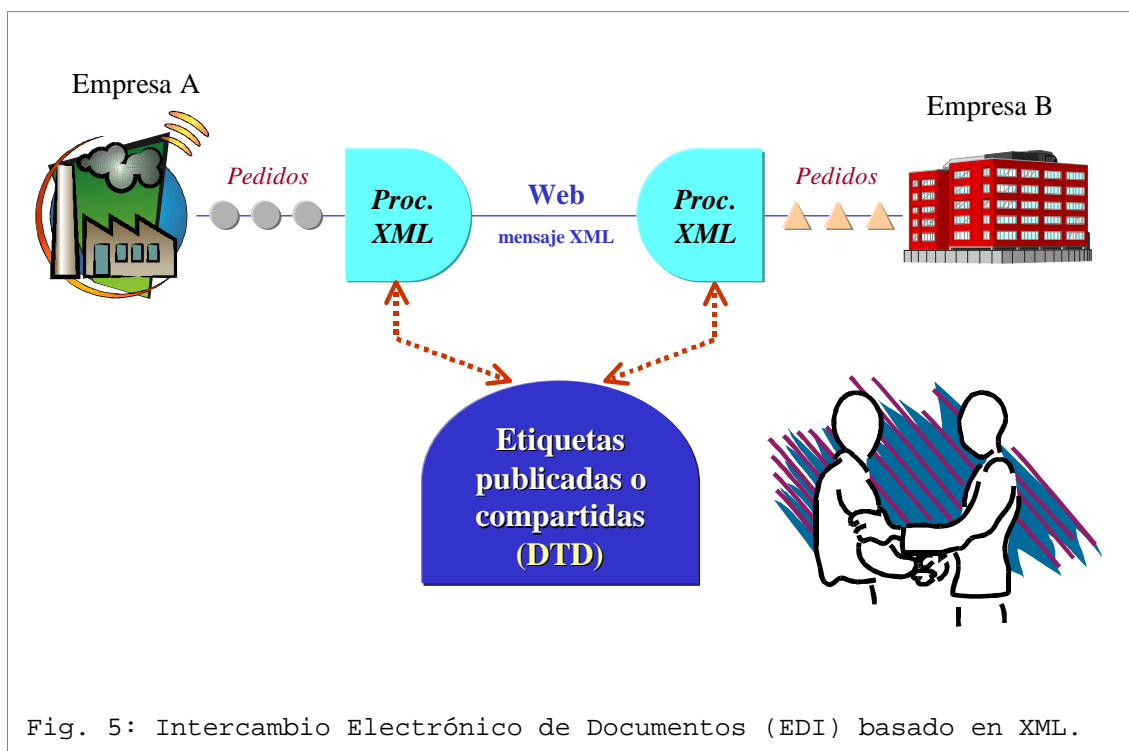


Fig. 5: Intercambio Electrónico de Documentos (EDI) basado en XML.

ñola existe ya una definición XML del formato del mensaje SICRES 2.0 dentro del proyecto "Pista. Ventanilla Unica" de los Ministerios de Fomento y de Administraciones Públicas.

En todos estos casos nos hemos referido al impacto que supone XML desde el punto de vista de las aplicaciones finales, pero también podemos hablar de una influencia similar en el terreno del software de base.

XML está alumbrando una nueva generación de herramientas de desarrollo, sistemas de gestión de bases de datos y productos "middleware" con un doble propósito:

- Aprovechar al máximo las posibilidades del estándar. Pensemos en la ventajas de un sistema de almacenamiento de contenidos y documentos nativos XML, en términos de rendimiento, respecto a soluciones "ad hoc" basadas en transformaciones internas a tablas SQL (imprescindibles dada la naturaleza jerárquica de XML). O en las ventajas, en términos de flexibilidad, de un "middleware" en el que formato interno de los mensajes fuese XML.
- Facilitar la integración con los actuales sistemas "no-XML". Existe una enorme cantidad de información en formato SQL o en formatos propietarios que por razones técnicas y económicas no puede ser ignorada ni reconstruida. Vía adaptadores ("wrappers") automáticos podemos ver estos datos como si fuesen estructuras XML.

Conclusiones

XML emerge como la opción más razonable para manejar la diversidad de datos, escenarios, dispositivos de acceso, etc. que demandan las nuevas aplicaciones de negocio y administración electrónicos.

Su capacidad para modelizar cualquier tipo de información; su naturaleza comprensible tanto para máquinas como para humanos; su independencia respecto a lenguajes de programación o plataformas software; y sus extensiones específicas para diferentes sectores económicos lo convierten en el único estándar realmente abierto y viable para el intercambio de información a través de Internet.