ximdex

Linked Open Data

Guía para la publicación de datos abiertos y enlazados

14.02.2014
Versión 04

Índice

1. Introducción

El presente documento desarrollado por Open Ximdex Evolution SL (OXE) proporciona una breve guía para la publicación de información en formato *Linked Open Data* (datos abiertos enlazados). Esta guía la usamos como punto de partida para analizar las posibilidades de una corporación para liberar información.

Responderemos principalmente a las preguntas de cómo y en qué formatos liberar información de interés generada por una **corporación ficticia** (ACME Corp.[1]), minimizando los costes de recolección y publicación y garantizando la plena reutilización de la información.

Es importante mencionar que con "datos" no sólo nos referimos a valores numéricos, sino que cualquier tipo de "documento" es susceptible de ser publicado como un "dato abierto".

1.1. Linked Open Data / Datos Abiertos Enlazados

El objetivo de *Linked Open Data* es mejorar la información publicada para eliminar barreras en su consumo por los usuarios finales. Dependiendo del grado de compromiso (estrellas) será más sencillo automatizar tareas de utilización de la información publicada, eliminándose intermediarios que aportan poco valor a la información.

Tim Berners-Lee, el padre de lo que hoy conocemos como protocolo HTTP y la Web, propone cinco niveles, ordenados de menos a más abiertos, para hacer una apertura progresiva de los datos.

Así, los cinco niveles hasta llegar a *Linked Open Data* son:

Nivel	Descripción
*	Publica tus datos en la Web en cualquier formato pero con una licencia libre (para que puedan denominarse *open data*)
**	Publica dichos datos en formatos estructurados legibles por máquinas (ej.: un documento MS excel en lugar de una imagen escaneada)
***	Utiliza formatos no propietarios (ej.: Tablas de Texto en lugar de MS excel)
****	Utiliza estándares del consorcio web (W3C) para la web semántica (ej.: RDF y SPARQL) para identificar las "cosas" para que puedan así apuntar a nuestra información desde el exterior (*Linked Open Data*).
*****	Enlaza tus datos con otros datos (de terceros) para proveer un contexto.

[1] http://es.wikipedia.org/wiki/Corporación_Acme

1.2. Necesidad de publicar Datos Abiertos

La publicación en formato “datos abiertos” es un valor en alza entre las instituciones públicas para facilitar el acceso a información pública completa, fiable y de calidad.

Liberar datos (o documentos) abre un abanico de posibilidades al usuario final para usarlos, reutilizarlos, redistribuirlos o integrarlos en aplicaciones propias, o de terceros, al trabajar directamente con la fuente de la información (“datos en bruto”, “en crudo” o “raw”) en lugar de con “productos” (ej.: tablas, gráficos, etc.) derivados de la misma.

1.3. Marco legal y normativa en España

La iniciativa de apertura de datos surge desde el ámbito europeo (Directiva europea 2003/98/CE[2]) para, entre otro objetivos, concienciar a los empleados públicos del valor que posee la información del sector público con la que trabajan. Basada en tres pilares fundamentales: la *transparencia*, *participación* y *colaboración*.

A nivel nacional se traduce en la **Ley del 37/2007**[3] de 16 de noviembre (RISP, Reutilización de la Información del Sector Público) y el **Real Decreto 1495/2011**[4], de 24 de octubre, que la desarrolla. La **Norma Técnica** se recoge en la resolución número 2380 de fecha 19 de Febrero de 2013.

[2] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:ES:HTML
[3] http://www.boe.es/buscar/doc.php?id=BOE-A-2007-19814
[4] http://www.boe.es/diario_boe/txt.php?id=BOE-A-2011-17560

2. ¿CÓMO LIBERAR?

Las principales requisitos que han de cumplir los datos a liberar son:

- Deben ser procesables automáticamente por máquinas (*machine readable*)
- Deben estar publicados en formatos abiertos que, a poder ser, no dependan de empresas concretas para favorecer así su reutilización (así, por ejemplo, es preferible publicar en formato CSV –*comma separated value*–, en lugar de formato Excel de Microsoft)
- Ha de facilitarse el acceso masivo a los mismos (normalmente "en bruto")
- Deben tener una "granularidad" suficiente para que resulten útiles al usuario.
- Deben ser fáciles de encontrar en un único punto de acceso en línea, y ser puntualmente publicados (compromiso con los usuarios potenciales) y en estándares abiertos.
- Deben poder ser descargables desde diferentes tipos de terminales y navegadores.

Con carácter general, todos estos puntos, que no son obligatorios pero sí recomendables, son los que debería reunir un conjunto de datos que desea ser liberado.

Además es recomendable establecer una licencia de uso de los datos publicados que permitan su reutilización gratuita para cualquier uso.

La publicación de una colección o conjunto de datos (*dataset*) en un formato específico (csv, pdf, etc.) constituye una "distribución". Es importante también proporcionar información de contexto para los datos publicados: geolocalización, recursos asociados, categoría, creador, idioma, frecuencia de actualización y plazos de validez de la información publicada, formatos, etc.

2.1. TIPOS DE LICENCIAS

La licencia que acompañe a los conjuntos de datos publicados deben ser claras, justas y transparentes, sin restringir la reutilización de los mismos o limitando la competencia.

También es conveniente indicar la duración de la licencia[5] y las distintas responsabilidades que se establecen en el uso de los mismos.

[5] http://opendatacommons.org

Las licencias de datos libres otorgan, básicamente, las siguientes libertades:

1. plena **libertad para compartir** (copiar, distribuir) **y utilizar los datos**,
2. plena **libertad para crear obras derivadas** de esos datos, y
3. plena **libertad para adaptarlos** (modificarlos y transformarlos)

La diferencia entre licencias proviene de las "obligaciones" que imponen al usuario de los datos, distinguiéndose dos tipos principales:

1. "*Dominio Público*", sin exigir ninguna contraprestación a cambio.
2. "***Open Data License***", exigiendo al usuario de los datos que:
    - se realice una **correcta atribución del origen de los datos**,
    - se **ofrezcan los datos derivados con el mismo tipo de licencia**,
    - **se ofrezca una versión "libre"** paralela a cualquier versión restringida.

De forma general, desde OXE sugerimos el uso de una Licencia tipo "Open Data" que garantice que los datos derivados tengan el mismo tipo de licencia para así poder beneficiarse de la nueva información creada por terceros y poder "medir" el uso que se hace de la información publicada.

En el caso de ser una corporación de carácter público habrá que regirse por la normativa RISP resumida a continuación.

2.2. NORMATIVA RISP

En el caso de organizaciones de carácter público se recomienda que los datos estén enlazados desde la "sede electrónica" de la corporación. El portal http://**datos.gob.es** agrega fichas-tipo para "publicitar" los datos publicados. El Real Decreto 1495/2011 de 24 de Octubre, que desarrolla la Ley 37/2007 sobre Reutilización de la Información del Sector Público (RISP), propone las siguientes observaciones:

- Se pondrán a disposición del público los documentos reutilizables por medios electrónicos, de una manera estructurada y usable, preferentemente en bruto, en formatos procesables y accesibles de modo automatizado.
- Se procurará que la información puesta a disposición se actualice en un tiempo razonable que permita el uso adecuado de dicha información, con frecuencia análoga con la que actualicen dicha información internamente.
- Las corporaciones informarán, preferentemente a través de «sede.gob.es/datosabiertos», sobre qué documentación es susceptible de ser

reutilizada, los formatos en que se encuentra disponible, las condiciones aplicables, la fecha de la última actualización, proporcionando, cuando esté disponible, información complementaria para su comprensión y procesamiento automatizado y facilitando al máximo la identificación, búsqueda y recuperación de los documentos disponibles mediante listados, bases de datos o índices.

- El acceso a documentos que contengan datos de carácter personal o referentes a la intimidad de las personas estará reservado a éstas, que podrán además ejercer sus derechos de rectificación, cancelación y oposición de acuerdo con lo previsto en la legislación de protección de datos personales y el artículo 37.2 de la Ley 30/1992, de 26 de noviembre. No obstante, siempre y cuando los medios técnicos y económicos lo permitan, **deberá procederse a la disociación de los datos personales**[6] a fin de permitir su reutilización por otras personas.

- Los órganos de la Administración General del Estado y los demás organismos y entidades del sector público estatal a que se hace referencia en el artículo 1.2 deberán adaptarse a las disposiciones de este real decreto en el plazo de un año desde su entrada en vigor. En el citado plazo de un año, aprobarán un plan propio de medidas de impulso de la reutilización de la información del sector público por medios electrónicos, que incluirá el compromiso de publicar, de una manera estructurada y usable y en bruto, en formatos procesables y accesibles de modo automatizado correspondientes a estándares abiertos, al menos cuatro conjuntos de documentos de alto impacto y valor en un plazo máximo de seis meses desde la finalización del plazo de adaptación previsto.

2.2.1. Norma Técnica RISP

La resolución número 2380 de fecha 19 de Febrero de 2013 aprueba la Norma Técnica[7] de Interoperabilidad de Reutilización de recursos de la información que establece condiciones comunes sobre los recursos de información elaborados o custodiados por el sector público para facilitar y garantizar el proceso de reutilización de la información (vocabularios a utilizar), asegurando su persistencia (indicando formato de la URI), el uso de formatos (estándares abiertos y que ofrezcan una representación semántica) y los términos y condiciones de uso adecuados.

2.2.2. Aviso Legal en Datos Bajo RISP

El aviso legal disponible bajo la dirección http://www.datos.gob.es/avisolegal enumera indicaciones sobre la responsabilidad y sobre las obligaciones contraídas por el uso y publicación de los datos (citar la fuente, fecha de actualización de la información, etc.)

[6] En los términos que se derivan de lo establecido en el artículo 3.f) de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal y en el artículo 5.1.e) del Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba su Reglamento de Desarrollo.
[7] http://datos.gob.es/datos/sites/default/files/NTI_Reutilizacion_Informacion_BOE-A-2013-2380.pdf

3. ¿QUÉ LIBERAR?

Analizando los datos que ofrece públicamente la ACME Corp. a sus usuarios en la actualidad, podemos clasificarlos en dos categorías: “informes” y “datos estadísticos”.

Para asignar prioridades para determinar los datos a liberar es aconsejable conocer qué información manejan los infomediarios (empresas que se dedican a procesar la información pública para revenderla en formatos diferentes) que actualmente procesan información de ACME Corp. y qué conjunto de datos consideran de mayor relevancia.

En todo caso, una máxima que siempre debemos cumplir es: garantizar la neutralidad del organismo liberador de datos públicos frente a los posibles usuarios de dichos datos.

3.1. ANÁLISIS DEL ESTADO ACTUAL

ACME Corp. publica actualmente Informes en formato PDF proporcionando relaciones entre algunos tipos de documentos. Toda la información de carácter personal es eliminada de forma previa a su publicación, no guardándose información interna sobre “trazabilidad” que permitiera mejorar las relaciones existentes.

ACME Corp. publica además información de tipo estadístico (gráficas, tablas) en un “Anuario”.

Se resumen a continuación los niveles (estrellas) indicando el grado de cumplimiento de ACME Corp. para cada uno de ellos:

1. *Publica los datos en web, sin importar el formato, con licencia libre (open data):*
    - ACME Corp. publica documentos (ej.: informes) en su web pero **NO** proporciona una licencia que indique como reutilizar la información publicada siendo relativamente sencillo asociar una licencia libre a los datos.
2. *Publica dichos datos en un formato estructurado (legible por máquinas):*
    - ACME Corp. publica **PARCIALMENTE** los documentos en formatos estructurados (informes en formato PDF)
3. *Procura que dicho formato estructurado NO sea propietario:*
    - ACME Corp. **NO** publica sus documentos en un formato no propietario, al utilizar PDF para su distribución. Sin embargo, a nivel interno se dispone de la información en formatos de tipo textual y no propietario por lo que es relativamente sencillo cumplir esta condición.

4. *Utiliza el estándar RDFa, para que se puedan emplear URLs para mapear los datos, convirtiéndolas en direccionables desde el exterior:*

    - ACME Corp. **NO** publica actualmente su información en formato RDFa, aunque al disponer de relaciones entre los datos es relativamente sencillo publicarlas en web en formato RDFa.

5. *Enlaza tus datos con otros datos para crear contextos (Linked Open Data):*

    - ACME Corp. no enlaza de forma sistemática y en formatos semánticos su información con la de otros organismos (para, por ejemplo, normalizar las poblaciones, organismos, etc.). Para cumplir este último nivel es necesario identificar "entidades" relevantes para utilizar las "anotaciones" ya existentes en la información utilizada a nivel interno.

A día de hoy podemos afirmar que con muy poco esfuerzo **ACME Corp.** alcanzaría 2 estrellas, siendo relativamente sencillo en tiempo y forma alcanzar la 3ª y 4ª estrella, al estar ya publicándose en la web las relaciones establecidas entre datos (relaciones entre informes) para ciertos tipos de documentos específicos.

Para alcanzar el nivel 4 se publicaría un listado, sin formato gráfico, para consumo por sistemas informáticos, utilizando enlaces en formato RDFa a los documentos relacionados.

Se revisan a continuación de forma somera los dos tipos de "datos" que podrían pasar a *Open Data* de forma casi inmediata:

3.2. INFORMES

Actualmente ACME Corp. publica en su portal web una ingente cantidad de informes y documentos relacionados en formato PDF.

Para alcanzar un nivel 3 habría que publicarlas en formatos libres (tipo textual, por ejemplo), asociando una licencia libre. Además de permitir su adscripción como "open data" se favorecería la búsqueda de información mediante la definición automática de listados o índices clasificados por tipos y fechas.

Para que los datos sean realmente reutilizables e interoperables hay que alcanzar el nivel 4. En el caso de la información publicable por ACME Corp., esto es aún más relevante por la cantidad de información relacionada con cada documento publicado susceptible de ser vinculada de forma sencilla.

Por último, para alcanzar el nivel 5 sería necesario utilizar las definiciones de "entidades" relevantes para la corporación (organismos, poblaciones) que ya se manejen a nivel interno en los flujos de trabajo, y relacionarlas con la información

publicada por otros organismos (ej: listado de poblaciones en formatos semánticos desde dbpedia.org)

3.3. Datos Estadísticos

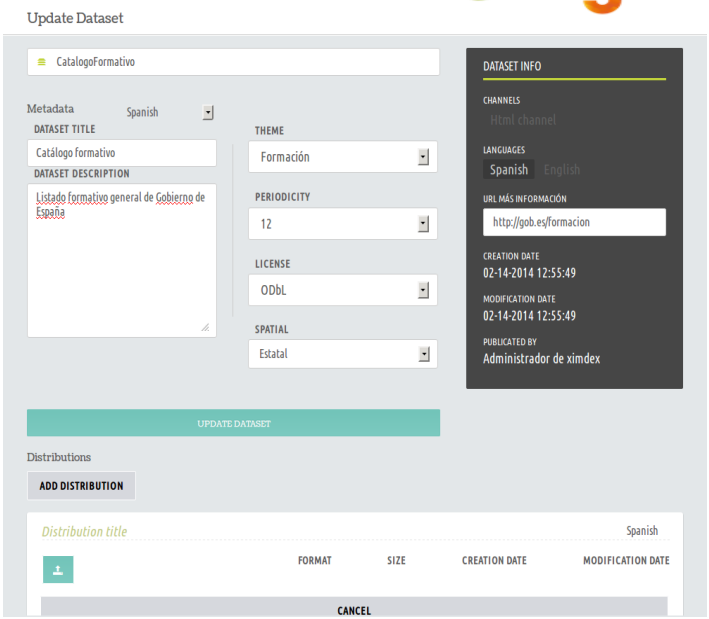
Las memorias anuales que publica ACME Corp. contienen datos estadísticos susceptibles de ser liberados. Se sugiere que los datos que originan las gráficas y tablas sean publicados “en bruto” (en texto plano, csv, …), dejando que el usuario final explote directamente la información, adquiriendo ésta un mayor valor al poder representarse y organizarse de la forma más adecuada para el “consumidor”.

4. MÓDULO XLYRE: DATOS ABIERTOS ENLAZADOS

El CMS semántico y libre XIMDEX ofrece el módulo *XLYRE* para la gestión visual de Datos Abiertos Enlazados (*Linked Open Data*) que facilita la creación de colecciones de datos publicables que, de forma automática, incluyen toda la información contextualizada disponible.

La gestión de la información se realiza de forma completamente visual a partir de la "agregación" de fuentes de datos.

Los datos pueden provenir bien del propio sistema XIMDEX, bien de otros repositorios bajo diferentes tecnologías: ximdex, apache solr, openlink virtuoso, CMIS, NoSQL, …



De forma resumida, el módulo XLYRE proporciona:

- Una gestión visual de colecciones de datos (*datasets*), su metainformación y sus relaciones,
- La transformación automática de datos para formar nuevos conjuntos de datos mediante filtros (utilizando SPARQL, RDFa, búsquedas sobre documentos no estructurados, ...),
- La transformación automática de datos para su publicación en distintos formatos,
- La generación automática de fichas-tipo para la publicación de una distribución (conjunto de datos en un formato específico) así como la generación de índices ordenados por diversos criterios,
- La generación automática de Portales Web con todos los datos y relaciones en distintos formatos e idiomas, así como su categorización y sindicación hacia portales de agregación de información.

La Figura siguiente recoge una ficha tipo para una colección de datos que ha sido publicada en un portal web generado automáticamente por Ximdex CMS atendiendo a las plantillas de diseño deseadas:

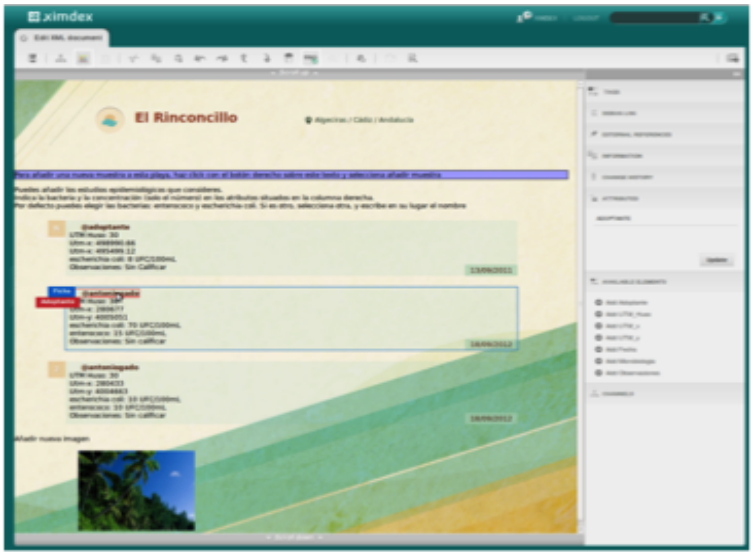


Figura 1: Portal Web de datos sintetizado por Ximdex CMS

4.1. Sobre Ximdex CMS

El entorno Ximdex sobre el que se ejecuta el módulo X-LYRE incluye un **editor de XML *wysiwyg/m***, completamente visual, que ayuda al usuario a proporcionar la meta-información y facilita la estructuración de la información mediante esquemas XML RNG enriquecidos semánticamente. La ilustración siguiente recoge Ximdex en el proceso de edición de un documento XML que es representado visualmente a partir del XML almacenado en el sistema:

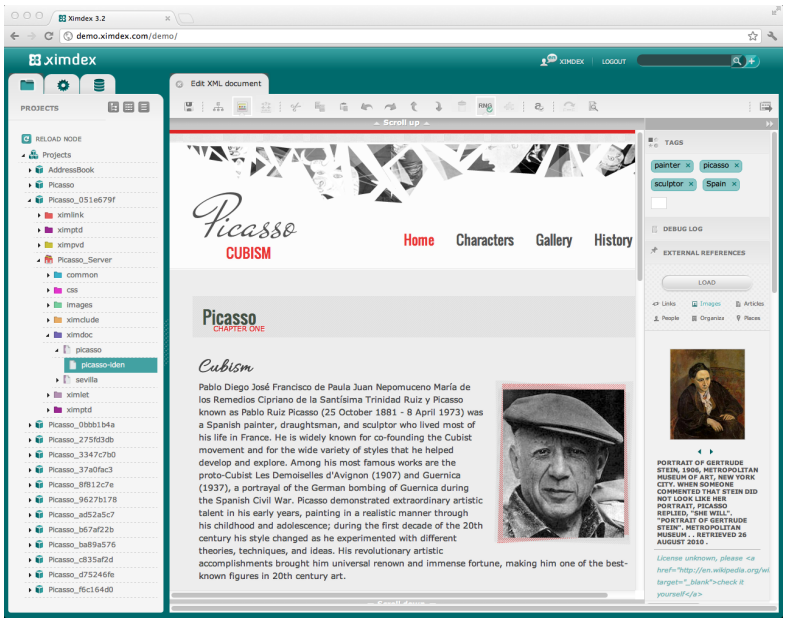


Figura 2: Ximdex CMS, editor WYSIWYG

Ximdex CMS utiliza el concepto de "**Publicación Desacoplada**", lo que garantiza una **plena neutralidad en la representación de la información**, permitiendo la transformación automática a los formatos y tecnologías existentes (XHMTL2, JSP, PHP, HTML5, ...) o incluso venideras. La publicación desacoplada utiliza ***cloud-computing* para facilitar una plena flexibilidad y escalabilidad**.

Otras funcionalidades de Ximdex CMS son, por ejemplo, la detección de intrusión en los contenidos publicados mediante el módulo **XHAWK**, el enriquecimiento automático de textos (*text enhacement*) y la generación automática de *Tags* (anotaciones semánticas) mediante el **módulo XOWL,** etc.

La Figura siguiente recoge la inserción en el documento de un conjunto de “tags” que han sido **automáticamente** sugeridas por Ximdex CMS mediante el módulo XOWL:

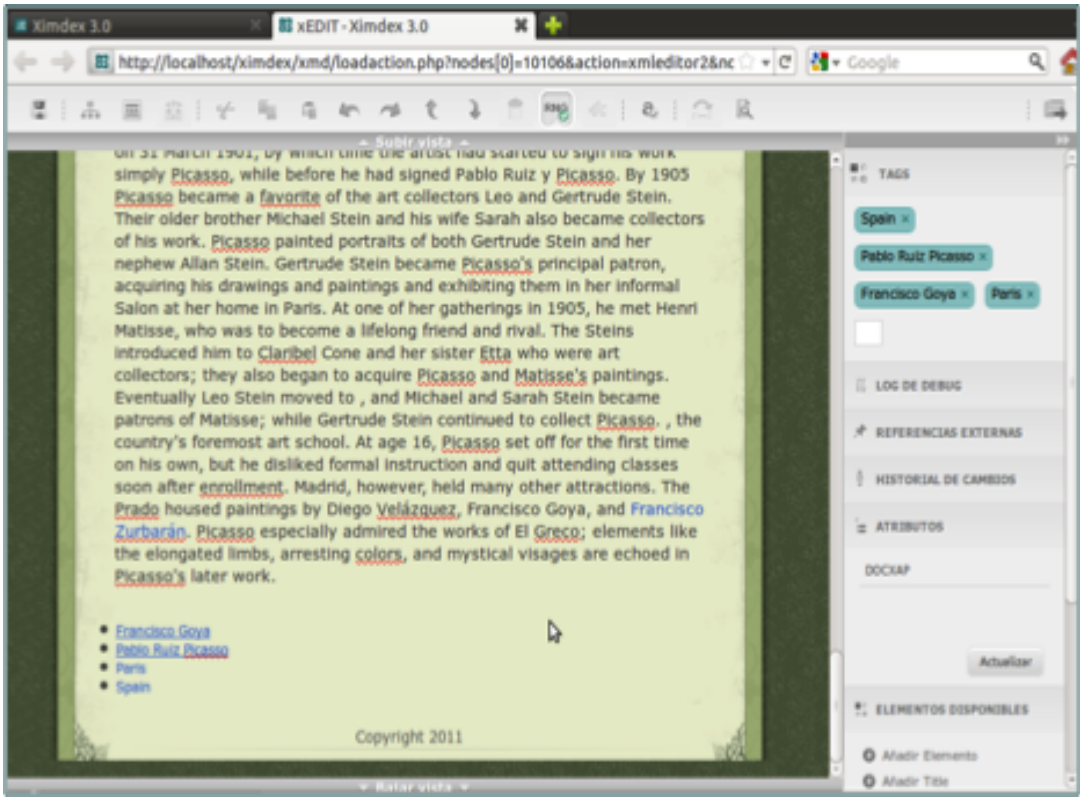


Figura 3: Generación automática de *tags*.

El CMS semántico y libre XIMDEX puede descargarse de su repositorio GIT en GITHUB (https://github.com/XIMDEX/ximdex). En la dirección http://demo.ximdex.com *puede accederse a demostradores del sistema completo y de sus módulos.*

5. Sobre Open Ximdex Evolution

Open Ximdex Evolution SL (OXE, www.ximdex.com) se centra en la gestión de la adquisición, transformación y publicación de datos, contenidos, servicios y procesos en formatos electrónicos. OXE une los paradigmas de la Web Semántica y de los Servicios Web en un mismo contexto para proporcionar soluciones de gestión semántica de contenidos en formato electrónicos para la publicación por terceros de portales de contenidos, datos y servicios.

Open Ximdex Evolution SL es actualmente la principal desarrolladora del CMS semántico XIMDEX (disponible en castellano, inglés, alemán y portugués). Versiones de nuestro entorno XIMDEX salen al mercado con licencia libre. Además, el libre acceso y la publicación de los formatos de intercambio y de transformación utilizados por el sistema, permite la modificación por el usuario de todos los aspectos que controlan la realización, personalización y explotación garantizando su independencia tecnológica.

Open Ximdex Evolution SL es miembro de la **Red Temática Española de Linked Data** (red.linkeddata.es). El módulo XLYRE para la gestión visual de datos libres del CMS Ximdex fue finalista del concurso celebrado en el evento **Open Data Sevilla** celebrado en Noviembre de 2012.

En Marzo de 2013, Open Ximdex Evolution SL ha sido elegida, de entre centenares de empresas, como una de las empresas tecnológicas más significativas a nivel Europeo por la Editorial Red Herring[8].

[8] http://www.redherring.com/events/red-herring-europe/2013_finalists/ (como XIMDEX CMS)