

Título	Adecuación de componentes existentes e implantación de un portal de datos estadísticos abiertos y enlazados		
Asunto	Datos abiertos y enlazados		
Categoría	Banco de datos		
Autor	Miguel Expósito Martín		
Unidad	Servicio de Informática y Banco de Datos		
Fecha creación	04/12/13	Fecha modificación	04/12/13
Palabras clave	Interoperabilidad, datos abiertos, reutilización, ENI		
Comentarios			
Documentos relacionados			

Sumario

El presente documento justifica el desarrollo e implantación de un portal de datos estadísticos abiertos y enlazados.

Índice de contenido

1. MOTIVACIÓN.....	3
2. ESTUDIO DE LA SITUACIÓN ACTUAL.....	4
3. ESPECIFICACIÓN DE REQUISITOS Y ENTREGABLES.....	5
3.1 Requisitos específicos del sistema portal opendata.....	5
3.2 Identificación de la información reutilizable.....	5
3.3 Identificación de recursos en datos enlazados (linked data).....	6
3.4 Definición del esquema de URIs.....	6
3.5 Gestión de identificadores de recursos linked data.....	6
3.6 Formatos y extensiones de los documentos.....	6
3.7 Metadatos.....	7
3.8 Requisitos para la adecuación de componentes	8
3.8.1 Componente gestor de publicaciones.....	8
3.8.2 Componentes banco de datos.....	8
4. DESCRIPCIÓN DE LA OFERTA.....	9
4.1 Resumen de la oferta técnica.....	9
4.1.1 Portal Opendata.....	9
4.1.2 Modificaciones en CKAN.....	9
4.1.2.1 Tema ICANE.....	9
4.1.2.2 Diferenciación de URI mediante redirección 303.....	10
4.1.2.3 Adaptación de metadatos a NTI RISP.....	10
4.1.3 Actualización de metadatos.....	11
4.1.4 Adecuación de componentes.....	11
4.2 Resumen de la Oferta Económica.....	11
5. GLOSARIO DE TÉRMINOS.....	12

1. MOTIVACIÓN

El acceso a la información estadística ofrecida por el Gobierno de Cantabria a través del ICANE puede mejorarse en su difusión, dentro del marco legal definido en la *Ley 37/2007, de 16 de noviembre, sobre **reutilización de la información del sector público***¹. En este sentido el proyecto de creación de un portal de datos abiertos estadísticos se encuadra también dentro del marco legislativo del *Real Decreto 4/2019, de 8 de enero, por el que se regula el **Esquema Nacional de Interoperabilidad** en el ámbito de la Administración Electrónica*², así como de la *Resolución de 19 de febrero de 2013, de la Secretaría de Estado de Administraciones Públicas, por la que se aprueba la **Norma Técnica de Interoperabilidad de Reutilización de recursos de la información***³ y sus posteriores guías de aplicación. La conveniencia y necesidad de definir una estrategia de apertura de datos estadísticos se apoya en la persecución de los siguientes objetivos:

- **Aumento de la interoperabilidad** dentro y fuera de la Administración.
- Incremento de la transparencia y productividad, **evitando duplicación de esfuerzos y reduciendo costes** para todos los usuarios de los datos públicos.
- Promoción de la estandarización de formatos, aumento de la facilidad de uso de dichos datos para desarrolladores o **infomediarios** y extensión de los servicios existentes en lugar de reemplazo.

Una política de apertura de datos potenciaría el sector infomediario en la región, entendido este como el conjunto de empresas que generan aplicaciones, **productos y/o servicios de valor añadido destinados a terceros** a partir de la información del sector público (según un estudio del ONTSI, La estimación del volumen de negocio asociado directamente a la actividad infomediaria de las empresas que constituyen el universo de dicho estudio se sitúa **entre 550 y 650 millones de euros**⁴).

El ICANE, dada su naturaleza y funciones en cuanto a la obtención, elaboración y difusión de datos estadísticos, entra en juego como un participante fundamental en las posibles políticas de liberación de datos. Por ello, con este proyecto pretende, utilizando los conjuntos de datos ya existentes, poner en marcha otro “escaparate de datos” o portal de datos estadísticos abiertos y enlazados con interfaces de programación de aplicaciones (APIs) consideradas como estándares “*de facto*” a nivel global, cumpliendo estrictamente con lo dispuesto en la Norma Técnica de Interoperabilidad de Reutilización de recursos de la información y sirviendo como referente para la liberación de conjuntos de datos abiertos y enlazados en la Administración Pública Regional.

1 <https://www.boe.es/buscar/doc.php?id=BOE-A-2007-19814>

2 <http://www.boe.es/boe/dias/2010/01/29/pdfs/BOE-A-2010-1331.pdf>

3 www.boe.es/boe/dias/2013/03/04/pdfs/BOE-A-2013-2380.pdf

4 http://datos.gob.es/datos/sites/default/files/files/estudio_infomediario/Info_sector%20infomediario_2012.pdf

2. ESTUDIO DE LA SITUACIÓN ACTUAL

Tras la puesta en marcha a principios de 2013 de las nuevas funcionalidades de datos estadísticos enlazados y APIs de datos y metadatos estadísticos, el ICANE ha detectado una oportunidad de negocio en el desarrollo de un portal de datos abiertos estadísticos, o “portal opendata” entendida como la posibilidad de ejecutar dicho proyecto a muy bajo coste gracias a las herramientas facilitadoras citadas ya disponibles.

La creación de un portal de datos abiertos estadísticos se encontraría, además, totalmente enmarcada dentro de la Norma Técnica de Interoperabilidad correspondiente (de Reutilización de recursos de la Información), que especifica cómo debe implementarse un portal de estas características.

Los conjuntos de datos a publicar se definirían a partir de las estadísticas ya difundidas por el ICANE, pasando a dar acceso a sus datos y metadatos a través de APIs estándar además de las ya existentes en la infraestructura web actual.

Una vez en marcha, este portal podría ser extendido con tantos conjuntos de datos como se considerase necesario, pudiendo convertirse así en la referencia de datos abiertos del Gobierno Regional.

3. ESPECIFICACIÓN DE REQUISITOS Y ENTREGABLES

Los requisitos para cada línea de actuación son los siguientes:

3.1 Requisitos específicos del sistema portal opendata

- Se creará un espacio dedicado con un índice que agrupe el acceso a los datos ya disponibles. Se incluirá asimismo **información sobre qué es la reutilización de datos públicos**, sus beneficios, información legal, etc.
- Se publicarán y **mantendrán actualizados** los documentos y recursos de información y los catálogos, tanto en su origen, como los espacios web de índice (catálogos de información pública reutilizable). Asimismo, se **informará de las condiciones de reutilización** aplicables sobre los conjuntos de datos.
- El portal se integrará con la infraestructura ya existente en ICANE **enlazando**, de la forma más transparente posible, **con las APIs de datos y metadatos** disponibles en el portal de difusión estadística actual.
- La tecnología subyacente será, preferiblemente, **CKAN**, por tratarse de un sistema abierto, estándar “*de facto*” y con reputada calidad dentro del ecosistema de datos abiertos (Reino Unido utiliza e impulsa esta tecnología)⁵. Se creará un catálogo DCAT utilizando al máximo los componentes ofrecidos en la instalación por defecto del portal y siguiendo lo establecido en la NTI RISP⁶ y su Guía de Aplicación⁷.
- En caso de ser necesario desarrollar a medida o añadir cualquier módulo de extensión al software base, se **procurará revertir los cambios** a través de los canales establecidos para la participación de la comunidad, con objeto de intentar, en la medida de lo posible, conseguir su integración en versiones posteriores del mismo.

3.2 Identificación de la información reutilizable

- En caso de que determinados conjuntos de datos, por su carácter, necesiten un procesamiento automatizado costoso y difícil, se consideraría como conjunto de datos reutilizable el inventario de los bancos de documentos o datos. De esta forma, se podría facilitar la reutilización, tanto del elemento final como de los metadatos que componen el catálogo.
- Para la construcción de identificadores de recursos uniformes (URIs) se tendrán en cuenta los **requisitos concretos mínimos** especificados en el punto IV.2 de la norma. En la implementación actual de icane.es se cumplen todos los requisitos, si bien el esquema de URIs no es conforme al definido en el anexo II de dicha norma. **Se tratará de encontrar una fórmula que permita adaptar al máximo el esquema actual al propuesto en el anexo, preferiblemente a través de mapeos de taxonomías y redirecciones.**

5 <http://data.gov.uk/>

6 http://www.boe.es/diario_boe/txt.php?id=BOE-A-2013-2380

7 http://administracionelectronica.gob.es/recursos/pae_000023387.pdf

3.3 Identificación de recursos en datos enlazados (*linked data*)

Además de la identificación abstracta de cada recurso semántico, existirán otros dos URI asociados que identifican al documento HTML o similar legible por personas y a la descripción semántica en RDFa través de cualquiera de las notaciones existentes.

Cualquier recurso semántico estará identificado por tres URIs:

- URI abstracto que identifica conceptualmente al recurso en sí.
- URI del documento para personas (HTML o similar).
- URI de la descripción semántica (expresado en cualquier notación de RDF).

Es deseable ofrecer, además de la notación de serialización RDF/XML ya implementada, las **notaciones turtle y N3**.

3.4 Definición del esquema de URIs

- Se realizará una propuesta de esquema con el mayor nivel de conformidad posible con el punto 6.1.2 del documento de Guía de Aplicación, así como su taxonomía (punto 6.2.1.5) **siendo obligatorio mantener invariables los elementos base y carácter**.
- Para la temática o sectores propuestos en dicha guía, se intentará establecer una correspondencia lo más fiel posible con la clasificación existente en la actualidad en icane.es.
- Se tendrá en cuenta también para la elaboración de dicha correspondencia el campo **dominio** descrito en la Guía, utilizando las pautas de creación de identificadores propuestas en la misma (punto 6.1.5).
- Se elaborará una propuesta para satisfacer la regla vii) del punto 6.1.5 de la Guía: los términos que componen los URI deberán ser legibles e interpretables por el mayor número de personas posible, por lo que **se utilizará el castellano** o cualquiera de las lenguas oficiales. Idealmente, se mantendrá también la estructura actual en idioma inglés utilizando redirecciones de manera que ambas puedan convivir.

3.5 Gestión de identificadores de recursos *linked data*

- Se implementará **negociación de contenido en el servidor** que gestiona los URI en función de la cabecera del agente que realiza la petición.
- Para ello, se utilizará el mecanismo de redirecciones **HTTP 303**.
- Se implementará un mecanismo para indicar que un recurso ha **desaparecido permanentemente a través del código HTTP 410**.

3.6 Formatos y extensiones de los documentos

- Documento RDF/XML: .rdf
- Documento RDF N3: .n3
- Notación turtle: .ttl

3.7 Metadatos

- Se integrarán en los ya existentes y en la medida de lo posible los especificados en el punto 6.2.1.1 de la Guía (dcat, dc, time, xsd).
- Se implementarán en el catálogo **todos los metadatos obligatorios** y todos aquellos no obligatorios en la medida de lo posible. Se aprovechará cualquier funcionalidad incorporada "por defecto" en el sistema de información utilizado para llevar a la práctica el portal opendata.
- Se implementarán todos los metadatos de conjuntos o recursos de información obligatorios no incluidos actualmente en icane.es, así como todos aquellos no obligatorios en la medida de lo posible. Concretamente:
 - descripción (dct:description)
 - temática/s(dcat:theme)
 - etiqueta/s(dcat:keyword)
 - identificador(dct:identifier)
 - fecha de creación(dct:issued)
 - idioma/s(dc:language)
 - organismo que expone y publica los datos(dct:publisher)
 - condiciones de uso(dct:license)
 - cobertura temporal (dct:temporal)
 - vigencia del recurso (dct:valid)
 - recursos relacionados (dct:references)
 - normativa(dct:confomsTo)
 - distribución/es(dcat:distribution).
- El ICANE añadirá los metadatos faltantes en su API de metadatos, y poblará todos aquellos de los que se disponga información. En caso de ser aprovechables, se utilizarán también los metadatos que ya existen.
- Se implementarán los metadatos no obligatorios que no estén incluidos actualmente en icane.es, en la medida de lo posible.
- Además de la **taxonomía de sectores**, se procurará utilizar también la **taxonomía de organismos** descrita en el punto 6.2.1.1 de la Guía. El ICANE actualizará su API de metadatos incluyendo el campo ID correspondiente utilizado por el [Directorio Común de Unidades Orgánicas y Oficinas](http://administracionelectronica.gob.es/ctt/verPestanaGeneral.htm?idIniciativa=238#.UqA5_Y15Yck)⁸.

8 http://administracionelectronica.gob.es/ctt/verPestanaGeneral.htm?idIniciativa=238#.UqA5_Y15Yck

3.8 Requisitos para la adecuación de componentes

3.8.1 Componente gestor de publicaciones

De cara a una posible y futura integración de las publicaciones estadísticas en formato PDF en el portal *opendata*, se hace necesaria una previa integración en la API de datos y metadatos estadísticos. Además, su integración permitirá tratar a todos los conjuntos de datos estadísticos, sean del tipo que sean, de manera homogénea. Para ello se fijan los siguientes requisitos:

- En la página actual de publicaciones (<http://www.ican.es/publications>) se mantendrán los tres nodos padre que se presentan en la actualidad (económicas, demográficas, síntesis), pero en vez de formar un árbol que cuelgue de ellos, se les enlazará con otra página con iconos similares a los del árbol de temas (<http://www.ican.es/economy>). A partir de ahí se implementará el árbol usando la API de metadatos del mismo modo que con las series, pero sin dividir en datos regionales, históricos y municipales.
- Se añadirá una categoría denominada "publicaciones" a las tres ya existentes y se generará el árbol a partir de dicho padre.
- Se añadirá también una nueva sección, denominada "síntesis".
- Se desarrollará un componente de *Liferay* que permita consultar las publicaciones de la misma manera que se hace con las series estadísticas, utilizando los cambios en la API de datos y metadatos estadísticos propuestos y a implementar por parte de los técnicos del ICANE.

3.8.2 Componentes banco de datos

Con objeto de mejorar la interacción entre el nuevo portal de datos abiertos y la ya existente API de datos y metadatos estadísticos (entendida esta como las capacidades de extracción de datos), se requieren las siguientes funcionalidades:

- Extensión de la API de datos y metadatos estadísticos para permitir su parametrización, esto es, pudiendo enviar consultas a medida con filtros sobre las distintas dimensiones de una serie a través de la URI.
- Selección de elementos de la consulta por defecto, aumentando así la flexibilidad para seleccionar únicamente miembros de primer nivel en las categorías jerárquicas.

Modificación de URI de fuente, quedando claramente diferenciadas las del proveedor y las de la encuesta o conjunto de datos fuente concretos.

4. DESCRIPCIÓN DE LA OFERTA

Para el desarrollo del proyecto se solicitó presupuesto a Alejandro Villar, Ingeniero de Telecomunicación profesional cualificado y con experiencia en entornos de *Business Intelligence*, web semántica y tratamiento de datos estadísticos ampliamente demostrada a través de su colaboración en otros proyectos relacionados con el banco de datos estadísticos del ICANE.

Según la oferta presentada, todas las modificaciones o extensiones propuestas serán realizadas en forma de librerías o clases adicionales, procurando alterar lo menos posible el código fuente de los componentes existentes, de forma que se faciliten las actualizaciones de la aplicación en el futuro.

El trabajo a entregar al ICANE consistirá en:

- Código fuente.
- Documentación técnica: memoria y manuales.

4.1 Resumen de la oferta técnica

A continuación se incluyen detalles relevantes sobre la propuesta técnica:

4.1.1 Portal Opendata

El Portal usa como sistema base el software CKAN, desarrollado por la *Open Knowledge Foundation*⁹ (OKFN), en su última versión estable en el momento de inicio de los trabajos (concretamente, la versión 2.2).

Los metadatos publicados para cada conjunto de datos (*dataset*) se han importado mediante un *script* que los obtiene directamente del punto de acceso a la API de Metadatos ICANE.

En conjunto con el ICANE, se han establecido equivalencias entre las distintas propiedades del modelo de metadatos local y el establecido por la Norma Técnica de Interoperabilidad de Reutilización de Recursos de la Información del Sector Público¹⁰ (NTI RISP).

4.1.2 Modificaciones en CKAN

Todas las modificaciones realizadas sobre la distribución estándar de CKAN se han llevado a cabo mediante el uso de extensiones, utilizando la API de *plugins* de CKAN, y facilitando de esta manera tanto la actualización futura del software base como la adición de módulos complementarios de terceros.

En concreto, se han desarrollado tres módulos, que se describen en las siguientes secciones:

4.1.2.1 Tema ICANE

Se ha llevado a cabo el desarrollo de un tema sencillo que adapta la apariencia de CKAN a los colores y estilos corporativos del ICANE.

9 <http://okfn.org/>

4.1.2.2 Diferenciación de URI mediante redirección 303

Este módulo complementa las capacidades de servicio de datos sobre *datasets* en formato RDF por parte de CKAN con un sistema de redirecciones HTTP 303, haciendo uso de un proceso de negociación de contenido¹¹.

El sistema intercepta las solicitudes de datos sobre *datasets* que no contengan en su ruta información de formato de serialización, y estudia la cabecera HTTP **Accept** enviada por el cliente para redirigirlo, usando un código 303, hacia la representación más adecuada.

4.1.2.3 Adaptación de metadatos a NTI RISP

Aquellos metadatos que CKAN no ofrece, pero que se exigen o recomiendan desde la NTI RISP, han sido adaptados al formato de esta última. En la siguiente tabla se pueden ver las equivalencias entre ambos modelos, donde un *dataset* CKAN se corresponde con un `dcat:Dataset`:

Modelo CKAN	NTI RISP
Descripción	<code>dct:description</code>
Etiquetas	<code>dcat:keyword</code>
URL Completa Dataset	<code>foaf:homepage</code>
Título	<code>rdfs:label</code>
URL Dataset	<code>dct:identifier</code>
Titulo	<code>dct:title</code>
Fecha de creación	<code>dct:issued</code>
Fecha de modificación	<code>dct:modified</code>
Extra 'accrualPeriodicity' ¹²	<code>dct:accrualPeriodicity</code>
Extra 'URI'	<code>dct:references</code>
Extra 'spatial'	<code>dct:spatial</code>
Recurso	<code>dcat:Distribution</code>
Autor	<code>dct:creator</code>
Texto fijo “es”	<code>dct:language</code>
Datos fijos contacto ICANE	<code>dct:publisher</code>
URL Fija Licencia ICANE	<code>dct:license</code>

El resto de propiedades extra no mencionadas arriba se añadirá como una `dct:relation` con sus correspondientes `rdfs:label` y `rdf:value`.

¹¹ Ver <http://avillar.es/cooluris#r303uri>.

¹² Sólo será necesario añadir el término de periodicidad del vocabulario Collection Description Frequency Namespace que va a continuación de “<http://purl.org/cld/freq/>”.

Adicionalmente, en caso de solicitar la lista de datasets (ruta `/dataset`) en formato RDF/XML o Notation3 (o, en su lugar, las rutas `/dataset.rdf` y `/dataset.n3`, respectivamente), el sistema genera un listado de los conjuntos de datos existentes en formato DCAT Catalog, de acuerdo con la Norma.

Se estudiará el rendimiento de este sistema y se evaluarán posibles optimizaciones, tales como el uso de un sistema de almacenamiento temporal de ficheros a modo de *cache*.

4.1.3 Actualización de metadatos

Para la actualización de metadatos se ha desarrollado un script capaz de leer en lote la información de *datasets* de la API de metadatos ICANE, y de crear *datasets* y recursos en el repositorio CKAN.

4.1.4 Adecuación de componentes

En ocasiones, el usuario puede hacer consultas al banco de datos en que seleccione categorías hijas cuyos miembros no sean únicos, sin incluir los padres de estas en la solicitud. Un caso muy claro son las jerarquías temporales: se pueden solicitar únicamente los datos mensuales, deseleccionando los años, que mostrarían los datos agregados anuales. En este caso, tendríamos una ristra de miembros [“Enero”, “Febrero”, ..., “Diciembre”] que se repetiría para cada año, pero sin información sobre a qué anualidad corresponde cada miembro.

Se añadirá, por tanto, la posibilidad de que estos miembros, cuando no figure en la consulta el miembro padre al que pertenecen, incluyan la etiqueta de este en la suya; para el ejemplo anterior, obtendríamos una lista del tipo [“2001 - Enero”, “2001 - Febrero”, ..., “2001 - Diciembre”].

4.2 Resumen de la Oferta Económica

Trabajos a realizar.....	9650,00 €
Retención (21%).....	2026,50 €
IVA (21%).....	2026,60 €
TOTAL.....	9650,00 €

El Servicio de Informática y Banco de Datos del ICANE selecciona y está conforme con la oferta de Alejandro Villar, dado que cumple con todos los requisitos (superando incluso los mínimos exigidos) a un coste muy competitivo en comparación con los costes de otros portales de similares características de ámbito municipal en la región.

5. GLOSARIO DE TÉRMINOS

- API: interfaz para programación de aplicaciones, del inglés *Application Programming Interface*. Se trata de un conjunto de funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción. En el caso del ICANE, su API de datos y metadatos estadísticos proporciona un conjunto de funciones para la descarga de datos y metadatos de forma automática.
- Business Intelligence: conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización.
- CKAN: plataforma de portales de datos abiertos líder a nivel mundial.
- DCAT: vocabulario RDF diseñado para facilitar la interoperabilidad entre catálogos de datos publicados en la Web.
- HTML: siglas de *HyperText Markup Language* (lenguaje de marcas de hipertexto), hace referencia al lenguaje de marcado para la elaboración de páginas web.
- HTTP: *Hypertext Transfer Protocol* o HTTP (en español protocolo de transferencia de hipertexto) es el protocolo usado en cada transacción de la *World Wide Web*.
- Infomediario: Conjunto de empresas que generan productos y/o servicios para su comercialización a terceros, a partir de la información del sector público. Esto incluye, tanto a las empresas que se han creado con esta finalidad como a aquellas que pese a no tener ésta como única finalidad, poseen un área y/o departamento específico dedicado a la creación y comercialización de nuevos productos y/o servicios basados en la información del sector público
- Interoperabilidad: capacidad de los sistemas de información, y por ende de los procedimientos a los que éstos dan soporte, de compartir datos y posibilitar el intercambio de información y conocimiento entre ellos.
- Java: lenguaje de programación orientado a objetos creado por la empresa *Sun Microsystems* a principios de los años 90. Plataforma abierta que constituye la base del desarrollo de software en gran número de Administraciones Públicas.
- Linked Open Data: la propuesta de datos enlazados (*linked data*) surge dentro de marco general de la Web semántica. El término "datos enlazados" hace referencia al método con el que se pueden mostrar, intercambiar y conectar datos a través de URIs desreferenciables en la Web.
- Liferay: gestor de portales web desarrollado en tecnología Java.
- Metadatos: datos que describen otros datos.
- Notation 3 ó N3: forma abreviada de serialización no-XML de modelos en RDF, diseñado pensando en la legibilidad por parte de humanos.

- ONTSI: Observatorio Nacional de las Telecomunicaciones y la Sociedad de la Información, es un órgano adscrito a la entidad pública empresarial Red.es, cuyo principal objetivo es el seguimiento y el análisis del sector de las Telecomunicaciones y de la Sociedad de la información.
- Opendata: filosofía y práctica que persigue que determinados datos estén disponibles de forma libre y abierta a todo el mundo, sin restricciones de copyright, patentes u otros mecanismos de control.
- RDF: *Resource Description Framework*, o marco de descripción de recursos, es un marco de trabajo para metadatos semánticos en la web desarrollado por el W3C.
- RDFa: conjunto de extensiones de XHTML propuestas por W3C para introducir semántica en los documentos HTML.
- Turtle: (*Terse RDF Triple Language*), formato para serializar RDF.
- URI: *Uniform Resource Indicator*, cadena de caracteres corta que identifica inequívocamente un recurso (servicio, página, documento, dirección de correo electrónico, enciclopedia, etc.).
- *World Wide Web*: sistema de distribución de información basado en hipertexto o hipermedios enlazados y accesibles a través de Internet.
- XML: *Xtensible Markup Language* (lenguaje de marcas extensible), es un lenguaje de marcas desarrollado por el W3C.

Fuentes: W3C, ONTSI, CKAN, Wikipedia.

En Santander, a 4 de Diciembre de 2013

EL JEFE DE SECCIÓN DE INFORMÁTICA ESTADÍSTICA Y BANCO DE DATOS

Fdo.: Miguel Expósito Martín