Estrategias de difusión web de datos estadísticos en Cantabria

Miguel Expósito Martín

Alberto Lezcano Lastra

Instituto Cántabro de Estadística

Dirección General de Organización y Tecnología

Gobierno de Cantabria

Índice de contenido

1. Introducción

La presente ponencia, elaborada con motivo de las XVIII Jornadas Estadísticas de las Comunidades Autónomas, pretende ofrecer una visión global de cuál ha sido y es en la actualidad la posición estratégica del Instituto Cántabro de Estadística en materia de difusión de datos estadísticos a través de la Web.

Dicha estrategia pasa por intentar hacer frente a algunos de los retos técnicos de la difusión de estadísticas oficiales, tales como: el abandono definitivo del papel, la proliferación de tecnologías móviles, el cambio de hábitos en el consumo de contenidos web por parte de los usuarios, la dificultad de identificación y contextualización de conjuntos de datos estadísticos (así como su de procesado automático), los problemas de interoperabilidad derivados entre productores y consumidores de datos estadísticos y la necesidad de ofrecer a la ciudadanía y a la sociedad herramientas útiles y suficientes para poder devolver los datos públicos y capacitar de cara a su correcta interpretación sin ambigüedad.

A lo largo de la ponencia se efectúa un recorrido de las distintas novedades que se han ido incorporando a esta área desde la puesta en práctica de la nueva filosofía de banco de datos estadísticos basado en tecnología OLAP[OLAP] en el año 2010 hasta las nuevas líneas de trabajo en nuevos estándares y movilidad para 2015, pasando por la difusión de datos estadísticos abiertos y enlazados.

2. Difusión básica: datos tabulares

Las bases del sistema de difusión de datos estadísticos existente en Cantabria en la actualidad se establecieron en el año 2011, con la finalización del proyecto de reestructuración del banco de datos del ICANE. La evolución del producto anterior quedó estancada en su primera versión, desarrollada principalmente sobre tecnología JSP; su gestión y mantenimiento resultaban costosos, presentando grandes barreras tecnológicas de cara a asegurar la calidad de los datos difundidos. Asimismo, su potencia de consulta era poco flexible y limitada en comparación con las nuevas tecnologías de análisis exhaustivo de datos.

En el momento de la reestructuración, destacaban dos corrientes tecnológicas en la difusión de datos estadísticos: sistemas basados en PC-Axis[PCX] (INE, ISTAC, IBESTAT) y sistemas basados en *Business Intelligence*[BI] (AEAT, Instituto de Estadística de la Rioja, ICANE). A continuación se expone cuál fue la estrategia adoptada por el ICANE a la hora de definir la arquitectura de un sistema básico de difusión de datos estadísticos.

2.1 PC-Axis y sistemas OLAP

PC-Axis es un formato de difusión, y no de almacenamiento, de datos estadísticos. Cualidades tales como su riqueza en características y metadatos, estructura flexible, relativa sencillez de proceso automático (hay formatos más sencillos para este propósito), rico ecosistema *software* asociado (visores como *jaxi* y derivados[JAXI]) y fuerte apoyo de muchas oficinas estadísticas lo convierten en un formato idóneo para la diseminación de datos estadísticos. Sin embargo, su naturaleza en forma de archivo de texto plano lo presenta como una solución más bien pobre a la hora de tratar el almacenamiento de datos. Además, carece de herramientas para dotar de significado a los datos y metadatos que incluye. Sirvan como ejemplos de carencias las siguientes:

- Imposibilidad de realización de consultas muy complejas a través de SQL.
- Inexistencia de índices.
- Mayor complejidad en su actualización a priori.
- Inexistencia de gestión de transacciones (ACID[ACID]) y accesos concurrentes multiproceso.
- Problemas de consistencia (estados válidos de los datos) y fiabilidad (capacidad de recuperación frente a errores).
- No es *programmer-friendly.*
- No es eficiente en lecturas y escrituras no secuenciales.
- No permite contextualizar los datos y metadatos que representa.

Como solución *win-win,* el ICANE optó por un sistema mixto que permitiese el almacenamiento de datos en sistemas gestores de bases de datos relacionales y su difusión en PC-Axis, entre otros formatos. Para ello, se hacía necesaria una capa de lógica de negocio que transformara al vuelo los datos provenientes de la base de datos a dicho formato de una forma dinámica: un motor de generación y visualización de estructuras de representación de datos multidimensionales. En concreto, se eligió el servidor OLAP *Mondrian*[MON], de código abierto, así como su visor HTML asociado *JPivot*[JPV]. Al tratarse de tecnología basada en *Java* y HTML/CSS/JS, se pudo configurar fácilmente un núcleo lo suficientemente sencillo como para permitir su extensión mediante módulos independientes, así como su integración con otros sistemas o servicios.

Para la correspondencia entre las visualizaciones de tablas cruzadas en HTML (u otros formatos) y los datos almacenados en bases de datos, *Mondrian* utiliza estructuras de representación multidimensionales (cubos), que se definen en base a una sintaxis en ficheros XML denominados *esquemas*. Entre sus posibilidades de especificación se encuentran la definición de jerarquías múltiples complejas y de miembros calculados en base a fórmulas que permiten realizar los más diversos ajustes de cara a obtener el resultado deseado.

En sus primeros años de existencia, el sistema base se extendió con los siguientes módulos:

- Selector de variables: extensión HTML para realizar una selección dinámica y personalizada de categorías y variables estadísticas conforme con la normativa actual en materia de accesibilidad[ACC] que incorpora funcionalidades de extracción de determinados metadatos estructurales desde los ficheros XML de esquemas OLAP. Además, incluye tecnología Javascript para implementar determinadas características avanzadas tales como: contador de celdas en tiempo real, limitación de la selección, posibilidad de formar una consulta pivotando filas y columnas y cambiando el orden de las variables y categorías, etc.
- Componente XLS: exportación con imágenes y estilos corporativos así como organización de la información en hojas separadas (datos, metadatos, gráficos).
- Componente SDMX[SDMX]: exportador compatible con SDMX 2.0 con formato *cross-sectional data*, dada la amplia variedad de tablas pivotantes que pueden ser generadas en tiempo real y el hecho de que no es posible garantizar que todas ellas puedan ser representadas como series temporales. Se implementa en una librería de abstracción SDMX-ML / XML.
- Componente PC-Axis: exportador compatible con el estándar PC-Axis.
- Componente JSON[JSON]: exportador con estructura anidada personalizada para el ICANE y en formato *Google Data Table*, permitiendo la representación directa utilizando *Google Chart Tools*[GCH].

– Otros componentes: exportador RDF[RDF], exportación directa de series, parametrización de exportación en base a URIs... Todos ellos se tratarán convenientemente en capítulos posteriores de esta ponencia.

Finalmente, se incluye una tabla resumen indicando cómo se resuelven distintas problemáticas más propias de la difusión estadística[AGY]:

Problemática OLAP	**Solución**
Cruces indiscriminados de datos	Forzar su imposibilidad mediante restricciones en miembros calculados
Escasa o nula integración de metadatos	Integración del sistema con un microservicio proveedor de metadatos
Pobre rendimiento o difícil escalabilidad	Uso de un servidor web caché y balanceo de carga a nivel de contenedor de aplicaciones
No aceptación de valores no numéricos	El sistema acepta valores no numéricos por defecto
Problemas de control de secreto estadístico	Integración con optimizadores lineales y sistemas de ocultación de celdas[AEAT]
Alto coste y alta dependencia tecnológica de empresas y/o servicios de informática	Modesto servicio de informática y presupuesto asociado al ICANE
Poca flexibilidad	Arquitectura modular orientada a microservicios

Tabla 1: Problemática y soluciones OLAP

3. APIs de datos y metadatos

3.1 Metadata

La demanda de servicios proveedores de metadatos por parte del cada vez más creciente ecosistema *software* de productores y consumidores de datos estadísticos generó en el ICANE una necesidad de normalización y centralización de los mismos, así como de desacoplar la producción de dichos metadatos de las tecnologías de sistemas de gestión de bases de datos subyacentes.

Para satisfacer estas necesidades y ofrecer una solución tecnológicamente solvente, se optó por implementar un microservicio[MSC] productor de metadatos. Su implementación consistió en un servicio web sin estado, o *RESTful Web Service*[RST], utilizando la librería *Java JAX-RS*[JAX] integrada en el *framework* de desarrollo rápido de aplicaciones *Grails*[GRA]. Las motivaciones que llevaron a esta elección fueron las siguientes:

- Filosofía *divide y vencerás*: eliminar la complejidad reduciendo un único sistema monolítico a numerosos sistemas simples.
- Despliegues, escalado y evolución individual de cada microservicio en función de las necesidades globales.
- Facilitación de la integración de distintas tecnologías (*Java, Python, PHP...*).
- División del mantenimiento del sistema global, permitiendo la especialización.

Si bien la semántica y sintaxis del microservicio de metadatos son relativamente complejas ya que responden a la organización de la información en el ICANE, se ha tratado de simplificar al máximo su funcionalidad sin perder por ello generalidad. Los metadatos se definen mediante un modelo de entidades con sus relaciones y atributos, implementado en una base de datos relacional. Se trata de una estructura jerárquica o en forma de árbol con los siguientes niveles generales:

Entidad	Descripción	Ejemplo
Categoría	Nivel de clasificación en base a naturaleza temporal o geográfica en su dimensión más general	datos regionales, datos históricos, datos municipales...
Sección	Primer nivel de clasificación dentro de una categoría, en base a naturaleza temática	economía, población

Entidad	Descripción	Ejemplo
Subsección	Segundo nivel de clasificación dentro de una categoría, o primero dentro de una sección	mercado de trabajo, cifras de población, etc.
Serie temporal	Entidad que representa un conjunto de observaciones ordenadas sobre una característica cuantitativa de un fenómeno individual o colectivo tomado en diferentes instantes de tiempo. Los elementos de esta clase se devuelven convenientemente clasificados y anidados en una estructura junto con agrupaciones, temas y estadísticas que carecen de valor semántico.	Producto interior bruto a precios de mercado, Revisiones anuales del Padrón Municipal de Habitantes

Tabla 2: Modelo de metadatos del ICANE

El acceso a los distintos metadatos asociados a estadísticas, series temporales, documentos, etc. se proporciona a través de URIs únicos que se traducen en peticiones GET HTTP, devolviéndose en formatos XML o JSON. En su diseño, se procuró simplificar al máximo la compleja estructura jerárquica existente, así como distintas peculiaridades del dominio de datos en el ICANE (como por ejemplo, la posibilidad de clasificación o existencia de una misma serie temporal en diversas categorías o secciones). Con el tiempo, el microservicio está evolucionando hacia la máxima sencillez posible, pasando de utilizar URIs del tipo:

http://www.icane.es/metadata/api/{category}/{section}/{subsection}/{uri-tag}

A URIs del tipo:

http://www.icane.es/metadata/api/{uri-tag}

Es decir, tan sólo sería necesario conocer la etiqueta identificativa de una estadística, serie o documento (o nodo en la jerarquía, de forma general) para recuperar los metadatos asociados a los mismos. Estas etiquetas se pueden obtener fácilmente a través de la API, capaz de producir listados de nodos o series con filtros predefinidos, como por ejemplo:

http://www.icane.es/metadata/api/{category}/{section}/{subsection}/time-series-list

Este URI devolvería un listado anidado de series temporales y todos sus metadatos para una categoría, sección y subsección dadas, entre los que se encuentran las etiquetas que las identifican de manera unívoca.

La API de metadatos se complementa con una documentación detallada ofrecida a través de métodos en la propia API y presentada mediante la librería *Swagger*[SWA] en su variante *Java*, que proporciona una especificación y marco de trabajo muy completos para describir, producir, consumir y visualizar servicios web *RESTful*.

3.2 Data

Si la provisión de metadatos a través de un microservicio web *Restful* fue un primer paso hacia una gran mejora en la facilitación de extracción de información estadística de la web del ICANE, parecía lógico dar el siguiente paso en la dirección de la provisión de datos a través de URIs únicos y *"des-referenciables"*[1].

Para ello, se optó por extender el producto básico *Mondrian* + *Jpivot* para que permitiera la exportación directa a través de URIs diferenciados que usaran una consulta establecida por defecto y en distintos formatos en función de los filtros ya incorporados como componentes. Dicha consulta se implementa bien a través de anotaciones ya existentes en los archivos XML con los esquemas de los cubos OLAP, bien a través de una anotación con la consulta especificada explícitamente en MDX.

La API de datos es mucho más sencilla sintácticamente que su homóloga de metadatos, pudiendo realizarse una petición de exportación directa de los datos de una serie temporal en un formato determinado utilizando el siguiente URI:

http://www.icane.es/data/api/{uri-tag}.{ext}

Es decir, tan sólo con la etiqueta identificativa de la serie y la extensión del formato deseado (XLS, JSON, SDMX, RDF, PC-Axis) el sistema devuelve los datos empaquetados según la preferencia del usuario.

3.3 Valoración de resultados

Era de esperar que los mayores consumidores de metadatos fuesen otros sistemas o servicios dentro del propio ICANE: portal web corporativo, banco de datos, sistemas de aseguramiento de la calidad de los datos, así como todos sus subsistemas. Para facilitar su integración, se desarrolló una pequeña librería cliente de la API (o *wrapper*) en *Java* que permitiera al resto de subsistemas consumir directamente del microservicio. La API de metadatos ha resultado ser flexible, *"cacheable"*, escalable, independiente y fácil de usar, mantener, integrar y extender. Por todo ello, ha permitido importantes ahorros en costes de desarrollo de otros productos *software*.

Por otra parte, la exportación directa de datos en distintos formatos a través de URIs mediante la API de datos ha fomentado la posibilidad de una mayor experimentación tecnológica con nuevas formas y herramientas de tratamiento de datos que consumen los datos al vuelo, los procesan y generan la salida deseada en forma de visualización, informe u similar. Ejemplos de dichas tecnologías son las aplicaciones de una sola página, o *Single Web Applications*, que se tratarán posteriormente en esta ponencia.

1 Mecanismo de recuperación de URIs que utiliza HTTP para obtener una copia de la representación del recurso al que identifica.

4. Web semántica

Debido al gran número de oficinas, agencias y organismos que recogen, procesan y publican datos estadísticos a nivel global, durante los últimos años han ido surgiendo varios estándares y metodologías para el intercambio de información (tales como SDMX), con el objeto de mejorar la interoperabilidad entre productores y consumidores de datos.

Por otra parte, desarrollándose de forma independiente, la Web Semántica (respaldada principalmente por el W3C[W3C]) se presenta como una útil herramienta para publicar tanto datos como metadatos contextualizados, haciéndolos fácilmente comprensibles y procesables por servicios de terceros y permitiendo también el establecimiento de relaciones asociadas a conceptos entre ellos.

Si bien es cierto que la idea de web semántica aún se encuentra en proceso de investigación y desarrollo, existen ya estándares y tecnologías que se están extendiendo y consolidando de forma global, tales como RDF, SKOS[SKO], SPARQL[SPQL] y especialmente en el ámbito de la estadística pública, *Data Cube*[DQ] como formato de difusión de datos estadísticos contextualizados y enlazados. En concreto, las distintas oficinas estadísticas están mostrando especial interés en cómo la web semántica podría facilitar a los técnicos y analistas el uso de datos estadísticos descritos correcta y completamente (en conjunto con otros tipos de datos, como geo-espaciales, científicos, etc.) y expresados semánticamente.

El ICANE no ha sido ajeno a esta evolución tecnológica ni a los posibles beneficios derivados de adaptar sus sistemas de difusión web al procesado automático de datos por parte de máquinas; por ello, tras la puesta en marcha de sus APIs de datos y metadatos, se ha detectado una oportunidad de negocio inmejorable para implementar y sentar las bases de una solución de datos abiertos y enlazados (*linked open data*[LKD]) en su portal web y banco de datos, poniendo sus datos y metadatos a la disposición no sólo de consumidores humanos, sino también de consumidores automáticos o máquinas.

4.1 Ámbito de actuación

Tras realizar una evaluación del estado tecnológico de la plataforma de difusión del ICANE, así como su estructura y contenido, se llegó a la conclusión de que era viable y efectivo en coste realizar las siguientes actuaciones:

- Generar un modelo RDF para la estructura de metadatos existente.
- Implementar una estrategia *linked open data* para servir entidades de metadatos.
- Desarrollar un filtro de exportación RDF *Data Cube* para el banco de datos.
- Proporcionar un punto de consulta de relaciones entre metadatos.

4.2 Modelo RDF

Para confeccionar el modelo RDF de metadatos el ICANE se definió previamente una ontología SKOS que permitiese representar las relaciones jerárquicas entre las distintas entidades. Además, se han utilizado vocabularios de uso común y extendido (tales como RDF, DCMI, FOAF...) para tipificar la mayoría de los metadatos:

Entidad	Propiedad	Valor / Descripción del metadato
Section	rdf:type	icane:Section, skos:ConceptScheme.
	skos:prefLabel y rdfs:label	Nombre o etiqueta de la sección.
Subsection	rdf:type	icane:Subsection, skos:ConceptScheme.
	skos:prefLabel y rdfs:label	Nombre o etiqueta de la subsección.
	icane:section, skos:inScheme	Sección a la que pertenece la subsección.
Category	rdf:type	icane:Category.
	rdfs:label	Nombre o etiqueta de la categoría.
	icane:acronym	Acrónimo de la categoría.
Folder	rdf:type	skos:Concept.
	skos:prefLabel, rdfs:label	Nombre o etiqueta de la carpeta.
	icane:section	Sección a la que pertenece la carpeta.
	icane:subsection	Subsección a la que pertenece la carpeta.
	skos:inScheme	Concept Schemes a los que pertenece la carpeta (sección y subsección de nuevo, en este caso)
	icane:category	Categoría a la que pertenece la carpeta.
	skos:broader / skos:narrower	Utilizado para crear una estructura jerárquica de carpetas.
Time series	rdf:type	icane:TimeSeries, qb:DataSet.
	rdfs:label, dcterms:title	Nombre o etiqueta de la serie temporal.
	dcterms:subject	Tema asociado a la serie temporal (normalmente el nodo padre en la jerarquía).
	icane:section	Sección a la que pertenece la serie temporal.
	icane:subsection	Subsección a la que pertenece la serie temporal.

Entidad	Propiedad	Valor / Descripción del metadato
	icane:category	Categoría a la que pertenece la serie temporal.
	dcterms:modified	Fecha de última actualización.
	dcterms:accrualPeriodicity	Periodicidad.
	dcterms:spatial	Reference area relative to this series.
	dcterms:temporal	Período temporal.
	dcterms:source	Fuente.
	rdfs:comment	Notas al pie.
	void:dataDump	URI de volcado de datos de la serie.
Reference area	rdf:type	icane:ReferenceArea, dcterms:Location.
	rdfs:label	Nombre o etiqueta del ámbito territorial.
Source	rdfs:label, dcterms:title	Nombre o etiqueta de la fuente.
	foaf:page	URL del sitio web principal de la fuente.

Tabla 3: Modelo RDF de metadatos del ICANE

4.3 Linked open data

Los principios de *linked open data* pueden resumirse con cuatro sencillas recomendaciones:

– Utilizar URIs para identificar los distintos recursos.

– Utilizar peticiones HTTP para el consumo de dichos URIs.

– Contextualizar los datos con información útil utilizando estándares como RDF y SPARQL.

– Incluir enlaces a otras URIs (o recursos) que permitan describir otro recursos asociados en la Web.

Definido y generado el modelo RDF, a partir de la información contenida en la base de datos del ICANE respecto a la estructura jerárquica de nodos y a los metadatos de las diversas series (entre los que se incluyen los enlaces HTTP que representan relaciones entre entidades), fue posible añadir marcado RDF para convertir las páginas tanto del portal web como del selector del Banco de Datos en *XHTML+RDFa* 1.1[RDFA], con etiquetado añadido para compatibilidad con clientes de la versión 1.0. De esta forma, el usuario humano no aprecia cambio alguno, mientras que un agente automatizado es capaz obtener toda la información semántica contenida en la página de forma transparente. Entre esta información semántica se encuentran tanto los metadatos ya disponibles para usuarios humanos como nuevos metadatos incluidos sólo para procesadores RDF. A continuación se incluye una tabla con los patrones de URIs disponibles, en base a etiquetas genéricas:

Entidad	Patrón de URI
Section	http://www.icane.es/{section}#section
Subsection	http://www.icane.es/{section}/{subsection}#section
Category (sólo RDF/XML)	http://www.icane.es/opendata/categories#{category}
Folder	http://www.icane.es/{section}/{subsection}#{category}-{folder}
Time Series	http://www.icane.es/data/{category}/{section}/{subsection}/{time-series}#timeseries
Reference area (sólo RDF/XML)	http://www.icane.es/opendata/reference-areas#{reference-area}
Source	http://www.icane.es/data/{category}/{section}/{subsection}/{time-series}#source_{id}

Tabla 4: Patrón de URIS del modelo de metadatos del ICANE

A través de un repositorio de enlaces almacenado en la base de datos de metadatos y expuesto a través de su correspondiente API, se seleccionaron recursos similares a las entidades existentes en el ICANE y se conectaron a través de las pertinentes propiedades RDF. Dada la heterogeneidad de los conjuntos de datos involucrados (todas las series del banco), la tarea de poblar los enlaces se realizó de forma manual. En la siguiente tabla se muestran las propiedades utilizadas para realizar los distintos enlaces así como el número de los mismos:

Entidad	Nº casos	Propiedad	Nº enlaces
Section	4	dcterms:subject	18
		rdfs:seeAlso	1
Subsection	27	dcterms:subject	141
		rdfs:seeAlso	43
Category	3	--	--
Folder	703	skos:closeMatch	161
		rdfs:seeAlso	199
Time Series	2707	--	--

Entidad	Nº casos	Propiedad	Nº enlaces
ReferenceArea	6	owl:sameAs	10
		rdfs:seeAlso	15
Source	2694	foaf:page	24

Tabla 5: Enlazado RDF según entidad

En cuanto a las organizaciones o entidades enlazadas desde el ICANE, se intentó seleccionar un conjunto razonable dentro de la nube LOD:

Base de datos	Nº enlaces
Geonames	4
DBpedia	45
Dbpedia española	47
INE	251 (no RDF)
Eurostat	22 (no RDF)
Lista de Encabezamientos de Materia (LEM) para las Bibliotecas Públicas	168
Lista de Encabezamientos de Materia de la Biblioteca del Congreso de EEUU	151

Tabla 6: Destino de los enlaces linked data

En el caso de los enlaces "no RDF", tan sólo se utilizaron las propiedades rdfs:seeAlso y foaf:page.

4.4 Exportación DataCube

El vocabulario RDF *Data Cube*[DQ] permite la publicación de datos multidimensionales (tales como datos estadísticos) en la web, de forma que puedan ser enlazados con conjuntos de datos o conceptos relacionados. Además, el modelo que subyace en este vocabulario es compatible con el modelo de cubo utilizado por SDMX.

El principal propósito de publicar todas las series estadísticas del ICANE en formato RDF – *Data cube* es poner dichos datos a disposición del público en un formato de fácil acceso y proceso (por ejemplo, por *crawlers* o a través de consultas en un punto SPARQL), hacerlos unívocamente identificables a nivel de registro mediante la asignación de URIs y, finalmente, facilitar la posibilidad de ser citados de tal manera que puedan ser enlazados desde fuentes externas.

Para su implementación, se decidió desarrollar otro filtro a modo de complemento del sistema de banco de datos que extendiera *Jpivot*. En dicho filtro, las series del Banco de Datos se definen como entidades de tipo qb:DataSet, según *The RDF Data Cube Vocabulary*, y sus URI también se construyen con identificadores de fragmento.

Dichas entidades contienen las siguientes propiedades:

dcterms:title para el nombre.

dcterms:subject con la URI del tema al que pertenece.

dcterms:source para la fuente de los datos.

dcterms:temporal para el rango temporal de la serie.

dcterms:accrualPeriodicity para la periodicidad.

dcterms:spatial para el ámbito territorial.

dcterms:modified para la fecha de actualización.

`rdfs:comment` para las notas al pie.

Asimismo, se representan las dimensiones, medidas y unidades de acuerdo con *The RDF Data Cube Vocabulary*, estableciéndose las siguientes correspondencias:

Sistema OLAP ICANE	**RDF Data Cube**
Cubo / serie	RDF *DataCube dataset*
Dimensión	*ConceptScheme* + *Concept class* + *Dimension Property*
Medida	RDF *DataCube MeasureProperty*
Celda	RDF *DataCube Observation*

Tabla 7: Correspondencia sistema OLAP - Data Cube

4.5 SPARQL

SPARQL[SPQL] es un protocolo y lenguaje de consultas de grafos RDF con el apoyo del W3C. La mayoría de las formas de consulta en SPARQL contienen un conjunto de patrones de tripleta (*triple patterns*) denominadas *patrón de grafo básico*. Los patrones de tripleta son similares a las tripletas RDF, excepto que cada sujeto, predicado y objeto puede ser una variable.

El objeto de desplegar un punto SPARQL en el portal del ICANE es permitir la realización de consultas avanzadas sobre los metadatos de las distintas series existentes en el banco de datos del ICANE, así como su organización jerárquica y relaciones internas.

Para su implementación se valoró utilizar una solución basada en la plataforma D2RQ[DRQ], si bien finalmente fue desechada debido a problemas de rendimiento derivados de una alta redundancia de consultas SQL producida por consultas SPARQL complejas así como a la gran variabilidad de los metadatos del banco, que implicaría frecuentes cambios en los modelos de correspondencias y re-despliegues.

Por ello, finalmente se optó por desarrollar una aplicación web a medida que genera un modelo RDF de *Apache Jena*[JNA] al vuelo utilizando la API de metadatos del ICANE y lo sirve mediante un servidor *Joseki* alojado en un contenedor *Apache Tomcat*[TCT]. La aplicación realiza comprobaciones programadas en busca de actualizaciones de metadatos, en cuyo caso actualiza el modelo en consecuencia.

Como última actuación y con objeto de aumentar la facilidad de uso entre los usuarios, se implementó un componente en el portal web con una interfaz gráfica y algunos ejemplos de consulta.

5. Portal de Datos estadísticos abiertos y enlazados

5.1 Motivación

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 plantea, 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 y cumpliendo estrictamente con lo dispuesto en la *Norma Técnica de Interoperabilidad de Reutilización de recursos de la información*[RISP] (en adelante, NTI RISP), sirviendo asimismo como referente para la liberación de conjuntos de datos abiertos y enlazados en la Administración Pública Regional de Cantabria.

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.

5.2 Requisitos

El requisito fundamental establecido para el desarrollo de un portal de datos estadísticos abiertos y enlazados en Cantabria es su conformidad con la NTI RISP de 2013, utilizando como herramienta su Guía de Aplicación[RISP]. Dicha norma tiene por objeto establecer el conjunto de pautas básicas para la reutilización de documentos y recursos de información elaborados o custodiados por el sector público. Principalmente, se centra en los siguientes aspectos técnicos:

- Selección, identificación y descripción de la información reutilizable.
- Formatos y puesta a disposición de recursos, así como condiciones de uso.
- Catálogo de información pública reutilizable.

En concreto, desde el ICANE se consideró conveniente hacer especial hincapié en la necesidad de resolver las siguientes cuestiones:

- Automatización de tareas de carga y actualización de conjuntos de datos.
- Implementación de negociación de contenido y redirecciones HTTP 303 y 410.
- Esquema de construcción de URIs en castellano, adaptando el esquema actual mediante el almacenamiento de correspondencias de etiquetas.
- Incorporación de pre-visualizadores de datos en tiempo real.

5.3 Solución

El Portal usa como sistema base el software CKAN[CKA], desarrollado por la *Open Knowledge Foundation*[OFKN], 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. Asimismo, se han establecido equivalencias entre las distintas propiedades del modelo de metadatos local y el establecido por la NTI RISP.

Si bien CKAN ofrece de serie la mayor parte de las funcionalidades necesarias para implementar un portal de datos abiertos (búsqueda, etiquetado y navegación de datos, correspondencias de metadatos, federación de portales, rico marco de trabajo de extensiones, API *RESTful* estándar *"de facto"*...), fue necesario realizar las siguientes modificaciones utilizando su API de *plugins*:

- Tema ICANE: un tema sencillo que adapta la apariencia de CKAN a los colores y estilos corporativos del ICANE.
- Negociación de contenido y redirecciones 303 y 410: 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.
- Adaptación de metadatos del ICANE 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	dct:description
Etiquetas	dcat:keyword
URL Completa Dataset	foaf:homepage
Título	rdfs:label
URL Dataset	dct:identifier

Modelo CKAN	**NTI RISP**
Titulo	dct:title
Fecha de creación	dct:issued
Fecha de modificación	dct:modified
Extra 'accrualPeriodicity'	dct:accrualPeriodicity
Extra 'URI'	dct:references
Extra 'spatial'	dct:spatial
Recurso	dcat:Distribution
Autor	dct:creator
Texto fijo “es”	dct:language
Datos fijos contacto ICANE	dct:publisher
URL Fija Licencia ICANE	dct:license

Tabla 8: Correspondencia metadatos CKAN / NTI RISP

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

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.

6. Visualización

En los tiempos de los "grandes datos" (o *Big Data*[BDT]), en los que la información bruta sin tratar es capaz de saturar la capacidad humana de proceso, es de vital importancia para los organismos públicos ser capaces de potenciar la capacidad de comunicar resultados de manera clara y sencilla. Uno de los mecanismos de comunicación más efectivos es la representación gráfica o visual de datos (en este caso, datos estadísticos). Tanto la parte estética como la funcional deben ir de la mano, intentando transmitir los aspectos clave de cualquier conjunto de datos más o menos complejo de manera intuitiva al público en general.

Si bien el banco de datos estadísticos de Cantabria contaba con interesantes capacidades de extracción automática de datos, las posibilidades de representación gráfica de los mismos estaban limitadas, desaprovechando un gran potencial de difusión. Por ello se propuso la incorporación al banco de datos estadísticos de nuevos módulos de generación gráfica que incluyesen también cartografía para complementar la oferta de datos municipales, uno de los puntos fuertes en la producción del ICANE.

6.1 Representación gráfica básica

El banco de datos del ICANE incluye de serie funcionalidades básicas de representación gráfica de datos altamente configurables. Sin embargo, dichas gráficas no son interactivas y sus posibilidades de representación son tan complejas que se hacen difíciles de utilizar por parte de los usuarios del banco.

Dada la naturaleza modular del sistema *software* que soporta el banco de datos, se consideró oportuno extenderlo mediante un nuevo complemento que se integrara con librerías gráficas abiertas, potentes y de uso extendido. De este modo, en caso de que los clientes que realicen consultas al banco tengan habilitada la ejecución de *JavaScript* en sus navegadores, se sustituye el sistema de representación de datos en gráficas mediante *JFreeChart*[JFC], que *JPivot* utiliza de forma nativa, por otro basado en la API de visualización de *Google* (*Google Visualization* API).

La implementación de la solución en el banco permite la elección, de forma sencilla por parte del usuario, de gráficas de barras, líneas, secciones circulares y áreas. En caso de intentar representar demasiados datos o de intentar utilizar una representación no adecuada para la naturaleza de los datos en cuestión, la librería muestra un aviso y se abstiene de representar la gráfica.

6.2 Visualización en mapas de coropletas

Al margen de las limitaciones existentes en las funciones de representación gráfica básica, en el banco de datos del ICANE existía una gran carencia en cuanto a la representación gráfica de datos municipales: la inexistencia de un mecanismo de visualización en mapas regionales.

Para cubrir este vacío, se ha desarrollado un sistema adicional de representación de datos estadísticos sobre mapas (*"GeoCharts"*[GEO]). El sistema dispone de dos tipos de mapas: coropletas a escala mundial y coropletas y círculos proporcionales a escala municipal. A nivel de cada serie o cubo, se puede configurar el tipo de mapa a utilizar mediante la adición de etiquetas en el *schema* XML, de la siguiente manera:

- El cubo puede contener una anotación de nombre *"geoChartType"* y valor *"world"* (para mapas mundiales) o *"cantabria-munip"* (para mapas municipales de Cantabria).
- Exactamente un nivel de alguna de las dimensiones del cubo debe contener una propiedad (<Property>) con nombre *"geoCode"* y una columna con el código de área, que a su vez será:
  - Un código de dos (ES) o tres (ESP) letras para indicar el país, en los mapas mundiales.
  - Un código de municipio INE (39005), en los mapas municipales de Cantabria.
- Adicionalmente, cada medida podrá tener una anotación de nombre *"geoChartStyle"* y valor *"choropleth"* o *"circle"*, para indicar, en los mapas municipales, si se usarán coropletas o círculos proporcionales. En caso de no existir esta propiedad, o de tener un valor no válido, se usarán coropletas por defecto.

El usuario tendrá la opción de modificar qué tipo de corte transversal quiere representar en el mapa mediante cajas de selección para todas las dimensiones utilizadas en la consulta (salvo la geográfica, que se usará para representación).

En el caso de los mapas mundiales, se usa la API de visualización de *Google* (en concreto, el *Google GeoChart* a escala mundial). Para los municipios de Cantabria, se ha desarrollado una solución de representación a medida basada en la librería de gráficos *JavaScript RaphäelJS*[RPH].

7. Líneas futuras

7.1 *Toolboxes* facilitadoras

Si bien las APIs de datos y metadatos han sido de gran utilidad para seguir creando productos de difusión estadística en el ICANE, se tiene poca constancia del uso de estos productos por parte de otros segmentos de consumidores (ciudadanos, empresas, Universidad, etc.). Uno de los numerosos retos de las oficinas estadísticas en cuanto a difusión estadística se refiere es intentar simplificar los servicios productores de datos hasta que su facilidad de uso anime a su utilización por parte de agentes externos. Para ello, el ICANE enfoca su estrategia en las siguientes actuaciones concretas:

- Liberación y puesta a disposición del público en general en *Github*[GHB] del cliente de metadatos desarrollado en *Java*.
- Fomento e impulso del desarrollo de clientes o *wrappers* en otras tecnologías, como *Python, R, Javascript*...
- Puesta a disposición del público de pequeños fragmentos de código o cuadernos con ejemplos de uso de los distintos productos ofrecidos por el ICANE y otras librerías de uso común en función de la tecnología.

7.2 Aplicaciones de una sola página (SPAs) en Javascript

Una aplicación de una sola página (o *Single Page Application*, SPA) es un tipo de aplicación web que se presenta en el navegador y no necesita de recarga de página durante su uso. Históricamente, han existido tres formas de SPA: *applets* en Java, aplicaciones de *Adobe Flash o Flex* y aplicaciones *Javascript*. Sin embargo, tradicionalmente, estas últimas no se han utilizado para desarrollar aplicaciones críticas por problemas derivados de falta de confianza en la propia tecnología. Hoy día, se puede afirmar que las debilidades arrastradas por *Javascript* como medio para desarrollar aplicaciones ricas en el navegador han sido superadas y el valor añadido por sus ventajas ha aumentado:

- Los navegadores web se han convertido en la aplicación de escritorio más utilizada a nivel global.
- El entorno de ejecución de *Javascript* en el navegador es uno de los más distribuidos en el mundo (gracias, en buena parte, a su inclusión en dispositivos móviles). Además, estos nuevos motores la han convertido en una tecnología sorprendentemente rápida y capaz de rivalizar, en ocasiones, con lenguajes compilados.
- Las tecnologías que normalmente acompañan a *Javascript*, tales como HTML5, SVG y CSS3, han avanzado lo suficiente como para estar a la altura en velocidad y calidad , en cuanto a sus capacidades gráficas se refiere, con las soluciones tradicionales.
- El ecosistema *Javascript* evoluciona: JSON, *Node.js*[NDE], *MongoDB*[MNG], *Jquery...* etc.

En lo que respecta a la difusión estadística, la proliferación de APIs de datos y metadatos y de librerías de visualización está fomentando el desarrollo de aplicaciones *Javascript* simples capaces de extraer esta información al vuelo, tratarla y producir resultados a modo de sumario con capacidades gráficas realmente interesante.

Entre las líneas de trabajo actuales y futuras del ICANE se encuentra el diseño e implementación de un producto *software* basado en *Node.js* y SPAs para ofrecer información sobre los municipios de Cantabria (evolución del anterior producto estático *"Fichas Municipales de Cantabria"*). Cada municipio tendría su SPA que consumiría datos y metadatos provenientes de las APIs del ICANE, pudiendo enviar para ello consultas parametrizadas a través de HTTP. Estos datos se utilizarían para elaborar representaciones tabulares sencillas y potentes visualizaciones dinámicas con cuadros de mando que permitan a los usuarios conocer los últimos indicadores actualizados relativos al municipio de interés.

7.3 Diseño web adaptable (Responsive Web Design)

En los últimos años, las ventas de dispositivos móviles han duplicado a las de computadoras personales; de hecho, ciertos estudios aseguran que las ventas de tabletas superarán a las de PCs en un par de años[2]. Asimismo, el tráfico en Internet es ya indiscutiblemente móvil, quedando el tráfico procedente de conexiones fijas relegado a un segundo plano.

Si bien los usuarios de computadores personales han ido mostrado preferencia por las aplicaciones web frente a las de escritorio, la tendencia en los terminales móviles viene siendo la opuesta.

La aparición de HTML 5 podría cambiar este escenario. HTML 5 no es sólo la nueva versión del omnipresente lenguaje de marcado, sino que define un nuevo estándar para el desarrollo de páginas y aplicaciones web. Está pensado para funcionar de forma transparente en cualquier plataforma y navegador, fijo o móvil. Las aplicaciones móviles desarrolladas en HTML 5 pueden instalarse en los terminales como si fueran una aplicación más, independientemente del sistema operativo utilizado; es decir, se desarrolla en una única tecnología y se despliega en cualquier terminal, reduciendo los costes y aumentando las posibilidades de difusión de la aplicación.

Existiendo antecedentes de aplicaciones móviles de difusión de datos estadísticos por parte de otras oficinas, el ICANE ha decidido continuar con su estrategia de seguir trabajando para disponer de una infraestructura tecnológica suficiente como para facilitar el desarrollo de una aplicación en una única tecnología. Observando la evolución de HTML5 y en la misma línea que las SPAs, parece razonable aprovechar las funcionalidades de extracción de información estadística de las APIs de datos y metadatos del ICANE, así como de otras APIs y librerías gráficas *Javascript* (como D3[D3]) existentes en el mercado para construir una *app* móvil de difusión de datos estadísticos en HTML5 compatible con la mayoría de terminales y navegadores.

2 Fuente: Gartner, IDC, Strategy Analytics, company filings, BI Intelligence estimates

Bibliografía

OLAP: Wikipedia, OLAP, http://es.wikipedia.org/wiki/OLAP
PCX: Estadística Suecia, Familia PC-Axis, http://www.scb.se/sv_/PC-Axis/Start/
BI: Wikipedia, Inteligencia empresarial, http://es.wikipedia.org/wiki/Inteligencia_empresarial
JAXI: UNECE, Plataforma JAXI, http://www1.unece.org/stat/platform/display/msis/Jaxi
ACID: Wikipedia, ACID, http://es.wikipedia.org/wiki/ACID
MON: Pentaho, Sitio web de Mondrian, http://community.pentaho.com/projects/mondrian/
JPV: Tonbeller, JPivot, http://jpivot.sourceforge.net/
ACC: Portal de Administración Electrónica, Normas sobre accesibilidad web, http://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Accesibilidad/pae_normativa/pae_eInclusion_Normas_Accesibilidad.html#.U6QViz-lUck
SDMX: sdmx.org, Statistical Data and Metadata Exchange, http://sdmx.org/
JSON: json.org, Introducing JSON, http://json.org/
GCH: Google, Google Chart Tools, https://developers.google.com/chart/
RDF: W3C, Resource Description Framework, http://www.w3.org/RDF/
AGY: Alberto González Yanes, Estrategia de difusión en PC-Axis, http://www.slideshare.net/algoya/estrategia-de-difusin-en-pcaxis
AEAT: Agencia Española de Administración Tributaria, Sistema de difusión estadística OpenJRubik, http://administracionelectronica.gob.es/ctt/verPestanaDescargas.htm?idIniciativa=397#.U6QW8j-lUck
MSC: Martin Fowler, Microservices, http://martinfowler.com/articles/microservices.html
RST: Wikipedia, Representational state transfer, http://en.wikipedia.org/wiki/Representational_state_transfer
JAX: Wikpedia, JAX-RS, http://es.wikipedia.org/wiki/JAX-RS
GRA: Pivotal, Grails, https://grails.org/
SWA: Reverb, Swagger, https://helloreverb.com/developers/swagger
W3C: W3C, W3C Semantic Web Activity, http://www.w3.org/2001/sw/
SKO: W3C, SKOS Simple Knowledge Organization System , http://www.w3.org/2004/02/skos/
SPQL: W3C, SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
DQ: W3C, The RDF Data Cube Vocabulary, http://www.w3.org/TR/vocab-data-cube/
LKD: W3C, Linked Data, http://www.w3.org/standards/semanticweb/data
RDFA: W3C, RDFa 1.1 Primer - Second Edition, http://www.w3.org/TR/xhtml-rdfa-primer/
DRQ: Richard Cyganiak, Accessing Relational Databasesas Virtual RDF Graphs, http://d2rq.org/
JNA: Apache Foundation, Apache Jena, https://jena.apache.org/
TCT: Apache Foundation, Apache Tomcat, http://tomcat.apache.org/
RISP: Portal de Administración Electŕonica, Reutilización de recursos de información, http://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Interoperabilidad_Inicio/pae_Normas_tecnicas_de_interoperabilidad.html#REUTILIZACIONRECURSOS
CKA: The Open Knowledge Foundation, The open source data portal software, http://ckan.org/
OFKN: Open Knowledge Foundation, Open Knowledge Foundation, https://okfn.org/
BDT: Wikipedia, Big Data, http://es.wikipedia.org/wiki/Big_data
JFC: Object Refinery, JFreeChart, http://www.jfree.org/jfreechart/
GEO: Google, Visualization: Geochart, https://developers.google.com/chart/interactive/docs/gallery/geochart
RPH: Dmitry Baranovskiy, https://developers.google.com/chart/interactive/docs/gallery/geochart, http://raphaeljs.com/
GHB: GitHub Inc., Github, https://github.com/
NDE: Joyent, Inc., Node.js, http://nodejs.org/
MNG: MongoDB, Inc., mongoDB, http://www.mongodb.org/
D3: Mike Bostock, Data-Driven Documents, http://d3js.org/