Islas Canarias
Del 15 al 19 de noviembre de 2021

# Catálogo de Datos Abiertos de Canarias

**Jesús Alberto González Yanes**
Instituto Canario de Estadística, ISTAC
jgonyanp@gobiernodecanarias.org

**Rita Díaz Adán**
Arte Consultores Tecnológicos S.L., ARTE
rdiaada@arte-consultores.com

**José Joaquín Escobar Gómez**
Arte Consultores Tecnológicos S.L., ARTE
jescgom@arte-consultores.com

**Francisco Javier Rodríguez García**
Arte Consultores Tecnológicos S.L., ARTE
frodgar@arte-consultores.com

**Miguel Jonás Bravo García**
Arte Consultores Tecnológicos S.L., ARTE
mbragar@arte-consultores.com

## Introducción

La Directiva de la Unión Europea relativa a la reutilización de la información del sector público especifica que los documentos elaborados por los organismos del sector público de los Estados miembros constituyen un conjunto amplio, diverso y valioso de recursos que pueden beneficiar a la economía del conocimiento. Esta Directiva se transpone en la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público.

Recientemente el Consejo de Europa ha adoptado una nueva normativa sobre los datos abiertos y la reutilización de la información del sector público. En esta nueva Directiva se introduce el concepto de conjuntos de **datos de alto valor**, que deberán facilitarse gratuitamente a través de una interfaz de programación de aplicaciones (API). El texto define seis categorías principales de conjuntos de datos de alto valor, e incluye a los **datos estadísticos** dentro de esta categoría. Asimismo, establece que los datos de investigación deben ser fáciles de encontrar, accesibles, interoperables y reutilizables (principios FAIR) y que los datos del sector público estarán disponibles de forma gratuita o a un coste muy

bajo.

El Decreto por el que se aprueba el Plan Estadístico de Canarias 2019-2022 (PEC-22), establece que durante su ejecución se impulsará la reutilización de los datos estadísticos de conformidad con la Ley sobre reutilización de la información del sector público. A su vez indica que la Infraestructura de Datos y Metadatos Estadísticos (**eDatos**) será el soporte para la difusión abierta e interoperable de los datos publicados por las operaciones del PEC-22, constituyéndose como el canal único para la difusión descentralizada de las estadísticas en las web corporativas del Gobierno de Canarias.

## Objetivos

Se pone de manifiesto la necesidad de crear un punto de acceso único a los conjuntos de datos publicados por organismos públicos, que además promueva la reutilización de los datos y dé lugar a un nuevo nicho de mercado para el sector infomediario.

En concreto se persiguen alcanzar los siguientes objetivos:

1. Disponer de un punto de acceso único a todos los recursos de la estadística pública canaria.
2. Disponer de una herramienta que permitiese llevar a cabo la gestión y ciclo de publicación de aquellos tipos de recursos que no contaban con una aplicación específica para tal fin.
3. Situar a la estadística pública canaria en el terreno internacional haciendo uso de los mecanismos de federación proveídos por los catálogos de datos.

## Metodología

Para cumplir con los objetivos señalados, desde el Instituto Canario de Estadística (ISTAC) se pone a disposición de la ciudadanía el portal de datos abiertos de la estadística pública en Canarias.

Entre las funcionalidades ofrecidas por dicho portal se sitúa el catálogo de datos abiertos. Este sistema agrupa los conjuntos de datos reutilizables generados por la actividad del ISTAC, permitiendo así el acceso a recursos que ofrecen información pública de alto impacto y valor.

A continuación iremos desgranando cómo se ha conseguido alcanzar cada uno de los objetivos propuestos anteriormente así como los retos tecnológicos que la implantación de este sistema ha supuesto.

**En busca de un punto de difusión único**

A día de hoy el ISTAC ya cuenta con toda una serie de herramientas de difusión que permiten que terceros (ciudadanos, empresas y/o reutilizadores) puedan consumir los recursos que elaboran y difunden. Sin embargo, cada una de esas herramientas está pensada para un tipo de recurso y, por lo tanto, difunde la información a través de un tipo de plataforma diferente. Por ejemplo, en el caso de los indicadores se difunden a través de widgets mientras que las tablas de datos lo hacen a través de un visualizador online

que permite la previsualización al vuelo de los conjuntos de datos pero también la descarga de los mismos. De esta forma, la infraestructura de Datos y Metadatos Estadísticos de Canarias (eDatos) a través de la cual se rige todo el ciclo de meta-documentación y difusión de los recursos cuenta a día de hoy con más de una veintena de sistemas entre los que encontramos API y sistemas de visualización / consulta de datos.

Partiendo de esta base, para determinados tipos de usuarios puede resultar complejo el buscar cuál es la información a la que desean acceder y/o cómo hacerlo. En este sentido, la implantación de un catálogo de datos abiertos facilita un punto de acceso único a todos los recursos del organismo. De esta manera, con independencia del que sea el sistema de gestión y/o visualización utilizado, el objetivo es llevar al catálogo todos los recursos.

No obstante, de cara a que el proyecto fuese sostenible y teniendo en cuenta el gran volumen de datos con el que se trabaja en una oficina estadística, se partí de la premisa de que la alimentación de los recursos en este catálogo no podía realizarse de forma manual puesto que esto supone grandes inconvenientes como:

- Consumiría demasiado tiempo.
- Se correría el riesgo de que los datos no estuviesen actualizado de forma puntual[1].
- Sería un proceso propenso a fallos.
- La calidad de los metadatos podría no ser adecuada. Incluso, se podría pecar de actualizar los datos pero no los metadatos.

Por tanto, partiendo de esta base, lo que se hizo fue comenzar a trabajar con el sistema de difusión de cubos estadísticos para la publicación de los mismos en el portal. La integración se realiza de tal forma que el sistema de difusión de cubos genera un fichero en formato RDF que sigue el vocabulario DCAT[2]. Al tratarse de un vocabulario diseñado para facilitar la interoperabilidad entre catálogos de datos publicados en la web, la generación de este fichero permite que cualquier otro sistema que implemente el vocabulario DCAT pueda consumir el mismo. En nuestro caso, será la propia aplicación del catálogo de datos quien se encargue de realizar el consumo de dicho documento RDF para, de esta forma, conseguir disponer de todos los conjuntos de datos actualizados de forma puntual y automática. Cabe mencionar también que el consumo de este fichero desde el catálogo de datos puede realizarse de forma manual o automática. Mediante el mecanismo automático, la carga está programada para que se realice una vez al día.

Este modelo de federación lleva funcionando más de un año y desde entonces se ha conseguido alimentar el catálogo de datos con más de siete mil conjuntos de datos que tienen también su correspondiente representación en el catálogo nacional (datos.gob.es).

Si bien es cierto que la puesta a disposición de los cubos de datos supone el principal volumen de los recursos generados el instituto, una vez superada esta primera fase se trabaja en la puesta a disposición de otro tipo de recursos. En este caso, se tiene en cuenta la experiencia de los recursos anteriores y se trata de mejorar la puesta a disposición. Si bien el mecanismo explicado anteriormente funciona adecuadamente, tiene el inconveniente de que dado que el catálogo expuesto en el fichero RDF es muy grande (más de 7000 conjuntos de datos) el tiempo que lleva su consumo, procesamiento y carga

[1] Principio 13. Oportunidad y puntualidad del código de buenas prácticas de las estadísticas europeas.

[2] DCAT permite a un editor describir conjuntos de datos y servicios de datos en un catálogo utilizando un modelo estándar y un vocabulario que facilita el consumo y la agregación de metadatos de varios catálogos.

también es elevado. Esto hace que la carga en el CKAN esté limitada a realizarse una vez al día.

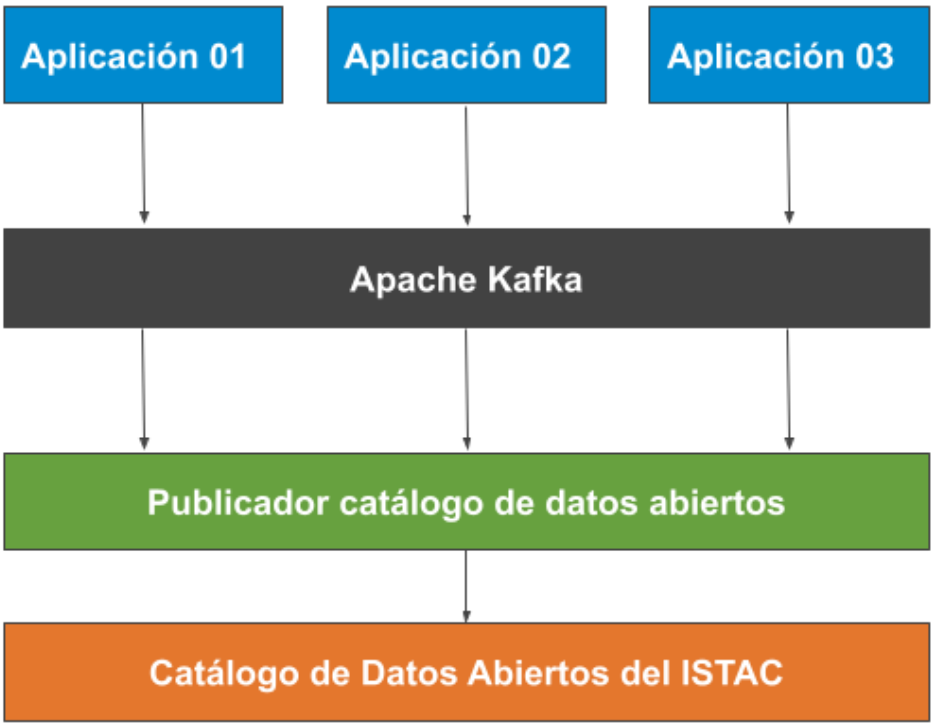
Llegados a este punto, de cara a cargar en el catálogo el resto de tipos de recursos, nos aprovechamos de que los sistemas de gestión que están detrás son más modernos y se opta por una alternativa diferente basada en la publicación en el catálogo mediante invocaciones en caliente a la API que ofrece la herramienta de catálogo. No obstante, de cara a evitar que todos los sistemas que gestionan recursos tengan que implementar las correspondientes invocaciones a las API, esta publicación se canaliza a través de un módulo central que conoce la anatomía de la API y sabe cómo proceder.
El funcionamiento general del sistema es el siguiente:

1. Los sistemas de gestión implementan un ciclo de vida en el que el estado final es "Publicado". A partir de este momento el recurso pasa a estar a disposición a través de diferentes canales como API o sistemas de visualización. Entre esos canales destaca un sistema denominado Apache Kafka que se encarga de notificar internamente que ese recurso ha pasado a ser público.
2. Apache Kafka se encarga de almacenar los mensajes sobre la publicación de los diferentes recursos para que aquellos sistemas interesados puedan consumir esos avisos.
3. Se crea un módulo de publicador de datos en el catálogo de datos abiertos. Este módulo actúa como consumidor de Kafka para conocer cuáles son los recursos sobre los que se debe notificar y, por cada uno de ellos, emite la correspondiente petición a la API del catálogo para dar de alta o actualizar el recurso en cuestión.

Es conveniente destacar que esta aproximación tiene la ventaja de que gracias al uso de Apache Kafka todo este procesamiento se lleva a cabo en tiempo real y, por lo tanto, se minimizan drásticamente los tiempos de puesta a disposición.

Todas estas soluciones están enfocadas en la filosofía seguida en todos los desarrollos del ISTAC de que el dato debe ser único. De esta forma, se publica una sola vez pero se reutiliza tantas como se desee a través de múltiples canales. Se persigue así llegar al número de usuarios más amplio posible y con las menores barreras posibles.

En la siguiente imagen puede verse el funcionamiento descrito para este tipo de carga en el catálogo de datos abiertos.




**Un sistema de gestión y publicación para determinados tipos de recursos.**

La infraestructura de Datos y Metadatos Estadísticos de Canarias (eDatos) cuenta no sólo con sistemas de reutilización y visualización sino también con sistemas que permiten llevar a cabo la elaboración, meta-documentación, ciclo de vida y publicación de los diferentes tipos de recursos.

De esta forma, dependiendo de la naturaleza del recurso que se trate (clasificación estadística, cubo de datos, indicador,...) existe un determinado tipo de sistema para la gestión y difusión del mismo. No obstante, existen determinados tipos de recursos para los que estos sistemas de difusión no han sido elaborados puesto que:

- Son recursos que tradicionalmente no habían sido publicados.
- Son recursos minoritarios.

Con la entrada en funcionamiento del catálogo de datos, se abre también la posibilidad de poder utilizar este sistema como gestor del ciclo de publicación de los recursos que hasta ahora no disponían de un sistema propio.

El catálogo de datos permite la carga manual de cualquier tipo de recurso así como la meta-documentación de los mismos. Por tanto, resulta el sistema idóneo para sin casi esfuerzo disponer de una herramienta para la gestión de estos recursos.

Utilizando este mecanismo ya son más de medio centenar de conjuntos de datos los que se encuentran gestionados a través de este sistema.

Algunos de los recursos que se han publicado mediante este mecanismo en el catálogo de datos abiertos son:

- Cartografía para uso estadístico.

    - Países del mundo.
    - Unidades Territoriales Estadísticas de la UE.
    - Islas Canarias.
    - Municipios de Canarias.
    - Comarcas y Grandes Comarcas de Canarias
    - Distritos y secciones de Canarias por anualidades.
    - Microdestinos turísticos de Canarias.
    - Malla estadística Eurostat adaptada a Canarias.
    - Directorio de Alojamientos Turísticos Colectivos de Canarias (ALOJATUR).
    - Directorio de aeropuertos.
- Datos estadísticos por mallas.
    - Indicadores demográficos para la malla estadística de Canarias 250 m de lado por anualidades.
    - Indicadores de afiliación a la Seguridad Social para la malla estadística de 250 m de lado por anualidades.

Conviene destacar que para estos tipos de recursos la carga no siempre tiene que ser complementamente manual sino que también se pueden llevar a cabo mecanismos que simplifiquen la publicación. En el caso en concreto del ISTAC se ha optado por automatizar mediante ETL la publicación en el portal de toda la cartografía.

**La federación como mecanismo clave para visualizar los datos.**

Uno de los objetivos que se quería lograr durante este proyecto era conseguir situar a la estadística pública canaria en el territorio internacional. En este sentido, contar con un catálogo de datos abiertos resulta especialmente ventajoso puesto que permite llevar a cabo mecanismos de federación mediante los cuales los recursos que se publican en un catálogo de datos pueden ser servidos también por otros catálogos de datos.

En el caso de Canarias, la Comunidad Autónoma ha trabajado[3] en el desarrollo de un Catálogo de Datos Abiertos de Canarias que, lejos de ser únicamente para la publicación de datos del gobierno autonómico, sirva como punto federado para la localización y reutilización de cualquier conjunto de datos de las administraciones públicas canarias. Cabe destacar que este catálogo de datos está a su vez federado con el catálogo nacional.

Por lo tanto, una vez completado el desarrollo del catálogo de datos abiertos del ISTAC, el siguiente paso fue la federación en el Portal de Datos Abiertos de Canarias. Alcanzado este primer paso, el objetivo inicial también estaba cumplido puesto que a través de la federación se consigue que los conjuntos de datos puedan consultarse a través del catálogo de datos abiertos de la Administración General del Estado o del portal oficial de datos europeos.

De esta forma, durante el pasado mes de enero de 2020, el ISTAC se convertía en el catálogo de datos abiertos que más conjuntos de datos aportaba al catálogo nacional (más del 20% de los conjuntos de datos publicados en el catálogo nacional correspondían a este organismo).

[3] En colaboración con el ISTAC.

De cara a poder llevar a cabo correctamente la federación, lo más importante es que debemos tener cumplimentados los metadatos de forma correcta siguiendo las indicaciones tanto de DCAT como de la NTI de Reutilización de recursos de información y su guía de aplicación.

**Base tecnológica.**

A nivel tecnológico se apostó por hacer uso del software CKAN para llevar a cabo la implementación del catálogo de datos abiertos.

CKAN es una plataforma de código abierto desarrollado por la Open Knowledge Foundation (OKF) que se ha convertido en un referente a nivel mundial en lo relativo a programas de apertura de datos. El éxito de esta plataforma reside en su código abierto, gratuito y libre que brinda una implementación rápida. Cabe destacar que más de cuarenta entidades a escala internacional han confiado en CKAN como soporte para sus portales Open Data.

Además, CKAN cuenta de forma nativa con determinado tipo de funcionalidades que se planteaban como claves para la elección del software. Estás son:

- Interfaz sencilla para la carga manual de recursos.
- Diferentes tipos de roles.
- API para carga de recursos de forma automática.
- Mecanismo de federación basado en DCAT para federar a terceros.
- Mecanismo de federación basado en DCAT para exponer nuestro catálogo y que terceros puedan federarnos.
- Posibilidad de ampliar la funcionalidad en base a extensiones personalizadas.

Por lo tanto, la elección de este software resultó sencilla y natural puesto que se ajustaba a los requerimientos y necesidades del ISTAC.

Partiendo de CKAN fue necesario llevar a cabo algunos tipos de mejoras que permitiesen cumplir al 100% con las necesidades planteadas. En este sentido, las mejoras iban encaminadas principalmente a dar respuesta a cuestiones relacionadas con la autenticación y autorización de usuarios así como a la identidad corporativa del catálogo.

**Autenticación y autorización**

Para el registro de los usuarios, CKAN dispone de un mecanismo de auto-registro mediante el cual, cualquier usuario que lo desee puede registrarse en la plataforma con un rol de “miembro”. De la misma manera, CKAN dispone de un mecanismo propio para la autenticación de los usuarios. Sin embargo, en un entorno corporativo y en aras de cumplir tanto la normativa propia como el Esquema Nacional de Seguridad, ninguna de estas dos opciones resultaban satisfactorias.

Por este motivo, se decidió llevar a cabo el desarrollo de una extensión propia que permitiese delegar las labores de autenticación en el CAS corporativo y que además, aprovechara la información devuelta por este sistema para registrar los usuarios en el CKAN (si no lo estaban ya) así como asignarles los roles correspondientes.

De esta forma, se sustituyen los mecanismos propios de CKAN de autenticación y

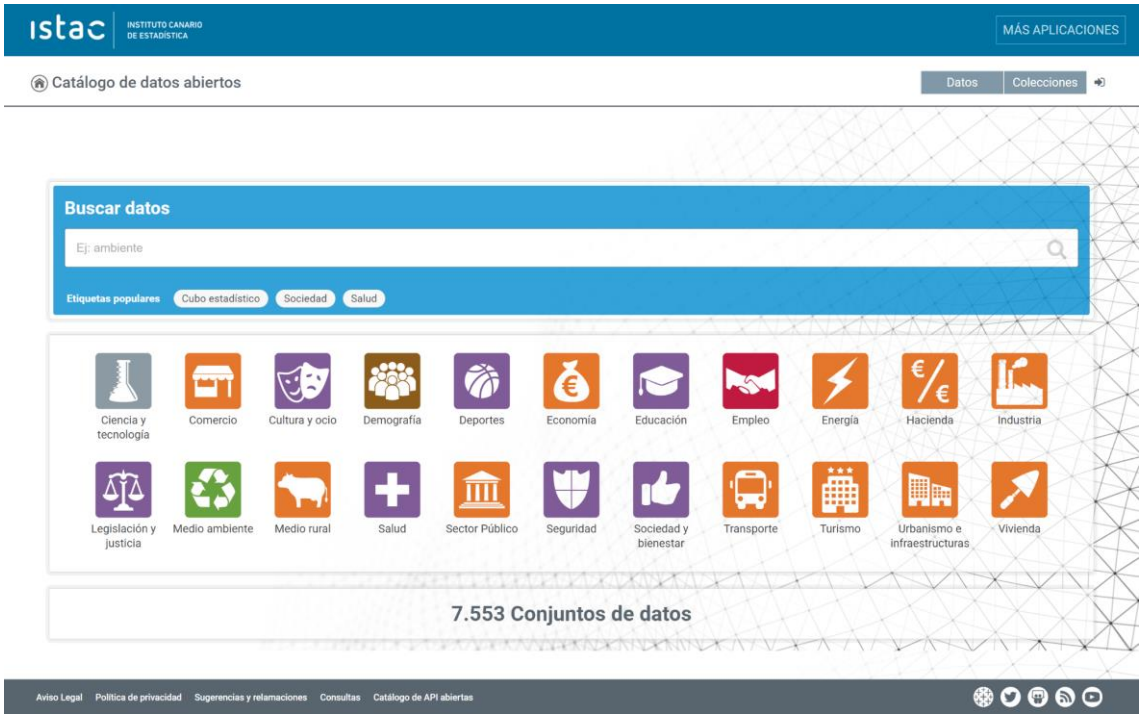
autorización por los existentes a nivel corporativo.

**Personalización catálogo**
Otro de los aspectos que resulta importante gestionar en gran parte de los sistemas institucionales es que cumpla con unos requisitos en lo que a la imagen corporativa se refiere.
En este sentido, se desarrolló una extensión para permitir adaptar la imagen del CKAN a la del resto de las aplicaciones del ISTAC. Es importante tener en cuenta que esta extensión no sólo nos sirve para modificar la estética sino que también permite llevar a cabo otro tipo de modificaciones interesantes:

- Eliminar opciones de la interfaz gráfica (por ejemplo la opción de auto-registro).
- Modificación de traducciones para adaptar el lenguaje a los términos usados en el resto de sistemas.
- Modificar la estructura de las páginas.

Mediante la inclusión de estas modificaciones se consigue disponer de un catálogo completamente adaptado a las necesidades del organismo.

En la siguiente imagen puede verse la estética que presenta actualmente la página de inicio del catálogo de datos abiertos del ISTAC.



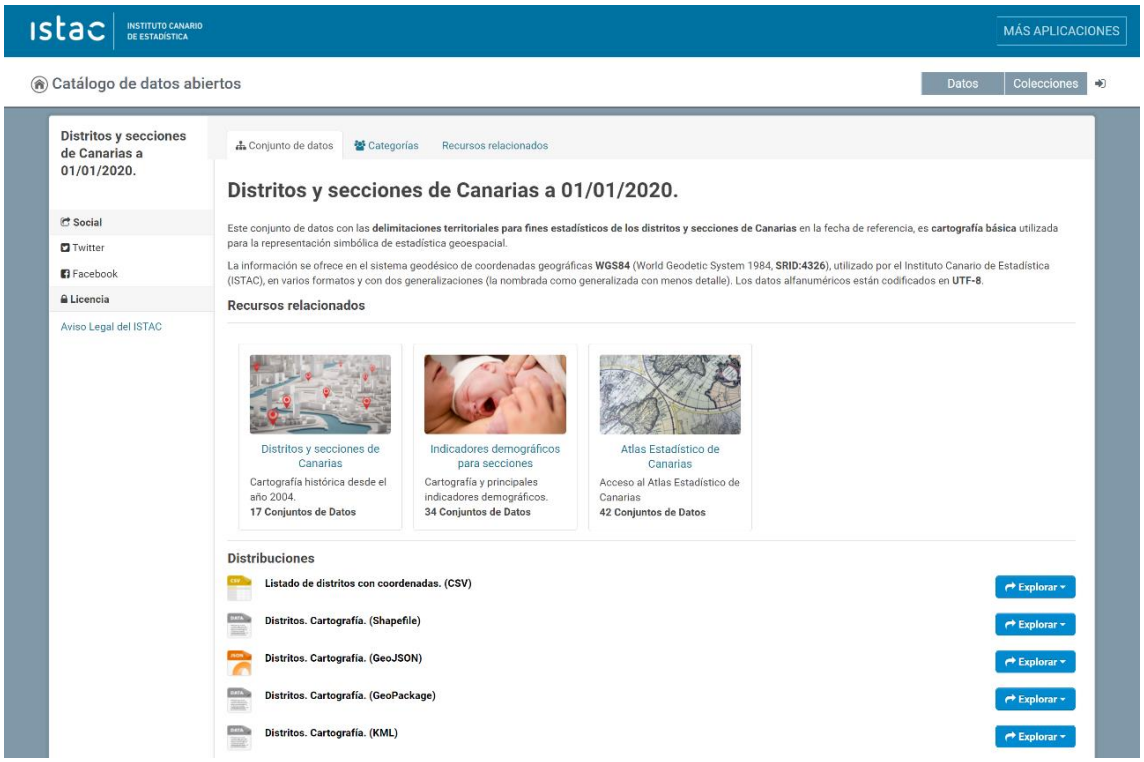

**Agrupación de conjuntos de datos**.
Cuando se maneja un gran volumen de información también comienzan a aparecer algunos problemas relacionados con la búsqueda efectiva de los recursos que se necesitan.

Para ayudar al usuario final en esta labor, se llevó a cabo también la incorporación de una extensión que permitiese dotar al CKAN de una nueva funcionalidad denominada “Colecciones” mediante la que se pueden agrupar conjuntos de datos similares.

De esta forma, cuando un usuario accede a un conjunto de datos se le muestran todas aquellas colecciones con las que se relaciona dicho conjunto de datos y que podrían resultar de interés.

Se implementa así un sistema nuevo de navegación basado en recomendaciones según similitudes de los conjuntos de datos.

En la siguiente imagen puede verse los recursos que se añaden como relacionados al consultar un conjunto de datos en concreto.




Por último, cabe mencionar que esta extensión está basada en una extensión ya existente y desarrollada por la comunidad CKAN denominada Showcase.

## Resultados

Durante esta anualidad el ISTAC ha alcanzado un nuevo hito en su apuesta por la apertura de datos consiguiendo publicar a través de su catálogo de datos abiertos más de 7.500 datasets. De esta forma, el catálogo de datos abiertos de la estadística pública canaria se sitúa como el catálogo de datos autonómico que más conjuntos de datos. No obstante, es importante destacar que no sólo se trata de una cuestión de volumen sino también de calidad. Todos estos conjuntos de datos son actualizados de forma puntual y sincronizada con el resto de sistemas de difusión del ISTAC.

Manejar un alto volumen de datos ha supuesto un gran reto a nivel de organización y gestión de la información que han desembocado en diversas líneas de trabajo.

El buen hacer por parte del ISTAC ha sido puesto de manifiesto en diferentes ocasiones durante el último año y muestra de ello son por ejemplo los siguientes artículos del portal de datos abiertos estatal.

- Datos Abiertos y Objetivos de Desarrollo Sostenible
- Los conjuntos de datos más demandados publicados por entidades locales

## Conclusión

El catálogo de datos abiertos del ISTAC se ha federado con el catálogo estatal y esto hace que sea la comunidad autónoma que más datos aporta al mismo. Por tanto, se ha conseguido visibilizar los datos de la estadística pública canaria dentro del espacio europeo.

Se ha conseguido disponer de un ecosistema sostenible y donde, a pesar del gran volumen de información, se es capaz de difundir de forma puntual y accesible. Con la calidad siempre presente se dispone de un nuevo sistema para la difusión de la información pero respetando la premisa del dato único.

## Principales referencias bibliográficas


[1] Portal de datos abiertos de la Estadística Pública Canaria - http://www.gobiernodecanarias.org/istac/datos-abiertos/

[2] Catálogo de API abiertas de la Estadística Pública Canaria - https://datos.canarias.es/api/estadisticas/

[3] Catálogo de datos abiertos de la Estadística Pública Canaria - https://datos.canarias.es/catalogos/estadisticas/.

[4] Catálogo de datos abiertos de la Administración Pública del Estado - https://datos.gob.es/es/catalogo

[5] Catálogo de datos abieros europeo - https://data.europa.eu/data/datasets

[6] NTI de reutilización de recursos de información - http://www.boe.es/boe/dias/2013/03/04/pdfs/BOE-A-2013-2380.pdf

[7] Reutilización de recursos de información. Guía de aplicación de la Norma Técnica de Interoperabilidad - https://datos.gob.es/sites/default/files/20160726_guia_de_aplicacion_de_la_nti_reutilizacion_recursos_de_informacion_l.pdf

[8] Vídeo explicativo relativo a eDatos - https://edatos.io/#plataforma

[9] Código de buenas prácticas de las estadísticas europeas - https://ec.europa.eu/eurostat/documents/4031688/9394048/KS-02-18-142-ES-N.pdf/e792b761-6f09-42a9-a1e0-3a3356a0de1c.

[10] DCAT - https://www.w3.org/TR/vocab-dcat-2/ .

[11] Apache Kafka - https://www.confluent.io/what-is-apache-kafka/

[12] CKAN - https://ckan.org/

[13] Showcase. CKAN extension - https://extensions.ckan.org/extension/showcase/

[14] Píldora formativa: Buscar en la API del catálogo de datos según etiquetas de los datasets - https://youtu.be/Dt7xNNTo23s

[15] Píldora formativa: ¿Cómo puedo reutilizar los datos de los Objetivos de Desarrollo Sostenible? - https://youtu.be/V6nN4j9NUME

[16] Datos Abiertos y Objetivos de Desarrollo Sostenible - https://datos.gob.es/es/blog/datos-abiertos-y-objetivos-de-desarrollo-sostenible

[17] Los conjuntos de datos más demandados publicados por entidades locales - https://datos.gob.es/es/noticia/los-conjuntos-de-datos-mas-demandados-publicados-por-entidades-locales