Revista Cartográfica 103 | julio-diciembre 2021 | Artículos
ISSN (impresa) 0080-2085 | ISSN (en línea) 2663-3981
DOI: https://doi.org/10.35424/rcarto.i103.1026

Revisión de la calidad de los conjuntos de datos abiertos sobre presupuestos

Quality Review of Budget Open Datasets

Manuel Antonio Ureña-Cámara[1]
Javier Nogueras-Iso [2]
Javier Lacasta[3]

Recibido 14 de marzo de 2021; aceptado 22 de mayo de 2021

Resumen

En este trabajo se presentan los resultados de la evaluación de la calidad de los conjuntos de datos abiertos sobre presupuestos disponibles en España. Para llevar a cabo la comparativa de evaluación se ha adoptado la Metodología de Evaluación de la Calidad de los Metadatos propuesta por el Portal de Datos Europeo (MQA). Se ha adaptado una metodología automática que aplica las cinco dimensiones de MQA separadas por la propiedad espacial y que es capaz de generar gráficas de descripción del conjunto de metadatos y otras gráficas comparativas siguiendo el ejemplo del ranking existente en el portal de MQA. Los resultados indican que, a pesar de las diferentes entidades que elaboran los metadatos, todos ellos alcanzan una puntuación similar limitada únicamente por la norma que define el diseño del portal de datos abiertos en España.

Palabras clave: *datos abiertos, Evaluación de Calidad, metadatos, presupuestos, MQA.*

Abstract

In this work, we present the results of quality evaluation of budget open datasets in Spain. To achieve this quality evaluation we have applied the Metadata Quality Assurance (MQA) methodology proposed to check the European open

[1] Universidad de Jaén, España, correo electrónico: maurena@ujaen.es.
ORCID: https://orcid.org/0000-0002-6373-4410

[2] Universidad de Zaragoza, España, correo electrónico: jnog@unizar.es.
ORCID: https://orcid.org/0000-0002-1279-0367

[3] Universidad de Zaragoza, España, correo electrónico: jlacasta@unizar.es.
ORCID: https://orcid.org/0000-0003-3071-5819

data portal. Following this, a methodology to test the five dimensions of MQA grouped by the spatial property has been developed. In addition, an automatic procedure to create comparative graphs, the first describing the spatial property of the corpus and the second following the MQA ranking. The results show that, even with some different dataset (and metadata) producers, the MQA value is similar in all the cases and are mainly limited by the policy that defines the design of open data portal in Spain.

Key words: *Open data, Quality Control, Metadata, Budget, MQA.*

1. Introducción

Durante los últimos años, las políticas de transparencia impulsadas por los distintos gobiernos, tanto a nivel local, regional, nacional o internacional, han promovido la creación de portales de datos abiertos donde se han ido facilitando datos de interés del sector público para diferentes propósitos tratando de alcanzar acuerdos desarrollados durante la década de 2000. Esta promesa se inició en el año 2008 y en su décimo aniversario ya son al menos 30 gobiernos (Web Foundation, 2018) los que han firmado la Carta de Datos Abiertos,[1] cuyos principios se alinean con algunas de los aspectos fundamentales de la calidad como la accesibilidad o la interoperabilidad, todo ello dentro de una gobernanza más cercana al ciudadano. Estos portales de datos abiertos se construyen sobre la base de catálogos de Datos Abiertos que exponen los metadatos (descripciones) de los conjuntos de datos (*dataset*) que se quieren hacer públicos utilizando estándares y especificaciones interoperables que acceden a un recurso (Neumaier *et al.*, 2016).

Desde el punto de vista de los ciudadanos, uno de los tipos de conjuntos de datos abiertos que despierta mayor interés es conocer con el mayor detalle posible los presupuestos públicos y cómo se ejecutan estos presupuestos con el fin de facilitar la transparencia y gobernanza (Kučera *et al.*, 2013). Por ello, aunque no sea un aspecto directamente contemplado por el Horizonte 2030, sí que permite comprobar que las diversas partes de administraciones y empresas que publican sus datos en abierto cumplen con uno o varios objetivos en esa línea. Eso sí, siempre que los metadatos cumplan con los estándares de calidad y que los datos cumplan con la información de los metadatos existentes.

El objetivo de este artículo es conocer la disponibilidad y calidad de los datos sobre presupuesto que se encuentran disponibles en España a través de su portal nacional de datos abiertos, https://datos.gob.es. Además, de realizar un análisis global de los datos de esta temática, también se pretende hacer un estudio comparativo de la calidad en las distintas comunidades autónomas y provincias. A pesar de que existen diversas metodologías para el control de calidad de los

[1] The International Open Data Charter: https://opendatacharter.net/

portales abiertos (Kubler *et al.*, 2018; Neumaier *et al.*, 2016; Nogueras-Iso *et al.*, 2021), para el propósito de este trabajo la evaluación de la calidad va a utilizar la metodología de Evaluación de la Calidad de los Metadatos (Unión Europea, 2020), también conocida por su nombre en inglés *Metadata Quality Assurance* (MQA). La metodología MQA es la base del cuadro de mandos que se ha integrado dentro del Portal Europeo de Datos para monitorizar los contenidos recolectados de los diferentes catálogos de datos abiertos (o portales de datos abiertos) que contribuyen al portal europeo. La metodología MQA está inspirada en los principios FAIR (Wilkinson *et al.*, 2016), los cuales proporcionan guías para mejorar la facilidad de localización (*findability*), la accesibilidad (*accessibility*), la interoperabilidad (*interoperability*), y la reutilización de recursos digitales. En particular, MQA propone 23 métricas distribuidas en cinco dimensiones que chequean el contenido de los metadatos que describen los conjuntos de datos publicados que serán desarrolladas en la sección 2.

El resto del artículo se estructura de la siguiente forma, la sección 2 describirá brevemente el estado del arte acerca de los portales de datos abiertos y de su control de calidad. La sección 3 describirá la metodología aplicada para el control de calidad de datos. La sección 4 explicará el *corpus* usado en el trabajo y los resultados obtenidos. La sección 5 dará unas conclusiones relativas al control realizado y la metodología aplicada.

2. Estado del arte sobre calidad de los portales de datos abiertos

Los datos abiertos son una tendencia actual desarrollada como parte de los procesos de transparencia y difusión de la información en todos los ámbitos que en la actualidad se está incrementando debido a la transformación digital (Oudkerk *et al.*, 2020). Desde el punto de vista de cualquier administración, los portales de datos abiertos son un método sencillo para la distribución y visualización de la información disponible de forma pública, accesible y transparente para los diversos actores, tanto otras entidades gubernamentales como las personas gobernadas. Por otra parte, este tipo de portales también permite que las empresas publiciten su información permitiendo tanto su reutilización como facilitar la difusión a las empresas.

Desde este punto de vista, se ha potenciado un desarrollo casi paralelo entre los portales de datos abiertos en Europa y América, tanto América del Norte como América Latina, consiguiendo, por tanto, una expansión en todos los países como un método barato de cumplir con los aspectos de transparencia que son de interés para diversos ámbitos y aspectos de la gobernanza. Según Sandoval (2019), existen más de 2 600 portales de datos abiertos disponibles a lo largo del planeta como puede verse en la web de *OpenDataSoft*.[2] De entre ellos destacan iniciativas como el *Open Government* de Estados Unidos[3] que muestra su capacidad de publicación con diversos portales del mundo alcanzando en la actualidad a más

[2] https://opendatainception.io/

[3] https://www.data.gov/open-gov/

de 300 catálogos distribuidos por todo el planeta (Figura 1). Asimismo podemos destacar iniciativas como el portal de datos abiertos del Reino Unido[4] que ha sido uno de los de mayor recorrido y el portal de datos abiertos de España[5] con diversos proyectos activos y un elevado número de conjuntos de datos.

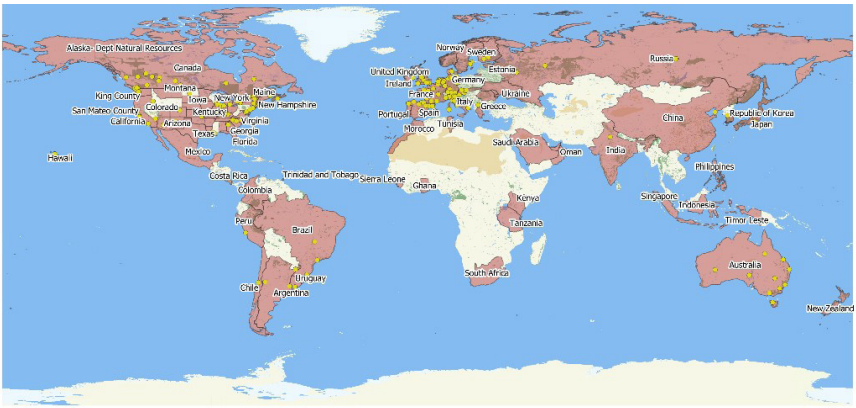

Figura 1. Portales internacionales y de diverso ámbito, accesibles y consultables desde el portal de datos abiertos del Gobierno de Estados Unidos.
Fuente: elaborado a partir de la información disponible en la web del Open Government sobre portales de datos abiertos. Mapa base: Heigit (WMS de OSM: https://www.osm-wms.de/).

Dentro del ámbito de Latinoamérica también hay un elevado interés en los portales de datos abiertos lo que se traduce en un número importante de portales auspiciados por los gobiernos de los diferentes países con el fin de aumentar la transparencia gubernamental. En la Tabla 1 se muestran algunos de ellos y otros adicionales pueden ser encontrados en Sandoval (2019).

Una de las evaluaciones de los portales de datos abiertos de gran relevancia es la llevada a cabo por la Fundación de Datos Abiertos que es parte de la Web Foundation.[6] En ella, de manera recurrente, se evalúan en diferentes portales de todo el planeta dimensiones relativas a la utilidad de los conjuntos de datos integrados, es decir, la preparación, implementación o impacto de los datos abiertos (Web Foundation, 2018). Pese a que un gran número de las iniciativas de datos abiertas evaluadas en este informe se encuentran fuera de América Latina, se puede observar como dentro de los países que más han mejorado desde el primer informe hasta el actual (Figura 2), 3 de los 6,

4 https://data.gov.uk/

5 https://datos.gob.es/

6 http://webfoundation.org/

Tabla 1. Algunos de los portales de datos abiertos en Latinoamérica

País	*URL del Portal de Datos Abiertos*	*API*	*Punto SPARQL*	*Otros requisitos*	*Puntuación informe Web Foundation 2018*
Argentina	http://datos.gob.ar/	CKAN + API Georreferenciación			47
Brasil	https://dados.gov.br/	CKAN			50
Chile	http://datos.gob.cl/	CKAN			40
Colombia	https://www.datos.gov.co/	Socrata			52
Ecuador	http://data.utpl.edu.ec/	Desconocida	Si		No evaluado
México	https://datos.gob.mx/	Rest Get / CKAN			69
Paraguay	https://www.datos.gov.py/	DKAN	No	Consultas sólo de usuarios registrados	34
Perú	https://www.datosabiertos.gob.pe/	Desconocida			No evaluado
Uruguay	https://catalogodatos.gub.uy/	CKAN			56
Venezuela	http://datos.gob.ve/	Desconocida			No evaluado

Nota: No se incluyen Costa Rica que está cambiando de portal por lo que está desactivado (comprobado en 2021).
Fuente: elaboración propia salvo las puntuaciones extraídas del ranking de Web Foundation, 2018.

pertenecen a esta región, concretamente México, Colombia y Uruguay. Esto denota tanto un gran interés en América Latina en apoyar esta tecnología, como la capacidad de alcanzar las mejores puntuaciones que ostentan países con un mayor desarrollo de este tipo de iniciativas como Canadá o Reino Unido.

A pesar de que los procesos de control de calidad propuestos por la Web Foundation, evalúan el impacto de los portales de datos abiertos en diversos aspectos del país como difusión, reutilización, etc., por sectores productivos, ninguno de ellos sería posible si no se realiza una revisión de los campos técnicos y semánticos que soportan al propio catálogo de datos abiertos. Por esta razón, metodologías como la MQA indicada en la Introducción son un paso obligatorio, y previo, para el análisis de la calidad de un catálogo.

MQA chequea el contenido de los metadatos de cada conjunto de datos utilizando 23 métricas distribuidas en cinco dimensiones:

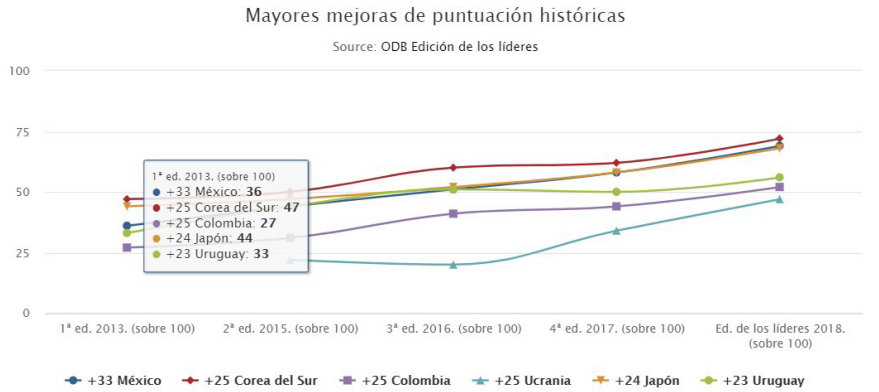

Figura 2. Mejores evoluciones en Portales de Datos Abiertos.
Fuente: Web Foundation, 2018.

- Facilidad de localización: se comprueba la disponibilidad de palabras clave, categorías, cobertura espacial y cobertura temporal.
- Accesibilidad: se comprueba la disponibilidad de URL de descarga, y el estado de accesibilidad de las URL de acceso y descarga.
- Interoperabilidad: se comprueba la conformidad de los metadatos respecto a la especificación DCAT-AP, la disponibilidad información sobre el formato de distribución, la utilización de formatos bien conocidos, la utilización de formatos no propietarios y la utilización de formatos procesables.
- Reusabilidad: se comprueba la disponibilidad de información sobre licencias, restricciones de acceso, puntos de contactos y editores. También se comprueba la utilización de licencias y restricciones de acceso internacionalmente reconocidas.
- Contextualidad: se comprueba la disponibilidad de información relativa a los derechos de distribución, el tamaño de las distribuciones, y las fechas de creación y distribución de los conjuntos de datos.

En la Tabla 2 se puede ver un resumen de esas métricas, así como la puntación máxima asociada a cada una. El objetivo final de MQA es proporcionar una puntuación de agrupaciones de uno o varios conjuntos de datos a través de sus metadatos. MQA establece una clasificación de la calidad de los metadatos en cuatro categorías en función de su puntuación: excelente (entre 351 y 405 puntos), buena (entre 221 y 350 puntos), suficiente (entre 121 y 220 puntos) y mala (entre 0 y 120 puntos).

Tabla 2. Métricas de la metodología MQA.

Dimensión	*Métrica*	*Puntos*
Facilidad de localización	Palabra clave disponible (Dataset/keyword)	30
	Categoría disponible (Dataset/theme)0	3
	Cobertura espacial disponible (Dataset/spatial)	20
	Cobertura temporal disponible (Dataset/temporal)	20
Accesibilidad	URL de acceso activa (Distribution/accessURL)	50
	URL de descarga disponible (Distribution/downloadURL)	20
	URL de descarga activa (Distribution/downloadURL)	30
Interoperabilidad	Formato disponible (Distribution/format)	20
	Información de formato MIME disponible (Distribution/mediaType)	10
	Utilización de formato conocido (Distribution/format o Distribution/mediaType)	10
	Utilización de formato no propietario (Distribution/format o Distribution/mediaType)	20
	Utilización de formato procesable (Distribution/format o Distribution/mediaType)	20
	Conformidad con especificación DCAT-AP (todas las entidades y propiedades)	30
Reusabilidad	Licencia disponible (Distribution/license)	20
	Utilización de licencia conocida (Distribution/license)	10
	Información sobre restricciones de acceso disponible (Dataset/accessRights)	10
	Utilización de tipo de restricciones de acceso conocido (Dataset/accessRights)	5
	Punto de contacto disponible (Dataset/contactPoint)	20
	Editor disponible (Dataset/publisher)	10
Contextualidad	Información sobre derechos de distribución disponible (Distribution/rights)	5
	Tamaño de la distribución disponible (Distribution/byteSize)	5
	Fecha de creación disponible (Dataset/issued o Distribution/issued)	5
	Fecha de modificación disponible (Dataset/modified o Distribution/modified)	5
	Total	*405*

Notas: se indica entre paréntesis la entidad y propiedad sobre la que se chequea

3. Metodología

Como se ha indicado antes, en este trabajo proponemos, y comprobamos la utilidad, un proceso para revisar la calidad de los conjuntos de datos a través de sus metadatos que consta de los siguientes pasos:

1. Análisis de la fuente de datos y sus mecanismos de acceso
2. Identificación de los conjuntos de datos de interés
3. Aplicación de MQA
4. Análisis comparativo de la calidad

Cada uno de los pasos anteriores se desarrollará en las posteriores subsecciones.

3.1. Análisis de la fuente de datos y sus mecanismos de acceso

Como fuente de obtención de los conjuntos de datos para evaluar según el MQA se ha seleccionado el portal de datos abiertos de España (https://datos.gob.es/). El portal de datos abiertos utiliza como esquema de metadatos el modelo propuesto en el Anexo III de la Norma Técnica de Interoperabilidad de Reutilización de recursos de la información (MINHAP, 2013).

Este esquema de metadatos está basado en el vocabulario DCAT (W3C, 2020). DCAT es el estándar de metadatos "de facto" en el contexto de los datos abiertos. DCAT es el acrónimo de *W3C's Data Catalog Vocabulary*, una recomendación de W3C para describir datos abiertos, desarrollado a través de un perfil de metadatos de *Dublin Core* basado en vocabularios RDF que se ha diseñado para facilitar la interoperabilidad entre catálogos de datos publicados en la web.

La Figura 3 muestra un diagrama UML con las entidades y propiedades incluidas en el esquema de metadatos español para datos abiertos. Como se puede observar los vocabularios basados en DCAT se centran en proporcionar información acerca de tres entidades principales: catálogos (*Catalog*), conjuntos de datos (*Dataset*) y distribuciones (*Distribution*). Las propiedades de un catálogo informan acerca de la organización a cargo de la publicación de conjuntos de datos abiertos. Por otra parte, las propiedades de un conjunto de datos proporcionan la información principal para su búsqueda y caracterización. Finalmente, las propiedades de una distribución se centran fundamentalmente en informar de los mecanismos para solicitar o descargar los conjuntos de datos.

Los portales de datos abiertos suelen implementarse mediante soluciones software como CKAN.[7] CKAN es la plataforma *Open Source* más utilizada para dar soporte a portales de datos abiertos e incluye los complementos necesarios para poder intercambiar metadatos basados en RDF, como es el caso de DCAT y del esquema de metadatos propuesto para el contexto español.

[7] https://ckan.org/

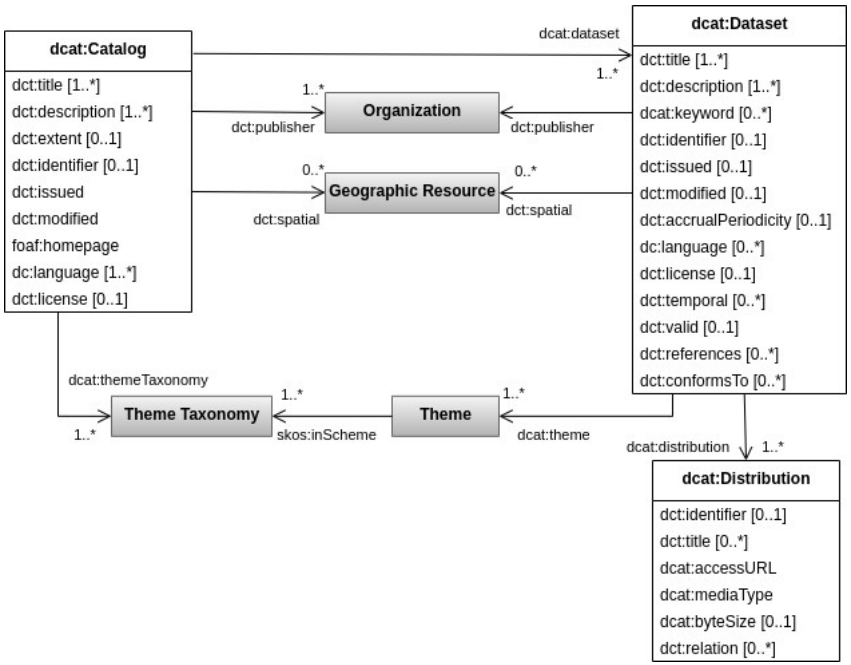

Figura 3. Entidades y propiedades del modelo de metadatos de https://datos.gob.es
Fuente: MINHAP, 2013.

En el caso del portal de datos abiertos del gobierno de España, se cuenta con un almacén de tripletas RDF donde se recolectan los metadatos que se han ido creando en portales datos abiertos de nivel regional y local. Estos portales que se han suscrito al portal nacional utilizan en la mayoría de los casos el software CKAN para generar los metadatos RDF propuestos en el esquema de metadatos nacional.

Respecto a las interfaces de acceso, el portal https://datos.gob.es, es accesible tanto a través de una interfaz (API)[8] diseñada de forma específica para el portal como a través de un punto SPARQL.[9] Por un lado, mediante una API REST, la interfaz específica proporciona operaciones concretas para filtrar metadatos de conjuntos de datos que cumplen ciertos criterios y descargarlos. Por otro lado, el punto SPARQL proporciona una interfaz de servicio que permite procesar consultas en lenguaje SPARQL para consultar los grafos RDF de los metadatos.

Para el propósito de este trabajo hemos utilizado ambas interfaces de acceso. Como se verá en la siguiente subsección, el punto SPARQL permite

[8] https://datos.gob.es/es/accessible-apidata

[9] https://datos.gob.es/es/accessible-sparq

una mayor flexibilidad para realizar filtros complejos de acuerdo a distintos criterios de los conjuntos de datos que se quieren recuperar. Una vez localizado el identificador o URI de un conjunto de datos de interés, la interfaz específica permite la descarga directa de los metadatos en el formato deseado.

3.2. Identificación de los conjuntos de datos de interés

De entre todos los conjuntos de datos existentes en el portal de datos abiertos, se han seleccionado los relativos a "presupuestos" por ser los que tienen una mayor influencia en la evolución de un país y para su población. En principio, para filtrar los conjuntos de datos relativos a los presupuestos tendría sentido estudiar el contenido de las propiedades relacionadas con la categoría del conjunto de datos (*dcat:theme*) o su palabra clave (*dcat:keyword*).

La utilización de la categoría parecía prometedora porque el esquema de metadatos obliga a utilizar un vocabulario seleccionado para las categorías (https://datos.gob.es/kos/sector-publico/sector) donde se puede acceder a los presupuestos a través de la categoría de "Sector Público". Sin embargo, tras realizar alguna consulta exploratoria a través del punto SPARQL, se descubrió que la categoría era poco discriminante para filtrar conjuntos de datos sobre presupuestos, ya que pueden haber sido anotados en distintos sectores como "Sector público", "Economía" o "Hacienda" y junto a otros conjuntos de datos que no están demasiado relacionados con presupuestos.

Por tanto, se optó por filtrar finalmente aquellos conjuntos de datos que contuviesen la palabra clave "presupuesto" en cualquiera de los idiomas oficiales de España. En la Figura 4 se muestra la consulta SPARQL utilizada para identificar las URI de todos los conjuntos de datos relacionados con presupuesto.

Además, dentro de este trabajo, como aspecto adicional, se propone analizar de forma separada la calidad de los conjuntos de datos disponibles en cada unidad administrativa para que, siendo entidades generadoras diferentes comprobar si el resultado de cumplimiento es similar. La Figura 5 muestra la consulta SPARQL que permite filtrar las URI de los conjuntos de datos relativos a presupuestos de una unidad administrativa concreta. Como también se puede dar el caso de que existan conjuntos de datos no asignados a ninguna unidad administrativa, la Figura 6 muestra la consulta SPARQL que permite filtrar los conjuntos de datos no asignados espacialmente.

Finalmente, para la descarga de los metadatos de un conjunto de datos previamente filtrado se ha hecho uso del API que permite descargar sus metadatos RDF en formato Turtle[10] si se conoce su identificador (o URI) a través de la siguiente petición GET: https://datos.gob.es/apidata/catalog/dataset/<ID>.ttl

10 https://www.w3.org/TR/turtle/

```
PREFIX dct:<http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
SELECT DISTINCT ?s WHERE {
    {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'aurrekontua', 'i')
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'presupuesto', 'i')
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'pressuposto', 'i')
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'orzamento', 'i')
    }
}
```

Figura 4. Consulta SPARQL para la descarga de conjuntos de datos que incluyen la palabra "presupuesto" en los idiomas oficiales en España.

```
PREFIX dct:<http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
SELECT DISTINCT ?s WHERE {
    {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'aurrekontua', 'i') .
        ?s      dct:spatial      <http://datos.gob.es/recurso/sector-
        publico/territorio/UNIDAD>
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'presupuesto', 'i') .
        ?s      dct:spatial      <http://datos.gob.es/recurso/sector-
        publico/territorio/UNIDAD>
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'pressuposto', 'i') .
        ?s      dct:spatial      <http://datos.gob.es/recurso/sector-
        publico/territorio/UNIDAD>
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'orzamento', 'i') .
        ?s      dct:spatial      <http://datos.gob.es/recurso/sector-
        publico/territorio/UNIDAD>
    }
}
```

Figura 5. Consulta SPARQL para la descarga de conjuntos de datos relativos a presupuesto de una unidad administrativa concreta (UNIDAD representa la entidad administrativa asignada).

```
PREFIX dct:<http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
SELECT DISTINCT ?s WHERE {
    {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'aurrekontua', 'i') .
        FILTER NOT EXISTS { ?s dct:spatial ?location }
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'presupuesto', 'i') .
        FILTER NOT EXISTS { ?s dct:spatial ?location }
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'pressuposto', 'i') .
        FILTER NOT EXISTS { ?s dct:spatial ?location }
    } UNION {
        ?s a dcat:Dataset .
        ?s dcat:keyword ?value .
        FILTER regex(str(?value), 'orzamento', 'i') .
        FILTER NOT EXISTS { ?s dct:spatial ?location }
    }

}
```

Figura 6. Consulta SPARQL para la descarga de conjuntos de datos relativos a presupuesto sin vinculación espacial.

3.3 Aplicación de MQA

Para calcular la valoración de un grupo de registros de metadatos, se desarrolló un programa en *Python* que carga en memoria el grafo RDF de los metadatos que se quieren evaluar siguiendo las dimensiones MQA. Respecto a la implementación concreta de las métricas, se han agrupado en cuatro categorías diferentes que han implementado de la siguiente forma:

- Para las métricas que chequean la disponibilidad de una propiedad, se han elaborado consultas SPARQL que contabilizan el número de entidades (*Dataset o Distribution*) que contienen esa propiedad respecto al número total de entidades que se están evaluando. Los porcentajes de disponibilidad de cada propiedad permiten prorratear la puntuación máxima que se puede alcanzar para cada una de estas métricas.
- Para las métricas que chequean la utilización de valores concretos dentro de una propiedad, se ha recuperado con consultas SPARQL los valores concretos de cada propiedad y se han validado respecto a los vocabularios específicos recomendados por el Portal Europeo de Datos (https://gitlab.com/european-data-portal/edp-vocabularies). El porcentaje de entidades conteniendo propiedades con valores incluidos en estos vocabularios respecto al número total de entidades permite prorratear la puntuación máxima que permite alcanzar cada una de estas métricas.
- Para las métricas que chequean el acceso de las URL, se ha recuperado con consultas SPARQL cada una de las direcciones y se han realizado las correspondientes peticiones HTTP para confirmar que están activas (código de estado entre 200 y estrictamente menor que 400). El porcentaje de

entidades conteniendo direcciones URL alcanzables respecto al número total de entidades permite prorratear la puntuación máxima que permite alcanzar cada una de estas métricas.

- Para chequear la conformidad con la especificación DCAT-AP, se ha validado el formato de los metadatos frente a los ficheros con restricciones SHAPE de la versión 2.0.1 de DCAT-AP elaborados por la Unión Europea (véase https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/201-0). Estos ficheros de restricciones permiten caracterizar formalmente las entidades y propiedades, así como sus tipos de datos y multiplicidades, de la especificación DCAT-AP utilizando el lenguaje SHACL,[11] una recomendación de W3C para validar grafos RDF.

3.4 Análisis comparativo de la calidad

Como para final para la evaluación de calidad y con el fin de realizar el análisis de los conjuntos de datos, tanto de forma global como separados por unidad administrativa se proponen dos conjuntos de gráficos. El primero para el análisis de la disponibilidad de la propiedad espacial de los metadatos y el segundo seguirá la propuesta del Portal Oficial de Datos Europeos para la comparación de catálogos de datos.

El gráfico para la comparación de la propiedad espacial quedará subdividido en tres secciones, una primera identificando los porcentajes de territorios sin disponibilidad de datos abiertos, otra sección identificando los porcentajes de datos con propiedad espacial o que incluyen propiedad espacial multivaluada y sin propiedad espacial que incluye aquellos no accesibles. Finalmente, la tercera sección incluye un gráfico que identifica el número absoluto de conjuntos de datos por cada marco territorial.

En cuanto al gráfico de comparación del MQA se desarrolló un programa de *Python* con capacidad para leer los resultados obtenidos del proceso anterior para acumular las puntuaciones por dimensión y la puntuación total y representarlo bajo un único gráfico donde cada conjunto de datos queda incluido en una línea. Finalmente, cada línea usa un conjunto de barras acumuladas mostrando el total posible para cada dimensión MQA y la valoración obtenida. De esta forma, se pueden comparar los valores acumulados respecto del máximo, entre ellos y ver qué dimensiones poseen un mayor cumplimiento del MQA.

4. Resultados

Del conjunto total de datos (datasets) de la web de datos abiertos de España se descargaron exclusivamente los metadatos relativos a los indicados en la sección anterior, es decir, los que contienen la palabra "presupuesto" dentro de alguno de los campos traducida a cualquier de los idiomas oficiales en España,

[11] https://www.w3.org/TR/shacl/

siguiendo lo indicado en la sección 3.2. Este conjunto de datos fue obtenido el 14 de noviembre de 2020, por lo que se considera una foto fija de la situación del portal para estos tipos de datos en un punto determinado del catálogo cuyo recuento y comparación se muestra en la Figura 7. El conjunto de datos asciende a 571 entradas de las que 1 no es accesible por lo que no se ha podido determinar ninguna medida de la calidad.

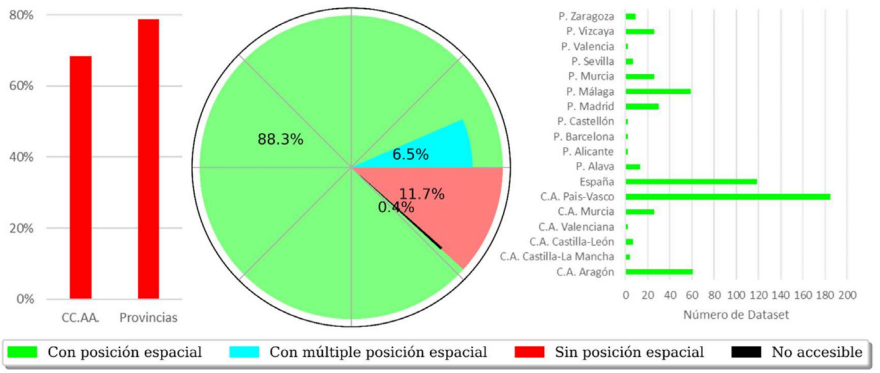

Figura 7. Resultados del MQA para distintas selecciones de los conjuntos de abiertos sobre presupuestos agrupadas por su atributo de localización espacial.

Antes de realizar una breve revisión de los datos, es necesario indicar que España tiene una división administrativa jerárquica. Cuando los datos indican España, se tratará de datos de todo el territorio, que a su vez está compuesto por Comunidades Autónomas y Ciudades Autónomas. Continuando esta estructuración, las Comunidades Autónomas están formadas por una o varias provincias. Desde un punto de vista administrativo, existe un nivel adicional, de manera que todas las provincias están compuestas por uno o varios municipios aunque éste nivel no se considera con el modelo de metadatos propuesto en España. Por lo que la cobertura espacial de un conjunto de datos está limitada al nivel de provincia.

Como primer aspecto de relevancia de los conjuntos de datos se puede notar, según la Figura 7, que pese a que una gran parte de los metadatos de los *datasets* tienen información espacial (88,3%), no todos la poseen. Este es un verdadero problema tratándose de información presupuestaria ya que limita la transparencia gubernamental en estos territorios y podría usarse en un futuro como posible mecanismo de valoración de este aspecto. En estos casos, los publicadores han priorizado la introducción de la información general del metadato, relegando el ámbito espacial que lo consideran reflejado en el título. Como ejemplo, el presupuesto de la Junta de Andalucía tiene como título (propiedad *dct:title*) "Presupuesto de la Comunidad Autónoma de Andalucía año

2019". Sin embargo no incluye ninguna propiedad espacial y como se puede observar en la gráfica de la derecha de la Figura 7, la Comunidad Autónoma Andaluza no dispone de ningún registro que incluya la localización ya que no es obligatorio según la Norma Técnica de Interoperabilidad (MINHAP, 2013). También es de destacar aquellos conjuntos de datos que poseen varias localizaciones de manera jerárquica, hasta tres localizaciones siguiendo los niveles administrativos de España antes comentados.

En cuanto a valores absolutos, del conjunto metadatos descargados, la Comunidad Autónoma del País Vasco es la que posee un mayor número absoluto (185) seguida por el conjunto de datos publicados por el Estado (119). El resto de territorios tiene un número muy inferior de datos abiertos accesibles relativos al presupuesto, recordando que la búsqueda se realizó usando el método de la palabra clave "presupuesto" en los diferentes idiomas oficiales en España.

Cambiando al análisis de los resultados obtenidos, antes de comenzar es necesario puntualizar que los metadatos publicados en https://datos.gob.es, solo pueden alcanzar 275 puntos en la valoración del MQA. Esto es debido al modelo de datos impuesto en la Norma Técnica de Interoperabilidad de Reutilización de Recursos de Información (MINHAP, 2013). Esta norma fuerza la limitación comentada de los metadatos elaborados en España para conjuntos de datos abiertos al eliminar ciertas propiedades respecto de DCAT-AP. Las puntuaciones que no se pueden alcanzar en España para la valoración de MQA con las limitaciones impuestas son las siguientes:

- Carencia de URL de descarga: 50 puntos en MQA.
- Carencia de información relativa a restricciones de acceso: 15 puntos en MQA.
- Carencia de información sobre punto de contacto: 20 puntos en MQA.
- Carencia de información sobre derechos de distribución: 5 puntos en MQA.
- Carencia de información de formato MIME (Nota: aunque esta propiedad es obligatoria en el esquema de metadatos español, solo se rellena la propiedad (*dct:format*): 10 puntos en MQA.

El conjunto anterior suma un total de 100 puntos, por lo que de forma general el test MQA sólo podría alcanzar un valor de 305. Sin embargo, los registros de metadatos del portal de datos abiertos no cumplen con el DCAT-AP porque el RDF generado no sigue exactamente las restricciones impuestas para algunas propiedades de los metadatos. Por ejemplo, el RDF de los metadatos relativo a las propiedades *dct:language*, *dct:spatial* o *dct:format* no define adecuadamente los recursos referenciados de idioma, localización o formato tal como espera la especificación de DCAT-AP. Esto resta otros 30 puntos más.

Teniendo en cuenta esa limitación en puntuación, se ha aplicado el MQA para todas las dimensiones y medidas a los metadatos que describen los 570 conjuntos de datos, recordamos que uno de ellos no era accesible. Los resultados de esta aplicación pueden verse en la Figura 8 donde las barras en gris representan la valoración máxima de MQA y los valores coloreados

identifican las valoraciones agregadas obtenidas para cada dimensión adaptando el ranking de valoraciones de catálogos disponible en la web del propio MQA (https://data.europa.eu/mqa/?locale=es).

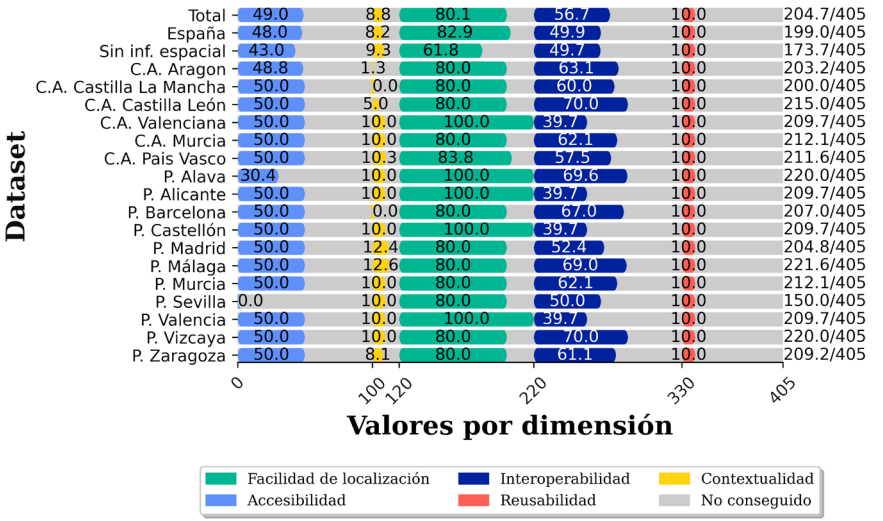

Figura 8. Resultados del MQA para distintas selecciones de los conjuntos de abiertos sobre presupuestos agrupadas por su atributo de localización espacial (Nota: la fila Total se refiere al MQA del total de los 570 registros sobre presupuesto del portal de datos abiertos de España).

En primer lugar, es necesario recordar que, tras la separación de los metadatos en los diferentes territorios, algunos de los conjuntos tienen menos de 10 registros, concretamente: las Comunidades Autónomas de Castilla-La Mancha, Castilla y León y la Comunidad Valenciana, además de las provincias de Alicante, Barcelona, Castellón, Sevilla, Valencia y Zaragoza (aunque ésta última dispone de nueve registros). Sin embargo, estos conjuntos de metadatos pueden tener un mayor número de comprobaciones para algunas de las medidas de cada dimensión porque provienen de propiedades multivaluadas, por ejemplo, el campo de formato (*dct:format*) es evaluado en un número mayor de ocasiones en múltiples metadatos.

Desde un punto de vista general, las puntuaciones son más o menos similares en los apartados de facilidad de localización, reusabilidad y accesibilidad (Figura 8, barras en verde, rojo y azul más claro), salvo el caso de los metadatos de la ciudad de Sevilla que obtienen una puntuación de 0. Este caso particular es de gran interés por dos razones, la primera porque todos los metadatos asociados a

la Provincia de Sevilla como presupuestos son los elaborados por la Universidad Pablo de Olavide, situada en la provincia de Sevilla, y que todos los datos sólo pueden ser accedidos mediante identificación de usuario y contraseña. Este último hecho hace que se incumpla uno de los criterios básicos de los datos abiertos. Sin embargo, también es interesante ver cómo la asociación espacial a una Universidad situada en una provincia altera los conjuntos de datos esperados que deberían ser relativos a los presupuestos provinciales, que en el caso de España serían presupuestos de las Diputaciones Provinciales.

Respecto de la dimensión de reusabilidad, se puede ver como todos los metadatos descargados para todas las configuraciones territoriales (Figura 8) se valoran con 10 puntos que son los relativos al publicador. Esto es fundamentalmente debido a las limitaciones impuestas por la Norma Técnica respecto del DCAT-AP al no incluir las propiedades relativas a restricciones de acceso (*dct:accessRights*) y punto de contacto (*dcat:contactPoint*). Por otra parte, la propiedad de licencia (*dct:license*), que se define como "URI que referencia al recurso que describe los términos de uso", tampoco está asociada a la entidad que chequea MQA. Mientras el esquema de metadatos de España propone asociarla a la clase *Dataset*, el MQA espera que la información de licencia se indique para cada tipo de distribución (clase *Distribution*).

En cuanto a la dimensión de interoperabilidad, la valoración del cumplimiento completo de DCAT-AP, como se ha indicado antes, reduce 30 puntos la valoración máxima al igual que la la no disponiblidad de información sobre el formato MIME asociado a la distribución. Aunque la información sobre el formato MIME (*dcat:mediaType*) es obligatoria según el esquema de metadatos de España (Figura 3), la realidad es que solo se rellena la propiedad *dct:format*. Este hecho reduce otros 10 puntos de valoración MQA dejando un máximo de 70, que es lo que alcanzan los metadatos de la Comunidad Autónoma de Castilla y León. En el otro extremo de esta valoración se encuentran territorios como la Comunidad Valenciana, donde sus metadatos fallan en la elección de un formato no propietario lo que reduce aún más su valoración en esta dimensión MQA.

Para terminar el análisis por dimensiones, en la contextualidad, todos los metadatos fallan en la propiedad de definición de los derechos (*dct:rights*) siguiendo las limitaciones indicadas al principio de la sección. El resto de las medidas arrojan diferentes valores que pueden ser más acordes con errores en la creación de los metadatos aunque de forma similar en todos los territorios.

Como resumen del análisis territorial de los metadatos, puede indicarse que todos los metadatos publicados pueden ser calificados de "suficientes" (al estar en el rango entre 121 y 220 puntos MQA) que de forma general serían "suficientes altos" por estar alrededor de los 200 puntos. La única excepción son los metadatos de la provincia de Málaga que superan los 220 puntos y que pueden calificarse de "buenos". Esto denota que la calidad de los metadatos del portal de datos abiertos de España para "presupuesto" como palabra clave, salvo en pocos casos, son debidas a las limitaciones propias de la Norma Técnica y al abuso de la propiedad *dct:format* frente a la recomendada *dcat:mediaType.* Este

análisis se ve corroborado por lo mostrado en la fila de "Total" que representa el conjunto de los 570 metadatos accedidos, tanto en las dimensiones como en la valoración global de 204 puntos ("suficiente alto").

5. Conclusiones

En este trabajo se ha propuesto un método para evaluar el estado de los conjuntos de datos de una temática concreta utilizando la metodología MQA, y poder dotarnos de un barómetro que mida su facilidad de localización, accesibilidad, interoperabilidad, reusabilidad y contextualización. En la propuesta, se ha definido un mecanismo para realizar el análisis MQA por territorio de forma automatizada y además se han planteado una representación gráfica adecuada para la realización de comparativas. El análisis por territorio es un aspecto de gran relevancia para los conjuntos de datos seleccionados, relativos a presupuestos, y de gran importancia para controlar aspectos como la transparencia gubernamental.

La propuesta ha sido probada en el portal de datos abiertos de España (https://datos.gob.es/) mostrando que, en el caso particular de este portal, las limitaciones en la valoración son un efecto secundario de las limitaciones impuestas por las Normas Técnicas que desarrollan el diseño del portal abierto y los metadatos integrados en el mismo. Por el contrario, aunque los metadatos hayan sido creados por diferentes entidades, de forma general, todas se han adaptado bien al perfil propuesto por las Normas Técnicas. La valoración global ha alcanzado el rango de "suficiente alto" para todos los conjuntos de datos seleccionados.

La metodología propuesta para el análisis de la calidad de los conjuntos de datos de una temática concreta sería aplicable para cualquier otro portal que contase con un punto SPARQL para el filtrado de información y una API REST para la descarga de los conjuntos de metadatos. En el caso de que un portal de datos no contase con punto de acceso SPARQL, se podrían implementar filtrados alternativos a través de los métodos de consulta disponibles en su API REST o realizar descargas masivas de los metadatos y posterior filtrado.

Finalmente, en el futuro nos planteamos extender el estudio realizado para los conjuntos de datos presupuestarios de España a los portales de Latino América y Europa para la comprobación de los cumplimientos básicos de calidad asociados a los portales de datos abiertos, ya que consideramos que es un paso previo para que se puedan realizar las evaluaciones de aplicabilidad y reutilización de la información propuestas por la Web Foundation.

Bibliografía

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, 35(1), 13-29. https://doi.org/10.1016/j.giq.2017.11.003

Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open Government Data Catalogs: Current Approaches and Quality Perspective. In A. Kő, C. Leitner, H. Leitold, & A. Prosser (Eds.), *Technology-Enabled Innovation for Democracy, Government and Governance* (152–166). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40160-2_13

MINHAP, (2013). 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. Ministerio de Hacienda y Administraciones Públicas (MINHAP), *Boletín Oficial del Estado*, lunes 4 de marzo de 2013. Recuperado de http://www.boe.es/boe/dias/2013/03/04/pdfs/BOE-A-2013-2380.pdf .

Neumaier, S., Umbrich, J, Polleres, A (2016). Automated Quality Assessment of Metadata across Open Data Portals. J. Data and Information Quality 8, 1, Article 2, 29 pp. https://doi.org/10.1145/2964909

Nogueras-Iso, J., Lacasta, J., Ureña-Cámara, M. A., & Ariza-López, F. J. (2021). Quality of Metadata in Open Data Portals, *IEEE Access*, 9, 60364-60382. https://doi.org/10.1109/ACCESS.2021.3073455

Oudkerk, F., Tinholt, D., van Steenbergen, E., Carrara, W., & Fischer, S. (2020). *Digital transformation and open data. European Data Portal, Oficina de Publicaciones*. ISBN: 978-92-78-41875-5. https://doi.org/10.2830/673557

Sandoval, F. (2019). Datos abiertos: oportunidades para la transformación social y digital en Venezuela. *Analecta Política*, 9(17), 295-315. http://dx.doi.org/10.18566/apolit.v9n17.a06

Unión Europea (2020). Metadata Quality Assessment Methodology. How EDP measures the quality of harvested metadata. Recuperado de https://www.europeandataportal.eu/mqa/methodology

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak A.,& Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship, *Scientific data*, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.18

W3C, (2020) Data Catalog Vocabulary (DCAT) - Version 2. W3C Recommendation 04 February 2020. Recuperado de https://www.w3.org/TR/vocab-dcat/

Web Foundation (2018). *El barómetro de los datos abiertos. Edición de los Líderes. De la promesa al progreso*. World Wide Web Foundation. Washington DC (EE.UU.), 28 pp. Recuperado de: https://opendatabarometer.org/doc/leadersEdition/ODB-leadersEdition-Report-ES.pdf