

#datasets de universidades españolas en Zenodo

a 31-12-2024

RESUMEN

Este trabajo describe un conjunto de datos que recopila información sobre los datasets de investigación depositados en **Zenodo** por investigadores afiliados a universidades públicas españolas entre 2014 y 2024. El dataset incluye, para cada universidad y año, el número de datasets identificados y la presencia del metadato '**subject**' para valorar la calidad de la descripción documental de los conjuntos de datos.

RECOPIACIÓN Y PROCESAMIENTO DE LOS DATOS

Los datos se obtuvieron mediante un 'script' que consulta **Zenodo** vía su API pública. La búsqueda se basa en identificar conjuntos de datos con autores o creadores afiliados a universidades públicas españolas. De cada uno se recopiló el año de depósito, la universidad y la presencia del metadato '**subject**', indicador del nivel básico de descripción documental. Cuando un conjunto de datos incluye autores de dos o más universidades, se contabiliza una unidad en el cómputo correspondiente a cada una. Por ello, los totales agregados no se interpretan como número de datasets únicos, sino como contribuciones institucionales. La captura de datos se realizó en la primera semana de enero de 2025.

SERIES DE DATOS ANALIZADOS (2014-2024)

- Número total de datasets de universidades públicas españolas depositados en **Zenodo**.
- Porcentaje de datasets que cuentan con la presencia del metadato '**subject**'.
- Número de datasets depositados en **Zenodo** desglosado por universidad.
- Porcentaje de datasets descritos con el metadato '**subject**' en cada universidad.

RESULTADOS GLOBALES

Año	Datasets depositados *	Datasets con metadato 'subject'
2014	75	64
2015	136	125
2016	164	143
2017	202	176
2018	429	311
2019	350	279
2020	571	393
2021	793	543
2022	1048	715
2023	1713	1463
2024	1463	527

AUTORES

Martínez Méndez, Francisco Javier, Universidad de Murcia, javima@um.es
López Carreño, Rosana, Universidad de Murcia, rosanalc@um.es
Castelló Cogollos, Lourdes, Universidad de Murcia, lourdes.castello@um.es
Baptista, Ana Alice, Universidade do Minho, analice@dsi.uminho.pt
Delgado Vázquez, Ángel M., Universidad Pablo de Olavide, adelvaz@upo.es

PALABRAS CLAVE

- Ciencia abierta
- Conjuntos de datos
- Investigación
- Universidades públicas
- España
- Zenodo

FECHA DEL ESTUDIO

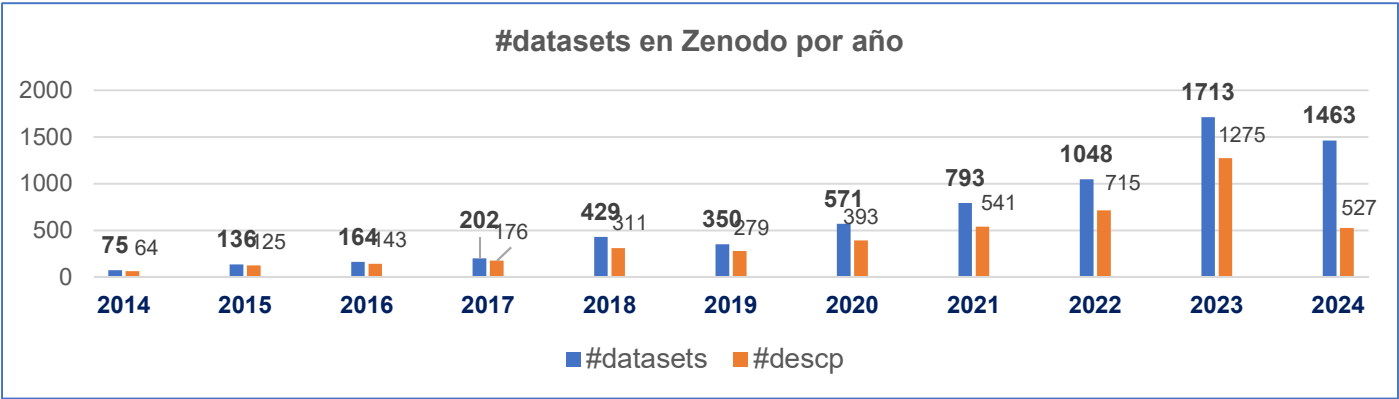
- Enero de 2025.

DATASET ASOCIADO

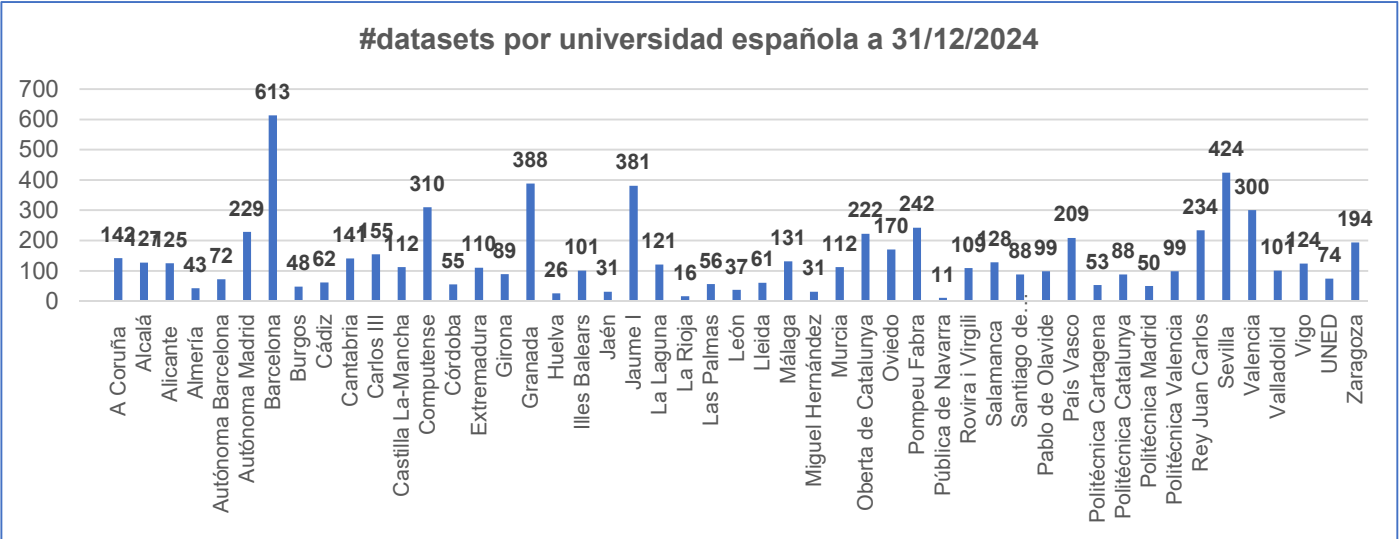
Mendez Martinez, F. J. (2025). *Datasets deposited by Spanish public universities in Zenodo up to 31 December 2024* [Dataset]. Zenodo.

<https://doi.org/10.5281/zenodo.18078549>

TOTAL DE CONJUNTOS DE DATOS DEPOSITADOS



Se observa que, a partir del año 2020, el aumento ha sido constante en cuanto al número de conjuntos de datos depositados en Zenodo por parte de las universidades públicas españolas. El descenso observado en 2024 debe interpretarse con cautela, ya que el año aún no estaba completamente consolidado en el momento de la recolección de datos. Lo que sí se observa es un **nivel menor de descripción** de los conjuntos de datos, problema que se visualiza preferentemente en la ausencia del metadato ‘subject’ (palabras clave o descriptores del conjunto de datos). El crecimiento sostenido del volumen de datasets coincide con una reducción relativa en el uso del metadato ‘subject’, lo que podría indicar tensiones entre volumen de depósito y calidad descriptiva.



Las universidades con más conjuntos de datos depositados son **Barcelona** (613), **Sevilla** (424), **Granada** (388), **Jaume I** (381), y **Valencia** (300). Por el lado contrario, las que menos conjuntos han depositado son **Pública de Navarra** (11), **La Rioja** (16), **León** (17), **Huelva** (26), **Jaén** (31), **Miguel Hernández** (31), **Almería** (43) y **Burgos** (48), todas con menos de **50 conjuntos de datos depositados**.

Estadísticos generales

- **Universidades analizadas:** 49
- **Total de datasets:** 6.944
- **Media de datasets por universidad:** 141,7
- **Mediana:** 110
- **Desviación estándar:** 119,77
- **Máximo:** 613 datasets (Universidad de **Barcelona**)
- **Mínimo:** 11 datasets (Universidad **Pública de Navarra**)

Distribución desigual: Aunque la media es de 141,7 datasets por universidad, la desviación estándar es alta (119,77), lo que indica una gran variabilidad entre universidades. Algunas superan ampliamente la media, mientras otras están muy por debajo.

Distribución por nivel de actividad

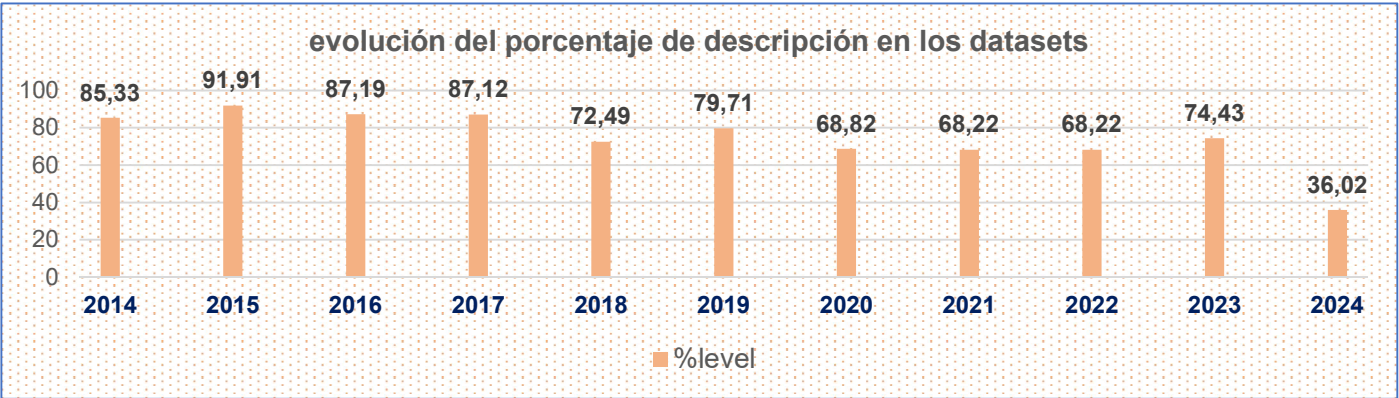
Nivel de actividad	Rango de datasets	Nº universidades	Porcentaje
Muy alta	≥ 400	2	4,1
Alta	200–399	9	18,4
Media	100–199	17	34,7
Baja	50–99	13	26,5
Muy baja	< 50	8	16,3

Interpretación

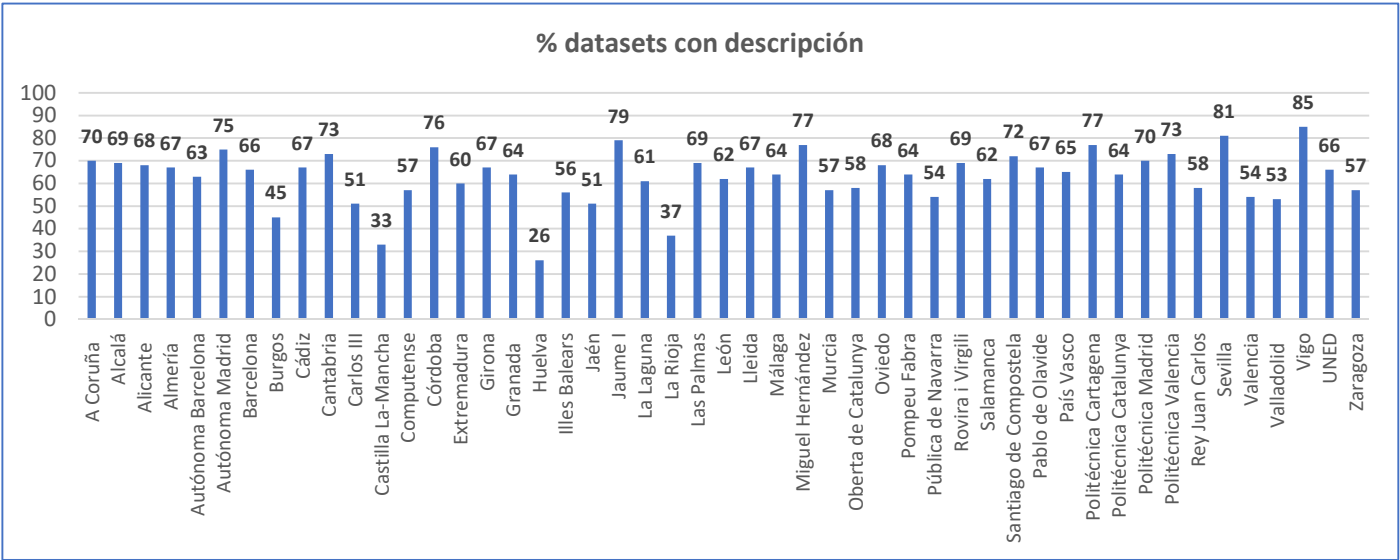
- **La mayoría de universidades (35%) tienen una actividad media** (entre 100 y 199 datasets).
- Solo un **pequeño grupo (4%)** alcanza el nivel **muy alto**, lo que señala una **concentración** de producción en pocas instituciones.
- Un **importante 43%** (baja + muy baja) está por debajo de 100 datasets, lo que indica un **reto pendiente en muchas universidades** en cuanto a políticas de datos abiertos.

Nivel de descripción por palabras clave en los conjuntos de datos depositados.

No todos los conjuntos de datos depositados en el repositorio aportan metainformación sobre palabras clave (descriptores) porque no tienen contenido en el metadato ‘subject’ (algo que resulta complicado de entender pero que ocurre).



El porcentaje de conjuntos de datos depositados en Zenodo con, al menos, una palabra clave superaba el 85% durante los primeros años para iniciar posteriormente una tendencia decreciente que volvió a cambiar en el año 2023 con un cierto ascenso. No tomamos en consideración los resultados del año 2024 porque no está completo aunque surge la duda alrededor del escaso valor calculado.



Las universidades que cuidan más la descripción por palabras clave de los conjuntos de datos son Vigo: (85%), Sevilla (81%), Jaime I (79%), Miguel Hernández (77%), Politécnica de Cartagena (77%), Córdoba (76%) y Autónoma de Madrid (75%). Por el otro lado, los porcentajes más bajos del nivel de descripción se dan en las siguientes universidades: Huelva (26%), Castilla La Mancha (33%) y La Rioja (37%).

Estadísticos generales

- **Universidades analizadas:** 49
- **Media de datasets descritos:** 63,04
- **Mediana:** 64%
- **Moda:** 67%
- **Desviación estándar:** 11,07
- **Máximo:** 85% (Universidad de Vigo)
- **Mínimo:** 26% (Universidad de Huelva)

Distribución por nivel de actividad

Nivel de actividad	Rango de datasets	Nº universidades	Porcentaje
Alta	>= 70%	12	24,50
Media	50 – 69,99%	33	67,34
Baja	< 50%	4	8,16

Interpretación

- **Desempeño general moderado:** La media (63%) indica un nivel medio de completitud en la descripción de datasets, con una mayoría de universidades situadas entre 57% y 69%. La mediana tiene un valor de 65%. La desviación estándar de 11,07%, por tanto, existe dispersión entre las instituciones, aunque mucho menor que en el análisis anterior.
- **Importancia de la documentación:** Las universidades con valores altos probablemente tienen políticas activas de ciencia abierta y gestión de datos. Las que están por debajo del promedio pueden necesitar reforzar su estrategia en este ámbito.

Limitaciones del estudio.

La identificación de universidades se basa en las afiliaciones declaradas en [Zenodo](#), cuya falta de normalización puede introducir ligeras variaciones en los recuentos. El nivel de descripción documental se ha estimado mediante la presencia del metadato ‘*subject*’, utilizado como indicador básico, sin evaluar la calidad ni adecuación de los términos empleados. Los datos correspondientes a 2024 son incompletos, al haberse recopilado unos días después del cierre del año natural. Además, el criterio de conteo asigna una unidad a cada universidad participante en conjuntos de datos con múltiples afiliaciones, por lo que los totales representan contribuciones institucionales y no conjuntos de datos únicos.

Ranking

Hemos adaptado la medida I_0 de **Borko** (Martínez Méndez, 2004, 70) creada originalmente para medir la efectividad de la recuperación de información en una medida que combina el volumen de conjuntos de datos publicados por cada universidad y el porcentaje de buenas prácticas aplicadas en la descripción (a partir del uso del metadado 'subject' para ello).

Este indicador no pretende medir calidad científica ni impacto, sino ofrecer un indicador sintético exploratorio del equilibrio entre volumen de depósito y prácticas básicas de descripción documental. Se presenta únicamente como un ejemplo de análisis derivado posible a partir del dataset y no constituye una evaluación institucional.

Para establecer nuestra clasificación o ranking, hemos seguido el **siguiente proceso**:

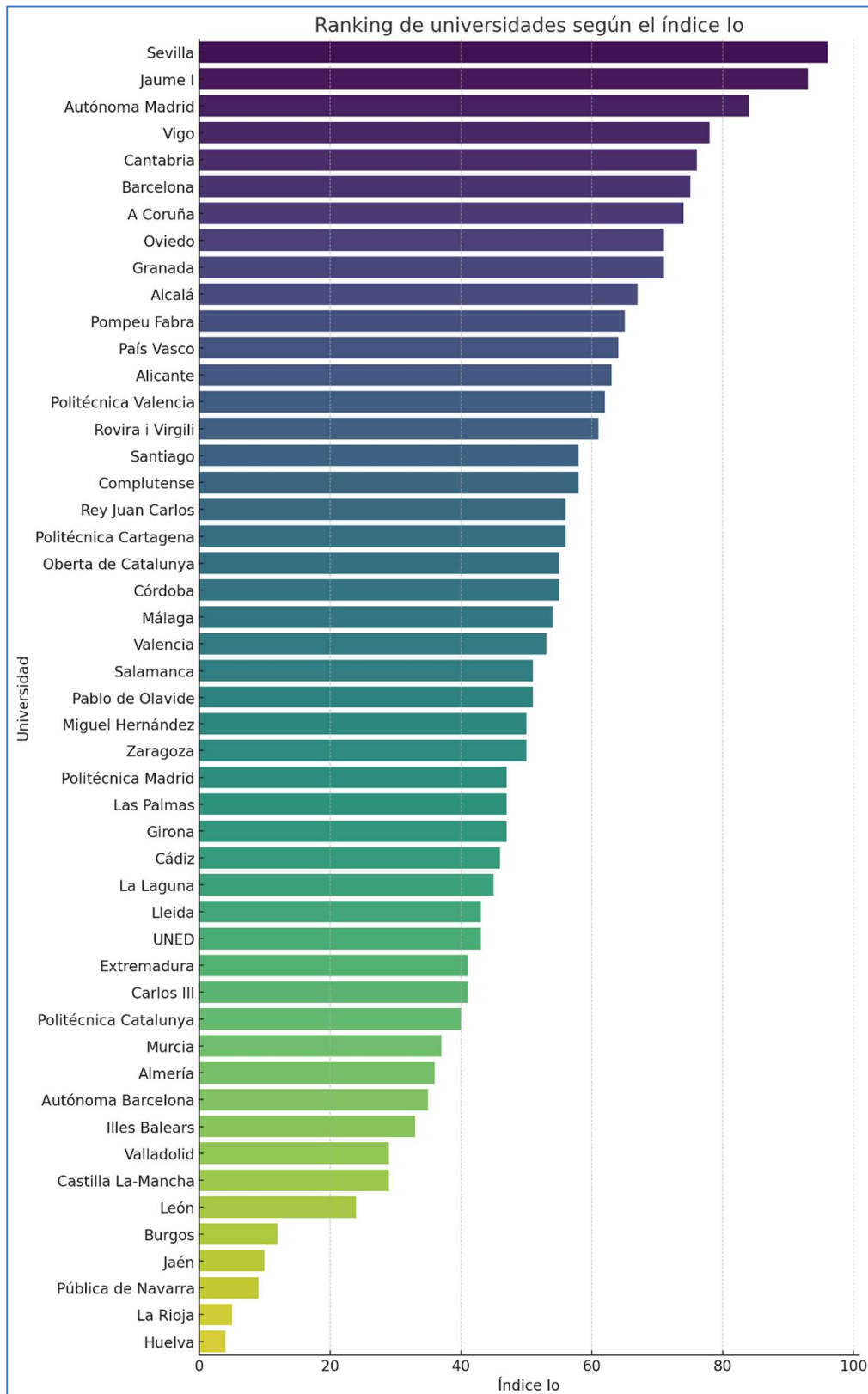
1. Ordenar las universidades de mayor a menor número de conjuntos de datos depositados en Zenodo.
2. Asignar una valor de 49 puntos ($N=49$, el número de universidades estudiadas) a la universidad con mayor número de conjuntos de datos depositados. A continuación, asignar 48 puntos a la que ocupa la segunda posición, 48 a la tercera y así, sucesivamente, 1 punto a la última (posición 49). Almacenamos esta valor en una columna llamada "P-Ordinal"
3. Ordenar las universidades de mayor a menor porcentaje de conjuntos de datos con, al menos, una palabra clave depositados en Zenodo.
4. Asignar una valor de 49 puntos ($N=49$, el número de universidades estudiadas) a la universidad con mayor porcentaje de descripción de sus conjuntos de datos depositados. A continuación, asignar 48 puntos a la que ocupa la segunda posición, 48 a la tercera y así, sucesivamente, 1 punto a la última (posición 49). Almacenamos esta valor en una columna llamada "P-Descripción".
5. En último lugar, se suma, para cada universidad, los valores asignados en los pasos 2 y 4 y almacenamos el resultado en la columna "I₀".

La medida original de **Borko** combinaba exhaustividad y precisión. Creemos posible establecer una analogía entre el número de conjuntos de datos depositados y la exhaustividad (son medidas de cobertura) y entre la precisión y el porcentaje de descripción (en este caso son medidas de efectividad). Por un lado se pondera el volumen depositado y, por el otro, un factor de calidad en su descripción documental, algo importante y trascendente en un repositorio.

Universidad	#datasets	P-Ordinal	%descripción	P-Descripción	Io
Sevilla	2	48	2	48	96
Jaume I	4	46	3	47	93
Autónoma Madrid	9	41	7	43	84
Vigo	21	29	1	49	78
Cantabria	16	34	8	42	76
Barcelona	1	49	24	26	75
A Coruña	15	35	11	39	74
Oviedo	13	37	16	34	71
Granada	3	47	26	24	71
Alcalá	19	31	14	36	67
Pompeu Fabra	7	43	28	22	65
País Vasco	11	39	25	25	64
Alicante	20	30	17	33	63
Politécnica Valencia	29	21	9	41	62
Rovira i Virgili	26	24	13	37	61
Santiago	32	18	10	40	58
Complutense	5	45	37	13	58
Rey Juan Carlos	8	42	36	14	56
Politécnica Cartagena	40	10	4	46	56
Oberta de Catalunya	10	40	35	15	55
Córdoba	39	11	6	44	55
Málaga	17	33	29	21	54
Valencia	6	44	41	9	53
Pablo de Olavide	30	20	19	31	51
Salamanca	18	32	31	19	51
Miguel Hernández	45	5	5	45	50
Zaragoza	12	38	38	12	50
Politécnica Madrid	41	9	12	38	47
Las Palmas	38	12	15	35	47
Girona	31	19	22	28	47
Cádiz	36	14	18	32	46
La Laguna	22	28	33	17	45
Lleida	37	13	20	30	43
UNED	34	16	23	27	43
Extremadura	25	25	34	16	41
Carlos III	14	36	45	5	41
Politécnica Catalunya	33	17	27	23	40
Murcia	24	26	39	11	37
Almería	43	7	21	29	36
Autónoma Barcelona	35	15	30	20	35
Illes Balears	27	23	40	10	33
Valladolid	28	22	43	7	29
Castilla La-Mancha	23	27	48	2	29
León	44	6	32	18	24
Burgos	42	8	46	4	12
Jaén	46	4	44	6	10
Pública de Navarra	49	1	42	8	9
La Rioja	48	2	47	3	5
Huelva	47	3	49	1	4

Estadísticos generales

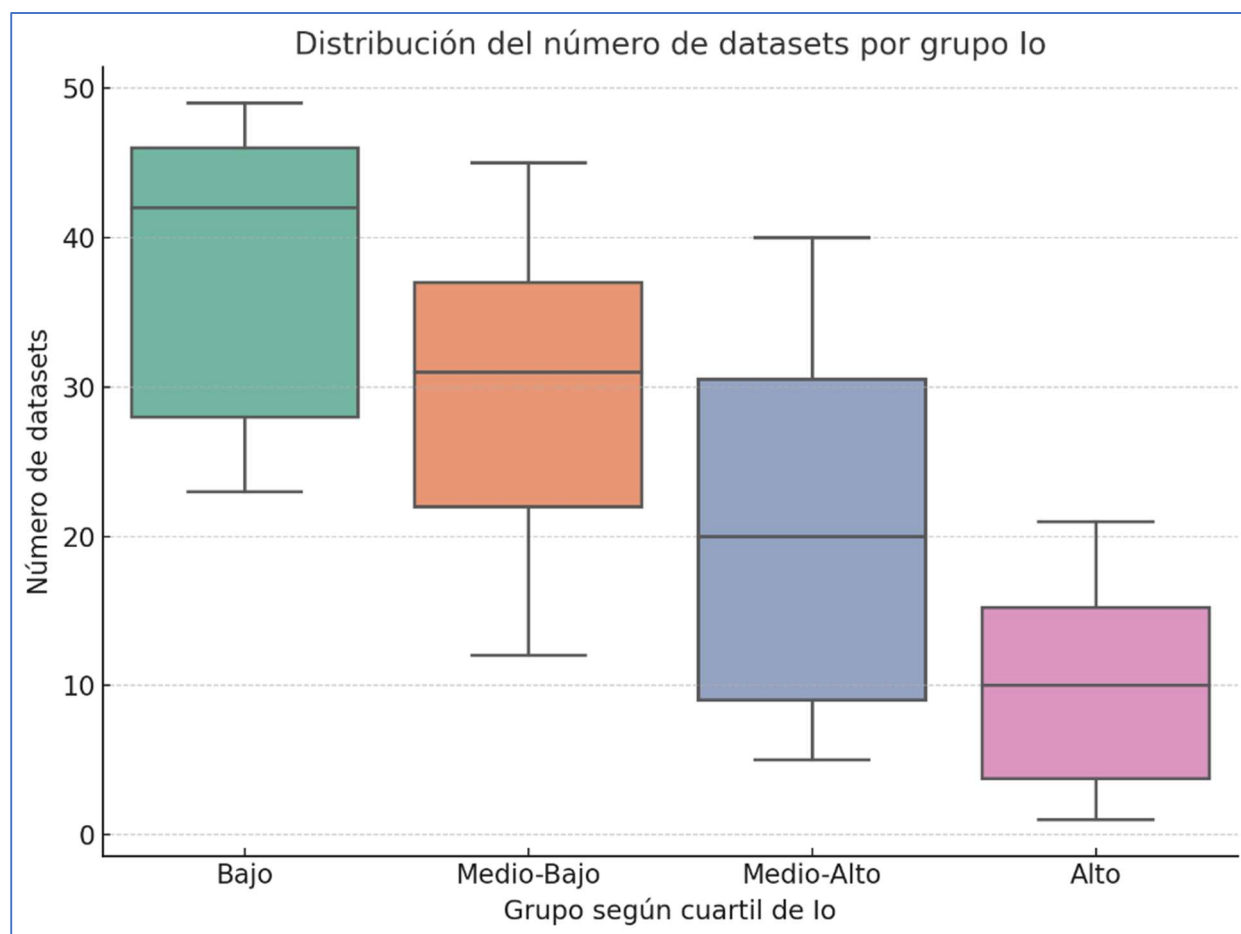
- **Universidades analizadas:** 49
- **Media de lo:** 50
- **Mediana:** 51
- **Moda:** 47
- **Desviación estándar:** 21,32
- **Máximo:** 96 (Universidad de **Sevilla**)
- **Mínimo:** 4 (Universidad de **Huelva**)



La medida de la gestión que las universidades españolas hacen de los datasets en Zenodo se puede distribuir en **cuartiles**.

Grupo lo	Rango lo	Universidades
Bajo	4 – 40	Politécnica Catalunya, Murcia, Almería, Autónoma Barcelona, Illes Balears, Valladolid, Castilla La-Mancha, León, Burgos, Jaén, Pública de Navarra, La Rioja, Huelva
Medio-Bajo	41 – 51	Pablo de Olavide, Salamanca, Miguel Hernández, Zaragoza, Politécnica Madrid, Las Palmas, Girona, Cádiz, La Laguna, Lleida, UNED, Extremadura, Carlos III
Medio-Alto	52 – 63	Alicante, Politécnica Valencia, Rovira i Virgili, Santiago, Complutense, Rey Juan Carlos, Politécnica Cartagena, Oberta de Catalunya, Córdoba, Málaga, Valencia
Alto	64 – 96	Sevilla, Jaume I, Autónoma Madrid, Vigo, Cantabria, Barcelona, A Coruña, Oviedo, Granada, Alcalá, Pompeu Fabra, País Vasco

Esta distribución da lugar al siguiente diagrama de cajas.



Licencias

Este informe y el dataset se acogen a la licencia [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/).

Referencias

Martínez Méndez, Francisco Javier. (2004). *Recuperación de Información: Modelos, Sistemas y Evaluación*. <http://hdl.handle.net/10201/4316>