

Análisis de la Madurez del Ecosistema de Datos Abiertos en España

Analysis of the Maturity of the Open Data Ecosystem in Spain



Trabajo de Fin de Máster
Curso 2025 - 26

Autor
Yulisme Zambrano Garrido

Director
Samer Hassan Collado

Master en Internet de la Cosas
Facultad de Informática
Universidad Complutense de Madrid

Análisis de la Madurez del Ecosistema de Datos Abiertos en España

Analysis of the Maturity of the Open Data Ecosystem in Spain

Trabajo de Fin de Master en Internet de las Cosas
Departamento de Ingeniería de Software e Inteligencia Artificial

Autor
Yulisme Zambrano Garrido

Director
Samer Hassan Collado

Convocatoria: Enero 2026
Calificación: 9

Master en Internet de la Cosas
Facultad de Informática
Universidad Complutense de Madrid

28 Enero de 2026

Dedicatoria

“A mi hija Isabella, mi inspiración y esfuerzo, gracias por tu paciencia, comprensión y amor. A mi esposo Guille, me has inspirado y apoyado en este proceso. Los amo. A toda mi familia, en la distancia, me han dado ánimos y apoyo. A mis amigos, y compañeros de estudios, hicimos un gran equipo.”

Agradecimientos

Expreso mi más sincero agradecimiento al Tutor y guía Samer Hassan por su orientación, paciencia, guía y constante apoyo.

A mis compañeros de máster hicimos un gran equipo.

A la Universidad. Profesores y personal, de alguna u otra forma contribuyen al buen desempeño que como alumno necesitamos.

Resumen

Análisis de la madurez del ecosistema de datos abiertos en España

Este proyecto propone evaluar de forma objetiva el nivel de madurez de los Portales de datos abiertos que hoy desempeñan un papel fundamental para los datos públicos, lo que implica un desafío, especialmente cuando se requiere un análisis comparable entre distintos portales, y conjunto de datos, a través de un enfoque metodológico a partir de los metadatos publicados y de un conjunto de métricas internas orientadas a la calidad, accesibilidad, trazabilidad e interoperabilidad de los datasets.

El trabajo se basa en la definición e implementación de un pipeline compuesto por varias fases, que incluye la selección de conjuntos de datos a partir de criterios homogéneos, la extracción y normalización de metadatos, la construcción de métricas, la validación manual y el análisis comparativo de resultados. La muestra analizada comprende diversos portales de datos abiertos en España, entre los que se incluyen el Portal Nacional, Datos abiertos Barcelona, Datos abiertos Aragón, Datos abiertos Murcia, Junta de Andalucía, Datos Comunidad de Madrid, y Analisis de Datos abiertos Castilla y León, seleccionando un volumen representativo de datasets pertenecientes a categorías temáticas comparables como transporte, salud, educación, medio ambiente y demografía.

A partir de los metadatos recopilados, se construyen indicadores que permiten evaluar aspectos como la disponibilidad de formatos reutilizables, la presencia de licencias abiertas, la antigüedad y frecuencia de actualización, la existencia de identificadores trazables y el grado de interoperabilidad técnica y semántica. Estos indicadores se agregan en métricas de madurez que facilitan la comparación entre portales y conjuntos de datos.

Como resultado, el trabajo ofrece una visión estructurada del estado de los portales analizados desde la perspectiva de sus metadatos y capacidades internas, proporcionando una base objetiva para el análisis comparativo y la identificación de buenas prácticas. La metodología propuesta es reproducible y extensible, y puede servir como apoyo para futuras evaluaciones de calidad y madurez en iniciativas de datos abiertos.

Palabras clave: análisis empírico, accesibilidad, calidad, datos abiertos, estándar dcat, interoperabilidad, madurez en portales, open data, portales de datos, reutilización de datos.

Abstract

Analysis of the Maturity of the Open Data Ecosystem in Spain

This project proposes an objective evaluation of the maturity level of open data portals, which currently play a fundamental role in the publication and use of public data. This task represents a significant challenge, particularly when a comparable analysis across different portals and datasets is required. To address this challenge, the study adopts a methodological approach based on the analysis of published metadata and a set of internal metrics focused on dataset quality, accessibility, traceability, and interoperability.

The study is based on the definition and implementation of a multi-phase pipeline that includes dataset selection using homogeneous criteria, metadata extraction and normalization, metric construction, manual validation, and comparative analysis of results. The analyzed sample comprises several open data portals in Spain, including the National Open Data Portal, Open Data Barcelona, Open Data Aragón, Open Data Murcia, the Andalusian Regional Government, the Community of Madrid Open Data Portal, and the Castilla y León Open Data initiative. A representative volume of datasets was selected from comparable thematic categories such as transport, health, education, environment, and demography.

Based on the collected metadata, a set of indicators is constructed to assess aspects such as the availability of reusable formats, the presence of open licenses, dataset age and update frequency, the existence of traceable identifiers, and the degree of technical and semantic interoperability. These indicators are aggregated into maturity metrics that enable systematic comparison across portals and datasets.

As a result, this work provides a structured overview of the current state of the analyzed portals from the perspective of their metadata and internal capabilities, offering an objective basis for comparative analysis and the identification of good practices. The proposed methodology is reproducible and extensible, and it may serve as a foundation for future quality and maturity assessments in open data initiatives.

Keywords: empirical analysis, accessibility, quality, open data, DCAT standard, interoperability, portal maturity, data portals, data reuse

Índice de contenidos

Índice de contenidos.....	8
Capítulo 1. Introducción.....	1
1.1. Motivación.....	3
1.2. Objetivos.....	4
1.2.1. Objetivo general.....	4
1.2.2. Objetivos específicos.....	5
1.3. Metodología resumida y plan de trabajo.....	6
1.4. Estructura del documento.....	7
Capítulo 2. Fundamentos Teóricos.....	9
2.1. Datos abiertos y ecosistema de datos.....	9
2.2. Infraestructuras de datos abiertos y portales de datos.....	9
2.3. Marco normativo y legislativo de los datos abiertos.....	10
2.4. Madurez de los portales de datos abiertos.....	11
2.5. Tecnologías y herramientas, Estándares en portales de datos abiertos.....	12
2.5.1. Infraestructura tecnológica del ecosistema de datos abiertos.....	12
2.5.2. Estándares y Metadatos: El papel de DCAT y DCAT-AP.....	13
2.5.3. Calidad Documental: Diccionarios de datos y Actualización.....	14
2.5.4. Formatos Semánticos, Formatos Abiertos y Principios FAIR.....	15
2.6. Métricas clave para la evaluación del ecosistema de datos abiertos.....	17
2.6.1. Interoperabilidad.....	17
2.6.2. Calidad de los datos.....	17
2.6.3. Accesibilidad.....	17
2.6.4. Trazabilidad.....	17
2.6.5. Reutilización.....	17
2.7. Interoperabilidad y estándares en la administración pública española.....	18
2.8. El paradigma del Open Data y su base conceptual.....	18
2.9. Síntesis conceptual final del capítulo.....	19
Capítulo 3. Estado del arte.....	20
3.1. Introducción al estado del arte.....	20
3.2. Evaluación de portales de datos abiertos.....	21
3.3. Enfoques conceptuales sobre Modelos de madurez de los datos abiertos.....	21
3.4. Informe europeo de madurez de datos abiertos.....	22
3.5. Marco legal y estratégico de los datos abiertos en la Unión Europea.....	23
3.6. Estudios de Madurez en el contexto español.....	23

3.7. Aportes desde América Latina y el Caribe al estudio de la madurez de los ecosistemas de datos.....	25
3.8. Aportes teóricos y metodológicos complementarios.....	27
3.9. Ecosistemas de datos e innovación social.....	29
3.10. Síntesis crítica del estado del arte y justificación de la contribución.....	29
Capítulo 4. Metodología y Tecnologías.....	31
4.1. Enfoque metodológico general.....	31
4.2. Arquitectura del pipeline de análisis.....	31
Fase 1. Extracción, selección y normalización de metadatos.....	31
Tabla 4.1. Selección, extracción de datasets, por Portal.....	32
Fase 2. Validación manual de una muestra.....	33
Fase 3. Cálculo de métricas de madurez.....	34
Fase 4. Generación de resultados.....	34
4.3. Selección de datasets y criterios de inclusión.....	35
4.4. Construcción de métricas de madurez de los portales.....	35
4.4.1. Operacionalización de los indicadores.....	36
4.4.2. Trazabilidad global.....	37
4.4.3. Interoperabilidad técnica.....	40
4.4.4. Interoperabilidad semántica.....	41
4.4.5. Accesibilidad.....	42
4.4.6. Calidad de los metadatos.....	43
4.4.7. Calcular el Índice global de madurez del portal.....	44
4.5. Generación de resultados y visualización.....	44
4.6. Implementación tecnológica y reproducibilidad.....	45
4.6.1. Librerías y Entorno de Procesamiento.....	45
4.6.2. Repositorio, documentación, herramientas.....	46
4.6.3. Automatización y Escalabilidad.....	46
4.7. Análisis exploratorio de trazabilidad externa (aporte metodológico complementario).....	46
Capítulo 5 - Diseño de la investigación y fuentes de datos.....	48
5.1. Introducción al diseño de la investigación.....	48
5.2. Selección de portales de datos abiertos.....	48
5.3. Proceso de selección y validación de datasets.....	49
5.4. Construcción del conjunto de datos finales.....	51
5.5. Métricas de madurez.....	51
5.6. Portales problemáticos y limitaciones encontradas.....	52
5.7. Tabla resumen de portales y datasets analizados.....	53

Tabla 5.1. Portales de datos abiertos, tecnologías utilizadas.....	54
Capítulo 6 - Preparación de datos e implementación.....	55
6.1. Extracción y consolidación de metadatos.....	55
6.2. Limpieza y normalización de los datos.....	57
6.2.1. Limpieza y normalización de metadatos.....	57
6.2.2. Normalización de formatos y accesibilidad técnica.....	58
6.2.3. Normalización de licencias y apertura legal.....	58
6.2.4. Relación con las métricas de madurez.....	58
6.3. Dificultades técnicas y decisiones de implementación.....	59
6.4. Implementación del pipeline y gestión de scripts.....	61
6.5. Síntesis del capítulo.....	62
Capítulo 7 - Resultados empíricos.....	63
7.1. Validación manual del pipeline y verificación de resultados.....	64
7.2. Análisis exploratorio de trazabilidad externa (aporte metodológico complementario).....	65
7.3. Resultados Métricas de Madurez de los Portales.....	66
7.3.1. Comparación global entre portales.....	66
7.3.2. Portal Barcelona.....	73
7.3.3. Portal Aragón.....	79
7.3.4. Portal Castilla y León.....	85
7.3.5. Portal Comunidad de Madrid.....	91
7.3.6. Portal Región de Murcia.....	98
7.3.7. Portal Junta de Andalucía.....	106
7.3.8. Portal Nacional (datos.gob.es).....	112
7.4. Análisis de la distribución de las métricas por portal.....	120
7.4.1. Distribución del Índice Global de Madurez por portal.....	120
7.4.2. Distribución de las dimensiones del índice por portal.....	121
7.5. Análisis estadístico de las métricas de madurez.....	124
Tabla 7.1. Resultados de ANOVA y Kruskal–Wallis por métrica y nivel de madurez.....	125
Tabla 7.2. Resultados de los contrastes post hoc de Tukey HSD.....	126
7.6. Interpretación de los resultados.....	127
Capítulo 8 – Discusión.....	128
8.1. Interpretación de los resultados a la luz del estado del arte.....	128
8.2. La trazabilidad como eje crítico de la madurez del ecosistema.....	129
8.3. Comparación con estudios previos y aportes diferenciales.....	131
8.4. Implicaciones para la evaluación de la madurez de los datos abiertos.....	133

8.5. Sesgos y limitaciones encontradas.....	134
8.6. Interpretación comparativa resultados descriptivos + boxplots + tests.....	137
Capítulo 9 – Conclusiones y líneas futuras.....	139
Líneas futuras.....	141
Contribuciones de este Trabajo de Fin de Máster.....	143
Chapter – Introduction.....	145
Introduction.....	145
1.1. Motivation.....	147
1.2. Objectives.....	148
1.2.1. General Objective.....	148
1.2.2. Specific Objectives.....	149
1.3. Methodology Overview and Work Plan.....	150
1.4. Structure of the Document.....	151
Chapter – Conclusions and Future Work.....	153
Future Research.....	155
Contributions of the Thesis.....	157
Referencias.....	158

Índice de figuras

Figura 4.1. Diagrama - Arquitectura del pipeline de análisis.....	33
Figura 7.1. Gráfica Índice Global de Madurez de los Portales abiertos....	67
Figura 7.2. Gráfica Promedio por dimensión de Madurez.....	68
Figura 7.3. Gráfica de Distribución Porcentual de Formatos abiertos y Cerrados/Propietarios.....	69
Figura 7.4. Gráfica Resultado Licenciamiento permitido (Dimensión Legal).....	70
Figura 7.5. Gráfica Frecuencia de actualización.....	71
Figura 7.6. Gráfica Accesibilidad vs Trazabilidad Global.....	71
Figura 7.7. Gráfica Calidad de metadatos vs Trazabilidad Global.....	72
Figura 7.8. Gráfica Top 3 de Categorías entre portales.....	73
Figura 7.9. Resultado Gráfico Índice de Madurez del Portal.....	74
Figura 7.10. Resultado Gráfico de componentes de Índice de Madurez del Portal.....	74
Figura 7.11. Resultado Clasificación de formatos para Interoperabilidad técnica.....	75
Figura 7.12. Resultado Licenciamiento permitido (Dimensión Legal).....	76
Figura 7.13. Resultado Gráfico Trazabilidad Global.....	77
Figura 7.14. Resultado Gráfico Métrica de Interoperabilidad Semántica.	78
Figura 7.15. Resultado Gráfico Métrica de Interoperabilidad Técnica...	78
Figura 7.16. Resultado Gráfico Métrica de Accesibilidad.....	79
Figura 7.17. Resultado Gráfico Métrica de Calidad de metadatos.....	79
Figura 7.18. Resultado Gráfico Índice de Madurez del Portal.....	80
Figura 7.19. Resultado Gráfico de componentes de Índice de Madurez del Portal.....	80
Figura 7.20. Resultado Clasificación de formatos para Interoperabilidad técnica.....	81
Figura 7.21. Resultado Licenciamiento permitido (Dimensión Legal).....	82
Figura 7.22. Resultado Gráfico Trazabilidad Global.....	83
Figura 7.23. Resultado Gráfico Métrica de Interoperabilidad Semántica.	84
Figura 7.24. Resultado Gráfico Métrica de Interoperabilidad Técnica...	84
Figura 7.25. Resultado Gráfico Métrica de Accesibilidad.....	85
Figura 7.26. Resultado Gráfico Métrica de Calidad de metadatos.....	85
Figura 7.27. Resultado Gráfico Índice de Madurez del Portal.....	86

Figura 7.28. Resultado Gráfico de componentes de Índice de Madurez del Portal.....	87
Figura 7.29. Resultado Clasificación de formatos para Interoperabilidad técnica.....	87
Figura 7.30. Resultado Licenciamiento permitido (Dimensión Legal).....	88
Figura 7.31. Resultado Gráfico Trazabilidad Global.....	89
Figura 7.32. Resultado Gráfico Métrica de Interoperabilidad Semántica.	90
Figura 7.33. Resultado Gráfico Métrica de Interoperabilidad Técnica...	90
Figura 7.34. Resultado Gráfico Métrica de Accesibilidad.....	91
Figura 7.35. Resultado Gráfico Métrica de Calidad de metadatos.....	91
Figura 7.36. Resultado Gráfico Índice de Madurez del Portal.....	92
Figura 7.37. Resultado Gráfico de componentes de Índice de Madurez del Portal.....	93
Figura 7.38. Resultado Clasificación de formatos para Interoperabilidad técnica.....	93
Figura 7.39. Resultado Licenciamiento permitido (Dimensión Legal).....	94
Figura 7.40. Resultado Gráfico Trazabilidad Global.....	95
Figura 7.41. Resultado Gráfico Métrica de Interoperabilidad Semántica.	96
Figura 7.42. Resultado Gráfico Métrica de Interoperabilidad Técnica...	96
Figura 7.43. Resultado Gráfico Métrica de Accesibilidad.....	97
Figura 7.44. Resultado Gráfico Métrica de Calidad de metadatos.....	98
Figura 7.45. Resultado Gráfico Índice de Madurez del Portal.....	99
Figura 7.46. Resultado Gráfico de componentes de Índice de Madurez del Portal.....	99
Figura 7.47. Resultado Clasificación de formatos para Interoperabilidad técnica.....	100
Figura 7.48. Resultado Licenciamiento permitido (Dimensión Legal)...	101
Figura 7.49. Resultado Gráfico Trazabilidad Global.....	102
Figura 7.50. Resultado Gráfico Métrica de Interoperabilidad Semántica.	103
Figura 7.51. Resultado Gráfico Métrica de Interoperabilidad Técnica.	104
Figura 7.52. Resultado Gráfico Métrica de Accesibilidad.....	105
Figura 7.53. Resultado Gráfico Métrica de Calidad de metadatos.....	106
Figura 7.54. Resultado Gráfico Índice de Madurez del Portal.....	107
Figura 7.55. Resultado Gráfico de componentes de Índice de Madurez del Portal.....	107

Figura 7.56. Resultado Clasificación de formatos para Interoperabilidad técnica.....	108
Figura 7.57. Resultado Licenciamiento permitido (Dimensión Legal)...	109
Figura 7.58. Resultado Gráfico Trazabilidad Global.....	110
Figura 7.59. Resultado Gráfico Métrica de Interoperabilidad Semántica.	111
Figura 7.60. Resultado Gráfico Métrica de Interoperabilidad Técnica.	111
Figura 7.61. Resultado Gráfico Métrica de Accesibilidad.....	112
Figura 7.62. Resultado Gráfico Métrica de Calidad de metadatos.....	112
Figura 7.63. Resultado Gráfico Índice de Madurez del Portal.....	113
Figura 7.64. Resultado Gráfico de componentes de Índice de Madurez del Portal.....	113
Figura 7.65. Resultado Clasificación de formatos para Interoperabilidad técnica.....	114
Figura 7.66. Resultado Licenciamiento permitido (Dimensión Legal)...	115
Figura 7.67. Resultado Gráfico Trazabilidad Global.....	116
Figura 7.68. Resultado Gráfico Métrica de Interoperabilidad Semántica.	117
Figura 7.69. Resultado Gráfico Métrica de Interoperabilidad Técnica.	118
Figura 7.70. Resultado Gráfico Métrica de Accesibilidad.....	119
Figura 7.71. Resultado Gráfico Métrica de Calidad de metadatos.....	120
Figura 7.72. Resultado Gráfico Distribución por portal.....	121
Figura 7.73. Resultado Gráfico Distribución Accesibilidad.....	122
Figura 7.74. Resultado Gráfico Distribución Interoperabilidad semántica	123
Figura 7.75. Resultado Gráfico Distribución Trazabilidad global.....	124

Índice de Tablas

Tabla 4.1. Selección, extracción de datasets, por Portal.....	32
Tabla 5.1. Portales de datos abiertos, tecnologías utilizadas.....	54
Tabla 7.1. Resultados de ANOVA y Kruskal–Wallis por métrica y nivel de madurez.....	125
Tabla 7.2. Resultados de los contrastes post hoc de Tukey HSD.....	126

Capítulo 1. Introducción

En los últimos años, los portales de datos abiertos (Open Data) han experimentado un crecimiento sostenido como parte de las estrategias de transparencia, innovación y gobierno abierto, impulsadas por las administraciones públicas y los organismos internacionales. Estos portales constituyen un punto de acceso fundamental para que ciudadanos, investigadores, empresas y organizaciones sociales puedan acceder a datos públicos con fines científicos, tecnológicos, económicos o sociales. En el contexto español, este ecosistema ha registrado una expansión significativa, con la existencia de portales de ámbito estatal, autonómico y municipal que publican miles de datasets en dominios tan diversos como la movilidad, la salud, la educación, el medio ambiente o la demografía.

El estado actual del conocimiento indica que la apertura de datos constituye únicamente el primer eslabón de una cadena de valor más amplia, donde el aprovechamiento de los conjuntos de datos depende directamente de sus niveles de madurez técnica, semántica y organizativa. En este sentido, la cuestión central no reside únicamente en el volumen de información publicada, sino en las condiciones de dicha publicación y en el grado de madurez operativa que presentan los datos para asegurar que sean accesibles, interoperables y trazables de manera sostenible.

El análisis de la madurez del ecosistema de datos abiertos requiere evaluar no solo la disponibilidad de *datasets*, sino también las condiciones estructurales bajo las cuales estos son publicados y gestionados por los distintos portales. En la práctica, muchos portales presentan debilidades significativas en aspectos fundamentales como la calidad y consistencia de los metadatos, la existencia de identificadores estables, la normalización semántica, la interoperabilidad técnica entre plataformas y la trazabilidad temporal de los datos. Estas carencias limitan la comparabilidad entre portales y dificultan la construcción de indicadores homogéneos que permitan evaluar de forma objetiva el grado de desarrollo del ecosistema. Como consecuencia, gran parte de los análisis existentes se han centrado en métricas descriptivas básicas desde la disponibilidad técnica de los datos, sin abordar de manera sistemática la evaluación integral de los componentes que determinan la madurez real de los portales.

En este contexto, el presente Trabajo Fin de Máster parte de la premisa de que la viabilidad técnica de los datos no puede evaluarse de forma fiable si previamente

no se analizan las condiciones de madurez del ecosistema. Por ello, el objetivo principal del estudio no es medir directamente el uso de los *datasets*, sino evaluar el grado de preparación de los portales y de sus metadatos para que cualquier interacción con ellos pueda producirse, ser trazada y, eventualmente, medida de forma reproducible. El trabajo se orienta, por tanto, al análisis de las condiciones estructurales, técnicas y semánticas que determinan la calidad, la accesibilidad, la trazabilidad y la interoperabilidad de los datos publicados en los portales de Open Data en España. Estas dimensiones se entienden como prerequisites fundamentales de la madurez y como elementos clave para habilitar estrategias futuras de medición empírica.

A partir de este enfoque, se diseña y aplica un marco metodológico estructurado en cuatro fases: La primera fase consiste en la selección y extracción automatizada de *datasets* a partir de diversos portales de datos abiertos de ámbito nacional, autonómico y municipal, entre ellos Portal Nacional (datos.gob.es), la Comunidad de Madrid, Datos abiertos Barcelona, Aragon, Region Murcia, Castilla y León, Junta Andalucía. Estos portales aportan un conjunto amplio y heterogéneo de *datasets*, organizados en categorías temáticas comparables como transporte, salud, educación, medio ambiente, ciencia y demografía. En esta fase se recopilan los metadatos necesarios para evaluar la madurez de los datos publicados. La segunda fase, es la validación manual de una muestra representativa para garantizar la fiabilidad de los datos extraídos. La tercera fase se centra en la construcción de métricas orientadas a evaluar la madurez de los portales, incorporando indicadores relacionados con la calidad de los metadatos, la accesibilidad, la interoperabilidad técnica y semántica, la trazabilidad a través de la consistencia temporal de los *datasets*. La cuarta fase aborda la visualización y el análisis comparativo de los resultados mediante gráficos que facilitan la comparación entre portales, categorías temáticas y dimensiones de madurez.

El aporte de este trabajo radica en que contribuye a cubrir un vacío existente en la evaluación empírica de los portales de Datos abiertos, proponiendo un enfoque metodológico reproducible centrado en la medición de la madurez del ecosistema. Los resultados obtenidos permiten formular recomendaciones prácticas orientadas a mejorar la calidad de los metadatos y la interoperabilidad, sentando así las bases para el desarrollo futuro de métricas de impacto más robustas. En definitiva, este estudio avanza hacia una comprensión más operativa del ecosistema, poniendo de manifiesto que la madurez expresada a través de la calidad y la trazabilidad de los metadatos es el elemento clave para garantizar la utilidad de los datos públicos.

1.1. Motivación

La apertura de datos públicos se ha consolidado como una herramienta fundamental para promover la transparencia institucional, fortalecer la participación ciudadana y fomentar la innovación basada en evidencia. En los últimos años, los portales de Open Data han incrementado de forma notable el volumen y la diversidad de la información publicada, ampliando el acceso a datos gubernamentales en múltiples dominios y niveles administrativos. Estos datos constituyen, además, un componente esencial en el desarrollo de infraestructuras digitales avanzadas, como las plataformas de **Internet de las Cosas (IoT)** y los servicios asociados a las **Smart Cities**, donde los datos abiertos se integran frecuentemente como fuentes externas para la toma de decisiones, la monitorización urbana y la prestación de servicios inteligentes.

No obstante, este crecimiento cuantitativo de los datos abiertos no siempre ha ido acompañado de una mejora equivalente en la calidad, la trazabilidad y la interoperabilidad de los datos publicados. En entornos como los sistemas IoT y las plataformas de ciudades inteligentes, donde los datos deben ser consumidos de forma automatizada, continua y fiable, la ausencia de metadatos consistentes y normalizados constituye una limitación crítica. Sin metadatos de calidad, los procesos de integración e interpretación automática de los datos fallan, comprometiendo el correcto funcionamiento de aplicaciones analíticas, servicios urbanos y sistemas de decisión basados en datos.

En la práctica, la mayoría de los portales de datos abiertos se limitan a ofrecer métricas internas como descargas o visualizaciones, las cuales aportan una visión parcial y poco concluyente sobre el impacto real de los datasets. Estos indicadores no permiten establecer vínculos fiables entre los datos publicados y su posible uso en infraestructuras tecnológicas complejas, como aplicaciones IoT, plataformas de análisis urbano, repositorios científicos o proyectos institucionales. Como consecuencia, resulta difícil evaluar de manera objetiva el valor público, científico o económico de los datos abiertos, generando una brecha entre los principios de la apertura y la evidencia empírica de su aprovechamiento efectivo.

Esta problemática es especialmente relevante en el contexto de los portales de datos abiertos en español, donde se observan de forma recurrente limitaciones estructurales como la ausencia de identificadores persistentes, la utilización de URLs inestables, la heterogeneidad de formatos y la falta de metadatos normalizados. Estas carencias dificultan la integración de los datos en infraestructuras de datos

abiertas utilizadas como soporte para servicios digitales avanzados, así como su eventual aprovechamiento por parte de terceros, sino también su integración en infraestructuras de datos abiertas utilizadas como soporte para servicios digitales avanzados, como los propios ecosistemas IoT y de Smart Cities.

La motivación central de este Trabajo Fin de Máster reside en la necesidad de realizar un diagnóstico profundo sobre el estado actual del ecosistema de datos abiertos en España. En lugar de centrar el análisis en el impacto externo de la información, la investigación se orienta a examinar la solidez de sus cimientos técnicos. El objetivo es determinar si el grado de madurez de las infraestructuras evaluado a través de la calidad, trazabilidad e interoperabilidad de los metadatos es suficiente para garantizar un entorno de datos fiable y preparado para las exigencias de la transformación digital.

De esta manera el proyecto se orienta a estudiar las condiciones técnicas y estructurales que deben cumplirse para que los datos abiertos puedan ser integrados de forma fiable en sistemas automatizados y plataformas inteligentes.

No obstante, el uso de técnicas automáticas de análisis y minería de texto permitirá evidenciar de forma objetiva las limitaciones actuales del ecosistema de datos abiertos, documentando los obstáculos técnicos que impiden una trazabilidad efectiva y una integración fiable de los datasets en infraestructuras digitales avanzadas.

En última instancia, la motivación de este proyecto radica en contribuir al desarrollo de un ecosistema de datos abiertos más maduro, sostenible y orientado a resultados, en el que la calidad de los metadatos, la interoperabilidad entre portales y la trazabilidad de los datasets se reconozcan como elementos clave no solo para la reutilización futura, sino también para el correcto funcionamiento de infraestructuras de datos abiertas utilizadas en contextos como las Smart Cities y los sistemas basados en IoT.

1.2. Objetivos

1.2.1. Objetivo general

Diseñar y evaluar un marco metodológico para el análisis de la madurez del ecosistema de datos abiertos en España, mediante el estudio sistemático de la calidad, accesibilidad, trazabilidad e interoperabilidad técnica y semántica de los metadatos de los datasets publicados en distintos portales de datos abiertos, con el

fin de diagnosticar el grado de desarrollo de sus infraestructuras de datos, identificar limitaciones técnicas y estructurales, y formular recomendaciones orientadas a la mejora de la interoperabilidad, la sostenibilidad y la preparación del ecosistema para su integración en infraestructuras digitales avanzadas.

1.2.2. Objetivos específicos

1. Seleccionar un conjunto representativo de portales de datos abiertos en España, considerando distintos niveles administrativos y tecnológicos, y extraer de forma automatizada los metadatos de sus datasets para su análisis comparativo.
2. Analizar la estructura de publicación y disponibilidad de metadatos de los datasets seleccionados, evaluando la presencia de información clave como identificadores, fechas, formatos, licencias, documentación y mecanismos de acceso.
3. Diseñar e implementar un pipeline automatizado de análisis, que permita aplicar de forma sistemática las métricas definidas, orientado a identificar las condiciones técnicas necesarias para la detección de madurez de portales y generar indicadores comparables entre portales.
4. Explorar de manera experimental la viabilidad de detectar señales de reutilización externa mediante fuentes abiertas como repositorios científicos, repositorios de software y plataformas de datos, evaluando las limitaciones prácticas de este enfoque en el contexto de los portales españoles.
5. Identificar y documentar las limitaciones técnicas, estructurales y semánticas que dificultan la accesibilidad y su trazabilidad automática, tales como la ausencia de identificadores persistentes, inconsistencias en URLs, falta de metadatos normalizados, escasa interoperabilidad y restricciones de las APIs.
6. Definir y proponer un conjunto de métricas para la evaluación de la madurez de los portales de datos abiertos, orientadas al análisis de la calidad de los metadatos (completitud, consistencia y coherencia), la accesibilidad de los datasets, la trazabilidad a lo largo de su ciclo de vida, así como la interoperabilidad técnica y semántica, de acuerdo con buenas prácticas y estándares de publicación de datos abiertos.
7. Formular recomendaciones prácticas dirigidas a responsables de portales de datos abiertos, orientadas a mejorar la calidad de los metadatos, la interoperabilidad entre portales y la preparación del ecosistema para su integración en sistemas digitales avanzados, como plataformas de análisis automatizado, infraestructuras IoT y servicios de Smart Cities.

1.3. Metodología resumida y plan de trabajo

La metodología de este Trabajo de Fin de Máster se articula como un estudio basado en evidencias, y el análisis comparativo orientado a analizar la madurez del ecosistema de datos abiertos en España, a partir de la evaluación sistemática de los metadatos publicados en distintos portales institucionales. El objetivo es caracterizar, de forma reproducible y comparable, el grado de preparación de estos portales para su integración en infraestructuras de datos abiertas y sistemas digitales avanzados.

El enfoque metodológico se basa en el análisis de dimensiones clave de la madurez de los datos abiertos, tales como la calidad de los metadatos, la accesibilidad, la trazabilidad y la interoperabilidad técnica y semántica. Para ello, se combinan técnicas de extracción automatizada de metadatos, procesos de normalización y validación, y la construcción de métricas comparables aplicables a distintos portales y categorías temáticas.

El plan de trabajo se organiza en cuatro etapas principales, alineadas con el pipeline metodológico desarrollado a lo largo del estudio:

- Extracción, selección y normalización de metadatos, a partir de los portales de datos abiertos seleccionados, utilizando los mecanismos de acceso disponibles (APIs y endpoints públicos) y aplicando criterios homogéneos de inclusión de datasets.
- Validación manual de una muestra representativa, orientada a complementar los procesos automatizados y a verificar la coherencia y accesibilidad efectiva de los metadatos más relevantes, garantizando la fiabilidad del conjunto de datos analizado.
- Cálculo de métricas de madurez, mediante la operacionalización de indicadores asociados a las dimensiones de accesibilidad, trazabilidad, interoperabilidad técnica y semántica, y calidad de los metadatos, permitiendo obtener puntuaciones comparables a nivel de dataset y de portal.
- Análisis y presentación de resultados, a través de representaciones gráficas y comparativas que facilitan la interpretación del estado de madurez de los portales, la identificación de patrones estructurales y la formulación de recomendaciones orientadas a la mejora del ecosistema de datos abiertos.

Este enfoque metodológico permite abordar el análisis de la madurez de los portales de datos abiertos desde una perspectiva estructural y empírica, asegurando la coherencia entre los objetivos planteados, la metodología aplicada y los resultados obtenidos a lo largo del trabajo.

1.4. Estructura del documento

La memoria se organiza en nueve capítulos que siguen una progresión lógica desde el marco conceptual hasta el análisis empírico y las conclusiones finales, con un enfoque centrado en la evaluación de la madurez técnica y estructural del ecosistema de datos abiertos en España.

- **Capítulo 1. Introducción:** Este capítulo presenta el contexto general del estudio, la motivación que lo sustenta, los objetivos de investigación, y el plan metodológico general, centrado en el análisis de la madurez de las infraestructuras de datos abiertos y de sus capacidades técnicas para integrarse en sistemas digitales avanzados.
- **Capítulo 2. Fundamentos teóricos:** El capítulo desarrolla los conceptos teóricos fundamentales que sustentan la investigación, incluyendo los principios del movimiento de datos abiertos, la noción de infraestructura de datos, la calidad de los metadatos, la accesibilidad, la trazabilidad y la interoperabilidad técnica y semántica. Se introducen también los enfoques conceptuales utilizados en la literatura para evaluar la madurez de ecosistemas de datos abiertos, proporcionando el marco necesario para la definición de métricas y dimensiones de análisis empleadas en el estudio.
- **Capítulo 3. Estado del arte:** Este capítulo revisa y analiza el estado del arte mediante el estudio de trabajos académicos, informes institucionales y propuestas técnicas relacionadas con la evaluación de portales de datos abiertos, los modelos de madurez, los indicadores estructurales y las metodologías empleadas para analizar la calidad, interoperabilidad y trazabilidad de los datos. Este capítulo permite identificar las limitaciones actuales y contextualiza la novedad y el alcance del presente trabajo.
- **Capítulo 4. Metodología y tecnologías:** En este capítulo se describe de forma detallada la metodología adoptada y las tecnologías empleadas para su implementación. Se presenta el pipeline propuesto, las herramientas utilizadas para la extracción y procesamiento de metadatos, las APIs disponibles en los portales analizados y los criterios generales aplicados para la selección y

evaluación de los datasets. Este capítulo constituye la base metodológica y técnica del estudio.

- **Capítulo 5. Diseño de la investigación y fuentes de datos:** El capítulo desarrolla el diseño de la investigación, detallando la selección de portales de datos abiertos, las categorías temáticas consideradas, los criterios de filtrado y selección de datasets y la definición operativa de las variables necesarias para el análisis. Asimismo, se describen las decisiones metodológicas adoptadas en función de la disponibilidad y características de los datos publicados en los portales españoles.
- **Capítulo 6. Preparación de datos e implementación:** Este capítulo documenta el proceso de preparación de los datos y la implementación técnica del pipeline. Incluye las tareas de extracción, limpieza, normalización y validación de los metadatos, así como la aplicación de los procedimientos necesarios para el cálculo de las métricas de madurez definidas. También se describen las dificultades técnicas y estructurales encontradas durante la implementación, derivadas de la heterogeneidad de los portales y de la calidad de los metadatos disponibles.
- **Capítulo 7. Resultados empíricos:** El capítulo séptimo presenta los resultados empíricos obtenidos a partir de la aplicación de las métricas de madurez. Se analizan los indicadores relacionados con la calidad de los metadatos, la accesibilidad, la trazabilidad y la interoperabilidad técnica y semántica de los datasets, así como las diferencias observadas entre portales. Los resultados se apoyan en gráficos comparativos que permiten caracterizar el nivel de madurez del ecosistema de datos abiertos en España.
- **Capítulo 8. Discusión:** Discusión e Interpretación de resultados, contraste con el estado del arte. Se analizan las implicaciones de los resultados para los gestores de portales y los responsables de políticas de datos abiertos, y se examinan las limitaciones metodológicas y estructurales del estudio, derivadas tanto del enfoque adoptado como de las características del ecosistema de datos abiertos en español.
- **Capítulo 9. Conclusiones y líneas futuras:** El último capítulo presenta las conclusiones finales del trabajo, sintetizando las principales aportaciones metodológicas y empíricas en relación con la evaluación de la madurez del ecosistema de datos abiertos en España. Se resumen los principales hallazgos, se formulan recomendaciones generales orientadas a la mejora de las infraestructuras de datos abiertos y se proponen líneas de trabajo futuro, centradas en la ampliación de métricas, la mejora de la interoperabilidad y la evolución hacia modelos más avanzados de evaluación del ecosistema.

Capítulo 2. Fundamentos Teóricos

2.1. Datos abiertos y ecosistema de datos

Los datos abiertos (open data) se definen como aquellos datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, sin más restricciones que la atribución de la fuente y, en algunos casos, la obligación de compartir bajo las mismas condiciones. Este enfoque se fundamenta en los principios de transparencia, acceso universal y reutilización de la información pública, promovidos por las políticas de gobierno abierto (Subirana de la Cruz, 2016).

Los datos abiertos son aquellos que cualquier persona puede acceder, utilizar y compartir libremente, siempre que estén disponibles en formatos legibles por máquinas y cuenten con una licencia que permita su reutilización, transformación y redistribución, incluso con fines comerciales (data.europa.eu.(s. f.))

Desde una perspectiva institucional, los datos abiertos constituyen un recurso estratégico para la innovación, el desarrollo económico y la participación ciudadana, siempre que se publican bajo licencias abiertas y en formatos técnicamente reutilizables (European Commission, 2019).

El concepto de ecosistema de datos abiertos amplía esta visión al considerar no solo la publicación de datos, sino el conjunto de actores, infraestructuras, normas, tecnologías y prácticas que permiten su producción, distribución y reutilización efectiva. Un ecosistema de datos abierto incluye administraciones públicas, reutilizadores (empresas, investigadores, ciudadanía), infraestructuras tecnológicas, estándares, marcos legales y mecanismos de gobernanza (Naser y Rosales, 2016).

En este contexto, la sola existencia de datasets publicados no garantiza su impacto. La calidad del ecosistema depende de la capacidad de los datos para ser encontrados, accedidos, comprendidos, reutilizados y trazados a lo largo del tiempo (Muñoz y Clavijo, 2022).

2.2. Infraestructuras de datos abiertos y portales de datos

Las infraestructuras de datos abiertos son los entornos tecnológicos que soportan el ciclo de vida de los datos abiertos, desde su publicación hasta su

reutilización. Estas infraestructuras se materializan principalmente a través de portales de datos abiertos, que actúan como puntos de acceso centralizados a los datasets publicados por una o varias entidades públicas (Aparicio García, 2016).

Los portales de datos abiertos no solo cumplen una función de catálogo, sino que incorporan servicios como APIs, visualizaciones, mecanismos de descarga, metadatos estructurados y, en algunos casos, métricas de uso. Su diseño y grado de estandarización influyen directamente en la interoperabilidad y reutilización de los datos (Curto-Rodríguez, 2021).

Diversos estudios han demostrado que los portales evolucionan de forma desigual, presentando distintos niveles de desarrollo en aspectos técnicos, semánticos y organizativos. Esta heterogeneidad dificulta la comparación entre portales y la evaluación global del ecosistema de datos abiertos (Criado et al., 2023).

2.3. Marco normativo y legislativo de los datos abiertos

El marco legal constituye un pilar fundamental del ecosistema de datos abiertos. En el ámbito europeo, la Directiva (UE) 2019/1024 sobre datos abiertos y reutilización de la información del sector público establece las bases para garantizar el acceso, la reutilización y la armonización normativa entre los Estados miembros (European Commission, 2019).

Esta directiva introduce conceptos clave como los datos de alto valor, la obligatoriedad de publicación en formatos legibles por máquina y el uso de licencias abiertas estandarizadas. En el contexto español, estas disposiciones se han integrado progresivamente en las políticas nacionales y autonómicas de datos abiertos, siendo el portal datos.gob.es un referente a nivel estatal (Datos.gob.es, 2019).

El cumplimiento normativo, sin embargo, no garantiza por sí mismo la reutilización efectiva de los datos. La legislación establece el marco mínimo, pero la madurez del ecosistema depende de factores adicionales como la calidad de los metadatos, la actualización de los datasets y su trazabilidad (Oviedo Blanco, 2016).

2.4. Madurez de los portales de datos abiertos

El concepto de madurez de los portales de datos abiertos hace referencia al grado de desarrollo alcanzado por un portal en relación con su capacidad para publicar datos de forma sostenible, interoperable y orientada a la reutilización.

En el ámbito de las políticas de datos abiertos, el concepto de madurez ha evolucionado desde una visión puramente cuantitativa basada en el volumen de información publicada hacia una perspectiva multidimensional que prioriza la calidad operativa y la utilidad técnica de las plataformas. La madurez de un portal se define como el grado de desarrollo y capacidad institucional para ofrecer funcionalidades avanzadas que satisfagan las necesidades de los usuarios, aseguren la sostenibilidad de la plataforma y promuevan el aprovechamiento efectivo de los recursos digitales. *Open Data Maturity Report* (European Data Portal, 2025)

Bajo este marco de referencia, la madurez se articula en torno a cuatro dimensiones estratégicas que permiten auditar el estado de un ecosistema de datos:

1. Características del portal: Representa la dimensión técnica y funcional. Evalúa la disponibilidad de herramientas avanzadas para el acceso a la información, tales como interfaces de programación de aplicaciones (APIs), puntos de acceso SPARQL, y funciones de búsqueda semántica. Asimismo, destaca la relevancia de los mecanismos de retroalimentación y la priorización de conjuntos de datos de alto valor (*High Value Datasets*).
2. Uso del portal: Se centra en la capacidad analítica de la plataforma para monitorear el comportamiento de los usuarios. Esta dimensión implica la optimización constante de la accesibilidad y el diseño de mejoras basadas en la experiencia de uso empírica.
3. Provisión de datos: Examina la madurez en la gestión de la cadena de suministro de datos. Incluye la integración de flujos de información procedentes de diversos niveles administrativos (nacional, regional y local), la gestión de datos dinámicos en tiempo real y el soporte técnico a los proveedores para garantizar la calidad desde el origen.
4. Sostenibilidad del portal: Analiza la viabilidad estratégica a largo plazo. Esta dimensión contempla la existencia de planes de promoción activa, la transparencia en la documentación técnica y el monitoreo sistemático del rendimiento operativo del portal.

Diversos modelos de madurez han sido propuestos en la literatura, incorporando dimensiones técnicas, organizativas y de impacto. Estos modelos permiten clasificar los portales en distintos niveles evolutivos, desde fases iniciales centradas en la publicación básica de datos hasta estadios avanzados orientados a la reutilización y generación de valor (Neira Paredes, 2019; Salas Rodríguez, 2023).

A nivel europeo, el Open Data Maturity Report se ha consolidado como un referente para evaluar la evolución de los países y sus portales, considerando dimensiones como políticas, impacto, calidad y uso de los datos (European Data Portal, 2025).

No obstante, la mayoría de estos enfoques se basan en indicadores declarativos o estructurales, y no en evidencias empíricas de reutilización real, lo que constituye una de las principales limitaciones detectadas en el estado del arte (Curto-Rodríguez, 2021).

En definitiva, la madurez de un portal actúa como un indicador de su capacidad para constituirse en un nodo central de acceso a la información pública. Desde esta perspectiva, un ecosistema maduro es aquel que trasciende la simple publicación para fomentar un entorno de transparencia e innovación, donde la calidad de los metadatos y la robustez técnica garantizan el impacto de los datos en los sectores científico y económico.

2.5. Tecnologías y herramientas, Estándares en portales de datos abiertos

La arquitectura tecnológica que sostiene el ecosistema de datos abiertos en España se articula mediante una combinación de plataformas de gestión, lenguajes de consulta y estándares de descripción. De acuerdo con (Maseda-Seco et al. 2017), la interoperabilidad técnica y semántica de estos portales no es solo una opción técnica, sino un requisito esencial para que los metadatos puedan ser consumidos por sistemas externos de forma eficiente.

2.5.1. Infraestructura tecnológica del ecosistema de datos abiertos

En la práctica, la mayoría de los portales españoles utilizan **CKAN**. La infraestructura tecnológica de los portales de datos abiertos se apoya en herramientas como CKAN, y se ha consolidado como referencia para la

implementación de catálogos de datos, facilitando la publicación, descubrimiento, uso y gestión de conjuntos de datos por parte de gobiernos, organizaciones y el público en general, ofreciendo funcionalidades como APIs y visualizaciones, su capacidad para exponer metadatos a través de APIs REST facilita que investigadores y desarrolladores realicen extracciones automatizadas. Para los portales de mayor madurez semántica, se integra el uso de lenguajes de consulta semántica como SPARQL, diseñado para interactuar con bases de datos RDF, esenciales para la explotación de datos enlazados. Como señala (Aparicio García, 2016), el despliegue de estas tecnologías en instituciones públicas y universidades españolas ha permitido transitar desde simples repositorios hacia infraestructuras de datos dinámicas y consultables.

2.5.2. Estándares y Metadatos: El papel de DCAT y DCAT-AP

DCAT-AP (*Data Catalog Vocabulary Application Profile*) es un perfil de aplicación basado en el estándar DCAT del W3C. Fue diseñado específicamente por la Unión Europea para homogeneizar la descripción de los catálogos de datos abiertos en todo el continente.

El estándar DCAT-AP se define como un perfil de aplicación basado en el vocabulario DCAT del W3C, diseñado por la Unión Europea para homogeneizar la descripción de los catálogos de datos abiertos en el continente (datos.gob.es, 2022)

El uso consistente de estándares es un factor clave para la automatización de procesos de análisis, como los pipelines desarrollados en este trabajo, orientados a la evaluación de la madurez y reutilización de los datos.

La descripción estructurada de los activos de información se apoya en el vocabulario. La interoperabilidad técnica y semántica de los portales de datos abiertos se apoya en el uso de estándares de metadatos. La estandarización es un componente crítico de la madurez, y en este ámbito el vocabulario DCAT (*Data Catalog Vocabulary*) se ha consolidado como el estándar de referencia para describir catálogos y datasets en la web (W3C, 2020). Este modelo permite normalizar elementos críticos como identificadores, licencias y formatos. En el contexto europeo y español, su perfil de aplicación DCAT-AP añade capas de especificidad que garantizan que los catálogos nacionales sean interoperables entre sí.

DCAT permite describir de forma estructurada elementos como identificadores, licencias, formatos, fechas de publicación y actualización, así como relaciones entre datasets. Su adopción facilita la interoperabilidad entre portales y el consumo automático de metadatos por aplicaciones externas. Lo que permite profundizar en la estructura multinivel de DCAT y muestran cómo los portales pueden mejorar la organización y descubrimiento de datasets mediante mapas semánticos. (Neumaier et al., 2016).

2.5.3. Calidad Documental: Diccionarios de datos y Actualización

La madurez de un dataset se mide por su documentación interna. El uso de diccionarios de datos y vocabularios controlados es fundamental para que el usuario final comprenda el significado de cada variable. La presencia de un diccionario de datos es un indicador de calidad documental superior, ya que reduce la ambigüedad en la interpretación. El diccionario de datos es un componente esencial para garantizar la comprensión y reutilización de los datasets.

El diccionario de datos se define como el repositorio centralizado que documenta los metadatos técnicos intrínsecos al almacenamiento y estructura de la información en un sistema de gestión de bases de datos (SGBD). Su propósito fundamental es eliminar la ambigüedad semántica y técnica, facilitando la interpretación unívoca de los activos de datos por parte de analistas y desarrolladores (datos.gob.es, 2022)

Complementariamente, el uso de vocabularios controlados y taxonomías normalizadas permite homogeneizar categorías y conceptos entre distintos portales, facilitando la interoperabilidad semántica y el análisis comparativo (W3C, 2020).

La ausencia de diccionarios de datos o el uso inconsistente de vocabularios limita significativamente la calidad del ecosistema y la posibilidad de análisis automatizados.

Asimismo, la frecuencia de actualización documentada asegura que la información sea oportuna y fiable, un aspecto que (Aleixandre-Benavent et al. 2021) resaltan como crítico para la investigación científica basada en datos abiertos.

2.5.4. Formatos Semánticos, Formatos Abiertos y Principios FAIR

En el ámbito de los datos abiertos, la distinción entre formatos permitidos (abiertos) y propietarios (cerrados) es determinante para la interoperabilidad y la reutilización de la información.

Clasificación de Formatos

- Formatos Abiertos (Permitidos): Aquellos cuyas especificaciones técnicas son públicas, no tienen restricciones de uso y son mantenidos por organizaciones de estándares (como el W3C). Su objetivo es garantizar que cualquier persona pueda acceder a los datos sin necesidad de adquirir software específico. Ejemplos comunes son CSV, JSON, XML, RDF y GeoJSON.
- Formatos Proprietarios o Cerrados: Archivos cuya estructura es propiedad de una empresa privada y suele requerir software de pago para su correcta lectura o edición. Aunque algunos son estándares de facto por su uso masivo, imponen barreras económicas y técnicas. Ejemplos típicos son XLSX (Microsoft Excel), DOCX (Microsoft Word) o PBIX (Power BI).

La evolución desde los formatos cerrados hacia estándares abiertos y semánticos constituye el eje central de la madurez tecnológica de los datos abiertos. Mientras que los formatos abiertos eliminan las barreras de entrada al software propietario, los formatos semánticos habilitan una capa de inteligencia que permite la conexión de datos heterogéneos. Finalmente, la evolución hacia el *Linked Open Data* se materializa en formatos como RDF y JSON-LD, que permiten que los datos no solo sean leídos por máquinas, sino que estén conectados entre sí. Este enfoque se alinea con los principios FAIR (Findable, Accesible, Interoperable, Reusable), los cuales son esenciales para garantizar que los datos, especialmente en sectores sensibles como las ciencias de la salud, mantengan su utilidad y trazabilidad a largo plazo. (García-Espinosa et al. 2022)

Los principios FAIR son un conjunto de directrices diseñadas para mejorar la reutilización de los datos científicos, haciéndolos más accesibles, interoperables y útiles tanto para humanos como para máquinas.

Los principios FAIR (por sus siglas en inglés: *Findability, Accessibility, Interoperability, and Reusability*) constituyen un marco de directrices técnicas y organizativas orientadas a optimizar el descubrimiento y la explotación de activos digitales. A diferencia de los modelos tradicionales de apertura, el enfoque FAIR pone

especial énfasis en la capacidad de las máquinas para encontrar y procesar datos de forma autónoma, exigiendo para ello metadatos ricos, identificadores persistentes y licencias de uso explícitas. En el contexto de los datos abiertos gubernamentales, estos principios actúan como el estándar de excelencia para determinar el grado de madurez técnica de una infraestructura de datos (Wilkinson y cols., 2016)

Un formato abierto (como CSV o JSON) garantiza que cualquier usuario, sin importar su capacidad económica para pagar licencias de software, pueda acceder al dato. Sin embargo, el salto hacia los formatos semánticos (como RDF o JSON-LD) representa la verdadera madurez. Mientras que un CSV es una tabla aislada, un formato semántico permite que el dato "se explique a sí mismo" y se conecte con otros. En tu investigación, esto es vital porque justifica por qué tu *pipeline* analiza si los portales ofrecen estas serializaciones para permitir consultas complejas vía SPARQL.

La clasificación de los activos de información según su formato es un indicador crítico de accesibilidad. Los formatos abiertos (como CSV o JSON) son aquellos cuyas especificaciones técnicas son públicas y no imponen restricciones de uso. (Open Knowledge Foundation, 2015). Por el contrario, autores como (Maseda-Seco et al. 2017) advierten que el uso de formatos propietarios (como XLSX o PBIX) introduce barreras económicas y técnicas, ya que su estructura pertenece a empresas privadas y exige software específico para su procesamiento. En el contexto español, el (Esquema Nacional de Interoperabilidad, 2010) prioriza el uso de estándares abiertos para asegurar la neutralidad tecnológica y la preservación de los datos a largo plazo.

En este marco, el uso consistente de estos estándares y tecnologías es lo que habilita la creación de herramientas de análisis avanzado, como los pipelines de procesamiento implementados en el presente estudio, permitiendo una evaluación objetiva de la madurez del ecosistema español.

En el ámbito semántico, (Zaveri et al. 2016) identificar métricas para evaluar la calidad de los datos enlazados, destacando la importancia de vocabularios estandarizados y metadatos ricos.

2.6. Métricas clave para la evaluación del ecosistema de datos abiertos

La evaluación de la madurez de un ecosistema de datos abiertos requiere la definición de métricas claras y reproducibles. En este trabajo se adoptan cinco dimensiones fundamentales, ampliamente reconocidas en la literatura:

2.6.1. Interoperabilidad

La interoperabilidad se refiere a la capacidad de los sistemas para intercambiar y reutilizar datos de forma efectiva. Incluye aspectos técnicos (formatos abiertos, APIs) y semánticos (uso de vocabularios controlados y estándares) (Maseda-Seco et al., 2017).

2.6.2. Calidad de los datos

La calidad de los datos abarca dimensiones como completitud, consistencia, actualidad y precisión. Una baja calidad limita significativamente la reutilización, incluso cuando los datos son abiertos (Batini y Scannapieco, 2016; Zaveri et al., 2016).

2.6.3. Accesibilidad

La accesibilidad evalúa la facilidad con la que los datos pueden ser encontrados y descargados, considerando la disponibilidad de URLs estables, formatos reutilizables y ausencia de barreras técnicas (De Sousa et al., 2025).

2.6.4. Trazabilidad

La trazabilidad hace referencia a la posibilidad de seguir el rastro de un dataset desde su publicación hasta sus reutilizaciones posteriores. Un dataset es trazable cuando dispone de identificadores persistentes y metadatos suficientes para ser referenciado en terceros sistemas, repositorios o publicaciones (Oviedo Blanco, 2016).

2.6.5. Reutilización

La reutilización constituye la dimensión de mayor impacto, ya que refleja el uso real de los datos en contextos como investigación científica, desarrollo de software, aplicaciones móviles o proyectos institucionales. La medición empírica de la reutilización sigue siendo un desafío abierto en la literatura, y representa el principal foco innovador de este TFM (Curto-Rodríguez, 2021).

2.7. Interoperabilidad y estándares en la administración pública española

La interoperabilidad es uno de los elementos más determinantes para evaluar la madurez de un ecosistema de datos abiertos. En España, el Esquema Nacional de Interoperabilidad (ENI) constituye el marco normativo que define los criterios técnicos, semánticos y organizativos necesarios para garantizar que la información pública pueda ser compartida y reutilizada entre distintas administraciones. El ENI establece principios como el uso de estándares abiertos, la normalización de metadatos, la interoperabilidad semántica mediante vocabularios comunes y la preservación de la información a largo plazo.

Este marco no solo asegura la compatibilidad técnica entre sistemas, sino que también actúa como un indicador clave para evaluar el grado de alineación de las administraciones con las buenas prácticas de gobierno digital. En la práctica, el cumplimiento del ENI condiciona directamente la calidad, consistencia y reutilización de los datos publicados en los portales españoles, convirtiéndose en un elemento indispensable para analizar la madurez del ecosistema nacional (Gobierno de España, 2010)

2.8. El paradigma del Open Data y su base conceptual

El paradigma del Open Data, o datos abiertos, se fundamenta en la filosofía de que determinados datos deben estar disponibles de forma libre para todo el público, sin restricciones de derechos de autor, patentes u otros mecanismos de control. De acuerdo con Subirana de la Cruz (2016), este movimiento no solo posee una dimensión técnica, sino que se erige como un pilar esencial para la transparencia y la rendición de cuentas en las administraciones públicas. La apertura de datos permite que la información generada con recursos públicos retorne a la sociedad, fomentando la creación de valor social y económico a través de su reutilización.

Desde un punto de vista conceptual, el Open Data se sustenta en una serie de principios que garantizan su utilidad real. Bello García (s. f.) sostiene que para que un conjunto de datos sea considerado efectivamente "abierto", debe cumplir con requisitos de accesibilidad y disponibilidad, lo que implica que la información sea suministrada en formatos que permitan su procesamiento automático. Esta base

conceptual sugiere que la sola publicación de documentos digitalizados no es suficiente; la interoperabilidad técnica es la que determina la capacidad de integración de estos datos en sistemas digitales avanzados.

La importancia de este paradigma reside en su capacidad para poner la información al alcance de todos. No se trata solo de publicar archivos, sino de derribar las barreras que separan a la administración del ciudadano, permitiendo que cualquier persona, sin importar sus recursos técnicos, pueda entender y reutilizar la información que le pertenece. Mientras que Subirana de la Cruz (2016) destaca el impacto en la gobernanza y la participación ciudadana, la literatura técnica refuerza que la base del Open Data es la eliminación de barreras, tanto legales como tecnológicas. Por tanto, el paradigma se define por la convergencia entre la voluntad política de transparencia y la implementación de estándares técnicos que aseguren que los datos sean completos, primarios, oportunos y, fundamentalmente, procesables por máquinas.

En este contexto, la madurez de un ecosistema de datos abiertos se mide por su capacidad para transitar desde la simple exposición de información hacia la construcción de infraestructuras de datos robustas. Como señala Bello García (s. f.), este proceso requiere una estructura semántica clara que facilite la trazabilidad y la calidad de los metadatos, elementos que constituyen el núcleo de la interoperabilidad en el entorno actual de las administraciones públicas españolas.

2.9. Síntesis conceptual final del capítulo

Este capítulo ha establecido los fundamentos conceptuales necesarios para comprender el análisis de la madurez del ecosistema de datos abiertos en España. Se han abordado los conceptos de datos abiertos, infraestructuras, marco legal, madurez de los portales, estándares tecnológicos y métricas clave, sentando las bases teóricas sobre las que se apoya la metodología desarrollada en los capítulos posteriores.

Capítulo 3. Estado del arte

3.1. Introducción al estado del arte

El capítulo define la revisión de proyectos de investigación a través de una revisión crítica de los estudios relevantes sobre la madurez de los ecosistemas de datos abiertos, con especial énfasis en el contexto español, así como latinoamericano. Se abordan modelos de evaluación, principios de datos abiertos, marco legal, dimensiones técnicas y sociales, así como enfoques institucionales que permiten comprender el grado de desarrollo de los portales de datos abiertos.

El análisis de los portales de datos abiertos y de los ecosistemas de datos abiertos ha sido objeto de un creciente interés académico e institucional durante los últimos diez años, especialmente en el contexto de las políticas de gobierno abierto y de reutilización de la información del sector público. Las obras y bibliografías existentes abordan este fenómeno desde múltiples perspectivas, incluyendo la evaluación de la calidad de los datos, la interoperabilidad, la accesibilidad, el impacto socioeconómico y los modelos de madurez de los portales (Curto-Rodríguez, 2021).

Sin embargo, gran parte de los estudios previos se han centrado en indicadores estructurales o declarativos, tales como el número de datasets publicados, la adopción de estándares de metadatos o la existencia de licencias abiertas. Aunque estos enfoques resultan útiles para describir el estado formal de los portales, ofrecen una visión limitada sobre el uso efectivo y la reutilización real de los datos por parte de terceros (Criado et al., 2023).

En este contexto, el presente capítulo tiene como objetivo revisar de manera crítica los principales trabajos académicos, informes institucionales y modelos existentes relacionados con la evaluación de portales de datos abiertos y su grado de madurez. A partir de esta revisión, se identifican las principales tendencias, limitaciones metodológicas y vacíos de investigación que justifican el enfoque empírico propuesto en este Trabajo Fin de Máster.

3.2. Evaluación de portales de datos abiertos

Los primeros estudios sobre portales de datos abiertos se centraron en analizar su grado de desarrollo técnico y organizativo, evaluando aspectos como la disponibilidad de datasets, la estructura de los catálogos y la presencia de mecanismos básicos de acceso y descarga. En este sentido, trabajos como el de (Curto-Rodríguez, 2021) propusieron análisis multidimensionales aplicados a portales autonómicos españoles, incorporando indicadores de calidad, interoperabilidad y actualización de los datos.

A nivel local y municipal, investigaciones recientes han profundizado en la comparación de portales, evidenciando una notable heterogeneidad en los niveles de madurez alcanzados, incluso dentro de un mismo país. Estos estudios subrayan la influencia de factores organizativos, políticos y técnicos en la evolución de los portales (Criado et al., 2023).

3.3. Enfoques conceptuales sobre Modelos de madurez de los datos abiertos

Con el objetivo de formar bases de conocimientos y sistematizar la evaluación de los portales, diversos autores han propuesto modelos de madurez que clasifican las iniciativas de datos abiertos en niveles evolutivos. Estos modelos suelen integrar dimensiones como gobernanza, tecnología, calidad de los datos y participación de los reutilizadores (Neira Paredes, 2019; Oviedo Blanco, 2016).

En el ámbito internacional, destacan también los informes periódicos del European Data Portal, que evalúan la madurez de los países europeos a partir de indicadores agregados y comparables (European Data Portal, 2025). Estos informes han contribuido a estandarizar criterios de evaluación, aunque presentan un enfoque principalmente macro y declarativo.

Una limitación común de estos modelos es su escasa capacidad para capturar dinámicas de reutilización empírica, ya que suelen basarse en encuestas, autoevaluaciones o métricas internas de los portales, sin rastrear evidencias externas como citas científicas, uso en software o aplicaciones.

La madurez de los portales de datos abiertos no solo depende de la cantidad de información publicada, sino de la calidad intrínseca tanto de los datos como de

las herramientas de gestión que el portal ofrece. El autor plantea que un modelo de madurez eficaz debe considerar la dicotomía entre publicación y consumo, asegurando que los datos públicos permitan una reutilización real que genere valor económico y social. En su investigación, Oviedo Blanco introduce un modelo cualitativo que evalúa criterios de calidad técnica y organizativa, los cuales fueron fundamentales para la creación de la norma técnica nacional en Costa Rica (INTE 27-02-02:2015). Asimismo, el autor concluye que la determinación del nivel de madurez es un mecanismo esencial para verificar si las instituciones están cumpliendo con los objetivos de transparencia y rendición de cuentas propuestos por las iniciativas de gobierno abierto (Oviedo Blanco, 2016)

La madurez de las iniciativas de datos abiertos no depende exclusivamente de la plataforma tecnológica, sino que requiere de un modelo de gobernanza que integre las dimensiones estratégicas y culturales de la institución. El autor propone un modelo de medición estructurado en tres ejes fundamentales: (i) la dimensión estratégica, que evalúa el respaldo político y la alineación del proyecto con los objetivos institucionales; (ii) la dimensión de capacidades técnicas y adaptativas, centrada en las competencias organizacionales y el cambio cultural necesario; y (iii) la dimensión ciudadana, que mide el grado de compromiso con la participación pública y la innovación. Neira Paredes concluye que el uso de una escala de madurez permite a las instituciones públicas identificar buenas prácticas y cerrar brechas de información, transformando los datos abiertos de un simple repositorio digital en una herramienta de colaboración entre el gobierno y la ciudadanía (Neira Paredes, 2019)

3.4. Informe europeo de madurez de datos abiertos

El *2025 Open Data Maturity Report*, elaborado por el European Data Portal, constituye la evaluación más completa y actualizada sobre el estado de los datos abiertos en los países miembros de la Unión Europea. Este informe analiza el nivel de desarrollo de los ecosistemas nacionales a través de cuatro dimensiones clave: política, impacto, calidad y portal, ofreciendo una visión comparada que permite identificar avances, brechas y tendencias comunes en la región. El estudio destaca que la madurez de los datos abiertos no depende únicamente de la publicación de conjuntos de datos, sino de la existencia de marcos normativos sólidos, estrategias de gobernanza, mecanismos de participación y herramientas que faciliten la

reutilización efectiva. Asimismo, el informe subraya que los países con mayores niveles de madurez son aquellos que han logrado integrar la apertura de datos en políticas públicas transversales, promoviendo la interoperabilidad, la estandarización y la generación de valor económico y social. Este análisis resulta especialmente relevante para el caso español, ya que permite situar su nivel de desarrollo dentro del contexto europeo y comprender los desafíos que aún persisten en términos de calidad, sostenibilidad y reutilización. (European Data Portal, 2025).

3.5. Marco legal y estratégico de los datos abiertos en la Unión Europea

La Unión Europea ha desarrollado un marco normativo y estratégico que orienta la apertura y reutilización de los datos públicos en los Estados miembros, estableciendo estándares mínimos y promoviendo la generación de valor económico y social. La Directiva (UE) 2019/1024 constituye el eje jurídico central de este marco, definiendo obligaciones vinculantes para las administraciones públicas, tales como la publicación en formatos abiertos, la no discriminación en el acceso y la obligatoriedad de liberar los denominados conjuntos de datos de alto valor. Este instrumento legal sitúa la apertura de datos como un mandato estructural dentro de las políticas de transparencia e interoperabilidad, influyendo directamente en la madurez de los portales nacionales y autonómicos. Por su parte, el informe *The economic impact of open data* complementa esta visión al analizar el valor económico derivado de la reutilización de datos en Europa, demostrando que los datos abiertos no solo cumplen una función normativa, sino que representan un activo estratégico capaz de impulsar la innovación, la eficiencia administrativa y el crecimiento económico. En conjunto, ambos documentos permiten comprender que la madurez de un ecosistema de datos abiertos depende tanto del cumplimiento del marco legal como de su capacidad para generar impacto económico y social. (European Commission, 2019; European Commission, 2020).

3.6. Estudios de Madurez en el contexto español

La evolución de los portales de datos abiertos en el ámbito local español, es analizada subrayando que estas plataformas son herramientas críticas para la transparencia y la coproducción de servicios. A través de un estudio empírico de 66 municipios de más de 50.000 habitantes, los autores identifican que existe una marcada predilección por el desarrollo de dimensiones técnicas orientadas a la

transparencia, mientras que la dimensión social y el fomento de la reutilización de datos permanecen en un estado incipiente. Los investigadores concluyen que, a pesar del avance tecnológico, la gestión pública inteligente en los ayuntamientos españoles aún enfrenta el reto de integrar plenamente dinámicas colaborativas y participativas que permitan aprovechar el potencial máximo de los datos públicos abiertos. (Criado, Jiménez Cid y Alcaide Muñoz (2023)

Curto-Rodríguez analiza la evolución y madurez de los portales de datos abiertos de las 17 Comunidades Autónomas de España, destacando un crecimiento exponencial en la oferta de conjuntos de datos (datasets) entre 2013 y 2019. El autor emplea un modelo de evaluación multidimensional que trasciende el simple conteo de archivos, integrando variables como el número de aplicaciones derivadas, las opciones de interacción ciudadana y la funcionalidad técnica del portal. Los resultados del estudio revelan una heterogeneidad significativa entre las regiones, identificando importantes "diferencias intracomunitarias" en cuanto al desarrollo de ecosistemas de reutilización. El investigador concluye que, si bien la infraestructura técnica ha madurado, el desafío actual reside en equilibrar la oferta de datos con herramientas que fomenten la colaboración y la creación de valor público real (Curto-Rodríguez (2021)

El estudio desarrollado por Aparicio García Ecosistema de datos abiertos en la Universidad de Alicante, analiza la construcción del ecosistema de datos abiertos de la Universidad de Alicante, en el que destaca la importancia de integrar políticas institucionales, infraestructura tecnológica y cultura organizacional para consolidar iniciativas de apertura. El autor define y muestra que la madurez de un ecosistema universitario depende de la capacidad de articular procesos internos de gestión de datos, promover la interoperabilidad y fomentar la participación de la comunidad académica en la reutilización de la información. Este caso resulta especialmente relevante para el contexto español, ya que evidencia cómo las instituciones públicas pueden avanzar hacia modelos de apertura sostenibles mediante estrategias de gobernanza y estandarización. (Aparicio García, 2016).

El Portal de Datos Abiertos de la Junta de Andalucía presenta los Principios del Open Data como una actualización y enfatiza la necesidad de garantizar la accesibilidad, la interoperabilidad, la calidad y la reutilización de los datos públicos. Este documento aporta una visión práctica de cómo una administración autonómica interpreta y operacionaliza los principios fundamentales de la apertura, alineándose con las directrices y marco legal europeo y con los estándares FAIR. Esta contribución al estado de arte radica en que ofrece un ejemplo concreto de implementación

institucional en España, mostrando cómo los principios teóricos se traducen en políticas, criterios técnicos y mecanismos de publicación que influyen directamente en la madurez del ecosistema regional. (Junta de Andalucía, 2023).

3.7. Aportes desde América Latina y el Caribe al estudio de la madurez de los ecosistemas de datos

Muñoz y Clavijo ofrecen una visión estructurada sobre el estado de los ecosistemas de datos en América Latina, destacando que la madurez de estos entornos depende de la interacción entre actores institucionales, capacidades técnicas y marcos normativos. Los autores subrayan que la consolidación de un ecosistema de datos requiere no solo infraestructura tecnológica, sino también gobernanza, coordinación interinstitucional y mecanismos de participación que permitan que los datos generen valor social y económico. Esta perspectiva resulta especialmente relevante para el análisis del caso español, ya que pone de manifiesto que la madurez no puede evaluarse únicamente desde la disponibilidad de datos, sino desde la articulación de un sistema que facilite su uso efectivo (Muñoz y Clavijo, 2022).

Muente-Kunigami y Serale presentan el desarrollo de los datos abiertos en América Latina y el Caribe, resaltando que la apertura de datos sólo alcanza su potencial cuando se acompaña de políticas públicas sólidas, marcos regulatorios claros y estrategias de sostenibilidad. Los autores enfatizan que la madurez de un ecosistema de datos no depende únicamente de la publicación de conjuntos de datos, sino de la creación de condiciones que incentiven su reutilización, como la estandarización, la disponibilidad de metadatos y la existencia de comunidades activas de usuarios. Este planteamiento es aplicable al contexto español, donde la evaluación de la madurez del ecosistema debe considerar no solo la oferta institucional, sino también la demanda y el impacto generado por los datos abiertos. Así, el análisis de Muente-Kunigami y Serale (2018) aporta un marco conceptual útil para interpretar los avances y desafíos de España en términos de gobernanza, sostenibilidad y valor público derivado de los datos abiertos. (Muente-Kunigami y Serale, 2018).

Salas Rodríguez sostiene que la medición de la madurez en el uso de datos abiertos no debe ser un fin en sí mismo, sino un medio para identificar mejoras en las políticas públicas que incrementen el beneficio social. En su investigación, el autor

destaca la relevancia de los datos abiertos espaciales (vinculados a referencias geográficas) como un subtipo crítico para la toma de decisiones en áreas como la salud pública y la respuesta ante desastres. Al analizar el contexto peruano, Salas Rodríguez identifica que las infraestructuras de datos espaciales a menudo funcionan únicamente como repositorios de información y no como estructuras de interoperabilidad. Por ello, propone un modelo de madurez que permite establecer una hoja de ruta para superar barreras tecnológicas y legales, facilitando que las instituciones públicas transiten hacia una gestión de datos más colaborativa y alineada con las buenas prácticas globales de gobierno electrónico (Salas Rodríguez, 2023).

De Sousa y Fonseca, argumentan que la madurez de un portal de datos abiertos no solo se mide por la disponibilidad de la información, sino por su capacidad para generar conocimiento a través de una usabilidad y accesibilidad óptimas. Los autores subrayan que para maximizar el valor de los datos en entornos urbanos, es imperativo abordar el rediseño de las plataformas, asegurando que los formatos y contenidos faciliten el procesamiento y fomenten la confianza ciudadana. La investigación concluye que un enfoque en el diseño centrado en el usuario es esencial para potenciar la transparencia y la rendición de cuentas, permitiendo que los datos abiertos se conviertan en verdaderos catalizadores de la planificación urbana y la participación social efectiva (De Sousa et al., 2025).

El informe elaborado por Naser y Rosales ofrece una visión integral sobre el desarrollo de los datos abiertos en América Latina y el Caribe, identificando avances significativos, pero también brechas estructurales que limitan la consolidación de ecosistemas maduros. Los autores destacan que, aunque la región ha experimentado un crecimiento notable en la creación de portales y en la adopción de políticas de transparencia, persisten desafíos relacionados con la calidad de los datos, la sostenibilidad institucional, la interoperabilidad y la falta de mecanismos que incentiven la reutilización. También resaltan que la madurez de los ecosistemas latinoamericanos depende de la capacidad de los gobiernos para articular marcos normativos sólidos, fortalecer capacidades técnicas y promover la participación ciudadana como eje de valor público. Este análisis resulta especialmente útil para contextualizar el caso español, ya que permite contrastar modelos de gobernanza, identificar buenas prácticas y comprender cómo factores institucionales y culturales influyen en el desarrollo de iniciativas de datos abiertos. (Naser & Rosales, 2016).

El trabajo de Rodríguez Herrera y Alvarado estudia los ecosistemas de innovación social en universidades latinoamericanas, destacando que la generación de valor público depende de la colaboración entre actores académicos, gubernamentales y comunitarios. Aunque no se centra exclusivamente en datos abiertos, su aporte es significativo porque demuestra que los ecosistemas basados en conocimiento requieren estructuras de gobernanza participativa, flujos de información abiertos y dinámicas de co-creación. Esta perspectiva complementa el análisis de la madurez de los datos abiertos al subrayar que la apertura no es únicamente un proceso técnico, sino un componente esencial para habilitar la innovación social y fortalecer la interacción entre instituciones y ciudadanía. (Rodríguez Herrera & Alvarado, 2015).

3.8. Aportes teóricos y metodológicos complementarios

La consolidación de ecosistemas de datos abiertos maduros requiere comprender cómo los datos deben ser gestionados para facilitar su reutilización, y en este ámbito los principios FAIR se han convertido en un referente internacional y establecen en sus principios que los datos deben ser encontrables, accesibles, interoperables y reutilizables, lo que permiten comprender que la madurez de un ecosistema de datos abiertos no depende únicamente de la publicación de información, sino de datos estructurados, de identificadores persistentes y mecanismos que garanticen su accesibilidad a largo plazo. Este marco conceptual aporta a la investigación una base sólida para evaluar si los portales españoles cumplen con estándares que permitan datos estructurados, situando la madurez como un atributo técnico, y a su vez como una condición que emerge de prácticas coherentes de gestión de datos (Wilkinson et al., 2016).

La calidad de los datos es un componente esencial para determinar el grado de madurez de un ecosistema de datos abiertos. Batini y Scannapieco desarrollan un marco metodológico que permite evaluar dimensiones como exactitud, completitud, consistencia, actualidad y accesibilidad, elementos que influyen directamente en la capacidad de los usuarios para transformar los datos en conocimiento. Su trabajo aporta a esta investigación una perspectiva crítica sobre la necesidad de que los portales no solo publiquen datos, sino que garanticen estándares mínimos de calidad que permitan su reutilización. Desde esta óptica, la madurez del ecosistema español

debe analizarse considerando si los datos publicados cumplen con criterios que aseguren su fiabilidad y utilidad (Batini y Scannapieco, 2016).

La interoperabilidad semántica es un desafío recurrente en los portales de datos abiertos, especialmente cuando se busca integrar información procedente de múltiples fuentes. Zaveri, donde realizan una revisión exhaustiva de métricas y métodos para evaluar la calidad de los datos enlazados, identificando problemas comunes como la falta de estandarización, la ausencia de metadatos adecuados o la inconsistencia en los vocabularios utilizados. Su aporte resulta fundamental para esta investigación, ya que permite comprender que la madurez de un ecosistema no depende únicamente de la cantidad de datos publicados, sino de su capacidad para conectarse y dialogar con otros conjuntos de datos. En el contexto español, este enfoque ayuda a evaluar si los portales avanzan hacia modelos de interoperabilidad que faciliten la creación de valor público. (Zaveri et al., 2016).

La estructura interna de los portales de datos es un elemento clave para garantizar la encontrabilidad y organización de los conjuntos de datos. Neumaier, Umbrich y Polleres, analizan cómo los portales implementan el estándar DCAT y proponen mapeos multinivel que permiten mejorar la navegación, clasificación y descubrimiento de datasets. Su trabajo aporta a esta investigación una comprensión más profunda de la importancia de los metadatos y de la arquitectura de la información en la madurez de los portales. En el caso español, donde DCAT-AP se ha adoptado como estándar europeo, este enfoque permite evaluar si los portales han logrado implementar estructuras que faciliten la reutilización y el acceso eficiente a los datos (Neumaier, Umbrich y Polleres, 2016).

El análisis de la interoperabilidad y los metadatos en el contexto español ha sido abordado por Maseda-Seco, Bueno-de-la-Fuente y Méndez, quienes estudian los datos abiertos de las bibliotecas municipales y evidencian la existencia de inconsistencias en la calidad de los metadatos, así como la falta de estandarización en los formatos utilizados. Su investigación aporta evidencia empírica relevante para esta tesis, ya que demuestra que la madurez del ecosistema español presenta avances desiguales entre sectores y administraciones. Además, proponen mejoras orientadas a la apertura y reutilización, lo que permite identificar brechas y oportunidades para fortalecer la calidad y la interoperabilidad en los portales públicos (Maseda-Seco, Bueno-de-la-Fuente y Méndez, 2017).

3.9. Ecosistemas de datos e innovación social

Los ecosistemas de datos abiertos no pueden entenderse únicamente desde una perspectiva técnica. Su madurez depende también de la capacidad de las instituciones para generar valor público y fomentar la colaboración entre actores.

Los ecosistemas de innovación social surgen cuando existe interacción entre instituciones, academia y ciudadanía, lo que resulta plenamente aplicable al ámbito de los datos abiertos. (Rodríguez Herrera y Alvarado, 2015)

En América Latina, existen estudios donde se evidencia que la madurez de los ecosistemas depende de factores institucionales, culturales y normativos, ofreciendo una perspectiva comparada útil para el caso español (Muñoz y Clavijo, 2022) y Naser y Rosales (2016)

En el ámbito nacional, Bello García (s. f.) destaca que los datos abiertos pueden fortalecer la participación social, mientras que (Finol, 2023) advierte sobre la necesidad de equilibrar apertura y privacidad. Estas perspectivas permiten comprender que la madurez de un ecosistema de datos abiertos no solo depende de su infraestructura tecnológica, sino también de su capacidad para generar impacto social, promover la transparencia y habilitar procesos de co-creación.

La dimensión social de los datos abiertos es un componente frecuentemente subestimado en los modelos de madurez. Bello García (s. f.) sostiene que los datos abiertos deben entenderse como herramientas para fortalecer la participación ciudadana y la colaboración entre gobierno y sociedad. Su análisis destaca que la madurez no se limita a la infraestructura tecnológica, sino que implica la capacidad de las instituciones para generar valor público a través de la transparencia, la interacción y la apertura institucional. Este enfoque complementa la perspectiva técnica de otros autores y permite situar la madurez del ecosistema español dentro de un marco más amplio de gobernanza democrática.

3.10. Síntesis crítica del estado del arte y justificación de la contribución

La revisión del estado del arte muestra que la madurez de los datos abiertos ha sido abordada desde enfoques académicos e institucionales que han contribuido a consolidar dimensiones ampliamente aceptadas como política, portal, calidad e impacto y, en el plano técnico, criterios relacionados con accesibilidad,

interoperabilidad, estandarización de metadatos y disponibilidad de formatos. En el caso europeo, los informes de madurez han permitido comparaciones agregadas entre países, mientras que los estudios en España han evidenciado avances desiguales entre administraciones y una heterogeneidad significativa en el desarrollo de los portales autonómicos y municipales.

Sin embargo, el análisis crítico revela un hueco recurrente: la mayoría de evaluaciones se apoyan en indicadores macro o declarativos, en métricas internas de los portales o en aproximaciones parcialmente cualitativas, lo que dificulta disponer de una evaluación reproducible, comparable y a nivel de dataset que permita caracterizar el estado real del ecosistema español con evidencia homogénea. En particular, la literatura muestra limitaciones para operacionalizar, de forma automatizada y consistente entre portales, aspectos técnicos clave que determinan la madurez efectiva: accesibilidad real de enlaces y recursos, calidad y completitud de metadatos, trazabilidad temporal (fechas y actualización) e interoperabilidad (formatos, vocabularios/estándares y consistencia semántica).

A partir de este vacío, el presente Trabajo Fin de Máster se justifica por una contribución original: el diseño y aplicación de un enfoque empírico y automatizable para medir la madurez del ecosistema de datos abiertos en España, basado en la extracción sistemática de metadatos y en la construcción de métricas comparables entre portales. Esta contribución no se limita a describir el estado formal de los portales, sino que materializa una evaluación cuantitativa reproducible, generando un conjunto de indicadores y un índice sintético que permiten comparar portales, identificar brechas por dimensión y establecer un ranking de madurez sustentado en evidencia.

De este modo, el trabajo aporta un marco metodológico que complementa las aproximaciones existentes: traduce dimensiones teóricas ampliamente aceptadas en medidas operativas, comparables y escalables, y ofrece resultados útiles tanto para investigación como para toma de decisiones públicas, al identificar dónde se concentra la falta de madurez técnica en el ecosistema español y qué dimensiones requieren priorización.

Capítulo 4. Metodología y Tecnologías

4.1. Enfoque metodológico general

La investigación se desarrolla bajo un enfoque empírico orientado a la evaluación de la madurez técnica y estructural de portales de datos abiertos en España, basado en el análisis sistemático de los metadatos publicados por distintos portales institucionales. La metodología se articula mediante un pipeline reproducible y escalable, diseñado para ser aplicado de forma homogénea a múltiples portales y a un volumen elevado de datasets.

El objetivo metodológico es caracterizar el grado de preparación de los portales para su integración en infraestructuras de datos abiertas y sistemas digitales avanzados, evaluando dimensiones clave como la accesibilidad, la trazabilidad, la interoperabilidad técnica y semántica, y la calidad de los metadatos. Este enfoque permite realizar comparaciones sistemáticas entre portales y analizar patrones estructurales comunes dentro del ecosistema de datos abiertos.

4.2. Arquitectura del pipeline de análisis

El análisis se articula mediante un pipeline estructurado en cuatro fases consecutivas, diseñado para transformar metadatos heterogéneos en métricas e indicadores de madurez comparables entre portales. La arquitectura del pipeline permite separar claramente las tareas de extracción, validación, análisis y generación de resultados, garantizando la reproducibilidad del proceso y la trazabilidad de cada etapa.

Fase 1. Extracción, selección y normalización de metadatos

En la primera fase se realiza el proceso automatizado para la extracción de los metadatos publicados por cada portal de datos abiertos, para su estudio, utilizando los mecanismos de acceso disponibles según la tecnología del portal (APIs y endpoints específicos), interactuando con las api. A partir de esta extracción inicial, se aplican los criterios de selección definidos para identificar datasets aptos para el análisis de madurez.

Los metadatos extraídos se someten a un proceso de normalización, orientado a homogeneizar estructuras, tipos de datos y denominaciones de campos. Este

proceso incluye la estandarización de identificadores, fechas de publicación y actualización, formatos, licencias, categorías temáticas y mecanismos de acceso, permitiendo construir tablas de metadatos consistentes y comparables entre portales.

PORTAL	URL de acceso	N.º datasets
Datos Portal Nacional	https://datos.gob.es/es/	450
Opendata Barcelona	https://opendata-ajuntament.barcelona.cat/	552
Ayuntamiento Junta Andalucía	https://www.juntadeandalucia.es/datosabiertos	600
Datos Comunidad Madrid	https://datos.comunidad.madrid/	600
Datos abiertos Región Murcia	https://datosabiertos.regiondemurcia.es/	598
Análisis de Datos abiertos Castilla y León	https://analisis.datosabiertos.jcyl.es/	419
Opendata Aragon	https://opendata.aragon.es/datos	600
TOTAL		3819

Tabla 4.1. Selección, extracción de datasets, por Portal

Como resultado de esta fase se obtiene, para cada portal, una tabla de metadatos consolidada que constituye la base del análisis posterior. Cada portal aporta aproximadamente **600 datasets**, seleccionados conforme a los criterios metodológicos establecidos.

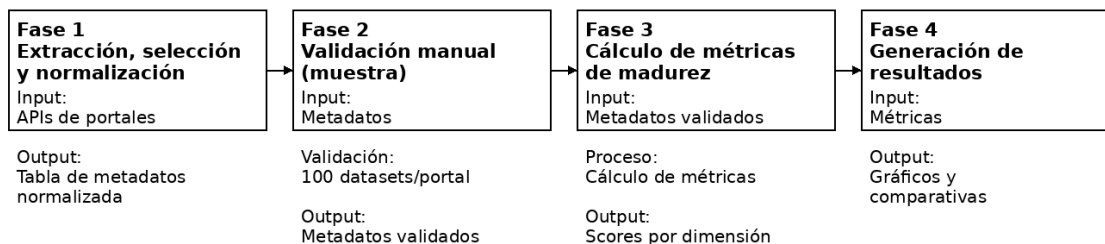


Figura 4.1. Diagrama - Arquitectura del pipeline de análisis

Figura 4.1. Arquitectura del pipeline de análisis de madurez de portales de datos abiertos: El pipeline se estructura en cuatro fases consecutivas: extracción y normalización de metadatos, validación manual de una muestra, cálculo de métricas de madurez y generación de resultados. Cada fase transforma los resultados de la etapa anterior, garantizando trazabilidad, reproducibilidad y comparabilidad entre portales.

Fase 2. Validación manual de una muestra

Con el fin de complementar los procesos automatizados y detectar incidencias no identificables mediante extracción programática, se incorpora una fase de validación manual de los resultados obtenidos en la fase anterior.

Para cada portal se selecciona una **muestra aleatoria de 100 datasets**, representativa del conjunto total analizado. La validación manual se centra en la revisión directa de los valores de los metadatos que conforman las tablas de resultados, verificando su coherencia y validez efectiva.

En particular, se comprueba manualmente la accesibilidad y consistencia de campos críticos como comprobar descargas `download_url`, comprobar accesos `dataset_uri`, verificar fechas de publicación (`issued`) y actualización (`modified`), licencias disponibles `license`, entidad publicadora `publisher`, frecuencia de actualización `update_frecuency`, título del dataset `title`, formatos disponibles `format`, categorías definidas `category`, y mecanismos de acceso mediante APIs. Esta fase permite identificar errores frecuentes como enlaces rotos, discrepancias entre formatos declarados y archivos accesibles, licencias ambiguas o metadatos

incompletos, aportando una capa adicional de control de calidad al análisis de madurez.

Fase 3. Cálculo de métricas de madurez

A partir de los metadatos normalizados y validados, se procede al cálculo de las métricas de madurez definidas para el estudio. Estas métricas permiten evaluar de forma sistemática las condiciones técnicas y estructurales, las dimensiones de interoperabilidad técnica, interoperabilidad semántica, accesibilidad, trazabilidad y calidad de los metadatos, proporcionando indicadores comparables a nivel de dataset y de portal.

Cada dimensión se operacionaliza mediante indicadores contruidos a partir de los valores observados en los metadatos, permitiendo obtener puntuaciones comparables a nivel de dataset y de portal. El objetivo de esta fase es caracterizar el grado de preparación estructural de los portales para integrarse en infraestructuras de datos abiertas y sistemas digitales avanzados.

El cálculo de las métricas se realiza de forma automatizada, garantizando consistencia en la aplicación de los criterios y permitiendo la replicabilidad del análisis en distintos portales o periodos temporales.

Fase 4. Generación de resultados

La fase final del pipeline se orienta a la síntesis, análisis y comunicación de los resultados, transformando las métricas calculadas en evidencias interpretables sobre el nivel de madurez de los portales de datos abiertos.

Los resultados se presentan mediante representaciones gráficas (gráficos de barras, distribuciones y comparativas), que permiten analizar el comportamiento de las distintas dimensiones de madurez, comparar portales entre sí y explorar diferencias por categorías temáticas y características de los datasets. Asimismo, se incluyen análisis agregados que facilitan la identificación de patrones recurrentes, fortalezas estructurales y debilidades comunes en el ecosistema de datos abiertos.

4.3. Selección de datasets y criterios de inclusión

Para garantizar la comparabilidad y el rigor metodológico del análisis, todos los portales se someten a un conjunto homogéneo de criterios de selección de datasets. Estos criterios priorizan la disponibilidad de metadatos accesibles, la existencia de información clave para el análisis de madurez y la representatividad temática.

Se incluyen datasets con metadatos disponibles vía API o descarga directa, licencias identificables, formatos estructurados preferentes (CSV, JSON, XML, RDF), Antigüedad mínima de publicación: Datasets con al menos 12 meses de antigüedad desde su publicación, suficiente para análisis temporal, presencia de campos claves como identificadores o URLs, fechas, publicador, formato, licencia y mecanismos de acceso. Categorías temáticas normalizadas: Los metadatos del portal se reclasifican en una taxonomía común. Selección dentro de categorías representativas y comparables entre portales (transporte, salud, educación, medio ambiente, Demografía y Estadística, ciencia y tecnología), para garantizar consistencia entre portales.

Adicionalmente, cuando los portales publican métricas internas de uso, se priorizan datasets con mayor actividad aparente. Se busca también diversidad de organismos publicadores y se valoran datasets con estructuras tabulares claras, metadatos semánticamente informativos y, cuando existe, documentación asociada como diccionarios de datos o el uso de estándares como DCAT y vocabularios controlados.

En términos operativos, se establece un volumen objetivo por portal, adaptado a la disponibilidad real de datasets que cumplen los criterios, con un rango de referencia aproximado (150–600 datasets), por cada Portal.

4.4. Construcción de métricas de madurez de los portales

En esta fase se procede al cálculo de las métricas de madurez del ecosistema de datos abiertos a partir de los metadatos previamente extraídos y normalizados. Dado el carácter heterogéneo de los metadatos que incluyen variables binarias, categóricas y textuales, fue necesario definir criterios explícitos de operacionalización, garantizando coherencia, reproducibilidad y comparabilidad entre datasets y portales.

Cada dimensión de la madurez se construye a partir de indicadores observables, que se transforman en variables binarias o porcentuales según su naturaleza. Posteriormente, estas variables se agregan mediante promedios simples para obtener puntuaciones normalizadas en una escala común de 0 a 100.

4.4.1. Operacionalización de los indicadores

Dado que los metadatos analizados presentan heterogeneidad estructural (variables binarias, textuales y categóricas), fue necesario definir reglas explícitas de operacionalización para cada indicador, con el objetivo de garantizar consistencia, comparabilidad y reproducibilidad.

Cada indicador se transforma en una variable binaria, donde:

- **El valor 1** indica que el criterio de madurez se cumple.
- **El valor 0** indica que el criterio no se cumple.
- **Valores ausentes, nulos o no definidos** se consideran explícitamente como **0**, al representar una ausencia de evidencia del criterio evaluado.

Esta decisión metodológica responde a un enfoque conservador, ampliamente utilizado en estudios de evaluación de calidad de datos abiertos, donde la ausencia de información se interpreta como una limitación estructural del ecosistema.

Ejemplo de reglas de codificación

Indicadores binarios nativos

(has_open_format, portal_has_api_rest, uses_controlled_vocab):

- 1 → el campo existe y su valor indica cumplimiento explícito.
- 0 → el campo existe pero indica no cumplimiento, o el campo no existe.

Indicadores textuales que toman valores binarios (publisher, title, description, category_defined):

- 1 → el campo contiene texto no vacío y semánticamente válido.

- 0 → el campo está vacío, es nulo o contiene expresiones como “no definido”, “consultar”, “none”, “null”.

Indicadores temporales

(modified, update_frequency):

- Se transforman primero en variables numéricas o categóricas.
- Posteriormente se evalúan mediante reglas explícitas (p. ej. antigüedad ≤ 6 meses).
- Si el valor no puede calcularse, el indicador toma valor 0.

Tratamiento de datos ausentes

En todos los casos, la ausencia de metadatos no se imputa ni se ignora, sino que se interpreta como una señal de baja madurez, asignando valor 0 al indicador correspondiente. Este enfoque permite capturar no solo la calidad de los valores presentes, sino también la completitud estructural del metadato, que es un elemento clave de la madurez del ecosistema.

4.4.2. Trazabilidad global

La trazabilidad global evalúa la capacidad de los datos para ser identificados, referenciados y contextualizados a lo largo de su ciclo de vida. Esta dimensión se concibe como un constructo multidimensional compuesto por trazabilidad de origen, trazabilidad única y trazabilidad temporal, cada una operacionalizada mediante indicadores específicos.

Esta trazabilidad considera la existencia de identificadores o URLs estables, la coherencia de la información temporal (fechas de publicación y actualización), la claridad de la procedencia del dataset y la consistencia de la información sobre el publicador.

a) Trazabilidad de origen

La trazabilidad de origen mide la capacidad de identificar de forma inequívoca quién publica el dataset y cómo se identifica. Se construye a partir de la presencia de los siguientes metadatos:

- publisher (entidad responsable de la publicación),

- identifier (identificador interno o externo del dataset),
- dataset_uri (URI persistente del recurso).

Dado que estos campos son de tipo textual, se consideran válidos cuando contienen información no vacía y semánticamente significativa. Para ello:

- Se normalizan los valores textuales.
- Se excluyen expresiones como "no definido", "consultar", valores nulos o cadenas vacías. (Tomará valor 0)
- Cada campo se transforma en una variable binaria:
1 si el texto está presente y es válido, 0 en caso contrario.

Cada indicador se codifica de forma binaria (1 si está presente y correctamente definido, 0 en caso contrario).

La puntuación de trazabilidad de origen se obtiene como el promedio del porcentaje de datasets que cumplen cada uno de estos indicadores:

$$\text{Trazabilidad}_{\text{origen}} = \frac{\text{publisher} + \text{identifier} + \text{dataset_uri}}{3} \times 100$$

b) Trazabilidad única

La trazabilidad única evalúa si el dataset dispone de identificadores persistentes o mecanismos de acceso estable de un dataset, que permitan su referencia inequívoca a lo largo del tiempo y entre sistemas.

Se consideran los siguientes indicadores:

- Presencia de un **DOI** u otro identificador persistente (doi),
- Disponibilidad de un **acceso público verificable** al recurso (public_access_ok o equivalente).

Ambos indicadores se tratan como variables binarias, dónde: presencia o cumplimiento, si encuentra el texto lo transforma en **1**. Si no lo encuentra lo nombra a

0 que indica ausencia.

La puntuación se calcula como el promedio simple del porcentaje de datasets que cumplen cada uno de estos criterios:

$$\text{Trazabilidad}_{\text{única}} = \frac{\text{doi} + \text{public_access_ok}}{2} \times 100$$

c) Trazabilidad temporal

La trazabilidad temporal mide la capacidad de conocer la vigencia y el mantenimiento de los datos (cuándo fue actualizado un dataset), combinando información objetiva sobre fechas con información declarativa sobre frecuencia de actualización, lo cual resulta esencial para evaluar su vigencia y fiabilidad.

Esta subdimensión se construye a partir de dos indicadores complementarios:

1. Antigüedad de la última actualización, calculada a partir del campo `modified`.
 - Se calcula el número de meses transcurridos desde la última actualización.
 - Se considera una condición madura cuando el dataset ha sido actualizado en los últimos 0–6 meses.
 - Este criterio se transforma en un porcentaje de datasets que cumplen dicha condición.
2. Frecuencia de actualización declarada, a partir del campo `update_frequency`.
 - Los valores textuales se normalizan y agrupan en categorías.
 - Se considera madura la pertenencia al Grupo 1, que incluye frecuencias altas o medias (instantánea, horaria, diaria, semanal, mensual, trimestral o semestral).

- Se calcula el porcentaje de datasets cuya frecuencia pertenece a este grupo.

La puntuación de trazabilidad temporal se obtiene como el promedio de ambos porcentajes.

$$\text{Trazabilidad}_{\text{temporal}} = \frac{\% \text{datasets con modified} \leq 6 \text{ meses} + \% \text{datasets en Grupo 1}}{2}$$

d) Cálculo de la trazabilidad global

La trazabilidad global se define calculando el promedio simple de las tres subdimensiones anteriores (origen, única y temporal), obteniendo una puntuación final normalizada entre 0 y 100.

$$\text{Trazabilidad}_{\text{global}} = \frac{\text{Trazabilidad}_{\text{origen}} + \text{Trazabilidad}_{\text{única}} + \text{Trazabilidad}_{\text{temporal}}}{3}$$

4.4.3. Interoperabilidad técnica

Evalúa la capacidad del portal para ofrecer acceso automatizado y consumo programático de los datos, considerando la disponibilidad de formatos abiertos, mecanismos de acceso estandarizados. Esta dimensión refleja la base técnica mínima necesaria para la integración de los datos en sistemas externos.

Esta dimensión se construye a partir de indicadores derivados de los metadatos técnicos de los datasets, los cuales son previamente normalizados y transformados en variables binarias para garantizar la comparabilidad entre portales.

El indicador **has_open_format** evalúa la disponibilidad de formatos abiertos en cada dataset. Para su cálculo, el campo format es normalizado y descompuesto en una lista de formatos individuales, los cuales son clasificados como abiertos o no abiertos en función de un conjunto predefinido de formatos abiertos (por ejemplo, CSV, JSON, XML, RDF, JSON-LD). El indicador adopta un valor binario, tomando el valor 1 si el dataset ofrece al menos un formato abierto y 0 en caso contrario o cuando el campo no está definido.

El indicador **portal_supports_dcat_dcatap** refleja la adopción de estándares de interoperabilidad semántica a nivel de portal y se trata igualmente como una variable binaria, tomando el valor 1 cuando el portal declara soporte para DCAT/DCAT-AP y 0 cuando dicho soporte no está presente o no puede ser verificado.

Ambos indicadores se tratan como variables binarias.

Se normaliza la lista de formatos.

- Se clasifica cada formato como abierto o no abierto (valores 0 /1)
- Se considera cumplimiento si existe al menos un formato abierto (valores 0 /1)

La puntuación de interoperabilidad técnica se calcula como el promedio del porcentaje de datasets que cumplen cada uno de estos indicadores, obteniendo así un valor normalizado en el rango [0, 100] a nivel de portal.

$$\text{Interoperabilidad}_{\text{técnica}} = \frac{\text{has_open_format} + \text{dcat}}{2} \times 100$$

4.4.4. Interoperabilidad semántica

Analiza el grado en que los metadatos permiten una interpretación coherente y compartida de los datos. Se consideran aspectos como el uso de vocabularios controlados, categorizaciones consistentes y el uso de formatos serializados, que faciliten la integración de datasets en distintos contextos.

Se construye a partir de los siguientes indicadores:

- **category**, es tratada como **texto válido** y se transforma en binaria derivada (**category_defined**) toma el valor **1**: si existe una categoría temática con texto válido, y **0**: campo vacío/no definido.
- **uses_controlled_vocab** (binaria): Toma el valor de **1** si el dataset/portal declara uso de vocabulario controlado. Al consultar directamente en la api, y valor **0**, cuando no lo declara o no está disponible.
- **has_semantic_serialization** (binaria): **1** si hay serialización semántica (p.ej. RDF/TTL/JSON-LD o flag equivalente), y valor **0** cuando hay ausencia o no lo cumple.

Los campos textuales se validan mediante presencia de contenido significativo, en el caso de categoría se normaliza en una nueva variable binaria `category_defined`, mientras que los campos técnicos se evalúan de forma binaria directamente. La puntuación final corresponde al promedio simple de los tres indicadores.

$$\text{Interoperabilidad}_{\text{semántica}} = \frac{\text{category} + \text{vocabulario} + \text{serialización}}{3} \times 100$$

4.4.5. Accesibilidad

Mide la facilidad con la que los datasets pueden ser accedidos de forma efectiva, incluyendo la estabilidad de los puntos de acceso, la disponibilidad real de los enlaces de descarga, las licencias permitidas, y la ausencia de barreras técnicas que limiten el consumo automatizado de los datos, evaluar las condiciones prácticas para que los datos puedan ser localizados, accedidos y reutilizados.

Se consideran los siguientes indicadores:

- Si un campo no existe en los metadatos, se toma como incumplimiento (0).
- No se imputan valores positivos.
- Se adopta un criterio conservador, habitual en evaluaciones de madurez.

En primer lugar, la disponibilidad de una API REST (**portal_has_api_rest**) indica si el portal ofrece mecanismos de acceso programático a los datasets; este indicador toma valor 1 cuando la API está disponible y 0 en caso contrario o cuando la información no está presente en los metadatos.

En segundo lugar, la licencia abierta (**license_open**) evalúa si el dataset se publica bajo una licencia que permite su reutilización; se asigna valor 1 cuando la licencia es abierta y 0 cuando es cerrada, ambigua o inexistente.

El tercer indicador corresponde al acceso público efectivo (**public_access_ok**), que refleja si los recursos asociados al dataset pueden ser accedidos sin barreras técnicas o restricciones de autenticación; toma valor 1 cuando el acceso es efectivo y 0 cuando no lo es o no puede verificarse.

Finalmente, la presencia de una URL de descarga (**download_url_present**) indica si existe al menos un enlace válido para la descarga directa de los datos, asignando valor 1 cuando dicho enlace está presente y 0 en caso contrario.

La puntuación de accesibilidad se calcula como el promedio del porcentaje de datasets que cumplen cada criterio.

La puntuación se calcula como:

$$\text{Accesibilidad}_{dataset} = \frac{\text{portal_has_api_rest} + \text{license_open} + \text{public_access_ok} + \text{download_url_present}}{4} \times 100$$

4.4.6. Calidad de los metadatos

Analiza la completitud, coherencia y precisión de la información descriptiva asociada a los datasets. Esta dimensión actúa de forma transversal, ya que condiciona directamente la accesibilidad, la interoperabilidad y la trazabilidad, siendo esencial para garantizar un uso efectivo y sostenible de los datos abiertos.

Se evalúa mediante:

- Presencia de un título (title - texto válido)
- Presencia de una descripción adecuada (description - (texto válido)

Estos indicadores, se normalizan a valores binarios, tomando como valor 1 cuando hay cumplimiento, es decir presencia del texto, y valor 0 cuando está vacío.

- Disponibilidad de un diccionario de datos (has_data_dictionary - variable binaria). Entre 0 y 1, este valor se extrae directamente de la api es decir 1 si usa diccionario de datos y 0 cuando no usa.

La puntuación se obtiene como el promedio simple de los tres indicadores.

$$\text{Calidad}_{metadatos} = \frac{\text{título} + \text{descripción} + \text{diccionario}}{3} \times 100$$

4.4.7. Calcular el Índice global de madurez del portal

El **Índice Global de Madurez** sintetiza las cinco dimensiones evaluadas en un único indicador agregado. Se calcula como el promedio ponderado de las dimensiones, asignando **pesos iguales (20%)** a cada una de ellas.

Este índice permite comparar portales de distintos niveles administrativos y tecnológicos, manteniendo una interpretación homogénea y transparente del proceso de cálculo.

$$\text{Índice Global} = \frac{\text{Trazabilidad}_{\text{global}} + \text{Interoperabilidad}_{\text{técnica}} + \text{Interoperabilidad}_{\text{semántica}} + \text{Accesibilidad} + \text{Calidad}_{\text{metadatos}}}{5}$$

Finalmente, el índice se clasifica en tres niveles de madurez:

- **Bajo:** 0–30
- **Medio:** 30–70
- **Alto:** 70–100

Esta clasificación facilita la interpretación comparativa de los resultados y permite identificar fortalezas y debilidades estructurales en los portales analizados.

4.5. Generación de resultados y visualización

La fase final se orienta a sintetizar y comunicar los resultados del análisis, transformando los indicadores calculados en evidencias comparables sobre el nivel de madurez de los portales. Se generan resultados agregados por portal y por categoría temática, permitiendo identificar fortalezas y debilidades estructurales en cada dimensión evaluada.

La visualización se implementa mediante gráficos orientados a: (i) mostrar el índice global de madurez, (ii) desagregar resultados por dimensión, (iii) comparar portales y categorías, y (iv) explorar relaciones temporales como antigüedad y frecuencia de actualización cuando la información se encuentra disponible en los metadatos.

En conjunto, esta fase cierra el pipeline metodológico proporcionando una visión integrada y comprensible del estado de madurez del ecosistema de datos abiertos, sirviendo como base para la interpretación crítica de los resultados, la formulación de recomendaciones y el análisis comparativo desarrollado en los capítulos posteriores.

4.6. Implementación tecnológica y reproducibilidad

El diseño del pipeline se basa íntegramente en el lenguaje Python (v3.10), seleccionado por su robustez en tareas de ciencia de datos y su capacidad de integración con fuentes gubernamentales. La implementación se estructura de forma modular para garantizar que cada fase sea independiente, auditable y fácilmente adaptable a diferentes interfaces de programación.

4.6.1. Librerías y Entorno de Procesamiento

La manipulación de datos se gestiona mediante la librería Pandas, que actúa como el motor de transformación (ETL), para la estructuración de datos , mientras que la comunicación con las API gubernamentales se realiza a través de Requests mediante protocolos HTTP/REST. Para el análisis textual y la normalización de metadatos, se emplean expresiones regulares (re) y el formato json.

El entorno de ejecución se ha centralizado en Anaconda (Jupyter Notebooks) por su versatilidad para el prototipado rápido y la visualización de resultados, con scripts adaptados a la tecnología de acceso de cada portal (CKAN, endpoints DCAT/SPARQL, Opendatasoft, APIs propias), y diseñados de forma modular para facilitar su extensión.

En cuanto a la presentación de hallazgos, el sistema genera dos tipos de productos:

1. Reportes Estructurados: Archivos en formatos CSV y Excel para el almacenamiento de métricas.
2. Visualizaciones: Gráficos estáticos con Seaborn e interactivos mediante Plotly, exportados en formato HTML para facilitar su análisis web.

El proceso completo se documenta y versiona para garantizar la reproducibilidad del análisis y la trazabilidad de las decisiones metodológicas adoptadas.

4.6.2. Repositorio, documentación, herramientas.

La integridad de la investigación se apoya en un repositorio público alojado en GitHub, que centraliza el código fuente del pipeline generado. Para garantizar la transferencia de conocimiento y la transparencia, el software se distribuye bajo la licencia MIT, permitiendo la libre reutilización y modificación del código siempre que se preserve la autoría original.

https://github.com/YuliZambrano/TFM_MIOT_YZG

En el proceso de desarrollo y estructuración lógica del código, se ha integrado el uso de asistentes de inteligencia artificial (Gemini y ChatGPT) como herramientas de apoyo para la optimización de algoritmos y la depuración de scripts.

4.6.3. Automatización y Escalabilidad

El pipeline destaca por su diseño modular y extensible. Cada portal de datos abiertos cuenta con un script independiente ajustado a sus especificaciones técnicas, pero todos mantienen una estructura de salida común. Este enfoque por lotes asegura que el sistema pueda escalar a nuevos portales sin alterar la lógica de cálculo de las métricas finales, garantizando una trazabilidad completa desde la extracción hasta el reporte final. El código incluye funciones específicas para verificar si el portal expone metadatos bajo el estándar DCAT/DCAT-AP, alineándose con las normativas europeas.

4.7. Análisis exploratorio de trazabilidad externa (aporte metodológico complementario)

De forma complementaria al análisis principal, se desarrolla un script exploratorio orientado a la detección de huellas de reutilización externa de los datasets publicados en portales de datos abiertos. Este análisis tiene un carácter estrictamente exploratorio y no forma parte del cálculo de las métricas ni del índice de madurez.

El objetivo de este aporte metodológico es investigar la viabilidad de detectar referencias externas a datasets mediante fuentes abiertas, así como documentar las limitaciones técnicas y estructurales que dificultan la trazabilidad externa en el ecosistema actual. Los resultados de este análisis se consideran como evidencia

contextual y como base para posibles líneas de investigación futuras, sin condicionar las conclusiones del análisis de madurez desarrollado en este trabajo.

Capítulo 5 - Diseño de la investigación y fuentes de datos

5.1. Introducción al diseño de la investigación

Tras haber definido la metodología en el capítulo anterior, este capítulo describe cómo dichas decisiones se materializaron en la práctica y el proceso seguido para ejecutar el estudio. Se detalla la implementación real del pipeline de análisis, la selección de los portales y datasets, así como las decisiones operativas adoptadas durante la recopilación y el proceso seguido para la preparación de los datos.

Este capítulo tiene como objetivo describir el proceso de construcción del conjunto de datos analizado, documentando las elecciones realizadas, las dificultades encontradas y las estrategias empleadas para resolverlas. De este modo, se garantiza la trazabilidad del proceso y la coherencia entre la metodología definida y su aplicación efectiva.

5.2. Selección de portales de datos abiertos

La selección de los portales de datos abiertos se realizó con el objetivo de representar la diversidad institucional, tecnológica y organizativa del ecosistema de datos abiertos en España. Se incluyeron portales de ámbito nacional, autonómico y local, priorizando aquellos con un volumen significativo de datasets y con mecanismos de acceso programático disponibles, con el fin de analizar cómo las diferencias organizativas, normativas y técnicas influyen en el grado de madurez de las infraestructuras de datos abiertos.

El conjunto final de portales analizados está compuesto por el portal nacional de datos abiertos, varios portales autonómicos, correspondientes a comunidades autónomas con estrategias consolidadas en materia de Open Data, y portales municipales de grandes ciudades, caracterizados por un elevado volumen de datasets y un uso intensivo de tecnologías de publicación de datos. Esta combinación permite evaluar el ecosistema desde una perspectiva comparativa y multinivel, evitando conclusiones basadas en un único tipo de administración.

En todos los casos, la inclusión de un portal estuvo condicionada por la disponibilidad efectiva de metadatos accesibles mediante APIs o endpoints públicos,

así como por la posibilidad de aplicar de forma homogénea los criterios metodológicos definidos en el estudio. Asimismo, se priorizaron portales con un volumen suficiente de datasets, estabilidad operativa y diversidad temática, garantizando la representatividad del análisis y la viabilidad del pipeline implementado.

5.3. Proceso de selección y validación de datasets

A partir de los portales seleccionados, se procedió a la extracción automatizada de los metadatos disponibles y a la aplicación de los criterios de inclusión definidos en la metodología. Un dataset se consideró válido para el análisis de madurez cuando cumplía, como mínimo, las siguientes condiciones:

- Disponibilidad de metadatos accesibles mediante API o descarga directa.
- Existencia de información básica identificable (título, publicador y mecanismo de acceso).
- Presencia de al menos un punto de acceso o referencia al recurso (URL, identificador).
- Licencia identificable o campo equivalente, incluso cuando su valor fuera incompleto o ambiguo.

Cuando alguno de los campos relevantes se encontraba ausente o vacío, el dataset no se descarta automáticamente. En estos casos, se procede a la fase de validación manual descrita en el Capítulo 4, con el objetivo de confirmar si la ausencia del metadato respondía a una limitación real del portal o a un problema puntual de extracción.

Los valores ausentes se mantuvieron explícitamente como vacíos en el conjunto de datos final, no como errores, sino como evidencia empírica de debilidades estructurales del portal en términos de calidad de metadatos, trazabilidad o accesibilidad. Este enfoque permite que la ausencia de información forme parte del análisis de madurez, en lugar de ser ocultada mediante imputaciones artificiales.

La automatización no descarta los dataset cuando hay un campo vacío o no definido, descarta cuando ninguno de los criterios definidos no se cumple. Cuando se

cumplen la mayoría de los criterios se mantiene. Esto es válido, lo que hace tener bases para poder analizar la madurez de los metadatos de cada Portal.

En el caso específico del portal de Aragón, se observó que la interfaz de programación (API) presentaba una actualización reciente centrada en conjuntos de datos publicados durante el año 2025. Tras una ejecución inicial del script, se constató que la aplicación del criterio de antigüedad mínima de 12 meses reducía drásticamente el volumen de la muestra, limitando la capacidad de evaluación estadística. Por este motivo, se procedió a modificar el mínimo de edad del dataset, permitiendo así la obtención de una muestra más representativa que garantizara el análisis de madurez del resto de las dimensiones de metadatos.

El proceso de gestión de la información prioriza la recuperación de datos frente a la eliminación sistemática de registros incompletos. Cuando se identifica la carencia de metadatos esenciales en la respuesta original de la API, tales como la categoría temática o la frecuencia de actualización, el sistema ejecuta un protocolo de búsqueda semántica basado en el análisis del título y la descripción.

Esta metodología permite salvar registros que, aun siendo técnicamente válidos en formato y accesibilidad, presentan vacíos en su catalogación oficial. En los casos donde la búsqueda semántica arroja un resultado positivo, se procede a asignar la categoría identificada para enriquecer el perfil del dataset y permitir su inclusión en los rankings temáticos. No obstante, si tras el análisis exhaustivo de los textos la información sigue siendo inexistente, el campo se etiqueta bajo el término "No definido". Este etiquetado asegura que el conjunto de datos permanezca dentro del cálculo de la estadística general de apertura, aunque su puntuación en las métricas de calidad documental se vea penalizada por la falta de precisión. Finalmente, se mantiene un estricto respeto por la integridad de la fuente original al evitar cualquier tipo de invención de datos; la persistencia de un campo vacío se interpreta como un hallazgo relevante sobre la transparencia del portal, evidenciando áreas de mejora en la gestión de metadatos de la administración pública.

5.4. Construcción del conjunto de datos finales

Como resultado del proceso de extracción, selección y validación, se construyó un conjunto de datos consolidado compuesto por **3.819 datasets**, distribuidos entre los distintos portales analizados. Para cada dataset se almacenan los metadatos necesarios para el cálculo de las métricas de madurez, manteniendo una estructura común que permitió su análisis comparativo.

La normalización de categorías temáticas, formatos, licencias y campos temporales fue clave para garantizar la coherencia del conjunto de datos finales. Este proceso permitió reducir la heterogeneidad inicial entre portales y sentar las bases para el análisis empírico presentado en el capítulo de resultados.

5.5. Métricas de madurez

Tras finalizar la extracción de metadatos sólidos y estables en la fase 1, en esta etapa se define el diseño conceptual de las métricas de madurez que se utilizarán para evaluar el ecosistema de datos abiertos en España. El propósito de este diseño es establecer un marco estructurado que permita transformar los metadatos de los datasets en indicadores comparables, alineados con las dimensiones clave de la madurez del ecosistema.

El diseño de las métricas parte de los metadatos extraídos, los cuales constituyen la base informativa del análisis. A partir de esta información, se definen cinco dimensiones de madurez: trazabilidad, accesibilidad, interoperabilidad semántica, interoperabilidad técnica y calidad. Estas dimensiones se seleccionan por su presencia recurrente en la literatura especializada y por su relevancia práctica para evaluar la madurez efectiva de los datos abiertos. La construcción de estas métricas responde a la necesidad de evaluar de forma objetiva la calidad y el grado de apertura de los datos publicados.

Cada dimensión se operacionaliza mediante un conjunto de indicadores observables en los metadatos, permitiendo una evaluación empírica y reproducible. El diseño evita el uso de métricas subjetivas o declarativas, priorizando variables verificables como fechas, formatos, licencias, identificadores y mecanismos de acceso.

5.6. Portales problemáticos y limitaciones encontradas

Durante la implementación del pipeline se identificaron portales que presentaron dificultades, pero fueron incluidos, porque otros criterios importantes tenían validez, esto se comprobó con la validación manual.

El caso de la Junta de Andalucía destaca como un ejemplo de portal de alta calidad que, sin embargo, presenta barreras técnicas significativas para la recolección automatizada de datos. A diferencia de plataformas basadas en CKAN estándar, su estructura ha demostrado ser históricamente más compleja de procesar mediante técnicas de *scraping*. Esta dificultad radica principalmente en que el portal no suele ofrecer archivos en formatos de descarga directa como CSV o JSON de manera uniforme, priorizando en su lugar la interoperabilidad a través de servicios geospaciales como WMS o WFS. Esta orientación obliga al reutilizador a emplear herramientas específicas para el tratamiento de datos geográficos, en lugar de peticiones web sencillas. Adicionalmente, durante el proceso de análisis se ha constatado que algunas URL de descarga extraídas desde el catálogo del portal corresponden a enlaces no operativos, como ocurre en determinados recursos publicados en formato ODS, lo que impide el acceso efectivo al contenido del dataset. Esta situación introduce una limitación adicional en la accesibilidad y fiabilidad de los metadatos, afectando negativamente a la automatización de procesos de recolección de datos, y en la práctica de la información publicada.

En el análisis del Portal de Datos Abiertos de Aragón, se han identificado determinadas problemáticas y limitaciones asociadas a los formatos de descarga de algunos conjuntos de datos, especialmente aquellos publicados en formato PX (PC-Axis). Aunque este formato es habitual en entornos estadísticos y permite una estructuración avanzada de la información, su uso presenta barreras relevantes desde la perspectiva de la interoperabilidad y accesibilidad de los datos. En particular, archivos como los disponibles en el catálogo del portal por ejemplo, el recurso en formato PX proporcionado por el Instituto Aragonés de Estadística requieren herramientas especializadas para su lectura y procesamiento, lo que dificulta su acceso por parte de usuarios no expertos y limita su integración directa en flujos automatizados o aplicaciones basadas en estándares abiertos ampliamente utilizados, como CSV o JSON. Esta situación evidencia una restricción práctica en la accesibilidad técnica, reduciendo parcialmente el grado de madurez del portal a pesar de la disponibilidad pública de la información.

En el caso del Portal Nacional, en las primeras versiones del script se han identificado algunos registros donde la `dataset_uri` (la página de aterrizaje o *landing page*) presenta un enlace roto, habitualmente debido a migraciones en los portales de origen que el federador nacional aún no ha actualizado. Sin embargo, la descarga directa del recurso sigue siendo válida, lo que indica que, aunque la capa de información humana está dañada, la capa de datos para máquinas permanece funcional. En versiones siguientes donde los endpoints eran dirigidos a otra API distinta, estos datos no se presentaron, quedando de forma completa los campos extraídos.

Por otro lado, el portal de la Región de Murcia presenta un escenario más crítico en cuanto a volumen, afectando a 119 datasets. Un ejemplo representativo es el recurso de la agenda de alcaldía de Lorca, donde la URI oficial devuelve un error de servidor (posiblemente por una caída del servicio de catálogo local o un error de configuración del servidor web), mientras que el enlace de descarga directa del archivo se mantiene operativo.

La inclusión de estas limitaciones en el análisis no se considera una debilidad del estudio, sino una característica inherente al ecosistema de datos abiertos, que refuerza la necesidad de evaluar la madurez de los portales a partir de evidencia empírica real y no de supuestos teóricos.

5.7. Tabla resumen de portales y datasets analizados

Con el fin de sintetizar las características principales de las fuentes de datos utilizadas, se construyó una tabla resumen que recoge, de forma estructurada, la información clave de cada portal analizado. Esta tabla actúa como referencia central para el análisis empírico desarrollado en los capítulos posteriores.

La tabla incluye las siguientes variables:

- Nombre del portal de datos abiertos.
- URL oficial del portal.
- Tecnología del portal (por ejemplo, CKAN, Opendatasoft, DCAT/SPARQL, APIs propias).
- Número total de datasets analizados por portal.
- Distribución de datasets por categoría temática.

Portal / Administración	URL de Acceso	Tecnología	Total Datasets (N)	Principales Categorías (Top 3)
Datos Portal Nacional	https://datos.gob.es/es/	API DATA+ SPARQL/DCAT-AP	450	Datos Geográficos y Medioambientales, Educación, Transporte
Opendata Barcelona	https://opendata-ajuntament.barcelona.cat/	CKAN	552	Ciencia y Tecnología, Transporte, Datos Geográficos y Medioambientales
Ayuntamiento Junta Andalucía	https://www.juntadeandalucia.es/datosabiertos	CKAN	600	Educación, Datos Demográficos y Estadísticos, Datos Geográficos y Medioambientales
Datos Comunidad Madrid	https://datos.comunidad.madrid/	CKAN	600	Educación, Salud, Datos Demográficos y Estadísticos
Datos abiertos Región Murcia	https://datosabiertos.regiondemurcia.es/	CKAN	598	Datos Geográficos y Medioambientales, Sector público, Hacienda
Análisis de Datos abiertos Castilla y León	https://analisis.datosabiertos.jcyl.es/	Opendatasoft	419	Salud, Sector público, Cultura
Opendata Aragon	https://opendata.aragon.es/datos	CKAN	600	Salud, Datos Demográficos y Estadísticos, Educación
TOTAL				3819

Tabla 5.1. Portales de datos abiertos, tecnologías utilizadas

La tabla resumen de portales y datasets presentada en este capítulo se utiliza como referencia directa en el Capítulo 7 (Resultados), donde se analizan las métricas de madurez y se comparan los portales evaluados. Esta conexión permite contextualizar los resultados y facilitar su interpretación a partir de las características estructurales de cada portal.

Capítulo 6 - Preparación de datos e implementación

Este capítulo describe el proceso de preparación de los datos y la implementación técnica del pipeline de análisis, a partir de los metadatos extraídos de los portales de datos abiertos seleccionados. Mientras que el Capítulo 4 definió la metodología y el diseño general del análisis, y el Capítulo 5 documentó la selección de fuentes y datasets, en este apartado se detalla cómo se llevaron a cabo las tareas de limpieza, normalización, validación e integración de los datos, necesarias para garantizar la fiabilidad del análisis empírico.

La preparación de los datos constituye una fase crítica del estudio, dado el alto grado de heterogeneidad técnica y semántica existente entre los portales analizados. Las decisiones adoptadas en esta etapa influyen directamente en la calidad de las métricas de madurez calculadas y en la comparabilidad de los resultados obtenidos.

6.1. Extracción y consolidación de metadatos

Una vez definidos los portales y los datasets a analizar, se procedió a la extracción automatizada de los metadatos mediante la ejecución de scripts específicos que conforman el pipeline, todos estos adaptados a la tecnología de cada portal, para interactuar con las APIs, ya fueran plataformas basadas en Api CKAN, OPENDATASOFT, API Data / SPARQL, u otros entornos tecnológicos empleados por administraciones públicas españolas. El objetivo del procedimiento inicial es obtener un conjunto homogéneo de metadatos que permite comparar portales y categorías temáticas bajo criterios uniformes. Los scripts implementados recorren de forma automática cada portal, obteniendo los resultados descargables en formatos (csv/xlsx) en una Lista o tabla que contiene los metadatos de cada portal, los resultados incluyen los criterios de selección definidos en capítulo 4, por cada dataset se filtran: la accesibilidad abierta, antigüedad igual o superior a doce meses, formatos abiertos, licencias presentes, la disponibilidad de un identificador suficientemente estable, url descargables, diccionario de datos y control de vocabulario presentes.

El resultado de este proceso (Tablas por portal) muestran las columnas de los metadatos requeridos para la siguiente fase, que constituyen el insumo fundamental para el resto del pipeline. Estas tablas contienen, para cada portal, una representación estructurada de su catálogo bajo un modelo uniforme, lo que incluye información como el título y la descripción del dataset, fechas de publicación y modificación, organismo publicador, formato principal, licencia declarada, URL de descarga, entre otros más. La importancia de estas tablas radica en que permiten estandarizar portales que, en condiciones normales, presentan diferencias significativas tanto en la estructura de sus APIs como en sus convenciones semánticas. Estas tablas funcionan como la base cronológica del estudio: permiten identificar el conjunto final de datasets por portal y sirven de punto de partida para la construcción de Métricas.

La preparación de datos realizada en la Fase 1 transforma los metadatos en bruto de cada API en variables de control. A continuación, se detallan los metadatos clave preparados y su propósito en la construcción de métricas:

- Validación de URI y Trazabilidad (**issued**, **modified**, **dataset_uri**): Se construye y verifica la persistencia del enlace. Un metadato preparado como "no trazable" indica una rotura en la Cadena de Suministro de Datos, una métrica crítica para la fiabilidad del portal.
- Normalización de Formatos: **format**, **n_format**, **has_open_format**: Se realiza una limpieza y estandarización (ej. de **.csv**, **text/csv**, **CSV**).
- Identificación de Serializaciones Semánticas (**has_semantic_serialization**): Se filtran específicamente formatos vinculados a la Web Semántica (RDF, Turtle, JSON-LD) para medir el cumplimiento del Nivel 4 y 5 de las 5 estrellas de Open Data.
- Verificación de Licencias (**license_open**, **license_present**): El script traduce nombres de licencias a valores booleanos (Abierta: Sí/No).
- Estandarización del Publicador (**publisher**): Se limpian los nombres de las organizaciones para evitar duplicidad (ej. "Dpto. de Salud" vs "Departamento de Salud"), permitiendo métricas de Desempeño Institucional por departamento.
- Categorización Temática (**category**): Mediante el motor de búsqueda por palabras clave, se asigna una categoría única. Esto permite el Análisis de Cobertura, para saber qué sectores de la administración están cumpliendo mejor con las directrices de apertura.

- **Cálculo de Vigencia** (`meets_age_criterion`, `age_month_issued`, `age_month_modified`): Se transforman las cadenas de texto de fecha en objetos de tiempo para calcular la Antigüedad del Dato. Esto permite medir si el portal mantiene un flujo de actualización constante o si el catálogo tiende a la obsolescencia.
- **Presencia de Diccionario de Datos** (`has_data_dictionary`): Se prepara este flag buscando documentación técnica asociada. Es la base para la métrica de Facilidad de Uso (Usability).

6.2. Limpieza y normalización de los datos

Tras la consolidación inicial, los metadatos fueron sometidos a un proceso sistemático de limpieza y normalización, orientado a reducir inconsistencias y facilitar el análisis comparativo entre portales. Esta etapa resulta esencial debido a la heterogeneidad en los esquemas de metadatos, formatos y convenciones utilizadas por los distintos portales de datos abiertos.

6.2.1. Limpieza y normalización de metadatos

Las principales tareas desarrolladas en esta fase incluyen:

- **Normalización de campos temporales**, unificando las distintas representaciones de las fechas de publicación (`issued`) y actualización (`modified`) en un formato estándar, lo que permite realizar análisis temporales consistentes.
- **Homogeneización de valores categóricos**, como licencias, formatos y frecuencias de actualización, mediante reglas de correspondencia definidas a partir de los valores observados en cada portal.
- **Limpieza de campos textuales**, eliminando caracteres especiales, etiquetas HTML, saltos de línea y valores redundantes que dificultan la interpretación y el procesamiento automático de los metadatos.
- **Transformación de estructuras complejas**, como listas o campos anidados, en representaciones tabulares planas, adecuadas para su análisis cuantitativo.

6.2.2. Normalización de formatos y accesibilidad técnica

Con el objetivo de analizar la accesibilidad técnica de los datasets, se construyeron variables relacionadas con los formatos disponibles y los mecanismos de acceso:

- **Clasificación del formato principal (open_format)**, identificando el tipo de formato predominante en cada dataset (por ejemplo, CSV, JSON, XML, PDF).
- **Disponibilidad de formatos recomendados (has_allowed_format)**, definida como una variable binaria que indica si el dataset incluye al menos un formato considerado adecuado para datos abiertos, tales como CSV, JSON, XML o formatos geoespaciales abiertos (por ejemplo, GeoJSON o Shapefile).

Estas variables permiten analizar la adecuación técnica de los datasets para su consumo automatizado y su integración en sistemas externos.

6.2.3. Normalización de licencias y apertura legal

La dimensión legal de los datos abiertos se abordó mediante la construcción de variables relacionadas con las licencias declaradas en los metadatos:

- **Detección de licencia abierta (license_open)**, definida como una variable binaria que identifica la presencia de términos asociados a licencias abiertas (por ejemplo, CC-BY, CC0 u Open Data License).
- **Clasificación del tipo de licencia (license_present)**, que agrupa los resultados en categorías generales como "Licencia abierta" o "Licencia restringida / no definida".

La ausencia o ambigüedad en la información sobre licencias se mantuvo como un elemento de análisis, al constituir una limitación relevante desde el punto de vista de la madurez estructural del portal

6.2.4. Relación con las métricas de madurez

Las variables derivadas obtenidas en las fases anteriores constituyen la base empírica para el cálculo de las métricas de madurez definidas en el estudio. En particular, estas variables alimentan las dimensiones de interoperabilidad técnica,

interoperabilidad semántica, accesibilidad, trazabilidad y calidad de los metadatos, que se integran posteriormente en el índice global de madurez.

La construcción explícita y trazable de estas variables permite garantizar la coherencia entre los metadatos originales y los indicadores utilizados en el análisis de resultados, reforzando la reproducibilidad y la transparencia del estudio.

Estas métricas de madurez, proporcionarán un diagnóstico comparativo de alto nivel sobre el estado en los diferentes portales analizados. En primer lugar, se visualizará una evaluación del cumplimiento de los estándares de interoperabilidad técnica y legal, permitiendo determinar qué tan preparados están los datos para ser usados. Este resultado es vital, ya que permite identificar las brechas entre la voluntad de transparencia y la capacidad técnica real de la administración.

En este contexto, la fase 3 se implementa con objetivo de medir aquellos atributos que sí son objetivamente observables y comparables en los portales: la antigüedad del dataset, la frecuencia de actualización, la apertura legal de la licencia, la disponibilidad de formatos abiertos, la trazabilidad técnica y la completitud de los metadatos. Para ello se construyeron indicadores específicos como la antigüedad en meses, la frecuencia de actualización, el índice de accesibilidad, entre otros.

6.3. Dificultades técnicas y decisiones de implementación

Durante la preparación de los datos y la implementación del pipeline se encontraron diversas dificultades técnicas, derivadas principalmente de la heterogeneidad de los portales y de la calidad desigual de los metadatos disponibles. Entre los problemas más frecuentes se incluyen APIs inestables, campos ausentes o mal documentados, formatos no normalizados y discrepancias semánticas entre portales.

Estas dificultades se abordaron mediante ajustes metodológicos orientados a garantizar la viabilidad del proceso, y adoptar estrategias de extracción diferenciadas en función de la infraestructura de cada plataforma, el uso de reglas de normalización conservadoras y la incorporación de validación manual. Lejos de ser un obstáculo, estas situaciones aportaron información valiosa sobre las

limitaciones estructurales del ecosistema de datos abiertos y reforzaron la pertinencia del enfoque de madurez adoptado en el estudio.

En cuanto a las dificultades técnicas identificadas, destaca el caso del portal nacional datos.gob.es, en el que fue necesario combinar dos mecanismos de acceso complementarios: la API REST APIDATA y el endpoint SPARQL. La consulta exclusiva a la API APIDATA permitía recuperar información básica de los datasets y sus distribuciones, pero presentaba limitaciones en campos clave como categorías normalizadas, información semántica y ciertos atributos de publicación. Por el contrario, el uso aislado del endpoint SPARQL proporcionaba metadatos enriquecidos conforme al modelo DCAT, pero no garantiza la recuperación completa de URLs de descarga ni de formatos efectivos. Por este motivo, se implementó una estrategia híbrida que integra ambas fuentes, permitiendo consolidar un esquema de metadatos más completo y consistente para el análisis posterior. En cuanto a [Opendata barcelona](https://opendata.barcelona.cat) tiene url de acceso inestable, el Portal en ocasiones permanece inactivo presentado un error de acceso, por lo que en las pruebas realizadas manualmente a cada portal se pudo detectar esta irregularidad.

De forma similar, el portal de Datos Abiertos de Castilla y León presenta particularidades técnicas derivadas del uso de la plataforma OpenDataSoft, lo que requirió la adaptación del pipeline a su API específica (Explore API v2.1). A diferencia de los portales basados en CKAN o en catálogos DCAT centralizados, este entorno expone los metadatos bajo una estructura distinta, con nomenclaturas y jerarquías propias, y con una disponibilidad desigual de campos como fechas de publicación, modificación o información semántica. Estas diferencias obligaron a redefinir los procesos de normalización y validación, así como a implementar verificaciones adicionales para garantizar la coherencia de los resultados. En conjunto, estas adaptaciones evidencian que, pese a la existencia de estándares abiertos, la interoperabilidad real entre portales sigue siendo limitada, lo que refuerza la necesidad de pipelines flexibles y específicos para cada tecnología subyacente.

También pudimos detectar deficiencias en la actualización y fechas. que hay portales como el de la Comunidad de Madrid que muestra a nivel de consulta la fecha de creación y de actualización invertida (no puede existir el caso en que la fecha de actualización del dataset sea menor que la fecha de publicación, lo que arrojó un punto débil de su portal. El Portal Nacional presenta dataset que cuando se actualiza el portal pierde su fecha de publicación, generando confusión para los usuarios es decir, un dataset que fue consultado el mes anterior tenía una fecha de

publicación distinta a la última actualización, y en ocasiones se mostró misma fecha de publicación con la de actualización.

En cuanto a las Licencias, en el Portal Junta andalucía aparecían Datos en Licencias que debían consultarse en organizaciones particulares (ejemplo Consultar a la Consejería de Turismo y Deporte), lo que llevó a mejorar y adaptar el script para poder identificar, cuando es una licencia abierta, o no abierta. En la auditoría del portal de datos abiertos de la Región de Murcia, se identifica que la apertura de licencias no se rige por una licencia estándar de Creative Commons, sino por un Aviso Legal propio que actúa como marco regulatorio. Este modelo, amparado en la Ley 37/2007, establece un régimen de 'licencia abierta' que permite la explotación de los datos tanto para fines comerciales como no comerciales de forma gratuita. La normativa autoriza actividades de reproducción, transformación y distribución global, siempre que el usuario cumpla con la atribución obligatoria de la fuente y preserve la integridad de los metadatos. Por tanto, aunque el portal emplee una URL de aviso legal personalizada, las condiciones de uso son funcionalmente equivalentes a una licencia de Reconocimiento (CC BY), garantizando la plena seguridad jurídica para el desarrollo de productos derivados".

6.4. Implementación del pipeline y gestión de scripts

La implementación del pipeline se realizó en Python, siguiendo un diseño modular, en el que cada fase del proceso (extracción, preparación, validación, cálculo de métricas y generación de resultados) se implementó mediante scripts independientes pero interconectados.

Cada portal cuenta con un script de extracción específico, adaptado a su tecnología de acceso, mientras que las fases de normalización, validación y cálculo de métricas se apoyan en funciones comunes reutilizables. Este enfoque permitió mantener una estructura homogénea de salida, independientemente de la fuente de datos, y facilitó la escalabilidad del pipeline a nuevos portales.

El uso de entornos de ejecución reproducibles y el versionado del código garantizaron la trazabilidad de los cambios realizados durante el desarrollo y la posibilidad de replicar el análisis en otros contextos.

Para garantizar la transparencia y la reproducibilidad de la investigación, el código fuente íntegro, la configuración de los entornos de ejecución y las instrucciones técnicas detalladas se encuentran disponibles en el repositorio público de GitHub. En dicho espacio, específicamente en el archivo README.md, se documentan los requisitos técnicos necesarios para la réplica del estudio, incluyendo la versión de Python empleada y el listado de librerías mediante un archivo de dependencias. Asimismo, el repositorio detalla la estructura del *pipeline*, los protocolos de ejecución y la vinculación directa con el repositorio de Zenodo para la obtención de los conjuntos de datos de entrada y salida, asegurando así que cualquier investigador pueda ejecutar el flujo de trabajo y validar los resultados obtenidos de manera autónoma.

6.5. Síntesis del capítulo

El proceso de preparación de datos e implementación descrito en este capítulo permitió construir un conjunto de datos consolidado, coherente y trazable, adecuado para el análisis empírico de la madurez de los portales de datos abiertos. Las decisiones adoptadas en esta fase aseguran la consistencia de las métricas calculadas y sientan las bases para la presentación y análisis de los resultados en el capítulo siguiente.

Lo que evidencia que la implementación del pipeline exigió una adaptación progresiva a las condiciones reales de los portales españoles, lo cual condujo a una redefinición metodológica que resulta coherente tanto con la práctica como con la teoría. El procesamiento de datos realizado en las fases 1 y 3, junto con la validación manual en la fase 2, constituye la base sobre la cual se construyen los resultados del capítulo siguiente. De esta manera, el Capítulo 6 funciona como un puente entre la formulación metodológica y la obtención efectiva de resultados, garantizando que los análisis subsiguientes se apoyan en información rigurosamente preparada y en una comprensión fundamentada de las limitaciones y posibilidades del ecosistema de datos abiertos en España.

Capítulo 7 - Resultados empíricos

El presente capítulo expone los resultados obtenidos a partir de la aplicación del pipeline metodológico descrito en los capítulos anteriores, orientado al análisis de la madurez del ecosistema de datos abiertos en España. A partir de las métricas definidas en los capítulos anteriores, los resultados muestran el análisis comparativo de la madurez del ecosistema de datos abiertos a partir de la evaluación de varios portales institucionales. Para ello, se construyó un Índice Global de Madurez (0–100), definido como el promedio simple entre las cinco dimensiones evaluadas con el mismo peso relativo (20% cada una): Trazabilidad, Interoperabilidad Semántica, Interoperabilidad Técnica, Accesibilidad y Calidad de los metadatos. Este índice actúa como ranking principal de desempeño, permitiendo comparar de forma sintética el grado de madurez alcanzado por cada portal.

Como referencia contextual para la interpretación de los resultados, se utiliza la tabla resumen [\(Tabla 5.1.\)](#) de portales y datasets presentada en el Capítulo 5, en la que se detallan los portales analizados, su nivel administrativo (nacional, autonómico y municipal), la tecnología utilizada, el número de datasets incluidos y su distribución por categorías temáticas. Esta información permite situar los resultados dentro de las características estructurales de cada portal y facilita la comparación entre ellos.

El análisis empírico se organiza en torno a las dimensiones de madurez definidas en el estudio: interoperabilidad técnica, interoperabilidad semántica, accesibilidad, trazabilidad y calidad de los metadatos. Para cada dimensión se presentan métricas agregadas a nivel de portal y de categoría temática, acompañadas de representaciones gráficas que permiten identificar patrones, diferencias y comportamientos recurrentes en el ecosistema analizado.

Los resultados que se exponen en este capítulo se limitan a la descripción y caracterización de las métricas obtenidas, sin introducir interpretaciones ni valoraciones críticas, que se abordan de forma específica en el capítulo de discusión.

La evidencia empírica derivada de la ejecución de los scripts se encuentra depositada en un repositorio público **Zenodo**, organizados en formatos interoperables (XLSX y CSV). Asimismo, se han incluido las menciones de autoría correspondientes a los portales de datos originales, dando

cumplimiento a las cláusulas de atribución exigidas en los avisos legales de las licencias españolas.

<https://doi.org/10.5281/zenodo.18246926>

7.1. Validación manual del pipeline y verificación de resultados

Con el fin de garantizar la fiabilidad y consistencia de los resultados obtenidos mediante los procesos automatizados, se incorporó en la fase 2 del capítulo 4 de la metodología, la estrategia de validación manual orientada a verificar la correcta ejecución del pipeline, en base a los resultados y la calidad de los metadatos analizados. Esta validación actúa como un mecanismo de control adicional que complementa los procedimientos automáticos y permite detectar incidencias no identificables mediante extracción programática.

Durante la fase de selección y extracción de datasets, se realizó una validación manual sistemática sobre muestras aleatorias de aproximadamente 100 datasets por cada portal analizado. El objetivo de esta revisión fue comprobar que los criterios de inclusión definidos se aplicaban de forma coherente y que los metadatos extraídos reflejaban de manera fiel la información publicada en los portales originales.

La validación manual se centró, en particular, en la comprobación de los siguientes aspectos:

- Accesibilidad efectiva de las URLs asociadas a los datasets.
- Existencia y tipo de licencia declarada en los metadatos.
- Disponibilidad real de formatos descargables.
- Número de formatos asociados a cada dataset.
- Coherencia entre las fechas de publicación (issued) y actualización (modified).
- Frecuencia de actualización declarada, cuando se encontraba disponible.
- Corrección y consistencia de los títulos de los datasets.
- Asignación de categorías temáticas.
- Viabilidad efectiva de los enlaces de descarga.

El proceso de validación puso de manifiesto diferencias significativas entre los portales en cuanto a la estructura y cobertura de los metadatos expuestos por sus APIs. Mientras que algunos portales priorizan la exposición de información relacionada con formatos y URLs de acceso, otros concentran los metadatos en campos legales, temporales o de atribución institucional. Esta heterogeneidad requirió un proceso iterativo de ajuste y refinamiento del pipeline, adaptando las consultas a las particularidades técnicas de cada portal.

A lo largo de este proceso iterativo, se identificaron y corrigieron inconsistencias específicas, mejorando progresivamente la capacidad de los scripts de extracción para generar metadatos coherentes, enlaces válidos y resultados reproducibles. La validación manual permitió, además, confirmar que las métricas calculadas se apoyan en información efectivamente accesible y verificable, reforzando la robustez del análisis empírico presentado en este capítulo.

7.2. Análisis exploratorio de trazabilidad externa (aporte metodológico complementario)

De manera complementaria, en el cuarto capítulo de la metodología, se plantea como análisis exploratorio de trazabilidad, el diseño de un script orientado a **detectar huellas de reutilización en fuentes externas**. Tras la ejecución del mismo, se identificaron limitaciones significativas en la identificación automática de dicha reutilización, factores que condicionan a futuras investigaciones. Se observó que las coincidencias obtenidas presentaban una correlación mínima con el uso efectivo de los conjuntos de datos originales; en su lugar, se detectaron citas que únicamente guardaban relación con fragmentos textuales o palabras clave vinculadas al título, sin representar una reutilización real del dato.

Las coincidencias detectadas entre los datasets analizados y posibles reutilizaciones externas resultaron muy escasas, incluso tras la verificación manual directa. Al hacer pruebas manuales, ingresando títulos exactos, identificadores, palabras claves, en cada una de las fuentes, se detectó que las reutilizaciones pueden existir, pero los reutilizadores no realizan la citación correspondiente. En los pocos casos en los que se identificaron coincidencias, estas fueron validadas mediante la comprobación directa del DOI o identificador resultante, confirmando

que la correspondencia detectada por el pipeline era correcta y no producto de coincidencias espurias o ruido textual.

La inclusión de esta validación manual fue un elemento clave para confirmar que la ausencia generalizada de evidencias de reutilización empírica no se debía a fallos del proceso automático, sino que reflejaba limitaciones estructurales del ecosistema de datos abiertos analizado, tales como la falta de identificadores persistentes, la escasa estandarización de los metadatos y la dificultad de trazar referencias externas de forma sistemática.

7.3. Resultados Métricas de Madurez de los Portales

En este apartado se presentan los resultados obtenidos a partir del análisis de los metadatos de los datasets seleccionados en los distintos portales de datos abiertos. Para cada portal se muestran un conjunto de visualizaciones que permiten caracterizar su nivel de trazabilidad, accesibilidad, calidad de metadatos, interoperabilidad, actualización y madurez, así como el análisis comparativo entre portales. Las figuras referenciadas corresponden a las visualizaciones generadas en la Fase 4 del pipeline por Portal.

7.3.1. Comparación global entre portales

El análisis comparativo evidencia diferencias claras entre los portales en términos de accesibilidad, trazabilidad e interoperabilidad. Aragón, Comunidad de Madrid y Portal nacional lideran el ranking global de interoperabilidad semántica, mientras que Castilla y León, Junta Andalucía, y Portal nacional presentan un nivel bajo en Trazabilidad. No obstante se destacan una combinación de formatos abiertos muy comparable con formatos cerrados y propietarios, así como un alto índice en Licencias de apertura total. En contraste, portales como Castilla y León, Junta Andalucía, y Comunidad de Madrid presentan un porcentaje muy igualado en Calidad de metadatos, pero en un nivel medio.

Asimismo, la comparación de indicadores clave por portal muestra que la interoperabilidad técnica suele estar más desarrollada que la semántica y una tendencia consistente con la literatura revisada en el estado del arte.

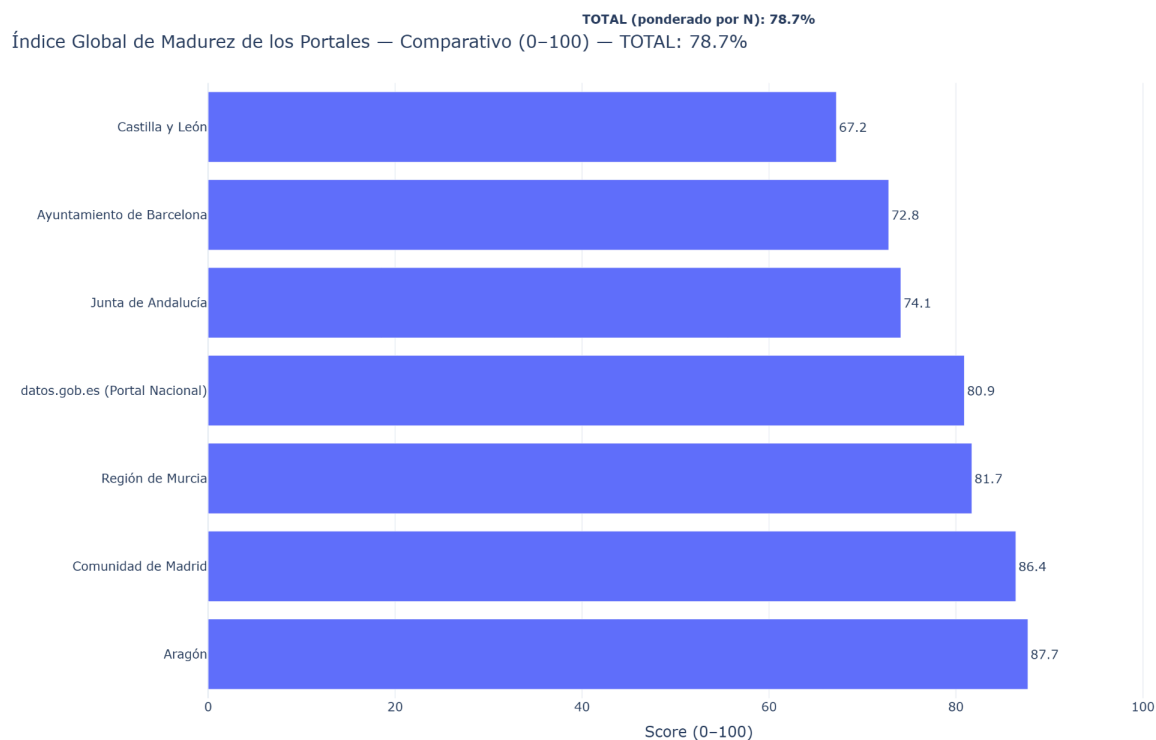


Figura 7.1. Gráfica Índice Global de Madurez de los Portales abiertos.

La figura muestra el Índice de Madurez Global, que es el promedio de 5 dimensiones (Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos).

Componentes del índice Global — Métricas Detallado

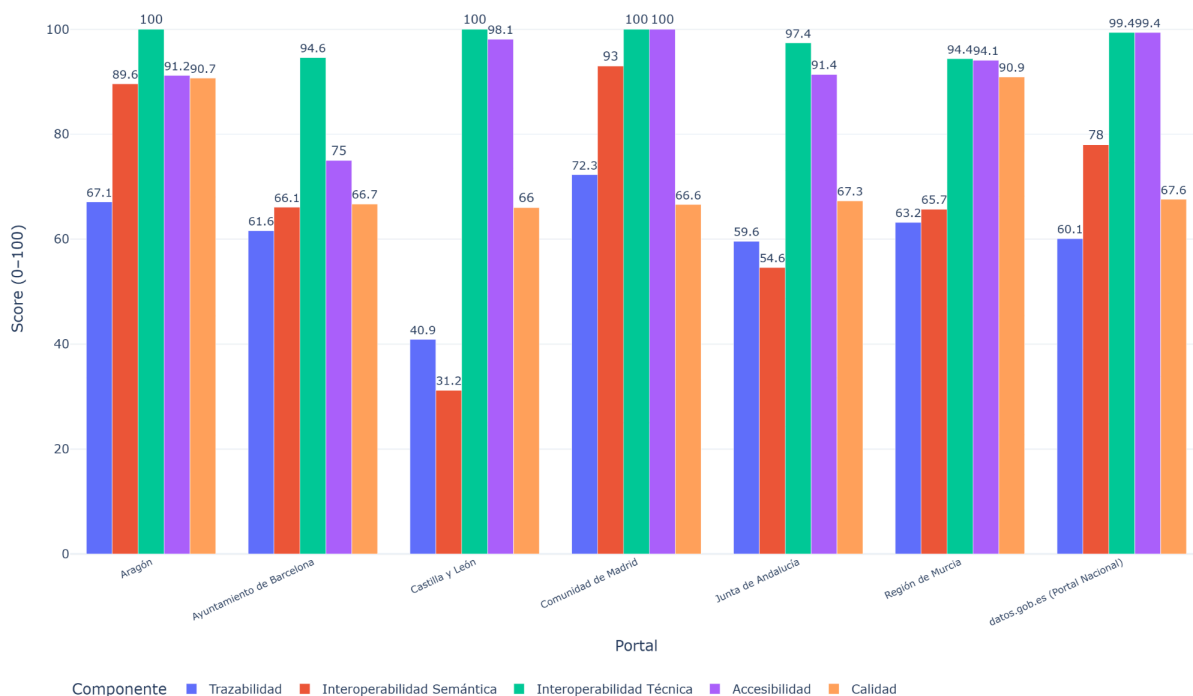


Figura 7.2. Gráfica Promedio por dimensión de Madurez

La figura muestra cada dimensión de componentes que forman el Índice Global, como resultado de cada portal, se calcula como el promedio simple del cumplimiento (%) de sus componentes a nivel dataset (presencia/ausencia).

Formatos (dataset-level) — % por portal (Abiertos vs Cerrados)

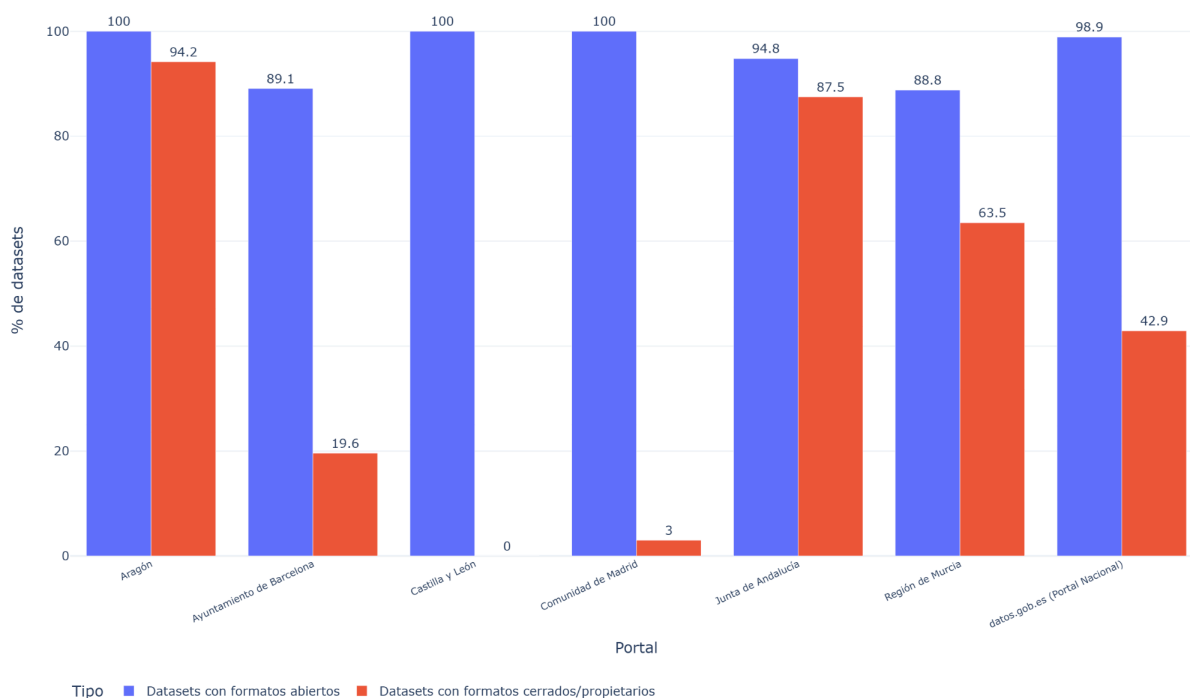


Figura 7.3. Gráfica de Distribución Porcentual de Formatos abiertos y Cerrados/Propietarios.

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado). Proporción calculada sobre el total de formatos reportados en los metadatos (Abiertos: CSV, JSON, GEOJSON, XML, TTL, TURTLE, N-TRIPLES, RDF, JSON-LD y Cerrados: ZIP, PDF, TXT, XLSX, DOCX, HTML).

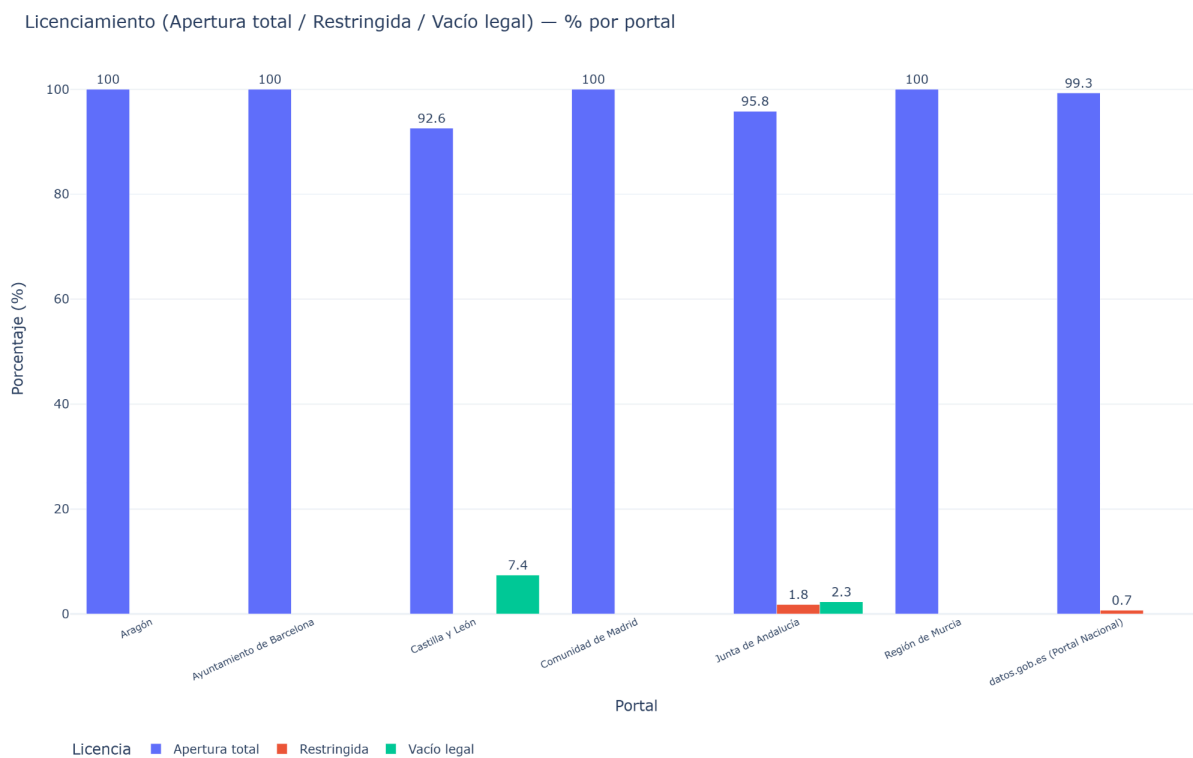


Figura 7.4. Gráfica Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar, en el dataset.

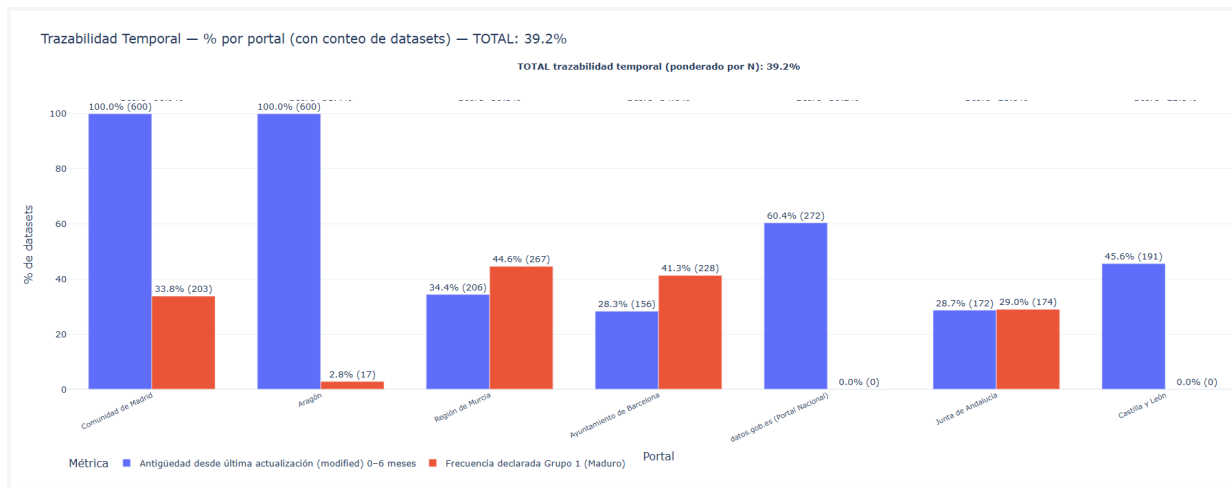


Figura 7.5. Gráfica Frecuencia de actualización.

La figura muestra el promedio en porcentaje promedio de datasets con frecuencia de actualización: Horas, Diaria y Semanal por portal, así como la antigüedad desde su última actualización. Nos da una perspectiva del % de actualización de estos rangos, que tienen los portales.

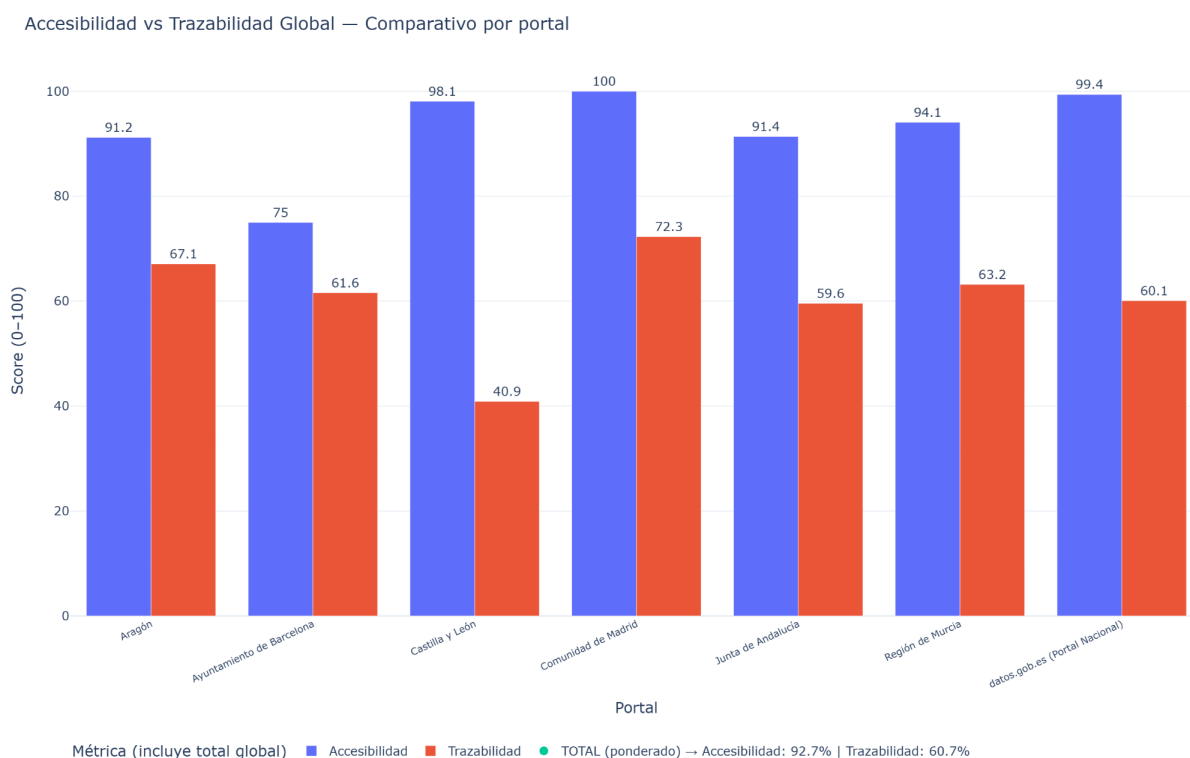


Figura 7.6. Gráfica Accesibilidad vs Trazabilidad Global.

La figura muestra comparativa entre el nivel de accesibilidad (Licencia abierta, URL descargas, api rest presentes en el dataset) con la la Trazabilidad (frecuencia de actualización, origen de los datos (publicador, identificador único, URL del dataset, antigüedad de actualización)

Calidad de Metadatos vs Trazabilidad Global — Comparativo por portal

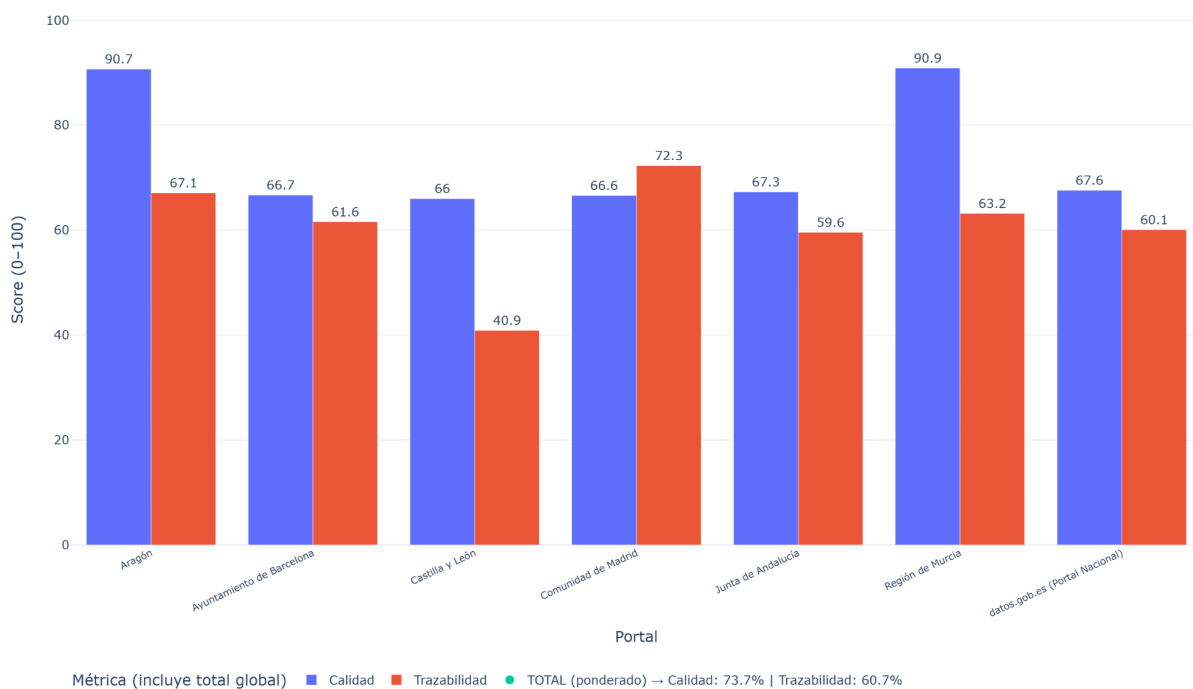


Figura 7.7. Gráfica Calidad de metadatos vs Trazabilidad Global.

La figura muestra comparativa el nivel de Calidad de metadatos que implica (Descripción, títulos y diccionario de datos) con la Trazabilidad (frecuencia de actualización, origen de los datos (publicador, identificador único, URL del dataset, antigüedad de actualización)

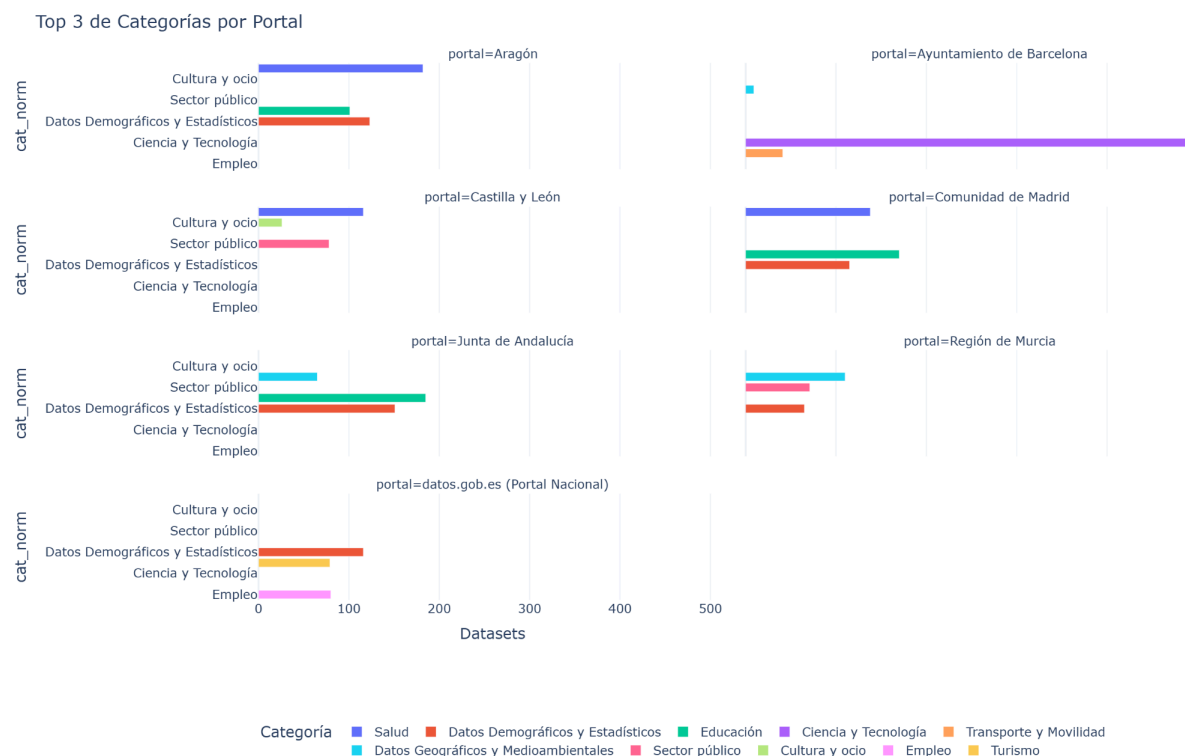


Figura 7.8. Gráfica Top 3 de Categorías entre portales.

La figura muestra Top 3 de categorías con más dataset, extraídas de la muestra de cada Portal.

7.3.2. Portal Barcelona

El análisis del portal Datos abiertos Barcelona se basa en una muestra de 552 conjuntos de datos, extraídos a través de su API institucional CKAN. Esta muestra abarca varias categorías temáticas, entre el top 3 destacan Ciencia y tecnología, Transporte y movilidad, Datos geográficos y medioambientales. Se observa una alta densidad de metadatos actualizados, lo que permite evaluar su madurez técnica bajo un marco de consistencia temporal elevado.

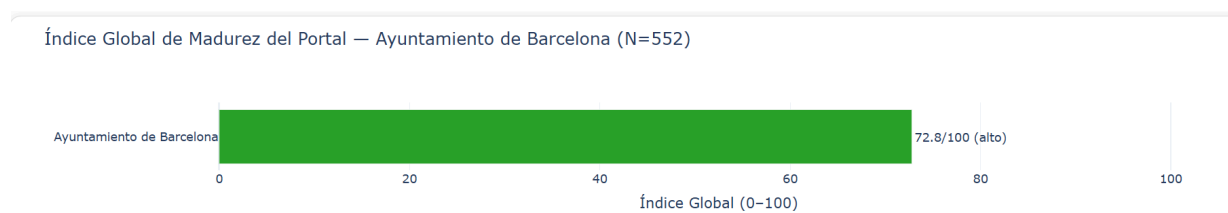


Figura 7.9. Resultado Gráfico Índice de Madurez del Portal.

La figura muestra el Índice global calculado como promedio **ponderado** de 5 componentes: Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos. Nivel definido Bajo 0–30, Medio 30–70, Alto 70–100.

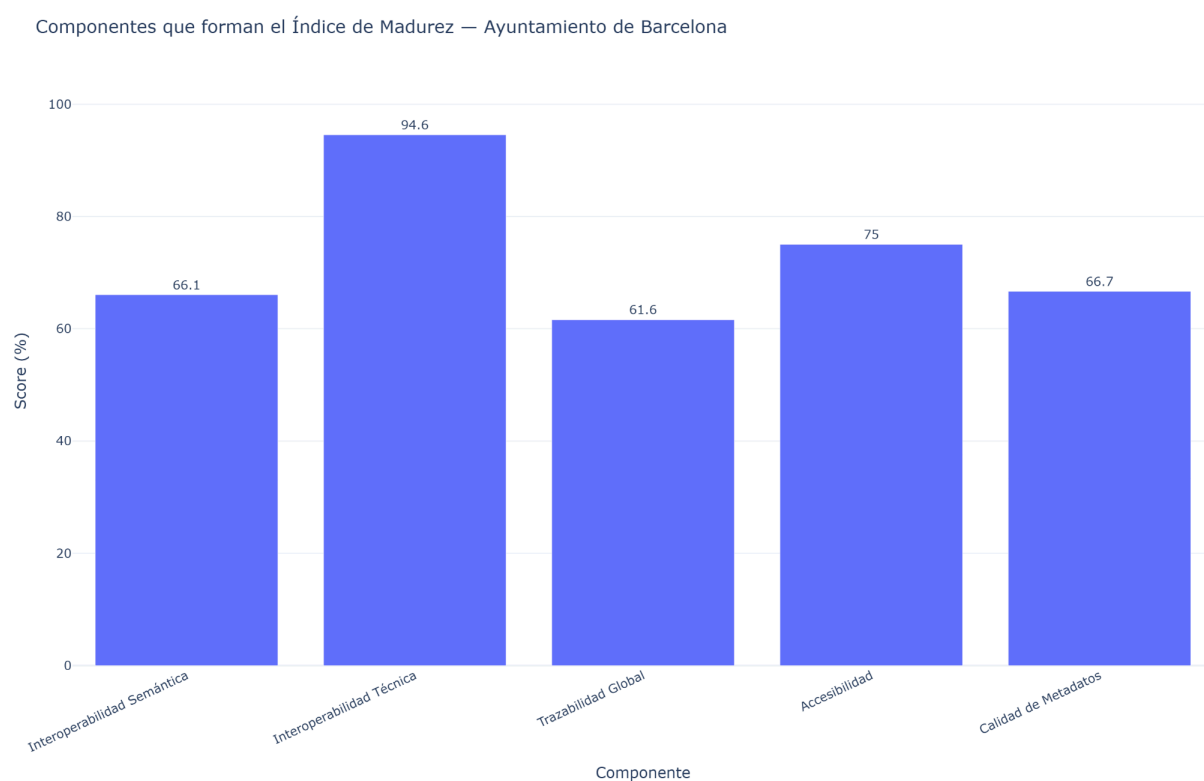


Figura 7.10. Resultado Gráfico de componentes de Índice de Madurez del Portal.

La figura muestra Cada score de dimensión se calcula como promedio del cumplimiento de los componentes que forman cada métrica (peso igual por componente). Los 5 componentes construyen el Índice de Madurez del Portal.

Interoperabilidad Técnica — Clasificación de formatos (datasets con presencia)

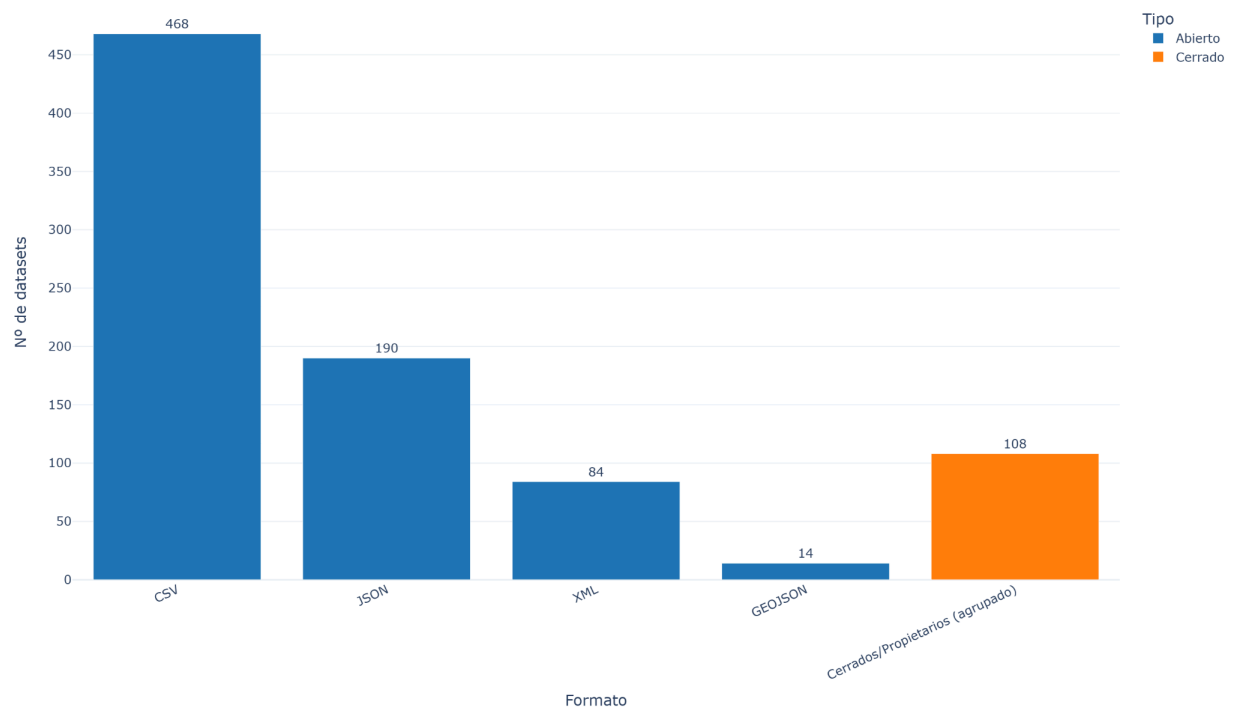


Figura 7.11. Resultado Clasificación de formatos para Interoperabilidad técnica

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado).

Accesibilidad (Dimensión legal) — Licenciamiento permitido

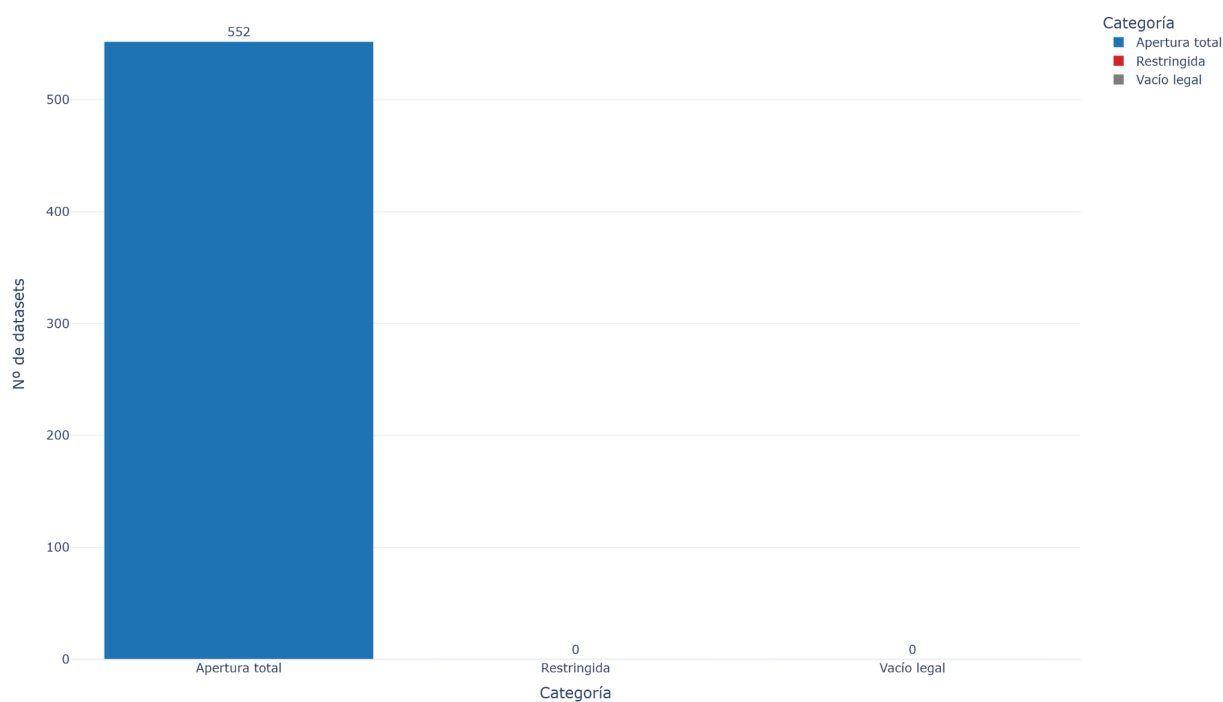


Figura 7.12. Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar.

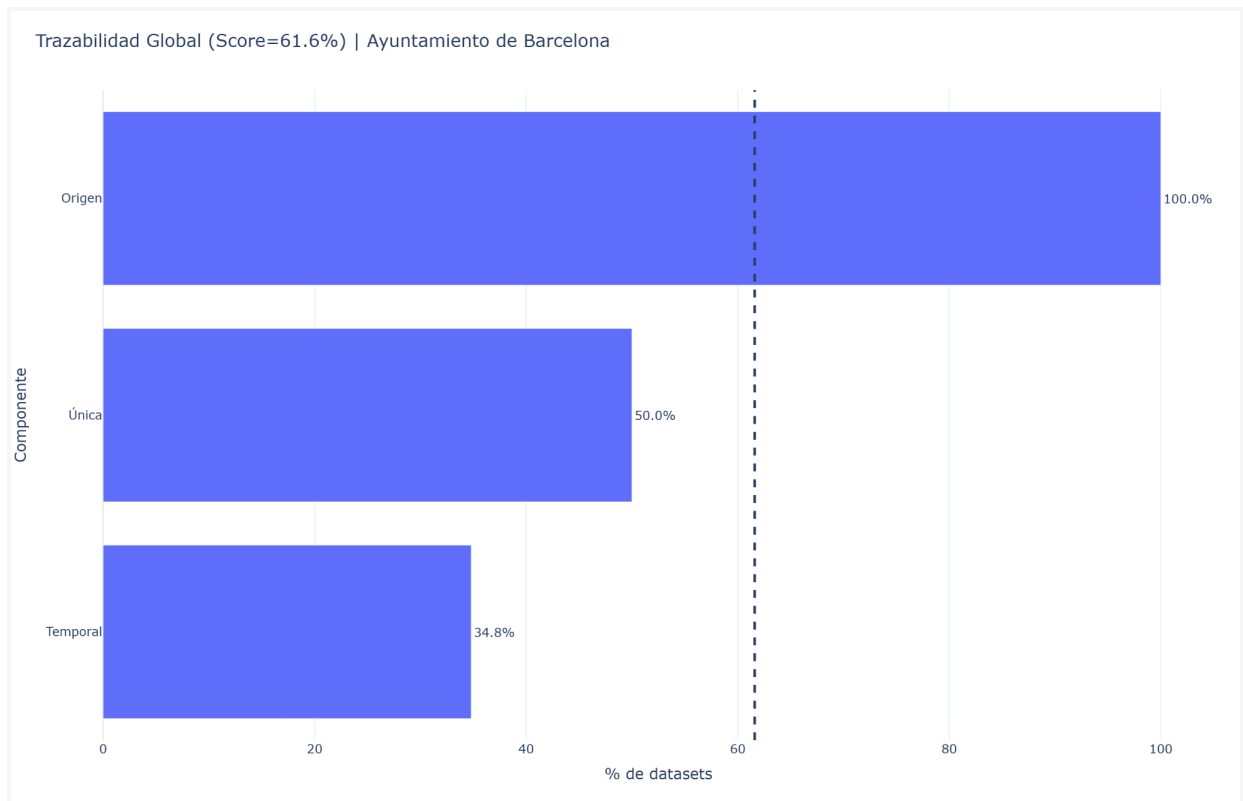


Figura 7.13. Resultado Gráfico Trazabilidad Global

La figura es el resultado global de tres componentes que lo forman porcentaje promedio de cumplimiento.

Trazabilidad Única: Lo conforman el DOI (Código único y permanente) y la url de acceso al dataset.

Trazabilidad Origen: Publicador, identificador (del dataset)

Trazabilidad Temporal: Frecuencia de actualización, Antigüedad del dataset desde su última actualización (valores medidos con actualización, por horas, diaria, semanal, mensual, hasta 6 meses) lo que hace un dataset trazable temporalmente.

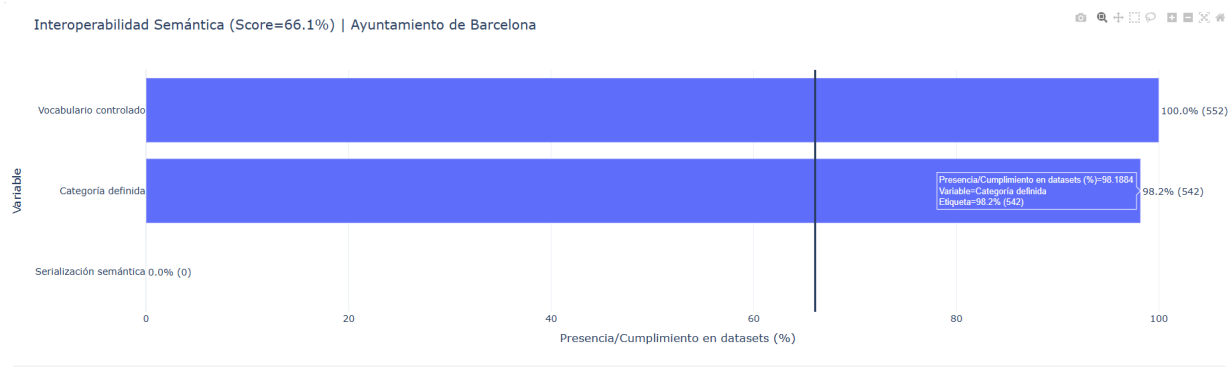


Figura 7.14. Resultado Gráfico Métrica de Interoperabilidad Semántica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 66.1%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.



Figura 7.15. Resultado Gráfico Métrica de Interoperabilidad Técnica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 94,6%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

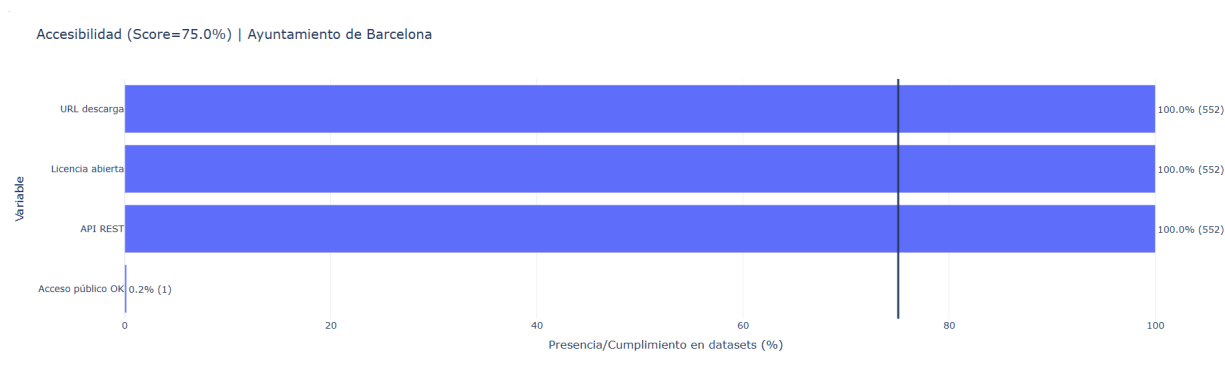


Figura 7.16. Resultado Gráfico Métrica de Accesibilidad

La figura muestra el promedio del total de las variables que lo forman (total de datasets 75%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

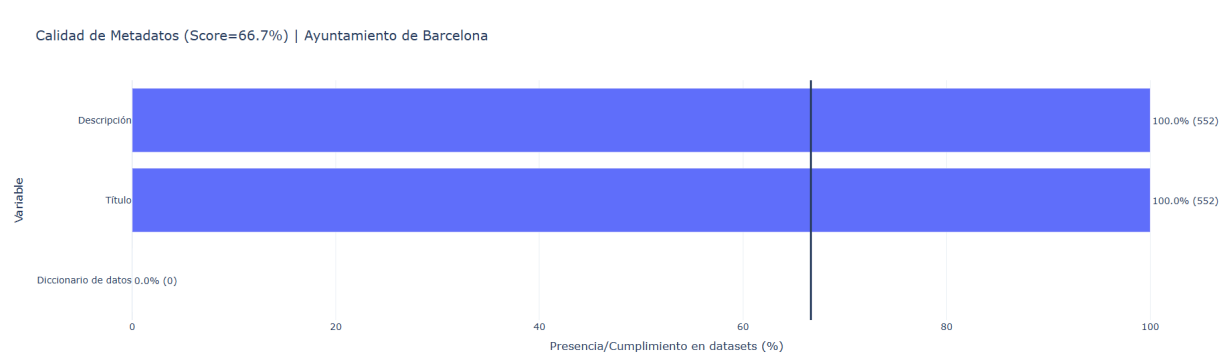


Figura 7.17. Resultado Gráfico Métrica de Calidad de metadatos.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 66.7%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

7.3.3. Portal Aragón

El análisis del portal Datos abiertos Aragón se basa en una muestra de 600 conjuntos de datos, extraídos a través de su API institucional CKAN. Esta muestra abarca varias categorías temáticas, entre el top 3 destacan Salud, Datos demográficos y estadísticos, Educación. Se observa una alta densidad de metadatos

actualizados, lo que permite evaluar su madurez técnica bajo un marco de consistencia temporal elevado.

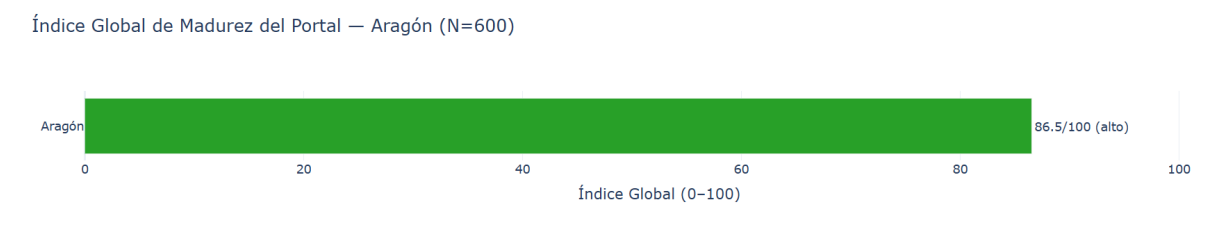


Figura 7.18. Resultado Gráfico Índice de Madurez del Portal.

La figura muestra el Índice global calculado como promedio **ponderado** de 5 componentes: Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos. Nivel definido Bajo 0–30, Medio 30–70, Alto 70–100.

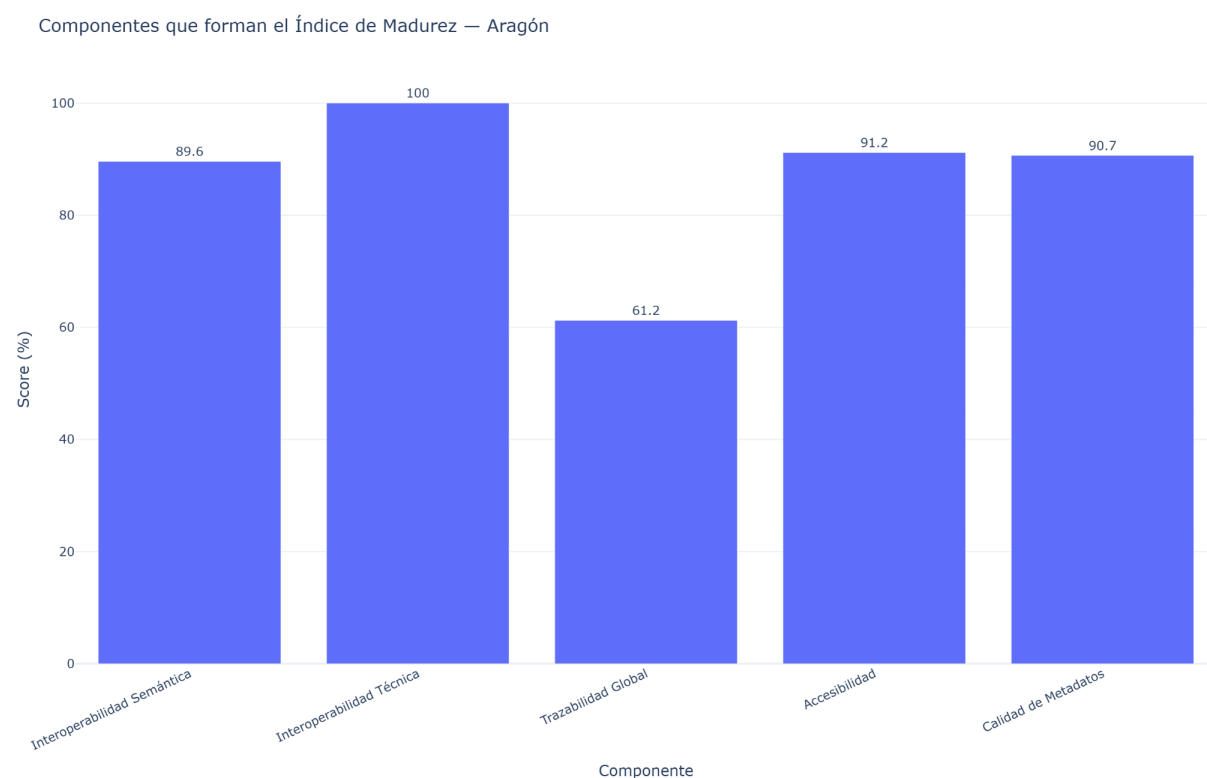


Figura 7.19. Resultado Gráfico de componentes de Índice de Madurez del Portal.

La figura muestra Cada score de dimensión se calcula como promedio del cumplimiento de los componentes que forman cada métrica (peso igual por componente). Los 5 componentes construyen el Índice de Madurez del Portal.

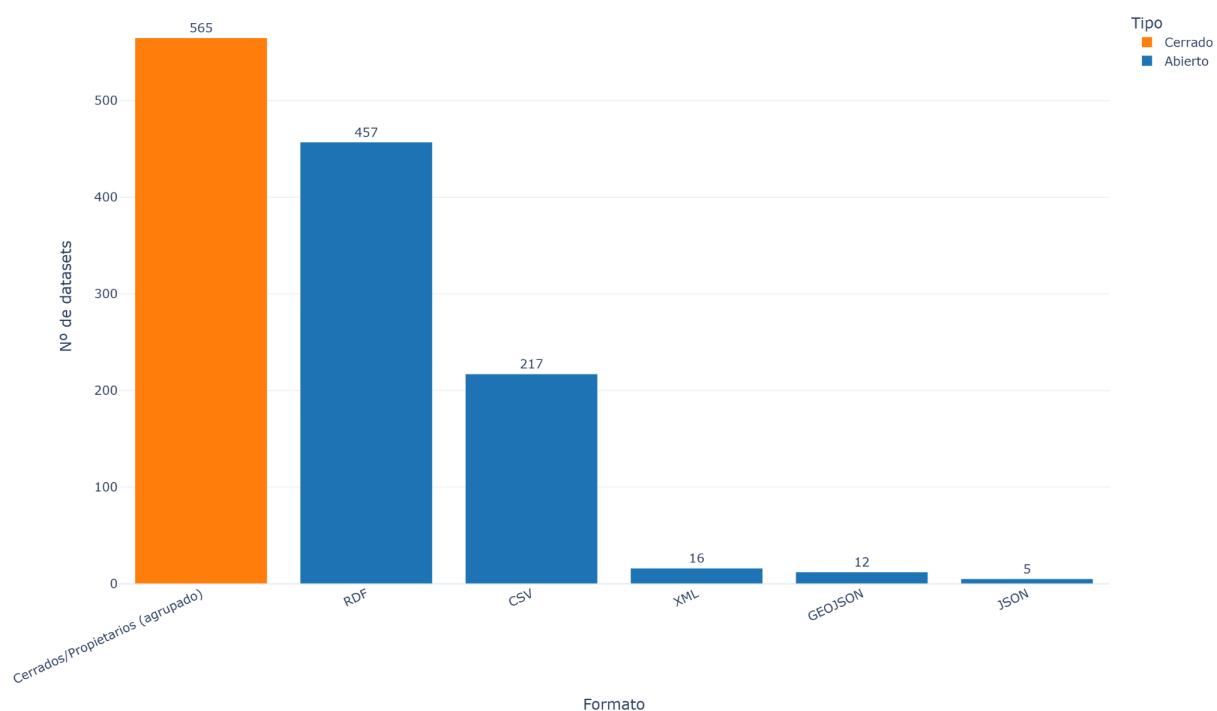


Figura 7.20. Resultado Clasificación de formatos para Interoperabilidad técnica

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado).

Accesibilidad (Dimensión legal) — Licenciamiento permitido

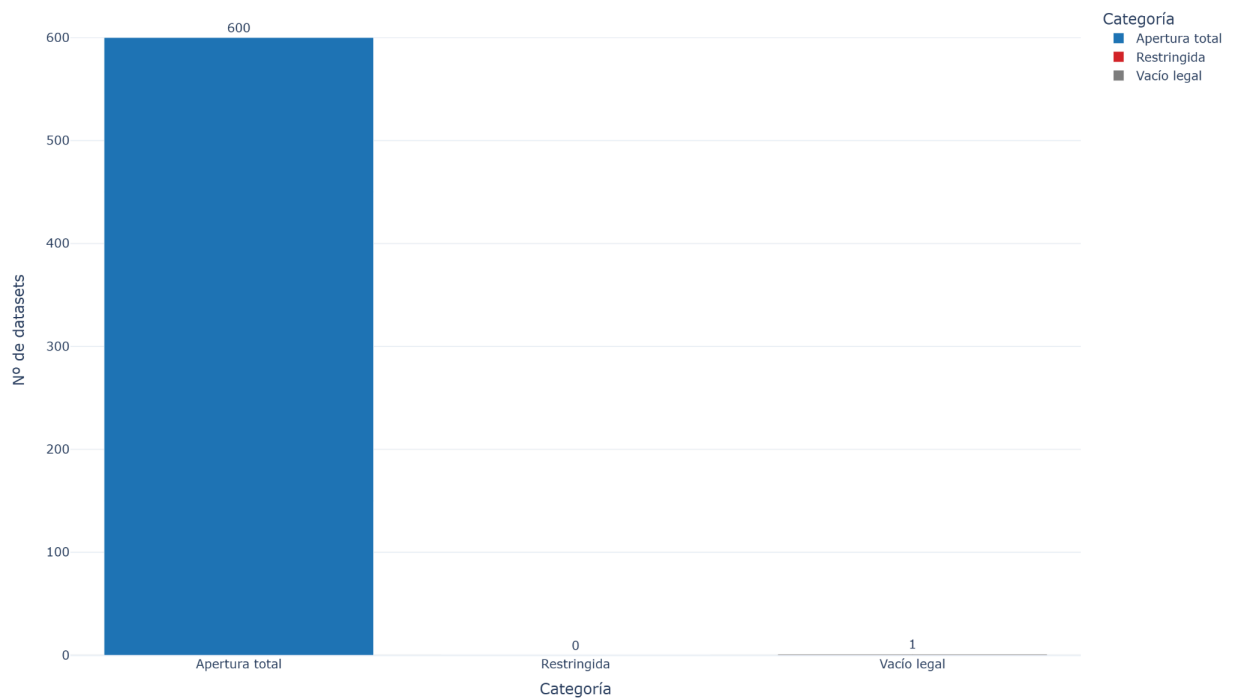


Figura 7.21. Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar.

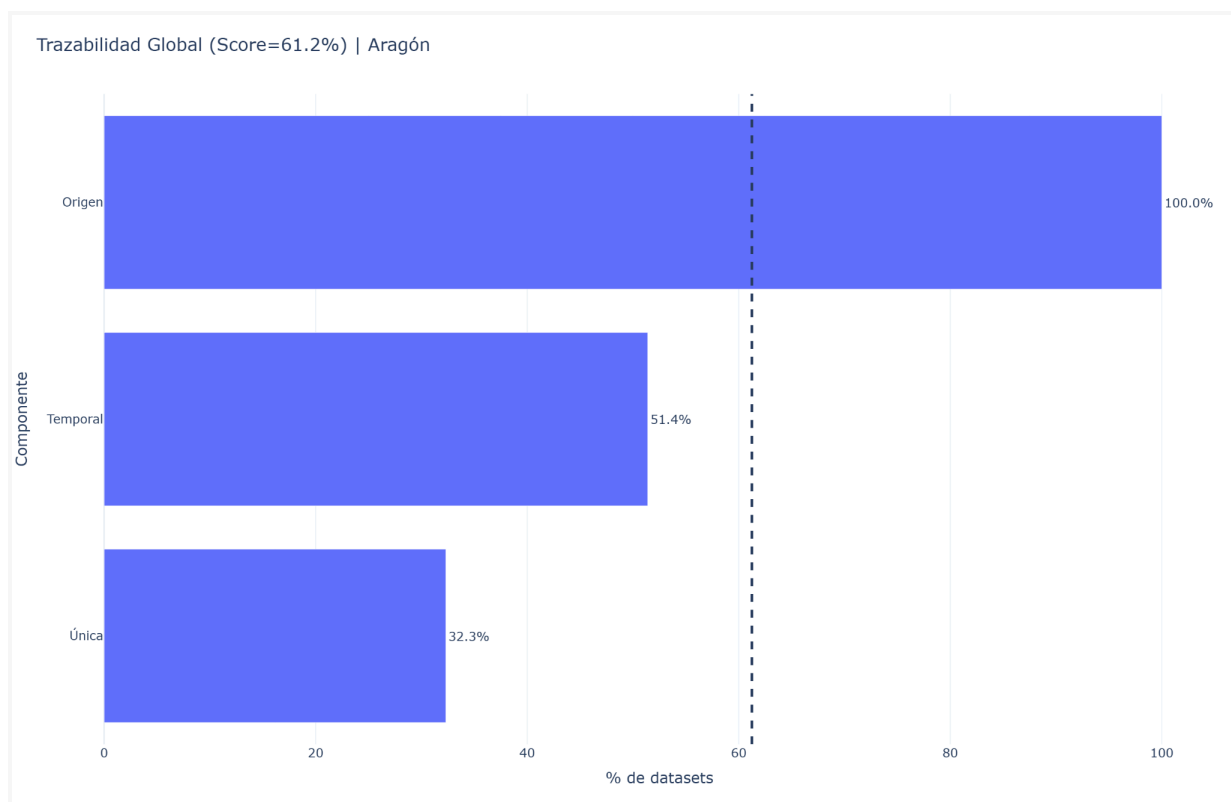


Figura 7.22. Resultado Gráfico Trazabilidad Global

La figura es el resultado global de tres componentes que lo forman porcentaje promedio de cumplimiento.

Trazabilidad Única: Lo conforman el DOI (Código único y permanente) y la url de acceso al dataset.

Trazabilidad Origen: Publicador, identificador (del dataset)

Trazabilidad Temporal: Frecuencia de actualización, Antigüedad del dataset desde su última actualización (valores medidos con actualización, por horas, diaria, semanal, mensual, hasta 6 meses) lo que hace un dataset trazable temporalmente.

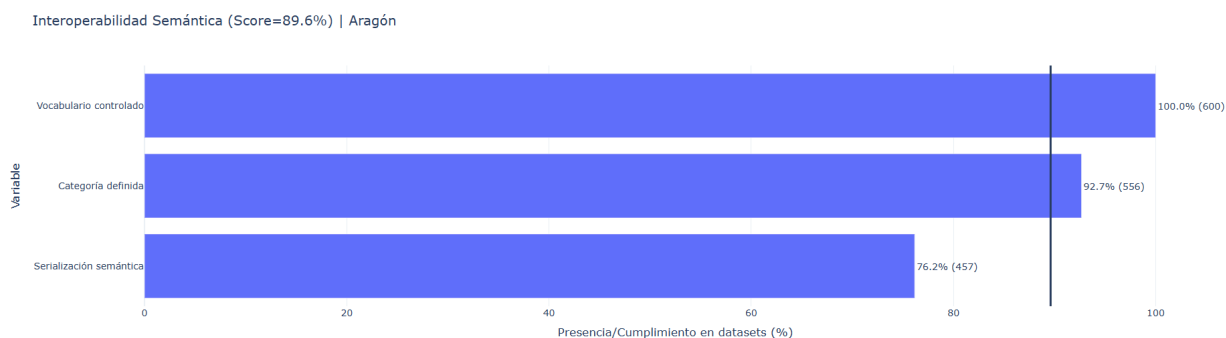


Figura 7.23. Resultado Gráfico Métrica de Interoperabilidad Semántica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 89.6%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

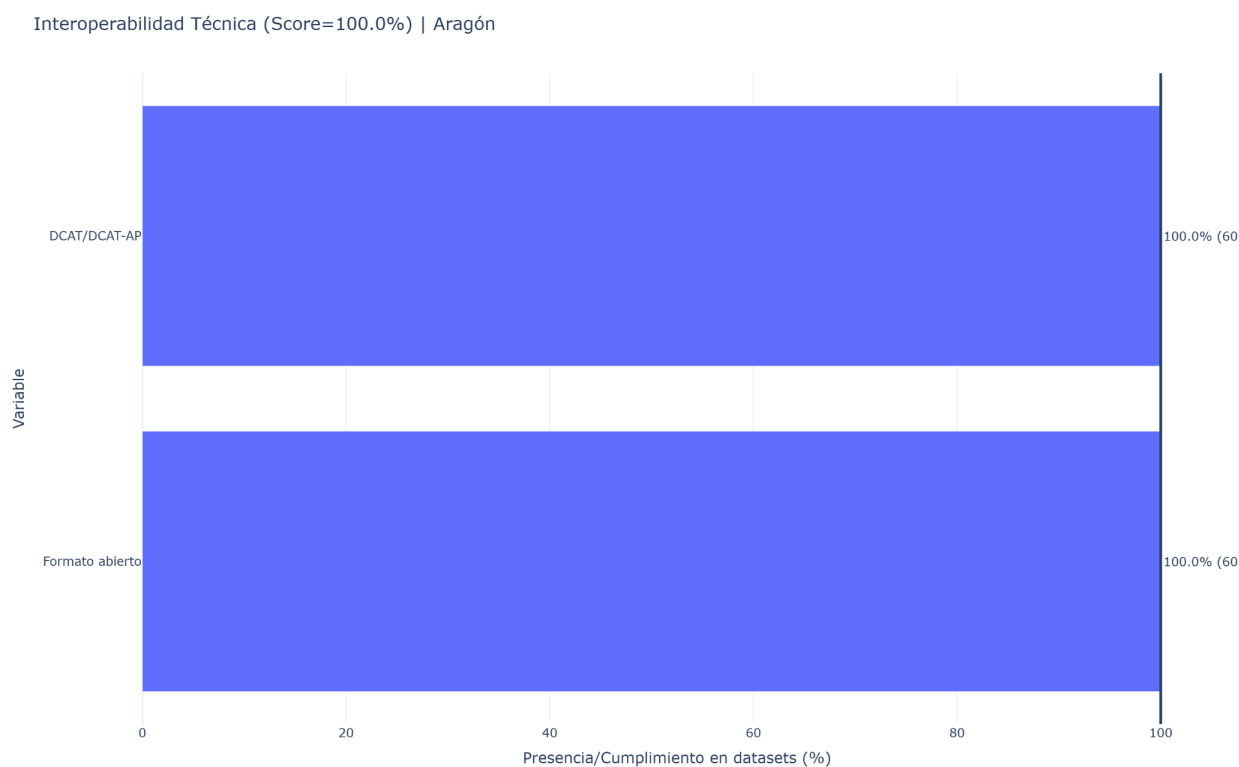


Figura 7.24. Resultado Gráfico Métrica de Interoperabilidad Técnica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 100%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

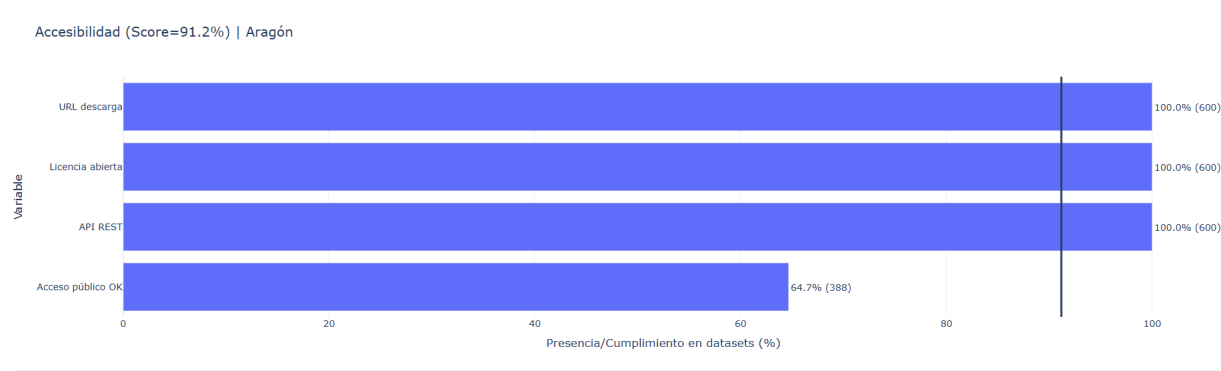


Figura 7.25. Resultado Gráfico Métrica de Accesibilidad

La figura muestra el promedio del total de las variables que lo forman (total de datasets 91.2%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

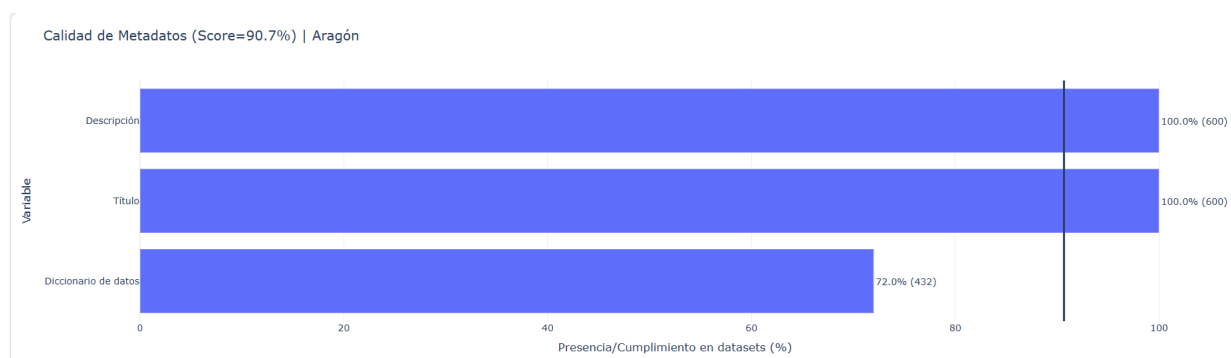


Figura 7.26. Resultado Gráfico Métrica de Calidad de metadatos.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 90.7%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

7.3.4. Portal Castilla y León

El análisis del portal Datos abiertos Castilla y León se basa en una muestra de 419 conjuntos de datos, extraídos a través de su API institucional Opendatasoft. Esta muestra abarca varias categorías temáticas, entre el top 3 destacan Salud, Sector

público, Cultura. Se observa una alta densidad de metadatos actualizados, lo que permite evaluar su madurez técnica bajo un marco de consistencia temporal elevado.

Índice Global de Madurez del Portal — Castilla y León (N=419)



Figura 7.27. Resultado Gráfico Índice de Madurez del Portal.

La figura muestra el Índice global calculado como promedio **ponderado** de 5 componentes: Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos. Nivel definido Bajo 0–30, Medio 30–70, Alto 70–100.

Componentes que forman el Índice de Madurez — Castilla y León

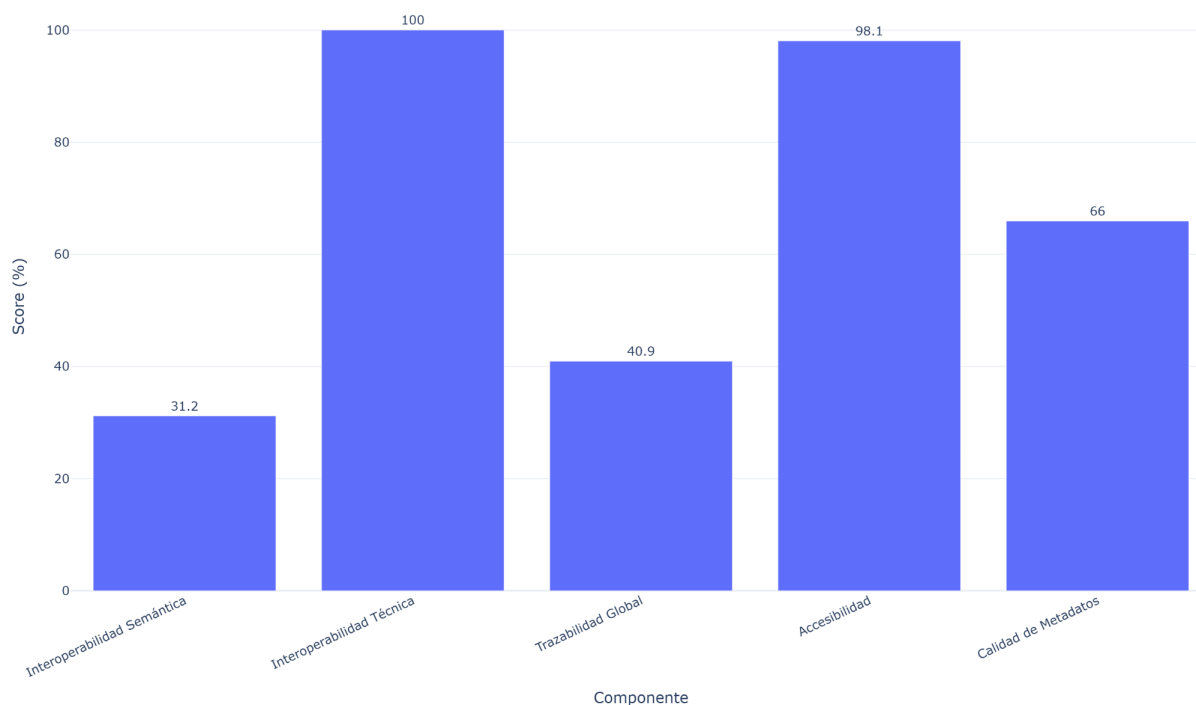


Figura 7.28. Resultado Gráfico de componentes de Índice de Madurez del Portal.

La figura muestra Cada score de dimensión se calcula como promedio del cumplimiento de los componentes que forman cada métrica (peso igual por componente). Los 5 componentes construyen el Índice de Madurez del Portal.

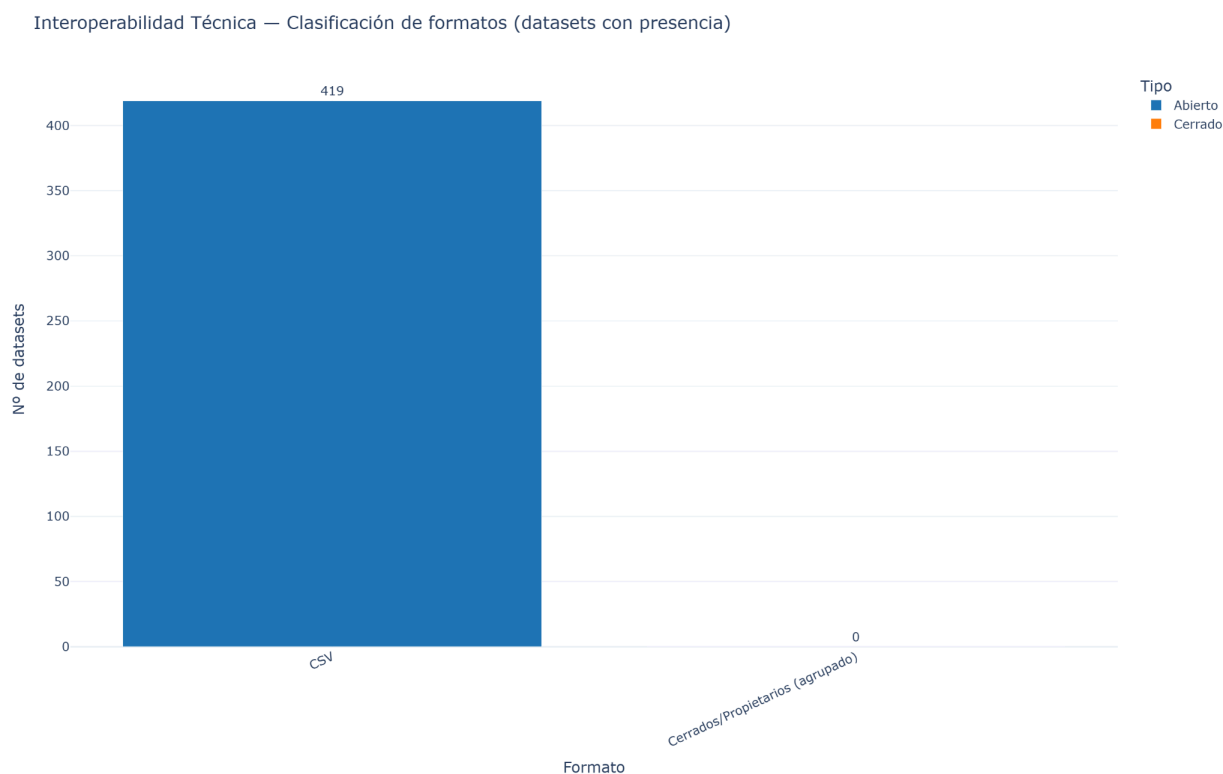


Figura 7.29. Resultado Clasificación de formatos para Interoperabilidad técnica

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado).

Accesibilidad (Dimensión legal) — Licenciamiento permitido

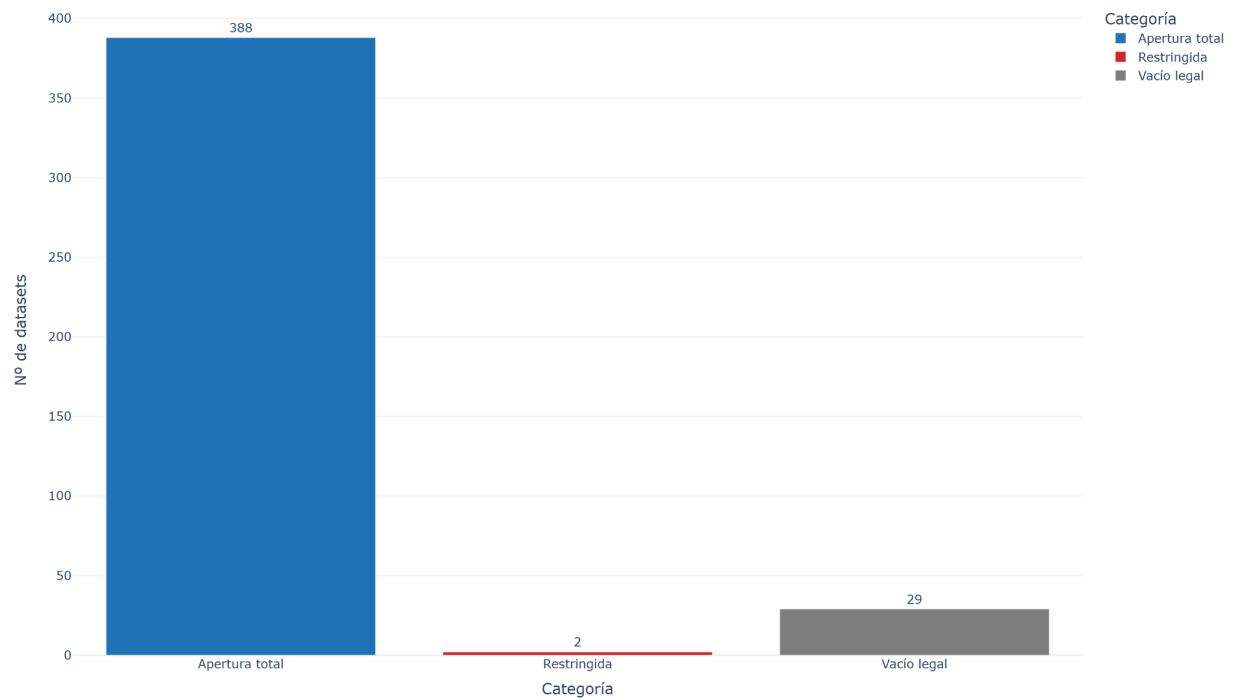


Figura 7.30. Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar.

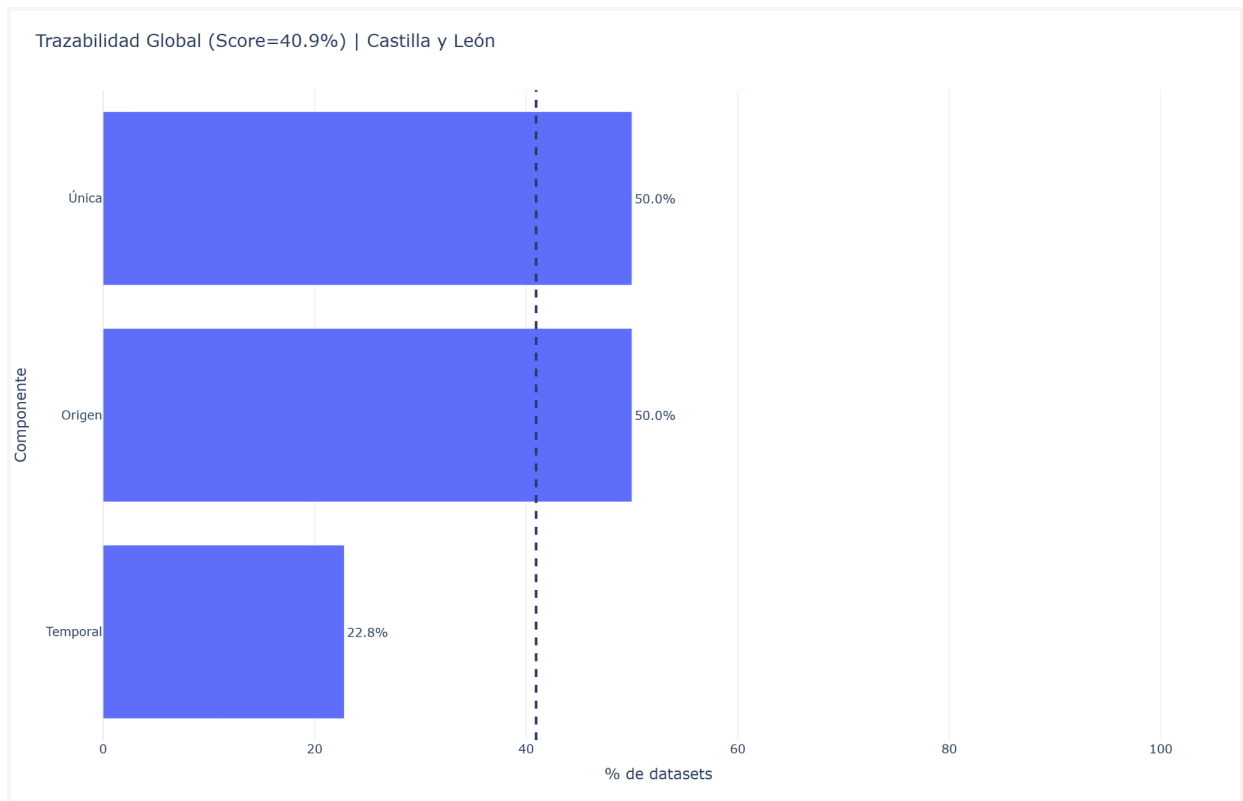


Figura 7.31. Resultado Gráfico Trazabilidad Global

La figura es el resultado global de tres componentes que lo forman porcentaje promedio de cumplimiento.

Trazabilidad Única: Lo conforman el DOI (Código único y permanente) y la url de acceso al dataset.

Trazabilidad Origen: Publicador, identificador (del dataset)

Trazabilidad Temporal: Frecuencia de actualización, Antigüedad del dataset desde su última actualización (valores medidos con actualización, por horas, diaria, semanal, mensual, hasta 6 meses) lo que hace un dataset trazable temporalmente.

Interoperabilidad Semántica (Score=31.2%) | Castilla y León

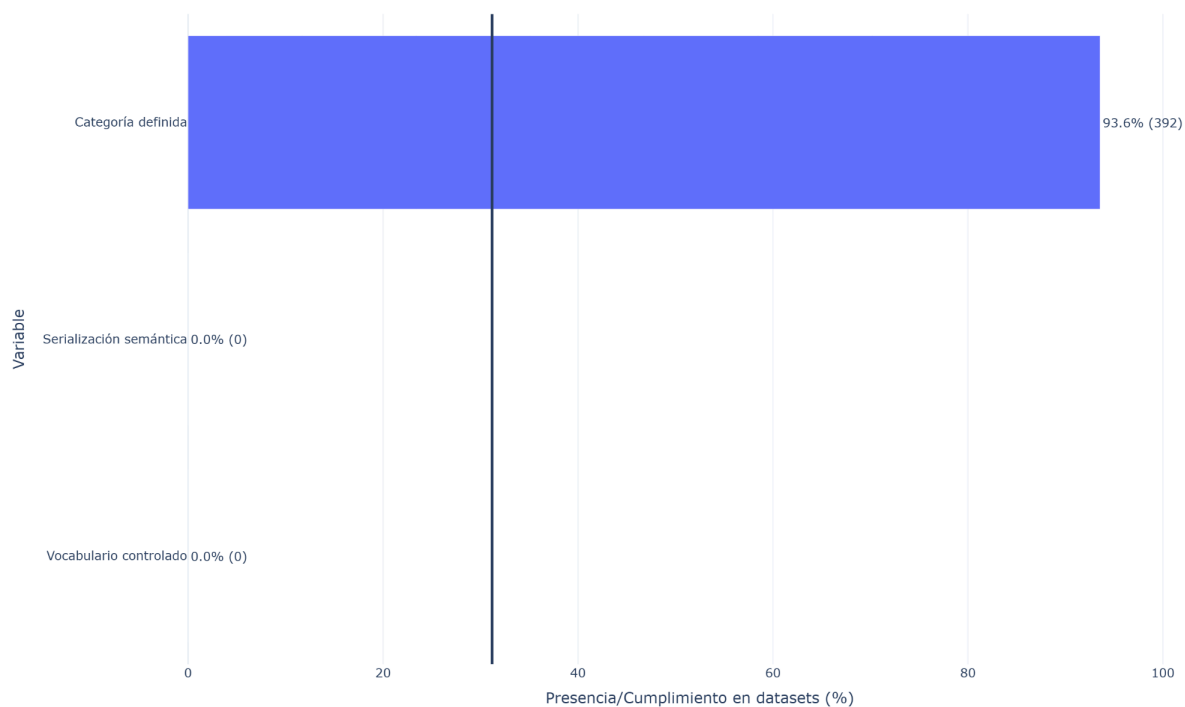


Figura 7.32. Resultado Gráfico Métrica de Interoperabilidad Semántica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 31.2%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

Interoperabilidad Técnica (Score=100.0%) | Castilla y León

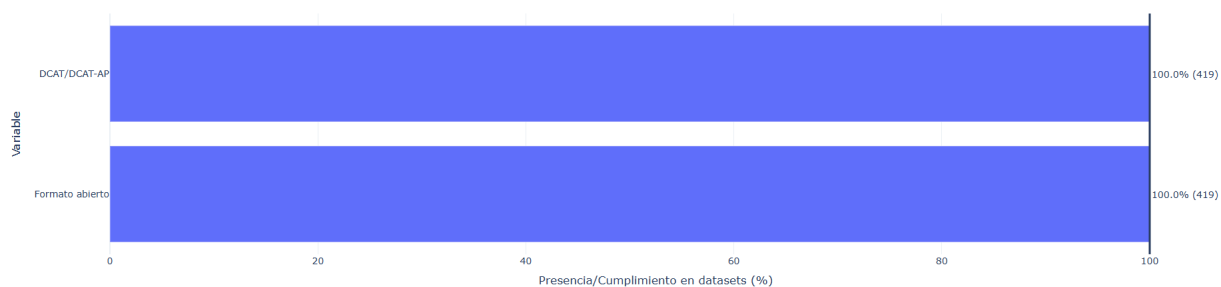


Figura 7.33. Resultado Gráfico Métrica de Interoperabilidad Técnica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 100%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

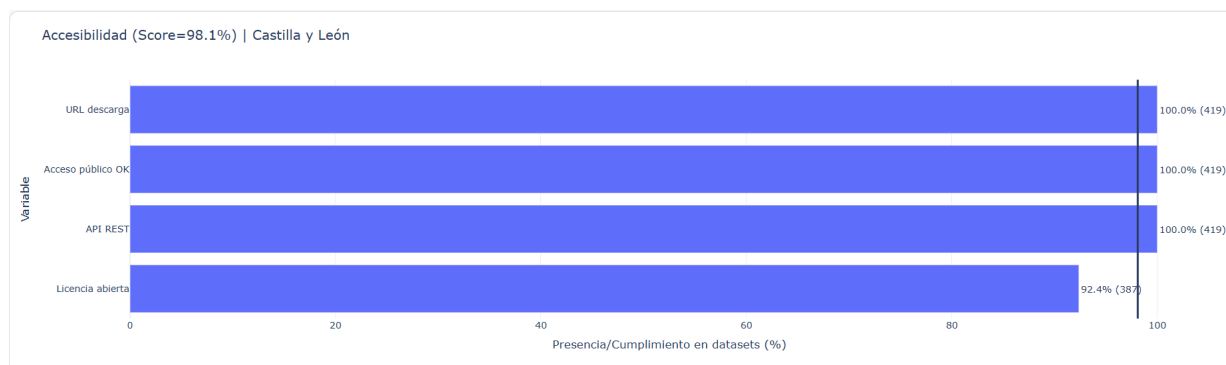


Figura 7.34. Resultado Gráfico Métrica de Accesibilidad

La figura muestra el promedio del total de las variables que lo forman (total de datasets 98.1%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

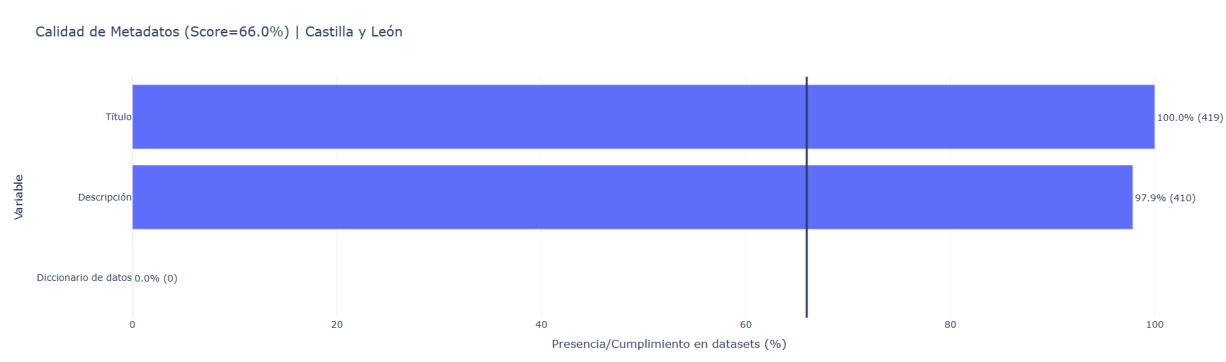


Figura 7.35. Resultado Gráfico Métrica de Calidad de metadatos.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 66.0%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

7.3.5. Portal Comunidad de Madrid

El análisis del portal Datos abiertos Comunidad de Madrid se basa en una muestra de 600 conjuntos de datos, extraídos a través de su API institucional CKAN. Esta muestra abarca varias categorías temáticas, entre el top 3 destacan Educación,

Salud, Datos demográficos y estadísticos. Se observa una alta densidad de metadatos actualizados, lo que permite evaluar su madurez técnica bajo un marco de consistencia temporal elevado.

Índice Global de Madurez del Portal — Comunidad de Madrid (N=600)

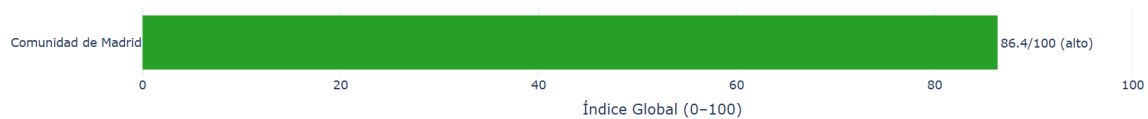


Figura 7.36. Resultado Gráfico Índice de Madurez del Portal.

La figura muestra el Índice global calculado como promedio **ponderado** de 5 componentes: Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos. Nivel definido Bajo 0–30, Medio 30–70, Alto 70–100.

Componentes que forman el Índice de Madurez — Comunidad de Madrid

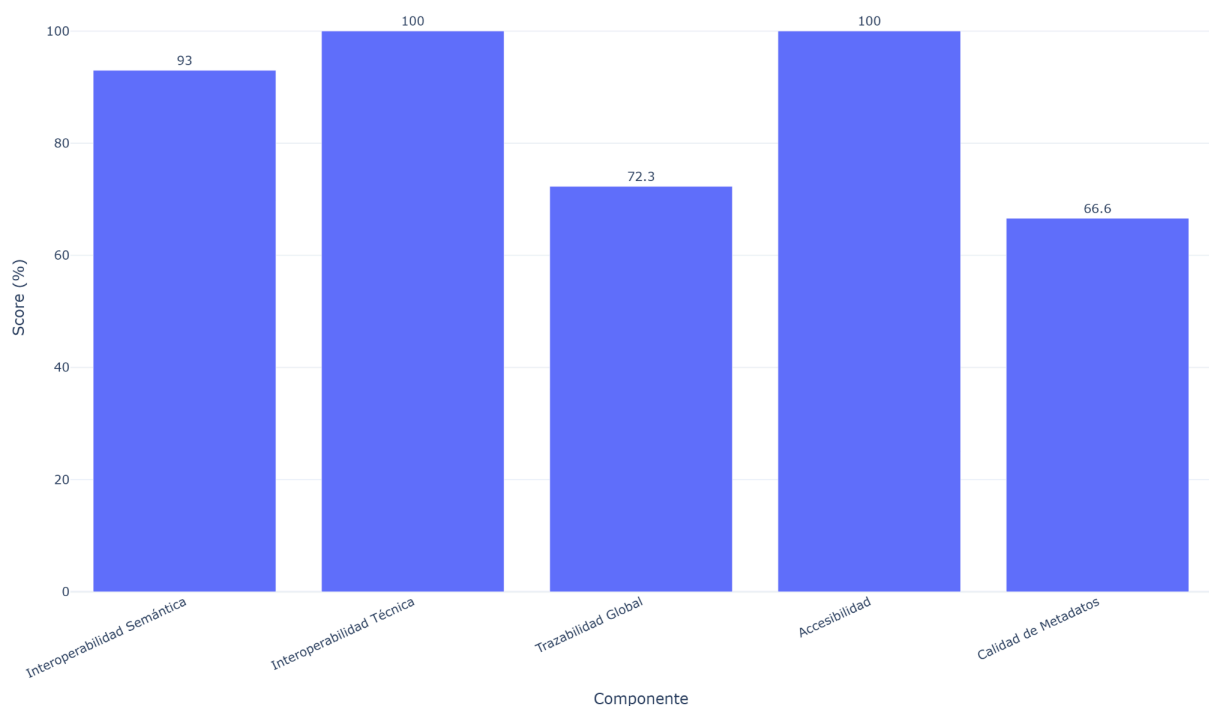


Figura 7.37. Resultado Gráfico de componentes de Índice de Madurez del Portal.

La figura muestra Cada score de dimensión se calcula como promedio del cumplimiento de los componentes que forman cada métrica (peso igual por componente). Los 5 componentes construyen el Índice de Madurez del Portal.

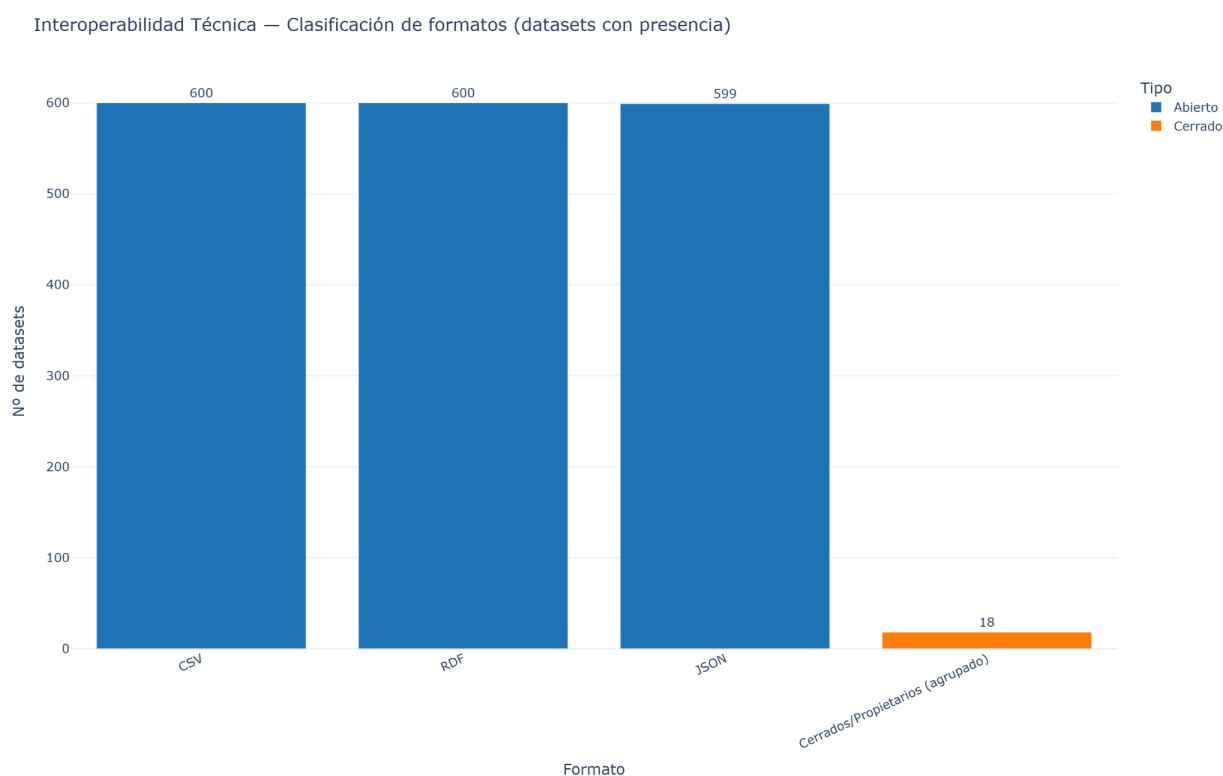


Figura 7.38. Resultado Clasificación de formatos para Interoperabilidad técnica

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado).

Accesibilidad (Dimensión legal) — Licenciamiento permitido

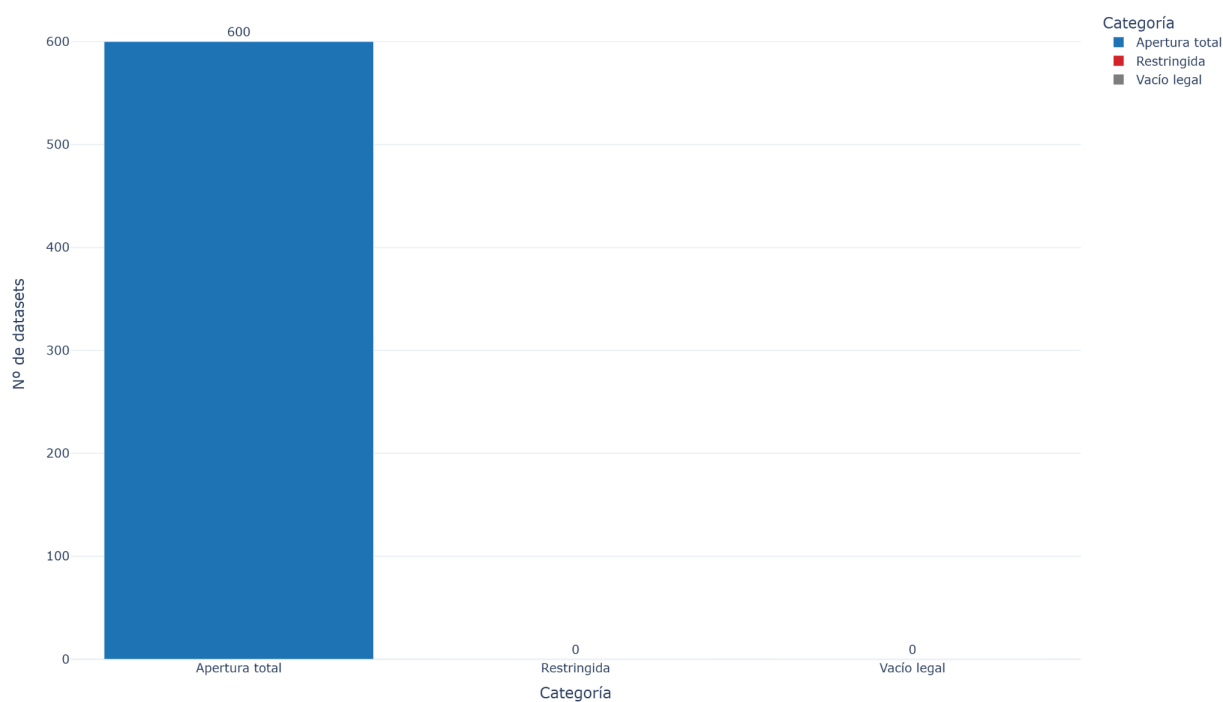


Figura 7.39. Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar.

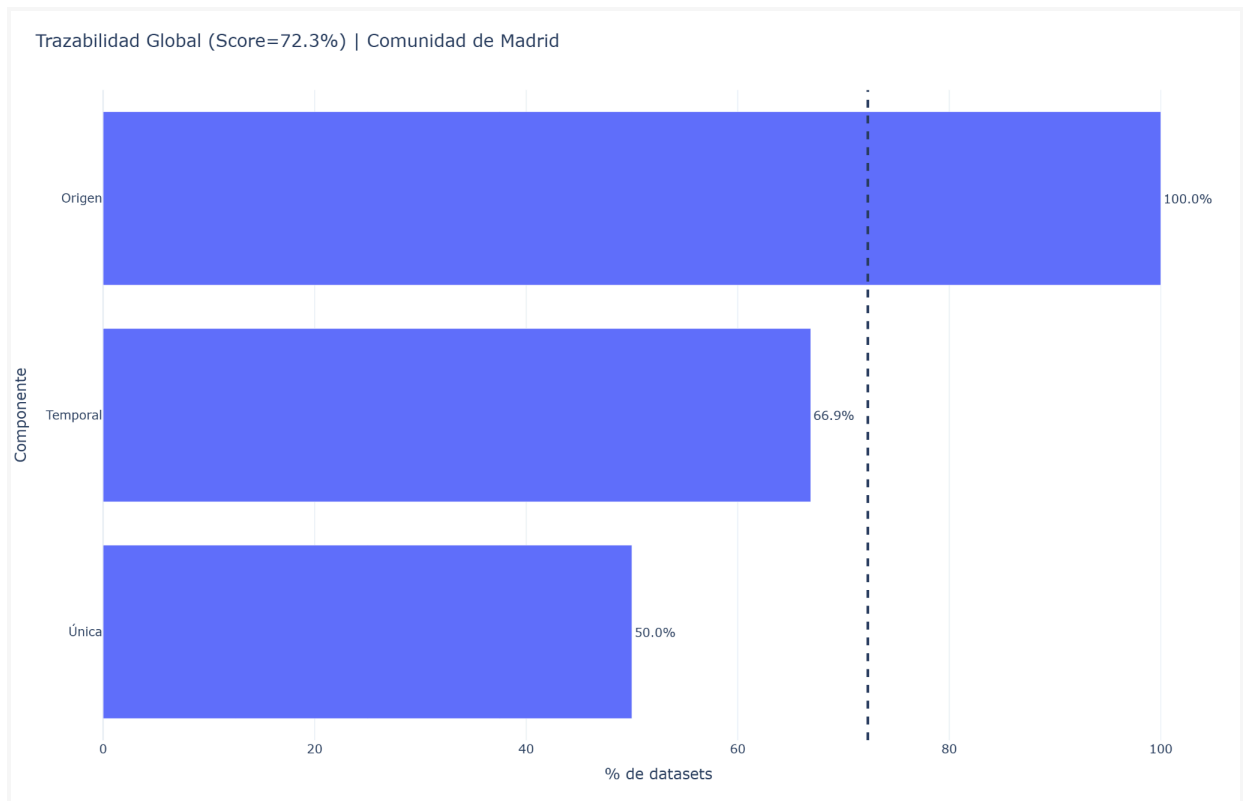


Figura 7.40. Resultado Gráfico Trazabilidad Global

La figura es el resultado global de tres componentes que lo forman porcentaje promedio de cumplimiento.

Trazabilidad Única: Lo conforman el DOI (Código único y permanente) y la url de acceso al dataset.

Trazabilidad Origen: Publicador, identificador (del dataset)

Trazabilidad Temporal: Frecuencia de actualización, Antigüedad del dataset desde su última actualización (valores medidos con actualización, por horas, diaria, semanal, mensual, hasta 6 meses) lo que hace un dataset trazable temporalmente.

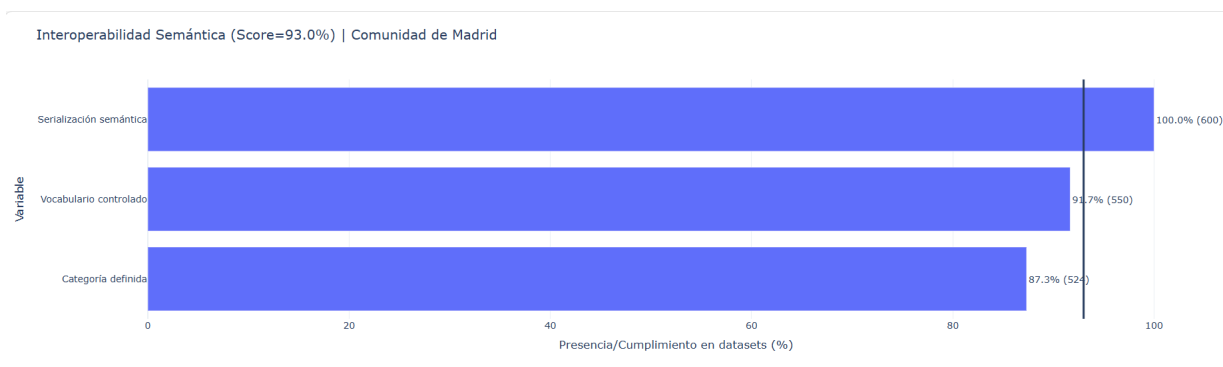


Figura 7.41. Resultado Gráfico Métrica de Interoperabilidad Semántica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 93%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

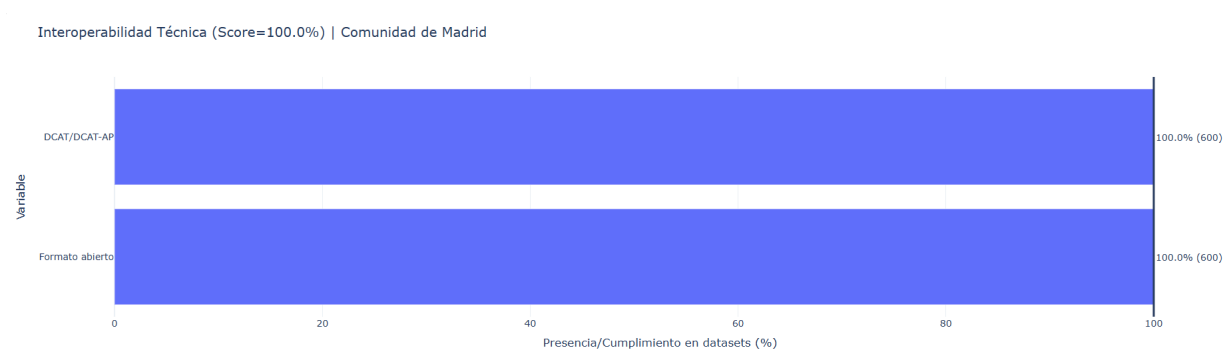


Figura 7.42. Resultado Gráfico Métrica de Interoperabilidad Técnica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 100%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

Accesibilidad (Score=100.0%) | Comunidad de Madrid

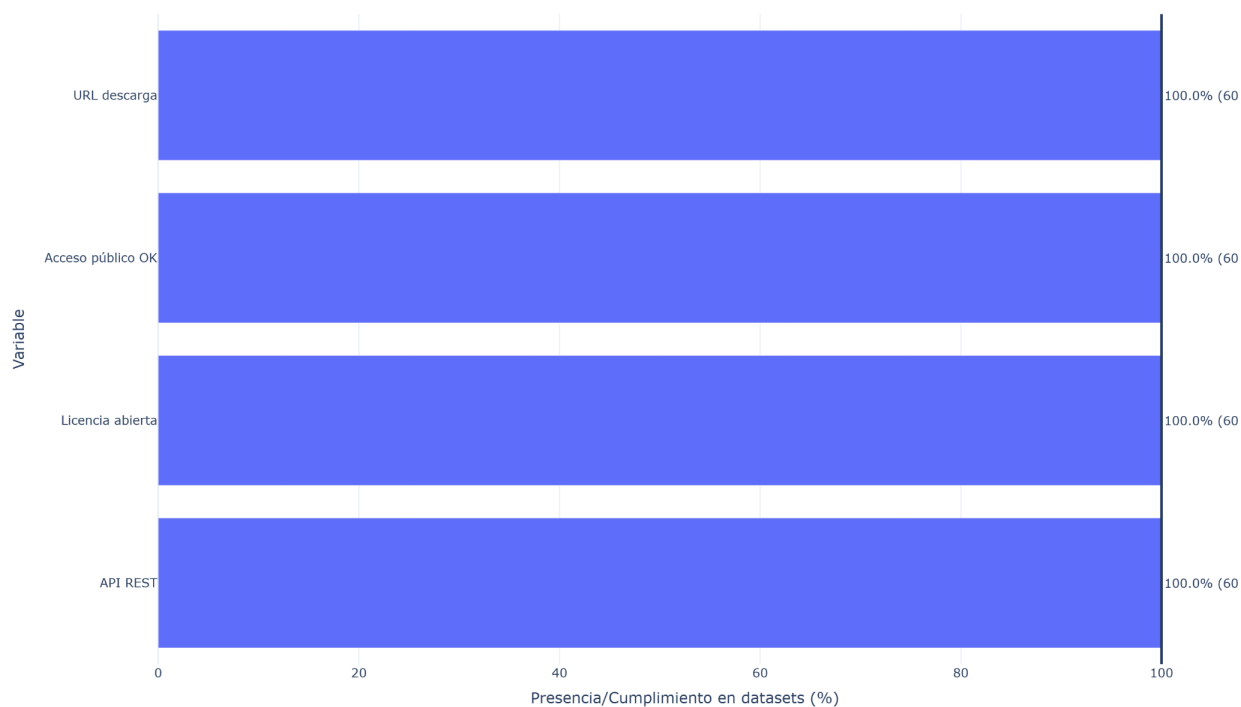


Figura 7.43. Resultado Gráfico Métrica de Accesibilidad

La figura muestra el promedio del total de las variables que lo forman (total de datasets 100%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

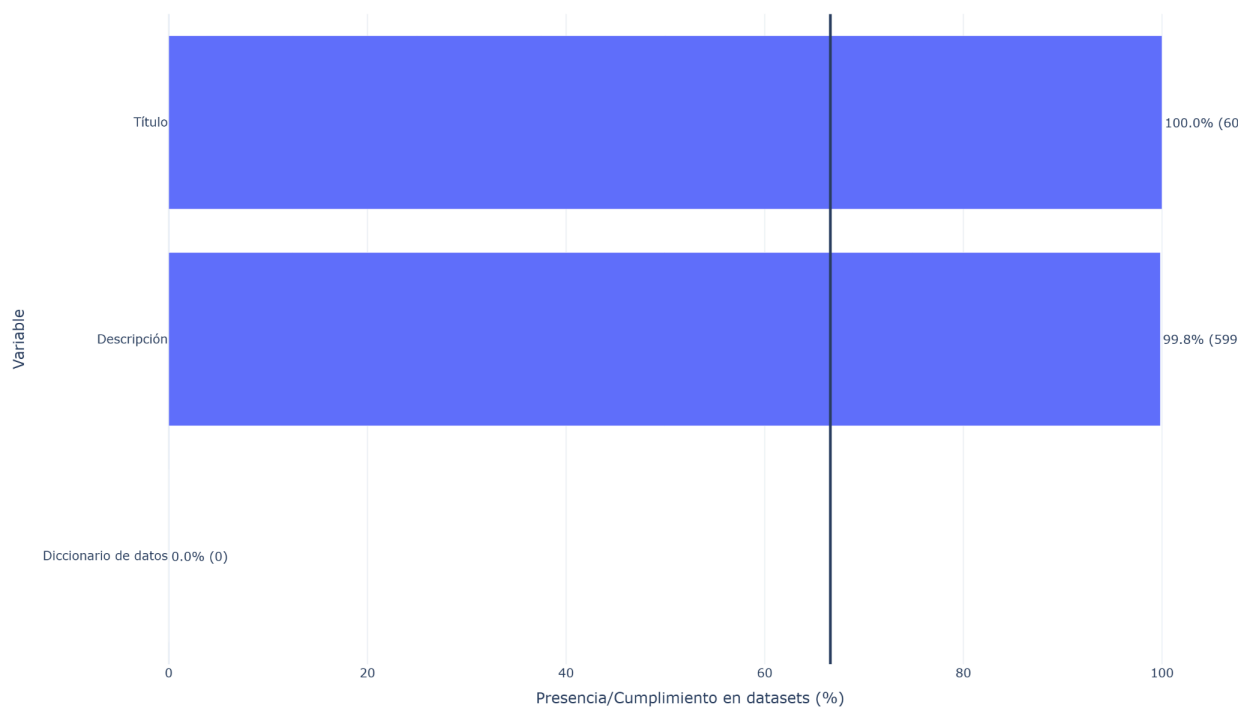


Figura 7.44. Resultado Gráfico Métrica de Calidad de metadatos.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 66.6%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

7.3.6. Portal Región de Murcia

El análisis del portal Datos abiertos Región Murcia se basa en una muestra de 598 conjuntos de datos, extraídos a través de su API institucional CKAN. Esta muestra abarca varias categorías temáticas, entre el top 3 destacan Datos geográficos y medioambientales, Sector público, Datos demográficos y estadísticos. Se observa una alta densidad de metadatos actualizados, lo que permite evaluar su madurez técnica bajo un marco de consistencia temporal elevado.

Índice Global de Madurez del Portal — Región de Murcia (N=598)



Figura 7.45. Resultado Gráfico Índice de Madurez del Portal.

La figura muestra el Índice global calculado como promedio **ponderado** de 5 componentes: Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos. Nivel definido Bajo 0–30, Medio 30–70, Alto 70–100.

Componentes que forman el Índice de Madurez — Región de Murcia

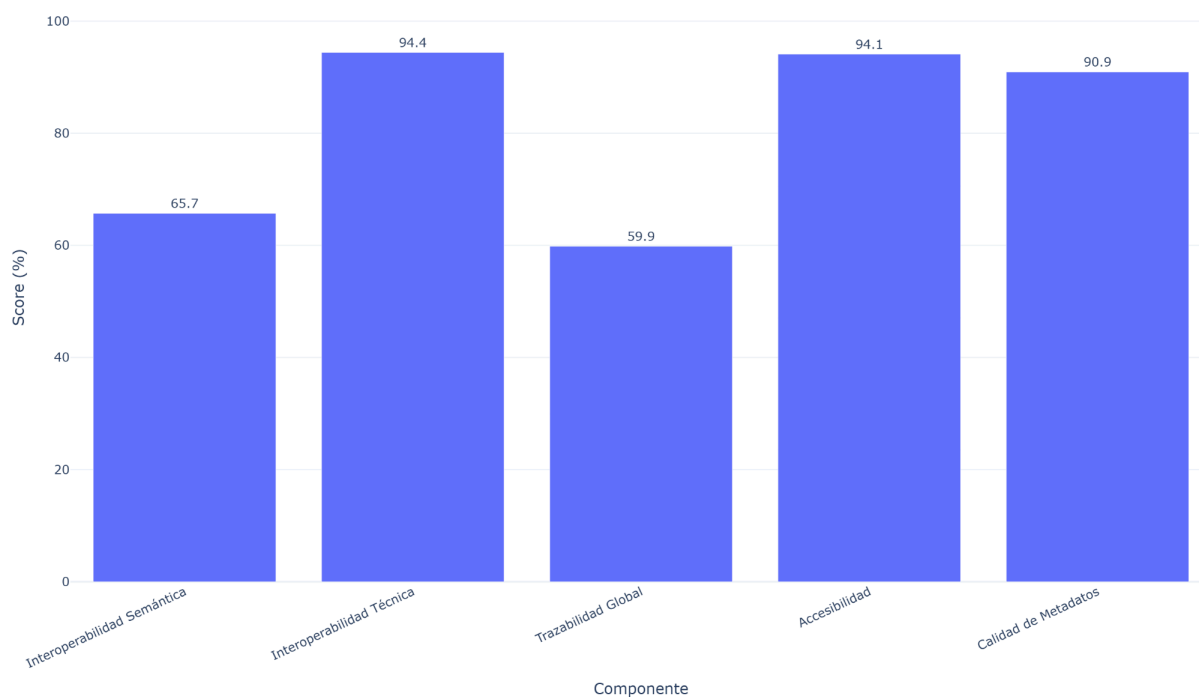


Figura 7.46. Resultado Gráfico de componentes de Índice de Madurez del Portal.

La figura muestra Cada score de dimensión se calcula como promedio del cumplimiento de los componentes que forman cada métrica (peso igual por componente). Los 5 componentes construyen el Índice de Madurez del Portal.

Interoperabilidad Técnica — Clasificación de formatos (datasets con presencia)

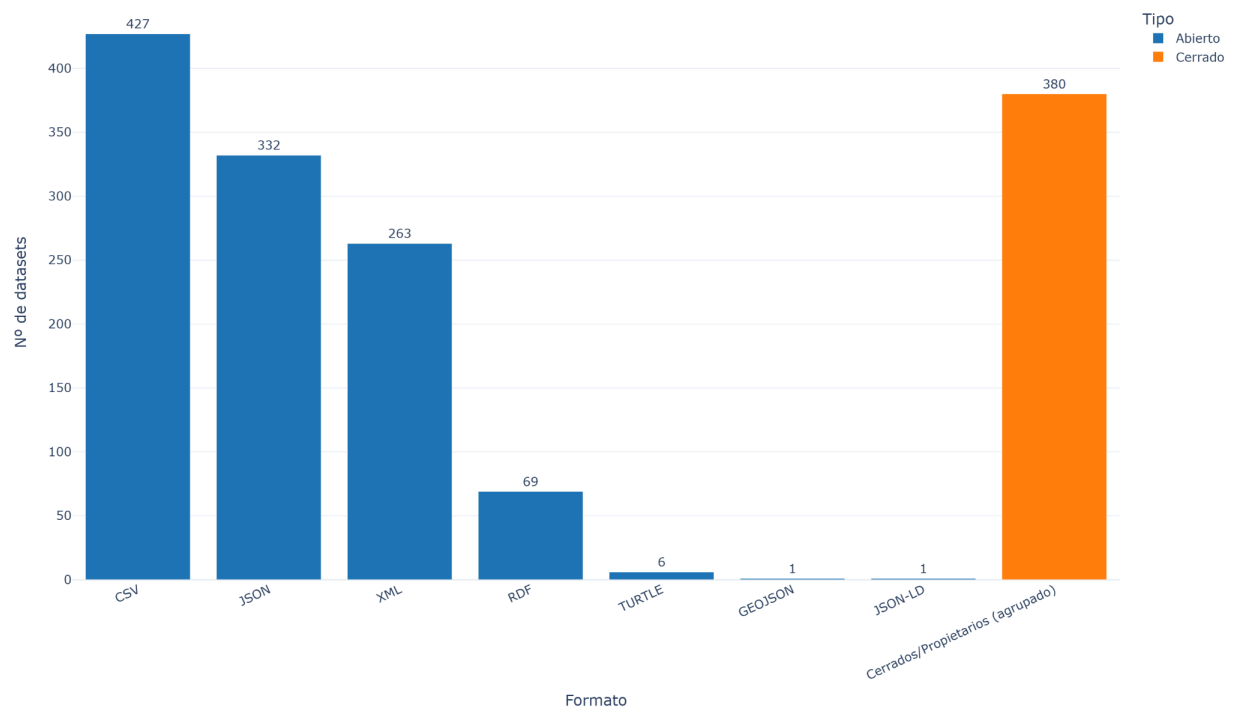


Figura 7.47. Resultado Clasificación de formatos para Interoperabilidad técnica

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado).

Accesibilidad (Dimensión legal) — Licenciamiento permitido

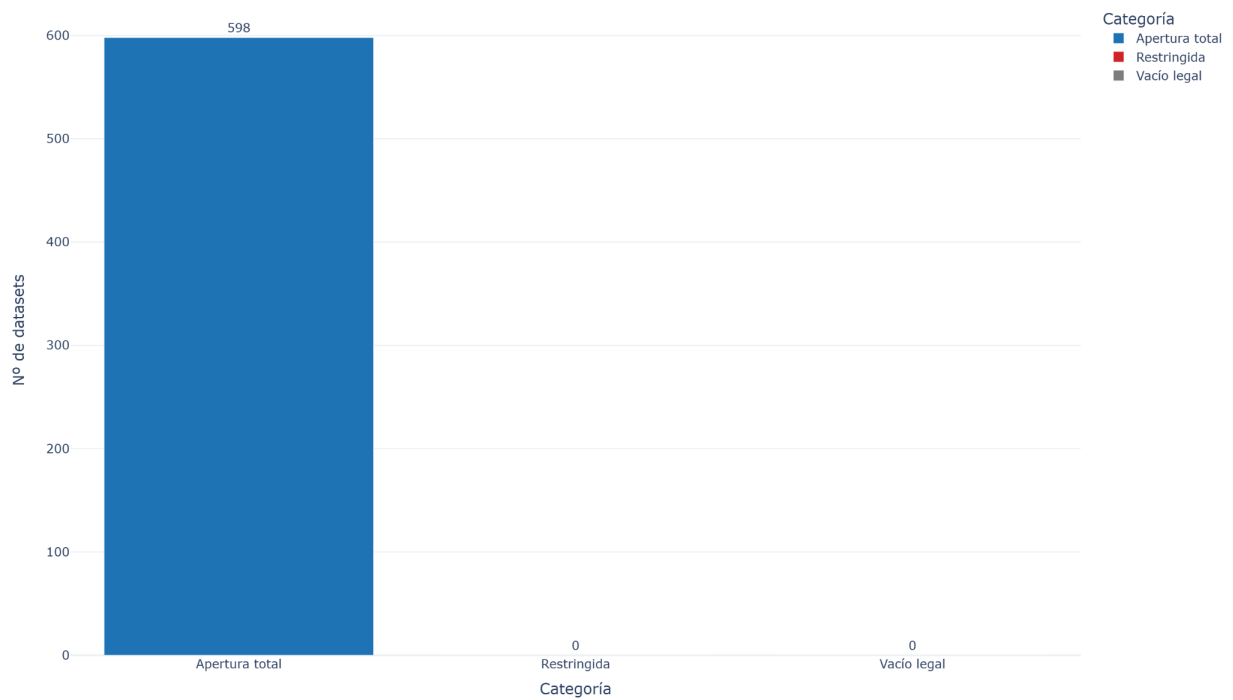


Figura 7.48. Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar.

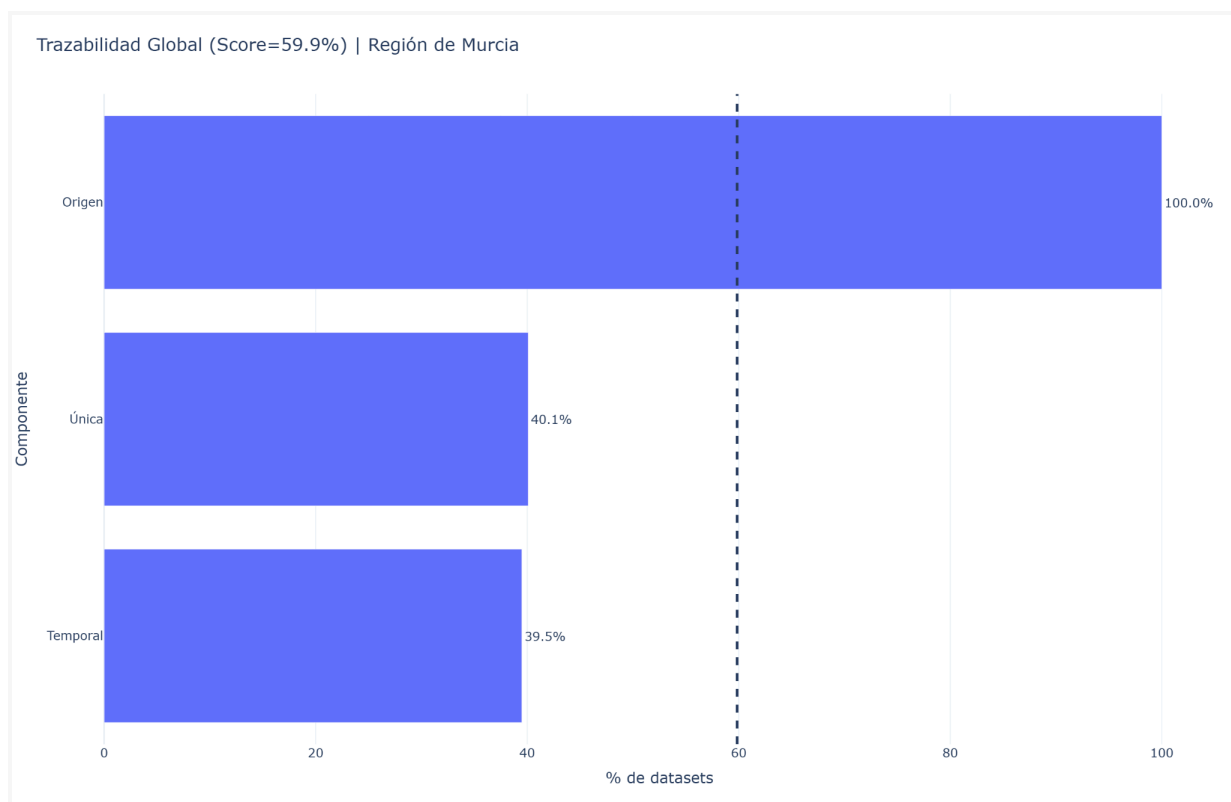


Figura 7.49. Resultado Gráfico Trazabilidad Global

La figura es el resultado global de tres componentes que lo forman porcentaje promedio de cumplimiento.

Trazabilidad Única: Lo conforman el DOI (Código único y permanente) y la url de acceso al dataset.

Trazabilidad Origen: Publicador, identificador (del dataset)

Trazabilidad Temporal: Frecuencia de actualización, Antigüedad del dataset desde su última actualización (valores medidos con actualización, por horas, diaria, semanal, mensual, hasta 6 meses) lo que hace un dataset trazable temporalmente.

Interoperabilidad Semántica (Score=65.7%) | Región de Murcia

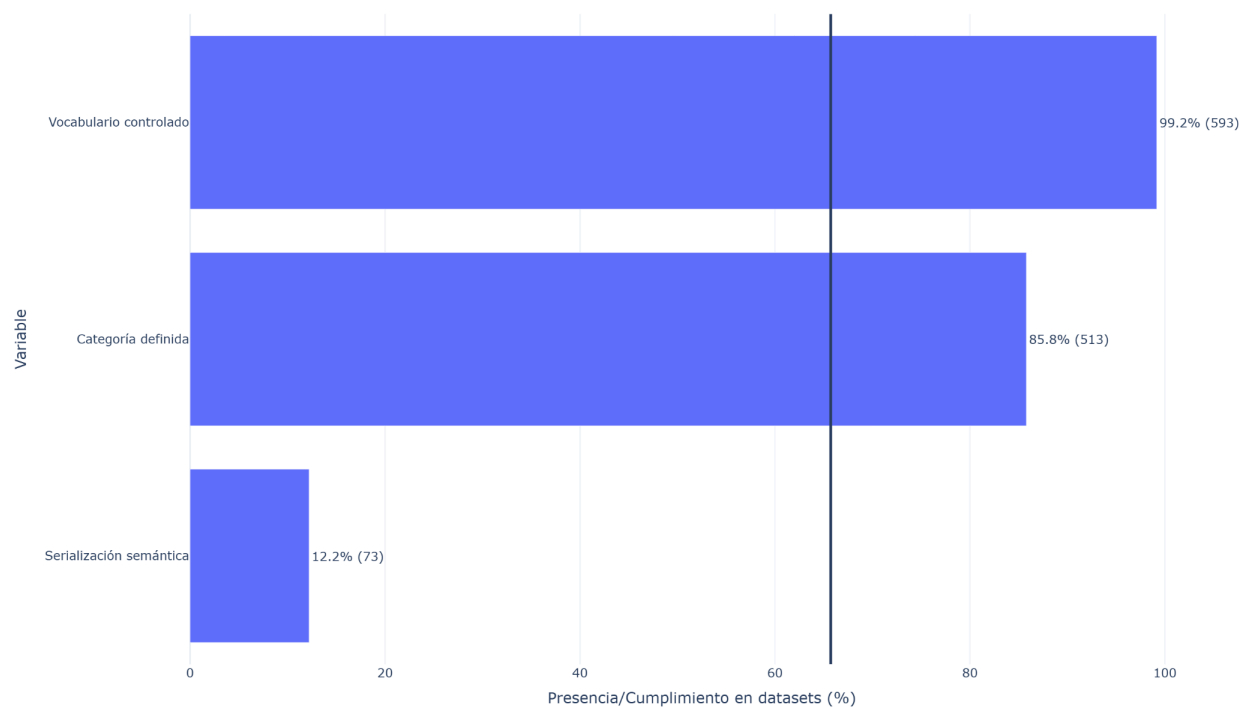


Figura 7.50. Resultado Gráfico Métrica de Interoperabilidad Semántica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

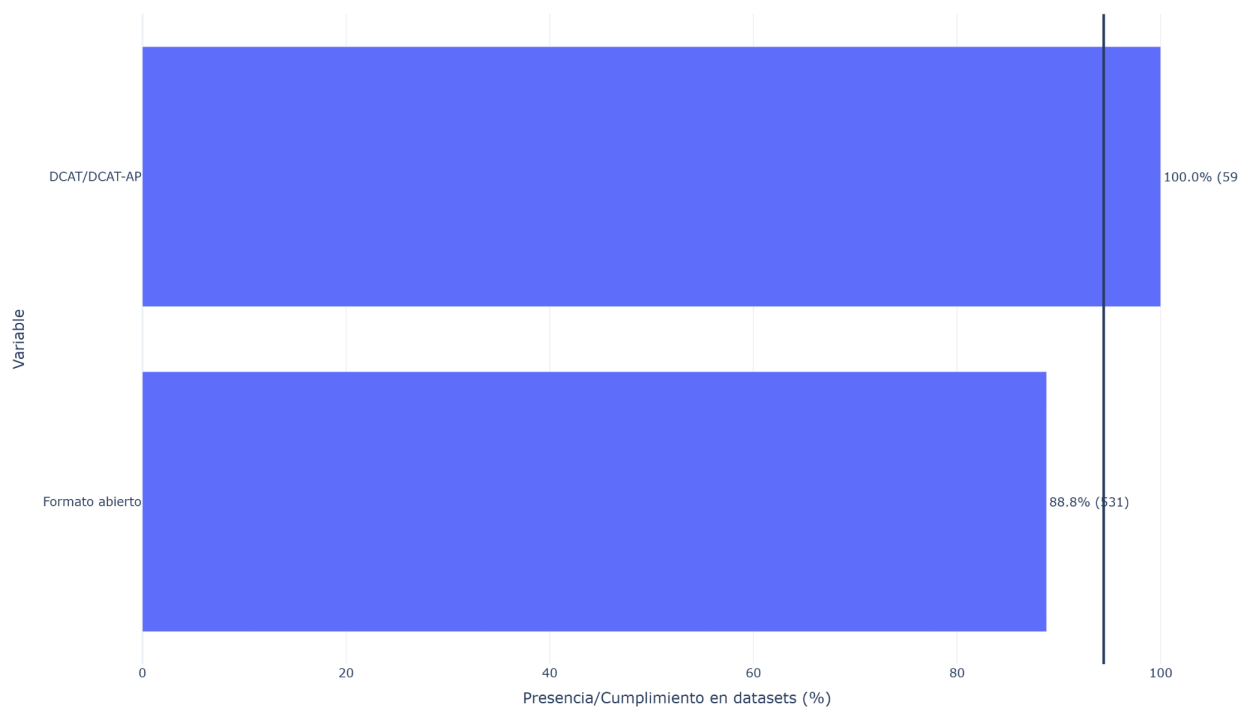


Figura 7.51. Resultado Gráfico Métrica de Interoperabilidad Técnica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

Accesibilidad (Score=94.1%) | Región de Murcia

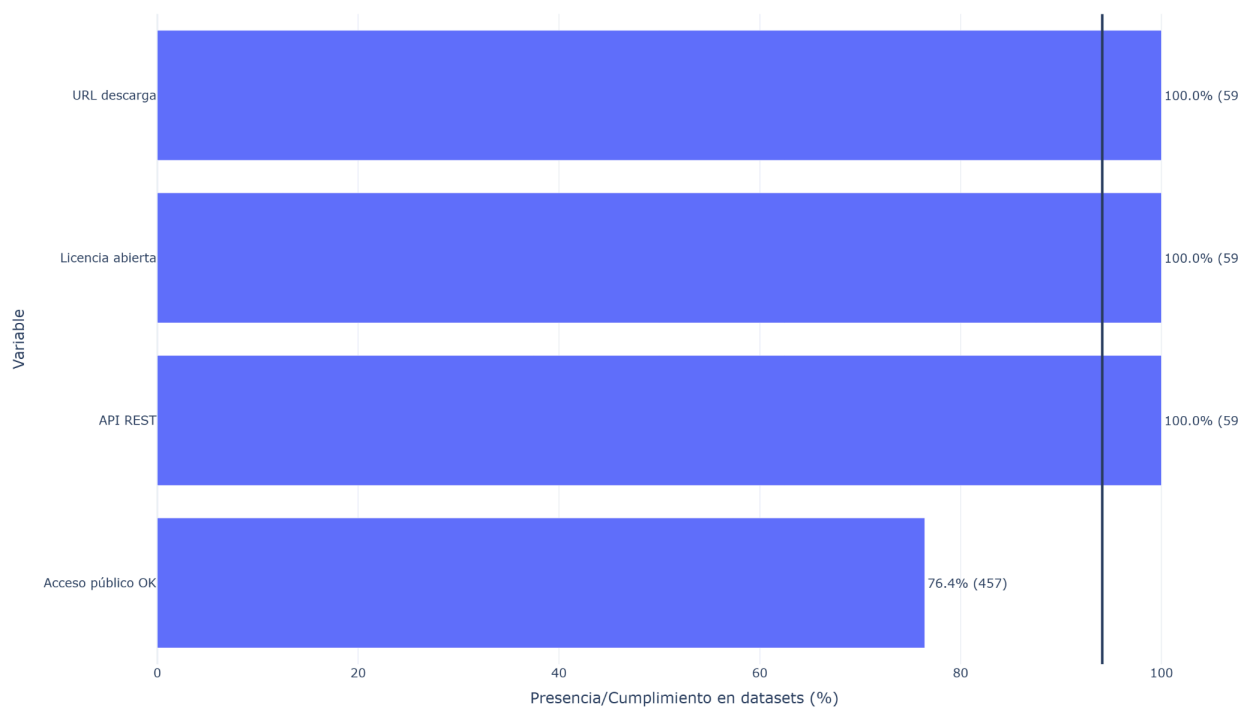


Figura 7.52. Resultado Gráfico Métrica de Accesibilidad

La figura muestra el promedio del total de las variables que lo forman (total de datasets), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

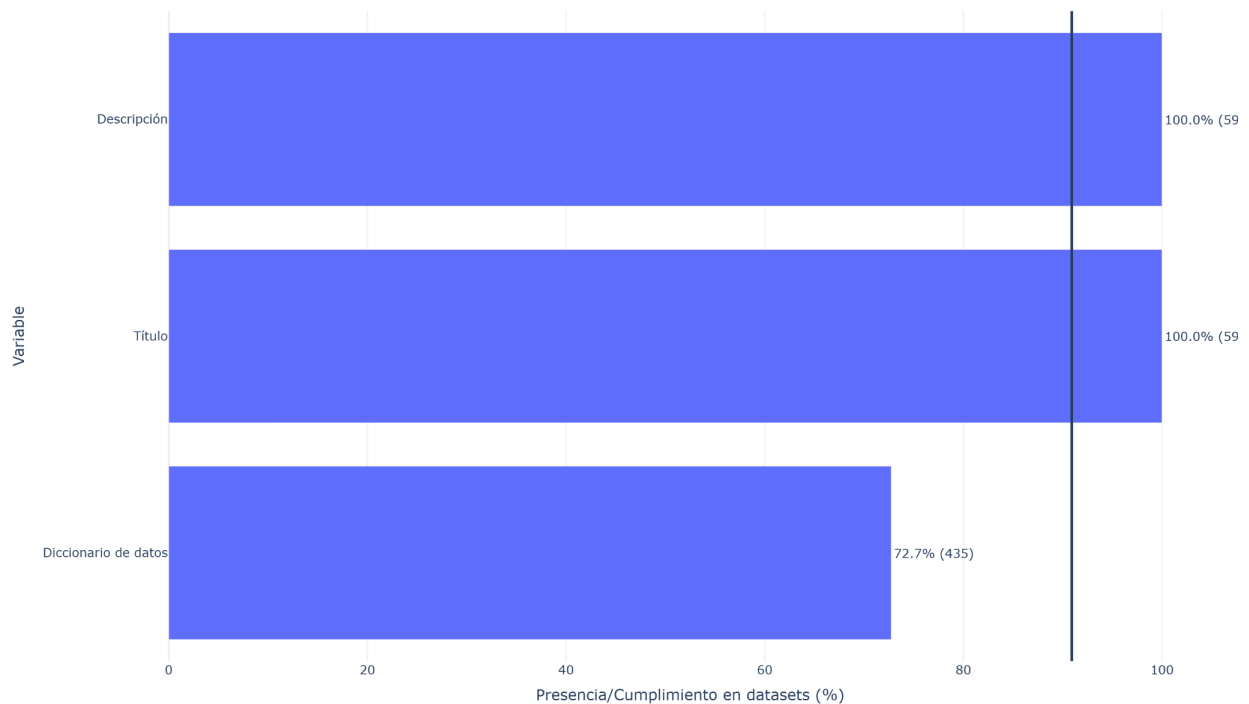


Figura 7.53. Resultado Gráfico Métrica de Calidad de metadatos.

La figura muestra el promedio del total de las variables que lo forman (total de datasets), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

7.3.7. Portal Junta de Andalucía

El análisis del portal Datos abiertos Junta Andalucía se basa en una muestra de 600 conjuntos de datos, extraídos a través de su API institucional CKAN. Esta muestra abarca varias categorías temáticas, entre el top 3 destacan Educación, Datos demográficos y estadísticos, Datos geográficos y medioambientales. Se observa una alta densidad de metadatos actualizados, lo que permite evaluar su madurez técnica bajo un marco de consistencia temporal elevado.

Índice Global de Madurez del Portal — Junta de Andalucía (N=600)

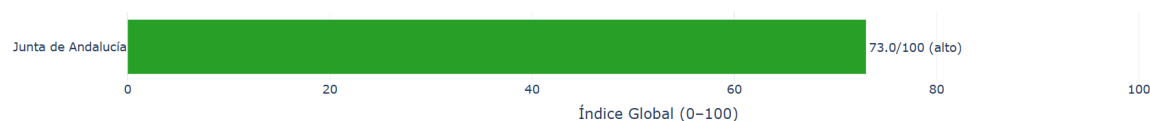


Figura 7.54. Resultado Gráfico Índice de Madurez del Portal.

La figura muestra el Índice global calculado como promedio **ponderado** de 5 componentes: Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos. Nivel definido Bajo 0-30, Medio 30-70, Alto 70-100.

Componentes que forman el Índice de Madurez — Junta de Andalucía

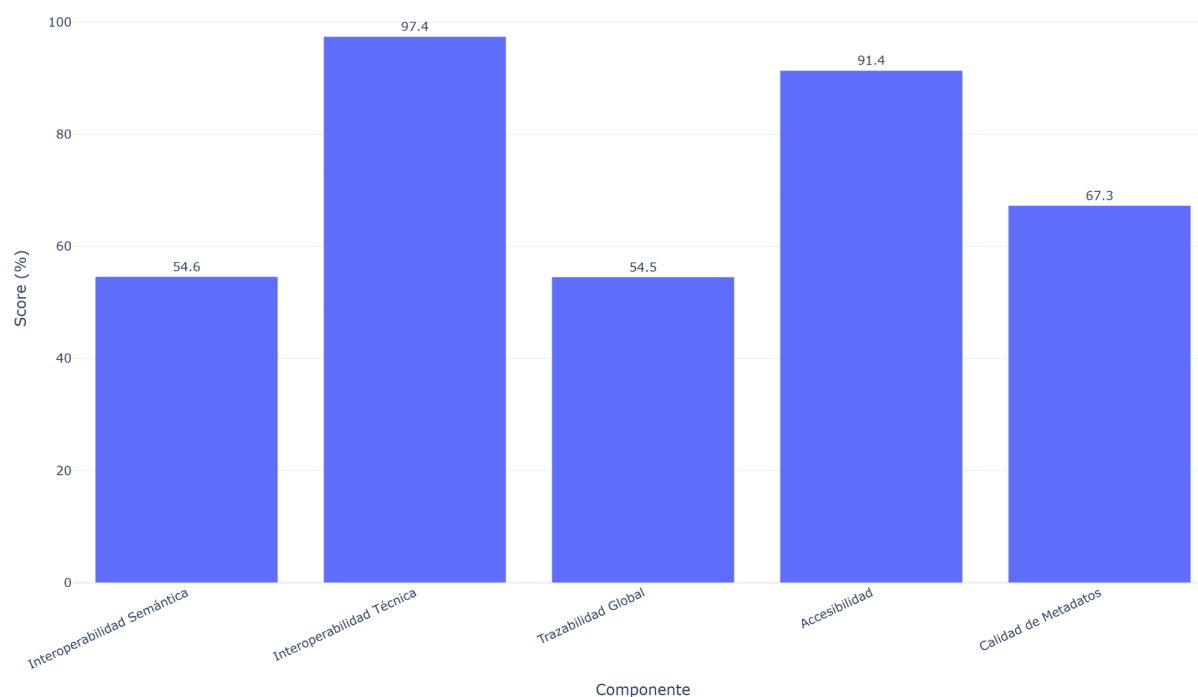


Figura 7.55. Resultado Gráfico de componentes de Índice de Madurez del Portal.

La figura muestra Cada score de dimensión se calcula como promedio del cumplimiento de los componentes que forman cada métrica (peso igual por componente). Los 5 componentes construyen el Índice de Madurez del Portal.

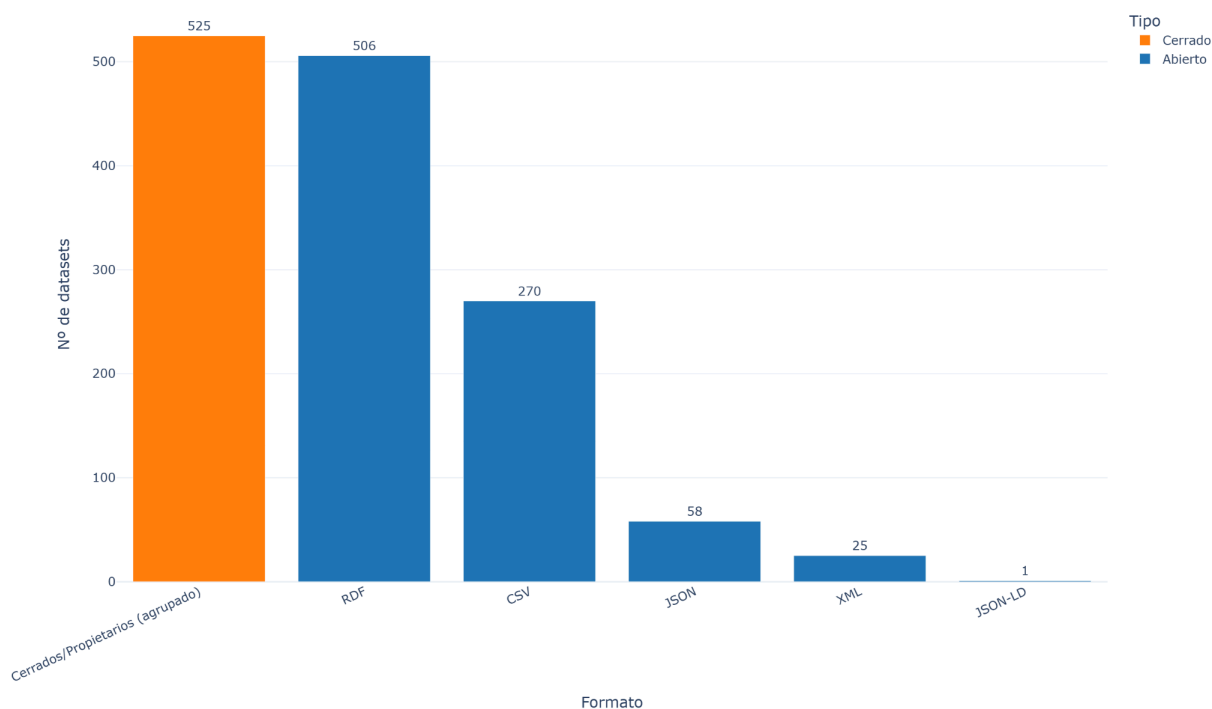


Figura 7.56. Resultado Clasificación de formatos para Interoperabilidad técnica

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado).

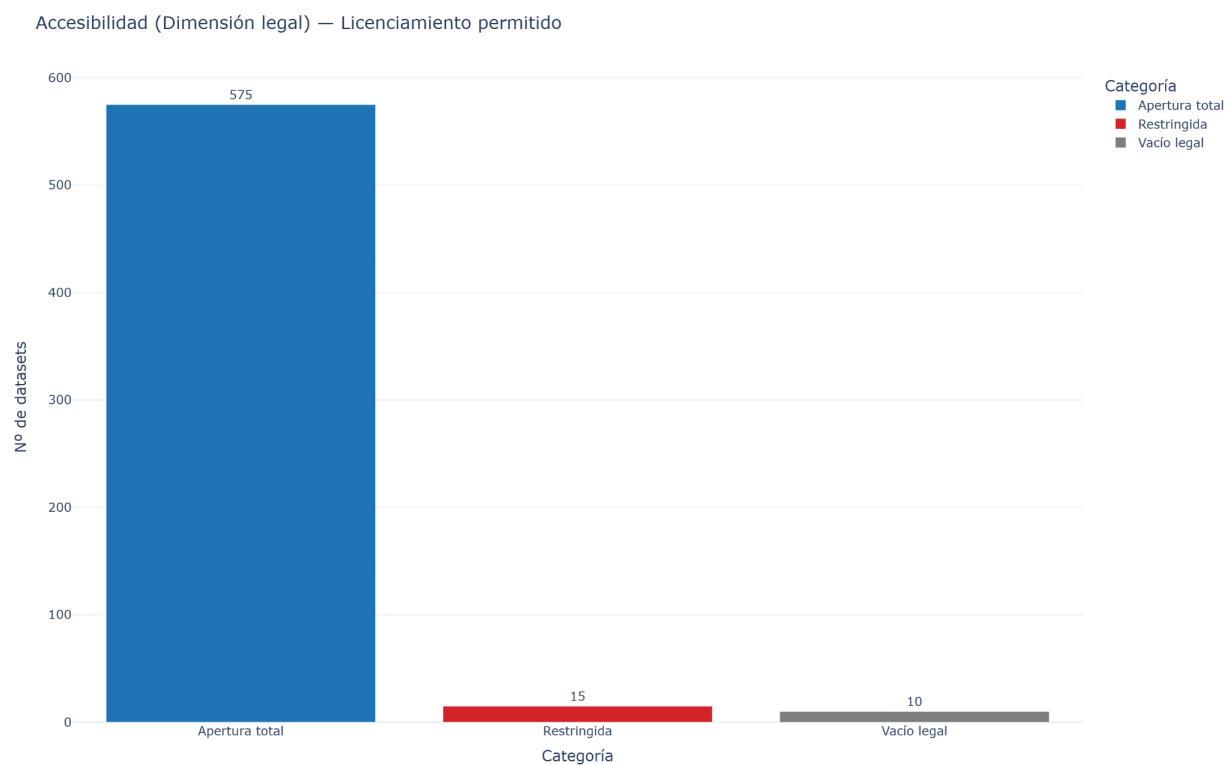


Figura 7.57. Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar.

Trazabilidad Global (Score=54.5%) | Junta de Andalucía

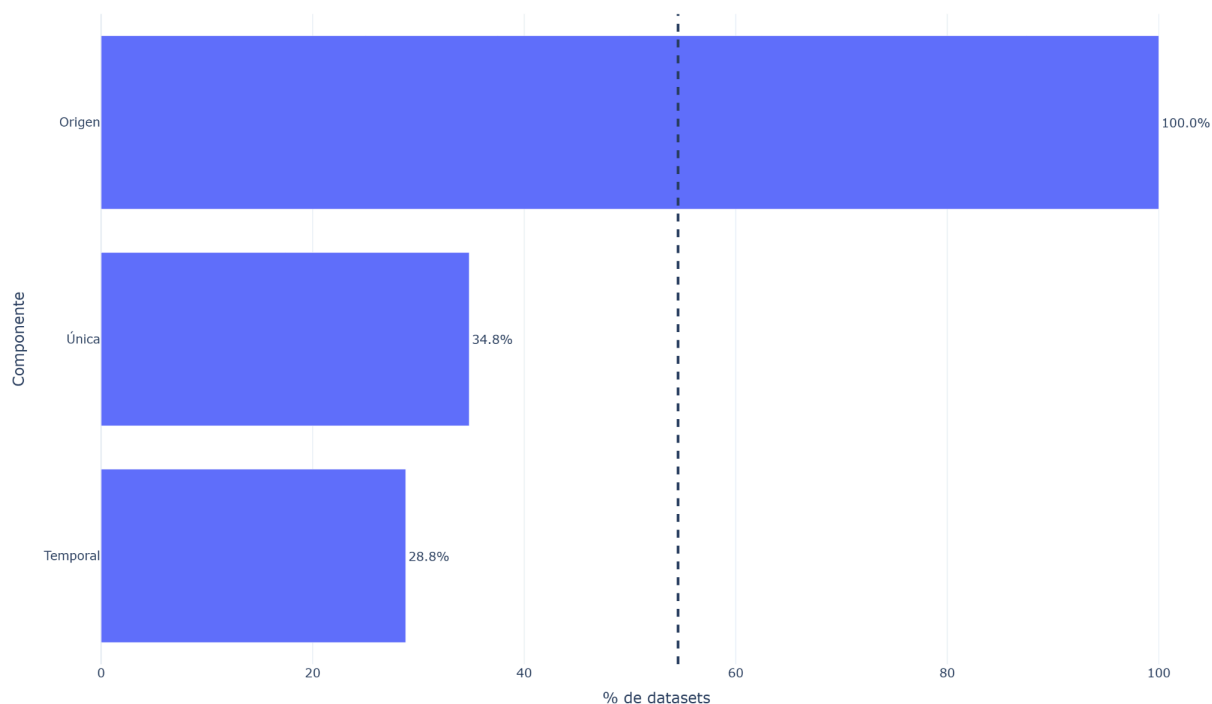


Figura 7.58. Resultado Gráfico Trazabilidad Global

La figura es el resultado global de tres componentes que lo forman porcentaje promedio de cumplimiento.

Trazabilidad Única: Lo conforman el DOI (Código único y permanente) y la url de acceso al dataset.

Trazabilidad Origen: Publicador, identificador (del dataset)

Trazabilidad Temporal: Frecuencia de actualización, Antigüedad del dataset desde su última actualización (valores medidos con actualización, por horas, diaria, semanal, mensual, hasta 6 meses) lo que hace un dataset trazable temporalmente.

Interoperabilidad Semántica (Score=54.6%) | Junta de Andalucía

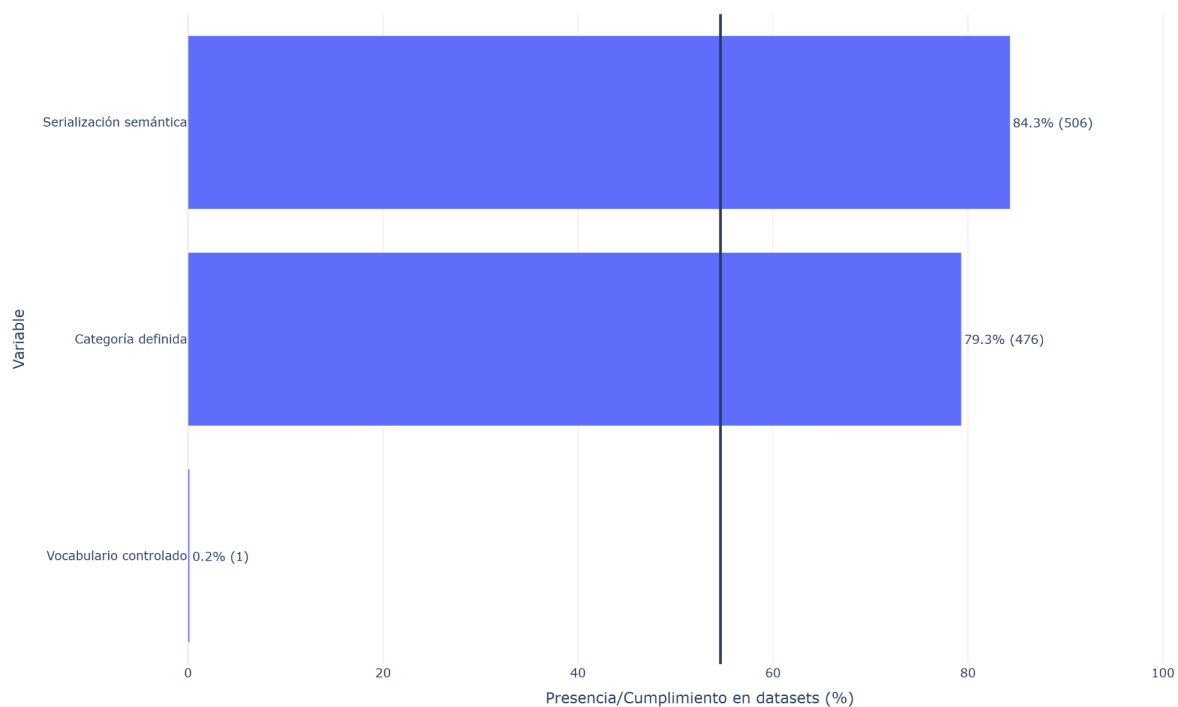


Figura 7.59. Resultado Gráfico Métrica de Interoperabilidad Semántica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 54.6%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

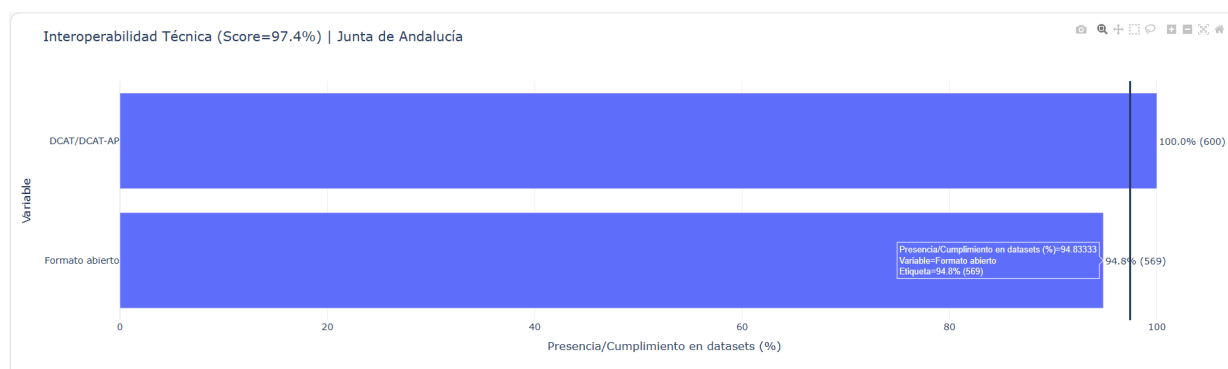


Figura 7.60. Resultado Gráfico Métrica de Interoperabilidad Técnica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 97.4%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

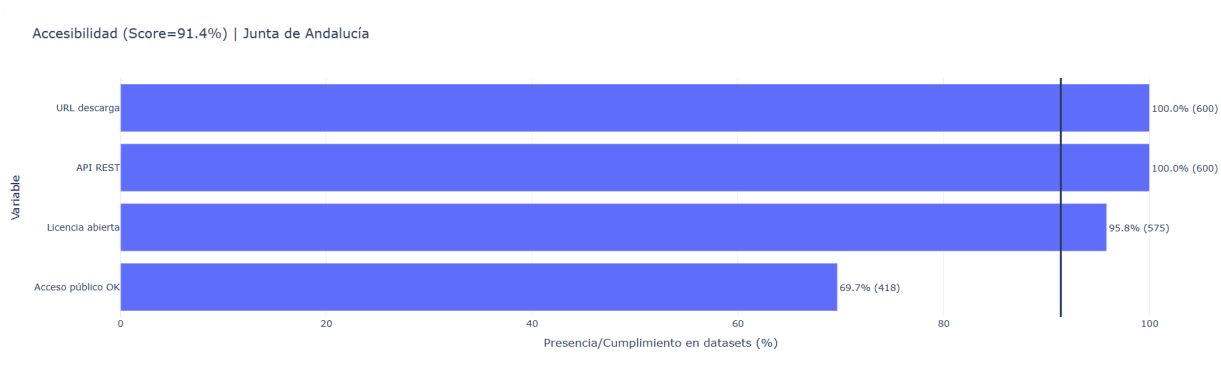


Figura 7.61. Resultado Gráfico Métrica de Accesibilidad

La figura muestra el promedio del total de las variables que lo forman (total de datasets 91.4%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

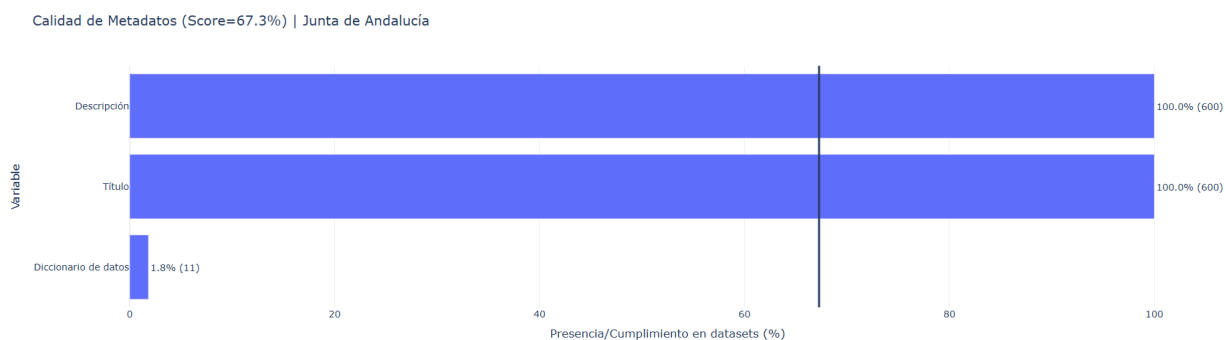


Figura 7.62. Resultado Gráfico Métrica de Calidad de metadatos.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 67.3%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

7.3.8. Portal Nacional (datos.gob.es)

El análisis de Datos abiertos del Portal nacional, se basa en una muestra de 450 conjuntos de datos, extraídos a través de su Tecnología APIDATA + SPARQL/DCAT-AP. Esta muestra abarca varias categorías temáticas, entre el top 3 destacan Datos demográficos y estadísticos, Empleo, Turismo. Se observa una alta densidad de metadatos actualizados, lo que permite evaluar su madurez técnica bajo un marco de consistencia temporal elevado.

Índice Global de Madurez del Portal — datos.gob.es (Portal Nacional) (N=450)

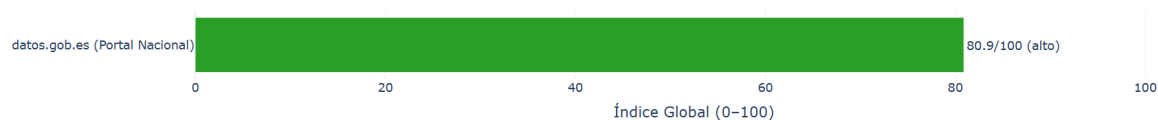


Figura 7.63. Resultado Gráfico Índice de Madurez del Portal.

La figura muestra el Índice global calculado como promedio **ponderado** de 5 componentes: Trazabilidad, Interoperabilidad semántica, Interoperabilidad técnica, Accesibilidad y Calidad de metadatos. Nivel definido Bajo 0–30, Medio 30–70, Alto 70–100.

Componentes que forman el Índice de Madurez — datos.gob.es (Portal Nacional)

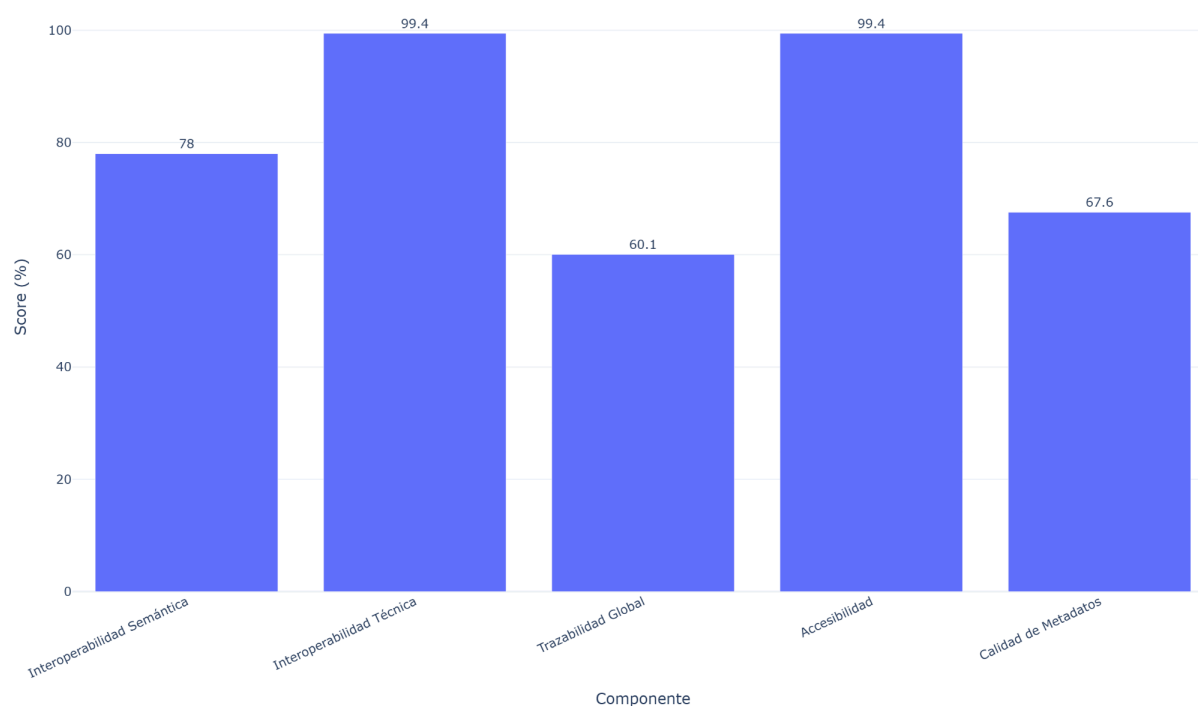


Figura 7.64. Resultado Gráfico de componentes de Índice de Madurez del Portal.

La figura muestra Cada score de dimensión se calcula como promedio del cumplimiento de los componentes que forman cada métrica (peso igual por componente). Los 5 componentes construyen el Índice de Madurez del Portal.

Interoperabilidad Técnica — Clasificación de formatos (datasets con presencia)

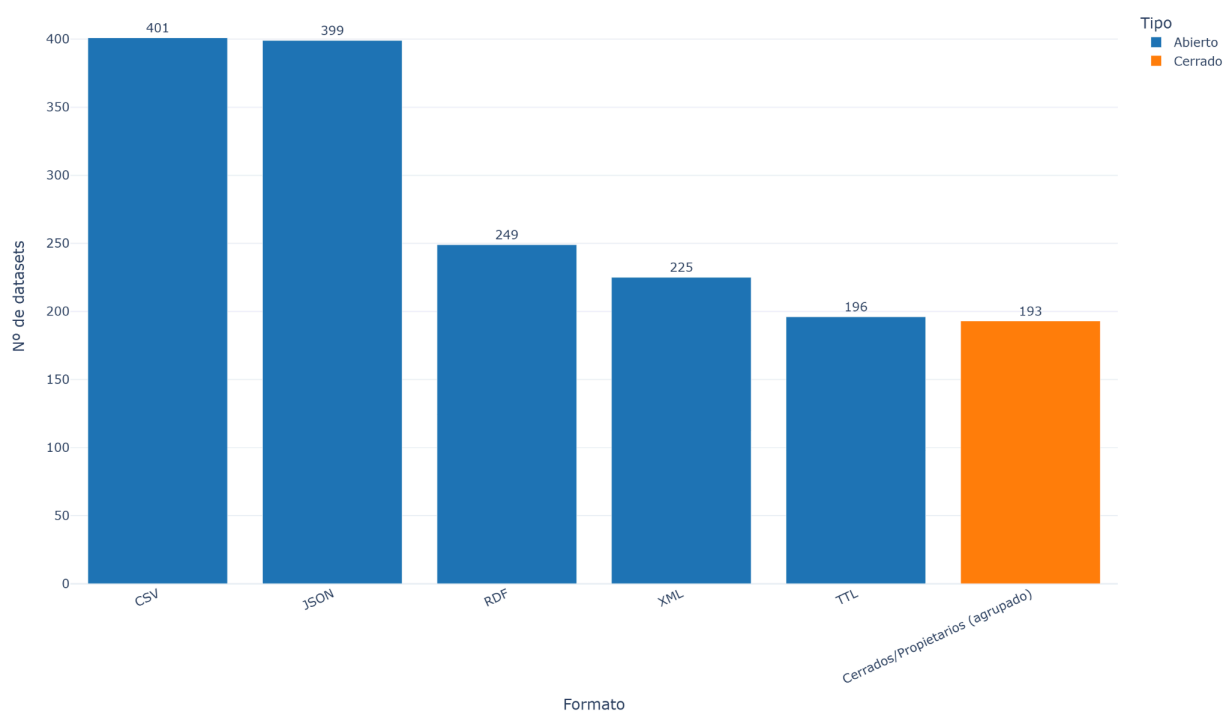


Figura 7.65. Resultado Clasificación de formatos para Interoperabilidad técnica

La figura muestra conteo de datasets con presencia de formatos abiertos (por tipo) y datasets con al menos un formato cerrado/propietario (agrupado).

Accesibilidad (Dimensión legal) — Licenciamiento permitido

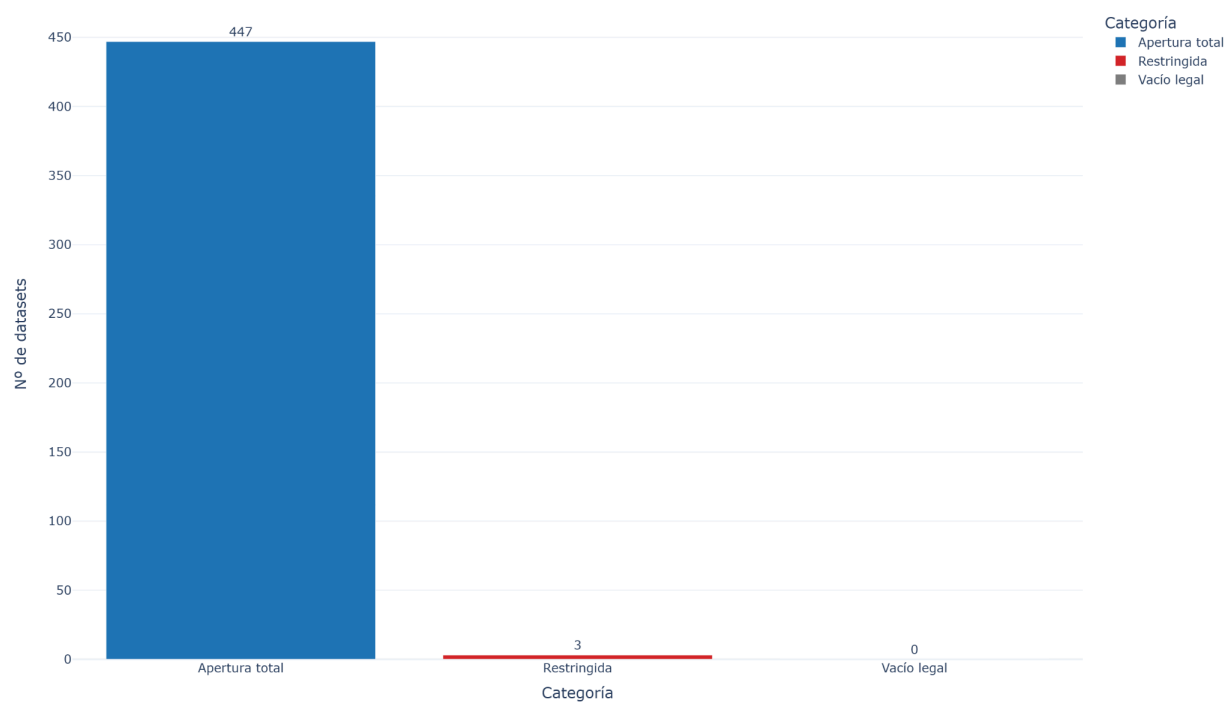


Figura 7.66. Resultado Licenciamiento permitido (Dimensión Legal)

La figura muestra las licencias encontradas dividida por 3 grupos: Apertura total (Todas las licencias abiertas Aviso Legal, Creative Commons. Restringida: NC/ND u otras restricciones. Vacío legal: licencia ausente/no definida/consultar.

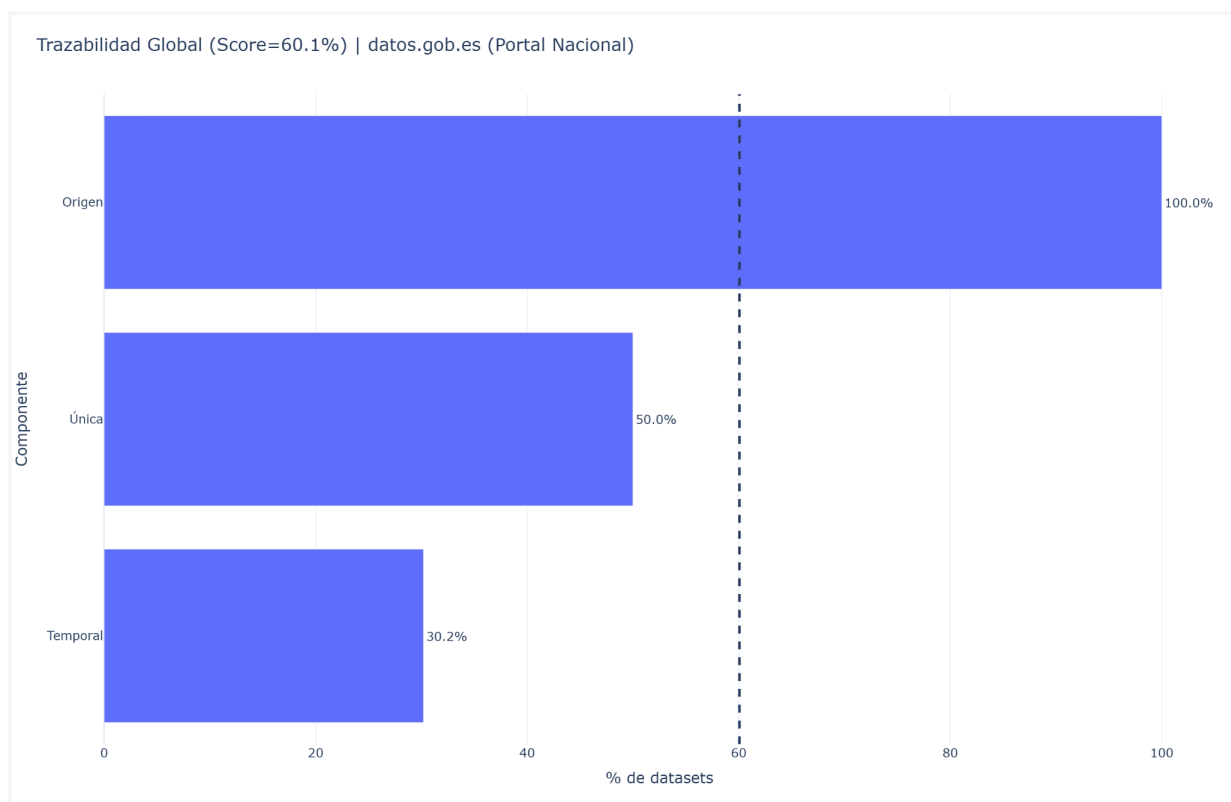


Figura 7.67. Resultado Gráfico Trazabilidad Global

La figura es el resultado global de tres componentes que lo forman porcentaje promedio de cumplimiento.

Trazabilidad Única: Lo conforman el DOI (Código único y permanente) y la url de acceso al dataset.

Trazabilidad Origen: Publicador, identificador (del dataset)

Trazabilidad Temporal: Frecuencia de actualización, Antigüedad del dataset desde su última actualización (valores medidos con actualización, por horas, diaria, semanal, mensual, hasta 6 meses) lo que hace un dataset trazable temporalmente.

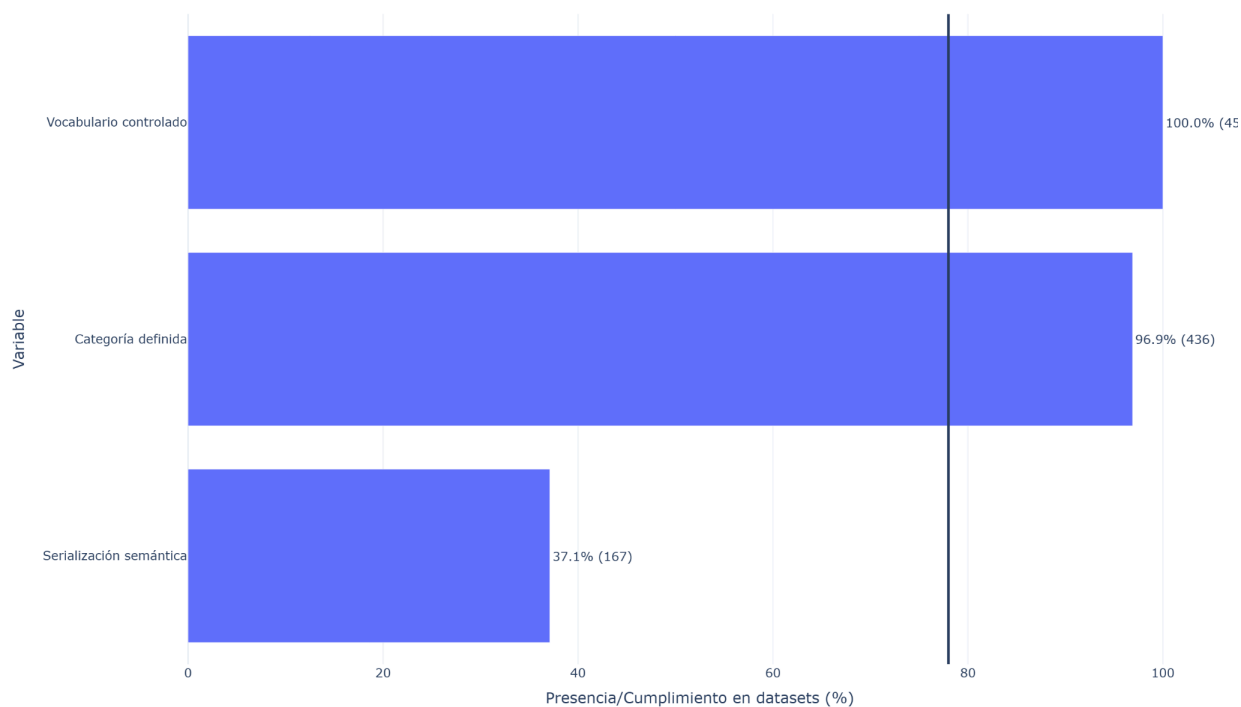


Figura 7.68. Resultado Gráfico Métrica de Interoperabilidad Semántica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 78%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

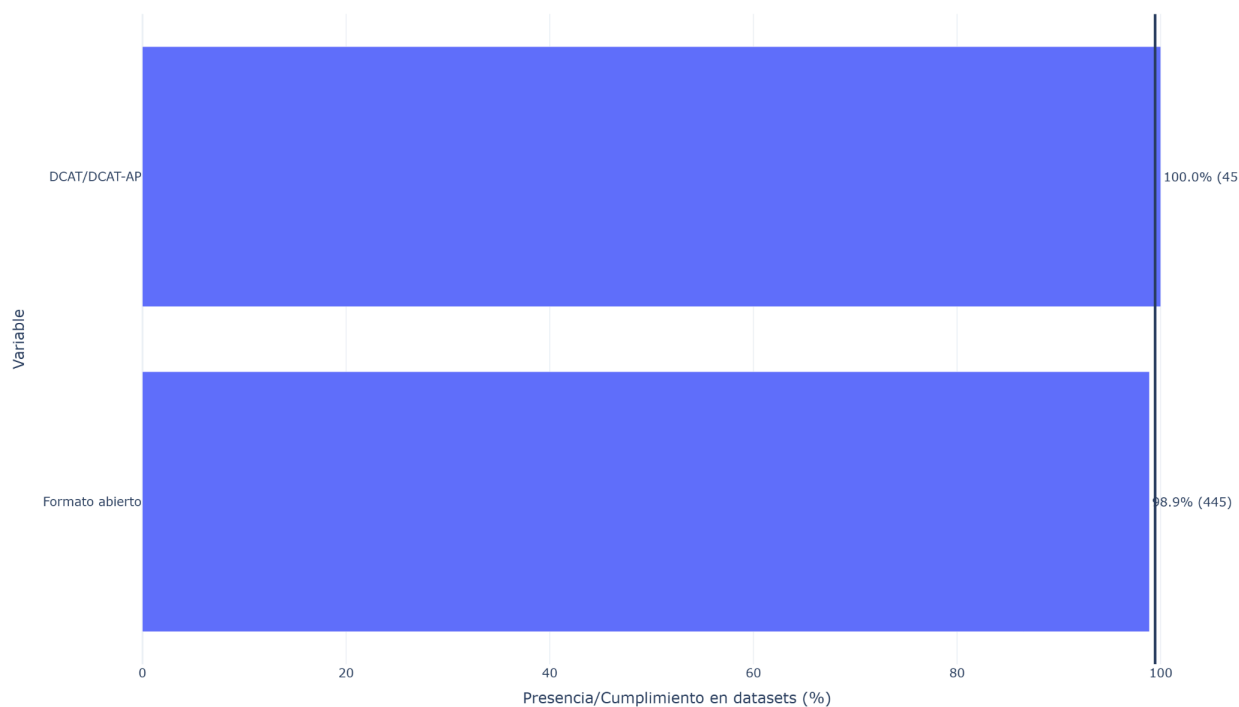


Figura 7.69. Resultado Gráfico Métrica de Interoperabilidad Técnica.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 99.4%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

Accesibilidad (Score=99.4%) | datos.gob.es (Portal Nacional)

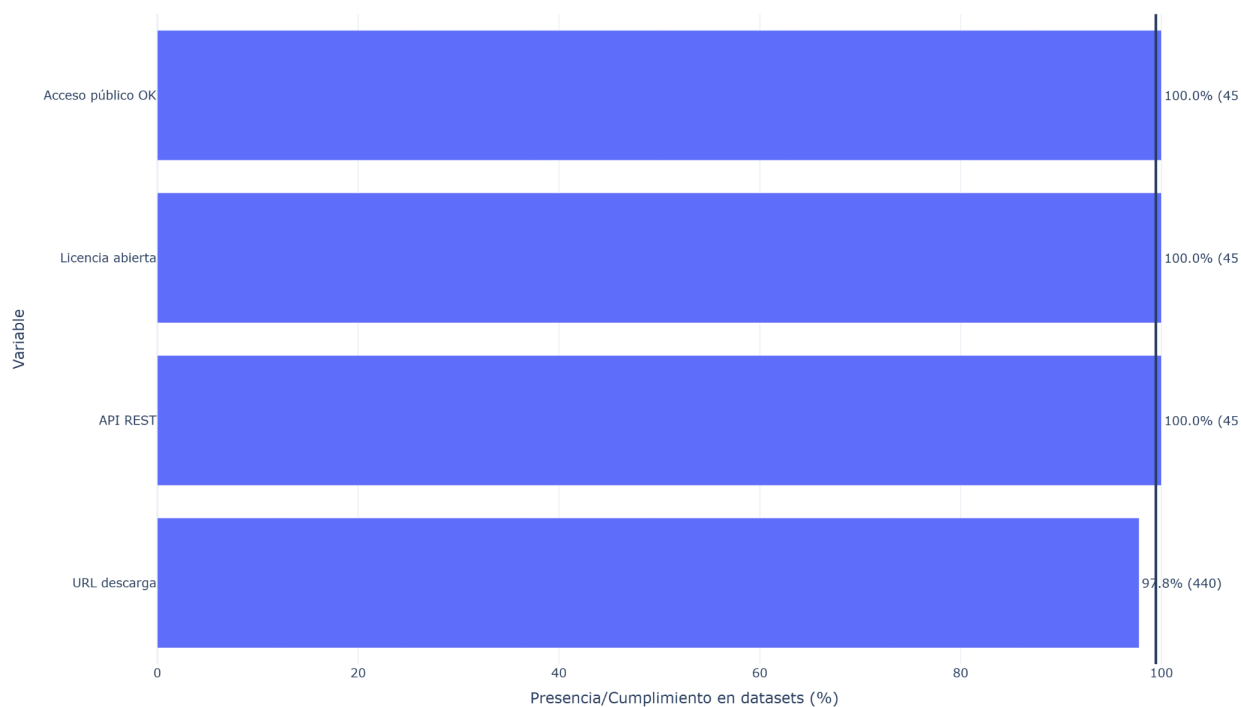


Figura 7.70. Resultado Gráfico Métrica de Accesibilidad

La figura muestra el promedio del total de las variables que lo forman (total de datasets 99.4%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

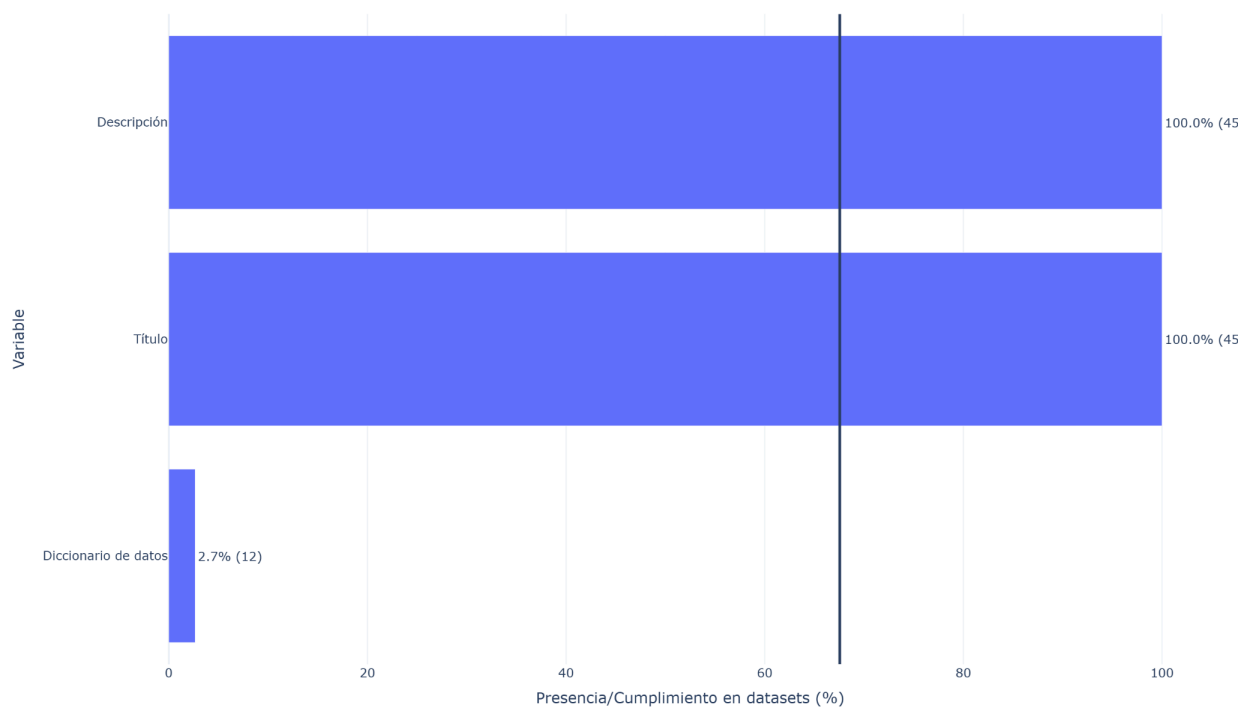


Figura 7.71. Resultado Gráfico Métrica de Calidad de metadatos.

La figura muestra el promedio del total de las variables que lo forman (total de datasets 67.6%), estas tienen un valor de presencialidad en los resultados de la tabla, siendo estas: 1 es un dato presente y contable, mientras que 0 es un valor ausente.

7.4. Análisis de la distribución de las métricas por portal

Con el objetivo de analizar no solo los valores agregados, sino también la variabilidad interna de las métricas a nivel de dataset, se estudió la distribución de cada dimensión del índice de madurez mediante diagramas de caja (boxplots), agrupados por portal.

Este enfoque permite identificar la dispersión, la presencia de valores atípicos y las diferencias en la consistencia interna de cada portal, complementando el análisis basado en valores medios.

7.4.1. Distribución del Índice Global de Madurez por portal

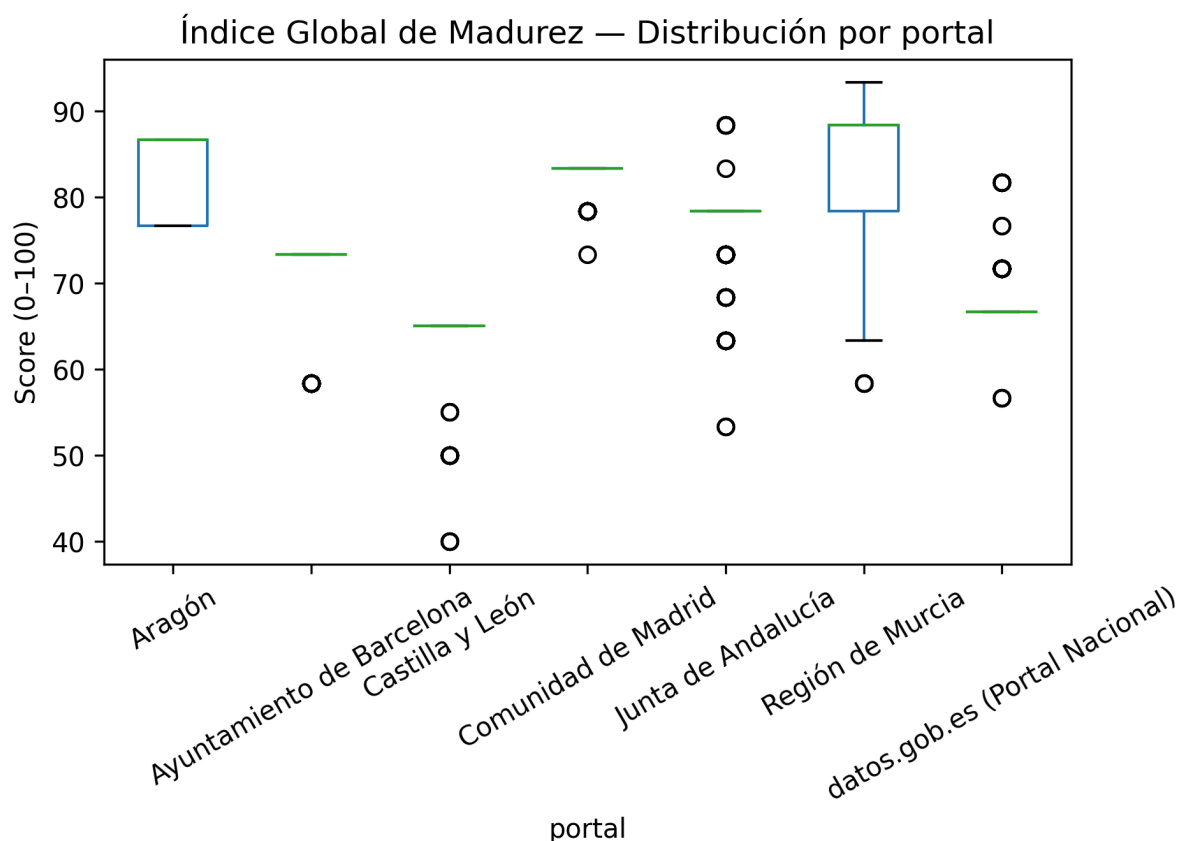


Figura 7.72. Resultado Gráfico Distribución por portal.

El diagrama de caja del Índice Global de Madurez muestra diferencias claras en la distribución de los scores entre portales. Mientras algunos portales presentan distribuciones más compactas, indicando un comportamiento homogéneo entre sus datasets, otros exhiben una mayor dispersión y la presencia de valores atípicos, lo que evidencia heterogeneidad interna en la calidad y madurez de los metadatos publicados.

7.4.2. Distribución de las dimensiones del índice por portal

a) Accesibilidad

En la dimensión de Accesibilidad, los boxplots evidencian distribuciones altamente concentradas en valores elevados para la mayoría de los portales, con una variabilidad limitada. No obstante, se observan valores atípicos en algunos casos, lo que indica la existencia de datasets que no

cumplen de forma consistente con los criterios de acceso definidos, pese a un alto desempeño medio del portal.

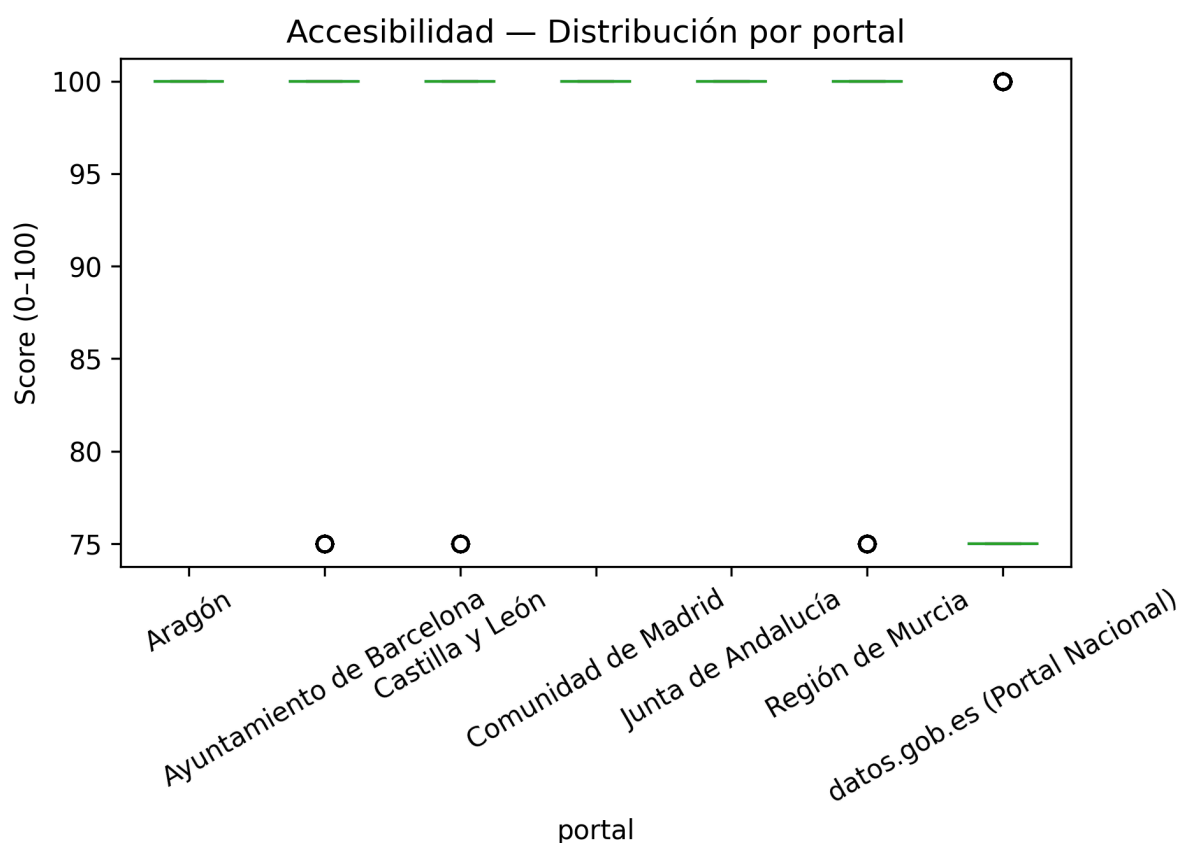


Figura 7.73. Resultado Gráfico Distribución Accesibilidad.

b) Interoperabilidad semántica

La Interoperabilidad Semántica presenta distribuciones más heterogéneas entre portales. Aunque varios portales alcanzan valores máximos en la mayoría de sus datasets, la presencia de cajas más amplias y valores atípicos inferiores revela diferencias en la aplicación consistente de vocabularios controlados, categorización temática y serialización semántica a nivel de dataset.

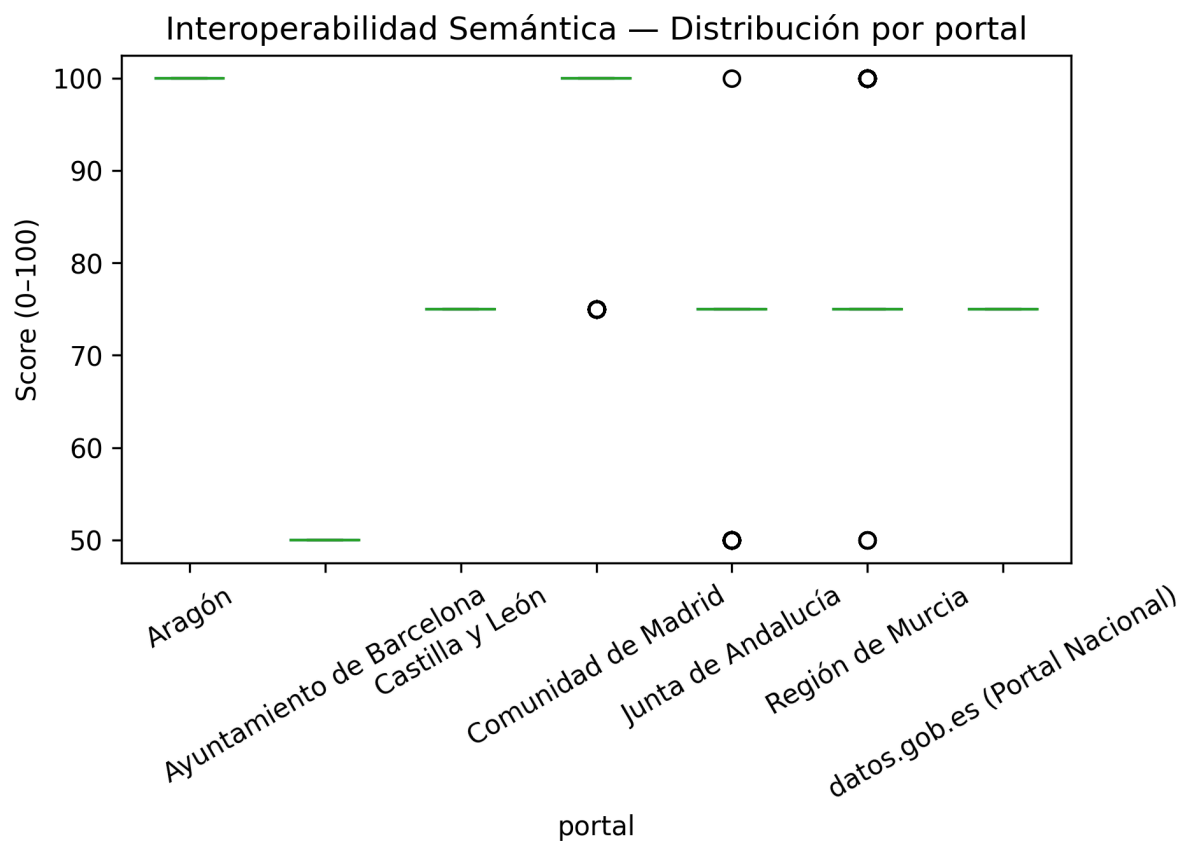


Figura 7.74. Resultado Gráfico Distribución Interoperabilidad semántica

c) Trazabilidad global

La distribución de la Trazabilidad Global muestra una elevada dispersión en algunos portales, con medianas moderadas y presencia de valores bajos. Este comportamiento refleja la coexistencia de datasets con información de origen, temporalidad y referencias persistentes bien definidas junto con otros que carecen de alguno de estos elementos, confirmando la trazabilidad como una dimensión crítica y desigual dentro del ecosistema analizado.

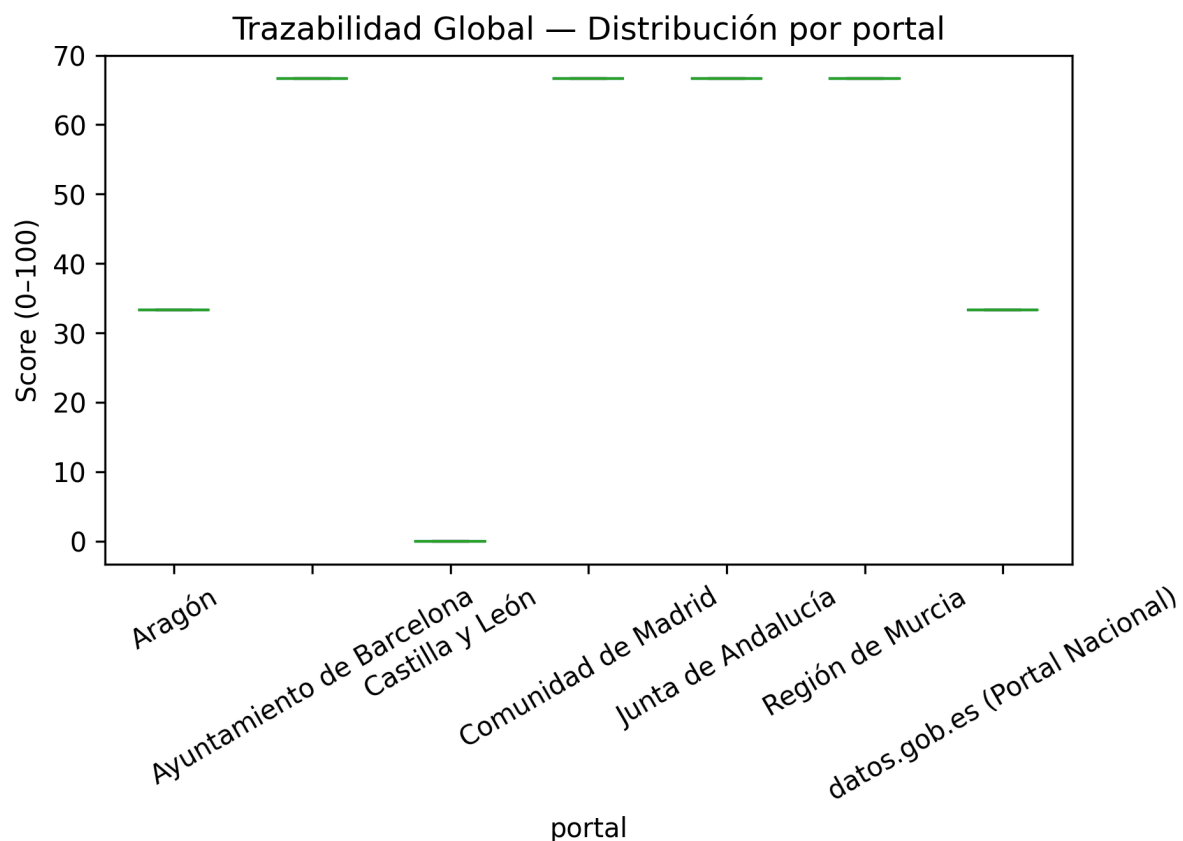


Figura 7.75. Resultado Gráfico Distribución Trazabilidad global

7.5. Análisis estadístico de las métricas de madurez

Las diferencias observadas en las distribuciones de las métricas sugieren que el comportamiento de los portales de datos abiertos no es homogéneo. Con el objetivo de evaluar si estas diferencias son estadísticamente significativas, se aplicaron contrastes inferenciales entre niveles de madurez del portal (bajo, medio y alto), considerando cada dataset como unidad de análisis.

En este apartado se presentan los resultados del análisis estadístico realizado sobre las métricas de calidad y madurez de los portales analizados. Se aplicaron pruebas paramétricas (ANOVA) y no paramétricas (Kruskal-Wallis), con el fin de garantizar la robustez de los resultados ante posibles desviaciones de los supuestos de normalidad y homocedasticidad.

Métrica	ANOVA p	Kruskal p	Media_bajo	Media_medio	Media_alto
Interoperabilidad Técnica	< 0.001	< 0.001	—	49.06	98.49
Índice Global	< 0.001	< 0.001	—	55.14	77.48
Accesibilidad	< 0.001	< 0.001	—	76.65	97.30
Interoperabilidad Semántica	< 0.001	< 0.001	—	58.96	79.30
Calidad Metadatos	< 0.001	< 0.001	—	46.70	61.97
Trazabilidad Global	0.01	0.26	—	44.34	50.36

Tabla 7.1. Resultados de ANOVA y Kruskal–Wallis por métrica y nivel de madurez

Interpretación de los contrastes globales.

La Tabla 7.1 recoge los resultados de los contrastes estadísticos aplicados a las métricas de madurez, utilizando como factor explicativo el nivel de madurez del portal. Los resultados muestran que, para la mayoría de las métricas analizadas, existen diferencias estadísticamente significativas entre niveles de madurez, con valores **inferiores a** ($p < 0.05$) tanto en las pruebas paramétricas como no paramétricas. Esto indica que el desempeño en dimensiones clave como la interoperabilidad técnica, la interoperabilidad semántica, la accesibilidad, la calidad de los metadatos y el índice global de madurez varía de forma sistemática en función del nivel de madurez del portal.

La interoperabilidad técnica presenta la diferencia más marcada entre niveles de madurez, con un valor medio de 49,06 en portales clasificados con madurez media frente a 98,49 en portales de madurez alta. Esta diferencia resulta altamente significativa, lo que evidencia que los portales más maduros disponen de una infraestructura técnica sustancialmente más sólida, especialmente en términos de uso de formatos abiertos, mecanismos de acceso normalizados y adopción de estándares técnicos.

En contraste, la métrica de trazabilidad global muestra un comportamiento diferenciado. Aunque el contraste paramétrico (ANOVA) sugiere la existencia de diferencias significativas entre niveles de madurez, el test no paramétrico de Kruskal–Wallis no confirma dicha relación. Esta falta de consistencia estadística indica que la trazabilidad no mejora de forma homogénea con el aumento del nivel de

madurez del portal, lo que refuerza su identificación como una de las principales debilidades estructurales del ecosistema de datos abiertos analizado.

Finalmente, la ausencia de valores en la columna **Media_bajo** para algunas métricas se debe a la escasa representación de datasets clasificados en el nivel de madurez bajo. Esta circunstancia refleja una característica empírica del ecosistema analizado y no una limitación del procedimiento estadístico aplicado.

Grupo 1	Grupo 2	Dif. medias	p-ajustada	IC inferior	IC superior	Rechaza H ₀	Métrica
Alto	Medio	-22.34	< 0.001	-23.86	-20.82	Sí	Índice Global
Alto	Medio	-6.01	0.01	-10.42	-1.61	Sí	Trazabilidad Global
Alto	Medio	-49.44	< 0.001	-51.11	-47.76	Sí	Interoperabilidad Técnica
Alto	Medio	-20.33	< 0.001	-23.57	-17.09	Sí	Interoperabilidad Semántica
Alto	Medio	-20.65	< 0.001	-22.14	-19.16	Sí	Accesibilidad
Alto	Medio	-15.27	< 0.001	-19.36	-11.18	Sí	Calidad de Metadatos

Tabla 7.2. Resultados de los contrastes post hoc de Tukey HSD

Interpretación de los contrastes post hoc.

La Tabla 7.2 presenta los resultados de los contrastes post hoc de Tukey HSD aplicados a aquellas métricas para las que el contraste ANOVA resultó estadísticamente significativo. Estos análisis permiten identificar entre qué niveles de madurez se concentran las diferencias observadas.

Los resultados indican que, en todas las métricas analizadas, las diferencias estadísticamente significativas se producen principalmente entre los niveles de madurez medio y alto, con valores de **p** ajustada inferiores a **0.05** en todos los casos.

En particular, la interoperabilidad técnica muestra la mayor diferencia media entre ambos niveles, seguida por la accesibilidad, la interoperabilidad semántica y la calidad de los metadatos. Estos resultados confirman que el incremento del nivel de madurez del portal se asocia con mejoras sustanciales y consistentes en las capacidades técnicas, semánticas y estructurales de los portales de datos abiertos.

7.6. Interpretación de los resultados

En conjunto, los resultados evidencian un ecosistema de datos abiertos técnicamente maduro y ampliamente accesible, pero con desafíos pendientes en la gestión y publicación de metadatos, especialmente en lo relativo a la trazabilidad temporal y la calidad informacional. Estos hallazgos subrayan la necesidad de orientar los esfuerzos futuros no solo a la infraestructura, sino también a la estandarización y gobernanza de los metadatos como elemento clave para una reutilización efectiva de los datos abiertos.

Capítulo 8 – Discusión

8.1. Interpretación de los resultados a la luz del estado del arte

Los resultados obtenidos en este trabajo confirman y amplían hallazgos previamente documentados en los proyectos de investigación y trabajos que analizan portales de datos abiertos. Estudios previos han analizado la madurez de estos ecosistemas principalmente desde enfoques normativos o estructurales, centrados en la adopción de estándares, la disponibilidad de metadatos o la calidad declarativa de los portales. En este contexto, modelos de referencia como los marcos de madurez de datos abiertos o los indicadores propuestos por organismos internacionales tienden a evaluar la capacidad potencial de reutilización, más que su trazabilidad operativa.

En contraste, el presente trabajo adopta un enfoque empírico y operativo de la madurez del ecosistema de datos abiertos, evaluándose a partir de cinco dimensiones clave: trazabilidad, accesibilidad, interoperabilidad semántica, interoperabilidad técnica y calidad de metadatos. Esta aproximación permite observar no solo si los portales cumplen formalmente con buenas prácticas, sino en qué medida dichas prácticas se traducen en condiciones reales de automatización, integración y uso sistemático de los datos como parte de infraestructuras digitales.

Los resultados muestran que, si bien la mayoría de los portales analizados presentan niveles aceptables de accesibilidad especialmente en lo relativo al uso de APIs estandarizadas y formatos abiertos, existen diferencias significativas en las dimensiones de interoperabilidad semántica y trazabilidad. Este hallazgo es coherente con trabajos previos que señalan la interoperabilidad como uno de los principales cuellos de botella en los ecosistemas de datos abiertos, particularmente cuando los metadatos no están alineados semánticamente o carecen de identificadores persistentes.

Desde una perspectiva operativa, estas limitaciones adquieren especial relevancia cuando los datos abiertos se integran en infraestructuras digitales complejas, como las utilizadas en servicios de Smart Cities y despliegues IoT, donde los datos públicos son consumidos de forma automática por plataformas de monitorización, sistemas de análisis en tiempo real y aplicaciones distribuidas. En estos contextos, la falta de trazabilidad efectiva en cuanto a mantener los portales con actualizaciones frecuentes, enlaces de descarga no resolubles, URLs inestables o

mecanismos de acceso no compatibles con la automatización, dificulta la escalabilidad de los sistemas y obliga a introducir validaciones manuales o soluciones específicas que reducen la eficiencia y la sostenibilidad de los despliegues.

De forma complementaria, las deficiencias observadas en la interoperabilidad semántica limitan la capacidad de integrar datasets procedentes de distintos portales en arquitecturas IoT heterogéneas. La ausencia de vocabularios controlados, diccionarios de datos y descripciones normalizadas incrementa los costes de integración y dificulta la reutilización sistemática de los datos en plataformas urbanas inteligentes, donde la alineación semántica es un requisito fundamental para la interoperabilidad entre sistemas.

En este sentido, los resultados del estudio sugieren que la madurez del ecosistema de datos abiertos en España no debe evaluarse únicamente a partir del volumen de datos publicados o del cumplimiento formal de estándares, sino atendiendo a la evolución del dato, desde su publicación hasta su consumo efectivo por infraestructuras digitales. El estado actual del ecosistema se caracteriza, por tanto, por un nivel elevado de apertura formal, pero con debilidades estructurales que condicionan su uso a gran escala en servicios de Smart Cities y despliegues IoT, especialmente cuando se requiere automatización, escalabilidad e integración continua.

8.2. La trazabilidad como eje crítico de la madurez del ecosistema

Uno de los principales aportes de este trabajo es la descomposición operativa del concepto de trazabilidad en tres componentes medibles: trazabilidad temporal, trazabilidad de origen y trazabilidad única. Esta desagregación ha permitido identificar con precisión cómo la trazabilidad condiciona la madurez real del ecosistema de datos abiertos, más allá de los valores agregados del Índice Global de Madurez.

Tal como se observa en la descomposición del índice global por dimensiones, los portales analizados presentan puntuaciones elevadas en accesibilidad e interoperabilidad técnica, mientras que la trazabilidad obtiene sistemáticamente los valores más bajos. Este desequilibrio estructural explica por qué un índice global aparentemente alto no se traduce en una madurez efectiva del ecosistema, sino en

una madurez funcional parcial, limitada a la fase de publicación pero insuficiente para sostener el ciclo de vida completo del dato.

La trazabilidad temporal, medida a través de la antigüedad desde la última actualización y la frecuencia de actualización, revela patrones relevantes sobre la sostenibilidad de los datasets. Los resultados indican que una proporción significativa de los datasets presenta largos periodos sin actualización o carecen de información clara sobre su periodicidad, lo que limita su valor para usos continuados, aun cuando sigan siendo técnicamente accesibles. Este hallazgo coincide con los trabajos de investigación, que destacan la actualización como un factor crítico para la reutilización efectiva, pero aporta una medición sistemática y comparable entre portales, ausente en muchos estudios anteriores.

La trazabilidad de origen basada en la identificación del publicador y de la existencia de referencias institucionales, muestra un comportamiento homogéneo entre portales, lo que sugiere una madurez institucional básica en la publicación de datos abiertos en España. No obstante, la ausencia de identificadores persistentes normalizados que permitan referenciar los datasets de forma estable a lo largo del tiempo.

Finalmente, la trazabilidad única, operacionalizada mediante la existencia de URLs de acceso a los datos que componen el dataset y los identificadores persistentes como DOI, presenta un resultado crítico: su valor es prácticamente nulo en todos los portales analizados. Esta ausencia no constituye un fallo metodológico, sino un resultado sustantivo del estudio. La trazabilidad única refleja una carencia estructural del ecosistema español de datos abiertos, alineada con lo señalado en trabajos recientes que reclaman mecanismos formales de citación de datasets. Al incorporar explícitamente esta dimensión aunque su resultado sea cero, el trabajo evidencia una carencia que permanece invisibilizada en evaluaciones más tradicionales, limitando severamente su integración en infraestructuras digitales avanzadas.

En este sentido, los resultados muestran que la madurez del ecosistema no puede inferirse únicamente a partir de altos niveles de accesibilidad o interoperabilidad técnica, sino que depende de la existencia de mecanismos de trazabilidad que garanticen la persistencia, identificabilidad y sostenibilidad de los datos a lo largo del tiempo.

En conjunto, el análisis confirma que la trazabilidad actúa como un eje crítico y limitante de la madurez del ecosistema de datos abiertos en España. Mientras no se aborde esta dimensión de forma sistemática, los portales seguirán mostrando un nivel elevado de apertura formal, pero una capacidad limitada para sostener usos avanzados, automatizados y escalables de los datos abiertos.

8.3. Comparación con estudios previos y aportes diferenciales

A diferencia de investigaciones previas centradas en evaluaciones estáticas o cualitativas de portales, este trabajo introduce un pipeline automatizado y reproducible, capaz de operar a escala sobre cientos de datasets y múltiples portales heterogéneos. Mientras que gran parte del estado del arte se apoya en enfoques estructurales y manuales, análisis de cumplimiento normativo, centrados en el volumen de datasets publicados, la adopción de estándares de metadatos o la existencia de licencias abiertas (Curto-Rodríguez, 2021; Criado et al., 2023; European Data Portal, 2025). Esta metodología propuesta permite medir empíricamente la madurez del ecosistema a partir de datos observables y comparables.

Cuando se comparan los valores elevados obtenidos en dimensiones como accesibilidad e interoperabilidad técnica, estos son coherentes con los diagnósticos recogidos en los informes europeos de madurez, que sitúan a España en posiciones avanzadas en términos de apertura formal y disponibilidad de infraestructuras de publicación. Los resultados de este trabajo confirman, por tanto, que los portales españoles han alcanzado un grado notable de madurez en lo relativo a la publicación de datos y al acceso técnico a los recursos.

Asimismo, la integración simultánea de dimensiones técnicas (formatos, APIs, licencias), semánticas (estructura de metadatos, diccionarios de datos), temporales (antigüedad y actualización) y organizativas (publicador, origen) constituye una contribución metodológica relevante. Esta nueva visión integrada supera enfoques parciales que analizan cada dimensión de forma aislada y permite construir una evaluación más global de la madurez del ecosistema.

Aunque, al evaluar el nivel operativo del dataset y analizar de forma sistemática dimensiones como la trazabilidad y la interoperabilidad semántica, los resultados revelan limitaciones que no quedan plenamente reflejadas en las evaluaciones agregadas. Esta observación coincide con lo señalado por diversos autores que advierten que la madurez de los ecosistemas de datos abiertos no puede evaluarse únicamente desde indicadores macro o declarativos, sino que

debe considerar la calidad efectiva de los metadatos y su capacidad para soportar procesos de integración y uso real (Oviedo Blanco, 2016; Batini y Scannapieco, 2016).

En particular, los problemas detectados en la estandarización semántica y en la ausencia de diccionarios de datos y vocabularios controlados coinciden y refuerzan los planteamientos de Zaveri et al. (2016) y Maseda-Seco et al. (2017), quienes identifican la interoperabilidad semántica como uno de los principales retos para la consolidación de ecosistemas de datos abiertos maduros. Los resultados de este trabajo muestran que, aunque los portales declaran el uso de estándares como DCAT-AP, su implementación es heterogénea y, en muchos casos, insuficiente para garantizar una integración automática entre portales.

En el análisis de la trazabilidad temporal y operativa de los datasets se aporta evidencia empírica que complementa los enfoques conceptuales revisados en el estado del arte. Mientras que las investigaciones existentes reconocen la importancia de la actualización y la calidad de los metadatos como factores de madurez (European Commission, 2019; Wilkinson et al., 2016), los resultados de este estudio muestran que una proporción significativa de los datasets que presenta frecuencias de actualización bajas, irregulares o no declaradas, así como mecanismos de acceso que dificultan su consumo automatizado. Este hallazgo permite pasar de una afirmación teórica sobre la importancia de la actualización a una medición sistemática y comparable entre portales.

Desde la perspectiva metodológica, el principal aporte diferencial de este proyecto resalta frente a los estudios revisados, y reside en la operacionalización empírica y automatizable de las dimensiones de madurez. Mientras que buena parte del estado de arte se apoya en evaluaciones manuales, encuestas institucionales o indicadores internos de los portales, el pipeline desarrollado permite medir de forma reproducible la madurez a nivel de dataset, capturando el comportamiento real de los portales durante procesos de extracción, validación y análisis automatizados. Facilitando la replicabilidad del estudio, permitiendo monitorear la evolución temporal. De este modo, el trabajo se posiciona no sólo como un análisis puntual, sino como una propuesta metodológica reutilizable para evaluaciones longitudinales y comparativas a nivel nacional o internacional.

La comparación con los estudios previos sugiere que la contribución de este Trabajo Fin de Máster no radica en contradecir los diagnósticos existentes, sino en complementarlos y profundizarlos. Al trasladar las dimensiones teóricas ampliamente aceptadas a métricas observables y comparables, el trabajo permite evidenciar que

la madurez del ecosistema de datos abiertos en España es elevada en términos formales, pero presenta debilidades estructurales en aspectos clave como la trazabilidad y la interoperabilidad semántica, que condicionan su uso efectivo en infraestructuras digitales avanzadas.

8.4. Implicaciones para la evaluación de la madurez de los datos abiertos

Los resultados de este trabajo sugieren que la madurez del ecosistema de datos abiertos en España no puede evaluarse únicamente a partir de la existencia de portales o del cumplimiento formal de estándares de publicación. Las debilidades detectadas en dimensiones clave como la trazabilidad operativa, la interoperabilidad semántica y la gestión temporal de los metadatos indican que la madurez alcanzada es funcional, pero incompleta, especialmente cuando se consideran escenarios de uso avanzados que requieren automatización, integración y escalabilidad.

Desde la perspectiva de las administraciones públicas proveedoras de datos, los hallazgos ponen de manifiesto la necesidad de complementar las estrategias actuales de apertura con políticas orientadas a la calidad operativa de los datos. En particular, la estabilidad de los enlaces de descarga, la coherencia de los metadatos temporales, la adopción de vocabularios controlados y la disponibilidad de diccionarios de datos emergen como factores críticos para consolidar los portales como verdaderas infraestructuras de datos abiertos. Estas implicaciones sugieren que la evaluación de la madurez debería incorporar indicadores que midan no sólo la publicación, sino también la capacidad del portal para sostener procesos automatizados y un consumo continuo de los datos a lo largo del tiempo.

Para la comunidad investigadora, este trabajo aporta una metodología empírica y reproducible que permite evaluar la madurez de los portales a nivel de dataset, superando enfoques basados en métricas agregadas o declarativas. Las métricas propuestas y el índice global de madurez facilitan comparaciones sistemáticas entre portales y abren la posibilidad de realizar estudios longitudinales que analicen la evolución del ecosistema. Asimismo, los resultados refuerzan la necesidad de que los modelos de evaluación de la madurez incorporen dimensiones operativas observables, alineadas con prácticas de ciencia de datos, interoperabilidad y gestión del ciclo de vida del dato.

Desde el punto de vista de la industria, los desarrolladores y los actores tecnológicos, los resultados evidencian que la madurez de los datos abiertos es un factor determinante para su integración efectiva en aplicaciones digitales, plataformas IoT y sistemas basados en analítica avanzada. Las limitaciones identificadas en interoperabilidad y trazabilidad incrementan los costes de integración y reducen la escalabilidad de las soluciones, lo que sugiere que los indicadores de madurez deberían reflejar explícitamente la capacidad de los portales para alimentar sistemas automáticos y arquitecturas distribuidas sin requerir adaptaciones manuales.

Finalmente, para la sociedad civil y los reutilizadores no especializados, la evaluación de la madurez tiene implicaciones directas en términos de confianza, transparencia y accesibilidad real. La consistencia de los metadatos, la claridad en la descripción de los datasets y la fiabilidad de los mecanismos de acceso influyen en la capacidad de los ciudadanos, organizaciones sociales y periodistas para utilizar los datos abiertos como herramientas de control, participación y generación de valor público. En este sentido, una evaluación más exigente de la madurez contribuye a fortalecer el papel de los datos abiertos como pilar de la gobernanza democrática.

En este sentido, el trabajo aporta evidencia empírica que refuerza la necesidad de evolucionar desde modelos centrados en la publicación hacia modelos orientados a la reutilización verificable y sostenible, donde la trazabilidad, la citabilidad y la interoperabilidad juegan un papel central.

8.5. Sesgos y limitaciones encontradas

Este proyecto presenta una serie de limitaciones y posibles sesgos que deben ser considerados para una correcta interpretación de los resultados, derivados tanto de las decisiones metodológicas adoptadas como de las características propias del ecosistema de datos abiertos analizado. No obstante, estas limitaciones no invalidan los hallazgos obtenidos, sino que delimitan su alcance y abren oportunidades claras para futuras investigaciones.

En primer lugar, este análisis se basa en tomar en cuenta la información que aparece públicamente en los metadatos y el comportamiento que muestran los portales mientras se ejecuta el pipeline automatizado. Por ello, no se incluyen aspectos internos de las administraciones públicas que podrían influir en cómo se

gestionan, mantienen o actualizan los datasets, como sus procesos organizativos, limitaciones técnicas internas o decisiones editoriales que no quedan reflejadas en los catálogos. Aunque esto deja fuera ciertos factores que podrían ayudar a entender mejor el contexto, el enfoque es coherente con el objetivo del estudio: evaluar la madurez de los portales tal y como la perciben los usuarios finales y los sistemas automáticos que consumen los datos.

En segundo lugar, la evaluación de la accesibilidad y la trazabilidad operativa se ve condicionada por la variabilidad técnica de los portales y por la existencia de mecanismos de acceso intermedios, redirecciones o controles que dificultan la descarga directa de los recursos. Aunque estas situaciones forman parte del análisis, su presencia puede introducir un sesgo al penalizar portales que, aun disponiendo de los datos, no están plenamente optimizados para el consumo automatizado. No obstante, este sesgo es consistente con el enfoque adoptado, que prioriza la automatización y la escalabilidad como criterios de madurez.

Otra limitación encontrada y relevante está relacionada con el uso de estándares entre portales. Si bien la mayoría de los portales analizados declara el uso de DCAT o DCAT-AP, las diferencias en su implementación dificultan la comparación completamente homogénea de algunos indicadores, especialmente en lo relativo a la interoperabilidad semántica y a la disponibilidad de diccionarios de datos o vocabularios controlados. Esta situación refleja una limitación estructural del ecosistema más que una debilidad del método, pero condiciona la precisión de ciertas métricas.

Por otro lado, el análisis de la frecuencia de actualización y de la antigüedad temporal de los datasets se apoya en los campos de fechas disponibles en los metadatos. En los casos en que esta información es incompleta, inconsistente o inexistente, ha sido necesario inferir o clasificar los datasets en categorías amplias, lo que puede reducir el nivel de detalle del análisis temporal. Esta limitación está directamente vinculada a las deficiencias de calidad de los metadatos identificadas en las conclusiones del trabajo.

Cabe destacar que desde el punto de vista del alcance del estudio, la selección de portales y categorías temáticas, aunque amplia y representativa del contexto español, no agota la totalidad del ecosistema de datos abiertos. Por tanto, los resultados no deben interpretarse como una evaluación exhaustiva de todos los portales existentes, sino como una caracterización empírica basada en una muestra significativa y metodológicamente controlada.

En este proyecto se realizó a términos de análisis exploratorio, ejecutar el pipeline y obtener la extracción de huellas de reutilización de datasets, por tanto como aporte metodológico resaltar las dificultades detectadas que proviene de la ausencia generalizada de identificadores persistentes como el DOI en los datasets publicados por los portales españoles. La carencia de este tipo de identificador complica enormemente la trazabilidad, ya que obliga a basar las búsquedas en elementos textuales como el título del dataset, la URL del portal o fragmentos descriptivos. Debido a ello, los motores de búsqueda académicos y científicos devuelven resultados que coinciden parcialmente, pero que no guardan relación con el dataset original. Esta situación genera un número significativo de falsos positivos, especialmente en plataformas como Google Scholar o OpenAlex, donde la coincidencia literal del título no implica necesariamente una reutilización real de los datos.

Otra limitación relacionada, deriva del diseño heterogéneo y poco estandarizado de los metadatos que acompañan a los datasets en los portales españoles. En muchas ocasiones, la información disponible se reduce a descripciones generales sin identificadores únicos, sin información de versión y sin un campo específico para su referencia formal. Esta falta de estandarización impide construir consultas precisas y reduce la eficacia de las búsquedas automatizadas, generando un sesgo hacia datasets con nombres muy específicos o menos ambiguos, mientras que aquellos cuyo título coincide con términos generales producen coincidencias irrelevantes.

Finalmente, la construcción del Índice Global de Madurez implica decisiones de agregación y ponderación que, aunque transparentes y justificadas, introducen un grado inevitable de simplificación. El uso de pesos iguales para todas las dimensiones responde a una decisión metodológica orientada a la comparabilidad, pero otros esquemas de ponderación podrían producir variaciones en el ranking final. Esta limitación no afecta a la validez de los hallazgos por dimensión, pero debe tenerse en cuenta al interpretar el índice sintético.

En conjunto, estas limitaciones ponen de manifiesto que la evaluación de la madurez de los datos abiertos es un proceso complejo, condicionado por factores técnicos, organizativos y contextuales. No obstante, lejos de restar valor al estudio, refuerzan la pertinencia del enfoque empírico adoptado y subrayan la necesidad de continuar avanzando hacia evaluaciones más profundas, estandarizadas y orientadas al uso efectivo de los datos abiertos en infraestructuras digitales avanzadas.

8.6. Interpretación comparativa resultados descriptivos + boxplots + tests

Los resultados obtenidos muestran que el ecosistema de datos abiertos analizado presenta un alto nivel de madurez técnica y accesibilidad, evidenciado por valores elevados y distribuciones concentradas en dimensiones como la interoperabilidad técnica, la interoperabilidad semántica y la accesibilidad. El análisis descriptivo por portal refleja un desempeño sólido en estas dimensiones, mientras que los diagramas de caja muestran una variabilidad limitada en la mayoría de los casos, lo que sugiere una aplicación relativamente homogénea de estándares técnicos y mecanismos de acceso a nivel de dataset.

El análisis de distribución mediante boxplots pone de manifiesto, no obstante, diferencias internas relevantes entre portales, especialmente en el Índice Global de Madurez y en las dimensiones relacionadas con la calidad de los metadatos y la trazabilidad. La presencia de valores atípicos y una mayor dispersión en estas métricas indica que, incluso en portales clasificados con niveles altos de madurez, coexisten datasets con grados desiguales de completitud y estructuración de los metadatos.

Estos patrones visuales son coherentes con los resultados del análisis estadístico inferencial. Las pruebas ANOVA y Kruskal–Wallis confirman la existencia de diferencias estadísticamente significativas entre niveles de madurez en la mayoría de las dimensiones analizadas, particularmente en interoperabilidad técnica, interoperabilidad semántica, accesibilidad y calidad de los metadatos. La magnitud de estas diferencias, reflejada en el incremento de las medias entre los niveles medio y alto, refuerza la asociación entre madurez institucional y capacidades técnicas efectivas de los portales.

En contraste, la trazabilidad global presenta un comportamiento diferenciado. Si bien el análisis paramétrico sugiere diferencias significativas entre niveles de madurez, el contraste no paramétrico no confirma esta tendencia de forma consistente. Esta falta de robustez estadística, junto con la elevada dispersión observada en los boxplots, indica que la trazabilidad en particular en sus componentes temporal y de identificación persistente constituye una de las principales debilidades estructurales del ecosistema, incluso en portales con altos niveles de madurez global.

En síntesis, los resultados evidencian que la madurez del ecosistema de datos abiertos no depende únicamente de la disponibilidad de infraestructuras técnicas o

mecanismos de acceso, sino de la calidad, coherencia y gobernanza de los metadatos. Los hallazgos subrayan la necesidad de orientar los esfuerzos futuros no solo a la mejora de las capacidades tecnológicas, sino también a la estandarización, actualización y fortalecimiento de la trazabilidad de los metadatos como condición esencial para una reutilización efectiva y sostenible de los datos abiertos.

Capítulo 9 – Conclusiones y líneas futuras

El presente Trabajo de Fin de Máster ha abordado el análisis de la madurez del ecosistema de datos abiertos en España mediante una aproximación empírica y reproducible, basada en el análisis sistemático de los metadatos publicados por distintos portales institucionales, con especial atención a las condiciones técnicas y estructurales que permiten la automatización, la integración y el uso sistemático de los datos en infraestructuras de datos abiertos. Dichas infraestructuras constituyen un componente fundamental de los servicios digitales avanzados, especialmente en el ámbito de las Smart Cities, donde los datos abiertos se integran de forma habitual con plataformas IoT y sistemas de analítica para la monitorización, la planificación y la toma de decisiones, y donde la calidad, estabilidad e interoperabilidad de los datos resulta determinante para su incorporación efectiva. Frente a enfoques centrados exclusivamente en la existencia de datos abiertos o en el cumplimiento formal de estándares, este trabajo propone una evaluación de la madurez orientada a las condiciones reales que permiten el uso correcto y efectivo de los datos. El análisis se ha centrado en el comportamiento real de los portales durante la ejecución de un pipeline automatizado de extracción y validación de metadatos, lo que ha permitido identificar limitaciones que no son evidentes desde una inspección puramente declarativa.

Los resultados empíricos obtenidos ponen de manifiesto varios hallazgos cuantitativos relevantes. En primer lugar, se observa que, aunque más del 70–80 % de los datasets analizados ofrecen formatos técnicamente accesibles, y en su variedad abiertos como csv, json, xml entre otros, la evidencia demuestra que la presencia de formatos cerrados o de difícil reutilización (ZIP, XLSX, ODS u otros contenedores) es comparable en volumen con los formatos abiertos. Este equilibrio entre formatos abiertos y no abiertos reduce la madurez técnica del ecosistema, ya que obliga a procesos adicionales de transformación y dificulta la interoperabilidad entre portales y sistemas. Es una muestra desalineada con respecto a lo que define el Esquema Nacional de Interoperabilidad (Gobierno de España, 2010) y a la normativa de reutilización (NTI-RISP). Este desequilibrio evidencia una brecha significativa entre accesibilidad formal y trazabilidad real.

Otro de los hallazgos encontrados en el estudio, es la existencia de problemas recurrentes de accesibilidad operativa, a pesar de que los portales declaran cumplir estándares de datos abiertos. Durante la fase de preparación de datos y validación manual del pipeline, se constató que una proporción significativa de URLs de

descarga no era funcional o no permitía una descarga directa del recurso, debido a enlaces rotos, redirecciones no gestionadas o mecanismos de acceso intermedio que impiden la automatización. Este comportamiento obliga a introducir validaciones manuales y reduce la fiabilidad de los portales como infraestructuras de datos orientadas a procesos automáticos.

En cuanto a la trazabilidad temporal de los datasets, los resultados muestran una baja madurez en la definición y mantenimiento de políticas de actualización. Aproximadamente un 40 % de los datasets analizados presenta frecuencias de actualización relativamente recientes, con intervalos que oscilan entre horas, días y hasta seis meses. Sin embargo, el resto de los datasets se caracteriza por actualizaciones anuales, irregulares, nunca o directamente inexistentes, o bien por la ausencia de información clara sobre su periodicidad en el Portal. Esta situación limita el uso de los datos en escenarios que requieren información actualizada, como la analítica continua o los sistemas de monitorización.

Asimismo, se observa un uso muy limitado de formatos semánticos y tecnologías orientadas a la interoperabilidad avanzada, como RDF o vocabularios enlazados, lo que restringe la capacidad de integración automática de los datasets. Aunque la mayoría de los portales analizados declaran alinearse con el estándar DCAT para la descripción de metadatos, este cumplimiento es en muchos casos superficial. Solo en torno al 30 % de los datasets se identificó la presencia de elementos asociados a una verdadera interoperabilidad semántica, como diccionarios de datos, vocabularios controlados o descripciones estructuradas de los campos. En última instancia, esta carencia de metadatos semánticos actúa como una barrera estructural para la madurez del ecosistema, impidiendo que los datos trasciendan de simples archivos descargables a activos de conocimiento plenamente integrados

En conjunto, los resultados evidencian que, si bien los portales analizados presentan avances significativos en accesibilidad técnica y disponibilidad de formatos abiertos, se concluye que la madurez en los ecosistemas españoles se encuentra condicionada no tanto por la adopción formal de estándares, sino por su implementación efectiva en términos de accesibilidad operativa, coherencia temporal, normalización semántica y soporte a la automatización. Estas limitaciones afectan directamente a la trazabilidad de los dataset y el potencial de los portales como infraestructuras de datos reutilizables en contextos tecnológicos avanzados.

En conjunto, este trabajo aporta una caracterización rigurosa del estado actual del ecosistema de datos abiertos en España, identificando fortalezas, debilidades y áreas de mejora. La metodología desarrollada sienta las bases para evaluaciones comparativas y longitudinales, contribuyendo a una comprensión más profunda de la madurez de los portales y de los factores que condicionan su efectividad.

Líneas futuras

A partir de los resultados obtenidos, se identifican diversas líneas futuras de actuación orientadas a mejorar la madurez del ecosistema de datos abiertos. En primer lugar, resulta prioritario reforzar la estabilidad y fiabilidad de los mecanismos de acceso a los datasets, garantizando que las URLs de descarga sean directas, persistentes y plenamente compatibles con procesos automatizados. La incorporación de validaciones automáticas periódicas sobre enlaces y recursos publicados permitiría reducir la aparición de enlaces rotos y mejorar la calidad operativa de los portales.

En segundo lugar, se recomienda avanzar hacia una definición más clara y homogénea de las políticas de actualización, incorporando metadatos temporales coherentes y verificables que permitan identificar la vigencia real de los datasets. Este aspecto es especialmente relevante para su integración en infraestructuras digitales, sistemas IoT y plataformas de análisis que dependen de datos actualizados de forma continua.

Otra línea futura fundamental consiste en reforzar el uso de formatos verdaderamente abiertos y semánticamente enriquecidos, reduciendo la dependencia de formatos cerrados o contenedores poco interoperables. En este sentido, la adopción más sistemática de formatos semánticos y el uso efectivo de vocabularios controlados facilitan la integración automática de los datos y su reutilización en entornos multiportal.

Asimismo, aunque la mayoría de los portales analizados declaran el uso del estándar DCAT, se plantea la necesidad de evolucionar desde un cumplimiento formal hacia una implementación profunda del estándar, que incluye diccionarios de datos accesibles, descripciones normalizadas de los campos y referencias explícitas a

vocabularios compartidos. Este avance permitiría mejorar significativamente la interoperabilidad semántica del ecosistema.

Desde el punto de vista metodológico, futuras investigaciones podrían ampliar el pipeline propuesto incorporando nuevas fuentes de información y técnicas de análisis. En particular, la integración de motores académicos adicionales, plataformas de proyectos europeos y análisis más avanzados de lenguaje natural permitiría detectar menciones implícitas de datasets que actualmente no dejan huellas formales explícitas. Asimismo, la combinación de enfoques automáticos con validaciones manuales o estudios de caso cualitativos podría contribuir a mejorar la precisión y cobertura del análisis de reutilización.

Como aporte metodológico y valor añadido, se propone la incorporación de mecanismos de citación y seguimiento del uso de los datasets directamente en los portales de datos abiertos, incluir procesos estandarizados de identificación y citación de datasets, tales como la asignación de identificadores persistentes, identifier y DOI (Identificador de Objeto Digital) los cuales identifican de manera persistente una publicación electrónica (como un artículo científico, libro o datos), asegurando que siempre se pueda encontrar y acceder al mismo, cuando un reutilizador hace uso de los datos, así como la inclusión explícita de instrucciones de citación visibles en las páginas de cada conjunto de datos. De modo que cuando un dataset sea utilizado en proyectos, investigaciones o desarrollos tecnológicos, dicha información pueda reflejarse en el propio portal. Tal como señalan trabajos previos, la mera disponibilidad técnica de estos mecanismos no garantiza su uso efectivo; por ello, resulta necesario que los portales promuevan activamente buenas prácticas de citación entre los reutilizadores, promover la adopción de identificadores normalizados en los portales de datos abiertos, siguiendo las prácticas de los repositorios de datos de investigación, se presenta como un requisito previo para avanzar hacia métricas empíricas robustas.

La generación automática de identificadores citables y su vinculación con referencias externas contribuiría a mejorar la trazabilidad y la visibilidad del impacto de los datos abiertos, reforzando así la madurez global del ecosistema.

También se sugiere que las administraciones públicas y organismos responsables de los portales consideren la incorporación de métricas internas de reutilización, tales como descargas, visualizaciones, referencias externas o proyectos derivados, que complementen las huellas de reutilización externas analizadas en este

trabajo. La combinación de métricas internas y externas permitiría obtener una visión más completa y realista del impacto y uso efectivo de los datasets publicados.

Este proyecto abre la puerta a líneas de investigación orientadas a evaluar el impacto de políticas públicas y directrices de citación de datos abiertos sobre la reutilización efectiva de los datasets. El análisis longitudinal de portales que adopten identificadores persistentes, normas de citación obligatorias o incentivos a la atribución permitiría estudiar si dichas medidas se traducen en una mayor trazabilidad y visibilidad del uso de los datos abiertos.

Contribuciones de este Trabajo de Fin de Máster

Contribuciones metodológicas

- Diseño de un pipeline automatizado y reproducible para la evaluación de la madurez de portales de datos abiertos, aplicable a múltiples tecnologías y contextos institucionales.
- Propuesta de un modelo de madurez operacionalizado en cinco dimensiones medibles a partir de metadatos reales, superando enfoques puramente normativos o declarativos.
- Descomposición explícita del concepto de trazabilidad en componentes temporal, de origen y reutilizable, permitiendo su evaluación empírica.

Contribuciones técnicas

- Implementación de scripts específicos adaptados a distintas tecnologías de portal (CKAN, OpenDataSoft, APIs REST, APIDATA + SPARQL), manteniendo criterios homogéneos de análisis.
- Desarrollo de un módulo independiente para la detección automatizada de huellas externas de reutilización (Fase exploratoria y metodológica), concebido como aporte metodológico reutilizable en contextos con mejores condiciones de trazabilidad.
- Normalización y comparación sistemática de metadatos heterogéneos procedentes de portales nacionales, autonómicos y locales.

Contribuciones empíricas

- Identificación de debilidades estructurales en la gestión de metadatos temporales, identificadores y estabilidad de acceso en portales relevantes del ecosistema español.
- Evidencia empírica de que la ausencia de identificadores persistentes constituye una barrera crítica para la reutilización avanzada de los datos abiertos.
- Generación de métricas comparables que permiten evaluar la madurez relativa de los portales analizados.

Chapter – Introduction

Introduction

In recent years, open data portals have experienced sustained growth as part of transparency, innovation, and open government strategies promoted by public administrations and international organizations. These portals constitute a fundamental access point for citizens, researchers, companies, and social organizations to access public data for scientific, technological, economic, or social purposes. In the Spanish context, this ecosystem has expanded significantly, with the existence of national, regional, and municipal portals that publish thousands of datasets across diverse domains such as mobility, health, education, the environment, and demography.

The current state of knowledge indicates that data openness represents only the first link in a broader value chain, in which the effective use of datasets depends directly on their levels of technical, semantic, and organizational maturity. In this regard, the central issue does not lie solely in the volume of information published, but rather in the conditions under which such information is published and in the degree of operational maturity exhibited by the data to ensure that they are accessible, interoperable, and traceable in a sustainable manner.

Analyzing the maturity of the open data ecosystem requires evaluating not only the availability of datasets, but also the structural conditions under which they are published and managed by different portals. In practice, many portals exhibit significant weaknesses in key aspects such as metadata quality and consistency, the existence of stable identifiers, semantic normalization, technical interoperability between platforms, and the temporal traceability of data. These shortcomings limit comparability across portals and hinder the construction of homogeneous indicators that allow for an objective assessment of ecosystem development. As a result, much of the existing research has focused on basic descriptive metrics related to technical availability, without systematically addressing a comprehensive evaluation of the components that determine the actual maturity of open data portals.

Within this context, the present Master's Thesis is based on the premise that the technical viability of data cannot be reliably assessed without first analyzing the maturity conditions of the ecosystem in which they are embedded. Therefore, the primary objective of this study is not to directly measure dataset usage, but rather to

evaluate the degree of preparedness of open data portals and their metadata to ensure that any interaction with the data can occur, be traced, and eventually be measured in a reproducible manner. Accordingly, the study focuses on analyzing the structural, technical, and semantic conditions that determine the quality, accessibility, traceability, and interoperability of data published on open data portals in Spain. These dimensions are understood as fundamental prerequisites of maturity and as key elements for enabling future empirical measurement strategies.

Based on this approach, a methodological framework structured into four phases is designed and applied. The first phase consists of the selection and automated extraction of datasets from various national, regional, and municipal open data portals, including the National Open Data Portal (datos.gob.es), the Community of Madrid, Open Data Barcelona, Aragón, the Region of Murcia, Castilla y León, and the Andalusian Regional Government. These portals provide a broad and heterogeneous collection of datasets, organized into comparable thematic categories such as transport, health, education, environment, science, and demography. During this phase, the metadata required to evaluate data maturity are collected. The second phase involves the manual validation of a representative sample to ensure the reliability of the extracted data. The third phase focuses on the construction of metrics aimed at assessing portal maturity, incorporating indicators related to metadata quality, accessibility, technical and semantic interoperability, and traceability through the temporal consistency of datasets. The fourth phase addresses the visualization and comparative analysis of results through graphical representations that facilitate comparisons across portals, thematic categories, and maturity dimensions.

The contribution of this work lies in addressing an existing gap in the empirical evaluation of open data portals by proposing a reproducible methodological approach focused on measuring ecosystem maturity. The results obtained enable the formulation of practical recommendations aimed at improving metadata quality and interoperability, thereby laying the foundations for the future development of more robust impact metrics. Ultimately, this study advances toward a more operational understanding of the open data ecosystem, highlighting that maturity expressed through metadata quality and traceability is a key element in ensuring the effective utility of public data.

1.1. Motivation

The opening of public data has become a fundamental instrument for promoting institutional transparency, strengthening citizen participation, and fostering evidence-based innovation. In recent years, open data portals have significantly increased both the volume and diversity of the information published, expanding access to governmental data across multiple domains and administrative levels. These data also constitute an essential component in the development of advanced digital infrastructures, such as Internet of Things (IoT) platforms and Smart City services, where open data are frequently integrated as external sources for decision-making, urban monitoring, and the provision of intelligent services.

However, this quantitative growth in open data availability has not always been accompanied by a corresponding improvement in the quality, traceability, and interoperability of the published data. In environments such as IoT systems and smart city platforms, where data must be consumed in an automated, continuous, and reliable manner, the absence of consistent and standardized metadata represents a critical limitation. Without high-quality metadata, automated data integration and interpretation processes are prone to failure, thereby compromising the proper functioning of analytical applications, urban services, and data-driven decision-making systems.

In practice, most open data portals restrict their evaluation to internal metrics such as downloads or visualizations, which provide only a partial and inconclusive view of the actual impact of datasets. These indicators do not allow for the establishment of reliable links between published data and their potential use in complex technological infrastructures, such as IoT applications, urban analytics platforms, scientific repositories, or institutional projects. As a result, it becomes difficult to objectively assess the public, scientific, or economic value of open data, creating a gap between the principles of openness and the empirical evidence of their effective reuse.

This issue is particularly relevant in the context of Spanish-language open data portals, where structural limitations are frequently observed, including the absence of persistent identifiers, the use of unstable URLs, heterogeneous data formats, and the lack of standardized metadata. These shortcomings hinder not only the reuse of data by third parties, but also their integration into open data infrastructures that support advanced digital services, such as IoT ecosystems and Smart City platforms.

The central motivation of this Master's Thesis lies in the need to conduct an in-depth diagnosis of the current state of the open data ecosystem in Spain. Rather than focusing on the external impact of published information, this research is oriented toward examining the robustness of its technical foundations. The aim is to determine whether the level of infrastructure maturity assessed through metadata quality, traceability, and interoperability, is sufficient to guarantee a reliable data environment capable of meeting the demands of digital transformation.

Accordingly, this project focuses on studying the technical and structural conditions that must be fulfilled for open data to be reliably integrated into automated systems and intelligent platforms.

However, the use of automated text analysis and mining techniques will enable an objective assessment of the current limitations of the open data ecosystem, documenting the technical obstacles that prevent effective traceability and the reliable integration of datasets into advanced digital infrastructures.

Ultimately, the motivation of this project is to contribute to the development of a more mature, sustainable, and results-oriented open data ecosystem, in which metadata quality, interoperability between portals, and dataset traceability are recognized as key elements not only for future data reuse, but also for the proper functioning of open data infrastructures used in contexts such as Smart Cities and IoT-based systems.

1.2. Objectives

1.2.1. General Objective

To design and evaluate a methodological framework for analyzing the maturity of the open data ecosystem in Spain through the systematic assessment of the quality, accessibility, traceability, and technical and semantic interoperability of the metadata associated with datasets published across different open data portals. The purpose is to diagnose the level of development of their data infrastructures, identify technical and structural limitations, and formulate recommendations aimed at improving interoperability, sustainability, and ecosystem readiness for integration into advanced digital infrastructures.

1.2.2. Specific Objectives

1. To select a representative set of open data portals in Spain, considering different administrative levels and technological contexts, and to automatically extract dataset metadata for comparative analysis.
2. To analyze the publication structure and metadata availability of the selected datasets, assessing the presence of key information such as identifiers, dates, formats, licenses, documentation, and access mechanisms.
3. To design and implement an automated analysis pipeline that systematically applies the defined metrics, with the objective of identifying the technical conditions required to assess portal maturity and generating comparable indicators across portals.
4. To experimentally explore the feasibility of detecting signals of external data reuse through open sources such as scientific repositories, software repositories, and data platforms, while evaluating the practical limitations of this approach in the context of Spanish open data portals.
5. To identify and document the technical, structural, and semantic limitations that hinder data accessibility and automated traceability, including the absence of persistent identifiers, URL inconsistencies, lack of standardized metadata, limited interoperability, and API-related constraints.
6. To define and propose a set of metrics for evaluating the maturity of open data portals, focused on metadata quality (completeness, consistency, and coherence), dataset accessibility, lifecycle traceability, and technical and semantic interoperability, in accordance with good practices and open data publication standards.
7. To formulate practical recommendations targeted at open data portal managers, aimed at improving metadata quality, interoperability between portals, and ecosystem readiness for integration into advanced digital systems, such as automated analytics platforms, IoT infrastructures, and Smart City services.

1.3. Methodology Overview and Work Plan

The methodology of this Master's Thesis is structured as an evidence-based study with a comparative analytical approach aimed at assessing the maturity of the open data ecosystem in Spain through the systematic evaluation of metadata published across different institutional open data portals. The objective is to characterize, in a reproducible and comparable manner, the degree of readiness of these portals for integration into open data infrastructures and advanced digital systems.

The methodological approach is based on the analysis of key dimensions of open data maturity, including metadata quality, accessibility, traceability, and technical and semantic interoperability. To this end, the study combines automated metadata extraction techniques, normalization and validation processes, and the construction of comparable metrics applicable across different portals and thematic categories.

The work plan is organized into four main stages, aligned with the methodological pipeline developed throughout the study:

- Metadata extraction, selection, and normalization, carried out across the selected open data portals using the available access mechanisms (APIs and public endpoints), and applying homogeneous dataset inclusion criteria.
- Manual validation of a representative sample, intended to complement automated processes and to verify the coherence and effective accessibility of the most relevant metadata elements, thereby ensuring the reliability of the analyzed dataset.
- Computation of maturity metrics, through the operationalization of indicators associated with accessibility, traceability, technical and semantic interoperability, and metadata quality, enabling the calculation of comparable scores at both dataset and portal levels.
- Analysis and presentation of results, using graphical and comparative representations that facilitate the interpretation of portal maturity levels, the identification of structural patterns, and the formulation of recommendations aimed at improving the open data ecosystem.

This methodological approach makes it possible to address the analysis of open data portal maturity from a structural and empirical perspective, ensuring consistency between the research objectives, the applied methodology, and the results obtained throughout the study.

1.4. Structure of the Document

This dissertation is organized into nine chapters that follow a logical progression from the conceptual framework to the empirical analysis and final conclusions, with a focus on evaluating the technical and structural maturity of the open data ecosystem in Spain.

- **Chapter 1. Introduction:** This chapter presents the general context of the study, the underlying motivation, the research objectives, and the overall methodological plan, with a focus on analyzing the maturity of open data infrastructures and their technical capacity to be integrated into advanced digital systems.
- **Chapter 2. Theoretical Foundations:** This chapter develops the fundamental theoretical concepts underpinning the research, including the principles of the open data movement, the notion of data infrastructure, metadata quality, accessibility, traceability, and technical and semantic interoperability. It also introduces the conceptual approaches used in the literature to assess the maturity of open data ecosystems, providing the necessary framework for defining the metrics and analytical dimensions employed in the study.
- **Chapter 3. State of the Art:** This chapter reviews and analyzes the state of the art through the examination of academic research, institutional reports, and technical proposals related to the evaluation of open data portals, maturity models, structural indicators, and methodologies used to assess data quality, interoperability, and traceability. This review allows for the identification of current limitations and contextualizes the novelty and scope of the present study.
- **Chapter 4. Methodology and Technologies:** This chapter provides a detailed description of the adopted methodology and the technologies used for its implementation. It presents the proposed pipeline, the tools employed for metadata extraction and processing, the APIs available in the analyzed portals, and the general criteria applied for dataset selection and evaluation. This chapter constitutes the methodological and technical foundation of the study.

- **Chapter 5. Research Design and Data Sources:** This chapter outlines the research design, detailing the selection of open data portals, the thematic categories considered, the dataset filtering and selection criteria, and the operational definition of the variables required for analysis. It also describes the methodological decisions adopted in response to the availability and characteristics of the data published by Spanish open data portals.
- **Chapter 6. Data Preparation and Implementation:** This chapter documents the data preparation process and the technical implementation of the pipeline. It includes tasks related to metadata extraction, cleaning, normalization, and validation, as well as the application of the procedures required to compute the defined maturity metrics. The chapter also describes the technical and structural challenges encountered during implementation, resulting from portal heterogeneity and the quality of the available metadata.
- **Chapter 7. Empirical Results:** The seventh chapter presents the empirical results obtained from the application of the maturity metrics. It analyzes indicators related to metadata quality, accessibility, traceability, and technical and semantic interoperability of the datasets, as well as the differences observed across portals. The results are supported by comparative visualizations that enable the characterization of the maturity level of the open data ecosystem in Spain.
- **Chapter 8. Discussion:** This chapter provides a discussion and interpretation of the results in relation to the state of the art. It examines the implications of the findings for open data portal managers and open data policy stakeholders, and it analyzes the methodological and structural limitations of the study, derived both from the adopted approach and from the characteristics of the Spanish open data ecosystem.
- **Chapter 9. Conclusions and Future Work:** The final chapter presents the conclusions of the study, synthesizing the main methodological and empirical contributions related to the evaluation of the maturity of the open data ecosystem in Spain. It summarizes the key findings, formulates general recommendations aimed at improving open data infrastructures, and proposes future research directions focused on metric expansion, improved interoperability, and the evolution toward more advanced ecosystem evaluation models.

Chapter – Conclusions and Future Work.

This Master's Thesis has addressed the analysis of the maturity of the open data ecosystem in Spain through an empirical and reproducible approach, based on the systematic examination of metadata published by different institutional portals, with particular attention to the technical and structural conditions that enable automation, integration, and systematic data use within open data infrastructures. Such infrastructures constitute a fundamental component of advanced digital services, especially in the context of Smart Cities, where open data is routinely integrated with IoT platforms and analytical systems for monitoring, planning, and decision-making. In these environments, data quality, stability, and interoperability are critical factors for effective integration. In contrast to approaches focused exclusively on the mere existence of open data or on formal compliance with standards, this study proposes a maturity assessment oriented toward the real conditions that enable the correct and effective use of data. The analysis focuses on the actual behavior of open data portals during the execution of an automated metadata extraction and validation pipeline, allowing the identification of limitations that are not evident through purely declarative inspection.

The empirical results reveal several relevant quantitative findings. First, although more than 70–80% of the analyzed datasets provide technically accessible formats, including a variety of open formats such as CSV, JSON, and XML, the evidence shows that the presence of closed or difficult-to-reuse formats (such as ZIP, XLSX, ODS, or other container formats) is comparable in volume to that of open formats. This balance between open and non-open formats reduces the technical maturity of the ecosystem, as it requires additional transformation processes and hinders interoperability between portals and systems. This situation is misaligned with the principles defined by the Spanish National Interoperability Framework (Esquema Nacional de Interoperabilidad; Government of Spain, 2010) and the regulations on data reuse (NTI-RISP). The observed imbalance highlights a significant gap between formal accessibility and actual traceability.

Another relevant finding concerns the existence of recurrent operational accessibility issues, despite portals declaring compliance with open data standards. During the data preparation and manual validation phases of the pipeline, it was observed that a substantial proportion of download URLs were not functional or did not allow direct access to the resource, due to broken links, unmanaged redirects, or intermediate access mechanisms that prevent automation. This behavior necessitates

the introduction of manual validation steps and reduces the reliability of portals as data infrastructures designed to support automated processes.

With respect to the temporal traceability of datasets, the results indicate a low level of maturity in the definition and maintenance of update policies. Approximately 40% of the analyzed datasets exhibit relatively recent update frequencies, ranging from hours and days to intervals of up to six months. However, the remaining datasets are characterized by annual, irregular, nonexistent, or undocumented updates, or by the absence of clear information regarding update periodicity within the portal. This situation limits the usability of the data in scenarios requiring up-to-date information, such as continuous analytics or real-time monitoring systems.

Furthermore, the study reveals a very limited use of semantic formats and technologies aimed at advanced interoperability, such as RDF or linked vocabularies, which constrains the automatic integration capabilities of datasets. Although most of the analyzed portals claim alignment with the DCAT standard for metadata description, this compliance is often superficial. Only around 30% of the datasets showed evidence of elements associated with genuine semantic interoperability, such as data dictionaries, controlled vocabularies, or structured field descriptions. Ultimately, this lack of semantic metadata constitutes a structural barrier to ecosystem maturity, preventing datasets from evolving beyond simple downloadable files into fully integrated knowledge assets.

Overall, the results indicate that, although the analyzed portals show significant progress in terms of technical accessibility and the availability of open formats, the maturity of the Spanish open data ecosystem is constrained less by the formal adoption of standards than by their effective implementation in terms of operational accessibility, temporal coherence, semantic normalization, and support for automation. These limitations directly affect dataset traceability and reduce the potential of open data portals to function as reusable data infrastructures in advanced technological contexts.

Taken together, this work provides a rigorous characterization of the current state of the open data ecosystem in Spain, identifying its strengths, weaknesses, and areas for improvement. The proposed methodology establishes a foundation for comparative and longitudinal evaluations, contributing to a deeper understanding of portal maturity and of the factors that condition their effectiveness as components of modern open data infrastructures.

Future Research

Based on the results obtained, several future lines of action are identified with the aim of improving the maturity of the open data ecosystem. First, it is essential to strengthen the stability and reliability of dataset access mechanisms by ensuring that download URLs are direct, persistent, and fully compatible with automated processes. The incorporation of periodic automated link and resource validation mechanisms would help reduce the occurrence of broken links and improve the operational quality of open data portals.

Second, progress is recommended toward a clearer and more homogeneous definition of dataset update policies, incorporating coherent and verifiable temporal metadata that allow the actual validity of datasets to be identified. This aspect is particularly relevant for integration into digital infrastructures, IoT systems, and analytical platforms that depend on continuously updated data.

Another key future line involves reinforcing the use of genuinely open and semantically enriched formats, reducing dependence on closed formats or poorly interoperable containers. In this regard, the more systematic adoption of semantic formats and the effective use of controlled vocabularies would facilitate the automatic integration of data and their reuse across multi-portal environments.

Although most of the analyzed portals declare compliance with the DCAT standard, there is a clear need to evolve from formal adherence toward a deeper implementation of the standard. This includes the provision of accessible data dictionaries, normalized field descriptions, and explicit references to shared vocabularies. Such progress would significantly enhance semantic interoperability across the ecosystem.

From a methodological perspective, future research could extend the proposed pipeline by incorporating additional data sources and analytical techniques. In particular, the integration of further academic search engines, European project platforms, and more advanced natural language processing methods would enable the detection of implicit dataset mentions that currently leave no explicit reuse traces. In addition, combining automated approaches with manual validation or qualitative case studies could improve the accuracy and coverage of reuse analysis.

As a methodological contribution and added value, this study proposes the incorporation of dataset citation and usage-tracking mechanisms directly within open data portals. This includes the standardized identification and citation of datasets through the assignment of persistent identifiers, such as dataset identifiers and Digital Object Identifiers (DOIs), which uniquely and persistently identify digital publications, including scientific articles, books, and datasets, ensuring that they can always be located and accessed when reused. Furthermore, the explicit inclusion of visible citation instructions on dataset landing pages would allow reuse in research projects or technological developments to be reflected directly within the portals themselves. As highlighted in previous studies, the mere technical availability of such mechanisms does not guarantee their effective use; therefore, open data portals must actively promote good citation practices among data reusers. Encouraging the adoption of standardized identifiers in open data portals, following the practices of research data repositories, should be regarded as a prerequisite for advancing toward robust empirical reuse metrics.

The automatic generation of citable identifiers and their linkage to external references would significantly improve dataset traceability and impact visibility, thereby reinforcing the overall maturity of the open data ecosystem.

It is also recommended that public administrations and portal managers consider incorporating internal reuse metrics, such as downloads, views, external references, or derived projects, to complement the external reuse traces analyzed in this study. The combination of internal and external metrics would provide a more comprehensive and realistic view of the actual impact and effective use of published datasets.

Finally, this project opens avenues for research aimed at evaluating the impact of public policies and open data citation guidelines on the effective reuse of datasets. Longitudinal analyses of portals that adopt persistent identifiers, mandatory citation rules, or attribution incentives would make it possible to assess whether such measures lead to increased traceability and visibility of open data usage.

Contributions of the Thesis

Methodological Contributions

- Design of an automated and reproducible pipeline for evaluating the maturity of open data portals, applicable across multiple technologies and institutional contexts.
- Proposal of a maturity model operationalized through five measurable dimensions derived from real metadata, going beyond purely normative or declarative approaches.
- Explicit decomposition of the concept of traceability into temporal, provenance-based, and reuse-related components, enabling its empirical assessment.

Technical Contributions

- Implementation of portal-specific scripts adapted to different technological environments (CKAN, OpenDataSoft, REST APIs, APIDATA + SPARQL), while maintaining homogeneous analytical criteria.
- Development of an independent module for the automated detection of external reuse traces (exploratory and methodological phase), conceived as a reusable methodological contribution in contexts with more favorable traceability conditions.
- Systematic normalization and comparison of heterogeneous metadata sourced from national, regional, and local open data portals.

Empirical Contributions

- Identification of structural weaknesses in the management of temporal metadata, identifiers, and access stability across relevant portals within the Spanish open data ecosystem.
- Empirical evidence demonstrating that the absence of persistent identifiers constitutes a critical barrier to advanced reuse of open data.
- Generation of comparable metrics that enable the assessment of the relative maturity of the analyzed open data portals.

Referencias

Ajuntament de Barcelona. (s. f.). *Open Data BCN: Portal de Datos Abiertos del Ayuntamiento de Barcelona.* <https://opendata-ajuntament.barcelona.cat/>

Aleixandre-Benavent, R., Sapena, A. F., & Peset, F. (2021). *Compartir los recursos útiles para la investigación: datos abiertos (open data).* *Educación Médica*, 22, 208-215. *Investigación* Edited by Jesús Millán Núñez-Cortés
<https://www.sciencedirect.com/science/article/pii/S1575181319301287>

Aparicio García, J. M. (2016). *Ecosistema de datos abiertos de la Universidad de Alicante.* Universidad de Alicante. Recuperado de
<https://rua.ua.es/entities/publication/2e7957fc-baa3-468a-95af-0324809fae03>

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality: Concepts, Methodologies and Techniques.*
<https://link.springer.com/book/10.1007/978-3-319-24106-7>

Bello García, A. (s. f.). Datos abiertos y participación en el gobierno social. *Economía Industrial*, 405, 95–104.
<https://www.mincotur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/405/BELLO%20GARCÍA.pdf> (mincotur.gob.es in Bing)

Comunidad de Madrid. (s. f.). *Portal de Datos Abiertos de la Comunidad de Madrid.*
<https://datos.comunidad.madrid/>

Criado, J. I., Jiménez Cid, C., & Alcaide Muñoz, L. (2023). Hacia la madurez de los portales de datos públicos abiertos en el sector público. Un análisis comparado del nivel local de gobierno en España. *Revista del CLAD Reforma y Democracia*, (85), 117-160. <https://doi.org/10.69733/clad.ryd.n85.a233>

Curto-Rodríguez, R. (2021). Análisis multidimensional de los portales de datos abiertos autonómicos españoles. *Revista Española de Documentación Científica*, 44(1), e289.
<https://doi.org/10.3989/redc.2021.1.1745>

Datos.gob.es. (2019). *Open Data: Publicación y reutilización de datos abiertos como iniciativa de Gobierno Abierto en la Administración (Guía).* Gobierno de España.
https://datos.gob.es/sites/default/files/doc/file/open_data_publicacion_y_reutilizacion_de_datos_abiertos_como_iniciativa_de_gobierno_abierto_en_la_administracion_compressed.pdf

datos.gob.es. (14 de junio de 2022). *¿Qué es un diccionario de datos y por qué es importante?* [Entrada de blog].
<https://datos.gob.es/es/blog/que-es-un-diccionario-de-datos-y-por-que-es-importante>

data.europa.eu (s. f.). *Módulo 1: ¿Qué son los datos abiertos?* [Módulo de curso en línea]. <https://data.europa.eu/elearning/es/module1/#/id/co-01>

De Sousa, F. K., Fonseca, K. V. O., Kozevitch, N. P., Liu, E., Naves, T., Patuzzo, G. V., & Beuren, A. T. (2025). *Rediseño de portales de datos abiertos: Mejorando la usabilidad y accesibilidad*. *Cuadernos del CIMBAGE*, 26(2), 26–40.
[https://doi.org/10.56503/CIMBAGE/Vol.2/Nro.26\(2024\)/3197](https://doi.org/10.56503/CIMBAGE/Vol.2/Nro.26(2024)/3197)

European Commission. (2019). *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information*. Official Journal of the European Union.
<https://eur-lex.europa.eu/eli/dir/2019/1024/oj/eng>

European Commission. (2020). *The economic impact of open data: Opportunities for value creation in Europe*.
<https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf>

European Data Portal. (2025). *2025 Open Data Maturity Report*. Publications Office of the European Union.
https://data.europa.eu/sites/default/files/2025-12/2025_odm_report_5.pdf

Finol, L. (2023). *Paradoja datos abiertos-privacidad como preámbulo de la política de acceso abierto impulsada desde universidades*. Gobierno y Administración Pública.
<https://doi.org/10.29393/GP5-3PDLF10003>

García-Espinosa, E., Rojas-Concepción AA, Vitón-Castillo AA, Lepez CO. (2022). Principios FAIR de gestión de datos de investigación en ciencias de la salud. *Revista Información Científica*, 101(6):e3865. [En línea]. Disponible:
<http://scielo.sld.cu/pdf/ric/v101n6/1028-9933-ric-101-06-e3865.pdf>

Gobierno de Aragón. (s. f.). Aragón Open Data. <https://opendata.aragon.es/datos>

Gobierno de España. (2010). Real Decreto 4/2010, de 8 de enero, por el que se regula el Esquema Nacional de Interoperabilidad en el ámbito de la Administración

Electrónica. Boletín Oficial del Estado. <https://www.boe.es/buscar/act.php?id=BOE-A-2010-1331>

Gobierno de España. (s. f.). *Portal de Datos Abiertos de España* (datos.gob.es). <https://datos.gob.es/es/>

Junta de Andalucía. (2023, 28 de abril). Principios del Open Data. Portal de Datos Abiertos de la Junta de Andalucía. <https://www.juntadeandalucia.es/datosabiertos/portal/actualidad/detalle/1519>

Junta de Andalucía. (s. f.). *Portal de Datos Abiertos de la Junta de Andalucía.* <https://www.juntadeandalucia.es/datosabiertos>

Junta de Castilla y León. (s. f.). *Análisis del Portal de Datos Abiertos de Castilla y León.* <https://analisis.datosabiertos.jcyl.es/>

Maseda-Seco, D., Bueno-de-la-Fuente, G., y Méndez, E. (2017). Análisis y categorización de los datos abiertos de las bibliotecas municipales españolas: Metadatos, interoperabilidad y propuesta para la apertura y reutilización. <https://doi.org/10.3145/epi.2017.may.05>

Muente-Kunigami, A., & Serale, F. (2018). *Los datos abiertos en América Latina y el Caribe.* Banco Interamericano de Desarrollo. <https://doi.org/10.18235/0001202>

Muñoz, H., & Clavijo, A. (2022). *El estado de los ecosistemas de datos en América Latina.* Cepei. <https://www.cepei.org/wp-content/uploads/2022/08/El-estado-de-los-ecosistemas-de-datos-en-America-Latina.pdf>

Naser, A., & Rosales, D. (2016). *Panorama regional de los datos abiertos: avances y desafíos en América Latina y el Caribe.* Comisión Económica para América Latina y el Caribe (CEPAL), ILPES. <https://www.cepal.org/es/publicaciones/40768-panorama-regional-datos-abiertos-avances-desafios-america-latina-caribe>

Neira Paredes, B. N. (2019). *Propuesta de modelo de madurez institucional para las iniciativas de datos abiertos del Gobierno de Chile* [Tesis de maestría, Universidad de Chile]. Repositorio Académico de la Universidad de Chile. <https://repositorio.uchile.cl/handle/2250/174693>

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Multi-level mappings for data portals: Exploring the high-level structure of DCAT. In H. Sack et al. (Eds.), *The Semantic Web: ESWC 2016 Satellite Events* (pp. 259–274). Springer.

https://doi.org/10.1007/978-3-319-46523-4_26

Oviedo Blanco, E. (2016). *Modelo de madurez para portales de datos abiertos e incorporación a la norma técnica nacional de Costa Rica* [Tesis de doctorado, Universidad de Alicante]. Dialnet.

<https://dialnet.unirioja.es/servlet/dctes?codigo=60727>

Quirós-García, E. (2021). *La huella digital y la protección de datos: su impacto en las culturas de alto contexto y alto control de incertidumbre en Latinoamérica*. InterSedes.

<https://www.redalyc.org/journal/666/66671361007/html/>

Región de Murcia. (s. f.). *Portal de Datos Abiertos de la Región de Murcia*.

<https://datosabiertos.regiondemurcia.es/>

Rodríguez Herrera, A., & Alvarado, H. (2015). *Ecosistemas de innovación social: el caso de las Universidades de América Latina*. Universidad de Los Andes. Recuperado de

https://d1wqtxts1xzle7.cloudfront.net/37404007/Ecosistemas_de_innovacion_social_-_el_caso_de_las_Universidades_de_America_Latina-libre.pdf

Salas Rodríguez, L. A. (2023). *Implementación de un modelo de medición del nivel de madurez en el uso de datos abiertos espaciales en el Estado Peruano* [Tesis de licenciatura, Pontificia Universidad Católica del Perú]. Repositorio Institucional PUCP.

<http://hdl.handle.net/20.500.12404/26145>

Seidor. (2025, octubre 16). *Hacia la convergencia tecnológica de los espacios de datos*. SEIDOR Blog.

<https://www.seidor.com/es-es/blog/hacia-la-convergencia-tecnologica-de-los-espacios-de-datos>

Subirana de la Cruz, S. (2016). Open Government: transparencia administrativa, derecho de acceso a la información pública, open data y reutilización de la información del sector público. *Revista Aranzadi Doctrinal*,. Disponible en:

<https://parlamento-cantabria.es/sites/default/files/dossieres-legislativos/Subirana%20de%20la%20Cruz.pdf>

W3C. (2020). *Data Catalog Vocabulary (DCAT) - Version 2*. W3C Recommendation. <https://www.w3.org/TR/vocab-dcat-2/>

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework, 7(1), 63–93. <https://doi.org/10.3233/SW-150175>