

Calidad de datos para datos abiertos

Cómo medir y mejorar la calidad utilizando Python

Presentación



César García Saez es **Ingeniero Técnico en Informática de Sistemas** por la Universidad Oberta de Cataluña, donde también cursó el **Máster Universitario de Ciencia de Datos**, especializándose en calidad de datos para datos abiertos. Trabaja como es **investigador independiente y formador en nuevas tecnologías**, especializado en Internet de las Cosas y la fabricación digital. Ha trabajado más de 13 años en Informática del Ayuntamiento de Madrid, dentro del departamento de Sistemas

Respecto a su experiencia en el ámbito de la impresión 3D y la fabricación digital, se graduó en en el año 2013 en León del **programa de fabricación digital Fabacademy**

Co-fundador de **Makespace Madrid**. Dirige el **podcast La Hora Maker**, donde recoge los proyectos y personas más interesantes del movimiento maker.

Forma parte de la **red de expertos de la Fundación COTEC**, estudiando el impacto de la impresión 3D a nivel económico y social.

Resumen



`df.isnull().sum()`

ISO
25012

Diez años de datos abiertos

Ley 19/2013, de 9 de diciembre de transparencia, acceso a la información pública y buen gobierno

Fuente: <https://www.boe.es/buscar/act.php?id=BOE-A-2013-12887>

Los países con mayores niveles en materia de transparencia y normas de buen gobierno cuentan con instituciones más fuertes, que favorecen el crecimiento económico y el desarrollo social. En estos países, los ciudadanos pueden juzgar mejor y con más criterio la capacidad de sus responsables públicos y decidir en consecuencia. Permitiendo una mejor fiscalización de la actividad pública se contribuye a la necesaria regeneración democrática, se promueve la eficiencia y eficacia del Estado y se favorece el crecimiento económico.

Preámbulo de la Ley de Transparencia, acceso a la información pública y las normas de buen gobierno

datos **abiertos**

¿Qué estás buscando?



#GRACIASMADRID

Tu ciudad más cerca

Gracias a nuestra plataforma de datos abiertos podrás encontrar todos los datos de Madrid que necesitas para tu proyecto

En portada

Acerca de Datos Abiertos

Catálogo de datos

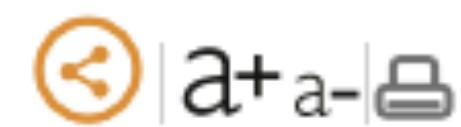
Colabora

Lo más visto

Localización de contenedores de r... / Mesas de tenis de mesa en instal... / Tasa criminalidad en barrios Madr...



En portada



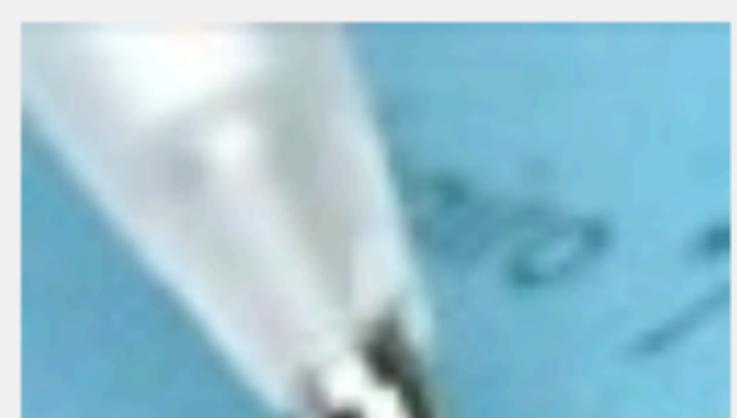
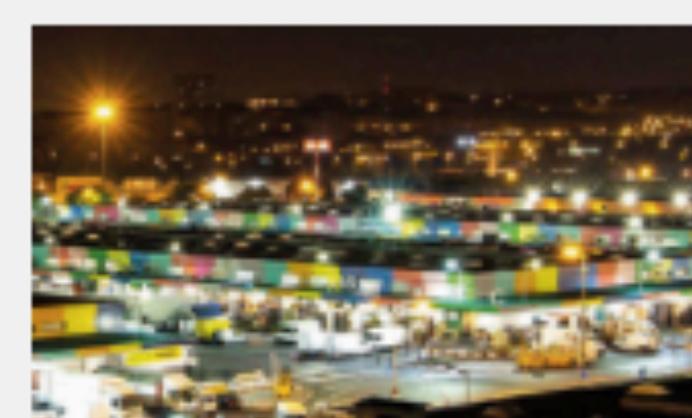
Conjuntos de datos destacados

¿Qué contiene el portal?

Datos de uso del portal

Novedades en el portal

Contacta



¿Qué estás buscando?



#GRACIASMADRID

TRANSPARENCIA

Para que todo lo que suceda esté a la vista de todas las personas

Portada

Organización

Recursos
humanos

Información
jurídica

Economía y
presupuestos

Relación con
la ciudadanía

Transparencia por
sectores

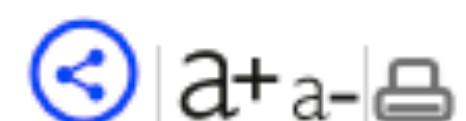
Acceso a la
información

Lo más visto

Compendio-2019 de las Normas Urba... / Retribuciones del personal funcio... / Niveles polínicos



Portada



¿Qué es transparencia?

¿Qué contiene el portal?

Evaluación de organismos
externos

Especial Covid-19 en Madrid

Datos, información y
resultados de la gestión
municipal del COVID-19



Datos de la
pandemia

Evolución gráfica de la
COVID-19 en la Ciudad

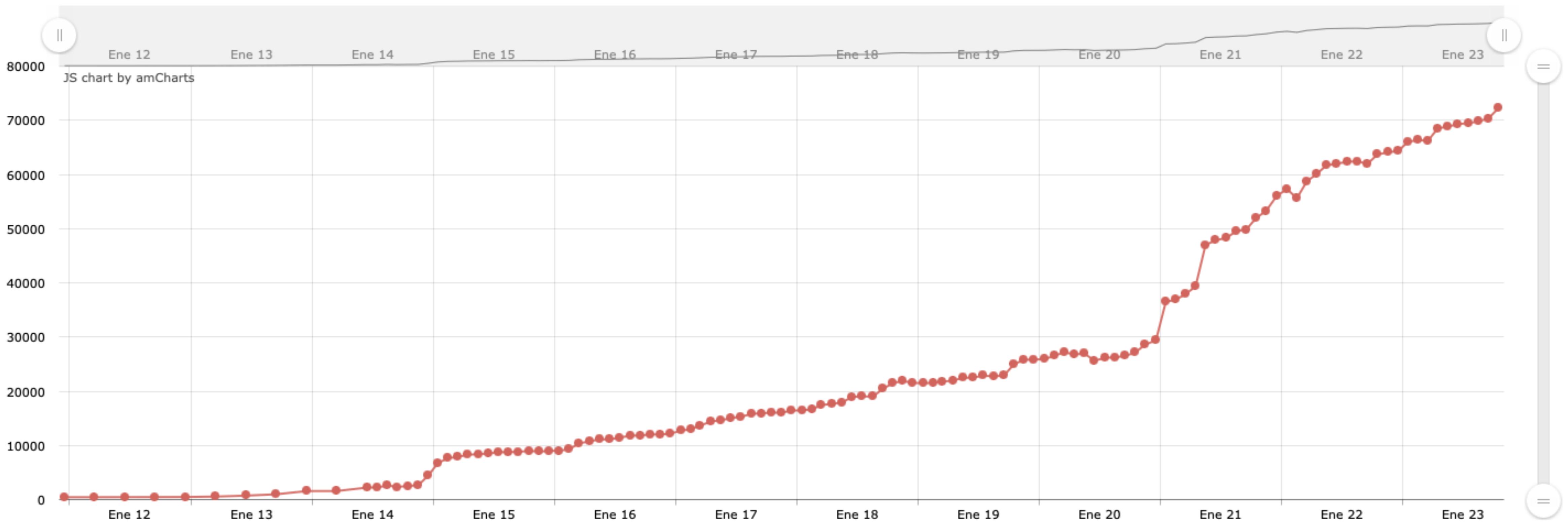


¿Qué son los datos abiertos?

Los datos abiertos son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen.

Crecimiento del número de datasets

Octubre 2023 - 72.230 datasets



Fuente: <https://datos.gob.es/es/dashboard>

1.527.490

Datasets en data.europa.eu

Estrategias futuras Union Europea

- Datasets de alto valor
(Deadline Junio 2024)
- Espacios de datos



**270 billones de EUR
 adicionales para 2028**

European Data Act (2022)

Fuente: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#:~:text=European%20data%20act,-With%20the%20European&text=The%20new%20rules%20are%20expected,to%20use%20products%20and%20services

**Cantidad >
Calidad**

NO PUEDES TENER MALA CALIDAD DE DATOS



SI NO LA MIDES

Calidad de datos

Estándares, normas y referencias

- ISO 25012
- ISO 25024
- DAMA DMBOK2
- Requisitos adicionales en la Propuesta de Ley de Inteligencia Artificial europea
- Artificial Intelligence - Data quality for analytics and machine learning (ISO/IEC 5259-X por desarrollar)

Calidad de datos

Estándares, normas y referencias nacionales

- Norma UNE 0077:2023 - Gobierno del dato
- Norma UNE 0077:2023 - Gestión del dato
- Norma UNE 0079:2023 - Gestión de la calidad del dato
- Norma UNE 0080:2023 - Guía de evaluación del Gobierno, Gestión y Gestión de la Calidad del Dato
- Norma UNE 0081:2023 - Guía de la evaluación de la Calidad de un Conjunto de Datos

Ahora mismo: El acceso a estas Especificaciones UNE está patrocinado por la Oficina del Dato de la SEDIA, siendo su descarga gratuita.

Grado en el que las características de los datos satisfacen las necesidades expresadas o implícitas cuando se utilizan bajo unas condiciones especificadas

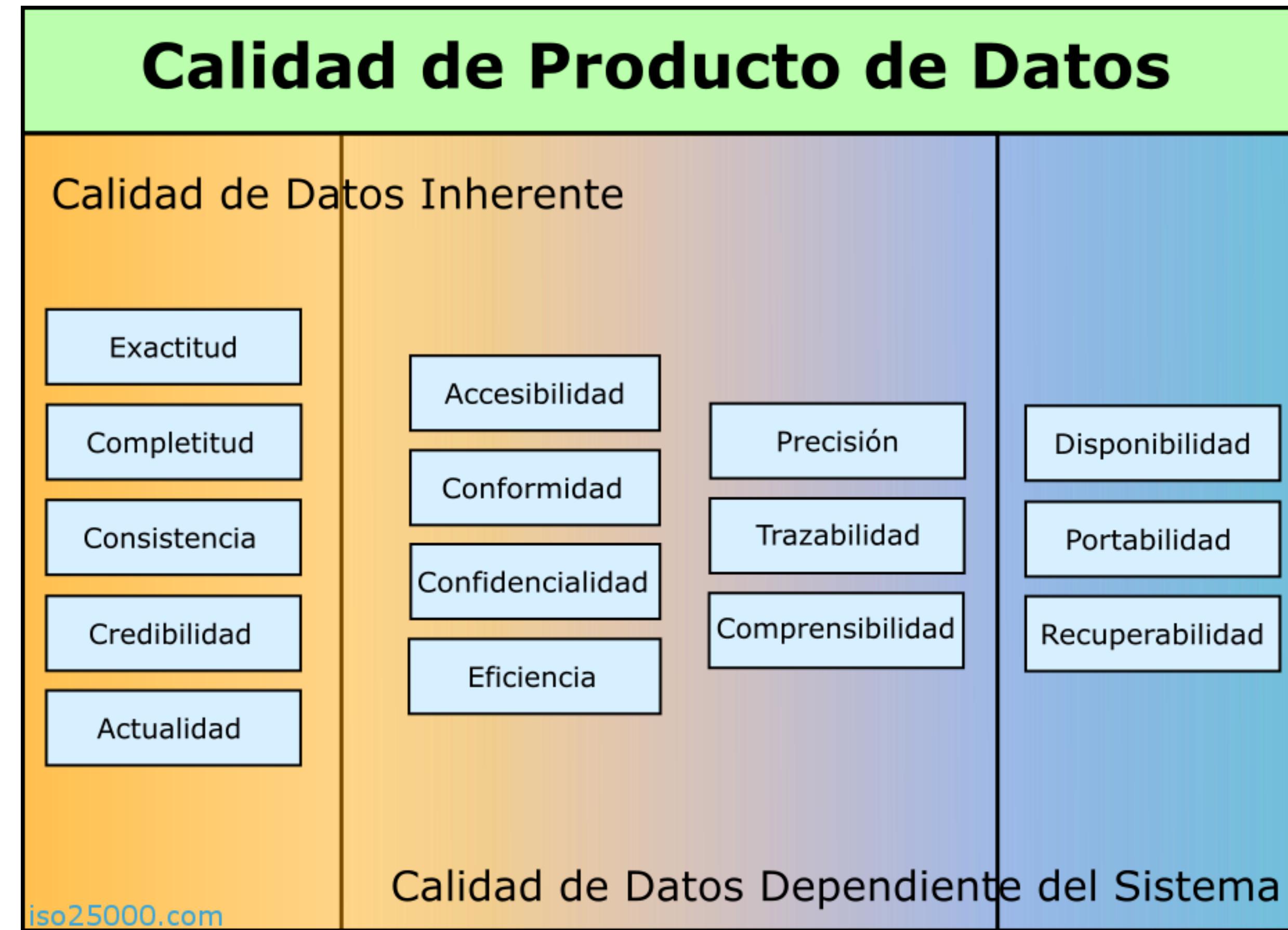
Norma UNE-ISO/IEC 25012

Grado en el que las características de los datos se ajustan al propósito

Resumen propio

Características ISO 25012

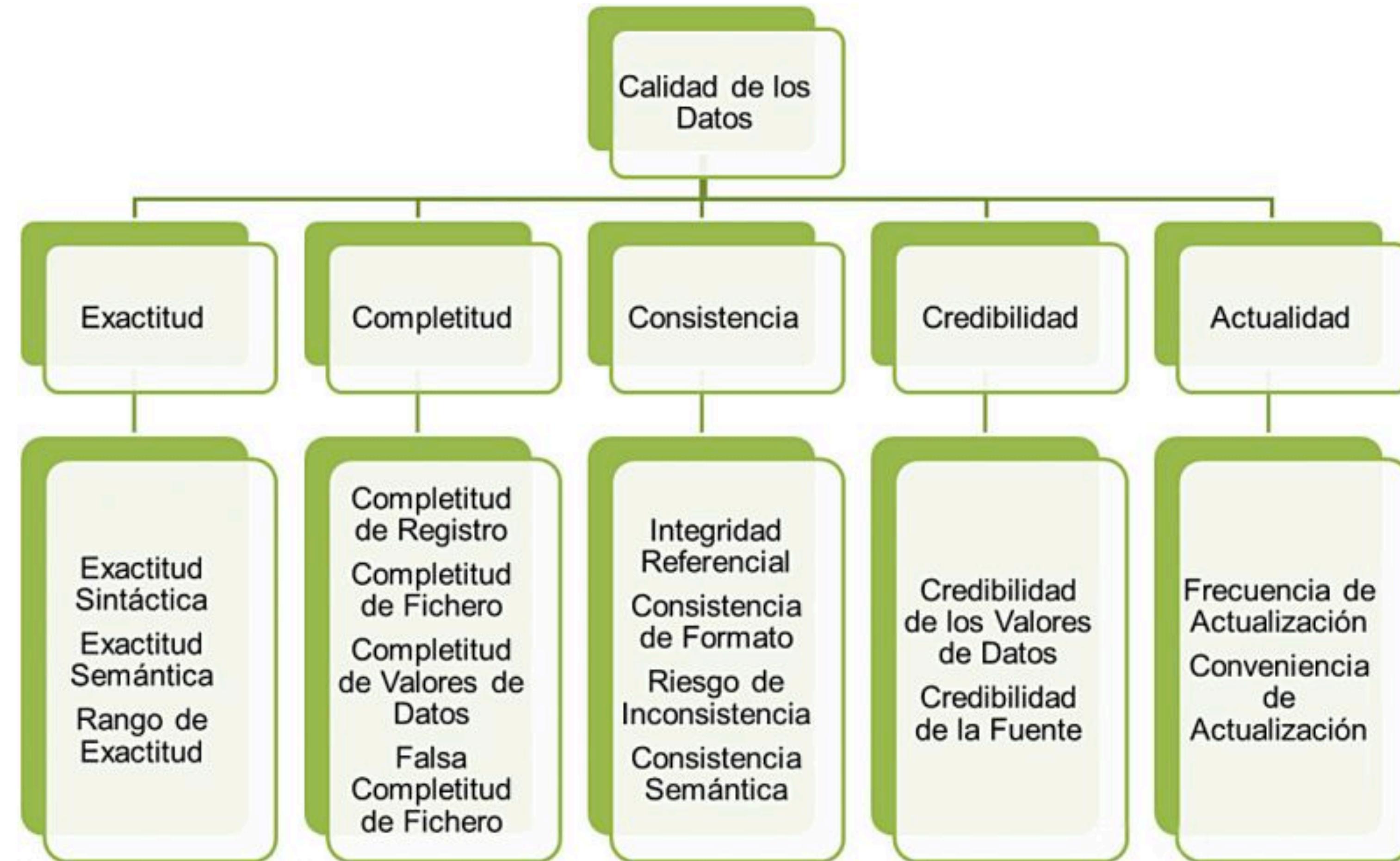
Inherentes y dependientes del sistema



Fuente: <https://iso25000.com/index.php/normas-iso-25000/iso-25012>

Métricas derivadas

Analizables desde los propios datasets



Referencia: <https://revista.aenor.com/358/gestion-en-las-organizaciones-basada-en-datos-de-calidad.html>

Proceso generación métricas

Calidad de un conjunto de datos

- Requisito de datos
- Requisito de calidad de datos
- Media ponderada

Proceso generación métricas

Calidad de un conjunto de datos

- Requisito de datos: Cada cuenta de usuario tendrá asociado un IBAN
- Requisito de calidad de datos:
 - El campo IBAN no será nulo para ninguna de las cuentas
 - Todos los IBAN tendrán un formato válido (regex)
 - (O alternativamente) el 95% de las cuentas tendrá un IBAN válido

**Ejemplos tomados de
datasets open data reales**

Formatos incorrectos

Exactitud semántica vs exactitud sintática

Año: 2023													
	CODIGO	ZONA VERDE / \nVIA PUBLICA	DISTRITO	DIRECCION	COMPLEMENTO_DIRECCION	COORD_X	COORD_Y	SISTEMA DE COORDENADAS	FECHA	DATOS	longitud	latitud	
0	FUE_01_0001	ZV CENTRO	PLAZA COMENDADORAS	PLAZA COMENDADORAS	439892.4997	4475419.798	ETRS89 / UTM	Zona 30N	45013	-3.708600	40.427243		
1	FUE_01_0002	ZV CENTRO	JARDINES CONDE DUQUE	JARDINES CONDE DUQUE	439766.5391	4475585.204	ETRS89 / UTM	Zona 30N	45013	-3.710101	40.428724		
2	FUE_01_0003	ZV CENTRO	PLAZA DE LAS SALESAS	PLAZA DE LAS SALESAS	441077.2599	4475077.763	ETRS89 / UTM	Zona 30N	45013	-3.694603	40.424246		
3	FUE_01_0004	ZV CENTRO	PLAZA VILLA DE PARIS	PLAZA VILLA DE PARIS	441281.789	4475192.9	ETRS89 / UTM	Zona 30N	45013	-3.692202	40.425298		
4	FUE_01_0005	ZV CENTRO	PLAZA VILLA DE PARIS	PLAZA VILLA DE PARIS	441199.3988	4475222.805	ETRS89 / UTM	Zona 30N	45013	-3.693176	40.425562		
Año: 2022													
	CODIGO	ZONA VERDE / \nVIA PUBLICA	DISTRITO	DIRECCION	COMPLEMENTO DE DIRECCION	COORD_X	COORD_Y	SISTEMA DE COORDENADAS	FECHA	DATOS	Longitud	Latitud	
0	FUE_01_0001	ZV CENTRO	PLAZA COMENDADORAS	PLAZA COMENDADORAS	439892.4997	4475419.798	ETRS89 / UTM	Zona 30N	44925.0	-3.708600	40.427243		
1	FUE_01_0002	ZV CENTRO	JARDINES CONDE DUQUE	JARDINES CONDE DUQUE	439766.5391	4475585.204	ETRS89 / UTM	Zona 30N	44925.0	-3.710101	40.428724		
2	FUE_01_0003	ZV CENTRO	PLAZA DE LAS SALESAS	PLAZA DE LAS SALESAS	441077.2599	4475077.763	ETRS89 / UTM	Zona 30N	44925.0	-3.694603	40.424246		
3	FUE_01_0004	ZV CENTRO	PLAZA VILLA DE PARIS	PLAZA VILLA DE PARIS	441281.7890	4475192.900	ETRS89 / UTM	Zona 30N	44925.0	-3.692202	40.425298		
4	FUE_01_0005	ZV CENTRO	PLAZA VILLA DE PARIS	PLAZA VILLA DE PARIS	441199.3988	4475222.805	ETRS89 / UTM	Zona 30N	44925.0	-3.693176	40.425562		
Año: 2021													
	CODIGO	ZONA VERDE / \nVIA PUBLICA	DISTRITO	DIRECCION	COMPLEMENTO_DIRECCION	COORD_X	COORD_Y	SISTEMA DE COORDENADAS	OBSERVACIONES	FECHA	DATOS	Longitud	Latitud
0	FUE_01_0001	ZV CENTRO	PLAZA COMENDADORAS	PLAZA COMENDADORAS	439892.4997	4475419.798	ETRS89 / UTM	Zona 30N	FUERA DE SERVICIO	31-oct.-21	-3.708600	40.427243	
1	FUE_01_0002	ZV CENTRO	JARDINES CONDE DUQUE	JARDINES CONDE DUQUE	439766.5391	4475585.204	ETRS89 / UTM	Zona 30N	EN SERVICIO	31-oct.-21	-3.710101	40.428724	
2	FUE_01_0003	ZV CENTRO	PLAZA DE LAS SALESAS	PLAZA DE LAS SALESAS	441077.2599	4475077.763	ETRS89 / UTM	Zona 30N	EN SERVICIO	31-oct.-21	-3.694603	40.424246	
3	FUE_01_0004	ZV CENTRO	PLAZA VILLA DE PARIS	PLAZA VILLA DE PARIS	441281.789	4475192.900	ETRS89 / UTM	Zona 30N	EN SERVICIO	31-oct.-21	-3.692202	40.425298	
4	FUE_01_0005	ZV CENTRO	PLAZA VILLA DE PARIS	PLAZA VILLA DE PARIS	441199.3988	4475222.805	ETRS89 / UTM	Zona 30N	EN SERVICIO	31-oct.-21	-3.693176	40.425562	
Año: 2020													

Referencia: Fuentes de Agua de Beber del Ayuntamiento de Madrid 2021 y 2022

Formatos incorrectos

Exactitud semántica vs exactitud sintática

COMPLEMENTO_DIRECCION	COORD_X	COORD_Y	SISTEMA DE COORDENADAS	FECHA DATOS	longitud	latitud
PLAZA COMENDADORAS	439892.4997	4475419.798	ETRS89 / UTM Zona 30N	45013	-3.708600	40.427243
JARDINES CONDE DUQUE	439766.5391	4475585.204	ETRS89 / UTM Zona 30N	45013	-3.710101	40.428724
PLAZA DE LAS SALESAS	441077.2599	4475077.763	ETRS89 / UTM Zona 30N	45013	-3.694603	40.424246
PLAZA VILLA DE PARIS	441281.789	4475192.9	ETRS89 / UTM Zona 30N	45013	-3.692202	40.425298
PLAZA VILLA DE PARIS	441199.3988	4475222.805	ETRS89 / UTM Zona 30N	45013	-3.693176	40.425562

COMPLEMENTO DE DIRECCION	COORD_X	COORD_Y	SISTEMA DE COORDENADAS	FECHA DATOS	Longitud	Latitud
PLAZA COMENDADORAS	439892.4997	4475419.798	ETRS89 / UTM Zona 30N	44925.0	-3.708600	40.427243
JARDINES CONDE DUQUE	439766.5391	4475585.204	ETRS89 / UTM Zona 30N	44925.0	-3.710101	40.428724
PLAZA DE LAS SALESAS	441077.2599	4475077.763	ETRS89 / UTM Zona 30N	44925.0	-3.694603	40.424246
PLAZA VILLA DE PARIS	441281.7890	4475192.900	ETRS89 / UTM Zona 30N	44925.0	-3.692202	40.425298
PLAZA VILLA DE PARIS	441199.3988	4475222.805	ETRS89 / UTM Zona 30N	44925.0	-3.693176	40.425562

COMPLEMENTO_DIRECCION	COORD_X	COORD_Y	SISTEMA DE COORDENADAS	OBSERVACIONES	FECHA DATOS	Longitud	Latitud
PLAZA COMENDADORAS	439892.4997	4475419.798	ETRS89 / UTM Zona 30N	FUERA DE SERVICIO	31-oct.-21	-3.708600	40.427243
JARDINES CONDE DUQUE	439766.5391	4475585.204	ETRS89 / UTM Zona 30N	EN SERVICIO	31-oct.-21	-3.710101	40.428724
PLAZA DE LAS SALESAS	441077.2599	4475077.763	ETRS89 / UTM Zona 30N	EN SERVICIO	31-oct.-21	-3.694603	40.424246
PLAZA VILLA DE PARIS	441281.789	4475192.900	ETRS89 / UTM Zona 30N	EN SERVICIO	31-oct.-21	-3.692202	40.425298
PLAZA VILLA DE PARIS	441199.3988	4475222.805	ETRS89 / UTM Zona 30N	EN SERVICIO	31-oct.-21	-3.693176	40.425562

Referencia: Fuentes de Agua de Beber del Ayuntamiento de Madrid 2021 y 2022

Campos crecientes - escalera invertida

Ajuntament de Barcelona CONTRACTES AMB MODIFICACIONS 2021 (1 de gener a 31 de desembre)										
GERÈNCIES I DISTRICTES										
Órgans contractant	Contracte núm.	Típus de contracte (obres, serveis, administració...)	Objecte del contracte	NOM ADJUDICATARI (RAÓ SOCIAL)	NIF ADJUDICATARI (Persones Físiques autoritzades)	Data formalització contracte	Preu contracte (sense IVA)	Preu contracte (IVA inclòs)	Data formalització modificació	Import modificació (IVA inclòs)

Ajuntament de Barcelona CONTRACTES AMB MODIFICACIONS 2020 (1 de gener a 31 de desembre)										
GERÈNCIES I DISTRICTES										
Órgans contractant	Contracte núm.	Típus de contracte (obres, serveis, administració...)	Objecte del contracte	Data formalització contracte	Preu contracte (sense IVA)	Preu contracte (IVA inclòs)	Data formalització modificació	Import modificació (sense IVA)	Import modificació (IVA inclòs)	Típus modificació (méspreu amb una "E")

Ajuntament de Barcelona CONTRACTES AMB MODIFICACIONS 2019 (1 gener a 31 desembre)										
GERÈNCIES I DISTRICTES DE L'AJUNTAMENT DE BARCELONA										
* Iva inclòs										
Órgans contractant	Contracte núm.	Típus de contracte (obres, serveis, administració...)	Objecte contracte	Data adjudicació / formalització	Preu* contracte	Data formalització modificació	Import * modificació	Típus modificació (méspreu amb una x)	% variació sobre el preu del contracte	
Gerència de Recursos	15002241-002	Privat d'Administració Pública	Cessió obra artística Bombers de Barcelona	20/10/15	40.000,00	NO CONSTA	90.000,00	SÍ	25,00	Ampliació de temps- Reajustament d'assiduitat

Ajuntament de Barcelona CONTRACTES AMB MODIFICACIONS 2018 (1 gener a 31 desembre)										
GERÈNCIES I DISTRICTES										
* Iva inclòs										
Órgans contractant	Contracte núm.	Típus de contracte (obres, serveis, administració...)	Objecte contracte	Data formalització	Preu* contracte	Data formalització modificació	Import * modificació	Típus modificació (méspreu amb una x)	% variació sobre el preu del contracte	OBSERVACIONS

Ajuntament de Barcelona CONTRACTES AMB MODIFICACIONS 2017 (1 gener a 31 desembre)										
* Iva inclòs										
ORGAN CONTRACTANT	Típus de contracte (obres, serveis, administració...)	Objecte contracte	Data adjudicació	Preu* contracte	Data acord modificació	Import * modificació	Típus modificació (méspreu amb una x)	% variació sobre el preu del contracte		
Gerència de Recursos	SERVEIS I OBRES	MANTENIMENT ELEMENTS VIA PÚBLICA DISTRICTES LOTS	17/12/14	1.156.050,00	26/10/17	86.703,75	SÍ	SÍ	7,50	

Formatos cambiantes

Cinco años, seis encabezados diferentes - Contratistas Barcelona

2021 Llistat Ampliat Contractistes de Contractes Pùblics Adjudicats en 2021 - Ajuntament de Barcelona (Gerències i Districtes)						
Nom Adjudicatari	NIF (per Publicar: PF anonimitzat)	Tipus Contracte	Objecte	Procediment (detall per a Publicar)	Import Adjudicació Net (sense IVA)	Import Adjudicació amb IVA
2020 Llistat Ampliat Contractistes de Contractes Pùblics Adjudicats en 2020 - Ajuntament de Barcelona (Gerències i Districtes)						
Nom Adjudicatari	NIF (per a Publicar: PF anonimitzat)	Tipus Contracte	Objecte	Procediment (detall per a publicar)	Import Adjudicació Net (sense IVA)	Import Adjudicació amb IVA
2019 Llistat Ampliat Proveïdors Contractes Pùblics Ajuntament de Barcelona (Gerències i Districtes)						
Nom Proveïdor	NIF (per Publicar: PF anonimitzat)	Tipus Contracte	Objecte	Procediment (per Publicar)	Import d'Adjudicació Net	Import d'Adjudicació +IVA
Nom Proveïdor	NIF Proveïdor	Tipus de Contracte	Objecte	Procediment (per Publicar)	Import Adjudicació Net	Import d'Adjudicació+IVA
Proveïdor	NIF Proveïdor	Tipus Contracte	Objecte	Procediment	Import Adjudicació	Nombre de contracts

Campos compuestos

✖ 0 300253-16-contratos-actividad-menores-2022.csv

TIPO DE CONTRATO	N. DE INVITACIONES CURSADAS	INVITADOS A PRESENTAR OFERTA
Servicios	3	(A79054748) TECNICOS ASOCIADOS INFORMATICA, S.A. PYME: SI (B83628339) REDCOM CIBERNÉTICO, S.L.. PYME: SI (B86061249) ZENER SOFTCONSULTING S.L.. PYME: SI
Suministros	3	(A79054748) TECNICOS ASOCIADOS INFORMATICA, S.A. PYME: SI (B86061249) ZENER SOFTCONSULTING S.L.. PYME: NO (B98198286) INFORMÁTICA FAER, S.L.. PYME: NO
Suministros	1	(B85635910) EDICIONES EL PAIS, S.L.. PYME: NO
Suministros	1	(A79102331) UNIDAD EDITORIAL; S.A.. PYME: NO
Suministros	1	(B81511834) DIARIO AS; S.L.. PYME: NO
Suministros	1	(A82031329) AUDIOVISUAL ESPAÑOLA 2000, S.A.. PYME: NO
Suministros	1	(B82824194) DIARIO ABC S.L.. PYME: NO
Suministros	1	(00837947B) ALBERTO VIDAL SILGADO. PYME: SI
Suministros	1	(A28760692) VALORIZA SERVICIOS MEDIOAMBIENTALES, S.A.. PYME: NO
Servicios	1	(B85653350) LEGAL PLANNING, SL. PYME: SI
Servicios	1	(B84943091) FALCONERS IBERIA SL. PYME: SI
Suministros	1	(A78510963) FORMULARIOS EUROPEOS, S.A.. PYME: SI
Servicios	3	(B61220521) SALVETTI LLOMBART, S.L.. PYME: SI (B82095522) INFORMATION RESOURCES ESPAÑA, S.L.. PYME: SI (G08557985) ASOCIACION ESPAÑOLA DE CODIFICACIONCOMERCIAL AECOC. PYME: SI
Obras	5	(A28057529) FERROVIAL S.A.. PYME: NO (A28649911) INLLAMA S A. PYME: SI (B28986511) CIMA EXTINTORES. PYME: SI (B82205170) NORMA FIRE, S.L.. PYME: SI (B84194034) INGENIERIA Y CONSERVACIÓN CONTRAINCENDIOS, S.L.. PYME: SI
Suministros	4	(A81286759) SIADDE SOLUCIONES S.A.. PYME: SI (B40165920) INV SISTEMAS Y SOLUCIONES DE SEGURIDAD, S. L.. PYME: SI (B63809560) WALLNER EUROPA S.L.. PYME: SI (B86184082) KOMANET TELECOMUNICACIONES S.L.. PYME: SI

Campos compuestos

122.csv

INVITADOS A PRESENTAR OFERTA

(A79054748) TECNICOS ASOCIADOS INFORMATICA, S.A. PYME: SI (B83628339) REDCOM CIBERNÉTICO, S.L.. PYME: SI (B86061249) ZENER SOFTCONSULTING S.L.. PYME: SI
(A79054748) TECNICOS ASOCIADOS INFORMATICA, S.A. PYME: SI (B86061249) ZENER SOFTCONSULTING S.L.. PYME: NO (B98198286) INFORMÁTICA FAER, S.L.. PYME: NO
(B85635910) EDICIONES EL PAÍS, S.L.. PYME: NO
(A79102331) UNIDAD EDITORIAL; S.A.. PYME: NO
(B81511834) DIARIO AS; S.L.. PYME: NO
(A82031329) AUDIOVISUAL ESPAÑOLA 2000, S.A.. PYME: NO
(B82824194) DIARIO ABC S.L.. PYME: NO
(00837947B) ALBERTO VIDAL SILGADO. PYME: SI
(A28760692) VALORIZA SERVICIOS MEDIOAMBIENTALES, S.A.. PYME: NO
(B85653350) LEGAL PLANNING, SL. PYME: SI
(B84943091) FALCONERS IBERIA SL. PYME: SI
(A78510963) FORMULARIOS EUROPEOS, S.A.. PYME: SI
(B61220521) SALVETTI LLOMBART, S.L.. PYME: SI (B82095522) INFORMATION RESOURCES ESPAÑA, S.L.. PYME: SI (G08557985) ASOCIACION ESPAÑOLA DE CODIFICACIONCOME
(A28057529) FERROVIAL S.A.. PYME: NO (A28649911) INLLAMA S A. PYME: SI (B28986511) CIMA EXTINTORES. PYME: SI (B82205170) NORMA FIRE, S.L.. PYME: SI (B84194034) I
(A81286759) SIADDE SOLUCIONES S.A.. PYME: SI (B40165920) INV SISTEMAS Y SOLUCIONES DE SEGURIDAD, S. L.. PYME: SI (B63809560) WALLNER EUROPA S.L.. PYME: SI (B8

Datos inconsistentes intrafichero

Contratos modificados Barcelona

Gerència Drets de Ciutadania, Participació i Transparència	17002834-001	Serveis	Gestió, impuls i dinamització Centre Recursos DH	31/01/18	249.853,60	15/03/19	0,00	Sí		0,00
Gerència Drets de Ciutadania, Participació i Transparència	18002708-001	Serveis	Servei atenció telefònica, gestió i tram. 010	19/02/19	11.021.334,34	NO CONSTA	-642.911,17	Sí		-5,83
Gerència Drets de Ciutadania, Participació i Transparència	19001461-001	Serveis	Serveis de traducció i correcció de textos	02/04/19	18.876,00	NO CONSTA	0,00	0		0,00
Gerència Drets de Ciutadania, Participació i Transparència	19002112-001	Subministraments	Trobada BCN Ciutat Diversa (2019-2020)	26/08/19	164.424,86	10/12/19	12.829,33	Sí		7,80
Gerència de Seguretat i Prevenció	18003713	Serveis	Neteja, descontaminació equips respiració SPEIS	31/7/19	503.536,10	17/10/19	177.706,70	Justament anualitats		35%
Gerència de Seguretat i Prevenció	19002901	Serveis	Suport organització comunitats veïns i veïnes	30/10/19	332.000,00	19/11/19	6916,66	Justament anualitats		2%
Gerència de Seguretat i Prevenció	19002770	Serveis	Mant.preventiu i correctiu eines hidràuliques SPEIS Lot 1	5/12/19	85.498,60	19/12/19	9406,88	Justament anualitats		11%
Gerència d'Ecologia Urbana	18002285	Subministraments	Subministrament de material d'impremta per a les campanyes de comunicació de l'Àrea d'Ecologia Urbana 2018-2019	08.01.2019	120.028,50 €	27.08.2019	24.005,70	X		20,00%
Gerència d'Ecologia Urbana	16001721	Serveis	Assistència Tècnica i Control de Qualitat del Contracte de Conservació de l'Enllumenat Públic de Barcelona (2016-19)	14.12.2019	775.473,85 €	31.10.2019	21.000,00	X		2,71%
Gerència d'Ecologia Urbana	18004961	Mixte	Subministrament i serveis de col·locació i retirada de plafons electorals durant el període 2019-2021	05.04.2019	73.928,79 €	25.10.2019	20.500,00	X		27,73%
Gerència d'Ecologia Urbana	16002516	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 1 (Ciutat Vella, Eixample i Sants-Montjuïc)	05.12.2016	17.374.923,70 €	09.12.2019	40.000,00	X		
Gerència d'Ecologia Urbana	16002517	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 2 (Les Corts, Sarrià-Sant Gervasi, Gràcia i Horta-Guinardó)	05.12.2016	17.641.863,60 €	09.12.2019	160.000,00	X		
Gerència d'Ecologia Urbana	16002518	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 3 (Nou Barris, Sant Andreu i Sant Martí)	05.12.2016	21.083.212,70 €	09.12.2019	20.000,00	X		
Gerència d'Ecologia Urbana	16005605	Serveis	Serveis de recollida i trasllat d'animals de companyia perduts o abandonats a la ciutat de Barcelona al Centre d'Accolliment d'Animals de Companyia de Barcelona (CAACB),	06.10.2017	323.000,00 €	16.12.2019	32.300,00	X		10,00%
Gerència d'Ecologia Urbana	14003093	Serveis	Manteniment del clavegueram de Barcelona (2015-2022)	16.02.2015	103.512.021,28 €	23.12.2019	54.132,10	X		0,05%
Gerència d'Ecologia Urbana	17006454	Serveis	Servei de Manteniment d'Escales Mecàniques, Ascensors Verticals i Ascensors Inclinats 2018-2020	13.07.2018	3.384.216,51 €	23.12.2019	100.000,00	X		2,95%

Referencia: https://ajuntament.barcelona.cat/transparencia/sites/default/files/modificacions_2019_ajuntamentbcn_0.xlsx

Datos inconsistentes intrafichero

Contratos modificados Barcelona

31/01/18	249.853,60	15/03/19	0,00	sí		0,00
19/02/19	11.021.334,34	NO CONSTA	-642.911,17	sí		-5,83
02/04/19	18.876,00	NO CONSTA	0,00	0		0,00
26/08/19	164.424,86	10/12/19	12.829,33	sí		7,80
31/7/19	503.536,10	17/10/19	177.706,79 ajustament anualitats			35%
30/10/19	332.000,00	19/11/19	6916,66 ajustament anualitats			2%
5/12/19	85.498,60	19/12/19	9406,88 ajustament anualitats			11%
08.01.2019	120.028,50 €	27.08.2019	24.005,70	X		20,00%
14.12.2019	775.473,85 €	31.10.2019	21.000,00	X		2,71%
05.04.2019	73.928,79 €	25.10.2019	20.500,00	X		27,73%
05.12.2016	17.374.923,70 €	09.12.2019	40.000,00	X		
05.12.2016	17.641.863,60 €	09.12.2019	160.000,00	X		
05.12.2016	21.083.212,70 €	09.12.2019	20.000,00	X		
06.10.2017	323.000,00 €	16.12.2019	32.300,00	X		10,00%
16.02.2015	103.512.021,28 €	23.12.2019	54.132,10	X		0,05%
13.07.2018	3.384.216,51 €	23.12.2019	100.000,00	X		2,95%

Referencia: https://ajuntament.barcelona.cat/transparencia/sites/default/files/modificacions_2019_ajuntamentbcn_0.xlsx

Hackers cívicos

Jaime Gómez-Obregón

Explorar Últimos contratos Ranking Nuevo Qué es esto En los medios Descarga los datos

El mapa de la **contratación pública** en Cantabria

Explora 33.309 contratos del Gobierno de Cantabria y más de 4.500 licitadores.
Una herramienta digital interactiva para la transparencia y el análisis
desarrollada por Jaime Gómez-Obregón. [\[Saber más\]](#)

×

FEDERACION CANTABRA DE BOLOS

IMPORTE **107.180 €** CONTRATOS **10**



25/4/2018 **Suministro de Juegos de Bolos**

15/2/2022 **Organizacion del V Concurso Mixto por Parejas de Bolo Palma "Bolos en Femenino"**

ADJUDICATARIO **FEDERACION CANTABRA DE BOLOS**

PRECIO **7260 €**

7/4/2021 **Organizacion del IV Concurso Mixto por Parejas de Bolo Palma "Bolos en Femenino"**

ADJUDICATARIO **FEDERACION CANTABRA DE BOLOS**

PRECIO **7260 €**

20/5/2016 **Suministro de Juegos de Bolos**

ADJUDICATARIO **POSTIGO FERNANDEZ LEANDRO JESUS, GOMEZ TRAPAGA RUBEN**

PRECIO **11.616 €**

Dirección General de Deporte > Consejería de Universidades, Igualdad, Cultura y Deporte

24/3/2021 **Suministro de Juegos de Bolos**

ADJUDICATARIO **GOMEZ TRAPAGA RUBEN**

PRECIO **7079 €**

Dirección General de Deporte > Consejería de Universidades, Igualdad, Cultura y Deporte

Diferentes puntos de vista

El propietario del dato determina la calidad



Todos los registros tienen una fecha. No hay fechas en blanco.



conjunto_datos_2019.xls

2019			
ID	Fecha	Importe	Adjudicatario
1	5/01/2019		
2	4 Marzo 2019		
3	12/31/2019		
4	1.Oct.19		



Los formatos de las fechas son todos diferentes y dificulta el análisis.



Pioneros

Open Data Toronto

The screenshot shows the Open Data Toronto website's Catalogue Quality Scores page. The top navigation bar includes links for OPEN DATA, Data Catalogue, Knowledge Centre, About, Gallery, Contact, a search bar, and a 'SEARCH' button. The main content area has a breadcrumb trail: OPEN DATA PORTAL HOME / OPEN DATA CATALOGUE / CATALOGUE QUALITY SCORES. On the left, there's a sidebar with dataset details: Details, Data quality score (beta, Gold), Data last refreshed (Dec 17, 2019), Refreshed (Weekly), Data type (Table), Topics (City government), License (Open Government License - Toronto). A section titled 'About Catalogue quality scores' explains the badge system based on dataset characteristics like usability, timeliness, and completeness. It also describes the collection method (CKAN API) and limitations, noting that static files are not scored. A 'DATA PREVIEW' table shows a single row of data with columns: _id, package, accessibility, completeness, freshness, metadata, usability, and score. The data row is: 186, air-conditioned-and-cool-, 1, 0.69, 0.5, 0.84, 0.86, 0.78.

OPEN DATA Data Catalogue Knowledge Centre About Gallery Contact

Search **SEARCH**

OPEN DATA PORTAL HOME / OPEN DATA CATALOGUE / CATALOGUE QUALITY SCORES

Details

Data quality score beta
Gold

Data last refreshed
Dec 17, 2019

Refreshed
Weekly

Data type
Table

Topics
City government

License
Open Government License -
Toronto

Publisher

Published by
Information & Technology

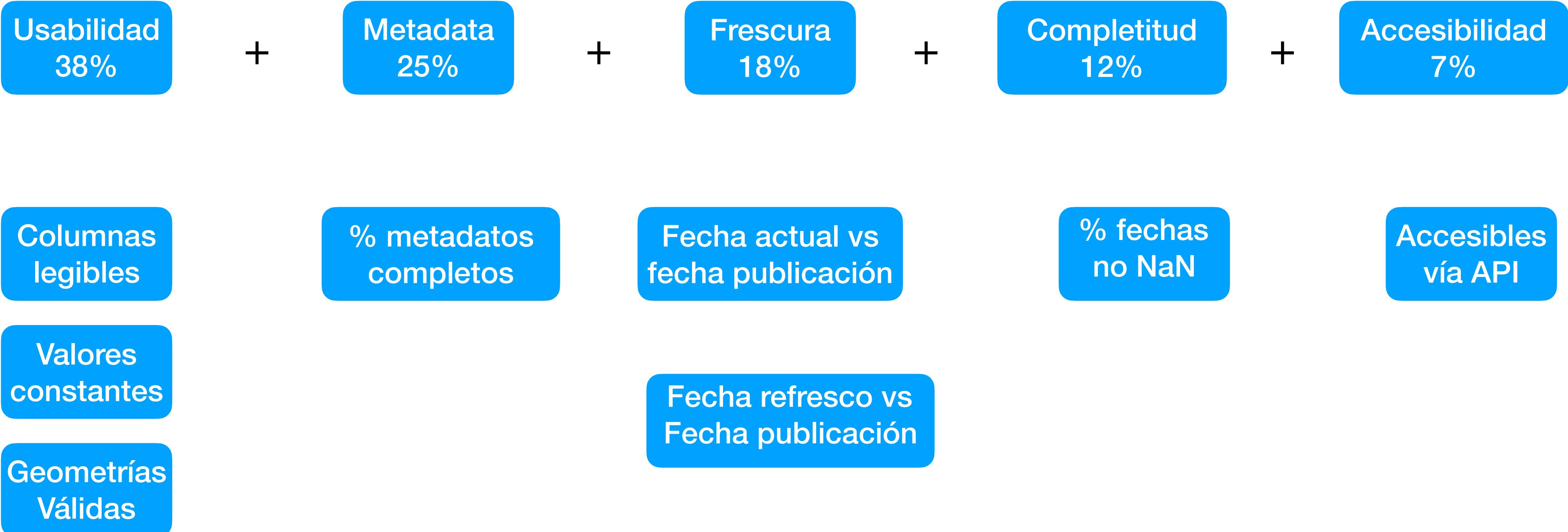
Contact
carlos.hernandez@toronto.ca

DATA PREVIEW

_id	package	accessibility	completeness	freshness	metadata	usability	score
186	air-conditioned-and-cool-	1	0.69	0.5	0.84	0.86	0.78

Ejemplo de medición

Open Data Toronto



Tidy Data

Publicación de Hadley Wickman (2014)

- Cada columna una variable
- Cada fila una observación
- Cada fichero un dominio de datos

Root cause remediation: Improving the quality of data goes beyond correcting errors. Problems with the quality of data should be understood and addressed at their root causes, rather than just their symptoms. Because these causes are often related to process or system design, improving data quality often requires changes to processes and the systems that support them.

DAMA DMBOK2 - Data Quality Programs principles

Many people assume that most data quality issues are caused by data entry errors. A more sophisticated understanding recognizes that gaps in or poor execution of business and technical processes cause many more problems than mis-keying. However, common sense says and research indicates that many data quality problems are caused by a lack of organizational commitment to high quality data, which itself stems from a lack of leadership, in the form of both governance and management.

DAMA DMBOK2 - Common Causes of Data Quality Issues

Paradoja de los datos abiertos

Inspirado en: https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/8_3_02_Nikiforova.pdf

Retos para la reutilización de los datos abiertos - sector público

Informe ASEDE 2023

- La información proporcionada en los datos no es homogénea (41.9%)
- Los conjuntos de datos no están disponibles en todas las CCAA o en todos los Ayuntamientos. (41%)
- La falta de actualización de los datos (38.1%)

Sector Académico

Ante la solicitud de identificar las tres barreras principales a la hora de acceder a los datos públicos para su reutilización por parte del Sector Académico se identifican:

- La falta de calidad de los datos (43%)
- Los conjuntos de datos no están disponibles en todas las CCAA o en todos los Ayuntamientos. (43%)
- La falta de actualización de los datos (39%)

Sector Privado

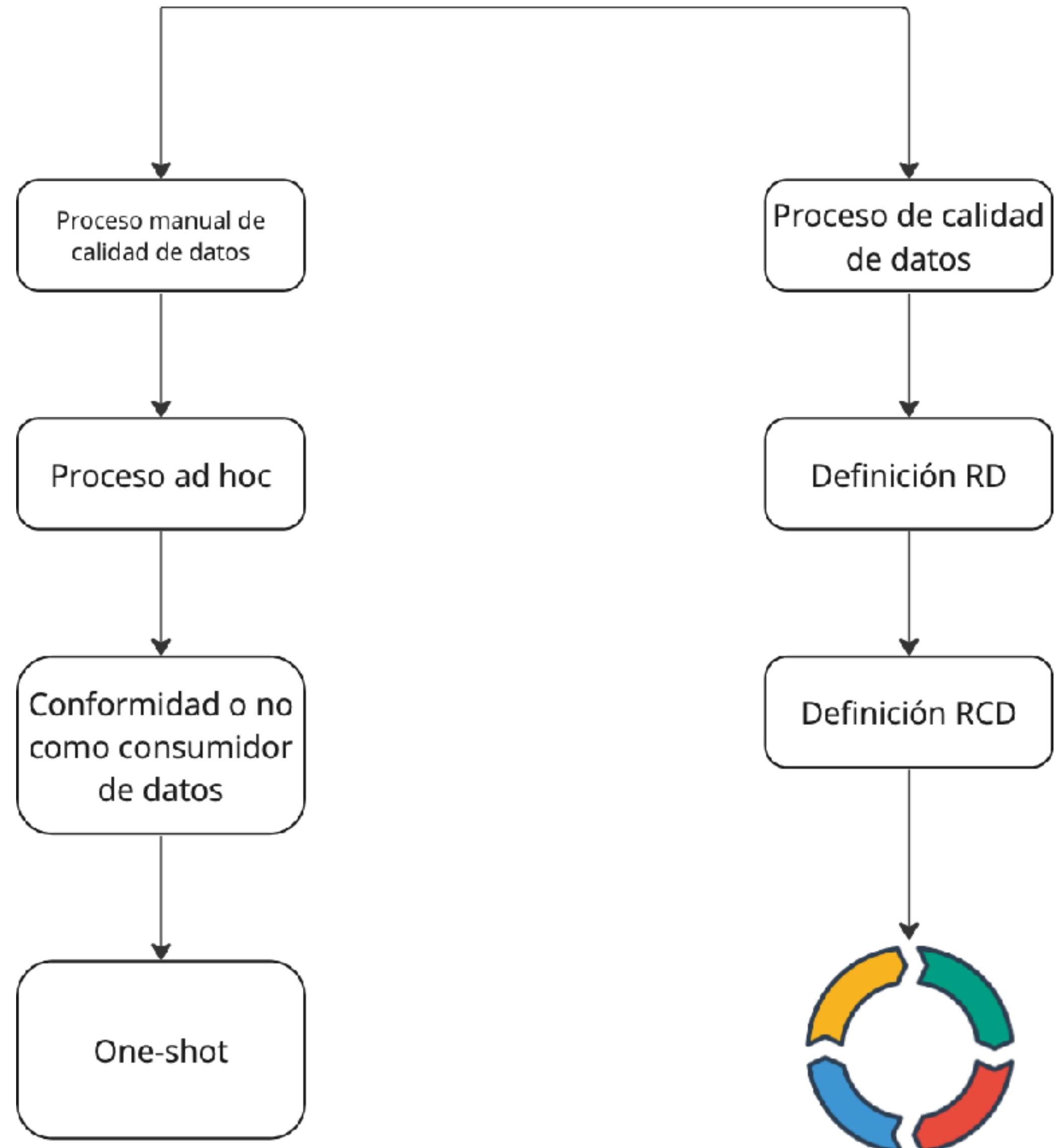
Ante la solicitud de identificar las tres barreras principales a la hora de acceder a los datos públicos para su reutilización por parte del Sector Privado son:

- La falta de actualización de los datos (49%)
- Los conjuntos de datos no están disponibles en todas las CCAA o en todos los Ayuntamientos. (46%)
- La dificultad en el acceso a los datos (43%)

¿Qué dice la norma?

Ciclo PDCA - ISO 8000:61

- Plan
- Do
- Check
- Act



Expectativas



`df.isnull().sum()`

`expect_column_values_to_not_be_null()`

Great Expectations

Descripción herramienta

Permite la generación de reglas de calidad de datos, llamadas expectativas, en lenguaje pseudo natural.

ej. `expect_columns_not_to_be_null()`

Ofrece funcionalidades para la gestión de la no conformidad de los datos, pero no para la reparación de los mismos.



Great Expectations



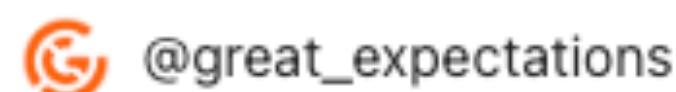
Expectativas al rescate

Librería Great Expectations

expect_column_values_to_not_be_null

- This expectation level is PRODUCTION

Contributors:



Tags:

core expectation

column map expectation

Metrics:

column_valuesnonnullunexpected_count

table.row_count

column_valuesnonnullunexpected_values

Backend support:

○ Pandas

○ Spark

○ SQLite

○ PostgreSQL

○ MySQL

○ MSSQL

○ Trino

○ Redshift

○ BigQuery

○ Snowflake

Alcance de la librería

Galería de expectativas - 324 en total

● [expect_column_distinct_values_to_be_continuous](#) (Contrib ColumnAggregateExpectation) 

Expect the set of distinct column values to be continuous.

Tags: core expectation, column aggregate ex... **Support:** 

Contribution status:

 Experimental	4 / 4
 Beta	3 / 3
 Production	1 / 2

[See More](#)

● [expect_column_distinct_values_to_be_in_set](#) (Core ColumnAggregateExpectation) 

Expect the set of distinct column values to be contained by a given set.

Tags: core expectation, column aggregate ex... **Support:**        +3

Contribution status:

 Experimental	4 / 4
 Beta	3 / 3
 Production	2 / 2

[See More](#)

● [expect_column_distinct_values_to_contain_set](#) (Core ColumnAggregateExpectation) 

Expect the set of distinct column values to contain a given set.

Tags: core expectation, column aggregate ex... **Support:**        +3

Contribution status:

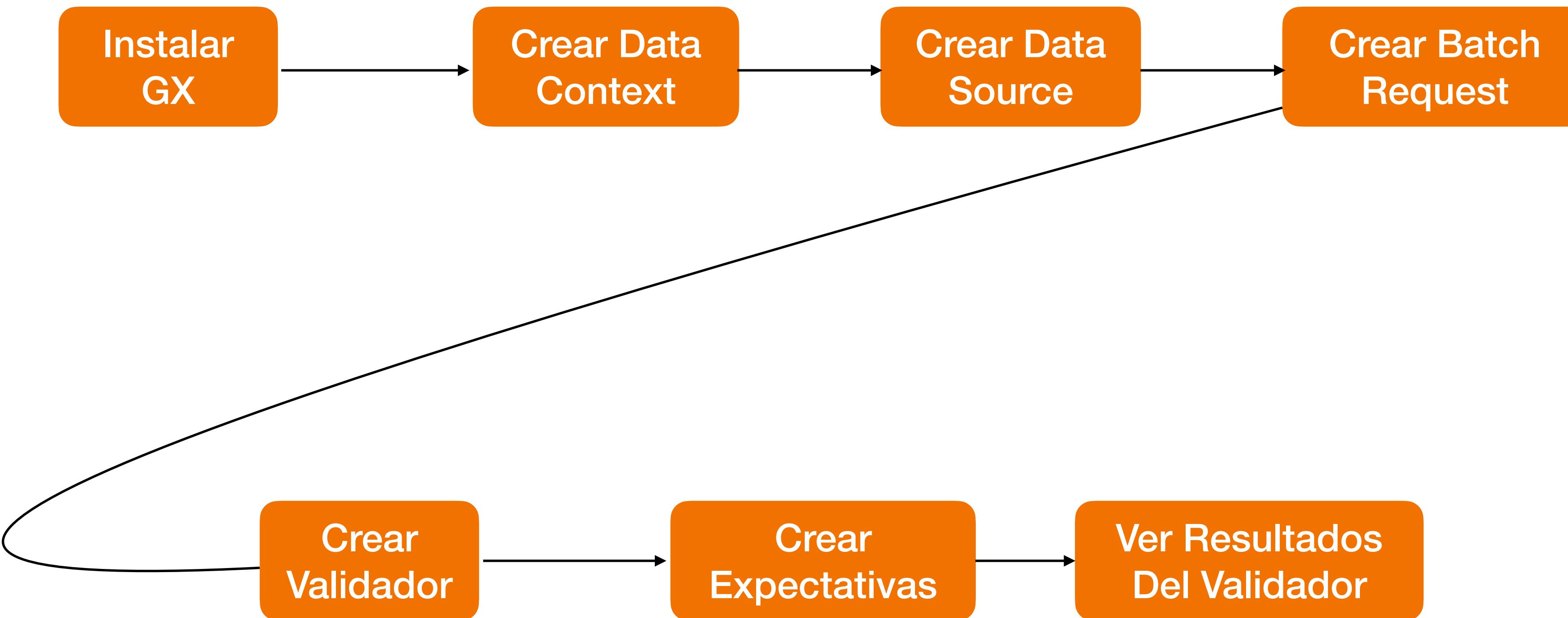
 Experimental	4 / 4
 Beta	3 / 3
 Production	2 / 2

[See More](#)

Alcance de la librería

Galería de expectativas - Multiples backends

Flujo de trabajo de validación de datos GX



Data Assistants

Version: 0.18.2

Data Assistants

Use the Onboarding Data Assistant to create Expectation Suites.



Create an Expectation Suite with the
Onboarding Data Assistant

Use the Onboarding Data Assistant to
Profile your data and automate the
generation of an Expectation Suite



Compare two tables with the Onboarding
Data Assistant

Use the Onboarding Data Assistant to create
an Expectation Suite that determines if two
tables are identical

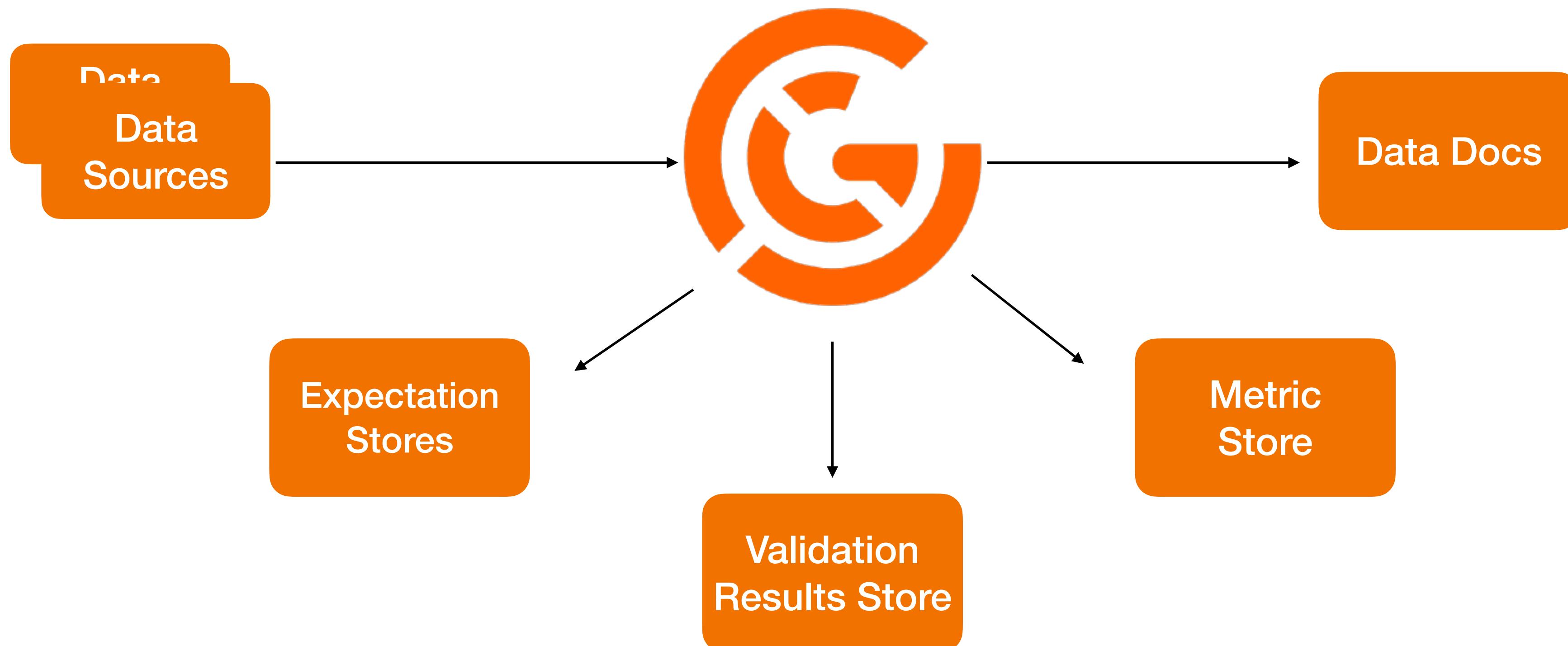


Create an Expectation Suite with the
Missingness Data Assistant

Use the Missingness Data Assistant to
Profile your data and automate the
generation of an Expectation Suite

Componentes de la solución

Almacenes



Ejemplo Código GX

Inicialización del contexto

```
import great_expectations as gx

context = gx.get_context()

datasource_name = "contratos_modificados_madrid"
path_to_folder_containing_csv_files = interim_output_folder

datasource = context.sources.add_pandas_filesystem(
    name=datasource_name, base_directory=path_to_folder_containing_csv_files
)

# Definimos originalmente el data_asset para el fichero 1 de 2022
contratos_2022_data_asset = datasource.add_csv_asset(
    name="contratos_2022_1",
    batching_regex="contratos_modificados_madrid_2022-1\.csv",
)

# Generamos un batch request sobre el data asset que sólo contiene la información de un único fichero (el primero de 2022)
contratos_2022_batch_request = contratos_2022_data_asset.build_batch_request()
```

Ejemplo Código GX

Definición suite expectativas

```
# Generamos una suite de expectativas sobre la que añadiremos de forma interactiva las expectativas que generemos
expectation_suite_name = "exp_suite_contratos_modificados_madrid_interativo"

expectation_suite = context.add_or_update_expectation_suite(
    expectation_suite_name=expectation_suite_name
)

# Creamos un validador para el batch request, pasándole la suite de expectativas que hemos creado

validator = context.get_validator(
    batch_request=contratos_2022_batch_request,
    expectation_suite_name=expectation_suite_name,
)
```

Ejemplo Código GX

Creación expectativas manualmente

```
# El campo ES_PYME debe contener los valores SI o NO
set_es_pyme = ['SI', 'NO']

validator.expect_column_values_to_be_in_set(column='ES_PYME', value_set=set_es_pyme, meta={"Dimension": "Accuracy", "QM": "Syntactic data accuracy"})

# Los valores del campo N. DE EXPEDIENTE deben comenzar por tres caracteres seguidos de una barra y una serie de caracteres alfanuméricos posteriores de longitud variable.
regex_list_expedientes = [
    '^\\d+\\/[a-zA-Z0-9/.-]+\\s*[a-zA-Z0-9/.-_]*$',      # Hace match a expresiones como '300/2019/00373', '711/2020/05971 L3-002'
    '^[A-Z]+\\/[a-zA-Z0-9/.-]+\\s*[a-zA-Z0-9/.-_]*$',    # Hace match a expresiones como 'MDC/2021/00080', 'PC-MER/2021/00011-ORD'
    '^[A-Z0-9-]+\\. [0-9.]+$',                            # Hace match a expresiones como 'PR22-0214.2'
]

validator.expect_column_values_to_match_regex(column='N. DE EXPEDIENTE', regex_list=regex_list_expedientes, meta={"Dimension": "Accuracy", "QM": "Semantic Accuracy"})
```

Ejemplo Código GX

Validación y resultados

```
# El campo Nº REGISTRO INCIDENCIA no puede estar en blanco
validator.expect_column_values_to_not_be_null(column='N. REGISTRO INCIDENCIA', meta={"Dimension": "Completeness", "QM": "Record completeness", "Priority": True})

# Definimos un checkpoint para recopilar todos los datos relacionados con la validación de las expectativas

checkpoint = gx.checkpoint.SimpleCheckpoint(
    name="checkpoint_2022_1",
    data_context=context,
    validations=[
        {
            "batch_request": contratos_2022_batch_request,
            "expectation_suite_name": "exp_suite_contratos_modificados_madrid_interativo",
        },
    ],
)

# Almacenamos la suite de expectativas en el validador
validator.save_expectation_suite(discard_failed_expectations=False)

# Ejecutamos el checkpoint para guardar los resultados de la validación
checkpoint_result = checkpoint.run()

# Generamos los datadocs correspondientes
context.build_data_docs()
```

Resultados - Data Docs

The screenshot shows the Great Expectations Data Docs interface. The top navigation bar includes the project name "great_expectations", a "Home" link, and a timestamp "2021-11-18T16:33:57Z". The main content area is titled "passenger_count".

Actions:

- Validation Filter:
 - Show All
 - Failed Only**
- [How to Edit This Suite](#)
- [Show Walkthrough](#)

Table of Contents:

- [Overview](#)
- [Table-Level Expectations](#)
- [passenger_count](#)

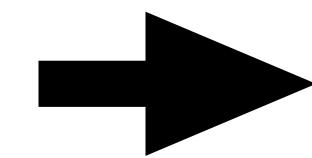
passenger_count

Status	Expectation	Observed Value																								
✓	minimum value must be greater than or equal to <code>1</code> and less than or equal to <code>1</code> .	1																								
✓	maximum value must be greater than or equal to <code>6</code> and less than or equal to <code>6</code> .	6																								
✓	mean must be greater than or equal to <code>1.5716</code> and less than or equal to <code>1.5716</code> .	1.5716																								
✓	median must be greater than or equal to <code>1.0</code> and less than or equal to <code>1.0</code> .	1																								
✓	quantiles must be within the following value ranges.	<table border="1"><thead><tr><th>Quantile</th><th>Min Value</th><th>Max Value</th><th>Quantile Value</th></tr></thead><tbody><tr><td>0.05</td><td>1</td><td>1</td><td>0.05 1</td></tr><tr><td>Q1</td><td>1</td><td>1</td><td>Q1 1</td></tr><tr><td>Median</td><td>1</td><td>1</td><td>Median 1</td></tr><tr><td>Q3</td><td>2</td><td>2</td><td>Q3 2</td></tr><tr><td>0.95</td><td>5</td><td>5</td><td>0.95 5</td></tr></tbody></table>	Quantile	Min Value	Max Value	Quantile Value	0.05	1	1	0.05 1	Q1	1	1	Q1 1	Median	1	1	Median 1	Q3	2	2	Q3 2	0.95	5	5	0.95 5
Quantile	Min Value	Max Value	Quantile Value																							
0.05	1	1	0.05 1																							
Q1	1	1	Q1 1																							
Median	1	1	Median 1																							
Q3	2	2	Q3 2																							
0.95	5	5	0.95 5																							
✓	values must belong to this set: <code>1 2 3 4 5 6</code> .	0% unexpected																								
✓	values must never be null.	100% not null																								
✓	fraction of unique values must be exactly <code>0.0006</code> .	0.0006																								
✓	value types must belong to this set: <code>INTEGER integer int int_ intB int16 int32 int64 uint8 uint16 uint32 uint64 INT INTEGER INT64 TINYINT BYTEINT SMALLINT BIGINT IntegerType LongType</code> .	int64																								

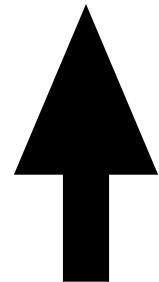
Flujo de trabajo de calidad de datos

Plan

Requisitos del dato

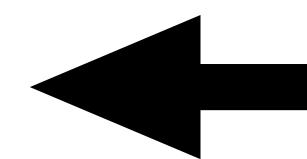


Requisitos calidad del dato

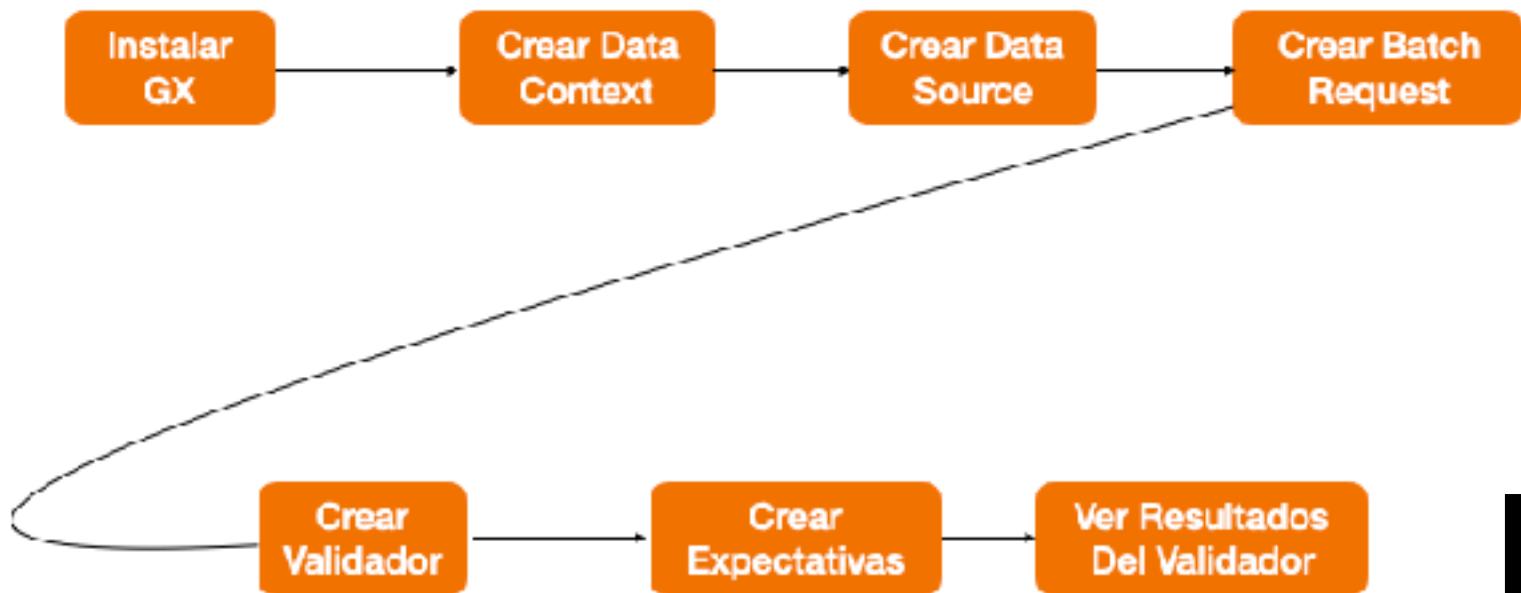


Act

Mejora de las reglas en base al feedback



Check



Do



Ventajas

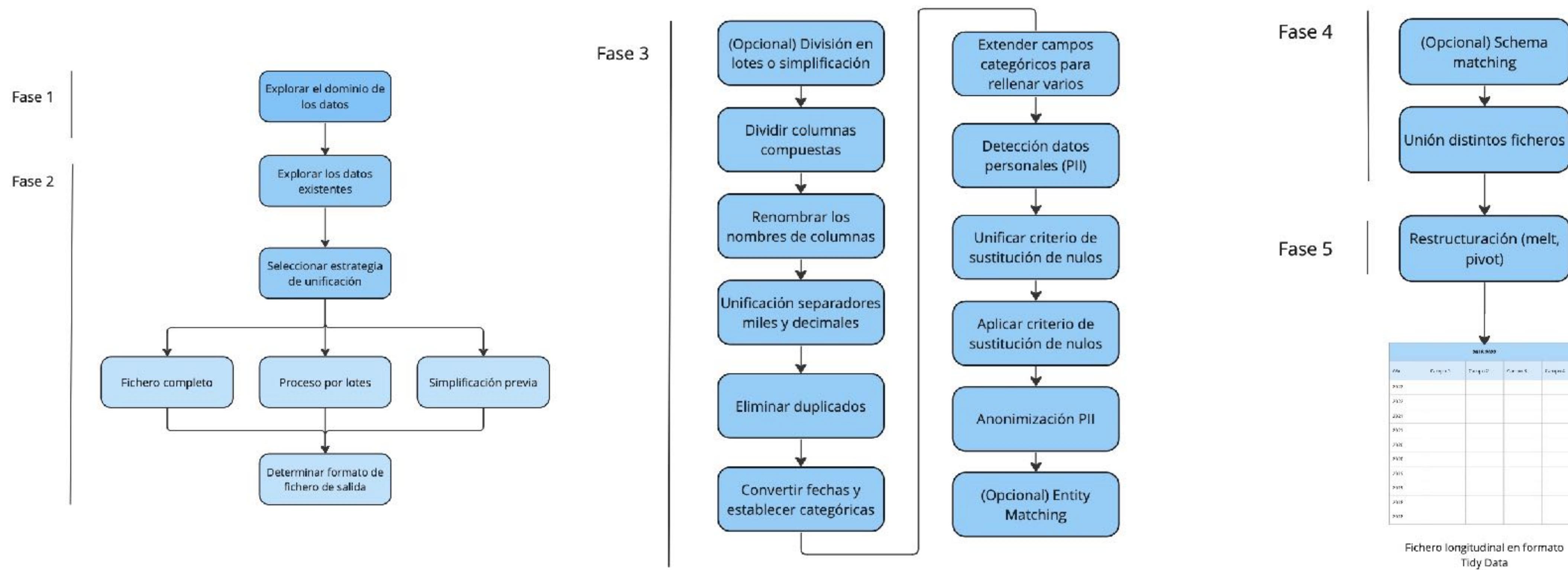
- Una sola función para validar contra múltiples backends: Pandas, Spark, SQLite, Postgresql, MySQL, MSSQL, Trino, Redshift, BigQuery, Snowflake
- Integraciones: Airflow, Databricks, Meltano, Datahub y la mayoría de servicios cloud.
- Legibilidad de las reglas
- Interpretabilidad de los resultados
- Solución Open Source
- GX Cloud para solución gestionada

Inconvenientes

- No es un entorno trivial de configurar
- Ha ido cambiando el formato de configuración con el tiempo
- Requiere de otros elementos para remediar los problemas de calidad del dato
- El perfilador automático añade una cantidad muy elevada de reglas
- No es fácil de utilizar por perfiles no técnicos

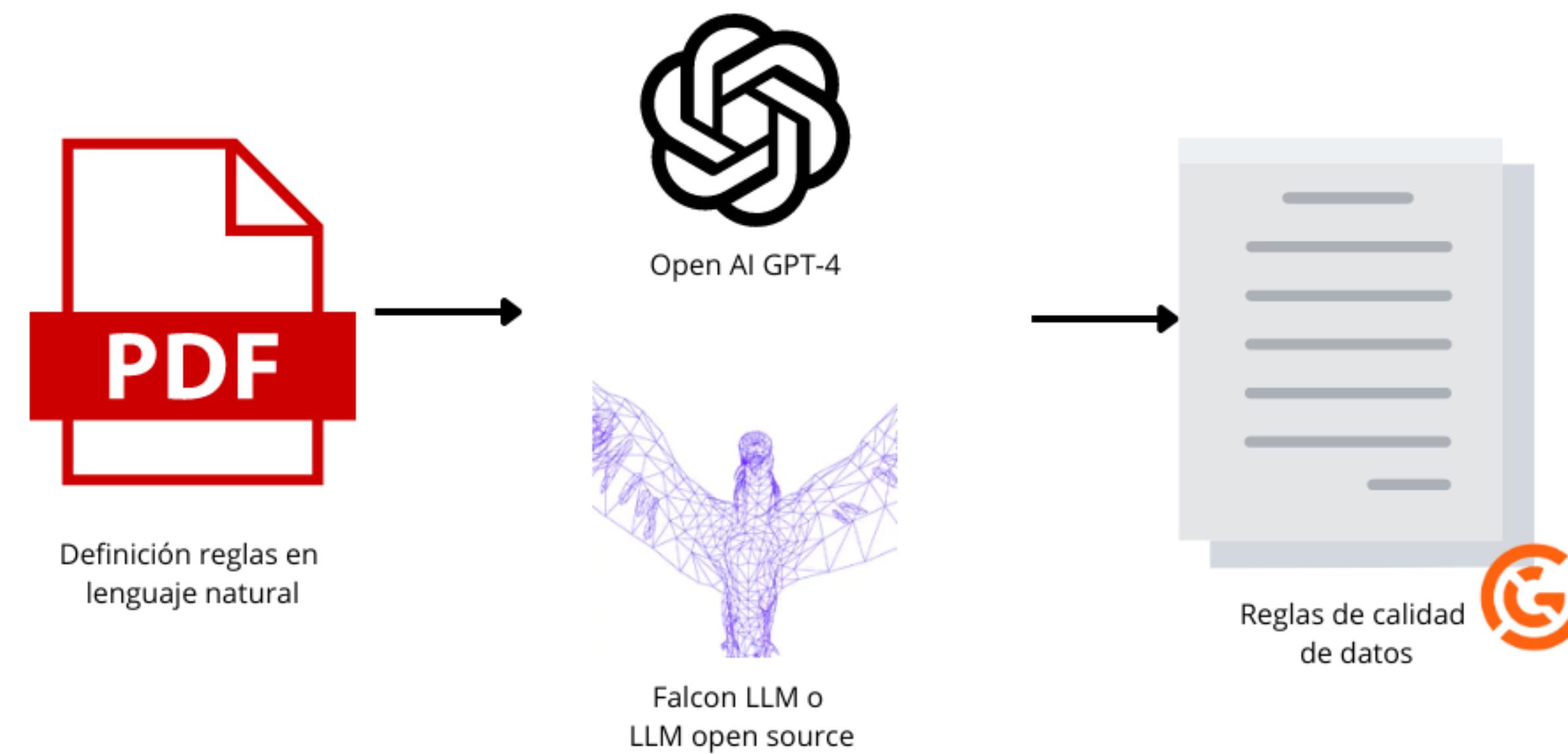
Metodología - Tidy Open Data

Creación datasets longitudinales



LLM para convertir texto en expectativas

Permite escalar la generación de reglas



PyData Global

6-8 Diciembre

- Tutorial: Improving Open Data Quality using Python
- Presentación aplicada sobre GX para limpieza de Open Data

**Calidad >
Cantidad**

Gracias por la atención

César García Sáez - @lahoramaker - cesar@lahoramaker.com