

Arquitectura para el descubrimiento y catalogación de datos con enriquecimiento semántico

Guido Velázquez Gonnet

Programa de Posgrado en Sistemas de Información y Tecnologías de Gestión de Datos
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Abril de 2026





Arquitectura para el descubrimiento y catalogación de datos con enriquecimiento semántico

Guido Velázquez Gonnet

Tesis de Maestría presentada al Programa de Posgrado en Sistemas de Información y Tecnologías de Gestión de Datos, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magíster en Sistemas de Información y Tecnologías de Gestión de Datos.

Director:
Dra. Adriana Marotta

Director académico:
Dra. Adriana Marotta

Montevideo – Uruguay
Abril de 2026

Velázquez Gonnet, Guido

Arquitectura para el descubrimiento y catalogación de datos con enriquecimiento semántico / Guido Velázquez Gonnet. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2026.

XIII, 94 p.: il.; 29, 7cm.

Director:

Adriana Marotta

Director académico:

Adriana Marotta

Tesis de Maestría – Universidad de la República, Programa en Sistemas de Información y Tecnologías de Gestión de Datos, 2026.

Referencias bibliográficas: p. 87 – 88.

1. metadatos, 2. catálogo de datos, 3. ontologías, 4. enriquecimiento semántico, 5. linaje de datos, 6. descubrimiento de datos, 7. DCAT, 8. gobierno de datos. I. Marotta, Adriana, . II. Universidad de la República, Programa de Posgrado en Sistemas de Información y Tecnologías de Gestión de Datos. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Dr. Raúl Ruggia

Dr. Guillermo Moncecchi

Dra. Edelweis Rohrer

Montevideo – Uruguay
Abril de 2026

Agradecimientos

La realización de este trabajo no hubiera sido posible sin el acompañamiento y el apoyo de quienes, de distintas maneras, formaron parte de este camino.

A mi esposa e hijos, por su paciencia, comprensión y afecto incondicional durante este proceso. Cada fin de semana postergado, cada plan familiar resignado y cada momento no compartido fue, generosamente, una forma de apoyo. Este logro es también de ellos.

A mi tutora y directora académica, Adriana, por sus consejos atinados y oportunos, su disponibilidad y su capacidad para identificar el camino cuando el rumbo no era claro. Su respaldo académico y personal fue un sostén fundamental a lo largo de todo el trabajo.

A los analistas del Banco Central del Uruguay, Alejandro, Cecilia, Mariana y Wendy, por su tiempo, predisposición y generosidad al compartir su conocimiento del dominio en las distintas instancias del trabajo. Su mirada desde la práctica fue indispensable tanto para comprender la problemática abordada como para validar los diseños y soluciones propuestos. Un agradecimiento especial a Carlos, actor clave desde el primer momento, cuando este trabajo no era más que una idea, por su apoyo, su confianza y su impulso para que tomara forma.

A todos ellos, mi sincero agradecimiento.

Declaración de uso de IA

Durante la elaboración del presente trabajo, el autor utilizó herramientas de inteligencia artificial generativa, en particular ChatGPT (OpenAI) y Claude (Anthropic), con el fin de realizar correcciones gramaticales y ortográficas, así como para mejorar el estilo de escritura, entre otros usos. Tras utilizar estas herramientas, el autor revisó y editó el contenido según fue necesario y asume plena responsabilidad por el contenido del presente documento.

RESUMEN

Las organizaciones contemporáneas gestionan volúmenes crecientes de datos distribuidos en múltiples sistemas, plataformas y áreas funcionales. En este contexto, los usuarios con conocimiento del dominio, analistas y expertos de negocio, enfrentan dificultades para identificar y comprender los activos de datos disponibles, dado que las herramientas de catalogación orientadas a metadatos técnicos no ofrecen mecanismos adecuados para vincular los datos físicos con los conceptos propios del dominio de negocio.

El presente trabajo propone una arquitectura de tres niveles para el descubrimiento y catalogación de datos con enriquecimiento semántico, aplicada al Banco Central del Uruguay (BCU) como caso de estudio. La arquitectura establece un puente entre la representación conceptual del conocimiento del dominio y los elementos físicos que lo sustentan, organizando los activos de información en un nivel conceptual, un nivel de metadatos y un nivel físico. Como parte del trabajo se realizó un análisis comparativo de herramientas de catalogación de datos disponibles en el mercado, en función del contexto institucional y tecnológico del BCU.

La propuesta fue validada mediante una prueba de concepto aplicada a un caso de estudio real en el Banco, cuyos resultados permitieron constatar la viabilidad del enfoque para facilitar el descubrimiento de datos por parte de usuarios no técnicos. Las principales conclusiones del trabajo destacan el rol de las ontologías como mecanismo de representación semántica y la pertinencia de estándares abiertos para la catalogación de datos, como es el caso del vocabulario Data Catalog Vocabulary (DCAT). Asimismo, se identifican desafíos y líneas de trabajo futuro orientadas a consolidar y escalar la propuesta en el contexto organizacional.

Palabras claves:
metadatos, catálogo de datos, ontologías, enriquecimiento semántico, linaje de datos, descubrimiento de datos, DCAT, gobierno de datos.

Lista de figuras

1.1 Organización de los datos en el Banco 5

3.1 Representación conceptual del ecosistema de datos organizacional 41
3.2 Niveles de abstracción 44
3.3 Vinculación entre ontologías y nivel físico 48

4.1 Esquema en los tres niveles 56
4.2 Ontología de dominio 57
4.3 Ontología de catálogo 58
4.4 Nivel Metadatos 60
4.5 Captura de OpenMetadata 61
4.6 Aplicación de DCAT 71
4.7 Flujo de procesamiento 73
4.8 Ejemplo dataset y sus distribuciones 78
4.9 Ejemplo linaje de datasets 79

Lista de tablas

2.1 Cuadro comparativo de herramientas 39

4.1 Vinculaciones entre los diferentes niveles 63
4.2 Conversión a sentencia turtle 74

Índice de fragmentos

1 Ejemplo de dataset definido en base a DCAT 70
2 Consulta SPARQL 76
3 Invocación a script 77

4 Ejemplo metadata cubo Datos de Fuente 91
5 Ejemplo de metadata cubo Dimensiones Públicas 94

Lista de siglas

Lista de siglas

BCU Banco Central del Uruguay vii, 3

DCAT Data Catalog Vocabulary vii, 65, 66, 67, 68, 69

IA Inteligencia Artificial vi

IRI Internationalized Resource Identifier 85

RDF Resource Description Framework 65, 67, 70, 71

W3C World Wide Web Consortium 54, 65

Tabla de contenidos

Lista de figuras VIII

Lista de tablas IX

Lista de siglas XI

1 Introducción 1

1.1 Motivación . . . 1
1.2 Objetivos . . . 2
1.3 Contribuciones . . . 3
1.4 Ámbito de aplicación . . . 3
1.5 Organización del documento . . . 7

2 Estado del Arte 8

2.1 Trabajos Académicos . . . 9
2.1.1 Trabajos de evaluación de herramientas . . . 9
2.1.2 Enriquecimiento con conocimiento del negocio . . . 11
2.2 Herramientas existentes . . . 14
2.2.1 Criterios de análisis de las herramientas . . . 16
2.2.2 OpenMetadata . . . 20
2.2.3 DataHub . . . 24
2.2.4 Microsoft Purview . . . 28
2.2.5 Collibra . . . 33
2.2.6 Comparación de herramientas . . . 37

3 Análisis y diseño de la solución propuesta 40

3.1 Problemática actual . . . 40
3.2 Linaje de los datos . . . 42
3.3 Propuesta de solución . . . 43

3.3.1 Nivel conceptual 44
3.3.2 Nivel de metadatos 46
3.3.3 Nivel físico 47
3.3.4 Modelo de vinculaciones 47

4 Prueba de concepto **50**
4.1 Caso de estudio y contexto institucional 50
4.1.1 Relevamiento y datos de interés 51
4.1.2 Necesidades relevadas 52
4.1.3 Selección del caso de estudio 53
4.1.4 Catálogo de datos desarrollado por el Banco 54
4.2 Implementación del caso de estudio 55
4.2.1 Arquitectura y componentes 56
4.2.2 Ontología de Catálogo 64
4.2.3 Análisis de resultados 79

5 Conclusiones **82**

Referencias bibliográficas **87**

Glosario **88**

Anexos **89**
Anexo 1 Descubrimiento de linaje 90

Capítulo 1

Introducción

«Los datos son el petróleo del siglo XXI». Esta afirmación, atribuida a Clive Humby (2006), sintetiza la importancia que los datos han adquirido en las organizaciones actuales. Autores posteriores han profundizado en esta analogía al enfatizar que el valor no reside únicamente en la posesión de datos, sino en la capacidad organizacional para explotarlos y transformarlos en conocimiento.[1] Del mismo modo que el petróleo requiere procesos de extracción y refinamiento para convertirse en combustible útil, los datos sin gestionarlos adecuadamente carecen de valor en términos de analítica y toma de decisiones informadas.

1.1. Motivación

En el marco del avance tecnológico y de los procesos de transformación digital, las organizaciones experimentan un crecimiento sostenido en el volumen y la complejidad de los datos que generan y almacenan. En numerosos casos, estos datos se gestionan de manera fragmentada a través de múltiples sistemas, plataformas y áreas funcionales, lo que dificulta su identificación, integración y aprovechamiento.

Frente a esta diversidad de fuentes y activos de información, surge la necesidad de organizarlos y catalogarlos de forma sistemática, de modo que usuarios no técnicos, pero con conocimiento experto del dominio, puedan utilizarlos con distintos propósitos analíticos y de integración. En este marco, se espera que los analistas dispongan de una plataforma que les permita navegar por los principales conceptos de negocio vinculados a su área, profundizar en su signi-

[1] https://www.linkedin.com/pulse/data-new-oil-llew-adamson/

ficado y de qué elementos se componen, y eventualmente acceder a consultas sobre los datos físicos que los sustentan o, al menos, entender su disposición y localización. De este modo, se favorecería el consumo de los datos y su integración con otros activos, especialmente en escenarios de análisis de información de carácter *ad hoc*.

No obstante, la revisión de la literatura evidencia que aún resulta incipiente la formulación de arquitecturas o metodologías que satisfagan plenamente esta demanda. En particular, no se identifican propuestas que ofrezcan a usuarios no técnicos un mecanismo accesible para vincular los principales conceptos de negocio con los elementos físicos que soportan los datos y su almacenamiento.

En este sentido, estudios recientes señalan que las herramientas de catálogo de datos[1] deberían, idealmente, abarcar la totalidad de los aspectos vinculados a la gestión de metadatos, desde la recopilación automatizada de metadatos técnicos y el modelado de datos para profesionales de TI, hasta la provisión de modelos de datos de alto nivel orientados al negocio para expertos del dominio. Sin embargo, muchas de las herramientas de catálogo de datos más avanzadas no implementan plenamente este amplio conjunto de funcionalidades, especialmente en lo que respecta a la expresividad de la semántica de los datos. [(9), Schrott].

1.2. Objetivos

El presente trabajo tiene como objetivo facilitar la exploración y el análisis autónomo de los datos por parte de usuarios no técnicos, promoviendo su valorización y aprovechamiento efectivo como activos estratégicos de la organización.

Para alcanzar este objetivo, el trabajo se propone los siguientes objetivos específicos:

- Definir una arquitectura que ordene y estructure los metadatos en el ámbito de una organización[2].

[1] Definición de catálogo de datos según IBM: ≪Un inventario con capacidad de consultar los activos de datos dentro de una organización que proporciona un entendimiento común y facilita el gobierno de datos. Permite a los usuarios explorar y localizar activos de datos relevantes en función de diversos criterios, tales como tipos de datos, atributos y glosarios de negocio.≫ [(1), Boufassil]

[2] Más adelante se describe la organización que constituye el ámbito de aplicación del presente trabajo.

- Incrementar la autonomía de los usuarios no técnicos en el descubrimiento, exploración y experimentación con nuevos activos de información.
- Establecer un marco de trabajo, junto con una herramienta de soporte, que sustente la implementación y evolución de un catálogo de datos.
- Proponer un modelo conceptual que represente las interrelaciones entre los datos y sus distintos niveles de agregación.

1.3. Contribuciones

Para alcanzar el objetivo propuesto, se analiza el estado del arte en herramientas de descubrimiento y catalogación de datos, junto con la literatura especializada sobre mecanismos para describir, documentar y enriquecer los datos existentes, con especial énfasis en su explotación analítica.

Sobre la base de estos conceptos, se propone una arquitectura de tres niveles para la descripción de los activos de información, incorporando ontologías como un lenguaje de mayor expresividad semántica para su representación y explicación. Finalmente, se desarrolla una prueba de concepto aplicada a un caso de estudio y se presenta una implementación en uno de los ámbitos de aplicación considerados, evidenciando la viabilidad de visibilizar los datos existentes y de identificar potenciales usos analíticos sobre los mismos.

Las contribuciones principales de este trabajo son las siguientes:

- Un análisis de las herramientas de descubrimiento y catalogación de datos existentes en el mercado, basado en criterios bien definidos.
- Una arquitectura de tres niveles para la descripción de los activos de información, la cual incorpora ontologías para organizar la información y dar semántica a los datos involucrados.
- Aplicación de la arquitectura propuesta a un caso práctico real.

1.4. Ámbito de aplicación

El presente trabajo se desarrolla en el ámbito del Banco Central del Uruguay (BCU) y surge de la necesidad de identificar, describir y documentar los activos de datos que la institución utiliza con fines analíticos y para la toma

de decisiones, considerados como recursos estratégicos para el cumplimiento de sus cometidos.

De acuerdo con su carta orgánica[1] , al Banco Central del Uruguay le corresponden múltiples funciones y responsabilidades, entre las que se destacan la preservación de la estabilidad de precios, la regulación y supervisión del sistema financiero y del sistema de pagos, así como el control del lavado de activos y la prevención del financiamiento del terrorismo en el ámbito financiero, entre otras.

En este marco, suele afirmarse que dentro del Banco coexisten múltiples «bancos» o líneas de negocio, lo cual refleja la diversidad de cometidos que ejerce la institución y la heterogeneidad de los procesos analíticos asociados a cada uno de ellos.

Desde el punto de vista organizacional, el Banco se estructura en varias áreas o líneas de reporte, cada una con objetivos específicos, tales como Política Monetaria, Superintendencia de Servicios Financieros, Asesoría Económica, Sistema de Pagos, Servicios Institucionales y Auditoría, entre otras. Desde el punto de vista de tecnología, el Banco se estructura de la siguiente forma. Los servicios de infraestructura tecnológica, por ejemplo, la provisión de servidores de aplicaciones y bases de datos, son compartidos transversalmente por todas las áreas. En contraste, las actividades de mantenimiento y desarrollo de aplicaciones se organizan en distintos departamentos, cada uno con su propio equipo y responsables, y alineados a áreas de negocio específicas, como Política Monetaria, Superintendencia de Servicios Financieros o Servicios Institucionales.

Esta estructura favorece la autonomía de las distintas líneas de negocio, permitiendo que cada área defina sus prioridades y enfoques analíticos. Así, algunas áreas se orientan a análisis detallados con foco en la entidad o sector correspondiente, por ejemplo sobre balances o instituciones individuales, mientras que otras priorizan análisis de tipo macro-económico, centrados en el desempeño agregado del sistema financiero y de la economía del país en su conjunto. Como consecuencia, cada área junto a su departamento de tecnología construyen sus propias bases de datos relacionales y, en muchos casos, desarrollan sus propios *data warehouses*, lo que hace necesario el intercambio de información entre áreas, bajo políticas específicas de acceso, uso y gobierno

[1] https://www.bcu.gub.uy/Acerca-de-BCU/Normativa/Documents/Carta-organica/carta_organica_toco.pdf

de los datos.

En la figura 1.1 se presenta una visión de alto nivel de la organización, los flujos y el intercambio de datos dentro del Banco Central del Uruguay. En el sector izquierdo de la figura se representa la recepción de información proveniente de entidades externas, tanto del sector financiero, en cumplimiento de la normativa vigente, como de otros organismos públicos, en el marco de convenios formales de intercambio de datos.

En este contexto, se destaca explícitamente el uso de Microsoft Excel, cuyo logotipo se ubica de forma específica a la altura de las bases de datos. Esta elección responde a su utilización intensiva como formato de recepción de información, y a que, en numerosos casos, los datos no solo son recibidos sino también almacenados y consolidados directamente en archivos de este tipo. Dichos archivos constituyen, en varias situaciones, fuentes primarias de información que alimentan periódicamente a sistemas convencionales de la institución, por ejemplo en procesos de carga mensual.

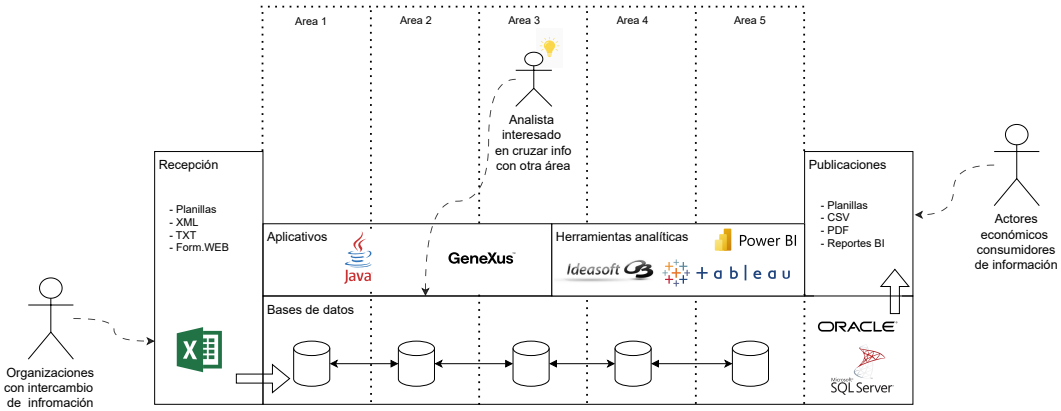


Figura 1.1: Organización de los datos en el Banco

En el extremo derecho de la figura se representa la fase final del ciclo de vida de determinados conjuntos de datos, los cuales, una vez procesados, validados y analizados, son difundidos en diversos formatos, incluidos nuevamente archivos Excel, como parte de las publicaciones periódicas de la institución. Estos conjuntos de datos se destinan a una diversidad de actores económicos, tanto del ámbito nacional como a organismos internacionales, contribuyendo de este modo al cumplimiento de los cometidos del Banco en materia de provisión de información estadística, rendición de cuentas y transparencia hacia los mercados.

En la zona central de la figura se representan los procesos mediante los cuales los datos son gestionados y generados por aplicaciones del propio Banco, ya sea a partir de la transformación de datos preexistentes, del ingreso de información por parte de los usuarios conforme a los procesos institucionales, o de la interacción con sistemas externos, como ocurre, por ejemplo, con las liquidaciones del sistema de pagos provenientes del resto del sistema financiero. Estas aplicaciones incluyen tanto desarrollos *in-house*, implementados principalmente mediante tecnologías como *Java* y *GeneXus*, como soluciones adquiridas a terceros, que son implantadas y mantenidas por la institución en conjunto con el proveedor correspondiente.

El recuadro de herramientas analíticas agrupa las soluciones de Business Intelligence, las cuales desempeñan un rol importante tanto en el análisis como en la publicación de información. *Ideasoft O3* se encuentra instalada en el Banco desde hace varios lustros y cuenta con múltiples cubos desarrollados, pertenecientes a diversas áreas y alimentados por varias bases de datos. En un período más reciente se incorporó *Tableau*, comprendiendo el desarrollo de tableros de visualización, y posteriormente *Power BI*, como la solución de adopción más actual. Cabe destacar que estas herramientas presentan distintos niveles de complejidad técnica para la construcción de cubos analíticos o tableros: desde Ideasoft O3, que demanda mayores conocimientos técnicos, hasta Power BI, que ofrece una experiencia de uso más intuitiva y accesible para usuarios no especializados. En cuanto a los procesos de transformación de datos (*ETL*) y a la implementación de lógica en la capa de datos, existe un uso extendido de desarrollos en *PL/SQL*.

Tal como se ilustra en las líneas verticales de la figura, las aplicaciones y las bases de datos se encuentran, en general, asociadas a áreas específicas del Banco. Estas aplicaciones operan mayoritariamente sobre datos estructurados, almacenados en bases de datos relacionales, y suelen ofrecer como funcionalidad la exportación de información a planillas Excel u otros formatos de datos semiestructurados. Los motores de bases de datos utilizados son principalmente *Oracle* y, en menor medida, *Microsoft SQL Server*.

En los últimos años, y siempre que resulta viable desde el punto de vista operativo y normativo, se ha habilitado el acceso de solo lectura a bases de datos relacionales para determinados usuarios analistas. Dichos accesos son nominativos y se conceden sobre bases de datos y tablas específicas, con el objetivo de facilitar tareas de análisis, como el desarrollo de reportes en he-

rramientas de *Business Intelligence*, por ejemplo *Microsoft Power BI*, o la automatización de salidas de información mediante lenguajes de *scripting*, como *R*. Actualmente, este tipo de acceso se otorga principalmente a las unidades responsables de los datos, cuyos analistas, si bien no conocen en detalle la implementación del modelo físico, poseen un conocimiento funcional suficiente para identificar la información de interés con apoyo limitado. No obstante, cuando el análisis requiere el uso de datos pertenecientes a otras áreas o unidades, surge la necesidad de facilitar la comprensión de los datos y de su modelo físico a usuarios que no disponen de documentación adecuada sobre los activos a los que acceden. Esta situación se ilustra en el caso de uso representado en la parte superior de la figura, donde un analista solicita acceso a datos de otra área con el fin de integrarlos o cruzarlos con información de su propia unidad o equipo de trabajo.

1.5. Organización del documento

El presente trabajo consta de cinco capítulos.

En el Capítulo 1, *Introducción*, se presentan la motivación, el contexto general y necesidades identificadas, más los objetivos del trabajo.

El Capítulo 2, *Estado del Arte*, hace una revisión de trabajos de la literatura actual relacionados a la catalogación de datos, posteriormente se evalúa la propuesta de la industria sobre cuatro soluciones seleccionadas y en base a criterios predefinidos.

En el Capítulo 3, *Análisis y diseño de la solución propuesta*, se ilustra la problemática actual y se desarrolla la arquitectura conceptual planteada, proponiendo dos modelos ontológicos para describir los datos, se detalla cada capa de la arquitectura y los vínculos entre las mismas.

El Capítulo 4, *Prueba de concepto*, describe la implementación aplicada a un caso de estudio y modelo seleccionado, incluyendo las decisiones de modelado, adopción de un estándar, las herramientas utilizadas y los resultados obtenidos.

Finalmente, el Capítulo 5, *Consideraciones finales*, presenta las conclusiones del trabajo, analiza sus aportes y limitaciones, y propone posibles líneas de trabajo futuro.

Capítulo 2

Estado del Arte

En el marco de este trabajo se analizaron diversos temas vinculados a la problemática de la gestión de datos orientada al análisis, incluyendo aportes relacionados con metadatos [(8), RavatZhao], [(7), Oukhouya, Lamya], modelo de metadatos *analysis-oriented* [(11), Zhao], soluciones de *data lake*[1], técnicas de *data discovery* [(5), Koutras], entre otros. No obstante, el estado del arte desarrollado se focaliza principalmente en el concepto de *catálogo de datos*, por tratarse del elemento más directamente vinculado tanto con el problema abordado como con las soluciones propuestas en esta tesis.

La catalogación es un concepto que se remonta a la Antigüedad, cuando bibliotecas como la de Alejandría ya utilizaban tablillas de arcilla para organizar sus colecciones. En bibliotecas de gran tamaño, la catalogación es esencial porque permite una organización y recuperación eficientes de los materiales. En la literatura existen múltiples intentos de definir el término "catálogo de datos"; sin embargo, no hay una definición ni una interpretación generalmente aceptadas sobre la funcionalidad requerida de las herramientas de catálogo de datos [(6), Kropshofer]. En el marco de este trabajo, adoptamos la definición de Gartner, según la cual un catálogo de datos es "Una colección de metadatos acerca de los activos de datos dentro de una organización, incluyendo bases de datos, tablas, archivos, informes y otros artefactos de datos. Su propósito es facilitar el descubrimiento de datos, entender su linaje y habilitar la analítica de autoservicio" [(1), Boufassil]. También a Gartner pertenece la siguiente afirmación: 'Son los metadatos los que convierten la información en un activo.' [2]

[1] https://www.amazon.com/Building-Data-Lakehouse-Bill-Inmon/dp/1634629663
[2] https://www.gartner.com/en/information-technology/glossary/metadata

2.1. Trabajos Académicos

En la literatura, el término catálogo de datos fue mencionado por primera vez por Franklin en 2005 dentro del contexto del concepto de dataspace. En sus inicios, el enfoque principal era la funcionalidad de catalogación propiamente dicha. Sin embargo, con el tiempo, se fueron incorporando otras capacidades, como la gestión de flujos de datos (data pipelines), el linaje de datos y funcionalidades de gobernanza. A pesar de la creciente popularidad de los catálogos de datos en el ámbito de las organizaciones, la literatura al respecto sigue siendo limitada. [(6), Kropshofer]

2.1.1. Trabajos de evaluación de herramientas

Las investigaciones orientadas a evaluar las funcionalidades de los catálogos de datos y a comparar las soluciones existentes son relativamente recientes en la comunidad científica. Actualmente no existe una aceptación generalizada sobre la definición o interpretación de las funcionalidades requeridas para una herramienta de catálogo de datos, lo cual provoca que no solo exista un gran número de herramientas llamadas de catálogo de datos sino también sea difícil para los profesionales de las organizaciones realizar un comparativo por funcionalidad. Un estudio reciente que se centra en las funcionalidades es [(6), Kropshofer], donde las principales contribuciones del estudio son (i) análisis de las funcionalidades más importantes de un catálogo de datos y (ii) evaluación sistemática de en qué medida las principales funcionalidades son implementadas por las herramientas de catálogo de datos actuales.

[(6), Kropshofer] sostiene que una herramienta de catálogo de datos ofrece al menos dos funcionalidades principales: (i) la recopilación y el almacenamiento de metadatos técnicos y (ii) la anotación de estos metadatos técnicos con conocimiento del dominio (contexto de negocio). El trabajo menciona que existe un solo estudio similar, con revisión por pares, que compara herramientas de catálogo de datos: [(4), JahnkeOtto]. Ambos estudios se complementan en términos de análisis de herramientas; mientras este último se centra en aspectos de integración de herramientas de gestión de metadata en la organización, [(6), Kropshofer] se focaliza en la capacidad de las herramientas para modelar y almacenar metadatos internamente, lo cual es particularmente importante para re-usar metadata en la organización, por ejemplo para chequeo de calidad de datos. Una valiosa contribución de [(6), Kropshofer] es la presentación de

los resultados en un formato tabular, que permite ordenar y filtrar por las diferentes herramientas y funcionalidades, habilitando a seleccionar la herramienta más adecuada según los requerimientos.

En la segunda mitad del año 2025 se publicó un nuevo estudio de análisis de herramientas de catálogo de datos: [(10), Tonnarelli]. El mismo evalúa 24 características (atómicas) en 39 herramientas a partir de un análisis sistemático de 86 fuentes, tanto de la literatura de la industria como de la literatura académica. Las 24 características atómicas o granulares las clasifica en cinco características de alto nivel que reflejan las etapas por las que pasan los metadatos en un entorno de catálogo de datos. Estas características de alto nivel son (i) Extracción o recolección de los metadatos, (ii) Enriquecimiento de metadatos, (iii) Uso de metadatos, (iv) Gobernanza de datos y (v) Colaboración. Esta última refiere a la interacción entre usuarios a través de la herramienta aportando evaluaciones y comentarios sobre los activos de datos. A su vez el estudio propone un modelo de madurez dividido en tres niveles , a) Nivel 1, como el nivel más básico, b) Nivel 2 y finalmente, c) Nivel 3, como el nivel más evolucionado, donde las herramientas de catálogo a partir de los datos de la organización crean un valor estratégico para el negocio.

A cada una de las 24 características atómicas le asigna uno de los tres niveles de madurez, y finalmente también a cada herramienta le asigna un nivel de madurez. De las 39 herramientas, 6 herramientas están alineadas al nivel 3 (superior) de madurez. El estudio aclara que los niveles de madurez no buscan medir confiabilidad ni adopción de las herramientas, sino alinearlas con las necesidades modernas de metadatos.

El estudio aunque es muy exhaustivo y permite comparar las herramientas desde diferentes características, gracias a la publicación de sus resultados en formato tabular, se encuentra desactualizado en algunas consideraciones de herramientas que se detectaron. El estudio le asigna a dos herramientas, CKAN y Truedat, el nivel 3 de madurez como herramientas open source, sin embargo Truedat ha cambiado su modelo de negocio y solo mantiene su núcleo como open source.[1] En el caso del nivel 3 de madurez, el mismo está compuesto por 5 características atómicas, (1) 'Data Product', (2) 'Data Quality Assessment', (3) 'Data Risk', (4) 'Texting' y (5) 'Sharing'. La primera está clasificada como característica de alto nivel (iii) Uso de metadatos, la segunda y tercera como (iv) Gobernanza de datos, y finalmente la cuarta y quinta característi-

[1] https://github.com/Bluetab/td-dd

cas atómicas clasificadas como (v) Colaboración. Sin embargo en la actualidad, desde mediados del año 2025, la herramienta Openmetadata tiene soporte tanto para la colaboración entre usuarios como para la característica de producto de datos, con lo cual se alinearía más al nivel 3 de madurez, a diferencia del nivel 2, que es el asignado por el estudio.

2.1.2. Enriquecimiento con conocimiento del negocio

Las herramientas de catálogo recolectan una gran cantidad de metadatos técnicos, en función de la infraestructura tecnológica disponible en la organización y del grado de soporte que cada herramienta ofrece para los distintos componentes de dicha infraestructura. Entre los elementos recolectados con mayor frecuencia se encuentran, por ejemplo, los metadatos asociados a tablas y campos de bases de datos relacionales, donde suele residir una parte sustancial de los datos organizacionales. Estos elementos existen en gran volumen dentro de las organizaciones y, en general, carecen de documentación que describa su propósito o significado desde una perspectiva de negocio.

En este contexto, para que los metadatos técnicos aporten valor a usuarios, como científicos de datos, analistas u otros perfiles orientados al análisis de información, resulta necesario enriquecerlos. Este enriquecimiento implica, en primer lugar, identificar y diferenciar cuáles de los elementos físicos poseen un valor efectivo para el análisis. En segundo lugar, supone vincular dichos elementos con términos y conceptos del negocio que les otorguen significado, los cuales forman parte de los procesos y objetivos de la organización. De este modo, el catálogo de datos se convierte de un inventario puramente técnico a una herramienta que facilita la interpretación, la exploración y el uso efectivo de los datos.

2.1.2.1. Soporte para documentación, etiquetado, clasificación y glosario de términos

En la valoración de herramientas respecto de su capacidad para representar el conocimiento del negocio en un catálogo de datos, la evidencia se concentra en el soporte para (i) documentación, (ii) etiquetado y/o clasificación y (iii) glosarios de términos. La literatura conceptualiza esta representación como enriquecimiento de metadatos [(10), Tonnarelli] o como anotación de metadatos técnicos con conocimiento del dominio (contexto de negocio) [(6), Kropshofer],

quien también evalúa el soporte de ontologías.

Según [(6), Kropshofer], la mayoría de las herramientas analizadas utilizan conceptos de expresividad limitada para modelar dicho conocimiento. En consecuencia, las evaluaciones comparativas suelen calificar a las herramientas según sus capacidades para agregar documentación y enriquecer los activos mediante etiquetas, clasificaciones, glosarios y, en menor medida, ontologías.

El estudio de [(2), Ehrlinger] también subraya la importancia de integrar los conceptos de negocio en los catálogos de datos. A través de una revisión sistemática de la literatura académica, los autores identifican el contexto de negocio como uno de los cuatro componentes a evaluar en las herramientas. El trabajo remarca que el verdadero público objetivo de un catálogo de datos suelen ser los usuarios de negocio, y no solo especialistas en datos o el área de TI. Por otra parte, el artículo propone lineamientos para la implementación de los catálogos de datos y analiza cómo su adopción puede acelerar las iniciativas de inteligencia artificial.

2.1.2.2. Producto de datos como soporte para el enriquecimiento

[(10), Tonnarelli] en su evaluación de herramientas de catálogo de datos identificó 5 características de alto nivel, y 24 características atómicas, entre las cuales se encuentra la característica 'producto de datos' y la clasifica dentro de la característica de alto nivel 'Uso de metadatos'. El concepto de producto de datos corresponde al segundo principio de la arquitectura *data mesh*[1], donde los datos son tratados como productos, es decir cada producto tiene un propósito claro, con métricas de calidad, soporte, entre otras consideraciones, se trata sin dudas de un uso de la metadata.

A su vez, los productos de datos pueden considerarse como un elemento relevante para el enriquecimiento de los metadatos, al instanciar conceptos de negocio que permiten identificar, por ejemplo, qué activos de datos físicos sirven a un propósito o caso de uso específico. Sin embargo, en la literatura revisada no se encontró evidencia de que los productos de datos u otros elementos definidos por los usuarios, como los propios conjuntos de datos (*datasets*) y sus atributos, sean considerados explícitamente como mecanismos para representar conceptos de negocio.

[1] https://en.wikipedia.org/wiki/Data_mesh

2.1.2.3. Ontologías para enriquecer el catálogo

Las herramientas de catálogo de datos, como ya se mencionó, no solo permiten descubrir y almacenar metadatos técnicos de los elementos físicos (por ejemplo, tablas y atributos), sino que también ofrecen mecanismos para incorporar contexto del conocimiento de dominio mediante anotaciones sobre la representación técnica de dichos elementos, a través de etiquetas (*tags*), glosarios de términos u ontologías. En cualquier catálogo, la ausencia de información contextual asociada a cada uno de sus ítems puede dificultar significativamente la localización y comprensión de los activos de datos.

En este sentido, [(6), Kropshofer] sostiene que la funcionalidad de anotar los metadatos técnicos con "contexto del negocio" resulta fundamental para mejorar la accesibilidad del catálogo y aumentar su valor para los usuarios del negocio. En esta misma línea, el autor argumenta que el uso de ontologías constituye una de las alternativas de mayor expresividad para incorporar conocimiento de dominio. Esta afirmación es consistente con lo planteado por [(3), FeilmayrWöß], quien señala que las ontologías representan la opción de mayor capacidad expresiva dentro de los modelos disponibles para la representación semántica.

Además de la expresividad de su modelo, las ontologías presentan otros beneficios relevantes: (i) facilitan el intercambio de conocimiento entre usuarios, por ejemplo al proveer un esquema común a datos heterogéneos; (ii) promueven la estandarización a partir de un lenguaje uniforme; y (iii) permiten una evolución continua mediante una gestión ágil de su estructura en grafo.[(3), FeilmayrWöß]

Aunque las tecnologías semánticas, como las ontologías y los grafos de conocimiento, existen desde hace varios años, su incorporación en las herramientas de catálogo de datos sigue siendo limitada. El desarrollo e implementación de una representación de conocimiento basada en ontologías para los catálogos de datos es un enfoque prometedor al que se le atribuye un gran potencial. [(6), Kropshofer]

2.1.2.4. Usos de ontologías

Es posible identificar dos usos principales de las ontologías para mejorar la expresividad de los metadatos de negocio: (i) describir el catálogo de datos mediante una ontología específica y (ii) modelar los conceptos de negocio y sus

relaciones a través de una ontología de dominio. El trabajo de [(9), Schrott] propone una ontología genérica para modelar catálogos de datos y utiliza ambas aproximaciones, (i) y (ii).

Uno de los objetivos de la prueba de concepto presentada en dicho trabajo es vincular las propiedades de la ontología de dominio con las columnas de bases de datos relacionales. Para lograrlo, se modela los metadatos técnicos de la herramienta de catálogo en una ontología que describe el catálogo (i), y se emplean ontologías adicionales para representar el dominio (ii). La prueba de concepto carga en la primera ontología las tablas y columnas del modelo relacional descubiertas por la herramienta de catálogo, y en la segunda incorpora las instancias de clases y propiedades propias de la ontología de dominio. Se establece la vinculación entre las instancias de propiedades de la ontología de dominio y las columnas de las tablas representadas en la ontología que describe el catálogo.

La prueba de concepto se estructura en torno a la pregunta central que el trabajo busca responder: qué clase de la ontología de negocio (el 'subject' en términos de RDF) se relaciona con qué otra clase de negocio (el 'object' en términos de RDF). En el caso instanciado en la prueba, el ejemplo corresponde a qué motocicletas están asociadas a qué reclamos de garantía. Esta información se obtiene a partir de bases de datos relacionales, repositorios NoSQL y otras fuentes, gracias a la vinculación previamente definida entre la metadata técnica y las ontologías de dominio

El trabajo se apoya en una organización que presenta un grado avanzado de catalogación de sus activos de datos, con una herramienta de catálogo capaz de recolectar metadata técnica desde diversas fuentes y que, además, dispone de ontologías de dominio que modelan distintos conceptos y relaciones de negocio.

Por razones de confidencialidad, el trabajo no publica los modelos intermedios ni las instancias de las ontologías utilizadas durante la prueba de concepto, dichos artefactos permitirían comprender con mayor precisión el proceso seguido y facilitarían la reutilización de la ontología genérica propuesta.

2.2. Herramientas existentes

Desde el comienzo de la iniciativa de gobierno de datos en el Banco, en el año 2020, se identificó como una necesidad facilitar y automatizar, mediante el uso de una herramienta, las tareas asociadas al gobierno de datos y al inventario

de los activos de información. En ese contexto, y con la colaboración de una consultora externa, se llevaron adelante las primeras instancias de trabajo, que comprendieron, en una etapa inicial, actividades de capacitación e intercambio orientadas a comprender la realidad institucional del Banco. Posteriormente, se avanzó en una prueba piloto de un repositorio central de datos y, finalmente, se realizó una introducción a cuatro herramientas de gobierno de datos de alcance internacional: Alation, Collibra, Precisely e Informatica.

Más recientemente, y en paralelo con este trabajo, el Banco mantuvo instancias de análisis y demostraciones de producto por parte de otras instituciones, tanto con el objetivo de conocer experiencias comparables como de evaluar las propuestas de distintos proveedores del mercado. En este marco, se consideró pertinente incorporar al estudio herramientas de código abierto, principalmente por la facilidad de acceso a entornos de prueba y porque, en determinados aspectos, estas soluciones pueden ofrecer funcionalidades experimentales o de laboratorio que aún no se encuentran disponibles en herramientas comerciales.

En este contexto, se definió evaluar, por un lado, herramientas de código abierto y, por otro, herramientas comerciales de clase mundial. En ambos casos, el foco estuvo puesto en identificar aquellas funcionalidades más relevantes y actuales que permitieran automatizar tareas de gobierno de datos, con especial énfasis en el catálogo de datos. A modo de ejemplo, se consideraron capacidades típicas de gobierno de datos, como la aplicación de políticas de acceso y clasificación de la información según su sensibilidad, y capacidades propias de catálogo de datos, como la gestión y enriquecimiento de metadatos orientados a describir los activos de datos y facilitar su descubrimiento y uso por parte de los usuarios del Banco.

En el caso de las herramientas de código abierto, se consideró como requisito necesario que se tratara de proyectos activos, con liberación periódica de nuevas versiones. Asimismo, se tuvo en cuenta el grado de adopción y visibilidad de cada herramienta, utilizando como indicador complementario la cantidad de estrellas asignadas a su repositorio de código fuente[1], entendidas como una métrica de popularidad y de interés de la comunidad en el proyecto asociado.

En función de estos criterios, y considerando además la disponibilidad de experiencias previas, las herramientas seleccionadas para su evaluación en este trabajo fueron OpenMetadata y DataHub dentro del conjunto de soluciones

[1] https://docs.github.com/en/get-started/exploring-projects-on-github/saving-repositories-with-stars

de código abierto, y Collibra y Microsoft Purview dentro del conjunto de herramientas comerciales de clase mundial con posibilidad de soporte local o regional.

Cabe destacar que, durante el proceso de análisis, se descartó la herramienta Apache Atlas, frecuentemente considerada como precursora de varias de las soluciones actuales de gobierno de datos. A partir de una instancia de demostración realizada en el Banco y de análisis posteriores, se concluyó que presenta una curva de aprendizaje elevada, tanto en lo relativo a su instalación y configuración como en su uso por parte de los usuarios finales. Adicionalmente, su diseño se encuentra fuertemente orientado al ecosistema Hadoop de Apache, y si bien ofrece soporte para bases de datos relacionales como Oracle, este no constituye el foco principal de la herramienta en su concepción original.

Luego de estos primeros análisis y evaluaciones, en la siguiente sección se describen los criterios finales seleccionados.

2.2.1. Criterios de análisis de las herramientas

Los criterios de evaluación de las herramientas de catálogo se definieron considerando tanto el contexto institucional y tecnológico del Banco, como el estado del arte académico y los objetivos del presente trabajo. Entre los aspectos relevantes se incluyeron la usabilidad para usuarios finales, la comunicación transversal entre áreas, los requerimientos de seguridad de la información y la orientación a resultados. A partir de este análisis, se establecieron doce criterios, los cuales se agruparon en cinco categorías con el objetivo de facilitar su interpretación. A continuación se describen estas categorías y sus criterios correspondientes.

(I) Funcionalidades esenciales del catálogo. Incluye capacidades básicas o generales esperadas en una herramienta de catálogo de datos.

- Interfaz y capacidades de búsqueda de activos.
- Soporte para social media y colaboración.
- Acceso mediante API.

(II) Capacidades de gobierno de datos. Agrupa funcionalidades vinculadas a políticas, calidad y consumo de datos confiables.

- Facilidad de consumo de los datos y políticas de acceso.
- Calidad, Datos confiables.

(III) Acceso y madurez de la herramienta. Refiere a aspectos rela-

cionados con comunidad, evolución, documentación y soporte.

- Comunidad, nuevas versiones, acceso a documentación y soporte.

(IV) Integración con la infraestructura tecnológica del Banco. Considera la facilidad de prueba, compatibilidad con la infraestructura existente y capacidades de linaje.

- Facilidad de la herramienta para probarla con datos del Banco.
- Soporte para con la infraestructura propia del Banco.
- Detección de relaciones de tipo linaje.

(V) Capacidades avanzadas de visualización de los activos. Esta categoría se centra en los objetivos específicos del presente trabajo, en particular en el soporte para el ingreso de conceptos definidos por el usuario y en su posterior representación gráfica, junto con las relaciones existentes entre los distintos activos de datos.

- Capacidades de descubrimiento de activos similares.
- Soporte para ingreso manual de datasets y sus relaciones con otros activos de datos.
- Visualización gráfica de mapas de activos de datos y sus relaciones.

A continuación se detallan los criterios de forma individual, en un orden independiente de su agrupación por categoría.

(1) Interfaz y capacidades de búsqueda de activos. Este criterio se enfoca en evaluar la facilidad de uso de las funcionalidades de búsqueda dentro del catálogo de datos de cada herramienta. La comparación se plantea en un escenario con decenas de miles de activos, donde solo algunos son de alto interés para el análisis de información mientras que los restantes son en general de uso técnico. La mayoría de los activos son artefactos de uso interno en los sistemas, con menor relevancia para un analista de datos o para los usuarios de herramientas de inteligencia de negocios (BI). Como parte de las facilidades de búsqueda el criterio evalúa qué expresiones son posibles en la búsqueda, con qué operadores, qué tipos de filtros son soportados. En cuanto a los resultados de la búsqueda, se analiza qué tipo de entidades se obtienen, además de los atributos, pestañas y acciones disponibles para cada elemento encontrado.

(2) Capacidades de descubrimiento de activos similares. Este criterio examina la capacidad de la herramienta para identificar activos o conjuntos de datos que sean idénticos o similares en cuanto a su esquema y contenido,

incluyendo la posibilidad de medir el grado de similitud. Esta característica permite al usuario buscar y comparar activos para determinar duplicidades de información o identificar cuál se ajusta mejor a sus necesidades de información.

(3) Facilidad de la herramienta para probarla con datos del Banco. Este criterio evalúa la capacidad de la herramienta o proveedor para conectarse a las fuentes de datos del Banco y analizar sus funcionalidades de descubrimiento, búsqueda y otras características sobre un conjunto de activos. Esto permite evaluar la información extraída automáticamente, incluyendo las relaciones identificadas entre los activos, tales como las relaciones de linaje.

Se valora la información obtenida automáticamente, incluyendo las relaciones identificadas entre los activos, como el linaje.

(4) Soporte para con la infraestructura propia del Banco. El Banco almacena sus datos en su propia infraestructura y red. Aunque utiliza servicios en la nube, estos sólo gestionan datos públicos o aquellos que cuentan con un acuerdo específico. Por lo tanto, cualquier solución de catálogo de datos destinada a extraer metadatos y, eventualmente, conjuntos de datos de ejemplo, tiene ventajas si los datos pueden permanecer dentro de la red interna del Banco. En cuanto a bases de datos y data warehouses donde el Banco gestiona sus activos de datos, generalmente se utilizan Oracle en las versiones 11g o 12c, aunque todavía se encuentran versiones antiguas como 9i y versiones más recientes como 18c y 19c. En lo referente a activos para análisis de información, como tableros y paneles de control, el Banco emplea Microsoft Power BI, tanto de forma local como en la nube en algunos casos, y Tableau exclusivamente de forma local

(5) Detección de relaciones de tipo linaje. El linaje es una información valiosa que las herramientas pueden extraer de manera automática. Es fundamental que las herramientas sean capaces de identificar el linaje en los activos del servicio del que están obteniendo datos, como las vistas de bases de datos y las tablas que sirven de fuentes de datos para los tableros. Además, un valor añadido que pueden ofrecer estas herramientas es la capacidad de interpretar mecanismos de ETL específicos del Banco, como la carga y las transformaciones de datos mediante Oracle PL/SQL, y los flujos de transformación a través de Pentaho Data Integration (PDI).

(6) Soporte para ingreso manual de data sets y sus relaciones con otros activos de datos. Este criterio se refiere a la capacidad de la herramienta para proporcionar información en el plano conceptual, permitien-

do describir conjuntos de datos relevantes para los analistas de datos y otros actores, y vincularlos a activos físicos descubiertos por la herramienta. Los conjuntos de datos, o data sets, son de propósito general y consisten en datos básicos. Algunas herramientas también ofrecen la opción de 'producto de datos', que difiere del conjunto de datos. Un producto de datos está vinculado a un caso de uso específico y tiene un objetivo definido para satisfacer una necesidad de información, como un informe financiero dentro de un contexto y análisis determinados. Por ejemplo, un producto de datos puede componerse de uno o varios conjuntos de datos.

(7) Visualización gráfica de mapas de activos de datos y sus relaciones. Este criterio tiene por objetivo evaluar si la herramienta ofrece funcionalidades que contribuya a una visión conceptual y general de los activos de datos, y que como mapa se espera que cuente con utilidades de ir de elementos generales a información más específica. Se apunta que los activos descubiertos más los elementos definidos por el usuario puedan ser vistos como diferentes capas de un mapa, ya sea por su trascendencia o para filtrar por distintos criterios de exploración, por ejemplo visualizar solo elementos del usuario junto al activo más lejano que se llegue por una relación de tipo linaje.

(8) Soporte para social media y colaboración. El soporte de la herramienta para la colaboración entre equipos y la publicación de información al estilo de una red social está diseñado para facilitar y fomentar la contribución de información valiosa por parte de los usuarios. Estos pueden opinar, mejorar descripciones, y asignar calificaciones sobre un activo. Las funcionalidades a analizar y evaluar en este aspecto incluyen la posibilidad de calificar un activo, modificar una descripción u otro atributo descriptivo de un activo, realizar comentarios, crear y asignar tareas de mejora de descripciones, y generar hilos de discusión sobre el activo, entre otras.

(9) Acceso mediante API. Este criterio evalúa si la herramienta cuenta con un API de acceso e integración, para luego evaluar sobre qué componentes y funcionales tiene alcance el API.

(10) Comunidad, nuevas versiones, acceso a documentación y soporte. Este criterio evalúa si la herramienta cuenta con una comunidad activa que le dé respaldo. Esto incluye la liberación periódica de versiones, así como la disponibilidad de documentación, información y acceso a demostraciones de la herramienta que faciliten la curva de aprendizaje para su adopción.

(11) Facilidad de consumo de los datos y políticas de acceso (Go-

bernanza). Este criterio no es estrictamente parte del catálogo de datos que es el objeto del trabajo pero se considera parte de una siguiente fase luego de contar con todos los datos catalogados. Una vez que todos los activos pudieron ser categorizados, asignada la relevancia adecuada, asociados a elementos del negocio, detectados o determinados los flujos de datos, comienza una etapa con la que se contaría con una disponibilidad "a un click" de todos los activos. De esta forma es importante evaluar las capacidades de las herramientas para controlar el acceso a los datos propiamente dichos, desde dos aspectos, primero si la herramienta ofrece una forma de consultar o acceder a un juego de datos de ejemplo del activo en cuestión y segundo si la herramienta ofrece un proceso de aprobación o forma de solicitud para que los roles adecuados eventualmente aprueben y den instrucciones para su acceso.

(12) Calidad, Datos confiables (Gobernanza). Una vez que se dispone de una gran cantidad de datos, es fundamental medir y mejorar su calidad. Para ello, es importante que la herramienta ofrezca reglas preestablecidas, como perfiles de datos en atributos y tablas, así como también reglas de consistencia y calidad. Además, estas reglas deberían ser personalizables mediante expresiones propias del usuario.

2.2.2. OpenMetadata

OpenMetadata es una plataforma unificada de metadatos diseñada para el descubrimiento, observabilidad y gobernanza de datos, nacida de la experiencia de Uber como una herramienta open source. Lanzada en 2021, OpenMetadata surge de la necesidad de establecer una única verdad de los metadatos, proporcionando un repositorio centralizado en contraste con las herramientas de calidad de datos, cada una con su propia copia y versión de metadatos. La plataforma busca establecer un estándar de metadatos, con un enfoque basado en esquemas, y ofrecer APIs que atienden diversos casos de uso, facilitando la integración tanto con herramientas proveedoras como consumidoras de metadatos, así como con la interfaz de usuario y las políticas de gobernanza de datos.[1] OpenMetadata destaca el representar la metadata como un grafo que no solo está centralizada y se almacena pasivamente, sino que también activamente organiza la metadata conectando servicios, herramientas, fuentes de datos, usuarios, equipos, actividad del usuario, workflows de procesamiento de

[1] https://blog.open-metadata.org/announcing-openmetadata-20399b816e60

datos, linaje, calidad, observabilidad y conectar con muchos otros contextos de datos. Para lograr esto requiere un estándar de metadatos con un vocabulario extensible, además de APIs diseñadas de manera centralizada y fáciles de usar. También ofrece una solución neutral respecto a proveedores y open source que habilita las características anteriores. En resumen, el enfoque de OpenMetadata es ofrecer un sistema de metadatos como producto, en lugar de como servicio, con una arquitectura sencilla y con un mínimo de dependencias para facilitar la implementación y operación a los usuarios. [2]

(1) Interfaz y capacidades de búsqueda de activos. OpenMetadata permite realizar búsquedas mediante palabras clave y devuelve resultados que incluyen activos como tablas, columnas y tableros. Además, muestra elementos de gobernanza de datos creados por el usuario, tales como productos de datos, etiquetas y glosarios de términos. El resultado de la búsqueda admite diferentes filtros, entre ellos tipo de activo, dominio o dueño de los datos, servicio o tipo de servicio que alimenta la metadata. Existe una búsqueda avanzada en la que es posible introducir los operadores lógicos `AND` y `OR`, y operadores como igualdad, contiene, es nulo, la negación de los anteriores o que cumpla una expresión regular. [1] Entrando en un activo en particular y dependiendo del mismo [2], es posible ver su linaje si el mismo fue descubierto, como también datos de ejemplos y otras tablas que se joinearon con el activo si el mismo también es una tabla, lo cual está disponible en función del `query-logs` del motor de base de datos.

(2) Capacidades de descubrimiento de activos similares. Según la documentación y las pruebas realizadas, no se encontró que OpenMetadata ofrezca soporte para identificar activos o conjuntos de datos similares.

(3) Facilidad de la herramienta para probarla con datos del Banco. Dada su arquitectura sencilla, la herramienta se instala y conecta a las fuentes de datos del Banco con relativa facilidad. Esto permite posteriormente visualizar los activos de datos del Banco y definir dominios, productos de datos, glosarios de términos y etiquetas, asociándolos a los activos descubiertos.

(4) Soporte para con la infraestructura propia del Banco. OpenMetadata incluye conectores para extraer metadatos de bases de datos Oracle, así como de tableros de Power BI y Tableau, que son las principales herramien-

[2] https://blog.open-metadata.org/why-openmetadata-is-the-right-choice-for-you-59e329163cac

[1] https://docs.open-metadata.org/latest/how-to-guides/data-discovery

[2] https://docs.open-metadata.org/latest/how-to-guides/data-discovery/details

tas de inteligencia de negocios utilizadas por el Banco. Sin embargo, Oracle es soportado solo desde la versión 12c[3], mientras que el banco aún utiliza varias bases de datos en la versión 11g. En cuanto a Tableau Server local, se pudo verificar su conexión. No obstante, esto no fue posible en el caso de Power BI Report Server, que es la solución local de Power BI para tableros.

(5) Detección de relaciones de tipo linaje. OpenMetadata es capaz de extraer linaje de metadatos existente en herramientas del Banco. Se pudo verificar con éxito el caso de la extracción de metadatos de Tableau Server local y la conexión del linaje de los tableros y activos en bases de datos Oracle. No se pudo verificar o encontrar información sobre otras herramientas que utiliza el Banco como son Oracle PL/SQL o procesos con transformaciones de Pentaho. Cuando las cargas de datos se realizan mediante una única sentencia SQL, bajo el formato `INSERT INTO table1 ... SELECT ... FROM table2`, se verificó que OpenMetadata es capaz de detectar el linaje correspondiente a partir de la información extraída de los *query-logs* de Oracle.

(6) Soporte para ingreso manual de data sets y sus relaciones con otros activos de datos. En OpenMetadata, actualmente no existe una entidad o concepto específico para 'conjunto de datos' (data set). Existe lo denominado 'producto de datos'[1], considerado como un informe o artefacto de salida dentro del dominio de datos correspondiente. El 'producto de datos' se refiere a datos más refinados, orientados a un caso de uso específico, mientras que el 'conjunto de datos' representa un concepto más básico con datos más crudos que pueden servir de base para otros conjuntos o productos de datos. Estos pueden verse como eslabones en una cadena, agregando valor a un producto final.

(7) Visualización gráfica de mapas de activos de datos y sus relaciones. Según las pruebas realizadas y la información disponible, no se identificaron elementos visuales que representen todos los activos de datos. OpenMetadata permite categorizar los activos en niveles, del 1 al 5, lo que habilita la opción de filtrar por estos niveles en las búsquedas.[2] Sin embargo, la asignación de importancia a cada activo, a través de estos niveles, parece ser únicamente manual. Además, no existe una representación visual que muestre estos niveles como capas en un mapa geográfico que permita seleccionar los niveles a

[3] https://docs.open-metadata.org/latest/connectors/database/oracle

[1] https://docs.open-metadata.org/latest/how-to-guides/data-governance/domains-&-data-products/data-products

[2] https://docs.open-metadata.org/latest/how-to-guides/data-insights/tiering

visualizar.

(8) Soporte para social media y colaboración. OpenMetadata fomenta la colaboración entre usuarios y equipos, rompiendo con los esquemas de silos de información y compartiéndola a lo largo de toda la organización.[1] Permite la creación de hilos de discusión alrededor de cualquier activo de información, donde un usuario puede realizar consultas y otros pueden responder para mejorar la descripción y recomendar el activo en cuestión u otros, basándose en su propia experiencia. Además de los hilos de discusión, se pueden crear y asignar tareas relacionadas con los activos, como solicitar la creación o mejora de una descripción, pedir una etiqueta o iniciar el flujo de aprobación de un nuevo término del glosario. También es posible generar anuncios para toda la organización, como informar sobre una nueva fuente de datos o cambios importantes inminentes.

(9) Acceso mediante API. OpenMetadata dispone de APIs REST para acceder a su repositorio de metadatos y utiliza esquemas JSON para modelar sus entidades. El acceso mediante APIs REST es la forma como el propio OpenMetadata permite alimentar el repositorio de metadatos a través de diversos conectores, desde fuentes internas y externas, como también explorar y buscar activos mediante la interfaz de usuario e integrarse con sistemas externos.[2]

(10) Comunidad, nuevas versiones, acceso a documentación y soporte. OpenMetadata tiene el respaldo de una sólida comunidad que frecuentemente proporciona documentación y webinars, libera versiones semanalmente en varios casos, y ofrece soporte activo a través de Slack. Además, el proyecto ha recibido 7,000 estrellas en GitHub.[3]

(11) Facilidad de consumo de los datos y políticas de acceso (Gobernanza). OpenMetadata ofrece la capacidad de mostrar datos de ejemplo a través de la pestaña 'Sample Data', disponible dentro de la visualización de activos como tablas o tópicos.[4] Sin embargo, no permite aplicar filtros ni realizar consultas para obtener datos distintos a los proporcionados como ejemplo. En cuanto a las políticas de acceso, no se integra con un flujo de aprobaciones que gestione las solicitudes de acceso a los activos de datos.

(12) Calidad, Datos confiables (Gobernanza). La calidad y observabilidad de los datos es un aspecto de relevancia para OpenMetadata. Desde su

[1] https://docs.open-metadata.org/latest/how-to-guides/data-collaboration
[2] https://docs.open-metadata.org/latest/main-concepts/high-level-design
[3] https://github.com/open-metadata/openmetadata
[4] https://docs.open-metadata.org/latest/how-to-guides/data-discovery/details

lanzamiento en 2021, la calidad de los datos está presente como un rol clave haciendo que los datos sean funcionales para las organizaciones. La herramienta ofrece perfiles de datos por columnas, ejecución de reglas de calidad y proporciona alertas mediante notificaciones o paneles de control sobre los resultados de estas ejecuciones.[5]

2.2.3. DataHub

DataHub es una herramienta de código abierto que, a lo largo de sus diferentes iteraciones, acumula más de 10 años de experiencia en la construcción de sistemas basados en metadatos en LinkedIn. En 2020, fue lanzada oficialmente bajo su nombre actual, tras ser completamente reescrita y compartida nuevamente con la comunidad open source. Dentro de la categoría de herramientas de código abierto, probablemente sea la herramienta con la comunidad más grande, ya que cuenta con más de 12,000 miembros en Slack, más de 10,000 estrellas en GitHub y supera los 6,000 despliegues. [1] [2] [3]

La herramienta se enfoca en el descubrimiento y la gestión de metadatos, la observabilidad y la gobernanza de datos. DataHub se distingue por su fuerte apuesta por establecer un estándar de facto para metadatos de datos y artefactos de inteligencia artificial. Su modelo permite incorporar tanto metadatos técnicos como de negocio. Gracias a los varios años de experiencia de su equipo, DataHub ha sabido aprender de errores pasados en el modelado de esta columna vertebral que es la metadata para descubrir, gobernar y monitorear los datos.

Otra característica destacable de DataHub es su potente motor de búsqueda de activos, que facilita la formulación de expresiones lógicas complejas, como también habilita diversos criterios de búsqueda y resultados a través de su interfaz GraphQL. Los resultados de la búsqueda pueden ordenarse según la popularidad de cada activo, medida por la cantidad de consultas mensuales en las que se utiliza. [4] [5]

(1) Interfaz y capacidades de búsqueda de activos. DataHub ofrece

[5] https://docs.open-metadata.org/latest/how-to-guides/data-quality-observability
[1] https://docs.datahub.com/docs/metadata-standards
[2] https://www.linkedin.com/blog/engineering/data-management/datahub-popular-metadata-architectures-explained
[3] https://medium.com/datahub-project/datahub-1-0-is-here-94805883135f
[4] https://docs.datahub.com/docs/how/search
[5] https://medium.com/datahub-project/using-datahub-for-search-discovery-fa309089be22

una interfaz de búsqueda moderna que facilita tanto búsquedas sencillas como avanzadas a través de su lenguaje de expresiones y operadores. La aplicación de filtros es intuitiva, mostrando la cantidad de elementos disponibles en cada opción. En las expresiones de búsqueda, se pueden emplear comillas para buscar frases exactas, excluir palabras con el signo `-`, definir precedencia con `()`, y usar los operadores `+` (AND) y `|` (OR). La búsqueda por atributos se realiza especificando su valor o solo su existencia mediante '`/q nombreAtributo:valor`' y '`/q _exists_:nombreAtributo`', respectivamente.

Entre los filtros disponibles, se encuentran opciones por tipo de activo, etiqueta, término del glosario asociado, dominio, propietario y sub-tipo. Adicionalmente, de forma gráfica, es posible generar expresiones para excluir un filtro o indicar si se aplican todos los filtros en modo AND u OR, es decir, mostrar los activos que cumplan con al menos uno de los criterios seleccionados.

La información asociada a cada activo en los resultados de búsqueda incluye el tipo de activo (Datasets, Dashboard, Data flow, Data job, etc.), la plataforma a la que pertenece (Tableau, Snowflake, Spark, MySQL, etc.), nombre, dominio y producto de datos (si aplica), documentación (descripción), etiquetas asignadas e indicadores de calidad como consultas realizadas sobre el activo, popularidad y frescura de los datos. Pestañas adicionales proporcionan información detallada sobre las columnas/atributos del activo, su linaje, estadísticas, validaciones recibidas e incidentes de calidad asociados. [1] [2]

(2) Capacidades de descubrimiento de activos similares. No se identificaron en la documentación disponible ni en la interfaz de usuario de DataHub funcionalidades para la detección o búsqueda de conjuntos de datos similares.

(3) Facilidad de la herramienta para probarla con datos del Banco. El acceso a la documentación y el software para una instalación rápida (quickstart) es accesible y sencillo de hacerla. Cuenta con una guía paso a paso y un cliente que realza la instalación en una instancia de DataHub utilizando `docker-compose`. Luego la configuración para conectar DataHub a fuentes del Banco es por medio de sintaxis YAML de la cual se dispone del esquema de la configuración según la fuente de datos. [3]

[1] https://docs.datahub.com/docs/how/search
[2] https://docs.datahub.com/docs/metadata-modeling/metadata-model/
[3] https://docs.datahub.com/docs/quickstart

(4) Soporte para con la infraestructura propia del Banco. DataHub proporciona conectores para Oracle, Tableau y Power BI, permitiendo la extracción de metadatos desde estas fuentes. Se verificó exitosamente la ingesta de metadatos para Oracle y Power BI Report Server (RS) mediante la configuración de parámetros en archivos con sintaxis YAML.

En el caso de Oracle, la conexión se estableció de forma inmediata. Para Power BI RS, en cambio, fue necesario realizar varios intentos y tareas de diagnóstico, incluyendo la búsqueda de errores y la posterior consulta en el canal oficial de soporte en Slack, debido a un parámetro que no está debidamente documentado. Finalmente, tras múltiples pruebas, se logró la extracción de activos desde Power BI RS. [1] [2]

(5) Detección de relaciones de tipo linaje. Según la documentación oficial, el conector de Tableau admite extracción de linaje a nivel de columna, lo que evidencia una capacidad más avanzada en comparación con el conector para Power BI Report Server que extrae linaje. En el caso de Oracle, el linaje recuperado se limitó a las relaciones entre vistas y las tablas subyacentes que estas consultan; no se comprobó extracción de linaje a partir de `query-logs`.

La extracción de linaje para el caso de Tableau, según la documentación oficial, el conector admite linaje a nivel de columna, lo cual indica una capacidad mayor del conector respecto al conector para Power BI RS. En el caso del conector para Oracle solo se extrajo el linaje específicamente identificado para las relaciones entre vistas y las tablas subyacentes que consultan, no hubo extracción de linaje del `query-logs`. [3]

(6) Soporte para ingreso manual de data sets y sus relaciones con otros activos de datos. En DataHub, el 'dataset' es una de las entidades más importantes dentro de su modelo de metadatos. Estas entidades dentro de DataHub representan una colección de datos que típicamente son una tabla, vista, archivo o flujo de una plataforma específica. Es importante mencionar que DataHub concibe el dataset como un metadato técnico, no como un concepto de negocio que se apoya en múltiples activos.

Actualmente, los datasets se pueden ingestar mediante el cliente de línea de comandos (CLI) o la API; no es posible crearlos directamente desde la interfaz web. Esto se debe a que el diseño de DataHub prioriza el descubrimiento

[1] https://docs.datahub.com/docs/generated/ingestion/sources/powerbi-report-server
[2] https://docs.datahub.com/docs/generated/ingestion/sources/oracle
[3] https://docs.datahub.com/docs/generated/ingestion/sources/tableau

automático de datasets a través de conectores de ingesta de metadatos. [4]

(7) Visualización gráfica de mapas de activos de datos y sus relaciones. De acuerdo con la documentación disponible y las pruebas hechas no se identificó en DataHub una funcionalidad o componente destinado a la visualización de un mapa con todos los activos.

(8) Soporte para social media y colaboración. DataHub permite enriquecer los metadatos de los activos mediante descripciones y enlaces a documentación externa, pero carece de funciones esenciales para la colaboración. No es posible agregar comentarios a los activos, compartir el trabajo de enriquecimiento con otros usuarios o asignar calificaciones. En esencia, la herramienta no provee de los medios como para trabajar de forma colaborativa y publicando información al estilo red social.

(9) Acceso mediante API. DataHub proporciona una API GraphQL[1] que permite interactuar de forma programática con las entidades y relaciones que conforman el grafo de metadatos de la organización. Entre las características destacadas del lenguaje de consulta GraphQL se encuentran su especificación formal y validada, su tipado fuerte, y su capacidad para estructurar consultas en formato JSON de manera accesible, flexible y eficiente.

Siguiendo una filosofía *API-first*, DataHub complementa su API con un SDK (Software Development Kit) y una interfaz de línea de comandos (CLI, Command Line Interface), lo cual facilita su adopción e integración por parte de desarrolladores. Estas características hacen que el API de DataHub se distinga por su enfoque amigable para el desarrollo y la automatización. [2] [3]

(10) Comunidad, nuevas versiones, acceso a documentación y soporte. Como se mencionó en la introducción, DataHub cuenta con una de las comunidades con más años de experiencia en el ámbito del modelado de herramientas de metadatos. Su comunidad es particularmente grande, con más de 12.000 miembros participando en los canales de Slack oficiales[4].

La liberación de versiones se realiza de forma periódica, con una frecuencia que varía entre mensual y trimestral, lo cual permite una evolución continua de la plataforma.[5]

[4] https://docs.datahub.com/docs/generated/metamodel/entities/dataset
[1] https://graphql.org/
[2] https://docs.datahub.com/docs/api/graphql/overview
[3] https://datahub.com/blog/5-features-to-look-out-for-in-a-modern-data-catalog/
[4] https://datahub.com/slack/
[5] https://github.com/datahub-project/datahub/releases

En cuanto a la documentación, DataHub ofrece un conjunto extenso de recursos. Además de permitir la instalación local, dispone de una instancia de demostración en línea y un recorrido guiado del producto, herramientas que resultan especialmente útiles para familiarizarse rápidamente con sus funcionalidades[6] [7].

(11) Facilidad de consumo de los datos y políticas de acceso (Gobernanza). DataHub no proporciona directamente acceso a datos de ejemplo o muestras de los mismos en su documentación ni a través de las pruebas realizadas. DataHub permite la definición de roles de acceso para los distintos actores, así como la configuración de políticas de control de acceso: a nivel de plataforma y a nivel de metadatos. Sin embargo, no se encontró funcionalidad que permita a un usuario solicitar acceso a un activo directamente desde la interfaz de la plataforma. [1]

(12) Calidad, Datos confiables (Gobernanza). Data observability es una de las características fundacionales de DataHub, junto a data discovery y data governance. DataHub permite asignar reglas (assertions) predefinidas sobre las columnas de cada activo, junto a la regla se establece con que frecuencia se ejecutan y luego es posible visualizar cuales fallaron y cuales cumplen los criterios establecidos, algunos de estos tipos de reglas son: contar la cantidad de valores nulos, su porcentaje, cantidad de valores únicos entre otros. Según la documentación a partir de reglas que no se cumplen es posible generar incidentes y enviar notificaciones, adicionalmente visualizar en un panel de control la salud general de los datos. Varios de estas funcionalidades de calidad y monitoreo están disponibles solo para la instalación en la nube. Por ejemplo la creación de de reglas asociados a columnas de una tabla no estaba disponible en la instalación local, llamada DataHub Core. [2] [3]

2.2.4. Microsoft Purview

Microsoft Purview es un conjunto integral de productos diseñado para abordar soluciones de gobernanza de datos, seguridad de datos, gestión de riesgos y cumplimiento normativo.[4] Dentro de las soluciones de gobernanza de da-

[6] https://demo.datahub.com/
[7] https://datahub.com/product-tour/
[1] https://docs.datahub.com/docs/authorization/access-policies-guide
[2] https://docs.datahub.com/docs/managed-datahub/managed-datahub-overview
[3] https://pages.datahub.com/data-quality-webinar?hs_preview=SkEEwXpw-187237002591
[4] https://learn.microsoft.com/en-us/purview/purview/

tos, destacamos el 'Catálogo Unificado', que será el foco de nuestro análisis. Además, se incluyen herramientas como Data Map y Data Governance Planning. En el ámbito de la seguridad de datos, Microsoft Purview ofrece productos tales como Protección de la Información, Administración de Riesgos Internos y Prevención de Pérdida de Datos. Por último, en el área de gestión de riesgos y cumplimiento normativo, se incluyen soluciones como Auditoría, Administración del Ciclo de Vida de los Datos y eDiscovery.

En conjunto, estos productos están diseñados para complementar otros servicios de Microsoft utilizados por sus clientes, como son Microsoft Azure, Office 365, Exchange Online y Microsoft Teams. Las soluciones de gobierno de datos ayudan a administrar servicios de datos en entorno local, multinube, y de software como servicio (SaaS). Esto incluye servicios que cuentan con almacenamiento de datos como Azure, Power BI, bases de datos como SQL Server o Hive, servicios de archivo como Amazon S3 y otros. Las soluciones incluyen:

- **Microsoft Purview Data Map** o Mapa de datos que registra la fuente de datos y la escanea identificando los activos y los datos sensibles.
- **Microsoft Purview Unified Catalog** para curar los datos, administrar la integridad y sacarle el mayor provecho posible a los datos.

Purview sugiere un plan de configuración e implantación para aprovechar mejor su catálogo de datos unificado.[1] En el mismo establece comenzar por crear uno o más dominios de datos, ya sean con una división por departamentos, funcional o de otro tipo, luego registrar una o más fuentes de datos con Data Map para el escaneo de activos de datos, crear un producto de datos o más, conectarlos con sus respectivos activos antes descubiertos, publicar el dominio luego de que en sucesivas iteraciones se obtuvo una retroalimentación satisfactoria de algunos usuarios elegidos para tal fin. Posteriormente los otros pasos en el plan son definir y ejecutar las reglas de calidad, crear una única verdad de los datos con el manejo los datos maestros, medir el nivel de madurez del gobierno de los datos con los controles de salud de los datos y por último democratizar los datos.

Con las primeras iteraciones del plan y el contar un grupo predeterminado de usuarios, Purview marca una estrategia de cómo implantar un gobierno de

[1] https://learn.microsoft.com/en-us/purview/unified-catalog-plan

datos, en particular un catálogo de datos, destaca que al principio no publicar el catálogo de datos a todos los usuarios de la organización, primero asegurarse que los activos de datos están en su nivel de entrega adecuado, por ejemplo que los productos de datos estén en el formato que lo necesitan los científicos de datos o que los reportes predefinidos están en la forma que los esperan los usuarios de negocio, porque de lo contrario los usuarios perderán la confianza en el catálogo.[2]

Otro punto importante que destaca Purview es la relevancia que le da a los conceptos de negocios y como conectar a estos con los activos de datos.[1] Los conceptos de negocio los define en estos 5 elementos 1. Dominios de gobernanza, 2. Productos de datos, Elementos de datos críticos (CDE por su sigla en inglés), 4. Términos del glosario y por último 5. Objetivos y resultados clave (OKR por su sigla en inglés).

(1) Interfaz y capacidades de búsqueda de activos. Purview cuenta con dos formas principales de búsqueda, una para buscar productos de datos y otra para buscar activos. En el caso de los productos los mismos pueden ser buscados por palabras claves como así también por lenguaje natural, escribiendo una frase que refiere al caso de uso para el cual se requiere el producto de datos. La búsqueda se realiza tanto sobre los atributos del producto de datos (nombre, descripción y caso de negocio) como así también sobre elementos asociados al producto como ser CDEs (elementos de datos críticos por su sigla en inglés), OKRs (objetivos y resultados claves) y glosario de términos, por último también sobre los atributos de los activos de datos contenidos en el producto de datos.

En el segundo caso de búsqueda, la misma es sobre activos de datos que se realiza por medio de palabras clave que ingresa el usuario y el resultado es una lista de diferentes activos de datos, tales como reportes, tablas, glosarios de términos, etc.[2] A las palabras clave del criterio de búsqueda también es posible agregarles filtros, indicando un operador lógico, `AND` o `OR`, para filtrar por ejemplo por tipo de activo, termino del glosario asignado al activo, etiqueta de seguridad, contacto que tiene el activo, etc. Todas las búsquedas consisten de palabras claves y operadores. A los operadores `AND` y `OR` mencionados anterior-

[2] https://learn.microsoft.com/en-us/purview/unified-catalog-plan#let-users-search-and-browse-your-governance-domains-and-data-products

[1] https://learn.microsoft.com/en-us/purview/data-governance-get-started#connect-your-data-with-business-concepts

[2] https://learn.microsoft.com/en-us/purview/unified-catalog-data-assets-search

mente se agregan también `NOT` para negar una palabra, agrupación por medio de `()`, frase específica por medio de comillas dobles `‘‘’’` y búsquedas de tipo `campo:palabraClave` para buscar palabras dentro de ciertos campos como por ejemplo `description:Deuda` que buscará todos los activos en cuya descripción tengan la palabra `Deuda`. El resultado de la búsqueda trae los datos básicos de cada activo como ser nombre, descripción, ubicación y nombre completo (FQN), entrando en un activo particular se puede observar clasificaciones que se aplican a los atributos del activo y secciones de linaje, esquema, glosarios de términos asociados, etc.

(2) Capacidades de descubrimiento de activos similares. En la documentación disponible, no se encontró evidencia de que las capacidades de descubrimiento ofrezcan soporte para la búsqueda de conjuntos de datos o activos similares. La única referencia identificada fue una pestaña denominada 'Related', ubicada en la visualización del activo, aunque no se proporciona una explicación detallada al respecto.

(3) Facilidad de la herramienta para probarla con datos del Banco. Al tratarse de una herramienta paga, a priori no es posible una demostración o prueba de conceptos con un juego de datos pertenecientes al Banco.

(4) Soporte para con la infraestructura propia del Banco. Data Map de Purview admite la ingesta de metadatos desde Oracle, abarcando desde la versión 6i hasta la 19c[1], lo que cubre una gran parte de los metadatos presentes en bases de datos relacionales. En cuanto a Tableau y Power BI, ambas ingestas son soportadas; Power BI Services por un lado y un servidor disponible en Internet para el caso de Tableau. Sin embargo, estas condiciones no se aplican actualmente en el contexto del Banco.

(5) Detección de relaciones de tipo linaje. El soporte de linaje en el Catálogo Unificado de Purview es relativamente reciente. La documentación actual sobre linaje está centrada en el Catálogo de Datos Clásico, el cual ya no se ofrece a nuevos clientes[2]. Esto se evidencia en la falta de detalles específicos sobre el linaje soportado en función de la fuente de datos. En el caso de Power BI, el linaje es compatible únicamente con las siguientes fuentes: Azure SQL Database, Azure Blob Storage, Azure Data Lake Store Gen1 y Gen2. Además, el linaje a nivel de columna para Power BI está soportado

[1] https://learn.microsoft.com/en-us/purview/register-scan-oracle-source

[2] https://learn.microsoft.com/en-us/purview/legacy/use-microsoft-purview-governance-portal

exclusivamente en Azure SQL Database[3]. No se encontró información detallada sobre las especificaciones del linaje capturado para otras fuentes de datos, como Tableau.

(6) Soporte para ingreso manual de data sets y sus relaciones con otros activos de datos. Purview permite la creación de productos de datos, cada uno con un caso de uso definido para ser compartido con otros usuarios interesados en datos. Cada producto de datos tiene un tipo asociado; un tipo posible es el conjunto de datos o 'data set'. Otros tipos pueden ser operacional, tableros o reportes, entre otros. Además, a cada producto de datos se le pueden asociar uno o más activos de datos, los cuales están relacionados para satisfacer el caso de uso del producto. Purview destaca los productos de datos, y en consecuencia los conjuntos de datos, como un concepto de negocio. Al asociarlos con sus activos, Purview conecta conceptos de negocio con datos físicos.

(7) Visualización gráfica de mapas de activos de datos y sus relaciones. No se encontró ninguna referencia en la información del producto que indique la existencia de un gráfico o alguna manera de visualizar un mapa de los activos de datos.

(8) Soporte para social media y colaboración. En la información disponible, no se evidenció que Purview promueva un entorno en el cual cualquier usuario consumidor de datos pueda modificar o sugerir una descripción para un activo. Tampoco parece permitir la asignación de calificaciones ni la creación de hilos de discusión a partir de un activo.

(9) Acceso mediante API. Purview ofrece, en modo 'preview', una API GraphQL, así como convencionalmente una API REST para acceder al Catálogo Unificado. La API GraphQL se distingue por su facilidad de uso, requiriendo un menor número de invocaciones en comparación con la API REST.[1]. El API esta pensada para acceder al catálogo de datos, gestionar los controles de acceso, como poder hacer búsquedas sobre el catálogo.[2]

(10) Comunidad, nuevas versiones, acceso a documentación y soporte. El catálogo unificado de Purview cuenta con un sólido respaldo, evidenciado por la abundante documentación y recursos disponibles, como videos y tutoriales. Además, es evidente que se trata de un producto en constante

[3] https://learn.microsoft.com/en-us/purview/concept-data-lineage

[1] https://learn.microsoft.com/en-us/purview/data-gov-api-graphql

[2] https://learn.microsoft.com/en-us/purview/data-governance-faq#how-do-i-use-the-rest-api-for-microsoft-purview-

desarrollo, con cambios recientes y una hoja de ruta para futuras actualizaciones[3]. Sin embargo, se pudo verificar la falta de acceso a demostraciones en línea donde sea posible interactuar con la interfaz de usuario.

(11) Facilidad de consumo de los datos y políticas de acceso (Gobernanza). La herramienta impulsa la incorporación del concepto de negocio de 'producto de datos' para facilitar la búsqueda de datos y activos de datos. Mientras este enfoque no se implemente en la organización, el acceso a los datos a través de activos de datos físicos estará limitado, o bien, podría no ser posible distinguir datos útiles o de calidad del resto de los activos de datos.[1] Además, no se evidenció la posibilidad de acceder a una muestra de datos de ejemplo de un activo.

(12) Calidad, Datos confiables (Gobernanza). Los datos de calidad o datos curados en la herramienta están sujetos a cómo se organice el gobierno de datos y se publiquen los conceptos de negocio en la misma.[2] Aunque existen reglas de calidad que se pueden aplicar a los datos, no se encontró documentación que explique su uso. Hay una función para obtener sugerencias de reglas de calidad a aplicar, pero actualmente está en modo 'preview' y tampoco se dispone de documentación al respecto. [3]

2.2.5. Collibra

Collibra[4] es una herramienta líder en gobernanza de datos, reconocida en el cuadrante mágico de Gartner por sus capacidades avanzadas en el cuadrante de plataformas de gobernanza y analítica de datos del año 2025. A diferencia de soluciones de código abierto como OpenMetadata y DataHub, Collibra es una plataforma paga que opera bajo un modelo de suscripción anual, con costos que suelen alcanzar cifras significativas.

En base a la documentación disponible el catálogo de datos es una de las características destacadas de Collibra, incluido el descubrimiento y manejo de metadatos.

(1) Interfaz y capacidades de búsqueda de activos. Collibra se destaca por una interfaz de usuario amigable e intuitiva, en particular para la

[3] https://learn.microsoft.com/en-us/purview/whats-new
[1] https://learn.microsoft.com/en-us/purview/unified-catalog-data-assets-search
[2] https://learn.microsoft.com/en-us/purview/unified-catalog
[3] https://learn.microsoft.com/en-us/purview/unified-catalog-responsible-ai-faq
[4] https://www.collibra.com/

búsqueda de activos de datos dentro de su catálogo de datos. Según la documentación disponible, permite ingresar una o más palabras claves y como resultado muestra cada uno de los activos que cumplen con el criterio de búsqueda. Estos activos resultantes, según los ejemplos de la documentación, pueden ser data sets, tablas, reportes, modelos analíticos, columnas, etc. En las expresiones de búsqueda no se pudo verificar que soporte operadores lógicos, o posibilidad gráfica de implementarlos, pero sí se puede apreciar que permite filtrar por data sets, reportes y tablas, y más importante permite filtrar por solo datos confiables o certificados como define Collibra. De cada activo resultante de la búsqueda se muestra, i) de que tipo es, ii) su nombre y descripción, comentario si lo hay, iii) datos jerárquicos de su ubicación, concretamente base de datos, esquema, sitio, carpeta o lo que aplique en cada caso, y iv) tags o etiquetas. Por último entrando en un activo particular, por ejemplo en el caso de una vista de base de datos, se muestra un resumen auto-generado que contiene la cantidad de columnas, los tipos en general de las columnas, y algo que es destacable, la clasificación de cada columna, como ser tipo e-mail, salario, centro de costos, estado del empleado, etc. Otras informaciones que se visualizan en el activo son una pestaña de datos similares, data profiling de columnas, por ejemplo valores distintos y porcentaje de valores vacíos, los responsables de los datos y acciones que se pueden realizar sobre el activo. Sino no se tiene acceso al activo, la única acción posible es solicitar acceso al mismo con lo cual se dispara un flujo de aprobación. Cuando ya se tiene acceso al activo, otros ejemplos de acciones sobre el mismo, son dejar hecha una pregunta a los expertos del data set o reportar un incidente del mismo.

(2) Capacidades de descubrimiento de activos similares. Collibra ofrece funcionalidades para identificar similitudes a nivel de conjuntos de datos, denominados "data sets" según la terminología utilizada por Collibra para este tipo de activo. En sus demostraciones guiadas, se incluye una pestaña dedicada a mostrar conjuntos de datos similares al examinar los detalles de un conjunto específico. La documentación señala que la búsqueda de similitudes se basa en el nombre y los atributos del conjunto de datos. [1]

No se ha encontrado documentación o información sobre el algoritmo o estrategia utilizada para identificar estas similitudes, ni tampoco si existe la posibilidad de que realice búsquedas basadas en los valores de los datos en

[1] https://productresources.collibra.com/docs/collibra/latest/Content/Catalog/co_matchers.htm

lugar de solo la metadata de los mismos.

(3) Facilidad de la herramienta para probarla con datos del Banco. Dado que Collibra es una herramienta de alcance mundial que requiere una licencia de uso, realizar una prueba con los datos del Banco resulta prácticamente inviable, a menos que el Banco tenga un interés real en adquirir la herramienta y un representante de la herramienta esté dispuesto a invertir en una prueba de conceptos.

(4) Soporte para con la infraestructura propia del Banco. Según documentación de Collibra, sobre la conexión a diversas fuentes de metadatos, Collibra admite bases de datos Oracle desde la versión 11g en adelante. Esto incluye la extracción de metadatos técnicos y el linaje. Además, Collibra proporciona soporte para extraer metadatos de tableros y reportes como Power BI y Tableau. [1]. La documentación no especifica para el caso de Power BI, si soporta la ingesta tanto de Power BI Services como de Power BI Report Server, este último es donde el Banco tiene mayor cantidad de tableros.

(5) Detección de relaciones de tipo linaje En los diferentes videos y presentaciones que Collibra publica, hace énfasis en las relaciones de tipo linaje y constantemente las nuevas versiones liberan soporte para extraer metadata de nuevas fuentes y nuevas formas de descubrir linaje, como por ejemplo en distintos objetos de las bases de datos como lo son procedimientos almacenados, triggers, database links, etc. En la interfaz de usuario, asociado por ejemplo a un activo de tipo vista de base de datos, se muestran dos pestañas correspondientes a como se alimenta el activo que son "Diagram" y "Technical Lineage", la primera refiere a donde se ubica gráficamente el activo, por ejemplo en que base de datos y por cuales reportes es consumida esta vista. La pestaña de "Technical Lineage" permite visualizar un gráfico con el linaje a nivel de cada columna de la vista, trazando una línea desde la columna origen en la tabla y a una columna en la vista como destino del flujo de datos. También en el ejemplo se muestra el código SQL con el objetivo de entender las transformaciones que reciben los datos entre origen y destino. Es importante mencionar que el linaje en Collibra es una característica que únicamente está presente en la nube, no on-premise, lo cual es una desventaja para el Banco. [2]

(6) Soporte para ingreso manual de data sets y sus relaciones con

[1] https://productresources.collibra.com/docs/collibra/latest/Content/CollibraDataLineage/TechnicalLineage/ref_technical-lineage-supported-data-sources.htm

[2] https://productresources.collibra.com/docs/collibra/latest/Content/to_data-lineage.htm

otros activos de datos. Collibra permite a los usuarios ingresar conjuntos de datos de forma manual y luego asociarles uno o más activos de datos. [3]

(7) Visualización gráfica de mapas de activos de datos y sus relaciones. En la documentación y recursos disponibles se encontraron diagramas asociados a un activo y sus relaciones con otros, pero no así un diagrama general de los datos. Lo que sí tiene Collibra es una diferenciación que hace entre 3 conjuntos de activos de información, i) Data awareness, ii) Data understanding y iii) Data use. Los primeros son de una escala de miles de millones, y los roles de TI son quienes saben que existen, los segundos (ii) son de escala de millones, y los roles de gobierno de datos como data stewards y data owners son los roles que tienen entendimiento de los mismos, por último, (iii) tenemos los activos que se usan, son de una escala de miles a cientos de miles de activos, usados por los analistas de negocio y consumidores de datos. Este último grupo de activos es lo que da Collibra en llamarle: el marketplace, que son una cantidad de activos más manejable y compuesta de datos confiables según el diseño del catálogo que hace Collibra. Esta clasificación, de separar lo que es secundario o auxiliar de lo que verdaderamente se usa, es una forma de poder contar con un mapa de los datos.

(8) Soporte para social media y colaboración. Collibra ofrece características que promueven un entorno colaborativo entre los usuarios y los equipos que consumen datos. Los usuarios pueden calificar un activo de datos con una puntuación de una a cinco estrellas, además de agregar comentarios. Posteriormente, otros usuarios pueden ver esta evaluación y ponerse en contacto con la persona que la realizó. La información disponible de Collibra no indica la posibilidad de crear hilos de discusión sobre un activo, como para debatir si existe otro activo más adecuado para un objetivo específico de información. Tampoco se menciona si un usuario puede crear y asignar tareas a otro usuario, como solicitar la mejora de la descripción de un activo.

(9) Acceso mediante API Es posible conectarse a Collibra mediante APIs de tipo REST. A través de este mecanismo, la plataforma ofrece diversas funcionalidades, como la creación, modificación y eliminación de recursos, incluyendo usuarios, activos y dominios. Además, permite importar grandes volúmenes de datos, interactuar con el motor de búsqueda del catálogo, con el linaje técnico, entre otras funcionalidades.

[3] https://productresources.collibra.com/docs/collibra/latest/Content/Catalog/DataSets/ta_create-data-set.htm

(10) Comunidad, nuevas versiones, acceso a documentación y soporte. Collibra es una herramienta reconocida mundialmente en el ámbito de los catálogos y la gobernanza de datos. Ofrece una amplia documentación sobre sus diversas funcionalidades y dispone de un sitio dedicado a la capacitación.[1] Además, promueve el conocimiento en este campo a través de artículos y conferencias, como se destaca en eventos específicos. [2]

(11) Facilidad de consumo de los datos y políticas de acceso (Gobernanza). Collibra proporciona una funcionalidad denominada Data Notebooks, la cual facilita y agiliza la exploración de los datos por parte del usuario. [1] Esta herramienta permite la creación de visualizaciones, como cuadros y gráficas, a partir de consultas SQL. El acceso al activo de datos se solicita mediante la ejecución de un flujo de trabajo preestablecido, sujeto a las correspondientes aprobaciones conforme a los procedimientos establecidos.

(12) Calidad, Datos confiables (Gobernanza) Collibra cuenta con un módulo de calidad y observabilidad que permite monitorear la salud de los datos mediante diferentes indicadores.[2] Estos indicadores incluyen, entre otros, indicadores de data profiling, reglas de validación específicas y un puntaje de calidad consolidado asociado a cada activo de datos, el puntaje es un porcentaje de 0 a 100, 100 % significa que en la ingesta no existieron incidentes de calidad. Cada uno de estos indicadores está vinculado a una dimensión de la calidad, lo que facilita la comunicación y comprensión de los incidentes relacionados con la calidad de los datos. En conjunto, este módulo, junto con las valoraciones y opiniones de los usuarios, genera la confianza en la integridad y validez de los datos almacenados.

2.2.6. Comparación de herramientas

El cuadro 2.1 presenta, en las filas, una lista de los criterios evaluados, mientras que las columnas resumen el análisis correspondiente a cada herramienta. Se asigna a cada celda un valor que refleja el grado de cumplimiento de cada criterio, determinado en forma relativa de comparar las distintas herramientas. Los valores que se asignan a cada criterio y herramienta según su

[1] https://university.collibra.com

[2] https://www.collibra.com/resources/expert-panel-maximizing-the-value-of-data-catalogs

[1] https://productresources.collibra.com/docs/collibra/latest/Content/DataNotebook/to_data-notebook.htm

[2] https://productresources.collibra.com/docs/collibra/latest/Content/UnifiedDataQuality/ref_data-quality-concepts.htm

valoración son los siguientes:

- "-" La funcionalidad o aspecto que se está evaluando no está presente en la herramienta o al menos no se pudo verificar que existiese según la información disponible al momento de la evaluación.
- **Bajo**. Se verifica que la funcionalidad o característica existe pero notoriamente carece de capacidades frente a las demás herramientas
- **Medio**. Claramente la funcionalidad está presente en la herramienta y está desarrollada de una manera razonable pero carece de una o más características comparada con las restantes herramientas.
- **Alto**. La funcionalidad o aspecto está implementada ampliamente, incluye características que otras herramientas no la tienen.

Criterio	**OpenMetadata**	**DataHub**	**Purview**	**Collibra**
(1) Interfaz y búsqueda de activos	Bajo	Medio	Medio	Alto
(2) Descubrimiento de activos similares	-	-	-	Bajo
(3) Facilidad de pruebas con datos del Banco	Alto	Medio	-	-
(4) Soporte para con la infraestructura del Banco	Bajo Oracle +12c	Bajo	Bajo PBI RS y Tableau local no verificados.	Bajo PBI RS y Tableau local no verificados.
(5) Detección relaciones tipo linaje	Bajo	Bajo	-	Bajo Solo disponible en la nube
(6) Soporte ingreso manual de data sets	- Limitado a producto de datos	-	Medio	Medio
(7) Visualización de mapas de activos	-	-	-	-
(8) Soporte para social media	Alto Permite hilos de discusión, asignación de tareas y publicación de anuncios	-	Bajo	Medio

Criterio	OpenMetadata	DataHub	Purview	Collibra
(9) Acceso mediante API	Medio	Alto	Medio	Medio
(10) Comunidad, nuevas versiones, documentación y soporte	Medio	Alto Más antigua	Bajo	Medio
(11) Facilidad de consumo de los datos y políticas de acceso	Bajo	-	Bajo	Medio
(12) Calidad, Datos confiables	Bajo	Bajo prioriza entorno nube	Bajo	Medio

Tabla 2.1: Cuadro comparativo de herramientas

Capítulo 3

Análisis y diseño de la solución propuesta

La solución propuesta busca promover la apertura y transparencia de los activos de datos dentro de la organización. En particular, pretende que cualquier integrante de la institución pueda identificar qué datos posee la organización, comprender su significado dentro del contexto de negocio y conocer su ubicación o fuente de origen. De este modo, se favorece un modelo de gestión del conocimiento basado en la democratización del acceso a la información, que facilita la toma de decisiones informadas, el aprovechamiento transversal de los datos y el fortalecimiento de una cultura organizacional orientada al dato ("data-driven culture").

3.1. Problemática actual

En la actualidad, los activos físicos de datos del Banco se encuentran distribuidos en múltiples bases de datos y herramientas informáticas heterogéneas. Si se intentara realizar un inventario exhaustivo de dichos activos, el resultado podría describirse metafóricamente como una "jungla de datos", caracterizada por la enorme cantidad, redundancia y heterogeneidad de los elementos existentes. Sin embargo, solo una fracción relativamente pequeña de estos activos contiene información de verdadero valor para el negocio.

Incluso entre aquellos conjuntos de datos que poseen alto potencial analítico, suele observarse una carencia significativa de documentación o metadatos descriptivos, como nombres o descripciones de tablas y campos, que faciliten

su comprensión e identificación. En consecuencia, los analistas y equipos de negocio enfrentan dificultades para interpretar adecuadamente el significado de los datos, identificar sus relaciones con otros activos o vincular códigos y denominaciones con conceptos del dominio. Esta situación limita el aprovechamiento efectivo de los activos de datos y evidencia la necesidad de un enfoque sistemático de catalogación, estandarización y contextualización semántica de los datos.

La figura 3.1 ilustra de manera conceptual el inventario actual de los activos de datos de la organización. En la parte inferior se observa una "jungla" compuesta por una gran cantidad de objetos desordenados, que simbolizan la multiplicidad de activos de datos alojados en diferentes repositorios. Se busca sintetizar la problemática del exceso de activos físicos con información realmente aprovechable para fines analíticos.

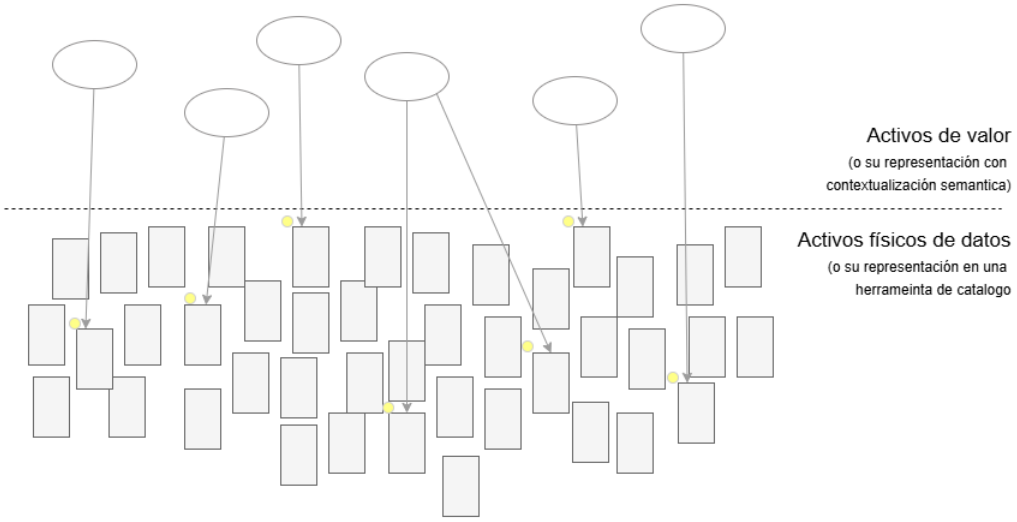


Figura 3.1: Representación conceptual del ecosistema de datos organizacional

En la parte superior, un conjunto reducido de óvalos representa aquellos activos que poseen alto valor para el negocio, es decir, los datos estratégicos o con mayor potencial analítico. Estos óvalos se conectan mediante flechas a determinados rectángulos, determinando cuál o cuáles son los activos físicos que soportan ese activo de valor. En los puntos de conexión aparece un pequeño círculo amarillo que simboliza la "iluminación" del dato, es decir, el momento en que un activo que estaba escondido adquiere contexto, significado y utilidad analítica o de otra índole para la organización.

3.2. Linaje de los datos

Desde las primeras etapas del relevamiento de los cubos se constató que identificar las fuentes de datos que los alimentan es fundamental para comprender íntegramente la información analizada. Una vez determinadas estas fuentes primarias, resulta igualmente relevante identificar las transformaciones intermedias y las fuentes que las preceden, conformando así una cadena de derivaciones sucesivas. Este encadenamiento constituye el flujo de datos o linaje, funcionalidad evaluada en las herramientas analizadas en el presente trabajo, bajo el criterio denominado **(5) Detección de relaciones de tipo linaje**, sección 2.2.1. Conocer esta traza contribuye a mejorar la confianza de los usuarios finales, ya que permite entender cómo los datos evolucionan a lo largo de las distintas etapas de transformación hasta conformar la información utilizada en el análisis.

En paralelo a la identificación y el análisis de los orígenes de datos de los cubos, se procedió a la instalación y evaluación de herramientas de catálogo de datos sobre las bases de datos que los alimentan, con el objetivo de verificar su capacidad para identificar orígenes y linaje entre distintos activos de datos, así como la evaluación de otras funcionalidades definidas en la sección 2.2.1. Las herramientas analizadas fueron OpenMetadata y DataHub, y los resultados de dicha evaluación se presentan en las secciones 2.2.2 y 2.2.3 respectivamente.

Las pruebas realizadas no arrojaron resultados satisfactorios en términos de linaje. Ello se debe, principalmente, a la tecnología utilizada en las transformaciones: procesos ETL implementados mediante PL/SQL en Oracle, frente a los cuales las herramientas evaluadas mostraron una capacidad limitada de descubrimiento y detección. Tampoco se obtuvieron resultados relevantes en la identificación automática de orígenes de datos para los cubos. El único caso exitoso se registró con OpenMetadata, que logró descubrir los activos (tablas y vistas) utilizados para un tablero de Tableau preparado específicamente para la prueba; sin embargo, esta tecnología no está siendo adoptada por el Banco, por lo cual no es un resultado que se pueda generalizar.

A partir de las pruebas y análisis realizados se constató que un linaje generado automáticamente por una herramienta, si bien tendría un valor potencial elevado, produciría un resultado excesivamente detallado: un grafo de gran escala que reflejaría las relaciones entre una multitud de activos físicos distribuidos en distintas bases de datos, muchos de ellos de carácter estrictamente

operativo.

Frente a esta complejidad, un enfoque top-down centrado en los conjuntos de datos del catálogo y en los flujos de información entre ellos resulta más adecuado para generar valor para el negocio. Este enfoque permite abstraer la "jungla" de elementos físicos y focalizar en las dependencias relevantes entre conjuntos, evitando la sobrecarga de detalle propia del linaje automatizado.

En el siguiente capítulo, sección 4.2, se desarrolla una posible aplicación práctica de este enfoque top-down para la identificación de los conjuntos de datos y los flujos que los interrelacionan.

3.3. Propuesta de solución

El presente trabajo propone un enfoque de descubrimiento de datos que parte desde un nivel alto de abstracción hasta llegar a los elementos físicos que sustentan dichos datos.

La figura 3.2 ilustra los tres niveles de abstracción que estructuran la arquitectura propuesta. El **nivel conceptual**, ubicado a la derecha, está conformado por dos tipos de ontologías representadas en planos diferenciados: la ontología del plano superior modela el conocimiento del dominio, mientras que la del plano inferior describe y contiene la estructura del catálogo de datos. A continuación, el **nivel de metadatos**, ubicado en el centro, actúa como una capa intermedia, compuesta principalmente por los metadatos de los activos descubiertos mediante una herramienta de catálogo de datos, que vinculan las entidades conceptuales con sus correspondencias operativas. Finalmente, el **nivel físico** agrupa los artefactos y herramientas que almacenan, procesan o analizan la información, materializando así las entidades definidas en los niveles superiores y cerrando el ciclo de representación entre el modelo conceptual y su implementación técnica.

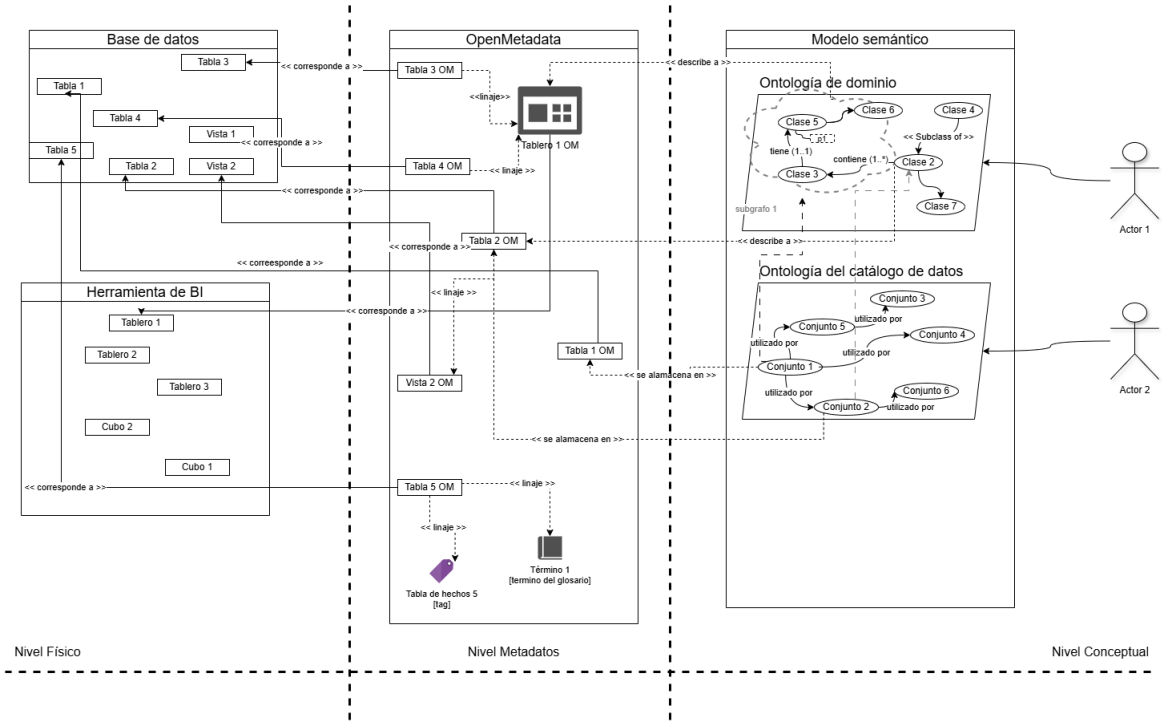


Figura 3.2: Niveles de abstracción

3.3.1. Nivel conceptual

El nivel conceptual permite relevar y modelar los principales datos de la organización. Existen dos perspectivas para realizar esto, materializadas en dos ontologías de enriquecimiento.

3.3.1.1. Ontología con conocimiento del dominio

La ontología representada en el plano superior permite establecer un lenguaje y vocabulario común para los diferentes conceptos que se intercambian al describir los datos. Mediante un trabajo conjunto entre los equipos de negocio y tecnología es posible relevar dichos conceptos y formalizarlos en una ontología compuesta por clases, propiedades y relaciones. Las clases describen las principales entidades del dominio, representadas en la figura mediante elipses, mientras que las flechas indican las relaciones entre ellas. Este modelo se construye utilizando una semántica formal la cual tiene una representación de tipo grafo como se muestra en la figura, lo que permite contrastar, validar y consensuar el modelo con los distintos actores involucrados, asegurando así una representación compartida y fácil de transmitir a otros actores de la

organización.

Otra ventaja del modelado semántico es la posibilidad de reutilizar ontologías consolidadas y ampliamente aceptadas como estándares, tales como *OWL*[1] y *RDFS*[2] para la definición estructural, o *SKOS*[3] para la representación de la terminología específica del dominio.

3.3.1.2. Ontología del catálogo de datos

La ontología representada en el plano inferior de la figura 3.2 describe el catálogo de datos y pertenece a una dimensión distinta de la ontología de dominio. En términos de modelado, es más sencilla, ya que incluye un número reducido de definiciones, con menos clases y relaciones. La clase principal del modelo representa un elemento del catálogo, es decir, un conjunto de datos, mientras que la relación central que se busca modelar en el alcance de este trabajo es el uso o la transformación de datos que ocurre entre un conjunto y otro. Otras clases complementarias incluyen el área o grupo responsable del conjunto de datos y su ubicación, entre otras.

A diferencia de la ontología de dominio, el objetivo principal de esta ontología es contener las instancias, también llamadas individuos, de las clases, actuando como el nivel donde se materializan los elementos concretos del catálogo.

Los conjuntos de datos, individuos de esta ontología, pueden describirse mediante una o más clases de la ontología de dominio del plano superior. Mientras la ontología del catálogo representa los conjuntos de datos desde la perspectiva de sus relaciones con otros conjuntos e individuos, como el grupo responsable o la ubicación, la ontología de dominio aborda el detalle de las entidades de negocio y los atributos que pueden estar presentes en cada conjunto de datos. Por esta razón, ambas ontologías se sitúan en planos o dimensiones conceptuales diferentes.

3.3.1.3. Vinculación entre ambas ontologías

En la propuesta, cada conjunto de datos, representado como un individuo en la ontología del catálogo, se vincula con uno o varios conceptos de la on-

[1] https://www.w3.org/TR/owl2-syntax/
[2] https://www.w3.org/TR/rdf-schema/
[3] https://www.w3.org/TR/skos-reference/

tología de dominio. Esta vinculación se establece con un subgrafo del dominio que incluye las clases relevantes, sus relaciones y sus propiedades de datos.

Así, un conjunto de datos queda descrito semánticamente por las entidades y atributos de negocio que representa. Esto permite entender qué conceptos del dominio están presentes en el conjunto y cómo se conectan entre sí, haciendo explícito el significado del contenido del conjunto de datos.

De esta forma, la ontología del catálogo aporta el inventario de conjuntos de datos, sus metadatos y las relaciones operativas que los vinculan, como el linaje entre datasets, mientras que la ontología de dominio proporciona el contexto semántico necesario para interpretarlos.

3.3.2. Nivel de metadatos

3.3.2.1. Activos y relaciones descubiertas

El nivel de metadatos está compuesto, en primer lugar, por la información extraída por una herramienta de catálogo de datos a través de su módulo de data discovery. Este módulo puede obtener metadatos de los activos físicos alojados en bases de datos relacionales, herramientas de inteligencia de negocio u otros tipos de sistemas de almacenamiento, según las capacidades de la herramienta y la compatibilidad con las tecnologías y versiones implementadas en la organización.

La capacidad del módulo de data discovery, en conjunto con la infraestructura tecnológica disponible, determina el alcance de la identificación de relaciones entre activos, tales como relaciones de linaje de datos, frecuencia de acceso o detección de procesos de transformación y carga.

3.3.2.2. Elementos creados por el usuario

Además de los activos descubiertos automáticamente, estas herramientas permiten incorporar elementos creados por el usuario que enriquecen los metadatos existentes. Como se muestra en la figura 3.2, y según las funcionalidades de cada herramienta, estos elementos pueden incluir etiquetas (tags) para categorizar activos o términos de glosario destinados a uniformar el significado de un activo, tabla o atributo.

Asimismo, es posible definir dominios que representan diferentes áreas de conocimiento dentro de la organización, a los cuales se pueden asociar pro-

ductos de datos como reportes o indicadores. Algunas herramientas también permiten aplicar clasificaciones estándar sobre los activos, por ejemplo, para identificar si un dato es personal o contiene información sensible.

Finalmente, es importante destacar que todos estos metadatos son gestionados y almacenados por la herramienta de catálogo, y que, al momento de este trabajo, no existe un modelo o estructura estándar que permita el intercambio o migración de estos metadatos entre distintas herramientas.

3.3.3. Nivel físico

Finalmente, en el área izquierda de la figura 3.2 se encuentra el nivel físico, compuesto por los activos físicos que conforman la base operativa de la arquitectura. Estos activos incluyen los artefactos que almacenan datos y aquellos que soportan los procesos de transformación, carga, análisis o publicación de información.

En términos de volumen, el nivel físico del Banco puede comprender desde cientos hasta miles de tablas, dependiendo del conjunto de bases de datos considerado. Esta magnitud permite dimensionar la complejidad del entorno y justifica que la estrategia de descubrimiento más eficiente consista en partir desde los conjuntos de datos o clases relevantes definidos en las ontologías de los niveles superiores, en lugar de analizar cada activo físico de manera individual. Este enfoque facilita identificar la relevancia de los activos en función de las entidades y datos gestionados por los equipos de negocio, optimizando el proceso de descubrimiento y priorización.

3.3.4. Modelo de vinculaciones

Además de identificar y abstraer los distintos activos de datos y sus relaciones en cada nivel, es fundamental vincular los elementos que componen los diferentes niveles de la arquitectura. Estos vínculos permiten conectar los objetos definidos en el nivel conceptual con su representación mediante metadatos en el nivel intermedio y, finalmente, con los elementos físicos que los materializan en el nivel inferior. A efectos de simplificar la explicación que sigue, el nivel de metadatos no se menciona de forma explícita: dado que existe una correspondencia biunívoca entre los elementos de dicho nivel y los elementos físicos, ambos pueden tratarse de manera indistinta sin pérdida de precisión.

La figura 3.3 resume las vinculaciones entre los elementos del nivel conceptual y con los elementos de los niveles inferiores, elementos físicos. Como se muestra en la figura, se pueden dar tres tipos de vinculaciones entre (i) los elementos de la ontología con conocimiento del dominio, (ii) con elementos de la ontología que describe el catálogo, y (iii) con elementos o artefactos que se encuentran en el nivel físico.

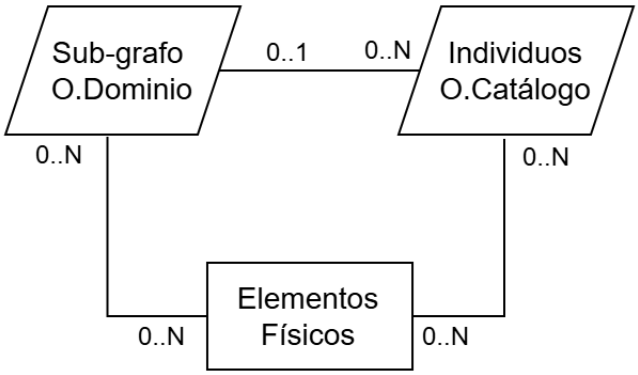


Figura 3.3: Vinculación entre ontologías y nivel físico

En el primer caso (i) un sub-grafo, compuesto por una o más de una clase de la ontología de dominio, se puede vincular con uno o más individuos de la ontología del catálogo (ii), donde cada individuo es un conjunto de datos, de esta forma un sub-grafo describe de forma semántica a uno o más conjuntos de datos, indicando relaciones entre entidades, y describiendo atributos de las mismas, que en términos ontológicos son *object properties* y *data properties* respectivamente. Por ejemplo un conjunto `instituciones-contables`, que contiene las instituciones del sistema contable, se puede vincular a un sub-grafo que consiste en las clases `institucion` y `cuenta`, donde ambas clases están relacionadas por la object property `pertenece`, de dominio la clase `cuenta` y rango la clase `institucion`, con cardinalidad `1..1` y `1..1` respectivamente, es decir a una cuenta pertenecer a una institución y solo una. Adicionalmente, asociadas a estas clases pueden existir las data properties `nombre` y `RUT` (identificador fiscal) para la clase `institucion` y las propiedades `nombre` e `identificador` para la clase `cuenta`, con dominio la clase correspondiente y literal como rango cada propiedad.

En el segundo caso (ii) un individuo de la ontología de catálogo, es decir uno o más conjuntos de datos se puede vincular a uno o más de un elementos físicos del nivel inferior. Con lo cual uno o varios conjuntos de datos, se puede vincular

a una tabla de base de datos, un tablero de análisis, un reporte publicado, entre otros. A través de esta vinculación se puede conocer por ejemplo sobre qué base de datos y qué tabla o tablas un conjunto de datos se aloja.

Por último, el tercer caso (iii), un elemento físico se puede vincular con un sub-grafo de clases de la ontología de dominio. En este caso una o más clases que describen un concepto de negocio se vinculan a uno o más elementos físicos. Mientras un individuo de la ontología de catálogo puede describir el dato del elemento físico vinculado, por ejemplo describir su granularidad mínima en el tiempo (día, mes, trimestre, etc) o desde cuando se tiene registros, una clase de la ontología de dominio describe los datos del elemento físico desde el punto de vista de sus relaciones y sus atributos, por ejemplo extendiendo una clase de una ontología de dominio reconocida o utilizando un vocabulario consensuado entre varios actores.

Capítulo 4

Prueba de concepto

La prueba de concepto tuvo como objetivo verificar, de manera simple y acotada, si el uso combinado de ontologías y metadatos, enmarcado en el esquema de tres niveles de abstracción presentado en el capítulo 3, permite identificar y describir los activos de datos del Banco. En este capítulo, en primer lugar, se delimita la problemática a un caso de estudio relevante para el trabajo, incorporando además la retroalimentación derivada de la iniciativa de gobierno de datos que el Banco se encuentra llevando adelante. A continuación, la segunda mitad del capítulo se centra en los aspectos de implementación de la propuesta, con especial énfasis en la obtención de una visualización en forma de mapa de datos. De este modo, el prototipo implementado busca dar respuesta a la pregunta compuesta: ¿qué datos posee la institución y cómo se relacionan entre sí?

4.1. Caso de estudio y contexto institucional

El caso de estudio se delimitó dentro del área de Política Monetaria, específicamente en el sector de Análisis Financiero. Este sector desarrolla múltiples análisis de información, algunos de ellos establecidos por la Carta Orgánica del Banco.[1]

Si bien diversas áreas de la institución presentan demandas significativas de información, incluyendo informes periódicos orientados a la toma de decisiones y al intercambio con organismos nacionales e internacionales, se seleccionó el sector de Análisis Financiero debido a la experiencia acumulada en el desarrollo

[1] https://www.bcu.gub.uy/Acerca-de-BCU/Normativa/Documents/Carta-organica/carta_organica_toco.pdf

de infraestructura analítica y a la importancia que asigna al uso y la gestión de datos en sus procesos.

4.1.1. Relevamiento y datos de interés

El sector de Análisis Financiero cuenta con múltiples cubos de información y un data warehouse que sustentan sus procesos analíticos. En el último año, ha incorporado además tableros de visualización, desarrollados directamente por los analistas, a diferencia de los cubos, cuya construcción estuvo a cargo del área de informática.

Para este trabajo, y con el objetivo de comprender el alcance de los análisis que realiza el sector e identificar un caso de estudio adecuado, se relevaron en detalle las dimensiones, medidas y fuentes de información asociadas a un subconjunto de dichos cubos.

Se identificaron las tablas y sentencias SQL del data warehouse utilizadas para la construcción de los cubos, los cuales se desarrollan en la herramienta Ideasoft O3[1]. Para determinados cubos se analizó en detalle el proceso mediante el cual se obtienen sus datos fuente, relevando los procedimientos de carga de las tablas de hechos y dimensiones que los sustentan. Este análisis permitió reconocer otras tablas del propio data warehouse, así como tablas de bases de datos externas que actúan como fuentes de información. La identificación de estos flujos y patrones de transformación se realizó con el apoyo de la contraparte informática del área. En términos generales, las cargas y transformaciones del data warehouse se implementan mediante procedimientos PL/SQL de Oracle, ejecutados tanto en la base del data warehouse como en otras bases Oracle accedidas a través de *database links*.[2] Los *database links* (dblinks) en Oracle, son objetos definidos en una base de datos que permiten acceder a objetos ubicados en otra base de datos como si fueran locales, por ejemplo consultar a tablas, invocar a procedimientos o acceder a otros objetos remotos de manera transparente para el usuario. Estas bases de datos remotas son parte en general de otros sistemas y difieren en la versión del motor de base de datos.

A medida que se avanzó en el relevamiento, el análisis se centró en dos cubos del sector. Además de identificar sus fuentes de información y procesos

[1] https://www.ideasoft.biz/productos/
[2] https://docs.oracle.com/database/apex-18.1/AEUTL/managing-db-links.htm

de carga, se buscó describir los análisis que habilitan sus medidas y dimensiones, cada uno con un orden aproximado de 15 a 20 elementos. Sin embargo, al tratarse de modelos multidimensionales, los cubos presentan información consolidada (agregada) por período. Por lo tanto, el dato original que alimenta esa consolidación no puede recuperarse directamente del cubo ni de las tablas de hechos asociadas; su trazabilidad exige retroceder en los procesos de transformación previos. La descripción de un dato consolidado resulta insuficiente para caracterizar la diversidad de registros que lo generan. Por ejemplo, una cuenta contable de pasivo no refleja por sí sola los distintos hechos económicos que la componen.

En consecuencia, en lugar de considerar los cubos como objeto de análisis, se optó por focalizar en determinados indicadores de negocio. Estos indicadores constituyen variables presentes tanto en reportes y publicaciones institucionales como en los propios cubos; sin embargo, a diferencia de estos últimos, responden a un criterio y una definición específica, siendo unidades de análisis más precisas y trazables.

4.1.2. Necesidades relevadas

Durante este proceso se evidenció que la infraestructura relacional del Banco contiene una gran cantidad y diversidad de elementos físicos (tablas, vistas y sentencias SQL asociadas), muchos de ellos destinados a fines operativos. Para evitar esta "jungla" (concepto presentado en la sección 3.1) de objetos y permitir la correcta selección de datos adecuados para el análisis, se vuelve necesario diferenciar, identificar y contextualizar explícitamente aquellos elementos que efectivamente sirven como insumo para procesos analíticos.

En paralelo al relevamiento y búsqueda del caso de estudio, el sector identificó dos necesidades complementarias: (1) mejorar la comunicación de sus análisis mediante visualizaciones más actuales que las provistas por Ideasoft O3, y (2) fortalecer la autonomía analítica, de modo que los propios analistas, en este caso, economistas, pudieran desarrollar tableros y reportes sin depender del área de tecnología. A partir de los avances del presente trabajo, fue posible proveer a estos analistas los orígenes de datos y metadatos relevados, facilitando la transformación de ciertos cubos de O3 en tableros de Microsoft Power BI orientados a nuevos requerimientos de visualización y gestión de los mismos.

4.1.3. Selección del caso de estudio

Una vez completada la etapa de relevamiento, en la que se analizaron en detalle algunas de las fuentes de información existentes, la lógica de transformación de estos datos y el aporte de los cubos multidimensionales al análisis (algunos ejemplos de este relevamiento se muestran en el anexo 1), se trabajó junto a los analistas en la identificación de los indicadores de mayor relevancia para el sector. De este proceso surgieron tres indicadores o conceptos candidatos: activos de reserva, base monetaria y deuda pública.

Los **activos de reserva** o reservas internacionales son los recursos financieros externos que el Banco mantiene para asegurar la estabilidad económica. La **base monetaria** es la cantidad total de dinero creada directamente por el Banco como autoridad monetaria del país. La **deuda pública** es el conjunto de obligaciones financieras que un Estado mantiene con sus acreedores.

Los primeros dos indicadores comparten un origen de datos similar, principalmente el Balance del Banco Central. La deuda pública, en cambio, se construye a partir de múltiples fuentes de información, algunas internas y otras externas al Banco. Además, para ciertos rubros no se dispone de todas las fuentes necesarias, por lo que parte de la información se completa mediante estimaciones.

Dado que el indicador de deuda pública depende de diversas fuentes, varias de ellas aún no incorporadas al data warehouse del sector y, por tanto, fuera del alcance del relevamiento realizado, su modelado exigiría un trabajo más exhaustivo que excede los objetivos de este estudio. En consecuencia, se optó por seleccionar uno de los otros dos indicadores, definiéndose como caso de estudio: el indicador de **base monetaria**.

Ontologías para base monetaria

La base monetaria, como indicador, cuenta con más de una publicación asociada con frecuencia diaria y dispone asimismo de un cubo principal que permite su análisis mediante datos consolidados al cierre de cada mes. Para avanzar en su modelado conceptual, se elaboró una ontología de dominio destinada a describir los datos que intervienen en su medición. Paralelamente, se construyó una ontología del catálogo orientada a representar los conjuntos de datos que conforman el indicador de base monetaria, así como aquellos que mantienen vínculos con el indicador.

En el proceso de describir los datos subyacentes mediante una ontología de dominio, se relevaron ontologías existentes aplicables al ámbito financiero. En esta búsqueda se identificó la ontología **FIBO** (Financial Industry Business Ontology)[1], desarrollada por la *Enterprise Data Management Council* en conjunto con la *Object Management Group*. FIBO abarca más de 60 módulos y contiene aproximadamente 3200 clases y 300 propiedades de datos.

No obstante, dentro de la documentación disponible, en particular en el módulo correspondiente al subdominio de balances contables, no se identificaron representaciones para algunos de los conceptos presentes en el balance del Banco, ni se encontraron explicaciones sobre las decisiones de modelado adoptadas por dicha ontología. En consecuencia, se descartó la adopción de FIBO en el presente trabajo.

Para la ontología de catálogo destinada a describir los distintos conjuntos de datos vinculados a la base monetaria, se identificó y adoptó el estándar *Data Catalog Vocabulary (DCAT)*[1] de *W3C*[2], utilizado en portales internacionales de datos.

4.1.4. Catálogo de datos desarrollado por el Banco

En el último año, y mediante un proceso consensuado entre diversas áreas, el Banco inició un proyecto para definir y construir un catálogo institucional de datos, con el objetivo de inventariar y describir los diferentes conjuntos de datos existentes. Durante este proceso, la institución mantuvo intercambios con otros bancos centrales que se encuentran en etapas más avanzadas de iniciativas similares.

De estos intercambios surgió una observación relevante: los primeros pasos de dichas instituciones fue que los propios usuarios de negocio lideraron la tarea de identificación de los datos y de los metadatos que resultan funcionales a sus objetivos. En consecuencia, el catálogo se construye desde una perspectiva centrada en el negocio, en contraste con un enfoque basado exclusivamente en los activos físicos identificados automáticamente por herramientas de catalogación.

Algunos de los atributos de metadatos definidos por el Banco para el catálogo incluyen un identificador, nombre y descripción para cada entrada, entendi-

[1] https://spec.edmcouncil.org/fibo/
[1] https://www.w3.org/ns/dcat
[2] https://www.w3.org/

da esta como un conjunto de datos. También se incorporaron la frecuencia de elaboración (diaria, mensual, trimestral, etc.) el estado de procesamiento (dato crudo o transformado), los roles asociados, como unidad de negocio dueña y unidad custodia, el nivel de sensibilidad, la presencia de datos personales, y la existencia de convenios interinstitucionales, entre otros. En total se establecieron 44 atributos de metadatos, los cuales se gestionan en una planilla Excel centralizada. Al momento de elaborar este informe, el catálogo superaba los 500 conjuntos de datos registrados.

Finalizando la construcción del catálogo, el Banco llevó adelante un proceso de reflexión estratégica orientado al próximo quinquenio, con participación transversal de todas las áreas. Una de las mesas de trabajo creadas se centró en el gobierno de datos y en la elaboración de insumos para la estrategia institucional en datos 2025–2030.

Como resultado, se definió una orientación estratégica hacia la toma de decisiones basada en datos, y entre las líneas de acción acordadas se incluyó la catalogación de los datos, priorizando en el corto plazo aquellos que el Banco recibe y aquellos que publica o comparte con otras instituciones, tanto locales como internacionales.

En la fase final del proceso se destacó la importancia de catalogar las "interrelaciones" entre los datos, priorizadas desde la visión del negocio. En el marco del presente trabajo, estas interrelaciones se interpretan como el flujo o linaje de los datos.

En las instancias de trabajo con los analistas referentes que participaron en la construcción del catálogo del Banco, se propuso incorporar una relación entre conjuntos de datos que permitiera representar el linaje entre ellos. Sin embargo, esta incorporación no se concretó debido a las restricciones de tiempo junto con que su definición requiere un relevamiento más profundo que el necesario para los demás atributos de metadatos. Adicionalmente, la herramienta utilizada para mantener el catálogo, una planilla Excel, resulta altamente manual y, en muchos casos, poco adecuada para gestionar relaciones entre entidades, lo que constituyó otra limitación para avanzar en esta etapa.

4.2. Implementación del caso de estudio

En conjunto con el relevamiento y análisis realizados para la selección del caso de estudio (sección 4.1), se fue estableciendo cómo cada elemento de meta-

datos del caso elegido puede representarse dentro de la propuesta de solución.

En las secciones siguientes se presenta, en primer lugar, un marco general de la implementación del caso de estudio y, posteriormente, se profundiza en uno de los componentes de metadatos concebidos para enriquecer la descripción de los datos: el linaje entre conjuntos de datos.

4.2.1. Arquitectura y componentes

El primer paso consiste en ubicar los distintos componentes del caso de estudio, base monetaria, dentro de los tres niveles de abstracción definidos en en el capítulo 3. La arquitectura de la Figura 3.3 se instancia para este caso de estudio en la Figura 4.1.

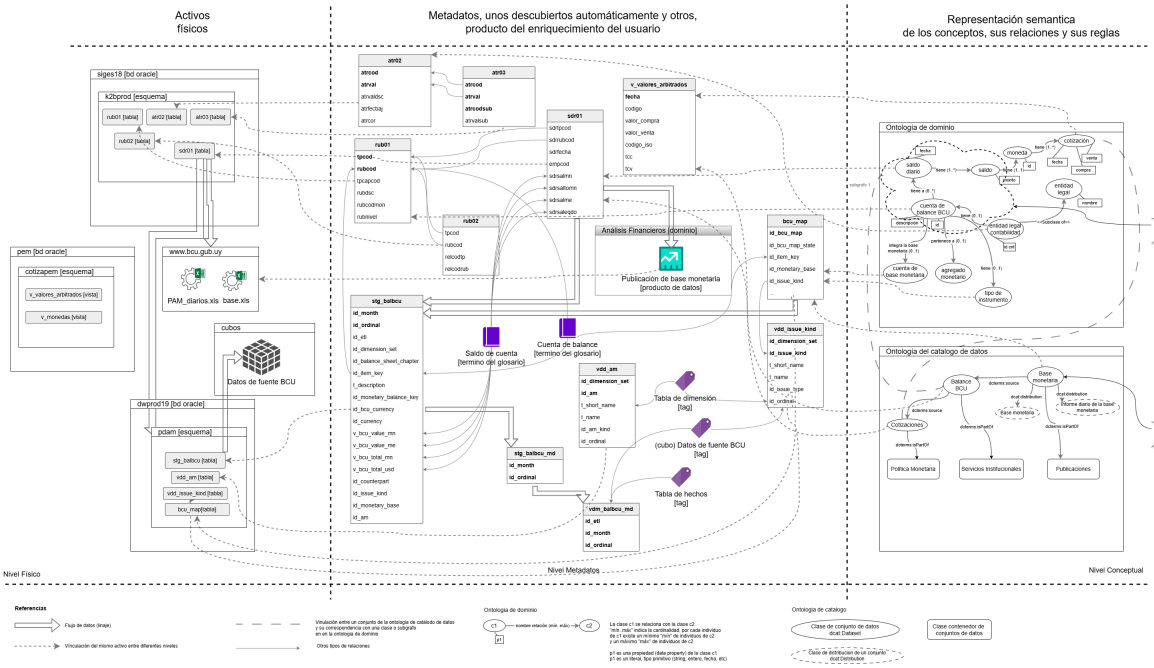


Figura 4.1: Esquema en los tres niveles

La figura 4.1 ilustra cómo un usuario analista puede navegar entre los distintos componentes de metadatos asociados al indicador de base monetaria, ya sea partiendo de la ontología de dominio, que describe mediante clases los datos que originan el indicador, o bien desde la ontología de catálogo, comenzando por alguno de los conjuntos de datos que corresponden a las publicaciones del indicador en el sitio web del Banco.

Ontología de dominio

En la ontología de dominio mostrada en la figura 4.2, el `Actor 1` inicia su análisis en el `subgrafo 1`, vinculado al conjunto `Balance BCU` de la ontología de catálogo. El balance se modela esencialmente a través de la clase `cuenta de balance BCU`, cuyos elementos, de acuerdo con la relación y restricción de cardinalidad `"tiene a (0..*)"`, cada cuenta de balance puede eventualmente tener uno o más `saldo diario`, y con su respectiva `fecha` como data property.

A su vez, la clase `saldo diario` puede presentar uno o más saldos, según la relación y restricción `"tiene a (1..*)"`, lo cual permite representar balances en distintas monedas.

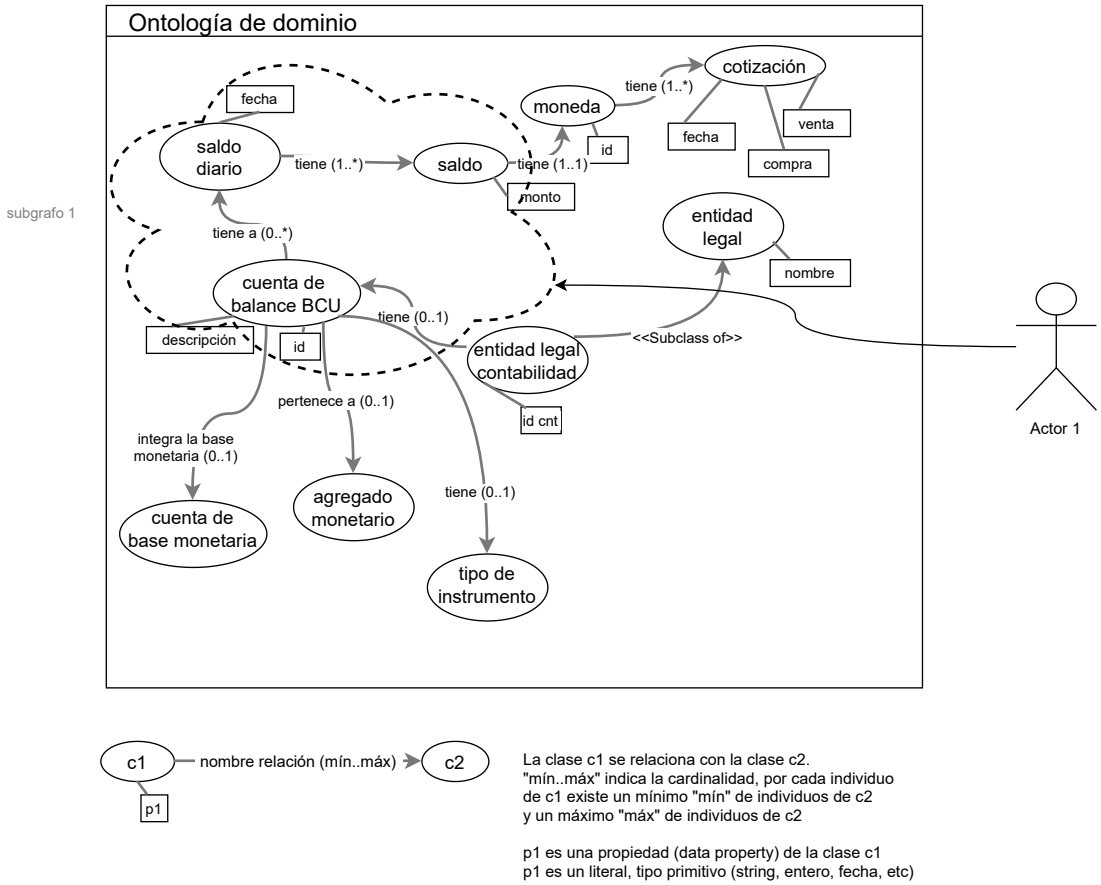


Figura 4.2: Ontología de dominio

Este modelado permite al analista comprender cómo se descompone el Balance del BCU en sus distintas entidades de datos. La identificación de las cuentas que integran la base monetaria se representa mediante la relación `"integra la base monetaria (0..1)"` y la entidad

"cuenta de base monetaria", las cuales señalan aquellas cuentas definidas para formar parte del indicador de base monetaria. El modelo de vinculaciones guía al usuario hacia la tabla bcu_map, cuya navegación se detalla en la sección 4.2.1.

Asimismo, la figura muestra la presencia de otras clases relacionadas con los componentes del balance. Por ejemplo, la clase entidad legal contabilidad, que representa el catálogo de instituciones del sistema contable. Mediante la relación «Subclass of», se indica que esta clase deriva de entidad legal, la cual representa una institución en sentido general, independiente del sistema contable, pero igualmente caracterizada como sujeto jurídico con obligaciones y derechos contractuales potenciales.

Ontología de catálogo

La figura 4.3 presenta los conjuntos de datos instanciados a partir de la ontología de catálogo aplicada al caso de estudio. Un usuario analista (Actor 2) puede iniciar su análisis desde la publicación del indicador Base monetaria e identificar, mediante la relación de tipo *object property* dcterms:source, que el conjunto Balance BCU constituye una de sus fuentes de información.

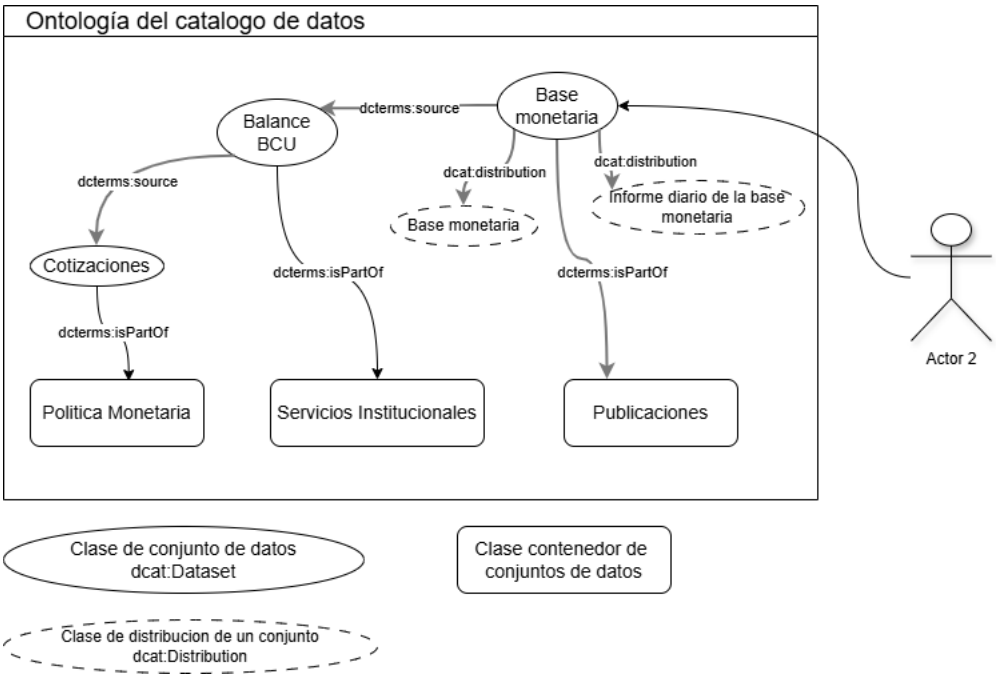


Figura 4.3: Ontología de catálogo

Del mismo modo, el modelo permite reconocer que el conjunto

Cotizaciones también actúa como fuente, evidenciando la existencia de un flujo de datos desde Cotizaciones hacia Balance BCU, y desde este último hacia Base monetaria. Esta representación facilita la comprensión de las interrelaciones entre conjuntos de datos desde una perspectiva centrada en el negocio. A su vez el modelo a través de la relación dcterms:distribution, muestra en contornos punteados, que base monetaria se representa en dos distribuciones, Base monetaria e Informe diario de la base monetaria, que conducen a dos archivos físicos diferentes que se publican.

En la parte inferior de la figura 4.3 se muestran tres rectángulos con esquinas curvas, que a través de la relación dcterms:isPartOf permite ubicar u organizar los conjuntos de datos: Cotizaciones, Balance BCU y Base monetaria, en los contenedores: Política Monetaria, Servicios Institucionales y Publicaciones respectivamente. Los dos primeros contenedores son áreas del Banco, mientras que "Publicaciones" conceptualmente refiere a un repositorio o lugar físico donde los conjuntos de datos son accesibles para todo público. En la sección 4.2.2.3 se detalla que son los contenedores en el contexto de este trabajo.

Nivel metadatos

En la figura 4.4 se presenta el nivel de metadatos, concretamente los metadatos de los elementos físicos del caso base monetaria elegido para el estudio. Como se describió la sección 3.3.2, parte de estos metadatos son extraídos automáticamente por la herramienta de catálogo y descubrimiento, mientras que otra parte corresponde a metadatos creados o enriquecidos manualmente por los usuarios.

Para las principales clases del balance, (subgrafo 1) en figura 4.2, se identifican las tablas que almacenan la información del mismo : rub01, rub02 y sdr01. Esta última se encuentra enriquecida con un glosario de términos incorporado por el usuario, denominado Saldo de cuenta. Dicho término describe que la tabla contiene cuatro campos de saldo, diferenciados según la moneda (nacional o extranjera) y una totalización, lo que permite interpretar mejor la estructura de los saldos registrados. De forma similar también se encuentra definido el glosario de término Cuenta de balance, que informa los identificadores de cuentas de balance con distintos nombres según el sistema o base de datos, rubcod en un caso e id_item_key en el otro.

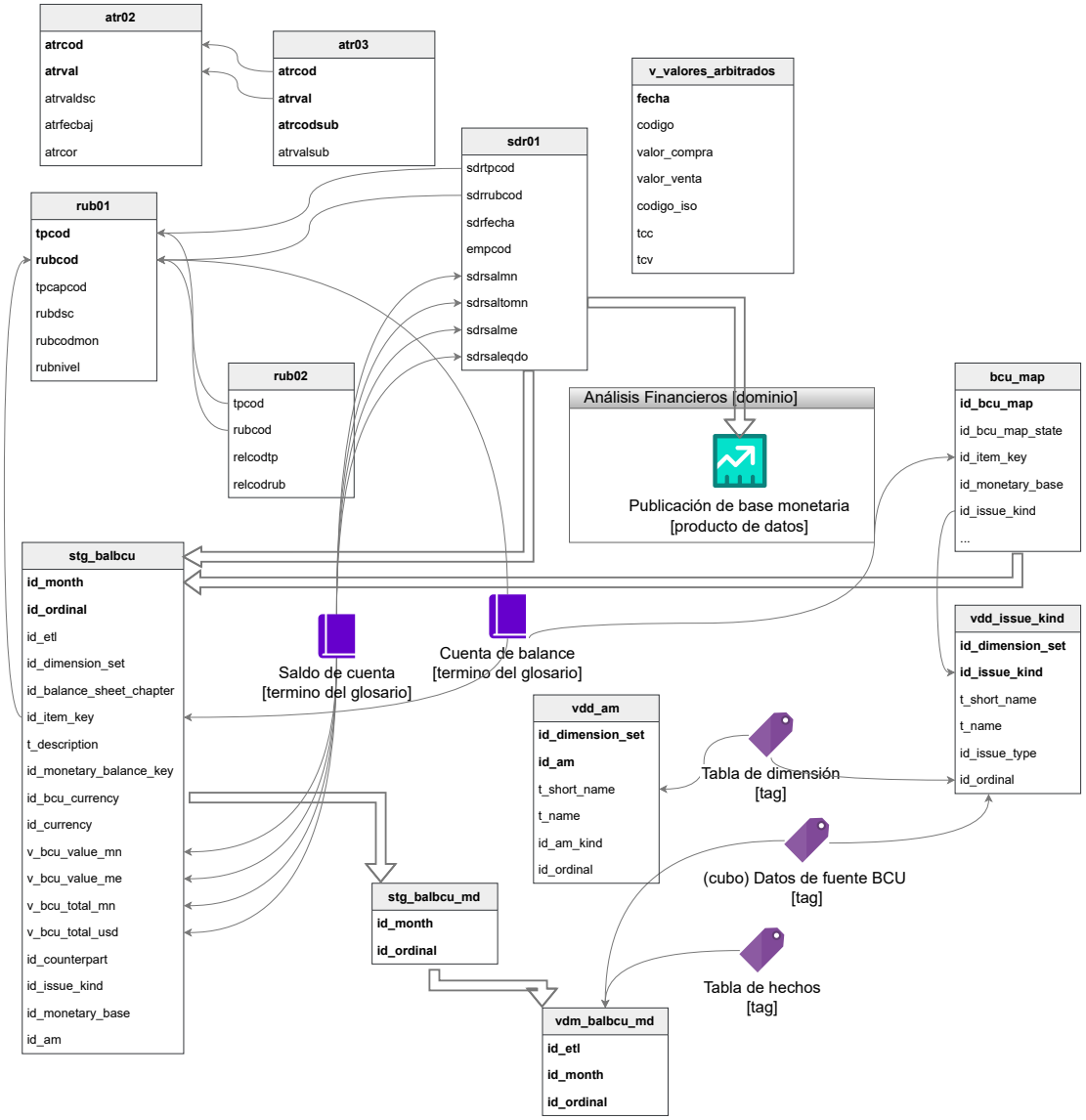


Figura 4.4: Nivel Metadatos

A partir de los vínculos definidos para el término `Saldo de cuenta` y la relación de tipo flujo de los datos (referencias descritas en la figura 4.1), se identifica que la tabla `stg_balbcu` se carga a partir de la tabla `sdr01`, también utilizando la tabla `bcu_map`. Esta última tabla, `bcu_map`, tiene el rol de asociar las cuentas de balance con otros conceptos, tales como si la cuenta integra o no el indicador de base monetaria, relacionarlas con tipos de instrumentos, entre otros conceptos. La tabla `stg_balbcu`, por su relación con otras tablas y las etiquetas visibles en el diagrama (`[tag]`), constituye una tabla de origen para la tabla de hechos del cubo `Datos de fuente BCU`, utilizado para analizar el

balance al cierre de cada mes.

Cabe destacar que esta información surge de un enriquecimiento manual realizado por un usuario con conocimiento tanto de la tecnología empleada como de las relaciones entre los elementos físicos, y no de un descubrimiento automático provisto por la herramienta de catalogación. El alcance de dichas herramientas se describe en la sección 2.2, y en particular el descubrimiento del flujo o linaje de los datos para el caso de estudio se analiza en la sección 3.2.

En la Figura 4.5 se presenta la interfaz de usuario de OpenMetadata. En ella se observa el menú principal de navegación y, en particular, la vista correspondiente a la tabla `vdm_balbcu`, junto con las distintas informaciones de metadatos que la herramienta ha extraído y asociado a dicho objeto.

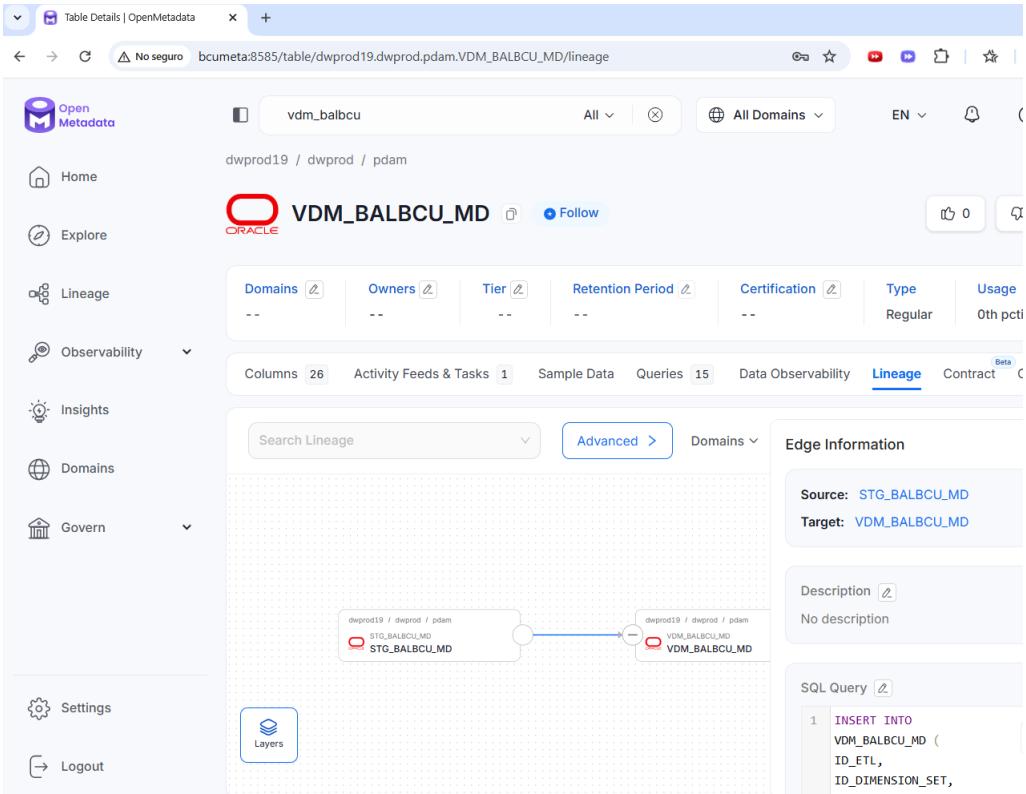


Figura 4.5: Captura de OpenMetadata

La figura permite apreciar el tipo de información que gestiona una herramienta de catálogo de datos, distinguiéndose entre metadatos generados automáticamente y aquellos incorporados manualmente por los usuarios. Entre estos últimos se destaca la sección `Govern`, donde se administran glosarios

de términos y etiquetas (*tags*), tal como se ilustra en la Figura 4.4. Asimismo, dentro del menú contextual de la tabla seleccionada, se observa la opción `Activity Feeds & Task`, correspondiente a funcionalidades de colaboración entre usuarios, en línea con el criterio (8) de soporte para *social media* definido en la sección 2.2.1. El indicador mostrado (valor 1) señala la existencia de una actividad reciente, en este caso, la incorporación de una etiqueta a la tabla.

En cuanto a los metadatos generados automáticamente, se destaca la pestaña `Columnas`, que presenta los atributos de la tabla, constituyendo una información esencial del catálogo. Por su parte, la pestaña `Queries` reúne las consultas *SQL* ejecutadas sobre la tabla, información que es extraída periodicamente a partir de los *query-logs* del motor de base de datos (Oracle). Finalmente, la pestaña `Lineage`, que se muestra en la figura, representa el linaje detectado a partir de las sentencias ejecutadas entre tablas. Si bien esta funcionalidad resulta de gran valor por su potencial analítico, su alcance es parcial; por ejemplo, no se visualiza el linaje con la tabla `stg_balbcu`. Para un mayor detalle de las opciones disponibles en la herramienta, se puede consultar la documentación oficial.[1]

Vinculación entre los elementos

En el cuadro 4.1 se puede observar cómo las clases y conjuntos de datos del nivel conceptual del caso de estudio se conectan con los elementos del nivel físico.

[1] https://docs.open-metadata.org/v1.12.x/how-to-guides/guide-for-data-users

Nivel Abstracto		Nivel Metadatos	Nivel Físico
Sub-grafo O.Dominio	**Individuos O.Catálogo**		
subgrafo 1 (clases): i) `cuenta balance BCU`, ii) `saldo diario`, iii) `saldo` otras clases: i) `cuenta de base monetaria`, ii) `tipo de instrumento`	Conjuntos: i) `Balance BCU` ii) `Base monetaria`	Tablas: `rub01`, `rub02`, `sdr01` `bcu_map` Publicaciones: `base.xls` y `PAM_diaros.xls` Cubo: `Datos de fuente BCU`	Ubicación: (bd.esquema): siges18.k2bprod dwprod19.pdam Ubicación: wwww.bcu.gub.uy[2] Ubicación: o3web
Clase: `cotización`	Conjunto: `Cotizaciones`	Vista: `v_valores_arbitrados`	Ubicación (bd.esquema): pem.cotizapem

Tabla 4.1: Vinculaciones entre los diferentes niveles

En la primera parte de la fila del cuadro se observa la vinculación del Balance de la contabilidad a través de sus distintos niveles, relacionando las clases del `subgrafo 1` con el conjunto de la ontología `Balance BCU` del Nivel Abstracto y, a su vez, con las tres tablas `rub01`, `rub02` y `sdr01` del Nivel Metadatos, presentadas en la tercera columna. En la cuarta columna, correspondiente al Nivel Físico, se indica cómo localizar dichas tablas mediante la nomenclatura «identificador de la base de datos».«nombre del esquema de base de datos».

La segunda parte de la primera fila hace referencia a la vinculación del concepto de Base Monetaria, que corresponde a un cálculo sobre un subconjunto de las cuentas del balance. La Base Monetaria se apoya en la clase `cuenta de base monetaria` y en el conjunto `Base monetaria`, ambos correspondientes al Nivel Abstracto de la arquitectura (primera y segunda columna del cuadro). En los niveles Metadatos y Físico, esta vinculación se materializa a través de la tabla `bcu_map`. Dicha tabla permite comprender cómo se implementa la relación `integra la base monetaria` definida en la ontología de dominio (véase el cuadro 4.2), ya que asocia cada cuenta del sistema de contabilidad (identificada mediante el atributo `id_item_key`), con distintas clasificaciones requeridas para el cálculo de indicadores en el sistema de *data warehouse*. Entre estas clasificaciones se encuentra la pertenencia a la Base Monetaria (atributo `id_monetaria_base`, con valores sí o no) o el tipo de instrumento (atributo `id_issue_kind`). Como se observa en la primera columna, también se hace referencia a la clase `tipo de instrumento`, la cual, si bien no forma parte del

cálculo de la Base Monetaria, se vincula a la misma tabla de relaciones `bcu_map` para asociar cada cuenta del sistema contable con el tipo de instrumento requerido por el sistema de *data warehouse*.

Finalmente, dentro de esta misma fila, se vinculan las clases y conjuntos asociados a la Base Monetaria con los productos de análisis de información resultantes, entre los que se encuentran las publicaciones diarias `base.xls` y `PAM_diario.xls`, así como uno de los cubos que emplea este indicador al cierre de cada mes, `Datos de fuente`.

En la segunda fila se presenta un ejemplo diferente, una clase y conjunto correspondiente a las cotizaciones del Banco, se trata de un conjunto de datos "reusable" por otros varios conjuntos, el cual se materializa como elemento físico en una vista de base de datos y su correspondiente localización en una base de datos

4.2.2. Ontología de Catálogo

Como parte del proceso de relevamiento e intercambio con diferentes analistas del Banco se entendió que la comunicación visual sobre los activos de datos y sus relaciones es de sumo valor para la organización. Este tipo de representación facilita una comprensión integrada de los artefactos de datos, reduce ambigüedades conceptuales y actúa como un punto común de referencia entre áreas técnicas y de negocio.

En las etapas iniciales del proceso de relevamiento y análisis se asumía que las herramientas de catálogo de datos podrían generar de forma automática un mapa de datos. En ese marco, uno de los criterios de evaluación considerados fue el criterio (7), definido en la sección 2.2.1, relativo a la capacidad de visualización de un mapa de datos. Sin embargo, a partir de los resultados obtenidos en dichas evaluaciones y del intercambio con otras instituciones, tal como se describe en la sección 4.1.4, se concluyó que, al menos para la infraestructura tecnológica del sector analizado, resulta complejo que estas herramientas logren capturar una visión global de los principales activos de datos de la organización y de sus relaciones más relevantes.

En este contexto, y dado que para el Banco, en el marco de su reflexión estratégica (sección 4.1.4), resulta prioritario inventariar y disponer de una representación de los principales activos de información y sus interrelaciones, se identificó la necesidad de hacer visibles dichos activos de forma sintética.

El objetivo es contar con la ayuda de una automatización que permita una comprensión rápida y clara de cuáles son los activos más relevantes y cómo se vinculan entre sí, evitando la necesidad de analizar listados exhaustivos y de difícil interpretación.

Con este propósito, en las siguientes secciones se toma como base el catálogo construido por el Banco. En la sección 4.2.2.3 se priorizan determinados aspectos del catálogo y se introducen ajustes alineados con el vocabulario DCAT. Posteriormente, el catálogo se sistematiza en un modelo relacional (sección 4.2.2.4) y, finalmente, en la sección 4.2.2.5 se muestra una representación gráfica que materializa una propuesta de mapa de datos, con la cual se completa el presente trabajo.

4.2.2.1. Búsqueda de un estándar de catálogo

En la etapa de búsqueda y revisión de literatura sobre catálogos de datos y sus modelos conceptuales, se identificó un estándar de W3C reconocido para el modelado de catálogos publicados en la Web. Este estándar es DCAT[1] , un vocabulario RDF[2] diseñado para facilitar la interoperabilidad entre distintos catálogos de datos.

DCAT permite que los conjuntos de datos sean descubiertos y consumidos con mayor facilidad por aplicaciones y sistemas, al proporcionar un conjunto de clases y propiedades RDF para describir datasets, distribuciones y servicios de datos dentro de un catálogo. El modelo se organiza entorno a siete clases principales, entre ellas: `dcat:Catalog`, `dcat:Dataset`, `dcat:DataService`, `dcat:DatasetSeries` y `dcat:Distribution`[3]

Un dataset en DCAT se define como una "colección de datos, publicada o gestionada por un único agente, y disponible para acceso o descarga en una o más serializaciones o formatos"[4]. Un dataset (`dcat:Dataset`) es una entidad conceptual, y puede estar representado por una o más distribuciones (`dcat:Distribution`) que serializan el dataset para su transferencia. Las distribuciones de un dataset pueden ser provistos por un servicios de datos `dcat:DataService`.

En el contexto del presente trabajo, la noción de *conjunto de datos* se

[1] https://www.w3.org/ns/dcat
[2] https://www.w3.org/RDF/
[3] https://www.w3.org/TR/vocab-dcat-3/images/dcat-all-attributes.svg
[4] https://www.w3.org/TR/vocab-dcat-3/#dcat-scope

adopta en un sentido más amplio que el definido por DCAT. No se restringe a una colección de datos, los cuales en general se tratan y se serializan como un archivo para su publicación en portales de datos, sino que puede abarcar también información obtenida a través de servicios como tableros de inteligencia de negocios o servicios de base de datos, los cuales pueden incluir múltiples colecciones de datos y, en consecuencia, dar lugar a más de una dimensión de análisis, como ocurre, por ejemplo, en el Balance contable del Banco.

4.2.2.2. Oportunidad de mejorar el catálogo desarrollado

El análisis del catálogo de datos elaborado por el Banco muestra que su interpretación resulta difícil para usuarios que no están familiarizados con los datos. A primera vista, varias entradas, cada una correspondiente a un conjunto de datos, presentan nombres, descripciones y atributos de metadatos muy similares, lo que genera duplicaciones aparentes y dificulta distinguir claramente entre los distintos conjuntos catalogados.

La ausencia de definiciones claras sobre qué se considera un conjunto de datos, así como la falta de criterios de granularidad y otras pautas conceptuales, genera ambigüedades en el proceso de incorporación de entradas al catálogo, el cual es realizado por diversos actores. En este contexto, la adopción de un estándar como DCAT resulta pertinente, ya que ofrece un modelo de metadatos internacionalmente reconocido, con estructuras y criterios ampliamente validados.

Asimismo, se identifica una marcada heterogeneidad en los valores asignados a los atributos de metadatos del catálogo actual. Existen solo dos posibilidades, mientras que algunos atributos se completan a partir de listas taxonómicas definidas, otros se registran como texto libre. A su vez en ambos casos existen valores como 'No aplica', que se introducen tanto para textos libres y tipos primitivos, como así también forman parte de listas desplegables, para informar que el dato fue ingresado pero no resulta aplicable al atributo de metadato correspondiente.

En contraste, DCAT propone un catálogo estructurado y, al mismo tiempo, flexible, donde se pueden utilizar relaciones definidas del modelo estándar, enriqueciendo semánticamente a cada elemento relacionado. El modelo permite utilizar predicados predefinidos del vocabulario para caracterizar de forma precisa los elementos asociados a un conjunto de datos, ta-

les como sus distribuciones, documentación o esquemas (`dcat:distribution`, `dcat:landingPage` y `dcterms:conformsTo`, respectivamente).[1] Cuando la relación específica no se encuentra tipificada, es posible recurrir al predicado genérico `dcterms:relation`, al cual pueden asociarse propiedades adicionales que enriquecen la semántica del vínculo, como descripciones, formatos o identificadores (`dcterms:description`, `dcterms:format` y `dcterms:identifier`, respectivamente). En este contexto, RDF puede visualizarse como una especificación que es similar al lenguaje natural en su construcción pero que también es estructurado, capaz de ser leído e interpretado por una máquina.

4.2.2.3. Ontología basada en DCAT

La ontología de catálogo se modeló tomando como base el estándar DCAT, con énfasis en representar aquellos elementos relevantes para los objetivos del presente trabajo. A continuación, se describen las definiciones y decisiones de modelado adoptadas para dichos elementos.

Relación de tipo linaje entre conjuntos. DCAT permite modelar múltiples relaciones entre recursos, tales como conjuntos generados por una actividad, conjuntos que reemplazan o sustituyen otros recursos, entre otras relaciones.[1] En particular, para representar relaciones de tipo linaje y trazabilidad, se analizaron las propiedades `prov:wasDerivedFrom`[2] y `dcterms:source`[3].

Para este trabajo se optó por utilizar `dcterms:source`, dado su carácter más general y su adecuación para expresar relaciones de dependencia entre conjuntos de datos sin imponer supuestos estrictos sobre cómo un conjunto se transforma en otro conjunto, o cómo utiliza otro conjunto de datos. De este modo, se modela que `Balance BCU` es fuente (`dcterms:source`) de `Base monetaria`, así como `Cotizaciones` es fuente de `Balance BCU`, relaciones que pueden observarse en la figura 4.3.

En formato RDF, utilizando sintaxis Turtle[4], la afirmación correspondiente a la primera relación se expresa de la siguiente manera:

```
bt:base-monetaria dcterms:source bt:balance-bcu .
```

[1] https://www.w3.org/TR/vocab-dcat-3/#qualified-forms
[1] https://www.w3.org/TR/vocab-dcat-3/#qualified-forms
[2] https://www.w3.org/TR/prov-o/#wasDerivedFrom
[3] http://purl.org/dc/terms/source
[4] https://www.w3.org/TR/turtle/

Ubicaciones o contenedores de los conjuntos de datos. En el análisis del flujo de datos, los conjuntos suelen asociarse o atravesar puntos intermedios en los que los datos se transforman, se combinan con otros o simplemente son utilizados como insumo para nuevos fines. La identificación de estos puntos resulta de especial relevancia para comprender el contexto en que se generan o transforman los datos, por ejemplo un conjunto de datos como es "Expectativas de inflación" se genera o confecciona en un punto o sector como lo es "Estadísticas". En el marco de este trabajo, dichos puntos se conceptualizan de manera amplia como ubicaciones o, de forma más general, como contenedores.

Estos contenedores pueden corresponder a una institución externa al Banco, representable, por ejemplo, mediante `foaf:Organization`, a un área o unidad interna, o bien a situaciones en las que un conjunto de datos trasciende los límites organizacionales, como cuando es publicado en un sitio web de acceso público. Asimismo, el concepto de contenedor puede referirse también a bases de datos o aplicaciones, que suelen ser la forma habitual en que los usuarios asocian un conjunto de datos en un determinado momento.

Para representar estas relaciones en la ontología de catálogo se adoptó el concepto de `contenedor`, y se utilizó la propiedad referenciada en DCAT: `dcterms:isPartOf`[1], para expresar que un conjunto de datos forma parte de una entidad mayor que actúa como ubicación o contenedor.

Aplicado al ejemplo de la figura 4.3, la siguiente afirmación en sintaxis Turtle representa que el conjunto `Base monetaria` forma parte del contenedor `Publicaciones`:

```
bt:base-monetaria dcterms:isPartOf bcucat:publicaciones .
```

Con la finalidad de organizar y ampliar el concepto de ubicaciones o contenedores, se establece que estos pueden anidarse, es decir, un contenedor puede formar parte de otro contenedor que es de mayor nivel de agregación. Por ejemplo, un tipo de instituciones como las *Instituciones de Intermediación Financiera* puede ser parte de un contenedor más amplio denominado *Instituciones externas* (externas al Banco). Esta relación de inclusión se representa también utilizando la propiedad `dcterms:isPartOf`.

Representación del catálogo del Banco. Como se describió en la sección 4.2.2.2 ciertas entradas del catálogo desarrollado por el Banco dan la

[1] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#isPartOf

apariencia de estar duplicadas, donde algunas de ellas solo se diferencian en parte del nombre de la misma. Para mejorar esto, tomando DCAT como base, varias entradas se representaron como el mismo dataset pero diferentes distribuciones, entendiendo que algunas entradas o filas de la planilla del catálogo del Banco son diferentes representaciones del mismo conjunto de datos. La clase `dcat:Distribution` se corresponde con una representación específica, por ejemplo según un formato y nivel de detalle determinados, y generalmente con una URL de descarga determinada[2]

En el fragmento 1 se muestra un ejemplo en Turtle con un conjunto de datos y sus distribuciones.

[2] https://www.w3.org/TR/vocab-dcat-3/#Class:Distribution

```
1  @prefix dcat: <http://www.w3.org/ns/dcat#> .
2  @prefix dcterms: <http://purl.org/dc/terms/> .
3  @prefix bt: <https://dcat.bcu.gub.uy/data/> .
4
5  bt:base-monetaria
6    a dcat:Dataset ;
7    dcterms:title "Base monetaria"@es ;
8    dcat:landingPage <https://www.bcu.gub.uy/Estadisticas-e-Indicadores/Paginas/
     Moneda-y-credito.aspx> ;
9    dcat:distribution bt:bm-043-base-xls ;
10   dcat:distribution bt:bm-046-informe-diario-xls ;
11 .
12
13 bt:bm-043-base-xls
14   a dcat:Distribution ;
15   dcterms:title "Base monetaria"@es ;
16   dcat:downloadURL <https://www.bcu.gub.uy/Estadisticas-e-Indicadores/MonedayC
     redito/Base-Monetaria/base.xls> ;
17   dcat:mediaType
     <https://www.iana.org/assignments/media-types/application/vnd.ms-excel> ;
18 .
19
20 bt:bm-046-informe-diario-xls
21   a dcat:Distribution ;
22   dcterms:title "Informe diario de la base monetaria del BCU"@es ;
23   dcat:downloadURL
     <https://www.bcu.gub.uy/Estadisticas-e-Indicadores/Informe.../infd...xls> ;
24   dcat:mediaType
     <https://www.iana.org/assignments/media-types/application/vnd.ms-excel> ;
25 .
```

Fragmento 1: Ejemplo de dataset definido en base a DCAT

`bt:base-monetaria` es la identificación del recurso (IRI[1]) para el conjunto de datos base monetaria, donde `bt` es el prefijo para el espacio de nombres (namespace) `https://dcat.bcu.gub.uy/data/`. En la línea 6 se especifica que el conjunto es de tipo `dcat:Dataset`, clase de DCAT, el `";"` al final de la línea, indica que existen más predicados RDF asociados asociados al mismo sujeto `bt:base-monetaria`. En las líneas 7 y 8 se establecen ciertas propiedades como el título o nombre del conjunto, y se establece que el idioma utilizado para el

[1] https://www.rfc-editor.org/rfc/rfc3987

mismo es español (`@es`). DCAT no establece obligatoriedad en las propiedades. En las líneas 9 y 10 se especifica que el conjunto de datos incluye dos distribuciones `bt:bm-043-base-xls` y `bt:bm-046-informe-diario-xls` cuyas propiedades se especifican en las sentencias RDF a partir de las líneas 13 y 20 respectivamente. La línea 11 con el `"."` marca el fin de la sentencia RDF que comenzó en la línea 5 definiendo el conjunto de base monetaria.

En la figura 4.6 se muestran, a modo de resumen, las clases y relaciones de la ontología de catálogo adoptando DCAT.

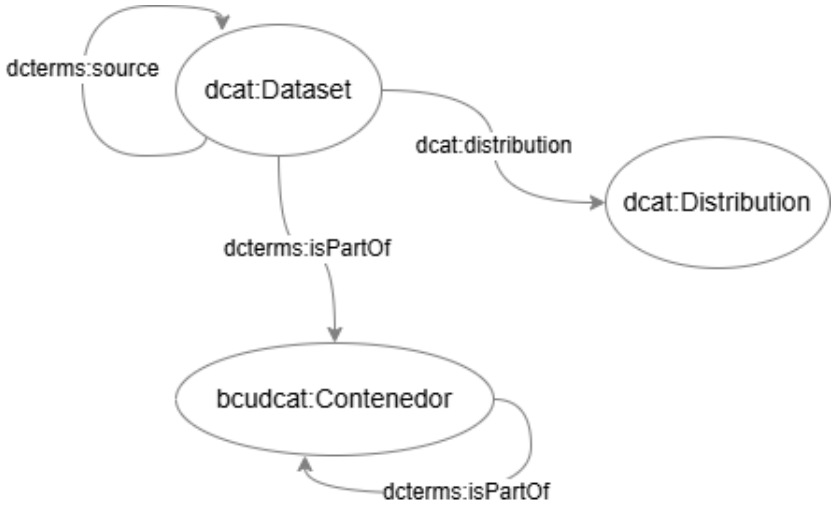


Figura 4.6: Aplicación de DCAT

El diagrama presenta dos clases definidas por DCAT, `dcat:Dataset` y `dcat:Distribution`, mientras que la clase `bcudcat:Contenedor` se define en el contexto de la presente propuesta. Asimismo, se identifican dos propiedades adoptadas y referenciadas en esta sección, `dcterms:source` y `dcterms:isPartOf`, las cuales establecen relaciones entre distintas instancias de una misma clase: `dcterms:source` entre instancias de `dcat:Dataset` y `dcterms:isPartOf` entre instancias de `bcudcat:Contenedor`.

4.2.2.4. Mantenimiento de la ontología en un modelo relacional

En el proceso de modelar las relaciones definidas en la ontología de catálogo, en particular, las relaciones de flujo o linaje entre conjuntos de datos y las relaciones entre un conjunto y su ubicación o contenedor, descritas en la sección 4.2.2.3, se identificó la necesidad de contar con una herramienta que facilitara el mantenimiento (altas, bajas y modificaciones) de dichas relacio-

nes. Un objetivo también buscado fue disponer de una solución que permitiera superar las limitaciones de la herramienta utilizada inicialmente para la construcción del catálogo del Banco (sección 4.1.4), por ejemplo en lo relativo al aspecto manual de la misma.

En este contexto, se optó por utilizar una herramienta ya disponible en la institución: `CMDBuild`[1] . Si bien el propósito principal de CMDBuild es la gestión de activos, entendidos como recursos tecnológicos (hardware y software), la plataforma ofrece un alto grado de flexibilidad para definir entidades, atributos y relaciones personalizadas. Esta capacidad permite modelar conjuntos de datos y sus vínculos asociados, aun cuando dichos conjuntos no constituyen el objeto de gestión original de la herramienta, adecuándose así a los requerimientos del presente trabajo.

CMDBuild permite definir clases con atributos de distintos tipos básicos (texto, numérico, fecha, entre otros). Cada clase se materializa en una tabla de base de datos relacional administrada íntegramente por la herramienta; a su vez, la incorporación de nuevos atributos se refleja automáticamente mediante la adición de los campos correspondientes en dicha tabla. Asimismo, CMDBuild posibilita la definición de `lookup` para representar listas acotadas de valores predefinidos, y de `dominios` para modelar relaciones de tipo 1–1, 1–N o N–N entre las clases definidas[2] .

Si bien CMDBuild no ofrece el nivel de flexibilidad de una aplicación web desarrollada a medida, presenta ventajas claras frente a un enfoque basado en Excel, el cual se limita al mantenimiento de un único tipo de entidad (sección 4.1.4). En contraste, CMDBuild habilita la gestión de múltiples tipos de entidades, por ejemplo, conjuntos de datos, sus distribuciones y contenedores o ubicaciones, junto con sus relaciones explícitas, aportando una estructura más representativa y adecuada a la evolución del catálogo.

Sobre esta base, la sección siguiente describe cómo, a partir de un modelo relacional sistematizado como el implementado en CMDBuild, se aplican una serie de transformaciones que permiten derivar, en una primera etapa, un modelo de ontología y, posteriormente, obtener una representación gráfica.

[1] https://www.cmdbuild.org/en/products/cmdbuild
[2] https://www.cmdbuild.org/en/project/features/database-modeling

4.2.2.5. Representación gráfica

Una vez que las principales relaciones fueron modeladas mediante una ontología estándar de catálogo de datos, y que se dispone de una aplicación que permite registrar, mantener y evolucionar el relevamiento de los conjuntos de datos, el próximo paso en esta prueba de concepto es centrarse en una representación gráfica de dichas relaciones.

En la Figura 4.7 se presentan los pasos seguidos para, a partir de una aplicación que permite mantener el catálogo de datos en una base de datos relacional, generar distintas representaciones gráficas de los elementos del catálogo y de sus relaciones. A continuación, se describen en detalle cada una de estas etapas.

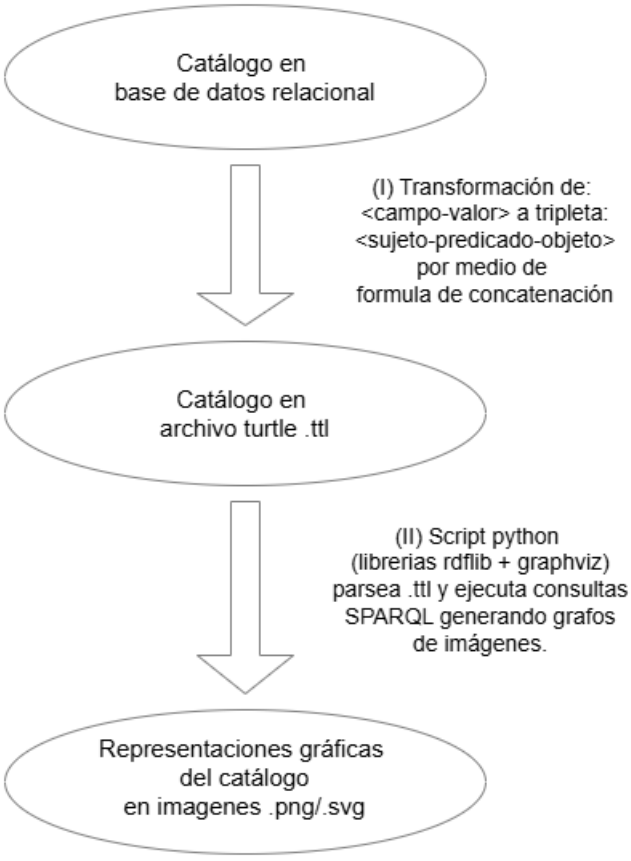


Figura 4.7: Flujo de procesamiento

(I) Transformación de modelo relacional <campo-valor>a sentencia RDF. El primer paso es transformar los datos alojados en una aplicación a la ontología de catálogo. La aplicación, *CMDBuild*, cuenta con la posibilidad de exportar a Excel cada una de las clases y relaciones entre las mismas. Entonces partiendo de esta funcionalidad, es posible agregarle columnas calculadas

a las planillas exportadas. En el cuadro 4.2 se puede observar un ejemplo de columna calculada.

Conjunto origen	**Conjunto destino**	**Fórmula de concatenación**
ds_1047	ds_1532	=”bt:”& B2 & ”dcterms:source ” & ”bt:”& A2 & ”.”

Tabla 4.2: Conversión a sentencia turtle

En el ejemplo del cuadro se exporta la relación de flujo de datos o linaje entre dos conjuntos de datos, en este caso los datos van desde el conjunto con identificador `ds_1047` al conjunto `ds_1532`, con lo cual el conjunto `ds_1532` tiene como fuente (`dcterms:source`) al conjunto `ds_1047`. `ds_1532` se ubica en una celda llamada `B2` mientras que `ds_1047` se ubica en una celda llamada `A2` como se puede observar en la fórmula del cuadro 4.2. En la fórmula se concatenan, el identificador del sujeto (valor de B2), el predicado que es la propiedad `dcterms:source`, y finalmente el identificador del objeto (valor de A2), con lo cual queda completada la sentencia RDF, queda generada de la siguiente forma:

```
bt:ds_1532 dcterms:source bt:ds_1047 .
```

(II) Script de consulta de la ontología y generación de representaciones gráficas. El segundo paso, el procesamiento de las transformaciones tiene como objetivo obtener una representación visual que abstraiga y sintetice las principales relaciones definidas en la ontología del catálogo de datos y, complementariamente, permita profundizar en vistas más detalladas, focalizadas en conjuntos o relaciones específicas según las necesidades del usuario final.

Para alcanzar este objetivo se evaluaron distintas alternativas, incluyendo herramientas y librerías de visualización gráfica, así como una combinación y selección de qué elementos de la ontología representar, dado que la visualización completa de todos los recursos y relaciones en una única imagen resulta poco viable y difícil de interpretar. En este proceso se analizaron soluciones de visualización de ontologías de propósito general, como *WebVOWL*[1] y *GraphDB*[2]

[1] https://github.com/VisualDataWeb/WebVOWL
[2] https://www.ontotext.com/products/graphdb/

, las cuales fueron descartadas por no ajustarse adecuadamente a la forma de representación buscada.

Finalmente, se optó por utilizar las librerías *RDFLib*[1] y *Graphviz*[2] , que, mediante el desarrollo de un *script* en *Python*[3] , proporcionan la flexibilidad necesaria para consultar la ontología, filtrar los elementos relevantes y generar múltiples representaciones gráficas adaptadas a distintos enfoques del usuario final.

RDFLib permite parsear ontologías en formato RDF, en particular el archivo generado en el paso anterior (I). La librería habilita consultar instancias y relaciones presentes en el grafo de la ontología y soporta la ejecución de consultas SPARQL[4] , las cuales se detallan más adelante.

Graphviz permite dibujar automáticamente un grafo en una representación visual legible, organizando automáticamente nodos y aristas de forma adecuada, y ofrece múltiples opciones de personalización gráfica. Asimismo, brinda la capacidad de agrupar nodos en subgrafos (*clusters*[5]), lo que posibilita asociar conjuntos de nodos a propiedades o relaciones compartidas entre varios conjuntos; por ejemplo, datos pertenecientes a una misma área o división de una organización.

Apoyado en estas dos librerías, *RDFLib* y *Graphviz*, el *script* consulta la ontología, recupera determinados conjuntos de datos junto con sus contenedores asociados, y representa los primeros como nodos, mientras que los segundos se visualizan como *clusters*. Posteriormente, se dibujan las aristas correspondientes a las relaciones de flujo o linaje de datos entre los nodos.

El script tiene dos modalidades de ejecución, la primera donde consulta todos los conjuntos de datos, sus relaciones de linaje más sus contenedores asociados, mientras que la segunda modalidad permite focalizarse en uno o más de un conjunto de datos. El centrarse en uno o más conjuntos se consigue filtrando por cada uno de dichos conjuntos y obteniendo los conjuntos ancestros más los descendientes, según las relaciones de linaje de cada uno.

Para el caso de la primera modalidad de ejecución que trae una visión completa de los conjuntos y sus relaciones de flujo, los conjuntos y relaciones disponibles no son en cantidad suficiente como para tener un ejemplo adecua-

[1] https://rdflib.readthedocs.io/en/stable/
[2] https://graphviz.org/
[3] https://www.python.org/
[4] https://www.w3.org/TR/sparql11-query/
[5] https://graphviz.org/Gallery/directed/cluster.html

do, sin embargo para la segunda modalidad, que se dispone de las relaciones obtenidas del relevamiento y ontología correspondiente (sección 4.2.1), se armaron dos ejemplos en función de dos conjuntos de datos que el Banco publica en su sitio web, **Base Monetaria** y **Deuda Pública**. Estos publicaciones, junto con los identificadores `43` y `83` respectivamente, son conjuntos de datos extraídos del catálogo de datos desarrollado por el Banco, descrito en la sección 4.1.4.

La segunda modalidad de ejecución del script, orientada a focalizar el análisis en uno o más conjuntos de datos y a obtener los conjuntos ancestros y descendientes en las relaciones de linaje, se basa en la ejecución de la consulta SPARQL que se presenta en el fragmento 2. Dicha consulta se encuentra embebida en el script en Python y actúa como mecanismo de consulta sobre la ontología, utilizando el soporte provisto por la librería RDFLib[1]. La consulta recibe como parámetro el identificador del conjunto de datos sobre el cual se desea focalizar el análisis (`conjunto_id`) y devuelve todos los conjuntos de datos que se encuentran a lo largo de las relaciones de linaje, ya sea como ancestros o descendientes del conjunto especificado. En particular, en el predicado `dcterms:source`, el operador `+` indica la transitividad de la relación, permitiendo extender la navegación desde relaciones directas de padre o hijo hacia todos los ancestros y descendientes definidos en la ontología.

```
1 PREFIX bt: <https://dcat.bcu.gub.uy/data/>
2 PREFIX dcterms: <http://purl.org/dc/terms/>
3
4 SELECT DISTINCT ?nodo_dataset
5 WHERE {{
6     {{ ?nodo_dataset dcterms:source+  bt:{conjunto_id} . }}
7   UNION
8     {{  bt:{conjunto_id} dcterms:source+ ?nodo_dataset . }}
9 }}
```

Fragmento 2: Consulta SPARQL

El fragmento 3 es un ejemplo de invocación a la segunda modalidad de ejecución del script focalizado en un conjunto de datos, donde como parte del flujo de ejecución se invoca a la consulta SPARQL 2, `ds_43` es el identificador

[1] https://rdflib.readthedocs.io/en/stable/intro_to_sparql/

del conjunto de datos **Base Monetaria** en el ámbito de la ontología. La parte numérica del identificador está dada por el usuario que gestiona el catálogo, mientras que el prefijo `ds` es un prefijo de la clase de la ontología que lo hace único respecto a otros identificadores de otras clases.

```
# Parámetros:
#   1. filename: nombre archivo imagen a generar.
#   2. format: tipo de imagen, por ejemplo 'png' o 'svg'.
#   3. conjuntos_id: identificadores de datasets donde hacer foco.
#   4. mostrarDist: Valor 'True' muestra las distribuciones de los datasets.
generar_grafo_focalizado('grafo_v07', 'png',  conjuntos_id={'ds_43'},
↪  mostrarDist=True)
```

Fragmento 3: Invocación a script

A continuación se ilustran las dos salidas gráficas correspondientes a cada uno de los ejemplos elaborados: Base Monetaria y Deuda Pública.

(1) Representación gráfica de Base Monetaria. Este ejemplo focaliza el análisis en el conjunto de datos `Base Monetaria` (figura 4.8), motivo por el cual el script resalta dicho conjunto en color amarillo.

En el ejemplo se incluyen también las distribuciones (`dcat:Distribution`), consideradas como distintas representaciones del conjunto de datos (`dcat:Dataset`). Los conjuntos de datos se representan mediante figuras elípticas de línea continua, mientras que las distribuciones se muestran con figuras de línea punteada; de forma análoga, las relaciones `dcat:distribution` se representan con líneas punteadas y las relaciones de linaje `dcterms:source` con líneas continuas.

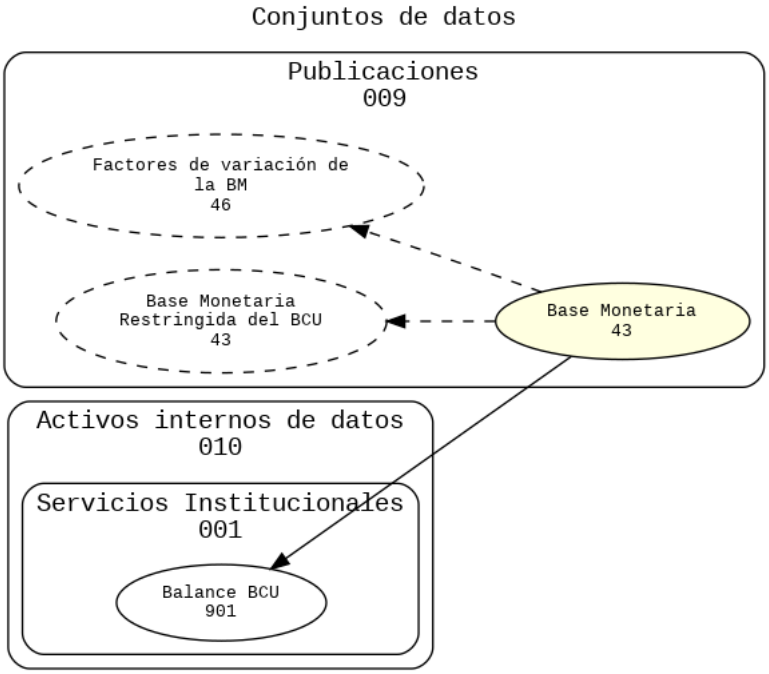


Figura 4.8: Ejemplo dataset y sus distribuciones

Las figuras contenedoras, `Publicaciones`, `Activos internos de datos` y `Servicios Institucionales`, junto con sus respectivos identificadores (009, 010 y 001), que corresponden a su identificación en la ontología, representan los contenedores o ubicaciones definidos por la clase `bcudcat:Contenedor`, tal como se describe en la sección 4.2.2.3.

Cabe señalar que el objetivo inicial era que los principales contenedores, i) *Instituciones externas*, ii) *Activos internos de datos* y iii) *Publicaciones*, se representaran siempre en una misma franja horizontal y dispuestos de izquierda a derecha, con el fin de reflejar el flujo general de los datos dentro del Banco: desde su recepción en la organización, su posterior transformación junto con otros activos de datos y, cuando corresponde, su publicación. No obstante, esta disposición no siempre se logra debido al algoritmo propio de distribución de los elementos gráficos que aplica en cada caso, implementado por la librería utilizada, Graphviz.

(2) Representación gráfica de Deuda Pública. La figura 4.9 muestra un mayor número de conjuntos de datos involucrados en las distintas transformaciones asociadas a este caso. Dado que el conjunto cuenta con múltiples distribuciones, y con el fin de preservar la claridad y legibilidad de la representación, estas no se visualizan; para ello, en la invocación al script se desactiva

el parámetro `mostrarDist`, tal como se ilustra en el fragmento 3.

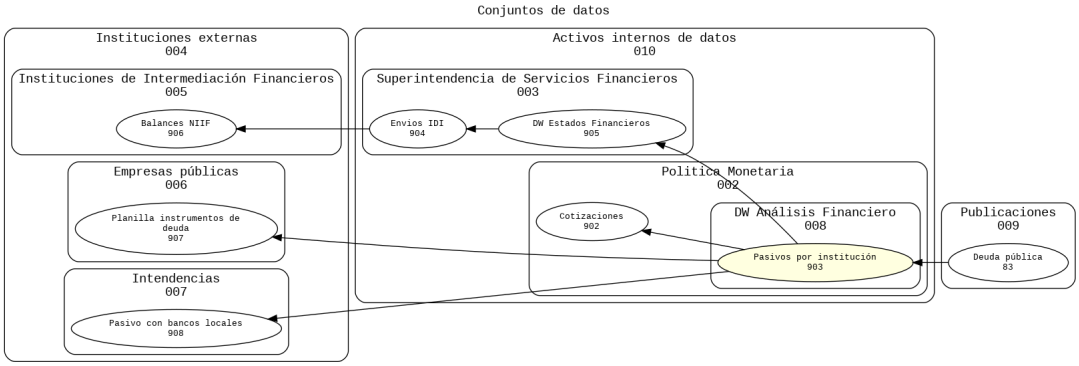


Figura 4.9: Ejemplo linaje de datasets

En este ejemplo, el análisis se focaliza en el conjunto de datos `Pasivos por institución`, resaltado en color amarillo, el cual actúa como fuente (`dcterms:source`) del conjunto `Deuda pública`, correspondiente a una publicación del Banco[1]. Tal como se observa, `Pasivos por institución` se alimenta de múltiples fuentes de datos, tanto internas como externas al Banco, y provenientes de distintas áreas de la institución.

En la medida en que se incorporan nuevos activos de datos y se explicitan sus relaciones, es posible construir y visualizar un mapa cada vez más completo de las interrelaciones entre los datos, aportando una respuesta concreta a una de las preguntas iniciales del trabajo, vinculada a la comprensión del linaje y las dependencias entre los principales activos de información del Banco.

En síntesis, el enfoque propuesto permite orientar el catálogo a las relaciones entre los datos, mediante el uso de un modelo de ontología estándar. Asimismo, posibilita mantener dichas relaciones de forma consistente y accesible para los usuarios, integrándolas en un sistema que facilita su gestión y consulta. Finalmente, este enfoque habilita una visualización en forma de mapa de datos, que contribuye a una comprensión más clara e integrada de los principales activos de información y de sus interrelaciones.

4.2.3. Análisis de resultados

A partir de las representaciones gráficas generadas sobre la prueba de concepto de la ontología de catálogo, se constató que la visualización facilita significativamente la identificación y validación de los principales conjuntos de datos

[1] https://www.bcu.gub.uy/Estadisticas-e-Indicadores/Paginas/Finanzas-Publicas.aspx

y de las relaciones entre ellos. En particular, la comunicación visual contribuye a hacer más comprensibles las interrelaciones y dependencias, favoreciendo una interpretación más intuitiva de la estructura del catálogo.

Si bien la prueba de concepto se limitó a dos conjuntos de datos y a dos relaciones entre ellos (`dcterms:source` y `dcterms:isPartOf`), resultó suficiente para evidenciar el potencial del enfoque propuesto, basado en la combinación de un modelo ontológico con representaciones gráficas. En este sentido, las ontologías pueden representarse mediante estructuras de grafos, lo que otorga flexibilidad para relevar y definir nuevas relaciones entre los datos y otros elementos del dominio. Asimismo, dichas definiciones se formalizan mediante estándares como *RDF*, lo que permite, a partir de este modelo formal, generar diversas visualizaciones y representaciones gráficas de los datos enriquecidos mediante herramientas que entienden el estándar.

Una vez disponible una versión funcional de la prueba de concepto, se realizó una instancia de demostración e intercambio con usuarios del sector al que pertenecen los casos de uso desarrollados. En dicha instancia se presentó el flujo de procesamiento (Figura 4.7), recorriendo cada una de sus etapas: la carga de conjuntos de datos y sus vinculaciones en el sistema, y las visualizaciones del linaje entre conjuntos obtenidas al final del procesamiento. La recepción por parte de los usuarios fue positiva y entusiasta; en particular, manifestaron interés en acceder al sistema de carga para realizar pruebas de ingreso, y se abrió un intercambio sobre la posibilidad de distribuir roles de carga entre distintas áreas de la organización. La instancia estuvo orientada a comprender las relaciones propuestas, el mecanismo de carga de información y las representaciones gráficas generadas, sin pretender constituir una evaluación formal de la aplicabilidad del enfoque a escala organizacional. No obstante, esta última dimensión representa uno de los principales desafíos del enfoque: el esfuerzo requerido para identificar los conjuntos de datos y sus relaciones de linaje es considerable, y su relevamiento exhaustivo requeriría la participación activa de referentes técnicos del área de tecnología con conocimiento profundo de donde y cómo se transforman los datos involucrados.

En cuanto a las herramientas utilizadas, sería deseable contar con una solución integrada que abarque tanto la construcción del modelo como la generación de sus correspondientes visualizaciones. No obstante, las herramientas disponibles resultan, en su mayoría, de propósito general, lo que limita su capacidad para interpretar adecuadamente las distintas relaciones y conceptos,

así como para representarlos mediante objetos visuales diferentes.

Finalmente, y como se ha señalado previamente, la flexibilidad y riqueza del modelo ontológico habilitan un enfoque *top-down*, que permite descomponer los conceptos y sus relaciones en conjuntos de datos progresivamente más concretos y vinculados a elementos tangibles. De este modo, es posible arribar a definiciones más precisas, tales como establecer que un conjunto de datos constituye la fuente de una determinada publicación y que se sustenta en tablas específicas de la base de datos.

Capítulo 5

Conclusiones

Una de las principales conclusiones del trabajo es que, de acuerdo con el estado del arte analizado, las herramientas orientadas a la extracción y gestión de metadatos a partir de elementos físicos, es decir, de *metadatos técnicos*, presentan limitaciones para facilitar el descubrimiento de datos desde la perspectiva de los usuarios con conocimiento del dominio. En particular, dichas herramientas no resultan suficientemente eficaces para identificar y presentar objetos o representaciones que permitan a estos usuarios localizar datos y comprender sus relaciones utilizando el lenguaje y el nivel de abstracción propios de su ámbito de conocimiento, al menos en el contexto organizacional analizado.

Como ejemplo, en el caso de estudio considerado, una búsqueda en un catálogo de metadatos técnicos utilizando términos como "balance" o "base monetaria" difícilmente devolvería elementos físicos tales como "sdr01" o "rub01". Aun en el caso de que dichos elementos fueran identificados, resulta improbable que una herramienta de estas características los presente de manera que el usuario pueda interpretarlos como "saldo diario de cuenta" o "cuentas de balance", respectivamente. Más complejo aún es que el sistema pueda mostrar de forma comprensible las relaciones relevantes entre estos elementos, por ejemplo, qué cuentas de balance componen la base monetaria o cómo se obtiene la cotización a partir del saldo de determinadas cuentas.

Frente a estas limitaciones, el trabajo propone abordar el descubrimiento de los datos a partir de los conceptos principales manejados por los expertos del dominio, utilizando para ello modelos ontológicos. Las ontologías constituyen uno de los mecanismos de representación más expresivos para describir dominios de conocimiento y, a juicio del autor, su incorporación en las he-

rramientas de catalogación de datos aún se encuentra poco desarrollada. Un modelo ontológico no solo permite representar las estructuras de datos de los elementos físicos, sino también dotar de significado a los conceptos del dominio y a las relaciones existentes entre ellos.

En el caso de estudio analizado, por ejemplo, una relación como "se calcula a partir de" permite representar de manera explícita qué cuentas de balance intervienen en el cálculo de un determinado indicador, como es el caso de la base monetaria. Este tipo de relación no corresponde a una relación estructural entre datos, como las que habitualmente se representan en los modelos relacionales, sino a una relación de naturaleza semántica, que describe el significado conceptual del vínculo entre los elementos involucrados.

Otro aspecto relevante al construir estos conceptos de negocio es la importancia de su vinculación con los objetos físicos que los materializan, a través de la metadata técnica que los representa en la herramienta de catálogo de datos. Esta vinculación otorga un sustento concreto a los conceptos abstraídos, permitiendo establecer un puente entre el nivel conceptual que manejan los expertos del dominio y los múltiples detalles técnicos presentes en los elementos físicos donde los datos son implementados y almacenados. De este modo, se favorece un proceso de ida y vuelta entre la representación conceptual del conocimiento y su implementación técnica.

Continuando con la arquitectura propuesta, una relación que se destaca para mejorar la comprensión de los datos en los tres niveles de la arquitectura propuesta es el linaje entre los conjuntos de datos en el nivel de abstracción, así como entre los elementos físicos en los niveles inferiores de la arquitectura. En el caso de estudio analizado, esta relación mostró ser especialmente útil para facilitar la interpretación de los datos y para identificar nuevas fuentes potencialmente relevantes para el análisis de información.

Un aporte concreto del trabajo fue la adopción del vocabulario Data Catalog Vocabulary (DCAT), estándar del W3C, como base para la ontología de catálogo. En particular, la propiedad `dcterms:source` demostró ser adecuada para representar relaciones de linaje entre conjuntos de datos, expresando dependencias de trazabilidad sin imponer supuestos sobre la naturaleza de las transformaciones involucradas. De este modo, el estándar no solo aportó una base semántica sólida para el catálogo, sino que también habilitó la generación de representaciones gráficas del flujo de datos a partir de consultas sobre la ontología, materializando una de las necesidades centrales identificadas en el

trabajo.

Además de los objetivos vinculados a la arquitectura y al modelo conceptual, el trabajo planteó dos objetivos específicos centrados en el usuario analista. Por un lado, incrementar su autonomía en el descubrimiento y exploración de activos de información y, por otro, establecer un marco de trabajo con una herramienta de soporte que sustente la implementación y evolución de un catálogo de datos. Si bien la propuesta contribuye en ambas direcciones, su materialización en la organización quedó acotada al alcance de la prueba de concepto realizada.

En relación al primer objetivo, durante los intercambios mantenidos con usuarios del sector fue posible suministrar, ante consultas concretas sobre conceptos de negocio, los elementos físicos correspondientes y sus relaciones, lo cual favoreció la autonomía del usuario no técnico en la identificación y comprensión de los datos de su área. En relación al segundo, la demostración del sistema de carga de conjuntos de datos en CMDBuild tuvo buena recepción como mecanismo para mantener y evolucionar el catálogo; no obstante, su aplicación quedó circunscripta a los conjuntos de datos y conceptos de negocio analizados en el presente trabajo.

Desafíos

A continuación se describen los principales desafíos identificados al escalar la propuesta más allá del caso de estudio y la prueba de concepto, hacia la construcción de conceptos de negocio y sus relaciones sobre los activos de datos de la organización.

Esfuerzo de modelado

Si bien la arquitectura propuesta y los elementos demostrados tuvieron una recepción favorable por parte de los usuarios, surgió una pregunta que ilustra un desafío central: «¿qué parte debemos cargar nosotros y qué parte realiza la herramienta?»

En el marco del presente trabajo, la identificación de atributos, la vinculación entre conceptos de negocio y elementos físicos, y el descubrimiento de relaciones de linaje en los distintos niveles de la arquitectura demandaron un esfuerzo considerable. Si bien OpenMetadata automatiza parcialmente algunas

de estas tareas, la correspondencia entre el nivel abstracto y el nivel físico, así como gran parte del relevamiento de relaciones de linaje, requiere conocimiento técnico y del dominio que no puede ser inferido por la herramienta. Este esfuerzo de modelado representa un costo no trivial para su adopción a escala organizacional.

Motivación y compromiso organizacional

En el ámbito del gobierno de datos, los resultados de las iniciativas de catalogación y documentación raramente son inmediatos ni fácilmente perceptibles por los actores involucrados. La presente propuesta no es una excepción: el relevamiento de definiciones, la identificación de conjuntos de datos y el mapeo de sus relaciones constituyen actividades analíticas cuyos beneficios se materializan de forma diferida. Esto representa un desafío para involucrar y sostener el compromiso de usuarios con conocimiento experto del dominio, cuya participación resulta indispensable para la construcción y mantenimiento del catálogo.

Trabajo futuro

A continuación se presentan algunas líneas de trabajo futuro identificadas a partir de los resultados y desafíos del presente trabajo.

Herramientas con soporte para ontologías

A partir de las evaluaciones realizadas, se concluye que las ontologías constituyen uno de los mecanismos más expresivos para el modelado del conocimiento de dominio, particularmente en lo que refiere a la representación de entidades de negocio, sus atributos y las relaciones entre ellas. En este sentido, resulta de interés analizar herramientas disponibles en el mercado que incorporen soporte nativo para modelos ontológicos.

En este sentido, se considera relevante explorar herramientas de catálogo de datos que permitan la importación de definiciones de entidades en lenguajes estándar como *RDFS*, y exporten identificadores de recursos (IRI[1]) como mecanismo estándar para referenciar elementos físicos y vincularlos a entidades del dominio, definidas por los usuarios en el nivel conceptual.

[1] https://www.rfc-editor.org/rfc/rfc3987

Descubrimiento automatizado de linaje

Una línea de trabajo de alto valor práctico consiste en la extracción automatizada de relaciones de linaje a partir de la infraestructura de transformación y carga (*ETL*) existente en la organización. Esto implicaría analizar el código fuente de los procesos ETL e integrar los resultados, mediante API o archivo, en una herramienta de catálogo como OpenMetadata, de modo que el linaje entre activos de datos sea inferido y representado de forma automática.

Asimismo, se considera de interés evaluar el uso de técnicas de inteligencia artificial para asistir en la interpretación de los procesos ETL y en la generación de relaciones de linaje. En particular, para el caso de una infraestructura basada en *PL/SQL* de Oracle, resulta pertinente analizar el costo-beneficio de incorporar anotaciones o etiquetas en el código fuente que faciliten su interpretación automatizada.

Aplicación de perfiles de DCAT

El estándar DCAT ha dado lugar a diversos perfiles de aplicación desarrollados en distintas jurisdicciones, entre los que se destacan DCAT-US[1], DCAT-AP[2] y DCAT-AP-ES[3]. Resulta de interés analizar comparativamente estos perfiles y evaluar su aplicabilidad en el contexto uruguayo, especialmente como marco para la publicación de datos abiertos y la interoperabilidad semántica entre instituciones del Estado.

[1] https://doi-do.github.io/dcat-us/
[2] https://op.europa.eu/en/web/eu-vocabularies/dcat-ap
[3] https://datosgobes.github.io/DCAT-AP-ES/

Referencias bibliográficas

[1] Boufassil, A., Bouhafer, F., Cherradi, M., y El Haddadi, A. (2023). Data Catalog: Approaches, Trends, and Future Directions. *2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 369-376. https://doi.org/10.1109/SITIS61268.2023.00067

[2] Ehrlinger, L., Schrott, J., Melichar, M., Kirchmayr, N., y Wöß, W. (2021). Data Catalogs: A Systematic Literature Review and Guidelines to Implementation. En G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky y S. Khan (Eds.), *Database and Expert Systems Applications - DEXA 2021 Workshops* (pp. 148-158). Springer International Publishing.

[3] Feilmayr, C., y Wöß, W. (2016). An analysis of ontologies and their success factors for application to business. *Data & Knowledge Engineering*, *101*, 1-23. https://doi.org/https://doi.org/10.1016/j.datak.2015.11.003

[4] Jahnke, N., y Otto, B. (2023). Data Catalogs in the Enterprise: Applications and Integration. *Datenbank-Spektrum*, *23*(2), 89-96. https://doi.org/10.1007/s13222-023-00445-2

[5] Koutras, C., Siachamis, G., Ionescu, A., Psarakis, K., Brons, J., Fragkoulis, M., Lofi, C., Bonifati, A., y Katsifodimos, A. (2021). Valentine: Evaluating Matching Techniques for Dataset Discovery. *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 468-479. https://doi.org/10.1109/ICDE51399.2021.00047

[6] Kropshofer, J., Schrott, J., Wöß, W., y Ehrlinger, L. (2025). A Survey on the Functionalities of Data Catalog Tools. *IEEE Access*, *13*, 83297-83319. https://doi.org/10.1109/ACCESS.2025.3568542

[7] Oukhouya, Lamya, El haddadi, Anass, Er-raha, Brahim y Asri, Hiba. (2021). A generic metadata management model for heterogeneous sour-

ces in a data warehouse. *E3S Web Conf.*, *297*, 01069. https://doi.org/10.1051/e3sconf/202129701069

[8] Ravat, F., y Zhao, Y. (2019). Metadata Management for Data Lakes. *ADBIS 2019: New Trends in Databases and Information Systems*, *ADBIS 2019 S*, 37-44. https://hal.science/hal-02397467

[9] Schrott, J., Weidinger, S., Tiefengrabner, M., Lettner, C., Wöß, W., y Ehrlinger, L. (2023). GOLDCASE: A Generic Ontology Layer for Data Catalog Semantics. En E. Garoufallou y A. Vlachidis (Eds.), *Metadata and Semantic Research* (pp. 26-38). Springer Nature Switzerland.

[10] Tonnarelli, M., Kumara, I., Driessen, S., Tamburri, D. A., van den Heuvel, W.-J., y Oor, P. (2025). Data catalog tools: A systematic multivocal literature review. *Journal of Systems and Software*, *230*, 112584. https://doi.org/https://doi.org/10.1016/j.jss.2025.112584

[11] Zhao, Y., Ravat, F., Aligon, J., Soule-dupuy, C., Ferrettini, G., y Megdiche, I. (2021). Analysis-oriented Metadata for Data Lakes. *Proceedings of the 25th International Database Engineering & Applications Symposium*, 194-203. https://doi.org/10.1145/3472163.3472273

ANEXOS

Anexo 1

Descubrimiento de linaje

El presente anexo ilustra el proceso de descubrimiento de linaje del cubo "DW - Datos de Fuente BCU" a partir del análisis de los procesos de extracción, transformación y carga (ETL) de datos.

Asimismo, se presentan fragmentos de la metadata del cubo, contenida en el archivo <nombre_de_cubo>_dump.txt y generada mediante una rutina desarrollada para tal fin a partir del modelo del cubo, que describe las dimensiones, medidas y fuentes del cubo, junto con los campos que permiten vincular dichos elementos.

Linaje de la tabla de hechos vdm_balbcu_md

La cadena de linaje identificada involucra las siguientes tablas, correspondientes al caso de estudio presentado en el trabajo:

(1) vdm_balbcu_md ⟶ (2) stg_balbcu_md ⟶ (3) stg_balbcu ⟶ (4) sdr01

En el fragmento de metadata 4, el campo v_value corresponde a la medida Valor del cubo, cuya agregación es de tipo LAST, es decir, toma el valor al cierre del período considerado (fin de mes o del período en cuestión). En contextos donde los paquetes PL/SQL son de considerable extensión, como ocurre en el presente caso, la identificación de un campo de este tipo resulta de utilidad para mantener la trazabilidad entre las tablas origen y destino en el flujo de datos.

```
DW - Datos de Fuente BCU_dump.txt

-- línea 1 - resumen
=============== Cube: 'DW - Datos de Fuente BCU'====================
        21 Dimensiones. Propias: 8  Referenciadas: 13
         2 Medidas
         9 Fuentes

-- línea 121 - definición de la medida Valor
    1 Valor     V_VALUE          LAST       MeasureDefFLSTB

-- línea 104 - consulta sobre la tabla de hechos (1) vdm_balbcu_md, campo
↪  V_VALUE de la medida Valor, y campo ID_ISSUE_KIND como ejemplo para
↪  dimensión Instrumento.
===  8 - Source: 'Datos Situacion' ===
    query: select...a.id_issue_kind,..., a.v_value    from
    ↪  pdam.vdm_balbcu_md a ..
    ...
    11-ID_ISSUE_KIND            12-ID_MATURITY_FRAME
    ...
    21-ID_EXPRESSED             22-V_VALUE
```

Fragmento 4: Ejemplo metadata cubo Datos de Fuente

Packages PL/SQL

Los siguientes fragmentos corresponden a sentencias Oracle PL/SQL pertenecientes a los *stored procedures* del proceso ETL tomado como ejemplo. En cada fragmento se identifica una sentencia de inserción (`INSERT`) y se traza, a partir de ella, la tabla de origen de los datos. El recorrido comienza en la tabla de hechos del cubo `(1) vdm_balbcu_md` y avanza eslabón por eslabón a lo largo de la cadena de linaje. En cada paso se resalta la sentencia de carga y se analiza el código del *stored procedure* hasta identificar la principal tabla fuente, que crea el eslabón siguiente en la cadena.

```
-- búsqueda de (1) vdm_balbcu_md en packages PL/SQL
k_vdw_balbcu.pck
    Line 1117:   insert into  vdm_balbcu_md  (
                    ...
                    select
                        ...
```

```
                        v_value,
                        v_month_result
                    from stg_balbcu_md a

-- búsqueda de (2) stg_balbcu_md en packages PL/SQL
k_vdw_balbcu.pck
    Line 2750:     insert into stg_balbcu_md
                        ...
                        nValueOri, --v_value
-- búsqueda nValueOri
k_vdw_balbcu.pck
   -- Convierte a millones
    Line 2719:       nValueOri := reg.v_value_ori;
    Line 2740:       nValueOri := nValueOri / 1000000;

-- búsqueda de reg
k_vdw_balbcu.pck
    Line 2575:   FOR reg IN
                  (select *
                   from stg_balbcu a
                   ...
-- búsqueda de (3) stg_balbcu
k_vdw_balbcu.pck
    Line 2126:     insert into stg_balbcu
                        ...
                           nValueOri, -- v_value_ori,
-- búsqueda de nValueOri
k_vdw_balbcu.pck
    Line 1978:       nValueOri := nBcuTotalMn;

-- búsqueda de nBcuTotalMn
k_vdw_balbcu.pck
    Line 1974:     nBcuTotalMn := nvl(reg.v_total_mn,0);

-- búsqueda de reg que da como resultado la tabla de origen (4) sdr01 por
↪  medio de dblink dlk2bprod2, también en el resultado se encuentra uno
↪  de los campos, v_total_mn, para el cálculo de la medida Valor.
k_vdw_balbcu.pck
```

```
Line 1452:  FOR reg in
             ...
Line 1465:    (select sdrtpcod, sdrrubcod, s.sdrfecha,
                  nvl(s.sdrsalmn,0)  v_value_mn, nvl(s.sdrsalme,0)
                   ↪  v_value_me,
                  nvl(s.sdrsaltomn,0) v_total_mn, nvl(s.sdrsaleqdo,0)
                   ↪  v_total_usd
                from k2bprod.sdr01@dlk2bprod2.world s
```

Linaje de la dimensión `vdd_issue_kind`

A continuación se ilustra la vinculación entre una dimensión del cubo y su tabla física, identificada a través del campo clave `ID_ISSUE_KIND` que la materializa. Cabe destacar que este campo se encuentra presente también en la tabla de hechos analizada en la sección anterior, fragmento 4, constituyendo así el vínculo entre ambos elementos del cubo.

Por su parte, `"DW - Dimensiones Publicas"` no es un cubo en sentido estricto, sino un objeto cuyo propósito es centralizar la definición de dimensiones compartidas, las cuales son consumidas por otros cubos, como es el caso de `"DW - Datos de Fuente BCU"` presentado en la sección anterior.

```
DW - Dimensiones Publicas_dump.txt

-- línea 1 - resumen
=============== Cube: 'DW - Dimensiones Publicas' ===============
     22 Dimensiones. Propias: 22  Referenciadas: 0
      0 Medidas
     21 Fuentes

-- línea 16  - definición de dimensión Instrumento (issue kind)
-- nivel inferior (hoja / nivel 4) es el que se relaciona con la tabla de
↪  hechos y permite también, armar la jerarquía, campo ID_ISSUE_KIND
===  4 - Dimension: 'Instrumento' ===
    Nivel 1 - ID_ISSUE_GROUP   ISSUE_GROUP_NAME  ISSUE_GROUP_SHORT_NAME
    Nivel 2 - ID_ISSUE_CLASS   ISSUE_CLASS_NAME  ISSUE_CLASS_SHORT_NAME
    Nivel 3 - ID_ISSUE_TYPE    ISSUE_TYPE_NAME   ISSUE_TYPE_SHORT_NAME
    Nivel 4 - ID_ISSUE_KIND    ISSUE_KIND_NAME   ISSUE_KIND_SHORT_NAME

-- línea 73 - búsqueda de campo ID_ISSUE_KIND
=== 13 - Source: 'Instrumento' ===
    query: select  ik.id_issue_kind, ... from  pdam.vdd_issue_kind ik,
    ↪  pdam.vdd_issue_type it, pdam.vdd_issue_class ic,
    ↪  pdam.vdd_issue_group ig where...

    01-ID_ISSUE_KIND  02-ISSUE_KIND_NAME  03-ISSUE_KIND_SHORT_NAME
    ↪  04-ISSUE_KIND_ORDINAL
```

Fragmento 5: Ejemplo de metadata cubo Dimensiones Públicas