BIREDIAL-ISTEC 2023

XII Conferencia Internacional sobre Bibliotecas y Repositorios Digitales

Del 18 al 20 de octubre de 2023

Modelo dimensional para la medición de la producción académica

de Albuquerque, Pablo, CESGI Comisión de Investigaciones Científicas, pablo@sedici.unlp.edu.ar; Villarreal, Gonzalo Luján, PREBI-SEDICI Universidad Nacional de La Plata y CESGI Comisión de Investigaciones Científicas, gonzalo@prebi.unlp.edu.ar ; De Giusti, Marisa Raquel, PREBI-SEDICI Universidad Nacional de La Plata y CESGI Comisión de Investigaciones Científicas, marisa.degiusti@sedici.unlp.edu.ar

Palabras clave

Data Warehouse ; Repositorios Institucionales ; Business Intelligence

Eje temático

Evaluación y métricas alternativas

Resumen

Este artículo se centra en la problemática de la integración y análisis conjunto de la producción académica de las instituciones argentinas, que se encuentra dispersa en diversas fuentes de información, como repositorios digitales, portales de revistas y sistemas CRIS.

Se propone un modelo dimensional basado en Kimball, independientemente de las fuentes de datos utilizadas, con el fin de tener un modelo que permita el análisis de datos provenientes de múltiples fuentes, a lo largo del tiempo.

Texto completo

Introducción

La gestión eficiente de grandes volúmenes de datos académicos es un desafío crucial en el contexto actual de la investigación y la educación superior. La creciente producción académica y la diversidad de fuentes y formatos de datos han llevado a la necesidad de desarrollar enfoques efectivos para organizar,

integrar y analizar esta información de manera sistemática (Xia et al., 2017). En el ámbito académico, la generación de datos involucra una amplia gama de entidades, incluyendo instituciones, colaboradores y publicaciones científicas. La cantidad y complejidad de datos vinculados a estas entidades requiere una gestión eficiente que permita su organización, consulta y análisis de manera efectiva.

Los desafíos asociados con la gestión de datos académicos son diversos. La diversidad de fuentes y formatos de datos dificulta la integración y el análisis holístico de la información. Además, la falta de estándares para la representación de entidades académicas, la necesidad de desambiguación de nombres y la dificultad para relacionar datos de diferentes fuentes agregan complejidad al proceso de gestión de datos académicos (de Albuquerque et al., 2021). Nuestro objetivo principal en este trabajo es diseñar un modelo dimensional que permita una gestión eficiente de datos académicos a lo largo del tiempo. Este enfoque se basa en la estructuración de la información relevante sobre instituciones, colaboradores y publicaciones en tablas estructuradas, aprovechando identificadores persistentes y vocabularios controlados para mejorar la integración y búsqueda de datos. Al utilizar este enfoque, es factible diseñar a futuro un esquema de Data Warehouse que integre las dimensiones y los hechos propuestos en este trabajo junto con procesos de ETL que permitan integrar, organizar y estructurar grandes volúmenes de datos, que luego podrán ser consultados y explotados por medio de herramientas de *Business Intelligence* como por ejemplo Power BI, Tableau o Google Looker Studio (ex Data Studio).

Metodología

En este artículo, nos basaremos en el Proceso de Diseño Dimensional de Cuatro Pasos propuesto por Ralph Kimball (Kimball & Ross, 2013). Este proceso está compuesto por 4 etapas que son:

1. Seleccionar el proceso de negocio
2. Declarar la granularidad
3. Identificar las dimensiones
4. Identificar los hechos

Selección del proceso de negocio

Los procesos de negocio son las actividades operativas realizadas por una organización, como por ejemplo tomar un pedido, procesar un reclamo o registrar estudiantes para una clase. En el enfoque dimensional de Kimball, la elección del proceso de negocio es importante porque define un objetivo de diseño específico y permite declarar la granularidad, las dimensiones y los

hechos correspondientes. La mayoría de las tablas de hechos se centran en los resultados de un solo proceso de negocio.

Además, cada proceso de negocio se correlaciona con una fila en la matriz de buses del almacén de datos empresariales. La matriz de buses es una herramienta utilizada para organizar y relacionar las diferentes dimensiones y procesos de negocio en un diseño dimensional coherente.

Declarar la granularidad

La granularidad se refiere al nivel de detalle con el que los datos son capturados en un modelo dimensional estableciendo qué representa exactamente una sola fila de la tabla de hechos. Es necesario definir la granularidad antes de elegir las dimensiones o los hechos, ya que todas las dimensiones o hechos potenciales deben ser consistentes con la granularidad declarada.

Kimball menciona el concepto de "atomic grain" o granularidad atómica. Esto se refiere al nivel más bajo en el que se capturan los datos en un proceso de negocio específico. Se recomienda enfocarse en los datos capturados a nivel atómico porque son más flexibles para responder a consultas impredecibles de los usuarios.

En un enfoque de granularidad atómica, los datos se representan en su forma más desglosada, capturando cada elemento individual. Por otro lado, la granularidad resumida implica la agregación o consolidación de datos en niveles más altos de abstracción, lo que implica una reducción en la especificidad de los detalles. Al declarar una granularidad resumida, es posible que solo se respondan las preguntas más frecuentes y se pierda la oportunidad de realizar un análisis más profundo. Si bien las granularidades resumidas o agregadas son útiles para optimizar el rendimiento, es importante tener en cuenta que pueden limitar la exploración de datos en detalle y restringir el alcance de las respuestas obtenidas. Esto implica que, al utilizar una granularidad resumida, es fundamental considerar cuidadosamente las preguntas comunes del negocio que se desean responder, ya que los niveles más detallados de los datos pueden no estar disponibles en la visualización o análisis.

Definir dimensiones para describir el contexto

Las dimensiones proporcionan el contexto de "quién, qué, dónde, cuándo, por qué y cómo" que rodea un evento del proceso de negocio. Las tablas de dimensiones contienen atributos descriptivos utilizados por las aplicaciones de inteligencia de negocios (BI) para filtrar y agrupar los hechos. Teniendo en

cuenta claramente la granularidad de la tabla de hechos, es posible identificar todas las dimensiones posibles.

Cuando una dimensión está asociada con una fila de hecho, preferiblemente debería tener un único valor. Las tablas de dimensiones contienen las etiquetas descriptivas que permiten aprovechar el sistema de DW/BI para el análisis de datos.

Definir hechos para medir

Los hechos son las mediciones que resultan de un evento del proceso de negocio y, en su mayoría, son valores numéricos. Una fila de la tabla de hechos tiene una relación uno a uno con un evento de medición tal como se describe en la granularidad de la tabla de hechos. Por lo tanto, una tabla de hechos corresponde a un evento físico observable y no a las necesidades de un informe en particular.

Extensiones y adaptaciones posibles en modelos dimensionales ante cambios en los datos

Los modelos dimensionales son resistentes a los cambios en las relaciones de datos. Estas extensiones se pueden implementar sin alterar las consultas o aplicaciones existentes de inteligencia de negocios (BI) y sin afectar los resultados de las consultas.

A continuación, se presentan los tipos de cambios que se pueden realizar en un modelo dimensional:

- Se pueden agregar nuevos hechos que sean consistentes con la granularidad existente de una tabla de hechos mediante la creación de nuevas columnas.
- Se pueden agregar nuevas dimensiones a una tabla de hechos existente mediante la creación de nuevas columnas de clave externa, siempre y cuando no alteren la granularidad de la tabla de hechos.
- Se pueden agregar nuevos atributos a una tabla de dimensiones existente mediante la creación de nuevas columnas.
- Se puede aumentar la granularidad de una tabla de hechos agregando atributos a una tabla de dimensiones existente y luego redefiniendo la tabla de hechos a una granularidad más baja. Es importante asegurarse de preservar los nombres de columna existentes en las tablas de hechos y dimensiones.

Estas extensiones permiten adaptar y ampliar los modelos dimensionales sin afectar las consultas y aplicaciones existentes.

BIREDIAL-ISTEC 2023

XII Conferencia Internacional sobre Bibliotecas y Repositorios Digitales

Del 18 al 20 de octubre de 2023

Propuesta

Selección del proceso de negocio

En el contexto de este estudio, el proceso de negocio seleccionado se centra en la publicación de recursos académicos generados por una institución científico-académica. El objetivo principal es medir y analizar la producción académica de la institución, estableciendo las bases para futuras evaluaciones del rendimiento y el impacto.

El proceso de publicación abarca una variedad de sistemas y plataformas utilizadas para difundir los recursos académicos. Se consideran fuentes de datos relevantes, como los repositorios institucionales, las revistas científicas, las conferencias, las bases de datos especializadas y los buscadores web, entre otros. Se prioriza el uso de servicios open source y fuentes de datos abiertas para promover la adopción de este modelo en diferentes instituciones científicas y académicas de Argentina.

El período de análisis no presenta limitaciones específicas, y se busca incluir la mayor cantidad de publicaciones posible. Esto permitirá obtener una visión global de la producción académica de la institución a lo largo del tiempo.

Entre los atributos a recopilar para cada publicación académica se encuentran:

- Autor: Identificación del autor o autores responsables del recurso académico.
- Institución: Afiliación institucional del autor o autores.
- Fecha de publicación: Información sobre el momento en que se realizó la publicación.
- Derechos de acceso (*access rights*): Licencia de uso asociada al recurso académico.
- Tipo de recurso: Clasificación del recurso académico, como artículos de investigación, libros, contribuciones a conferencias, entre otros.

Declaración de granularidad

El nivel de detalle deseado para cada evento de medición es a nivel de cada recurso académico individual. Esto implica capturar información específica sobre cada recurso, que puede estar compuesto por archivos de texto, licencias, imágenes, entre otros. Además, se busca representar a cada autor y las relaciones de dependencia entre ellos, para tener en cuenta casos en los que un autor esté afiliado a múltiples instituciones. En este caso, la granularidad está definida por la publicación de recursos identificables en la web a través de identificadores persistentes, como DOI o Handle en el caso de publicaciones,

ORCID para los autores y ROR para las instituciones. Estos identificadores permiten la intersección de datos provenientes de diferentes fuentes, facilitando así la medición y el análisis (Sompel et al., 2014).

La información sobre los eventos de producción académica se captura y registra a través de sistemas y bases de datos específicas. En particular, en una primera instancia, se pueden utilizar los repositorios institucionales propios de cada institución, donde se define qué es relevante para cada una de estas. Estos repositorios cuentan con personal dedicado a revisar, corregir y normalizar los recursos provenientes de otras fuentes de datos. Esto garantiza un nivel mínimo de calidad de los datos para su posterior procesamiento. Además, se aprovecha la interoperabilidad y las directrices definidas a nivel nacional para asegurar la disponibilidad de los metadatos a través de OAI-PMH. Una vez explotados los repositorios propios, pueden considerarse nodos nacionales o sistemas agregadores, en busca de recursos de autores propios de la institución cargados en otros repositorios, ya sea en casos de múltiple dependencia o en coautoría con investigadores de otras instituciones. En una tercera instancia, pueden considerarse fuentes externas como Unpaywall, Crossref o Scopus, prestando especial atención a la disponibilidad de herramientas de consulta y acceso a datos (OAI-PMH, API REST, OpenSearch, etc).

Existen restricciones en cuanto a la disponibilidad de datos a nivel de granularidad deseado. En principio se busca utilizar fuentes de datos abiertas siempre que sea posible, y en el caso particular de Argentina, los repositorios institucionales que forman parte del SNRD, ofrecen la exposición de metadatos de su producción académica (Directrices SNRD, 2015). Existen modelos similares en otros países de la región que también pueden aprovecharse, como por ejemplo el SIC en Chile, KIMUK en Costa Rica, REMERI en México o SILO en Uruguay. E incluso es viable considerar a LA Referencia como fuente de acceso a recursos provenientes de 12 nodos nacionales (LA Referencia - Nodos, s. f.). Sin embargo, se reconoce la utilidad de fuentes de datos cerradas propias de cada institución, que agilicen la ingesta y el procesamiento de los datos académicos. Aunque es importante tener en cuenta que estas fuentes cerradas pueden requerir adaptaciones para su reutilización en otros contextos.

Identificación de dimensiones

Antes de adentrarnos en la identificación de las dimensiones relevantes para medir la producción académica de la institución, es importante destacar que en nuestro modelo hemos adoptado una convención de nombres para las tablas. Todas las dimensiones se representan en la base de datos con el prefijo 'dim_' seguido de un nombre descriptivo. Esta convención nos permite identificar

fácilmente las tablas relacionadas con las dimensiones en el esquema de la base de datos. Además, hemos optado por utilizar nombres en inglés para mantener coherencia con las buenas prácticas de modelado dimensional y la nomenclatura ampliamente utilizada en el campo de la ciencia de datos.

Las dimensiones identificadas son:

- Institución
- Colaborador
- Disponibilidad o nivel acceso
- Tipo de recurso

Asimismo, en todas las dimensiones, se registran datos adicionales como la fecha de creación del registro ('created') en la fuente de datos, la fecha de modificación ('changed') una vez incorporado al modelo, y un valor arbitrario que refleja el grado de confianza ('confidence') basado en el origen y calidad de los datos. Además, en este modelo se han incluido tablas auxiliares con el sufijo '_helper' que desempeñan un papel fundamental en el proceso de ingesta de datos, desambiguación y normalización. Estas tablas auxiliares contienen atributos adicionales que describen a cada registro en mayor detalle, como el nombre y apellido de un colaborador o el nombre y sigla de una institución.

Además de los atributos descriptivos, las tablas auxiliares también incluyen los campos 'source' y 'source_id', que indican el nombre de la fuente de datos de origen (en un formato legible para el usuario) y el identificador correspondiente en dicha fuente de datos. Estas tablas auxiliares desempeñan un papel crítico en garantizar la calidad y coherencia de los datos de las dimensiones en el modelo dimensional. Al incluir información adicional en estas tablas, se simplifica y agiliza el proceso de carga y transformación de los datos, brindando un soporte fundamental para la integración de datos provenientes de diferentes fuentes. Además, al separar esta información auxiliar de la tabla principal, se facilita la visualización y la interpretación directa por parte del usuario final, centrándose en los atributos relevantes para el análisis de la producción académica de la institución.

Dimensión Institución

La dimensión 'Institución' se encuentra representada en la tabla 'dim_institution'. Esta dimensión es fundamental para capturar información descriptiva sobre las instituciones académicas. En dicha tabla, se almacenan atributos como el nombre, sigla y tipo de institución. Además, cabe destacar que la tabla

'dim_institution' cuenta con seis atributos que hacen referencia a identificadores de instituciones que son consideradas como padres para cada registro. Esta relación jerárquica permite establecer la estructura organizativa de las instituciones. Asimismo, se han incluido otros seis atributos que almacenan los nombres normalizados de las instituciones, facilitando su interpretación por parte de los usuarios finales. Un campo adicional, denominado 'doble_dependencia', indica si una institución tiene más de un padre directo, brindando información relevante sobre las relaciones de dependencia entre las instituciones.

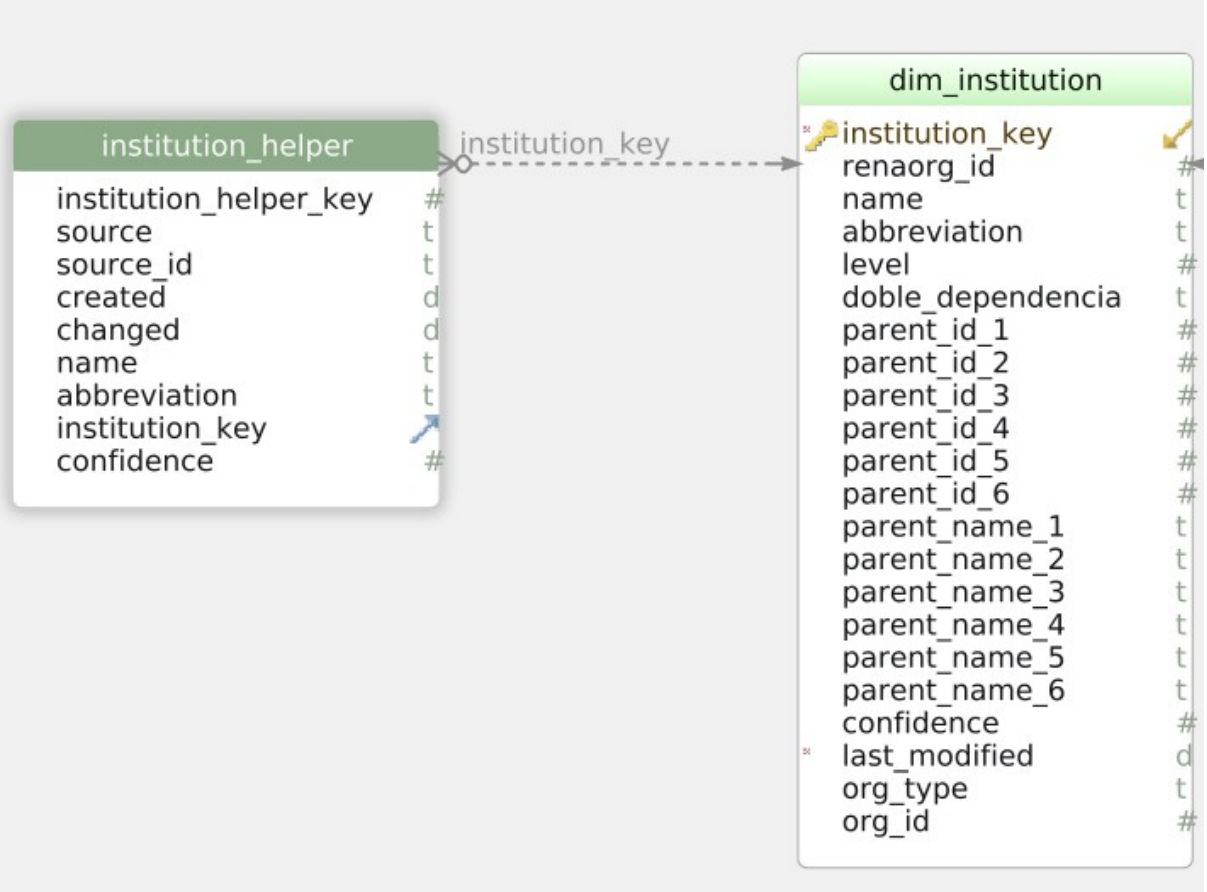

Imagen 1. Esquema de la de la dimensión ‘Institución’.

La dimensión 'Institución' utiliza el Registro de Organizaciones de Investigación (ROR) como la clave principal o business key para la identificación única de las instituciones. ROR es un registro a nivel mundial, impulsado por la comunidad, que proporciona identificadores persistentes y abiertos para las organizaciones de investigación. La elección de ROR como fuente de identificación se basa en su amplia adopción en la comunidad académica y científica, así como en su

disponibilidad como recurso de acceso abierto (French, 2023). Esto facilita la reutilización de los datos y garantiza la integridad de las referencias a las instituciones en el modelo dimensional. Asimismo, para el caso de instituciones Argentinas, es posible tomar el identificador nacional RENAORG como business key complementaria.

Además, se ha creado una tabla auxiliar denominada 'institution_helper' para almacenar datos adicionales que son relevantes en el proceso de ingesta de datos, desambiguación y normalización, pero que no son necesarios para la visualización o interpretación directa por parte del usuario final.

Dimensión Colaborador

La dimensión 'Colaborador' representa a las personas que han contribuido en la producción académica, ya sea como autores principales, coautores, directores de tesis de postgrado, traductores u otros roles relevantes en el proceso. Esta dimensión registra datos en la tabla “dim_contributor” y almacena información sobre los colaboradores involucrados en la producción académica, permitiendo un análisis detallado de su participación. Tanto los autores principales como los colaboradores desempeñan un papel fundamental en la generación del conocimiento y es importante incluirlos en el modelo dimensional para comprender completamente la producción académica de la institución.

La *business key* utilizada para identificar de manera única a cada colaborador en la dimensión 'Colaborador' es el identificador ORCID (Open Researcher and Contributor ID). ORCID es un sistema globalmente reconocido y adoptado por la comunidad académica que proporciona identificadores persistentes y únicos para investigadores y colaboradores. La elección de ORCID como *business key* se debe a su amplia adopción y a su disponibilidad en acceso abierto, lo que facilita la integración de datos y promueve la reutilización de la información académica (Petro, 2020). Mediante el uso de ORCID, se asegura la integridad y la trazabilidad de los colaboradores en el análisis de la producción académica de la institución.

La dimensión 'Colaborador' registra datos como el nombre, apellido y número de teléfono, así como identificadores ampliamente adoptados en otros servicios como Google Scholar, Publons y Scopus. Si bien existe una tabla auxiliar asociada a la dimensión para referenciar nuevos orígenes de datos, los identificadores de Google Scholar, Scopus, Publons, y las páginas web personales de los colaboradores se almacenan directamente en la tabla 'dim_contributor'. Esta elección busca aprovechar que las personas ya están familiarizados con el uso de estos identificadores y por lo tanto simplificarán las tareas de generación de informes y reportes a partir de los datos alojados.

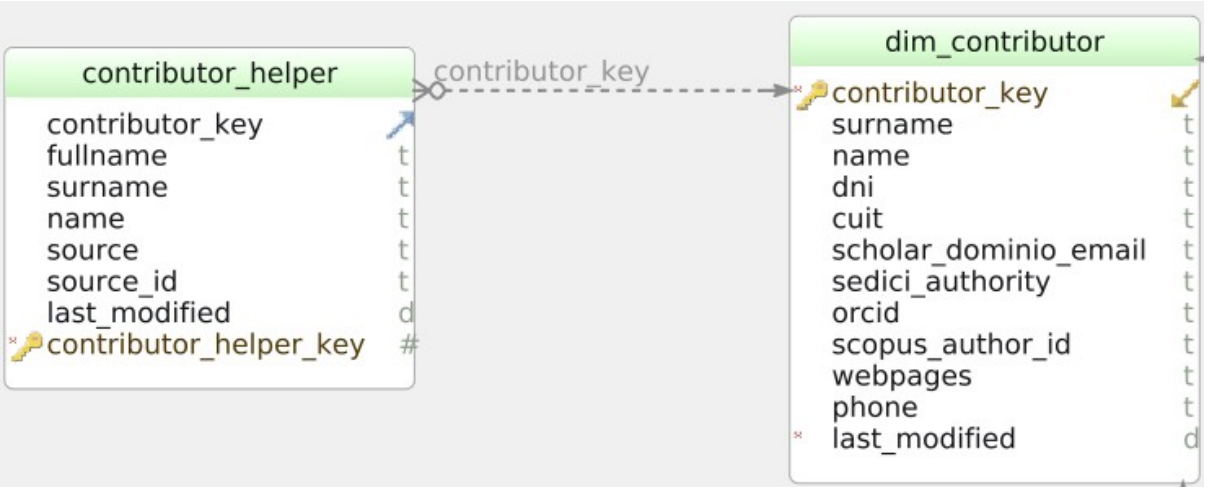

Imagen 2. Esquema de la dimensión ‘Colaborador’

Rol de colaborador en una publicación

Una misma persona puede tener diferentes roles en distintas publicaciones. Por ejemplo, puede ser autor de un artículo, director de una tesis y compilador de un libro. Para representar el rol desempeñado por una persona en una publicación se utiliza la tabla 'bridge_contributor_publication'. Además de contener las claves foráneas que hacen referencia al colaborador y su respectiva publicación, esta tabla incluye un atributo denominado 'role', el cual almacena inicialmente el rol en forma de una cadena de texto.

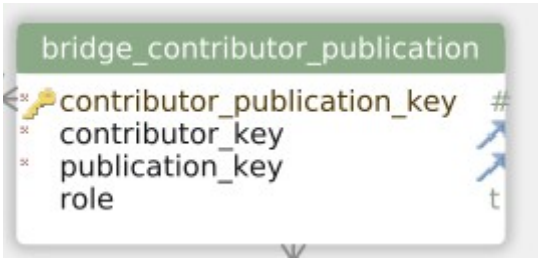

Imagen 3. Esquema del rol de colaborador en una publicación

Dimensión de Disponibilidad o derechos de acceso

La dimensión de ‘Disponibilidad’ se encuentra representada en la tabla 'dim_access_right'. La misma se utiliza para modelar los diferentes niveles de acceso que pueden aplicarse a un recurso académico. Para establecer la business key de esta dimensión, se ha tomado como referencia el vocabulario controlado definido por COAR (COAR Controlled Vocabularies Interest Group,

2022). En dicha tabla se registran valores como embargoed access (acceso con embargo), metadata only access (acceso solo a metadatos), open access (acceso abierto) y restricted access (acceso restringido). Estos atributos permiten representar de manera precisa y estandarizada las políticas de disponibilidad de los recursos académicos en la web.

Esta dimensión está compuesta por varios atributos que proporcionan información adicional sobre los niveles de acceso de un recurso académico. Entre estos atributos se encuentran 'label', que contiene el nombre legible en inglés para el usuario final, y 'label_es', que corresponde a su traducción al español para una mejor comprensión. Además, se incluye el atributo 'uri', que almacena la URI del recurso en la web, permitiendo su fácil acceso.

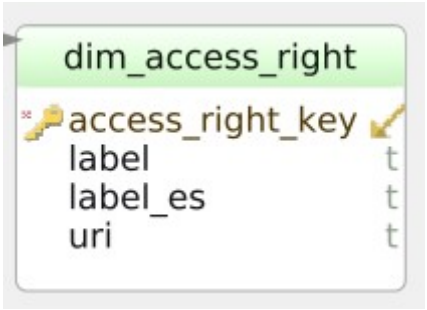

Imagen 4. Esquema de la dimensión de ‘Disponibilidad’

Dimensión de Tipo de Recurso

La tabla 'dim_resource_type' es utilizada para representar los diferentes tipos de recursos que se generan como parte de la producción académica de una institución. En esta dimensión, se almacenan los datos relacionados con los tipos de recursos, siguiendo un modelo basado en el vocabulario controlado de COAR (COAR Controlled Vocabularies Interest Group, 2022).

La dimensión 'dim_resource_type' cuenta con varios atributos que enriquecen su contenido. Entre ellos se encuentran 'label' y 'label_es', que contienen las etiquetas en inglés y español respectivamente, para una mejor comprensión por parte del usuario final. También se incluye 'definition_en', que proporciona una descripción del alcance del tipo de recurso en inglés.

Además, se registra la URI del recurso en la web, permitiendo su acceso directo. Asimismo, se maneja una estructura jerárquica en este vocabulario controlado, por lo que se incluye la URI del padre, para reflejar las relaciones entre los diferentes tipos y subtipos de documentos.

Imagen 5. Esquema de la dimensión de ‘Tipo de Recurso’.

Identificación de hechos

Hecho publicación

El hecho de publicación se registra en la tabla "fact_publication" y contiene información detallada sobre la publicación de un recurso académico específico. Para garantizar la unicidad de cada publicación, se utiliza una business key preferentemente basada en el identificador persistente conocido como DOI (Digital Object Identifier), el cual es mantenido por Crossref (Hendricks, 2021). En caso de que una publicación no disponga de un DOI, se utiliza el atributo "authority" para almacenar otro identificador válido, como por ejemplo "handle", ampliamente utilizado en plataformas como DSpace (Donohue, 2021).

Con el fin de facilitar la integración de nuevas publicaciones y mejorar la interoperabilidad, se ha desarrollado una tabla auxiliar llamada "helper" que contiene otros identificadores relacionados con las publicaciones. Estos identificadores adicionales se almacenan para asistir en el proceso de integración de datos de nuevas publicaciones, siguiendo la misma estrategia aplicada en las dimensiones de institución y colaborador

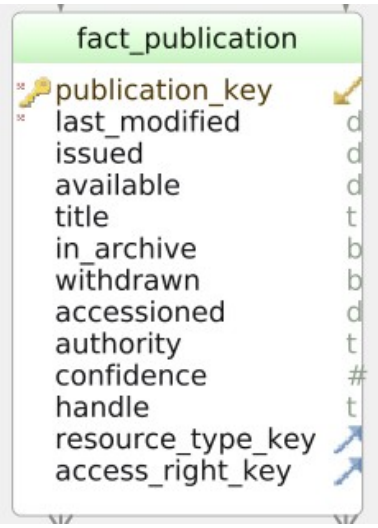

Imagen 6. Esquema del hecho ‘Publicación’.

Hecho filiación

La tabla de hechos de ‘Filiación’, representado por la tabla "fact_affiliation", busca capturar la relación existente entre un colaborador (contributor) y una institución en un período de tiempo específico. Para lograr esto, se utilizan claves foráneas que hacen referencia a las dimensiones de colaborador (contributor) e institución (institution), junto con los atributos de fecha de inicio y fecha de finalización para indicar la validez de esta relación en el tiempo.

En este caso, no se han identificado *business keys* específicas que permitan identificar la filiación de una persona de manera única. Sin embargo, se puede utilizar la información de las publicaciones asociadas a un colaborador para calcular los valores de inicio y fin de la filiación. Por ejemplo, si una persona publica un artículo donde establece su afiliación con una institución X en el año 2018, podemos inferir que esa filiación existió al menos a partir de dicho año; si luego se procesa una publicación del mismo autor vinculada a la misma institución con año 2016, podrá registrarse que la persona ya pertenecía a la institución desde al menos 2016.

De esta manera, la tabla de hechos "fact_affiliation" proporciona una representación precisa de la relación entre colaboradores e instituciones en un contexto temporal, permitiendo un análisis más completo de la producción académica y las colaboraciones entre diferentes actores. Este modelo también resulta útil para representar la pertenencia a múltiples instituciones de una misma persona en un momento dato.

Imagen 7. Esquema de ‘Filiación’

Conclusión

En este trabajo, hemos abordado el desafío de diseñar un esquema dimensional para gestionar de manera eficiente un gran volumen de datos académicos a lo largo del tiempo. Nuestro enfoque se ha centrado en la representación estructurada de información relevante sobre instituciones, colaboradores y publicaciones en un modelo dimensional. Mediante la implementación de tablas como 'dim_institution', 'dim_contributor' y 'fact_publication' en un sistema de bases de datos relacional, hemos logrado representar de manera precisa la información clave relacionada con estas entidades.

Nuestra contribución a la gestión de datos académicos radica en la propuesta de un modelo dimensional que mejora la eficiencia en la consulta y organización de la información. Utilizando identificadores persistentes ampliamente adoptados como DOI y ORCID, y aprovechando vocabularios controlados como los proporcionados por COAR, hemos facilitado la integración y búsqueda de datos académicos.

Al basarnos en la metodología propuesta por Kimball, nuestro modelo dimensional ofrece beneficios significativos en términos de estructura de datos, precisión en la identificación de instituciones y colaboradores, y eficiencia en el análisis de publicaciones académicas. Además, hemos destacado la importancia de considerar tablas auxiliares, como 'institution_helper' y 'contributor_helper', para apoyar tareas de ingesta de datos y desambiguación.

Como trabajo futuro, planeamos expandir este modelo para incorporar los hechos y dimensiones adecuados que permitan medir no solo la producción científica, sino también el impacto que esta genera, de acuerdo a los intereses de cada institución. Sumado a esto, como se mencionó previamente, este proyecto requiere una adecuada selección de fuentes de datos, considerando especialmente el nivel de confianza o la calidad de los datos que ofrece, la definición de patrones de ingesta y procesamiento de datos (ETL, ELT o ETLT) (Gupta et al., 2020) y la generación de una infraestructura capaz de ejecutar

estos procesos, almacenar los datos de manera eficiente y generar reportes para usuarios finales.

Bibliografía

1. Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big Scholarly Data: A Survey. IEEE Transactions on Big Data, 3(1), 18-35. https://doi.org/10.1109/TBDATA.2016.2641460

2. de Albuquerque, P. C., Villarreal, G. L., & De Giusti, M. R. (2021). Proposal of a Data Warehouse for Scholarly Institutions built on Institutional Repositories. IX Jornadas de Cloud Computing, Big Data & Emerging Topics (Modalidad virtual, 22 al 25 de junio de 2021). http://sedici.unlp.edu.ar/handle/10915/125161

3. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit, 3rd Edition. Kimball Group. https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-toolkit/

4. Sompel, H. V. de, Sanderson, R., Shankar, H., & Klein, M. (2014). Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. International Journal of Digital Curation, 9(1), Art. 1. https://doi.org/10.2218/ijdc.v9i1.320

5. Directrices para proveedores de contenido del Sistema Nacional de Repositorios Digitales. (2015). [ebook] Buenos Aires: Sistema Nacional de Repositorios Digitales. Ministerio de Ciencia, Tecnología e Innovación Productiva. Disponible en: https://repositoriosdigitales.mincyt.gob.ar/files/Directrices_SNRD_2015.pdf . Recuperado 29 de mayo de 2023

6. LA Referencia—Nodos. (s. f.). Recuperado 29 de mayo de 2023, de https://www.lareferencia.info/es/nodos

7. French, A. (2023). Case-studies. Research Organization Registry (ROR). Recuperado 29 de mayo de 2023, de https://ror.org/categories/case-studies/

8. Petro, J. (2020, octubre 20). 10M ORCID iDs! 10M ORCID IDs! https://info.orcid.org/10m-orcid-ids/

9. COAR Controlled Vocabularies Interest Group. (2022, septiembre 29). Controlled Vocabularies for Repositories: Access Rights. https://vocabularies.coar-repositories.org/access_rights/

10. COAR Controlled Vocabularies Interest Group. (2022, septiembre 29). Controlled Vocabularies for Repositories: Resource Types. https://vocabularies.coar-repositories.org/resource_types/

11. Donohue, T. (2021). Handle.Net Registry Support—DSpace 7.x Documentation—LYRASIS Wiki. https://wiki.lyrasis.org/display/DSDOC7x/Handle.Net+Registry+Support

12. Hendricks, G. (2021, diciembre 11). The research nexus [Website]. Crossref. https://www.crossref.org/documentation/research-nexus/

13. Gupta, A., Sahayadhas, A., & Gupta, V. (2020). Proposed Techniques to Design Speed Efficient Data Warehouse Architecture for Fastening Knowledge Discovery Process. 2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 200-201. https://doi.org/10.1109/AIKE48582.2020.00039