ANÁLISIS PRELIMINAR SOBRE CATALOGACIÓN DE METADATOS EN BASE A UN MUESTREO DEL CATÁLOGO DE IDERA

Luis Reynoso

Facultad de Informática, Universidad Nacional del Comahue, Neuquén,
luis.reynoso@fi.uncoma.edu.ar

Dirección Provincial de Catastro e Información Territorial, Neuquén,
lreynoso@neuquen.gov.ar

1. INTRODUCCIÓN

El web scraping, aplicado al catálogo nacional de metadatos de IDERA, ofrece una herramienta invaluable para obtener conclusiones significativas sobre el proceso de catalogación. Los metadatos son esenciales para la organización y recuperación eficiente de la información, pero su calidad y coherencia pueden variar considerablemente, afectando la utilidad y precisión del catálogo. Aquí es donde el web scraping juega un papel crucial para revelar su estado y oportunidades de mejora. El propósito de este resumen extendido es mostrar los resultados preliminares luego de seis meses de trabajo en la aplicación de técnicas de web scraping en el catálogo de metadatos de IDERA utilizando el lenguaje R con los paquetes ows4R, xml2 y xml. Al automatizar la extracción de datos del catálogo nacional, el web scraping permite recopilar grandes volúmenes de metadatos de manera sistemática. La extracción es costosa, los catálogos no permiten la extracción de grandes volúmenes en masa. Otro aspecto crucial del estudio es la capacidad de comparar los metadatos extraídos con estándares y normativas internacionales, como Dublin Core o ISO 19115 para datos geoespaciales. Esto facilita la evaluación de la conformidad del catálogo de IDERA con las mejores prácticas internacionales y proporciona recomendaciones concretas para mejorar los procesos de catalogación y mantenimiento de metadatos. En resumen, el uso de técnicas de web scraping en el catálogo de metadatos de IDERA no solo permite una recopilación eficiente y sistemática de información, sino que también facilita análisis detallados que pueden mejorar significativamente la calidad y utilidad del catálogo, asegurando que cumpla con los estándares requeridos.

2. MATERIALES Y MÉTODOS

2.1 MUESTRA DEL CATÁLOGO DE METADATOS DE IDERA Y ELEMENTOS ANALIZADOS

Como insumos para el objetivo planteado se utilizaron los siguientes dos documentos publicados en la página de IDERA: (1) Perfil de Metadatos para Datos Vectoriales IDERA v2.0: Contiene la descripción del Perfil de Metadatos IDERA (PMIDERA) para datos vectoriales utilizando como base el estándar ISO

19115 y la aplicación técnica ISO 19139 (IDERAa, 2024); (2) Perfil de Metadatos para Imágenes Satelitales v1.0: Contiene la descripción de referencia y discusión del Perfil de Metadatos IDERA (PMIDERA) para imágenes satelitales utilizando como base el estándar ISO 19115-2 y la aplicación técnica ISO 19139-2. (IDERAb, 2024). El paquete básico de elementos propuesto como el Perfil de IDERA (PMIDERA) consiste de un subconjunto mínimo de elementos considerados necesarios e indispensables del Núcleo (CORE) de la norma ISO 19115. El Anexo I de ambos documentos citados incluye la descripción de cada elemento con "su descripción básica, su implementación en XML, un ejemplo que referencie al dato utilizado, la cantidad máxima que puede aparecer el elemento y su tipo de dato, indicando el dominio permitido". El estudio realizado implicó el análisis de una muestra significativa del catálogo de metadatos de IDERA. La muestra consiste en 2852 registros de un total de 4644 registros que contiene el catálogo de metadatos. Es decir la muestra representa el 61,41% del catálogo. El estudio se focalizó en analizar aquellos tipos de elementos del Perfil de metadatos de IDERA correspondientes a las clases A y B: "A. Información de Identificación"; "B. Información del Sistema de Referencia". Prestando especial atención en los elemento que se consideran obligatorios, esto esta categorizado en ambos perfiles como tipo de elemento "A. Obligatorio" y algunos elementos categorizados como opcionales "B. Opcional" (IDERAa, 2024; IDERAb, 2024).

2.2. TÉCNICA DE WEB SCRAPING, LENGUAJE DE PROGRAMACIÓN Y PAQUETES EMPLEADOS

El web scraping (Vaughan, 2028) es una técnica automatizada ampliamente utilizada para extraer información precisa de páginas y catálogos web mediante el uso de programas o scripts. Se programó en lenguaje R un proceso que, aplicando esta técnica, extrae datos estructurados en XML del catálogo de metadatos, utilizando los paquetes SF, ows4R, XML y xml2 de R. SF (simple features for R) permite consultar el modelo de geometrías de características simples (SFA), un estándar (ISO 19125) desarrollado por OGC (Pebesma, 2018). El paquete ows4R (Blondel, 2024) proporciona una interfaz para servicios web de OGC, incluyendo WFS, WCS, CSW, y WPS. Los paquetes XML y xml2 se usaron para extraer valores de nodos XML de acuerdo con las normas ISO 19115 y 19139. El proceso implementado en R genera un tablero de control sobre la muestra obtenida, con una tabla de datos interactiva y una serie de gráficos estadísticos (utilizando los paquetes reactable y plotly).

3. RESULTADOS

El tablero concentra los resultados más significativos (por el volumen tarda 15 segundos en abrir):

https://opendata.fi.uncoma.edu.ar/IDERA/TableroMetadatos_2024.html

TEMA DE LOS METADATOS: Los 2852 registros de nuestra muestra, conformes a las normas ISO 19115 y 19139, proporcionan una visión detallada sobre la distribución de categorías temáticas en el catálogo. Los resultados obtenidos muestran una amplia diversidad de temas en los registros del catálogo. Cabe destacar que cada metadato puede incluir más de un tema. Los datos recuperados del catálogo de metadatos revelan una preponderancia de registros relacionados con la categoría de **ubicación (22.12%)**, seguida por **límites (15.43%)** y **sociedad (9.96%)**. Estas tres categorías combinadas representan más del 47% de los registros totales, indicando una fuerte orientación del catálogo hacia temas de georreferenciación y delimitación espacial. Es notable la presencia de registros sin valor en la categoría temática (7.22%), lo cual sugiere posibles áreas de mejora en la calidad y completitud de los metadatos registrados. La categoría de **economía (5.08%)** también tiene una representación significativa, al igual que **información geo científica (4.03%)**, mostrando la relevancia de estos temas en el contexto del catálogo. Además, algunas categorías reflejan la intersección de múltiples temas, como **límites y economía (2.10%)**, **estructura y sociedad (1.23%)**, y **límites y sociedad (0.46%)**, lo cual destaca la complejidad y la multifuncionalidad de ciertos registros. En resumen, la distribución temática del catálogo de metadatos proporciona una visión clara de las prioridades y enfoques en la recopilación y gestión de información geoespacial. La alta representación de temas relacionados con ubicación y límites, junto con la diversidad de otros temas, subraya la importancia de estos aspectos en el análisis y utilización de datos geoespaciales en Argentina.

ESTADO DE LOS METADATOS: El análisis del estado de los metadatos en el catálogo revela importantes observaciones sobre la calidad y completitud de la información registrada. Aunque opcional, este elemento proporciona un indicador clave sobre la vigencia y utilidad de los datos geoespaciales disponibles. De los 2852 registros analizados, el 81.22% no tiene un valor asignado para el estado del metadato, lo que sugiere una falta de información crítica que dificulta la evaluación de la vigencia y relevancia de los datos. Solo el 18.78% de los registros se clasifican como completos, indicando que una parte significativa de los metadatos carece de la información necesaria para ser considerados confiables. Además, no se encontraron registros en las categorías de Archivo histórico, Obsoleto, En curso, Planeado, Requerido o En desarrollo. Estos resultados destacan la necesidad de mejorar la documentación y actualización de los metadatos en el catálogo. La falta de información sobre el estado de los metadatos puede impactar negativamente en la gestión y uso eficiente de los datos geoespaciales. Además, la ausencia de valores en las categorías de estado sugiere una oportunidad

para establecer procesos más robustos de actualización y mantenimiento de los metadatos. Implementar prácticas para asegurar que los estados de los metadatos se registren y actualicen regularmente podría mejorar significativamente la calidad del catálogo. En resumen, el análisis del estado de los metadatos revela áreas críticas de mejora en la gestión de la información geoespacial. Asegurar que los metadatos tengan valores completos y actualizados es esencial para garantizar su utilidad y relevancia, facilitando una gestión más eficiente.

FRECUENCIA DE MANTENIMIENTO DE LOS METADATOS: El análisis revela una significativa disparidad en los valores asignados, lo que sugiere variaciones en las prácticas de actualización y gestión de los datos geoespaciales. La mayoría de los metadatos (2046 registros) indican una frecuencia de mantenimiento "según necesidad", lo que podría reflejar una falta de programación regular en la actualización de los datos debido a recursos limitados, prioridades cambiantes y la falta de políticas claras. Un número significativo de metadatos (512 registros) no especifica la frecuencia de mantenimiento, lo que sugiere una posible falta de atención o desconocimiento sobre la importancia de documentar esta información, afectando la confiabilidad y utilidad de los datos. Además, la diversidad de valores como "continuo", "no planificado", "anualmente", y "diariamente" indica una disparidad en las prácticas de mantenimiento de los metadatos, atribuida a diferentes capacidades organizativas y la variación en la relevancia de los datos. Las categorías "desconocido" (32 metadatos) y "no planificado" (129 metadatos) resaltan la necesidad de mejorar la planificación y la claridad en las prácticas de mantenimiento de los metadatos. La disparidad en la frecuencia de mantenimiento y la preponderancia de la categoría "según necesidad" sugieren la necesidad de establecer procedimientos más uniformes para la actualización de los metadatos. Establecer prácticas de mantenimiento más consistentes y documentadas podría mejorar la integridad y la utilidad del catálogo de metadatos.

TIPO DE REPRESENTACION ESPACIAL DE LOS METADATOS: El análisis de la muestra de 2852 registros revela tendencias significativas, una abrumadora mayoría, 2189 registros (76.74%), se clasifican como tipo *vector*. Esta predominancia indica una fuerte preferencia por datos vectoriales, que son esenciales para representar con precisión la geometría de objetos geográficos como puntos, líneas y polígonos. Por otro lado, 198 registros (6.94%) se identificaron como tipo *grid*, lo que señala un uso moderado de datos rasterizados. Estos datos son cruciales para aplicaciones que requieren análisis espacial continuo, como estudios climáticos y topográficos. Es notable que solo 27 registros (0.95%) se clasificaron como tabla de texto (texttable), lo que sugiere que la representación textual de datos

espaciales no es una práctica común en los geoservicios analizados. Un aspecto preocupante es que 436 registros (15.28%) no especifican un tipo de representación. En resumen, estos hallazgos destacan la predominancia del formato vectorial, la presencia moderada de datos tipo grid, la escasa representación de tablas de texto y nula representación de otros tipos (ej. tin, video, etc). Esta conclusión está en línea con las respuesta del informe anual de IDERA 2024 sobre metadatos que revela que el 67.6% de las IDE cargan metadatos, principalmente para datos vectoriales (ISO 19115) y en menor medida para datos raster (ISO 19139). Las interfaces más usadas son Geonode y Geonetwork, con 16 respuestas cada una, seguidas por QGIS y PDF. El 54.5% de los encuestados utiliza la plantilla de metadatos vectoriales de IDERA, mientras que el 45.5% no la usa.

ESCALA DE LOS METADATOS: El análisis de las escalas asociadas con los metadatos recuperados a través de web scraping revela una variabilidad significativa en los niveles de detalle geográfico de los datos geoespaciales disponibles en el catálogo. La mayoría de los metadatos indican escalas grandes, lo que permite una alta precisión en la representación espacial. Por ejemplo, las escalas más comunes son 1/1.000 con 784 metadatos, 1/10.000 con 203 metadatos y 1/5.000 con 198 metadatos. Estos datos son detallados y útiles para análisis de alta precisión, necesarios en estudios locales y específicos como planificación urbana, infraestructura y mapeo detallado de pequeñas áreas. También se observan metadatos con escalas más generales, útiles para estudios regionales o nacionales, como 1/1.200.000 con 165 metadatos, 1/50.000 con 100 metadatos y 1/250.000 con 70 metadatos. Estas escalas sugieren que los datos geoespaciales del catálogo también son adecuados para análisis a mayor escala, proporcionando una visión más amplia de las áreas geográficas.

EXTENSIÓN GEOGRÁFICA: Se obtuvieron los valores de latitud y longitud de los metadatos para generar polígonos que representan su extensión geográfica. Se separaron 588 metadatos cuya extensión geográfica no pertenece al polígono envolvente de la República Argentina para su revisión. Luego, de 2068 registros incluidos en la envolvente y se mapeó la extensión geográfica utilizando ArgenMap Oscuro, revelando patrones, esto es, donde se concentran la mayor cantidad de metadatos: provincias como NOA (Tucumán, Salta), San Luis, Chaco, Santa Fé, Formosa y Neuquén.

4. CONCLUSIONES

El análisis del estado de los metadatos del catálogo de IDERA muestra logros y áreas de mejora en la gestión de la información geoespacial. La mayoría de los registros están en formato vectorial, lo que resalta su importancia en la georreferenciación y el análisis espacial. Sin embargo, la frecuencia de

mantenimiento de los metadatos es dispar, predominando la categoría "según necesidad", lo que sugiere la necesidad de directrices uniformes y procedimientos de actualización regulares. Muchos metadatos carecen de valores asignados para su estado, afectando la confiabilidad y utilidad de los datos. Mejorar la completitud y actualización de los metadatos es esencial. La variabilidad en las escalas geográficas y la falta de especificación en la representación espacial indican la necesidad de un enfoque equilibrado y una mejor estandarización para maximizar la utilidad de los metadatos. Los hallazgos subrayan la necesidad de políticas claras, recursos adecuados y procedimientos robustos de actualización. Implementar estas mejoras puede aumentar la calidad y efectividad del catálogo de IDERA. Además, es fundamental invertir en la capacitación del personal y desarrollar directrices claras para asegurar la uniformidad y calidad de la información. Procesos de revisión y auditoría periódica pueden ayudar a identificar y corregir valores ausentes y mejorar la precisión de los metadatos.

Futuras investigaciones pueden aplicar técnicas de minería de datos y análisis estadístico para detectar inconsistencias en la estructura de los metadatos. A pesar de los avances, existen oportunidades significativas para mejorar la especificación, mantenimiento y documentación de los metadatos, lo que permitirá una gestión más eficiente y efectiva de la información geoespacial en Argentina.

5. REFERENCIAS BIBLIOGRÁFICAS

Blondel, E. (2024). ows4R: R Interface to OGC Web-Services (0.4). Zenodo. [Accedido 12-07-2024]: https://doi.org/10.5281/zenodo.12544282

Edzer Pebesma (2018). Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal . [Accedido 12-07-2024]: https://journal.r-project.org/archive/2018/RJ-2018-009/index.html

IDERAa (2024) Perfil de Metadatos para Datos Vectoriales IDERA . [Accedido 12-07-2024]: v2.0 https://www.idera.gob.ar/images/stories/downloads/estandares/PMIDERA_Perfil_Metadatos_p_Datos_Vectoriales_IDERA_V2_0.pdf

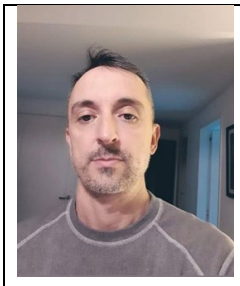
IDERAb (2024) Perfil de Metadatos para Imágenes Satelitales v1.0 . [Accedido 12-07-2024]: https://www.idera.gob.ar/images/stories/downloads/estandares/Perfil_Metadatos_Imagenes_Satelitales_V_1.0.pdf

ISO (2003) ISO 19115 Geographic information – Metadata. 2003.

ISO (2003) ISO 19115 CORE Geographic information – Metadata. 2003;

Vaughan, D. (2018). Web Scraping with R: A Tidy Approach. CRC Press.

Sobre el Autor:



Luis Reynoso es Licenciado en Tecnología Educativa por la Universidad Tecnológica Nacional (UTN), Licenciado en Ciencias de la Computación y Magister en Ciencias de la Computación por la Universidad Nacional del Sur, Doctor por la Universidad de Castilla-La Mancha (UCLM), España. investigador y director del proyecto 04/F023: Tecnologías de Datos Espaciales, Visualización y Realidad Virtual. Coordinador del proyecto: “Mejora de la Infraestructura Territorial Catastral (ITC) de la Dirección Provincial de Catastro e Información Territorial de la Provincia del Neuquén” en el marco del Programa de Fortalecimiento de la Gestión Provincial, BID 3835/OC-AR. Es coordinador del grupo de trabajo de Academia y Ciencia en IDERA, referente designado por la Universidad Nacional del Comahue y por la provincia del Neuquén y representante regional por la región Patagonia en IDERA. Ex-fellow del Instituto de Tecnología de Software en la Universidad de Naciones Unidas (IIST-UNU). Es autor de numerosas publicaciones a nivel nacional e internacional.