

Smart corporate inspection: RAG applied to the Public Registry of commercial companies

Eckert Karina B.^{1,2} , Ruidías Héctor J.¹ , Rodoni Lucas M.¹ , and
Miralles Camila G.^{2,3} 

¹ Universidad Gastón Dachary (UGD), Posadas, Misiones, Argentina

² Universidad Nacional de Misiones (UNaM), Posadas, Misiones, Argentina

³ Dirección General de Personas Jurídicas y Registro Público de la Provincia de Misiones, Posadas, Misiones, Argentina

karinaeck@gmail.com, hruidias@ugd.edu.ar, lucasrodoni@gmail.com,
camilagmiralles@gmail.com

Abstract. This work presents the application of Retrieval-Augmented Generation (RAG) in the field of the Public Registry of commercial companies, following the guidelines for reporting suspicious transactions (ROS) issued by the Financial Information Unit (UIF) regarding obligated entities and corporate structures. It sets out the main technical and legal considerations for querying digitized records, addressing the challenges posed by the condition of historical documents and the need for integration between software and critical human supervision.

Keywords: RAG, LLM, money laundering, UIF

Fiscalización societaria inteligente: RAG aplicado al Registro Público de Sociedades

Abstract. En este trabajo se presenta la aplicación de la Generación Aumentada por Recuperación (RAG) en el ámbito del Registro Público de Sociedades comerciales siguiendo los lineamientos para el reporte de operaciones sospechosas (ROS) de la Unidad de Información Financiera (UIF) sobre sujetos obligados y estructuras societarias. Se plantea las principales consideraciones técnicas y legales para consultar protocolos digitalizados, abordando los desafíos que presenta el estado de los documentos históricos y la necesidad de una integración entre el software y la supervisión humana crítica.

Keywords: RAG, LLM, lavado de dinero, UIF

1 Introducción

El presente trabajo describe un escenario de aplicación de Generación Aumentada por Recuperación (RAG) en el Registro Público de Sociedades, orientado a fortalecer las capacidades de fiscalización y cumplimiento de los lineamientos de la Unidad de Información Financiera (UIF) del estado nacional. La fiscalización de estructuras societarias es un pilar fundamental en la prevención del lavado de activos y la evasión fiscal, marco regulado en Argentina a través de la Ley 25.246 “InfoLeg - Información Legislativa”, 2000.

En dicho contexto, el Reporte de Operación Sospechosa (ROS) constituye la herramienta clave para informar transacciones que, de acuerdo con los estándares del GAFI (Grupo de Acción Financiera Internacional o FATF, por su sigla en inglés) “GAFI”, 2016, carecen de justificación económica o jurídica lícita. Históricamente, la consulta de protocolos (libros de actas) ha sido una tarea manual y lenta debido a la dispersión física de los registros.

En el año 2022 el Ministerio de Gobierno de la Provincia de Misiones inició un proceso de Modernización de la Dirección de Personas Jurídicas y Registro Público. Como parte de dicho proceso se llevó a cabo la digitalización del acervo documental de Protocolos de un período de más de 50 años, de información histórica de asociaciones civiles y sociedades comerciales.

Con un estado de conservación de moderado a bueno, se pudo obtener una versión en imágenes de tales documentos, de los que luego se extrajo el texto a través de OCR (del inglés, Reconocimiento Óptico de Caracteres) con TesseractSmith, 2007, con la finalidad realizar búsquedas dentro de los documentos e implementar técnicas de Procesamiento de Lenguaje Natural (NLP, del inglés, Natural Language Processing) para optimizar la identificación de irregularidades, donde se utilizaron diversas técnicas de identificación de entidades nombradas (NER por Name Entity Recognition) en forma automática Barlaug and Gulla, 2021 Getoor and Machanavajjhala, 2012.

No obstante, no todos los nombres de personas físicas y jurídicas resultan del todo legibles ya en el material de origen, y a esto se suma que no siempre se asientan los nombres de la misma forma, o que no se registra toda la información relevante para la identificación de los mismos (por ejemplo: no se carga el DNI), los nombres de las personas físicas por otro lado se registran de distintas maneras a lo largo de un mismo protocolo (por ejemplo en un acta de constitución), vicios de forma que dificulta la identificación. Pese a que la digitalización facilitó el acceso a los documentos, persistieron dificultades para detectar irregularidades de manera efectiva.

Con el auge de los Grandes Modelos de Lenguaje (LLM, del inglés Large Language Model) que se dio en los últimos años, y posteriormente el surgimiento de los RAG que sumó la posibilidad de integrar documentos propios a las respuestas brindadas por tales modelos, se vio la oportunidad de aprovecharlos para poder sortear las limitaciones mencionadas, tanto de los errores de OCR, como la diversidad de formas en las que se presenta la información relativas a las personas físicas y jurídicas.

2 Diseño propuesto para la Fiscalización con RAG

En esta sección se propone el diseño de arquitectura básica que permite que el analista pueda indagar por relaciones de interés en la base documental a través del RAG.

Cabe mencionar que debido a que los LLM son modelos de inteligencia artificial que son esencialmente de naturaleza estocástica, son propensos a un efecto conocido como *alucinación*, por lo que en contexto sensibles hay que diseñar soluciones que integren salvaguardas técnica y operativas.

Es por ello que en este contexto se propone que el sistema opere bajo un paradigma de HITL (Human-in-the Loop), donde el LLM actúa como un motor de descubrimiento masivo pero la decisión administrativa final permanece supeditada al análisis del experto.

1. **Consulta (Entrada):** El analista parte de un criterio específico de sospecha (ej: "Identificar sociedades con cambios recientes en su composición accionaria que compartan un mismo domicilio").
2. **Ingesta y Vectorización:** Los PDF procesados por OCR se fragmentan y convierten en embeddings. El uso de vectores es crítico para mitigar errores de lectura del OCR, permitiendo identificar términos semánticamente cercanos a pesar de ruidos en la imagen (ej: "Juan Pérez" vs. "Juan P6rez").
3. **Recuperación Semántica (Retrieval):** El sistema busca en la base de datos vectorial los fragmentos de texto más relevantes para la hipótesis planteada.
4. **Generación con Trazabilidad:** El LLM sintetiza la información encontrada. Dada la naturaleza jurídica, cada respuesta debe incluir hipervínculos directos al PDF original, resaltando el Tomo y Folio de donde se extrajo la información. Esto permite al funcionario validar la evidencia antes de remitir cualquier informe a la UIF.

3 Discusión sobre la sensibilidad de Datos y la soberanía tecnológica

Un aspecto crítico del diseño planteado descansa sobre el manejo información sensible, no solamente sobre inferencias incorrectas que pueda realizar el modelo, sino también de si el resguardo de la misma es el apropiado seguridad informática. Aunque los documentos están alojados localmente, el uso de modelos en la nube plantea desafíos para la privacidad, y conviene revisar esto a la luz de la legislación vigente, tal como es el caso de la Ley 25326 de Datos Personales "PROTECCION DE LOS DATOS", 2000. Es por ello que plantea una línea de trabajo con referentes de la DGPJyRP para abordar la dos problemáticas antes mencionadas:

1. Seguridad de la Información: El sistema debe tratar los resultados como "sospechas de coincidencia" y no como verdades absolutas, requiriendo siempre la supervisión humana crítica.

2. Soberanía de Datos: Se plantea como discusión la necesidad de avanzar hacia modelos de lenguaje de código abierto ejecutados on-premise. Esto garantizaría que los datos de protocolos no abandonen el perímetro de seguridad del Estado, manteniendo el control total sobre la información sensible durante el proceso de generación.

4 Conclusiones y Trabajo Futuro

La aplicación de RAG sobre el patrimonio documental digitalizado representa un salto cualitativo en la capacidad de respuesta clave para Sujetos Obligados que cuentan con un aservo documental crítico como lo es la DGPJyRP. El enfoque adoptado en este trabajo transforma una tarea antes inabarcable en un proceso ágil y auditable. Sin embargo, dada la complejidad de buscar sobre documentos muy heterogéneos tanto en estado de conservación como en la naturaleza de su confección, puede no resultar suficiente depender del RAG convencional, por lo que se perfila como trabajo futuro, su evolución hacia GraphRAGEdge et al., 2025, integrando grafos de conocimiento para detectar automáticamente vínculos camuflados entre personas y sociedades que no poseen identificadores unívocos en los registros históricos.

References

- Barlaug, N., & Gulla, J. A. (2021). Neural Networks for Entity Matching: A Survey. *ACM Trans. Knowl. Discov. Data*, 15(3), 52:1–52:37. <https://doi.org/10.1145/3442200>
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2025, February). From Local to Global: A Graph RAG Approach to Query-Focused Summarization [arXiv:2404.16130 [cs]]. <https://doi.org/10.48550/arXiv.2404.16130>
- GAFI. (2016, October). Retrieved April 17, 2026, from <https://www.argentina.gob.ar/uif/internacional/gafi>
- Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. *Proc. VLDB Endow.*, 5(12), 2018–2019. <https://doi.org/10.14778/2367502.2367564>
- InfoLeg - Información Legislativa. (2000, April). Retrieved April 17, 2026, from <https://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=62977>
- PROTECCION DE LOS DATOS. (2000). Retrieved April 18, 2026, from <https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm>
- Smith, R. (2007). An Overview of the Tesseract OCR Engine [ISSN: 2379-2140]. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2, 629–633. <https://doi.org/10.1109/ICDAR.2007.4376991>