SECOND SOUTH AMERICAN NLP SCHOOL · BUENOS AIRES · AGOSTO 2026

Nexo-Memoria

Un grafo de conocimiento para la Memoria, Verdad y Justicia.

Extracción de información con LLMs sobre sentencias por crímenes de lesa humanidad cometidos durante la última dictadura cívico-militar (1976–1983).

Nicolás Rubinstein · **Carolina Villella** · **Edgar Altszyler** · **Luciano Del Corro**

01 Un corpus grande y fragmentado

Miles de páginas de sentencias públicas registran víctimas, perpetradores, centros clandestinos y cadenas de mando. Por su volumen y dispersión, el abordaje de este corpus ha sido fragmentario, tanto para fines académicos como para la extracción de información relevante para el propósito de organizaciones como Abuelas de Plaza de Mayo.

02 De dónde salen los datos

CONSEJO DE LA MAGISTRATURA

Repositorio de Causas de Lesa Humanidad

`sentencias completas · PDF`

SECRETARÍA DE DERECHOS HUMANOS

Juicios de Lesa Humanidad

`datos estructurados · estado procesal, imputados, víctimas`

+ PRÓXIMAS FUENTES

03 Por qué un grafo

Un grafo modela el corpus de una forma especialmente adecuada a las preguntas del dominio: como una red, con los vínculos en primer plano. Cada hecho extraído es un **triplet** — un formato que además se acopla bien a la extracción con LLMs:

`«Juan Pérez»` — `fue_detenido_en` → `«ESMA»`

En RDF, sobre esos triplets se *razona* con una ontología propia del dominio, interoperable con estándares preexistentes (OWL-Time, GeoSPARQL):

`evento` → `ESMA` ⊂ `CABA` ⊂ `Argentina`

«eventos en Argentina» recupera el hecho aunque ningún documento lo diga literalmente: la lógica vive en la jerarquía, no en cada triplet.

Y la proyección a LPG habilita el análisis topológico — caminos, vecindarios, centralidad, comunidades. Los traversals son operaciones nativas que permiten descubrir conexiones indirectas, a varios saltos (multi-hop), que no constan explícitas en ningún documento individual.

nexo-memoria.abuelas.org.ar

El corpus en números

936	211k	≈125M
DOCUMENTOS	PÁGINAS	TOKENS (EST.)

391 veredictos y resoluciones (prom. 15 págs.) · 545 fundamentos de sentencia (prom. 376 págs.)

477 expedientes · 7.018 víctimas · 693 crímenes · 1.500 imputados

Según metadatos del Consejo de la Magistratura y la Secretaría de DD. HH.

04 Extracción: del texto a los triplets

Preprocesamiento `Docling`

PDF → texto semi-estructurado: secciones, notas, paratexto. Preserva la ubicación de cada hecho para trazabilidad.

Clasificación `LLM`

Se jerarquizan chunks ricos en información sustantiva.

Extracción semántica `LLM`

1 · Entidades personas, organizaciones, lugares, fechas

2 · Relaciones predicados del vocabulario formal

3 · Eventos hechos complejos con participantes, lugar y tiempo

+ openIE en paralelo: capa exploratoria fuera del grafo canónico. Detecta déficits del vocabulario.

Consolidación `determinístico + LLM`

Resolución de correferencia; deduplicación de entidades intra-documento y entre documentos.

Validación `SHACL`

Reglas estructurales del dominio.

GraphDB verdad semántica · RDF · OWL · SHACL → **Neo4j** proyección operacional · traversals · GDS · UI + **Postgres** procedencia · revisión · estado del pipeline

05 Revisión humana y procedencia

La extracción automática se complementa con revisión experta: el equipo de investigación de Abuelas valida y corrige los datos en una interfaz dedicada. Las correcciones se registran en un grafo separado del canónico, con autoría y fecha, y cada hecho conserva la referencia al documento y fragmento de origen.

`abierta` → `en revisión` → `decidida` → `implementada` / `descartada`

Ciclo de vida de cada anotación — comentarios, correcciones, merges y splits de entidades — con autoría y tiempos persistidos.

06 Resolución de entidades

La unificación de menciones combina similitud léxica (Jaro-Winkler, Jaccard) con evidencia contextual: aliases explícitos en el texto, coocurrencia espacio-temporal y diccionarios del dominio.

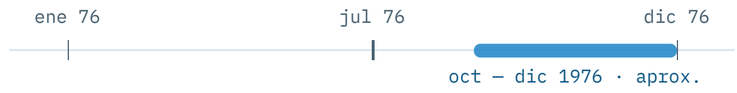
reject	candidate → LLM	auto_merge
descartado	same / different / uncertain	unificación directa

Los pares ambiguos se difieren a un resolver LLM y, en última instancia, a revisión humana.

07 Tiempo y lugar, hechos interoperables

Las fechas de los expedientes rara vez son precisas. Un parser de patrones regex en castellano las convierte en intervalos OWL-Time con precisión e incertidumbre explícitas; las 13 relaciones de Allen razonan sobre solapamientos.

«a fines de 1976»



Los lugares se canonicalizan contra un diccionario curado (CONADEP, RUVTE, Memoria Abierta): «El Atlético» ≡ «Club Atlético». El diccionario asocia coordenadas a los centros clandestinos, lo que habilita razonamiento espacial — pertenencia a jurisdicciones, proximidad — mediante GeoSPARQL.

08 Una ontología del dominio

OWL propio, diseñado para el contexto argentino y su aparato judicial — con tipos específicos de centros clandestinos, jerarquías de fuerzas y relaciones procesales entre causas — procurando compatibilidad con vocabularios estándar del dominio, como HURIDOCS.

`Sobreviviente` `NacimientoEnCautiverio` `InscripciónFraudulenta` `CentroClandestino`

`45 clases · 46 propiedades de objeto · 52 atributos · 29 shapes SHACL`

09 Modos de consulta

INVESTIGADORA

¿Hay personas que hayan coincidido en más de un centro de detención?

HILARIO · AGENTE

Sí. La mayoría de las coincidencias múltiples son en realidad sectores de un mismo centro («Capucha», «Sótano» dentro de la ESMA). El caso más claro entre centros distintos: C. M. V. y H. R. P. coincidieron en la Base de Buzos Tácticos de Mar del Plata y después en la ESMA.

- → **Agente conversacional** — lenguaje natural → SPARQL/Cypher; ve el grafo en pantalla.
- → **Búsqueda y filtros** — por nombre, alias o tipo (explorador Sigma.js).
- → **Análisis estructural** — centralidad, comunidades, caminos entre entidades.
- → **API y MCP** — endpoints SPARQL y Cypher para uso programático.

La evaluación usa un ground truth construido a partir de la tesis de maestría de Carolina Villella (Apropiación de niños y niñas durante el terrorismo de Estado en Argentina, UNL): sobre los primeros documentos procesados, **96,6%** *de recall en personas y* **100%** *en eventos. La configuración de extracción surge de un benchmark propio de 12 LLMs.*

PRÓXIMOS PASOS

- ▸ Procesar el corpus público completo (936 documentos)
- ▸ Integración con datos internos de Abuelas
- ▸ Expansión de la ontología con el feedback de uso

.UBA EXACTAS Facultad de Ciencias Exactas y Naturales · ABUELAS DE PLAZA DE MAYO · Quantit