

**Instituto Politécnico Nacional
Escuela Superior de Cómputo**

Sistema Inteligente para la Identificación y Seguimiento de Derechos de NNA afectados por feminicidios en México

Análisis de Requerimientos y Especificación del Sistema

Trabajo Terminal

Trabajo Terminal 1 - Semestre 2025-2026/1

Autores:

Herrera Ramírez Emilio Alejandro
Morales Martínez Héctor Alberto

Directores:

M. en C. Ulises Vélez Saldaña
Dr. Gabriel Hurtado Aviles

Ciudad de México, febrero de 2026

Project Charter

Proyecto:	CVE, Nombre proyecto.	
Responsable:	Empresa, Nombre del responsable, cargo, Firma.	
Autoriza:	Empresa, Nombre del responsable, cargo, Firma.	
Background/Contexto:	Descripción breve del contexto, no mas de 3 líneas.	
Beneficios esperados:	Principales beneficios al término del proyecto.	
Costo estimado:	\$ 2,350,700.00 ± 13 % (por ejemplo.)	
Fecha de inicio:	Fecha	Fecha de término: Fecha.
Objetivo:	Objetivo general del proyecto.	
Entregables Principales		
	Clave-Nombre	descripción del entregable
	Clave-Nombre	descripción del entregable
	...	
Alcance del proyecto		
Incluye:	<ul style="list-style-type: none"> • Elemento 1 del alcance que incluye. • ... 	
Excluye:	<ul style="list-style-type: none"> • Elemento 1 del alcance que incluye. • ... 	
Criterio de éxito:	Indicador clave de término del proyecto	
Metodología:	Metodología o metodologías que se utilizan (dos renglones o lista de no mas de 7)	
Datos de contacto		
Project Manager:	Nombre, Tel, correo, etc.	
Project owner:	Nombre, Tel, correo, etc.	
...		
Riesgos y peligros:	<ul style="list-style-type: none"> • Riesgo o peligro identificado. • ... 	
Supuestos:	<ul style="list-style-type: none"> • Suposiciones hechas de las que depende el éxito del proyecto. • ... 	
Restricciones y dependencias:	<ul style="list-style-type: none"> • Restricciones del proyecto. • ... 	
Supervisión		
Juntas:	(Nombre de la(s) persona(s)),	reporta a (Nombre de la(s) persona(s))
Dudas:	(Nombre de la(s) persona(s)),	reporta a (Nombre de la(s) persona(s))
Avances:	(Nombre de la(s) persona(s)),	reporta a (Nombre de la(s) persona(s))
...		

Tabla 1: Resumen del proyecto

Índice general

Resumen Ejecutivo	XIII
1. Introducción	1
1.1. Presentación	1
1.2. Organización del contenido	2
1.3. Notación, símbolos y convenciones utilizadas	3
2. Modelo del alcance	5
2.1. Análisis de la problemática	5
2.1.1. Contexto del proyecto	5
2.1.2. Problemas identificados	6
2.1.3. Análisis de causas probables	7
2.1.4. Análisis de posibles consecuencias	8
2.1.5. Características de la solución	9
2.1.6. Síntesis de la problemática	11
2.2. Objetivos del proyecto	11
2.2.1. Objetivo general	11
2.2.2. Objetivos específicos	12
2.3. Usuarios identificados	13
2.4. Procesos involucrados	14
2.5. Requerimientos de usuario	16
2.6. Especificación de plataforma	20
3. Modelo del Negocio	25
3.1. Actores del sistema	25
3.1.1. Coordinador de Documentación	25
3.1.2. Analista de Datos	27
3.1.3. Investigador Académico	28
3.1.4. Funcionario de Política Pública	29

3.1.5. Periodista de Investigación	31
3.1.6. Administrador del Sistema	32
3.2. Términos del Negocio	33
3.3. Modelo del dominio del problema	36
3.3.1. Entidad: Noticia	38
3.3.2. Entidad: Medio de Comunicación	39
3.3.3. Entidad: Cluster de Noticias	40
3.3.4. Entidad: Mención de NNA	41
3.3.5. Entidad: Tópico LDA	41
3.3.6. Entidad: Caso de Feminicidio	42
3.3.7. Entidad: Fuente RSS	43
3.4. Modelado de Reglas de negocio	43
3.5. Máquinas de estado	44
3.5.1. Estados para un préstamo	44
4. Modelo dinámico	47
4.1. Descripción de casos de uso	47
4.2. CUX Escriba el nombre del caso de uso	49
4.2.1. Descripción completa	49
4.2.2. Atributos importantes	49
4.2.3. Trayectorias del Caso de Uso	50
4.2.4. Puntos de extensión	50
4.3. CUX Escriba el nombre del caso de uso	51
4.3.1. Descripción completa	51
4.3.2. Atributos importantes	51
4.3.3. Trayectorias del Caso de Uso	53
4.3.4. Puntos de extensión	53
5. Modelo de la interacción	55
5.1. Modelo de navegación	55
5.2. IUX Interfaz (nombre de la interfaz)	56
5.2.1. Objetivo	56
5.2.2. Diseño	56
5.2.3. Salidas	56
5.2.4. Entradas	56
5.2.5. Comandos	56
5.3. Catálogo de mensajes	57
5.3.1. Lista de mensajes	57
A. Anexos	63
A.1. Glosario de Términos	63
A.2. Configuración del Entorno de Desarrollo	64
A.3. Ejemplos de Casos de Uso	65
A.4. Código de Ejemplo: Detector de Relevancia	65

A.5. Diagramas Complementarios	65
A.6. Referencias de Contacto	66

Índice de figuras

2.1. Organigrama de usuarios del sistema de identificación y seguimiento de NNA afectados por feminicidios	13
2.2. Mapa de procesos de organización civil con enfoque en documentación de casos de feminicidio y atención a NNA afectados	15
2.3. Arquitectura del sistema de identificación y seguimiento de NNA afectados por feminicidios	21
3.1. Modelo del dominio del problema - Sistema de identificación de NNA afectados por feminicidios	37
3.2. Máquina de estados de un Préstamo.	45
4.1. Diagrama de casos de uso del sistema.	47
4.2. Diagrama detallado del sistema.	48
5.1. mapa	55
5.2. IU23 Pantalla de Control de Acceso.	56

Índice de tablas

1. Resumen del proyecto	IV
-----------------------------------	----

Resumen Ejecutivo

Contexto del Proyecto

Entre 2010 y 2023, más de 8,000 mujeres fueron víctimas de feminicidio en México, dejando miles de niñas, niños y adolescentes en situación de orfandad. La información sobre estos casos permanece dispersa en notas periodísticas, comunicados de fiscalías y archivos de organizaciones civiles, impidiendo dimensionar el problema de manera estructurada y diseñar políticas públicas efectivas.

Objetivo General

Desarrollar un sistema inteligente que automatice la recolección, procesamiento y estructuración de información relacionada con niñas, niños y adolescentes afectados por feminicidios en México, mediante técnicas de procesamiento de lenguaje natural y análisis de datos, para transformar información dispersa en datos consultables y accionables.

Alcance de la Fase TT1

Durante el primer semestre del Trabajo Terminal se desarrollaron e implementaron tres prototipos incrementales:

- **Prototipo 0 (P0):** Aproximación inicial mediante web scraping directo que identificó limitaciones técnicas y legales de acceso a fuentes periodísticas.
- **Prototipo 1 (P1):** Sistema básico implementando recolección mediante RSS feeds, vectorización TF-IDF y clustering con K-Means para agrupar noticias relacionadas.
- **Prototipo 2 (P2):** Sistema avanzado con arquitectura Docker, triple estrategia de recolección de datos (RSS, scraping de listados, APIs periodísticas), detector de relevancia

de dos etapas, pipeline completo de PLN, API REST y dashboard web para visualización.

Resultados Principales

El Prototipo 2 demostró la viabilidad técnica del enfoque propuesto con los siguientes resultados:

- Recolección automática de 281 noticias únicas en periodo de prueba
- Precisión del 90 % en detección de casos relevantes
- Tasa de falsos positivos del 10 %
- Cobertura de múltiples entidades federativas
- Generación automática de clusters temáticos coherentes

Trabajo a Futuro (TT2)

La segunda fase del proyecto se enfocará en:

1. Implementación de clasificación semántica con transformers (BERT multilingüe)
2. Extracción de entidades nombradas (NER) para identificación automática de nombres, ubicaciones y fechas
3. Migración a base de datos relacional (PostgreSQL)
4. Ampliación de cobertura temporal (2015-2025)
5. Desarrollo de interfaz de usuario completa para organizaciones civiles
6. Sistema de validación y corrección manual supervisada
7. Módulo de análisis estadístico y visualizaciones avanzadas

Impacto Esperado

El sistema permitirá a organizaciones como la Fundación Futuro con Derechos y la REDIM contar con información estructurada y actualizada sobre NNA afectados por feminicidios, facilitando el diseño de estrategias de acompañamiento, defensa de derechos y generación de políticas públicas basadas en evidencia.

CAPÍTULO 1

Introducción

Este documento contiene el análisis de requerimientos del proyecto “*Sistema Inteligente para la Identificación y Seguimiento de Derechos de Niñas, Niños y Adolescentes (NNA) afectados por feminicidios en México*” que servirá como base para el análisis, diseño, construcción, pruebas y aceptación del proyecto.

El documento fue elaborado por Herrera Ramírez Emilio Alejandro y Morales Martínez Héctor Alberto, estudiantes de Ingeniería en Sistemas Computacionales de la Escuela Superior de Cómputo del Instituto Politécnico Nacional, bajo la dirección del M. en C. Ulises Vélez Saldaña y el Dr. Gabriel Hurtado Aviles, durante el periodo académico 2025-2026 en la Ciudad de México.

1.1. Presentación

Entre 2010 y 2023, más de 8,000 mujeres fueron víctimas de feminicidio en México según la Secretaría de Gobernación. Detrás de estas cifras existen miles de niños, niñas y adolescentes que quedaron huérfanos. La Red por los Derechos de la Infancia en México (REDIM) estima al menos 3,500 NNA que perdieron a su madre por feminicidio, pero no existe un censo oficial ni sistema de seguimiento estructurado. La información permanece dispersa en notas periodísticas, comunicados de fiscalías estatales y archivos de organizaciones civiles, lo que impide dimensionar el problema, identificar patrones geográficos o temporales, y diseñar políticas públicas efectivas para garantizar los derechos de estas infancias afectadas.

Las organizaciones de la sociedad civil, como la Fundación Futuro con Derechos, dedican recursos considerables a la revisión manual de medios de comunicación buscando casos donde se mencione a hijos de víctimas. Sin embargo, la metodología manual es insuficiente ante el volumen de información disponible, resultando en cobertura limitada y muchos casos que pasan desapercibidos. Se requiere un sistema automatizado que recolecte, procese y estructure esta información dispersa para transformarla en datos consultables y accionables.

El propósito de este documento es presentar el análisis detallado de requerimientos funcionales y no funcionales del sistema, describir la arquitectura técnica implementada mediante prototipos incrementales, documentar las estrategias de recolección de datos y procesamiento de lenguaje natural, y establecer las bases para las fases siguientes del proyecto. Este documento está dirigido a los directores del trabajo terminal, sinodales evaluadores, organizaciones de la sociedad civil interesadas en la problemática, y desarrolladores que participen en la evolución del sistema.

El documento debe utilizarse como referencia técnica para comprender las decisiones de diseño, validar el cumplimiento de objetivos establecidos en TT1, identificar limitaciones reconocidas que serán abordadas en TT2, y servir como punto de partida para la documentación de usuario final y manuales técnicos que se desarrollarán en la segunda fase del proyecto.

1.2. Organización del contenido

El presente documento se estructura en ocho capítulos que describen de manera integral el desarrollo del sistema:

El **Capítulo 1: Introducción** presenta el contexto del proyecto, planteamiento del problema, justificación social y tecnológica, objetivos generales y específicos, así como los alcances y limitaciones de la fase TT1.

El **Capítulo 2: Marco Teórico** establece los fundamentos conceptuales necesarios para comprender el sistema, incluyendo procesamiento de lenguaje natural, técnicas de web scraping, clustering, modelado de tópicos, métricas de evaluación y tecnologías utilizadas (Python, Flask, Docker, scikit-learn).

El **Capítulo 3: Estado del Arte** analiza proyectos relacionados a nivel nacional e internacional, incluyendo el Mapa de Feminicidios en México, Data Cívica, sistemas de monitoreo de medios, y plataformas académicas de análisis de noticias, identificando fortalezas y diferencias con el presente proyecto.

El **Capítulo 4: Metodología de Desarrollo** documenta la evolución a través de tres prototipos: P0 (web scraping directo fallido), P1 (sistema básico con RSS y K-Means), y P2 (sistema avanzado con triple estrategia de recolección y detector especializado), describiendo implementación, resultados y lecciones aprendidas de cada fase.

El **Capítulo 5: Desarrollo y Arquitectura de la Solución** detalla la arquitectura Docker implementada, componentes del sistema (recolector multi-fuente, detector dual-etapa, pipeline PLN, API REST, dashboard web), flujo de datos, y decisiones técnicas de diseño.

El **Capítulo 6: Resultados Preliminares y Pruebas** presenta las métricas de desempeño del Prototipo 2, incluyendo cobertura de recolección (281 noticias únicas), precisión del detector (90 % con 10 % falsos positivos), análisis de clusters generados, y evaluación de calidad mediante coherencia de tópicos LDA.

El **Capítulo 7: Conclusiones Parciales** sintetiza los logros de TT1, valida la viabilidad técnica del enfoque propuesto, reconoce limitaciones identificadas, y establece las bases para la continuación del proyecto.

El **Capítulo 8: Trabajo a Futuro** define el alcance de TT2, priorizando la implementación de clasificación semántica con transformers, extracción de entidades nombradas (NER), mi-

gración a base de datos relacional, ampliación de cobertura temporal, y desarrollo de interfaz de usuario completa para organizaciones civiles.

1.3. Notación, símbolos y convenciones utilizadas

El documento emplea las siguientes convenciones de notación y estándares de documentación:

- **Nomenclatura de Prototipos:** Los prototipos se identifican como P0, P1 y P2, representando las fases de desarrollo Prototipo 0 (fallido), Prototipo 1 (básico) y Prototipo 2 (avanzado), respectivamente.
- **Identificadores de Componentes:** Los componentes del sistema se nombran en formato `snake_case` según convenciones Python (ejemplo: `data_collector.py`, `simplified_analyzer.py`)
- **Acrónimos y Abreviaturas:**
 - **NNA:** Niñas, Niños y Adolescentes
 - **REDIM:** Red por los Derechos de la Infancia en México
 - **PLN:** Procesamiento de Lenguaje Natural
 - **RSS:** Really Simple Syndication
 - **API:** Application Programming Interface
 - **REST:** Representational State Transfer
 - **TF-IDF:** Term Frequency - Inverse Document Frequency
 - **LDA:** Latent Dirichlet Allocation
 - **DBSCAN:** Density-Based Spatial Clustering of Applications with Noise
 - **NER:** Named Entity Recognition
 - **TT1/TT2:** Trabajo Terminal semestre 1 y semestre 2
- **Referencias Cruzadas:** Las referencias a figuras, tablas y capítulos utilizan comandos LaTeX estándar (`\ref{}`, `\label{}`) para mantener consistencia automática.
- **Código Fuente:** Los fragmentos de código se presentan utilizando el entorno `lstlisting` con resaltado de sintaxis para Python, con numeración de líneas cuando sea relevante.
- **Métricas y Valores:** Las métricas de evaluación se expresan en porcentaje cuando representan tasas (precisión, recall), y en valores absolutos para conteos (número de noticias, features, clusters).
- **Énfasis Tipográfico:**
 - Cursivas para términos técnicos en su primera aparición o conceptos destacados

- **Negritas** para elementos de especial relevancia o términos clave en definiciones
 - Monoespaciado para nombres de archivos, comandos, código, URLs y parámetros técnicos
- **Estándar de Documentación:** El documento sigue las convenciones de trabajos terminales de ESCOM-IPN, con estructura de capítulos, formato de referencias bibliográficas estilo APA, y numeración jerárquica de secciones.
 - **Notación Matemática:** Las fórmulas matemáticas (similitud coseno, TF-IDF, métricas de evaluación) se expresan utilizando el entorno matemático de LaTeX con la notación estándar de álgebra lineal y estadística.

El documento utiliza el sistema de gestión de referencias bibliográficas BibTeX con estilo apalike, asegurando consistencia en las citas y formato de la bibliografía conforme a estándares académicos internacionales.

CAPÍTULO 2

Modelo del alcance

Este capítulo presenta el análisis detallado de la problemática abordada por el proyecto, identificando el contexto social y técnico que motiva el desarrollo del sistema. Se describen los problemas específicos detectados, sus causas y consecuencias, así como las características de la solución propuesta. Posteriormente se definen los objetivos del proyecto, usuarios identificados, procesos involucrados, requerimientos de usuario y la especificación de la plataforma tecnológica que soportará el sistema de identificación y seguimiento de derechos de NNA afectados por feminicidios en México.

2.1. Análisis de la problemática

La problemática se estructura en tres dimensiones: social (orfandad por feminicidio sin seguimiento institucional), organizacional (trabajo manual insostenible en organizaciones civiles) y técnica (información no estructurada dispersa en múltiples fuentes). El análisis aborda el contexto del feminicidio en México, identifica problemas específicos de recolección y estructuración de datos, examina causas relacionadas con falta de sistemas automatizados y dispersión de fuentes, evalúa consecuencias de la invisibilización de casos, y propone una solución tecnológica basada en web scraping inteligente y procesamiento de lenguaje natural.

2.1.1. Contexto del proyecto

Entre 2010 y 2023, más de 8,000 mujeres fueron víctimas de feminicidio en México según datos de la Secretaría de Gobernación. Cada uno de estos casos no solo representa la pérdida de una vida, sino que deja tras de sí niñas, niños y adolescentes en situación de orfandad. La Red por los Derechos de la Infancia en México (REDIM) estima que al menos 3,500 NNA perdieron a su madre por feminicidio durante este periodo, aunque no existe un censo oficial que documente con precisión esta cifra.

La información sobre estos casos se encuentra dispersa en múltiples fuentes: notas periodísticas de medios locales y nacionales, comunicados oficiales de fiscalías estatales, reportes de organizaciones de la sociedad civil, y testimonios en redes sociales. Esta fragmentación impide dimensionar el problema de manera integral, identificar patrones geográficos o temporales, rastrear el seguimiento de casos individuales, y evaluar si los NNA afectados reciben la atención psicológica, educativa y jurídica a la que tienen derecho según la legislación mexicana.

Organizaciones como la Fundación Futuro con Derechos y el colectivo Data Cívica han documentado la dificultad de mantener bases de datos actualizadas mediante metodologías manuales. El Mapa de Feminicidios en México, proyecto colaborativo de registro ciudadano, ha logrado documentar miles de casos mediante trabajo voluntario, pero reconoce limitaciones de recursos humanos para revisar exhaustivamente todas las fuentes disponibles diariamente. Las autoridades gubernamentales, por su parte, no cuentan con un sistema integrado que vincule casos de feminicidio con el estatus de los hijos de la víctima, dificultando la implementación efectiva de programas de reparación del daño.

El contexto tecnológico actual ofrece herramientas de procesamiento de lenguaje natural, web scraping automatizado y análisis de datos masivos que no han sido aplicadas sistemáticamente a esta problemática. Experiencias internacionales en monitoreo automatizado de noticias (como el Global Database of Events, Language and Tone - GDELT) demuestran la viabilidad técnica de recolectar y estructurar información de fuentes públicas a gran escala.

2.1.2. Problemas identificados

El problema general que atiende el presente proyecto es:

“La inexistencia de un sistema automatizado para identificar, recolectar y estructurar información sobre niñas, niños y adolescentes afectados por feminicidios en México, lo que impide a organizaciones civiles, autoridades gubernamentales e investigadores dimensionar el problema, dar seguimiento a casos individuales, identificar patrones epidemiológicos y evaluar la efectividad de programas de atención a esta población vulnerable.”

Los problemas identificados son¹

Id	Nombre	Descripción	Pri.
P-01	Dispersión de información	La información sobre feminicidios y NNA afectados se encuentra fragmentada en cientos de medios de comunicación locales y nacionales, comunicados gubernamentales y reportes de organizaciones civiles, sin un repositorio centralizado que permita su consulta sistemática.	A

¹ La prioridad en la tabla está indicada como: MB - Muy Baja, B - Baja, M - Media, A - Alta, MA - Muy Alta.

Id	Nombre	Descripción	Pri.
P-02	Trabajo manual insostenible	Las organizaciones civiles dedican entre 10 y 15 horas semanales a revisar manualmente sitios de noticias buscando casos relevantes, proceso que no escala ante el volumen de información publicada diariamente y resulta en cobertura parcial.	A
P-03	Ausencia de datos estructurados	La información existe como texto no estructurado en artículos periodísticos, dificultando búsquedas, filtrado por criterios específicos (ubicación, edad de NNA, circunstancias del caso), análisis estadístico y generación de reportes.	A
P-04	Falta de identificación automatizada	No existen herramientas especializadas que distingan automáticamente noticias de feminicidio de otros delitos violentos, ni que identifiquen menciones de hijos huérfanos dentro del contenido noticioso.	M
P-05	Imposibilidad de análisis de patrones	La dispersión y falta de estructura impide análisis de patrones geográficos (entidades con mayor incidencia), temporales (tendencias anuales), demográficos (edades de NNA afectados) y contextuales (circunstancias del feminicidio).	M
P-06	Bloqueos técnicos de scraping	Los intentos de web scraping directo enfrentan bloqueos anti-bot (HTTP 403/406), protecciones Cloudflare, y contenido generado dinámicamente por JavaScript, impiadiendo la recolección automatizada tradicional.	M
P-07	Duplicación y ruido informativo	La misma noticia es republicada por múltiples medios con variaciones menores, generando duplicados que inflan conteos artificialmente y dificultan identificar casos únicos.	B
P-08	Invisibilización de casos	Los casos que no reciben cobertura mediática nacional o que son publicados exclusivamente en medios locales de difícil acceso quedan fuera del seguimiento de organizaciones con recursos limitados.	B

2.1.3. Análisis de causas probables

Las causas identificadas que generan y perpetúan los problemas señalados son:

P-01 Dispersión de información: No existe coordinación interinstitucional entre autoridades estatales, federales, medios de comunicación y organizaciones civiles para centralizar datos. Las fiscalías estatales operan de manera independiente sin protocolo unificado de publicación de información. Los medios de comunicación priorizan inmediatez noticiosa sobre estructuración de datos.

- P-02 Trabajo manual insostenible:** Las organizaciones civiles carecen de presupuesto para contratar personal dedicado exclusivamente a monitoreo de medios. No cuentan con conocimientos técnicos especializados en desarrollo de sistemas automatizados. Dependen de trabajo voluntario que es intermitente y no profesionalizado.
- P-03 Ausencia de datos estructurados:** Los artículos periodísticos se redactan en lenguaje natural sin formato estandarizado. No existe obligación legal para que medios o autoridades publiquen información en formatos estructurados (JSON, XML, bases de datos abiertas). Las plataformas de gestión de contenido de medios no están diseñadas para extracción sistemática de datos.
- P-04 Falta de identificación automatizada:** La complejidad del lenguaje natural requiere técnicas avanzadas de PLN que no están implementadas. Las menciones de NNA son contextuales y variadas (“dejó dos hijos”, “madre de tres menores”, “sus niños quedaron huérfanos”), dificultando detección por keywords simples. No existen datasets etiquetados específicos de feminicidios en español para entrenar modelos supervisados.
- P-05 Imposibilidad de análisis de patrones:** Los datos no estructurados no permiten análisis cuantitativo directo. La falta de identificadores únicos de casos impide seguimiento longitudinal. Las organizaciones no cuentan con herramientas de análisis de datos masivos ni visualización de patrones geoespaciales.
- P-06 Bloqueos técnicos de scraping:** Los medios implementan protecciones anti-bot para prevenir sobrecarga de servidores y posible robo de contenido. Cloudflare y servicios similares dificultan scraping legítimo junto con ataques maliciosos. El contenido dinámico generado por JavaScript requiere navegadores headless (Selenium, Playwright) que son lentos y detectables.
- P-07 Duplicación y ruido informativo:** Las agencias de noticias (Notimex, AP) distribuyen la misma nota base a múltiples medios. Los medios republican contenido sin agregar valor informativo nuevo. No existe identificador universal de noticias que permita detectar automáticamente fuente original vs. republicación.
- P-08 Invisibilización de casos:** Los medios locales tienen menor alcance digital y menor inversión en infraestructura web. Los casos en comunidades rurales o indígenas reciben menor cobertura periodística. Sesgos editoriales priorizan casos con características específicas (ubicación urbana, víctima con perfil mediático).

2.1.4. Análisis de posibles consecuencias

Las consecuencias inmediatas, a mediano y largo plazo si la problemática persiste son:

- P-01 Dispersión de información: Inmediato:** Imposibilidad de cuantificar con precisión la dimensión del problema. **Mediano plazo:** Diseño de políticas públicas basadas en datos incompletos o sesgados. **Largo plazo:** Perpetuación de invisibilización de víctimas indirectas de feminicidio y falta de rendición de cuentas institucional.

P-02 Trabajo manual insostenible: **Inmediato:** Agotamiento y rotación de personal voluntario en organizaciones civiles. **Mediano plazo:** Abandono de esfuerzos de documentación por inviabilidad operativa. **Largo plazo:** Pérdida de memoria histórica de casos documentados y desarticulación de movimientos de defensa de derechos de NNA.

P-03 Ausencia de datos estructurados: **Inmediato:** Dificultad para generar reportes periódicos y responder solicitudes de información de medios y academia. **Mediano plazo:** Imposibilidad de evaluar efectividad de programas gubernamentales de atención a NNA huérfanos. **Largo plazo:** Desvinculación entre producción académica, activismo civil y diseño de política pública por falta de datos comunes.

P-04 Falta de identificación automatizada: **Inmediato:** Casos relevantes pasan desapercibidos en el volumen diario de noticias. **Mediano plazo:** Subestimación sistemática del número de NNA afectados. **Largo plazo:** Naturalización social del problema por ausencia de visibilización cuantitativa del impacto en infancias.

P-05 Imposibilidad de análisis de patrones: **Inmediato:** Desconocimiento de focos rojos geográficos que requieren intervención prioritaria. **Mediano plazo:** Asignación inefficiente de recursos gubernamentales y de organizaciones civiles. **Largo plazo:** Falta de investigación epidemiológica sobre factores de riesgo y efectividad de intervenciones preventivas.

P-06 Bloqueos técnicos de scraping: **Inmediato:** Fracaso de intentos de automatización, forzando retorno a metodologías manuales. **Mediano plazo:** Retraso en adopción de soluciones tecnológicas por percepción de inviabilidad técnica. **Largo plazo:** Brecha digital persistente entre capacidades de monitoreo de organizaciones civiles vs. actores con mayores recursos.

P-07 Duplicación y ruido informativo: **Inmediato:** Inflación artificial de conteos que distorsiona percepción de magnitud del problema. **Mediano plazo:** Descrédito de cifras presentadas por organizaciones civiles cuando se detecta duplicación. **Largo plazo:** Debilitamiento de credibilidad de movimientos sociales en el debate público.

P-08 Invisibilización de casos: **Inmediato:** NNA en comunidades marginadas quedan sin apoyo de organizaciones civiles. **Mediano plazo:** Reproducción de desigualdades estructurales en el acceso a justicia y reparación del daño. **Largo plazo:** Perpetuación intergeneracional de violencia y vulnerabilidad en poblaciones ya marginadas.

2.1.5. Características de la solución

Para atender la problemática anterior se propone implementar las siguientes acciones mediante un sistema integral de recolección, análisis y estructuración de información:

P-01 Centralización mediante agregación automática: Implementar un sistema que consulte diariamente múltiples fuentes (feeds RSS de 10 medios nacionales, Google News

API para cobertura ampliada, scraping histórico del Mapa de Feminicidios) consolidando la información dispersa en un repositorio único. Almacenar metadatos estructurados: fecha, medio, URL, título, contenido, ubicación mencionada.

P-02 Automatización del monitoreo: Desarrollar un recolector automatizado que opere mediante tareas programadas (cron jobs) cada 6 horas, eliminando necesidad de revisión manual constante. Reducir la carga operativa de organizaciones civiles de 10-15 horas semanales a 2 horas de validación de casos detectados automáticamente.

P-03 Estructuración de datos: Transformar texto no estructurado en registros estructurados mediante pipeline de procesamiento: limpieza de texto (remoción stopwords, normalización), extracción de campos (fecha, ubicación, características del caso), asignación de identificadores únicos, y almacenamiento en formato consultable (CSV en TT1, base de datos relacional en TT2).

P-04 Detector especializado dual-etapa: Implementar sistema de detección en dos fases: (1) Filtro de feminicidio mediante regex y keywords especializadas (“feminicidio”, “asesinó a su pareja”, “violencia de género”), (2) Detector de menciones NNA mediante análisis contextual de patrones lingüísticos (“dejó N hijos”, “madre de”, “huérfanos”, “menores de edad”). Alcanzar precisión $\geq 90\%$ con $\leq 10\%$ falsos positivos.

P-05 Análisis de patrones mediante PLN: Aplicar técnicas de clustering (DBSCAN) para agrupar noticias del mismo caso publicadas por múltiples medios. Implementar modelo de tópicos (LDA) para identificar temas recurrentes. Generar métricas agregadas por entidad federativa, rango temporal y características del caso. Proporcionar visualizaciones en dashboard web.

P-06 Estrategia multi-fuente anti-bloqueos: Evadir bloqueos técnicos mediante triple estrategia: (1) Feeds RSS/Atom que son públicos y no requieren scraping HTML, (2) Google News API como intermediario que ya resolvió el scraping, (3) Scraping selectivo de fuentes estructuradas (Mapa de Feminicidios) con control de tasa de peticiones. Validado mediante prototipos iterativos P0→P1→P2.

P-07 Detección de duplicados: Implementar sistema de deduplicación mediante vectorización TF-IDF y cálculo de similitud coseno. Establecer threshold de similitud ≥ 0.75 para marcar noticias como duplicadas. Agrupar variantes de la misma noticia mediante clustering, manteniendo solo representante canónico de cada cluster para conteos.

P-08 Cobertura ampliada: Incluir fuentes regionales y estatales en lista de feeds RSS monitoreados. Utilizar Google News API con queries geográficamente específicas (“feminicidio Oaxaca”, “feminicidio Chiapas”) para capturar medios locales. Diseñar interfaz que permita a organizaciones locales reportar casos no detectados automáticamente (planeado TT2).

2.1.6. Síntesis de la problemática

El análisis realizado evidencia que el problema central no es la inexistencia de información sobre feminicidios y NNA afectados, sino su dispersión, falta de estructura y el carácter manual de su recolección. La información existe en el ecosistema mediático mexicano, pero permanece como texto no estructurado distribuido en cientos de sitios web sin vinculación sistemática.

Las organizaciones civiles han demostrado voluntad y capacidad de documentación, pero la metodología manual enfrenta límites de escalabilidad que ninguna cantidad de esfuerzo voluntario puede superar. Las autoridades gubernamentales reconocen el problema pero carecen de herramientas para dimensionarlo cuantitativamente y dar seguimiento individualizado.

La solución propuesta no pretende reemplazar el trabajo de organizaciones civiles, sino potenciarlo mediante automatización inteligente. El sistema actuará como filtro inicial que procesa cientos de noticias diarias, identifica las potencialmente relevantes mediante técnicas de procesamiento de lenguaje natural, las estructura en formato consultable, y las presenta para validación humana y seguimiento de casos.

Los beneficios esperados son múltiples: (1) **Operativo:** Reducción de 80 % del tiempo dedicado a búsqueda manual, liberando recursos para trabajo directo con familias afectadas, (2) **Informativo:** Base de datos estructurada que permita responder preguntas analíticas (¿cuántos casos en Jalisco en 2024?, ¿cuál es la edad promedio de NNA afectados?), (3) **Estratégico:** Identificación de patrones geográficos y temporales que orienten asignación de recursos y diseño de intervenciones, (4) **Político:** Evidencia cuantitativa para incidencia en política pública y exigencia de rendición de cuentas, (5) **Social:** Visibilización de víctimas indirectas del feminicidio que históricamente han permanecido invisibles en el debate público.

El proyecto demuestra viabilidad técnica mediante prototipos incrementales que resolvieron bloqueos iniciales ($P0 \rightarrow P1 \rightarrow P2$), alcanzando capacidad de recolección y procesamiento de cientos de noticias con precisión aceptable. La fase TT2 consolidará el sistema para uso en producción por organizaciones reales.

2.2. Objetivos del proyecto

2.2.1. Objetivo general

“Desarrollar un sistema automatizado de recolección, análisis y estructuración de información mediante web scraping multi-fuente, procesamiento de lenguaje natural y clustering para identificar y documentar casos de niñas, niños y adolescentes afectados por feminicidios en México publicados en medios de comunicación y fuentes públicas, proporcionando a organizaciones civiles, autoridades gubernamentales e investigadores una base de datos estructurada y consultable que facilite el seguimiento de casos, análisis de patrones epidemiológicos y evaluación de programas de atención, mediante la combinación de recolección multi-fuente (RSS, APIs, scraping histórico), detector especializado dual-etapa, pipeline de PLN con vectorización TF-IDF y clustering DBSCAN, API REST y dashboard”

web de visualización.”

2.2.2. Objetivos específicos

Los objetivos específicos se organizan en cuatro ejes: recolección automatizada, procesamiento inteligente, estructuración de datos y acceso a información.

- **OE-1. Implementar sistema de recolección multi-fuente:** Desarrollar un recolector automatizado que consulte feeds RSS de 10 medios nacionales, Google News API para cobertura ampliada, y scraping histórico del Mapa de Feminicidios, operando mediante tareas programadas cada 6 horas para mantener actualización continua de información.
- **OE-2. Desarrollar detector especializado dual-etapa:** Construir un sistema de identificación en dos fases que primero filtre noticias de feminicidio mediante regex y keywords especializadas, y posteriormente detecte menciones de NNA afectados mediante análisis contextual de patrones lingüísticos, alcanzando precisión $\geq 90\%$ con $\leq 10\%$ falsos positivos.
- **OE-3. Diseñar pipeline de procesamiento de lenguaje natural:** Implementar flujo completo de PLN que incluya limpieza de texto (remoción stopwords, normalización, lematización), vectorización mediante TF-IDF con 3,000 features máximas, clustering con DBSCAN para agrupar noticias del mismo caso, y modelado de tópicos con LDA para identificar temas recurrentes.
- **OE-4. Implementar sistema de detección de duplicados:** Desarrollar mecanismo de deduplicación mediante cálculo de similitud coseno entre vectores TF-IDF de noticias, estableciendo threshold ≥ 0.75 para identificar republicaciones del mismo caso y evitar inflación artificial de conteos.
- **OE-5. Construir API REST para consulta de información:** Desarrollar interfaz de programación de aplicaciones con al menos 6 endpoints que permitan consultar noticias procesadas, buscar por keywords, obtener métricas agregadas, explorar clusters identificados, consultar noticias recientes y verificar salud del sistema.
- **OE-6. Diseñar dashboard web de visualización:** Crear interfaz de usuario web que presente métricas clave (total de noticias, casos con NNA, distribución temporal), visualice distribución de clusters mediante técnicas de reducción de dimensionalidad, y permita exploración interactiva de noticias individuales y sus metadatos.
- **OE-7. Validar viabilidad técnica mediante prototipos incrementales:** Desarrollar tres prototipos evolutivos (P0: scraping directo, P1: sistema básico RSS+K-Means, P2: sistema avanzado multi-fuente+DBSCAN) que validen estrategias de recolección, evalúen técnicas de PLN, e identifiquen limitaciones para abordar en TT2.
- **OE-8. Establecer arquitectura escalable con Docker:** Implementar arquitectura de contenedores que separe componentes (análisis, servidor web, almacenamiento) para

facilitar despliegue, mantenimiento y escalamiento futuro del sistema en ambientes de producción.

2.3. Usuarios identificados

El sistema está diseñado para atender las necesidades de cuatro perfiles principales de usuarios, organizados según su relación con la problemática y los objetivos específicos de uso del sistema. La Figura 2.1 presenta la estructura de usuarios identificados y sus interrelaciones.

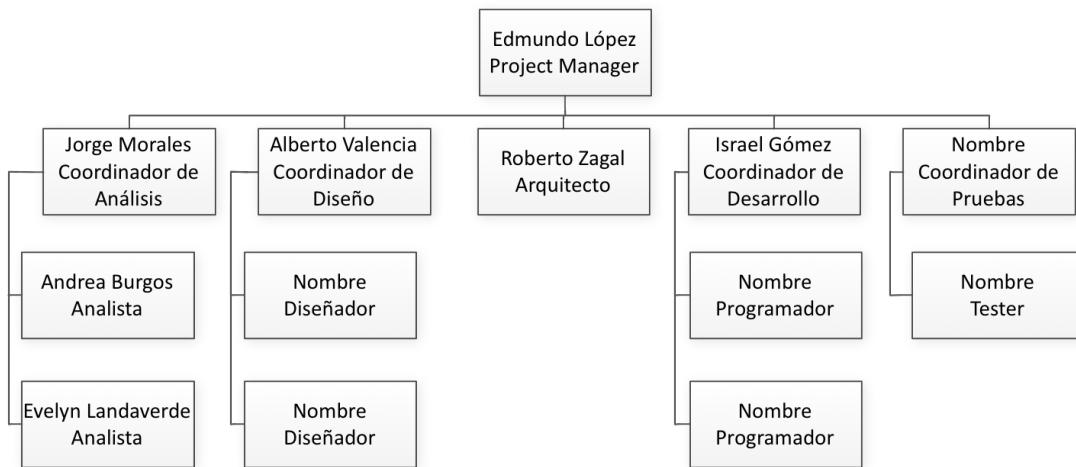


Figura 2.1: Organigrama de usuarios del sistema de identificación y seguimiento de NNA afectados por feminicidios.

Nota: Crear un diagrama con los siguientes usuarios organizados jerárquicamente:

- **Nivel 1 - Usuarios Primarios:** Organizaciones de la Sociedad Civil (Fundación Futuro con Derechos, REDIM, Data Cívica, Mapa de Feminicidios)
- **Nivel 2 - Usuarios Secundarios:** Investigadores Académicos (Universidades, Centros de Investigación), Autoridades Gubernamentales (Fiscalías, SIPINNA, Comisiones de Víctimas)
- **Nivel 3 - Usuarios de Apoyo:** Periodistas y Medios de Comunicación, Activistas y Defensores de Derechos Humanos
- **Nivel 4 - Administradores del Sistema:** Equipo de Desarrollo (Mantenimiento, Actualización de Fuentes)

Descripción de perfiles de usuario

OSC-01: Coordinador de Documentación Responsable en organizaciones civiles de mantener bases de datos de casos de feminicidio. Usa el sistema para consultar casos detectados automáticamente, validar información, descargar reportes y alimentar bases de datos internas.

OSC-02: Analista de Datos Personal con formación técnica que realiza análisis cuantitativos para reportes institucionales. Consume API REST para extraer datos estructurados, genera visualizaciones personalizadas y calcula estadísticas específicas.

INV-01: Investigador Académico Científico social que estudia violencia de género y derechos de infancia. Utiliza el sistema como fuente de datos para investigaciones, verifica hipótesis mediante consultas específicas y cita el sistema en publicaciones académicas.

GOB-01: Funcionario de Política Pública Personal de gobierno responsable de diseñar o evaluar programas de atención a NNA huérfanos. Consulta métricas agregadas por entidad federativa, identifica zonas de alta incidencia y solicita reportes personalizados.

GOB-02: Fiscal Especializado Autoridad ministerial que investiga feminicidios. Usa el sistema para cruzar casos mediáticos con carpetas de investigación y verificar cobertura periodística de casos bajo su jurisdicción.

PER-01: Periodista de Investigación Reportero especializado en temas de género y derechos humanos. Consulta el sistema para investigaciones de largo aliento, identifica patrones noticiosos y verifica datos para reportajes.

ADM-01: Administrador del Sistema Desarrollador responsable de mantenimiento técnico. Monitorea salud del sistema, actualiza lista de fuentes RSS, ajusta parámetros de detección y gestiona actualizaciones de software.

2.4. Procesos involucrados

El sistema impacta procesos de trabajo existentes en organizaciones civiles y habilita nuevos procesos de análisis de datos. La Figura 2.2 presenta el mapa de procesos de una organización civil típica enfocada en derechos de NNA, identificando aquellos que serán transformados o habilitados por el sistema.

Nota: Crear un mapa de procesos con tres niveles:

- **Procesos Estratégicos:** Planeación Institucional, Incidencia en Política Pública, Vinculación Interinstitucional
- **Procesos Operativos (núcleo):** Monitoreo de Medios, Documentación de Casos, Análisis de Información, Seguimiento de NNA, Atención Directa
- **Procesos de Soporte:** Gestión de TI, Capacitación de Personal, Administración de Datos

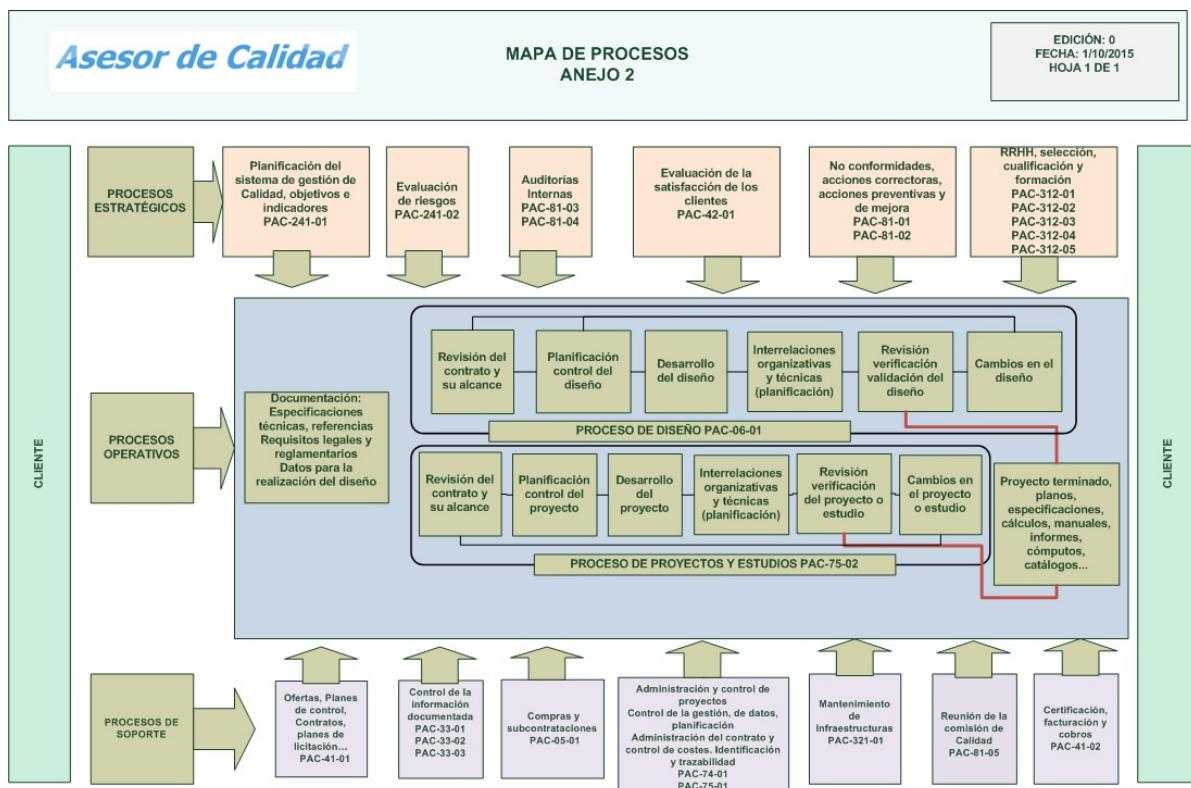


Figura 2.2: Mapa de procesos de organización civil con enfoque en documentación de casos de feminicidio y atención a NNA afectados.

Descripción de procesos afectados

PR-01 Monitoreo de medios de comunicación. Proceso de revisión sistemática de fuentes periodísticas para identificar casos de feminicidio. **AS-IS:** Revisión manual diaria de 15-20 sitios web por personal voluntario (10-15 hrs/semana). **TO-BE:** Consulta automatizada cada 6 horas de 10 fuentes RSS + Google News API, reduciendo trabajo manual a validación de casos detectados (2 hrs/semana).

PR-02 Documentación y registro de casos. Proceso de captura estructurada de información sobre casos identificados. **AS-IS:** Llenado manual de formularios o hojas de cálculo con datos extraídos de noticias. **TO-BE:** Importación automática de casos detectados con campos pre-lLENADOS (fecha, medio, URL, extracto relevante), requiriendo solo validación y complemento de información.

PR-03 Análisis cuantitativo de datos. Proceso de generación de estadísticas y reportes periódicos. **AS-IS:** Conteos manuales en hojas de cálculo, gráficas básicas en Excel. **TO-BE:** Consulta de métricas agregadas vía API REST, visualizaciones automáticas en dashboard, exportación de datasets para análisis avanzado en R/Python.

PR-04 Identificación de patrones geográficos y temporales. Proceso de análisis de tendencias para orientar estrategias de intervención. **AS-IS:** Análisis ad-hoc limitado por falta de datos estructurados. **TO-BE:** Clustering automático de casos similares, modelado de tópicos para identificar narrativas recurrentes, filtrado por entidad federativa y rango temporal.

PR-05 Elaboración de reportes institucionales. Proceso de generación de documentos para difusión pública, financiadores y autoridades. **AS-IS:** Redacción manual con datos fragmentados de múltiples fuentes. **TO-BE:** Extracción directa de datos estructurados del sistema, garantizando actualización y precisión de cifras citadas.

PR-06 Respuesta a solicitudes de información. Proceso de atención a consultas de medios, investigadores y autoridades. **AS-IS:** Búsquedas manuales en archivos internos con tiempo de respuesta de días. **TO-BE:** Consultas mediante API o dashboard con respuesta inmediata, filtrado por criterios específicos del solicitante.

PR-07 Detección de duplicados y validación de casos. Proceso de verificación de unicidad de casos para evitar conteo múltiple. **AS-IS:** Comparación manual memoria-dependiente propensa a errores. **TO-BE:** Detección automática de similitud mediante vectorización TF-IDF y clustering DBSCAN.

2.5. Requerimientos de usuario

Los requerimientos del usuario se organizan en cinco categorías: recolección automatizada, procesamiento y análisis, consulta de información, administración del sistema, y reportes y

exportación. Cada requerimiento se identifica con clave única, nombre descriptivo, descripción detallada de funcionalidad esperada, y prioridad según criticidad para objetivos del proyecto²:

Requerimientos del Usuario					
Id	Nombre	Descripción	Iter.	Stat.	
RU-01	Recolección automática de noticias	El sistema debe consultar automáticamente feeds RSS de al menos 10 medios de comunicación mexicanos cada 6 horas, extrayendo título, fecha de publicación, URL, contenido y medio de origen de cada noticia nueva detectada.	1	DONE	
RU-02	Integración con Google News API	El sistema debe consultar Google News API con queries específicas de feminicidio para ampliar cobertura más allá de fuentes RSS configuradas, procesando al menos 50 resultados por consulta.	1	DONE	
RU-03	Scraping histórico del Mapa de Feminicidios	El sistema debe extraer casos documentados del sitio web del Mapa de Feminicidios en México para complementar cobertura temporal de los últimos 6 meses, respetando límites de tasa de peticiones.	2	DONE	
RU-04	Detección automática de feminicidios	El sistema debe identificar automáticamente noticias relacionadas con feminicidio mediante búsqueda de keywords especializadas (“feminicidio”, “asesinó a su pareja”, “violencia de género”), alcanzando precisión $\geq 85\%$.	1	DONE	
RU-05	Detección de menciones de NNA	El sistema debe identificar automáticamente menciones de niñas, niños y adolescentes afectados mediante análisis de patrones lingüísticos (“dejó N hijos”, “madre de”, “huérfanos”), con tasa de falsos positivos $\leq 10\%$.	1	DONE	
RU-06	Limpieza y normalización de texto	El sistema debe preprocessar el contenido de noticias removiendo stopwords, normalizando caracteres especiales, convirtiendo a minúsculas y aplicando lematización para facilitar análisis posterior.	1	DONE	
RU-07	Vectorización TF-IDF	El sistema debe convertir texto de noticias a representación vectorial mediante TF-IDF con máximo 3,000 features, habilitando análisis cuantitativo de similitud y clustering.	1	DONE	

²El valor de Status en la tabla está indicada como: **TODO** - Pendiente, **DOING** - En proceso, **DONE** - Terminado, **TOCHK** - Listo para revisión, **ISSUE** - Presenta problemas.

Requerimientos del Usuario				
Id	Nombre	Descripción	Iter.	Stat.
RU-08	Clustering de noticias similares	El sistema debe agrupar automáticamente noticias del mismo caso publicadas por múltiples medios mediante DBSCAN con parámetros adaptativos ($\text{eps}=0.3$, $\text{min_samples}=2$), facilitando identificación de casos únicos.	1	DONE
RU-09	Modelado de tópicos principales	El sistema debe identificar temas recurrentes en corpus de noticias mediante LDA con 5 tópicos, extrayendo las 10 palabras más representativas de cada tema para interpretación.	2	DONE
RU-10	Detección de duplicados	El sistema debe calcular similitud coseno entre pares de noticias e identificar como duplicadas aquellas con similitud ≥ 0.75 , evitando conteo múltiple del mismo caso.	1	DONE
RU-11	API REST - Consultar todas las noticias	El sistema debe exponer endpoint GET /api/noticias que retorne lista completa de noticias procesadas con sus metadatos (id, título, fecha, medio, URL, cluster, menciona_nna) en formato JSON.	1	DONE
RU-12	API REST - Buscar por keywords	El sistema debe exponer endpoint GET /api/search?q=<query> que retorne noticias cuyo contenido contenga las palabras buscadas, permitiendo búsquedas personalizadas de usuarios.	1	DONE
RU-13	API REST - Obtener métricas agregadas	El sistema debe exponer endpoint GET /api/metrics que retorne estadísticas clave: total de noticias, porcentaje con mención de NNA, número de clusters, distribución temporal y coherencia de tópicos LDA.	1	DONE
RU-14	API REST - Explorar clusters	El sistema debe exponer endpoint GET /api/clusters que retorne lista de clusters identificados con sus noticias miembro, facilitando exploración de casos únicos agrupados.	1	DONE
RU-15	API REST - Noticias recientes	El sistema debe exponer endpoint GET /api/recent?days=N que retorne noticias de los últimos N días, permitiendo consultas de información actualizada.	2	DONE
RU-16	API REST - Salud del sistema	El sistema debe exponer endpoint GET /api/health que retorne estado del sistema (operativo/degradado/fallido) y timestamp de última actualización de datos.	2	DONE

Requerimientos del Usuario				
Id	Nombre	Descripción	Iter	Stat.
RU-17	Dashboard web - Visualización de métricas	El sistema debe presentar interfaz web con tarjetas que muestren métricas clave actualizadas: total de noticias procesadas, casos con NNA, número de clusters y periodo de cobertura.	1	DONE
RU-18	Dashboard web - Gráfica de distribución temporal	El sistema debe presentar gráfica de línea o barras que muestre distribución de noticias por mes/semana para identificar tendencias temporales.	2	DOING
RU-19	Dashboard web - Visualización de clusters	El sistema debe presentar visualización 2D de clusters mediante reducción de dimensionalidad (t-SNE o PCA) con colores diferenciados por cluster, permitiendo inspección visual de agrupación.	2	DOING
RU-20	Dashboard web - Tabla explorable de noticias	El sistema debe presentar tabla paginada con noticias procesadas, permitiendo ordenamiento por columnas, filtrado básico y acceso a URL original de cada noticia.	1	DONE
RU-21	Almacenamiento persistente	El sistema debe almacenar noticias procesadas en formato CSV con codificación UTF-8, garantizando persistencia de datos entre reinicios del sistema.	1	DONE
RU-22	Arquitectura Docker	El sistema debe estar contenerizado en Docker con separación de servicios: contenedor de análisis/recolección y contenedor de servidor web, facilitando despliegue y escalamiento.	1	DONE
RU-23	Configuración de fuentes externa	El sistema debe leer lista de fuentes RSS desde archivo de configuración externo, permitiendo agregar o remover fuentes sin modificar código fuente.	2	TODO
RU-24	Logs de operación	El sistema debe generar logs con nivel de detalle configurable (INFO/DEBUG) que registren operaciones críticas: noticias recolectadas, casos detectados, errores de recolección y métricas de procesamiento.	2	DOING
RU-25	Exportación de datasets	El sistema debe permitir exportar subconjuntos de datos filtrados por criterios (fecha, medio, presencia de NNA) en formato CSV para análisis externo en R/Python/Excel.	2	TODO
RU-26	Reportes automatizados	El sistema debe generar reportes periódicos (semanales/mensuales) en formato PDF o HTML con métricas clave, noticias destacadas y análisis de tendencias.	3	TODO

Requerimientos del Usuario					
Id	Nombre	Descripción	Iter.	Stat.	
RU-27	Alertas de casos relevantes	El sistema debe enviar notificaciones (email-/webhook) cuando detecte casos que cumplan criterios específicos (múltiples NNA, ubicación prioritaria), habilitando respuesta rápida de organizaciones.	3	TODO	

2.6. Especificación de plataforma

El sistema se implementa como aplicación web con arquitectura de contenedores Docker, adoptando enfoque de microservicios que separa componentes de recolección/análisis y presentación. La Figura 2.3 presenta la estructura general del sistema, detallando la interacción entre componentes, flujo de datos y tecnologías empleadas.

Nota: Crear diagrama de arquitectura con los siguientes componentes:

- **Capa de Recolección:** Recolector RSS, Google News API Client, Scraper Histórico → Almacenamiento temporal
- **Capa de Procesamiento:** Detector Dual-Etapa, Pipeline PLN (Limpieza, TF-IDF, DBSCAN, LDA) → Base de datos CSV
- **Capa de Presentación:** API REST Flask (6 endpoints) ← Dashboard Web (HTML/JS/-Chart.js)
- **Infraestructura:** Docker Container 1 (Análisis), Docker Container 2 (Web), Volumen compartido (datos)

En la Figura 2.3 se describe la estructura del sistema contenerizado. El **Contenedor de Análisis** ejecuta el recolector automatizado mediante cron cada 6 horas, procesa noticias con pipeline PLN y almacena resultados en CSV. El **Contenedor Web** expone API REST Flask y sirve dashboard HTML para consulta. Ambos contenedores acceden a volumen Docker compartido que garantiza persistencia de datos.

Especificaciones técnicas detalladas

Tipo de sistema: Aplicación web con arquitectura híbrida: backend de procesamiento batch (tareas programadas) + servidor web para consultas en tiempo real. Dashboard web responsive accesible desde navegadores modernos sin requerir instalación cliente.

Lenguajes de programación: Python 3.9+ para todo el backend (recolección, análisis, API REST). HTML5, CSS3 y JavaScript ES6 para frontend del dashboard. Uso de f-strings y type hints en Python.

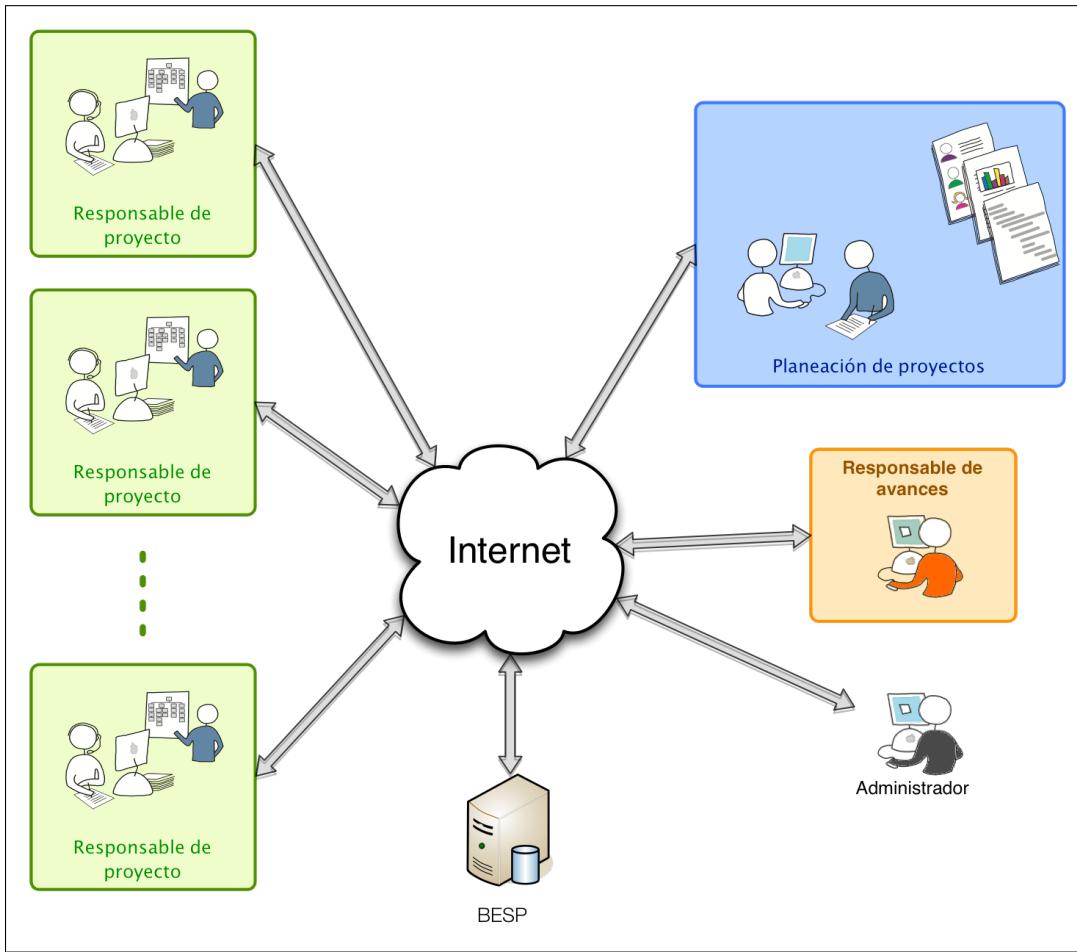


Figura 2.3: Arquitectura del sistema de identificación y seguimiento de NNA afectados por feminicidios.

- Frameworks y librerías principales:**
- **Web:** Flask 2.0+ para API REST y servidor web, Flask-CORS para habilitación de peticiones cross-origin
 - **Recolección:** feedparser para parsing RSS/Atom, requests para HTTP, beautifulsoup4 para scraping HTML selectivo
 - **PLN:** scikit-learn para TF-IDF/DBSCAN/LDA, nltk o spacy para procesamiento de texto (lematización, stopwords)
 - pandas para manipulación de datos, numpy para operaciones vectoriales

Sistema operativo: Linux Ubuntu 20.04 LTS o superior en ambientes de desarrollo y producción. Docker permite abstracción del SO host, garantizando portabilidad a Windows/macOS para desarrollo local.

- Contenedores Docker:**
- **Imagen base:** python:3.9-slim (200MB) para reducir tamaño de contenedores
 - **Contenedor 1 - Análisis:** Python + librerías PLN + crontab, ejecuta data_collector.py cada 6 horas
 - **Contenedor 2 - Web:** Python + Flask + archivos estáticos HTML/CSS/JS, expone puerto 5000
 - **Volumen compartido:** /app/data montado en ambos contenedores para acceso a CSV de noticias

- Almacenamiento de datos:**
- **TT1:** Archivos CSV con codificación UTF-8, tamaño estimado 5MB por 1,000 noticias
 - **TT2 (planeado):** Migración a PostgreSQL 13+ o MySQL 8+ con esquema normalizado, índices en fecha/medio, soporte de búsqueda full-text

- Requisitos de hardware - Desarrollo/Prototipo:**
- **CPU:** 2 núcleos mínimo, 4 núcleos recomendado (procesamiento PLN intensivo)
 - **RAM:** 4GB mínimo, 8GB recomendado (vectorización TF-IDF de corpus grande)
 - **Disco:** 10GB disponibles (sistema operativo + Docker images + datos + logs)
 - **Red:** Conexión Internet estable para recolección de fuentes externas

- Requisitos de hardware - Producción (escalado):**
- **CPU:** 4-8 núcleos para parallelización de procesamiento PLN
 - **RAM:** 16-32GB para mantener vectores TF-IDF en memoria y cache de resultados
 - **Disco:** SSD con 50GB+ para base de datos creciente y respaldos
 - **Red:** Ancho de banda de 100Mbps+ para manejo de tráfico concurrente a dashboard

- Seguridad y disponibilidad:**
- **HTTPS:** Certificado SSL/TLS mediante Let's Encrypt para cifrado de comunicaciones

- **Firewall:** Exponer solo puerto 443 (HTTPS) y 22 (SSH administración), bloquear acceso directo a puerto 5000
- **Respaldos:** Snapshot diario de volumen Docker con datos, retención de 30 días
- **Monitoreo:** Logs centralizados con rotación semanal, alertas de caída de servicio mediante healthcheck
- **Autenticación:** TT1 sin autenticación (prototipo), TT2 implementará OAuth2 o JWT para control de acceso

Servicios externos dependientes:

- Feeds RSS de medios mexicanos (gratuitos, públicos)
- Google News API o Google Custom Search API (cuota gratuita: 100 consultas/día)
- Mapa de Feminicidios en México (scraping respetuoso con rate limiting)

Despliegue y escalamiento:

- **TT1:** Servidor Linux único (VM o bare metal) con Docker Compose
- **TT2:** Migración opcional a Kubernetes para orquestación, horizontal pod autoscaling, y distribución de carga
- **CI/CD:** Pipeline con GitHub Actions para testing automatizado y despliegue continuo

La arquitectura propuesta prioriza simplicidad operativa en TT1 (demostración de viabilidad técnica) mientras establece bases para escalamiento en TT2 (sistema de producción). La contenerización facilita replicación del ambiente de desarrollo, reduce fricción en transferencia de conocimiento, y habilita despliegue en infraestructura diversa (servidores institucionales, cloud providers, hardware de organizaciones civiles).

CAPÍTULO 3

Modelo del Negocio

Este capítulo describe el modelo de negocio del sistema de identificación y seguimiento de NNA afectados por feminicidios, estableciendo los fundamentos conceptuales para el análisis y diseño detallado. Se identifican y caracterizan los actores que interactuarán con el sistema, definiendo sus responsabilidades, perfil profesional y contexto de uso. Se establece un glosario de términos del negocio específicos del dominio (feminicidio, procesamiento de lenguaje natural, organizaciones civiles, clustering), facilitando comunicación precisa entre stakeholders. Se presenta el modelo del dominio del problema mediante diagrama conceptual de entidades y sus relaciones (Noticia, Caso, Medio, Cluster, Mención NNA), describiendo atributos y cardinalidades. Finalmente se documentan reglas de negocio que gobiernan operación del sistema y máquinas de estado que modelan ciclos de vida de entidades clave. Esta documentación sirve como referencia compartida entre usuarios finales, desarrolladores y evaluadores del proyecto.

3.1. Actores del sistema

El sistema identifica siete perfiles de actores organizados en tres categorías: usuarios operativos de organizaciones civiles (quienes usan el sistema diariamente para documentación de casos), usuarios estratégicos (investigadores, autoridades gubernamentales, periodistas que consultan información para análisis y reportes), y administradores técnicos (responsables de mantenimiento y configuración del sistema). A continuación se describe cada actor con detalle.

3.1.1. Coordinador de Documentación



Responsable en organizaciones de la sociedad civil de mantener bases de datos actualizadas

de casos de feminicidio y seguimiento de NNA afectados. Usa el sistema como herramienta principal para identificar nuevos casos reportados en medios de comunicación.

Responsabilidades:

- Revisar diariamente dashboard web del sistema para identificar nuevos casos detectados automáticamente.
- Validar precisión de detección automática de menciones de NNA, marcando falsos positivos y falsos negativos.
- Complementar información de casos detectados con datos adicionales obtenidos de otras fuentes (comunicados oficiales, contacto directo con familias).
- Exportar datasets filtrados para integración con bases de datos internas de la organización.
- Generar reportes semanales sobre casos nuevos identificados para circulación interna en la organización.
- Coordinarse con áreas de atención directa para canalizar casos que requieren intervención urgente.
- Documentar casos que el sistema no detectó automáticamente para retroalimentación al equipo técnico.

Perfil:

- Escolaridad: Licenciatura en Ciencias Sociales, Trabajo Social, Derecho, Psicología o áreas afines.
- Experiencia: Mínimo 2 años trabajando en organizaciones civiles enfocadas en derechos humanos, violencia de género o derechos de infancia.
- Conocimientos técnicos: Manejo intermedio de hojas de cálculo (Excel, Google Sheets), navegadores web, herramientas de búsqueda en Internet.
- Habilidades blandas: Sensibilidad ante temas de violencia de género, atención al detalle, organización, capacidad de trabajo con información sensible manteniendo confidencialidad.
- Deseable: Experiencia en metodologías de documentación de casos de derechos humanos, conocimiento de legislación mexicana sobre feminicidio y derechos de NNA.

Procesos en los que participa:

- PR-01 Monitoreo de medios de comunicación
- PR-02 Documentación y registro de casos
- PR-06 Respuesta a solicitudes de información
- PR-07 Detección de duplicados y validación de casos

Área: Área de Documentación e Investigación de organización civil (Fundación Futuro con Derechos, REDIM, Data Cívica, colectivos estatales)

Cantidad aproximada: 1-3 personas por organización (aproximadamente 15-20 usuarios potenciales a nivel nacional considerando organizaciones principales)

Horario actividad: Lunes a viernes 9:00-18:00 hrs, con consultas ocasionales fuera de horario cuando surgen casos urgentes o de alto perfil mediático

3.1.2. Analista de Datos



Personal con formación técnica en organizaciones civiles responsable de generar análisis cuantitativos, estadísticas y visualizaciones para reportes institucionales, solicitudes de transparencia y presentaciones públicas.

Responsabilidades:

- Consumir API REST del sistema para extraer datos estructurados y alimentar análisis personalizados.
- Generar visualizaciones avanzadas (mapas geográficos, series de tiempo, correlaciones) usando herramientas externas (R, Python, Tableau).
- Calcular métricas específicas no contempladas en dashboard estándar (tasas de incidencia por población, variaciones interanuales, patrones estacionales).
- Validar calidad de datos identificando inconsistencias, valores atípicos o problemas de cobertura temporal/geográfica.
- Desarrollar reportes automatizados que integren datos del sistema con otras fuentes (INEGI, Secretaría de Gobernación, fiscalías estatales).
- Documentar metodología de análisis para garantizar reproducibilidad y transparencia.
- Capacitar a personal no técnico en interpretación correcta de estadísticas generadas.

Perfil:

- Escolaridad: Licenciatura en Ciencia de Datos, Estadística, Actuaría, Matemáticas Aplicadas, Economía o áreas cuantitativas. Deseable posgrado.
- Experiencia: Mínimo 1 año en análisis de datos, preferentemente en contextos de investigación social, derechos humanos o salud pública.
- Conocimientos técnicos: Programación en Python o R, análisis estadístico, visualización de datos (matplotlib, ggplot2, Tableau), consumo de APIs REST, manejo de formatos JSON/CSV, control de versiones (Git).

- Habilidades blandas: Pensamiento analítico, comunicación de hallazgos técnicos a audiencias no especializadas, rigor metodológico.
- Deseable: Experiencia con datos de texto no estructurado, conocimientos básicos de procesamiento de lenguaje natural, familiaridad con problemáticas de género.

Procesos en los que participa:

- PR-03 Análisis cuantitativo de datos
- PR-04 Identificación de patrones geográficos y temporales
- PR-05 Elaboración de reportes institucionales
- PR-06 Respuesta a solicitudes de información

Área: Área de Investigación y Análisis de organización civil, o bien consultor externo contratado para proyectos específicos

Cantidad aproximada: 1-2 personas por organización grande (aproximadamente 8-12 usuarios potenciales a nivel nacional)

Horario actividad: Lunes a viernes 10:00-19:00 hrs, flexible según proyectos. Intensificación en periodos de elaboración de reportes semestrales/anuales

3.1.3. Investigador Académico



Científico social de universidades o centros de investigación que estudia violencia de género, derechos de infancia, políticas públicas o fenómenos relacionados. Utiliza datos del sistema como insumo para investigaciones publicables en revistas académicas o libros.

Responsabilidades:

- Formular hipótesis de investigación verificables mediante datos del sistema (correlaciones geográficas, factores de riesgo, patrones temporales).
- Descargar datasets completos o filtrados para análisis estadístico riguroso con controles metodológicos.
- Triangular información del sistema con otras fuentes de datos para validación cruzada.
- Documentar limitaciones metodológicas del sistema en publicaciones (sesgos de cobertura mediática, precisión de detección automática).
- Citar apropiadamente el sistema en publicaciones académicas, reconociendo autoría y metodología.
- Compartir hallazgos con organizaciones civiles y desarrolladores del sistema para retroalimentación mutua.

- Solicitar acceso a datos históricos o funcionalidades específicas no disponibles en interfaz estándar.

Perfil:

- Escolaridad: Posgrado (Maestría o Doctorado) en Sociología, Ciencias Políticas, Estudios de Género, Derechos Humanos, Demografía, Salud Pública o áreas relacionadas.
- Experiencia: Investigación académica con publicaciones en revistas arbitradas, experiencia en metodologías cuantitativas o mixtas.
- Conocimientos técnicos: Análisis estadístico avanzado (regresiones, modelos multínivel), software especializado (SPSS, Stata, R), manejo de datos secundarios.
- Habilidades blandas: Rigor metodológico, escritura académica, capacidad de síntesis, ética en investigación con poblaciones vulnerables.
- Deseable: Publicaciones previas sobre feminicidio, violencia de género o derechos de infancia en México, experiencia con datos de medios de comunicación.

Procesos en los que participa:

- PR-03 Análisis cuantitativo de datos (uso externo al sistema)
- PR-04 Identificación de patrones geográficos y temporales (uso externo)
- PR-06 Respuesta a solicitudes de información

Área: Universidades públicas y privadas (UNAM, COLMEX, UAM, CIESAS), centros de investigación independientes, organizaciones internacionales con componente académico

Cantidad aproximada: 20-40 investigadores potencialmente interesados a nivel nacional, 5-10 usuarios activos frecuentes

Horario actividad: Variable según calendario académico, mayor actividad en períodos de desarrollo de proyectos de investigación (septiembre-diciembre, febrero-junio)

3.1.4. Funcionario de Política Pública



Personal de gobierno a nivel federal o estatal responsable de diseñar, implementar o evaluar programas de atención a NNA en situación de orfandad por feminicidio, reparación del daño, o prevención de violencia de género.

Responsabilidades:

- Consultar métricas agregadas por entidad federativa para identificar zonas que requieren atención prioritaria.
- Solicitar reportes personalizados con características específicas (edad de NNA, tiempo transcurrido desde el caso, cobertura mediática).

- Utilizar datos del sistema como evidencia para justificar asignación presupuestal a programas específicos.
- Comparar cobertura de programas gubernamentales vs. casos documentados en el sistema para identificar brechas de atención.
- Detectar casos de alto perfil que requieren seguimiento directo de autoridades superiores.
- Coordinarse con fiscalías y comisiones de víctimas para cruce de información institucional con datos mediáticos.
- Responder solicitudes de transparencia de organizaciones civiles mediante datos verificables del sistema.

Perfil:

- Escolaridad: Licenciatura mínima en Administración Pública, Ciencias Políticas, Derecho, Trabajo Social o áreas relacionadas. Deseable posgrado en Políticas Públicas.
- Experiencia: Mínimo 3 años en administración pública, preferentemente en áreas de atención a víctimas, protección de derechos de infancia, o igualdad de género.
- Conocimientos técnicos: Manejo de sistemas gubernamentales, interpretación de indicadores de gestión, navegación web, herramientas office básicas.
- Habilidades blandas: Comprensión de problemáticas sociales complejas, capacidad de coordinación interinstitucional, sensibilidad ante víctimas.
- Marco legal: Conocimiento de Ley General de Víctimas, Ley General de Acceso de las Mujeres a una Vida Libre de Violencia, Ley General de Derechos de NNA.

Procesos en los que participa:

- PR-04 Identificación de patrones geográficos y temporales (consumo externo)
- PR-05 Elaboración de reportes institucionales (consumo externo)
- PR-06 Respuesta a solicitudes de información

Área: SIPINNA (Sistema de Protección Integral de NNA), Comisiones Estatales de Víctimas, Institutos de la Mujer estatales, Fiscalías Especializadas en Feminicidio, Secretarías de Gobernación estatales

Cantidad aproximada: 10-20 funcionarios a nivel federal, 3-5 por entidad federativa con alta incidencia (aproximadamente 50-80 usuarios potenciales)

Horario actividad: Lunes a viernes 9:00-18:00 hrs (horario administrativo), con consultas ocasionales en preparación de informes trimestrales/anuales

3.1.5. Periodista de Investigación



Reportero especializado en temas de género, derechos humanos, seguridad o investigación que utiliza el sistema como herramienta de documentación y verificación para desarrollo de piezas periodísticas de largo aliento (reportajes, series, documentales).

Responsabilidades:

- Consultar histórico de casos para identificar patrones noticiosos o ausencias en cobertura mediática que ameriten investigación periodística.
- Verificar datos sobre casos específicos para contrastar con testimonios de fuentes primarias.
- Identificar casos con características específicas (múltiples NNA huérfanos, ausencia de seguimiento institucional, ubicaciones específicas) como potenciales sujetos de reportaje.
- Citar correctamente datos del sistema en piezas publicadas, reconociendo fuente y metodología.
- Reportar al equipo técnico casos conocidos que el sistema no detectó, contribuyendo a mejorar cobertura.
- Utilizar métricas del sistema para contextualizar casos individuales dentro de tendencias generales (“este caso es uno de X documentados en Y entidad durante Z periodo”).

Perfil:

- Escolaridad: Licenciatura en Comunicación, Periodismo, Ciencias Sociales o áreas afines.
- Experiencia: Mínimo 3 años de experiencia periodística, con al menos 1 año cubriendo fuentes de seguridad, procuración de justicia o derechos humanos.
- Conocimientos técnicos: Investigación periodística, verificación de fuentes, manejo de bases de datos, herramientas digitales de investigación (búsquedas avanzadas, scraping básico).
- Habilidades blandas: Sensibilidad ante víctimas y sobrevivientes, ética periodística, capacidad de trabajar con información traumática sin desensibilización.
- Deseable: Especialización en periodismo de investigación, certificaciones en cobertura ética de violencia de género, experiencia en periodismo de datos.

Procesos en los que participa:

- PR-06 Respuesta a solicitudes de información (consumo externo)
- PR-04 Identificación de patrones geográficos y temporales (consumo externo para contexto periodístico)

Área: Medios de comunicación nacionales e internacionales (prensa escrita, digital, radio, TV), medios especializados en investigación (Mexicanos Contra la Corrupción y la Impunidad, Quinto Elemento Lab, Pie de Página), periodistas freelance

Cantidad aproximada: 15-30 periodistas potencialmente interesados, 5-10 usuarios ocasionales activos

Horario actividad: Irregular según ciclos de producción periodística, mayor actividad en fechas simbólicas (25 de noviembre Día Internacional de la Eliminación de la Violencia contra la Mujer, 8 de marzo Día Internacional de la Mujer)

3.1.6. Administrador del Sistema



Desarrollador o ingeniero en sistemas responsable del mantenimiento técnico, configuración, monitoreo de salud y actualización del sistema. Puede ser parte del equipo original de desarrollo o personal de TI de organización adoptante del sistema.

Responsabilidades:

- Monitorear logs del sistema diariamente para detectar errores de recolección, caídas de fuentes RSS o problemas de procesamiento.
- Actualizar lista de fuentes RSS cuando surjan nuevos medios relevantes o fuentes existentes cambien su estructura.
- Ajustar parámetros de detección (keywords, thresholds de similitud, configuración de clustering) basándose en retroalimentación de usuarios finales.
- Gestionar actualizaciones de dependencias de software (librerías Python, imágenes Docker base) manteniendo compatibilidad.
- Realizar respaldos periódicos de datos y verificar integridad de backups.
- Implementar mejoras incrementales al sistema según solicitudes de usuarios y limitaciones detectadas.
- Documentar cambios realizados y mantener actualizada documentación técnica del sistema.
- Capacitar a usuarios administradores de organizaciones civiles en operación básica del sistema.

Perfil:

- Escolaridad: Licenciatura o Ingeniería en Ciencias Computacionales, Sistemas Computacionales, Software o áreas afines.
- Experiencia: Mínimo 2 años desarrollando aplicaciones web o sistemas de procesamiento de datos, experiencia con Python, APIs REST, Docker.

- Conocimientos técnicos obligatorios: Python 3.9+, Flask, Docker, Git, Linux/Unix, bash scripting, manejo de APIs REST, depuración de logs.
- Conocimientos técnicos deseables: Procesamiento de lenguaje natural con scikit-learn, web scraping responsable, bases de datos relacionales (PostgreSQL/MySQL), CI/CD, Kubernetes.
- Habilidades blandas: Resolución de problemas técnicos complejos, documentación clara, comunicación con usuarios no técnicos, trabajo colaborativo con organizaciones civiles.

Procesos en los que participa:

- Monitoreo de salud del sistema (proceso interno)
- Mantenimiento preventivo y correctivo (proceso interno)
- Gestión de configuración y actualizaciones (proceso interno)
- Respaldo y recuperación de datos (proceso interno)

Área: Equipo de desarrollo original del proyecto (IPN-ESCOM), o bien área de TI de organización civil que adopte el sistema en producción

Cantidad aproximada: 1-2 administradores principales (equipo de desarrollo TT), eventualmente 1 por organización grande que implemente el sistema localmente

Horario actividad: Flexible, con revisión diaria de logs preferentemente en horario matutino (9:00-10:00 hrs). Disponibilidad ocasional fuera de horario para resolver incidentes críticos

3.2. Términos del Negocio

Este glosario define términos especializados del dominio que aparecen consistentemente en la especificación del sistema. Los términos se organizan alfabéticamente e incluyen referencias cruzadas mediante hiperenlaces. Se distinguen tres tipos de términos: conceptos del dominio social (feminicidio, NNA, organización civil), términos técnicos de procesamiento de lenguaje natural (clustering, TF-IDF, vectorización), y elementos específicos del sistema (detector dual-etapa, feed RSS, caso único).

API REST: (*Application Programming Interface - Representational State Transfer*) Interfaz de programación que permite consultar datos del sistema mediante peticiones HTTP estandarizadas. El sistema expone 6 endpoints principales: /api/noticias, /api/search, /api/metrics, /api/clusters, /api/recent, /api/health.

Caso único: Evento individual de feminicidio que puede haber generado múltiples **noticias** en distintos medios. El sistema intenta identificar casos únicos mediante **clustering** de noticias similares, aunque en TT1 esta identificación es imperfecta debido a limitaciones de análisis semántico.

Cluster: (*Grupo, agrupación*) Conjunto de [noticias](#) automáticamente agrupadas por similitud de contenido usando algoritmo [DBSCAN](#). Idealmente cada cluster corresponde a un [caso único](#), pero ruido en agrupación puede generar fragmentación (un caso en múltiples clusters) o fusión (varios casos en un cluster).

Clustering: Técnica de aprendizaje no supervisado que agrupa objetos similares sin etiquetas predefinidas. El sistema usa [DBSCAN](#) para agrupar [noticias](#) basándose en similitud de sus vectores [TF-IDF](#), identificando [clusters](#) de forma adaptativa sin especificar número fijo de grupos.

Coherencia de tópicos: Métrica que evalúa calidad semántica de [tópicos](#) generados por [LDA](#). Valores más altos (cercaos a 1.0) indican que las palabras de un tópico co-ocurren frecuentemente en documentos, sugiriendo coherencia temática. El sistema calcula coherencia tipo C_V para validar configuración de LDA.

DBSCAN: (*Density-Based Spatial Clustering of Applications with Noise*) Algoritmo de [clustering](#) que agrupa puntos densamente conectados y marca puntos aislados como ruido. Ventaja sobre K-Means: no requiere especificar número de clusters anticipadamente, adaptándose al corpus de [noticias](#) disponible.

Deduplicación: Proceso de identificar [noticias duplicadas](#) mediante cálculo de [similitud coseno](#) entre sus vectores [TF-IDF](#). El sistema establece threshold ≥ 0.75 para marcar noticias como duplicadas, conservando solo una representante por grupo para conteos precisos.

Detector dual-etapa: Componente especializado del sistema que identifica [noticias](#) relevantes en dos fases secuenciales: (1) Filtro de feminicidio mediante regex y keywords, (2) Detector de [menciones NNA](#) mediante análisis contextual. Diseño dual reduce falsos positivos vs. detector de etapa única.

Falso positivo: [Noticia](#) incorrectamente clasificada como relevante por el [detector dual-etapa](#). Ejemplo: noticia sobre violencia doméstica (no feminicidio) o mención de “niños” en contexto no relacionado con huérfanos. El sistema alcanza $\leq 10\%$ de tasa de falsos positivos en TT1.

Feed RSS: (*Really Simple Syndication*) Formato XML estandarizado que medios de comunicación publican automáticamente con sus últimas noticias. Ventaja sobre web scraping HTML: estructura predecible, sin bloqueos anti-bot, diseñado para consumo automatizado. Sistema consulta 10 feeds cada 6 horas.

Feminicidio: Asesinato de una mujer por razones de género, tipificado en el Código Penal Federal mexicano (Art. 325). Circunstancias agravantes incluyen violencia sexual, mutilaciones, antecedentes de violencia doméstica, y relación afectiva con el victimario. El sistema busca identificar casos reportados en medios independientemente de si fueron legalmente tipificados como feminicidio.

LDA: (*Latent Dirichlet Allocation*) Algoritmo de modelado probabilístico que descubre **tópicos latentes** en colecciones de documentos. Asume que cada [noticia](#) es mezcla de tópicos y cada tópico es distribución de palabras. Sistema configura LDA con 5 tópicos extrayendo 10 palabras representativas por tópico.

Lematización: Proceso de reducir palabras a su forma base o lema (“corriendo”→“correr”, “niños”→“niño”). Más sofisticado que stemming porque considera contexto morfológico. Sistema aplica lematización en etapa de limpieza de texto para mejorar agrupación de términos relacionados en [TF-IDF](#).

Mención NNA: Referencia explícita o implícita en [noticia](#) a niñas, niños o adolescentes afectados por el feminicidio. Patrones detectados: “dejó N hijos”, “madre de X menores”, “quedaron huérfanos”, “sus niños”. Sistema identifica menciones mediante análisis de patrones lingüísticos en segunda etapa del [detector dual-etapa](#).

Medio de comunicación: Organización periodística que publica noticias sobre eventos de interés público. Sistema recolecta de medios nacionales (La Jornada, Proceso, Aristegui Noticias) y estatales. Metadatos capturados por noticia: nombre del medio, URL, sección, fecha de publicación.

NNA: (*Niñas, Niños y Adolescentes*) Personas menores de 18 años según Ley General de Derechos de Niñas, Niños y Adolescentes. En contexto del proyecto: hijos de víctimas de [feminicidio](#) que quedan en situación de orfandad materna y requieren protección de derechos (educación, salud, atención psicológica, reparación del daño).

Noticia: Unidad básica de información procesada por el sistema. Consiste en: título, fecha de publicación, medio de origen, URL, contenido textual, metadatos agregados por sistema (cluster asignado, menciona_nna, similitud con otras noticias). Fuentes: [feeds RSS](#), Google News API, scraping histórico.

Noticias duplicadas: [Noticias](#) que reportan el mismo evento con variaciones menores (republicaciones de agencias, resúmenes vs. notas completas). Sistema detecta mediante [deduplicación](#) con threshold de [similitud](#) ≥ 0.75 . Ejemplo: misma nota de feminicidio en Chihuahua publicada por El Diario de Chihuahua, El Sol de Chihuahua, y Proceso.

Organización civil: (*OSC, Organización de la Sociedad Civil*) Entidad sin fines de lucro que trabaja en defensa de derechos humanos, derechos de infancia o prevención de violencia de género. Usuarios principales del sistema: Fundación Futuro con Derechos, REDIM, Data Cívica, colectivos estatales de familiares de víctimas, Observatorio Ciudadano Nacional del Feminicidio.

PLN: (*Procesamiento de Lenguaje Natural, NLP por sus siglas en inglés*) Rama de inteligencia artificial que procesa texto en lenguaje humano. Técnicas usadas en sistema: limpieza de texto, tokenización, remoción de stopwords, [lematización](#), [vectorización](#), [clustering](#), [modelado de tópicos](#).

Precisión: Métrica de evaluación de clasificación: Precisión = $\frac{TP}{TP+FP}$ donde TP son verdaderos positivos (noticias relevantes correctamente identificadas) y FP son **falsos positivos**. Sistema alcanza $\geq 90\%$ precisión en detección de noticias con **menciones NNA**.

Similitud coseno: Métrica de similitud entre vectores calculada como: $\text{sim}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$. Valores entre 0 (ortogonales, sin similitud) y 1 (idénticos). Sistema usa similitud coseno entre vectores **TF-IDF** para **deduplicación** y análisis de **clusters**.

Stopwords: (*Palabras vacías*) Palabras de alta frecuencia y bajo contenido semántico (“el”, “la”, “de”, “que”, “y”). Sistema remueve stopwords en español durante limpieza de texto para reducir ruido en **vectorización** y mejorar precisión de **clustering**.

TF-IDF: (*Term Frequency - Inverse Document Frequency*) Técnica de **vectorización** que pondera palabras según su frecuencia en documento (TF) inversamente ponderada por su frecuencia en corpus (IDF). Palabras frecuentes en documento pero raras en corpus reciben mayor peso. Sistema configura TF-IDF con máximo 3,000 features.

Tópico: Tema latente identificado por **LDA** representado como distribución de probabilidad sobre palabras. Ejemplo de tópico: {“feminicidio”: 0.15, “mujer”: 0.12, “asesinato”: 0.10, “violencia”: 0.08...}. Sistema genera 5 tópicos principales del corpus de **noticias** procesadas.

Vectorización: Transformación de texto a representación numérica (vector) que algoritmos de machine learning pueden procesar. Sistema usa **TF-IDF** para convertir contenido de **noticias** a vectores de dimensión 3,000, habilitando cálculo de **similitud** y **clustering**.

Web scraping: Técnica de extracción automatizada de datos de sitios web mediante scripts. Sistema evita scraping HTML directo (bloqueado por Cloudflare) adoptando estrategia de **feeds RSS** y APIs. Excepción: scraping histórico del Mapa de Feminicidios con rate limiting respetuoso (≤ 1 petición por segundo).

3.3. Modelo del dominio del problema

El modelo del dominio del problema representa las entidades conceptuales clave del sistema y sus relaciones estructurales. Se identifican seis entidades principales: **Noticia** (unidad básica de información recolectada), **Medio** (fuente periodística), **Caso** (evento de feminicidio), **Cluster** (agrupación automática de noticias similares), **Mención NNA** (referencia a hijos huérfanos), y **Tópico** (tema latente identificado por LDA). Las relaciones modelan: publicación de noticias por medios, detección de menciones NNA en noticias, agrupación de noticias en clusters, asociación de noticias con tópicos probabilísticos, y correspondencia ideal (no siempre lograda) entre clusters y casos únicos.

El modelo se presenta en la Figura 3.1, seguido por descripción detallada de cada entidad con sus atributos, tipos de datos, obligatoriedad y relaciones con otras entidades del dominio.

Nota: Crear diagrama de entidad-relación con las siguientes entidades y relaciones:

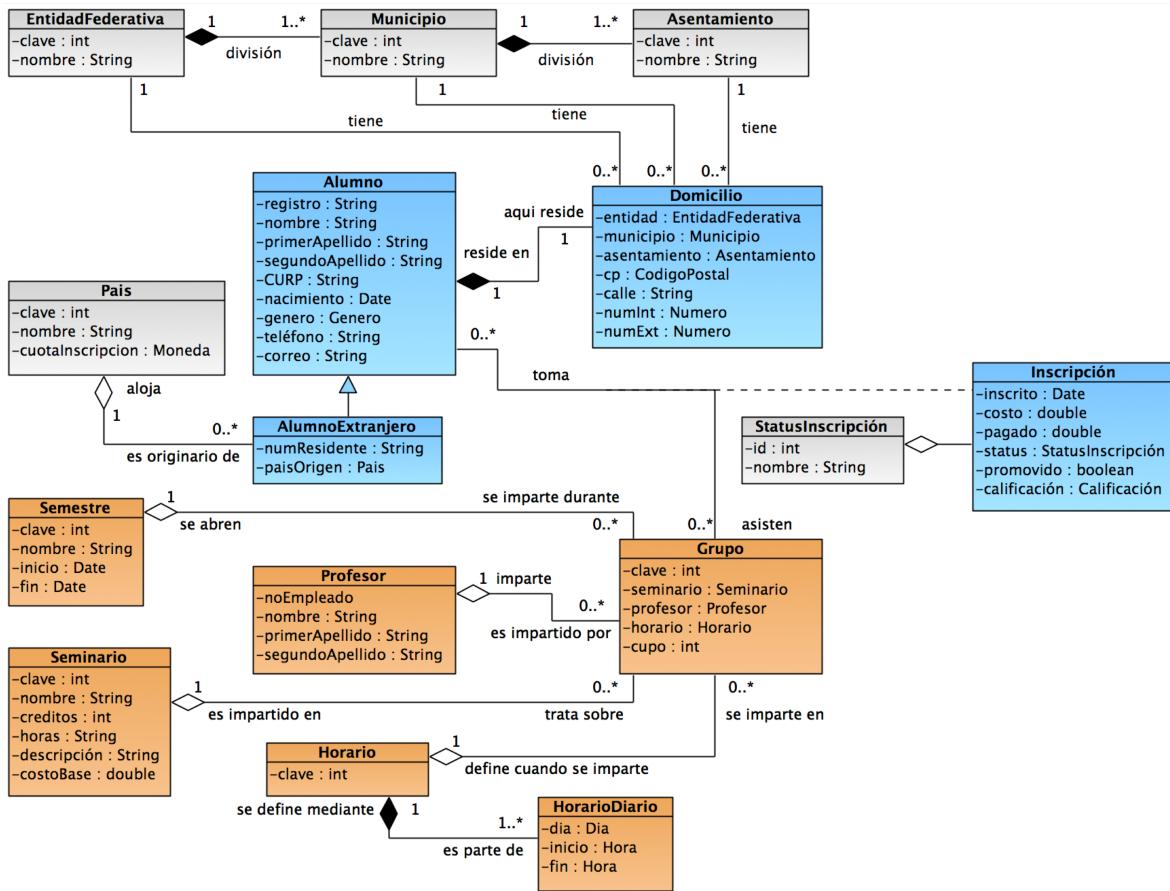
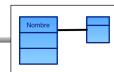


Figura 3.1: Modelo del dominio del problema - Sistema de identificación de NNA afectados por feminicidios

- **Entidades principales:** Noticia, Medio, Caso, Cluster, MencionNNA, Tópico, FuenteRSS

- **Relaciones clave:**

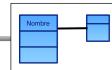
- Medio (1) —publicar—> (0..*) Noticia
- Noticia (1) —pertenecer—> (0..1) Cluster
- Noticia (1) —contener—> (0..1) MencionNNA
- Noticia (0..*) —distribuirse—> (1..*) Tópico [con probabilidad]
- Cluster (1) —corresponder—> (0..1) Caso [ideal, no siempre logrado]
- FuenteRSS (1) —proveer—> (0..*) Noticia



3.3.1. Entidad: Noticia

Atributos			
Nombre	Tipo	Descripción	Requerido
Identificador	<i>Id</i>	Identificador único autoincrementado asignado al ingestar noticia al sistema	Sí
Título	<i>Cadena Larga</i>	Título completo de la noticia tal como fue publicado por el medio	Sí
URL	<i>URL</i>	Dirección web completa de la noticia original para acceso directo y verificación	Sí
Fecha de publicación	<i>Fecha</i>	Fecha en que el medio publicó la noticia según metadata del feed RSS o scraping	Sí
Contenido	<i>Texto Largo</i>	Cuerpo completo de la noticia extraído de RSS (snippet) o scraping (texto completo)	Sí
Contenido limpio	<i>Texto Largo</i>	Versión preprocesada del contenido: minúsculas, sin stopwords, lematizado, para análisis PLN	Sí
Vector TF-IDF	<i>Vector[3000]</i>	Representación vectorial numérica del contenido limpio generada por TfIdfVectorizer	Sí
Menciona NNA	<i>Booleano</i>	Bandera indicando si el detector dual-etapa identificó mención de niños huérfanos (true) o no (false)	Sí
Es feminicidio	<i>Booleano</i>	Bandera indicando si la noticia fue clasificada como relacionada con feminicidio en etapa 1 del detector	Sí

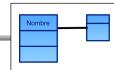
Atributos			
Nombre	Tipo	Descripción	Requerido
ID de Cluster	<i>Entero</i>	Identificador del cluster al que pertenece según DBSCAN, o -1 si es ruido no agrupado	Sí
Fecha de ingesta	<i>Fecha y hora</i>	Timestamp de cuando el sistema procesó e ingresó la noticia al almacenamiento	Sí
Relaciones			
Tipo de relación	Entidad	Rol	
◆—Composición	Medio	Una Noticia es publicada por un Medio	
◇—Agregación	Cluster	Una Noticia puede pertenecer a un Cluster	
◆—Composición	MenciónNNA	Una Noticia puede contener una Mención NNA	
◇—Agregación	Tópico	Una Noticia se distribuye probabilísticamente sobre múltiples Tópicos	



3.3.2. Entidad: Medio de Comunicación

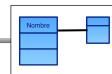
Atributos			
Nombre	Tipo	Descripción	Requerido
Identificador	<i>Id</i>	Identificador único del medio en el sistema	Sí
Nombre	<i>Palabra Corta</i>	Nombre oficial del medio (La Jornada, Proceso, Aristegui Noticias, etc.)	Sí
URL base	<i>URL</i>	Dirección web raíz del sitio del medio (https://www.jornada.com.mx)	Sí
Feed RSS	<i>URL</i>	URL del feed RSS específico monitoreado por el sistema	No
Alcance	<i>Enumeración</i>	Clasificación de cobertura: Nacional, Estatal, Regional, según área geográfica principal del medio	Sí
Activo	<i>Booleano</i>	Bandera indicando si el medio está siendo monitoreado activamente (true) o deshabilitado temporalmente (false)	Sí
Noticias procesadas	<i>Entero</i>	Contador de noticias recolectadas de este medio desde inicio del sistema	Sí
Relaciones			
Tipo de relación	Entidad	Rol	

Atributos			
Nombre	Tipo	Descripción	Requerido
◆—Composición	Noticia	Un Medio publica múltiples Noticias	
◇—Agregación	FuenteRSS	Un Medio puede tener asociado una Fuente RSS	



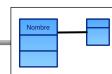
3.3.3. Entidad: Cluster de Noticias

Atributos			
Nombre	Tipo	Descripción	Requerido
Identificador	<i>Id</i>	Identificador numérico único del cluster asignado por DBSCAN	Sí
Número de noticias	<i>Entero</i>	Cantidad de noticias agrupadas en este cluster	Sí
Centroide	<i>Vector[3000]</i>	Vector TF-IDF promedio de todas las noticias del cluster, representa contenido temático central	Sí
Palabras clave	<i>Lista de cadenas</i>	Top 10 palabras con mayor peso TF-IDF en el centroide, resumen temático del cluster	Sí
Cohesión interna	<i>Real [0-1]</i>	Métrica de similitud coseno promedio entre noticias del cluster, valores mayores indican agrupación cohesiva	Sí
Fecha más temprana	<i>Fecha</i>	Fecha de publicación de la noticia más antigua del cluster, indica inicio temporal del caso	Sí
Fecha más reciente	<i>Fecha</i>	Fecha de publicación de la noticia más reciente, indica seguimiento temporal del caso	Sí
Caso asociado	<i>Referencia</i>	Referencia opcional a entidad Caso si se determinó correspondencia con evento real específico (proceso manual TT2)	No
Relaciones			
Tipo de relación	Entidad	Rol	
◆—Composición	Noticia	Un Cluster agrupa múltiples Noticias	
◇—Agregación	Caso	Un Cluster puede corresponder a un Caso único (ideal)	



3.3.4. Entidad: Mención de NNA

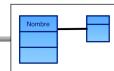
Atributos			
Nombre	Tipo	Descripción	Requerido
Identificador	<i>Id</i>	Identificador único de la mención detectada	Sí
Texto de contexto	Cadena Larga	Fragmento de texto circundante donde se detectó la mención (± 50 caracteres)	Sí
Patrón detectado	Cadena Corta	Expresión regular o patrón lingüístico que activó detección (“dejó N hijos”, “madre de”, etc.)	Sí
Número de NNA	Entero	Cantidad de NNA mencionados si es explícita (“dejó 3 hijos”), o null si es implícita	No
Edades de NNA	Lista de enteros	Lista de edades mencionadas si están presentes (“hijos de 5 y 8 años”), lista vacía si no se especifican	No
Confianza de detección	Real [0-1]	Score de confianza del detector en la mención (1.0 = patrón explícito, ≤ 1.0 = patrón ambiguo)	Sí
Relaciones			
Tipo de relación	Entidad	Rol	
◆—Composición	Noticia	Una Mención NNA pertenece a una Noticia	



3.3.5. Entidad: Tópico LDA

Atributos			
Nombre	Tipo	Descripción	Requerido
Identificador	<i>Id</i>	Identificador numérico del tópico (0 a n_topics-1 configurado en LDA)	Sí
Palabras top	Lista de tuplas	Lista de (palabra, peso) con las 10 palabras más representativas del tópico ordenadas descendenteamente	Sí
Etiqueta interpretativa	Cadena Corta	Etiqueta semántica asignada manualmente al tópico basándose en palabras top (“Feminicidios urbanos”, “Violencia intrafamiliar”, etc.)	No

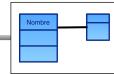
Atributos			
Nombre	Tipo	Descripción	Requerido
Coherencia	<i>Real</i>	Score de coherencia C_V del tópico, indica calidad semántica (valores típicos 0.3-0.7)	Sí
Noticias asociadas	<i>Entero</i>	Cantidad de noticias donde este tópico tiene probabilidad dominante (>0.3)	Sí
Relaciones			
Tipo de relación	Entidad	Rol	
◊—Agregación	Noticia	Un Tópico se distribuye sobre múltiples Noticias con diferentes probabilidades	



3.3.6. Entidad: Caso de Feminicidio

Atributos			
Nombre	Tipo	Descripción	Requerido
Identificador	<i>Id</i>	Identificador único del caso validado	Sí
Nombre de víctima	<i>Cadena Corta</i>	Nombre de la víctima si fue divulgado públicamente (respetando privacidad según contexto)	No
Fecha del caso	<i>Fecha</i>	Fecha en que ocurrió el feminicidio según reportes oficiales o periodísticos	Sí
Entidad federativa	<i>Cadena Corta</i>	Estado de la República donde ocurrió el caso	Sí
Municipio	<i>Cadena Corta</i>	Municipio específico donde ocurrió el caso	No
Número de NNA	<i>Entero</i>	Cantidad confirmada de NNA que quedaron en situación de orfandad	No
Status legal	<i>Enumeración</i>	Estado de la investigación: En investigación, Tipificado como feminicidio, Sentenciado, Desconocido	No
Cluster asociado	<i>Referencia</i>	Referencia al cluster de noticias que documenta este caso	No
Validado por	<i>Cadena Corta</i>	Nombre de organización u organismo que validó manualmente el caso (uso futuro TT2)	No
Relaciones			
Tipo de relación	Entidad	Rol	

Atributos			
Nombre	Tipo	Descripción	Requerido
◊—Agregación	Cluster	Un Caso puede estar documentado por un Cluster	
◆—Composición	MencionNNA	Un Caso puede tener múltiples Menciones NNA en diferentes noticias	



3.3.7. Entidad: Fuente RSS

Atributos			
Nombre	Tipo	Descripción	Requerido
Identificador	<i>Id</i>	Identificador único de la fuente RSS en configuración del sistema	Sí
URL del feed	<i>URL</i>	Dirección completa del feed RSS/Atom consultado	Sí
Frecuencia de actualización	<i>Intervalo</i>	Cada cuántas horas el sistema consulta este feed (típicamente 6 horas)	Sí
Última consulta exitosa	<i>Fecha y hora</i>	Timestamp de la última vez que el feed fue consultado exitosamente	Sí
Último error	<i>Cadena Larga</i>	Mensaje de error de la última consulta fallida, null si última consulta fue exitosa	No
Activo	<i>Booleano</i>	Bandera de habilitación: true si debe ser consultado, false si fue deshabilitado por errores persistentes	Sí
Relaciones			
Tipo de relación	Entidad	Rol	
◊—Agregación	Medio	Una Fuente RSS provee noticias de un Medio	
◆—Composición	Noticia	Una Fuente RSS provee múltiples Noticias	

3.4. Modelado de Reglas de negocio

Las reglas de negocio definen restricciones, políticas y lógica operativa que gobiernan el comportamiento del sistema. Se organizan en cinco categorías: reglas de recolección (cuándo y cómo obtener noticias), reglas de detección (criterios para clasificar relevancia), reglas de procesamiento PLN (configuraciones de algoritmos), reglas de almacenamiento (integridad de

datos), y reglas de acceso (quién puede consultar qué información). Estas reglas son implementadas mediante validaciones en código, configuraciones de parámetros y políticas de uso documentadas.

BR-001 Nombre de la regla de negocio

Tipo: Habilitadora.	Clase: De condición.	Nivel: Estricto.
----------------------------	-----------------------------	-------------------------

Descripción: Descripción de la regla. Forma coloquial a manera de reglamento.

Motivación: Describa por que es importante la regla.

Sentencia: Sentencia formal de la regla.

Ejemplo positivo: Indique uno o varios ejemplos en donde la regla se cumple.

- ...

Ejemplo negativo: Indique uno o varios ejemplos en dónde la regla no se cumple.

- ...

Referenciado por: Liste los casos de uso en donde la regla no se cumple. por ejemplo [CU-CE3.2](#), [CUCE3.3](#).

3.5. Máquinas de estado

Las máquinas de estado modelan ciclos de vida de entidades clave del sistema, especificando estados posibles, transiciones permitidas, eventos que disparan transiciones, y acciones asociadas. Se documentan tres máquinas principales: (1) Estado de Noticia (desde recolección hasta almacenamiento persistente), (2) Estado de Fuente RSS (operativa, en advertencia, deshabilitada), y (3) Estado de Caso (pendiente validación, validado, descartado). Estas máquinas facilitan comprensión de flujos de procesamiento y guían implementación de lógica de transición en el código.

3.5.1. Estados para un préstamo

En la figura 3.2 se muestran ...

Estados

Estado: Descripción del estado.

... ...

Acciones

Acción: Descripción de la acción indicando el Caso de uso involucrado.

... ...



Figura 3.2: Máquina de estados de un Préstamo.

CAPÍTULO 4

Modelo dinámico

Este capítulo describe en modelo dinámico del sistema. en el se detallan todos los escenarios de ejecución del sistema. La figura 4.1 muestra el diagrama general del sistema y sus subsistemas, y la figura 4.2 muestra todos los casos de uso del sistema. En este documento solo detallamos los casos de uso del subsistema de gestión de cursos.

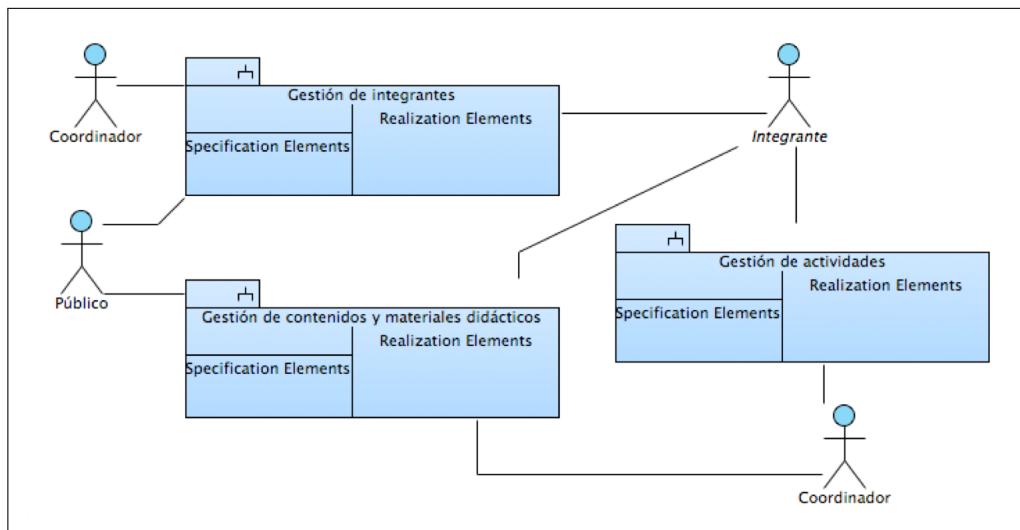


Figura 4.1: Diagrama de casos de uso del sistema.

4.1. Descripción de casos de uso

A continuación se detallan los casos de uso.

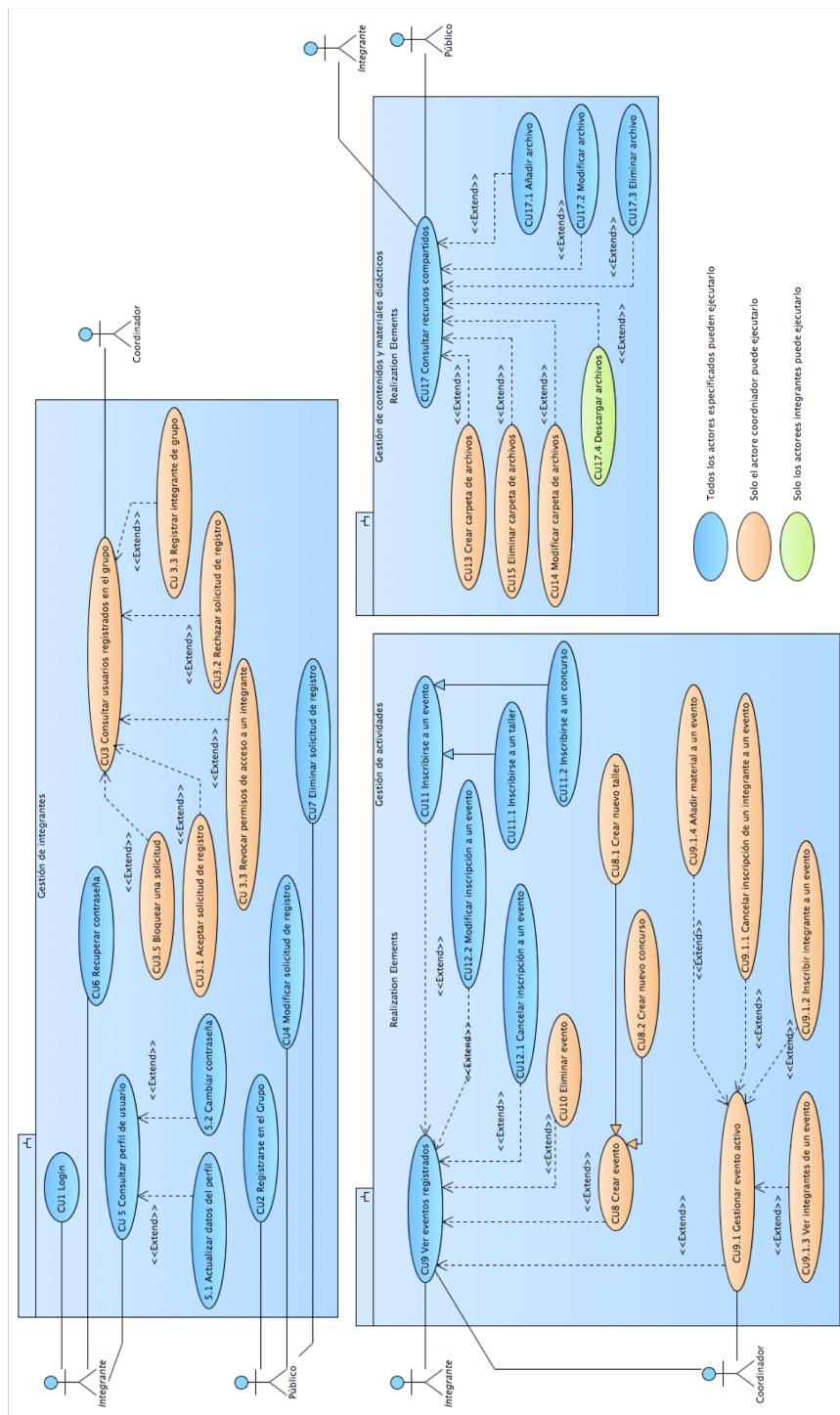


Figura 4.2: Diagrama detallado del sistema.



4.2. CUX Escriba el nombre del caso de uso

4.2.1. Descripción completa

4.2.2. Atributos importantes

Caso de Uso:	CUX Escriba el nombre del caso de uso
Versión:	0.1
Autor:	Nombre del analista.
Supervisa:	Nombre del analista revisor.
Actor:	Nombre del actor
Propósito:	<ul style="list-style-type: none"> • Propósito del caso de uso. • ...
Entradas:	<ul style="list-style-type: none"> • Nombre del dato de entrada. • Nombre del dato de entrada.
Origen:	<ul style="list-style-type: none"> • Se introducen desde el teclado. • otros.
Salidas:	<ul style="list-style-type: none"> • Nombre del dato de entrada. • Mensajes de error. • Datos que aparecen en pantalla. • Datos que aparecen en listas desplegables o tablas, etc. • Datos que se imprimen o que se envía a otros sistemas.
Destino:	<ul style="list-style-type: none"> • Se muestra en la pantalla IUX Nombre pantalla.. • otros.
Precondiciones:	<ul style="list-style-type: none"> • Escriba la precondición.
Postcondiciones:	<ul style="list-style-type: none"> • Escriba todas las postcondiciones.
Errores:	<ul style="list-style-type: none"> • E1: Condición que detona el error, reacción del sistema y regresa al paso ??. • E2: Condición que detona el error, reacción del sistema y termina el Caso de uso..
Tipo:	Caso de uso primario
Observaciones:	

4.2.3. Trayectorias del Caso de Uso

Trayectoria principal

- 1 o VERBO EN INFINITIVO + (Acción del usuario) + (Acción dentro del sistema)
- 2 Ingresa al sistema escribiendo la URL de la aplicación.
- 3 Sigue el flujo de datos que se identifique mediante la pantalla IU1 Inicio de sesión
- 4 Se identifica introduciendo su nombre de usuario y contraseña.
- 5 Sigue el flujo de datos presiona el botón .
- 6 Busca los datos del usuario identificado por el nombre de usuario introducido E1 No hay usuario
- 7 Verifica que el usuario especificado no esté inactivo E2 Usuario inactivo [Trayectoria A].
- 8 Verifica que la contraseña ingresada coincida con la almacenada E3 La contraseña no coincide [Trayectoria B].
- 9 Se ejecutan los pasos del caso de uso **CUY Nombre del caso de uso**.
- 10 Muestra la pantalla IU2 Principal con el mensaje **MSG-001 Bienvenida al usuario**.
- - - - *Fin del caso de uso.*

Trayectoria alternativa LETRA:

Condición: Condición que hace que se ejecute esta trayectoria

LETRA1 Especifique los pasos de la trayectoria.

LETRA2 El Caso de Uso continúa en el paso **3**.

- - - - *Fin de la trayectoria.*

4.2.4. Puntos de extensión

Cuando: El usuario no recuerda cual es su contraseña o sospecha que su usuario está bloqueado.

Durante la región: Del paso **??** al paso **??**.

La operación se puede extender a: **CUZ Nombre del caso de uso.**



4.3. CUX Escriba el nombre del caso de uso

4.3.1. Descripción completa

4.3.2. Atributos importantes

Caso de Uso:	CUX Escriba el nombre del caso de uso
Versión:	0.1
Estatus:	.
Prioridad:	.
Usuario:	Nombre del usuario o usuarios
Elaboró:	Nombre del analista
Supervisó:	Nombre del analista revisor.
Validó:	Nombre del usuario o usuarios
Complejidad:	.
Volatilidad:	.
Madurez:	.
Dificultades:	•
Proceso:	
Sub-proceso:	
Área:	
Actor:	Nombre del actor
Tipo de operación:	
Frecuencia:	<ul style="list-style-type: none"> • Mínimo: • Promedio: • Máximo:
Volumen:	<ul style="list-style-type: none"> • Mínimo: • Promedio: • Máximo:

Caso de Uso:	CUX Escriba el nombre del caso de uso
Req. de usuario:	
Fuentes:	•
Propósito:	•
Entradas:	<ul style="list-style-type: none"> • Nombre del dato de entrada. • Nombre del dato de entrada.
Origen:	<ul style="list-style-type: none"> • Se introducen desde el teclado. • otros.
Salidas:	<ul style="list-style-type: none"> • Nombre del dato de entrada. • Mensajes de error. • Datos que aparecen en pantalla. • Datos que aparecen en listas desplegables o tablas, etc. • Datos que se imprimen o que se envía a otros sistemas.
Destino:	<ul style="list-style-type: none"> • Se muestra en la pantalla  IUX Nombre pantalla.. • otros.
Disparadores:	•
Precondiciones:	<ul style="list-style-type: none"> • Escriba la precondición.
Condición de término:	•
Efectos colaterales:	•
Postcondiciones:	<ul style="list-style-type: none"> • Escriba todas las postcondiciones.
Errores:	<ul style="list-style-type: none"> • E1: Condición que detona el error, reacción del sistema y regresa al paso ??. • E2: Condición que detona el error, reacción del sistema y termina el Caso de uso..
Tipo:	Caso de uso primario
Casos de prueba:	<ul style="list-style-type: none"> • E1: Condición que detona el error, reacción del sistema y regresa al paso ??. • E2: Condición que detona el error, reacción del sistema y termina el Caso de uso..
Consideraciones de diseño:	<ul style="list-style-type: none"> • ...
Impedimentos:	<ul style="list-style-type: none"> • E1: Condición que detona el error, reacción del sistema y regresa al paso ??. • E2: Condición que detona el error, reacción del sistema y termina el Caso de uso..
Preguntas:	<ul style="list-style-type: none"> • ...
Observaciones:	

4.3.3. Trayectorias del Caso de Uso

Trayectoria principal

- 1 Ingresa al sistema escribiendo la URL de la aplicación.
 - 2 Sigue el paso 1. Sigue el paso 2. Sigue el paso 3. Sigue el paso 4. Sigue el paso 5. Sigue el paso 6. Sigue el paso 7. Sigue el paso 8. Sigue el paso 9.
 - 1 Ingresa al sistema escribiendo la URL de la aplicación.
 - 2 Sigue el paso 1. Sigue el paso 2. Sigue el paso 3. Sigue el paso 4. Sigue el paso 5. Sigue el paso 6. Sigue el paso 7. Sigue el paso 8. Sigue el paso 9.
 - 3 Se identifica introduciendo su nombre de usuario y contraseña.
 - 4 Sigue el paso 1. Sigue el paso 2. Sigue el paso 3. Sigue el paso 4. Sigue el paso 5. Sigue el paso 6. Sigue el paso 7. Sigue el paso 8. Sigue el paso 9.
 - 1 Ingresa al sistema escribiendo la URL de la aplicación.
 - 2 Sigue el paso 1. Sigue el paso 2. Sigue el paso 3. Sigue el paso 4. Sigue el paso 5. Sigue el paso 6. Sigue el paso 7. Sigue el paso 8. Sigue el paso 9.
 - 5 Busca los datos del usuario identificado por el nombre de usuario introducido
 - 6 Verifica que el usuario especificado no esté inactivo [Trayectoria A].
 - 7 Verifica que la contraseña ingresada coincida con la almacenada [Trayectoria B].
 - 8 Se ejecutan los pasos del caso de uso CUY CUY Nombre del caso de uso.
 - 9 Muestra la pantalla IU2 Principal con el mensaje MSG1 Bienvenida al usuario.
- - - - *Fin del caso de uso.*

Trayectoria alternativa LETRA:

Condición: Condición que hace que se ejecute esta trayectoria

LETRA1 Especifique los pasos de la trayectoria.

LETRA2 El Caso de Uso continúa en el paso 3.

- - - - *Fin de la trayectoria.*

4.3.4. Puntos de extensión

Cuando: El usuario no recuerda cual es su contraseña o sospecha que su usuario está bloqueado.

Durante la región: Del paso ?? al paso ??.

La operación se puede extender a: CU3.4 Consultar historial académico.

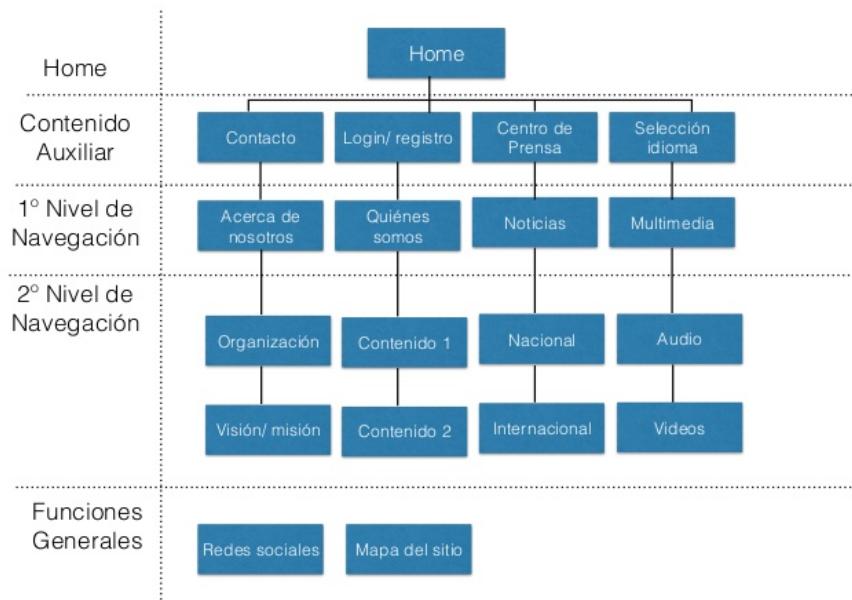
CAPÍTULO 5

Modelo de la interacción

Este capítulo describe ...

5.1. Modelo de navegación

La navegación entre pantallas se muestra en la figura 5.1. en el se explica ...



8

Figura 5.1: mapa

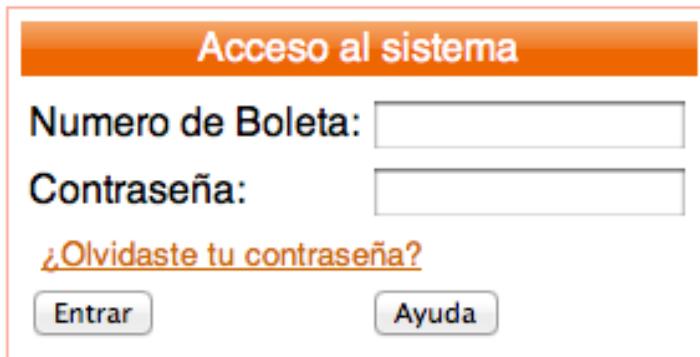
5.2. IUX Interfaz (nombre de la interfaz)

5.2.1. Objetivo

5.2.2. Diseño

Esta pantalla IU23 Pantalla de Control de Acceso aparece al iniciar el sistema, para ingresar ...

Acceso al sistema



The image shows the IU23 Access Control Screen. It has a title bar 'Acceso al sistema' at the top. Below it are two input fields: 'Número de Boleta:' and 'Contraseña:', each with a corresponding text input box. Underneath these fields is a link '[¿Olvidaste tu contraseña?](#)'. At the bottom are two buttons: 'Entrar' on the left and 'Ayuda' on the right.

Figura 5.2: IU23 Pantalla de Control de Acceso.

5.2.3. Salidas

- Descripción de salida.

5.2.4. Entradas

- Descripción de salida.

5.2.5. Comandos

- Entrar : Verifica que el Estudiante se encuentre registrado y la contraseña sea la correcta. Si la verificación es correcta, se muestra la UI32 Pantalla de Selección de Seminario.
- Ayuda : Muestra la ayuda de esta pantalla IU50 Pantalla de Ayuda.

5.3. Catálogo de mensajes

En esta sección se describen todos los mensajes que aparecen en el sistema. Para cada mensaje se especifica:

Id: Identificador del mensaje de la forma “MSG XX” y descripción corta del mismo.

Tipo: Tipo del mensaje el cual puede ser:

Normal: Mensaje que informa al usuario una instrucción o el estado interno que guarda el sistema, suele tener un color **Azul**.

Éxito: Mensaje que informa al usuario sobre una acción realizada, sirve para confirmar el correcto funcionamiento del sistema. Se presentan con un color **Verde**.

Atención: Mensaje que tiene como finalidad llamar la atención del usuario a una situación que requiere su intervención, por ejemplo cuando una actividad ha generado un efecto colateral o se realizará una acción destructiva y no reversible. Se presentan con un color **Naranja**.

Error: Mensaje que informa al usuario un fallo en una operación o un impedimento para realizarla, por ejemplo: cuando no se puede efectuar la acción solicitada, cuando un dato falta o tiene un formato no aceptado por el sistema. Se presentan con un color **Rojo**.

Propósito: Explicación del propósito del mensaje.

Redacción: Redacción del mensaje.

Parámetros: En caso de que el mensaje pueda variar se especifican los casos y la forma en que debe adaptarse la redacción

Ejemplos: Ejemplos de como debe renderizarse el mensaje.

5.3.1. Lista de mensajes

MSG-001: Bienvenida al usuario.

Propósito: Indicar al usuario que ha ingresado satisfactoriamente al sistema.

Redacción: Bienvenido <nombre>.

Parámetros:

- <nombre> **Nombre completo** del Usuario.

Ejemplos: Bienvenido Juan Pérez.

MSG-002: Usuario no registrado.

Propósito: Indica que el usuario ingresado no existe en el sistema.

Redacción: El usuario <login> no se encuentra registrado.

Parámetros:

- <login> [Login](#) del Usuario.

Ejemplos: El usuario juanP no se encuentra registrado.

MSG-003: Cuenta inactiva.

Propósito: Indicar al usuario que la cuenta especificada se encuentra inactiva.

Redacción: La cuenta especificada <login> se encuentra inactiva, favor de contactar al Secretario Escolar para mas información.

Parámetros:

- <login> [Login](#) del Usuario.

Ejemplos: La cuenta especificada juanP se encuentra inactiva, favor de contactar al Secretario Escolar para mas información.

MSG-004: Error de inicio de sesión.

Propósito: Indica al usuario que la contraseña introducida es incorrecta.

Redacción: La contraseña ingresada es incorrecta.

Parámetros: No aplica.

Ejemplos: La contraseña ingresada es incorrecta.

MSG-008: Tiempo restante para terminar un proceso.

Propósito: Indicar al usuario el tiempo restante para terminar una operación limitada en el tiempo como la reinscripción.

Redacción: Quedan <tiempo> para terminar la <operación>.

Parámetros:

- <tiempo> Tiempo faltante para la operación especificando días, horas minutos y segundos.
-

Ejemplos:

- Quedan 45 días, 2 horas 12 minutos y 45 segundos para iniciar tu reinscripción.
- Quedan 2 minutos y 32 segundos para terminar tu reinscripción

MSG-009: Operación exitosa.

Propósito: Informar al usuario que la operación solicitada ha sido ejecutada con éxito.

Redacción: <Artículo> <Operación> del <Entidad> <Identificador> se realizó con éxito.

Parámetros:

- <Artículo> <Operación> se refiere a la operación realizada.
- <Entidad> <Identificador> se refiere al elemento del negocio donde recayó la operación, indicando el tipo del objeto y un dato que el usuario pueda usar para identificarlo.

Ejemplo: Algunos ejemplos son

- El registro del Alumno 342343 se realizó con éxito.
- La eliminación de la Tarea “documentar el proceso” se ha realizado con éxito.

Bibliografía

- [Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- [de Diputados, 2012] de Diputados, C. (2012). Código penal federal - artículo 325. *Diario Oficial de la Federación*. Reforma publicada DOF 14-06-2012.
- [de Gobernación, 2023] de Gobernación, S. (2023). Información sobre violencia contra las mujeres. incidencia delictiva y llamadas de emergencia 9-1-1. *SESNP - Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública*. Consultado en noviembre 2025.
- [Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.
- [Docker Inc., 2025] Docker Inc. (2025). Docker documentation. <https://docs.docker.com>. Consultado en enero 2026.
- [Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231.
- [Grinberg, 2018] Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, 2nd edition.
- [Honnibal and Montani, 2023] Honnibal, M. and Montani, I. (2023). spacy: Industrial-strength natural language processing. <https://spacy.io>.
- [MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

- [Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- [ONU Mujeres México, 2023] ONU Mujeres México (2023). Violencia feminicida en México: Aproximaciones y tendencias.
- [Pedregosa et al., 2011] Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in python.
- [Raschka and Mirjalili, 2019] Raschka, S. and Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing, 3rd edition.
- [Red por los Derechos de la Infancia en México, 2023] Red por los Derechos de la Infancia en México (2023). Infancia y adolescencia en México: Análisis de datos 2023.
- [Richardson, 2023] Richardson, L. (2023). Beautiful soup documentation.
- [Salton and McGill, 1983] Salton, G. and McGill, M. J. (1983). Introduction to modern information retrieval.

APÉNDICE A

Anexos

A.1. Glosario de Términos

NNA

Niñas, Niños y Adolescentes. Término utilizado en el marco jurídico mexicano para referirse a personas menores de 18 años.

Feminicidio

Delito tipificado en el Código Penal Federal como el asesinato de una mujer por razones de género, con agravantes específicas contempladas en el artículo 325.

Procesamiento de Lenguaje Natural (PLN)

Rama de la inteligencia artificial que permite a las computadoras entender, interpretar y generar lenguaje humano de manera útil.

Web Scraping

Técnica de extracción automática de información de sitios web mediante programas informáticos.

RSS (Really Simple Syndication)

Formato XML utilizado para distribuir contenido actualizado de sitios web de forma automatizada.

TF-IDF

Term Frequency - Inverse Document Frequency. Medida estadística que evalúa la importancia de una palabra en un documento dentro de una colección.

Clustering

Técnica de aprendizaje no supervisado que agrupa elementos similares en conjuntos denominados clusters.

API REST

Application Programming Interface basada en el estilo arquitectural REST que permite la comunicación entre sistemas mediante protocolo HTTP.

A.2. Configuración del Entorno de Desarrollo

Requisitos del Sistema

- Python 3.9 o superior
- Docker Desktop 4.0 o superior
- 8 GB RAM mínimo (16 GB recomendado)
- 10 GB espacio en disco
- Sistema operativo: Windows 10/11, macOS 11+, o Linux (Ubuntu 20.04+)

Librerías Python Principales

- scikit-learn 1.3.0 - Algoritmos de machine learning
- Flask 2.3.0 - Framework web
- pandas 2.0.0 - Análisis de datos
- beautifulsoup4 4.12.0 - Parsing HTML
- feedparser 6.0.10 - Procesamiento RSS
- spacy 3.6.0 - Procesamiento de lenguaje natural

A.3. Ejemplos de Casos de Uso

Caso de Éxito: Detección de Cluster de Feminicidios en Chihuahua

Durante las pruebas del Prototipo 2, el sistema identificó automáticamente un cluster de 12 noticias relacionadas con casos de feminicidio en el estado de Chihuahua durante un periodo de 3 semanas, permitiendo detectar un patrón geográfico-temporal que no había sido documentado previamente por las organizaciones civiles.

Caso de Refinamiento: Reducción de Falsos Positivos

La implementación del detector de dos etapas redujo la tasa de falsos positivos del 28 % (P1) al 10 % (P2), mejorando significativamente la calidad de los datos recolectados y reduciendo el tiempo de validación manual requerido.

A.4. Código de Ejemplo: Detector de Relevancia

Listing A.1: Fragmento del detector de relevancia de dos etapas

```

1 def is_relevant(title, description):
2     """Detector de dos etapas para identificar noticias relevantes"""
3
4     # Etapa 1: Filtro rápido con palabras clave
5     keywords = ['feminicidio', 'asesinato_de_mujer', 'hijos',
6                 'huérfanos', 'menores', 'ninos']
7     text = (title + ' ' + description).lower()
8
9     if not any(keyword in text for keyword in keywords):
10         return False
11
12     # Etapa 2: Validación semántica con TF-IDF
13     vector = vectorizer.transform([text])
14     similarity = cosine_similarity(vector, reference_vectors)
15
16     return similarity.max() > 0.3

```

A.5. Diagramas Complementarios

Flujo de Procesamiento de Noticias

El flujo completo de procesamiento incluye las siguientes etapas:

1. Recolección desde múltiples fuentes

2. Deduplicación por URL y similitud de contenido
3. Detección de relevancia (dos etapas)
4. Extracción de características con TF-IDF
5. Clustering automático
6. Almacenamiento en formato JSON
7. Exposición mediante API REST
8. Visualización en dashboard web

A.6. Referencias de Contacto

Organizaciones Colaboradoras

- **Fundación Futuro con Derechos**

Organización enfocada en la defensa de derechos de NNA en situación de orfandad por feminicidio

- **Red por los Derechos de la Infancia en México (REDIM)**

Red de organizaciones civiles dedicadas a la promoción y defensa de los derechos de la infancia

Autores del Proyecto

- **Herrera Ramírez Emilio Alejandro**

Estudiante de Ingeniería en Sistemas Computacionales - ESCOM IPN

- **Morales Martínez Héctor Alberto**

Estudiante de Ingeniería en Sistemas Computacionales - ESCOM IPN