

Alcance ético del proyecto

Proyecto: JEM-Full — archivo del Jurado de Enjuiciamiento de Magistrados preparado para IA **Fase:** 1 del plan de [JEM.md](#) **Estado:** borrador para adopción por el equipo

Usa la convención de etiquetado descrita en [marco-legal.md](#) §0: [Dato], [Inferencia], [Propuesta], [Verificar].

Este documento fija **qué hace y qué no hace el proyecto**. Su función es que las decisiones de alcance estén escritas antes de que la presión de un plazo las tome por nosotros.

1. Objeto declarado

[Propuesta] El proyecto estudia **la preparación de un archivo documental público para su uso por sistemas de IA**. Su unidad de análisis es el documento y la causa, no la persona.

El proyecto no es, y debe rechazar activamente ser leído como:

- una evaluación del desempeño del JEM como órgano;
- una auditoría de la conducta de magistrados individuales;
- un sistema de perfilamiento de jueces, fiscales o denunciantes;
- una herramienta de predicción de decisiones judiciales.

[Inferencia] Esta delimitación no es sólo prudencia institucional: es lo que hace que el trabajo sea metodológicamente honesto. Un corpus con la tasa de error de OCR aún sin medir y con la extracción de votos sin construir **no puede sostener ninguna afirmación sobre conducta individual**. Afirmarlas sería un error técnico antes que ético.

2. Superficie de recolección: qué sí y qué no

2.1 Incluido

[Propuesta] Sólo la **superficie pública de consulta**:

Fuente	Naturaleza
Resoluciones publicadas (autos interlocutorios, sentencias definitivas)	Publicación de oficio
Dictámenes	Publicación de oficio
Expedientes escaneados de acceso público	Publicación de oficio
Órdenes del día y actas de sesión	Publicación de oficio
Informes estadísticos institucionales	Publicación de oficio
Marco normativo publicado por el JEM	Publicación de oficio

2.2 Excluido deliberadamente

[Propuesta] Queda **fuera del alcance**, y su exclusión debe declararse en el artículo:

Fuente	Motivo de exclusión
Sistema transaccional autenticado (presentación de escritos, seguimiento por las partes)	Requiere usuario y contraseña. Automatizar el acceso sería acceso indebido, con independencia de que las credenciales existieran.
"Trámite Digital de Antecedentes"	Datos personales sensibles de postulantes, ajenos al objeto del estudio.
Cualquier documento obtenido por error de configuración del servidor	Que sea técnicamente alcanzable no lo hace público. Si aparece, se reporta al JEM y no se incorpora.
Datos de partes obtenidos por cruce con fuentes externas	Multiplicaría el riesgo de reidentificación sin aportar al objeto de estudio.

[Propuesta] **Regla de decisión ante la duda:** si para obtener un documento hace falta sortear un control de acceso —autenticación, CAPTCHA, límite de tasa, cabecera de robots— el documento está fuera de alcance. La barrera técnica se interpreta como manifestación de voluntad del órgano, aunque sea débil.

3. Conducta técnica frente a los servidores del JEM

[Propuesta] Toda recolección futura (Fase 2) debe:

1. Identificarse con un `User-Agent` honesto, que incluya el nombre del proyecto y una forma de contacto.
2. Respetar `robots.txt` aunque no sea jurídicamente vinculante.
3. Limitar la concurrencia y la tasa de petición a niveles que no degraden el servicio. [Inferencia] Los portales corren sobre JSF con estado de sesión en el servidor; la concurrencia agresiva puede agotar sesiones y afectar a usuarios reales.
4. Preferir la sincronización incremental al re-scraping completo.
5. Registrar cada descarga con URL, fecha, hora y código de respuesta —requisito de procedencia que hoy **no se cumple** (ver [../../MANIFEST.md](#) §7).

[Propuesta] Ejecutar la recolección en horario de baja demanda y detenerla ante cualquier señal de degradación del servicio.

4. Personas identificables en el corpus

4.1 Qué contiene realmente

[Dato] Medido sobre el corte vigente:

- Los nombres propios aparecen **en las rutas de archivo**, no sólo en el texto: el inventario contiene entradas del tipo `Auto_Interlocutorio_126-2020 - Investigación Preliminar <NOMBRE Y APELLIDOS>`.
- **49 archivos** tienen `adolescencia` en su ruta y **63** tienen `Niñez`.
- **325 documentos** con texto extraído mencionan `adolescencia`; **289** mencionan `Niñez`.

[Inferencia] Las menciones a la jurisdicción de la niñez y la adolescencia corresponden mayoritariamente a la **competencia del magistrado enjuiciado**, no a que el menor sea parte. Pero **no está verificado**, y hasta verificarlo debe asumirse el caso más restrictivo.

4.2 Categorías de personas y tratamiento diferenciado

[Propuesta] No todas las personas nombradas están en la misma posición:

Categoría	Posición	Tratamiento
Magistrados y agentes fiscales enjuiciados	Funcionarios públicos, en ejercicio de función pública, sometidos a un procedimiento público	Menor expectativa de privacidad. Tratamiento estándar.
Integrantes del JEM	Ídem, en ejercicio de función pública	Ídem. Son el objeto de la extracción de votos.
Denunciantes	Particulares que activaron un procedimiento público	Expectativa intermedia. Evaluar seudonimización en derivados publicados.
Terceros mencionados incidentalmente (testigos, víctimas, familiares)	No eligieron estar en el expediente	Mayor protección. No deben ser entidades normalizadas ni indexadas como tales.
Menores de edad	Protección reforzada	Exclusión de cualquier derivado publicado.

[Propuesta] La capa Gold normaliza entidades **sólo para las dos primeras categorías**. Extender el reconocimiento de entidades a todo nombre propio del texto convertiría el corpus en un índice de personas — exactamente el efecto de agregación que el proyecto identifica como riesgo.

4.3 El efecto de agregación

[Inferencia] El riesgo principal no es que un documento sea público —ya lo es— sino que **4.627 documentos dispersos se conviertan en una base consultable**. La búsqueda semántica sobre el corpus completo produce perfiles que no existían cuando los documentos estaban separados. Es un cambio de naturaleza, no de grado.

[Propuesta] Convertirlo en **indicador medible** dentro del índice FAIR ampliado (Fase 9), tal como sugiere JEM.md §3.9: tasa de menciones a terceros no involucrados directamente, y tasa de documentos que contienen categorías de datos sensibles. Medir el riesgo es preferible a declararlo.

5. Publicación de resultados

[Propuesta] Escalonar lo que se publica según su riesgo:

Artefacto	Publicación		Condición
Código, esquemas, scripts de evaluación	Sí, abierta	Sin condición	
Manifiesto e inventario agregado (conteos, distribuciones)	Sí, abierta	Sin condición	
Métricas de calidad (CER/WER, precisión, FAIR)	Sí, abierta	Sin condición	
Banco de preguntas de evaluación	Sí, abierta	Sin datos personales en los enunciados	
Metadatos por documento (tipo, número, fecha, causa)	Sí	Revisión previa de carátulas, que contienen nombres	

Texto extraído por OCR

Decisión abierta

Requiere la Fase 6 cerrada y decisión sobre seudonimización

Entidades y votos estructurados

Decisión abierta

Requiere la Fase 8 cerrada. Ver §6

[Propuesta] **Ninguna cifra sobre votos se publica antes de que la Fase 8 mida su fiabilidad.** JEM.md §11 lo identifica como riesgo reputacional y tiene razón: una estadística de "decisividad" de un integrante del JEM, derivada de un extractor sin auditar, es una afirmación sobre una persona real sostenida por un número que todavía no sabemos si es correcto.

6. Tratamiento de la incertidumbre

[Propuesta] Regla no negociable, heredada de JEM.md §13:

NO_DETERMINABLE nunca se convierte en cero, en ausencia, ni en abstención.

[Inferencia] Es la salvaguarda ética más importante del diseño técnico. Si el extractor no puede determinar el sentido de un voto, la respuesta correcta es "no determinable", no "no votó". La diferencia entre ambas, agregada sobre cientos de resoluciones, es la diferencia entre describir un archivo y difamar a una persona.

[Propuesta] Los estados de determinación (confirmado, probable, ambiguo, no_determinable, no_aplicable) deben sobrevivir a toda agregación posterior: si un gráfico del artículo no puede representar la incertidumbre, se cambia el gráfico, no el dato.

7. Relación con el JEM

[Propuesta]

- El proyecto **no requiere autorización** del JEM para estudiar documentos que el órgano publica por mandato legal, y no debe presentarse como avalado por él.
 - **Notificar al JEM** antes de publicar el artículo es una cortesía institucional razonable y una oportunidad de corrección de errores fácticos. No implica someter los hallazgos a aprobación.
 - Si se detectan **errores en los datos publicados por el JEM** (documentos corruptos, metadatos inconsistentes, documentos faltantes), reportarlos. El proyecto encontró 2 archivos vacíos, 2 inválidos y 1 corrupto; esa información es útil para el órgano.
 - Si se detecta una **exposición indebida de datos personales** en el sitio del JEM, reportarla de forma privada y **no incluirla en el artículo** hasta que el órgano haya podido responder.
-

8. Límites que el equipo se impone

[Propuesta] El proyecto **no hará**, aunque sea técnicamente posible:

1. Publicar rankings de magistrados, integrantes del JEM o cualquier persona.
 2. Entrenar modelos predictivos sobre decisiones o conductas individuales.
 3. Cruzar el corpus con registros externos (redes sociales, padrones, registros públicos de otra naturaleza) para enriquecer perfiles.
 4. Publicar el texto completo de documentos que contengan datos de menores.
 5. Usar el corpus para fines comerciales sin una revisión ética específica.
-

6. Presentar métricas derivadas de una fase no auditada como si fueran hallazgos establecidos.

[Propuesta] Revisar esta lista en la Fase 11 y de nuevo antes de la entrega (Fase 12). Si alguna restricción resultara incompatible con el trabajo, la salida correcta es discutirla y documentar el cambio —no ignorarla en silencio.

Documentos relacionados

- [marco-legal.md](#) — Leyes 5282/2014 y 7593/2025
- [licencia-derivados.md](#) — condiciones de reutilización
- [../../MANIFEST.md](#) — corte vigente y su limitación de procedencia