¿Cómo preparamos nuestros repositorios para un ecosistema transformado por la inteligencia artificial?

Francisco José García-Peñalvo
Grupo GRIAL
Dpto. Informática y Automática
Instituto Universitario de Ciencias de la Educación (https://ror.org/00xnj6419)
Universidad de Salamanca (https://ror.org/02f40zc51), Salamanca, España
fgarcia@usal.es https://orcid.org/0000-0001-9987-5584
http://grial.usal.es
https://twitter.com/frangp

Disponible en:
https://d66z.short.gy/IjjFA0

Repositorios digitales: de la infraestructura al liderazgo de la Ciencia Abierta en América Latina
ICI Academy
11 de noviembre de 2026

Imagen generada con Gemini Nano Banana 2

CC BY NC SA

El repositorio, en la encrucijada

La IA cambia a la vez quién descubre, quién consume y qué carga soporta la infraestructura [1]

Quién descubre

La búsqueda pasa de palabras clave a sistemas que sintetizan la respuesta y pueden ocultar la fuente original

Quién consume

Agentes y modelos leen, fragmentan y reutilizan la colección a una escala que ninguna persona alcanza

Qué carga soporta

El rastreo masivo tensiona servidores pensados para lectores humanos y presupuestos públicos

La meta no es añadir un chatbot. Es lograr un repositorio legible por máquinas, verificable y gobernable

Abierto para leer, no para extraer

EL REFERENTE EUROPEO

Art. 3 CDSM

Minería de textos y datos para investigación científica en organismos de investigación. **No se puede anular por contrato** [2]

Art. 4 CDSM

Minería para cualquier fin, también comercial, **salvo reserva de derechos legible por máquina** (opt-out) [2]

AI Act, art. 53

Los proveedores de modelos de propósito general deben **identificar y respetar esas reservas** [3]

Fuera de la UE, cada país necesita su propio análisis jurídico: la lógica europea sirve como referencia

La paradoja

El mandato de acceso abierto y las licencias permisivas dejan vía libre al entrenamiento comercial si no existe una reserva explícita. Las respuestas generadas omiten la cita y desvían el tráfico [5]

Dos palancas

- **Reserva legible por máquina:** TDMRep, robots.txt, ai.txt
- **Derecho sui generis sobre la base de datos:** protege la colección sin cerrar el acceso documento a documento [4]

El coste oculto de los bots de IA

>90 %

de los 66 repositorios que respondieron a la encuesta de COAR (2025) sufre crawlers agresivos, con ralentizaciones y caídas [7]

Las descargas automatizadas distorsionan las estadísticas COUNTER [9]

Táctica del bot	**Respuesta** [8]
Proxies residenciales: miles de IP domésticas	→ **WAF (*Web Application Firewall*) con análisis de comportamiento**
Búsquedas multifaceta que esquivan la caché	→ **Limitar profundidad y frecuencia de facetas**
User-Agent falsificado como navegador	→ **Verificación ligera en el navegador**
Descargas paralelas de PDF a texto completo	→ **Rate limiting por sesión y descarga masiva aislada**

Bloquear a todos los robots protege el servidor, pero reduce la capacidad de descubrimiento. **Defensa en capas, no cierre** [10]

AI-ready: el orden importa

4 **Capa de IA**
Búsqueda semántica, RAG, curación asistida

3 **Confianza y preservación**
PREMIS, checksums, TRUST, CoreTrustSeal [13]

2 **Interoperabilidad**
OAI-PMH y REST/JSON, schema.org en JSON-LD, Signposting, COAR Notify

1 **Base informacional**
PIDs (ORCID, ROR, DOI), metadatos ricos, licencias legibles por máquina, procedencia y versiones [12]

EVIDENCIA: PILOTO USRN [11]

≈ 1 de cada 2

repositorios del piloto tenía OAI-PMH mal configurado

700.000+

registros pasaron a ser recolectables al mejorar la interoperabilidad

Una capa inteligente sobre una base incompleta amplifica sus defectos

Dos públicos: personas y máquinas

De SEO a GEO

Estructurar la información para que los motores generativos interpreten, resuman y atribuyan bien la fuente [14]

Fuente de verdad para RAG

El repositorio aporta los documentos autenticados que fundamentan la respuesta. No elimina las alucinaciones, pero muestra la evidencia

Semántica explícita

schema.org (ScholarlyArticle, Dataset), malla de PID (*Persistent Identifier*) y licencias legibles por máquina

CURACIÓN ASISTIDA [15]

La IA propone

↓

El repositorio valida

↓

Una persona acepta

Cada metadato declara su procedencia: autor, registro externo, algoritmo o validación humana

AI-ready, no AI-dependent: objetos, metadatos, PID y preservación deben seguir funcionando si cambia o desaparece el proveedor de IA

Gobernanza ética: FAIR + CARE

FAIR para las máquinas, CARE para las comunidades

FAIR [12]

Que los agentes legítimos puedan procesar y reutilizar

F	A	I	R
Findable	Accessible	Interoperable	Reusable

CARE [16]

Que las comunidades conserven el control y el beneficio

C	A	R	E
Collective benefit	Authority to control	Responsibility	Ethics

Riesgo: los agregadores comerciales priorizan escala y normalización frente a minorías, lenguas y saberes locales [18]

Medir el error de la IA por lengua y disciplina, no solo en promedio [17]

Un responsable institucional para cada proceso automatizado



RECOMENDACIONES

Apertura gobernada en cuatro frentes

Jurídico

- La política de depósito aclara que acceso abierto no autoriza el raspado masivo
- Reserva de derechos legible por máquina
- Derecho sui generis sobre la colección

Técnico

- WAF perimetral junto a los servicios de TI
- Rate limiting, baneo progresivo y control de facetas
- Claves API y descarga masiva para usos legítimos

Arquitectura

- Malla de PIDs: ORCID, ROR, DOI
- schema.org en JSON-LD, Signposting, licencias legibles
- COAR Notify y repositorios de nueva generación

Gobernanza

- FAIR + CARE en la curación asistida
- Comité de repositorio e IA con biblioteca, TI, jurídico e investigación
- Acción en red: COAR y redes nacionales y regionales

Por dónde empezar

PRIMEROS PASOS

1. Auditar metadatos, PID, OAI-PMH, licencias, API y preservación
2. Validar la recolección frente a los agregadores
3. Aprobar una política institucional de IA
4. Medir el tráfico de agentes y fijar límites

ANTES DE LLEVAR IA A PRODUCCIÓN

Calidad

Por campo, lengua y disciplina

Coste

Inferencia, supervisión y corrección

Riesgo

Errores de atribución y exposición de datos

Reversibilidad

Apagar el modelo y seguir funcionando

Imagen generada con Gemini Nano Banana 2



Abiertos al conocimiento, no al extractivismo

Repositorios AI-ready, no AI-dependent

¿Cómo preparamos nuestros repositorios para un ecosistema transformado por la inteligencia artificial?

Gracias



Imagen generada con Gemini Nano Banana 2

Referencias



Imagen generada con ChatGPT Plus

Referencias

1. COAR, "AI and Repositories," Confederation of Open Access Repositories. Disponible en: https://coar-repositories.org/what-we-do/ai-and-repositories/.
2. European Parliament y The Council of the European Union, "Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market," Official Journal of the European Union, L 130, 2019. Disponible en: https://eur-lex.europa.eu/eli/dir/2019/790/oj.
3. European Parliament y The Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act)," Official Journal of the European Union, 2024. Disponible en: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
4. "Protecting University Repositories from Aggressive Web Scraping: Using Database Rights to Retain Control over Academic Content," Kluwer Copyright Blog. Disponible en: https://legalblogs.wolterskluwer.com/copyright-blog/protecting-university-repositories-from-aggressive-web-scraping-using-database-rights-to-retain-control-over-academic-content/.
5. "Tensions Between Digital Libraries and Generative AI," Project MUSE. Disponible en: https://muse.jhu.edu/article/992004.
6. "Generative AI and intellectual property: a systematic literature review of possible issues and mitigations." Disponible en: https://www.researchgate.net/publication/408197294.
7. COAR, "Report of the COAR Survey on AI Bots," June 2025. Disponible en: https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf.
8. "Mitigating Aggressive Crawler Traffic in the Age of Generative AI," The Code4Lib Journal. Disponible en: https://journal.code4lib.org/articles/18489.
9. LibLynx, "Bot wars: Will AI scraping change the Scholarly landscape forever?" Disponible en: https://www.liblynx.com/news/bot-wars-will-ai-scraping-change-the-scholarly-landscape-forever.
10. COAR, "Mitigating the impact of AI bots." Disponible en: https://coar-repositories.org/news-updates/mitigating-the-impact-of-ai-bots/.
11. COAR, "USRN Discovery Pilot: Outcomes and Next Steps." Disponible en: https://coar-repositories.org/news-updates/usrn-discovery-pilot-outcomes-and-next-steps/.
12. M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, art. 160018, 2016. doi: 10.1038/sdata.2016.18.
13. D. Lin et al., "The TRUST Principles for digital repositories," Scientific Data, vol. 7, art. 144, 2020. doi: 10.1038/s41597-020-0486-7.
14. "Generative Engine Optimization in digital repositories," Infonomy. Disponible en: https://infonomy.scimagoepi.com/index.php/infonomy/en/article/view/115.
15. Digital Preservation Coalition, "AI-powered metadata workflows at the ICAEW Digital Archive." Disponible en: https://www.dpconline.org/blog/ai-powered-metadata-workflows-at-icaew-craig-mccarthy.
16. S. R. Carroll et al., "The CARE Principles for Indigenous Data Governance," Data Science Journal, vol. 19, art. 43, 2020. doi: 10.5334/dsj-2020-043.
17. ESIP, "Are Your Data Ready? Take Stock With ESIP's New AI-Ready Checklist." Disponible en: https://www.esipfed.org/checklist-ai-ready-data/.
18. "Bridging FAIR and CARE in ETD Metadata: An LLM assisted Cross [...]." Disponible en: https://dcpapers.dublincore.org/article/952576294.

Cita recomendada

F. J. García-Peñalvo, "¿Cómo preparamos nuestros repositorios para un ecosistema transformado por la inteligencia artificial?," presentado en Repositorios digitales: de la infraestructura al liderazgo de la Ciencia Abierta en América Latina (ICI Academy, 11 de noviembre de 2026). Salamanca, España: Grupo GRIAL, 2026. Disponible: https://d66z.short.gy/IjjFA0. doi: 10.5281/zenodo.24136235.

¿Cómo preparamos nuestros repositorios para un ecosistema transformado por la inteligencia artificial?

Francisco José García-Peñalvo
Grupo GRIAL
Dpto. Informática y Automática
Instituto Universitario de Ciencias de la Educación (https://ror.org/00xnj6419)
Universidad de Salamanca (https://ror.org/02f40zc51), Salamanca, España
fgarcia@usal.es https://orcid.org/0000-0001-9987-5584
http://grial.usal.es
https://twitter.com/frangp

Disponible en:
https://d66z.short.gy/IjjFA0

Repositorios digitales: de la infraestructura al liderazgo de la Ciencia Abierta en América Latina
ICI Academy
11 de noviembre de 2026

Imagen generada con Gemini Nano Banana 2

CC BY NC SA