



I Seminario Internacional de Biblioteca Académicas y
Especializadas
"Retos para el cambio social y transformación digital"

Repositorios digitales peruanos: retos y oportunidades en el marco de la inteligencia artificial

Libio Huaroto ([0000-0003-4667-0993](tel:0000-0003-4667-0993))
Universidad Peruana de Ciencias Aplicadas (UPC)
Lima, Perú

Noviembre 15, 2023



Temario

- Situación actual de repositorio
- Servicios de información IA
- Aplicaciones IA en repositorios
- Comentarios

Repositorios en Perú I

- Uso del software **DSpace 5, 6 y 7** para RI.
- **DSpaceCRIS** y **PURE** para CRIS (#PeruCRIS).
- **Recolectores** (ALICIA y RENATI).
- Uso de **Dataverse** para repositorio de datos.
- Uso de **REST API** - integración con sistemas internos.
- Sistemas de **geolocalización**.
- Integración **Altmetrics**.
- Incremento servicios en la **nube** para RI.

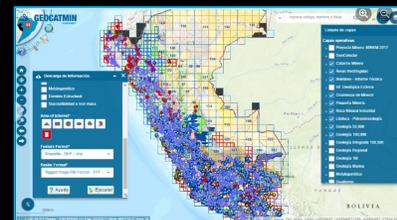


API REST



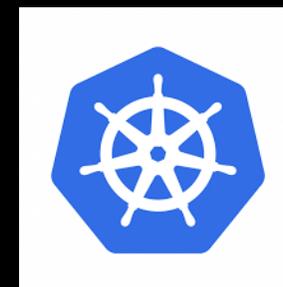
Repositorios en Perú II

- **Dashboard de métricas** para RI y Recolectores (Big Data).
- **Reportadores RI**
- Arquitectura de **contenedores** en RI.
- Integración **identificadores ORCID**.
- Aplicación de **estándar RDF**.
- Aplicación de **datos estructurados** (Schema.org).
- **Linked data** (DSpaceCris).
- **Apps RI** (móvil).



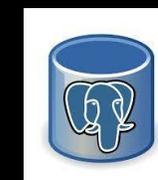
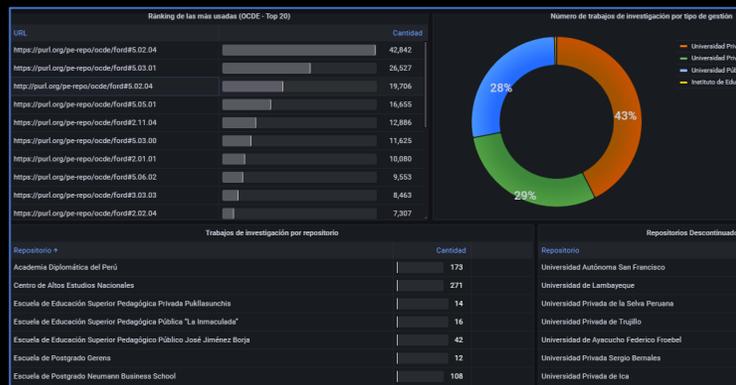
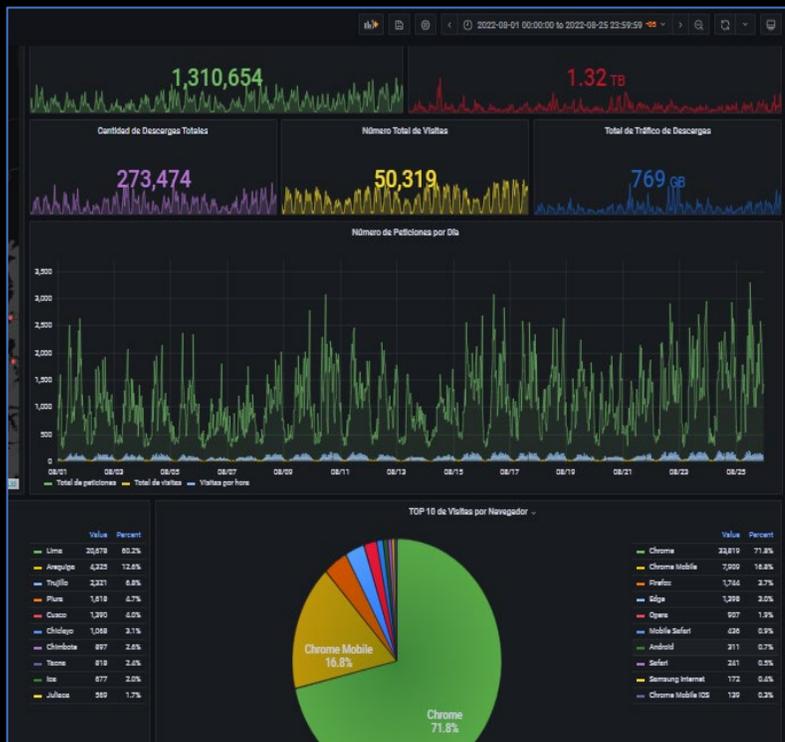
ORCID

isni



Repositorios en Perú III

Dashboard para Recolectores y RI

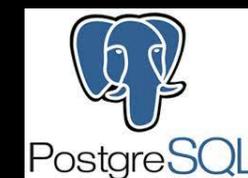
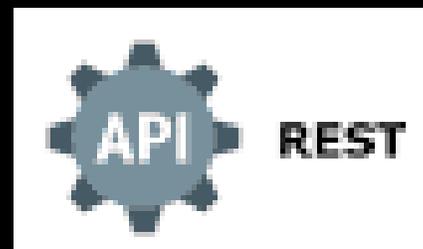


Repositorios en Perú IV

Apps RI

Servicios de búsqueda

Dashboard para plataformas OJS



Servicios de inteligencia artificial

 **OpenAI** 
@OpenAI · [Seguir](#) 

Announcing ChatGPT for Android! The app will be rolling out to users next week, and you can pre-order in the Google Play Store starting today:



play.google.com
ChatGPT - Apps on Google Play
With ChatGPT, find instant answers, professional input, and creative inspiration

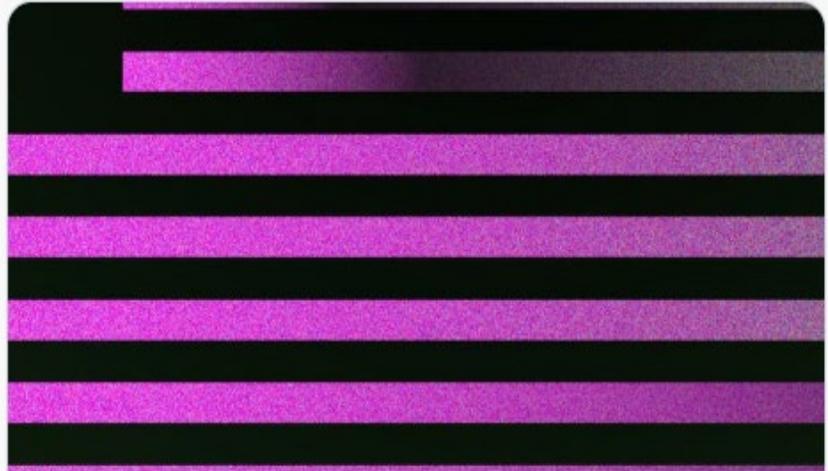
2:59 p. m. · 21 jul. 2023 

 8 mil  Responder  Compartir

[Leer 534 respuestas](#)

 **OpenAI**  @OpenAI · 30 nov. 2022 

Try talking with ChatGPT, our new AI system which is optimized for dialogue. Your feedback will help us improve it.



openai.com
Introducing ChatGPT
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ...

 1.315  4.413  13,5 mil 

Servicios de inteligencia artificial

2023 LifeArchitect.ai data (shared) ☆ 📄 🌐

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda

Menús 🖨️ 📄 100% Solo comentarios

A1 Permalink:

1	Permalink:	Timeline view: https://lifecycle.ai/timeline					The Memo: https://lifecycle.ai/memo				
2	Model	Lab	Playground	Parameters (B) trained (B)	Tokens (B)	Ratio Tokens:Params (Chinchilla scaling20:1)	ALScore "ALScore" is Sqr Root of Training dataset	Announced	Public?	Paper / Repo	Notes
3	Meta GPT	Meta AI	https://www.theverge.c	2000			W 📄 🌟 🧑	TBA		https://arch Training in Q1 2024 on H100s. No projec	
4	Ajax GPT	Apple	https://www.theverge.c	200			W 📄 🌟 🧑	TBA		Internal use.	
5	GPT-5	OpenAI	https://lifecycle.ai/g	2000			W 📄 🌟 🧑	TBA		Due mid-2024.	
6	Claude-Next	Anthropic	https://techcrunch.com	2000			W 📄 🌟 🧑	TBA		2T params estimated	
7	OLMo	Allen AI	https://allenai.org/olmc	70			W 📄 🌟 🧑	TBA		Open source. Due Q1 2024.	
8	G3PO	OpenAI	https://lifecycle.sub	70			W 📄 🌟 🧑	TBA		Open source version, probably 70B para	
9	Arrakis (GPT-4.5?)	OpenAI	https://lifecycle.ai/gpt-5/							True multimodal (anything in > anything	
10	Gobi (GPT-5?)	OpenAI	https://lifecycle.ai/gpt-5/								
11	Gemini	Google Dee...	https://lifecycle.ai/g	8000	65000	9:1	76.0 W 📄 🌟 🧑	TBA		https://arxiv Due Dec/2023. 8T params estimated	
12	jina-embeddings-v2	Jina AI	https://huggingface.co/	0.435			🌟	Oct/2023	🟢	https://jina Alternative to text-embedding-ada-002.	
13	Fuyu	Adept	https://huggingface.co/	8			W 📄 🌟 🧑	Oct/2023	🟢	https://www VLM. 8B available under open licence, N	
14	ERNIE 4.0	Baidu		1000			W 📄 🌟 🧑	Oct/2023	🟡	https://reut No architecture information. English-dul	
15	Zephyr	Hugging Face H4	https://huggingface.co/	7.3	800	110:1	0.3 W 📄 🌟 🧑	Oct/2023	🟢	https://hug Mistral with 'aligned' data removed from	
16	PaLI-3	Google Dee...		5			🧑	Oct/2023	🔴	https://arxiv VLM. Next iteration of PaLI via Pathways	
17	NVIDIA	Retro 48B		48	1200	25:1	0.8 W 📄 🌟 🧑	Oct/2023	🟢	https://arxiv the largest LLM pretrained with retrieva	
18	Lemur	XLANG Lab	https://github.com/Ope	70	2090	30:1	1.3 W 📄 🌟 🧑	Oct/2023	🟢	https://arxiv Llama 2 + The Stack (code)	
19	AceGPT	KAUST/Shenzhen	https://huggingface.co/	13	2010	155:1	0.5 W 📄 🌟 🧑	Oct/2023	🟢	https://gith Arabic. Llama 2 + RLAI	
20	Yasa-1	Reka AI	https://reka.ai/announcing-our-multimodal-ai-assistant/				🧑	Oct/2023	🟡	https://reka Multi-modal. No public arch info. Resear	

Servicios de biblioteca basados en inteligencia artificial

Table 1. Use cases of AI in libraries.

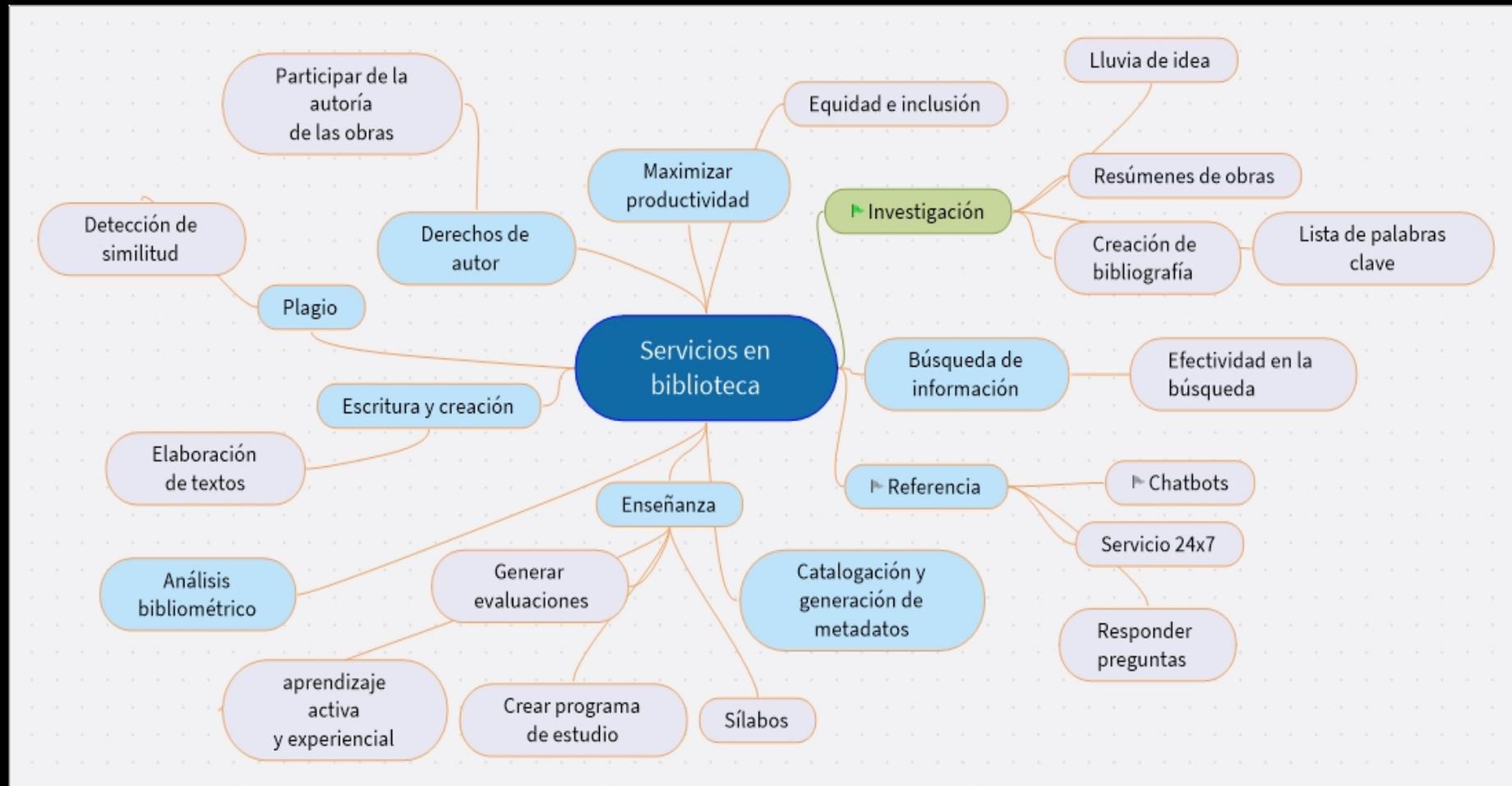
A. AI use cases	B. Area of library operation	C. Skills and knowledge required	D. Drivers	E. Barriers	F. Likelihood
1. Library back-end processes: RPA	Routine administration	Programming and RPA tools Workflow analysis	Efficiency	Benefits of RPA marginal	High in all libraries

Defining artificial intelligence for librarians

Andrew M. Cox; Suvodeep Mazumdar (2022)

1. Library back-end processes: ASRS/ other robotic solutions	Manual tasks	Normal manual implementation	2. Library services: Chatbots and digital assistants	Reference queries and other user interactions	Building knowledgebase Understanding likely navigation paths of information requests	24/7, consistent service	No off shelf solutions yet Equal access could be hard to achieve Dehumanisation Staff resistance	High – in mass institutions
2. Library services for users: Knowledge discovery	Collection management	Understand the project Engineering Statistical Project	3. Data scientist community creation – as lead or as participant	User liaison	Community building skills Relevant knowledge about finding, licensing and preserving data	Need for support on data location, data preservation	Alternative leadership	Very high in academic institutions – especially if have rich collections for digital humanities.
			4. Data and AI literacy	Information literacy and other user training	Pedagogy General knowledge of data and AI	Risks around uses of AI in wider society – ethics as driver	Library role not clear	Very high – in all libraries
			5. User management	Strategy and planning	Statistics Computational methods	To build user need driven services	Ethics – as barrier	High in all

Servicios de biblioteca basados en inteligencia artificial



Fuentes de información científica - IA

AI SCIENCE SEARCH ENGINES EXPLODE IN NUMBER

Tools powered by large language models are intended to help researchers digest or do science.

By Katharine Sanderson

As large language models (LLMs) gallop ever onwards – among them GPT-4, OpenAI's latest incarnation of the technology behind ChatGPT – scientists are beginning to make use of their power. The explosion of tools powered by artificial intelligence (AI) includes several search engines that aim to make it easier for researchers to grasp seminal scientific papers or summarize a field's major findings. Their developers claim the apps will democratize and streamline access to research.

But some tools need more refinement before researchers can use them to help their studies, say scientists who have experimented with them. Clémentine Fourrier is a Paris-based researcher who evaluates LLMs at Hugging Face, a company in New York City that develops open-source AI platforms. She used an AI search engine called Elicit, which uses an LLM to craft its answers, to help find papers for her PhD thesis. Elicit searches papers in the

first go-to search for academic search," he says. "In terms of relevancy, I had the opposite experience [to Fourrier] with Elicit. I normally get roughly the same relevancy as Google Scholar – but once in a while, it interprets my search query better."

These discrepancies might be field-dependent, Tay suggests. Fourrier adds that, in her research area, time is crucial. "A year in machine learning is a century in any other field," she says. "Anything prior to five years is completely irrelevant," and Elicit doesn't pick up on this, she adds.

Full-text search
Another tool, scite, whose developers are based in New York City, uses an LLM to organize and add context to paper citations – including where, when and how one paper is cited by

"A year in machine learning is a century in any other field."

Nature | Vol 616 | 27 April 2023

ARTIFICIAL-INTELLIGENCE SEARCH WRANGLES LITERATURE

Developers want to free scientists to focus on discovery and innovation by helping them to draw connections from a massive body of publications. By Amanda Heidt

Work / Technology & tools

Collectively, the platforms facilitate many of the early steps in the writing process. Critics risk, however, that the programs remain not actively untested and run the risk of perpetuating existing biases in the academic publishing process.

The teams behind these tools say they built them to connect information overload and to pull connections from the tangled web of the scientific literature. Some programs, such as Consensus, give researchers backed answers to yes or no questions, such as Semantic Scholar, Elicit and InSight, and as digital assistants – by using up bibliographies, suggesting new papers and generating research summaries.

Collectively, the platforms facilitate many of the early steps in the writing process. Critics risk, however, that the programs remain not actively untested and run the risk of perpetuating existing biases in the academic publishing process.

The teams behind these tools say they built them to connect information overload and to pull connections from the tangled web of the scientific literature. Some programs, such as Consensus, give researchers backed answers to yes or no questions, such as Semantic Scholar, Elicit and InSight, and as digital assistants – by using up bibliographies, suggesting new papers and generating research summaries.

Nature | Vol 620 | 10 August 2023

<https://doi.org/10.1038/s41591-023-1111-1>

Large language models should be used as scientific reasoning engines, not knowledge databases

Large language models (LLMs) such as GPT-4 have emerged into the public view in recent months. These artificial intelligence (AI) models boast impressive capabilities – and can analyze and write texts with human-like performance, including for medical and scientific applications – and can be used to respond to students assignments and to scientific and medical examination questions, and to create text for research proposals or scientific publications. Yet LLMs often fall short when

information from the Internet. Alternatively, the LLM can be provided with detailed text containing specific information, for example, as part of the prompt, which the LLM then interprets, compares and uses as context to produce its output. Providing such contextual information to the input prompt enables what is called 'in-context learning'. Scientific texts (abstracts or short full-text articles), treatment guidelines developed by professional societies or regulatory guidelines can be included in the prompts verbatim before a

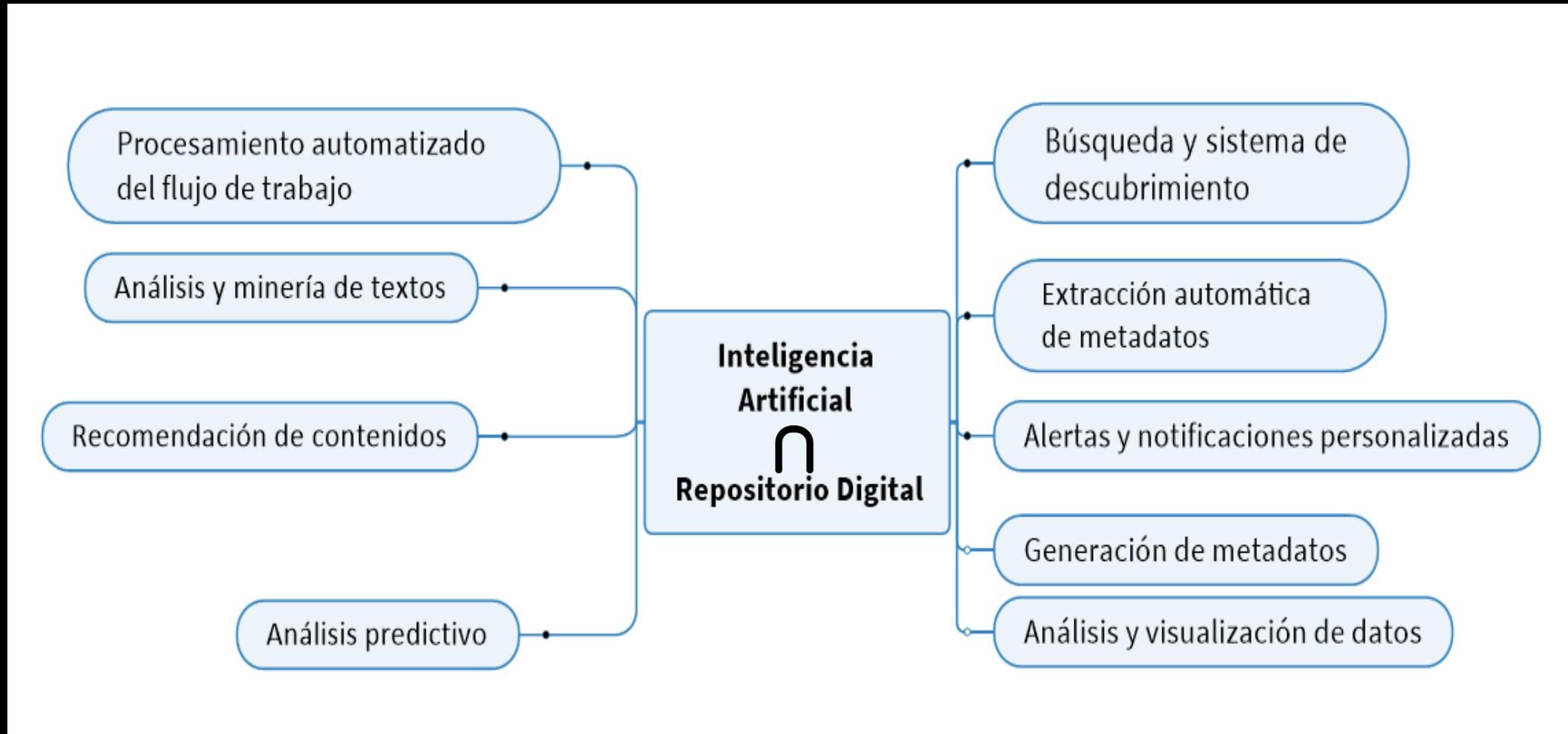
appropriate prompts. GPT-4 yields and specific suggestions for additional trials, including details of sampling, disease severity and patient performance. This ability to evaluate and interpret data, generate new hypotheses and suggest pertinent experiment time-saving for clinical trials.

The deployment of LLMs as and reasoning systems in medicine holds potential value. Rather than ing with conventional informati

Nature Human Behaviour volume 7



Inteligencia artificial – Repositorios digitales



Inteligencia artificial – repositorios



Motor de búsqueda impulsado por **inteligencia artificial** que utiliza algoritmos de aprendizaje automático para extraer información clave de artículos de investigación.

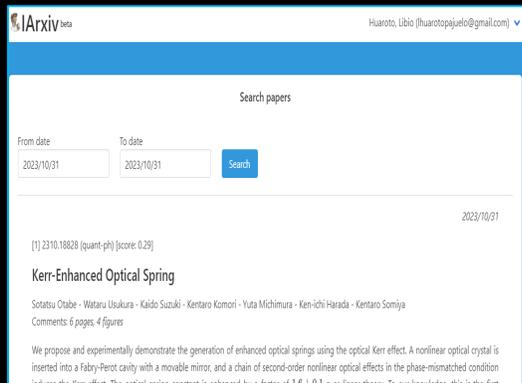


Servicio de IA que ofrece recomendaciones de artículos de repositorios institucionales. Utiliza algoritmos de aprendizaje automático para analizar el comportamiento de los usuarios, patrones de citas, entre otros.



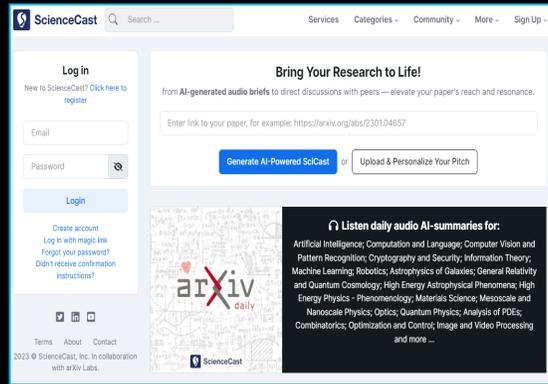
Plataforma de información sobre investigación que emplea **técnicas de IA** para analizar y curar los resultados de investigación de los repositorios institucionales.

Inteligencia artificial – Repositorios digitales



SCIENCE CAST

Plataforma de acceso libre, mediante una IA genera los resúmenes en audio de los artículos publicados en arXiv.



iArxiv

A.I. customized arXiv email notifications

Organiza bibliografía científica según las preferencias de cada usuario-



<https://arxiv.org/>

Repositórios Latinoamericanos – BIREDIAL 2023

- CRIAÇÃO DE UM **CHATBOT** PARA O ARCA - REPOSITÓRIO INSTITUCIONAL DA FIOCRUZ Brasil - Outubro 2023
- Desenvolvido com framework RASA.
- RASA, software livre que utiliza linguagem de Python, permite criar chatbots com capacidades de processamento de linguagem natural y aprendizagem automático.

The image displays two overlapping screenshots. The top screenshot shows a presentation slide titled "CRIAÇÃO DE UM CHATBOT PARA O ARCA - Repositório Institucional da Fiocruz" with the "CIÊNCIA BERTA" logo and authors Valéria Machado da Costa and Cláudia Fernandes de Queiroz. The bottom screenshot shows the ARCA website interface with a chatbot window open. The chatbot window, titled "Wal", displays a message: "Arca - Repositório Institucional da Fiocruz visa aumentar a visibilidade a produção intelectual da Instituição. Seguimos também a política de acesso aberto da instituição. Em que posso te ajudar?". Below the message are several blue buttons: "Software utilizado", "Comunidades do repositório", "Tipos de documentos", "Depósito de documentos", "Pesquisa de documentos", and "Política de Acesso Aberto". The website background shows a search bar and a list of items.

Inteligencia artificial – Análisis de datos para CRIS

Formamos líderes responsables para el mundo

"Experiencias de Inteligencia Artificial en Bibliotecas Universitarias"

Uso de ChatGPT en la determinación de la relación de las investigaciones de la UP con los Objetivos de Desarrollo Sostenible (ODS)

up.edu.pe



"Experiencias de Inteligencia Artificial en Bibliotecas Universitarias"

Diseño de los prompts de aprendizaje y análisis ODS

UNIVERSIDAD DEL PACÍFICO

Proceso de entrenamiento:

Se le instruyó a ChatGPT que debe codificar el resultado de su análisis.

1. ODS1: Fin de la pobreza
2. ODS2: Hambre cero
3. ODS3: Salud y bienestar
4. ODS4: Educación de calidad
5. ODS5: Igualdad de género
6. ODS6: Agua limpia y saneamiento

Ahora vamos a asignar un código único a cada uno de los objetivos de la lista. El primer objetivo será ODS1, el segundo será ODS2 y así sucesivamente hasta el último objetivo.

iPerfecto! A continuación, te muestro la lista de los Objetivos de Desarrollo Sostenible (ODS) con su respectivo código único asignado:

1. ODS1: Fin de la pobreza
2. ODS2: Hambre cero
3. ODS3: Salud y bienestar
4. ODS4: Educación de calidad
5. ODS5: Igualdad de género
6. ODS6: Agua limpia y saneamiento
7. ODS7: Energía asequible y no contaminante
8. ODS8: Trabajo decente y crecimiento económico
9. ODS9: Industria, innovación e infraestructura
10. ODS10: Reducción de las desigualdades
11. ODS11: Ciudades y comunidades sostenibles
12. ODS12: Producción y consumo responsables
13. ODS13: Acción por el clima
14. ODS14: Vida submarina
15. ODS15: Vida de ecosistemas terrestres
16. ODS16: Paz, justicia e instituciones sólidas
17. ODS17: Alianzas para lograr los objetivos

Estos códigos te permitirán identificar de manera única cada uno de los Objetivos de Desarrollo Sostenible.

up.edu.pe

Inteligencia artificial – Generación de metadatos keywords

SOIL, 9, 155–168, 2023
 https://doi.org/10.5194/soil-9-155-2023
 © Author(s) 2023. This work is distributed under the Creative Commons Attribution 4.0 License.

SOIL 

Potential of natural language processing for metadata extraction from environmental scientific publications

Guillaume Blanchy¹, Lukas Albrecht², John Koestel^{2,3}, and Sarah Garre¹

¹Department of Plant, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium
²Soil Fertility and Soil Protection, Agroscope, Reckenholzstrasse 191, 8046, Zurich, Switzerland
³Department of Soil and Environment, Institute for Soil and Environment, Swedish University of Agricultural Sciences, Box 7014, 75007 Uppsala, Sweden

Correspondence: Guillaume Blanchy (guillaume.blanchy@ilvo.vlaanderen.be)

Received: 23 June 2022 – Discussion started: 5 July 2022
 Revised: 27 January 2023 – Accepted: 3 February 2023 – Published: 14 March 2023

Abstract. Summarizing information from large bodies of scientific literature is an essential but work-intensive task. This is especially true in environmental studies where multiple factors (e.g., soil, climate, vegetation) can contribute to the effects observed. Meta-analyses, studies that quantitatively summarize findings of a large body of literature, rely on manually curated databases built upon primary publications. However, given the increasing amount of literature, this manual work is likely to require more and more effort in the future. Natural language processing (NLP) facilitates this task, but it is not clear yet to which extent the extraction process is reliable or complete. In this work, we explore three NLP techniques that can help support this task: topic



LH Hello, suggest 10 keyword for this abstract:

 Certainly, generating appropriate keywords for your abstract can assist in enhancing its visibility and accessibility in academic databases. Here are 10 keywords that encapsulate the essential themes and methodologies presented in your work:

- 1. Meta-Analysis:** The study hinges on the quantitative summarization of large bodies of scientific literature, a classic meta-analytic approach.
- 2. Environmental Studies:** Your research specifically targets environmental studies, focusing on factors like soil, climate, and vegetation.
- 3. Natural Language Processing (NLP):** A core methodology employed, particularly relevant

REPOSITORIO ACADÉMICO UPC

INICIO

Listar

Todos los contenidos

Búsquedas

Contenidos

Elija una de las categorías para ver sus contenidos

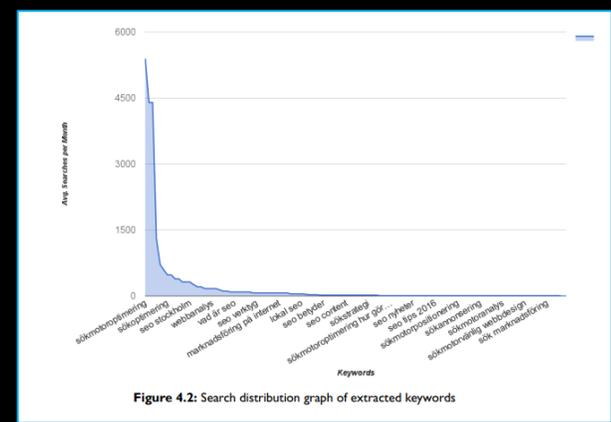
Artículos científicos [1889]
 Sección que corresponde a la producción científica registrada en bases de datos internacionales y que corresponden a los miembros de la universidad.

Artículos académicos [278]
 Artículos publicados en revistas académicas cuya afiliación corresponde a la universidad.

LH suggest 05 descriptors for this abstract

- 1. Quantitative Meta-Analysis in Environmental Science:** This descriptor encapsulates the methodological framework and subject matter, highlighting the quantitative approach to meta-analysis specifically within the realm of environmental studies.
- 2. NLP-Assisted Data Extraction and Synthesis:** This focuses on the technological innovation introduced in the study, specifically the use of Natural Language Processing for data extraction and synthesis in scientific literature.
- 3. Soil and Water Management Variables:** Given that the study delves into factors such as hydraulic conductivity and the impact of agricultural practices on sustainable water management, this descriptor is pertinent for researchers interested in these specific variables.
- 4. Topic Modeling and Relationship Extraction:** This descriptor captures the specialized NLP techniques like topic modeling and the shortest dependency path method that are utilized for text mining in your research.

Long Tail keywords

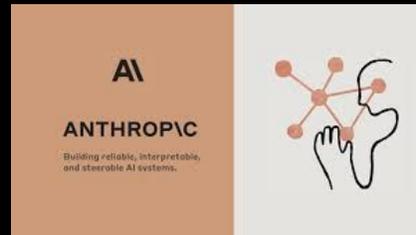


Política de posicionamiento web (Webometrics)

Inteligencia artificial – Extracción de metadatos

Extracción de metadatos

1. Recopilación de información
2. Identificación de metadatos
3. Extraer texto
4. Limpiar texto
5. Organización metadatos
6. Formato de salida
7. Mejorar (salida)



bring information about dublin core metadata standar

LH

according to this information, extract the metadata from this file

LH

Here is the Dublin Core metadata extracted from the PDF file:

Title: High altitudes, population density, and poverty: Unraveling the complexities of COVID-19 in Peru during the years 2020–2022

Creator: David A. Vizcardo, Jorge R. Aranibar, César Vladimir Munayco Escate

Subject: COVID-19, Peru, altitude, population density, poverty, SARS-CoV-2

Description: This ecological study analyzed associations between altitude, population density, poverty and COVID-19 incidence and case fatality rate across 196 provinces in Peru from 2020-2022 using government data.

Publisher: Elsevier Inc.

Contributor: Preventive Medicine Reports

Date: 2023

Type: Text

Format: PDF

Identifier: <https://doi.org/10.1016/j.pmedr.2023.102423>

Based on the Dublin Core metadata extracted from the PDF research article, here are the corresponding Dublin Core elements and tags:

```
dc:title
High altitudes, population density, and poverty: Unraveling the complexities of COVID-19 in Peru during the years 2020–2022
</dc:title>
```

```
dc:creator
David A. Vizcardo, Jorge R. Aranibar, César Vladimir Munayco Escate
</dc:creator>
```

```
dc:subject
COVID-19, Peru, altitude, population density, poverty, SARS-CoV-2
</dc:subject>
```

```
dc:description
This ecological study analyzed associations between altitude, population density, poverty and COVID-19 incidence and case fatality rate across 196 provinces in Peru from 2020-2022 using government data.
</dc:description>
```

```
dc:publisher
Elsevier Inc.
</dc:publisher>
```

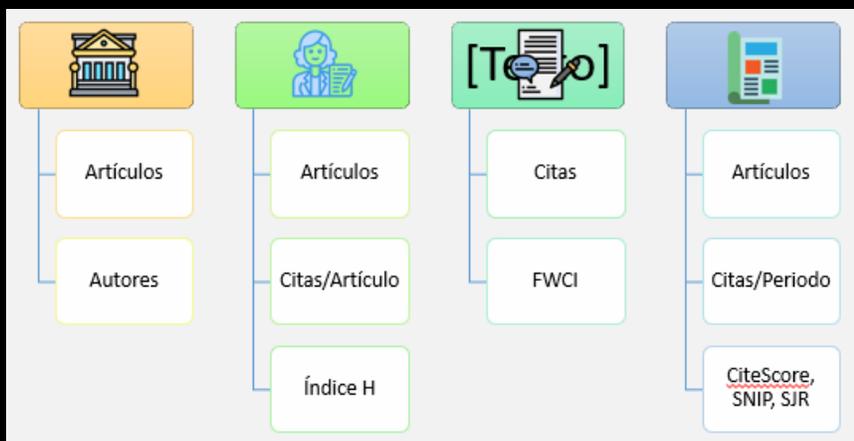
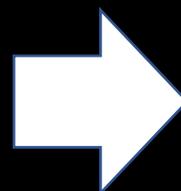


Inteligencia artificial – Análisis de datos

- Analizar la producción investigadora
- Análisis de citas
- Métricas de autoría
- Clasificación e impacto de las revistas
- Evaluación comparativa:



Universidad 1
Top 10 journals order by citation



Rank	Author	Total Citations
1	Yáñez J.A., Jahanshahi A.A., Alvarez-Risco A.,...	125.0
2	Quispe-Cañari J.F., Fidel-Rosales E., Manrique...	114.0
3	Awoyera P.O., Kirgiz M.S., Viloría A., Ovallos...	98.0
4	Quesada A.E., Medeiros L.J., Clemens M.W., Fer...	85.0
5	White C.M., Pasupuleti V., Roman Y.M., Li Y., ...	65.0
6	Benites-Zapata V.A., Ulloque-Badaracco J.R., A...	64.0
7	Alvarez-Risco A., Del-Aguila-Arcentalés S., Yá...	58.0
8	Palanisamy M., Kolandasamy P., Awoyera P., Gob...	55.0
9	Jeri-Yabar A., Sanchez-Carbonel A., Tito K., R...	52.0
10	Piscoya A., Ng-Sueng L.F., del Riego A.P., Cer...	51.0

Comentarios

- Promover y participar en espacios de discusión y formación sobre los temas de la inteligencia artificial.
- Adoptar y promover las recomendaciones de UNESCO en favor de la ciencias abierta - repositorios de datos.
- Fortalecer las competencias del equipo de trabajo en biblioteca para los aspectos de IA.
- Adopción de políticas que involucren aspectos éticos, servicios de información IA, entre otros.

Muchas gracias