

ISIDORE 2030 (eng. version)

ISIDORE 2030 is a research and engineering program in the humanities and social sciences (SHS) aimed at renewing the functionalities of the academic search engine and assistant isidore.science. Launched in 2010, isidore.science, like other discovery tools, is in constant evolution. The proliferation of generative and selective artificial intelligence (AI) technologies significantly impacts research instruments, which must evolve while remaining functional for the communities they serve. After 13 years of indexing and document enrichment, isidore.science has aged and is at a crossroads: how to renew it and orient it while continuing to exploit it? The intervention will outline the main projects of ISIDORE 2030 while reviewing the steps taken in recent years.

AI in ISIDORE: From Classification to *Retrieval Augmented Generation*

ISIDORE has been using AI since 2010, primarily for classifying, categorizing, and enriching data and metadata. Developed with Antidot SA, the various AI technologies have evolved and enabled the Huma-Num team to gain solid experience in AI for SHS (Silvestre de Sacy et al., 2024).

Since 2023, within the HN Lab of Huma-Num, several initiatives have been launched to incorporate weighted generative AI into ISIDORE. This involves conceptualizing and implementing the redesign of the search engine by 2030. The aim is to improve the current functionalities of the engine while integrating the latest technologies offered by the advent and democratization of large language models (Pouyllau et al., 2024).

In this context, the HN Lab team has been working for several months on a specific functionality, RAG (for *Retrieval Augmented Generation*). RAG is an innovative technique that combines natural language generation and information retrieval capabilities to enhance the performance and interpretability of large language models. This technique attempts to minimize their black box aspects by relying on a specific corpus provided by the user to generate responses (Silvestre de Sacy et al., 2024).

When a query is submitted, the system retrieves relevant information from the user-provided database, combines it with the user's query, and then passes the augmented query to the Large Language Models (LLMs) to generate a context-based response.

This approach aims to minimize issues such as hallucinations, outdated knowledge, and reasoning limitations in LLMs. By integrating external knowledge sources, RAG improves the overall quality, accuracy of LLM responses, and provides interpretative context to the response while keeping the data corpus up-to-date and enrichable (Maronet & Truc, 2024).

Expected Functionalities

- Content analysis and corpus alignment;
- Dashboards for creating state-of-the-art proposals on scientific questions;
- Automatic summaries and document syntheses;
- Translation and quality analysis of translations (rare, ancient languages, etc.);
- Exploration of scientific communities;
- Detection of emerging themes and communities;
- Scientific monitoring;
- Detection of long-tail themes;
- Qualitative improvement of metadata;

• ...

Bibliography

Silvestre de Sacy, A., Faci, A., Pouyllau, S., & Maronet, L. (2024, October 18). Pre-targeted-RAG - Retrieval Augmented Generation on pre-targeted groups of research article communities. ColDoc, Université Paris-Nanterre. Zenodo. <https://doi.org/10.5281/zenodo.13950650>

Pouyllau, S. (2024, October 4). ISIDORE 2030: Adapting AI to the needs of document and data research in SHS. Conference at GF2i (GF2i), PARIS. HN LAB. <https://doi.org/10.5281/zenodo.13892964>

Pouyllau, S., Silvestre de Sacy, A., Maronet, L., & Faci, A. (2024). Capitalizing on experience to experiment and innovate: feedback and reflection on the future of the Huma-Num research infrastructure. (1.0). Digital Humanities in the Nordic and Baltic Countries (DHNB), Reykjavik. HN Lab. <https://doi.org/10.5281/zenodo.13889742>

Maronet, L., & Truc, A. (2024). Improving workflows in digital art history: sharing annotations for patrimonial images segmentation and object detection. Transformations, A DARIAH Journal, 1. <https://doi.org/10.5281/zenodo.13947909>

Maronet, L., & Truc, A. (2024, June 16). Improving workflows in digital art history: the usefulness of patrimonial images segmentation. Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities, Lisbon, Portugal. Zenodo. <https://doi.org/10.5281/zenodo.11863661>

Silvestre de Sacy, A., Faci, A., Maronet, L., & Pouyllau, S. (2024). Note on the experience of AI within the Huma-Num Lab (huma-num version) (1.1). ACFAS 2024 (ACFAS), Ottawa. HN Lab. <https://doi.org/10.5281/zenodo.10846773>

Additional References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Lewis, P., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Stoyanov, V. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

These additional references provide a theoretical and practical context for natural language generation and information retrieval techniques, as well as the language models used in the ISIDORE 2030 project.