Developing a Smart Archival Assistant with Conversational Features and Linguistic Abilities

THE ASK_ARCHILAB INITIATIVE

Prof Dr. Basma Makhlouf Shabou
Head of Master Programme Information Sciences

Dr. Lamia Friha
Resp. Cellule RD, University of Geneva

Wassila Ramli
Corporate analytics and digital project officer, Contexa

Agenda

01
Context/Problem

02
Research Goals

03
Technical Design

04
Implementation

05
Validation

06
Future Direction

07
Conclusion

Get
Started →

ARCHiLAB

h e g
Haute école de gestion
Genève

Hes·SO GENÈVE
Haute Ecole Spécialisée
de Suisse occidentale

UNIVERSITÉ
DE GENÈVE

InterPARES Trust AI

PIAF PORTAIL INTERNATIONAL ARCHIVISTIQUE FRANCOPHONE

ICA INTERNATIONAL COUNCIL ON ARCHIVES
CONSEIL INTERNATIONAL DES ARCHIVES

01 — CONTEXT/PROBLEM

Digital transformation has fundamentally changed how **archival knowledge** is managed, accessed, and shared.

01 — CONTEXT/PROBLEM

Digital transformation has fundamentally changed how **archival knowledge** is managed, accessed, and shared.

The Challenge

Contemporary archival management faces unprecedented challenges due to exponential growth of digital records and complex data formats.

Financial
Limited budgets for standards and training

Technical
Diverse formats, legacy system integration

Cognitive
Complex terminology across languages

Academic & professional
Restricted access to practices and innovative research

Sociocultural
Varying interpretations of principles

Explore User Needs →

02 — RESEARCH GOALS

Project Objectives

1 Develop Advanced Conversational AI

Multi-agent systems and generative AI for natural interactions by automated multilingual processing

2 Fast and Accurate Search

Within archives and archival resources by intelligent context-aware search

3 Contextual Responses

Relevant answers grounded in archival knowledge through personalized user experiences

4 Open Archival Knowledge

Accessible to wide range of users using a cross-collection knowledge discovery

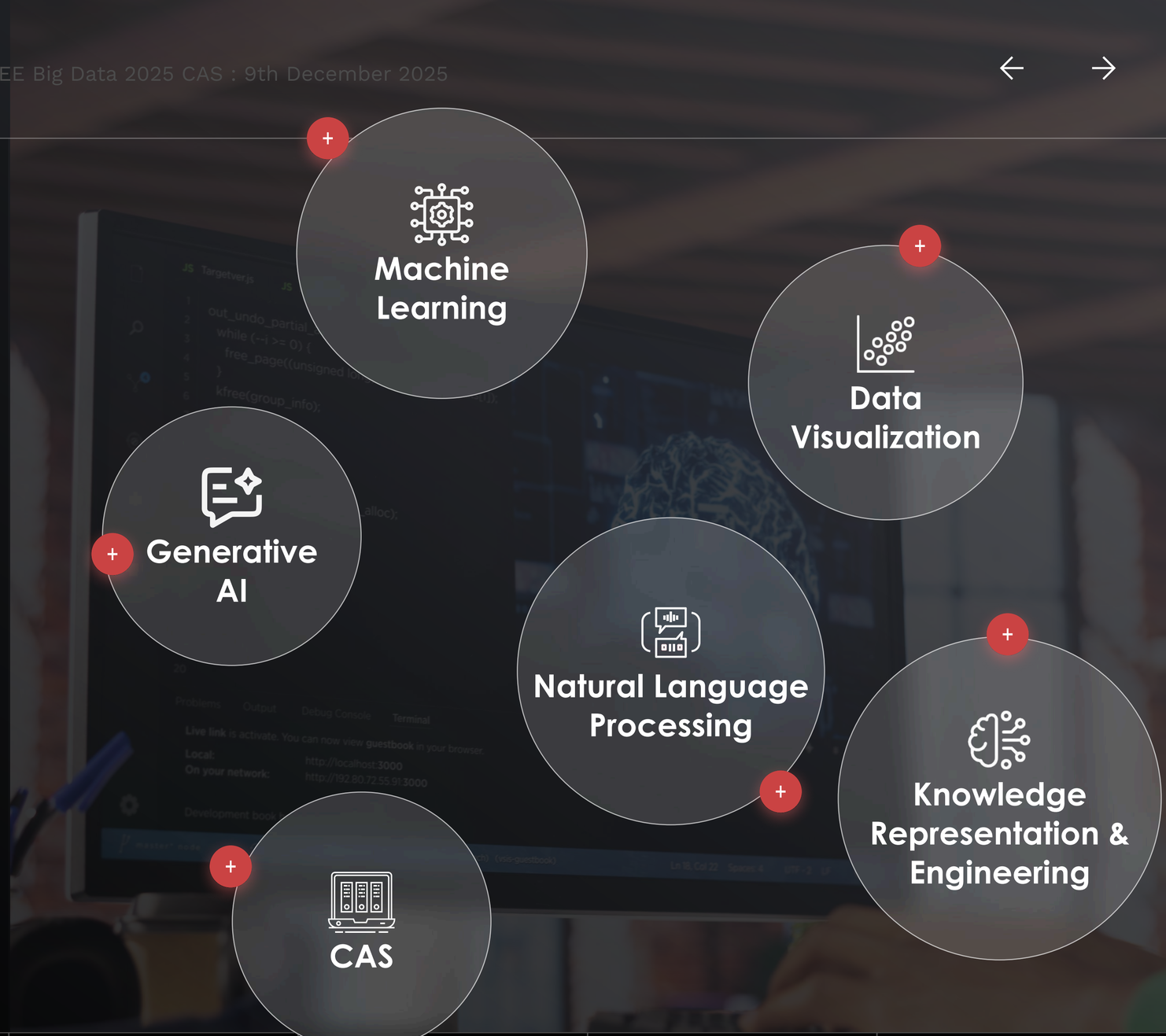
SEE RESEARCH DOMAINS →

02 — RESEARCH GOALS

Project Objectives

Core Objectives →

Research Domains →



Machine Learning

Role in Ask_ArchiLab

Machine Learning powers the intelligent retrieval and pattern recognition capabilities, enabling the system to understand context and deliver relevant archival information.

Key Technologies

Embeddings | Vector Search | Semantic Similarity | BM25 | Dense Retrieval

Data Visualization

Role in Ask_ArchiLab

Advanced visualization techniques make complex archival data accessible and understandable, revealing patterns and relationships that enhance user insight

Key Technologies

Knowledge Graphs | Interactive Dashboards | Relationship Ma | Network Visualization

Knowledge Representation & Engineering

Role in Ask_ArchiLab

Knowledge representation forms the foundation of our intelligent archival assistant, enabling structured and semantic understanding of archival data.

Key Technologies

RDF Triples | SPARQL | Ontologies | Linked Data | TripleStore

Natural Language Processing

Role in Ask_ArchiLab

NLP enables the system to understand user queries i natural language, process multilingual content, and extract meaning from archival texts.

Key Technologies

Tokenization | POS Tagging | Named Entity Recognition | Query Reformulation | Multilingual Processing

Generative AI

Role in Ask_ArchiLab

Generative AI enables natural conversational inter and intelligent synthesis of archival knowledge, cr contextually relevant responses

Key Technologies

RAG | GraphRAG | LLM | Multi-Vector RAG | Context Injection

Computational Archival Science

Role in Ask_ArchiLab

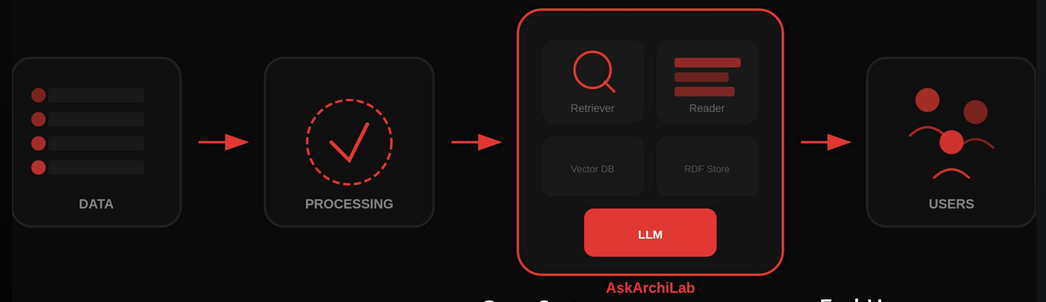
CAS automates the collection and processing of archival data. It helps transform raw archival resources into structured, usable knowledge at scale while maintaining archival standards.

Key Technologies

Data Mining | Knowledge Graphs | Metadata Standards | LLM Integration | Web Scraping | Embeddings

03 — TECHNICAL DESIGN

System Architecture



Data Collection

Restricted access datasets (standards, reports, courses) and unrestricted datasets (dictionaries, terminology, open portals) → flows into DataWarehouse

Data Processing

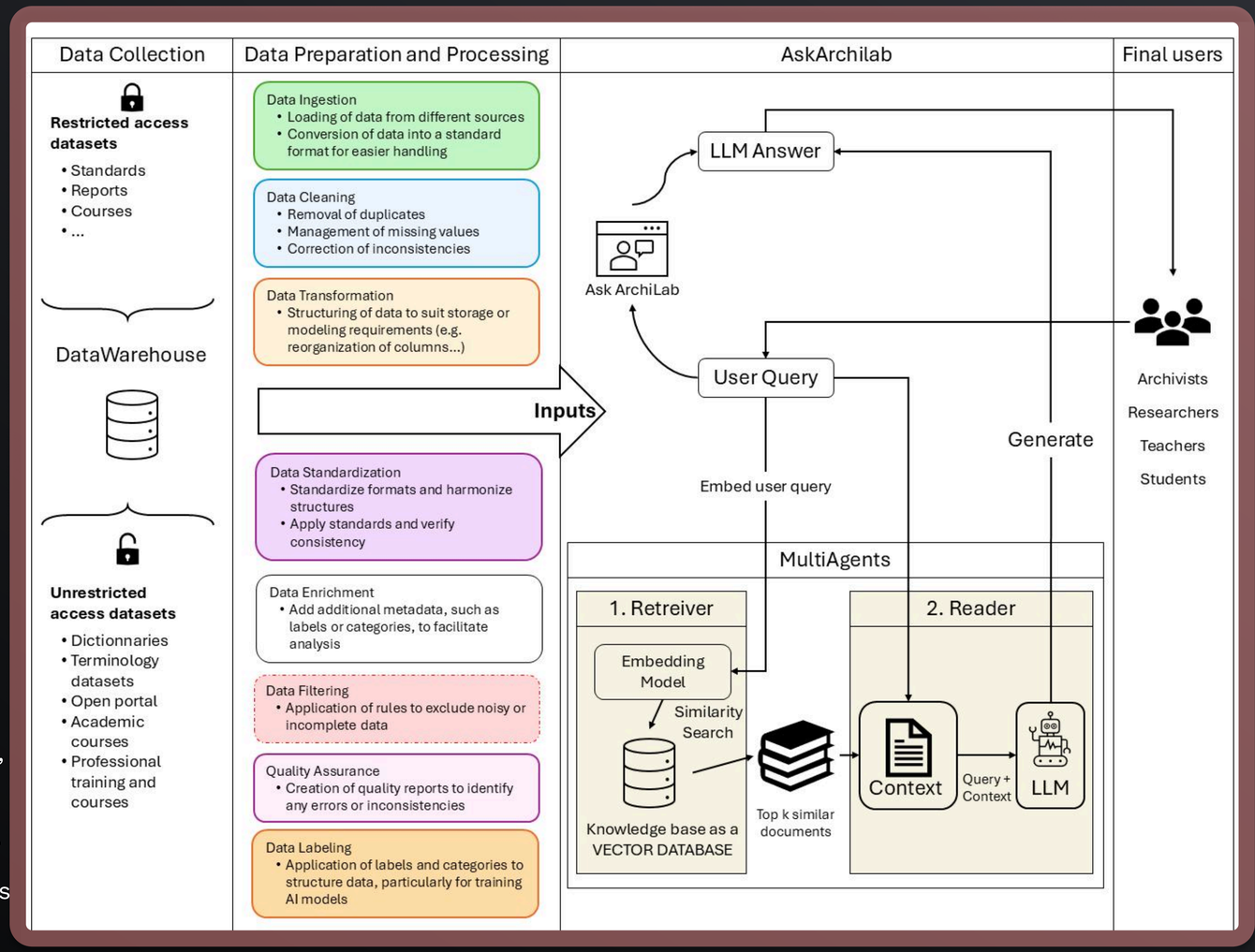
Sequential pipeline: ingestion, cleaning, transformation, standardization, enrichment, filtering, quality assurance, and labeling

Core System (AskArchiLab)

- User query processing through embedding model
- Triplestore for RDF representation + Vector database for knowledge storage
- Retrieval system finding top k similar documents
- Dual-path architecture: direct LLM answering + context-enhanced responses

End Users

Web-based user interface for query submission and response retrieval. Users (Professors, archivists, researchers, and students) can interact with the system, ask questions, and receive contextualized answers



04 — IMPLEMENTATION

Practical Workflow

Six-layer modular architecture for seamless query processing

1

User Interaction

Manages UI, captures queries, validates inputs

2

Query Preprocessing

Tokenization, POS tagging, semantic embeddings

3

Hybrid Retrieval

Dense vector + RDF triples

6

Response Delivery

Output formatting, logging, feedback collection

5

Post-Processing

Citations, verification, archival standards compliance

4

LLM Processing

Context fusion with retrieved data, grounded generation

05 — VALIDATION

Preliminary Results

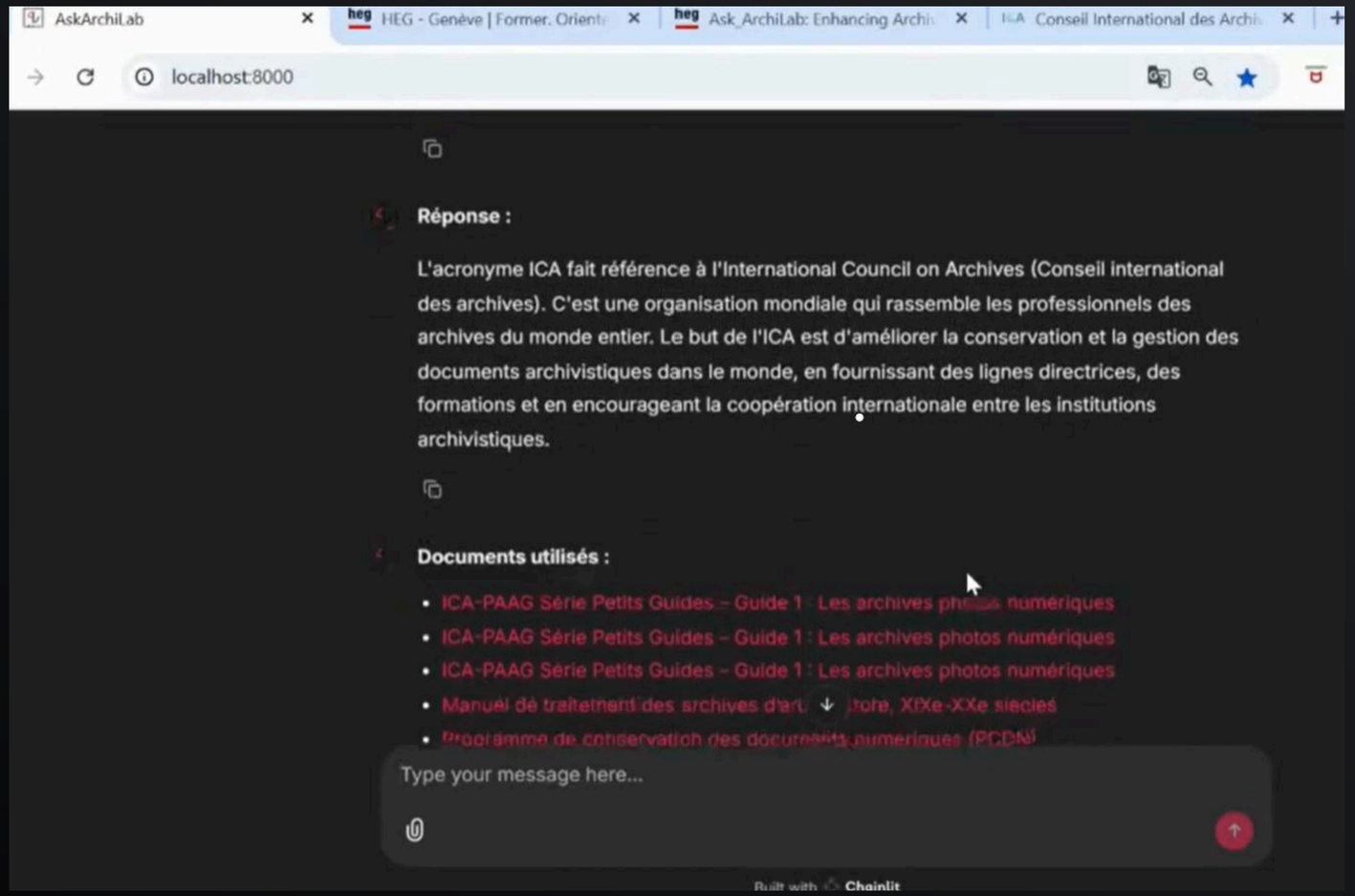
Case 1: ICA

International Council on Archives

Tested on ICA Online Resource Centre content including toolkits, guides, manuals, standards, glossaries, directories, and bibliographies. System successfully retrieved relevant documents (ISAD(G) guidelines) and grounded responses in verified sources.

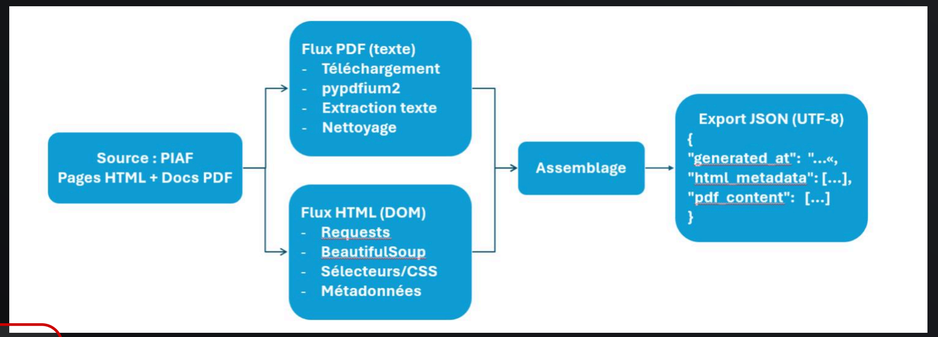
✓ Coherent responses ✓ Source verification

✓ High relevance



05 — VALIDATION

Preliminary Results



Case 2: PIAF

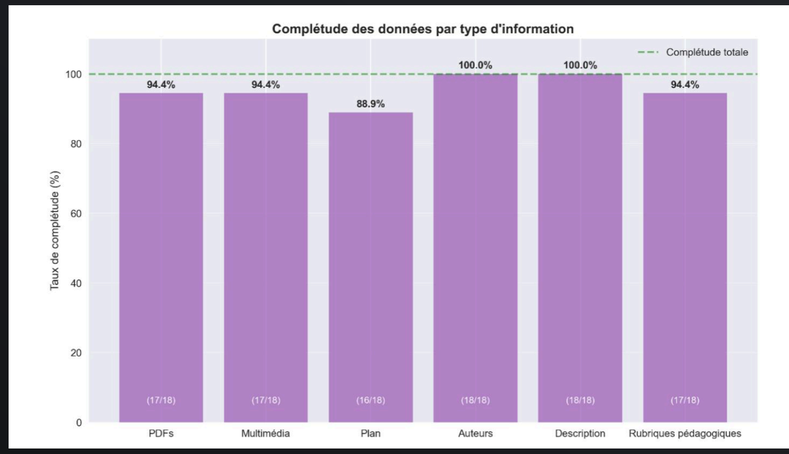
Portail International Archivistique Francophone

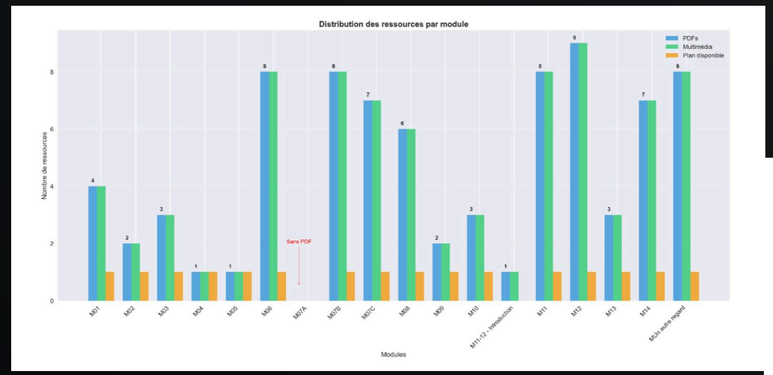
Complete transformation of PIAF's educational content covering the entire archival management cycle. Rigorous preprocessing pipeline applied to multilingual training modules from international experts.

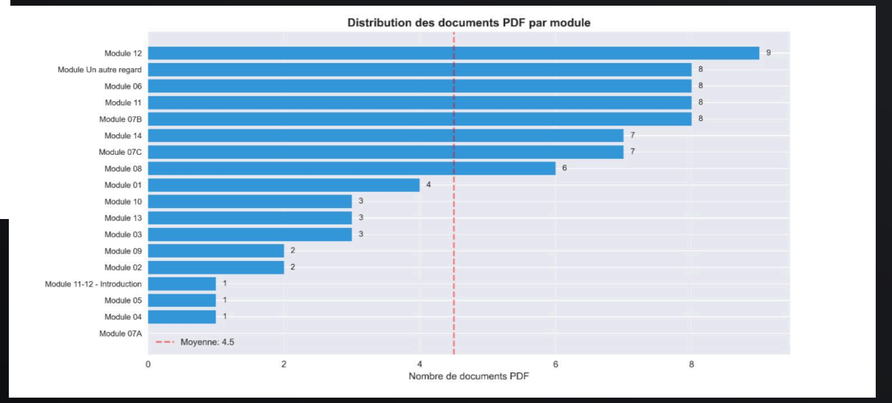
90%+ Coverage

Source traceability

Structured dataset



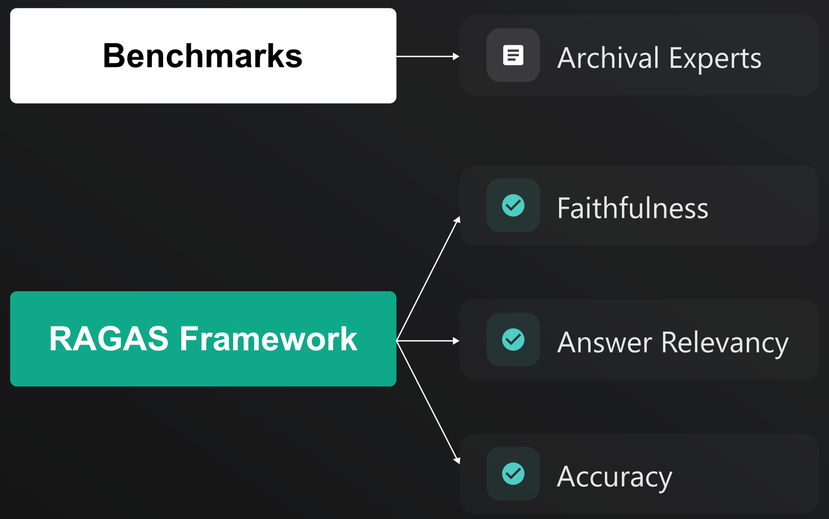




05 — VALIDATION

Ongoing Evaluation Process

End User Testing



Technical Testing Approach (Jan-March 2026)

- Query syntax validation and accuracy testing
- Generated queries verification
- Overall system performance assessment
- Automated tests on retrievers and generators (upcoming)
- Distributed vs monolithic agent comparison (upcoming)
- Other LLM & RAG frameworks for assessment

06 — FUTURE WORKS

Roadmap & Next Steps

Multi-Agent System

Our Multi-Agent approach will

- **Augment prompting capabilities:** An intelligent agent will be responsible for reformulating, enriching, contextualizing, and optimizing user prompts.
- **Augment retrieval capabilities:** An intelligent agent—or a coordinated set of agents—will handle this complex task by leveraging multilingual processing and complementary retrieval strategies.
- **Augment reliability, traceability, explainability, and guidance:** An intelligent agent will gather user feedback and maintain a structured prompt library to support consistent, transparent, and well-guided interactions.

1 — Phase 1 **Initial Agent** (Video Demo here)

Conversational agent setup with ICA dataset. Single-agent system with basic RAG implementation

2 — Phase 2 **Heterogeneous Data**

Expand capabilities with heterogeneous data sources with RDF knowledge representation

3 — Phase 3 **Multi-Agent System**

Implementation of distributed intelligent agents with parallel processing and collaborative problem-solving

4 — Phase 4 **Extended Use Cases**

Exploration and validation of diverse archival scenarios, international collaboration expansion, and real-world deployment.

07 — CONCLUSION

Latest advances in Generative AI combined with Knowledge Engineering have made innovation possible to face unprecedented challenges in contemporary archival practices.

This work introduces an innovative approach combining Advanced RAG and knowledge graphs techniques to maximize context and retrieve precise information

Thank you

Q & A

Special thanks to our collaborators and partners

Aurèle Nicolet and ArchiLab team,

Pierre-Yves Burgi from University of Geneva,

Luciana Duranti and **Muhammad Abdul-Mageed**, InterPares Trust AI coleaders,

Our partners at **ICA** and **PIAF** for their support and contributions,

Maryam Ablouâ for her work on the PIAF dataset,

Aïchatou Orou Bade for the Demo Video.

Basma Makhlouf Shabou
basma.makhlouf-shabou@hesge.ch
HES-SO Geneva

Lamia Friha
lamia.friha@unige.ch
University of Geneva

Wassila Ramli
Wassila.RAMLI@contexa.ch
Contexa

REFERENCES

- [1] G. Colavizza, T. Blanke, C. Jeurgens, and J. Noordegraaf, "Archives and AI: an overview of current debates and future perspectives," J. Comput. Cult. Herit., vol. 15, no. 1, pp. 1–15, Feb. 2022, doi: 10.1145/3479010.
- [2] A. Hawkins, "Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web," Arch. Sci., vol. 22, no. 3, pp. 319–344, Sept. 2022, doi: 10.1007/s10502-021-09381-0.
- [3] A. Marquet, "A maturity model for measuring digital transformation of archives and libraries," Qualitative and Quantitative Methods in Libraries, vol. 10, no. 3, pp. 269–282, Oct. 2021.
- [4] J. Sheridan and C. Foster, "'digitalising a national archive': interview with john sheridan, digital director at the national archives, UK," AI Soc., vol. 39, no. 2, pp. 665–668, Apr. 2024, doi: 10.1007/s00146-022-01510-2.
- [5] I. Schellnack-Kelly and M. Modiba, "Developing smart archives in society 5.0: leveraging artificial intelligence for managing audiovisual archives in Africa," Inf. Dev., vol. 41, no. 3, pp. 626–641, Sept. 2025, doi: 10.1177/02666669241286224.
- [6] S. J. Russell and P. Norvig, Eds., Artificial Intelligence: A Modern Approach, 4th ed. 2022. [Online]. Available: https://aima.cs.berkeley.edu/global-index.html
- [7] A. Z.Tsague, E. T. Fute, and L. P. Fotso, "Modeling of an environment for electronic archiving using a multi-agent approach," Int. J. Appl. Inf. Syst., vol. 5, no. 7, pp. 32–37, May 2013, doi: 10.5120/ijais13-450941.
- [8] J. Pellegrino, M. Maggiora, and W. Allasia, "A multi-agent approach for autonomous digital preservation," in 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Turin, Italy: IEEE, June 2015, pp. 1–6. doi: 10.1109/ICMEW.2015.7169866.
- [9] S. Aryal et al., "Leveraging multi-AI agents for cross-domain knowledge discovery," Apr. 12, 2024, arXiv: arXiv:2404.08511. doi: 10.48550/arXiv.2404.08511.
- [10] Portail International Archivistique Francophone, "Tous les cours," Portail International Archivistique Francophone. Accessed: Oct. 31, 2025. [Online]. Available: https://www.piaf-archives.org/tous-les-cours
- [11] W. Ramli, Ask_archilab tool, (June 09, 2025). [Online Video]. Available: https://www.youtube.com/watch?v=NNoAd_HMDRM
- [12] Multi-Agent GraphRAG: A Text-to-Cypher Framework for Labeled Property Graphs
- [13] X. Comtesse et al. "Bots and Robots", october 2025, BOTS AND ROBOTS | Georg Editeur