

Dormant Digital Collections, AI Triage

*Lessons from the Islam West Africa Collection
and beyond*

Frédéric Madore

Data Curator

Africa Multiple Cluster of Excellence



 [fmadore](#)  [frederickmadore.com](#)  [0000-0003-0959-2092](#)  [frederickmadore](#)  frederick.madore@uni-bayreuth.de

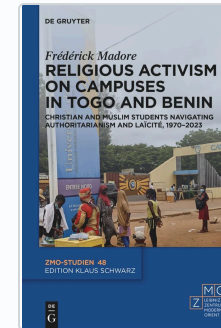


Scan for these slides

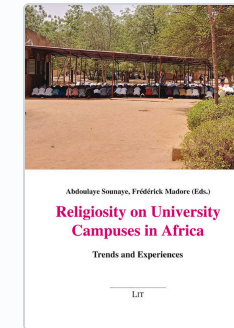
slides.frederickmadore.com

A historian, working with AI

- Data Curator, Cluster of Excellence "Africa Multiple", University of Bayreuth
- Historian of Islam in Francophone West Africa since the 1960s — youth, women, politics, education
- Developer of the **Islam West Africa Collection (IWAC)** since 2021
- Coordinator, *Digital Humanities and AI in African Studies*
- Edited volume forthcoming, Bielefeld University Press (2027)



Monograph · 2025



Co-edited · 2023

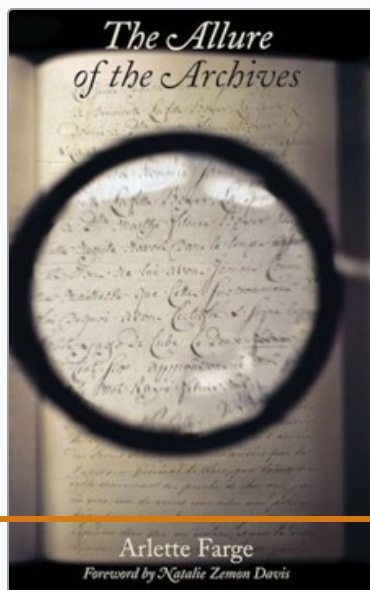


DH & AI initiative

— THE ALLURE OF THE ARCHIVE

From the reading room to the "scanning party"

THE READING ROOM



THE SHIFT

Thousands of JPEGs and hours of recordings, half named, untouched for years — affect moves from slow intimacy to frantic accumulation.

THE SCANNING PARTY



An unsorted backroom, photographed in three days.

— FROM BACKLOG TO TRIAGE

One historian, no team

- Fieldwork across West Africa, plus libraries in Berlin, Paris, Washington
- Public side: the open-access IWAC — islam.zmo.de
- Cataloguing it all by hand is impossible

THE QUESTION I PUT TO AI

Not "read this for me" — but **where is close reading worth my time?**

— A DIGITAL HOARDER, IN NUMBERS

ITEMS COLLECTED (ZOTERO)	30,000+
STILL UNPROCESSED	~15,000
OPEN ACCESS (IWAC)	14,500+
HISTORIAN, NO TEAM	1

Six uses of AI in the research workflow

1 Text extraction

OCR and HTR with multimodal models

3 Entity & topic extraction

who, where, what, about what

5 First-pass analysis at scale

sentiment, classification, triage

2 Audio & video transcription

including under-resourced languages

4 "No-code" visualisation

through AI-assisted coding

6 NotebookLM

chat with your own sources

Multimodal LLMs read the page like a person



Layout recognition failure

Layout recognition failure: traditional OCR scrambles a multi-column page.

- Traditional OCR breaks on multi-column papers, tables, handwriting, quick phone scans
- Even a 5–10% error rate matters: one typo in a name hides a document
- "Ouédraogo" → "Oueclaoqo": the source is there, but no one will find it
- Multimodal AI handles it in one pass: printed & handwritten, French & Arabic, messy layout

STRONGEST SO FAR

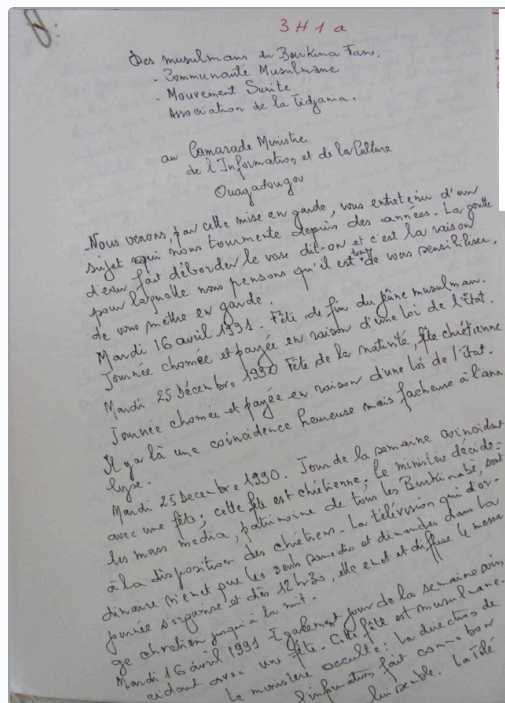
🌟 Google Gemini 3 Pro — the best model I have tested for this.



Enye Togo yeyea fofo avafiawodzifia Gnassingbe Eyadema fe didi
vevi be amesiame nana dzidzo kple nutifafa me lle yefe Denyigba

↓ Printed Ewé, diacritics and all — scroll for the full extraction.

Reading cursive French



Handwritten letter, Burkina Faso, 1991.

TRANSCRIBED TEXT · FRENCH

3 H 1 a

Des musulmans du Burkina Faso,

- Communauté Musulmane
- Mouvement Sunite
- Association de la Tidjania.

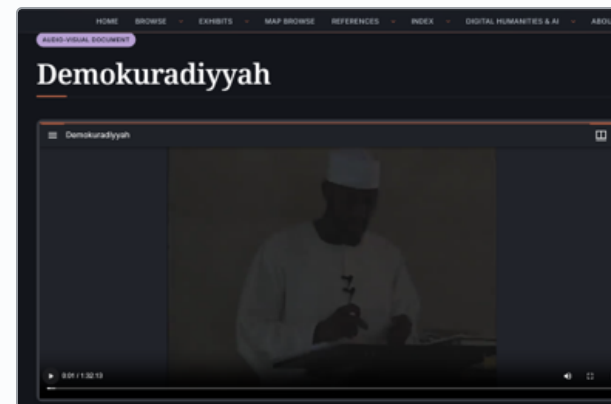
au Camarade Ministre de l'Information et de la Culture, Ouagadougou

Nous venons, par cette mise en garde, vous entretenir d'un sujet qui nous tourmente depuis des années. La goutte d'eau fait déborder le vase dit-on et c'est

 Highlighted: the writer's own spelling slips, kept verbatim — the model didn't silently correct them.

Audio and video transcription

- The same multimodal logic extends to audio and video
- Transcripts return with **timestamps and speaker labels**
- IWAC corpus: **~155 hours** of Nigerian preachers, mostly Hausa & Arabic
- Early tests in Hausa, Arabic, Kurmanji Kurdish — promising as a first draft



```
CONTENT
=====
[Part 1]
=====

Transcription of: iwac-video-0000001-1.mp4
Generated using: Google gemini-3-pro-preview
=====

[Segment 1]
[00:00:00] Speaker 1: Alhamdulillah nahmaduhu wa nasta'inuhu wa nastaghfiruh. Wa na'udhu billahi min
shururi anfusina wa min sayyi'ati a'malina. Man yahdihillahu fala mudilla lah, wa man yudlil fala hadiya
lah. Wa ashhadu alla ilaha illallah wahdahu la sharika lah, wa ashhadu anna Muhammadan 'abduhu wa
rasuluh.

[00:00:18] Speaker 1: Ya ayyuha alladhina amanu ittaqullaha haqqa tuqatihi wala tamutunna ila wa
antum muslimun. Ya ayyuha annasu ittaqu rabbakum alladhi khalaqakum min nafsni wahida wa khaleqa
minha zawjaha wa baththa minhuma rijalan kathiran wa nisa'a. Wattaqullaha alladhi tasa'aluna bihi wal-
arham innallaha kana 'alaykum raqiba.

[00:00:39] Speaker 1: Ya ayyuha alladhina amanu ittaqullaha wa qulu qawlan sadida. Yuslih lakum
a'malakum wa yaghfir lakum dhunubakum. Wa man yuti'llaha wa rasulahu faqad faza fawzan 'azima.
Amma ba'du fa inna asdaqal hadithi kitabullah, wa ahsanal hadiyi hadyu Muhammadin. Wa sharral umuri
mubdatatiba. fa inna kulla mubdatatin hid'a, wa kulla hid'atin dalala, wa kulla dalalatin fin-nar
```

A sermon and its machine transcript — timestamps, speaker labels.

— A CAUTION

Cleaner than the source

- AI-OCR and audio-to-text open previously unusable material
- The model silently **modernises spellings**, standardises toponyms, smooths regional variants
- It can also **hallucinate**: invent text that was never there
- The failure mode is harder to spot: a fluent paragraph with invented detail

THE DISCIPLINE

Keep the original image alongside the text. **Verify against the scan before you quote.**

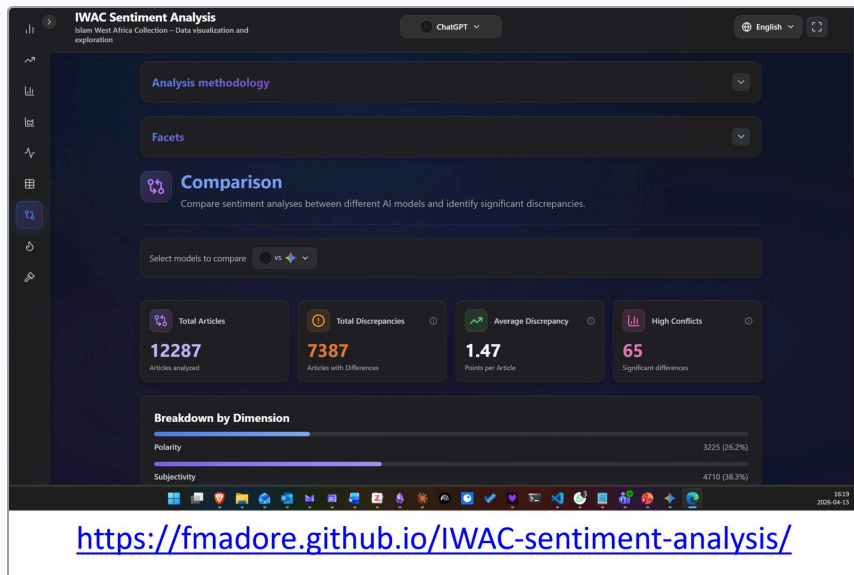


-

[illegible]

"Find me every document that discusses **secularism** or the **pilgrimage**."

Sentiment analysis, without writing code



Three models compared, article by article.

fmadore.github.io/IWAC-sentiment-analysis ↗

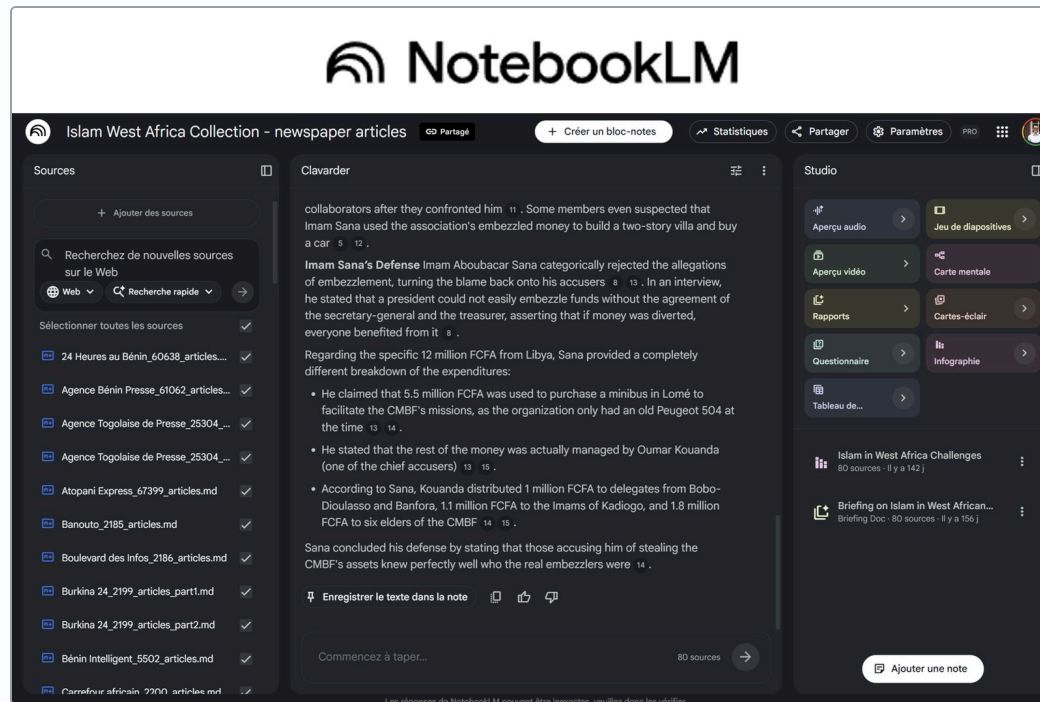
- 12,000+ articles, three LLMs, three axes: polarity towards Islam, subjectivity, centrality of religion
- ~€35 and 24 hours of batch processing — months of manual coding avoided
- Triage, not full reading; model disagreement is itself a signal

A PATTERN I'D HAVE MISSED

Positive portrayals of Muslim communities *inside* Sahelian terrorism coverage.

— USE 6 · CHAT WITH YOUR SOURCES

NotebookLM: the lowest-friction entry point



NotebookLM over the IWAC newspaper corpus.

- Only knows what you upload; **cites the exact passage** for every answer
- Your sources are not used to train models
- Free tier: 50 sources per notebook (PDF, web, YouTube, audio)
- Multilingual: ask in your language, it searches sources in any of 80+



The same workflows, in your browser

1 OCR & HTR

PDFs and images, printed or handwritten

2 Audio & video transcription

interviews and lectures, with speaker labels

3 Summaries & keywords

batch a whole spreadsheet of texts

- Google Colab notebooks — nothing to install
- A free Gemini API key; results save to your Drive
- Open to anyone; written for researchers, not developers



github.com/fmadore/zmo-ai-pipelines ↗

— THE STAKES

Western tools, non-Western sources

THE COSTS

- **Training data** overwhelmingly Western — concepts read through European lenses
- **Labour:** Kenyan annotators paid ~\$5/day; similar conditions elsewhere
- **Infrastructure:** strongest models in US firms; only China credible as an alternative
- **Opaque** and non-reproducible: the same input can yield different answers

WHAT IT MAKES POSSIBLE

Research at **scales**, and with **documents**, not previously feasible — degraded scans, under-resourced languages, corpora too large to read.

HELD IN TENSION

Neither uncritical adoption nor refusal. Use the tools, and disclose what they cost.

— IN CLOSING

The scans on our hard drives are the raw material

- The technology is no longer the obstacle — small teams, no deep coding required
- Models you can **download and run on a university server** — evolving fast
- AI organises; **we interpret. Always.**
- Experiment with sources you know well, so you can catch the errors



Scan for these slides
slides.frederickmadore.com

 [fmadore](#)  frederickmadore.com  [0000-0003-0959-2092](#)  [frederickmadore](#)  frederick.madore@uni-bayreuth.de

