

ENKRS 2026 • BERLIN

Heritage and Horizons: From Historical Collections to the Age of AI

AI-Driven Preservation & Discovery: Unlocking Hidden Archives

Park So-Yeon

Preservation and Research Center, NLK



National Library of Korea

Presentation Overview

01

Preservation
Crisis

02

Paradigm
Shift

03

Technical
Validation

04

AI Discovery
Vision

05

Roadmap



| National Identity & Global Role

Founded in **1945**, the NLK is mandated by **Library Act Article 19** as the supreme repository of the nation's intellectual heritage.



303 Specialized Experts



94 Billion KRW Budget (2026 Strategic Plan)



Global Hub for Sovereign AI Training Data



National Library of Korea

Massive Scale: National Data Assets

The library holds over **37 million** items.

15M physical volumes accumulated over 80 years.

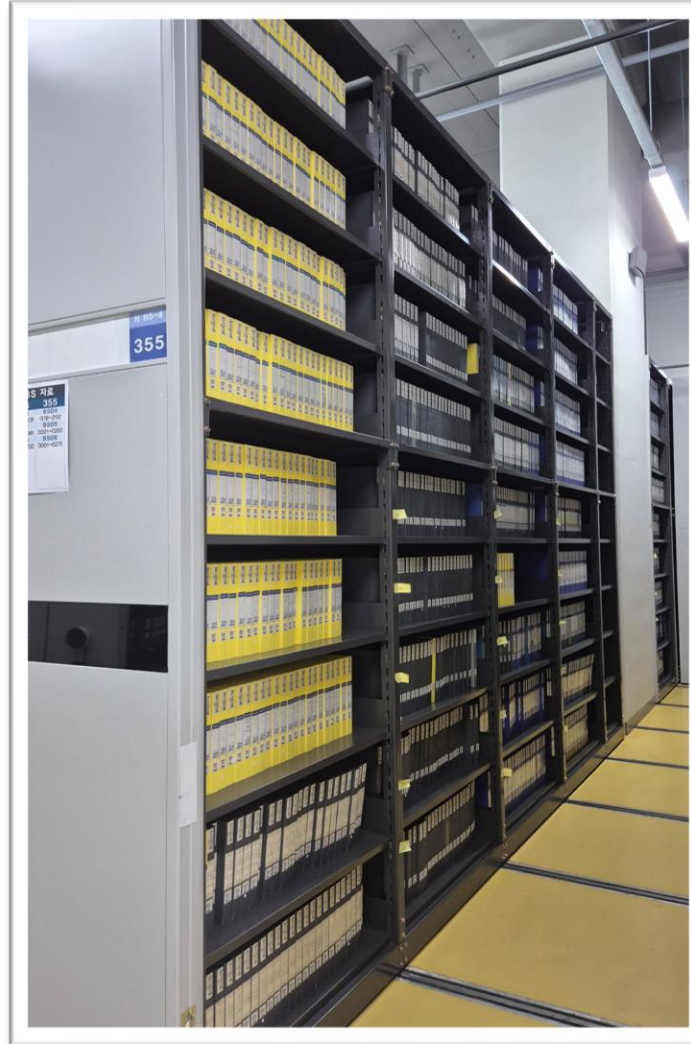
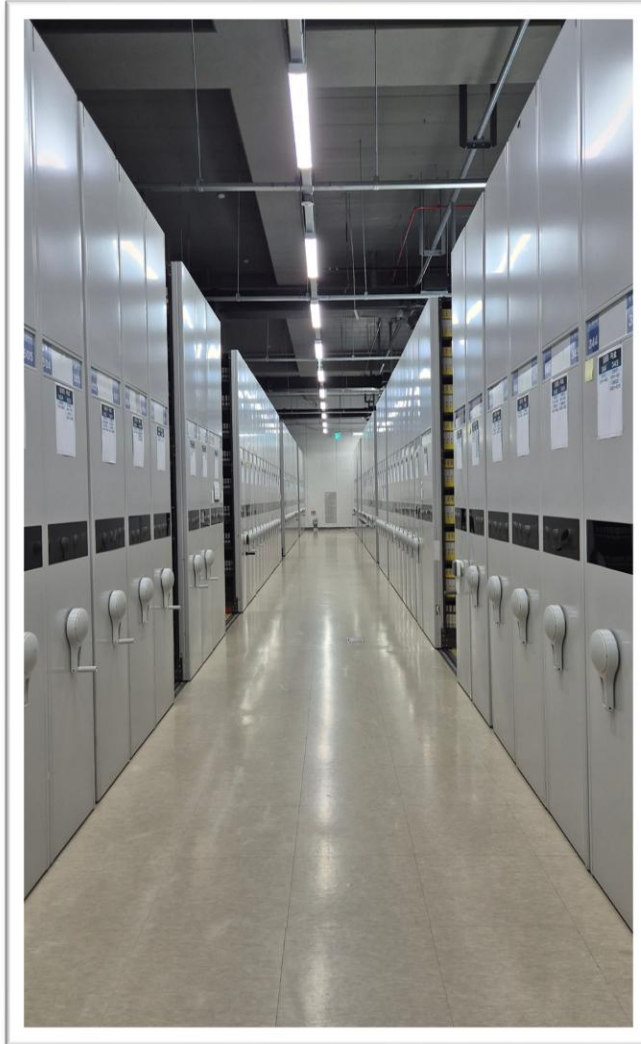
22M+ digital items, acquired under the Online Legal Deposit mandate (est. 2016).



⚡ Online(Digital) : e-Journals · e-Books · Websites · Digitized Files

⚡ Legacy Broadcast Archives : KBS & SBS BETA TAPES 680,000+ Units (Trusteeship)

Status of Non-Book storage at NLK



Institutional AI Readiness : NLK's Strategic Framework



AI Future Strategy

Internal AI Task Force established. AI Data Service Future Strategy published (Oct. 2025).



Culture Tech AX

NLK leads a MCST-funded project(5.9B KRW) developing AI agent technology for facilitator-led debates and knowledge linking.



Scalable Infra

Leveraging MSIT-supported GPU resources for internal data pipeline development and AI experimentation .



The Crisis of Preservation

Why traditional methods are failing national archives



Audiovisual Heritage: Collective DNA

AV records capture non-reproducible fragments of history that text alone cannot convey. They form the living pulse of national memory.



Aural Reality

The actual voices of historical figures, capturing raw human emotion and tone that documents miss.



Visual Evidence

Authentic records of vanished landscapes, architectures, and the evolution of daily lifestyles.



Social Identity

The cultural DNA of a nation, anchoring identity through shared visual history.



Physical Collapse: Sticky-Shed Syndrome

HYDROLYTIC DESTRUCTION

Irreversible delamination

Binder hydrolysis causes the magnetic layer to peel from the substrate. Tapes are physically “deteriorating”, rendering them unplayable.

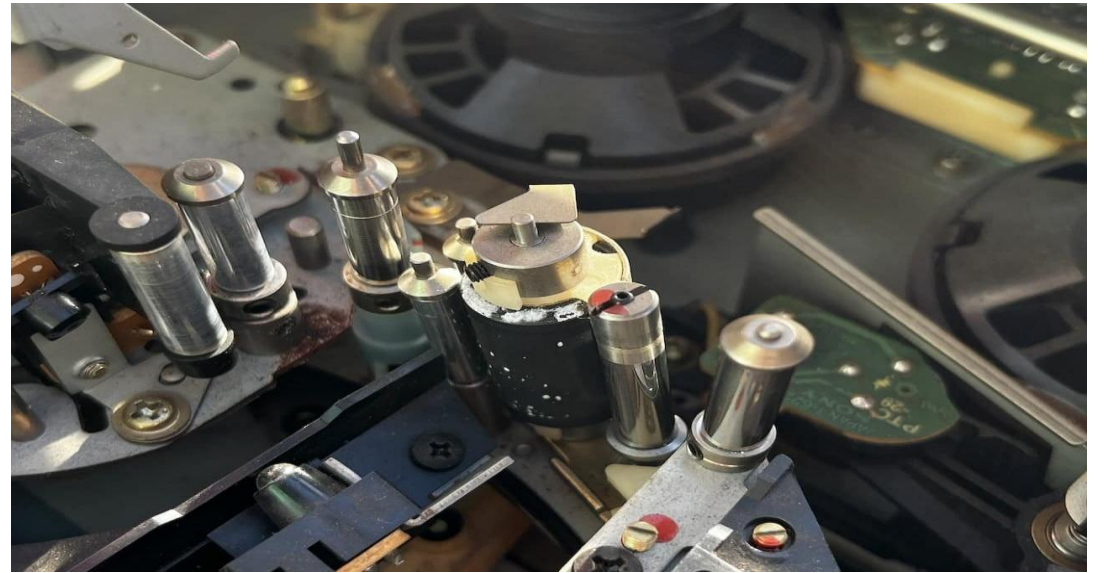
*** Attempting playback in this state results in total data loss.**



Hardware Obsolescence

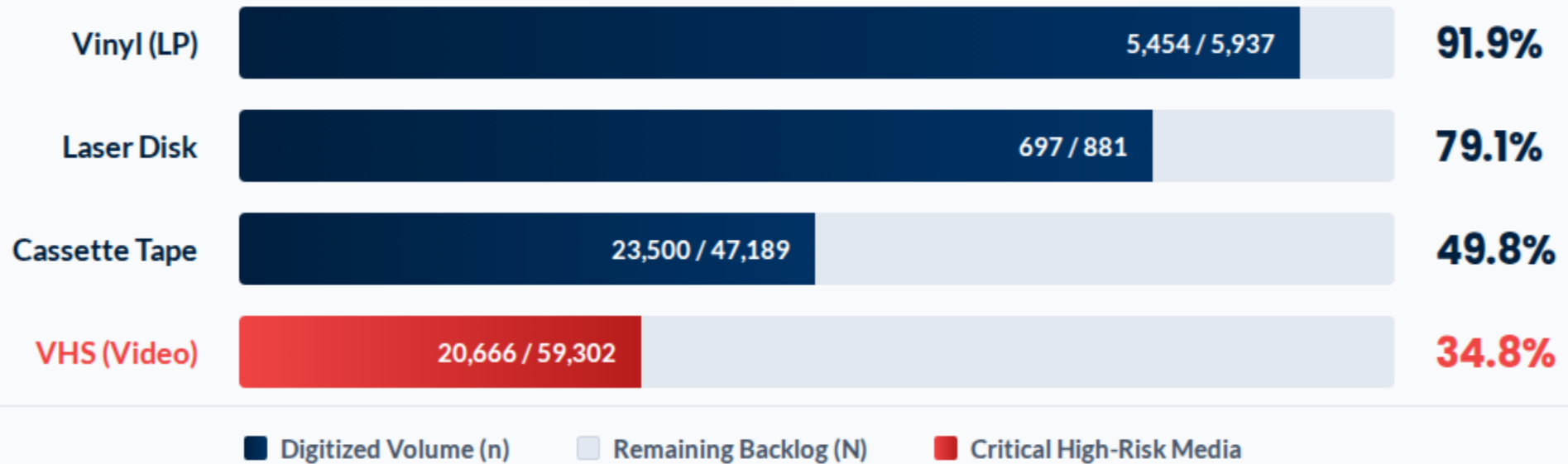
The Vanishing Infrastructure

Production of Betamax, Betacam, and U-matic components have ceased globally. Playback devices have become as rare as the materials they preserve.



Analog AV Digitization Status

Detailed Breakdown of Carrier-to-File Conversion Progress



! Critical Alert : As playback hardware becomes increasingly scarce and magnetic media continues to deteriorate, the 34.8% completion rate for VHS creates a high risk of permanent heritage loss.

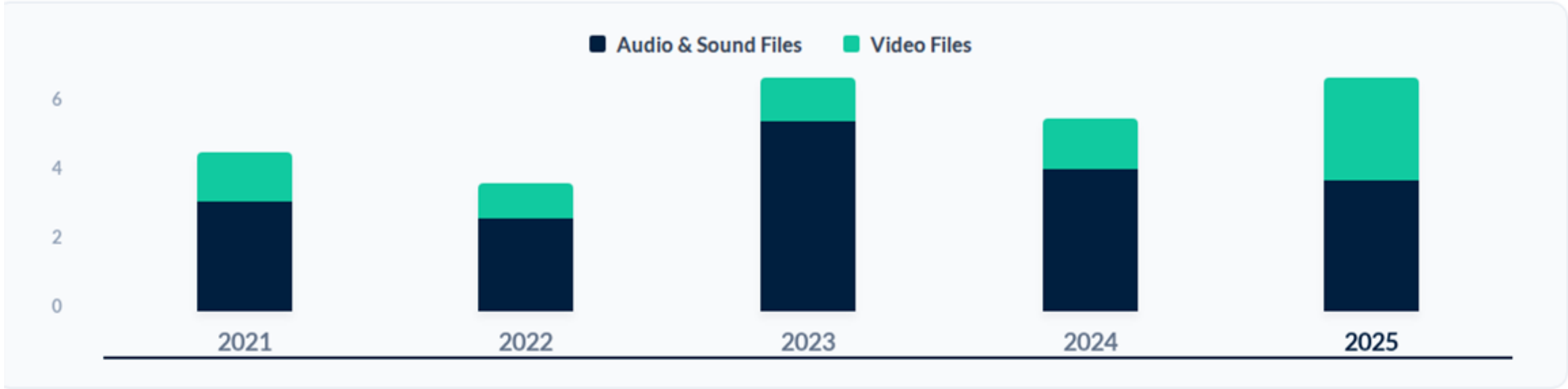


The Paradigm Shift

From manual cataloging to AI interpretation



Born-Digital Acquisition Status



Category	2021	2022	2023	2024	2025
Audio & Sound Files	20,752	17,380	38,694	26,688	29,474
Video Files 	9,074	6,716	8,827	9,432	23,025



The 400,000-Hour Mission

400k

HOURS OF NATIONAL HERITAGE

220,000h (Analog Backlog)

VHS and Cassette tapes awaiting digitization

"The Preservation Mission."

180,000h (Born-Digital Assets)

Online acquisitions and legacy digital files
lacking metadata for discovery.

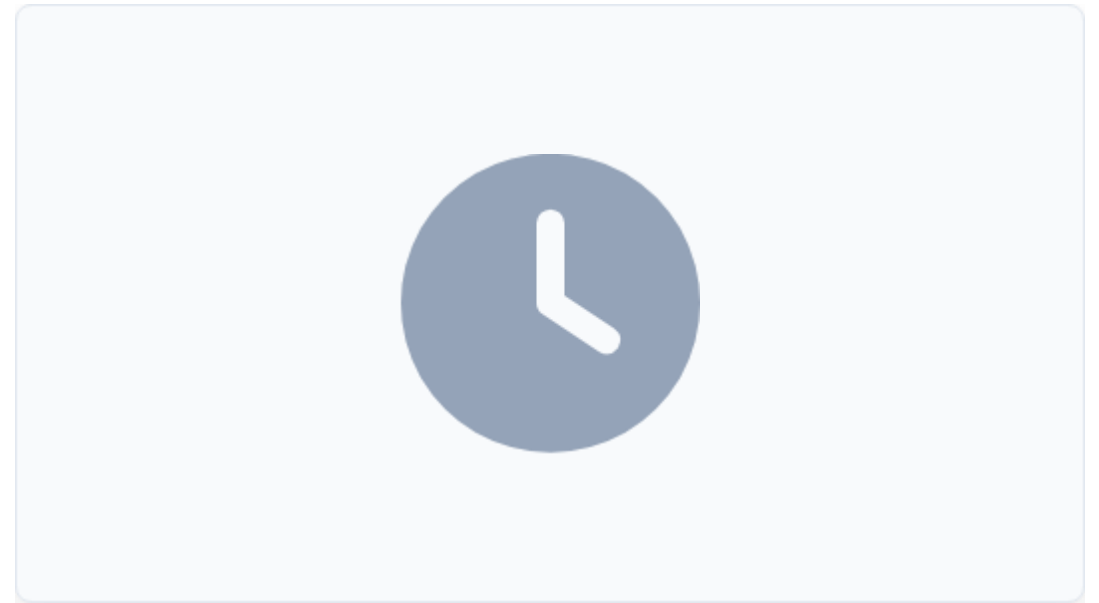
"The Discovery Mission."



| The Indexing Wall

137

Years of Manual Review



To manually review 400,000 hours of digital files at 8 hours/day, 365 days a year — it would take 137 years. Human indexing alone is not an option.



National Library of Korea

Evolution of Discovery Systems



1945~ Card Catalog

Physical indexing requiring manual curation
and drawer discovery.



1990~ Web Search

Electronic databases based on exact
keyword matching.



2024~ AI Discovery

Semantic intelligence understanding intent
and context.



The Limit of Keyword Search

Query: "Find Seoul street scenes from 1988"

✕ Search Result: 0 Results

If "Seoul street scenes from 1988 " isn't in the title or curator-entered tags, the archive stays silent. Traditional search fails historical intent.

The screenshot shows the National Library of Korea website. The search bar at the top contains the query "1988년 서울의 거리" (Streets of Seoul in 1988). Below the search bar, the results are displayed in a list format. The first result is titled "1. 기사 서울과 바르셀로나, 38년의 거리" (Article: Seoul and Barcelona, 38 Years Apart). The second result is titled "2. 기사 레저가이드: 김승환-홍리나: 서울에서 1시간거리, 주말 데이트 코스 8" (Article: Leisure Guide: Kim Seung-hwan-Hong Rina: Seoul 1-hour distance, weekend date course 8). The third result is titled "3. 기사 해외에서 보내 온 신년메시지: '고국의 동포여러분, 안녕하십니까: 브라질: 서울올림픽이 이곳의 화제거리'" (Article: New Year's message sent from abroad: 'Hello to our compatriots in the motherland: Brazil: Seoul Olympics are the topic of conversation here'). Each result includes a brief description and a link to view the full text.



Paradox of the Dark Archive



Intelligence Strategy: Concurrent Process

"Beyond mere storage — toward discoverable assets"



Physical Rescue

Real-time signal capture and digital master generation
before hardware obsolescence



Intelligent Awakening

AI pipeline ingestion concurrent with digitization;
real-time content analysis and structuring



Technical Validation

Measuring STT (Speech-to-Text) precision and performance gains



Innovation: Specialized Architecture



Dual Scribe

WhisperX & Faster-Whisper parallel validation.



Word-Merge

Automated word-level merging algorithm.



LLM Post-Edit

Llama-3 context correction & preservation.



Technical Validation: Metrics & Taxonomy

Indicator Definition

CER: Character-level reliability (Optimized for Korean).

WER: Word-level global standard.

$$WER = \frac{S + D + I}{N}$$

Variable Matrix

S (Substitutions): Incorrect character mapping.

D (Deletions): Signal omission in archives.

I (Insertions): Model-generated hallucination.

N (Total Count): Reference ground-truth count.



Performance Leap: 82% to 90%+

PHASE 1: BASELINE

82%

Open-source Whisper results

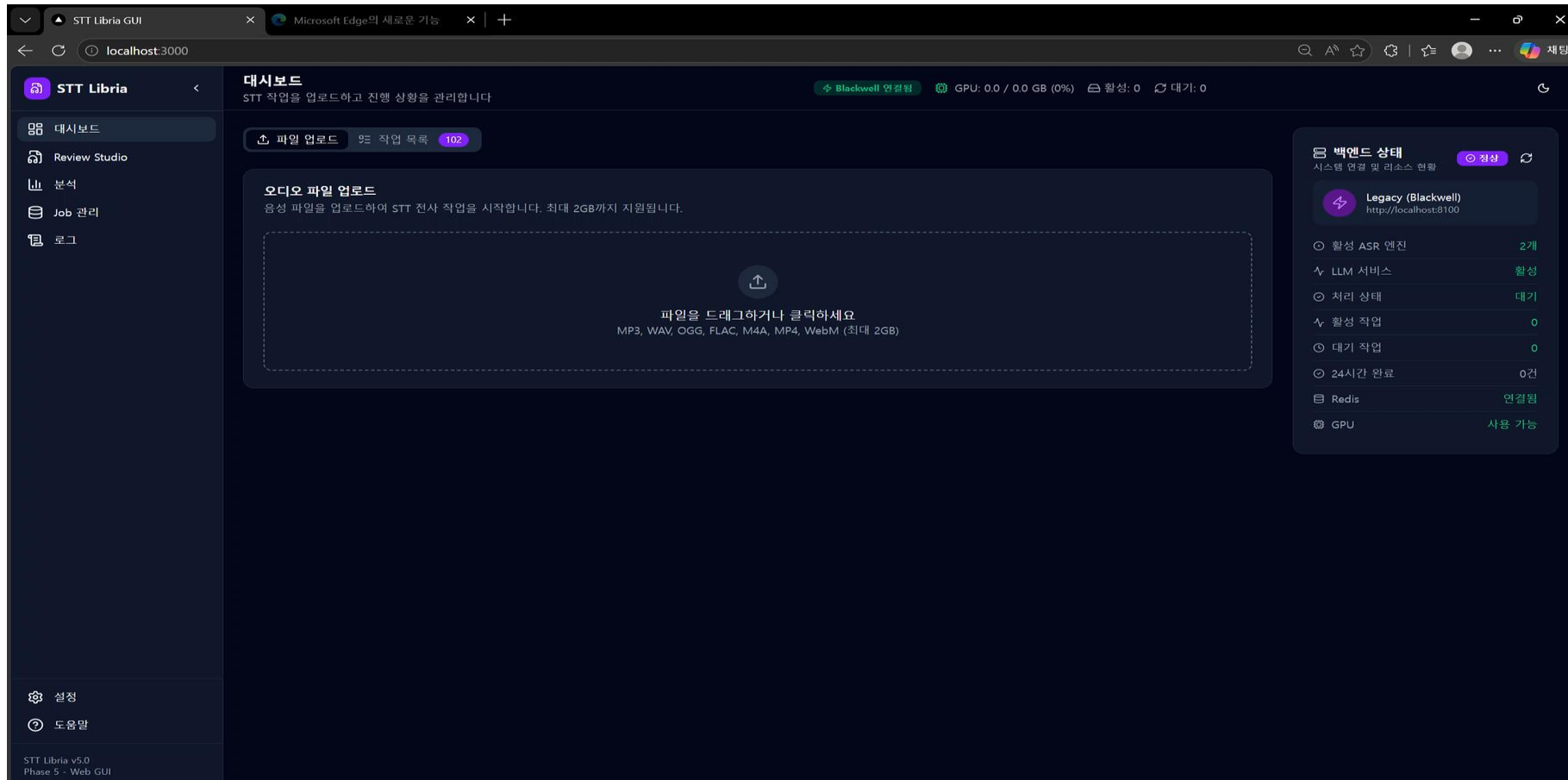
PHASE 2: SPECIALIZED

90% +

ETRI Ensemble Model results



Validation: STT Libria Platform



STT Precision Management: Analysis Dashboard

Dash

127.0.0.1:8050

출력하기 가져가기 국립중앙도서관 등... Google NAVER ChatGPT 공적자료열람실 i:나루

텍스트 비교 및 오류 분류

원본 JSON 업로드

수정된 JSON 업로드

비교 실행

오류 세부 사항

텍스트 비교

정확도

비교 텍스트

1년 82년 무렵이 되면 이제 조선정
란, 1882년에 발생한 임오군란, 임
이다라는 것이 명시가 됩니다바로

한 주권상실의 내용⁶²은 크게
세 가지로 나뉘어 구분될 수 있

시기를 공화국의 개국은 조영공공에
거주지 밖 10박심⁶⁶리까지 토지와
주요한 거기 고자 선의의 자으로

오류 유형

○ 비슷한 발음으로 잘못 전사

○ 완전히 다른 발음으로 잘못 전사

○ 반복 전사

○ 번역하여 전사

○ 특수문자 줄격

○ 특이사항

○ 행동사/관행사/부사

○ 조사/어미

○ 감탄사

○ 여러 어절

○ 오음란

○ 자음란

○ 자음과 오음

○ 사자성어/속문/관공표현

○ 해당 없음

상세 표

전체 평가

Metric	C	S	D	I	오류율(%)	정확도(%)
CEP	9998	168	58	271	4.909	95.091
NER	2006	128	19	59	9.568	90.432

SDI Precision Tagging

Sub (S): Proper Noun Tracking

Corrected AI Result

Del (D): Noise Signal Omission

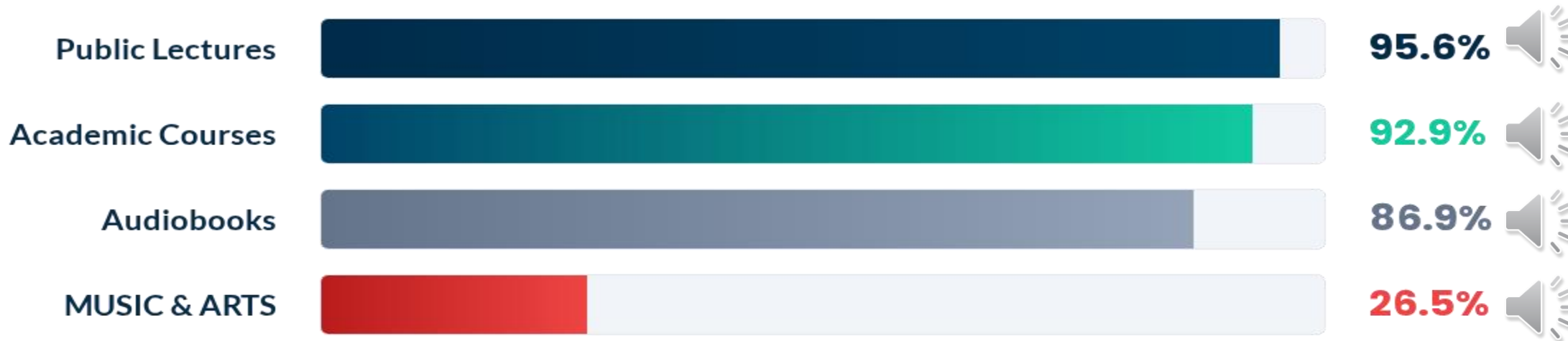
Missing Text

Ins (I): Hallucination Filtering

Hallucination

 National Library of Korea

Technical Validation: Domain STT Accuracy



The Critical Bottleneck in Music Domains

Speech-centered collections achieved over 90% STT accuracy, while the Music & Arts domain dropped to 26.5% due to melodic overlap and background interference. This highlights the need for specialized MIR (Music Information Retrieval) and multimodal AI technologies for national music archives.



Error Impact Evaluation

Error Type	Contextual Impact	Curator Response
Proper Nouns	Critical (Historical Distortion)	Manual Review via SDI Dashboard
Technical Terms	Causes Discovery Failure	Build Domain-Specific Lexicon
Interjections	Negligible	Automated removal via LLM post-edit



Vision: AI Discovery

Expanding textual success to the visual pixel



From STT Pilot to AV Discovery

3.8M Record Pilot

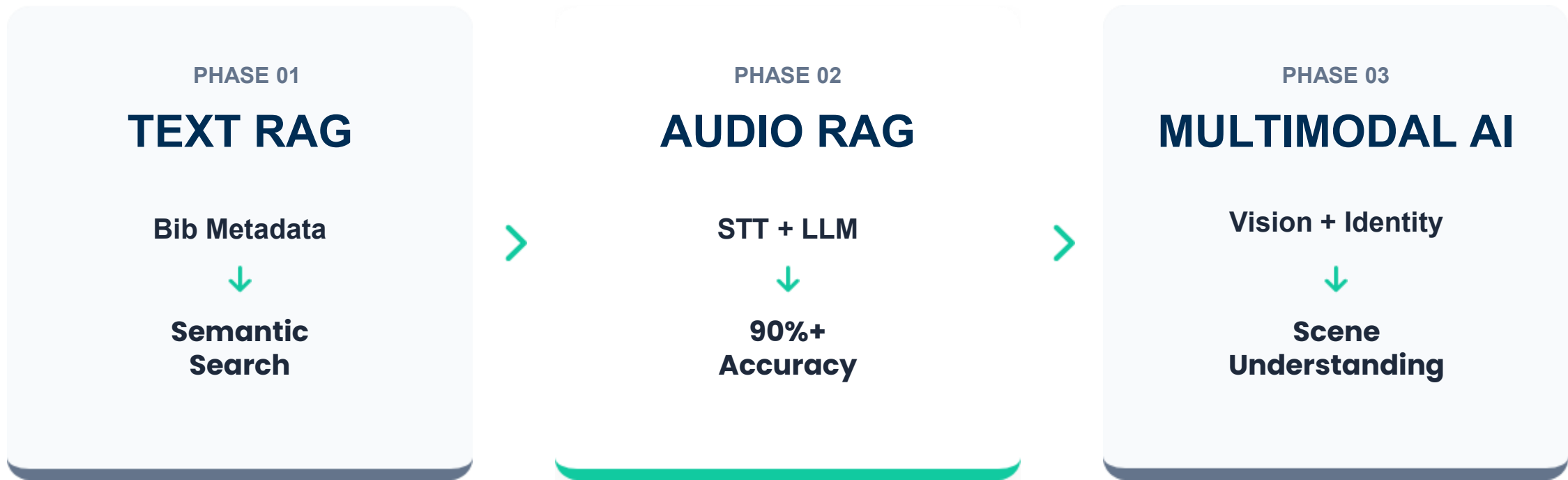
A pilot study by Kim (2025) demonstrated the feasibility of intent-based book discovery using Bib-text RAG.

The AV Challenge

Extending STT findings to full AV content — where speech ends and visual, musical context begins.



RAG Evolution: From Text to Multimodal



Multimodal Discovery Architecture

Resource Type	Next-Gen AI Engine	Structured Metadata Output
 Oral History	STT + LLM (NER)	• Auto-indexing people & dates • Time-stamped summaries
 Music Archives	MIR + Audio Separation	• Genre, BPM, & Key tagging • Precision lyric-to-audio sync
 News & Records	VLM + Object Re-ID	• Historical figure recognition • Landmark & Geo-location

Our Goal

Our strategy converts unstructured audiovisual streams into multi-layered, granular metadata, enabling the next generation of RAG-based discovery.

Roadmap

Infrastructure & service deployment: 2027–2029



Infrastructure Roadmap 2027–2029

2027

Foundation

- MAM-AI Pilot Setup
- GPU Infrastructure
- Pipeline Testing

2028

AI Scaling

- Full Scene Indexing
- Backlog Processing
- Model Refinement

2029

Service Launch

- RAG Discovery Beta
- Public Web Portal
- Hub Integration

Goal: A seamless transition from Infrastructure setup to Nationwide Intelligent Search.

Pixel Discovery: 2002 World Cup

Intelligent Query: "Moment of the goal celebration against Italy."

98% SIMILARITY | RECORD #2002-WC

Timestamp 01:12:45: Golden Goal event, specific player recognition (Ahn J.H.), and stadium atmosphere detected.

AI identifies historical events through **visual intelligence**, even without text metadata.



National Library of Korea

Object Intelligence as Evidence



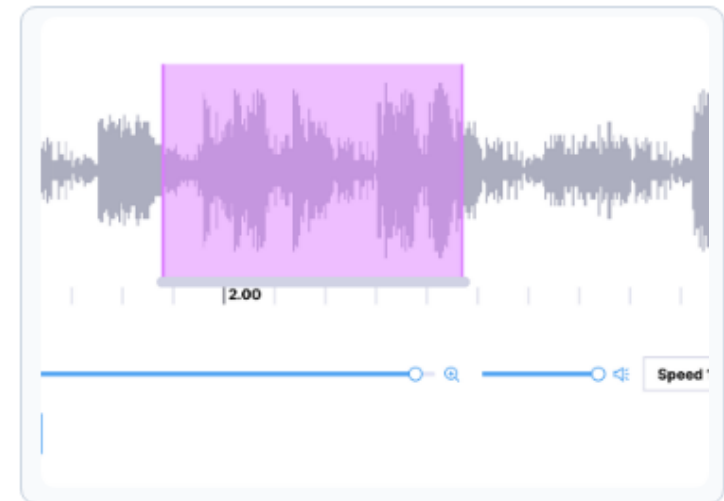
Cultural Symbols

Identifying 'Red Devils' shirts indexes the 2002 era automatically.



Key Figures

Recognizing Coach Hiddink confirms match context in unlabeled reels.



Event Detection

Detecting 'Golden Goal' event generates per-second metadata.



Pyeongchang 2029 Vision

National Repository for Library Collection



Repurposed Legacy

Transformed from the 2018 Pyeongchang Winter Olympics International Broadcasting Centre (IBC)



Massive Capacity

Target capacity of 14 million volumes over the next 30 years, with an integrated data center



Multimedia Hub

Central hub for long-term digital preservation and AI-driven multimedia resource management



National Repository for Library Collection · Pyeongchang, Gangwon-do



National Library of Korea

Audiovisual Records Yet to Be Found

A Discovery at the U.S. National Archives (NARA)

Voice of America liberation-era recordings — broadcast to Korea around the time of liberation — have been discovered at the U.S. National Archives (NARA).

***Somewhere, hidden records
are still waiting to be found.***

NLK is committed to **surveying, restoring, and digitizing** Korean AV heritage held overseas — and we look forward to **collaborating with you**.



Voice of America liberation recordings · U.S. National Archives

- Surveying unclassified Korean Audiovisual materials overseas
- Technical restoration & digitization support
- Building a shared discovery network



***"Records preserved but not discovered
are as good as silent."***

AI is the **new intelligence** transforming
dark archives into **living national assets**.



Thank You for Your Attention

감사합니다

• Danke schön

Preservation and Research Center, NLK

Park So-Yeon, Conservator | vet16@korea.kr

www.nl.go.kr

