OPEN ACCESS

ISSN 3022-6198

Korea Special Library Association

KSLA Insight

2026. August. Vol. 6

Google Scholar 색인 개선과 AI 기반 연구성과 관리 자동화: ETRI 도서관 사례

한국전자통신연구원 **배성진**

KSLA 한국전문도서관협의회
Korea Special Library Association

OPEN ACCESS

Korea Special Library Association

KSLA Insight

Google Scholar 색인 개선과 AI 기반 연구성과 관리 자동화: ETRI 도서관 사례

Improving Google Scholar Indexing and AI-Based Automation of Research Output Management: A Case Study of the ETRI Library

배 성 진 한국전자통신연구원

Seongjin Bae(Electronics and Telecommunications Research Institute, ETRI)

[초 록]

한국전자통신연구원(이하 ETRI) 이 기관리포지터리 지식공유플랫폼의 Google Scholar 색인을 제고하기 위해 수행한 메타데이터 품질 개선과 연구성과 업무 자동화 경험을 정리하였다. 2024년에는 Google Scholar 색인 문제를 자료 단위의 메타데이터 품질 문제로 규정하고, 저자명·논문명 정합성, 초록·DOI·원문 PDF·메타태그 보강과 색인 모니터링을 통해 리포지터리 성과물의 색인 문제를 개선하였다. 2025년에는 외부 DB에서 초록을 확보하기 어려운 학술대회 논문을 대상으로 OCR과 LLM을 결합한 원문 기반 초록 추출 절차를 설계하고, 요약 생성이 아니라 원문에 존재하는 초록 구간을 식별하는 방식의 실무 적용 가능성을 확인하였다. 2026년부터는 Codex 기반 에이전트 워크플로와 RPA를 활용해 학술대회 프로그램 탐색, 근거 자료 수집, 검증 게이트를 분리한 학술대회 성과 입력·검증 자동화를 추진하고 있다. 이 과정은 도서관의 AI 활용이 새로운 시스템을 별도로 구축하는 방식만이 아니라, 기존 업무의 메타데이터를 정비하고 반복적인 확인·입력 절차를 자동화하며 사람이 검증할 수 있는 구조를 마련하는 방식으로도 충분히 실질적인 변화를 만들 수 있음을 보여 준다.

This article presents the ETRI Library's efforts to improve the Google Scholar discoverability of its institutional repository, the ETRI Knowledge Sharing Platform(KSP), through metadata quality improvement

※ 본 원고는 2025년 전국도서관대회 발표 자료 및 2026년 국립중앙도서관·전문도서관협의회 협력교육 자료를 재가공, 정리한 것임을 밝힙니다.

and workflow automation. It describes three related initiatives: improving item-level metadata and indexing conditions in 2024, developing an OCR- and LLM-based workflow for extracting existing abstracts from paper PDFs in 2025, and building an AI- and RPA-assisted process for entering and verifying conference output records. The case shows that practical AI adoption in libraries can begin with metadata cleanup, automation of repetitive verification tasks, and workflows that keep human review in the loop.

키워드 : 기관리포지터리, 메타데이터 품질, 초록 자동 추출, Institutional repository, Metadata quality, Automated abstract extraction, Google Scholar, Large Language Model(LLM), Robotic Process Automation(RPA), AI agent

들어가며: 색인되지 못하는 기관리포지터리 성과물

한국전자통신연구원(이하 ETRI)은 정보·통신·전자·방송 분야를 연구하는 정부출연연구기관이며, ETRI 도서관은 2018년부터 지식공유플랫폼(ETRI Knowledge Sharing Platform, 이하 KSP)을 운영해 왔다. KSP에는 논문, 연구보고서 등 5만 건 이상의 연구성과가 공개되어 있다. KSP는 ETRI가 자체 구축한 기관리포지터리(Institutional Repository, IR) 로서, 여러 학술지에 흩어진 연구자의 논문은 물론, 상용 DB에서 잘 드러나지 않는 학술대회 논문과 연구보고서까지 연구자 및 부서 단위로 통합하여 제공한다. 이를 통해 연구성과의 가시성과 접근성을 높여 대외 확산과 공유를 지원하고, 연구자·부서별 성과 정보를 바탕으로 원내 전문가와 협력 가능 분야를 탐색할 수 있는 기반을 제공한다.

그러나 2024년 당시 Google Scholar에서 확인되는 KSP 성과물의 색인 건수는 수십 건에 불과했다. Google Search Console에서 확인할 수 있는 Googlebot의 일일 크롤링 요청도 300여 건 이하에 머물렀다. 원문과 서지정보를 공개하고 있었지만, 학술 검색엔진이 이를 충분히 읽고 신뢰할 수 있는 형태로 받아들이지 못하고 있었다.

이 문제는 단순히 KSP의 방문자 수가 적다는 차원의 문제가 아니었다. IR에 축적된 기관의 연구성과가 학술 검색 환경에 제대로 드러나지 않는다는 뜻이었고, 이는 곧 IR 본연의 기능이 충분히 작동하지 못하고 있음을 의미했다. 따라서 IR의 색인과 노출을 개선하는 일은 기관의 연구성과가 검색 가능한 지식 자원으로 활용되도록 만드는 일이다. 본 고에서는 그 과정을 세 단계로 나누어 정리한다. 첫째는 Google Scholar 색인 개선, 둘째는 PDF 논문 초록 자동 추출, 셋째는 학술성과 관리 자동화이다.

Google Scholar 색인 개선

색인 문제의 진단

Google Scholar는 색인 알고리즘을 상세히 공개하지 않으며, 개별 사이트의 색인 문제에 대한 직접 문의도 제한적이다. 또한 Google Scholar는 일반 Google 검색과 구별되어, Google Search Console에서 페이지 색인이 확인되더라도 Google Scholar에 반드시 색인되는 것은 아니고, 반대로 Scholar에 보이는 자료가 Search Console에 색인되지 않는 경우도 있다. 그래서 초기에는 크롤러 접근성, 사이트맵, 페이지 속도, PDF 다운로드 경로 같은 기술 요소를 하나씩 확인하는 방식으로 접근했다. 그러나 Google Scholar의 공개 가이드라인과 리포지터리 색인 설명을 검토하면서, 핵심은 단순한 웹 접근성보다 학술 콘텐츠로 판별될 수 있는 자료 단위 메타데이터의 신뢰성에 있다는 점을 확인하였다(〈표 1〉 참조).

[표 1] Google Scholar 색인 요건과 권장 메타태그

Google Scholar 색인 요건

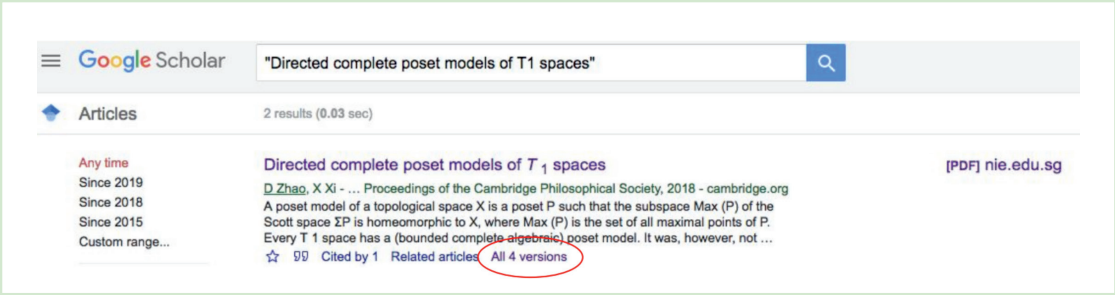
- 콘텐츠는 학술 콘텐츠로 한정(journal, conference 등)
- 초록을 볼 수 있어야 함
- 광범위한 메타데이터 오류를 가진 웹사이트는 색인 대상에서 제외
- 웹페이지에 HTML 메타태그가 있어야 하며, 메타태그 정보는 원문 PDF내 정보와 일치 필요
- 원문 PDF 텍스트와 동일한 언어로 논문 메타데이터, 메타태그 정보 작성 필요

출처: Google Scholar Inclusion Guidelines for Webmasters(https://scholar.google.com/intl/en/scholar/inclusion.html)
Google Scholar Indexing for Institutional Repositories(https://www.carl-abrc.ca/wp-content/uploads/2021/01/Google_Scholar_webinar_Jan2021.pdf)

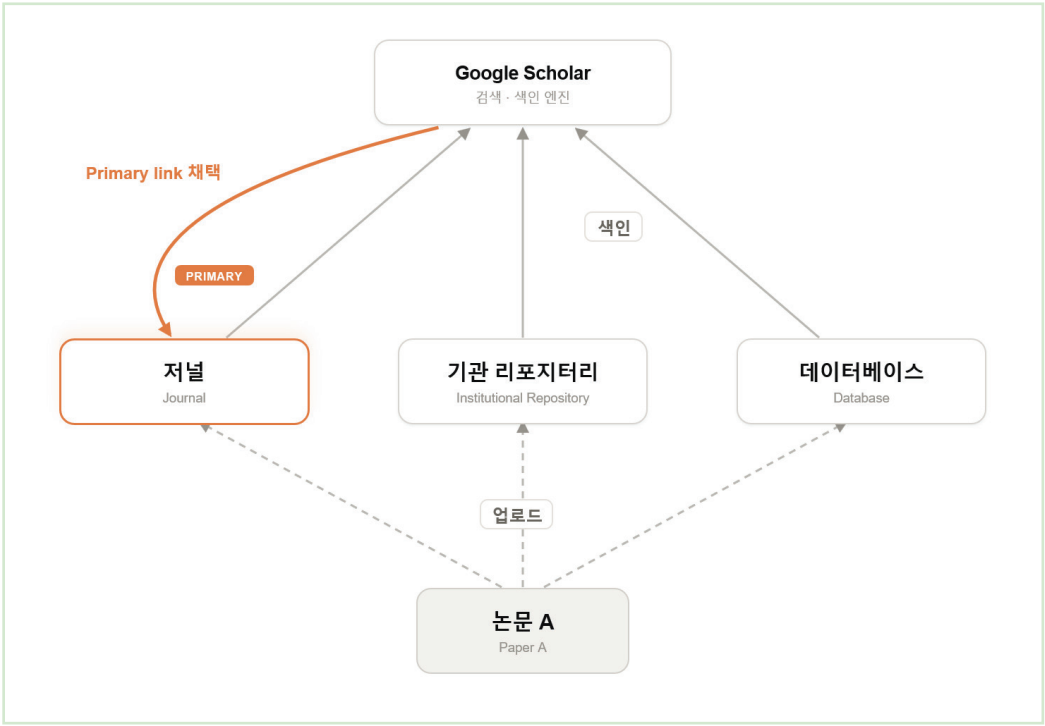
Google Scholar에서 요구하는 메타태그 예시

```
<meta name="citation_title" content="The testis isoform of the phosphorylase kinase catalytic subunit (PhK-T) plays a critical role in regulation of glycogen mobilization in developing lung">
<meta name="citation_author" content="Liu, Li">
<meta name="citation_publication_date" content="1996/05/17">
<meta name="citation_journal_title" content="Journal of Biological Chemistry">
<meta name="citation_volume" content="271">
<meta name="citation_issue" content="20">
<meta name="citation_firstpage" content="11761">
<meta name="citation_lastpage" content="11766">
<meta name="citation_pdf_url" content="http://www.example.com/content/271/20/11761.full.pdf">
```

Google Scholar의 색인 방식을 이해하는 것도 중요했다. Google Scholar는 리포지터리에서 제공되는 메타데이터와 원문 PDF를 바탕으로 해당 자료가 어떤 논문인지 판별하고, 출판사 웹사이트, 데이터베이스, 리포지터리 등 여러 웹사이트에 올라간 동일 논문을 하나로 묶어 보여 준다(〈그림 1〉 참조). 이때 최초 검색 결과에 노출되는 링크가 primary link(대표 링크)이고, 'All versions' 안에 포함되는 나머지 링크가 secondary link(보조 링크)이다(〈그림 2〉 참조). 결국 리포지터리 자료가 출판사 사이트 논문이나 다른 데이터베이스 자료와 같은 논문으로 인식되려면 제목, 저자명, 발행정보, DOI, 초록, 원문 PDF 등 기본 메타데이터가 원문과 일치해야 한다(Westin 2021).



〈그림 1〉 Google Scholar 검색 결과의 All versions 링크



〈그림 2〉 Google Scholar 색인과 Primary link 선택 구조

이러한 내용을 파악한 뒤 점검의 초점을 웹페이지 구조에서 자료 단위의 품질로 옮겼다. 색인된 논문과 색인되지 않은 논문을 나란히 놓고 메타데이터와 HTML 메타태그를 비교하였다. 같은 논문이 출판사 사이트, KSP, 데이터베이스에 동시에 존재하는 경우에는 Google Scholar가 어느 쪽을 primary link로 보여 주는지도 함께 보았다. 그 결과 '페이지에 접근 자체가 가능한가' 보다 '해당 페이지가 원문과 일치하는 정확한 학술 메타데이터를 제공하는가'가 더 중요한 문제라는 판단에 이르렀다. 도서관 입장에서는 검색엔진 최적화보다 메타데이터 품질 관리의 문제에 가까웠다.

KSP를 이 기준으로 점검하자 여러 문제가 드러났다. 논문에 기재된 저자명이 아니라 기관 인사 데이터베이스의 영문명이 노출되는 경우가 있었고, 실제 논문의 원제목 대신 연구자가 제출한 ETRI 내부 결재문서의 제목이나 별도 영문 제목이 표출되는 경우도 있었다. 언어, 출판사, 키워드, DOI, 초록 관련 메타태그도 누락되거나 잘못 입력된 사례가 많았다. 또한 KSP 논문의 초록 보유율은 26%, 원문 PDF 대외공개 비율은 4%에 그쳤고, 오기입된 초록은 1,800여 건, 오기입된 저자명은 800여 건, 중복 표출되는 논문은 100건 이상이었다.

메타데이터 정비와 색인 개선 결과

개선은 우선 Google Scholar 색인에 직접 영향을 주는 항목부터 진행하였고, 본 고에서는 메타데이터 정합성, 외부 데이터 연계, 색인 모니터링 및 메타태그 중심으로 정리하였다(〈그림 3〉, 〈그림 4〉 참조).

첫째, 논문 원문에 적힌 정보를 우선하도록 표기 방식을 바꾸었다. 기존에는 웹사이트의 언어 설정에 따라 저자명과 논문명이 한글과 영문으로 변환되었고, 저자명은 인사 DB에 등록된 영문명이 노출되는 경우가 있었다. 개선 후에는 모든 저자명과 논문명을 실제 논문에 기재된 형태를 기준으로 노출하도록 조정하였다.

둘째, 외부 메타데이터 API를 활용해 KSP 논문 메타데이터의 누락값과 오류를 일괄 보정하고, 가능한 한 많은 요소의 메타데이터를 제공하도록 하였다. Scopus, Crossref, Unpaywall, 한국학술지인용색인(KCI), ScienceON, DBpia 등 API를 제공하는 외부 출처를 활용하여 KSP 메타데이터와 대조하고, 오류값을 수정하거나 누락 항목을 보완하였다. 국내 학술대회 논문처럼 외부에서 초록을 구하기 어려운 자료는 GPT API를 활용해 PDF 원문에서 직접 초록을 추출하였다. 2024년 당시에는 비용 문제로 gpt-3.5 기반의 제한적인 자동 추출에 머물렀지만, 이후 대량 추출 파이프라인을 설계하는 출발점이 되었다.

셋째, KSP의 Google Scholar 색인 상태를 지속적으로 확인하기 위해 자동 모니터링 코드를 구축하였다. 이 코드는 하루 두 차례 site:ksp.etri.re.kr 검색을 실행하여 KSP 성과물이 Google Scholar에서 primary link로 색인된 논문 목록을 수집하고, 새로 색인된 논문과 색인에서 제외된 논문을 로그로 기록하였다. 모니

터링 결과, 색인에서 제외되는 KSP 논문들은 대부분 메타데이터 오류를 가지고 있었다. 반면 제목, 저자명, 초록, DOI 등 정확한 메타데이터를 제공하고 KSP에서 독점적으로 제공하는 자료는 primary link로 노출될 가능성이 높았다.

Google Scholar Search Results

Total results: 38
Task performed on: 2024-10-16 20:00:00 KST

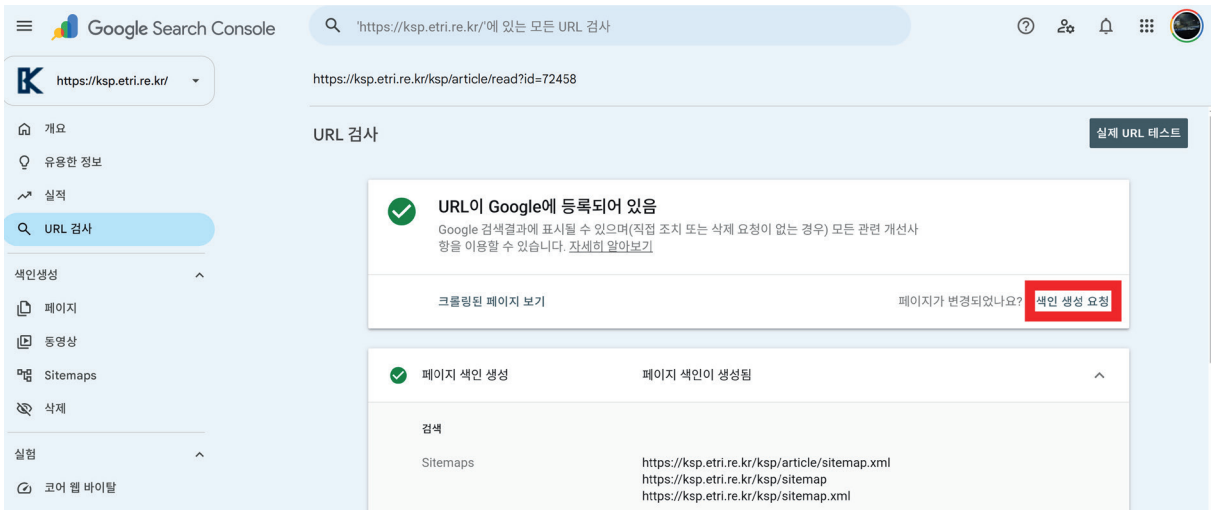
Newly Added Entries (4):

Title	Authors	Journal	Year	Cited by	URL	Previously Indexed	Removed Date
Similarity Enhancement of Face Recognition Performance between Real Faces and Facial Videos through Adjustment of a Display Device	조미영 , 정영숙 , 전병태	N/A	2016	0	Link	No	N/A
Low Complexity Spectrum Sensing Algorithm with Integrated Preprocessing for Multi-Channel Detection	H Jung, J Um, BJ Jeong	N/A	2013	0	Link	No	N/A
2D Microwave Image Reconstruction of Breast Cancer Detector Using a Simplex Method and Method of Moments	KC Kim, BD Cho, TH Kim, JM Lee, SI Jeon	N/A	2010	0	Link	No	N/A
국제경쟁력 강화를 위한 정보통신 원천기술연구	A ETRI	N/A	NA	0	Link	No	N/A

Removed Entries (2):

Title	Authors	Journal	Year	Cited by	URL	First Seen
Low Complexity Spectrum Sensing Algorithm with Integrated Preprocessing for Multi-Channel Detection	H Jung, J Um, BJ Jeong	N/A	2013	0	Link	2024-10-14
2D Microwave Image Reconstruction of Breast Cancer Detector Using a Simplex Method and Method of Moments	김기채 , 조병두 , 김태홍 , 백정기	N/A	2010	0	Link	2024-10-09

<그림 3> Google Scholar 색인 모니터링 결과 화면



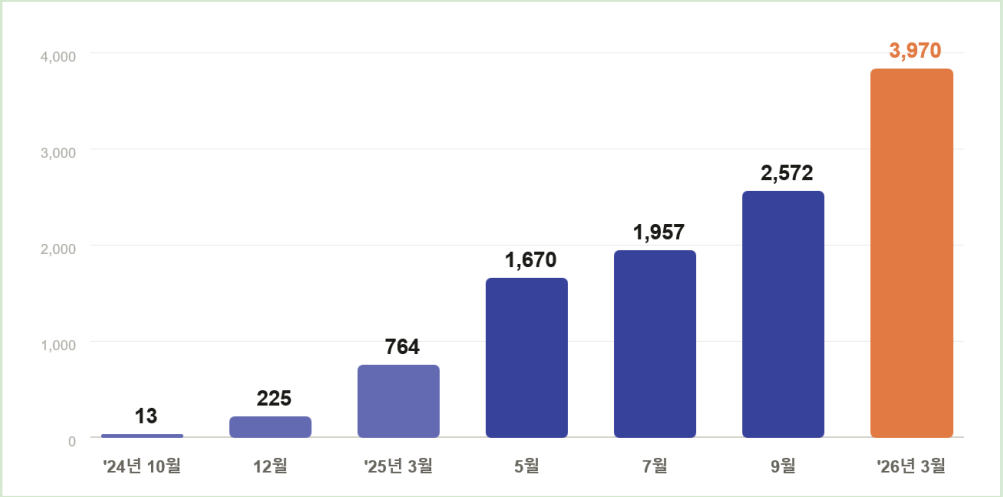
<그림 4> Google Search Console 색인 생성 요청 화면

마지막으로 기술적 정비도 병행하였다. Googlebot이 논문 페이지를 학술 콘텐츠로 안정적으로 해석할 수 있도록 Google Scholar에서 권장하는 Highwire Press 메타태그를 추가하고, 불필요한 HTML 정보를 줄였으며, 원문 PDF URL은 실제 PDF 파일임을 알 수 있는 형태(~.pdf)로 수정하였다. 사이트맵에는 데이터 수정일이 반영되도록 바꾸고, Google Scholar 색인이 불필요한 KSP의 검색결과 목록 페이지에는 noindex 메타태그를 삽입하였다. 이는 Googlebot이 웹사이트를 읽되, 색인해야 할 학술 자료와 색인하지 않아도 되는 목록 페이지를 구분하도록 돕기 위한 조치였다(〈표 2〉 참조).

[표 2] KSP Google Scholar 색인 문제점 및 수행 조치

구분	확인된 문제	수행 조치
콘텐츠 품질	초록·출판사 등 필수 필드 누락, 초록·저자명 오기입, 중복 입력	초록 1,800여 건과 저자명 800여 건 수정, 중복논문 150여 건 검출·삭제, GPT·웹크롤링·API로 13,000여 건 초록 반입
	OA 정보 누락·오류로 인한 원문 PDF 제공 비율 저조	OA 정보 1,200여 건 수정, 원문 제공 논문 비율 4%에서 11%로 확대
	인사 DB 영문명과 실제 논문 저자명 불일치	20만 건의 저자명 매핑을 정비하고, KSP에서는 논문 저자명을 사용
웹사이트	필수 메타태그 누락, HTML 내 불필요 정보 포함	Google Scholar 필수 메타태그 추가, 불필요 정보 삭제·정비
	동적 URL, 사이트맵 수정일 미반영, 삭제 논문 빈 화면, 모달 중복 페이지	원문 PDF URL을 .pdf 형태로 수정, 데이터 수정 시 사이트맵 수정일 반영, 모달 페이지 noindex 및 canonical link 추가
	언어 설정에 따라 제목·저자명이 한글/영문으로 바뀌고, Google Scholar 비색인 자료가 함께 노출	성과물 자체의 언어를 기준으로 제목·저자명을 표기하도록 변경

프로젝트를 시작한 뒤 약 6개월 후 Google Scholar 색인 수에 가시적인 변화가 나타나기 시작했다. 10여 건에 불과했던 색인 수는 2024년 12월 말 200여 건을 넘어섰으며, 2025년 3월에는 KSP 논문이 Google Scholar에 secondary link로도 색인되기 시작하였다. 또한 figshare에 공개된 전 세계 기관리포지터리 Google Scholar 색인 수 데이터셋 기준으로 KSP의 순위는 2024년 3,056위(272건)에서 2026년 1,105위(3,970건)로 상승하였다(Aguillo 2025a, 2025b, 2026). Google Search Console에서 확인되는 Googlebot의 크롤링 요청도 일일 300여 건 이하에서 2,000~3,000건 수준으로 증가하였다(〈그림 5〉 참조).



<그림 5> KSP Google Scholar 색인 수 변화(2024.10.-2026.3.)

물론 이 수치를 단일 원인의 결과로 해석해서는 안 된다. 표기 정합성 개선, 메타태그 보강, 초록 반입, 원문 공개, 중복 제거, 색인 요청이 동시에 이루어졌기 때문이다. 다만 이 경험을 통해 분명해진 점은 있다. Google의 비공개 알고리즘을 추정하거나 시중의 SEO(Search Engine Optimization) 도구들을 이용하기보다는 Google 공개 가이드라인이 요구하는 신뢰 조건을 하나씩 충족시키는 접근이 현실적이라는 점이다. 그리고 그 신뢰 조건의 대부분은 결국 메타데이터 품질과 연결되어 있었다.

이번 개선 작업은 Google Scholar 색인 문제를 단순한 검색엔진 최적화 문제가 아니라, 기관리포지터리가 축적해 온 메타데이터 품질을 다시 점검하는 계기가 되었다. 특히 수작업 입력 과정에서 누적된 메타데이터 오류는 검색엔진이 KSP 자료를 정확한 학술 콘텐츠로 인식하는 데 장애가 될 수 있음을 확인하였다. 동시에 ChatGPT를 활용한 바이브코딩은 문제를 빠르게 진단하고 반복적으로 개선하는 데 도움이 되었다. 프로그래밍을 전문적으로 배운 사람이 아니더라도, 해결하려는 업무 문제와 데이터 구조를 알고 있다면 자동화 코드를 만들고 검증할 수 있다는 가능성도 확인하였다.

PDF 논문 초록 자동 추출

초록 누락과 원문 기반 추출

2024년 Google Scholar 색인 개선 작업에서 가장 자주 부딪힌 문제는 초록 누락이었다. 이는 단순히 메타데이터가 일부 비어 있는 문제가 아니었다. Google Scholar가 학술 논문 페이지를 식별하고 색인 대상으로 판단할 때, 제목·저자·발행정보와 함께 초록은 논문의 내용을 확인하는 핵심 메타데이터로 작동한다. 따라서 초록이 없는 논문은 검색 결과에서 잘 보이지 않는 수준을 넘어, 애초에 Google Scholar의 기본적인 색인 요건을 충분히 충족하지 못해 기관의 연구성과로 노출될 기회 자체를 잃을 수 있었다.

앞서 설명한 프로젝트에서 2024년 약 1만 3천여 건의 초록을 추가 반입하기 전까지 KSP의 Google Scholar 색인 논문 수는 두 자릿수 수준에 머물렀다. 그러나 초록 반입 이후 약 1년 남짓한 기간 동안 색인 건수는 수천 건 규모로 증가하였다. 물론 이 변화가 초록만의 효과라고 단정할 수는 없지만, 초록이 Google Scholar에서 논문을 발견하고 색인하는 데 영향을 미치는 핵심 조건이라는 판단은 충분히 가능하였다. 다시 말해 초록 누락은 개별 논문의 정보 부족 문제가 아니라, 기관리포지터리에 축적된 연구성과가 외부 학술 검색 환경에서 보이지 않게 되는 구조적인 장애 요인이었다.

그런데 2024년에 수행한 개선 작업에도 불구하고 초록 미보유 논문은 여전히 많았다(〈표 3〉 참조). 2025년 기준 KSP의 학술지·학술대회 논문 38,242건 중 초록 미보유 건수는 14,341건이었다. 특히 학술대회 논문의 경우 전체 25,183건 중 12,307건에 달했다. 초록 미보유 논문들은 상용 학술 DB의 커버리지 밖에 있는 회색문헌 성격이 강해, KCI·Scopus·Web of Science 같은 외부 출처에서 초록을 일괄 반입하는 방식이 통하지 않았다. 또한 수작업으로 초록을 직접 입력하거나 개별적으로 반입하는 것도 비용과 시간이 과도하게 들어 현실적이지 않았다.

[표 3] KSP 논문 초록 보유 현황(2025.3.5. 기준)

구분	초록 보유(건)	초록 미보유(건)	합계(건)
학술지 논문	11,025	2,034	13,059
학술대회 논문	12,876	12,307	25,183
합계	23,901	14,341	38,242

결국 필요한 것은 외부 DB에서 초록을 가져오는 방식이 아니라, PDF 원문 안에 이미 존재하는 초록을 자동으로 찾아내는 방식이었다. 논문 첫 페이지를 텍스트화한 뒤, 그 안에서 초록에 해당하는 부분만 정확히 추출할 수 있다면 대량의 초록 미보유 논문을 훨씬 효율적으로 보완할 수 있었다. 처음에는 비교적 단순한 방식으로 접근했다. PDF에서 텍스트를 추출한 뒤 정규표현식으로 초록을 나타내는 라벨 텍스트 'Abstract'과 논문 본문이 시작되는 제목 텍스트 'Introduction' 사이를 잘라내면 될 것이라고 생각했다. 일부 논문에는 이 방식이 통했지만, 범위를 넓히자 바로 한계가 드러났다. 초록 라벨이 'Abstract', 'Summary' 등으로 제각각이었고, 초록 라벨 없이 첫 문단이 바로 시작되는 경우도 있었다. 'Keywords'가 'Abstract' 앞에 있는 경우, 학술대회 논문처럼 본문 시작 표시가 없는 경우도 섞여 있었다. 규칙을 하나 추가하면 한 유형은 맞지만 다른 유형에서 문제가 발생하는 현상이 반복되었다.

이 지점에서 LLM(Large Language Model)을 활용할 가능성을 확인하였다. LLM은 단순히 특정 문자열을 찾는 방식이 아니라, 주어진 문맥 안에서 문단의 기능과 구조를 판단할 수 있다. 논문에서 초록은 전체 내용을 요약하는 독립 단락이며, 서론이나 본문과는 기능적으로 구분된다. 따라서 LLM의 문맥 이해와 문서 구조 인식 능력은 초록 추출 과업과 자연스럽게 연결된다. 즉, Abstract라는 단어가 있는지 여부만 보는 것이 아니라, 해당 문단이 논문의 요약 역할을 하는지, 본문 서론에 해당하는지, 키워드나 저자정보 같은 다른 영역인지를 함께 판단할 수 있다는 점에서 정규표현식 기반 방식의 한계를 보완할 수 있었다.

다만 이번 작업에서 LLM을 사용한 목적은 논문 내용을 새로 요약하는 것이 아니었다. 기관리포지터리 메타데이터로 활용하려면 초록은 원문 PDF 안에 실제로 존재하는 문장을 기반으로 해야 했다. LLM이 그럴듯한 요약문을 새로 만들어낸다면, 이는 메타데이터 보완이 아니라 원문에 없는 정보를 생성하는 문제가 된다. 따라서 이번 작업의 핵심은 LLM을 '요약 생성 도구'가 아니라 '원문 기반 초록 식별·추출 도구'로 활용하는 것이었다. 초록이 존재하는 경우에는 해당 부분만 추출하고, 초록이 존재하지 않는 경우에는 임의로 내용을 만들어내지 않고 '초록 없음'을 의미하는 응답값 'None'을 반환하도록 지시하였다.

이러한 문제의 해결 방법을 고민하던 중 2025년 OpenAI는 GPT-4.1, GPT-4.1 mini, GPT-4.1 nano를 API 모델로 발표하였다(OpenAI 2025a). 이 모델군은 긴 컨텍스트 처리, 지시 준수, 코딩 성능 개선을 특징으로 제시되었고, 실제 업무 테스트에서도 이전 모델보다 원문 기반 추출 지시를 안정적으로 따르는 편이었다. 2024년에는 gpt-3.5를 활용하여 단순히 프롬프트만으로 일부 논문의 초록을 추출하는 수준에 머물렀으나, 2025년에는 새로 출시한 GPT-4.1 mini 모델을 활용하여 논문 성과 입력 업무에 실제 적용 가능한 초록 추출 코드를 개발하는 것을 목표로 삼았다. 구체적으로는 논문 PDF의 초기 페이지를 OCR로 텍스트화하고, 그 결과에서 기존에 존재하는 초록 부분만 추출하도록 설계하였다.

그러나 실제 적용과정에서는 PDF에서 초록만을 추출하는 작업은 생각보다 훨씬 까다로웠다. 텍스트 레이어가 없는 이미지 PDF가 많았고, 텍스트 PDF라 하더라도 편집상 실수로 보이지 않는 숨은 텍스트가 남아 있는 경우가 많았다. 또한 PDF의 레이아웃은 논문마다 달랐고, 학술지 논문과 학술대회 논문의 형식도 일정하지 않았다(〈그림 6〉 참조). 결국 초록 자동 추출은 단순히 PDF에서 문자열을 읽어 오는 문제가 아니었다. 다양한 PDF 형식과 불완전한 텍스트 추출 결과 속에서, 원문에 실제로 존재하는 초록만을 안정적으로 식별해야 하는 문제였다.

〈그림 6〉 논문의 다양한 레이아웃

이에 실제 업무에 적용할 수 있도록 추출 절차를 재설계하였다. 초록 추출을 '초록 시작 문자열과 초록 끝 문자열을 식별해 자르는 문제'가 아니라, '문서에서 초록이라는 구간을 판별하는 문제'로 보았다. 또한, 다양한 OCR 제품을 테스트하여 당시 상대적으로 논문 레이아웃 인식 능력이 우수한 Google Cloud Vision API를 사용하기로 하였다. 그리고 ChatGPT를 통한 바이브코딩으로 초록 추출 파이프라인을 세 단계로 설계하였다(〈표 4〉 참조). 첫째, 이미지 PDF와 텍스트 PDF를 별도로 구분하지 않고 모든 PDF의 첫 2페이지를 OCR 처리하여 텍스트를 추출하였다. 둘째, gpt-4.1-mini API를 호출하여 OCR 텍스트를 입력하고 아래 프롬프트와 같이 초록을 원문 그대로 반환하도록 지시하였다(〈표 5〉 참조). 초록이 없거나 불확실한 경우에는 None을 반환하게 하여, 모델이 임의로 요약문을 생성하지 못하도록 하였다.

[표 4] OCR·LLM 기반 초록 추출 파이프라인

단계	주요 기능	입력	출력	사용 기술
1	PDF 전처리 및 OCR	논문 PDF 1~2페이지	페이지별 텍스트	Google Cloud Vision API
2	LLM 기반 초록 추출	OCR 텍스트 전체	추출 초록 또는 None	gpt-4.1-mini API
3	성능 평가 및 로그 기록	추출 초록 + 공식 초록	유사도 지표, 분류 결과	유사도 계산, 통계 분석

[표 5] 초록 추출 프롬프트

시스템 프롬프트(sys_msg)

You are an AI assistant that EXTRACTS exactly and ONLY the main body of the ABSTRACT from scientific documents **only if appropriate**. The document may NOT have an abstract at all. Use internal step-by-step reasoning but output only the abstract verbatim, same language.

STRICT EXTRACTION RULES:

1. Only extract a single, contiguous block of text that is explicitly presented as the abstract. Do NOT select or merge sentences from different locations. Do NOT combine, summarize, rephrase, or invent any content.
2. If the abstract is extremely short (e.g., just one or two sentences), output exactly those sentences and absolutely nothing else. DO NOT add, supplement, or concatenate sentences from other parts of the text.
3. The 'Introduction' section, or any section labeled 'Introduction', '1. Introduction', '서론', '개요', or similar, is NOT an abstract and must NEVER be extracted as the abstract under ANY circumstances.
4. The text provided is from ONLY the first 2 pages of the document. You must return "None" if it is determined that the text has no abstract with a probability greater than 50%.
5. If the abstract is missing or absent, or you are unsure, return 'None'.

NEGATIVE EXAMPLE:

- If there is no section labeled as abstract, and no block clearly marked or separated as abstract, return 'None'. Do NOT extract an introduction.
- If the only available text is an introduction or a body paragraph, return 'None'.

유저 프롬프트(user_msg)

The text below is from the first 2 pages of the document. First, internally reason why a block is the abstract (do not show your reasoning), then output ONLY the pure abstract body without any headings, keywords, or index terms. If there is no clear abstract, respond with 'None'. Here is the text:{ocr_text}

약 한 달간의 초록 추출 테스트와 프롬프트 개선을 반복하여 프롬프트의 핵심을 '연속된 단락만 추출할 것, 서로 떨어진 부분을 이어 붙이거나 요약·재작성하지 말 것, Introduction·서론·개요 등 서론에 해당하는 구간은 초록으로 간주하지 말 것, 초록이 없거나 불확실하면 'None'을 반환할 것'으로 확정하였다. 또한, 사용자 메시지에는 입력 텍스트가 첫 2페이지에서 추출된 것임을 밝히고, 최종 출력에는 제목·키워드·색인어를 제외한 초록 본문만 반환하도록 지시하였다.

평가 결과와 한계

초록 추출의 정확성 평가는 2025년 6월 기준 KSP 보유 논문 중 Scopus에 색인된 영문 논문을 대상으로 했다. KSP 메타데이터와 Scopus 레코드를 매칭하여 PDF 원문과 Scopus 초록을 모두 확보할 수 있는 경우만 선별했고, 국문 논문, PDF 원문 미보유 자료, Scopus 초록이 비어 있는 레코드는 제외하였다. 이 절차를 거쳐 총 13,563편의 PDF–초록 쌍을 최종 데이터셋으로 구축하였다. 데이터셋의 게재연도별 분포는 〈표 6〉과 같으며, 자료 유형별로는 학술대회 자료가 7,496편(55.3%), 학술지 논문이 6,067편(44.7%)으로 학술대회 자료가 다소 높은 비중을 차지하였다(〈표 7〉 참조).

[표 6] 데이터셋의 게재연도별 분포

구분	건수	비율
~2000년	52	0.4%
2001~2010년	4,705	34.7%
2011~2020년	6,785	50.0%
2021~2025년	2,021	14.9%
합계	13,563	100.0%

[표 7] 데이터셋의 자료 유형별 분포

구분	건수	비율
학술대회	7,496	55.3%
학술지	6,067	44.7%
합계	13,563	100.0%

한편, 전체 13,563편 가운데 PDF 1~2페이지 내에 초록이 존재하지 않는 논문이 89편(약 0.7%) 확인되었다. 이들 논문은 초록이 3페이지 이후에 위치하거나, 별도의 초록 없이 곧바로 서론(Introduction)으로 시작하는 형식을 취하고 있었다. 본 연구에서 구현한 파이프라인은 PDF의 첫 2페이지에서 추출한 텍스트만을 LLM 입력으로 사용하도록 설계되어 있으므로, 이러한 논문에 대해서는 LLM이 초록을 추출하지 않고 '초록 없음(None)'을 반환하는 것이 기대되는 동작으로 간주하였다.

파이프라인을 통해 추출한 초록과 Scopus 초록을 비교한 결과, 완전 일치는 12,261건(90.4%), 유사 일치

는 1,122건(8.3%), 불일치는 180건(1.3%)이었다(〈표 8〉 참조). 완전 일치는 Scopus 공식 초록과 동일해 리포지터리에 그대로 반입해도 되는 수준이다. 유사 일치는 수식, 띄어쓰기, 하이픈, 온점 같은 표면 형식에서만 차이가 있는 경우로, 검색·색인 관점에서는 대부분 실무적으로 활용 가능한 품질로 보았다. 정량 평가에서는 문자 수준 유사도와 단어 수준 정밀도(precision), 재현율(recall), F1 점수를 기본 지표로 삼았다. 또한 추출된 초록이 OCR 텍스트 안에 그대로 포함되는지도 확인하여, 원문에 없는 문장을 새로 만들어낸 사례를 걸러내고자 했다.

[표 8] 초록 추출 정확도와 오류 유형

분류	건수	비율	설명
완전 일치	12,261	90.4%	공식 초록과 동일
유사 일치	1,122	8.3%	수식, 띄어쓰기, 하이픈, 온점에서 차이
불일치	180	1.3%	초록 추출 실패 또는 무관 텍스트 포함

불일치 사례는 크게 '초록 추출 실패'와 '무관 텍스트 포함'으로 구분되었다(〈표 8〉 참조). 초록 추출 실패는 입력 범위에 초록이 존재하지만 파이프라인이 이를 찾지 못해 None을 반환하거나 초록의 일부만 추출한 경우를 말한다. 무관 텍스트 포함은 초록을 추출했으나 서론 제목, 저자 정보 또는 다른 본문 문장이 함께 포함된 경우를 의미한다. 이는 원문에 초록이 없어 None을 정확하게 반환한 사례와는 구분된다. 이러한 불일치 사례의 상당수는 LLM 자체의 오류라기보다 OCR 단계의 구조 인식 오류에서 비롯되었다. 대표적으로 초록이 왼쪽 단에 있고 오른쪽 단 상단에서 I. Introduction이 시작되는 2단 레이아웃에서, OCR이 초록 첫 문장 직후 오른쪽 단의 제목을 이어 붙인 뒤 다시 왼쪽 단의 초록 문장을 인식하는 현상이 일부 나타났다. 이처럼 초록 중간에 서론 제목이 삽입되면 프롬프트에서 Introduction을 제외하도록 지시하더라도 문장의 연속성이 훼손되어 LLM이 초록의 경계를 정확하게 판단하기 어렵다.

[표 9] 유사 일치와 불일치 사례의 대표 증상

사례 유형	대표 증상
유사 일치	수식·그리스 문자·점·지수 표기가 깨지지만 문장 의미는 유지
2단 레이아웃 불일치	초록 중간에 "I. Introduction"이 삽입되는 등 레이아웃 파싱 실패
수식 중심 논문	OCR 처리에서 기호 왜곡

이 연구의 핵심 성과는 LLM을 활용해 원문 PDF 안에 실제로 존재하는 초록을 90% 이상의 정확도로 자동 추출할 수 있음을 확인했다는 점이다. 기존에는 초록 누락 논문을 보완하려면 외부 DB에 의존하거나 사람이 직접 확인해 입력해야 했지만, OCR과 LLM을 결합한 방식은 PDF 원문 자체를 기반으로 대량의 초록을 자동 식별·추출할 수 있는 가능성을 보여 주었다. 특히 LLM을 요약 생성 도구가 아니라 원문에 존재하는 초록 구간을 판별하는 추출 도구로 제한하고, None 반환 조건과 원문 그대로 추출하는 규칙을 명확히 부여함으로써 메타데이터 품질을 훼손할 수 있는 환각 위험을 줄였다. 다만 레이아웃과 수식의 오인식과 같은 문제는 여전히 별도 보정이나 수동 검토가 필요한 영역으로 남았다.

AI 에이전트 활용 학술성과 입력·검증 자동화

학술성과 검증 자동화의 설계

2026년 ETRI 도서관은 기존의 초록 자동 추출 작업에서 한 단계 더 나아가, 학술성과 입력·검증 업무 전반을 AI와 RPA(Robotic Process Automation) 기반으로 자동화하는 프로젝트를 추진하고 있다. 본 프로젝트는 현재 진행 중인 과제로, 여기서는 그중 학술대회 발표 사실 확인과 프로그램 정보 수집 자동화 사례를 중심으로 소개하고자 한다. 이 절에서 '학술성과'는 학술지 논문과 학술대회에서 발표된 논문·확장초록·발표초록·포스터 등의 성과물을 포괄하는 의미로 사용된다.

ETRI 도서관의 학술성과 입력 업무는 연구자가 학술지 논문 또는 학술대회 성과물에 관한 정보와 원문 파일을 포함하여 결재문서를 기안하면, 담당자가 이를 확인한 뒤 관련 메타데이터를 전자도서관에 입력하고 기관리포지터리에 공개하는 방식으로 이루어진다. 입력된 학술성과 정보는 기관의 연구성과 관리와 직원 내부 평가에도 활용되기 때문에, 정확한 메타데이터 입력과 검증이 매우 중요하다. 학술성과 입력데이터의 수집원과 활용흐름을 도식화하면 〈그림 7〉과 같다.

특히 학술대회 성과물은 학술지 논문처럼 출판사 페이지나 DOI 기반 메타데이터가 명확하게 제공되지 않는 경우가 많으며, full paper 형태의 프로시딩 논문부터 확장초록, 발표초록, 포스터, 프로그램 수록 정보까지 형식이 다양하다. 그러나 학술대회 프로그램, 발표 일정표, 공식 웹페이지 등을 통해 발표 사실이 확인되면 기관 성과로 인정될 수 있다. 따라서 담당자는 연구자가 제출한 정보와 원문 PDF를 확인하는 것에 그치지 않고, 실제로 해당 성과물이 학술대회에서 발표되었는지 확인할 수 있는 공식 근거 자료를 찾아야 한다.

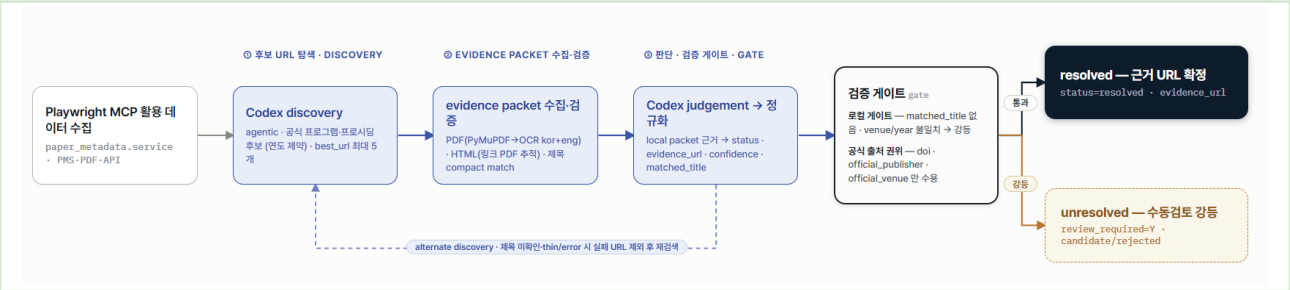
<그림 7> 학술성과 입력 데이터의 수집원과 활용 흐름

그러나 학술대회 웹사이트는 구조와 제공 방식이 매우 다양하다. 어떤 학술대회는 별도 프로그램 페이지를 운영하지만, 일부 국내 학술대회는 독립적인 학술대회 웹사이트 없이 학회 홈페이지 공지사항에 프로그램 파일을 첨부하는 방식으로 운영된다. 경우에 따라서는 텍스트 추출이 어려운 이미지 PDF나 스캔 파일 형태로 프로그램이 제공되기도 한다. 이처럼 발표 근거 자료의 위치와 형식이 일정하지 않기 때문에, 담당자가 개별 학술대회 사이트를 직접 방문하여 프로그램 정보를 찾고 논문 제목을 대조하는 작업에는 많은 시간과 노력이 소요된다.

이러한 업무 부담을 줄이기 위해 ETRI 도서관은 OpenAI의 코딩 에이전트인 Codex를 활용하여 학술대회 프로그램 정보를 자동으로 탐색하는 PubFinder(가칭)를 개발하고 있다(〈그림 8〉 참조). Codex는 사용자가 자연어로 작업을 지시하면 코드 작성, 수정, 실행, 오류 점검 등을 보조하는 AI 기반 개발 도구이다. 특히 사용자의 작업 폴더 안에서 자동화 코드를 읽고 수정하며 필요한 명령을 실행할 수 있어, Python 기반 업무 자동화 코드를 빠르게 작성하고 개선하는 데 활용할 수 있다(OpenAI 2025b, 2026).

PubFinder의 핵심 구조는 검색과 판단을 하나의 단계로 처리하지 않고, 역할이 다른 두 개의 Codex 에이전트로 분리했다는 점이다. 먼저 discovery agent는 웹검색을 통해 연구자가 제출한 논문 제목, 학술대회명, 발표처 등의 정보를 바탕으로 공식 학술대회 프로그램 페이지, 프로그램 PDF, 발표 일정표 등 후보 URL을

탐색한다. 이 단계에서는 가능한 한 많은 웹페이지를 무작정 수집하는 것이 아니라, 공식 프로그램·프로시딩 후보와 같이 학술성과 검증에 활용될 가능성이 높은 URL을 우선적으로 선별한다.



<그림 8> PubFinder의 후보 URL 탐색·근거 수집·검증 흐름

다음으로 수집된 후보 URL에 대해서는 evidence packet을 구성한다. evidence packet에는 후보 페이지의 URL, PDF 텍스트 추출 결과, 필요한 경우 OCR 결과, HTML 본문, 제목 매칭에 필요한 요약 정보 등이 포함된다. 이처럼 판단에 필요한 근거 자료를 하나의 묶음으로 정리한 뒤, 별도의 Codex judgement agent가 이를 검토한다. judgement agent는 해당 근거 자료가 실제로 제출 논문의 발표 사실을 뒷받침하는지 판단하고, 근거 URL, 매칭된 제목, 신뢰도, 검증 상태 등을 구조화된 값으로 정리한다.

이 구조는 AI가 검색 결과를 곧바로 정답으로 확정하지 않도록 하기 위한 장치이다. discovery agent는 후보를 찾는 데 집중하고, judgement agent는 후보가 실제 근거로 사용될 수 있는지를 따로 판단한다. 즉, '찾기'와 '판단'을 분리함으로써 검색 과정에서 발생할 수 있는 오류와 검증 과정에서 발생할 수 있는 오류를 구분하고, 사람이 확인해야 하는 사례를 별도로 걸러낼 수 있도록 설계하였다. 이는 생성형 AI를 업무에 적용할 때 필요한 human in the loop 원칙을 운영 절차 안에 반영한 것이다.

또한 최종 단계에는 검증 게이트를 두어, judgement agent의 판단 결과를 그대로 수용하지 않고 규칙 기반으로 한 번 더 점검하도록 했다. 예를 들어 논문 제목이 충분히 매칭되지 않거나, 학술대회명·발표연도 정보가 제출 정보와 충돌하는 경우에는 자동 확정하지 않고 수동 검토 대상으로 분류한다. 반대로 공식 학술대회 웹사이트, 공식 프로그램 PDF, DOI 또는 출판사 공식 페이지처럼 신뢰할 수 있는 출처에서 논문 제목과 발표 정보가 확인되는 경우에는 근거 URL을 확정한다. 이처럼 AI 판단과 규칙 기반 검증을 함께 적용함으로써 자동화의 효율성과 검증 업무의 신뢰성을 동시에 확보하고자 했다.

적용 범위와 확장 계획

현재 개발 과정에서는 Python 기반 자동화 코드와 Codex의 코드 작성·실행 보조 기능을 결합하여, 학술대회 프로그램 페이지나 공식 PDF를 찾는 반복 작업을 자동화하고 있다. 담당자가 직접 검색엔진을 이용해 학술대회명을 검색하고, 여러 페이지를 열어 프로그램 파일을 찾고, PDF 안에서 논문 제목을 확인하던 과정을 discovery agent, evidence packet, judgement agent, 검증 게이트의 흐름으로 재구성한 것이다.

해외 학술대회의 경우 별도 웹사이트나 프로그램 페이지가 체계적으로 운영되는 경우가 많아 비교적 높은 비율로 공식 프로그램 정보를 확인할 수 있었다. 반면 국내 학술대회의 경우 독립적인 학술대회 웹사이트가 없거나, 학회 홈페이지 공지사항에 프로그램 파일을 첨부하는 방식이 많아 해외 학술대회와 동일한 방식으로 처리하기 어렵다. 또한 일부 프로그램 파일은 이미지 PDF 형태로 제공되어 일반적인 텍스트 검색만으로는 논문 제목을 확인하기 어려운 경우도 있다. 이에 따라 국내 학술대회에 대해서는 공지사항 탐색, 첨부파일 다운로드, PDF 텍스트 추출, OCR 처리 등 다양한 예외 상황을 고려한 보완 기능이 필요하다.

향후에는 학술대회별 웹사이트 구조와 자료 형식의 차이, 공지사항 첨부파일, 이미지 PDF 등 다양한 예외 상황에 대응할 수 있도록 프로그램 탐색과 검증 기능을 고도화할 계획이다. 아울러 근거자료와 판단 결과를 기록하고 불확실한 사례는 담당자가 최종 확인하는 구조를 강화하여 자동화의 정확성과 신뢰성을 높이고자 한다.

나오며

도서관의 AI 활용은 기술 도입의 문제가 아니라, 기존 업무의 문제를 정의하고 개선하는 과정에서 구체화된다. 이번 KSP 사례에서 가장 먼저 효과를 낸 것은 제목, 저자명, 초록, DOI, 원문 PDF, 메타태그와 같이 도서관이 오래전부터 다루어 온 메타데이터를 정확하게 정비하는 일이었다. Google Scholar 색인 개선, PDF 초록 자동 추출, 학술성과 입력·검증 자동화는 서로 다른 프로젝트처럼 보이지만, 모두 신뢰할 수 있는 메타데이터를 만들고 이를 반복적으로 확인하는 문제에서 출발했다는 공통점을 가진다.

자동화의 필요성도 이 지점에서 분명해진다. 연구성과 관리 업무에는 자료마다 예외가 많고, 여전히 사람이 판단해야 할 부분이 남아 있다. 그러나 검색, 대조, OCR, API 조회, 로그 기록, 오류 후보 선별처럼 반복 가능한 절차는 RPA, AI 에이전트가 충분히 분담할 수 있다. 자동화는 사서와 담당자의 판단을 대체하는 것이 아니라, 사람이 판단해야 할 사례를 더 빨리 찾고 그 판단에 필요한 근거를 더 잘 정리하도록 돕는 역할을

한다. 따라서 도서관의 자동화는 '전부 자동 처리'를 목표로 하기보다, 자동 수집–자동 검증–사람의 최종 판단까지 이어지는 구조로 설계될 필요가 있다.

AI는 전문 IT팀이나 대규모 예산이 없는 도서관에도 업무 개선의 현실적인 가능성을 열어 준다. 담당자가 전산에 대한 지식이 풍부하지 않더라도 해당 도메인에 대한 이해도가 충분하다면, Codex, Claude Code와 같은 코딩 에이전트를 활용해 작은 자동화 스크립트를 만들고, 실패 사례를 확인하며 코드를 고치고, 반복 업무를 단계적으로 줄여 나갈 수 있다. 중요한 것은 처음부터 완성도 높은 시스템을 구축하려 하기보다, 도서관이 매일 업무 흐름 속에서 자동화할 수 있는 지점을 찾아 하나씩 개선해 나가는 일이다. 그렇게 축적된 자동화는 이용자에게 접근성과 활용도가 높은 연구성과를 제공하고, 연구자에게 더 정확한 성과 관리를 지원하며, 도서관에는 더 지속 가능한 업무 방식으로 자리 잡을 것이다.

사사

본 글에서 소개한 활동의 일부는 2025~2026년 국가정책정보협의회 분과위원회, 2026년 한국전문도서관협의회 제1차 연구·조사 과제 및 2026년 한국전자통신연구원 성과창출지원사업(26OE1200)의 지원을 받아 수행되었다.

참고문헌

Aguillo, I. F. "Transparent ranking of institutional open access repositories according to number of items indexed by Google Scholar 2018–2024" *figshare*, 2025a, https://doi.org/10.6084/m9.figshare.28295870

Aguillo, I. F. "Transparent ranking of repositories 2025" *figshare*, 2025b, https://doi.org/10.6084/m9.figshare.29099234

Aguillo, I. F. "Transparent ranking of repositories 2026" *figshare*, 2026, https://doi.org/10.6084/m9.figshare.31833976

Google Scholar. "Inclusion guidelines for webmasters." 2026, https://scholar.google.com/intl/en/scholar/inclusion.html

OpenAI. "Introducing GPT-4.1 in the API." 2025a, https://openai.com/index/gpt-4-1/

OpenAI. "Introducing Codex." 2025b, https://openai.com/index/introducing-codex/

OpenAI. "Codex CLI." 2026, https://developers.openai.com/codex/cli

Westin, M. "Google Scholar indexing for institutional repositories" *Canadian Association of Research Libraries*, 2021, https://www.carl-abrc.ca/wp-content/uploads/2021/01/Google_Scholar_webinar_Jan2021.pdf

KSLA 한국전문도서관협의회
Korea Special Library Association

Korea Special Library Association

KSLA Insight

발 행 일 2026년 8월 22일

발 행 인 황재영

편집위원 KSLA 편찬위원회
위원장 황혜경 **부위원장** 이경미 **위원** 신정은, 김은진, 박지은, 윤지연, 황선경, 권성국

발 행 처 한국전문도서관협의회

I S S N 3022-6198