

생성형 AI와 정보조직

발제: 문헌정보학전공 박사수료 박미진

발제논문

- 김선욱, 이해경 (2025). 서지 메타데이터 자동 생성 성능 비교 연구 - 국내외 대규모 언어모델을 중심으로 -. 한국비블리아학회지, 36(4), 303-331.
<http://dx.doi.org/10.14699/kbiblia.2025.36.4.30>
- 이해경, 이용구 (2025). 생성형 AI를 활용한 지식그래프 구축: 구술생애사를 중심으로. 정보관리학회지, 42(3), 237-255.
<http://dx.doi.org/10.3743/KOSIM.2025.42.3.237>
- 이정미 (2026). 생성형 AI 기반 메타데이터 자동 생성의 편향과 공정성에 대한 정보학적 분석. 한국비블리아학회지, 37(1), 467-488.
<http://dx.doi.org/10.14699/kbiblia.2026.37.1.467>

서지 메타데이터 자동 생성 성능 비교 연구*

- 국내외 대규모 언어모델(Large Language Model)을 중심으로 -

A Comparative Study of Automatic Bibliographic Metadata Generation Performance: Focusing on Domestic and International Large Language Models (LLMs)

김 선 욱 (SeonWook Kim)**

이 혜 경 (Hyekyung Lee)***

초 록

본 연구는 국내 소버린 AI와 글로벌 LLM을 비교하여 도서관 메타데이터 자동 생성을 위한 활용 가능성 파악하는 것을 목적으로 한다. 이를 위해 GPT, Gemini, Grok, HyperCLOVA, EXAONE, AX 총 6종의 생성형 대규모 인공지능 언어모델을 대상으로 국내외 도서 40권의 MARC 레코드를 생성하게 하고, 완전성, 정확성, 규칙성의 세 가지 척도에 따라 필드 단위 성능을 평가하였다. 분석 결과, 첫째, GPT, Gemini, Grok의 글로벌 LLM 3종은 필드 누락이 적고 지시기호, 식별기호 등 형식 요소를 비교적 안정적으로 처리하여, 전반적으로 국내 소버린 AI 모델보다 높은 성능을 보였으나, 국내 도서로 전환될 경우, 필드 구성과 작성방식 등에서 오류를 보이며 성능이 저하되는 경향이 확인되었다. 둘째, HyperCLOVA, EXAONE, AX의 국내 소버린 AI 모델은 MARC21 및 KORMARC 기술 모두에서 전반적인 성능 수준이 낮았고, 국내 도서에 대해서도 뚜렷한 성능 향상을 보이지 못하였다. 셋째, 필드별로는 표제와 책임표시사항(245) 처리 대부분의 모델이 비교적 안정적으로 생성하는 영역이 있는 반면, 총서사항(490/830)이나 기본표목의 설정 등 규칙 의존도가 있는 필드에서 모델 간 성능 격차 및 MARC21의 총서 처리 방식을 KORMARC에 기계적으로 적용하는 등 서지 작성 규칙 구조에 대한 이해 부족을 드러냈다. 이에 따라 현시점에서 생성형 인공지능을 도서관 메타데이터 업무에 도입할 때, 전면적인 자동목록 도구로의 전환 보다, 서지 레코드 초안 생성과 오류 탐지, 보완을 지원하는 보조 도구로 활용하는 것이 타당함을 시사하며, 아울러 국내 소버린 AI의 성능 안정성을 확보하기 위해서는 KORMARC를 포함한 국내 서지 데이터를 기반으로 한 체계적인 학습이 필요할 것으로 보였다. 또한 도서관용 소버린 AI를 구축하기 위해서는 학습 데이터의 선별이 주요한 과제로 요구된다.

연구의 배경 및 목적

- 대규모 언어 모델(LLM), 소버린 AI의 등장과 영향력 확대
- 뉴노멀에 대비하기 위한 인공지능 서비스 관리 능력을 함양한 사서 양성 필요
- 도서관에서 적재적소에 인공지능 기반 서비스를 배치 및 운용할 수 있는 역량 필요

KORMARC과 MARC21 작성 규칙에 부합하는 서지 메타데이터 생성 품질의 현황과 한계 파악

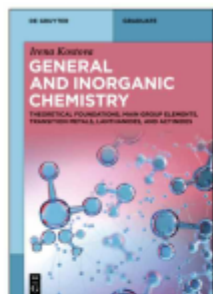
- 글로벌 상용 LLM 3종 (GPT, Gemini, Grok) vs. 국내 소버린 AI 3종 (EXAONE, A.X, HyperCLOVA)

→ 도서관 현장에서 합리적인 생성형 AI 모델 선택과 업무 프로세스 개발을 위한 지침을 마련하는 데 기초자료 제공

연구방법

1. 데이터 수집 및 처리

- 한국어 / 영어 신간도서(2025년 7월 4주차 ~ 8월 1주차) 중 문학분야 / 과학분야 도서 40종
- 표지, 판권기, MARC 레코드(국가도서관)
- Gemini를 사용한 도서 이미지 처리



표지 이미지



판권기 이미지

(가) 도서 이미지

You are an expert librarian specializing in cataloging and creating MARC records. The attached images are key pages from a book, including the front cover and the colophon. Please Analyze the images and extract all relevant text using OCR. Due to data processing, do not add any comment irrelevant comments in the response.

(나) 서지정보 추출 프롬프트

```

### Image 1: Front Cover
**Publisher:** DE GRUYTER
**Series/Audience:** GRADUATE
**Author:** Inena Kostova
**Title:** GENERAL AND INORGANIC CHEMISTRY
**Subtitle:** THEORETICAL FOUNDATIONS, MAIN GROUP ELEMENTS, TRANSITION METALS, LANTHANIDES, AND ACTINIDES
**Publisher Logo:** DE G
***

### Image 2: Colophon
**Author**
Prof. Inena Kostova
Department of Chemistry
Faculty of Pharmacy
Medical University
1000 Sofia
Bulgaria
inena.kostova@yahoo.com

**ISBN** 978-3-11-171222-2
**e-ISBN (PDF)** 978-3-11-171224-6
**e-ISBN (EPUB)** 978-3-11-171254-3

**Library of Congress Control Number:** 2025931334

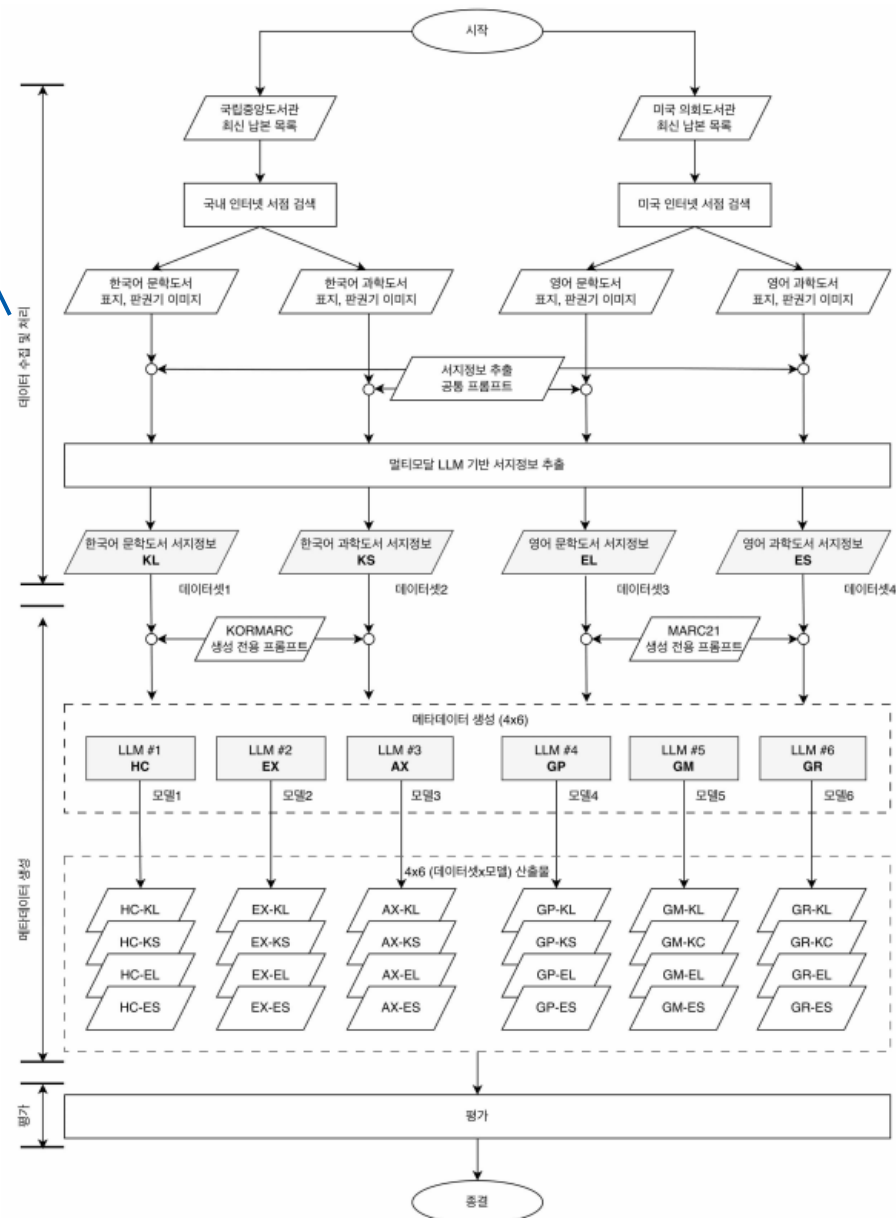
**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at [http://dnb.dnb.de]/[http://dnb.dnb.de].

© 2025 Walter de Gruyter GmbH, Berlin/Boston, Genthiner Straße 13, 10785 Berlin
**Cover image:** MF3dE+/Getty Images
**Typesetting:** Integra Software Services Pvt. Ltd.
**Printing and binding:** CPI books GmbH, Leck

[www.degruyter.com][https://www.degruyter.com]
Questions about General Product Safety Regulation:
productsafety@degruyterbrill.com
    
```

(다) 추출 결과

<그림 2> 서지정보 추출 예시



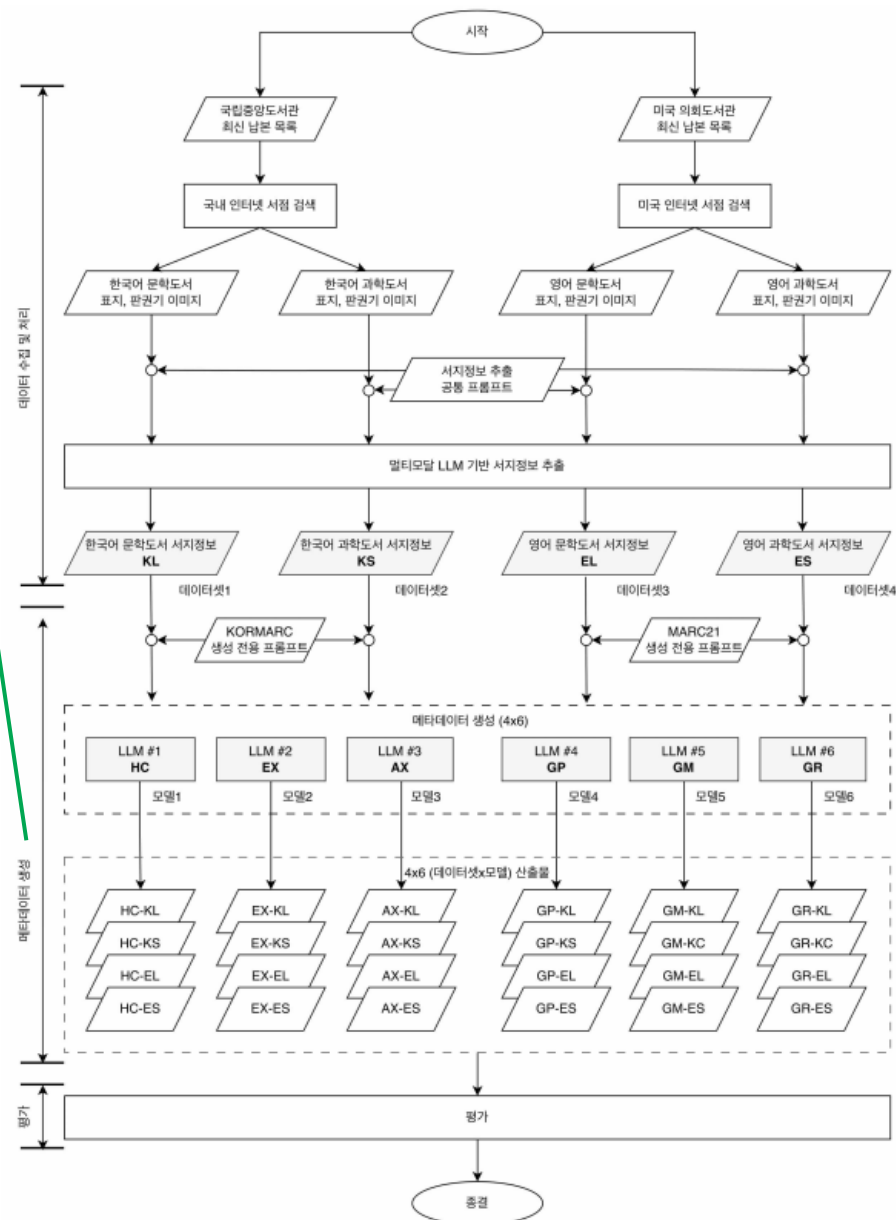
<그림 1> 연구 설계

2. 메타데이터 생성

- 국내/외, 문학/과학도서 4세트씩
6개의 언어모델에서 생성(총 24세트)

〈표 4〉 언어모델별 MARC 레코드 생성 결과 예시

AI type	field	ind	content	
HC	020		978-3-11-171222-2 2025	
	260	1	de gruyter berlin/boston genthiner straÙe 13 10785 berlin	
	500		general and inorganic chemistry: theoretical foundations, main group elements, transition metals, lanthanides, and actinides / irena kostova	
	650	0	#general and inorganic chemistry#graduate	
	990	0	#minneapolis minnesota	
EX	100	1	\$a Kostova, Irena	
	245	0	\$a General and inorganic chemistry: theoretical foundations, main group elements, transition metals, lanthanides, and actinides / \$c Irena Kostova.	
	260	0	\$a Berlin: Walter de Gruyter GmbH, \$c [2025].	
	264	1	\$a Berlin/Boston: \$b Walter de Gruyter GmbH, \$c [2025].	
	520	0	\$a CPI books GmbH, Leck.	
AX	020		\$a978-3-11-171222	
	100	1	\$aKostova, Irena, \$e author.	
	245	10	\$aGeneral and inorganic chemistry: theoretical foundations, main group elements, transition metals, lanthanides, and actinides / \$c Irena Kostova.	
	260		\$aBerlin: \$aBoston: \$b Walter de Gruyter GmbH, \$c [2025].	
	700	2	\$aWalter de Gruyter GmbH, \$c [2025].	
GR	020		\$a 978-3-11-171222	
	100	1	\$a Kostova, Irena, \$e author.	
	245	10	\$a General and inorganic chemistry: theoretical foundations, main group elements, transition metals, lanthanides, and actinides / \$c Irena Kostova.	
	264	1	\$a Berlin: \$a Boston: \$b Walter de Gruyter GmbH, \$c [2025].	
	830	0	\$a De Gruyter graduate.	



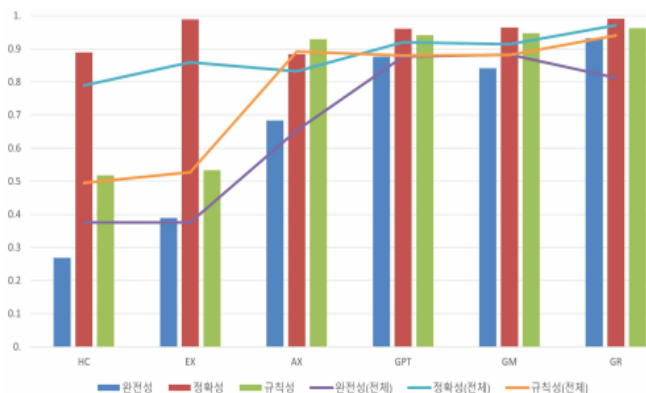
〈그림 1〉 연구 설계

3. 평가

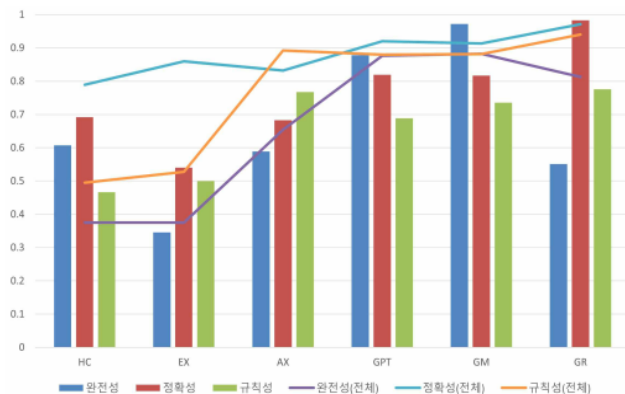
- 국가도서관이 작성한 MARC 레코드와 비교 → 완전성, 정확성, 규칙성 평가

〈표 7〉 언어모델이 생성한 MARC 레코드를 국내외로 구분한 성능

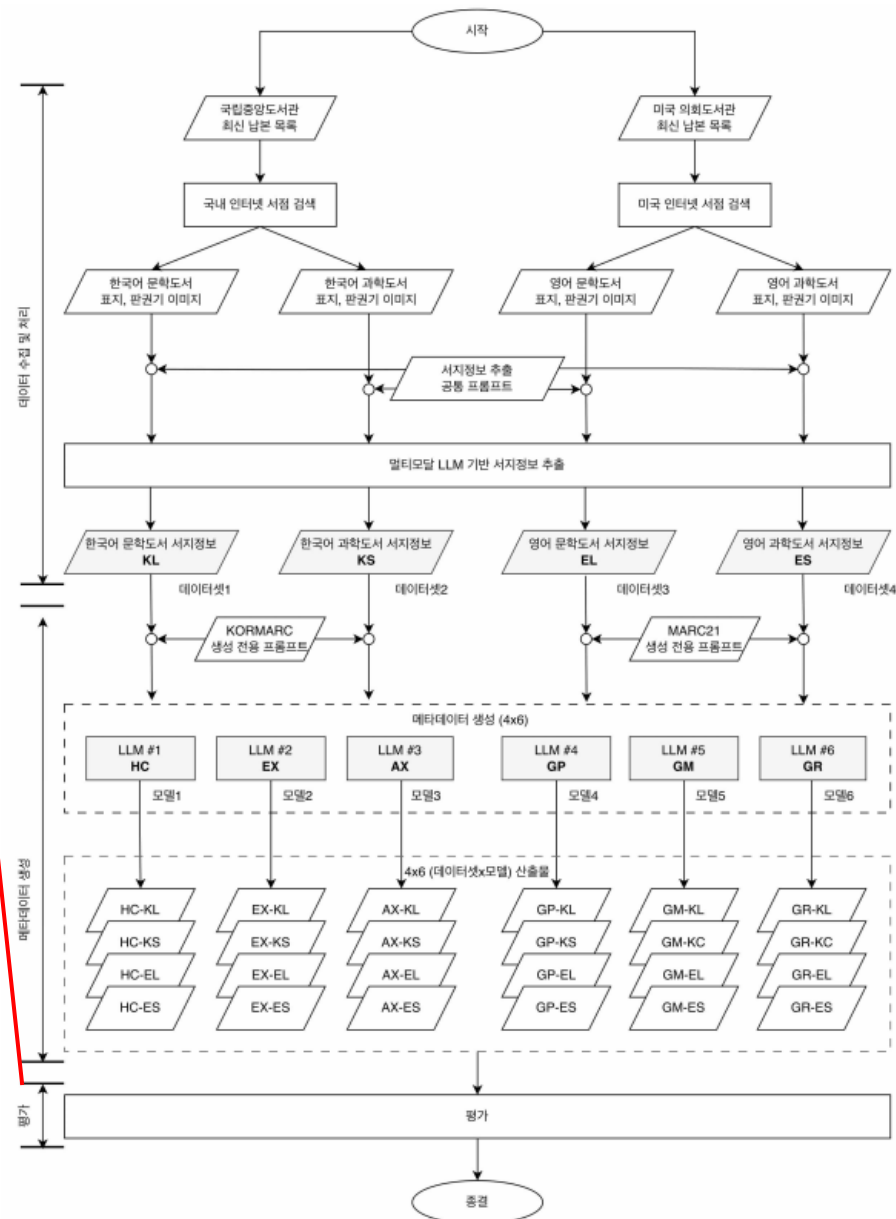
척도	AI 유형	HC	EX	AX	GP	GM	GR
		MARC21	완전성 0.2692 정확성 0.8889 규칙성 0.5179	0.3889 0.9890 0.5333	0.6838 0.8844 0.9293	0.8761 0.9610 0.9416	0.8419 0.9645 0.9474
KOR MARC	완전성 정확성 규칙성	0.6075 0.6923 0.4667	0.3458 0.5405 0.5000	0.5888 0.6825 0.7674	0.9720 0.8173 0.7353	0.8785 0.8191 0.6883	0.5514 0.9830 0.7759



MARC21



KORMARC



〈그림 1〉 연구 설계

〈표 2〉 생성형 AI에서 확인 할 수 있는 기술 필드

필드	구분	MARC21 (RDA 적용)	KORMARC (KCR4)
1XX	기본표목	○	-
245	표제 및 책임표시사항	●	●
246	여러 형태의 표제	○	○
250	판사항	○	○
260	발행, 배포, 간사사항	●	●
264	생산, 발행, 배포, 제작, 저작권 표시	●	●
490	총서사항	○	○
8XX	총서부출	○	○
5XX	주기사항	○	○
6XX	주제명 표목	○	○
7XX	부출표목	○	○
020	입수사항	○	○

〈표 8〉 주요 필드별 성능

필드	AI유형 척도	HC	EX	AX	GP	GM	GR
		020	완전성	0.5000	0	0.6571	0.9857
	정확성	0.8857	0	0.9130	0.9565	0.9259	0.9855
	규칙성	0.4839	0	1	0.8333	0.7900	0.9853
100	완전성	0.0556	0.9444	0.9444	0.8889	0.9444	0.9444
	정확성	0	0.9412	0.9412	1	1	1
	규칙성	0	0.5938	0.8750	1	1	1
245	완전성	0	1	1	1	1	1
	정확성	0	1	0.9750	1	1	1
	규칙성	0	0.5125	0.7179	0.6875	0.700	0.7250
260	완전성	0.9500	0.8500	0.7500	1	1	0.0500
	정확성	0	0	0.9333	1	1	1
	규칙성	0	0	1	0.9750	0.9750	1
264	완전성	0	0.6538	0.0385	1	0.9231	0.8462
	정확성	0	1	1	1	1	1
	규칙성	0	0.5	1	0.9808	0.9792	1
490	완전성	0.0625	0	0.0625	0.8125	0.9375	0.9375
	정확성	0	0	0.5000	1	1	1
	규칙성	0	0	0	0.6923	0.7000	0.5333
830	완전성	0.0833	0	0	0	1	0.3333
	정확성	0	0	0	0	0.8333	1
	규칙성	0	0	0	0	0.9000	1

〈표 9〉 규칙 오류의 대표 사례

	생성 MARC21		정답 MARC21
	필드	내용	내용
지시기호 오류	245	14\$aScratching the surface: \$bexploring Earth's layers / \$cKate Allen Fox : illustrated by Erin Brown.	10\$a Scratching the surface: \$b exploring earth's layers / \$c by Kate Allen Fox : illustrated by Erin Brown.
식별기호 오류	100	1b\$a Falligant, Erin, author.	1b\$a Falligant, Erin, \$e author.
지시기호 식별기호 오류	245	00\$a Energy reactions in the kitchen \$i Ann McCallum Staats.	10\$a Energy reactions in the kitchen / \$c Ann McCallum Staats, M.Ed.
	생성 KORMARC		정답 KORMARC
	필드	내용	내용
지시기호 오류	490 490	1b \$a 그냥 훑치면 되지 bb \$a 그냥 훑치면 되지	10▼a워픽 =▼aWefic :▼v89
식별기호 오류	020 020	bb \$a 9791192549484 \$c 03370 bb \$a 978-89-6090-938-6 \$d 03810	bb▼a9791192549484▼g03370:▼c\17500 bb▼a9788960909380▼g03810:▼c\17000
지시기호 식별기호 오류	245 245 260	10 \$a 바르셀로나의 유서 / \$c 백세희. 41 \$a앞으로 무엇을 더 이렇게 : \$v91 00 \$a김학원 \$a(주)휴머니스트출판그룹	00▼a바르셀로나의 유서 /▼d지은이: 백세희 00▼a고백의 시대 /▼d지은이: 이현석 bb▼a서울: ▼bH(휴머니스트출판그룹),▼c2025

연구결과

- **LLM 3종(GPT, Gemini, Grok) vs. 소버린 AI 3종(HyperCLOVA, EXAONE, A.X)**
 - 단, 국내도서에 대하여는 LLM의 성능이 크게 저하됨
- 표제와 책임표시사항(245)은 안정성이 높은 반면,
- 발행사항(260), 총서사항(490/830) 등 규칙 의존도가 높은 필드에서는 모델 간 성능이 차이를 보임

현시점에서 생성형 AI를 도서관 메타데이터 업무 도입 시, 전면적인 자동목록 도구보다는
서지레코드 초안 생성, 오류 탐지, 보완을 지원하는 보조도구로 활용하는 것이 타당함

생성형 AI를 활용한 지식그래프 구축: 구술생애사를 중심으로

Constructing a Knowledge Graph Using Generative AI: Focusing on Oral Life Histories

이혜경 (HyeKyung Lee)*

이용구 (Yong-Gu Lee)**

초 록

본 연구의 목적은 구술생애사 자료를 대상으로 생성형 AI를 활용하여 RDF 기반 지식그래프 구축의 가능성을 탐색하는 데 있다. 이를 위하여 실제 파독 간호사 7인의 구술생애사를 대상으로 생성형 AI에 적용하였으며, RDF 트리플을 자동 생성하도록 하였다. 이후 7인에 대한 RDF 트리플을 통합하여 지식그래프를 형성하고자 하였다. 그 결과, 첫째, 생성형 AI는 비정형 데이터인 구술생애사에서 일정 수준의 RDF 트리플을 생성하여 이러한 자료에 대한 지식 구조화 가능성을 확인할 수 있었다. 둘째, 동일하거나 유사한 의미의 술어를 일관성 있게 생성하지 못하는 한계가 드러났으며, RDF 구축을 위해서는 통합화된 술어 체계가 필요함을 시사하였다. 셋째, 통합 지식그래프를 도출함으로써 개별 구술생애사 간 RDF 연계가 가능하고, 이를 통해 시각화하여 다각적 관점에서 데이터를 파악할 수 있음을 확인하였다. 다만, 구술생애사 자료의 구어적 특성, 맥락 의존적 표현, 사투리 등은 AI의 개체명 인식에 어려움을 주었으며, 객체 수가 불필요하게 증가하는 문제점 등이 한계로 나타났다. 그럼에도 불구하고 본 연구는 구술생애사와 지식그래프를 연결한 탐색적 시도라는 점에서 의의가 있으며, 복잡한 절차를 거쳐야 했던 지식그래프 구축 과정을 생성형 AI를 통해 간소화할 수 있는 가능성을 제시하였다.

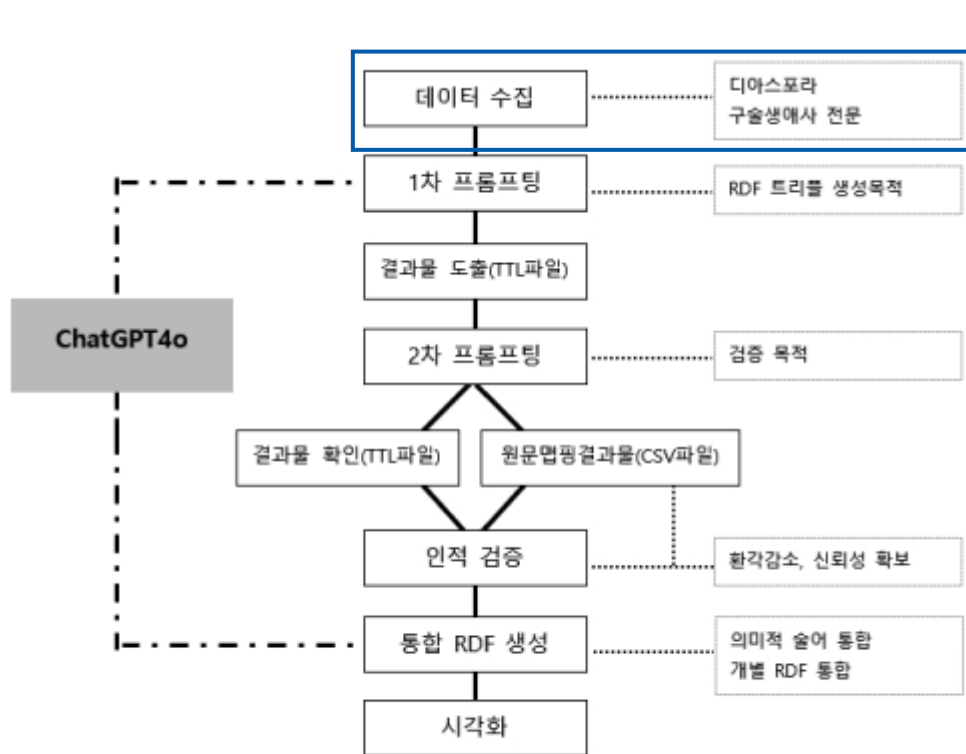
연구의 배경 및 목적

- 의미와 맥락을 담은 지식 자원으로서 텍스트 데이터가 지니는 가치
- 개체와 관계를 통해 데이터 간 연관성을 명시화하는 지식그래프(Knowledge Graph)
- AI의 등장과 발전으로 자연어 분석과 활용 가능성 증대

생성형 AI를 활용한 RDF 기반 지식그래프 구축 가능성 탐색

- 한인 디아스포라로서의 파독 간호사 구술생애사 기록을 대상으로

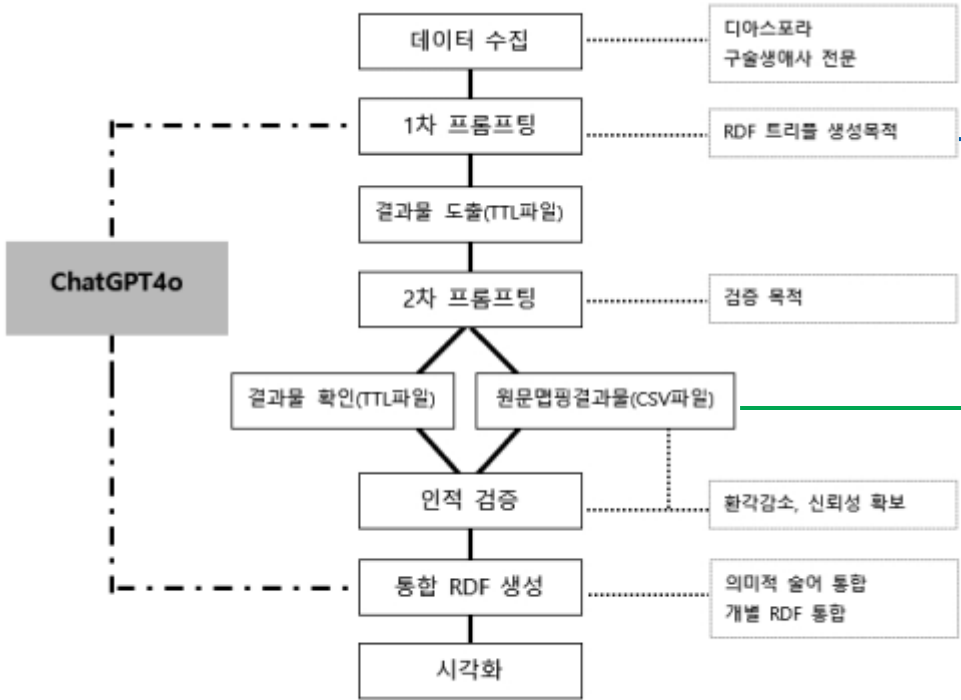
→ 구술생애사 연구에서 생성형 AI의 새로운 활용 가능성 제시 및 지식그래프 구축을 위한 방법론적 통찰 제공



〈그림 1〉 연구 절차 개요도

〈표 1〉 RDF 대상 텍스트의 문단 및 문장 통계

대상 인물	문단	문장
1번	71	339
2번	104	720
3번	90	501
4번	47	213
5번	116	819
6번	84	400
7번	111	572
합계(평균)	623(89)	3564(509)



〈그림 1〉 연구 절차 개요도

지시문:
 다음에 제시되는 텍스트는 구술생애사이며 특히 한인 디아스포라에 해당합니다.
 이 구술생애사 텍스트에서 인물, 장소, 사건, 조직 등의 주요 개체(Entity)와 개념을 추출하고, 이를 RDF 트리플(subject, predicate, object) 형식으로 변환하여 주세요.
 1) 결과는 Turtle(TTL) 형식의 파일로 주세요.
 2) 엑셀파일로 RDF 트리플의 근거 문장을 함께 제시해주세요, 시간을 갖고 천천히 작업하고 결과는 매우 다양하고 깊이있게 처리되어야 합니다.
안내문:
 다음은 독일로 파견된 간호사의 일대기를 담은 구술사입니다.

〈그림 3〉 1차 프롬프트 지시문

```

ex:Grandfather_DM a foaf:Person ;
  rdfs:label "담양의 조부(부동)"@ko ;
  schema:additionalType "부동"@ko .

ex:KODCO a schema:Organization ;
  rdfs:label "해외개발공사"@ko .

ex:ParkJaeSoon a foaf:Person ;
  ex:visitedKoreaAfterGwangju true ;
  schema:affiliation ex:Family_Business_Cloth ;
  schema:alternateName "박재순"@ko ;
  schema:alumniOf ex:JeonnamGirlsMiddleSchool,
    ex:JeonnamNursingSchool ;
  schema:award "일본 시장 표창(근속·업적)"@ko ;
  schema:birthDate "1943"^^xsd:gYear ;
  schema:birthPlace ex:DamYang ;
  schema:children ex:Son1,
    ex:Son2 ;
  schema:educationalCredentialAwarded "조선원 교육 수료"@ko ;
  schema:endDate "1963"^^xsd:gYear ;
  schema:event ex:AbuDhabi_Performance,
    ex:Account_Scandal ;
  schema:familyName "박"@ko ;
  schema:graduationYear "1961"^^xsd:gYear ;
  schema:grewUpIn ex:GwangJu ;
  schema:hasOccupation "간호사"@ko,
    "양호교사"@ko ;
  
```

〈그림 4〉 ChatGPT가 생성한 RDF 트리플 사례

Subject	Predicate	Object	SourceText
http://example.org/KimJeongJa	http://xmlns.com/foaf/0.1/name	김정자	난 1940년생이고 동생은 1945년생이에요.
http://example.org/KimJeongJa	http://example.org/birthPlace	일본 야마구치현	일본에서 태어났어요.
http://example.org/KimJeongJa	http://example.org/birthYear	1940	난 1940년생이고...
http://example.org/KimHoJa	http://xmlns.com/foaf/0.1/name	김호자	동생은 1945년생이에요.
http://example.org/KimHoJa	http://example.org/birthPlace	일본 야마구치현	일본에서 태어났어요.
http://example.org/KimHoJa	http://example.org/birthYear	1945	동생은 1945년생이에요.
http://example.org/KimJeongJa	http://example.org/sibling	http://example.org/KimHoJa	김정자와 김호자 자매
http://example.org/KimJeongJa	http://example.org/parent	http://example.org/Father	아버님은 경남 고성에서 태어나서...
http://example.org/KimJeongJa	http://example.org/parent	http://example.org/Mother	어머니가 아버지 찾아서 연락선을 타고...
http://example.org/baptism	http://example.org/year	1962	1962년도에 영세를 받았어요.
http://example.org/KimJeongJa	http://example.org/participatedIn	http://example.org/baptism	영세를 1962년도에 받았어요.
http://example.org/KimJeongJa	http://example.org/workPlace	콜론 시립양로원	콜론시립양로원으로 가게 되었어요.

〈그림 5〉 RDF 트리플과 추출에 사용된 근거자료 원문(CSV 파일 일부)

〈표 3〉 슬어의 생애사적 범주 군집화

인물	인물의 기본속성	가족관계	교육과 학력	직업과 소속	지리 및 장소	이동과 이주	사건과 활동	시간과 기간
1번	birthDate birthPlace knowsLanguage nationality religion	children familyName numberOfChildren parent siblingCount spouse	educationalCredent tialAwarded graduationYear	affiliation hasOccupation jobTitle memberOf worksFor	homeLocation studyLocation	conscriptedTo	award event participatedIn	endDate startDate hasBeginning
2번	agent birthDate birthPlace religion	parentOf spouse		affiliation member memberOf	address addressCountry fromLocation toLocation	wasDerivedFrom arrivalTime departureTime		
3번	birthPlace birthYear deathYear nationality religion	child parent sibling spouse	education	memberOf occupation workPlace workYearEnd workYearStart workedAs	relatedTo eventYear participatedIn	relatedTo	eventYear participatedIn	baptizedYear year
4번	birthDate birthPlace hasReligion lifeStage nationality officialBirthPlace	hasChild spouse	attendedSchool	memberOf occupation workedAt	livedIn currentCity	movedTo	event	attendedInstitution
5번	birthDate birthPlace knowsLanguage nationality	children parent sibling spouse		employer hasOccupation jobTitle memberOf worksFor	hometown location			endDate startDate
6번	birthPlace birthYear causeOfDeath durationUntilDeath religion	child reason relatedPerson spouse		jobRole organization	currentResidence country place	emigratedTo	hasEvent	baptismYear date pprofessionStartYear
7번	agent birthDate birthPlace deathCause deathDate marriageDate nationality	hasSibling children parent 1spouse		employmentHistory hasOccupation jobTitle memberOf worksFor	addressCountry location			startDate

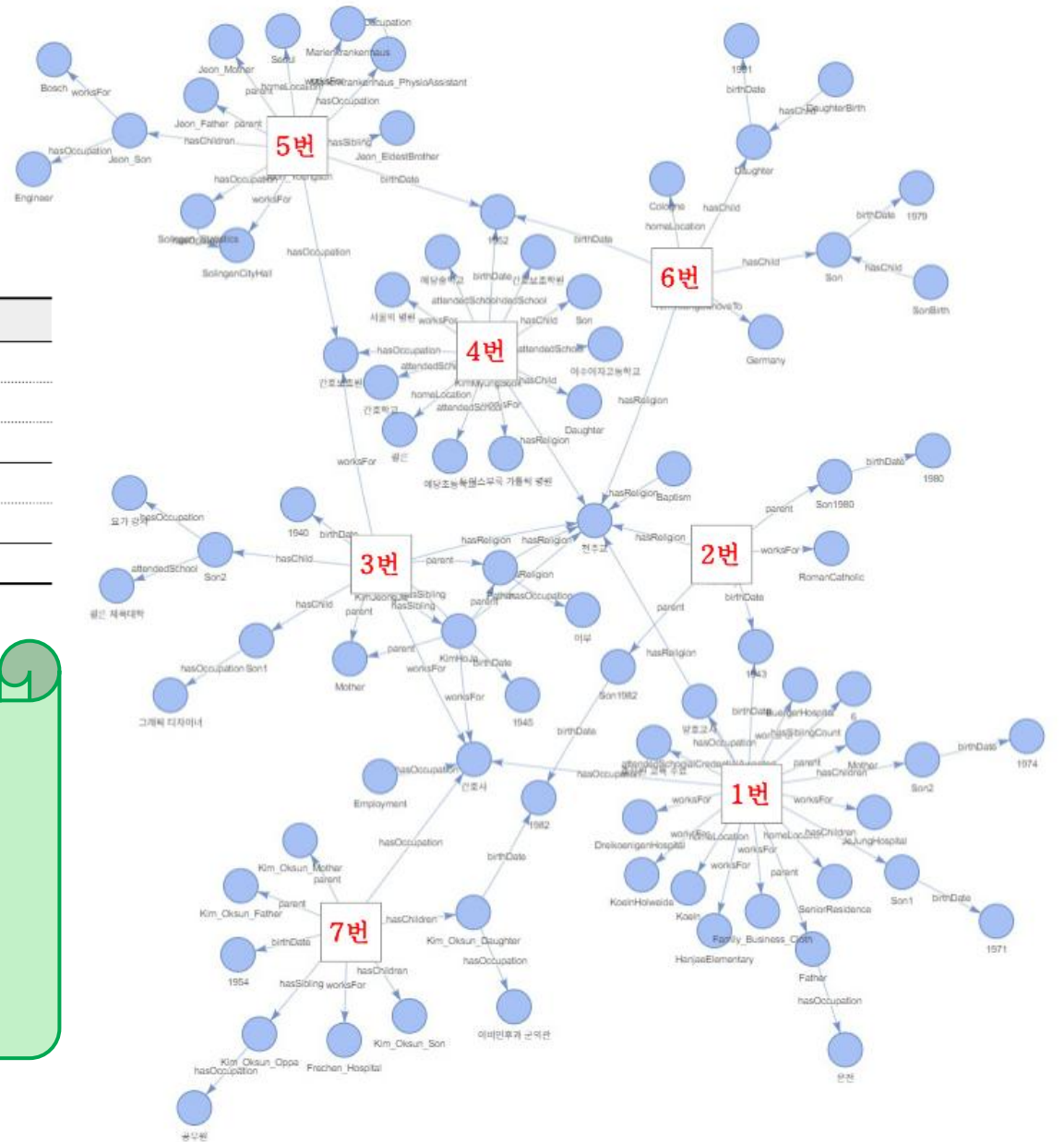
〈표 5〉 유사 슬어의 통합화 내역

슬어 정의	슬어	선정된 슬어
생년	birthDate / birthYear	birthDate
가족	parent / parentOf	parent
	sibling / hasSibling	hasSibling
자녀	child / children / hasChild / numberOfChildren	hasChild
교육	attendedSchool / attendedInstitution / education	attendedSchool
종교	religion / hasReligion	hasReligion
직업	jobTitle / hasOccupation / occupation / employer / jobRole	hasOccupation
소속	worksFor / workedAs / affiliation / workedAt	worksFor
이주 이동	toLocation / movedTo / emigratedTo	moveTo
거주	homeLocation / currentResidence / currentCity / hometown	homeLocation

연구결과

〈표 6〉 인물의 연관관계

범주 구분	객체	인물
생년	1943	1번, 2번
	1952	4번, 5번, 6번
	1982	2번의 자녀, 7번의 자녀
직업	간호사	1번, 2번, 3번, 3번의 동생, 7번
	간호보조원	3번, 4번, 5번
종교	천주교	1번, 2번, 3번, 3번의 아버지, 3번의 동생, 4번, 6번



- 생성형 AI는 구술생애사의 서술적 특징을 반영한 RDF 트리플을 생성 가능
- 그러나 술어의 일관성에서 한계 → 통합된 술어 집합 구성 필요
- 이를 통해 다차원적 관점에서 생애사를 시각화하여 탐색할 수 있음
- **연구의 한계**) 구축된 지식그래프에 대한 체계적인 평가 미실시

생성형 AI 기반 메타데이터 자동 생성의 편향과 공정성에 대한 정보학적 분석*

An Analysis of Bias and Fairness in Generative AI-Based Metadata Automation: An Information Science Perspective

이 정 미 (Jeong-Mee Lee)**

초 록

본 연구는 AI에 기반한 자동 메타데이터 생성에서 발생하는 편향과 공정성 문제를 정보조직과 자원발견 관점에서 재구성하고자 하는 연구이다. 이를 위해 최근의 국내외 문헌을 체계적으로 분석하여, AI 자동화가 정보조직의 어느 단계에 개입하는지, 그리고 그 과정에서 의미 단순화, 대표성 불균형, 비동제어위 확산, 분류 위계 왜곡과 같은 구조적 편향이 어떻게 형성되는지를 검토하였다. 또한 이러한 편향이 검색 결과의 가시성, 주제 접근성, 자원 노출 기회에 어떠한 영향을 미치는지를 자원발견의 공정성 측면에서 논의하였다. 연구 결과, AI 자동 메타데이터 생성은 효율성 증대라는 장점도 있으나 학습 데이터의 통계적 분포와 알고리즘적 추론 방식에 기반하여 특정 주제나 집단을 과대·과소 포장할 가능성을 내포하고 있음을 확인하였다. 이는 지식이 조직되고 노출되는 구조 자체를 재편하는 정보학적 쟁점으로 이해되어야 한다. 따라서 AI 자동 생성 메타데이터는 성능 지표를 넘어, 지식조직 체계와의 정합성 및 자원발견의 형평성을 함께 고려하는 통합적 분석 프레임이 요구된다. 본 연구는 AI 자동 메타데이터 편향을 정보 흐름의 구조적 불균형이라는 관점에서 설명하여 정보학적 분석에 기반한 자원발견의 공정성을 위한 이론적 기반을 제시하는 데 의의가 있다.

연구의 배경 및 목적

- AI의 등장 이후 도서관 정보조직 분야에서 AI의 활용되며 대체까지 거론되는 상황
- 그러나 AI 도입은 효율성 이면에 편향성이라는 문제를 야기할 수 있음
- 그간 ‘정확성’과 ‘완전성’에만 집중되어 축적된 선행연구
- 정보조직의 핵심: 지식의 체계화와 의미적 연결

정보조직과 자원의 발견이라는 정보학의 핵심 개념 속에서
AI 자동 메타데이터 생성의 편향 문제를 재구성

→ AI의 편향 문제를 정보윤리가 아니라 지식이 조직되고 노출되는 방식의 변화로 분석하는 관점 제시

AI가 메타데이터 생성의 핵심 단계에서 만들어내는 편향의 구체적인 양상과 정보학적 특성은 무엇인가?

AI의 통계적 추론 기반 키워드 생성 방식은 도서관의 전통적인 통제어휘 체계 및 지식조직체계(KOS)의 의미적 위계와 어떻게 충돌하며 이를 재구성하는가?

자동 생성된 메타데이터의 편향이 자원 발견의 공정성(정보 접근의 구조적 형평성)에 어떠한 영향을 미치며, 이를 완화하기 위한 정보전문가의 역할은 무엇인가?

• 질적 문헌분석

- 2020년 이후 발표된 국내외 학술 논문, 학위 논문, 주요 사례 연구 수집
 - 수집 키워드: Cataloging and metadata creation, AI metadata creation, automatic metadata creation, metadata generation, metadata extraction, knowledge organization in library, resource discovery+library, AI algorithm + metadata creation, Algorithmic Bias, representation bias
- 수집된 180여 편의 논문을 초록과 본문을 검토하여 연구주제가 AI 기반 메타데이터 생성, 자동 주제 분석, 자동 분류, 자원발견 또는 지식조직과 직접 관련된 문헌만 선정
 - 단순 시스템 개발 보고, 기술 소개 문헌 등 제외
- 메타데이터 생성 사례에 관한 문헌 우선 선정
- 최종 40편 내외의 문헌 검토

〈표 2〉 연구의 내용과 분석 흐름

정보조직 요소	자동화 개입 내용	주요 편향 유형	자원발견 영향	대응 연구질문
주제 분석	텍스트 기반 주제 자동 추론	맥락 단순화, 핵심 주제 누락	검색 적합도 저하, 주제 왜곡	연구문제 1
개념 추출	빈도·확률 기반 핵심어 추출	소수 개념 과소대표, 일반 개념 과대표	주제 대표성 불균형	연구문제 1
범주화	자동 분류 및 군집화	과도한 일반화, 경계 모호화	브라우징 경로 왜곡	연구문제 1 연구문제 2
용어 선택	비통제어휘 생성 및 사용	표준 용어 불일치, 동의어 분산	검색 누락·중복	연구문제 2
관계 설정	자동 관계 추론 및 연결	의미 위계 왜곡, 관계 과잉/결손	탐색 맥락 붕괴	연구문제 2
메타데이터 적용	검색·정렬·추천에 활용	특정 자원 지속적 비가시화	검색·발견 공정성 저해	연구문제 3

연구결과

연구문제1. AI 메타데이터 생성에서 편향의 구체적 양상과 정보학적 특성

- 정보의 단순화와 맥락의 누락
 - 정보조직에서의 주제 분석: 문헌의 종합적 검토, 저자의 문제의식 고려, 학문의 특징, 이용자의 잠재적 요구 등을 다각도로 고려한 해석적 판단 ▶ 맥락적
 - AI의 주제 분석: 텍스트 내 단어의 통계적 빈도나 벡터 공간 내의 유사도 ▶ 통계적
 - 발생빈도를 중심으로 빈도가 낮은 개념이나 명칭들을 통계적 노이즈로 처리하여 메타데이터 구조에서 제외
 - 동시출현 기반 네트워크를 형성하여 범주화 단계에서 세부 개념들이 상위 개념으로 포괄되어 버리는 경향이 관찰 → 과도한 일반화
- 학습데이터의 편중과 통계적 추론 방식의 한계
 - 주로 영미권 데이터 위주로 학습되기 때문에 다른 언어나 문화권에서의 적용성 / 정확도가 떨어짐
 - 사실 검증이 아닌 학습된 패턴을 조합하여 작업을 수행하는 AI의 특성 상, 희소 주제나 예외 규칙은 노이즈로 처리

연구결과

연구문제2. AI의 통계적 추론 방식과 기존 지식조직체계의 충돌 양상

- 통제어휘와 자동 생성 용어 간의 불일치
 - 통제어휘집에 없는 자연어 태그 생성, 인명전거를 주제명으로 잘못 사용하는 등
 - AI가 생성한 용어는 기존 표준화된 지식조직체계와 결합되지 못하는 ‘의미적 파편화’ 가능성 → 상호운용성 저해
- 분류 체계 상의 범주나 위계의 왜곡
 - 텍스트 내 특정 키워드의 편향은 전체적인 맥락 해석을 저해하여 잘못된 상위 범주에 할당하는 오류 발생
- 표준화 원리와 자동화 유연성 간의 긴장
 - 풍부한 자연어 검색 제공 vs. 기술의 일관성, 접근의 안정성, 상호운용성

연구결과

연구문제3. AI 메타데이터 생성의 편향이 자원 발견의 공정성에 미치는 영향

- 검색 결과 가시성의 불균형을 초래 ▶ 자원 발견의 공정성 훼손
 - 결과물의 품질 편차와 가시성의 불균형
 - AI 생성 메타데이터의 오류는 자원의 노출 가능성을 직접적으로 제한할 수 있음
 - 해당 정보 자원의 가시성에 직접적으로 영향을 미칠 수 있음
 - 도메인 편향과 주제 영역별 가시성의 불균형
 - 구조적 학문 분야 vs. 감정적, 해석적 분야에서의 AI의 성능 차이
 - AI 운용 기관 역량의 차이로 인한 불균형
 - 기술 인프라, 데이터 정합성, 재정 자원 측면에서의 격차가 그대로 반영
 - 데이터 기반이 탄탄한 기관일수록 더 높은 품질의 결과를 생산할 가능성이 매우 높음

연구결과

• 자원 발견 공정성의 정보학적 재정의

표상(representation)에 따른 공정성

- 특정 집단, 주제, 문화권에 대한 학습데이터 부족 및 편향은 검색 결과 내에서 해당 자원의 의미적 위치 약화 ▶ 가시성 저하
- 적절한 의미 표상의 보장 필요

의미 위계에 따른 공정성

- 인류의 지식구조체계가 가진 본질적 위계질서와 상충되는 AI 생성 지식 ▶ 지식 위계의 재편성
- 동등한 위계적 배치 가능성 보장 필요

검색 환경에서의 접근 기회에 따른 공정성

- 자동 생성된 메타데이터의 오류 → 검색 질의와의 매칭 저하, 추천 알고리즘에서의 배제 ▶ 이용자의 접근 기회 감소
- 동등한 발견 가능성 보장 필요

제도적 통제와 함께 하는 공정성

- ‘보조도구’로서의 AI의 활용 = 메타데이터 검증 체계 존재 여부가 공정성에 지대한 영향을 미침
- 도메인별 정합성, 모니터링, 기술 격차 완화를 위한 정책적 지원 필요

※ 연구방법 비교

논문 1.

메타데이터 자동 생성 성능 비교

- 데이터 선정 및 전처리
- 메타데이터 생성
- 생성 결과 평가
 - 생성 필드의 수치상 비교
 - 반복 생성 시의 안정성
 - 오류의 수정 가능성

논문 2.

구술생애사 지식그래프 구축

- 대상 선정 및 데이터 수집
- RDF 구축 및 정규화
- 구축 결과에 대한 검증
- 시각화
- 결과에 대한 체계적인 평가

논문 3.

AI 메타데이터 자동 생성 편향과 공정성 논의

- 분석 문헌 선정 및 수집
 - 정보조직에서 AI의 개입
 - AI 생성 데이터의 구조적 편향
 - 편향의 영향

※ 토론 질문

- AI가 여러분들의 생활/연구활동에 어떤 영향을 미치고 있나요?
- 도서관의 업무과정에서의 AI 활용은 어떻게 적용되어야 할까요?
 - Technical Services
 - Public Services
- 도서관에서 AI 활용을 위해 고려해야 할 / 준비해야 할 제반사항에는 어떤 것들이 있을까요?