

# 안지은 (Jieun Ahn)

AI · 데이터 엔지니어링

문헌정보학(LIS)과 소프트웨어(SW)를 함께 공부하며, LLM · 지식그래프 · 데이터 품질을 다룹니다.

📍 Seoul, Korea    ✉ aje20010827@gmail.com

 GitHub

 LinkedIn

 Blog

 Email

## About

인공지능 기술을 깊이 이해하고 정리하는 것을 좋아합니다. 수학 기초부터 LLM(RAG·파인튜닝), 지식그래프(온톨로지·SPARQL), 데이터 품질 평가까지 폭넓게 학습하며 기술 블로그 '전자두뇌만들기'를 운영합니다. (이 소개 문구는 자유롭게 수정하세요.)

## Skills

### LLM / NLP

LangChain    RAG    Fine-tuning

Prompt Engineering

### Knowledge Graph

Ontology (Protégé)    Neo4j    SPARQL

RDF/OWL

### Data

Python (pandas)    데이터 품질 평가 (AIRD)

데이터 정제

### Web / Tooling

Git/GitHub    VitePress    Vue

# Projects

---

기타

[address\\_management\\_tool](#)

---

[quality\\_check\\_tool](#)

---

[vannevar](#)

---

## FAIR 데이터 원칙을 적용한 국내 디지털 아카이브 평가와 개선 방향 · 2024 · 공동저자

탐색성·접근성·상호운용성·재사용성(FAIR) 관점에서 디지털 아카이브를 평가하는 방법론을 제시하고, 국내 디지털 아카이브 261개를 조사해 최종 선정한 69개를 대상으로 FAIR 데이터 원칙 준수 여부를 정량 분석한 연구다. ARDC FAIR Self Assessment Tool의 세부 지표를 디지털 아카이브 특성에 맞게 재해석·구체화해 적용했으며, 국내 디지털 아카이브가 전반적으로 FAIR 원칙을 반영하지 못하고 특히 상호운용성과 재사용성이 매우 낮음을 확인했다. · 성과: 한국기록관리학회지 24(4), 113-134 게재 (2024.11, DOI 10.14404/JKSARM.2024.24.4.113). 국내 디지털 아카이브 261개 조사 및 69개 정밀 평가 수행. 조사 대상 261개 중 74%가 탐색·접근에 제한, 평가 대상 69개의 FAIR 평균 점수 50.43점. 세부 지표 해석 전문을 깃헙에 공개해 재현가능성 확보.

[ARDC FAIR Self Assessment Tool](#)   [FAIR 데이터 원칙](#)   [구글 검색 기반 수집](#)   [DADA 웹 크롤링](#)

[다중 평가자 교차검증](#)   [FAIR 지수 정량화\(가중치 기반 100점 산정\)](#)

---

## 공공데이터 주소 품질 향상을 위한 절차적 정제 프레임워크(AQRF): 설계와 실증 · 2025 · 제1저자 (안지은, 중앙대학교 문헌정보학과 석사과정 / 교신저자 김학래 교수)

공공데이터의 주소 품질을 체계적으로 개선하기 위해 의미 단위 분해·구성요소 보완·행정정합성 검증·오류 수정의 4단계로 구성된 절차 기반 프레임워크(AQRF)를 제안한 연구다. 보건의료 5개 공공데이터세트 약 14,000건을 대상으로 실증한 결과 평균 정제 성공률 99.18%를 달성했고, KS X ISO 19157 기준의 구성요소 인식과 행정정합성이 일관되게 향상되었다. 주소 품질을 필드 수준 정확성 문제가 아니라 데이터세트 간 구조적 관계로 재개념화하고 품질관리 단위를 프로세스 자체로 확장한 점이 핵심 기여다. · 성과: 「지적과 국토정보」(Journal of Cadastre & Land Information) 제55권 제2호(한국국토정보공사, 2025.12, p.157-172) 게재, DOI 10.22640/lxsiri.2025.55.2.157. 보건의료 5개 데이터세트 약 14,217건 대상 평균 정제 성공률 99.18%(전 데이터세트 98% 이상), 타 도메인 5개 기관 교차검증 평균 95.33% 달성. 2024년도 중앙대학교 CAU GRS 지원 과제.

정규표현식 기반 구조 분석    의미기반 검증(하이브리드 접근)    Levenshtein distance 유사도 탐색

Address Class 표준 입력 포맷    행정안전부 도로명주소 API    통계청 법정동코드 DB    도로명주소DB

KS X ISO 19157 품질지표 체계

## 공공데이터 재사용성 향상을 위한 9차원 품질 진단 프레임워크의 설계와 실증 · 2026 · 제1저자 (안지은, 중앙대학교 문헌정보학과 석사과정; 교신저자 김학래 교수)

공공데이터 CSV 파일의 재사용성을 자동으로 진단하는 9차원 품질평가 프레임워크를 제안한 연구다. 행정안전부 4대 지표와 ISO/IEC 25024 메트릭 체계를 기반으로 구조완결성·스키마안정성·의미중복도·재사용적합성·표준연계성 5개 차원을 추가하고, Gate 메커니즘과 참조데이터 기반 규칙엔진으로 자동화 진단 파이프라인을 구현했다. 산업통상자원부 산하 19개 기관의 CSV 1,258건을 실증분석하여 적시성(D4) 취약과 기관 간 품질 편차를 확인하고 기관 맞춤형 큐레이션의 필요성을 제시했다. · 성과: 정보관리학회지 제43권 제2호(329-352면) 게재. 산업통상자원부 산하 19개 기관 CSV 1,258건 실증분석(전체 평균 77.7점 B등급)으로 김학래(2020) 91건 대비 약 14배 규모 분석 수행. 재사용적합성(D8)·표준연계성(D9)을 독립 차원으로 분리한 9차원 품질지표 체계 정의. 2024년도 중앙대학교 CAU GRS 지원.

Python    ISO    IEC 25024 측정 메트릭    Gate 메커니즘    참조데이터 기반 규칙엔진    CSV 파싱

인코딩 탐지(UTF-8    EUC-KR    CP949)    병렬처리(동시처리 30)    SQL    Semantic

Constant Metric 분류    임계값 기반 채점

## 데이터 품질 검사 스킬 (data-quality-check-skill)

테이블 형식 데이터 파일(CSV·TSV·XLSX·XLS)의 품질을 AIRD Part 2 품질 엔진으로 진단하는 Claude Skill 이다. 자체 평가 로직 없이 외부 quality\_check 엔진을 호출·오케스트레이션하며, 검사 도중 선택창 (AskUserQuestion)으로 참조데이터·의존성 규칙 등 사용자 정의 입력을 수집한다. 진단 결과는 마크다운 표 리포트와 데이터 정제 가이드, 그리고 W3C DQV 기반 TTL로 산출한다.

Python 3.10+ pandas numpy openpyxl Claude Skill AIRD Part 2 품질 엔진 W3C DQV  
TTL RDF

### linklear · 2025년 2학기 (지식표현연구 기말 과제)

입력한 링크의 웹페이지를 크롤링·정제한 뒤 LLM으로 주제를 분석하고, 이를 그래프 데이터베이스에 저장해 질의응답까지 지원하는 AI 참고자료 정리 서비스다. 크롤링·그래프저장·저장·질의를 담당하는 4개 에이전트로 파이프라인을 분리했고, 사용자-링크-주제를 노드와 SAVED·ABOUT 관계로 표현하는 지식표현 구조를 실제 서비스에 적용했다. pip 설치형 패키지로 배포해 Colab 등 외부 환경에서도 재사용할 수 있게 했다. · 성과: 정량적 성과 지표는 자료에 없음. 확인되는 결과물: GitHub 설치형 패키지로 배포해 Colab 등 외부 환경에서 재사용 가능, Railway(백엔드)+Vercel(프론트) 배포 구성 완료, 기말 발표자료 산출

Nuxt 3 FastAPI uvicorn Neo4j Aura Groq (Llama 3.1 70B) langchain neo4j 드라이버  
aiohttp httpx openpyxl python-dotenv Railway Vercel

### Linklear — URL 정리 에이전트 (지식표현연구 기말과제) · 개인 프로젝트 기획·설계·구현 (정보학 석사 지식표현연구 기말과제)

저장한 URL을 체계적으로 정리·재탐색하지 못하는 문제에서 출발해, 링크를 붙여넣기만 하면 AI가 메타 정보를 수집하고 주제를 자동 분류·저장하는 URL 정리 에이전트 'Linklear'를 제안·구현했다. 채팅 질의로 질의어-링크 유사도를 계산해 원하는 자료와 연관 자료를 탐색하며, 그래프 구조 저장으로 주제 간 연관성까지 시각화한다. 역할을 분리한 4개 LLM 에이전트의 협업 구조와 Neo4j 지식그래프를 결합한 엔드투엔드 시스템이다. · 성과: 역할을 분리한 4개 에이전트(CrawlerAgent·QueryAgent·StorageAgent·GraphAgent) 협업 구조 설계·구현, Neo4j 스키마 설계 및 규칙 기반 Cypher 생성으로 그래프 DB를 검색·추천에 적용, 질의어→주제어 추출→그래프 탐색→연관성 기반 우선순위 산출 검색 흐름 구현, 유튜브 링크 크롤링 예외와 유사·포괄 주제 오분류 문제를 프롬프트 개선(제목 핵심어 우선·기존 주제 목록 참조·구체 키워드 지향)으로 해결

Nuxt Vue.js FastAPI Groq LLM Chain llama3.3 70B Neo4j Aura DB Cypher

링크를 입력하면 웹 크롤링으로 페이지 주제를 분석하고 Neo4j 그래프 DB에 저장한 뒤, 저장된 자료를 근거로 질의에 응답하는 AI 레퍼런스 정리 서비스. 좌측 채팅창에서 링크 추가·질문을 하고 우측 디렉토리에서 주제별로 정리된 자료를 확인하는 흐름으로 구성된다.

Nuxt 3    FastAPI    uvicorn    Neo4j Aura    Cypher    Groq(Llama 3.1 70B)    LangChain  
aiohttp    openpyxl    python-dotenv    Python 3.12    Railway    Vercel

**character\_kb\_chatbot — 온톨로지 기반 캐릭터 챗봇** · 개인별 역할 분담은 자료에 명시되지 않음. 자료가 근거하는 작업 범위: 온톨로지 설계·표준 어휘 매핑, 백엔드(FastAPI)·그래프 DB(KuzuDB) 스키마, 프론트(Vue), 패키지화 리팩토링, 스모크/QA 검증 전반

그래프 DB(KuzuDB)에 캐릭터·사용자·세계·이벤트·감정·기억 등을 온톨로지 스키마로 저장하고, LLM 대화 중 이 지식 그래프를 조회·갱신하는 캐릭터 챗봇 패키지다. FastAPI 백엔드와 Vue 3/Vite 프론트로 구성되며, 대화 오케스트레이션·정보 추출·SHACL 기반 그래프 편집 검증을 갖췄다. 감정(Ekman 6대·PAD)·성격(Big Five) 모델과 한국어 형태소 분석을 포함하고, 구조는 ROAD 온톨로지에서 영감을 받았다. · 성과: 패키지화 1단계 완료·검증(11개 파일 import 전환 + \_\_init\_\_.py 1개, 잔존 bare 내부 import 0건, compileall exit 0, 전 스모크 통과); 동작 보존 확인(validate.py 결과가 베이스라인과 동일 — Characters 2/Users 1/World·Loc 1·1/Event 1/EmotionalState 1/KNOWS 2); import 성능 측정(engine 0.9s, backend.runtime 0.8s, backend.chatlog 1.5s, ConversationRuntime 인스턴스화 0.0s); SHACL 무결성(shapes.ttl 파싱 OK, 103 트리플); 런타임 이슈 원인 규명(import backend.app 45s+ 행 → google.generativeai 워밍업의 GIL 점유로 국소화, ONTO\_SKIP\_WARMUP 가드로 회피); 대형 모듈 진단(runtime.py 995줄/chat.py 933 줄/extraction.py 683줄 식별 및 분리 계획 수립)

Python 3.13    FastAPI    uvicorn    Pydantic v2    KuzuDB    OpenAI    Anthropic  
Google Generative AI    rdflib    pyshacl(SHACL)    TTL    Turtle    kiwipiepy    Vue 3    Vite

## Experience

Researcher · HIKE Lab. 2023-2026

## Awards

제13회 산업통상부 공공데이터 활용 아이디어 공모전 (대상) · 산업통상부 2025

HIKE Lab. 으로 참가 / 아이디어 기획 부문 대상

데이터 활용 아이디어&시각화 경진대회 · 한국데이터산업진흥원 2021

시각화 스토리텔링 부문 우수상

## Certifications

SQLD(SQL 개발자) · 한국데이터산업진흥원 2026.03.27

자격번호: SQLD-060014854

## Education

중앙대학교 2020-2024

문헌정보학과, 소프트웨어학부(복수전공)/학사

중앙대학교 2024-2026

문헌정보학과/석사