

데이터저널리즘

이론과 실무

이종혁 (경희대 미디어학과 교수)

2026

Part 1

데이터 저널리즘의 개념과 역사

데이터 저널리즘이란?

개념 정의

"데이터를 수집하고(collecting), 분석하고(analyzing), 시각화해서(visualizing) 새로운 사실과 숨겨진 진실을 밝혀내는 것"
— 김태형 (2015)

"데이터에서 유용한 정보를 추출하고, 이를 기반으로 기사를 작성하며, 인터랙티브 시각화를 통해 독자의 이해를 돕는 과정" — Veglis & Bratsas (2017)

특징

- CAR(Computer-Assisted Reporting)와 정밀저널리즘(Precision Journalism)에서 진화
- 빅데이터 축적과 공개 경향, 컴퓨터 기술의 향상이 배경
- 데이터 수집으로 근거 대기 → 데이터에서 새로운 의미 찾기
- 사회과학 방법론 → 컴퓨터공학 기술 적용으로 확장

Part 1. 개념과 역사

데이터저널리즘으로의 진화

- CBS, UNIVAC 컴퓨터로 미 대선 예측 (1952)
- 필립 마이어(Philip Meyer)의 디트로이트 인종 폭동 보도 → 정밀저널리즘 (1967)
- Philadelphia Inquirer 사법시스템 탐사보도 (1972)
- Atlanta Journal-Constitution 은행 대출 인종차별 관행 보도 (1988)
- NICAR(National Institute for Computer-Assisted Reporting) 설립 (1989)
- 제민일보 4.3 사건 관련 보도 (1990년대 초)
- 동아일보 <정권 따라 검찰 요직 춤 쳤다> (2000)
- KBS 탐사보도팀 창설, <법은 평등한가> (2005~2006)
- 2010년대 뉴스룸에 데이터팀 구축 → 데이터저널리즘 본격화
- 2020년대 AI-LLM 도입 → 에이전트 기반 저널리즘으로 진화 중



영·호남출신검사 검찰요직 등
점유비율 (단위:%)



컴퓨터이셔널 저널리즘(Computational Journalism)

데이터 저널리즘

데이터 수집→분석→시각화를 거쳐
사회적 의미 도출

로봇 저널리즘

데이터를 바탕으로 소프트웨어가
뉴스를 자동 작성 (예: 지진봇, 스포츠 리캡)

AI 저널리즘

수용자와 커뮤니케이션하며 원하는 뉴스와 답을 제공. LLM 기반 에이전트 활용 확대

* Stanford Computational Journalism Lab: 컴퓨터이셔널 방법을 활용해 책임 있는 보도를 가능하게 하는 연구

Part 2

데이터 저널리즘의 방법 수집 · 분석 · 시각화

데이터 수집

기존 데이터 발굴

- 데이터 아카이브 활용 (공공데이터포털, 국가통계포털)
- 정보공개청구 (open.go.kr)
- 기관에 직접 요구·입수

현장 관찰·인터뷰·설문

- 직접 데이터 구성 (설문조사, 현장 관찰)
- AI 녹취·번역 도구 활용 (클로바노트, Otter 등)

웹 등에서 데이터 추출

- Web scraping / crawling (Python, R)
- API 활용 (공공 API, SNS API 등)
- AI 기반 자동 수집 (LLM + MCP 등)

주요 데이터 소스

- data.go.kr / kosis.kr / 통계청 / DART
- 빅카인즈 / 정보공개포털 / 환경통계포털

주요 데이터·통계 사이트

분야	사이트	URL
종합	공공데이터포털	data.go.kr
종합	국가통계포털 (KOSIS)	kosis.kr
종합	정보공개포털	open.go.kr
뉴스	빅카인즈	bigkinds.or.kr
경제	한국은행 경제통계시스템	ecos.bok.or.kr
기업	전자공시시스템 (DART)	dart.fss.or.kr
환경	환경통계포털	stat.me.go.kr
교통	국토교통통계누리	stat.molit.go.kr
노동	노동통계정보시스템	laborstat.moel.go.kr
방송	방송통계포털	mediastat.or.kr

* 사이트 모음: https://jonghxxx.github.io/2026_1/datajour_source.html

데이터 분석

통계 분석

- 기술통계, 회귀분석, 상관분석
- Python (Pandas, SciPy)
- R (tidyverse, ggplot2)
- Excel 피벗테이블

텍스트·NLP 분석

- 형태소 분석, 감성 분석
- 토픽 모델링 (LDA)
- BERT, LLM 기반 분류
- 문서 요약·구조화

머신러닝·AI

- 분류(Classification)
- 이미지·영상 분석 (Vision AI)
- 패턴 탐지 (Change Detection)
- **LLM API 대규모 분석**

데이터 시각화

주요 시각화 유형 (Stalph, 2018 기반)

시각화 유형	비중	대표 도구
Bar Chart	32.7%	D3.js, Datawrapper, Flourish
Map	16.3%	QGIS, Mapbox, Leaflet, Kepler.gl
Line Chart	11.6%	Datawrapper, Flourish, Chart.js
Pie/Donut Chart	8.0%	Datawrapper, Google Charts
Table	6.5%	DataTables, AG Grid
Dashboard	8.2%	Tableau, Flourish, Streamlit
Infographic	4.1%	Canva, Napkin AI, Figma
기타 (네트워크, 워드클라우드 등)	12.6%	Gephi, Observable, D3.js

Part 3

AI 시대의 데이터 저널리즘

1.0 → 2.0 → 3.0

AI는 기자의 "두 번째 뇌(Second Brain)"

멀티플랫폼+AI 시대의 도전

- 웹·유튜브·SNS 동시 대응 필수
- 소수 인력으로 더 깊이 있는 데이터 생산 요구
- 심층기획 + 데이터 + 시각화로 경쟁력 강화
- 지속 가능한 뉴스 프로젝트 필요: 실시간 대시보드, 챗봇, RAG DB

취재 워크플로우 9단계와 AI

- 1) 아이디어 발굴 및 가치 평가
- 2) 정보수집·리서치·모니터링
- 3) 데이터 수집·분석·시각화
- 4) 인터뷰·현장취재 (녹취·번역)
- 5) 기사 작성·요약·검토
- 6) 사실검증 (Fact Check)
- 7) 여론 반응 분석
- 8) 숏폼/SNS 제작
- 9) 지식 관리 (Knowledge DB)

핵심 개념: SaaS → PaaS → IaaS

구분	SaaS (Software as a Service)	PaaS (Platform as a Service)	IaaS (Infrastructure as a Service)
비유	완성된 음식 배달	밀키트 (재료+주방)	전문 주방 구축 + 시장보기
역할	사용자 (User) 사용만 함	활용자 (Utilizer) 조합 및 개발	설계자 (Architect) 인프라 구성·운영
예시	Gmail, Tableau, Excel	OpenAI API, Vercel	Agent System, 챗봇, DB
데이터 저널리즘	1.0 전통적 도구 활용	2.0 AI 활용 저널리즘	3.0 에이전트 설계 저널리즘

데이터 저널리즘 1.0: 전통적 도구 활용

특징

- 정보수집: 포털 검색, 정보공개청구
- 데이터: 통계청 다운로드 → CSV 변환
- 인터뷰: 녹음기 + 수동 전사(타이핑)
- 분석: 엑셀 피벗 테이블, R 통계

한계점

- 데이터 수집/입력이 대부분 수작업
- 속도와 규모의 제약
- 정적(Static) 그래픽 중심
- 실시간 인터랙티브 구현 어려움

대표 도구: Excel, Google Sheets, Tableau, Datawrapper, R

데이터 저널리즘 2.0: AI 활용 저널리즘

기존 워크플로우 위에 LLM과 AI 도구를 얹어 생산성 극대화. 반복 업무는 AI에게, 해석·판단은 기자에게.

분야별 도구

- 리서치: Perplexity, Genspark, NotebookLM, Deep Research
- 녹취/요약: 클로바노트, Otter, 다글로
- 시각화: Napkin AI, Flourish, Datawrapper
- 코딩: Claude, Cursor, GitHub Copilot
- 이미지 생성: DALL-E, Midjourney

AI 기술 적용 영역

- NLP: 문서 구조화, 요약, 텍스트 분석
- Vision AI: 이미지·영상, 항공사진, CCTV 분석
- Audio AI: 음성 기록 텍스트화
- API: 대규모 데이터 분석에 LLM 적용
- **프롬프트 엔지니어링이 핵심 기술**

데이터 저널리즘 3.0: 에이전트 설계 저널리즘

코드 리터러시를 바탕으로 취재·분석·검증 시스템을 직접 구축

시스템 구성요소

- 데이터 레이어: 기사 아카이브, 팩트체크 DB
- LLM 레이어: 자체 미세조정 모델, RAG
- 오케스트레이션: 에이전트 (LangGraph, CrewAI)
- MCP(Model Context Protocol)로 도구 연결

기자의 새로운 역할

- Agent: 취재 보도 등 특수 역할 수행
- Architect: 워크플로우 설계 및 모니터링
- Knowledge Engineer: RAG DB 구축 관리
- 'journalist-programmer' 하이브리드 역할 부상

종합 비교: 1.0 vs 2.0 vs 3.0

구분	1.0 (SaaS)	2.0 (PaaS)	3.0 (IaaS)
핵심 가치	정확한 분석 (Analyst)	효율적 생성 (Generator)	시스템 설계 (Architect)
개발 필요	없음 (Tool 사용)	중간 (Prompting)	필수 (Coding/System)
코딩 방식	N/A	Vibe Coding (위험)	Augmented Coding
결과물	정적 기사/그래픽	텍스트·이미지·음성 대규모 분석	지속 가능한 시스템/DB
지속성	일회성 보도	프로젝트 단위	항시적 모니터링

Part 6

사례 모음

국내외 데이터 저널리즘

1.0 사례: 대규모 DB 구축의 끈기

경찰 비위 기록 통합 DB (USA Today, 2019)

- 미국 전역의 경찰 비위 기록 추적·공개
- 흩어진 문서를 스크래핑, PDF를 일일이 확인하여 데이터 입력
- 기술적 난이도보다 '데이터 수집의 끈기'가 돋보인 사례
- 유형: 대규모 DB 구축

ProPublica: The Prescribers (2013)

미국 정부의 처방 데이터를 분석해 의사별 처방 패턴 공개. 의사 개인 조회 가능.

Buzzfeed: Spies in the Skies (2016)

비행 데이터를 분석해 미 정부 감시 비행기의 경로 추적·시각화.

2.0 사례: AI로 '없던 데이터' 만들기

우크라이나 전쟁 피해 지도 (Texty.org.ua, 2024)

- 위성 사진을 ML 모델에 학습, 파괴 건물 자동 식별·매핑
- 인간이 불가능한 규모의 반복 노동을 AI가 대체

아마존 불법 채굴 자동 탐지 (Amazon Mining Watch, 2025)

- 위성 사진 Change Detection으로 숲→채굴장 변화 감지
- 실시간 모니터링 플랫폼으로 감시 영역 확장

패션 잡지 피부색 분석 (The Pudding, 2019)

- 얼굴 인식 AI로 Vogue 표지 모델 피부색 추출·데이터화
- 이미지를 사회학적 증거로 변환하는 '문화적 데이터 저널리즘'

AI 데이터센터 전기료 분석 (Bloomberg, 2025)

- 25,000개 그리드 노드를 분석, 데이터센터 인근 전기 요금 최대 267% 급등 확인
- AI 산업의 사회적 비용을 데이터로 검증

3.0 사례: 시스템을 구축하는 저널리즘

부패 감지 AI 'Funés' (Ojo Público, 2020)

- 공공 계약 데이터에서 업체 간 관계·정치 후원금 분석
- 회사명 입력 → '부패 위험 점수' 계산, 취재 타겟 추천

페이스북 알고리즘 감시 (The Markup, 2021)

- 성별·인종·지역·정치성향별 패널의 뉴스피드 자동 수집·분석
- 기자가 직접 갈 수 없는 곳에 '디지털 대리인' 파견

코로나 전파 시뮬레이터 (Washington Post, 2020)

- 사회적 거리두기 효과를 인터랙티브 시뮬레이션으로 설명
- 독자 참여형 시뮬레이션의 대표 사례

AI로 수백 시간 인터뷰 분석 (New York Times, 2025)

- 상원 인사청문 영상과 줌 미팅 수백 건을 LLM으로 분석
- "물리적으로 불가능한 청취"를 AI가 해결 — 'Cheat Sheet'라는 내부 AI 도구 개발

한국 데이터 저널리즘 사례

뉴스타파: 가짜 학회 참여 학자 공개 (2018-2019)

- OMICS Group, WASET 등 가짜 학회 참여 학자 데이터 분석·공개

중앙일보: 우리 동네 의회 살림 (2018)

- 2014~2018 기초의회 가계부 검사, 인터랙티브 조회 시스템 제공

2024 한국 데이터 저널리즘 어워드 수상작

- 자녀 살해 후 자살 : 비극을 기록하다 (데이터 기반 탐사보도 상)
- 당신이 버린 옷의 최후 (데이터 시각화 상)
- 명태균 게이트, 판도라의 상자를 열다 (오픈 데이터 상)
- 이태원 기억담기 캠페인 (오픈 데이터 특별상)
- 매니페스토 – 내일을 바꾸는 약속 (데이터저널리즘 혁신 상)
- 같은 더위는 없다 – 폭염 불평등 실태 보고서 (기후위기 특별상)

* http://datajournalismawards.kr/award_list/

Part 7

관련 단체 · 리소스 · 데이터 소스

주요 단체와 협업 네트워크

IRE (Investigative Reporters & Editors)

1975년 설립. 5,000여명 회원. 27,000여건 탐사보도 사례. NICAR(1989) 산하 운영. ire.org

GIJN (Global Investigative Journalism Network)

2003년 설립. 72개국 163개 회원사. 데이터저널리즘 지원. SIGMA Awards 시상. [Data Journalism Top 10](http://DataJournalismTop10.gijn.org). gijn.org

ICIJ (International Consortium of Investigative Journalists)

1997년 설립. 60개국 160여명. 파나마 페이퍼스(2016) 등 세계적 협업 탐사보도. icij.org

한국탐사저널리즘센터 (뉴스타파)

비영리 독립언론. 심층 탐사보도와 데이터저널리즘 특화. newstapa.org

DataJournalism.com

European Journalism Centre에서 2018년 설립. 데이터 자료·교육·커뮤니티 제공.

취재보도용 AI 도구 가이드 (2025)

단계	도구	용도
리서치	Perplexity, Genspark, NotebookLM, Deep Research	검색·요약·배경조사
데이터 수집	Python (BeautifulSoup, Selenium), LLM API	웹스크래핑·API·자동수집
데이터 분석	Python (Pandas), R, Claude/ChatGPT Code Interpreter	통계·NLP·ML 분석
녹취·번역	클로바노트, Otter, 다글로, Whisper	음성→텍스트 자동전환
시각화	Datawrapper, Flourish, Napkin AI, D3.js	차트·맵·인터랙티브
기사 작성·검토	Claude, ChatGPT, Gemini	초고 작성·교정·요약
사실검증	ClaimBuster, Google Fact Check Tools	주장 검증·교차확인
에이전트	LangGraph, CrewAI, Claude Code, Gemini CLI	자동화 파이프라인 구축

* AI 도구 모음 : jonghxxx.github.io/2026_1/datajour_ai.html

Part 8

제언

"AI를 사용했느냐?"가 아니라,
"AI 덕분에 더 깊이 취재했는가?"

"AI는 기사를 대체하지 않습니다.
하지만 **시스템을 구축하는(3.0) 기자**가
도구만 쓰는(1.0) 기자와 AI를 활용하는(2.0) 기사를
대체할 수 있습니다."