creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

석 사 학 위 논 문

중소기업 인공지능 내재화를 위한 데이터세트 구축 자동화 시스템 개발 연구

HANSUNG UNIVERSITY

2026년

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합컨설팅전공

장 진 호

HANSUNG
UNIVERSITY

석 사 학 위 논 문
지도교수 최동준

중소기업 인공지능 내재화를 위한 데이터세트 구축 자동화 시스템 개발 연구

Development of an Automated System for Dataset Generation:
A Consulting Approach for SME AI Internalization

HANSUNG UNIVERSITY

2026년 6월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합컨설팅전공

장 진 호

석 사 학 위 논 문
지도교수 최동준

중소기업 인공지능 내재화를 위한 데이터세트 구축 자동화 시스템 개발 연구

Development of an Automated System for Dataset Generation:
A Consulting Approach for SME AI Internalization

위 논문을 컨설팅학 석사학위 논문으로 제출함

2026년 6월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합컨설팅전공

장 진 호

장진호의 컨설팅학 석사학위 논문을 인준함

2026년 6월 일

심사위원장 주 형 근 (인)

심 사 위 원 정 진 택 (인)

심 사 위 원 최 동 준 (인)

HANSUNG UNIVERSITY

국 문 초 록

중소기업 인공지능 내재화를 위한 데이터세트 구축 자동화 시스템 개발 연구

한성대학교 지식서비스&컨설팅대학원
스 마 트 융 합 컨 설 팅 학 과
스 마 트 융 합 컨 설 팅 전 공
장 진 호

전 세계 산업 분야에서 인공지능 기술의 활용이 급증하고 있다. 인공지능 기술은 사람이 하는 일의 일부를 대신해 주는 도구에 머물지 않고 조직의 운영 구조와 판단 체계를 스스로 학습하고 개선하는 지능형 운영시스템으로 빠르게 발전하고 있다. 산업 분야와 직무에 따라 인공지능 기술 사용 수준에 차이가 있겠으나, 인공지능 기술을 내재화(內在化)한 기업과 그렇지 않은 기업 간의 격차가 크게 벌어지고 있다. 인공지능 전환(AX)과 활용은 비즈니스 환경에서 기업의 생존을 결정하게 될 것이 틀림없다.

인공지능 기술 내재화를 위해 우리 중소기업은 무엇을 준비하면 좋을까? 인공지능 기술 내재화는 인공지능 운영에 필요한 지식과 기술 경험을 조직 내부에 축적하여 외부 의존도를 낮추고 자체 관리 및 발전 역량을 갖추는 과정이다. 연구자는 중소기업이 지금 시작할 수 있는 인공지능 내재화의 시작점은 데이터와 정보의 축적, 즉 인공지능 학습을 위한 지식의 내재화가 인공지능 전환을 위해 해야 할 일이라고 판단했다. 중소기업의 재직 근로자는 필요할 때 언제든지 인공지능 도구를 활용할 수 있는 역량을 확보해야 하고, 중소기업은 이때 조직의 목표 달성에 필요한 데이터를 보유하고 있어야 한다는 의미이다. 이를 위해서는

항시 기업이 분석, 활용할 수 있는 최신 데이터를 자동으로 확보하여 보유할 수 있는 체계를 갖추는 것이 매우 중요하다. 이 연구는 가장 보편적인 데이터 확보 방식인 인터넷 웹사이트에 게시된 정보, 첨부파일 등 데이터를 수집하여 AI 활용을 위한 데이터세트로 저장, 보관하는 기능을 내재화할 수 있는 정보시스템을 구현하고 실제 동작시켜 실증결과를 제시하고 전문가 검토 등을 거쳐 시사점을 제시하였다.

【주요어】 인공지능 기술 내재화, 인공지능 학습 데이터세트, 웹사이트 정보 수집, 인공지능 전환 컨설팅

HANSUNG UNIVERSITY

<목 차>

Ⅰ. 서론 ····· 1

1.1 연구의 필요성 ····· 1

1.2 연구의 목적 ····· 2

1.3 연구 방법 및 논문의 구성 ····· 4

1.3.1 연구 방법 ····· 4

1.3.2 논문의 구성 체계 ····· 5

Ⅱ. 선행연구 고찰 ····· 6

2.1 기술 내재화 연구 ····· 6

2.2 웹사이트의 특징 및 유형 연구 ····· 8

2.2.1 온라인 자료, 웹사이트, 홈페이지의 정의 ····· 8

2.2.2 웹사이트 구동 환경 ····· 10

2.2.3 웹사이트 구성 요소 ····· 11

2.3 웹사이트 정보 수집 연구 ····· 13

2.3.1 웹사이트 정보의 자원화, 웹기록물 개요 ····· 13

2.3.2 웹기록물 수집 방법 ····· 14

2.3.3 웹 기록물 보존 포맷 ····· 14

2.4 인공지능 학습과 검색 증강(RAG) ····· 17

2.4.1 인공지능 학습 일반 개념 ····· 17

2.4.2 인공지능 데이터세트 연구 ····· 17

Ⅲ. 연구 방법 및 절차 ····· 20

3.1 연구 방법 ····· 20

3.2 연구 절차 ····· 20

3.2.1 (1단계) 대상 웹사이트 및 수집 기술 등 선정 ····· 20

3.2.2 (2단계) 웹사이트 수집 시스템 개발 ····· 30

3.2.3 (3단계) 웹사이트 정보 수집 실증 ····· 35

3.2.4 (4단계) 데이터세트 생성 및 인공지능 활용 ····· 36
3.2.5 전문가 그룹 리뷰 ····· 37

Ⅳ. 데이터세트 구축 자동화 시스템 ····· 38
4.1 웹사이트 수집 및 전처리 기술개발 ····· 38
4.1.1 웹사이트 수집 시스템 개발 ····· 38
4.1.2 전처리 및 학습 데이터세트 생성 기술개발 ····· 40
4.2 웹사이트 정보 수집 및 AI 학습 데이터세트 생성 ····· 43
4.3 AI 학습 데이터세트 품질 검토 ····· 45
4.4 전문가 그룹 리뷰 ····· 49
4.4.1 인터뷰 개요 ····· 49
4.4.2 인터뷰 참여자 ····· 49
4.4.3 분석 결과 검토 의견 ····· 50
4.5 소결 ····· 53

Ⅴ. 결론 및 제언 ····· 55
5.1 연구 결과 요약 및 시사점 ····· 55
5.1.1 연구 결과 요약 ····· 55
5.1.2 시사점 ····· 56
5.2 정책 제언 및 연구의 한계 ····· 57
5.2.1 정책 제언 ····· 57
5.2.2 연구의 한계 ····· 58

참 고 문 헌 ····· 59

표 목차

[표 1-1] 연구 방법 ····· 4
[표 1-2] 논문의 구성 ····· 5
[표 2-1] AI 활용 촉진(Accelerator) 확보 ····· 7
[표 2-2] 내재화, 기술 내재화, AI 기술 내재화 정의 ····· 7
[표 2-3] 웹사이트, 홈페이지 정의 ····· 8
[표 2-4] 표면 웹, 심층 웹, 다크웹 비교 ····· 10
[표 2-5] 웹사이트 구동 환경 ····· 11
[표 2-6] 사이트 구성 요소별 포맷 ····· 12
[표 2-7] 웹사이트 및 웹기록물 정의 ····· 13
[표 2-8] 웹 기록물 보존 포맷 ····· 16
[표 2-9] 공공데이터 제공기관 데이터 제공 포맷 ····· 18
[표 3-1] 연구 방법 및 절차 ····· 20
[표 3-2] 수집 대상 웹사이트 사전 조사 ····· 21
[표 3-3] 그룹A, 정책 자료, 지원 사업 웹사이트 URL ····· 22
[표 3-4] 그룹B, 정책자금, 융자, 보증 사업 웹사이트 URL ····· 23
[표 3-5] 그룹C, 시장동향, 기술표준, 기술 트렌드 웹사이트 URL ····· 24
[표 3-6] 그룹D, 해외시장, 해외마케팅 웹사이트 URL ····· 25
[표 3-7] 데이터세트 포맷별 토큰 사용량[단위: 토큰] ····· 27
[표 3-8] AI 학습 데이터세트 포맷 비교 ····· 29
[표 3-9] 웹 크롤링 레이어 주요 기능 ····· 32
[표 3-10] 전처리 레이어 주요 기능 ····· 33
[표 3-11] 데이터 세트 레이어 주요 기능 ····· 33
[표 3-12] 웹사이트 수집 시스템 개발 언어 ····· 34
[표 3-13] 웹사이트 수집 시스템 동작 환경 ····· 34
[표 4-1] 수집 대상 관리를 위한 핵심 엔티티 구조 ····· 39
[표 4-2] 데이터세트 생성 결과 ····· 45
[표 4-3] 품질 검토 결과 ····· 48

[표 4-4] 인터뷰 참여자 정보 ····· 49
[표 4-5] (인터뷰 요약1) 데이터세트 구축 자동화 시스템 의견 ····· 50
[표 4-6] (인터뷰 요약2) 웹 아카이빙 포맷과 데이터세트 포맷 의견 ···· 51
[표 4-7] (인터뷰 요약3) 시스템 활용 확산을 위해 필요한 조언 ····· 52

그림 목차

[그림 3-1] 데이터세트 구축 자동화 시스템 - 3 레이어 ····· 31
[그림 4-1] WARC 파일 헤더 예시 ····· 41
[그림 4-2] JSON 구조 예시 ····· 42
[그림 4-3] 전처리 및 직렬화 결과 ····· 44

HANSUNG UNIVERSITY

Ⅰ. 서론

1.1 연구의 필요성

전 세계 산업 분야에서 인공지능 기술의 활용이 급증하고 있다. 인공지능은 10년 만에 약 10만 배 발전 했으며 이는 무어의 법칙이 예측한 100배 향상을 크게 능가하는 속도이다(성동훈, 2025). 인공지능은 사람이 하는 일의 일부를 대신해 주는 도구에 머물지 않고 조직의 운영 구조와 판단 체계를 스스로 학습하고 개선하는 지능형 운영시스템으로 빠르게 발전하고 있다(LG CNS, 2025).

이재명 대통령은 2025년 11월, 2026년도 정부 예산안 국회 시정연설에서 “AI 시대, 하루 늦으면 한 세대 뒤처지게 된다”며, “농경 사회에서 산업 사회로, 산업 사회에서 정보 사회로 전환한 것처럼 AI 사회로의 전환은 피할 수 없는 필연”이라 강조했다(국정브리핑, 2025). 대기업 L사는 자체적으로 개발한 AI 플랫폼을 이용해 기존 약 22개월이 소요되던 신물질 및 신약 개발 연구 과정을 하루만에 완료했다고 밝혔다(박성완, 2026). L사는 이 신소재와 신약 개발을 지원하는 ‘AI 연구동료’가 논문과 특허, 분자구조, 이미지 등 다양한 형태의 멀티모달 데이터를 분석해 기존보다 수십 배 빠른 속도로 유망한 후보 물질을 발굴함으로써 개발 기간을 획기적으로 단축했다(박성완, 2026). 위 L사의 사례를 산술적으로 계산해 보면 ‘AI 연구 동료’를 가진 기업이 하루에 할 수 있는 일을 그렇지 않은 기업은 22개월이 소요된다는 의미인데, 쉬지 않고 일하는 인공지능의 1년(365일)으로 환산해 보면 AI 없이 이용하는 기업은 AI 연구 동료를 가진 기업과 같은 결과를 얻기 위해서 669.2년(365일×22월÷12월)의 시간이 필요하다는 의미가 된다.

산업 분야와 직무에 따라 AI 사용 수준에 차이가 있겠으나, AI 기술을 활용할 수 있는 기업 즉 AI를 내재화한 기업과 그렇지 않은 기업은 하루에 22개월 만큼 뒤처지고 있다고 볼 수 있다. 시간이 지날수록 더 큰 차이가 생기게 된다. AI 전환은 글로벌 비즈니스 환경에서 뒤처지지 않기 위해서 지체할 수 없는 현실이다. 인공지능 시대에 AI 내재화는 기업의 생존을 좌우하는 핵심 요소가 될 것으로 전망된다.

1.2 연구의 목적

한국 정부에서는 소버린 AI의 중요성을 강조하고 있다. 소버린 AI에 대한 일치된 정의는 부재하나 '특정 국가나 기업에 대한 기술적 종속을 회피하고 자국민의 데이터와 알고리즘 규범 환경에 대한 통제권을 유지하려는 시도'로 통용된다(윤정현, 2025). 가트너(Gartner社)는 소버린 AI를 "인프라, 데이터, 모델, 표준 전반에 걸쳐 AI의 개발, 배포, 활용을 독립적으로 통제할 수 있는 국가 또는 조직 능력"으로 정의하는데, 단순히 데이터를 자국 내에서 저장하는 것만으로 주권을 충족하지 못하며, 외국 법률의 적용을 배재할 수 있는 인프라 통제권이 수배되야 한다는 의미이다(이나연, 2025). 종합하면 소버린 AI는 국가·기업의 인공지능 기술의 종속성 없는 독립적 운영과 통제를 유지하는 능력과 활동으로 정리할 수 있다.

한국산업기술진흥원에서는 2026년 2월, 글로벌 메가트렌드와 정책·기술 이슈를 분석하여 'AI 내재화 시대 10대 유망산업'을 발표했다. 동 내용에는 '사용자 지시 없이 스스로 계획·실행하여 복잡한 업무를 자동 수행하는 AI 서비스 산업'을 '자율 에이전트 AI 산업'이라 정의 했다(KIAT, 2026). 인공지능 전환 시대 AI를 활용하는 기업이 갖추어야 할 준비 상황은 기업이 처한 여건에 따라 다를 수 있다. AI 관련 하드웨어나 솔루션·서비스 또는 시스템 통합 사업자 등 공급자가 아닌 기업, 특히 우리나라 기업의 대다수를 차지하는 중소기업에서는 무엇을 내재화해야 하는 것일까? 연구자는 앞서 L사의 사례에서 AI 연구 동료가 "다양한 자료"를 분석하여 결과를 냈다는 점에 주목해야 한다고 생각했다. 최신 인공지능 모델을 탑재한 고성능 하드웨어와 인공지능 모델을 갖추었다고 하여도 인공지능이 활용할 수 있는 데이터가 없다면 AI 도구를 활용하기 어렵기 때문이다. 똑똑한 인공지능 동료를 만들려면 개인의 머릿속에만 있는 지식을 체계적으로 정리하여 필요시 인공지능 도구를 학습 시켜 이용할 수 있어야 한다.

인터넷에는 전 세계 사용자들이 등록한 정보가 게시되어 있다. 인터넷에는 수천억 개의 방대한 웹페이지가 검색엔진 등에 색인되어 사용자가 키워드 검색 등의 방법으로 직접 열람하는 방식으로 사용해 왔다. 이제 인공지능 기술이 학습을 위해 방대한 인터넷 정보를 이용한다. 그러나, 인터넷 정보는 시간이 지남에

따라 지속적으로 사라지는 특징이 있다. 퓨 리서치 센터의 2024년 연구에 따르면 2013년부터 2023년 사이에 색인된 웹페이지의 상당수가 지금은 접속할 수 없다(Athena Chapekis et. al. ,20240). 오래된 웹페이지의 경우 더 심각하다. 2013년에 존재했던 웹페이지의 약 38%는 현재 접속할 수 없는 반면, 2023년에 접속한 웹페이지의 8%를 2024년에 접속할 수 없다. 위키백과 작성의 근거나 참고 자료로 연결되어 있는 링크 웹페이지의 54%는 더 이상 존재하지 않는다.

연구자는 중소기업이 지금 시작할 수 있는 AX 전환은 분산된 데이터와 정보의 축적, 즉 지식의 내재화가 AX 경쟁력 확보를 위해 해야 할 일이라고 판단했다. 중소기업 재직 근로자는 필요할 때 언제든지 AI를 이용하고 활용하는 역량을 확보해야 하고, 중소기업은 이때 조직의 목표 달성에 필요한 데이터를 보유하고 있어야 한다. 이를 위해서는 항시 기업이 활용해야 할 최신 데이터 보유 체계를 갖추는 것이 필요하다. 이 연구는 가장 보편적인 데이터 확보 방식인 인터넷 웹사이트에 게시된 정보, 첨부파일 등 데이터를 수집하여 AI 활용을 위한 데이터세트로 저장, 보관하는 기능을 내재화할 수 있는 정보시스템을 구현하고 실제 동작시켜 실증결과를 제시하는 것을 범위로 한다.

1.3 연구 방법 및 논문의 구성

1.3.1 연구 방법

중소기업 AI 내재화를 위한 인공지능 데이터세트 구축 시스템 개발 연구는 총 4단계로 진행한다. 첫 번째 단계는 중소기업이 필요로 하는 데이터 수집 대상 인터넷 웹사이트를 선정하는 절차이다. 두 번째 단계는 인터넷 웹사이트의 정보를 수집하고 데이터세트를 생성하는 시스템을 개발하는 절차이다. 세 번째 단계는 위 시스템을 동작시켜 중소기업이 필요한 웹사이트와 웹사이트에 저장된 첨부파일 등을 수집하는 절차이다. 이 단계에서는 지식정보를 제공하는 동적 웹페이지를 특정하여 진행한다. 네 번째 단계는 수집한 데이터를 가공하여 인공지능 데이터세트를 저장하는 절차이다. 구축한 인공지능 데이터세트는 실제 웹사이트의 게시물과 비교하여 정상 수집을 확인한다. 이후 최종적으로 전문가의 검토를 거쳐 동 시스템의 이용에 관한 시사점을 도출한다.

[표 1-1] 연구 방법

1단계	2단계	3단계	4단계
대상 웹사이트 선정	**웹사이트 수집 시스템 개발**	**웹사이트 정보 수집 실증**	**데이트세트 생성 및 인공지능 활용**
① 중소기업 지원 웹사이트 조사 ② 선정 기준 수립 ③ 웹사이트 선정 ④ 데이터세트 표맷 선정	① 웹사이트 수집 시스템 설계 ② 웹사이트 수집 시스템 개발	① 대상 웹사이트 정보 수집 ② 정상 수집 확인·재수집	① 데이터 전처리 ② 인공지능 데이터세트 생성 ③ 인공지능 학습 활용 테스트

⇩

전문가 검토 (개선 의견 및 활용 확대 방안 등)

1.3.2 논문의 구성 체계

이 연구는 총 5장으로 구성하였다. 제1장 서론에서는 연구의 필요성, 연구의 목적, 연구 방법과 논문의 구성을 다룬다. 제2장 선행연구 고찰에서는 기술 내재화 연구, 웹사이트 특징 및 유형 연구, 웹사이트 정보 수집 연구, 인공지능 학습과 검색 증강에 관한 선행연구를 다룬다. 제3장에서는 연구 방법과 절차를 다룬다. 제4장에서는 인공지능 데이터세트 구축 시스템 구축, 웹사이트 정보 수집 활용을 실증한다. 제5장 결론 및 제언에서는 연구 결과를 요약하고 시사점을 다룬다.

[표 1-2] 논문의 구성

1	**서론**		
연구의 필요성	연구의 목적	연구방법 및 논문의 구성	
⇩			
2	**선행연구 고찰**		
기술 내재화 연구	웹사이트의 특징 및 유형 연구	웹사이트 정보 수집 연구	인공지능 학습과 검색 증강(RAG)
⇩			
3	**연구 방법 및 절차**		
수집 대상 웹사이트 및 데이터세트 포맷 선정	웹사이트 수집 시스템 설계 및 개발	웹사이트 정보 수집 실증 및 확인	인공지능 데이터세트 생성 및 활용 테스트
⇩			
4	**인공지능 데이터세트 구축 시스템 연구**		
웹사이트 정보 수집 시스템 설계 및 개발	대상 웹사이트 정보 수집 및 결과 확인	인공지능 데이터세트 생성	인공지능 데이터세트 활용 테스트
전문가 검토			
⇩			
5	**결론 및 제언**		
연구 결과 요약 및 시사점			
정책 제언 및 연구의 한계			

Ⅱ. 선행연구 고찰

2.1 기술 내재화 연구

내재화(Internalization, 內在化)란 어떤 성질 따위가 사물이나 일정한 범위의 안에 들어 있게 되는 것을 의미한다(국립국어원). 민간기업 뿐 아니라 공공기관의 경우 경영전략이나 전략과제 등 새로운 제도 등이 조직에 소개되고 정착되는 과정을 내재화라고 한다(이향수, 이성훈, 20016). 내재화란 제도와 규정의 이념과 근본 의도가 함께 전달되어 수용되는 과정으로 새로운 제도나 경영전략 등이 성공적으로 정착되어 조직의 성과 제고로 연계되기 위해서는 반드시 내재화 과정이 필요하다(이향수, 이성훈, 2016). 기술 내재화(Technology Internalization)란 단순히 새로운 기술을 도입하는 것에 그치지 않고, 프로세스, 인력, 도구의 통합을 통해 해당 기술을 조직 내부에 깊숙이 정착시켜 조직의 역량을 강화하는 일련의 과정을 의미한다(Xian Zhi Tian, 2014).

산업 AI 내재화는 AI를 활용하는 수요 기업의 능력을 제고하는 활동을 의미한다. 이는 제조 지식과 AI 코딩 지식의 융합을 통해 이루어지며, 특히 코딩 지식이 없는 제조 전문가도 AI 솔루션을 설계하고 활용할 수 있도록 지원하여 기업 내부에서 AI 역량을 강화하는 것을 목표로 한다(이창선, 홍점범, 2026). 한국 정부는 산업 AI 내재화를 해 「산업디지털 전환 촉진법」에 따라 「산업 인공지능(AI) 내재화 전략」을 수립했다(관계부처 합동, 2023). 동 전략 수립의 배경에는 한국의 강점인 “주력사업+ICT 경쟁력”을 토대로 산업 전반에 AI를 내재화하여 튼튼한 실물경제, 꺾이지 않는 수출강국 실현 등 “산업의 대전환”을 추진하는 것이다. AI 내재화 전략의 비전은 ‘산업 AI 내재화를 통한 글로벌 산업 강국 도약’으로 산업 전반에 AI를 내재화하여 AI 활용기업을 1%에서 2030년까지 30%로 확대 하기로 했다(관계부처 합동, 2023). 동 전략에서는 수요기업의 AI 활용을 용이하게 하고, 공급기업의 기술 역량을 강화할 수 있는 주요 AI 기반 기술 확보 계획을 수립했다. 주요 기반 기술은 ① 데이터 전처리, ② AI 활용 디지털 협업 Tool, ③ Low코드·No코드가 제시 되었다.

[표 2-1] AI 활용 촉진(Accelerator) 확보

구분	주요 내용
데이터 전처리	○ 수요기업 Raw 데이터를 AI가 학습 가능한 형태로 가공하는 기술개발 * 산업데이터 컨설팅 사업('23년, 10억원) 활용, 추후 예산 확대 추진
AI 활용 디지털 협업 Tool	○ 현장 작업자의 AI 활용 및 협업 지원하는 디지털트윈, AR/VR 개발 * 공장간·기업간 데이터 이동·협업 가능한 가상공장 구현('23~'25년)
Low코드 No코드	○ 非전문인력이 AI를 용이하게 활용할 수 있도록 Low코드·No코드 개발 * 전사적 DX 기술개발('23년~), 업종별·분야별 AI 기술 접목 기술개발 등

출처: 산업통상자원부(2023), 산업 인공지능(AI) 내재화 전략 심의 확정 보도자료

제조 기업의 AI 기술 내재화에 필요한 직무와 기술을 FGI로 분석한 연구에서는 AI 기술 내재화에 필요한 핵심 직무로 데이터 분석가, 데이터 엔지니어를 1순위로 선정하였고, 필요한 기술 우선순위로 컴퓨터 비전 기술, 생성형 AI 기술을 1위로 꼽았다(손승진, 2025). 필요 기술 스택 우선순위는 1순위 LLM, 2순위 파인튜닝과 파이썬을 선정했다(손승진, 2025).

AI 기술 내재화란 AI 시스템 운영에 필요한 지식, 기술, 경험을 조직 내부에 축적하여 자체적으로 관리하고 발전시킬 수 있는 역량을 갖추는 과정이다(알체라, 2026).

[표 2-2] 내재화, 기술 내재화, AI 기술 내재화 정의

구분	용어	용어
국립국어원 (우리말샘)	내재화	어떤 성질 따위가 사물이나 일정한 범위의 안에 들어 있게 되는 것
이향수,이 성훈(2016)	내재화	새로운 제도와 규정의 이념과 근본 의도가 함께 전달되어 수용되는 과정
Xian Zhi Tian(2014)	기술 내재화	새로운 기술을 도입하는 것에 그치지 않고, 프로세스, 인력, 도구의 통합을 통해 해당 기술을 조직 내부에 깊숙이 정착시켜 조직의 역량을 강화하는 일련의 과정
이창선, 홍정범 (2023)	산업 AI 내재화	AI를 활용하는 수요 기업의 능력을 제고하는 활동
알체라	AI 기술 내재화	AI 시스템 운영에 필요한 지식, 기술, 경험을 조직 내부에 축적하여 자체적으로 관리하고 발전시킬 수 있는 역량을 갖추는 과정

출처: 각 기관, 선행연구 정의 정리

2.2 웹사이트의 특징 및 유형 연구

2.2.1 온라인 자료, 웹사이트, 홈페이지의 정의

‘온라인 자료’에 관한 정의는 「도서관법」에서 확인할 수 있다. 「도서관법」 제3조(정의)에서 ‘온라인 자료’란 정보통신망(「정보통신망 이용촉진 및 정보보호 등에 관한 법률」 제2조제1항제1호의 정보통신망을 말한다. 이하 같다)을 통하여 공중송신(「저작권법」 제2조제7호의 공중송신을 말한다)되는 자료를 의미한다고 규정하고 있다. 같은법 시행령 제16조(온라인 자료의 수집)에서는 온라인 자료는 전자적 형태로 작성된 웹사이트, 웹자료 등으로 분류하고 있다.

웹사이트는 브라우저로 접속하는 인터넷 기반 자원이지만 운영 목적에 따라 분류되기도 한다. 구체적으로는 접근 권한과 기능에 따라 기관 내부망인 인트라넷, 협력사 간 연계망인 익스트라넷, 대중에게 개방된 공공 웹사이트의 3가지 유형으로 구분하고 각 용도를 규명한 연구가 존재한다(이지은, 2005).

[표 2-3] 웹사이트, 홈페이지 정의

구분	용어	용어
도서관법 제2조 9호	온라인 자료	“온라인 자료”란 정보통신망(「정보통신망 이용촉진 및 정보보호 등에 관한 법률」 제2조제1항제1호의 정보통신망을 말한다. 이하 같다)을 통하여 공중 송신(「저작권법」 제2조제7호의 공중 송신을 말한다. 이하 같다) 되는 자료를 말한다.
행정안전부 (2021)	웹 페이지	인터넷을 통해 텍스트, 그림, 영상, 음성 등의 정보를 제공하기 위해 만들어진 웹 브라우저상의 문서
	웹사이트	특정 서비스를 위해 구성된 웹페이지의 집합체
	홈페이지	인터넷상에서 어떤 웹사이트에 접속할 때 제일 먼저 보여지는 웹페이지를 말하나, 우리나라에서는 웹사이트를 지칭
국립국어원	홈페이지	개인이나 단체가 월드 와이드 웹에서 볼 수 있게 만든 하이퍼텍스트 * 순화대상어: 홈페이지, 순화어: 누리집, 원어 homepage(영)

출처: 각 기관, 연구자 정리

공공기관의 웹기록물 관리에 관한 근거는 「공공기록물관리에관한법률」에 근거를 두고 있다. 동법 시행령 제2조 제2호에서 '전자기록물이란, 전자적인 형태로 작성하여 송수신 또는 저장되는 전자문서, 웹기록물 및 행정정보 데이터세트 등의 기록을 정보자료'로 규정하며 전자기록물의 범위 내에 웹기록물을 포함하고 있다. 동 시행령 제2조의 제10호에서는 '웹기록물이란, 공공기관에서 운영 활동하는 웹사이트, 블로그, 소셜네트워크서비스(SNS) 등 웹을 기반으로 생산된 기록정보자료와 웹사이트 운영 및 구축과 관련된 관리정보'로 규정하고 있다. 전자기록물과 웹기록물의 정의에 근거하여 요약하면, 웹기록물은 전자기록물과 동등한 정보자료로, 공공기관에서 웹을 기반으로 생산한 자료와 운영과 구축에 관련된 정보를 포함하는 개념이다. 이는 웹의 특징으로 웹사이트는 다양한 프로그래밍 언어를 기반으로 만들어진 정적 또는 동적인 구조의 웹사이트에 콘텐츠가 탑재된 형태로 전산정보처리시스템에 등록되면, 온라인상의 사용자가 웹 브라우저 등의 도구를 이용하여 접속하여 이용하게 되는 구조이다.

웹을 기록물 관점에서 접근한 연구에서는 웹기록물을 '표면 웹기록물'과 '심층 웹기록물'로 구분하여 정의한 바 있다(차승준 외, 2009). '표면 웹기록물'을 사용자의 접속에 따라 동일하게 표현되는 정적인 문서의 집합, '심층 웹기록물'을 사용자의 요구 변경에 따라 갱신되는 웹기록물로 규정한다(차승준 외, 2009).

국제정보시스템감사통제협회에서는 웹사이트를 표면웹, 심층웹, 다크웹 3단계로 구분한다. 표면웹은 검색엔진 등을 이용하여 접속할 수 있는 일반적인 웹사이트와 소셜네트워크 서비스 등을 의미하고, 심층웹은 연구기관이나 정부의 데이터베이스 및 비용을 지불하고 사용하는 데이터베이스 등을 의미하고, 다크웹은 검색엔진으로 접속할 수 없는 인터넷의 일부 영역으로 광범위한 불법 거래 활동 영역으로 설명하였다(Wanda Archy, 2018). 최근 보고서에서는 웹을 5가지 또는 7가지 수준으로 구분하였다. 소셜네트워크, 포털, 이메일 서비스를 가상 상위 계층인 Level S로 분류하고 Clearnet이라 호칭한다. 일반적인 웹사이트와 공공기관·행정기관 웹사이트가 이 계층에 해당된다. 이후 일반사용자의 제한 영역에서 Level 1~5 까지 분류한다. Level 1의 웹을 '표면 웹'으로 분류하고 제한된 소셜네트워크, 아마존, 이베이 등으로 설명하며, Level 3의 웹을 '심층 웹'으로 분류하여 민감한 콘텐츠 및 데이터베이스 등으로 규정한다(장진호 외, 2022).

[표 2-4] 표면 웹, 심층 웹, 다크웹 비교

구분	용어	용어
차승준, 이윤정, 이규철 (2009)	표면 웹기록물	표면 웹 기록물은 접근할 때마다 동일하게 표현되는 정적인 문서의 집합으로 구성된 웹사이트로 기관 소개, 연혁 등의 자료가 이에 속함
	심층 웹기록물	사용자 요구가 변경될 때마다 저장된 데이터베이스 내용도 갱신되는 웹사이트로 키워드 검색, 공지 사항, 게시물, 자료실 등이 이에 속한다. (표면 웹의 부분이 아닌 웹의 컨텐츠로 일반적 검색엔진으로 접근되지 않는 것)
ISACA (2018)	표면 웹 (Surface Web)	검색엔진으로 접속하여 사용하는 대부분의 일반 웹사이트 * 검색엔진(구글, 빙 등), 소셜 네트워크(페이스북, 트위터, 넷플릿스 등), .com, .org와 같은 흔한 도메인 웹사이트
	심층 웹 (Depp Web)	웹 브라우저와 같은 일반적인 도구를 이용하여 접근할 수 있는 웹사이트로, 엄청난 양의 데이터를 제공하고 있으나 찾기 어려움 * 연구기관 및 정부의 데이터베이스, 구독하거나 비용을 지불해야 하는 웹사이트, 데이터 제공 웹사이트 등
	다크 웹 (Dark Web)	기존 검색엔진으로 액세스할 수 없는 인터넷의 일부이며, 액세스하려면 특별한 익명화 소프트웨어가 필요

출처: 각 기관, 연구자 정리

2.2.2 웹사이트 구동 환경

웹사이트가 구동되기 위해서는 기본적으로 컴퓨팅 구동 장치인 하드웨어가 필요하다. 최근에는 가상화 기술로 인해 물리적인 하드웨어뿐 아니라 클라우드 서버를 이용하기도 한다. 공공영역에서는 상당수의 공공기관이 국가정보자원관리원에서 제공하는 물리 또는 가상장치를 이용한다. 공공부문의 웹사이트는 업무 특성에 따라 자체 운영, 국가정보자원관리원, 산하기관 운영 등의 방식으로 운영되고 있다. 컴퓨팅 구동 장치는 운영체계, 웹서버, 애플리케이션 서버, 데이터베이스와 같은 기반 소프트웨어가 탑재된다. 이후 다양한 프로그래밍 언어로 개발된 응용시스템이 구동된다. 공공영역에서는 행정안전부에서 배포하는 전자정부 프레임워크를 이용한 JAVA 언어가 주로 사용된다. 또한 소스 코드를 직접 등록하는 대신 컴파일 또는 빌드 파일을 배포하는 구조로 진행되므로 체계적인 소스 통

합·형상 관리 및 컨테이너 오케스트레이션 도구 등을 이용하게 된다. 그 외 서버 보안, 백업, 모니터링 및 백신 등 보안 영역에서 탑재, 운영되는 요소는 매우 다양하고 서비스 특성에 따라 적절한 수단의 관리 도구가 탑재되어 운영된다. 보안을 위해 운영하는 DDoS 방지 장치, 침입탐지, 방화벽, 웹방화벽 등의 장치도 중요한 요소이다.

[표 2-5] 웹사이트 구동 환경

구분	내용
컴퓨팅 장치	하드웨어 서버, 클라우드 서브
가상화 도구	ZEN, Vmware, Hyper-V, Oracle VM, PROXMOX 등
운영체계	RedHat Linux, CentOS, Windows OS 등
웹서버	Apache Web, NginX, WebtoB 등
애플리케이션 서버	Tomcat, JBoss, JEUS 등
데이터베이스	MySQL, MariaDB, CUBRID 등
개발언어	ASP, PHP, JAVA 등
컨테이너 오케스트레이션 도구	Kubernetes, Docker Swarm, HashiCorp Nomad 등

출처: 이씨플라자(2022). 연구자 수정

2.2.3 웹사이트 구성 요소

가장 대표적인 웹사이트의 구성 요소는 HTML이다. 다양한 프로그래밍 언어로 개발되어 운영되는 웹사이트는 웹 브라우저로 이용하면서 HTML 코드로 재현된다. 또한 웹사이트 내외부의 타 시스템과 연동을 위해 XML을 사용하기도 한다. 최근에는 타 기관이 제공하는 웹사이트의 정보를 제공하기 위해 이용하는 공개된 API(OPEN API)가 다양하게 사용되는데 XML 구조와 JSON 방식이 주를 이루고 있다. 그 외에는 문서, 이미지, 동영상, 사운드, 압축파일, 기타 응용 시스템 연동 포맷 등이 사용된다. 새로운 응용프로그램의 등장에 따라 새로운 포맷이 개발되고 웹에 탑재되어 운영되고 있으나 최근에는 Non-Plugin을 선호

하고 있으므로 다양한 포맷을 그대로 웹사이트에 게시하지 않고 '서비스 포맷'으로 변환하여 웹 브라우저만으로 열람할 수 있도록 구축되어 있다. 최근까지도 다양한 포맷을 서비스에 제공하기 위해 플래시 기술을 이용하기도 하였다. 다양한 포맷의 파일이 플래시 플레이어를 경유하여 영상, 문서, 만화, 게임 등을 제공했으나, 플래시 기술 자체가 웹 브라우저에서 지원이 중단되면서 현재 플래시 기반 웹콘텐츠는 아예 이용이 불가능하게 되었다.

관련된 요소 들은 공공기관이 해당 파일을 현용 웹사이트에서 운영하고 있으면 '웹사이트 구성 요소'이나, 기록관리 목적으로 웹사이트의 정보자료를 보존하는 경우 '웹기록물 구성 요소'가 된다. 연구자는 현용과 보존 측면에서 웹사이트와 웹기록물을 구분하고 있으므로 공공기관의 자체 관리와 외부(중앙기록물관리기관 등)의 관리로 구분되는 개념은 아니다. 해외에서는 중앙행정기관 등의 웹사이트를 민간에서 수집·보존하는 서비스가 운영되고 있다.

최근에는 웹을 구성하는 다양한 콘텐츠가 제공되면서 동영상 자막파일(SMI, SRT), VR(VR tour), 3D 사진 등 기존 웹을 이루는 기술과 다른 방식도 등장하고 있다. 이 경우 응용시스템과 함께 별도의 구동 솔루션 등이 탑재되어 운영된다. 웹사이트 또는 웹기록물의 유형별 파일 포맷은 다음과 같으며 한정되지 않는다.

[표 2-6] 사이트 구성 요소별 포맷

구분	연구
웹문서	HTML, XML, CSS 등
문서	DOC(X), PPT(X), XLS(X), HWP(X), PDF, TIF 등
이미지	TIF, JPEG, GIF, BMF, PNG 등
동영상	MP4, AVI, MOV, ASF 등
사운드	MP3, MIDI, WAV, RIFF 등
압축파일	ZIP, RAR, ARJ, WAR, TAR 등
기타	RIS, VR Tour

출처: 국가기록원(2023). 연구자 정리

2.3 웹사이트 정보 수집 연구

2.3.1 웹사이트 정보의 자원화, 웹기록물 개요

웹 정보자원은 쉽게 변경되거나 사라지기 쉬운 휘발적 성격을 갖고 있으나, 현대 사회에서 지식정보 공유와 커뮤니케이션을 위한 웹에 대한 의존도는 거의 절대적으로 증가하고 있다(김희정, 2010). 따라서 가치 있는 웹 자원에 대해서는 별도의 아카이빙 작업을 수행하는 것이 필요하다(김희정, 2010).

[표 2-7] 웹사이트 및 웹기록물 정의

구분	용어	내용
도서관법	온라인 자료	“온라인 자료”란 정보통신망(「정보통신망 이용촉진 및 정보보호 등에 관한 법률」 제2조제1항제1호의 정보통신망을 말한다. 이하 같다)을 통하여 공중 송신(「저작권법」 제2조제7호의 공중 송신을 말한다. 이하 같다) 되는 자료
행정안전부 웹사이트 구축 운영 가이드	웹페이지	인터넷을 통해 텍스트, 그림, 영상, 음성 등의 정보를 제공하기 위해 만들어진 웹 브라우저상의 문서
	웹사이트	특정 서비스를 위해 구성된 웹페이지의 집합체
	홈페이지	인터넷상에서 어떤 웹사이트에 접속할 때 제일 먼저 보여지는 웹페이지를 말하나, 우리나라에서는 웹사이트를 지칭
	행정 공공 웹사이트	행정기관 등이 제공하는 정보 및 서비스를 인터넷에서 제공하는 가상공간으로, 특정 콘텐츠, 서비스 등을 위한 웹페이지들의 집합
국립국어원	홈페이지	개인이나 단체가 월드 와이드 웹에서 볼 수 있게 만든 하이퍼텍스트 * 순화대상어: 홈페이지, 순화어: 누리집, 원어 homepage(영)
이지은 (2006) NAA.GOV.AU (2000)	웹사이트	웹사이트는 인터넷의 특정 도메인을 가지며, 브라우저 기술을 이용하여 찾을 수 있는 링크된 정보자원의 조직 * 인트라넷(intranet), 익스트라넷(Extranet), 공공 웹사이트(Public website)로 구분
	공공 웹사이트	공공 웹사이트는 공공기관의 행정적 통제 아래에서 함께 링크되어 있는 전자파일 컬렉션으로, 월드와이드웹을 통해서 공공에게 접근 가능하도록 만들어져 있음

출처: 각 기관, 연구자 정리

2.3.2 웹기록물 수집 방법

현재 운영 중인 공공기관 및 중앙행정기관의 웹사이트 수집(아카이빙) 도구로 다양한 소프트웨어가 검토되었다. 주로 제시된 웹 아카이빙 도구는 Heritrix, Browsertrix, HTTrack, DeepARC, Wayback Machine 등이다. Heritrix는 오픈소스 기반의 웹크롤러로 약 20년간 국내외 다양한 조직에서 이용하고 있다. 웹사이트를 포괄적으로 수집·저장하는데 용이한 URI(Uniform Resource Identifier) 분석 기반 자동 페이지 수집기능이 매우 효율적이다. 다만, 웹사이트는 사용자 환경의 재미 또는 메시지 전달 등 최적의 효과와 경험을 전달하기 위해 사용자 환경의 웹 브라우저에서 동작하는 스크립트 의존도가 점점 증가하고 있다. 동적으로 렌더링 되는 소셜미디어 등에서는 이러한 현상이 더 분명하게 나타난다. 예를 들어 페이스북이나 트위터와 같은 소셜미디어의 경우 사용자의 스크롤을 인식하여 정해진 수량의 콘텐츠가 로딩된다. 이때, URL이나 URI의 변화가 없다. IT 기술자들은 '무한 스크롤'이라고 표현하기도 하는데, 이러한 구조는 Heritrix와 같이 URI를 이용하여 접근하는 웹사이트에서는 필요한 링크나 접속 경로를 확보할 수 없다.

Archive-it은 이러한 문제를 해결하기 위해 Heritrix를 보완하는 Umbra를 개발하였다. 동작 구조를 설명하면, Heritrix가 URL에 접근하는 방식으로 웹페이지를 수집하면서 Umbra를 이용하여 클라이언트 측 스크립트를 실행한다. 이에 Heritrix 단독으로 확보할 수 없었던 URL을 감지하여 크롤링할 수 있다.

2.3.3 웹 기록물 보존 포맷

OASIS는 국립중앙도서관에서 2001년 11월 '온라인저작물 수집·보존 시스템 구축 ISP 및 시범서비스 개발'을 기점으로 2005년 웹기반 OASIS 시스템 구축사업을 완료하고 2006년부터 웹 아카이빙 시스템 OASIS를 운영 중이다. 더블린코어(Dublin Core)의 15개 요소 외 내부 관리 메타데이터를 추가하여 총 32개의 메타데이터로 구성되어 있다. OASIS의 디지털 아카이빙 메타데이터는 설명적, 구조적, 관리적, 보존적 메타데이터로 구성되며, 웹 자원의 보존 포맷으로는

WARC가 활용된다(오상훈·최영선, 2009). 그러나 기존의 전자기록물 정보 패키지(SIP, AIP, DIP) 규격은 텍스트 기반의 전자문서에 편중되어 있어, 향후 웹기록물과 데이터세트 등 새로운 전자기록물 유형을 수용할 수 있는 패키지 규격 재설계가 필수적이다(김명옥·이상용, 2010).

이러한 신규 유형 보존을 위해 표준 웹기록물용 보존 포맷인 KoSurWeb(Korea Surface Web records)과 심층 웹기록물용 보존 포맷인 KoDeWeb(Korea Deep Web records)이 정의된 바 있다(유니닥스, 2009; 박병주 외, 2010). 이 중 KoSurWeb 포맷은 기록물 건, 심층 웹, 기술 메타데이터의 3개 메타데이터 영역과 1개의 콘텐츠 구조로 이루어져 있으며, 최종적으로는 WARC 포맷과 이를 설명하는 메타데이터가 결합된 형태를 취한다(유니닥스, 2009; 박병주 외, 2010). 웹 기록물은 ISO 28500으로 규정된 WARC를 표준으로 사용하고 있다. 2009년 5월 최초 규정 후 현재 2017년 8월에 개정된 Edition 2가 최신버전이다. 웹 사이트를 저장하고 보존하는 기술적 포맷 중에서 유일하게 국제 표준으로 채택된 것이 WARC이다.

이렇게 다양한 방식으로 수집된 웹기록물DB와 WARC는 수집 시점마다 전체를 수집함에 따라 필연적으로 중복이 발생한다. Heritrix의 경우 수집 로그에 URL(Uniform Resource Locator) 별로 고유한 해시값을 제공하며, 기존 수집한 자료를 유지하고, 신규 수집 웹기록물 만을 저장할 수 있다. 웹크롤러는 웹사이트 전체를 수집하는 동작을 반복함에 따라 스토리지 효율화를 위해서 중복제거 프로세스는 중요하다. 이연수외(2013)의 연구에서는 별도의 중복대상 리스트 목록을 DB로 관리하여, 수집된 URI의 해시값과 비교하는 방식으로 웹크롤러의 중복 필터링 기능을 개선하였다. 동 연구는 실증결과를 제시하였는데, 1차 웹기록물 수집 후 N차 웹기록물 수집 과정에서 테스트 대상 웹사이트에 대해 90% 이상의 중복 처리 결과를 제시하였다. 한편, 해당 연구에서는 행정안전부 14개의 상이한 URI 해시값이 동일한 현상이 발생하였다고 하였다(이연수 외, 2012). 제시한 2개 웹사이트의 공통점은 데이터를 입력하고 조회하는 화면으로 일반적인 웹페이지가 아니어서 정상 사례로 보기 어렵다. 동 연구는 웹기록물 수집시스템이 수집된 웹기록물과 기존 보유 웹기록물의 중복을 비교하여 중복을 회피하는 방식으로 매우 높은 성공률을 보였다.

[표 2-8] 웹 기록물 보존 포맷

구분	용어	내용	비고
박병주, 차승준, 이규철 (2010)	KoSurWeb (Korea Surface Web)	표면 웹기록물을 보존하기 위한 문서보존포맷 (DC 메타데이터 참조 + 기술 메타데이터), XML 구성	
	KoDeWeb (Korea Deep Web)	심층 웹기록물을 보존하기 위한 문서보존포맷 (SIARD 메타데이터 참조), XML 구성, ZIP64 파일 포맷으로 압축	
저장포맷	WARC	웹기록물 보존포맷 국제표준(ISO 28500)	
	K-WARC	WARC에 필요 메타데이터를 추가하여 개발한 한국 포맷	
	SingleFile (github 공식)	SingleFile은 Chrome, Firefox, Microsoft Edge, Vivaldi, Brave, Waterfox, Yandex Browser 및 Opera와 호환되는 웹 확장 도구 ※ 하나의 HTML 파일에 완전한 웹 페이지 저장 가능	
	SIARD (윤성호, 2021)	스위스 연방 기록원(Swiss Federal Archives, SFA)에서 개발한 관계형 데이터베이스 보존포맷으로, 소프트웨어로부터 독립해 데이터를 보존할 수 있는 파일 포맷	
	PDF (Adobe)	Adobe사가 개발한 개방형 표준 문서. ISO 32000 https://www.adobe.com/kr/acrobat/about-adobe-pdf.html	
	mht (Microsoft)	인터넷 익스플로러에서 웹 페이지를 저장하는 아카이브 포맷으로, 웹페이지를 하나의 파일로 저장할 수 있음	

출처: 각 기관, 연구자 정리

2.4 인공지능 학습과 검색 증강(RAG)

2.4.1 인공지능 학습 일반 개념

GPT-3와 GPT-4 등 대규모 언어 모델(LLM)은 뛰어난 자연어 처리 능력을 보여주나, 사전 학습된 내부 파라미터에만 의존해야 하는 근본적인 제약을 지닌다(윤여찬, 김수균, 2025). 학습 정보가 부족할 경우에는 정보의 최신성이 결여되고 사실과 다른 정보를 생성하는 환각(Hallucination) 현상을 유발할 위험이 높다. 이를 보완기 위해 문서 검색(Document retrieval)과 텍스트 생성을 결합하여 문맥이 풍부한 응답을 도출하는 검색 증강 생성((Retrieval-Augmented Generation, RAG) 기술이 주요한 대안으로 대두되었다(Bose, 2025).

RAG 시스템은 외부 소스(웹사이트 API, 데이터베이스, 문서 등)에서 수집한 데이터를 임베딩 언어 모델을 통해 벡터(Vector)로 변환하고, 이를 벡터 데이터베이스에 저장하여 검색 가능한 지식 라이브러리를 구축하는 방식으로 동작한다(SK.COM, 2025). 벡터화한 후 유사도 계산을 통해 가장 관련성 높은 문서를 검색(Retrieval)하고, 추출된 데이터를 LLM에 제공하여 정확한 문맥 기반의 응답을 생성하게 된다(SK.COM, 2025).

중소기업이 웹사이트 정보를 기반으로 AI를 내재화할 때, RAG 시스템은 제한된 자원 내에서 데이터 보안과 운영 효율성을 동시에 확보할 수 있는 전략적 가치를 제공한다. 기업은 외부 유출을 방지해야 하는 민감한 내부 데이터를 다룰 때 조직의 보안 요구사항을 충족하는 온프레미스(On-Premise) 기반의 RAG 아키텍처를 구축하여 정보 접근을 효과적으로 제어할 수 있다(조민수, 양진홍).

2.4.2 인공지능 데이터세트 연구

인공지능 지식 내재화를 위해 초기 데이터를 구축할 때는 관련 분야 공공데이터를 확보하는 것이 우선시 된다. 일반적으로 공공데이터 포털, AI 허브 등을 이용하여 데이터를 직접 내려받아 등록하거나 OPEN API를 이용하여 정보처리시스템 간 직접 데이터를 다운로드 받아 처리할 수 있다.

공공데이터 포털(www.data.go.kr)은 공공기관이 생성 또는 취득하여 관리하고 있는 공공데이터를 한 곳에서 제공하는 통합 창구로, 국민이 쉽고 편리하게 공공데이터를 이용할 수 있도록 파일데이터, 오픈 API, 시각화 등 다양한 방식으로 제공하고 있으며, 누구라도 쉽고 편리한 검색을 통해 원하는 공공데이터를 빠르고 정확하게 찾을 수 있다(공공데이터 포털, 2026).

공공, 민간 구분 없이 가장 많은 데이터를 확보할 수 있는 웹사이트는 One-윈도우(www.data1window.kr)이다. One-윈도우는 한국의 다양한 데이터를 한 곳에서 손쉽게 접근하고 활용할 수 있는 국가 데이터 인프라의 포털 서비스로, 민간과 정부·공공기관이 제공하는 방대한 데이터를 통합하여, 연구자 · 기업 · 정부 관계자 · 일반 시민들이 필요로 하는 정확하고 신뢰성 있는 정보를 쉽고 빠르게 제공하는 것을 목표로 한다(ONE-윈도우, 2026).

[표 2-9] 공공데이터 제공기관 데이터 제공 포맷

구분	제공 데이터 포맷
공공데이터 포털 (행정안전부)	특정 프로그램에 종속되지 않고 기계 판독(Machine-Readable)이 가능한 오픈 포맷(Open Format) 형태로 데이터를 제공
ONE-윈도우 (NIA)	데이터 제공 플랫폼과 연계하여 연구 및 분석에 바로 활용할 수 있도록 CSV, XLS(X), PDF, JSON 등 포맷으로 데이터 제공
AI 허브 (NIA)	데이터에 대한 메타 정보, 태그 등이 포함된 구조화된 데이터로 주로 JOSON 형식으로 제공
인공지능 데이터 공유·활용 서비스 (KISTI)	* 다운로드 방식: .json, .jonl, txt. csv, tex(LatTeX), jpg 등 * OPEN API : JSON/REST API
산업데이터 서비스 포털 (KOIIA)	파일 기반 데이터 : CSV, XLSX 등 시각화 통계자료 : 이미지(PNG, JEPG 등), 인터랙티브 웹 문서 리포트 : PDF , API/데이터 스페이스 : JSON, RESTful

출처: 각 기관, 연구자 정리

AI 허브(aihub.or.kr)는 과학기술정보통신부와 한국지능정보사회진흥원이 운영하는 국가 AI 개발 지원 플랫폼으로, AI 기술 및 제품·서비스 개발에 필요한 AI 인프라(AI 데이터, AI SW API, 컴퓨팅 자원)를 대한민국 국민이면 누구나(연구자, 기업, 공공기관 등) 활용할 수 있도록 제공 및 지원한다(AI허브, 2025). 지능

정보산업 인프라 조성을 위해 14개 분야 AI 학습용 데이터와 국내외 기관·기업에서 보유한 AI 학습용 데이터를 공개하고 있다. 한국과학기술정보연구원의 인공지능 데이터 공유·활용 서비스(aida.kisti.re.kr)에서는 RAG 데이터세트를 공개하고 있다. 동 데이터세트는 RAG 기반 전문 도메인 질의응답(QA) 시스템을 위한 학습 데이터를 제공한다. 각 샘플은 주어진 질문에 대한 논문 기반의 추론형 정답과 전문 도메인에 특화된 핵심 문구 및 핵심 문장으로 구성하여 전문 지식에 기반한 RAG 모델이 더욱 정밀하고 정보 집약적인 응답을 생성할 수 있도록 지원할 수 있도록 했다(한국과학기술정보연구원, 2025).

HANSUNG UNIVERSITY

Ⅲ. 연구 방법 및 절차

3.1 연구 방법

이 연구는 중소기업 지원 웹사이트를 대상으로 웹사이트 수집 시스템을 개발하고, 선별한 URL을 대상으로 지원 정보를 수집하며, 수집된 데이터를 전처리하여 인공지능(AI) 학습 데이터세트를 구축하는 것을 목적으로 한다. 이를 달성하기 위한 구체적인 연구 방법 및 절차는 대상 웹사이트 선정, 웹사이트 수집 시스템 개발, 웹사이트 정보 수집, 학습 데이터세트 구축의 총 4단계로 구성되며, 구축된 데이터의 신뢰성을 확보하기 위해 품질 검증과 전문가 검토를 수행한다.

[표 3-1] 연구 방법 및 절차

1단계	2단계	3단계	4단계
대상 웹사이트 선정	**웹사이트 수집 시스템 개발**	**웹사이트 정보 수집 실증**	**데이트세트 생성 및 인공지능 활용**
① 중소기업 지원 웹사이트 조사 ② 선정 기준에 따른 웹사이트 선정 ③ 데이터세트 표맷 선정	① 웹사이트 수집 시스템 설계 ② 웹사이트 수집 시스템 개발	① 대상 웹사이트 정보 수집 ② 정상 수집 확인·재수집	① 데이터 전처리 및 인공지능 데이터세트 생성 ② 인공지능 학습 활용 테스트

⇩

전문가 검토 (개선의견 및 활용확대 방안 등)

3.2 연구 절차

3.2.1 (1단계) 대상 웹사이트 및 수집 기술 등 선정

3.2.1.1 중소기업 지원 웹사이트 조사

중소기업이 영위하는 주된 산업 영역에 따라 수집 대상 웹사이트의 성격은 상이하다. 이 연구에서는 중소기업 지원 정책 및 정보 제공 플랫폼을 중심으로, 중소기업의 AI 내재화 구현에 필요한 데이터를 보유한 웹사이트 현황을 포괄적으로 조사하여 최종 분석 대상을 선정하였다.

대상 웹사이트 분류는 첫 번째, 정부 및 지자체의 지원 사업과 정책 자료는 중소기업의 생존 및 자금 확보에 직결되는 시의성 높은 데이터로서 기업 경영, 기술 개발, 인력 확보 등 전방위적 기업 활동을 견인하는 필수 동력이다. 두 번째 시장 동향, 기술 표준, 규제 정보 및 기술 트렌드 데이터는 중소기업의 신제품 기획과 연구개발(R&D) 방향성을 수립하는 핵심 지표이다. 세 번째, 마케팅 전략 수립 및 AI 기반 가격 제안 모델 구축을 위해서는 경쟁사 제품 데이터, 전문 플랫폼의 동종 제품 정보, 외부 커뮤니티의 고객 반응 및 VOC 데이터가 요구된다.

[표 3-2] 수집 대상 웹사이트 사전 조사

분류	기관명	URL	검토	선정
정책 자료, 지원 사업	중소벤처기업부	www.mss.go.kr		○
	기업마당	www.bizinfo.go.kr	구조상 수집불가	
	중소벤처24	www.smes.go.kr		○
	중소기업중앙회	www.kbiz.or.kr		○
	중소벤처기업연구원	www.kosi.re.kr		○
정책자금, 융자, 보증	중소벤처기업진흥공단	www.kosmes.or.kr		○
	기술보증기금	www.kibo.or.kr	구조상 수집불가	
	신용보증기금	www.kodit.co.kr	구조상 수집불가	
	소상공인시장진흥공단	www.semas.or.kr		○
	신용보증재단중앙회	www.koreg.or.kr		○
시장동향, 기술표준, 기술 트렌드	중소기업기술정보진흥원	www.tipa.or.kr		○
	중소기업 기술개발사업 종합관리시스템	www.smtech.go.kr		○
	한국산업기술진흥원	www.kiat.or.kr		○
	한국산업기술진흥협회	www.koita.or.kr		○
	중소기업뉴스	www.kbiznews.co.kr	구조상 수집불가	
해외 시장 정보, 해외 마케팅	대한무역투자진흥공사	www.kotra.or.kr		○
	한국무역협회	www.kita.net		○
	GEP 글로벌전시플랫폼	www.gep.or.kr		○
	한국전시산업진흥회	akei.or.kr		○

사전조사를 통해 수집 대상으로 선정하지 않은 웹사이트는 iframe 구조로 제작되었거나 아코디언 UI(펼치기와 접기)를 사용, 기술적으로 크롤링이 불가능한 구조로 분석되어 최종 대상에서 제외하였다.

3.2.1.2 수집 대상 웹사이트 기준 수립 및 웹사이트 선정

이 연구의 목적을 달성하기 위한 실질적인 데이터 수집 단계에서는, 선정된 대상 웹사이트의 전체 웹 데이터 중 연구의 본질적 취지와 부합하지 않는 단순 안내성 정보, 기관 소개 등의 일반 정보를 배제하였다. 대신 자연어 처리(NLP) 및 인공지능 모델 학습에 실질적으로 기여할 수 있는 정형·비정형 텍스트 기반의 유의미한 콘텐츠를 포함한 핵심 URL만을 선별하여 수집 대상으로 확정하였다. 이는 수집 과정에서의 데이터 노이즈를 최소화하고, 궁극적으로 중소기업 AI 내재화 모델의 학습 효율성과 예측 정확도를 제고하기 위함이다.

1) 정책 자료, 지원 사업 인터넷 정보 자원

중소기업 웹사이트 정보 내재화 대상 중 정책 자료, 지원 대상 정보 수집 대상은 4개 웹사이트 12개 게시물 메뉴로 정하였다. 대상 기관은 중소벤처기업부, 중소벤처24, 중소기업중앙회, 중소벤처기업연구원이다.

[표 3-3] 그룹A, 정책 자료, 지원 사업 웹사이트 URL

구분	분류 코드	메뉴명	URL
중소벤처기업부	A-01	중소벤처기업부 > 사업공고	https://www.mss.go.kr/site/smba/ex/bbs/List.do?cbIdx=310
중소벤처24	A-02	중소벤처24 > 사업공고	https://www.smes.go.kr/main/sportsBsnsPolicy
	A-03	중소벤처24 > 정책금융	https://www.smes.go.kr/main/plcyFnl
중소기업중앙회	A-04	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 중소기업현황	https://www.kbiz.or.kr/ko/contents/bbs/list.do?mnSeq=323
	A-05	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 중소기업월별동향	https://www.kbiz.or.kr/ko/contents/bbs/list.do?mnSeq=324
	A-06	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 중소기업분석통계	https://www.kbiz.or.kr/ko/contents/bbs/list.do?mnSeq=325

구분	분류 코드	메뉴명	URL
	A-07	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 기타간행물	https://www.kbiz.or.kr/ko/contents/bbs/list.do?mnSeq=1622
중소벤처기업 연구원	A-08	중소벤처기업연구원 > 발간물 > 정책연구 보고서	https://www.kosi.re.kr/front/functionDisplay?menuFrontNo=599966874772379832292720698908024275 49&menuFrontURL=front/basicResearchList
	A-09	중소벤처기업연구원 > 발간물 > KOSI 중소기업 동향	https://www.kosi.re.kr/front/functionDisplay?menuFrontNo=8541227368831382748464407005008270452 0&menuFrontURL=front/economicTrendsList
	A-10	중소벤처기업연구원 > 발간물 > 해외 중소기업 정책동향	https://www.kosi.re.kr/front/functionDisplay?menuFrontNo=1279566317136465117478376578350753167 25&menuFrontURL=/front/policyTrendList
	A-11	중소벤처기업연구원 > 발간물 > 중소기업 이슈 포커스	https://www.kosi.re.kr/front/functionDisplay?menuFrontNo=3224239192277005966330614802756471165 8&menuFrontURL=front/focusList
	A-12	중소벤처기업연구원 > 발간물 > Journal of SME Policy(등재지)	https://www.kosi.re.kr/front/functionDisplay?menuFrontNo=1145302226406496969015116148507326936 71&menuFrontURL=/front/policyList

2) 정책자금, 융자, 보증 인터넷 정보 자원

중소기업 지원 유관 기관들이 자체 발간한 연구보고서, 정책자금 정보, 국내외 연구조사보고서를 대상으로 3개 웹사이트 6개 게시물 메뉴로 정하였다. 대상 기관은 중소벤처기업진흥공단, 기술보증기금, 소상공인시장진흥공단, 신용보증재단중앙회이다.

[표 3-4] 그룹B, 정책자금, 융자, 보증 사업 웹사이트 URL

구분	그룹 코드	메뉴명	URL
중소벤처기업 진흥공단	B-01	중소벤처기업진흥공단 > 발간자료 > KOSME 연구	https://www.kosmes.or.kr/nsh/SH/IFS/SHIFS016M0.do
	B-02	중소벤처기업진흥공단 > 발간자료 > 연차보고서	https://www.kosmes.or.kr/nsh/SH/IFS/SHIFS017M0.do
	B-03	중소벤처기업진흥공단 > 발간자료 > 중진공 소식지	https://www.kosmes.or.kr/nsh/SH/IFS/SHIFS020M0.do
소상공인 시장진흥공단	B-04	소상공인시장진흥공단 > 유관공고	https://www.semas.or.kr/web/board/webBoardList.kmdc?bCd=2002&pNm=BOA0122
신용보증재단 중앙회	B-05	신용보증재단중앙회 > 알림마당 > 발간자료 > 조사연구	https://www.koreg.or.kr/koreg/na/ntt/selectNttList.do?mi=1093&bbsId=1041
	B-06	신용보증재단중앙회 > 알림마당 > 발간자료 > 해외 조사연구	https://www.koreg.or.kr/koreg/na/ntt/selectNttList.do?mi=1091&bbsId=1047

3) 시장동향, 기술표준, 기술 트렌드 인터넷 정보 자원

중소기업 지원사업, R&D 사업 및 자료, 기술동향, 우수사례 등 기술 개발 관련 자료를 대상으로 4개 웹사이트 9개 게시물 메뉴로 정하였다. 대상 기관은 중소기업기술정보진흥원, 한국산업기술진흥원, 한국산업기술진흥협회이다.

[표 3-5] 그룹C, 시장동향, 기술표준, 기술 트렌드 웹사이트 URL

구분	그룹 코드	메뉴명	URL
중소기업기술 정보진흥원	C-01	중소기업기술정보진흥원 > TIPA뉴스레터	https://www.tipa.or.kr/s040201
	C-02	중소기업기술정보진흥원 > 간행물	https://www.tipa.or.kr/publications
중소기업 기술개발사업	C-03	중소기업 기술개발사업 종합관리시스템 > 정보마당 > 자료마당 > 기술로드맵	https://www.smtech.go.kr/front/ifg/tr/tecRoadmap_list.do
	C-04	중소기업 기술개발사업 종합관리시스템 > 정보마당 > 자료마당 > 우수사례집	https://www.smtech.go.kr/front/ifg/eb/bestCaseBook.do
한국산업 기술진흥원	C-05	한국산업기술진흥원 > 사업・자료 > 사업공고	https://www.kiat.or.kr/front/board/boardContentsListPage.do?board_id=90&MenuId=b159c9dac684471b87256f1e25404f5e
	C-06	한국산업기술진흥원 > 사업・자료 > 정책간행물	https://www.kiat.or.kr/front/board/boardContentsListPage.do?board_id=71&MenuId=878cb9b6d5ec41bf914ad5c0f590ed14
	C-07	한국산업기술진흥원 > 사업・자료 > KIAT 공개자료 > 산업기술통계	https://www.kiat.or.kr/front/board/boardContentsListPage.do?board_id=105&MenuId=098bdacec586452faf9ebfbb547f62be
한국산업기술 진흥협회	C-08	한국산업기술진흥협회 > 최신기술동향	https://www.koita.or.kr/board/commBoardPublct003List.do
	C-09	한국산업기술진흥협회 > 산업기술혁신연구원 > 발간자료	https://www.koita.or.kr/board/commBoardInoPublicList.do

4) 해외시장 정보, 해외마케팅 인터넷 정보 자원

해외시장 정보, 해외마케팅 정보를 대상으로 중소기업 지원사업, R&D 사업 및 자료, 기술동향, 우수사례 등 기술 개발 관련 자료를 대상으로 대상으로 4개 웹사이트 12개 게시물 메뉴로 정하였다. 대상 기관은 한국무역투자진흥공사(KOTRA), 한국무역협회, 한국전시산업진흥회이다.

[표 3-6] 그룹D, 해외시장, 해외마케팅 웹사이트 URL

구분	그룹 코드	메뉴명	URL
KOTRA	D-01	kotra 해외시장뉴스 > 뉴스	https://dream.kotra.or.kr/kotranews/cms/com/index.do?MENU_ID=70
	D-02	kotra 해외시장뉴스 > 상품・산업	https://dream.kotra.or.kr/kotranews/cms/com/index.do?MENU_ID=170
	D-03	kotra 해외시장뉴스 > 국가・지역정보 > 진출전략	https://dream.kotra.or.kr/kotranews/cms/com/index.do?MENU_ID=250
	D-04	kotra 해외시장뉴스 > 국가・지역정보 > 출장자료	https://dream.kotra.or.kr/kotranews/cms/com/index.do?MENU_ID=260
	D-05	kotra 해외시장뉴스 > 보고서	https://dream.kotra.or.kr/kotranews/cms/com/index.do?MENU_ID=280
한국무역협회	D-06	한국무역협회 > 공지・뉴스 > 무역뉴스	https://www.kita.net/board/totalTradeNews/totalTradeNewsList.do
	D-07	한국무역협회 > 공지・뉴스 > 해외시장뉴스	https://www.kita.net/board/overseasMarketNews/overseasMarketNewsList.do
	D-08	한국무역협회 > 공지・뉴스 > 해외시장동향	https://www.kita.net/board/overseasMarket/overseasMarketList.do
	D-09	한국무역협회 > 공지・뉴스 > 보고서	https://www.kita.net/researchTrade/report/reportMain/reportMainList.do
GEP 글로벌 전시플랫폼	D-10	GEP 글로벌전시플랫폼 > 전시회정보	https://www.gep.or.kr/gept/ovrss/exbi/exbiInfo/allExbiList.do
한국전시산업 진흥원	D-11	한국전시산업진흥회 > 알림마당 > 산업 및 정책동향	https://akei.or.kr/bbs/board.php?bo_table=trend
	D-12	한국전시산업진흥회 > 알림마당 > 자료실	https://akei.or.kr/bbs/board.php?bo_table=referenceRoom

3.2.1.3 인공지능 학습 데이터세트 포맷 선정

1) 학습 데이터세트와 토큰

대형 언어모델을 사용하는 인공지능 서비스는 토큰(Token) 단위로 계산한다. 토큰은 글자나 단어의 조각을 의미한다. 범용적으로 사용하는 대형 언어모델에서는 개인 또는 사용자별로 토큰 한도를 정한다. 토큰 한도는 특정 계정 또는 사용자가 일정 기간 동안 프롬프트와 같이 글자나 단어를 보낼 수 있는 입력값과 인공지능 서비스가 돌려 줄 수 있는 출력값을 합친 값의 상한선을 의미한다. 구체적으로 질문에 해당하는 프롬프트를 길게 쓰면 입력 토큰이 증가한다. 답변이 길고 자세하면 출력 토큰이 증가한다. 토큰 소모량이 늘어나고 누적되어 개인 한도에 도달하면 다시 토큰이 확보되기 전까지 해당 서비스를 이용할 수 없다. 시간적 여유가 있다면 일정 시간 기다려야 하고, 시간적 여유가 없다면 비용을 지불하고 토큰을 확보해야 한다. 따라서 대기시간을 최소화하여 단절 없는 서비스를 구현하고, 비용 효과적인 인공지능 시스템을 운용하기 위해 토큰 소요를 효율화하려면 최적화된 학습 데이터세트 포맷을 선정해야 한다.

2) 인공지능 학습 데이터세트 포맷 선정

대형 언어모델에서는 여러 개의 파일을 업로드 하면 파일별로 텍스트를 추출하고 추출된 텍스트를 토큰으로 변환한다. 이 토큰 사용량을 파일 단위, 세션 단위로 내부에서 기록하고 세션당 최대 파일 개수, 파일당 최대 크기, 전체 토큰 한도와 같은 정책을 운영한다. 이 과정에서 여러 개의 파일을 한 번에 등록하더라도 각 파일이 각각 얼마나 토큰을 사용하였는지는 인공지능 서비스가 알려주지 않는다. 인공지능 서비스가 내부적으로는 파일 단위 사용량(즉, 토큰)으로 관리하지만 그 값을 사용자에게 보여주는 기능은 확인하지 못하였다. 따라서, 직접 테스트가 필요하다.

비즈니스 환경에서 가장 많이 사용하는 문서 포맷은 MS워드의 DOCX 파일 포맷, 한국에서는 한컴 HWPX 파일 포맷이다. 동 문서에는 문자, 표, 그림 등 다양한 콘텐츠를 포함하여 인간이 열람할 수 있는 형태로 작성되어 있다. 이렇게 전문 워드프로세스로 작성한 문서를 제3자에게 보내거나 보관할 때 PDF 파일을 생성하여 사용한다. PDF는 ISO에 규격이 공개되어 있어 다양한 뷰어를 이

용하여 이용할 수 있는 장점이 있기 때문이다. 최근에는 인공지능이 생성한 HTML 파일을 학습데이터로 이용하기도 한다. 공공부문에서 인공지능 학습데이터를 제공하는 AI HUB 웹사이트에서는 JSON 포맷으로 데이터를 제공한다. JSON 포맷으로 데이터를 다운로드 받거나 일부 데이터는 API 이용하여 연계 이용할 수 있다.

대규모 언어모델의 인공지능 서비스는 문서에서 글자나 단어들을 읽어들여 학습 하므로 전처리 작업에 토큰이 소비되는 것을 최소화 하면 토큰을 효율적으로 사용할 수 있다. 연구자는 동일한 내용을 담은 문서를 MS워드, 흔글, HTML, JSON으로 만들고, 그중 MS워드와 흔글은 PDF로 변환하여 인공지능에서 사용하는 토큰 소비량을 측정했다. 즉, 원문 포맷과 이를 PDF로 변환한 포맷을 비교하여 PDF 포맷간의 차이도 비교하기 위함이다. 측정 방법은 한성대학교에서 학생들에게 제공하는 LLM 서비스 헬피를 이용했다. 헬피는 ChatGPT, Claude, Gemini 등 LLM 서비스의 다양한 모델을 단일 UI에서 이용할 수 있는 인공지능 서비스이다. 2026년 4월 등록된 LLM 모델 중 최신 모델 3개를 선정했다.

웹사이트의 데이터를 인공지능 모델의 학습 데이터세트로 가공하기 위해 인공지능 모델이 학습 및 프롬프트 입력 시 각 파일 포맷이 가지는 토큰(Token) 소모 효율성과 기술적 특성을 비교 분석하였다. 인공지능 모델의 인공지능 학습 및 추론 과정에서 토큰 소모량은 연산 비용뿐만 아니라 콘텍스트 창(Context Window)의 활용 제한과도 직결되므로, 데이터 전처리 단계에서의 포맷 최적화는 필수적이다. 이에 따른 포맷별 분석 결과는 [표 3-7]과 같다.

[표 3-7] 데이터세트 포맷별 토큰 사용량[단위: 토큰]

구분	DOCX (MS워드)	HWPX (흔글)	PDF 1 (워드2PDF)	PDF 2 (흔글2 PDF)	HTML	JSON Lines
GPT-5.5	6	5	9	7	2	1
Claude Opus 4.7	7	5	10	8	2	1
Gemini 3.1 pro	4	3	6	4	1	1

JSON(JavaScript Object Notation)과 일반 텍스트(Plain Text)는 토큰 효율성 측면에서 가장 우수하며 AI 학습을 위한 최적의 포맷으로 나타났다. 일반 텍스트는 서식 및 메타데이터가 전혀 없는 순수 데이터 형태로 존재하므로 글자 수 대비 토큰 소모가 가장 적어 비용 효율적이다. 반면 JSON은 키-값(Key-Value) 구조를 통해 데이터의 속성과 맥락을 명확히 정의함으로써 줄글 형태인 일반 텍스트의 한계를 보완한다. JSON은 필드명의 반복으로 인해 미세한 토큰 추가 소모가 발생할 수 있으나, 데이터 구조화를 통해 AI의 환각 현상(Hallucination)을 최소화할 수 있어 RAG(검색증강생성) 및 미세조정(Fine-tuning) 데이터세트 구축에 가장 최적화된 포맷으로 평가된다.

HTML은 원본의 하이퍼링크 구조를 보존하고 마크다운(Markdown)으로의 변환이 용이하다는 장점이 있어 정제 시 높은 효율성을 보인다. 그러나 스크립트, CSS, 마크업 태그 등을 사전에 제거하지 않고 사용할 경우, 본문 외 노이즈 데이터가 대량의 토큰을 낭비하게 만드는 현상이 발생할 수 있으므로 HTML 파서를 통한 전처리가 필수이다.

오피스 문서 포맷인 Word(DOCX)와 HWP(아래아한글)는 중간 단계의 토큰 효율성을 보인다. Word는 XML 기반의 개방형 구조를 취하고 있어 python-docx 등의 라이브러리를 통해 순수 텍스트 및 표 데이터를 안정적으로 파싱할 수 있다. 하지만 여전히 학습데이터로 사용하기에는 불필요한 XML 태그와 스타일이 존재하여 정제를 하여야 사용이 가능하다. 국내 공공 표준인 HWP의 경우 개방형 포맷인 HWPX로 전환하더라도 오픈소스 파서 생태계의 한계로 인해 텍스트 추출 시 띄어쓰기 오류나 줄바꿈 깨짐 현상이 빈번히 발생한다. 이러한 오류는 인공지능 모델이 단어를 비정상적으로 쪼개어 인식하는 '토큰 분절(Token Splitting)' 현상을 유도하여 의도치 않은 비용 상승과 모델의 이해도 저하를 초래하므로 정밀한 맞춤법 및 인코딩 정제가 요구된다.

[표 3-8] AI 학습 데이터세트 포맷 비교

포맷 구분	토큰 효율성	장점	단점
JSON JSON Lines	매우 높음	· 데이터 구조화(Key-Value) · AI의 환각(Hallucination) 최소화 · 파인튜닝 (Fine-tuning) 규격에 최적	· Key 값(필드명)이 반복되어 미세한 토큰 추가 소모 · 정교한 DB 설계 및 파싱 필요
일반 텍스트	매우 높음	· 텍스트 순도 100% (노이즈 낮음) · 글자 수 대비 비용 저렴 · AI 가독성 높음	· 데이터 간의 관계/맥락 표현 불가 · 표(Table)나 이미지 레이아웃 소실
HTML	높음	· 원본의 하이퍼링크 구조 보존 · 마크다운 (Markdown) 변환이 매우 용이	· 정제 안 할 경우 태그, 스크립트, CSS가 토큰을 대량으로 낭비함
DOCX (MS워드)	보통	· XML 기반이라 텍스트 추출이 매우 안정적.· 표(Table) 구조를 AI가 읽기 쉽게 파싱 가능.	· 문서 스타일, 글꼴 등 불필요한 메타데이터 포함 · 대용량 일괄 처리 시 파서 리소스 소모
HWP (한컴)	보통	· 국내 공공·정부 표준 문서 서식 보존 · 정부 간행물의 원본 맥락 유지	· 오픈소스 파서 부족으로 텍스트 깨짐 발생 · 띄어쓰기 오류로 인한 토큰 분절(Splitting) 및 비용 증가 위험
PDF	매우 낮음	· 인간이 시각적으로 확인하기 가장 좋음	· 텍스트 레이어 꼬임 (다단 구조 오독) · 스캔본의 OCR 오탈자로 토큰 폭탄 유발 · 멀티모달 입력 시 비용이 수십 배 폭증

PDF는 AI 학습 관점에서 가장 비효율적인 포맷으로 분석되었다. PDF는 인간의 시각적 가독성 및 보존성에는 최적화되어 있으나, 구조적인 한계로 문맥 오독을 유발하며 특히 이미지 형태의 스캔본 PDF는 OCR 변환 과정이 필수이며 OCR 변환 과정에서 발생하는 오인식 글자로 인해 토큰 소모량이 증가하며, 최근의 멀티모달(Vision) 입력을 활용하더라도 이미지 타일당 토큰이 추가 과금되므로 텍스트 추출 방식 대비 수십 배의 토큰 비용을 발생시킨다는 한계가 있다.

따라서 이 연구에서는 가장 토큰 효율성이 높고, 데이터 구조화에 용이하며 파인튜닝에 최적화된 것으로 평가받는 JSON 및 JSON Lines 포맷을 학습 데이터로 채택하였다.

3.2.2 (2단계) 웹사이트 수집 시스템 개발

3.2.2.1 웹사이트 수집 시스템 설계

이 연구에서 제안하는 웹 아카이빙 시스템의 핵심 기능 중 하나는 대상 웹사이트의 동적 등록 및 수집 주기 스케줄링 설정이다. 실제 운영 환경이나 다수 부서가 협업하는 조직 구조에서는 업무 범위와 보안 권한이 상이한 다년도·다계층 사용자를 지원하기 위해, 사용자 및 그룹별로 대상 웹사이트를 격리·제어하는 권한 관리 아키텍처가 요구된다. 그러나 이 연구는 웹 자원의 실시간 수집 메커니즘과 AI 학습 데이터세트 생성의 기술적 타당성을 검증하는 데 목적이 있으므로, 시스템 복잡도를 완화하고 핵심 프로토타입의 가용성을 극대화하기 위해 독립된 단일 사용자(Single User) 구조로 시스템 범위를 제한하여 설계 및 구현하였다.

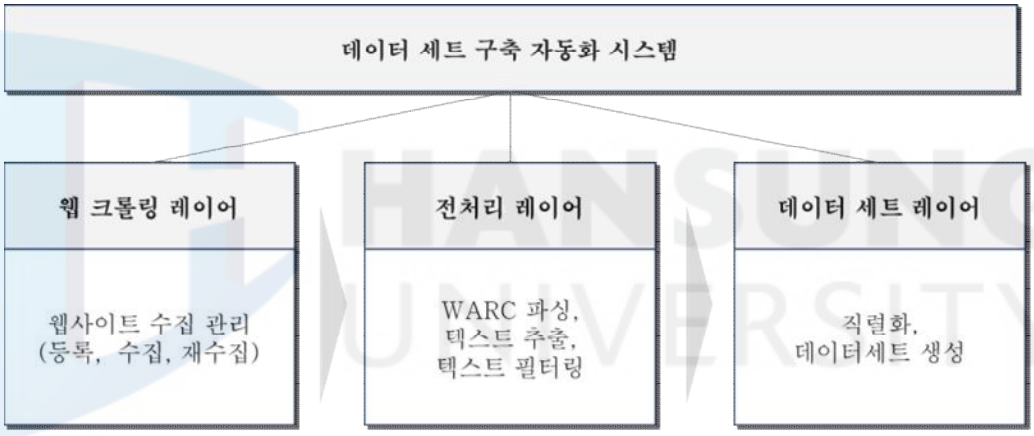
이 연구에서 제안하는 전체적인 데이터세트 구축 자동화 시스템의 논리적 아키텍처는 그림3-1에 제시된 바와 같다. 설계된 시스템은 대규모 비정형 웹 자원으로부터 순도 높은 정형 데이터를 추출하기 위해 명확한 역할 분담을 지닌 3대 서브시스템(Sub-system)으로 명명한 웹 크롤링 레이어(Web Crawling Layer), 전처리 레이어(Pre-process Layer), 데이터세트 레이어(Dataset Layer)로 구성되며, 데이터의 흐름(Data Flow)에 따라 순차적 연쇄 구동 체계로 구성한다.

웹 크롤링 레이어는 파이프라인의 최전단에 위치하여 원천 데이터의 수집 및 인프라 관리를 전담하는 모듈이다. 이 레이어에서는 수집 대상 플랫폼의 메타데이터를 등록하는 '등록' 기능, 웹 아카이빙을 수행하는 '수집' 기능이 가동된다. 특히, 웹 환경의 네트워크 지연 및 서버 오류 등으로 발생하는 데이터 누락에 대응하고 웹 크롤러가 인지하지 못하고 수집에 실패하는 동적 URL을 확보하기 위해 '재수집' 메커니즘을 내장함으로써 일차 수집 단계에서의 데이터 완전성을 확보하는 중추적 역할을 수행한다.

전처리 레이어는 웹 크롤링 레이어로부터 산출된 아카이브 포맷인 WARC(WACZ) 파일을 입력값으로 삼아 데이터의 순도를 높이는 변환 및 정제 영역이다. 본 레이어는 대용량 파일 스트림을 라인 단위로 읽어 들이는 'WARC 파싱' 공정을 시발점으로 하여, 마크업 구조 내부에서 태그를 분리하는 '텍스트

추출' 작업을 거친다. 이후 노이즈 제거 및 최소 문자열 길이 기준 검증에 기반한 '텍스트 필터링' 알고리즘을 연속적으로 적용하여 연속적인 공백, 무의미한 태그와 스타일시트, 자바스크립트 등을 소거한다.

데이터세트 레이어는 정제 공정이 완결된 데이터를 최종적인 기계학습 모델 학습 규격으로 정형화하는 종단 레이어이다. 이 레이어에서는 추출된 문자열과 메타데이터(URL, 수집 일시 등)를 사전에 정의한 스키마 키-밸류(Key-Value) 구조에 맞추어 변환하는 '직렬화(Serialization)'를 수행하며, 최종적으로 하나의 라인에 하나의 독립된 JSON 객체가 매핑되는 JSON Lines(.jsonl) 포맷의 '데이터세트 생성'을 완료하여 디스크에 저장한다.



[그림 3-1] 데이터세트 구축 자동화 시스템 - 3 레이어

웹 크롤링 레이어는 운영의 효율성과 시스템 관리의 편의성을 극대화하기 위하여 4개의 기능으로 세분화하여 구현한다. 각 기능은 MariaDB를 데이터 저장소로 사용하며 오픈소스 기반 웹사이트 수집 엔진인 브라우저트릭스(BrowserTrix)와 연동되며, 유저 인터페이스(UI) 레이어와 백엔드 제어 레이어가 분리된 컴포넌트 기반 아키텍처를 취하고 있다.

접근 관리 기능은 수집 파이프라인의 자원 보호와 작업 주체별 권한 격리(Role-Based Access Control)를 수행하는 모듈로 사용자 등록, 사용자 상세조회, 로그인, 로그아웃의 기능을 담당한다. 사용자 등록 및 상세조회 기능은 시스

템을 운영할 연구자 및 관리자의 계정을 생성하는 기능으로, 패스워드는 일방향 암호화 해시 함수를 거쳐 데이터베이스에 저장된다. 상세조회 기능을 통해 사용자의 소속, 작업 권한 등급, 최근 접속 이력 등의 메타데이터를 정밀하게 관리한다. 로그인 및 로그아웃 기능은 인증 및 인가(Authentication & Authorization) 프로세스를 처리한다. 로그인이 성공하면 상태 세션을 유지하며 사용자 영역의 접근을 허용하고 로그아웃 시 세션 토큰을 무효화하여 비인가자의 불법적인 크롤러 가동 및 데이터 접근을 원천 차단한다. 수집 대상 관리 기능은 실제 수집이 될 웹사이트의 메타정보(시드 URL, 기관명, 분류값)를 데이터베이스에 저장 관리하며 등록된 수집 대상 사이트의 현재 연결 상태와 크롤링 상태를 모니터링한다. 또한 URL이나 기관명을 수정할 수 있도록 구현한다. 수집 관리는 브라우저트릭스 엔진과 연계되어 크롤링 프로세스의 실행, 중단 등 라이프 사이클(Lifecycle)을 실시간으로 제어하는 역할을 하며 자바스크립트로 구현된 사용자 인터렉티브 요소에서 동적 URL을 탐지하는 기능을 구현한다. 수집 결과 관리는 수집이 완료된 웹사이트의 결과를 상태 값과 뷰어를 통해 확인할 수 있도록 한다.

[표 3-9] 웹 크롤링 레이어 주요 기능

구분	내용
접근 관리	사용자 등록, 사용자 상세조회, 로그인, 로그아웃
수집 대상 관리	수집 대상 등록, 수집 대상 상세조회, 수집 대상 수정
수집 관리	수집 실행, 수집 중단, 동적 URL 탐지
수집 결과 관리	결과 조회

전처리 레이어는 3가지 주요 기능으로 구성된다. WARC 파싱 기능은 대용량 WARC 파일을 일시에 메모리에 적재할 때 발생하는 시스템 오버헤드를 방지하기 위해, 파일 I/O를 레코드(Record) 단위로 순차 탐색하는 스트리밍 구조로 구현한다. 텍스트 추출 기능은 읽어들인 레코드에서 HTML 인코딩 엔티티를 표준 문자로 변환하고 시각 디자인 요소를 소거하여 순수 텍스트 영역만을 추출하는 역할을 한다. 텍스트 필터링 기능은 추출된 순수 텍스트에서 개행문자, 연속 공백을 치환하는 화이트스페이스 (Whitespace) 정규화 처리를 하고 사전에 정의한

토큰 임계치 미만의 저밀도 레코드를 데이터세트에서 배제하는 과정을 거쳐 순도 높은 텍스트로 정제한다.

[표 3-10] 전처리 레이어 주요 기능

구분	내용
WARC 파싱	스트리밍 I/O, WARC 파서
텍스트 추출	HTML/XML 마크업 태그 소거, 텍스트 추출
텍스트 필터링	화이트스페이스 정규화, 임계치 적용

데이터세트 레이어는 전처리 파이프라인의 최종 종단(Terminal Stage)에 위치하여, 노이즈가 제거된 정제 텍스트와 가용 메타데이터를 기계 학습 모델이 즉각적으로 파싱 및 학습할 수 있는 표준 명세 구조로 정형화하고 스토리지에 저장하는 역할을 담당한다. 본 레이어는 데이터의 아키텍처적 유연성과 하드웨어 I/O 효율성을 극대화하기 위해 직렬화, 데이터세트 생성의 연결된 연산 공정으로 구체화되며, 각 모듈의 기술적 기능 명세는 다음과 같다.

직렬화 모듈은 메모리 상에 객체(Object) 형태로 존재하는 데이터를 데이터 스키마 규칙에 맞추어 키-밸류(Key-Value) 구조로 매핑한다. 이 과정에서 유일성을 보장하는 문자열 형태의 data_id, 출처 추적성을 지닌 source_url, ISO 8601 표준 규격의 collected_at 타임스탬프, 그리고 정제된 본문인 cleaned_content 가 하나의 JSON 객체 레코드로 결합된다.

데이터세트 생성 모듈은 직렬화가 완료된 개별 JSON 레코드들을 각 객체 간의 경계를 개행 문자(\n)로 구분하는 JSON Lines (.jsonl) 파일 구조로 저장한다.

[표 3-11] 데이터 세트 레이어 주요 기능

구분	내용
직렬화	스키마 매핑, 인코딩 적용, JSON 객체 생성
데이터세트 생성	JSON Lines 포맷팅, 데이터 세트 생성

3.2.2.2 웹사이트 수집 시스템 및 데이터세트 생성 프로그램 개발

이 연구에서는 Java와 Python 프로그래밍 언어를 사용한다. Java 언어는 웹사이트 수집 대상을 등록, 수정, 관리하는 웹 기반의 UI를 개발하는 데 사용한다. Python 언어는 오픈소스 기반 수집 엔진인 브라우저트릭스의 실행을 제어하고 전처리 레이어, 데이터세트 레이어를 구현하는 영역에서 사용한다.

[표 3-12] 웹사이트 수집 시스템 개발 언어

구분	언어	내용
개발 언어1	JAVA	웹 UI
개발 언어2	Python	수집 엔진 제어, 전처리 레이어, 데이터세트 생성 레이어

이들 언어는 각각 자바 가상 머신(Java Virtual Machine, JVM)과 파이썬 가상 머신(Python Virtual Machine, PVM)을 통해 바이트코드(Bytecode) 형태로 실행되므로, 특정 운영체제(OS)나 하드웨어 아키텍처에 종속되지 않는 플랫폼 독립성(Platform Independence)을 보장한다. 따라서 본 시스템은 도입 조직의 인프라 환경(Windows, Linux, macOS 등)에 구애받지 않고 유연하게 이식 및 운영이 가능하다. 이 연구에서는 다음과 같은 환경에서 동작 시킨다.

[표 3-13] 웹사이트 수집 시스템 동작 환경

구분	내용
운영체계	Rocky Linux 9.5 (Blue Onyx)
JDK	OpenJDK 21
Python	Python 3.9.21
웹크롤러	Browsertrix 1.22.10
구동서버	Proxmox 가상화 기반 VM / 1CPU 4Core, MEM 4GB

3.2.3 (3단계) 웹사이트 정보 수집 실증

3.2.3.1 수집 대상 웹사이트 정보 수집

수집 대상 웹사이트 정보 수집 단계는 앞서 정의한 18개의 중소기업 지원 플랫폼을 대상으로, 이 연구에서 개발한 '데이터세트 구축 자동화 시스템'을 가동하여 핵심 정책, 기술 트렌드 공고 등의 게시물 텍스트 데이터와 이에 종속된 첨부파일 자원을 획득하고 스토리지에 저장하는 절차이다. 본 공정의 목적은 단순히 웹 문서의 스크래핑(Scraping)을 넘어, 원천 소스의 형태를 국제 표준 규격으로 보존함으로써 다운스트림(Downstream) 전처리 및 소형언어모델(SLM) 파인튜닝 단계에 순도 높은 학습 데이터세트를 공급하는 데 있다.

3.2.3.2 웹사이트 정보 수집 확인 및 재수집

웹사이트 정보 수집 확인 및 재수집 절차는 웹 환경의 물리적 네트워크 지연 및 서버 오류로 인한 정적 데이터 누락에 대응하고, 일반적인 크롤러가 인지하지 못하는 자바스크립트 기반의 은닉된 동적 URL을 능동적으로 탐지하여 복원하기 위하여 '동적 URL 탐지 및 자동 재수집 알고리즘'을 적용한다. 이 프로세스는 수집 과정에서 발생하는 타임아웃 및 HTTP 프로토콜 에러 자원을 수집 오류 로그에 기록하는 동시에, javascript: 또는 onclick= 등 인라인 이벤트 핸들러 내에 숨겨진 동적 목적지 인자값을 가상 돔(DOM) 분석을 통해 정형 URL 체계로 복원하여 수집 시드를 재생성한다. 이후 일정 시간 대기 후 재수집 루프를 실행하며, 최종적으로 수집된 페이로드는 이미 수집 생성된 WARC 파일에 병합하여 저장한다. 이러한 재수집 절차는 데이터 수집의 사각지대를 해소하고 데이터셋의 완전성을 확보함으로써, 향후 인공지능 모델 파인튜닝 시 발생할 수 있는 데이터 편향성을 완화하는 타당성을 제공한다.

3.2.4 (4단계) 데이터세트 생성 및 인공지능 활용

3.2.4.1 데이터 전처리 및 인공지능 학습 데이터세트 생성

이 연구에서 구현한 데이터세트 구축 자동화 시스템은 원본 웹 자원을 파악하고 수집, 저장하는 단계와, 이를 인공지능 학습 자원으로 가공하는 정제 및 구조화 단계의 절차로 구성된다. 먼저, 웹 크롤러 아키텍처를 통해 수집된 웹사이트 정보는 네트워크 계층의 트랜잭션 데이터와 원본 콘텐츠의 무결성을 동시에 보존하기 위해 국제 표준 규격인 WARC(Web ARChive, ISO 28500) 파일 포맷으로 저장된다. WARC 파일은 단순한 웹페이지의 텍스트 스냅샷을 넘어, 수집 당시의 HTTP 요청/응답 헤더(HTTP Request/Response Header) 정보와 메타데이터 정보를 독립된 블록 구조와 해시값을 포함해 저장하므로 원본성(Originality)을 유지하며 데이터를 구성하는 원시 데이터 저장소 역할을 수행한다.

그러나 WARC 포맷은 인간 중심의 가독성 및 재현을 전제로 한 원형 보존에 최적화되어 있어, 거대언어모델(LLM), 소형언어모델(SLM)을 비롯한 인공지능 기반 알고리즘이 직접 학습하거나 콘텍스트를 이해하기에는 비효율적이다. 이를 극복하기 위해 동 시스템 내에 WARC-to-JSON 데이터 변환을 위한 전처리 레이어와 데이터세트 레이어를 설계하였다. 이러한 가공 공정을 통해 최종 산출된 JSON Lines 데이터세트는 인공지능 모델의 토큰(Token) 소모 효율성을 극대화하고 문맥 이해도를 연산 비용 대비 최상으로 끌어올리는 인공지능 최적화 원천 데이터로 기능하게 된다.

3.2.4.2 AI 학습 데이터세트 품질 검토

구축된 데이터세트의 무결성과 학술적 신뢰성을 담보하기 위하여, 파이프라인의 최종 종단에서 데이터 수집 수율과 본문 정제 상태를 다각적으로 검증하는 'AI 학습 데이터세트 품질 검토(Quality Assurance)' 절차를 수행한다. 이 검토 공정은 수집 완료된 대상 URL의 통계를 산출하고, 전처리 과정을 거쳐 생성 완료된 최종 JSON Lines(.jsonl) 객체의 '성공 건수'와 404 에러 페이지 및 최소

문자열 길이 임계치 미달로 배제된 '필터링 건수'를 대조 분석하여 전체 처리 공정의 수율(Yield)을 진단한다. 나아가, 생성된 텍스트 데이터세트의 품질을 검증하기 위해 무작위 표본 추출(Random Sampling)을 통한 사후 정성 검토를 병행하여, 필터링 검증과 인코딩 에러가 존재하지 않는지 판별한다. 이와 같이 수율 진단과 정성적 검증을 결합한 다각적 품질 검토는 데이터세트의 완전성을 높여여준다. 이는 결과적으로 인공지능 모델 학습에 최적화된 데이터세트를 생성한다는 기술적 시도의 근거가 된다.

3.2.4.3 인공지능 학습 활용

이 연구를 통해 산출된 데이터세트는 인공지능 모델의 지식 내재화 및 하위 공정 최적화를 위한 기계 학습(Machine Learning) 프로세스에 직접적으로 적용·활용된다. 순도 높은 텍스트 데이터세트와 표준화된 메타데이터가 결합된 JSON Lines(.jsonl) 형태의 데이터세트는 스트리밍 I/O 인터페이스를 통해 인공지능 모델의 학습 파이프라인에 병목 현상 없이 공급된다. 또한 인공지능 모델이 중소기업 지원 정책, 융자 및 R&D 지원 등 공공 도메인의 목적성에 맞는 맥락을 인지할 수 있도록 미세조정 학습 데이터로 기능하며, 토크나이저(Tokenizer) 최적화 과정을 거쳐 인공지능 모델의 가중치(Weights) 연산에 반영된다. 특히 전처리 레이어와 재수집 알고리즘을 거치며 마크업 노이즈 및 동적 URL 누락에 따른 왜곡 요인이 사전에 차단되었기 때문에, 인공지능 모델의 환각(Hallucination) 현상을 최소화하고 중소기업에 특화된 질의응답 및 답변 생성 과정의 성능을 향상하는 자원으로 활용될 수 있다.

3.2.5 전문가 그룹 리뷰

이 연구 결과를 웹사이트 수집 및 아카이빙 분야 전문가, 인공지능 학습 분야 전문가 각 3인을 대상으로 인터뷰를 실시하고 기능의 고도화 및 활용확산에 대한 의견을 수렴한다,

Ⅳ. 데이터세트 구축 자동화 시스템

4.1 웹사이트 수집 및 전처리 기술개발

4.1.1 웹사이트 수집 시스템 개발

4.1.1.1 웹사이트 수집 기능 설계 및 개발

1) 시스템 아키텍처

이 절에서는 데이터세트 구축 자동화 시스템의 핵심 요소인 '웹사이트 수집 시스템(Web Collection Subsystem)'의 구현 체계를 기술한다. 웹사이트 수집 시스템은 크롤링 대상 웹사이트의 등록, 수집 실행 그리고 수집된 비정형 웹 자원을 국제 표준 WARC(Web ARChive, ISO 28500) 기반의 파일 패키징 구조화를 수행하는 주체가 된다.

전체적인 개발 환경은 호스트 운영체제(OS)의 이기종 커널 종속성을 차단하고 시스템 이식성(Portability)을 극대화하기 위해 Java 및 Python 언어를 혼용하여 개발되었으며, 웹사이트 수집 엔진은 브라우저트릭스(Browsertrix)를 이용하였고 수집 엔진의 제어 및 수집 대상 웹사이트의 구조를 분석하여 동적 URL을 해석하고 시드를 추가하는 부분은 Python 언어로 개발하였다. 개발한 코드는 자바 가상 머신(JVM)과 파이썬 가상 머신(PVM)의 가상 런타임 환경 위에서 동작하도록 설계되었다. 또한 수집 대상을 저장하고 관리하기 위한 웹 기반의 관리 도구를 개발하였고 데이터의 관리를 위해 MariaDB DBMS를 사용하였다.

2) 데이터베이스 스키마 설계

대상 웹사이트의 유연한 관리와 수집 스케줄링을 제어하기 위해, 관계형 데이터베이스(RDBMS)를 사용해 수집 대상 도메인을 관리하는 스키마를 설계하였다. 수집 엔진은 구동 시 데이터베이스의 TARGET_SITES 테이블을 참조하여 활성화된 시드 URL(Seed URL)과 수집 깊이(Crawl Depth), 그리고 수집 주기 정책을 메모리에 적재한다.

또한 조직 내 부서별 업무 도메인 분리와 다중 사용자 제어에 따른 오버헤드를 최소화하고, 웹 수집 및 데이터세트 자동화 시스템 구축이라는 본연의 과제에 자원을 집중하기 위해, 동 시스템은 단일 관리자 권한 구조(Single Administrator Structure)로 설계하였다. 수집 대상 관리를 위한 핵심 엔티티 구조는 다음과 같다.

[표 4-1] 수집 대상 관리를 위한 핵심 엔티티 구조

구분	내용
SITE_ID (PK)	수집 대상을 식별하기 위한 고유의 인덱스 번호 (UUID)
SEED_URL	크롤러가 최초로 접근하여 하이퍼링크를 탐색할 정규화된 URL 주소
CRAWL_INTERVAL	크론 표현식(Cron Expression)을 추상화한 수집 주기 스케줄링 파라미터 (예: 매주, 매월, 실시간 배치 등)
DEPTH_LIMIT	최초 시드 페이지로부터 파고들 하위 하이퍼링크의 최대 탐색 깊이 임계값
STATUS	현재 해당 사이트의 수집 활성화 여부 및 크롤러 매핑 상태 필드

3) 스케줄러 기반 멀티스레드 크롤링 메커니즘

설정된 수집 주기에 따라 자동으로 크롤링 프로세스를 실행하기 위해 가상 머신 내부의 타이머 스케줄러 알고리즘을 도입하였다. 시스템의 백엔드에서 동작하는 스케줄링 엔진은 백그라운드 데몬(Daemon) 형태로 상주하며, 각 수집 대상 웹사이트의 타임스탬프를 실시간으로 모니터링한다. 지정된 수집 주기에 도달하면, 엔진은 시스템의 병목 현상을 방지하고 자원 효율성을 제고하기 위해 웹사이트별로 독립된 워커 스레드(Worker Thread)를 할당하는 멀티스레딩(Multi-threading) 기법을 수행한다.

각 워커 스레드에 할당된 크롤러는 지정된 DEPTH_LIMIT 범위 내에서 웹 페이지 내부의 DOM(Document Object Model) 트리 구조를 선형 탐색하며, 정적 HTML 소스뿐만 아니라 하이퍼링크 앵커 태그(<a href="...">)를 동적으로 추출하여 수집 큐(Queue)를 갱신한다. 이 과정에서 도메인 이탈을 방지하기 위한 화이트리스트(Whitelist) 필터링이 수행되며, 웹 서버의 부하 경감을 위한 알고리즘을 적용하여 요청 간격을 유연하게 통제하였다.

4.1.1.2 웹사이트 수집 데이터 저장 기술개발

웹 크롤링을 통해 수집되는 데이터는 정부의 정책 공고(정형/반정형), 기술 트렌드 분석 보고서(비정형 텍스트 및 PDF), 외부 커뮤니티의 VOC(비정형 텍스트) 등 매우 다양한 스키마(Schema)와 구조를 지닌다. 따라서 이 연구에서는 이처럼 상이한 형태를 가진 대규모 데이터의 정합성을 유지하고, 향후 인공지능 모델의 학습 데이터세트로 원활히 활용하기 위하여 크롤링 엔진이 웹 서버로부터 응답받은 원시 바이너리와 텍스트 자원은 즉시 디스크에 저장되지 않고, 데이터의 무결성을 위해 WARC 파일 포맷 빌더(Builder)를 통과하게 된다. 수집 시스템은 단일 트랜잭션마다 발생하는 네트워크 레이어의 정보, 즉 HTTP 요청 헤더(HTTP Request Header), HTTP 응답 헤더(HTTP Response Header), 그리고 실제 웹페이지의 응답 본문(Response Body) 페이로드를 레코드 단위로 분리하여 WARC 파일에 블록으로 저장한다.

이렇게 물리적으로 생성된 WARC 파일은 압축(GZIP) 상태로 디스크 저장공간에 저장된다. 이 단계까지가 웹 자원을 물리적으로 확보하는 단계이며, 이후 구축된 WARC 파일의 각 페이로드 블록을 선형 검색하여 불필요한 마크업 노이즈를 제거하고 머신러닝에 친화적인 JSON 데이터세트로 변환하는 후속 전처리 레이어로 전달한다.

4.1.2 전처리 및 학습 데이터세트 생성 기술개발

4.1.2.1 AI 학습 데이터세트 구조 정의

앞선 과정에서 웹사이트로부터 수집된 대규모 원천 데이터를 인공지능 모델의 학습 데이터세트로 생성하기 위하여 표준 웹 아카이빙 포맷인 WARC(Web ARChive)를 1차 결과물로 생성하였다. WARC 파일은 웹 페이지의 HTML, HTTP 헤더, 메타데이터를 블록의 형태로 보존하는 특징이 있으나, 그림 4-1과 같이 다량의 노이즈와 학습 데이터로서 무의미한 마크업 태그를 포함하고 있어 인공지능 모델의 입력 데이터로 직접 활용하기에는 한계가 존재한다.



[그림 4-1] WARC 파일 헤더 예시

따라서 수집된 WARC 파일로부터 순수 텍스트 및 메타데이터를 추출(Extraction), 정제(Cleaning)하고, 이를 인공지능 모델이 즉각적으로 학습할 수 있는 표준 JSON(JavaScript Object Notation) 포맷의 데이터세트로 구조화하였다. 데이터세트 구조 정의의 핵심 목적은 수집한 데이터의 원천 속성을 유지하면서도, 자연어 처리(NLP) 및 다양한 인공지능 내재화 알고리즘(분류, 감성 분석, 개체명 인식 등)에 유연하게 매핑될 수 있는 표준 피처 스키마(Feature Schema)를 정의하는 데 있다. 그림 4-2는 원천 웹 데이터가 필터링 공정을 거쳐 인공지능의 학습에 직접 투입될 수 있는 표준 JSON 형태로 구조화된 구체적인 결과물을 보여준다.

```
{
  "data_id": "warc_rec_20260606_00123",
  "source_url": "https://www.bizinfo.go.kr/web/lay1/bbs/S1T122C128/AS/74/view.do",
  "collected_at": "2026-06-06T01:20:00Z",
  "cleaned_content": "2026년도 중소기업 AI 내재화 지원 사업 공고. 본 사업은 제조 혁신을 위해 중
  "target_features": {
    "policy_support": {
      "support_category": "R&D_and_Tech",
      "eligibility_criteria": ["창업 3년 이상", "매출액 10억 원 이상", "중소 제조기업"]
    },
    "market_tech": {
      "keyword_tags": ["AI 내재화", "제조 혁신", "인공지능 솔루션"],
      "industry_code": "C"
    }
  }
}
```


[그림 4-2] JSON 구조 예시

4.1.2.2 전처리 기술개발

수집한 결과물인 WARC(Web ARChive) 파일의 원천 구조는 웹 표준 아카이빙 규격(ISO 28500)에 의거하여 헤더(Header) 영역과 페이로드(Payload) 영역의 쌍으로 구성되어 있다. 구체적으로는 WARC 블록 내에서 실제 웹 자원의 크기를 나타내는 바이너리 길이는 Content-Length 필드로 정의되며, 해당 구문이 종료되는 시점 이후부터 명시된 바이트(Byte) 크기만큼 페이로드 데이터가 공간을 차지하게 된다. 이때 HTTP 응답 페이로드 내에 포함된 자원들은 Content-Type 헤더의 명세에 따라 text/html, text/css, application/javascript, image/* 등 웹사이트를 구성하는 다양한 MIME 타입(Multipurpose Internet Mail Extensions)으로 분류된다. 그러나 이처럼 다양한 데이터 형식 중 텍스트 마이닝 기반의 인공지능 학습에 실질적으로 유의미한 정보 자원은 text/html 포맷으로 한정된다.

데이터 전처리 레이어의 첫 번째 단계에서는, WARC 파일 내에 존재하는 모든 레코드를 순차적으로 탐색(Iterate)하면서 Content-Type 헤더가 text/html인 유효한 레코드만을 선별적으로 추출하여 후속 공정으로 전달하는 데이터 필터링을 수행하였다. 두 번째 단계는 추출된 페이로드 내에서 인공지능 학습의 노이즈로 작용하는 비정형 요소를 단계적으로 제거하는 과정이다.

마지막으로 텍스트 추출 과정에서 발생하는 불필요한 비인쇄 제어 문자, 연속

된 탭, 중복 줄바꿈 등의 화이트 스페이스(Whitespace)를 단일 공백으로 치환하는 문자열 정규화를 수행함으로써 자연어 처리(NLP) 모델이 즉각적으로 임베딩할 수 있는 순수 문자열만을 남기게 된다.

4.1.2.3 AI 학습 데이터세트 생성 기술개발

정제 공정을 통과한 순수 텍스트 데이터는 인공지능 모델의 유효한 학습 단위를 보장하기 위해 최종 적재 전 정량적 품질 검증을 거치게 된다. 추출된 문자열의 최종 길이가 사전에 정의한 최소 길이 임계치를 초과하는 경우에 한해서만 실질적인 인공지능 학습 데이터세트의 인스턴스로 채택되며, 기준 미달의 유실 데이터나 정보 밀도가 낮은 페이지는 파이프라인 상에서 원천 배제된다. 품질 검증을 통과하여 최종 채택된 데이터는 타 기계학습 모델 및 데이터 프레임워크와의 상호 운용성을 극대화하기 위하여 사전에 정의된 표준 JSON 객체 구조로 포맷팅된다. 개별 JSON 객체는 데이터의 유일성과 결함 추적을 보장하기 위해 고유한 해시값을 가지는 data_id, 해당 데이터의 수집 출처이자 원천 웹 자원의 주소인 source_url, 원천 데이터가 아카이빙 된 시점을 시계열로 관리할 수 있는 collected_at, 그리고 최종 노이즈 정제 공정이 완결된 텍스트인 cleaned_content 필드로 구조화된다. 최종적으로 이와 같이 구조화된 개별 JSON 객체들은 파일 내에서 한 행(Line)에 하나의 완결된 JSON 객체를 저장하는 JSON Lines(.jsonl) 포맷으로 변환되어 디스크에 저장된다. 이러한 객체 독립적 저장 전략은 대용량 파일 처리 시 파일 전체를 메모리에 로드하지 않고 라인 단위의 스트리밍(Streaming) 처리를 가능하게 하여 처리 환경 내에서의 I/O 병목을 최소화하고 인공지능 모델의 데이터 공급 효율성을 최적화하는 기틀이 된다.

4.2 웹사이트 정보 수집 및 AI 학습 데이터세트 생성

이 연구에서 제시한 데이터 구축 자동화 시스템을 가동하여 중소기업 지원 유관 웹사이트를 대상으로 실증 실험을 수행한 결과, 최종 선정된 총 39개 웹사이트로부터 일차적으로 117,693건의 원천 URL 자원을 수집하였다. 웹 크롤링 레이어를 통해 생성된 원천 WARC 파일은 총 110개 세그먼트

(Segment)로 분할되어 저장되었다. 이후 전처리 레이어에서 HTML 태그 제거, 스크립트 제거 및 최소 문자열 임계치 필터링 알고리즘을 적용한 결과, 전체 원천 데이터의 약 70.69%에 해당하는 83,197건이 인공지능 학습 데이터세트(JSON Lines)로 성공적으로 생성되었다. 반면, 앞서 분석한 404 에러 페이지 및 MIME 타입 불일치 등의 노이즈 데이터로 판명되어 전처리 과정에서 필터링된 무효 데이터는 총 34,496건(약 29.31%)으로 집계되었다.



[그림 4-3] 전처리 및 직렬화 결과

이 연구에서 전처리 알고리즘을 통해 필터링된 데이터를 사후 분석한 결과, HTTP 응답 헤더와 실제 데이터 페이로드(Payload) 간의 불일치 현상이 크게 세 가지 유형으로 관측되었다. 가장 높은 빈도를 보인 사례는 HTTP 응답 헤더의 Content-Type은 정상적인 text/html로 응답되었으나, 실제 콘텐츠는 유효한 정보를 포함하지 않는 '404 Not Found', 로그인 페이지 등의 제어 URL인 경우였다. 이는 일차적인 헤더 검증만으로는 무효 데이터를 완전히 배제할 수 없음을 시사한다. 또 다른 유형은 웹 크롤러가 요청한 웹 자원(URI)의 실제 포맷은 이미지,바이너리 등의 다른 포맷임에도 불구하고, 대상 웹 서버의 MIME 타입 설정 오류(Misconfiguration)로 인해 응답 헤더에는 text/html로 오응답(Misclassification)된 비정상 데이터가 식별되었다. 마지막 유형은 페이로드를 통해 얻어진 텍스트의 문자열 길이가 임계치로 설정한 100을 넘지 않는 경우이다. 결과적으로 이 연구에서 도입한 문자열 길이 기반의 임계치 필터링 및 태그 소

거 알고리즘은 이러한 헤더 수준의 왜곡 데이터와 노이즈를 유기적으로 차단함으로써, 구축되는 인공지능 학습 데이터세트의 순도와 정밀성을 확보하는 데 기여했음을 보여준다.

4.3 AI 학습 데이터세트 품질 검토

[표4-2]는 시스템을 실행하여 웹사이트로부터 데이터를 수집하고 전처리를 거쳐 최종 생성된 데이터세트의 결과를 나타낸다. C-04, D-06, D-07, D-08, D-09는 수집 진행 중 대상 기관의 시스템에서 웹크롤러의 IP를 차단하여 수집 진행이 되지 않아 결과 값에서 제외하였다.

[표 4-2] 데이터세트 생성 결과

그룹 코드	메뉴명	수집URL (건)	성공 (건)	필터링 (건)	수율(%)	비고
A-01	중소벤처기업부 > 사업공고	465	464	1	99.78%	
A-02	중소벤처24 > 사업공고	890	889	1	99.89%	
A-03	중소벤처24 > 정책금융	169	169	0	100.00%	
A-04	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 중소기업현황	5	5	0	100.00%	
A-05	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 중소기업월별동향	129	129	0	100.00%	
A-06	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 중소기업분석통계	72	67	5	93.06%	
A-07	중소기업중앙회 > 정보마당 > 조사연구통계 > 간행물 > 기타간행물	9	9	0	100.00%	
A-08	중소벤처기업연구원 > 발간물 > 정책연구 보고서	512	500	12	97.66%	

그룹 코드	메뉴명	수집URL (건)	성공 (건)	필터링 (건)	수율(%)	비고
A-09	중소벤처기업연구원 > 발간물 > KOSI 중소기업 동향	210	184	26	87.62%	
A-10	중소벤처기업연구원 > 발간물 > 해외 중소기업 정책동향	61	61	0	100.00%	
A-11	중소벤처기업연구원 > 발간물 > 중소기업 이슈 포커스	390	390	0	100.00%	
A-12	중소벤처기업연구원 > 발간물 > Journal of SME Policy(등재지)	23	23	0	100.00%	
B-01	중소벤처기업진흥공단 > 발간자료 > KOSME 연구	219	198	21	90.41%	
B-02	중소벤처기업진흥공단 > 발간자료 > 연차보고서	21	17	4	80.95%	
B-03	중소벤처기업진흥공단 > 발간자료 > 중진공 소식지	27	16	11	59.26%	
B-04	소상공인시장진흥공단 > 유관공고	62	62	0	100.00%	
B-05	신용보증재단중앙회 > 알림마당 > 발간자료 > 조사연구	24	24	0	100.00%	
B-06	신용보증재단중앙회 > 알림마당 > 발간자료 > 해외 조사연구	13	13	0	100.00%	
C-01	중소기업기술정보진흥원 > TIPA뉴스레터	221	220	1	99.55%	
C-02	중소기업기술정보진흥원 > 간행물	5	4	1	80.00%	
C-03	중소기업 기술개발사업 종합관리시스템 > 정보마당 > 자료마당 > 기술로드맵	62	60	2	96.77%	
C-04	중소기업 기술개발사업 종합관리시스템 > 정보마당 > 자료마당 > 우수사례집	-	-	-	-	IP 차단됨
C-05	한국산업기술진흥원 > 사업·자료 > 사업공고	1,285	1,283	2	99.84%	
C-06	한국산업기술진흥원 > 사업·자료 > 정책간행물	1,204	1,202	2	99.83%	
C-07	한국산업기술진흥원 > 사업·자료 > KIAT 공개자료 > 산업기술통계	15	15	0	100.00%	

그룹 코드	메뉴명	수집URL (건)	성공 (건)	필터링 (건)	수율(%)	비고
C-08	한국산업기술진흥협회 > 최신기술동향	3,554	3,553	1	99.97%	
C-09	한국산업기술진흥협회 > 산업기술혁신연구원 > 발간자료	28	28	0	100.00%	
D-01	kotra 해외시장뉴스 > 뉴스	60,490	42,594	17,896	70.41%	
D-02	kotra 해외시장뉴스 > 상품・산업	44,059	28,188	15,871	63.98%	
D-03	kotra 해외시장뉴스 > 국가・지역정보 > 진출전략	1,141	681	460	59.68%	
D-04	kotra 해외시장뉴스 > 국가・지역정보 > 출장자료	126	87	39	69.05%	
D-05	kotra 해외시장뉴스 > 보고서	608	468	140	76.97%	
D-06	한국무역협회 > 공지・뉴스 > 무역뉴스	-	-	-	-	수집 중 IP 접속 차단
D-07	한국무역협회 > 공지・뉴스 > 해외시장뉴스	-	-	-	-	
D-08	한국무역협회 > 공지・뉴스 > 해외시장동향	-	-	-	-	
D-09	한국무역협회 > 공지・뉴스 > 보고서	-	-	-	-	
D-10	GEP 글로벌전시플랫폼 > 전시회정보	1,238	1,238	0	100.00%	
D-11	한국전시산업진흥회 > 알림마당 > 산업 및 정책동향	337	337	0	100.00%	
D-12	한국전시산업진흥회 > 알림마당 > 자료실	19	19	0	100.00%	
합계		117,693	83,197	34,496	70.69%	

이 절에서는 구축된 중소기업 도메인 특화 데이터세트의 무결성과 신뢰성을 확보하기 위해, 동 시스템의 최종 종단 파이프라인에 구현된 정량적·정성적 품질 검토 모듈의 구동 방식과 실제 테스트 환경에서의 검증 결과를 기술한다. 본 검토 공정은 수집 수율을 진단하는 정량적 분석과 텍스트의 정제 상태를 검사하는 정성적 검증으로 이원화하여 수행되었다. 정량적 수율 진단 및 필터링 메커니즘 검증시스템은 일차 수집 및 재수집 공정을 거쳐 저장된 원천 WARC 파일과 최종 직렬화된 데이터세트 간의 변환 수율(Yield)을 로그분석을 통해 추적한다. 로그파일에 기록된 정보를 통해 각 웹사이트별로 수집된 총 URL 수, 필터링된 유실 건수, 데이터세트 생성 건수를 교차 대조하여 데이터세트의 완전성을 정량적으로 평가하였다. 또한 최소 문자열 길이 임계치를 파이프라인 내부 검증 로직에 적용하였다. 시스템은 HTML 태그가 소거된 순수 텍스트의 길이가 100자 미만인 레코드를 제어 페이지(404 Not Found, 로그인 등) 중심의 무효한 자원으로 분류하여 최종 데이터세트에서 배제하도록 설계되었다. 정책자료나 보고서 등 책자가 첨부된 게시물의 경우 본문 내용이 100자 미만으로 작성된 경우가 많아 임계치를 넘지 않으므로 해당 데이터도 데이터세트에서 배제되었다.

시스템의 가동을 통해 39개 웹사이트로부터 추출된 데이터세트의 그룹별 및 종합 품질 검토 결과는 [표 4-3]와 같다.

[표 4-3] 품질 검토 결과

그룹 코드	타겟	수집 URL(건)	성공 (건)	필터링 (건)	수율 (%)
A	정책 자료, 지원 사업 인터넷 정보 자원	2,935	2,890	45	98.47%
B	정책자금, 융자, 보증	366	330	36	90.16%
C	시장동향, 기술표준, 기술 트렌드	6,374	6,365	9	99.86%
D	해외시장 정보, 해외마케팅 정보	108,018	73,612	34,406	68.15%
합계	전체 도메인	117,693	83,197	34,496	70.69%

4.4 전문가 그룹 리뷰

웹사이트의 웹페이지, 웹게시물의 정보를 수집하여 인공지능 학습을 위한 데이터세트를 자동적으로 구축하는 이 연구 결과를 정보시스템 활용 관련 오랜 경험을 가진 전문가를 개별 인터뷰하고 그 결과를 정리하였다.

4.4.1 인터뷰 개요

연구자는 전문가에게 인공지능 전환 배경 등을 설명하고, 중소기업이 현재 할 수 있는 업무로 도메인 업무에 필요한 양질의 데이터세트 구축의 필요성을 설명하였다. 향후 중소기업 정보화 지원 컨설턴트가 수진 기업으로서 중소기업의 인공지능 전환 컨설팅을 진행할 경우를 가정하고 의견을 받아 정리하는 방식으로 진행하였다. 답변의 범위를 축소하기 위해 정보시스템, 데이터 포맷, 활용 확산을 메인 주제로 진행하였다.

- 질의1. 데이터세트 자동 구축 시스템 수집 결과에 관한 검토 의견
- 질의2. 웹페이지 아카이빙 포맷과 데이터세트 포맷에 관한 의견
- 질의3. 동 시스템 활용 확산을 위해 필요한 조언

4.4.2 인터뷰 참여자

이 연구의 인터뷰 참여자는 정보시스템 구축, 온라인 서비스 분야에서 오랫동안 활동한 3명의 전문가에게 이 시스템의 용도와 수집 결과, 데이터세트 포맷 등에 관하여 설명하고 의견을 정리하였다.

[표 4-4] 인터뷰 참여자 정보

구분	경력	관련 산업	참여기관(향후 비실명화)
전문가1	PM (경력 35년)	공공/글로벌 분야	컨설팅 전문가 상무
전문가2	PM (경력 26년)	공공/민간 분야	공공 / 중견 SI 상무
전문가3	PM (경력 27년)	국방정보화 분야	공공 / 중견 SI 부사장

4.4.3 분석 결과 검토 의견

1) 질의1. 데이터세트 구축 자동화 시스템

데이터세트 구축 자동화 시스템의 배경과 활용에 대한 질의에서 전문가들은 데이터 수집 목적이 명확해야 필요한 정보를 수집하여 활용할 수 있어야 한다고 판단했다. 중소기업의 정보화 수준에 대한 우려도 포함되었다.

[표 4-5] (인터뷰 요약1) 데이터세트 구축 자동화 시스템 의견

구분	전문가 의견
전문가1	○ 중소기업은 내수 중심의 기업과 수출 중심의 기업으로 구분할 수 있는데. 수출 중심의 기업의 경우 내수 중심 기업보다 훨신 많은 정보를 수집, 분석 하기 위해 인적자원과 시간이 소요된다. 동 시스템은 국내 정보 수집 및 활용 뿐 아니라 해외 국가의 시장 정보, 품목 시세 변동 정보, 정책 변화 등 데이터를 활용하면 활용 가치가 높다. ○ 예를 들어, 런던 금속 거래소(LME) 웹사이트에서는 비철금속의 가격 변동 정보가 공개되는데. 우리나라의 많은 비철금속 생산, 유통 채널 종사자들이 해당 정보를 수집하여 국내 가격을 정한다. LME는 오늘자 데이터는 무료로 제공하고 과거 데이터를 이용하려면 요금제 회원으로 가입해야 한다. 이런 경우 의미가 있다. 특정 수요를 발굴하여 적용하면 데이터세트의 내재화가 가능할 수 있다.
전문가2	○ 신용정보회사나 권리조사 회사에서는 매일 아침마다 시스템이 자동으로 인터넷등기소에 접속하여 대상 부동산 물건을 조회하고 소유권이나 소유권 이외의 권리변동을 파악하는 시스템을 본적이 있다. 또한 부동산등기부등본 파일을 저장하고 갑구와 을구의 데이터를 파싱하여 데이터로 활용하는 사례도 보았다. ○ 즉, 이 시스템은 웹페이지의 정보와 파일의 다운로드에 그치고 있지만, 비즈니스 로직 상 동적인 웹페이지의 정보도 수집해야 하고, 수집한 첨부파일도 학습 데이터로만 사용하는 것이 어니라 필요한 데이터를 직접 꺼내서 활용할 수 있는 기능도 필요하다. ○ 단, 웹사이트의 데이터가 무한하게 유지되지 않고, 보유하지 않고 링크만 가지고 있으면 시간이 지날수록 이용 가능성이 낮아 진다는 점에 분명히 동의하므로, 데이터세트 구축을 자동으로 처리하는 이 시스템은 의미가 있다.
전문가3	○ 웹사이트 수집, 웹페이지 수집, 웹 크롤링, 로봇 자동화(RPA) 등 개념적으로 이 연구 결과와 유사한 기술의 제품이 시중에서 사용되고 있으나, 이 연구는 인공지능 학습을 위한 데이터세트 자동 구축이라는 점에서 타 시스템과 구분된다. ○ 그러나, 유사 기능을 가진 솔루션 벤더사에서 추가적으로 데이터세트 생성 또는 변환 기능을 개발하면 동 시스템과 같은 결과를 낼 수 있으므로 차별화 요소가 필요하다. 설치형 시스템의 개발 대신 SaaS 방식을 이용하는 등 중소기업 정보화 환경을 고려한 쉬운 접근이 필요하다.

2) 질의2. 웹페이지 아카이빙 포맷과 데이터세트 포맷에 관한 의견

정보시스템 환경이 열악한 중소기업 경영 환경에서 JSON 데이터세트 단위 저장 방식은 장단점이 있다. 글로벌·국내 AI 도구 활용 시 직접 정보시스템을 구축하는 대신 저렴하게 이용할 수 있으므로 동 시스템에 대한 효익이 우려된다. JSON은 AI 허브 등에서 이미 널리 쓰여 중소기업 데이터 내재화 포맷으로 적합하나, WARC 포맷의 JSON 변환 구조는 증분 데이터가 여러 파일로 분산돼 학습 효율 저하 가능성이 있다는 의견을 받았다.

[표 4-6] (인터뷰 요약2) 웹 아카이빙 포맷과 데이터세트 포맷 의견

구분	전문가 의견
전문가1	○ 동 시스템을 제안한 배경과 기능을 충분히 이해하지만 우려되는 부문이 있다. ○ 다차원적으로 데이트를 관리하기 위해서는 데이터 웨어하우스, 데이터 레이크 등에 데이터를 저장할 수 있으나, 중소기업 현장은 정보시스템이 부재한 경우가 많아서 데이터 저장소도 운영하지 않을 가능성이 높다. 이 배경하에 데이터베이스, SQL이 아닌 데이터세트 파일(JSON) 단위로 데이터를 저장하는 이 방식은 긍정적이기도 하고 부정적이기도 하다 ○ 글로벌 빅테크 기업의 인공지능 도구를 이용하거나, 국내 기업의 도구를 이용하는 경우에도 중소기업의 제품 또는 서비스와 관련된 데이터를 보유하고 있으면 활용할 수 있는 구조는 이해 하였으나, 구글과 같이 막대한 인터넷 데이터베이스를 축적한 기업이 이미 학습한 데이터일 가능성을 배제할 수 없다. 이 경우 별도 시스템의 구축 및 운영 비용이 경쟁력이 있을지 우려가 된다.
전문가2	○ 현재 시스템 구조는 매번 수집 시 마다 웹 아카이빙 WARC 포맷을 생성하고, 이후 WARC 포맷의 정보를 데이터세트 JSON으로 변경하는 것으로 보인다. 이 경우, 증분 데이터가 계속해서 별도의 파일로 존재하게 되는데 실제 AI 학습 단계에서 여러 개의 데이터세트를 등록하는 것은 패킷의 사용이나 학습 데이터의 활용 면에서 효율 저하가 우려된다. ○ 향후 동 시스템의 상용화를 고려한다면, 웹 페이지 또는 특정 데이터세트의 목적에 따라 하나의 데이터세트에 새로운 데이터가 계속 축적되는 구조가 필요하다. (연구자는 이 의견에서 패킷 최적화 측면에서 기 학습된 데이터세트의 불필요성을 어필하고, 증분 데이터의 학습이 더 효과적임을 설명하였다)
전문가3	○ 인터넷 웹사이트, 웹페이지 아카이빙 영역은 사실 생소하다. 일반적으로 사용하지 않는 포맷으로 어떤 뷰어나 어떤 전-후처리를 해야 동 포맷을 사용할 수 있을지 인지하지 못하고 있다. 다만, WARC 포맷을 변환하여 목표 데이터세트 포맷 JSON으로 사용한다면, 중간 과정인 WARC 포맷은 중소기업 입장에서 특별히 인지할 필요는 없다. ○ 국가에서 운영하는 AI 허브 웹사이트에서도 JSON 포맷으로 데이터를 제공한다. 이미 데이터세트 규격으로 많이 활용하고 있으므로, 인공지능 시대를 대비하는 중소기업 데이터 내재화 포맷으로 적합하다.

3) 질의3. 기타 동 시스템의 활용에 관한 의견

중소기업의 현실은 부족한 게 많다. 디지털 & 인공지능 전환을 위한 역량과 예산 그리고 인력이 부족해 자체 시스템 구축이 어렵다. 양질의 데이터 필요성에 대한 인식도 낮다. 이 연구 시스템을 중소기업이 이용하게 하려면 무료 가입 후 데이터 양이나 수집 대상 웹페이지의 범위 등에 따라 과금하는 구독형 모델이 바람직하다. 시스템 확산을 위해서는 중소기업 재직자 역량 강화와 활용 사례 제시가 중요하며, AI 챗봇 시스템 등과 연동하여 질의응답 또는 민원 게시판 데이터를 데이터세트로 만들어 활용하면 새로운 부가가치가 발생할 수 있다는 의견 등을 제시했다.

[표 4-7] (인터뷰 요약3) 시스템 활용 확산을 위해 필요한 조언

구분	전문가 의견
전문가1	○ 중소기업 현장 컨설턴트는 실제 중소기업 방문 횟수가 많지 않다. 그리고 정보화를 지원하기 위한 컨설팅을 제공하는 경우는 흔하지 않다. 일반적으로 월 일정 비용을 지불하고 이용하는 구독형 웹서비스를 이용하고, 자체 시스템을 구축할 만한 여력이 없는 경우가 많다. ○ 동 시스템을 상용화하는 경우, Free/Standard/Premium 등 등급으로 구분하여 일단 중소기업이 무료로 가입하여 이용할 수 있게 하고 이후 수집 대상 웹페이지나 데이터 용량이 일정 기준을 초과할 때 요금 등급을 상양하도록 하는 접근 방식을 추천한다.
전문가2	○ 중소기업은 자체 예산과 자체 인력으로 정보화 사업을 진행하기 어렵다. 인공지능 기술의 이용도 중요하지만 현재 인공지능 역량이 현저히 부족해서 양질의 데이터세트가 구축되어도 활용하기 어렵다. ○ 디지털 전환, 인공지능 전환 시대의 핵심은 양질의 데이터 보유이다. 그러나, 중소기업은 정보시스템, 데이터 저장 시스템, 데이터 처리 인력도 부재하다. ○ 데이터세트 구축 자동화 시스템의 필요성을 인지하기 위해서는 우선 중소기업 재직자의 역량을 강화시키고 필요한 데이터를 지속적으로 확보하는 것이 중요한 일인지에 대해 인지시켜야 동 시스템의 활용 확산이 가능할 것이다.
전문가3	○ 이 연구는 중소기업의 인공지능 내재화를 위해 데이터세트를 자동으로 구축하는 시스템을 개발하였으나, 내재화 이후 무엇을 할 수 있을지에 대한 활용 사례 제시가 필요하다. 예를 들면 데이터 수집 대상이 외부가 아니라 자체적으로 운영하는 문의 게시판, 민원 게시판의 경우 AI 챗봇에 자동으로 학습할 수 있다. 이 경우 관리자가 새로운 사례를 추가하지 않아도 민원 처리가 가능하게 된다.

4.5 소결

이 연구는 인공지능 기술의 활용을 위하여 중소기업이 필요로 하는 외부 데이터를 내재화할 수 있는 시스템을 구축하는 과정을 실증하고 유효성을 검증하는 것이다. 이 과정에서 도출된 전산학적 성과와 데이터 엔지니어링 측면의 결론을 요약하면 다음과 같다.

중소기업 정책정보, 금융, 기술 정보, 해외시장 정보 등을 아우르는 19개 기관을 선정하고 각 기관의 업무 유형에 따라 정책자료 및 지원 사업 정보, 정책자금 및 융자 정보, 시장동향 및 기술 정보, 해외시장 및 해외 마케팅 정보의 네 그룹으로 분할하고 총 39개의 URL을 대상 도메인으로 최종 선별하였다. 선별된 URL은 웹 크롤러의 Seed 파일에 게시물, 첨부파일 등 목표로 하는 수집 자원으로 설정하였다.

동 시스템은 수집된 원천 데이터의 무결성과 데이터 원본의 보존을 위해 WARC 빌더를 통해 HTTP 요청·응답 헤더 및 페이로드를 레코드 단위로 저장한다. 수집 과정에는 비동기 AJAX 및 자바스크립트 이벤트 핸들러 뒤에 은닉된 동적 URL을 탐지하고 목적 URL을 추출하기 위해 플레이라이트 기반의 동적 URL 탐지 프로세스를 개발하여 정적 크롤러 대비 수집 성능을 유의미한 수준으로 향상시켰다. 이후 실증 실험을 통해 39개 웹 URL로부터 총 117,693건의 원천 데이터를 수집하여 국제 표준 WARC(ISO 28500) 파일 포맷으로 저장하였다. 저장된 WARC 원천 데이터를 선형 검색하는 전처리 레이어를 가동해 불필요한 마크업 노이즈를 제거하고, 인공지능 모델 학습에 즉각 활용 가능한 기계 판독형 JSON 데이터세트로 정형화 및 가공하는 체계를 마련하였다. 그 결과, 전체 데이터의 70.69%에 달하는 총 83,197건의 데이터세트를 스트리밍 I/O에 최적화된 JSON Lines(.jsonl) 포맷으로 생성하였다.

동 시스템과 동작 실험에 대해 외부 전문가 3인을 대상으로 진행한 인터뷰 및 질의에 대한 내용은 다음과 같이 정리한다.

‘[질의 1] 데이터세트 구축 자동화 시스템의 타당성 및 차별성’에 대해서 전문가들은 URL 유실에 따른 데이터 휘발성을 방지하는 자동 수집 시스템의 필요성에 전적으로 공감하였다. 특히 글로벌 시장 시세 변동, 해외 정책 변화 등 전용

인프라가 필요한 특정 수요를 발굴하여 적용할 때 중소기업의 데이터 내재화 가치가 극대화될 것으로 진단하였다. 또한 시중의 일반 크롤러나 로봇 자동화(RPA) 도구와 차별화하기 위해서는 단순 웹페이지 수집을 넘어 첨부된 문서의 내부 데이터를 직접 파싱·추출하는 기능 고도화가 필수적이며, 정보화 수준이 낮은 중소기업 환경을 고려하여 설치형 시스템 대신 SaaS(서비스형 소프트웨어) 방식으로의 전환을 통해 접근성을 높여야 한다고 제언했다.

‘[질의 2] 웹 아카이빙(WARC) 및 데이터세트 포맷(JSON)의 적합성’에 대해서 전문가들은 자체 데이터 레이크나 인프라를 운영하기 어려운 중소기업의 경영 현실을 고려할 때, 파일 단위의 JSON 저장 방식은 매우 실용적이라고 동의했다. JSON 포맷은 국가 AI 허브 등에서도 표준 규격으로 널리 쓰이고 있어 내재화 포맷으로 가장 적합하며, 난해한 중간 표준인 WARC 포맷을 인지할 필요 없이 백엔드에서 변환되도록 설계된 구조 역시 긍정적으로 평가받았다.

‘[질의 3] 시스템의 비즈니스 모델 및 활용 확산 방안’에 대해서 전문가들은 자체 예산 부족으로 시스템 구축 여력이 없는 중소기업의 현실을 반영하여, 초기에는 무료로 서비스를 제공하되 수집 대상 웹페이지 범위나 데이터 용량에 따라 과금 단계를 상향 적용하는 등급별 구독형 요금제 도입이 가장 바람직하다고 제안하였다. 또한 시스템의 확산을 위해서는 양질의 데이터 확보가 왜 중요한지 인지할 수 있도록 중소기업 재직자를 대상으로 한 역량 교육이 선행되어야 하며 나아가 데이터 내재화 이후의 명확한 활용 사례를 제시해야 하는 등 즉각적인 부가가치를 창출할 수 있는 비즈니스 시나리오 제시가 동반되어야 함을 강조했다.

Ⅴ. 결론 및 제언

5.1 연구 결과 요약 및 시사점

5.1.1 연구 결과 요약

인공지능 도구를 잘 이용하는 기업과 그렇지 않은 기업의 격차가 점차 커지는 가운데, 연구자는 우리 중소기업이 시급히 내재화해야 할 대상으로 '신뢰성 높은 양질의 인공지능 데이터세트' 확보로 판단하고 공공분야 중소기업 지원 웹사이트의 정보(웹페이지, 게시물, 첨부파일 등)를 자동으로 수집하여 인공지능 학습 데이터세트를 생성하는 시스템 개발 연구를 진행하였다. 인공지능 전환 시대에 인공지능 역량의 내재화는 경쟁환경에 있는 모든 주체에게 중요하다. 특히, 기술 변화에 대응력이 낮은 한국 중소기업의 생존을 위해 매우 중요하고 시급하다.

이 연구의 결과는 다음과 같이 네 가지로 요약한다. 첫 번째, '인공지능 데이터세트 구축 자동화 시스템'은 중소기업이 겪는 가장 큰 페인포인트인 '학습용 데이터 부족'과 '데이터 가공 전문 인력 부재' 문제를 동시에 해결할 수 있는 내재화 핵심 시스템이 될 수 있다는 것이다. 이 시스템의 완성도를 높여 중소기업이 도입하면 매번 필요한 정보가 있을 때마다 인터넷 정보를 검색하는 데 시간을 소모하는 문제, 인터넷에서 사라진 정보를 이용하지 못하는 문제, 최신 데이터나 최신 소식을 확인하지 못해서 발생하는 문제 등을 해결할 수 있다.

두 번째, 인공지능 학습 데이터세트 확보의 목표는 경쟁사 동향 파악, 시장 정보 모니터링, 공공 입찰 정보 획득, 분야별 정부 R&D 지원 사업 등 목적을 명확하게 하고 대상 웹사이트의 정보를 수집하여 내재화할 때 활용 가치를 높일 수 있다. 동 시스템의 활용을 극대화하기 위해서는 목적에 따른 대상 웹사이트와 정보를 명확하게 해야 한다.

세 번째, 제품과 서비스 설명서나 유의 사항 등을 학습하고 인터넷에서 이 제품과 관련된 고객의 소리 등을 연결하면 고객 서비스 품질을 높일 수 있다. 전문가의 의견에서 확인한 바와 같이 대고객 서비스 측면에서 활용할 수 있다.

AICC, AI 챗봇 시스템을 이용하여 고객 상담, 응대 서비스를 제공하는 중소기업이라면 새로운 데이터의 학습으로 스스로 똑똑해지는 인공지능 상담원이 반가울 수밖에 없다.

네 번째, 컨설팅 현장에서 권장되는 인공지능 학습 데이터세트 포맷은 JSON이다. 현재 가장 많이 사용하는 학습 데이터 포맷은 PDF 파일 이다. PDF 파일은 MS워드(DOC), MS파워포인트(PPT), 한컴오피스(HWP) 포맷의 문서를 변환하여 사용할 수 있고, 원본과 동일한 화면을 재현하므로 인간이 식별하기 좋은 포맷이다. 그러나 PDF 파일에서 데이터만을 추출하는 과정에서 연산 비용을 낭비하게 된다. 순수한 디지털 정보에 가까운 방식으로 데이터세트를 보유하면 토큰 사용을 줄여 인공지능 연산 비용을 효율화할 수 있게 된다.

5.1.2 시사점

이 연구는 웹사이트에서 개별적으로 관리되어 온 웹 자원을 웹 크롤링 기술을 사용하여 하나로 통합하고, 이를 인공지능이 즉각 학습할 수 있는 정형 데이터세트로 생성했다는 점에서 학술적·기술적으로 중요한 시사점을 지닌다. 학술적으로는 자바스크립트 기반 동적 URL 누락 문제를 가상 DOM 및 헤드리스 브라우저로 해결하는 방법론적 대안을 제시하였다. 기술적 관점에서는 텍스트 정제 가이드라인과 최소 문자열 길이 임계치 검증 레이어를 실증 시스템에 구현함으로써, 인공지능 모델의 토크나이저 통과 시 발생하는 토큰 부풀림(Token Bloat) 현상과 노이즈를 원천적으로 제한하는 JSON Lines(.jsonl) 직렬화 공정을 증명하였다. 이는 가용한 자본과 인프라가 부족한 중소기업이 거대언어모델(LLM)의 API 의존성에서 벗어나, 고유의 도메인 지식을 반영한 독자적인 소형언어모델(SLM)로도 AI 내재화(AX) 생태계를 저비용·고효율로 구축할 수 있다는 가능성을 시사한다. 궁극적으로 이 연구의 메커니즘은 도메인 특화 데이터세트 구축의 표준적 아키텍처로 기능하여, 데이터 구축 및 정제 전반의 비용 절감과 AI 모델의 환각 현상 완화에 크게 기여할 수 있을 것이다.

5.2 정책 제언 및 연구의 한계

5.2.1 정책 제언

이 연구에서 개발한 '중소기업 인공지능 내재화를 위한 데이터세트 구축 자동화 시스템'은 분산된 중소기업 지원 기관들의 웹 자원을 AI 친화적인 정형 데이터세트로 구축하는 기술적 타당성을 입증하였다. 본 연구 결과를 바탕으로 향후 대한민국 중소기업의 인공지능 전환(AX) 및 AI 내재화 생태계를 국가 차원에서 견인하기 위한 정책적 제언을 다음과 같이 제시한다.

현재 중소벤처기업부, 중소벤처기업진흥공단, 중소기업기술정보진흥원 등 수많은 유관 공공기관의 지원 사업 정보와 기술 및 마케팅 정보는 인간 수요자를 위한 웹페이지 및 문서파일(PDF, HWP) 형태로 생산·소비되고 있다. 그러나 이러한 비정형 구조는 인공지능 모델이 즉각적으로 지식을 학습하기에 심각한 노이즈와 토큰 부풀림(Token Bloat) 문제를 야기한다. 따라서 정부는 향후 공공 데이터를 개방할 때 단순한 API나 정적 문서 개방을 넘어, 본 연구에서 제시한 JSON Lines(.jsonl) 형태의 '기계 학습 전용 직렬화 표준 포맷'을 제도적으로 지원해야 한다. 원천 데이터 생산 단계부터 AI 친화적 규격을 표준화한다면 데이터 정제에 소요되는 사회적 연산 비용을 획기적으로 절감할 수 있을 것이다.

현재의 정부 AI 바우처 및 스마트서비스 지원 정책은 거대언어모델(LLM) API 사용료 지원이나 단발성 상용 솔루션 구매 보조에 치우쳐 있어, 영세한 중소기업이 자체적인 기술 자생력을 확보하기 어렵다. 기술적·자본적 격차에 따른 디지털 불균형을 근본적으로 해소하기 위해서는, 본 연구에서 구축한 텍스트 기반 데이터세트와 같이 특정 산업 도메인에 특화된 소형언어모델(SLM)을 중소기업이 직접 미세조정(Fine-Tuning)하고 자사 인프라에 내재화할 수 있도록 돕는 실질적인 기술 이전 및 고도화 지원으로 정책 패러다임을 전환해야 한다. 이는 기업의 내부 데이터 유출 리스크를 원천 차단하면서도 독립적인 AX 역량을 강화하는 지속 가능한 대안이 될 수 있다.

5.2.2 연구의 한계

이 연구는 분산된 중소기업 지원 기관들의 웹사이트를 대상으로 수집 및 전처리, 데이터세트 생성에 이르는 시스템을 구축하여 10만 건 이상의 정형 데이터세트를 확보하였으나, 몇 가지 기술적·방법론적 한계를 지닌다.

첫 번째, 가상 DOM 분석 알고리즘을 통해 자바스크립트 기반의 미인지 동적 URL을 능동적으로 탐지하였음에도 불구하고, 무한 스크롤, 다중 레이어 등 고도화된 웹 UI의 구현 메커니즘을 우회하는 데는 한계가 있어 일부 플랫폼의 데이터 유실을 방지하지 못했다.

두 번째, 데이터세트 품질 검토에서 본문 텍스트의 정제 상태와 화이트스페이스 정규화는 성공적으로 적용하였으나, 원천 데이터 내에 존재하는 표(Table) 구조나 이미지 내의 텍스트(OCR 대상 자원) 등 시각적·구조적 비정형 요소를 손실 없이 수집하고 데이터화하는 데 부족함을 보였다.

세 번째, 정형화된 JSON Lines(.jsonl) 데이터셋의 효용성을 인공지능 언어모델(LLM)의 토크나이저 압축 효율 관점에서만 정량적으로 검증하였을 뿐, 실제 미세조정 학습을 거친 AI 모델의 추론 성능과 환각 감소율을 실증적인 지표로 측정하지 못했다는 한계가 있다.

마지막으로 수집 과정에서 일부 대상 기관 시스템의 보안 정책으로 인해 웹크롤러의 IP가 차단되는 현상이 발생하여 해당 도메인(C-04, D-06, D-07, D-08, D-09)의 자원을 최종 결과값에서 제외해야 하는 기술적 한계를 나타냈다. 이와 같이 공공기관 웹사이트가 취하고 있는 봇 차단(Anti-Bot) 정책이나 방화벽 체계를 우회하여 모든 수집 대상의 수집 완전성을 확보하는 데는 한계가 존재하였으며, 향후 대규모 사이트 수집의 연속성을 보장하기 위해서는 IP 프록시 순환, 부하 분산 크롤링 아키텍처, 차단 알고리즘 우회 전략 등에 대한 추가 연구가 요구된다.

참 고 문 헌

1. 국내 문헌

김희정. (2010). 웹 아카이빙 인터페이스 유형 및 특성 분석. 『한국기록관리학회』. 10(2), 한국기록관리학회지, 2010.12, 147-170.

김명목, 이상용. (2010). 전자기록물의 장기보존을 위한 기능요소 연구. 『한국기록관리학회』. 10(2), 한국기록관리학회지, 2010.12, 101-126.

박병주, 차승준, 이규철. (2010). 웹기록물 보존을 위한 전자기록물 장기보존 포맷 확장 설계. 『한국디지털산업학회』, 15(4), 한국전자거래학회지, 2010.11, 33-47.

손승진. (2025). 제조 기업의 AI 내재화를 위한 인력 수요 및 기술 역량: 국내 가전 대기업 사례를 중심으로. 『한국지식경영학회』, 26(3), 지식경영연구, 2025.9, 65-90.

오상훈, 최영선. (2009). ISO 14721 OAIS 참조모형을 활용한 웹 아카이빙의 메타데이터 구조 및 요소 정의. 『한국정보처리학회』, 16(5), 정보처리학회논문지, 2009.8, 651-660.

윤여찬, 김수균. (2025). 생성형 AI를 위한 Retrieval-Augmented Generation (RAG) 기술 동향 및 전망. 『한국컴퓨터교육학회』, 28(2), 컴퓨터교육학회 논문지, 2025.2, 69-80.

이연수, 남성운, 윤대현. (2013). 대규모 웹기록물의 원격수집을 위한 콘텐츠 중복 필터링 개선 연구. 『한국기록학회』, 35, 기록학연구, 2013.1, 133-160.

이창선, 홍정범. (2023). DUI AI 솔루션을 활용한 산업 AI 내재화. 『한국생산제조학회』. 2023 한국생산제조학회 학술발표대회 논문집, 2023.7, 103-103.

이향수, 이성훈. (2016). 경영전략 내재화가 공공기관의 발전에 미치는 영향.

『한국디지털정책학회』, 14(5), 디지털융복합연구, 2016.5, 1-10.
장진호, 권혁상, 이규모, 최동준. (2022). 원격수집 방식의 웹기록물 관리를 위한 웹수집기 성능 비교 연구. 『한국기록학회』, 74, 기록학연구, 2022.10, 155-186.
장진호, 권혁상, 이규모, 최동준. (2022). 공공부문 웹아카이빙을 위한 원격 수집 친화적인 웹사이트 연구. 『한국차세대컴퓨팅학회』, 18(6), 차세대컴퓨팅학회 논문지, 2022.12, 21-31.
조민수, 양진홍. (2024). 정보접근 제어 기능을 가진 On-Premise기반 RAG 시스템 구축 방안 연구. 『한국정보전자통신기술학회』, 17(6), 한국정보전자통신기술학회 논문지, 2024.12, 494-502.
차승준, 이규철. (2008). 웹기록물 아카이빙 기반기술 연구개발. 『한국디지털산업학회』. 2008 지식정보산업연합회 창립기념 학술대회, 2008.12, 359-368.
관계부처 합동. (2023). 『산업 AI 내재화 전략, 제1차 산업 디지털 전환 종합계획』. 보고서
윤정현. (2025). 『한국형 소버린 AI 국가전략의 모색』. 국가안보전략연구원, 349, INSS전략보고, 2025.11, 1-24.
이씨플라자. (2022). 『웹기록물 관리 및 활용 모델 연구』. 행정안전부 국가기록원. 2022.12.
유니닥스. (2009). 『웹기록물 아카이빙 기반기술 연구 및 적용 시험(유니닥스/정준선)』. 국가기록원 연구결과보고서, 2009.11.
국정브리핑. (2025). 이 대통령 "AI 고속도로 구축해 도약과 성장의 미래 열어야". 웹사이트.
박성완. (2026). 1년 걸리던 시물질 연구, AI로 하루만에... LG, 신소재 개발 플랫폼 특허, 노컷뉴스.
알체라. AI 기술 내재화는 한 번의 도입 아닌 '지속 가능한 역량. 웹사이트

연합뉴스. (2026). 한국산업기술진흥원, ‘AI 내재화 시대’ 10대 유망산업 발표.
행정안전부. (2017). 「제18대 대통령기록물 1,106만건 대통령기록관으로 이관」. 보도자료, 2017.5.11.
행정안전부. (2022). 「제19대 대통령기록물 1,116만건 대통령기록관으로 이관」. 보도자료, 2022.5.12.
SK.COM. RAG 시스템을 위한 문서 전처리 가이드: AI가 이해하기 쉬운 형태로 만들기. 웹사이트.
LG CNS. AI OS의 출현, IT 기업의 경쟁력은 AI 내재화에 있다. 웹사이트
행정안전부 공동데이터 포털. 웹사이트.
https://www.data.go.kr/data/15070395/fileData.do
한국지능정보사회진흥원(NIA), 국가데이터통합플랫폼, ONE 윈도우.
https://www.data1window.kr/noti/businessInfo
한국지능정보사회진흥원(NIA), AI HUB. 웹사이트.
https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71812
한국과학기술정보연구원(KISTI), DATAON 인공지능 데이터 공유·활용 서비스.
https://aida.kisti.re.kr/about/intro
한국산업지능화협회(KOIIA), 산업데이터 서비스 포털. 웹사이트.
https://www.idx.or.kr/portal/main

2. 국외문헌

Bose, R. (2025). Introduction to Retrieval-Augmented Generation (RAG). In: Mastering Retrieval-Augmented Generation. Apress, Berkeley, CA. 3-32.
Tian, X. Z. (2014). Information-Applied Technology in Exploration-Internalization Mode. Advanced Materials Research, 886, 660-663.

ABSTRACT

Development of an Automated System for Dataset Generation:
A Consulting Approach for SME AI Internalization

Chang, Jin-Ho

Major in Smart Convergence Consulting

Dept. of Smart Convergence Consulting

Graduate School of Knowledge Service & Consulting

Hansung University

The application of artificial intelligence(AI) technology is rapidly surging across global industrial sectors. Rather than remaining a mere tool that substitutes for portions of human labor, AI is swiftly evolving into an intelligent operating system capable of autonomously learning and optimizing an organization's operational structures and decision-making frameworks. While the degree of AI utilization varies, a significant divide is emerging between enterprises that have internalized AI capabilities and those that have not. Undoubtedly, AI Transformation(AX) and its practical application will dictate corporate survival in the modern business environment.

To achieve this AI internalization, what preparations must small and medium-sized enterprises (SMEs) in South Korea undertake? AI

internalization is the process of accumulating the requisite knowledge and technical expertise for AI operations within an organization, thereby mitigating reliance on external resources and cultivating independent management and developmental capabilities. This study posits that the most actionable initial step for SMEs embarking on AX is the aggregation of dispersed data and information—fundamentally, the internalization of knowledge. Employees within SMEs must secure the competency to leverage AI tools on demand, while the organizations themselves must possess the requisite data to achieve their strategic objectives. To facilitate this, it is critical to establish a systemic architecture that automatically and continuously acquires up-to-date data, ensuring its availability for corporate analysis and application.

Consequently, this research developed and implemented an information system designed to internalize the capability of extracting ubiquitous web-based data—such as website text and attached files—and structuring it into datasets for AI utilization. By demonstrating the operationalization of this system, presenting empirical results, and incorporating expert reviews, this study provides actionable implications for South Korean SMEs pursuing AI internalization.

【Keyword】 Internalization of AI technology, AI training datasets, Web data extraction, AI transformation consulting