

SyncCrawl™ : 적응형 웹 데이터 수집 및 RAG 지식 자산화 플랫폼

수집 대상 웹사이트의 화면 구조가 바뀌어도 문맥을 파악해 수집을 이어가고, 모아온 데이터를 깔끔하게 정제해 사내 분석·AI 지식베이스(RAG)로 만들어 주는 플랫폼입니다.

EXECUTIVE SUMMARY · 도입 핵심 방향

복잡한 수작업과 부서 간 대기 시간을 줄이고, 현업 담당자가 쉽고 안전하게 업무를 처리할 수 있는 실무 중심 환경을 제안합니다.

01. 한눈에 보는 도입 기대 지표

Key Highlights

자율 보정

사이트 개편 시 선택자 복구
화면 구조 변경에도 수집 유지

4대 컨테이너

Server · Agent · Console 분
리
Kubernetes(AKS) 대규모 분산 수
집

RAG 직결

노이즈 제거 및 벡터 자산화
광고·태그 제거 후 pgvector 저장

실증 완료

대기업 엔터프라이즈 구축 실적
아모레퍼시픽·효성ITX 구축 완료

02. 도입 부서 주요 고민 및 SyncCrawl™ 해결 방안 조건표

Requirement & Solution Mapping

업무 영역	현업에서 자주 겪는 고민	SyncCrawl™ 해결 방법
사이트 개편 시 수집 중단	외부 사이트의 디자인이나 HTML 태그가 조금만 바뀌어도 크롤러가 멈춰서 개발자가 매번 고쳐야 함	고정된 태그 위치만 보지 않고 페이지 안의 글자와 배치를 분석해 바뀐 데이터 위치를 스스로 다시 찾음
운영자 관리 편의성	수집 일정을 바꾸거나 왜 에러가 났는지 보려면 개발자가 서버에 들어가 로그를 직접 뒤져야 함	웹 기반 운영 콘솔(smart-crawling-console)에서 시나리오 등록, 즉시 실행, 상태·로그 확인을 한눈에 처리
대규모 수집 및 차단 대응	여러 사이트에서 많은 데이터를 동시에 모을 때 세션이 꼬이거나 봇 차단에 걸려 수집이 실패함	시나리오 생성 워커와 실행 워커를 분리하고, 동시 세션 중재 및 프록시 제어로 안정적인 수집 유지
수집 데이터 AI 활용	긁어온 원본 HTML에 광고와 불필요한 메뉴가 섞여 있어 사내 AI(RAG)에 바로 넣으면 엉뚱한 답을 함	본문과 핵심 필드만 깨끗하게 골라내고 의미 단위로 나누어 (Chunking) 벡터 DB에 바로 저장

Quick Access

온라인 문서 및 체험 안내

기술 문서 포털(<https://doc.empasy.com/synccrawl/>) 및 통합 포털(<https://sync.empasy.io>)에서 상세 정보를 확인하실 수 있습니다.

구축형 엔터프라이즈 솔루션 (별도 시연 지원)

도입 배경 및 기존 업무 방식과의 비교

현재 업무에서 반복되는 불편함을 짚어보고, 도입 후 업무 흐름이 어떻게 간편해지는지 비교합니다.

01. 기존 업무 환경에서 자주 겪는 4가지 어려움

Current Pain Points

왜 기존 방식으로는 업무 속도와 품질을 동시에 높이기 어려울까요?

- 경쟁사 가격, 글로벌 SNS 반응, 공공 공고 등 외부 데이터를 모을 때 사이트 구조가 바뀔 때마다 수집기가 멈춰 데이터가 비게 됩니다.
- 크롤링 스크립트가 개발자 PC나 개별 서버에 흩어져 있어, 지금 어떤 수집 작업이 돌고 있고 어디서 실패했는지 관리하기 어렵습니다.
- 로그인 세션이 필요한 사이트나 동적 화면(SPA)은 일반 수집 도구로 가져오면 빈 화면(백색 화면)만 저장되는 경우가 많습니다.
- 수집한 데이터를 정제해 사내 분석 DB나 AI 지식 저장소에 넣는 과정이 수작업으로 나뉘어 있어 최신 정보 반영이 늦어집니다.

02. 기존 처리 방식(AS-IS) vs SyncCrawl™ 도입 후(TO-BE) 흐름 비교

Process Transformation

기존 수작업 중심 방식 (AS-IS)

반복 수작업 · 대기 발생

1 개별 스크립트 수동 코딩

사이트마다 HTML 태그 위치를 찾아 코드 작성

2 사이트 개편 시 수집 멈춤

클래스명 변경 시 빈 값 수집 또는 에러 중단

3 서버 접속 후 수동 디버깅

개발자가 로그를 확인하고 선택자 코드 재배포

4 수작업 데이터 정제·가공

불필요한 광고·태그를 수동으로 걸러낸 뒤 DB 적재

SyncCrawl™ 도입 후 (TO-BE)

간편 처리 · 안전 검증

1 웹 콘솔에서 시나리오 설정

수집 대상 주소와 가져올 항목(상품명·가격·본문) 지정

2 분산 에이전트 자동 수집

전용 워커가 렌더링 확인 및 화면 변경 시 위치 자동 보정

3 콘솔에서 실시간 상태·로그 확인

DB 직접 접속 없이 웹 화면에서 성공·실패 즉시 점검

4 자동 정제 및 RAG 지식 저장

노이즈 제거 후 정형 JSON 및 벡터 DB로 즉시 적재

03. 핵심 추진 방향 요약

Key Direction

방향 01 · 쉬운 사용성

복잡한 명령어 없이 직관적 처리

개발 지식이 없는 현업 담당자도 친숙한 웹 화면과 일상적인 언어로 원하는 작업을 바로 실행할 수 있습니다.

방향 02 · 안전한 통제

사람이 확인하고 언제든지 복구

중요한 변경은 담당자가 눈으로 확인한 뒤 반영하며, 변경 직전 상태를 저장해 두어 실수가 생겨도 바로 되돌립니다.

방향 03 · 비용 효율화

반복 인건비 및 AI 요금 절약

수작업 공수를 줄이고, 캐시 재사용과 사내 경량 모델을 활용해 운영 비용 부담을 낮춥니다.

목표 시스템 구성도 (3-Layer 아키텍처)

사용자 화면층 · 애플리케이션 및 AI 엔진층 · 데이터 및 인프라층의 3개 계층으로 분리된 안정적인 구조입니다.

01. SyncCrawl™ 3-Layer 통합 아키텍처 다이어그램

Target Architecture



02. 기존 사내 시스템과의 유연한 연동 구조

Seamless Integration

연동 구분	지원 방식 및 표준 규격	도입 고객사 이점
단독 도입 및 단계적 확장	필요한 제품 1종만 먼저 단독 도입 가능하며, 추후 타 Sync 제품군과 표준 MCP로 즉시 연결	초기 도입 부담 없이 가장 급한 업무부터 순차 적용 가능

연동 구분	지원 방식 및 표준 규격	도입 고객사 이점
사내 기간계(ERP/SSO/DB) 연동	표준 REST API(/api/v1/...), OAuth2/JWT 통합 인증, JDBC 표준 DB 연결 지원	기존 운영 중인 사내 시스템을 갈아엎지 않고 그대로 연결 해 활용

핵심 작동 프로세스 및 4대 주요 기능

실제 업무가 어떤 순서로 처리되는지와 실무에서 바로 체감할 수 있는 4가지 핵심 기능을 안내합니다.

01. 4단계 업무 처리 프로세스 (Step-by-Step)

End-to-End Workflow

STEP 01

운영 콘솔에서 수집 시나리오 정의

소요시간: 수 분 내

관리자 웹 콘솔에서 수집할 사이트 주소와 추출할 항목(제목, 본문, 가격, 날짜 등), 실행 주기를 등록합니다.

STEP 02

시나리오 에이전트의 화면 분석 및 규칙 생성

소요시간: 자동 구성

시나리오 에이전트가 대상 페이지의 구조를 분석해 수집 규칙을 만들고, 사이트 개편 시에는 문맥을 보고 선택자를 보정합니다.

STEP 03

크롤링 에이전트 분산 실행 및 렌더링 검증

소요시간: 병렬 수집

독립된 크롤링 에이전트 컨테이너들이 동시에 수집을 수행하며, 화면이 완전히 났는지 확인해 빈 화면 수집을 방지합니다.

STEP 04

노이즈 정제 및 분석 DB · 벡터 저장소 적재

소요시간: 실시간

불필요한 광고와 HTML 태그를 걸어내고 깔끔한 정형 데이터와 벡터 임베딩으로 변환해 분석 시스템과 AI에 공급합니다.

02. SyncCrawl™ 4대 핵심 기능 상세

Core Capabilities

핵심 기능 01

Server · Scenario Agent · Crawl Agent 분리 구조

관리 서버와 시나리오 분석기, 실제 브라우저 실행 워커를 독립 컨테이너로 분리해 대규모 수집 시에도 안정적으로 확장됩니다.

핵심 기능 02

웹사이트 개편에 대응하는 적응형 선택자 복구

대상 사이트의 레이아웃이나 태그 이름이 바뀌어도 주변 텍스트와 시각적 구조를 파악해 수집 파이프라인이 끊기지 않게 돕습니다.

핵심 기능 03

동시 세션 중재 및 백색 화면 오탐 방지

여러 워커가 동시에 둘 때 로그인 세션이 충돌하지 않도록 조율하고, 페이지 로딩 지연을 정확히 판별해 누락 없는 수집을 지원합니다.

핵심 기능 04

사내 AI 정확도를 높이는 RAG 정제 파이프라인

단순히 웹페이지를 긁어오는 데 그치지 않고, 문맥을 살려 문단을 나누고 (Chunking) 벡터화하여 사내 AI가 바로 참고할 수 있게 만듭니다.

주요 화면 구성 및 기존 방식과의 비교

실무 담당자가 사용하는 주요 화면 구성과 기존 수작업·타사 도구 대비 달라지는 점을 정리했습니다.

01. 실무자용 주요 화면 및 메뉴 구성

Key Screens & Menus

화면 01

smart-crawling-console

스마트 크롤링 통합 운영 콘솔

등록된 수집 시나리오 목록, 실행 성공/실패 현황, 스케줄 일정을 한 화면에서 관리합니다.

화면 02

/scenarios/builder

실사이트 기반 시나리오 빌더

수집 대상 웹페이지를 미리 보면서 가져올 데이터 항목과 파라미터를 직관적으로 설정합니다.

화면 03

/executions/logs

작업 실행 상태 및 실시간 로그 뷰어

각 수집 작업의 진행 단계와 시스템/잡 로그를 웹에서 바로 확인해 문제 원인을 빠르게 찾습니다.

화면 04

/knowledge/pipeline

정제 데이터 및 RAG 지식 적재 현황

수집된 원본 데이터와 노이즈 제거 후 변환된 정형 JSON·벡터 청크 결과를 비교 확인합니다.

02. 기존 운영 방식 대비 SyncCrawl™ 비교표

Before vs After Comparison

비교 항목	기존 수작업 · 일반 도구 환경	SyncCrawl™ 도입 환경
대상 사이트 UI 변경 시	선택자 오류로 수집 중단 (개발자 수동 코드 수정 필요)	문맥 기반 선택자 자동 보정으로 수집 연속성 확보
수집 현황 및 로그 확인	개발자가 서버나 DB에 직접 접속해 로그 파일 확인	운영자 전용 웹 콘솔에서 실행 상태·로그 실시간 조회
보안 및 비밀번호 관리	설정 파일에 계정·API 키 평문 저장 위험	KeyVault CSI 연동 및 SSRF 내부망 접근 차단 기본 적용
AI(RAG) 데이터 연동	수집 후 별도 정제 프로그램을 다시 개발해야 함	노이즈 제거·청킹·벡터 DB 적재까지 한 파이프라인으로 처리

보안 관리 대책 및 운영 비용(TCO) 절감 방안

기업 내부 데이터를 안전하게 보호하는 보안 체계와 AI-인프라 운영 비용을 합리적으로 줄이는 방안입니다.

SECURITY & COST OPTIMIZATION

사내 폐쇄망·전용 클라우드 설치로 데이터 유출 걱정을 덜고, 사내 모델·캐시 활용으로 고정비 안에서 경제적으로 운영합니다.

01. 엔터프라이즈 3대 보안 및 안전 관리 대책

Security & Governance

보안 대책 01

SSRF(서버 측 요청 위조) 차단 및 URL 검증

크롤러가 허용되지 않은 내부망 주소나 비정상 리다이렉트로 접근하지 못하도록 URL 검증기가 사전 차단합니다.

보안 대책 02

KeyVault CSI 기반 시크릿 관리 및 ABAC 권한 제어

인증 정보와 암호키를 YAML에 평문으로 두지 않고 보안 볼트에서 주입하며, 속성 기반 접근 제어(ABAC)로 데이터를 보호합니다.

보안 대책 03

CodeQL · Semgrep 정적 보안 검증 및 폐쇄망 지원

소스코드 보안 취약점 점검을 상시 수행하며, 외부망과 분리된 사내 환경에서도 내부 LLM과 연동해 안전하게 동작합니다.

02. 운영 비용(TCO) 및 리소스 절감 비교

Cost Optimization Analysis

비용 항목	기존 방식 (수작업 · 외부 유료 API 전역 의존)	SyncCrawl™ 최적 운영안 (자동화 · 캐시/사내GPU 결합)
크롤러 유지보수 개발 인건비	사이트 개편 때마다 수집 스크립트 수정에 상시 인력 투입	적응형 선택자 보정과 웹 콘솔 관리로 유지보수 공수 대폭 절감
데이터 정제 및 전처리 비용	원시 HTML 정제 및 RAG 변환용 별도 파이프라인 구축	수집과 동시에 정형화·벡터화까지 일괄 처리해 구축비 절약
인프라 운영 효율성	수집기마다 개별 가상머신(VM) 상시 가동	Kubernetes 기반 멀티 레플리카 구조로 수집 시간대에만 탄력 운영

03. 고객사 환경에 맞춘 3가지 배포(설치) 옵션

Deployment Options

옵션 A · 사내 구축형

온프레미스 (폐쇄망)

외부 인터넷과 분리된 고객사 자체 서버 및 사내 GPU 환경에 독립 설치하여 보안 규정을 준수합니다.

옵션 B · 전용 클라우드

고객사 전용 VPC

AWS, Azure, NCP 등 고객사 전용 격리 네트워크 안에 컨테이너(Docker/K8s) 기반으로 신속하게 구축합니다.

옵션 C · 하이브리드

사내 DB + 클라우드 연동

민감한 고객 원장은 사내에 두고, 필요한 연산 기능만 암호화 통신으로 연결해 초기 구축 부담을 낮춥니다.

단계별 도입 추진 일정(WBS) 및 산출물 안내

사전 검증(PoC)으로 실효성을 먼저 확인한 뒤 본 시스템 구축과 현업 이관까지 안전하게 진행합니다.

01. 3단계 표준 도입 일정 및 주요 과업 (WBS)

Implementation Roadmap

추진 단계	소요 기간	주요 수행 과업 내용	제공 산출물
1단계: 수집 대상 분석 및 시범 구성	1~2주차 (2주)	핵심 타겟 사이트 5~10곳 선정 · 수집 스키마 정의 및 시나리오 에이전트 검증	수집 대상 분석서 · 시범 수집 데이터
2단계: 분산 클러스터 및 콘솔 구축	3~6주차 (4주)	Server/Agent/Console 컨테이너 배포 · KeyVault 보안 적용 · RAG 벡터 DB 연동	아키텍처 구성도 · 연동 명세서
3단계: 안정화 및 운영자 교육	7~8주차 (2주)	대량 수집 스케줄 리허설 · 운영 콘솔 시나리오 관리 및 로그 점검 실습 교육	운영자 매뉴얼 · 완료보고서

02. 단계별 품질 점검 기준 (Quality Gate)

Quality Assurance

GATE 01 · 착수/설계

요구사항·아키텍처 점검

고객사 업무 환경과 연동 대상 API, 보안 요건을 사전 확인하고 설계서를 확정합니다.

GATE 02 · 실행데이터 연동

실환경 동작 검증

가짜 데이터가 아닌 실제 테스트 DB와 연동하여 기능 동작과 데이터 정합성을 확인합니다.

GATE 03 · 통합 테스트

성능·권한·복구 리허설

역할별 접근 권한, 예외 상황 시 스냅샷 복원, 동시 접속 응답 속도를 종합 점검합니다.

GATE 04 · 변화관리/이관

현업 실습 교육 및 오픈

운영자와 현업 담당자가 직접 화면을 조작해 보는 실습 교육 후 안정화 지원에 돌입합니다.

03. 전담 수행 조직 및 역할 분담

Project Organization

역할 구분	주요 담당 업무	고객사 협업 사항
PM / 솔루션 아키텍트	전체 일정·품질 관리, 사내 시스템 연동 아키텍처 및 보안 설계 총괄	도입 목표 범위 확정 및 주간 진척 점검
도메인·AI 엔지니어	SyncCrawl™ 코어 구축, 업무 규칙 및 프롬프트·모델 최적화, 인터페이스 개발	연동 대상 시스템 API/DB 명세 공유
QA / 변화관리 전담	실환경 시나리오 검증, 운영자·현업 사용자 매뉴얼 작성 및 대면 실습 교육 진행	현업 시범 사용자(UAT) 참여 및 검수

도입 기대 효과(ROI), 주요 활용 사례 및 기술 지원

도입 전후의 실질적인 업무 개선 효과와 안정적인 운영을 뒷받침하는 3단계 기술 지원 체계입니다.

01. 도입 전·후 기대 효과 비교 (ROI)

Expected Benefits

측정 항목	도입 전 (기존 방식)	SyncCrawl™ 도입 후	개선 효과
신규 사이트 수집 시나리오 등록	사이트당 2~3일 (개발자 스크립트 코딩 및 배포)	30분 내외 (콘솔 및 시나리오 에이전트 활용)	구축 속도 대폭 향상
사이트 개편에 따른 수집 중단 시간	오류 인지부터 코드 수정 배포까지 수일 공백	문맥 기반 선택자 보정 및 콘솔 즉시 조치	유지보수 공수 80% 절감
외부 데이터 사내 AI(RAG) 반영 주기	수집 후 수동 전처리를 거쳐 주/월 단위 반영	수집-정제-벡터 적재 자동 연결로 일 단위 최신화	최신 시장 정보 확보

02. 추천 도입 대상 및 주요 활용 시나리오

Use Cases & References

활용 분야 01

아모레퍼시픽 글로벌 SNS 크롤링 구축 (실증 완료)

국내외 주요 SNS 및 뷰티 채널의 소비자 반응·트렌드 데이터를 안정적으로 수집·분석하는 파이프라인 구축

활용 분야 02

효성ITX 스마트 크롤링 시스템 구축 (실증 완료)

대규모 웹 데이터 수집 시나리오를 중앙 콘솔에서 관리하고 고가용성 분산 에이전트로 실행하는 엔터프라이즈 환경 구축

03. 3단계 전담 기술 지원 및 유지보수 체계 (SLA)

Maintenance & Support

LEVEL 1 · 헬프데스크

15분 내 초동 응답

온라인·유선 기술 지원 창구를 통해 사용법 문의 및 설정 변경을 빠르게 안내합니다.

LEVEL 2 · 전담 엔지니어

1시간 내 원인 분석·복원

연동 이상 발생 시 로그 분석과 스냅샷 원상 복원을 통해 업무 지연을 최소화합니다.

LEVEL 3 · 코어 연구소

정기 점검 및 기능 패치

분기별 성능·보안 정기 점검 리포트를 제공하고 코어 엔진 업데이트를 지원합니다.

CONTACT & LIVE DEMO

주식회사 엠파시 (Empasy Inc.) | 살아 있는 소프트웨어는 엠파시가 만듭니다

서울특별시 금천구 가산디지털1로 205-27 · 도입 문의: poh@empasy.com / 0507-1360-8169

공식 홈페이지: <https://empasy.io>

통합 데모 포털: <https://sync.empasy.io>

기술 문서 포털: <https://doc.empasy.com>