

출판환경 변화 대응 R&D 2 단계(2026)

저작권 침해요소 아카이빙 실무 진행 방안

작성자: 엄진섭

작성 목적: 2 단계 R&D 를 안전하게 종료하기 위해, 저작권 침해 케이스 아카이빙(컴복스)과 분석 사례 리포트(바이칼)의 분류체계 및 진행 계획을 정리한다.

대상 과제: RS-2024-00442061 「출판환경 변화 대응을 위한 생성형 AI 기반 출판 콘텐츠 분석 및 공유 플랫폼 기술 개발」

해당 단계: 2 단계 (2026.01.01~2026.12.31)

문서 버전: ver. 1.2 (그룹별 매핑 재정렬 반영)

I. 사업 개요

본 과제는 1 단계(2024.07.01~2025.12.31)와 2 단계(2026.01.01~2026.12.31)로 구성된 30 개월 R&D 이며, 2026 년은 단일 회계연도 내에서 과제를 종료해야 한다. 저작권 침해 방지 관련 산출물은 1 단계 보고서 "2 단계 목표"에 명시되어 있다.

II. 2026 년 수행 과제

컴복스 담당

- 저작권 침해요소 케이스 아카이빙 (DB 구축)

바이칼에이아이 담당

- 저작권 침해 정보 케이스 분석
- 저작권 침해요소별 메타데이터 분석 사례 리포트

오투오 담당: 정량 KPI (공인인증)

- 표절여부판별 정밀도(precision): 1 단계 96% → 2 단계 목표 97%, 가중치 10%

III. 안전 종료 4 대 원칙

- **원칙 1. 범위 확장 금지** — 1 단계 보고서에 명시된 산출물만 정확히 채운다.
1 단계에서 완성된 "콘텐츠 구성요소 정의서 1 식"과 "표절 감지 모델"을 모수로 활용해 케이스 누적·정리에 집중한다.
- **원칙 2. 분류체계 1Q 확정** — 아카이빙과 분석 리포트는 동일 분류체계를 공유해야 한다. 2026 년 1~2 월 컴복스·바이칼 합동으로 ver.1.0 을 확정하고 이후 변경하지 않는다.(IV 분류체계에 분류)
- **원칙 3. 공인인증 일정 역산** — 시험기관 일정상 8~12 주 소요 예상되므로, 9 월까지 모델·테스트셋 동결 → 10~11 월 공인시험 → 12 월 결과 반영 보고서 제출.
- **원칙 4. 정량화된 증빙** — 아카이빙은 정성 산출물이므로 "케이스 N 건 이상, 케이스당 메타필드 M 개" 등 계량 기준을 미리 정해 보고서에 명시한다.

IV. 침해유형 분류체계 (메타 태그)

법령상 권리 + 저작인격권 + 위원회 상담사례집상 표절 실무유형을 결합하되, 나누구는 텍스트·전자책·종이책 중심 어문저작물 서비스이므로, 어문저작물에 적용되지 않는 권리(공연·전시·대여)는 분류 외로 처리한다.

1. 채택 카테고리 (총 10 중)

구분	카테고리	적용 범위
저작권재산권	복제권	핵심
저작권재산권	공중송신권(전송 포함)	핵심
저작권재산권	배포권	종이책·큰글자책 등 유형물 한정
저작권재산권	2 차적저작물작성권	핵심 (2 차 창작 기능)
저작인격권	공표권	핵심
저작인격권	성명표시권	핵심 (AI 집필 결과물 작성자 표시)
저작인격권	동일성유지권	핵심 (교정·교열에 따른 변형)

표절 실무	인용 표시 누락	핵심
표절 실무	자기 창작인 양 표시	핵심
표절 실무	2 차적저작물 수준 미달 가공을 새 창작으로 표시	핵심

2. 제외 카테고리 (총 3 종)

카테고리	제외 사유
공연권	텍스트·전자책 송신 중심 서비스로 공연 행위 부재
전시권	대법원 2010 도 4468 판결에 따라 어문저작물에 미적용 (단, 표지·삽입 이미지는 별도 관리)
대여권	한국 저작권법상 상업용 음반·일부 프로그램 한정, 어문저작물 미적용

V. 일정 (안)

시기	컴복스	바이칼
1Q 2026	분류체계 ver.1.0 확정 합동 작업	분류체계 합동 작업, 표절 모델 튜닝 착수
2Q 2026	1 단계 OCR 데이터 기반 케이스 1 차 누적	분석 사례 리포트 초안, 모델 튜닝 완료
3Q 2026	케이스 누적 완료, DB 공유 기능 구현	통합 표절 검출 시스템 구현, 9 월 모델 동결
4Q 2026	최종 아카이빙, 보고서 작성	10~11 월 공인시험, 12 월 결과 반영 보고서

VI. 위험 요인 및 대응

- 위험 1 — 표절 감지 모델 완성도 확인 필요. KPI 미달 (목표 97%): 8 월 자체 모의 공인시험 → 9 월 모델 동결 → 10~11 월 공인시험
- 위험 2 — 분류체계 후반 변경: 합동 분류체계가 후반에 흔들리면 양사 산출물 정합성이 무너짐. 대응: 1Q 확정 후 변경 금지를 합동 회의록에 기록.

- 위험 3 — 아카이빙 정량 기준 불명확: 평가자가 달성도를 판단 어려움. 대응: 1Q 분류체계 확정 시 케이스 수·필드 수 목표치 동시 확정.

VII. 표절 검출 알고리즘 개발 방향

1. 검출 계층 구조

저작권 침해는 "의거성 + 실질적 유사성"의 2 요건으로 판단되므로, 기술적 체크는 실질적 유사성을 정량화하는 것이 핵심이다. 텍스트 비교는 통상 3 개 계층으로 나뉜다.

계층	잡아내는 침해 양태	대표 알고리즘
어휘·표면	그대로 복붙, 부분 베끼기	n-gram Shingling, MinHash, SimHash, LSH, Jaccard/Dice, Rabin-Karp, TF-IDF + 코사인
의미	패러프레이즈, 어순 변경, 유의어 치환	Sentence-BERT(SBERT), SimCSE/KoSimCSE, DPR, Cross-encoder 재랭킹
서사·구조	줄거리 도용, 인물·사건 차용	스토리아크 추출 비교, 엔티티 그래프 비교, 콘텐츠 구성요소 기반 비교

2. 1 단계 자산 (재활용 대상)

- 1 단계 보고서(p.11)에 따르면, "콘텐츠 구성요소 추출 모델"을 기반으로 "콘텐츠 구성 요소 기반 표절 감지 모델"이 완성되어 있고, 단순 표현 유사성뿐 아니라 서사 구조 수준의 표절까지 탐지하도록 고도화되었다. 1 단계 성능 그래프상 임계값 0.75~0.8 구간에서 F1·정확도가 최고치를 보였으며, 표절 정밀도는 1 단계 96%(KPI 95%) → 2 단계 목표 97%이다.
- 즉, 의미·구조 계층은 이미 자체 모델이 있으므로 추가 R&D 모델 개발 없이 정밀도 1%p 향상에 집중하는 것이 안전한 종료 경로이다.

3. 권장 아키텍처 — 3 단 캐스케이딩

단계	모듈	역할	비고
----	----	----	----

1 차 필터	MinHash + LSH	신규 원고가 들어오면 기존 자서전 코퍼스에서 후보 N 건을 ms 단위로 추림	오픈소스 datasketch 활용
2 차 정밀	KoSimCSE/KoSBERT 임베딩 + 코사인 유사도	후보 N 건 대상 문단별 정밀 비교	sentence-transformers 활용
3 차 최종	1 단계 콘텐츠 구성요소 기반 모델	의심 케이스를 서사 구조 수준에서 최종 판정	1 단계 자산 그대로

이 구조는 신규 R&D 모델 개발이 아닌 검증된 오픈소스 + 1 단계 자산의 조합이므로, 평가 시 "신규 과제 추가"가 아닌 2 단계 산출물 중 '통합 표절 검출 시스템'의 구체화로 방어 가능하다.

4. 자서전 특수성 — 오탐(False Positive) 방지 장치

자서전은 공통 경험(초등학교 입학, 군 입대, 결혼식 등)으로 인해 표면 유사도가 자연스럽게 높아지므로, 별도 보완이 없으면 거짓 양성이 폭증해 정밀도 KPI 를 깎아먹는다.

- 공통 표현 사전: "에 입학하였다", "남동생이 태어났다" 같은 자서전 빈출 패턴을 비교 대상에서 제외
- 엔티티 마스킹 후 비교: 인명·지명·날짜는 1 단계 NER 자산(메타데이터 추출 F1 81%) 활용해 마스킹
- 보수적 임계값 설정: 정밀도 우선으로 0.85 이상 사용 (재현율 일부 손실 감수)

5. 비교 대상 코퍼스 — 권리관계 명확한 범위로 한정

- 나누구 플랫폼 내 원고 (이용약관상 비교 동의 확보 필수)
- 컴북스 자체 보유 자서전 (1 단계 OCR 처리 10 여 종)
- 공공·저작권 free 코퍼스 (국립중앙도서관 공공누리 자료 등)
- 외부 상용 자서전은 사전 크롤링·DB 화 금지, 인용·도용 의심 시 사후 수동 확인으로 처리

6. 안전 종료 관점 결론

- 새로운 표절 알고리즘 R&D 는 추가하지 않는다. "콘텐츠 구성요소 기반 표절 감지 모델"의 정밀도 97% 달성에 집중하되, 운영 단계 효율을 위해 MinHash+LSH 1 차 필터와 KoSimCSE 2 차 정밀 비교를 사전·사후 모듈로 결합한다.
- 본 구성은 2 단계 계획의 "통합 표절 검출 시스템"을 구현하는 자연스러운 형태로, 평가 시 신규 R&D 추가가 아닌 기존 계획의 구체화로 설명 가능하다.

VIII. 본문 텍스트 표절 검출의 실효성과 범위 설정

1. 문제 제기

나누구는 "저자가 자기 자서전을 쓰는 사이트"이므로, 표절 검출의 비교 대상이 플랫폼 내부 데이터로 한정될 경우 실효성 한계가 분명하다. 본 장은 이 한계를 정직하게 인정하고, R&D 트랙과 상용 서비스 트랙을 분리해 관리하기 위한 범위 설정 기준을 정리한다.

2. 내부 코퍼스 한정 검출의 한계

자서전 저자가 베껴올 가능성이 있는 출처를 정리하면, 내부 코퍼스만으로 잡을 수 있는 범위는 단 한 가지(다른 나누구 사용자 자서전) 뿐이다.

표절 원천 (베껴오는 출처)	내부 코퍼스로 검출 가능 여부
유명한 자서전·회고록 (출간 도서)	불가
신문·잡지 기사, 인터넷 칼럼	불가
시·소설·수필 (문학 작품)	불가
위키백과·블로그·SNS	불가
ChatGPT·Claude 등 AI 생성물	불가
가족·지인의 사적 기록(일기·편지)	불가 (비공개)
다른 나누구 사용자의 자서전	가능

또한 자서전은 본질적으로 본인 이야기이므로 다른 사용자 자서전을 베낄 동기 자체가 약하다. 즉, 사업적 침해 방지 효용은 제한적이다.

3. 외부 코퍼스 비교의 현실적 제약

실제로 위험한 케이스(시·노래 가사 무단 인용, 신문기사 옮겨오기, 유명 자서전 베끼기, AI 생성물 표절)는 모두 외부 코퍼스 비교가 필요하지만 다음 제약이 따른다.

- 법적 리스크: 출간 도서를 통째로 DB 화하는 것 자체가 복제권 침해
- 규모 문제: 위키·웹·뉴스·도서·SNS 커버 시 검색엔진 수준 인프라 필요
- AI 생성물: 원본 코퍼스가 존재하지 않아 비교 자체가 불가능

Turnitin·Copyleaks·KCI 등 상용 표절 검출 서비스가 고비용인 이유가 코퍼스 확보 때문이며, 단일 R&D 과제로 자체 구축하는 것은 현실성이 낮다.

4. R&D 과제 관점의 유효성

다행히 2 단계 R&D 종료 관점에서는 내부 코퍼스만으로 충분히 KPI 를 충족할 수 있다.

- KPI 는 모델 성능 지표: 표절 정밀도 97%는 주어진 테스트셋 기준 정확도이며, 공인시험도 시험기관 자체 테스트셋으로 평가
- 산출물 정의가 시스템 단위: 2 단계 명세는 "통합 표절 검출 시스템 개발"이며, "전 세계 콘텐츠 커버 서비스 런칭"이 아님
- 평가 기준과의 정합성: 정부 R&D 평가에서 단일 과제에 글로벌 코퍼스 보유를 기대하지 않으며, 모델 성능·시스템 동작 증명으로 충분

5. 트랙 분리 전략

구분	R&D 트랙 (2 단계 종료용)	상용 서비스 트랙 (R&D 외)
비교 코퍼스	나누구 내부 + 컴복스 OCR 자서전 10 여 종 + 공공 코퍼스	외부 상용 저작물 (별도)
목표	표절 정밀도 97% 공인인증	침해 신고 접수·삭제, 사후 대응
산출물	통합 표절 검출 시스템 (모델·아키텍처)	약관, 신고·심사 절차, 외부 솔루션 연동
처리 방식	기술적 검출	약관·면책·운영 절차

상용 트랙의 구체 수단

- 업로드 시 "타인 저작물을 허락 없이 포함하지 않았음" 동의 체크
- 저작권 침해 의심 신고 접수·심사 절차(notice & takedown)
- 명백한 침해 발견 시 출간 거부·삭제 권한 약관 명시
- 외부 코퍼스 비교가 필요한 경우 Turnitin-Copyleaks 등 유료 솔루션 API 연동 검토
- AI 사용 사실의 약관 고지 및 AI 산출물 저작권 귀속 조항 명문화

6. 보고서·평가 대응 문구

평가자가 "내부 데이터만으로 실제 침해 방지가 되는가?"라고 물을 가능성에 대비해, 보고서 본문에 다음 취지의 한 문장을 명시하는 것이 안전하다.

"본 시스템은 플랫폼 내부 자서전 간 표절 검출과 공공 코퍼스 대비 검증을 범위로 하며, 외부 상용 저작물 전체를 커버하는 침해 방지 서비스는 본 과제의 범위 외이다. 외부 저작물에 대한 보호는 이용약관, 신고·심사 절차, 필요 시 외부 표절 검출 서비스 연동으로 별도 처리한다."

이 명시는 R&D 과제 범위와 상용 서비스 범위를 분리하여, 평가 시에는 R&D 산출물로 방어하고 상용 런칭 시에는 약관·운영으로 보완하는 이중 안전망 역할을 한다.

7. 결론

내부 코퍼스 한정 표절 검출은 사업적 의미는 제한적이지만, R&D 과제 종료 관점에서는 KPI 충족이 가능하다. 상용 서비스의 실질적 침해 방지는 기술이 아니라 약관·신고 절차·외부 솔루션 연동으로 보완한다. 이 트랙 분리가 안전 종료 원칙과 부합하며, R&D와 상용 서비스 양쪽 모두를 무리 없이 끌고 가는 경로이다.

IX. 침해 케이스 ↔ 메타 태그 매핑 총괄표 (ver. 1.2)

1. 매핑 목적

IV 장에서确定的한 10 종 메타 태그(저작권재산권 4 + 저작인격권 3 + 표절 실무 3)와 자서전에서 발생할 수 있는 침해 케이스를 1:1 매핑하여, 컴복스 아카이빙(케이스 DB)과 바이칼 분석 사례 리포트가 동일 메타 라벨을 공유하도록 한다. 본 표는 케이스 라벨링 워크시트로 그대로 활용 가능하다.

- 그룹별 No 체계 도입 (A=저자 가해, B=저자 피해, C=플랫폼, D=다른 사용자, E=유족, X=분류체계 외)
- 각 그룹 내에서 매체·자산 유형별 묶음
- 향후 케이스 추가는 해당 그룹 끝번호로 부여 (분류체계 동결 원칙 유지)
- "구 No" 컬럼으로 ver.1.0/1.1 과의 추적성 확보

2. 매핑 총괄표 (38 개 케이스 × 10 종 메타 태그)

범례: 주(붉은색 굵게)는 가장 직접적으로 침해되는 권리, 보조(회색)는 함께 발생할 수 있는 권리.

No	구 No	케이스 유형	행위자	복제권	공중송신권	배포권	2 차적저작 물	공표권	성명표시 권	동일성유 지권	인용 누락	자기 창작인 양	미달 가공
그룹 A. 저자(가해) — A-1. 외부 텍스트 인용·수록													
A1	1	사·노래 가사 본문 무단 인용	저자(가해)	주	보조(전자책)	보조(종이책)					주		
A2	2	소설·수필 본문 발췌 무단 수록	저자(가해)	주	보조(전자책)	보조(종이책)					주		

No	구 No	케이스 유형	행위자	복제권	공중송신권	배포권	2차적저작물	공표권	성명표시권	동일성유지권	인용 누락	자기 창작인 양	미달 가공
A3	3	신문·잡지 기사 전문 옮겨 적기	저자(가해)	주	보조(전자책)	보조(종이책)					주		
A4	4	인터넷 블로그·SNS 글 무단 수록	저자(가해)	주	보조(전자책)				보조		주		
A5	5	위키백과·백과사전 본문 그대로 사용	저자(가해)	주	보조(전자책)				보조		주		
A-2. 타인 자서전·회고·사적 기록													
A6	6	유명인 자서전 일부 베끼기	저자(가해)	주	보조(전자책)	보조(종이책)						주	
A7	7	다른 자서전 줄거리·서사 변형 차용	저자(가해)	보조			주					보조	주
A8	8	친구·가족 회고를 본인 글로 옮김	저자(가해)	주			보조		주			보조	
A9	9	가족 일기·편지 무단 수록 (미공표)	저자(가해)	주	보조(전자책)			주	보조				
A10	10	부모·조부모 자서전 통째 재수록	저자(가해)	주		보조(종이책)	주	보조	보조				
A11	11	학창시절 친구의 시·편지 수록	저자(가해)	주				보조	주				
A12	12	동료 이메일·업무문서 인용	저자(가해)	주				주	보조				
A-3. 구술·강연·녹음													
A13	36	부모님·스승의 강연·설교 녹음해 본문에 옮김	저자(가해)	주	보조(전자책)			주	보조				
A-4. 학술·교육 자료													
A14	38	본인 과거 논문·학위논문 발췌 무단 수록	저자(가해)	주	보조(전자책)	보조(종이책)			보조		보조		
A15	39	교과서·교재 본문 그대로 수록	저자(가해)	주	보조(전자책)	보조(종이책)					주		

No	구 No	케이스 유형	행위자	복제권	공중송신권	배포권	2 차적저작 물	공표권	성명표시 권	동일성유 지권	인용 누락	자기 창작인 양	미달 가공
A-5. 번역물													
A16	31	외국 자서전·도서 직접 번역해 본인 글로 수록	저자(가해)	보조	보조(전자책)	보조(종이책)	주					주	
A-6. 이미지·시각 자산													
A17	13	인터넷 옛 사진·포스터 본문 삽입	저자(가해)	주	보조(전자책)				보조				
A18	14	졸업앨범 단체사진 무단 사용	저자(가해)	주									
A19	15	신문 스크랩·잡지 표지 사진 사용	저자(가해)	주	보조(전자책)				보조				
A20	44	만화 캐릭터·브랜드 로고 본문 삽입	저자(가해)	주	보조(전자책)				보조				
A-7. 음원·영상													
A21	16	오디오북 BGM 에 상업 음원 사용	저자(가해)	주	주				보조				
A-8. 디지털 사적 통신													
A22	49	동창회 카톡방·단톡방 대화 캡처 수록	저자(가해)	주	보조(전자책)			주	보조				
A-9. 사후·고인 자료													
A23	51	사망한 친구·동료의 유작·일기 본문 수록	저자(가해)	주				주	주				
A-10. AI 도구 사용													
A24	17	AI memorization 반환 결과 사용	저자(가해, 비의도)	주	보조(전자책)							보조	
A25	18	AI 생성물 우연 유사 (기존 저작물)	저자(가해, 비의도)	주	보조(전자책)								
A26	19	AI 결과물을 본인 저작인 양 표시	저자(가해)						보조			주	

No	구 No	케이스 유형	행위자	복제권	공중송신권	배포권	2차적저작물	공표권	성명표시권	동일성유지권	인용 누락	자기 창작인 양	미달 가공
A-11. 대필													
A27	20	대필 작가 작성분을 저자 단독 명의 출간	저자·플랫폼						주			보조	
그룹 B. 저자(피해)													
B1	22	다른 사용자가 내 자서전 베끼	저자(피해)	주	보조(전자책)	보조(종이책)			보조	보조		보조	
B2	23	다른 사용자가 내 서사·구조 차용	저자(피해)				주		보조	보조			주
B3	24	외부 사이트·SNS의 내 자서전 무단 게재	저자(피해)	주	주				보조				
B4	25	외부의 AI 학습 데이터로 무단 수집	저자(피해)	주	주								
그룹 C. 플랫폼													
C1	21	편집자·교정자 손이 많이 들어간 경우	플랫폼				보조		보조	주			
C2	42	유료 폰트 무단 사용 (본문·표지)	저자·플랫폼	주		보조(종이책)							
C3	54	플랫폼이 사용자 자서전을 AI 학습 데이터로 사용	플랫폼	주	주		보조						
그룹 D. 다른 사용자													
D1	26	2차 창작 기능에서 원저자 동의 없는 변형	다른 사용자	보조			주		주	주			
그룹 E. 유족													
E1	27	저자 사후 유족이 본문 임의 수정	유족	보조			보조			주			
E2	28	사후 유고 자서전에 미공표 일기 포함	유족	주				주	보조				
그룹 X. 분류체계 외 — 처리 방안 비교													

No	구 No	케이스 유형	행위자	복제권	공중송신권	배포권	2차적저작물	공표권	성명표시권	동일성유지권	인용 누락	자기 창작인 양	미달 가공
X1	29	자서전 등장 제 3 자의 사적 정보 노출	저자(가해)	사생활 침해 영역, 약관·운영 절차로 처리									
X2	30	자서전 등장 제 3 자의 명예 훼손 묘사	저자(가해)	명예훼손 영역, 약관·운영 절차로 처리									

3. 그룹별 통계

그룹	케이스 수	비중	주 권리 영역
A. 저자(가해)	27 건	71.1%	본문 텍스트 표절 검출의 핵심 대상
B. 저자(피해)	4 건	10.5%	신고·삭제 절차(notice & takedown) 대상
C. 플랫폼	3 건	7.9%	약관·라이선스·동의 절차 대상
D. 다른 사용자	1 건	2.6%	2 차 창작 기능 운영 정책 대상
E. 유족	2 건	5.3%	사후 권리 가이드라인 대상
X. 분류체계 외	2 건	5.3%	별도 약관·운영 절차
합계	38 건	100%	

4. 메타 태그별 등장 빈도 (38 개 케이스 누적)

메타 태그	주 태그 등장	보조 태그 등장	합계
복제권	24	3	27
공중송신권	5	18	23
성명표시권	5	14	19
2 차적저작물작성권	5	5	10
공표권	6	3	9
배포권	0	8	8
자기 창작인 양 표시	3	5	8
인용 표시 누락	6	1	7
동일성유지권	4	3	7
2 차적저작물 미달 가공	2	0	2

5. A 그룹 세부군별 검출 가능성

세부군	케이스 수	검출 가능성 (1 단계 자산 기준)
A-1 외부 텍스트	5 건	본문 텍스트 표절 모델로 검출 가능
A-2 타인 자서전·회고	7 건	외부 코퍼스 한정. 내부 코퍼스 검출 가능성 낮음
A-3 구술·강연 녹음	1 건	검출 불가, 약관·신고 절차
A-4 학술·교육	2 건	외부 코퍼스 필요
A-5 번역물	1 건	다국어 임베딩 필요, 검출 난이도 높음
A-6 이미지·시각 자산	4 건	텍스트 모델로 검출 불가, 별도 영역
A-7 음원·영상	1 건	텍스트 모델로 검출 불가, 라이선스 영역
A-8 디지털 사적 통신	1 건	검출 불가, 약관·신고 절차
A-9 사후·고인 자료	1 건	검출 불가, 약관·신고 절차
A-10 AI 도구 사용	3 건	A24·A25 만 일부 검출 가능 (memorization·우연 유사)
A-11 대필	1 건	검출 불가, 계약·약관 영역

시사점: A 그룹 27 건 중 본문 텍스트 표절 모델로 검출 가능한 케이스는 A-1 군(5 건) + A-2 군 일부(외부 출간 자서전 제외하면 0~1 건) + A-10 일부(2 건) = 약 7~8 건이다. 나머지 19~20 건은 모두 약관·신고 절차·외부 솔루션 영역으로, VIII 장 트랙 분리 원칙의 매핑 근거가 된다.

6. 가중 위험 케이스 (주 태그 2 개 이상)

분쟁 시 손해배상액·형사 처벌이 가중되므로 아카이빙 표본 추출 우선순위 상위로 잡는다.

No	케이스 유형	주 태그
A10	부모·조부모 자서전 재수록	복제권, 2 차적저작물작성권
A13	강연·설교 녹음	복제권, 공표권
A21	오디오북 BGM 상업 음원	복제권, 공중송신권
A22	단톡방 대화 캡처	복제권, 공표권
A23	사망 친구·동료 유작·일기	복제권, 공표권, 성명표시권
B3	외부 사이트 무단 게재	복제권, 공중송신권
B4	외부 AI 학습 데이터 수집	복제권, 공중송신권
C3	플랫폼의 AI 학습 데이터 사용	복제권, 공중송신권
D1	2 차 창작 변형	2 차적저작물작성권, 성명표시권, 동일성유지권

A 그룹 5 건, B 그룹 2 건, C 그룹 1 건, D 그룹 1 건. A23·D1 이 주 태그 3 개로 최고 위험 케이스.

7. 분류체계 검증 시사점

- 복제권·공중송신권·성명표시권 3 대 핵심 태그가 압도적 우위 → 표본 추출 시 이 세 태그 결합 케이스를 다양하게 확보
- 공표권이 주 태그로 6 건 확보 → 미공표 저작물(강연·단톡방·고인 유작·유족 유고) 영역이 자서전 침해의 중요한 축임을 매핑이 확인
- 2 차적저작물 미달 가공은 등장 빈도 낮음(2 건) → 학술 사례·판례로 보강 필요
- 배포권은 주 태그 0 건, 보조만 등장 → IV 장 "종이책·큰글자책 한정" 적용 범위 좁힘이 매핑 결과로 검증됨

8. 운영 권장 사항

- 본 표를 ver. 1.2 로 1Q 2026 합동 회의에서 확정. 이전 버전과 추적성은 "구 No" 컬럼으로 보장
- 라벨링 워크시트로 활용: 신규 케이스는 그룹(행위자) 결정 → 세부군(매체) 결정 → 끝번호 부여 순
- 우선순위 배치: 가중 위험 케이스 9 건을 아카이빙 표본 추출 우선순위 상위로 잡음
- 열(메타 태그) 구조는 변경하지 않음. 행 추가 시에만 No 부여 (분류체계 동결 원칙 유지)
- 케이스 추가가 누락 영역(번역·구술·시각자산·사적통신·사후권리)에서 발생하면 해당 세부군 끝번호 + 1 로 부여

9. 결론

본 매핑표는 IV 장 분류체계와 자서전 침해 케이스 간의 정합성을 확보하고, 컴복스 아카이빙과 바이칼 분석 리포트의 라벨링 일관성을 보장하는 운영 워크시트다. 1Q 2026 합동 회의에서 ver.1.2 로 확정된 뒤, 케이스 추가 시에만 행을 추가하고 열(메타 태그) 구조는 변경하지 않는다. R&D 산출물(통합 표절 검출 시스템)과 상용 서비스 운영(약관·신고·외부 솔루션) 양쪽 모두에서 일관된 분류 기준으로 작동한다.

출처

제공 자료 (프로젝트 내부)

- 1 단계보고서_출판환경_커뮤니케이션북스 1128.pdf — 2 단계 목표, 성능지표 표(표절 정밀도 96→97%), 콘텐츠 구성요소 기반 표절 감지 모델(p.11), 메타데이터 추출 F1 81%, 자서전 OCR 10 여 종
- 2024-05 출판환경변화_커뮤니케이션북스_1 단계발표자료_20251208.pdf — 출판 포맷 구성
- 자서전사업개요 20240924_20250519.pdf — 사업 발전 단계

외부 자료

- 국가법령정보센터 「저작권법」 제 5 조(2 차적저작물·번역), 제 11~13 조(저작인격권), 제 14 조 제 2 항(사후 인격권), 제 16~22 조(저작재산권), 제 25 조(공표된 저작물의 인용), 제 136 조(벌칙) — <https://www.law.go.kr/lsEflInfoP.do?lsiSeq=148848>
- 한국저작권위원회 「네티즌이 알아야 할 저작권」 — <https://www.copyright.or.kr/information-materials/common-sense/knowledge-for-netizen/index.do>
- 한국저작권위원회 「저작권 상담사례집」 (표절 실무 3 유형) — <https://www.copyright.or.kr/>
- 찾기쉬운 생활법령정보 「저작재산권」(의거성·실질적 유사성 기준) — <https://easylaw.go.kr/CSP/CnpClsMain.laf?popMenu=ov&csmSeq=695&ccfNo=2&cciNo=1&cnpClsNo=2>
- 대법원 2010. 9. 9. 선고 2010 도 4468 판결 (어문저작물에 대한 전시권 미적용)
- Frontiers in Computer Science 「Plagiarism types and detection methods: a systematic survey」(2025) — <https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1504725/full>
- RETSim, arXiv:2311.17264 — <https://arxiv.org/html/2311.17264>

- Milvus Blog 「MinHash LSH in Milvus」(2025.05) — <https://milvus.io/blog/minhash-lsh-in-milvus-the-secret-weapon-for-fighting-duplicates-in-llm-training-data.md>
- DZone 「Shingling for Similarity and Plagiarism Detection」 — <https://dzone.com/articles/shingling-for-similarity-and-plagiarism-detection>
- arXiv 1702.03082 「Using Word Embedding for Cross-Language Plagiarism Detection」 — <https://arxiv.org/pdf/1702.03082>
- Turnitin (상용 표절 검출) — <https://www.turnitin.com/>
- Copyleaks (AI 생성물 탐지 포함 상용 솔루션) — <https://copyleaks.com/>
- KCI 문헌 유사도 검사 서비스 — <https://check.kci.go.kr/>
- 한국콘텐츠진흥원 — <https://www.kocca.kr/> (공고번호 2-24-D000-790 호)