Biweekly Report
통권 제87호(2026. 8-1)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제87호(2026. 8-1)

- AI 시대의 파일 진위를 넘어선 신뢰성 검증 기술
- IPTC C2PA 안내서로 본 뉴스 콘텐츠 권리정보의 기록과 검증
- 음향 식별 기술과 딥페이크 시대의 콘텐츠 신뢰

AI 시대의 파일 진위를 넘어선 신뢰성 검증 기술

뉴스 브리프

사진, 영상, 음성, 문서가 편집 도구와 AI로 손쉽게 조작되면서, 우리가 접하는 디지털 파일을 그대로 신뢰하기 어려운 환경이 형성되고 있다. 이에 파일의 진위와 신뢰성을 판별하는 검증 기술이 주목받고 있으며, 최근 미국의 위조 탐지 소프트웨어 기업 애티스티브가 사진과 문서, 음성, 영상을 하나의 화면에서 통합 검증하는 플랫폼을 선보였다. 이 기술은 여러 AI 모델로 조작 흔적을 0에서 100 사이의 점수로 수치화하고, 파일마다 고유한 지문을 부여해 변경과 재사용 여부를 추적한다. 여기서 제기되는 쟁점은 파일이 진짜인지를 가리는 것을 넘어, 그 파일을 특정 결정에 신뢰하고 써도 되는지를 어떻게 판별하는가에 있다. 본 보고서는 검증 기술의 작동 원리를 살펴보고, 앞으로 필요한 기술적 대응 방향을 전망한다.

뉴스 플러스

Ⅰ. 서론: 디지털 파일의 진위 검증 기술과 중요성 증대

- **위조 미디어 확산과 파일 검증의 필요성**

사진과 영상, 음성, 문서를 사실처럼 바꾸는 편집 도구와 딥페이크 기술이 빠르게 확산되면서, 우리가 눈으로 확인한 파일을 그대로 사실로 받아들이기 어려운 상황이 이어지고 있다. 이러한 흐름 속에서 미국의 위조 탐지 소프트웨어 기업 애티스티브(Attestiv)는 2026년 7월 제출된 파일을 자동으로 검증하는 새로운 플랫폼 딥스캔(DeepScan)을 공개했다. 딥스캔은 개별 파일의 조작 및 위조 여부를 개별적으로 하나씩 가려내던 기존의 방식에서 한 걸음 나아가, 그 파일을 특정 청구나 신청, 거래와 같은 구체적인 업무 맥락 안에서 신뢰할 수 있는지 자체를 판별하는 방향으로 검증의 초점을 옮겼다. 사진과 문서, 음성, 영상을 하나의 화면에서 함께 분석하고, 기업이 정한 규칙과 임계값에 따라 통과, 보류, 추가 확인 및 상향 처리로 나누어 판정한다.

조작된 파일은 단순한 기술적 결함에 그치지 않고, 허위 정보의 확산과 신뢰의 훼손, 정교한 피싱 공격, 잘못된 지급과 심사 같은 실질적 손실로 이어질 수 있다. 그동안 위조 탐지는 파일의 진위성을 가려내는 데 주로 집중해 왔으나, 이것만으로는 해당 파일을 계약, 거래, 청구 등과 같은 결정에 활용해도 되는지까지 판단하기 어렵다는 한계가 지적되고 있다. 결국 문제의 핵심은 파일이 '진짜인가'라는 질문을 넘어, 그 파일을 특정 청구나 신청, 거래에 신뢰할 수 있는가 라는 질문으로 옮겨가고 있다. 이는 창작물의 원본이 훼손되거나 무단으로 재사용되는 상황과도 직접 맞닿아 있으며, 위조를 탐지하고 진본을 확인하는 기술이 어떻게 저작권 보호와 침해 판별의 문제로까지 확장되는지, 그 구체적인 기술적 원리를 본론에서 살펴볼 필요를 제기한다.

• 진위 검증 기술의 등장 배경과 의미

디지털 파일의 조작이 점점 쉬워지고 AI 생성물이 빠르게 늘면서, 사람의 눈이나 단일 검사만으로는 위조를 가려내기 어려워졌다. 이에 여러 AI 모델을 함께 동원해 조작 흔적을 점수로 나타내고, 파일마다 고유한 지문을 부여한 뒤 이를 원본과 대조하여 진위를 가리는 방식의 새로운 검증 기술이 등장하였다. 애티스티브의 기술은 사진, 영상, 음성 및 문서를 여러 개의 모델로 함께 분석하여 그 결과를 종합해 0에서 100 사이의 점수를 매기고, 각 파일마다 고유한 지문을 만들어 이후 변경이 불가능한 원장에 따로 보관해 두는 방식을 기본으로 취한다.

애티스티브가 공개한 기술의 의미는 개별 파일이 진짜인지 가려내는 것을 넘어, 그 파일이 언제 만들어졌고 이후 어떻게 바뀌었는지를 추적할 수 있다는 데 있다. 파일에 부여된 지문은 원본이 무엇인지 식별하는 근거가 될 뿐 아니라, 같은 파일이 다른 곳에서 다시 쓰이거나 여러 번 중복해서 등장하는지를 함께 가려내는 유용한 단서로도 작동한다. 이러한 특성은 제출된 파일의 신뢰성이 심사나 지급, 보도 등 핵심 결과에 직접적인 영향을 미치는 만큼, 보험, 금융, 언론 등의 주요 산업군에서 최근 도입 필요성이 크게 주목받고 있는 상황이다. 결국 검증 기술은 파일의 진위를 판별하는 것을 넘어, 원본을 보존하고 그 변경 이력을 확인하는 기술적 토대로 자리 잡아 가고 있다.

II. 본론: 다중 모델 검증과 지문 원장의 작동 원리

• 이상 점수 산정과 다중 모델 판별 방식

이상 점수(anomaly score)는 파일에서 감지된 조작의 흔적이 어느 정도인지를 여러 모델의 분석을 종합해 0에서 100 사이의 하나의 수치로 나타낸 값이다. 쉽게 말해 이 숫자가 낮을수록 파일이 원본에 가깝다는 의미이고, 숫자가 높을수록 어딘가 조작되었을 가능성이 크다는 뜻이다. 애티스티브의 검증 방식은 이렇게 산출된 하나의 점수를 통해 파일의 현재 상태를 한눈에 간단하게 확인할 수 있도록 만들어져

있다. 실제로 점수가 40 미만이면 대체로 안전한 파일로 보고, 60을 초과하면 조작이 의심되는 파일로 나누어 구분하며, 그 사이에 놓여 있는 애매한 값은 사람이 직접 한 번 더 자세히 살펴보고 최종적으로 다시 확인해야 하는 경우로 처리하게 된다.

이 점수가 산출되는 전체 과정은 서로 다른 역할을 맡은 여러 AI 모델이 하나의 파일을 동시에 나누어 분석하는 데에서 시작된다. 각각의 모델은 파일이 지닌 서로 다른 특징을 저마다 살펴보고, 조작이 의심되는 정도가 어느 만큼인지를 개별적으로 판단하여 그 결과를 내놓는다. 이렇게 서로 다른 여러 모델에서 나온 각각의 판단은 다시 하나로 종합되어, 최종적으로 그 파일 전체의 상태를 대표하는 하나의 값, 즉 0에서 100 사이의 이상 점수로 산출되는 것이다. 이때 임계값은 정상과 의심을 나누는 기준이 되는 특정한 점수를 뜻하며, 다시 말해 파일의 점수가 그 값을 넘어서면 사람이 한 번 더 직접 확인하도록 시스템이 자동으로 걸러내 주는 하나의 경계선에 해당한다고 볼 수 있다.

판별 결과는 히트맵(heatmap)으로도 함께 제공되는데, 히트맵은 파일에서 조작이 의심되는 위치를 색으로 표시해 보여주는 화면이다. 다시 말해 파일에서 조작이 의심되는 위치를 사람이 눈으로 직접 확인할 수 있도록 나타낸 것이다. 이러한 자동 판별은 사람이 파일을 하나씩 일일이 검사하지 않고도 많은 양을 빠르게 확인할 수 있게 해주어, 보험 청구나 금융 심사처럼 대량의 파일을 다루는 분야에서 널리 활용되고 있다. 또한 임계값을 각자의 위험 수준에 맞추어 조정할 수 있어, 원본과 견주어 파일이 변경되었는지를 판별하는 기준을 분야마다 다르게 적용할 수 있다는 점도 이 방식이 지닌 장점으로 꼽힌다.

• 조작 유형별 탐지 모델의 작동 원리

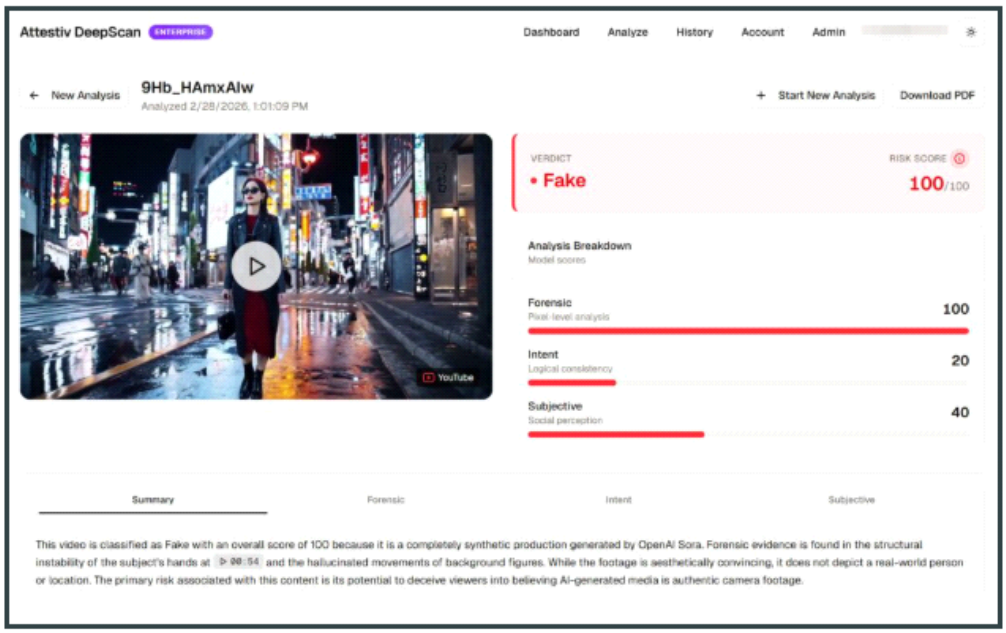
애티스티브의 검증은 하나의 모델이 모든 위조를 잡아내는 것이 아니라, 서로 다른 조작 유형을 각각 전담하는 여러 모델을 함께 쓰는 방식이다. 이 가운데 메타데이터(metadata) 모델은 파일에 담긴 정보를 살펴 편집이나 AI 처리 도구가 남긴 흔적을 추적한다. 메타데이터는 파일이 언제 어떤 기기나 프로그램으로 만들어지고 수정되었는지를 담은 부가 정보를 뜻한다.편집 도구가 이 정보를 건드리거나 지운 흔적이 있으면, 모델은 이를 조작 가능성의 단서로 삼는다.

얕은 위조(Shallow Fakes) 모델은 포토샵이나 PDF 편집기와 같은 일반적인 도구로 수정된 파일을 찾아낸다. 얕은 위조는 고도의 AI 없이도 손쉽게 만들 수 있는 단순한 형태의 조작을 뜻한다. 즉 이미지의 일부를 지우거나 문서의 숫자를 바꾸는 것처럼, 흔한 편집 도구만으로 이루어지는 변형이다. 이 모델은 파일에 이전 버전이 남아 있으면 이를 함께 대조해 어느 부분이 달라졌는지 확인한다. 반면 딥페이크(deepfake) 모델은 파일 자체에 남은 AI 조작의 흔적이나 확장자 불일치, 손상된 내용처럼 더욱 정교한 위조의 신호를 잡아낸다.

재사용(Reuse) 모델은 화면이나 다른 사진을 다시 촬영한 이미지를 식별한다. 재사용은 이미 존재하는 파일을 새로 만든 것처럼 다시 제출하는 행위를 뜻한다. 다시 말해 원본을 그대로 또는 약간 바꾸어 다른

상황에 다시 사용하는 경우다. 여기에 더해 역방향 이미지 검색을 수행하는 모델은 같은 이미지가 인터넷의 다른 곳에 이미 올라와 있는지를 확인한다. 이러한 재사용 탐지와 역방향 검색은 하나의 파일이 원래 어디에서 왔는지, 그리고 무단으로 다시 쓰이고 있는지를 기술적으로 가려내는 근거가 된다.

[그림 1] 애티스티브 딥스캔(DeepScan)의 실제 분석 결과 화면



출처: Attestiv, “Deepfake Video and Audio Detection Software”, 2026.08.03. 접속 기준, https://attestiv.com/deepfake-video-detection-software/

• 디지털 지문과 원장을 통한 진본 보존

디지털 지문(digital fingerprint)은 파일의 내용을 계산해 만들어 낸 고유한 식별값을 뜻한다. 다시 말해 파일마다 서로 다르게 부여되는, 그 파일만의 고유한 서명과 같은 값이다. 이 지문은 파일이 촬영되거나 시스템에 들어오는 시점에 부여되어, 변경이 불가능한 원장(ledger에 따로 보관된다. 원장은 여기서 한번 기록되면 이후 내용을 되돌리거나 고칠 수 없는 저장 공간을 뜻한다. 이렇게 지문을 원장에 남겨 두면, 나중에 언제든 원본과 대조해 파일이 그사이 변경되었는지를 확인할 수 있다.

애티스티브는 파일의 지문 상태를 세 가지로 나누어 표시한다. 첫 번째는 촬영 시점에 지문이 부여된 ‘인증’, 두 번째는 사후 시스템에 들여온 ‘반입’, 세 번째는 검증되지 않았거나 지문이 없는 ‘변경’으로 구분한다. 이러한 방식은 파일에 부여된 지문을 원본과 대조함으로써, 원본이 그대로 유지되고 있는지 아니면 누군가 손을 댔는지를 기술적으로 판별하는 근거가 된다. 또한 같은 지문이 여러 번 나타나는지를 확인하면, 하나의 파일이 중복해서 쓰이거나 재사용되는 상황도 함께 가려낼 수 있다. 결국 지문과 원장은 파일의 원본을 식별하고 그 변경 이력을 추적하는 기술적 토대로 작동한다.

III. 결론: 진위 검증 기술의 과제와 방향

• 공통 신뢰 기반으로의 확장 과제

결국 이번에 살펴본 검증 기술은 파일이 진짜인지 아닌지를 가리는 데 머물지 않고, 그 파일이 언제 만들어졌고 이후 어떻게 바뀌었는지까지 추적하는 방향으로 나아가고 있다. 여러 AI 모델이 조작 흔적을 점수로 나타내고, 파일마다 고유한 지문을 부여해 원본과 대조하는 방식은 사람의 눈만으로는 잡아내기 어려운 위조를 자동으로 걸러낸다. 이는 파일의 진위를 판단하는 일이 개인의 직관이 아니라 수치와 기록을 통해 이루어지는 단계로 옮겨가고 있음을 의미한다. 이러한 점에서 검증 기술은 원본을 식별하고 그 변경 이력을 확인하는 기술적 토대로 자리 잡아 가고 있다. 이를 통해 위조 탐지와 원본 보존은 서로 분리된 문제가 아니라 하나의 흐름 안에서 다루어지게 된다.

다만 이 기술이 넓게 자리 잡기 위해서는 아직 풀어야 할 과제가 남아 있다. 서로 다른 도구와 기기가 만들어 내는 지문과 이력이 하나의 표준 위에서 호환되어야 할 것으로 보인다. 또한 조작 기술이 정교해질수록 이를 탐지하는 모델도 함께 발전해야 하므로, 검증의 정확도를 꾸준히 높여 가는 기술적 보완이 요구된다. 따라서 앞으로의 방향은 개별 기업의 검증을 넘어, 파일의 진위와 이력을 사회 전반이 신뢰할 수 있는 공통 기반으로 넓혀 가는 데 있다. 이는 검증 기술이 특정 산업의 도구를 넘어, 디지털 파일 전반의 신뢰를 뒷받침하는 바탕으로 확장되어야 함을 의미한다.

참고문헌

• Attestiv, “Attestiv Launches DeepScan, Extending Deepfake Detection into File Validation”, GlobeNewswire, 2026.07.08., https://www.globenewswire.com/news-release/2026/07/08/3324028/0/en/attestiv-launches-deepscan-extending-deepfake-detection-into-file-validation.html
• Attestiv, “Deepfake Video and Audio Detection Software”, 2026.08.03. 접속 기준, https://attestiv.com/deepfake-video-detection-software/

IPTC C2PA 안내서로 본 뉴스 콘텐츠 권리정보의 기록과 검증

뉴스 브리프

뉴스 콘텐츠의 권리정보 문제는 저작권 표시나 이용조건을 파일에 적어 두는 것뿐 아니라, 그 정보가 CMS와 CDN, 플랫폼을 거치는 과정에서 유지·검증되는지와도 연결된다. IPTC가 공개한 C2PA 콘텐츠 크리덴셜 FAQ는 제작·편집 이력을 기록하는 기본 구조부터 발행자 신원과 저작권 표시, 이용조건, 인공지능 학습 거부 정보를 연결하는 방식까지 정리하고 있다. 다만 누가 어떤 정보를 제시했고 발행 이후 변경됐는지는 확인할 수 있지만, 서명자가 실제 권리자인지 해당 이용이 허락된 것인지까지 판정하기는 어렵다. 본 보고서는 IPTC FAQ를 중심으로 C2PA의 권리정보 기록·검증 구조와 언론사의 도입 조건을 살펴보고, 유통 과정의 메타데이터 삭제와 개인정보 보호 조치가 원출처 정보에 미치는 영향을 분석한다.

뉴스 플러스

I. 서론: 뉴스 콘텐츠 유통 과정에서 분리되는 권리정보

• 반복되는 복제·변환 과정에서 사라지는 출처정보

뉴스 이미지와 영상은 카메라와 편집 프로그램에서 제작된 뒤 언론사 콘텐츠관리시스템(Content Management System, 이하 CMS)과 콘텐츠전송네트워크(Content Delivery Network, 이하 CDN)를 거쳐 발행되고, 이후 포털과 소셜미디어, 영상 플랫폼으로 반복해 복제·변환된다. 이 과정에서 파일에 담긴 제작자, 발행자, 촬영일, 출처와 이용조건 등의 메타데이터(metadata)는 삭제되거나 콘텐츠와 분리될 수 있다.

권리정보가 유지되지 않으면 이용자는 콘텐츠를 최초로 발행한 언론사와 편집 여부를 확인하기 어렵다. 권리자 역시 성명표시나 이용허락 조건을 콘텐츠와 함께 전달하기 어려워진다. 허위 서사의 확산, 권위에 대한 신뢰 하락과 생성형 인공지능의 부상으로 콘텐츠가 어디에서 왔고 어떤 과정을 거쳤는지를 확인할 수 있는 수단의 필요성도 커지고 있다.

• 국제언론통신협의회, 언론사 대상 C2PA 안내서 공개

국제언론통신협의회(International Press Telecommunications Council, 이하 IPTC)는 2026년 7월 8일 방송사, 뉴스 발행사와 뉴스 통신사를 대상으로 콘텐츠 출처·진위 연합(Coalition for Content Provenance and Authenticity, 이하 C2PA)의 콘텐츠 크리덴셜(Content Credentials)*에 관한 FAQ 안내서를 공개했다.

이 문서는 새로운 기술 규격이나 구현 기준을 제정한 것이 아니라, C2PA 도입을 검토하는 언론사의 주요 질의에 답하는 형식을 취한다. 기본 개념뿐 아니라 도입 비용, 서명 도구, 발행자 신원 확인, CMS와 CDN의 메타데이터 삭제, 보안과 개인정보 보호 등 실제 뉴스룸에서 발생하는 문제를 다룬다. IPTC는 별도의 구현 가이드도 운영하고 있으며, 최신판은 2025년 6월 공개된 1.1버전이고 새로운 버전을 준비하고 있다고 밝혔다.[1)]

C2PA를 적용한 사례는 이미 여러 분야에서 확인된다. 프랑스 공영방송 프랑스 텔레비지옹(France Télévisions)은 일일 뉴스 방송 영상을 C2PA 인증서로 서명하고 있으며, 이 작업으로 유럽방송연맹(EBU) 혁신상을 받았다. 캐논(Canon)은 언론사용 진정성 이미징 시스템(Authenticity Imaging System)을 적용한 EOS R1과 EOS R5 Mark II를 출시했으며, 소니(Sony)는 카메라 진정성 솔루션(Camera Authenticity Solution)과 함께 스틸카메라와 캠코더 제품을 출시했다. 플랫폼에서는 틱톡(TikTok)이 업로드 시 C2PA 매니페스트(manifest)를 확인해 인공지능 생성 콘텐츠를 표시하고, 링크드인(LinkedIn)은 게시물에 첨부된 이미지와 영상 파일에 콘텐츠 크리덴셜이 삽입된 경우 이를 표시한다.[1)]

다만 IPTC는 C2PA가 콘텐츠의 진위나 인공지능 생성 여부를 직접 판별하는 기술은 아니라고 설명한다. C2PA는 제작자나 발행자가 기록한 정보를 확인하고, 콘텐츠와 메타데이터가 발행 이후 변경됐는지를 검증하는 방식이다. 이러한 기능적 범위를 전제로 볼 때, C2PA는 저작권 관점에서 다른 측면에서도 주목할 만하다. C2PA 기본 규격이 기록하는 제작·편집 이력에 더해, 별도 규격을 통해 발행자와 출처, 저작권 표시, 이용조건과 인공지능 학습 거부 정보를 콘텐츠에 연결할 수 있기 때문이다. IPTC는 이러한 정보가 실제 뉴스 제작·유통 과정에서 유지되기 위해 필요한 발행자 인증, CMS·CDN 설정, 메타데이터 삭제 대응과 개인정보 보호 방안도 함께 설명한다. 이에 본론에서는 IPTC FAQ를 중심으로 C2PA가 뉴스 콘텐츠의 권리정보를 기록·검증하는 구조와, 언론사가 이를 도입하는 과정에서 발생하는 주요 쟁점을 살펴본다.

* 콘텐츠 크리덴셜(Content Credentials): 미디어·기술기업이 함께 참여하는 표준화 조직 C2PA의 기술 규격을 기반으로 미디어 파일에 삽입되거나 데이터베이스에 저장되는 검증 가능한 메타데이터를 가리키는 이용자용 명칭

1) IPTC, “IPTC Frequently Asked Questions on C2PA Content Credentials”, 2026.07.08., https://iptc.org/std/guidelines/media-provenance/C2PA-FAQs/IPTC-C2PA-FAQs-v1.0-2027-07-08.pdf

II. 본론: C2PA 기반 권리정보 관리와 언론사 도입 쟁점

• C2PA의 권리정보 기록·검증 구조

C2PA는 콘텐츠에 관한 개별 정보를 어설션(assertion)이라는 단위로 기록한다. 여기에는 제작에 사용된 기기와 소프트웨어, 자르기, 크기 변경, 형식 변환과 편집 작업, 인공지능 이용 여부 등이 포함될 수 있다. 어설션은 클레임(claim)에 묶여 디지털 서명을 받고, 서명된 클레임과 관련 어설션은 매니페스트를 구성한다. 매니페스트는 미디어 파일에 삽입하거나 외부 저장소에 보관할 수 있다.

콘텐츠와 매니페스트를 연결하는 기본 방식은 하드 바인딩(hard binding)이다. 콘텐츠의 데이터를 해시 함수로 계산한 값을 매니페스트에 담고, 검증 시점에 다시 계산한 값과 비교하는 구조다. 콘텐츠가 변경되면 두 값이 일치하지 않으므로 서명 이후 변경 여부를 확인할 수 있다. 웹사이트 보안 인증서에도 활용되는 공개키 암호 방식을 사용하며, IPTC는 기반 기술이 미국 국립표준기술연구소(NIST)와 미국, 영국, 호주, 뉴질랜드, 캐나다의 정보기관으로 구성된 파이브 아이즈(Five Eyes)의 검토를 거쳤다고 설명한다.

적용 대상은 초기의 정지 이미지에서 확대돼 텍스트, 영상 파일, 오디오, PDF와 전자책 등 대부분의 미디어 형식을 포괄한다. 실시간 스트리밍 관련 규격도 공개됐으며 일부 기업이 개념검증(PoC)을 시연했지만, 실제 서비스에 적용된 사례는 아직 확인되지 않는다.

여기에 발행자가 전달하려는 저작권과 이용조건 등의 정보는 창작자 어설션 작업반(Creator Assertions Working Group, 이하 CAWG)*이 마련한 별도 규격을 통해 추가된다. CAWG 메타데이터 어설션을 이용하면 캡션, 출처, 제작일과 위치정보 같은 제작 관련 정보와 저작권 표시, 이용권한과 이용조건 등 권리정보를 서명된 메타데이터에 포함할 수 있다. 대체 텍스트 등 접근성 정보와 콘텐츠 제작 과정의 인공지능 이용 여부도 함께 기록할 수 있으며, 인공지능 학습 거부 의사는 이용조건의 일부로 표현된다.

IPTC 신뢰정보 가이드는 최소 권고 항목으로 저작권 표시, 크레디트 라인(credit line), 제작자, 접근성 정보와 관련 신뢰지표를 제시한다. 인공지능 이용 여부는 디지털 소스 유형 값으로 표현할 수 있다. 인공지능 학습 거부와 관련해서는 CAWG 학습·데이터마이닝 어설션이 별도로 마련돼 있다. 이는 콘텐츠를 인공지능 학습이나 텍스트·데이터마이닝(TDM)에 이용하지 말라는 의사를 기계가 읽을 수 있는 형태로 전달하기 위한 규격이다. IPTC는 이 방식이 유럽연합 인공지능법(EU AI Act)에 따른 TDM 거부 프로토콜을 확정하기 위한 유럽연합 집행위원회의 절차에서 평가 대상 가운데 하나로 검토되고 있으며, 최종 목록은 2026년 말 발표될 예정이라고 설명했다. 즉 C2PA 생태계는 콘텐츠의 제작 이력을 기록하는 데서 나아가 권리자가 전달하려는 저작권 정보와 이용조건, 인공지능 학습 거부 의사를 콘텐츠에 연결하는 방향으로 범위를 넓히고 있다.

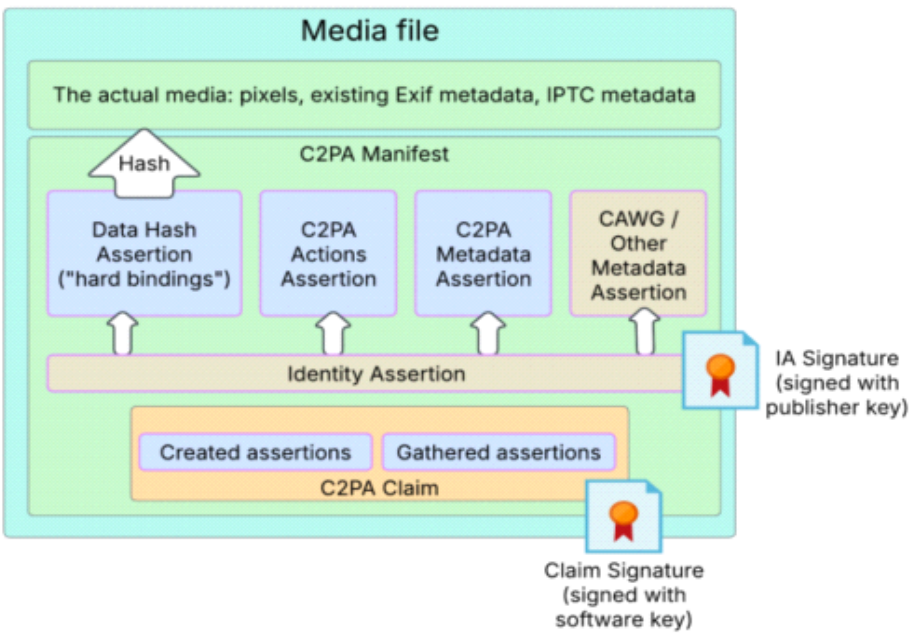
* 창작자 어설션 작업반(Creator Assertions Working Group, CAWG): 탈중앙 신원 재단(Decentralized Identity Foundation, DIF) 산하 조직으로, C2PA 규격 위에 적용되는 추가 규격을 개발

• 발행자 신원 검증과 언론사의 도입 조건

C2PA 신뢰체계는 콘텐츠에 서명할 수 있는 주체를 점차 엄격히 제한하는 방향으로 바뀌었으며, 콘텐츠에 서명하는 도구와 콘텐츠를 제작·발행한 주체의 신원을 구분한다. 적합성 검증 프로그램(Conformance Program) 도입 이후 공식 신뢰체계에서 인정되는 C2PA 매니페스트 서명은 검증을 통과한 하드웨어와 소프트웨어가 담당하며, 제작자와 발행자의 신원을 밝히는 역할은 콘텐츠를 제작하거나 발행한 주체의 신원정보를 C2PA 매니페스트와 연결하는 기능인 CAWG 신원 어설션이 담당한다.

언론사는 자체 조직 인증서로 신원 어설션에 서명해 콘텐츠를 발행한 주체를 표시한다. 결과적으로 하나의 콘텐츠에는 두 개의 서명 층위가 적용된다. C2PA 매니페스트는 도구 키로 서명되고, 발행자 신원과 편집 메타데이터는 발행자 키로 별도 서명된다.

[그림 1] C2PA 매니페스트와 발행자 신원의 이중 서명 구조



출처: IPTC, “IPTC Frequently Asked Questions on C2PA Content Credentials”, 2026.07.08., https://iptc.org/std/guidelines/media-provenance/C2PA-FAQs/IPTC-C2PA-FAQs-v1.0-2027-07-08.pdf

발행자 인증서가 실제 언론사의 것인지를 확인하는 수단으로는 IPTC가 운영하는 검증 뉴스 발행사 목록(Verified News Publishers List)이 사용된다. CAWG는 검증 도구가 발행자의 신원 인증서를 확인할 수 있도록 IPTC의 검증 뉴스 발행사 목록을 표준적인 신원 확인 수단 가운데 하나로 지정했다.언론사가 C2PA 생태계에서 발행자 신원을 검증받기 위한 주요 경로 중 하나인 셈이다. 다만 이 목록은 해당 조직이 저널리즘 뉴스 콘텐츠를 발행한다는 사실만 확인하며, 보도의 품질이나 사실성은 판단하지 않는다.

도입 경로는 두 가지다. 언론사가 자체 제작·서명 도구를 적합성 검증 프로그램에 통과시키거나, 이미 검증을 통과한 상용 도구를 뉴스룸 시스템에 통합하는 방식이다. 신원 어설션 서명에 필요한 인증서 비용은 발급기관에 따라 연간 수백~수천 초반 달러 또는 유로 수준이며, IPTC는 검증 뉴스 발행사 목록의 신원 어설션 서명용 인증서 발급기관으로 디지서트(DigiCert), 글로벌사인(GlobalSign), SSL.com과 트루포(Trufo)를 현재 인정하고 있다. 어느 경로를 택하더라도 비용의 대부분은 인증서 자체보다 서명 시스템을 기존 제작·발행 업무 과정에 통합하는 데서 발생한다.

권고 방식과 실제로 선택할 수 있는 상용 제품 사이에는 간극도 존재한다. FAQ 공개 시점 기준 IPTC는 적합성 검증을 통과한 상용 서명 도구 일곱 종을 제시했지만, 이 목록은 확정적·포괄적 명단이 아니며 추가 제품에 대한 제보를 요청하고 있다. 당시 IPTC가 파악한 제품 가운데 권고 방식인 서명된 CAWG 신원 어설션을 내장 지원한 것은 트루포의 제품뿐이었다. 상용 도구를 통합하는 경로에 한정하면 언론사가 선택할 수 있는 제품이 제한적이었던 셈이다.

• 유통 과정에서 사라지는 메타데이터와 대응

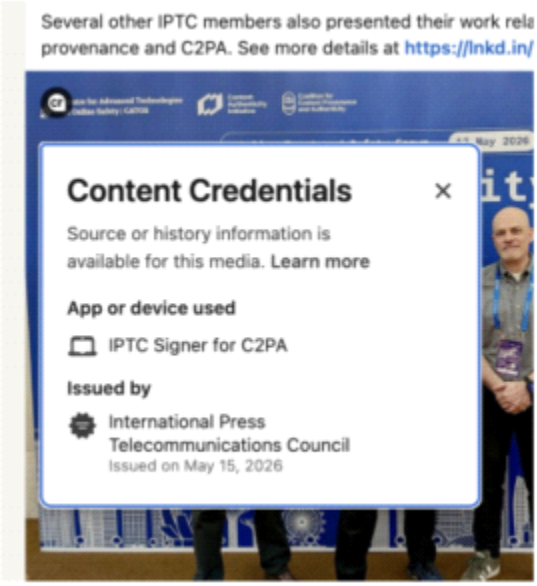
언론사가 콘텐츠에 매니페스트를 적용해도 이후 처리 과정에서 메타데이터가 삭제될 수 있다. IPTC가 제시하는 대표적 상황은 CMS가 기기와 화면 크기에 맞춰 여러 버전의 이미지를 생성하는 경우다. 원본만 서명하고 변형본을 서명하지 않으면 실제 이용자에게 전달되는 파일에는 출처정보가 남지 않을 수 있다.

IPTC는 CMS와 CDN이 메타데이터를 삭제하지 않도록 설정하고, 생성된 모든 이미지 변형본에 빠짐없이 서명하며, 삭제를 피할 수 없으면 별도 수단을 병행하도록 권고한다. 대부분의 CDN 도구는 전체 메타데이터를 삭제하지 않도록 설정할 수 있으며, 전체 정보의 보존이 어렵더라도 최소한 C2PA 메타데이터만은 삭제하지 않도록 별도로 구성할 수 있다.

그러나 콘텐츠가 언론사 시스템을 떠난 뒤에는 발행자가 처리 방식을 통제할 수 없다. 현재 콘텐츠 크리덴셜은 기본적으로 이용자에게 자동으로 노출되지 않으며, 어도비 콘텐츠 진정성(Adobe Content Authenticity) 크롬 확장 프로그램과 같은 브라우저 확장 기능을 설치한 경우 'cr' 표시를 확인할 수 있다. 이와 별도로 링크드인 등 일부 웹사이트는 자체적으로 유효한 콘텐츠 크리덴셜이 포함된 이미지에 'cr' 표시를 노출한다.

그러나 링크드인이 실제 이용자에게 전달하는 이미지 파일에서는 C2PA 메타데이터가 삭제돼 있어, 해당 파일을 검증 도구에 넣어도 출처정보를 확인할 수 없다. 콘텐츠 크리덴셜의 존재는 사이트 화면에 표시하면서 콘텐츠 크리덴셜이 포함된 파일은 전달하지 않는 구조다.

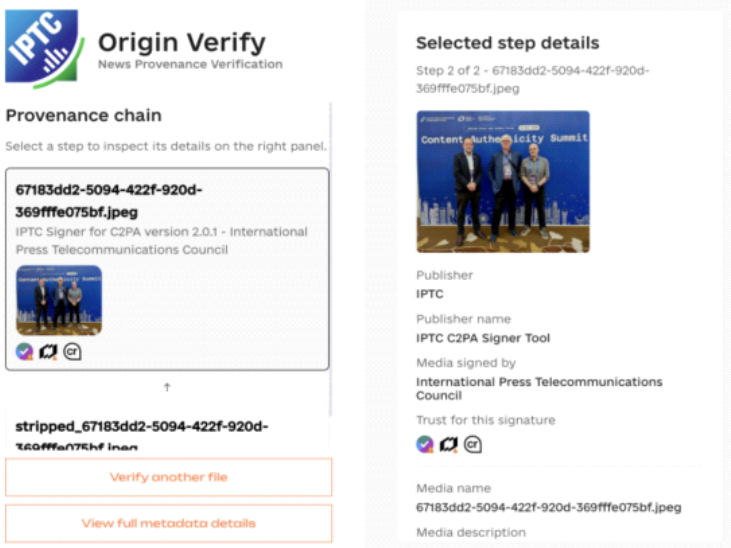
[그림 2] 링크드인의 콘텐츠 크리덴셜 표시 화면



출처: IPTC, "IPTC Frequently Asked Questions on C2PA Content Credentials", 2026.07.08., https://iptc.org/std/guidelines/media-provenance/C2PA-FAQs/IPTC-C2PA-FAQs-v1.0-2027-07-08.pdf

반면 서명된 원본 파일을 검증 도구에 직접 넣으면 전체 출처정보를 확인할 수 있다. IPTC는 뉴스 콘텐츠의 출처 검증을 위한 오리진 베리파이(Origin Verify) 도구를 제공한다.

[그림 3] IPTC 오리진 베리파이의 출처 검증 화면



출처: IPTC, "IPTC Frequently Asked Questions on C2PA Content Credentials", 2026.07.08., https://iptc.org/std/guidelines/media-provenance/C2PA-FAQs/IPTC-C2PA-FAQs-v1.0-2027-07-08.pdf

메타데이터가 삭제된 경우에 대비한 수단은 소프트 바인딩(soft binding)이다. 콘텐츠에서 산출해 등록소에 보관하는 지문(fingerprint)과 콘텐츠에 삽입하는 비가시적 워터마크가 대표적이며, 자르기, 색상 변경, 회전과 화면 캡처 같은 단순 변형을 견디도록 구현할 수 있다.

콘텐츠 진정성 이니셔티브(Content Authenticity Initiative, CAI)는 C2PA와 워터마크·지문을 병행하는 방식을 지속형 콘텐츠 크리덴셜(Durable Content Credentials)이라고 부른다. 유럽연합 인공지능법 실행규범도 인공지능 생성 콘텐츠에 이러한 병용 방식을 권고한다. 구글은 제미나이(Gemini)에 C2PA와 자사 워터마크 기술인 신스아이디(SynthID)를 함께 적용하고 있으며, 오픈AI도 신스아이디를 이용한 지속형 콘텐츠 크리덴셜 방식을 도입할 계획이라고 밝혔다.

다만 소프트 바인딩에는 구조적 제약이 있다. 지문과 워터마크는 통계적 방법을 사용하므로 오탐지와 미탐지를 완전히 없앨 수 없다. IPTC는 워터마크 알고리즘이 공개되면 역설계될 수 있어 관련 시스템이 독점·상용 기술에 의존하게 되며, 이를 완전한 공개표준으로 구현하기는 어렵다고 설명한다. 워터마크와 대응 메타데이터를 보관할 등록소도 필요해 운영 비용이 발생한다. IPTC가 특정 워터마킹 사업자를 추천하지 않는 이유이기도 하다. 어도비가 공개한 트러스트마크(TrustMark)는 공개 규격이어서 우발적인 메타데이터 삭제에 대응하는 용도로는 활용할 수 있지만, 악의적인 삭제까지 방지하지는 못한다.

• 개인정보 보호와 원출처 정보의 단절

C2PA 규격에는 위치나 정확한 촬영 시각처럼 민감할 수 있는 정보를 콘텐츠의 제작·편집·발행 이력이 이어지는 출처 사슬(provenance chain)에서 선택적으로 제거하는 리댁션(redaction) 기능이 규정돼 있다. 악의적인 행위자가 위치와 시간정보를 이용해 취재원, 피촬영자와 취재 인력의 사생활이나 안전을 위협할 수 있어 언론사에는 중요한 기능이다. 그러나 IPTC는 이 기능이 아직 여러 도구에 충분히 구현되지 않았다고 평가한다.

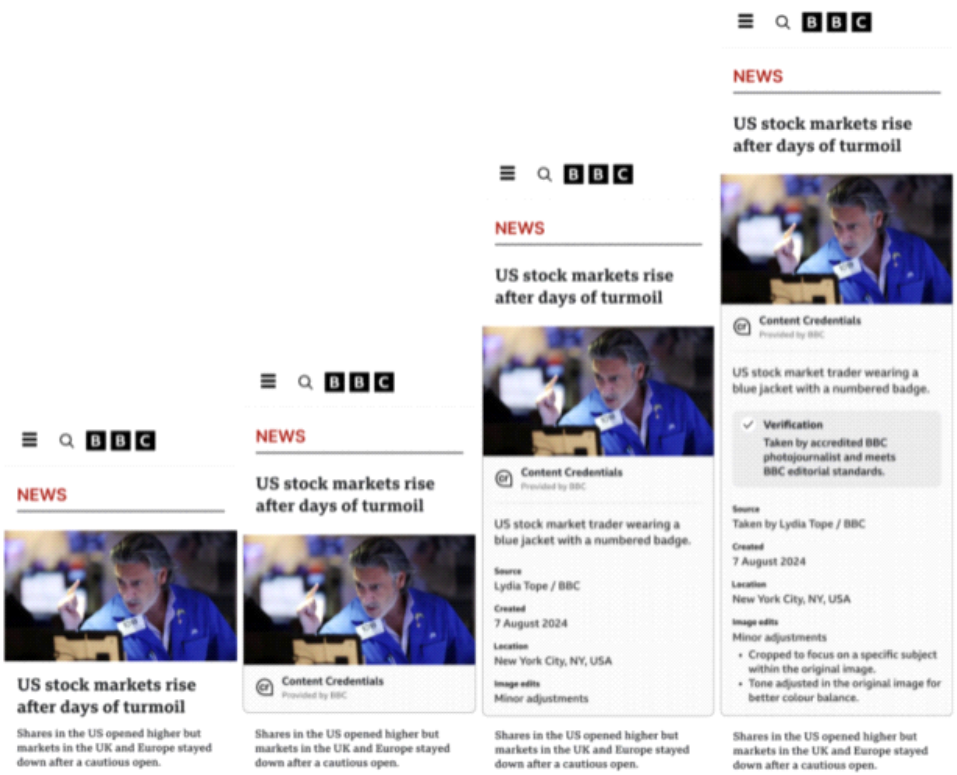
이에 따라 IPTC는 리댁션 기능이 보편화되기 전까지의 대안으로 클린 슬레이트 스탬핑(clean slate stamping)을 권고한다. 발행 시점에 기존 C2PA 메타데이터를 모두 제거한 뒤, 언론사가 책임지고 증명할 의사가 있는 정보만 새로운 매니페스트에 담아 자체 신원 어설션으로 다시 서명하는 방식이다.

IPTC FAQ에 인용된 전 게티이미지(Getty Images) 컨설턴트 폴 레이니츠(Paul Reinitz)는 제작자의 신원을 완전히 공개하지 않은 상태에서 출처까지 검증하는 데에는 한계가 있지만, 제작자가 신뢰할 수 있는 언론사에 신원을 확인시키고 언론사가 검증된 도구로 콘텐츠에 서명하면 제작자의 신원을 보호하면서 콘텐츠의 출처를 확인할 수 있다고 설명한다.

다만 이를 권리정보 관점에서 보면 반대 방향의 결과도 발생할 수 있다. C2PA 매니페스트에 어떤 정보를 입력할지는 강제되지 않으며, 클린 슬레이트 스탬핑 과정에서는 프리랜서, 통신사와 제보자 등 원제작자가 입력한 출처·귀속정보가 제거되고 언론사 정보로 대체될 가능성이 있다. 이 경우 이용자가 확인할 수 있는 출처 사슬은 언론사가 새로 서명한 발행 시점부터 시작한다. 권리정보의 지속성과 개인정보 보호가 항상 같은 방향으로 작동하지는 않는 셈이다.

IPTC의 일부 협력 기관은 C2PA와 CAWG 메타데이터가 언론사 웹사이트에서 이용자에게 어떻게 표시될 수 있는지를 구상한 화면도 제시했다.

[그림 4] 언론사 웹사이트의 콘텐츠 크리덴셜 노출 구상



출처: IPTC, “IPTC Frequently Asked Questions on C2PA Content Credentials”, 2026.07.08., https://iptc.org/std/guidelines/media-provenance/C2PA-FAQs/IPTC-C2PA-FAQs-v1.0-2027-07-08.pdf

III. 결론: 권리정보 유통 수단으로서 C2PA의 의미와 한계

• 뉴스 콘텐츠 권리정보의 검증 기반 마련

IPTC의 FAQ 공개는 새로운 구현 기준을 제정한 것이 아니라, 언론사가 C2PA를 실제 제작·발행 환경에 적용할 때 발생하는 도입 비용, 신원 확인, 메타데이터 삭제와 개인정보 보호 문제를 정리했다는 데 의미가 있다.

C2PA는 콘텐츠의 발행 주체, 제작·편집 도구, 저작권과 이용조건 등의 정보를 콘텐츠에 연결하고, 발행 이후 변경 여부를 확인할 수 있도록 한다. 이에 따라 뉴스 콘텐츠의 권리정보 관리는 파일에 정보를 표시하는 수준에서 정보에 서명하고 이를 검증하는 방식으로 확대될 가능성이 있다.

• **진위·권리 판단과 이용 통제에는 한계**

다만 C2PA는 콘텐츠의 사실성, 실제 권리자 여부나 이용의 적법성을 판단하지 않는다. 서명된 정보는 항상 제3자 검증을 거친 사실은 아니며, 기본적으로 발행자가 제시한 주장에 해당한다. C2PA는 그 주장이 누구의 것이고 이후 변경됐는지를 확인하는 층위에 있다. 이용조건을 전달하지만 이용 자체를 차단하지는 않는다는 점에서 기술적보호조치나 복제방지 기술과도 구별된다.

클린 슬레이트 스탬핑 과정에서 원제작자의 출처정보가 단절될 수 있고, 소프트 바인딩은 상용 워터마크와 외부 등록소에 의존한다는 한계도 있다. 결국 C2PA의 실효성은 기술 규격 자체보다 언론사, 제작 도구, CMS·CDN과 플랫폼이 권리정보를 얼마나 지속적으로 기록·유지·표시하는지에 달려 있다.

참고문헌

- Coalition for Content Provenance and Authenticity(C2PA), “Content Credentials: C2PA Technical Specification”, Version 2.4, 2026.04., https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html
- International Press Telecommunications Council(IPTC), “Expressing Trust and Credibility Information in IPTC Standards”, Version 1.1, 2025.05.08., https://iptc.org/std/guidelines/trust-and-credibility/
- International Press Telecommunications Council(IPTC), “IPTC Frequently Asked Questions on C2PA Content Credentials”, Version 1.0, 2026.07.08., https://iptc.org/std/guidelines/media-provenance/C2PA-FAQs/IPTC-C2PA-FAQs-v1.0-2027-07-08.pdf
- International Press Telecommunications Council(IPTC), “IPTC publishes Frequently Asked Questions on C2PA Content Credentials”, 2026.07.08., https://iptc.org/news/iptc-c2pa-content-credentials-faqs-released/

기술용어

순번	용어	설명
1	콘텐츠 크리덴셜 (Content Credentials)	미디어·기술기업이 함께 참여하는 표준화 조직 C2PA의 기술 규격을 기반으로 미디어 파일에 삽입되거나 데이터베이스에 저장되는 검증 가능한 메타데이터를 가리키는 이용자용 명칭
2	창작자 어설션 작업반 (Creator Assertions Working Group, CAWG)	탈중앙 신원 재단(Decentralized Identity Foundation, DIF) 산하 조직으로, C2PA 규격 위에 적용되는 추가 규격을 개발

음향 식별 기술과 딥페이크 시대의 콘텐츠 신뢰

뉴스 브리프

AI가 만들어낸 음성이 실제 목소리와 구분하기 어려운 수준에 이르면서, 콘텐츠의 진위를 가려내려는 기술이 주목받고 있다. 일본에서 출시된 사인드사운드는 그 흐름을 보여주는 서비스다. 소리에 식별 정보를 심는 워터마크와 소리의 특징을 추출해 대조하는 핑거프린트를 결합해, 콘텐츠를 미리 등록해 두고 나중에 원본과 맞대어 진위를 확인한다. 여기에 AI가 생성한 음성인지를 판정하는 기능을 더해, 조작이나 사칭에 대응한다. 다만 심어 둔 정보는 재가공 과정에서 손상될 수 있고, 생성 기술이 정교해질수록 판정도 계속 갱신해야 하는 과제가 남는다. 본 보고서는 사인드사운드가 어떤 원리로 콘텐츠를 등록하고 진위를 가려내는지 그 작동 방식을 살펴보고, 기술적 보완과 과제를 함께 전망한다.

뉴스 플러스

I. 서론: 디지털 콘텐츠 진본성 확보의 과제

• 음성 딥페이크 확산과 진위 판별의 한계

AI를 활용해 사람의 목소리를 그대로 복제해 만들어내는 음성 기술이 빠르게 발전하면서, 실제와 구분하기 어려운 음성이 온라인 곳곳에서 나타나고 있다. 실제 인물이 직접 말한 음성과 AI가 생성한 음성을 사람의 귀만으로 구분하는 일은 이제 점점 더 어려워지고 있는 상황이다. 유명인의 얼굴과 목소리를 흉내낸 딥페이크(deepfake)* 영상이나 공공기관을 사칭한 음성이 온라인 공간에 빠르게 퍼지면서, 무엇이 실제로 촬영되고 녹음된 원본이고 무엇이 조작된 것인지 일반 이용자가 스스로 판단하기 힘든 사례가 계속 이어지고 있다. 실제와 생성물의 경계가 흐려지는 이러한 현상은 개인의 명예 훼손을 넘어 사회 전반의 정보 신뢰와 공적 소통의 기반을 흔드는 문제로 이어진다는 점에서 그 사회적 파장이 점차 커지고 있다.

이러한 흐름에 대응해 콘텐츠의 진위와 출처를 제3자가 대신 확인하고 증명해 주는 서비스가 등장하고 있다. 일본의 음향통신기술 기업인 에빅사(Evixar)와 전자결재 서비스로 사업을 넓혀 온 기업인 샤치하타(Shachihata)가 함께 선보인 음성·영상 콘텐츠 증명 서비스 사인드사운드(SIGNED SOUND)가 대표적인 사례다. 다만 조작 여부를 사후에 가려내는 판별 방식은 생성 기술이 정교해질수록 정확도가 떨어지고, 새로운 생성 방식이 나올 때마다 판별 기술이 그 뒤를 다시 쫓아가야 하는 한계를 안고 있어 대응이 언제나 늦어진다는 지적을 받는다. 이는 조작된 콘텐츠를 사후에 적발하는 접근만으로는 빠르게 발전하는 생성 기술 앞에서 콘텐츠의 신뢰를 지키기 어렵다는 점을 의미한다.

* 딥페이크(deepfake): 인공지능을 이용해 실제 인물의 얼굴이나 목소리를 합성해 만들어낸 생성물을 가리킴. 실제로 말하거나 행동하지 않은 내용을 실제처럼 보이거나 들리게 만든 영상과 음성을 뜻함

• 음향 식별 기술의 등장 배경과 사인드사운드의 출현

조작된 콘텐츠를 사후에 판별하는 방식이 한계를 보이면서, 콘텐츠가 만들어지는 시점에 미리 식별 정보를 남겨 두려는 사전적 접근이 대안으로 주목받고 있다. 이 접근의 중심에 음향 식별 기술이 있다. 소리에 사람이 알아채지 못하는 정보를 심어 두는 음향 워터마크(audio watermark)와, 소리 자체의 특징을 뽑아내 원본과 대조하는 음향 핑거프린트(audio fingerprint)가 대표적인 방식이다. 두 기술은 원래 소리로 기기를 제어하고 콘텐츠를 인식하는 분야에서 쓰여 왔으나, 최근에는 콘텐츠의 진위와 출처를 증명하는 수단으로 그 쓰임이 넓어지고 있다.

이러한 전환의 배경에는 국제 사회의 정책 흐름도 자리한다. 주요 7개국 정상회의(Group of Seven Summit)에서 2023년 합의한 히로시마 AI 프로세스(Hiroshima AI Process)*는 워터마크와 콘텐츠 인증 기술의 개발과 보급을 명시했고, 콘텐츠의 출처와 이력을 검증하기 위한 국제 표준을 마련하려는 움직임도 이어지고 있다. 에빅사의 음향 식별 기술은 일본 총무성(総務省)이 허위정보와 오정보에 대응하는 기술을 개발하고 실증하는 사업에 연속으로 채택되며 국가 차원의 연구개발 과정에서 검증을 거쳐 왔다. 이렇게 검증된 기술을 실제 서비스로 구현한 것이 에빅사와 샤치하타가 함께 선보인 사인드사운드다. 뒤에서는 이 서비스가 어떤 원리로 콘텐츠를 등록하고 진위를 가려내는지 그 작동 방식을 살펴본다.

* 히로시마 AI 프로세스(Hiroshima AI Process): 2023년 G7이 일본의 주도로 합의한 국제 규범 논의로, 발전된 AI 시스템을 안전하고 신뢰할 수 있게 개발·활용하기 위한 국제 행동 규범을 담음. 여기에는 워터마크와 콘텐츠 인증 기술의 개발·보급을 권고하는 내용이 포함

II. 본론: 사인드사운드의 작동 원리와 콘텐츠 증명

• 소리로 진위를 가리는 두 가지 기본 원리

사인드사운드가 콘텐츠의 진위를 가리는 방식은 음향 워터마크와 음향 핑거프린트라는 두 가지 원리 위에 서 있다. 이 가운데 음향 워터마크는 음성 신호에 암호화한 문자 정보 등을 사람이 알아채지 못하게 심어 두는 기술이다. 소리에 식별 정보를 직접 넣어 두었다가 나중에 그 정보를 다시 꺼내면, 누가 언제 만든 콘텐츠인지를 소리 안에서 바로 확인할 수 있다. 에빅사의 워터마크는 매체를 거쳐도 정보가 유지되는 성질과 심은 정보가 겉으로 드러나지 않는 성질, 그리고 잔향과 잡음이 있는 환경에서의 내성과 음질 저하 성능이 검증되어 있어 다양한 용도에 맞춰 쓰인다.

다른 하나인 음향 핑거프린트는 소리를 신호처리로 분석해, 서로 다른 내용을 구별할 수 있는 최소한의 정보를 뽑아내는 기술이다. 이렇게 뽑아낸 정보를 특징량(特徵量)이라 한다. 특징량을 부호화해 데이터베이스에 두었다가 나중에 들어온 소리를 같은 방식으로 부호화해 맞대어 보면 같은 콘텐츠인지 인식할 수 있으며, 소리 전체가 아니라 구별에 필요한 부분만 남기므로 포맷이나 코덱에 관계없이 특징량을 만들어 빠르게 대조할 수 있다. 게다가 부호화된 데이터는 원래 소리로 되돌릴 수 없고 콘텐츠의 의미 자체를 인식하지도 않아, 부호화된 데이터를 다루는 과정에서 원음이 그대로 노출될 위험을 줄여 준다.

두 원리는 콘텐츠를 다루는 접근 방향이 서로 다르다. 워터마크는 콘텐츠에 정보를 심어 두었다가 그 정보를 꺼내 확인하는 반면, 핑거프린트는 콘텐츠에 손을 대지 않고도 특징량을 대조해 같은 소리인지 알아낸다. 사인드사운드는 이 두 원리를 하나로 묶어, 콘텐츠를 등록하는 단계에서 워터마크를 심는 동시에 핑거프린트를 데이터베이스에 등록해 두고, 이후 진위를 확인할 때 두 방식을 겹쳐 써서 어느 한쪽이 손상되어도 다른 쪽이 판정을 뒷받침하도록 한다.

• 사인드사운드의 등록과 조회 작동 방식

사인드사운드의 작동은 등록과 조회라는 두 단계로 나뉜다. 먼저 등록 단계에서는 배포하려는 정규 콘텐츠에 이력의 표식을 남기는데, 음성이나 동영상 파일에 음향 워터마크로 이력 정보를 심고 동시에 그 음성의 특징량을 뽑아 핑거프린트로 데이터베이스에 등록해 둔다. 이렇게 하면 배포한 쪽이 자신의 정규 콘텐츠에 표식을 남겨, 시청자나 이차 확산처가 언제든 등록해 둔 원본과 맞대어 볼 수 있는 상태가 된다.

이 등록 처리는 실제 운용을 견딜 만큼 빠르게 이뤄진다. 콘텐츠 관리 화면에서 동영상을 올리면 음성에 워터마크가 심기는 동시에 핑거프린트 데이터가 만들어져 데이터베이스에 등록되며, 한 시간 분량의 동영상도 몇 분 정도면 끝난다. 처리가 오래 걸리면 일상적인 업무에 부담이 되어 도입이 어려워지므로, 이러한 속도는 서비스가 실제 현장에서 쓰이기 위한 조건이 된다.

조회 단계에서는 의심스러운 콘텐츠가 나타났을 때 한 가지 방법에만 의존하지 않고 서로 다른 방법으로 거듭 확인한다. 사인드사운드는 워터마크를 검출하고 핑거프린트 데이터베이스와 대조하며 AI가 생성한 음성인지를 판정하는 확인을 함께 수행하는데, 워터마크가 검출되고 원본과 일치하면 정규 콘텐츠로 볼 근거가 되고 표식이 없거나 대조에 실패하면 원본으로 보기 어렵다는 실마리가 된다. 많은 파일을 대조하는 조건 하에서 일치한 콘텐츠의 파일 이름과 시간 정보를 찾는다면, 제출한 영상과 원본 파일을 비교할 수 있게 된다.

이렇게 여러 방법을 함께 쓰는 이유는 한 방법이 뚫리거나 손상되어도 다른 방법이 이를 메우도록 하기 위해서다. 음원이 재가공되며 워터마크가 훼손되더라도 핑거프린트 대조가 판정을 뒷받침할 수 있고, 등록되지 않은 콘텐츠라도 AI가 생성한 음성인지를 판정해 진위를 가리는 실마리를 얻을 수 있다. 이렇게 사인드사운드는 무엇이 가짜인지를 사후에 쫓기보다, 무엇이 원본인지를 미리 표식으로 남겨 두고 그 표식의 유무와 일치 여부로 진위를 가리는 방식으로 작동한다.

• AI 생성 음성 판정과 콘텐츠 보호 기능

사인드사운드에는 등록된 원본과 대조하는 방식과 별개로, 음성 자체를 분석해 AI가 생성한 음성인지를 판정하는 기능이 들어 있다. 이 기능은 등록되지 않은 콘텐츠에도 진위를 가릴 실마리를 준다는 점에서 워터마크나 핑거프린트의 부족한 부분을 채우는데, 주목할 점은 이 판정 기술 자체가 AI에 기대지 않는 방식이라는 데 있다. 덕분에 계산 부담이 가벼워, 고성능 연산 장치가 없는 저가 스마트폰에서도 단말 자체에서 판정을 처리할 수 있다.

판정 성능은 여러 생성 모델을 대상으로 검증되었다. 당초 목표로 삼은 다섯 개 모델을 크게 웃도는 25개의 음성 생성 AI 모델을 대상으로 판정 시험을 실시했고, 특정 계열을 제외한 24종에서 양호한 판정 정확도를 얻었다. 또한 최종 시험 소재에서 지난해 60% 정도이던 판정 성공률이 올해 92%로 올라, 진위를 가려내는 능력이 크게 개선되었다. 다만 생성 기술이 정교해질수록 판정도 계속 갱신해야 하므로, 이러한 성능 향상은 앞으로 다양한 생성 음성에 대응하기 위한 바탕이 되는 동시에 지속적인 갱신이라는 과제도 함께 남긴다.

이러한 판정과 대조 기능은 콘텐츠 보호로 이어진다. 음원의 추적 가능성을 확보하고 부정한 이용을 감지하면, 콘텐츠를 만든 쪽의 권리를 지키는 근거를 마련할 수 있기 때문이다. 실제로 라디오 콘텐츠에 이 기술을 적용하는 방안을 검토하는 과정에서, 방송된 콘텐츠의 진위를 담보하고 개변이나 부정 이용을 추적할 수 있게 함으로써 보도와 정보 발신의 신뢰성을 높이는 방향이 제시되었다. 이렇게 사인드사운드는 진위를 가리는 데 그치지 않고, 콘텐츠가 부정하게 쓰이는 것을 막는 보호 수단으로도 기능한다.

III. 결론: 신뢰 가능한 정보 유통의 과제

• 사전적 콘텐츠 보호로의 전환과 남은 과제

사인드사운드가 보여주는 것은 콘텐츠의 진위를 다루는 방식이 사후 판별에서 사전 증명으로 옮겨가고 있다는 점이다. 조작을 뒤늦게 적발하는 대신 원본에 미리 표식을 남겨 두고 그 표식으로 진위를 가리는 것으로, 결국 이 방식이 겨냥하는 것은 무엇이 가짜인지 일일이 쫓는 부담을 덜고 무엇이 원본인지를 증명해 신뢰의 기준을 세우는 일이다. 이는 콘텐츠를 지키는 무게중심이 사후 대응에서 사전 대비로 이동하고 있음을 의미한다. 이러한 점에서 음향 식별 기술은 딥페이크와 AI 생성 음성이 늘어나는 환경에서 콘텐츠의 신뢰를 떠받치는 바탕이 된다.

다만 이 방식이 자리 잡으려면 풀어야 할 과제도 남아 있다. 심어 둔 표식은 재가공 과정에서 손상될 수 있고, 생성 기술이 정교해질수록 판정 기능도 계속 갱신해야 하므로, 기술의 내성과 판정 능력을 함께 끌어올리는 노력이 이어져야 한다. 또한 원본을 미리 등록하는 방식이 힘을 발휘하려면 콘텐츠를 만드는 쪽이 표식을 남기는 일이 자연스러운 관행으로 퍼져야 하고, 이를 뒷받침할 제도적 기반과 표준도 함께 마련되어야 한다. 따라서 앞으로는 기술적 보완과 더불어 이러한 증명 방식을 사회가 함께 받아들이고 운용하는 틀을 갖추는 일이 과제로 남으며, 이를 통해 콘텐츠의 진위를 둘러싼 신뢰가 개별 기술을 넘어 사회적 기반으로 자리 잡을 수 있을 것으로 보인다.

참고문헌

- Evixar, "シヤチハタ株式会社と共同で、音声・動画コンテンツの第三者証明サービス「SIGNED SOUND（音のしるし）」を正式提供開始", 2026.07.07., https://www.evixar.com/news/20260707_01/
- Evixar, "ディープフェイクの悪用防止へ エヴィクサー、シヤチハタが提携", Yahoo News Japan, 2026.03.16., https://news.yahoo.co.jp/articles/29c15401b57b71bbec256046a36b420b8bb34518
- Evixar, "音響透かしと音響フィンガープリントを用いた偽・誤情報対策クラウドシステムの開発・実証 成果報告書", 일본 총무성(総務省), 2026.03.19., https://www.soumu.go.jp/main_content/001068725.pdf

기술용어

순번	용어	설명
1	딥페이크 (deepfake)	인공지능을 이용해 실제 인물의 얼굴이나 목소리를 합성해 만들어낸 산출물을 가리킴. 실제로 말하거나 행동하지 않은 내용을 실제처럼 보이거나 들리게 만든 영상과 음성을 뜻함
2	히로시마 AI 프로세스 (Hiroshima AI Process)	2023년 G7이 일본의 주도로 합의한 국제 규범 논의로, 발전된 AI 시스템을 안전하고 신뢰할 수 있게 개발·활용하기 위한 국제 행동 규범을 담음. 여기에는 워터마크와 콘텐츠 인증 기술의 개발·보급을 권고하는 내용이 포함