AI시대의 보증서, 디지털 워터마크

권오현 계간 스켑틱 편집자

예나 지금이나 인간 삶에서 중요한 가치 중 하나는 신뢰이다. 어떤 사람이나 사물이 믿을 만한가 그렇지 않은가를 가려내는 것은 우리의 생존과 직결되는 문제였다.

워터마크의 기원

워터마크는 인간의 생산물에 신뢰성을 부여하는 한 가지 방법으로서 그 물건에 조작이 불가능한, 누가 생산했는지를 표시하는 특별한 이미지를 삽입하는 것이다. 그 역사는 매우 오래돼 1292년 이탈리아 파브리아노 지역에서 워터마크가 삽입된 종이가 발견될 정도다.

제지 공장이 모여 유럽 최대의 종이 생산 지역이었던 파브리아노 지방의 종이는 그 질도 최고급이었는데, 그만큼 다른 지역에서 만든 가짜도 많았다. 이에 제지공들은 13세기 말부터 젖은 종이를 거르는 그물망에 수를 놓듯이 그림이나 글씨를 새기고 그 위에 종이를 얇게 올려 특정 부분을 반투명하게 만들었다. 오늘날 우리가 위조지폐를 감별하듯 빛에 비추면 그 부분이 드러나는 것이다. 파브리아노의 제지 공장은 워터마크 기술을 통해 종이에 이름, 형태, 품질 같은 다양한 저작권 정보를 삽입했다.

그 후 워터마크 생성 방식은 발전을 거듭해 굵고 가는 와이어로 덮인 롤러 방식인 댄디 롤, 롤 자체 표면에 부조 영역을 만들어 음영을 표현하는 실린더 몰드 방식으로 이어졌다. 또한 워터마크는 곧 종이뿐만 아니라 지폐, 우표, 여권 등에도 사용돼 제품의 원본성과 제작자의 신원을 보증하고 저작권으로 보호하는 신뢰의 증표로서 널리 사용되었다.

워터마크, 디지털로 확장되다

인터넷의 등장과 디지털 콘텐츠의 폭발적인 성장으

로 그림, 동영상, 오디오, 텍스트 등 각종 디지털 콘텐츠의 저작권 보호 필요성이 대두됐다. 이에 워터마크 역시 '디지털 워터마크'라는 개념으로 변모했다.

디지털 워터마크를 그 용도에 따라 분류하자면 크게 강성(robust), 연성(fragile), 핑거 프린팅(finger printing), 스테가노그래피(steganography) 네 가지로 나눌 수 있다. 강성 워터마크는 '강건하다'라는 영어 단어 뜻에서도 알 수 있듯 원본 워터마크를 변조하려는 외부 시도가 있을 때 데이터의 품질이 심각하게 훼손되기 전에는 워터마크가 깨지지 않도록 설계한다. 즉 콘텐츠를 쓸모없게 만들지 않고서는 워터마크를 깰 수 없도록 강건하기 때문에 저작권을 보호하고 신뢰성을 담지할 수 있다. 이런 강성 워터마크 기법을 활용하는 곳이 각종 민원 서류를 포함한 다양한 증명서의 온라인 발급과 인터넷 서명이다.

연성 워터마크는 병원에서 찍은 환자의 임상 관련 이미지나 정부 문서, 기밀 문서의 원본 텍스트와 같이 파일의 외부 유출을 막아야 하는 경우에 사용한다. 연성 워터마크는 데이터에 변형을 가하면 그 즉시 워터마크가 깨지면서 원본 콘텐츠도 동시에 훼손된다.

핑거프린팅은 이름처럼 '지문'이나 '바코드'와 유사한 개념으로 고유번호나 식별자를 콘텐츠에 삽입하는 것이다. 이 기술을 적용하면 누구에게 어떤 경로로 콘텐츠가 배포됐는지 알 수 있으므로 제품이나 콘텐츠를 분류하고 그 전송 경로를 확인할 수 있다. 또한 불법적인 유통이 발생할 경우 배포자를 추적할 수 있다.

스테가노그래피는 '감추어져 있다'는 뜻의 그리스어인 'stegano'와 '쓰다, 그리다'라는 뜻의 'graphos'가 결합된 용어로, 사진, 그림, 동영상 같은 일반적인 파일 안에 데이터를 숨기거나 다른 형태로 위장해 주고받는 일종의 '암호통신' 기술에 해당한다. 동영상, 사진, 오디오의 디지털 파일의 노이즈(noise)를 다른 메시지로 대체해도 사람이 인지할 수 없다는 사실을 이용한다. 이처럼 디지털 워터마크는 인지하기 어려워 일상적인 사용을 방해하지 않는다는 특성을 갖는다.

생성형 인공지능 시대, 디지털 워터마크는 어디로 가는가?

최근 생성형 인공지능 서비스가 우리 일상 속에 침투해 활용되면서 디지털 워터마크의 중요성은 더욱 커지고 있다. 인공지능으로 만든 딥페이크 사진과 영상이 정치적, 사회적, 문화적 악영향을 끼치며 신뢰의 가치를 잠식하고 있기 때문이다.

이에 구글은 2023년 인공지능 생성 이미지에 보이지 않는 워터마크를 삽입하는 '신스ID(SynthID)'라는 도구를 공개했다. 신스ID는 인공지능만 인식하는 픽셀 단위 흔적을 남겨 이미지를 구별한다. Chat-GPT를 만든 오픈 AI 역시 자사의 그림 생성 인공지능 '달리3'로 만든 이미지에 '콘텐츠 출처와 진위 확인을 위한 연합(C2PA)'의 워터마크를 부착한다. C2PA는 마이크로소프트, 어도비, 인텔 등의 기업이 주도하는 개방형 기술 표준으로 이미지의 메타데이터 형태로 존재하며 시각적으로는 볼 수 없다.

우리 정부를 포함한 미국, 유럽은 제도적으로 인공지능 생성물에 디지털 워터마크를 넣는 입법안을 마련하고 있다. 과학기술정보통신부는 2023년 10월 제4차 인공지능 최고위 전략대화에서 인공지능 기반 서비스 가이드라인을 마련하고 자율적인 검인증 제도를 도입하기로 했다.

하지만 해결해야 할 문제도 있다. 워터마크가 있다는 말은 이를 우회하거나 깨려는 누군가가 있다는 뜻이다. 따라서 보안 분야와 마찬가지로, 워터마크도 창과 방패의 싸움이 끝없이 이어진다. AI의 등장으로 정교한 디지털 워터마크도 쉽사리 파훼될 수 있는 현실에 대처하려면 전세계적인 연구와 노력이 필요해 보인다.