AI 학습을 위한 저작물 이용과 저작권 문제

| 김경숙 | 저작권보호심의위원회 심의위원장_상명대학교 인문콘텐츠학부 지적재산권전공 교수 |

1. 들어가며

인공지능 기술은 현재 다양한 분야, 특히 문화와 예술 창작에 많은 영향을 미치고 있다. 이러한 현상은 인공지능이 생성한 작품에 대한 저작권 보호 범위 및 인공지능 학습을 위한 저작물 이용에 대한 저작권 침해 문제에 대한 논의를 촉발시켰다.

인공지능은 인간과의 협업이 요구되는 인공지능과 그렇지 않은 생성형 AI(generative AI)으로 분류될 수 있다. 전자는 자동번역, 음악 작곡용 소프트웨어 도구, CAD를 이용한 엔지니어링이나 건축설계 등과 같이 특정한 결과를 달성하기 위한 자동화된 프로세스를 포함한다. 반면에 생성형 AI는 특정한 결과를 미리 정하지 않고 독립적으로 작동하여 결과를 도출한다. 이는 인간이 창작하는 방식과 매우 유사하다.

생성형 AI는 자체 학습 능력을 활용하여, 새로운 소설을 작성하거나 멜로디를 작곡하는 등 인간이 만들어 낼 수 있는 창작물을 생성할 수 있다. 그러나 이 과정에서는 대량의 데이터를 필요로 하며, 이 데이터에는 저작권이 있는 저작물이 포함될 수 있다. 이로 인해 저작권자의 동의 없이 이러한 저작물이 사용되는 경우가 있어, 많은 저작권자들이 이에 대한 우려를 표현하고 있다. 실제로, 이러한 문제로 인해 해외에서는 몇몇 소송이 진행되고 있다.

이러한 문제들을 예측하여 생성형 AI가 이용하는 저작물에 대한 면책 규정에 대한 논의가 진행되고 있다. 유럽과 일본은 이를 법률로 도입한 반면, 미국은 기존의 공정 이용 조항을 활용하여 이 문제를 해결하려는 방향을 보이고 있다.

우리나라도 생성형 AI와 관련하여 AI 학습 데이터에 사용되는 저작물의 원활한 이용 방안을 모색하고 있는 중이다. 이에 해외에서의 AI 관련된 소송 사건과 면책규정들을 통하여 TDM(Text and Data Mining) 면책규정 도입의 방향을 제시하고자 한다.

2. AI 학습을 위한 저작물 이용

(1) AI 학습 과정

생성형 AI 모델을 학습시키기 위하여 AI는 인터넷의 책, 웹사이트, 기사 및 다양한 문학 자료에서 대량의 데이터를 수집하고 광범위한 훈련을 거치면서 알고리즘(모델)에 대량의 데이터를 공급한다. 이때 인터넷에서 자료를 끌어와 봇에게 '공급'하는 것에는 복제와 송신이 수반된다. 그뿐만 아니라, 학습과정에서 저작권자들이 부착한 권리관리정보들이 제거·변경·변조될 수도 있다.

예컨대, 이미지 생성형 AI 모델인 경우는 이렇게 입력된 데이터를 바탕으로 이미지 등을 학습한다. AI 모델로 이미지를 학습하여 생성하는 과정을 살펴보면 대부분의 이미지 생성 솔루션은 '잠재 확산(latent diffusion)'이라는 기술을 사용한다. 이것은 노이즈(noise)를 사용하는 알고리즘을 생성하는 방법이다. 먼저, 소프트웨어는 원본 이미지를 노이즈로 바꾼다. 즉, 픽셀의 무작위 패턴(예: 화이트 노이즈 또는 눈)으로 변환하는 것이다. 그런 다음 컴퓨터는 수많은 시행착오를 거치며 원본 이미지를 점차 재현하려고 시도하면서, 거의 완벽할 때까지 학습을 한다.

이 같은 학습 과정 동안 프로그램은 주로 특정 물체에 대해 특정 색상의 픽셀이 서로 어디에 위치하는지에 대한 통계적 관계를 찾으려고 시도한다. 현재의 컴퓨터로는 이 과정을 수십억 개의 이미지에 대해 자동으로 수행할 수 있다. 스테빌러티 AI(Stability AI)사의 최신 시스템 중 하나인 스테이블디퓨전(Stable Diffusion)은 인터넷에서 수집한 50억 개의 그래픽으로 구성된 LAION-5B 데이터베이스를 사용한다.

중요한 것은 스테빌러티 AI 모델(Stability AI model)이 LAION-5B의 이미지들을 바탕으로 훈련되는 동안 이 이미지들은 스테빌러티 AI 모델에 저장되어 복제되는 것은 아니다. 대신 스테빌러티 AI 모델은 예를 들어 고양이 이미지 학습을 위해 LAION-5B의 무수히 많은 고양이 이미지 사이의 유사성을 분석하고, 이러한 이미지의 패턴이나 구조적 유사성에 대한 정보를 저장하여 고양이에 대한 기준에 부합하는 전혀 새로운 이미지를 식별하거나 생성해낸다. 이러한 점은 우리 인간들이 이전에 본 적 있는 모든 고양이를 완벽하게 사진과 같이 기억하고 있지는 않지만, 몇 마리를 본 후 고양이들의 일반적인 특징을 기억하여 추후 하나의 고양이 이미지를 다른 이미지와 구분할 수 있는 것과 같다. 즉, Stability AI 모델의 학습과 이미지 생성과정은 인간의 기억 과정과 유사하다.

상기 사례에서처럼 학습 과정에서 이미지 생성형 AI 모델은 원본 이미지를 노이즈로 바꾸기 때문에 원본인 저작물을 이용하는 것이 되며 비록 원본을 복제하여 저장하지 않는다 하더라도 실질적 유사성 판단에서 의식 또는 무의식의 접근으로 볼 수 있으므로 원저작물에 접근하였다고 볼 수 있다. 이점은 저작권 침해 판단에서 중요한 요소이기도 하다.

(2) 생성형 AI 관련 사례

1) UAB Planner 5D v. Facebook, Inc.[1)]

2019년, 리투아니아 회사인 UAB Planner 5D는 Facebook, Inc., Facebook Technologies, LLC 및 프린스턴 대학교의 이사회를 상대로 미국 캘리포니아 북부 지방 법원에 저작권 및 영업 비밀을 침해하였다고 주장하며 소송을 제기했다.

Planner 5D는 테이블, 의자 및 소파와 같은 가상 데이터들의 데이터셋을 운용하며 이를 바탕으로 가상 인테리어 디자인을 생성할 수 있는 홈 디자인 웹사이트를 운영하면서 이에 대한 저작권을 주장하였다. Planner 5D는 자사가 보유한 저작물들을 프린스턴 대학의 컴퓨터 과학자들이 허락 없이 전체 데이터를 다운로드했다고 주장하였다. 그리고 링크를 Facebook에 게시함으로써 공중이 쉽게 접근 가능하도록 한 것에 대해 Facebook도 책임이 있다고 주장하였다. 이에 대하여, 프린스턴 대학 측은 연구 목적으로 Planner 5D 데이터를 이용했을 뿐이라고 항변하였다.

2020년 7월, Planner 5D가 저작물이라고 주장하는 가상 데이터들이 저작권 보호 대상인지 증명하지 못하여 저작권 침해에 대한 소는 기각되었다.

2) Thomson Reuters Enter. Ctr. GmbH v. ROSS Intelligence Inc.[2)]

2020년 5월, 원고인 Thomson Reuters Enterprise Centre GmbH(톰슨 로이터)와 West Publishing Corporation(West)은 ROSS Intelligence Inc.(ROSS)을 미국 델라웨어 지방 법원에서 저작권 침해로 소를 제기했다.

원고들은 법률 산업 전반에서 널리 사용되는 법률 검색 플랫폼인 웨스트로(Westlaw)를 운영하고 마케팅하는 회사이다. 그리고 ROSS는 인공지능을 활용한 새로운 법률 검색 플랫폼을 개발하였으며, ROSS의 검색 도구를 개선하기 위하여 LegalEase Solutions, LLC와 협력 관계를 맺었다.

원고측은 LegalEase가 "봇을 사용하여 [원고의] 독점 정보를 대량으로 다운로드하고 저장한 후 ROSS에 제공했다"라며 저작권 침해를 주장하였다.

이에 대해 피고인 ROSS는 공정이용의 각 요소에 대하여, (1) Westlaw 콘텐츠의 이용은 기능적이고 변형적 이용이었고, (2) 복제된 Westlaw 자료에 대한 저작권 보호는 좁으며(thin), (3) ROSS 플랫폼에는 최종적으로 저작권 자료가 포함되어 있지 않기 때문에 이용된 양은 중요하지 않으며, (4) ROSS의 생성물들은 Westlaw의 시장을 대체하지 않았다고 항변하며 공정이용을 주장하였다.

그러나, 델라웨어 지방 법원은 피고의 이용행위는 공정이용에 해당하지 않으며 저작권 침해를 구성한다고 판단하였다.

---

1) UAB "Planner 5D" v. Facebook, Inc., 534 F. Supp. 3d 1126 (N.D. Cal. 2021)
2) Thomson Reuters Enter. Ctr. GmbH v. ROSS Intelligence Inc., 529 F. Supp. 3d 303 (D. Del. 2021)

3) Doe 1 v. GitHub Inc., N.D. Cal.[3)]

2022년 11월 3일, 미국 캘리포니아 북부 지방법원에 익명의 프로그래머 그룹이 마이크로소프트, GitHub(마이크로소프트 자회사), 그리고 OpenAI를 상대로 DMCA 제1202조를 위반한 것으로 소를 제기하였다. 이 조항은 침해를 유도하거나 은폐하기 위해 거짓 저작권 관리 정보(CMI)를 제공하거나 배포하는 것을 불법으로 규정하고 있다.

원고 측은 마이크로소프트와 GitHub이 오픈소스 라이선스 조건을 준수하지 않고 원고들의 자료를 사용하여, 원고들의 저작권이 있는 코드를 불법적으로 복제하고 라이선스에 따른 각종 표시 요건을 위반하여 Codex와 Copilot를 개발했다고 주장한다. 두 시스템 모두 공개적으로 접근 가능한 소프트웨어 코드와 다른 자료들, 그리고 원고들이 만들었다고 주장하는 침해 소프트웨어 코드를 포함한 대량의 자료로 학습한 보조 AI 기반 시스템으로서 소프트웨어 프로그래머에게 제공되고 있다.

2023년 1월, 마이크로소프트와 OpenAI는 이 사건에서 기각 이유를 주장하는 이유서를 제출했다. 이유서에서 원고들은 자신들의 이용행위로 인해 특정 손해를 입었다는 입증을 하지 못하여 이 사건 소송이 부적합하며, 또한 원고 측은 피고가 남용하였다고 주장하는 저작물과 계약을 위반하였다는 부분을 특정하지 못하였다고 주장하였다.

이 사건은 현재 진행 중이다.

4) Sarah Andersen v. Midjouney[4)]

2023년 1월 13일, 수상 경력이 있는 시각 예술가인 사라 앤더슨(Sarah Andersen), 켈리 맥커난(Kelly McKernan), 칼라 오르티즈(Karla Ortiz) 및 그 외 아티스트들이 미국 캘리포니아 북부 지방 법원 샌프란시스코 지부에 스타빌리티 AI Ltd., 스타빌리티 AI, Inc., 미드저니(Midjourney), Inc., 디비언트아트(DeviantArt), Inc.를 상대로 집단 소송을 제기했다.

원고들은 소장에 Stable Diffusion (Stability AI), DreamStudio (Stability 제작), Midjourney Product (Midjourney), DreamUp (DeviantArt) 등이 자신들의 작품들을 동의 없이 스크랩하여 AI 알고리즘을 학습하는 데 사용하였다고 주장하였다.

원고들이 주장한 피고의 법률 위반행위는 미국 저작권법 17 U.S.C. 제106조의 직접 침해, 제501조에 따른 대위 침해(Vicarious Copyright Infringement), 제1201조에서 제1205조에 따른 DMCA 위반(DMCA Violations), 캘리포니아 민법 제3344조에 따른 퍼블리시티권 침해(Right of Publicity Violations), 그리고 부정경쟁방지법(Unfair Competition law, Cal. Bus. & Prof.) 제17200조이다.

원고들은 "피고들의 인공지능 시스템은 원고들의 작품을 포함한 수십억 개의 저작권이 있는 이미지를 허가 없이 다운로드하거나 획득하여 정교한 '학습 이미지(Training Images)'를 만들었다. 이러한 이미지를 기반으로 AI 모델을 학습시켜 이 이미지들을 인공지능 시스템 내에 압축된 복사본으로 저장한다. 이 작업은 아티스트들의 동의 없이 이루어졌으며, 그들에게 어떠한 보상도 제공하지 않았다. 이용자의 요청에 기반하여 생성되는 이미지는 전적으로 학습 이미지에 기반하고 있으며, 특정 출력물을 생성할 때 인공지능 모델이 사용하는 특정 이미지의 파생 작품이다. 그 결과, AI 이미지 모델은 '저작권이 보호된 트레이

3) DOE 1 et al v. GitHub, Inc. et al 4:2022cv06823
4) Sarah Andersen v. Midjouney 3:23-cv-00201

닝 이미지의 압축된 복사본'을 보유하고 이를 '재결합'하여 생성하는 '21세기 콜라주 도구'로서의 기능에 불과하다. 피고는 인공지능으로 생성된 이미지 제품을 허락없이 이용함으로써, 상당한 상업적 이익을 얻었다"라고 주장한다.

현재 이 사건 소송도 진행 중이다.

5) Getty Images vs Stability AI[5)]

Getty Images는 미국과 영국에서 두 건의 평행한 사건을 제기했다.

2023년 2월 초, Getty Images는 이미지 생성기 'Stable Diffusion AI'의 개발자인 Stability AI를 상대로 미국 델라웨어 지방 법원에 소송을 제기했다. Getty Images는 Stability AI가 저작권이 있는 사진을 침해하고, 저작권 관리 정보(CMI)를 제거하거나 변경하고, 부정확한 저작권 관리 정보를 제공하며, 상표를 침해했다는 주장을 제기했다. Getty Images는 Stability AI가 명시적으로 사용을 금지한 조건에도 불구하고 그의 웹 사이트에서 사진을 복사하여 1200만 장 이상의 이미지와 관련 메타데이터를 'Stable Diffusion' 학습에 사용했다고 주장한다.

저작권 관리 정보(CMI) 관련 주장에 대해 Getty Images는 Stable Diffusion이 생성한 결과물에서 종종 Getty Images의 변형된 워터마크를 발견할 수 있으므로, "Stability AI가 무단으로 복사한 저작권 있는 이미지와 그 모델이 생성하는 결과물 간의 명백한 관련성을 강조한다"라고 주장한다.

(3) 소결

상기 사례들을 통하여 파악할 수 있는 AI 저작권문제는 직접 침해, AI 플랫폼의 대위책임, 그리고 권리관리정보의 변경 또는 제거에 따른 DMCA 책임등이다. 이에 대하여 피고 측은 공정이용을 통하여 저작권침해에 대한 항변을 하고 있다.

AI 학습을 위한 저작물 이용에서는 먼저 저작물인지를 판단하여야 하는데 UAB Planner 5D v. Facebook, Inc. 사건에서와 같이 이용된 원고의 데이터가 저작물인지 여부가 불명확한 경우에는 저작권 침해 판단을 하기 어려울 것이다. 반면 저작물을 상업적으로 사용하였음이 명백한 경우에는 Thomson Reuters Enter. Ctr. GmbH v. ROSS Intelligence Inc. 사건에서와 같이 공정이용이 부정되어 저작권침해가 될 수도 있음을 보여주고 있다.

또한 관할권도 문제 될 수 있는데 AI 모델의 훈련이 EU 관할 지역 내에서 이루어졌다면, 디지털 싱글 마켓 지침(CDSM)의 저작권 제4조에 따른 텍스트와 데이터 마이닝에 대한 새로운 예외 조항에 따라 허용되었을 수 있다. Getty Images 사건의 경우는 미국과 영국에서 함께 소송이 진행 중이므로 관할에 따라 달리 판단될 가능성도 크다.

최근 생성형 AI와 관련된 여러 사건들이 법원에서 다투어지고 있는 상황이므로 이들에 대한 판결 결과에 따라 추후 생성형 AI에 대한 저작권 분쟁에 큰 영향을 미칠 것으로 보인다.

5) Getty Images v Stability AI 1:23-cv-00135-UNA

3. TDM 면책규정

학습 단계나 TDM 과정에서는 학습 데이터를 생성하고, 이를 인공지능 학습용 소프트웨어에 입력하는 과정에서 데이터의 복제가 발생한다. 인공지능의 기계학습(머신러닝)을 위해서는 타인의 저작물을 이용하는 것은 불가피하다. 특히 인공지능 학습은 전처리(pre-processing) 된 데이터의 대부분을 학습 데이터로 활용하는 경우가 많아, 데이터를 적법하게 이용할 권한이 없다면 광범위한 저작권 침해가 발생할 수 있다.

원칙적으로, 인공지능 개발자는 저작권자의 허락을 받아 해당 저작물을 학습 데이터로 이용하는 것이 가장 바람직하지만, 데이터 마이닝 같은 기술을 사용할 때 개별 작품에 대한 저작권자의 허락을 얻는 것은 쉽지 않다. 또한 작품의 저작자가 불명확하거나 원본 소스가 알려져 있지 않은 경우에는 저작권자의 허락을 얻는다는 것은 거의 불가능한 일이다.

설사, 개별적으로 저작권자로부터 이용허락을 받을 수 있다 하더라도, 그러한 허락을 얻기 위한 노력과 비용이 데이터 마이닝에서 추출된 정보의 가치를 초과한다면, 그러한 이용허락은 큰 의미가 없다. 이 때문에 데이터 마이닝을 통해 가치 있는 정보를 추출하는 데 적극적인 노력을 포기할 가능성이 있다.

인공지능 분야의 발전을 위한 학습 데이터 확보를 위하여, 머신러닝으로 제공되는 저작물에 대해서는 저작권 제한이 필요하다고 보아 EU와 일본과 같은 해외 국가에서는 저작권 개정을 통해 데이터 마이닝 활동에서의 저작권 예외 및 제한을 규정하기도 하였다. 또한, 별도의 규정을 둔 것은 아니지만 미국은 데이터 마이닝을 공정 이용으로 판단하고 있다. 한국은 TDM 면책규정에 관한 법안이 상정 중이다. 해외 TDM 면책 사유들을 통하여 현재 개정법안의 TDM 면책규정에 대한 타당성을 이하에서 살펴본다.

(1) 미국

미국 저작권법에는 데이터 마이닝에 대한 구체적인 예외 조항이 명시되어 있지 않다. 대신, 디지털 기술과 관련된 일련의 사례에서 연방 법원은 공정 이용(fair use)이 허용되어 컴퓨팅 분석과 디지털 아카이브 생성을 가능하게 하고 검색 서비스를 제공할 수 있다고 판단한 바 있다. 따라서, 저작물의 이용이 공정한 경우, 이용자는 저작권자의 허락이나 동의를 받을 필요가 없다.[6)]

공정한 이용(fair use) 여부는 저작권 목적에 맞추어 제107조에 규정한 4가지 요소들을 고려하여 사안에 따라 결정된다.[7)] 이하 컴퓨터 분석을 위한 이용행위에 관하여 공정이용 여부가 다투어진 사례들을 바탕으로 하여 각 요소별로 공정이용의 판단 근거를 살펴본다.

첫 번째 요소인 TDM에 이용되는 저작물 이용의 목적 및 성격과 관련하여 법원은 TDM과정에서 검색 가능한 데이터베이스나 검색 엔진의 생성은 매우 변형적이어서 공정이용이 될 가능성이 크다고 판단하고 있다. TDM관련 변형적 이용으로 판단한 사례로는 다음과 같다.

6) FireSabre Consulting LLC v. Sheehy, No. 11-cv-4719 (CS), 2013 WL 5420977 (S.D.N.Y. Sept. 2013).
7) Arrow Productions, Ltd., v. The Weinstein Company, .Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 549, 105 S.Ct. 2218, 85 L.Ed.2d 588 (1985), Bill Graham Archives v. Dorling Kindersley Limited).

① Authors Guild v. HathiTrust 사건[8]에서 제2순회구는 전체 텍스트 검색 가능 데이터베이스의 작성은 본질적으로 변형적인 이용이고 단어 검색의 결과 문자, 표현, 의미, 그리고 그 메시지가 변형적으로 이용되었다고 판단하였다.

② Fox News Network, LLC v. TVEyes 사건[9]에서는 TVEyes가 텔레비전과 라디오 방송의 전체 내용을 녹화하고, 클로즈드 캡션과 텍스트 음성 변환 기술을 사용하여, 콘텐츠 검색이 가능한 데이터베이스를 만드는 과정에서 프로그램 스크립트의 일부가 포함되었더라도 변형적 이용에 해당되어 비록 상업적 목적으로 이용되었더라도 공정이용에 해당한다고 판단하였다.

③ Authors Guild v. Google 사건[10]에서는 Google이 파트너 라이브러리 컬렉션의 책을 디지털 방식으로 스캔하여 학자와 연구자가 사용할 수 있는 검색 가능한 데이터베이스에 통합하는 과정에서 8분의 1페이지 길이의 "발췌문(snippet)"이 포함된 것에 대하여, 이 프로젝트는 "새로운 분야의 데이터 마이닝과 텍스트 마이닝을 포함한 실질적인 연구를 목적으로 책 텍스트를 데이터로 변환하여 새로운 연구 분야를 열게 했다"라고 언급하면서, 책에 나오는 단어들은 지금까지 쓰이지 않았던 방식의 변형적 이용이라고 판단하였다.

④ A.V. v. iParadigms, LLC 사건[11]에서는 iParadigms가 교사가 사이트를 통해 제출한 학생의 작품과 이전에 서비스에 제출된 논문뿐 아니라 인터넷에서 이용할 수 있는 콘텐츠 등과 비교하여 학생들의 작품이 표절되었는지 여부를 판단할 수 있도록 해주는 TurnItIn이라 불리는 데이터베이스를 구축한 사안에 대해 TurnItIn 서비스의 상업적 특성에도 불구하고, 그 사용은 "매우 변형적"이라고 판단했다.

⑤ Perfect 10 v. Amazon 사건[12]에서는 Google이 검색 엔진에 저작권으로 보호된 이미지의 "썸네일"을 통하여 이용자가 원고 웹사이트에 있는 전체 이미지로 접속할 수 있도록 "인라인 링크"를 한 것에 대하여 제9순회구는 "전자 참조 도구"로서의 목적이 매우 변형적이라고 판단하였다.

⑥ Field v. Google 사건[13]에서 Google이 저자의 원래 웹 콘텐츠의 복사본을 웹사이트 캐시에 제공하였고, 캐시된 링크가 웹 비교 또는 검색어 식별을 위해 보관용 등 여러 가지 이유로 이용된 것에 대해 변형이용으로 판단하였다.

⑦ Kelly v. Arriba Soft 사건[14]에서는 검색 엔진 회사인 Arriba Soft 회사가 사진작가의 웹사이트에 호스팅된 이미지에 썸네일로 인라인링크하면서, Arriba Soft의 검색 엔진이 인터넷 상의 이미지에 대한 액세스를 색인화하여 접근을 용이하게 하는 도구로 사용한 것에 대하여 변형적 이용으로 판단하였다.

⑧ White v. West 사건[15]에서는 Westlaw와 LexisNexis 두 출판사가 소송 자료들을 포함한 법적 파일들을 데이터베이스로 복사하여, 복사된 법률 파일에 메타데이터를 추가해 대화형 법률 연구 도구를 만든 것에 대하여, 그 작업규모가 크고 검색 결과에는 법률 자료 전문이 포함되었다 하더라도 변형적 이용에 해당한다고 판단하였다.[16]

---

8) Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014).
9) Fox News Network, LLC v. TVEyes, Inc., 43 F. Supp. 3d(S.D.N.Y.2014).
10) Authors Guild v. Google, 770 F.Supp.2d 666 (S.D.N.Y. 2011).
11) A.V. ex rel. Vanderhye v. iParadigms, L.L.C.,562 F.3d 630 (4th Cir. 2009).
12) Perfect 10 v. Amazon, 508 F.3d 1146 (9th Cir. 2007).
13) Field v. Google, 412 F.Supp.2d 1106 (D. Nv. 2006).
14) Kelly v. Arriba Soft, 336 F.3d 811 (9th Cir. 2003).
15) White v. West Pub. Corp. No. 12-cv-1340-JSR (S.D.N.Y. Jul. 3, 2014).
16) See Roxana Robinson, "How Google Stole the Work of Millions of Authors", http://www.schoolinfosystem.org/2016/02/24/how-google-stole-the-work-of-millions-of-authors/).

두 번째 요소인 저작물의 성질과 관련하여 법원은 TDM의 공정이용을 결정한 대부분의 사건들이 일반적으로 보호를 받는 창작물들을 포함하고 있다고 보았다. 이런 경우 일반적으로 법원은 원고에게 더 유리하다고 판단해왔다. 그러나, TDM과 관련해서는 TDM 분석 결과에서 원 저작물을 감득하기 어렵기 때문에, 저작물의 성질에 관한 요소에는 큰 비중을 두지 않고 있다.

세 번째 요소인 이용된 저작물의 양과 질의 판단에 있어서는, 연구자는 데이터 및 텍스트 전체를 복사하지 않으면 연구 등에 필요한 내용을 분석하지 못하기 때문에 TDM을 효과적으로 사용하기 위해서는 전체 텍스트 또는 저작물 데이터베이스를 전체적으로 복사할 필요가 있다. 따라서, 이 점에 대해서 White v. West 사건에서 법원은 TDM을 위한 데이터베이스의 전체를 완전히 복제할 수밖에 없으므로, TDM에 관한한 이 요소는 중립적이라고 판단하였다.[17)]

네 번째 요소인 잠재적 시장에의 영향은 이차적인 이용이 원저작물의 대체물로 작용함으로써 초래되는 손해에만 초점을 맞추고,[18)] 변형적 이용으로 인한 경제적 손해는 계산하지 않는다. 변형적 이용은 정의상 원저작물의 대체물로 보지 않기 때문이다.[19)] 그리고, 네 번째 요소를 판단할 때는 첫 번째 요소와의 밀접한 연관성을 고려하여야 하는데, 원저작물의 목적과 다른 목적을 달성하기 위해 복제를 할수록 원저작물의 만족스러운 대체요소로 작용할 가능성이 적기 때문이다.[20)] 결과적으로, TDM의 매우 변형적인 특성은 저작권이 있는 저작물을 대체할 가능성이 낮기 때문에 TDM으로 인해 저작물의 원본시장에 부정적인 영향을 미칠 가능성이 낮다.[21)]

이 네 가지 요소를 종합적으로 고려할 때, 법원은 TDM이 저작물을 전부 복제하더라도 변형적인 이용에 해당되고, 원저작물의 시장에 미치는 영향이 크지 않아 시장을 대체하지 않는다고 보아, 공정이용이 될 가능성이 크다고 보고 있다.

(2) EU

CDSM[22)]은 TDM을 '디지털 형태의 텍스트와 데이터를 분석하여 패턴, 추세 및 상관 관계와 같은 정보를 생성하는 것을 목표로 하는 모든 자동화된 분석 기법'(제2조(2))으로 정의하며, 새로운 기술에 의해 가능해진 '텍스트, 소리, 이미지 또는 데이터와 같은 디지털 형태의 정보에 대한 자동화된 계산 분석'(전문 8)을 의미한다. 제2조(2)는 대량의 데이터를 자동 또는 반자동으로 분석할 수 있는 도구의 잠재력을 적절하게 식별하는 포괄적인 정의이다.[23)]

유럽 연합은 데이터베이스 지침(96/9/EC)에 따라 별도의 독자적인 권리로서 데이터베이스 권리를 가지고 있으며, 이는 데이터를 획득, 검증 또는 표시하기 위해 상당한 투자가 이루어진 데이터베이스의 콘텐츠에 적용된다.[24)] 이 독특한 데이터베이스 권리는 투자자에게 부여되며, 투자자는 회사 또는 소위 데이터베이스 제작자가 될 수 있다. 따라서 데이터셋과 저작물이 TDM 과정에서 사용될 때, 데이터베이스 제작자의 권리와 저작자의 권리가 함께 침해될 수 있다.

---

17) White v. West Pub. Corp. No. 12-cv-1340-JSR (S.D.N.Y. Jul. 3, 2014), at 6-7.
18) HathiTrust, 755 F.3d at 99 (citing Campbell v. Acuff-Rose Music, Inc., 510 U.S.569, 591 (1994)).
19) Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 614(2d Cir. 2006)).
20) Authors Guild, Inc. v. Google, Inc. (Google Books), 804 F.3d 202 (2d. Cir. 2015), at 222-223.
21) Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 90 (2d Cir. 2014). at 97 (characterizing building the search index as a "quint essentially transformative use").
22) EU DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.
23) Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.
24) Thomas Margoni & Martin Kretschmer, "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology" GRUR International, Volume 71, Issue 8, August 2022, Pages 685-701, https://doi.org/10.1093/grurint/ikac054, Published: 26 July 2022.

데이터 마이닝 과정에서 저작물과 데이터셋의 원활한 활용을 위해, EU는 CDSM에서 TDM에 대한 두 가지 예외 조항(제3조 및 제4조)을 도입하였다. 제3조는 의무조항으로서, 그 목적은 과학 연구 목적으로 TDM을 수행하기 위해 연구 기관 및 문화유산 기관에 의해 수행되는 복제 및 추출 행위를 면책하기 위한 것이다.[25] 또한, 계약보다 과학 연구를 위해 연구 및 문화 기관에서 수행되는 TDM의 면책이 우선하며, TDM을 위한 기술 보호 조치의 우회도 허용된다(제7조(2)).[26]

제4조는 제3조와 유사하지만 중요한 차이점이 있다. 제4조는 저작권이 있는 작품을 TDM에 사용할 수 있도록 허용하지만, 권리 소유자가 'opt-out'(선택적 철회) 또는 'contract-out'(계약 철회)을 명시적으로 유예할 수 있다. 다시 말해, 'opt-out' 또는 'contract-out'로 다른 자들의 TDM을 막을 수 있다.[27]

결과적으로, TDM에 의한 저작물의 이용이 권리자에 의해 'opt-out'(선택적 철회) 또는 'contract-out'(계약 철회)로 명시적으로 유예되면, 연구 목적으로 활동하는 연구 및 문화 기관이 아닌 기업, 정부, 시민들은 AI 개발을 위해 권리자로부터 구체적인 이용허락을 받아야 한다. 'opt-out' 또는 'contract-out'이 없는 경우에만, TDM이 가능하다(제4조(2)).

요약하면, 연구 및 문화 기관에 의한 과학 연구를 위한 제3조는 연구 및 문화 기관이 저작권이 있는 저작물에 접근, 복사, 추출하는 것을 방지하기 위한 기술 보호 조치(TPM) 등을 허용하지 않으며, 계약 철회 또는 선택적 철회 또한 허용하지 않는다. 반면, 연구 목적 이외의 이용에 대해서는, TPM과 계약으로 제4조(2)를 무효화하여 저작권자가 저작물 등의 이용에 대한 권리를 유보할 수 있도록 하고 있다. CDSM은 이용 목적에 따라 조항을 달리하여 규정되어 있다는 것을 특징으로 한다.

이미 영국(제29A조), 독일(제60d조), 스위스(제24d조)와 같은 유럽 국가들은 이들 CDSM 조항들을 이행하여 저작권법에 TDM 저작권 예외와 제한을 도입했다.[28] 단, 현행 영국 CDPA 제29A조의 텍스트 및 데이터 마이닝 예외는 비영리 목적의 연구만으로 제한하고 있으며, 폭넓은 텍스트 및 데이터 마이닝(TDM) 예외 규정의 도입을 계획했다가 2023년 3월 철회한 바 있다.[29]

(3) 일본 저작권법 제30조의4

일본의 TDM 저작권 예외는 2009년에 처음 도입되었고(2018년 개정 이전의 제47조의7), 2018년 개정에 따라 조문 변경과 내용의 수정이 있었다(제30조의4). 개정을 통해 '컴퓨터를 사용하여'라는 요건이 삭제되고 '어떤 방법으로든 이용'이라는 표현이 추가되었다.[30] 따라서 현행 일본 저작권법은 제30조의4에서 TDM에 대한 예외 조항을 명시적으로 규정하고 있다.

제30조의4는 저작물에 표현된 아이디어나 감정을 '향수하기' 위한 것이 아닌 경우, 사용자가 권리자의 허락 없이 저작물을 원활하게 이용할 수 있도록 한다. 제30조의4는 작품에서 표현된 아이디어나 감정을 즐기지 않기 위한 사용 유형을 다음과 같이 열거하고 있다.

---

25) Christophe Geiger, Giancarlo Frosio, Oleksandr Bulayenko, "The EXCEPTION FOR TEXT AND DATA MINING (TDM) IN THE PROPOSED DIRECTIVE ON COPYRIGHT IN THE DIGITAL SINGLE MARKET – LEGAL ASPECTS" Center for International Intellectual Property Studies Research Paper No. 2018-02

26) Ibid.

27) Maria Bottis Marinos Papadopoulos, Christos Zampakolas & Paraskevi Ganatsiou, "Text and Data Mining in the EU Acquis Communautaire Tinkering with TDM & Digital Legal Deposit" (2019) 12 Erasmus L. Rev. 190, at 196-198.

28) FireSabre Consulting LLC v. Sheehy, No. 11-cv-4719 (CS), 2013 WL 5420977, at 7 (S.D.N.Y. Sept. 2013).

29) https://www.lexology.com/library/detail.aspx?g=cb6087d4-ba48-474a-81c5-4f1f85eba01e

30) Act No 30 of 25 May 2018. See in detail Japan Copyright Office (JCO), 'Outline of the Amendments to the Copyright Act in 2018' (2019)4 Patents & Licensing 10, 12 (footnote 8).Referring to the translation see Japan Copyright Office (n 30) 11-12.

제30조의 4 (저작물에 표현된 사상 또는 감정의 향수를 목적으로 하지 않는 이용) 저작물은 다음의 경우 기타 해당 저작물에 표현된 사상 또는 감정을 스스로 향수하거나 타인에게 향수시킬 것을 목적으로 하지 않는 경우에는 그 필요하다고 인정되는 한도에서 어떠한 방법에 의하든 사용할 수 있다. **다만, 해당 저작물의 종류 및 용도, 해당 이용 양태에 비추어 저작권자의 이익을 부당하게 침해하는 경우에는 그러하지 아니하다.**

1. 저작물의 녹음, 녹화 기타 이용에 관한 기술의 개발 또는 실용화를 위한 시험용으로 제공하는 경우
2. 정보 해석 (다수의 저작물 기타 대량의 정보로부터 해당 정보를 구성하는 언어, 소리, 영상 기타 요소에 관한 정보를 추출, 비교, 분류 기타 분석을 실시하는 것을 말한다. 제47조의 5 제1항 제2호에서 같다)의 용도로 제공하는 경우
3. 제1호 및 제2호에 정한 경우 외에 저작물의 표현에 대한 사람의 지각에 의한 인식을 수반하지 않고 해당 저작물을 전자계산기에 의한 정보처리 과정에서의 이용, 기타 이용 (프로그램 저작물에 있어서는 해당 저작물의 전자계산기의 실행을 제외한다)에 제공하는 경우

제30조의4 제1항 제2호에서의 "정보 해석"이란 대량의 정보에서 그 정보를 구성하는 언어, 소리, 영상 기타 요소에 관련된 정보를 추출하여, 비교, 분류 기타 분석을 하는 것을 말한다(정보 해석의 기술에는 화상해석, 음성해석, 언어 해석, 웹 해석 등의 기술 분야가 있고, 본인 인증, 자동 번역, 사회 동향 조사, 정보 검색 등에 이용되어왔다).[31] 정보를 해석하기 위해서는 대량의 데이터(타인의 저작물을 포함한다)를 축적할 필요가 있으나, 이는 저작물 그 자체를 향유하기 위한 목적이 아니기 때문에 저작권자의 이익을 침해하지 않는다. 이에, 정보 해석의 목적으로 정보를 컴퓨터에 축적하는 행위는 저작권 침해로 보지 않는 것이다.[32]

제30조의4 본문은 정보해석이 저작권제한규정에 해당하기 위해서는 먼저 해당 저작물에 표현된 사상 또는 감정을 스스로 향수하거나 타인에게 향수시킬 것을 목적으로 하지 않는 경우에 해당되어야 하는 것으로 규정하고 있다. 또한, '어떤 방법으로든지' 이용할 수 있으므로, 복제에 한정하지 않고 번안이나 공중 송신등도 할 수 있다.[33] 번안은 정보 해석 과정에서 저작물의 구성 요소를 추출하여 통계 처리에 적합한 형태로 바꾸는 행위를 말한다. 해당 저작물의 종류 및 용도, 해당 이용의 양태에 비추어 저작권자의 이익을 부당하게 해치게 되는 경우는 인정되지 않는다. 정보 해석의 목적이라면 저작물을 이용하는 주체에 한정은 없고, 이용할 수 있는 저작물에도 한정이 없다. 이로써 사람이 아닌 컴퓨터가 이른바 데이터 마이닝이나 텍스트 마이닝 등을 하는 경우에, 분석 대상이 되는 데이터 등이 저작물에 해당한다고 하더라도 저작권 침해를 염려하지 않고 데이터 등을 수집, 기록, 분석할 수 있게 되었다.[34] 저작인접권에 대해서도 제한 규정을 마련하고 있으므로, 실연, 음반, 방송 또는 유선방송도 기록할 수 있다(일본 저작권법 제102조 제1항).

31) 文化庁著作権課 「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定に関する基本的な考え方」（令和元年10月24日), 6頁。
32) Id.
33) Id.
34) 차상육(2018), "빅데이터의 지적재산법상 보호", 법조 제67권 제2호, 법조협회, 118면.

(4) 우리나라

개정안은 컴퓨터를 이용하여 저작물이 포함된 대량의 정보를 분석하기 위하여 저작물을 복제·전송하는 경우를 저작재산권 제한 사유로 신설한다고 제안이유를 설명하고 있다.[35]

> 제43조(정보분석을 위한 복제 · 전송)
> ① 컴퓨터를 이용한 자동화 분석기술을 통해 다수의 저작물을 포함한 대량의 정보를 분석(규칙, 구조, 경향, 상관관계 등의 정보를 추출하는 것)하여 추가적인 정보 또는 가치를 생성하기 위한 것으로 저작물에 표현된 사상이나 감정을 향유하지 않는 경우에는 필요한 한도 안에서 저작물을 복제 · 전송할 수 있다. 다만, 해당 저작물에 적법하게 접근할 수 있는 경우에 한정한다.
> ② 제1항에 따라 만들어진 복제물은 정보분석을 위하여 필요한 한도에서 보관할 수 있다.

현행 저작권법에는 정보 분석을 위한 복제 등에 관한 직접적인 저작재산권 제한규정은 없으나, 데이터마이닝을 위한 일련의 저작물이용 행위에는 포괄적인 저작재산권 제한 규정인 '공정이용 조항'이 적용될 여지가 있다. 그러나 이는 저작자의 이익을 부당하게 해치지 않는 범위 내에서 저작물의 공정한 이용이 가능하다고 포괄적으로 규정하고 있을 뿐이어서 데이터 분석이 해당 규정에 따라 면책되는지 여부가 불확실한바, 개정안은 ① 대량의 정보를 분석하여 추가적인 정보 또는 가치를 생성하기 위한 목적으로, ② 저작물에 표현된 사상이나 감정을 향유하지 않고, ③ 이용하고자 하는 저작물에 적법하게 접근할 수 있는 경우, 저작물의 복제·전송을 명시적으로 허용하여 관련 사업자의 법적불확실성을 해소하고 빅데이터 산업의 발전을 도모하려는 취지이다.[36]

개정안은 저작물을 복제·전송할 수 있는 요건으로 ① 저작물에 표현된 사상이나 감정의 향유를 금지하여 인간이 분석과정에 참여하는 것을 허용하지 않고, ② 저작물에 적법하게 접근할 것을 규정하여 해킹, 불법 다운로드 등을 통한 복제는 배제하고 있다는 점에서 저작권 제한 범위를 합리적으로 설정한 것이라는 견해가 있다. 그럼에도 불구하고 일반적으로 공익성이 강한 분야에 도입되어 있는 다른 저작권 제한 규정과 달리, 정보 분석을 위한 저작권 제한 규정은 관련 산업의 발전을 직접적 목적으로 하고, 영리적 목적의 이용도 허용하고 있어 이에 대한 권리자들의 반발이 예상된다.[37]

35) 문화체육관광위원회, 「저작권법 전부개정법률안 검토보고〈창작자 권리 보호 및 기술환경 변화에 따른 저작권 제도 개선〉」, (2021.2.), 30-31면.
36) Id.
37) Id.

(5) 소결

TDM의 대상이 되는 데이터는 주로 출판물, 그 외 저작권으로 보호되는 저작물이 되는 경우가 많다. 이때 TDM 과정에서 발생하는 복제 및 전송에 대해 저작권이 제한될 것인가가 문제 된다. 이를 위해 해외에서도 TDM을 위한 저작권 제한규정이 EU를 비롯하여 미국, 일본 등에 도입되었고, 우리나라도 2020년 저작권법 개정안에 TDM면책규정을 도입하였다. 그러나, 개정안[38]의 내용에는 TDM의 허용범위를 "해당 저작물에 적법하게 접근할 수 있는 경우에 한정한다."라고 되어 있어 접근에 대한 적법성이 문제 될 것으로 보인다.

TDM의 허용범위에 대하여 EU, 미국, 일본도 각각 다른 입장을 보이고 있다. EU의 경우에는 DSM저작권지침으로 연구 및 비상업적 목적을 위하여 기술적보호조치를 우회하여 TDM을 하는 경우에도 저작권이 제한될 수 있으나, 계약으로 이를 배제할 수 있도록 하고 있다. 이에 반하여 미국의 경우에는 상업적 비상업적 목적을 불문하고 TDM의 결과의 저작물이용형태가 공정이용에 해당한다면 저작권이 제한될 수 있도록 설계하고 있다. 일본은 상업적인 목적의 경우에도 TDM이 가능하다고 하고 있으나, 기술적보호조치가 되어 있는 경우에는 이를 우회하여 이루어지는 저작물 이용행위는 저작권이 제한되지 않음을 법으로 규정하고 있다.

상기 비교법적인 관점에서 살펴보았을 때, 현재 우리나라 저작권법 개정안은 일본과 흡사하다고 볼 수 있으나, 적법한 접근이라는 점에서 차이가 있다고 할 것이다.

4. 결론

인공지능(AI)의 급격한 발전과 확산은 우리 사회의 많은 분야에서 변화를 가져오며 새로운 저작권 문제를 초래하고 있다. 현행 저작권법은 인간의 창작 활동에 초점을 두고 있지만, AI 생성물은 인간의 개입이 없다는 점에서 생성물에 대한 저작권 문제가 한편에서는 논의되고 있다. 그리고 그 전제로서 AI 학습을 위해 대량의 데이터 이용과정에서의 저작권 침해가 문제 되고 있다.

이런 문제를 해결하기 위해, TDM을 위한 면책규정들이 해외에서 도입되었고 국내에서도 개정안이 만들어진 상황이다. 그러나 개정안을 두고도 그 타당성 여부를 두고 논의가 지속적으로 이루어지고 있는 바 국내 AI 산업 발전에도 기여할 수 있는 제도의 도입이 이루어질 수 있도록 현재 진행 중인 해외 소송들과 TDM면책 규정들을 면밀히 살펴볼 필요가 있다.

38) 도종환의원 대표발의 저작권법 전부개정법률안 (의안번호 7440)

• 본 내용은 저작권보호심의위원회 심의위원의 개인적 견해로, 한국저작권보호원의 공식적인 의견이 아님을 알려드립니다.