# C STORY

it's safe

KOREA COPYRIGHT PROTECTION AGENCY | 2023. 12

**줌 인 포커스**

저작권 특별사법경찰 도입부터
저작권 범죄 과학수사대 출범까지

**저작권 시야 넓히기**

K-콘텐츠 저작권 감시대상국,
우선감시대상국, 우선협상국은?

**저작권 보호 이슈 던지기**

공공누리 제도(制度)와
제도(製圖) 사이

**새로운 창작, 저작권 확립에서 시작됩니다.**

K-콘텐츠 저작권을
지키는 구원투수로
문화매력국가 대한민국을
만들어가겠습니다.

한국저작권보호원
KOREA COPYRIGHT PROTECTION AGENCY

VOL.41

저작권 보호 전문지

C STORY

통 권 41호
발 행 일 2023년 12월
발 행 인 박정렬
발 행 처 한국저작권보호원
주 소 서울특별시 마포구 월드컵북로 400(상암동1002) 서울경제진흥원 4, 9, 10층
기획·제작 홍보협력부(편집부)
디자인·인쇄 문화공감
등록번호 마포 마-00057
전 화 1588-0190
팩 스 02-3153-2719
구독신청 홍보협력부 서영훈 (02-3153-2745)

CONTENTS

2023
DECEMBER
Vol.41

## 깊게 보는 저작권

04
**줌 인 포커스**
저작권 특별사법경찰 도입부터
저작권 범죄 과학수사대 출범까지

08
**저작권 기술 동향**
AI의 시대, 무단 크롤링과
크롤링 방지 기술

홈페이지 바로가기 블로그 바로가기

**한국저작권보호원은 저작권법 제122조의2에 의거**하여 국내외 저작권 보호기반 조성 및 저작권 분야의 국제 경쟁력 강화를 목적으로 설립되었습니다. 체계적, 효율적인 **저작권 보호 시스템구축과 국내 콘텐츠 산업 발전을 위해** 유관기관과 유기적으로 협력하며 최선의 노력을 다하고 있습니다. **C STORY는 저작권 보호 의식 제고를 위해 한국저작권보호원에서 무료로 배포하고 있으며, 홈페이지를 통해서도 보실 수 있습니다.**

* 본지에 실린 글의 내용은 한국저작권보호원의 의견과 다를 수 있습니다.


14
**해외 저작권 보호 동향**
해외 저작권 보호 전문정보가 '쏙쏙'

18
**저작권법으로 세상 읽기**
저작권법상 정보제공청구와
정보통신망법상 정보제공청구제도의
비교


## 우리 곁의 저작권


24
**저작권 변천사**
1986년 저작권법 11
: 보호기간 변천사

28
**저작권 시야 넓히기**
K-콘텐츠 저작권 감시대상국,
우선감시대상국, 우선협상국은?
저작권 침해 종합대응시스템
구축과 저작권 보호 방안

32
**C STORY가 만난 사람**
이제 저작권 보호를 넘어 저작권
진흥으로 발전해가야 합니다
정홍택 한국저작권단체연합회
(前)이사장

36
**이야기로 보는 심의사례**
소프트웨어의 '크랙' 파일을
제공 중인 게시물

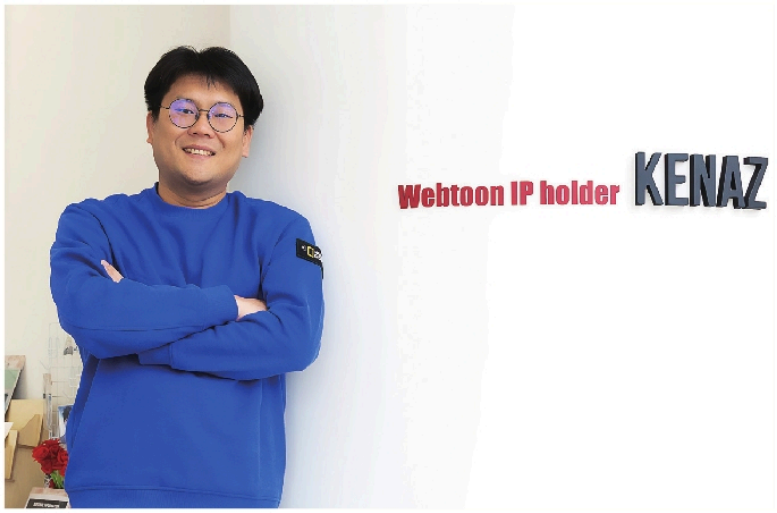
38
**저작권 보호 이슈 던지기**
공공누리 제도(制度)와
제도(製圖) 사이




46
**저작권 보호, 혼자가 아니야**
❶ 케나즈의 힘, 지식재산권의
소중한 가치를 나누다
박영준㈜케나즈 부대표
❷ 저작권 보호 법률 컨설팅 사례
온라인 쇼핑몰 운영과
저작권 관련 유의사항(2)

53
**생생저작권 보호 현장**
❶ 2023 저작권 보호 유관기관 열린포럼
❷ 2023 한국 · 필리핀 저작권 포럼 등 개최
❸ 제11회 SW산업보호대상 시상식 개최

56
**단신뉴스**

줌 인 포커스

# 저작권 특별사법경찰 도입부터 저작권 범죄 과학수사대 출범까지

**필자는 어느새 15년이란 시간을 저작권특별사법경찰을 업으로 삼고 있다. 10년이면 강산이 변한다고 이 글을 쓰기 위해 회고해 보니 저작권특별사법경찰 환경과 위상에 대한 많은 변화가 일어났다는 것을 새삼 느끼게 된다.**

**글** 김찬 문화체육관광부 저작권보호과

## 특사경의 도입(배경)

2008년 저작권법과 컴퓨터프로그램보호법의 통합으로 당시 정보통신부의 특별사법경찰이 문체부에 이체되어 저작권특별사법경찰(이하 '저작권경찰')이 발족하게 되었다. 모든 정부 부처에서 정책의 실효성과 전문성을 겸비한 특별사법경찰제도를 도입하던 시기였다. 한편, 콘텐츠, 방송, 통신서비스 등 IT를 통한 디지털컨버전스가 한창일 무렵 SW에 한정된 수사 권한이 모든 저작권(콘텐츠) 범죄로 확대 개편된 것이다.

2008년 9월, 상암동 문화콘텐츠센터에서 개최된 '저작권경찰 발대식'을 통해 저작권경찰의 대외적인 활동이 시작되었다. 이날 '불법저작물의 유통은 저작권자의 재산권 행사에 막대한 지장을 초래하고, 창작의지를 감소시켜 결과적으로 문화산업 후퇴로 이어질 수밖에 없다'는 유인촌 장관님의 연설과 함께 저작권경찰은 창작자 보호의 첨병이 될 것임을 천명하는 자리였다.

당시 SW 고소 사건 일색이던 특사경에게 저작권 수사는 미지의 영역이었다. 오프라인을 통한 신학기 출판물, DVD, CD 등 길거리 단속과 온라인 음원 전송으로 뜨겁게 달구던 소리바다 사건을 비롯한 그간의 저작권 수사를 되짚어 글을 이어가고자 한다.

## 활동과 성과(웹하드 → 토렌트 → 웹툰 → 링크, TVPAD → 스트리밍 등)

저작권경찰 도입 초기의 핫이슈는 단연 웹하드였다. 모든 디지털콘텐츠 이용의 저변 확산에 일면 일조하였다고 주장하는 이들도 있었으나, 웹하드 등록제 시행 이전에 음원, 영상, 심지어 성인물까지 모든 저작물 불법 유통의 메카였다. 저작권경찰의 첫 수사 대상은 자연스레 웹하드로 향하게 되었다.

2010년 웹하드 업체 첫 영장 집행은 10여 년이 지난 지금도 너무 생생하다. 채증과정에 엄청나게 큰 스토리지(저장용량)와 서버의 테이블 구조를 파악하고, 분석하다 보면 밤을 꼬박 새우기는 다반사이고 수십만 개의 저작물 목록을 작성하는데도 하루가 꼬박 걸리곤 했기 때문이다. 이 과정은 앞으로 이어질 저작권경찰 수사의 험난함을 예고

하는 시발점이 아니었나 생각된다.

이때 웹하드에서 단 몇 개월 운영으로 10억 이상의 수익을 취하고 있음을 확인했을 때 권리자의 아픔도 공감되면서 저작권경찰의 자부심으로 쾌감도 느낄 수 있었다. 이 첫 사건으로 범죄수익도 환수되었고, 당시 사건을 지휘하였던 중앙지검에 저작권경찰의 존재감을 심어준 계기가 되었다. 다음 해 검찰과의 합동(기획)수사가 대대적으로 이루어져 웹하드에 자정의 바람도 일으키는 성과도 이루게 되었다.

또한 웹하드 등록제, 필터링 기술을 이용한 기술적보호조치 도입, 범죄수익환수 대상범죄 법령개정에도 일조하였다고 자부한다. 모든 사례를 나열할 수는 없지만 이후 저작권경찰은 토렌트, 게임사설서버, 링크, TVPAD 등 최초라는 수식어를 달고 수사 사례를 만들어 가면서 업무를 수행하게 되었기 때문이다.

특히, 저작권경찰이 2015년 기획수사를 통해 이끌어낸 '링크행위의 방조 유죄 대법원 판결(2021.9.9., 대법원 2017도19025 저작권법위반방조)'은 저작권경찰의 빛나는 성과 중 하나다. 최근 누누티비 등 현재까지 횡행하고 있는 '불법 스트리밍(실시간재생) 링크사이트'를 예견한 듯, 이 대법 판례는 링크(사이트) 행위의 법리 다툼을 일단락 시켰다. 현재까지 법령개정을 통해 링크행위의 위법성을 입법하려는 노력에 앞서 모든 불법 OTT형 서비스 수사에 대응 가능한 판례를 이끌어 냈기 때문이다. (당시 수사담당자 전진덕 사무관(현, 한국예술종합학교 입학관리과장)님의 노고에도 다시 한번 감사의 마음을 전한다.)

이후 불법 토렌트 사이트, 웹하드 모바일 서비스, 게임 사설서버, 링크사이트, 출판 등 해마다 사이버 불법 유통의 온상지를 기획수사 대상을 선정하여 저작권범죄 수사에 박차를 가하였다. 이는 국내 서버를 이용한 온라인 불법저작물 유포행위가 사라지게 된 원인이기도 하다.

## 한류 열풍과 함께 온라인 불법 유통의 판도 변화

이후 2018년 경찰청과 합동수사를 통해 불법 웹툰 사이트 밤토끼 운영자가 검거되는 등 활기를 띠었으나, 그 이후 한류 확산과 K-콘텐츠의 인기가 더해지며 저작권 침해행위가 국경을 넘나들며 발생되었고 업계와 창작자의 수사요구 또한 거세지고 있었다.

온라인 침해 불법저작물 유포행위는 계속 진화하여 국제화, 지능화되어 수사의 한계 극복을 위한 노력에도 불구하고 눈앞에 보이는 범죄행각에도 대부분의 사건들이 장기화되어 답보 상태에 머물게 된다. 이런 상황을 타개하기 위해 제3국의 수사기관과 국제공조가 더욱 절실한 상황이었다. 궁즉통이라 했던가! 2020년부터 국제공조를 위해 경찰청, 인터폴, 학계, 국회 등 많은 만남과 노력에 보답하듯 인터폴과의 협업체계 구축이 현실화되었다.

2021년 4월, 문체부-국제형사경찰기구(인터폴)간 국제수사 협력 '온라인 저작권 침해대응(I-SOP: INTERPOL - Stop Online Piracy)' 팀을 구성하였고, 2022년 3월에는 인터폴 공조를 통해 해외(모로코)에서 국내 웹툰을 번역, 유포하는 사이트 운영자를 현지 수사기관과 공조를 통한 피의자 검거, 사이트 폐쇄 등 국제공조수사를 통한 첫

성과를 이루었다.

인터폴과 I-SOP이 구성되지 않았다면 국내에서 최초유포자로 확인된 10여 명의 외국인 노동자를 끝으로 반쪽짜리 수사로 마무리될 사건이었다.

이 사건(스카이망가 사건)은 성공리에 종결되었으나 몇 가지 중요 과제를 던져주었다.

1. **공조의 신속성** - 사건 종결까지 1년(불법사이트 평균 수명 1년 이내)
2. **상대국에 따른 유연성** - 국가별 법, 제도, 사법시스템 이해 부족
3. **국제공조 대응조직 부재** - 인터폴, 국제형사사법공조, 범죄인인도 등 국제공조의 다채널 활용·운영 전담(지원) 조직 부재 등이다.

일부는 해결이 된듯하지만 앞으로 다가올 많은 수사를 통해서 저작권경찰이 풀어나가야 할 숙제이기도 하다.

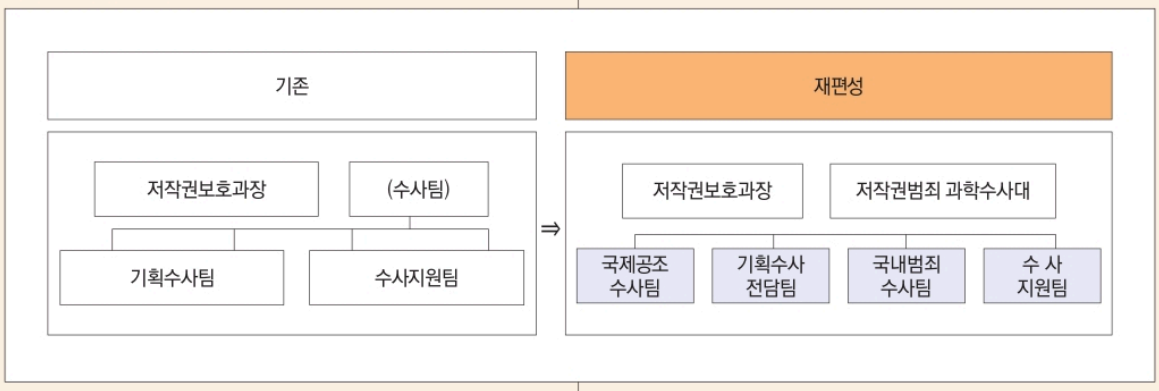
## 저작권범죄 과학수사대 출범

**배경(국내수사의 한계 극복)**

저작권경찰 출범 이후 대부분이 온라인 저작권 범죄와의 전쟁이었다. 그 이유는 온라인 불법행위가 국제화, 지능화, 조직화 되어 해외에서 불법서버 구축, VPN(접속우회 은닉), 클라우드플레어(보안서버), CDN서버, 가상화폐, 차명계좌 이용 등 수사를 회피할 수 있는 모든 환경과 접목되었기 때문이다. 특히 사회적으로 이슈를 불러일으켰던 사건은 이들 침해기술을 활용한 대규모 콘텐츠 불법유통사이트인 '누누티비'의 등장이었다. 업계 추산으로 5조원의 피해를 줄 만큼 심각했던 상황이었다.

이를 타개하기 위해 특정 국가나 부서에 한정된 수사력 한계에 대한 논의와 더불어 국가간, 부서간, 민관간에 밀착된 공조체계가 필요하다는 여러 의견들이 모아졌고 결국 대

저작권범죄과학수사대




통령 지시사항에 따라 2023년 3월 6개 부처(법무부, 과기정통부, 방통위, 경찰청, 외교부 등)가 참여한 범부처 대응 협의체가 결성되었다. 수개월 동안 치열한 고민과 논의 끝에 'K-콘텐츠 불법유통 근절 종합대책'을 마련하여 지난 8월 말 발표하였다. 주요 골자는 국제적 채널과 범부처, 민간역량을 총동원하여 K-콘텐츠 불법유통을 근절하기 위한 4대 SCSC전략 즉, '속도와 엄정함(Speed & Strict)', '공조(Cooperation)', '과학(Science)'. '변화(Change)'로 모니터링부터 수사, 처벌에 이르기까지 촘촘한 세부 과제들을 담았다.

그 대책의 일환으로 기존 저작권경찰 조직을 국제공조 및 포렌식 강화를 위해 저작권범죄분석실을 개소하고 기능별 '국제공조수사팀, 기획수사팀, 국내수사팀, 지원팀' 등 4개로 개편하였다.

각각의 업무체계는 명칭 그대로 국내수사, 기획수사, 국제공조수사, 수사지원 업무로 구분하고 전문성을 갖추고 체계적인 수사를 위한 역할 분담에 따른 것이다.

과학수사대 출범 이후 인도네시아 현지 수사기관과의 공조를 통해 한류콘텐츠를 불법유통한 IP TV 서비스 운영자를 검거하고 서비스를 종료시켰다.

저작권경찰의 수사역량과 K-콘텐츠 보호를 위한 국제공조 강화의 필요성을 증명해준 수사 성과다.

또한 저작권경찰이 다루는 모든 사건을 체계적으로 관리할 수 있도록 저작권경찰 전용 수사지원시스템(CCIS)을 구축하여 시범운영 중으로 2024년부터 전면 도입을 앞두고 있다.

이글을 기고하면서 많은 일들, 함께한 동료들이 스쳐 지나간다. 지난 일은 추억이 되고 감사하게 된다는 말이 맞나보다. 15년이 훌쩍 지난 오늘, 그 시작을 같이 한 장관님을 두 번 모시게 된 것이 우연이 아니란 생각이다. 2024년 저작권경찰에게 새로운 변화의 동력을 불어넣어 줄 것을 기대하며 더욱 발전된 저작권경찰의 미래를 소망해 본다.

콘텐츠가 문화가 되고 산업이 되고 국력이 된다. 콘텐츠의 시작은 창작자, 곧 저작권의 보호에서 비롯된다.

"저작권 보호, 바로 지금!!" 그 일선에 저작권경찰이 앞장서고자 한다.

# AI의 시대, 무단 크롤링과 크롤링 방지 기술

글 정순한 에너지경제신문 디지털콘텐츠국 국장

최근 소개된 Chat Gpt에 대한 세상의 관심이 뜨겁다. Chat GPT를 이용하면 수백 페이지의 문서를 학습하여 해당 문서의 내용의 요약정리가 가능하며, 대화가 가능하고 관련된 이미지 생성 또한 가능하다. 사람이 정리하고 이해한 후에 대화하는 지능을 이제는 인공지능이 가지게 되었다. 사람이 생각하고 실행하는 행위를 기계가 학습하고 실행하는 시대에서 사람이 하는 행위와 기계가 하는 행위의 적법한 경계를 정하기가 어려워지고 있다.

예전에 기독교단체에서 주관하는 성경필사전시를 본 적이 있다. 많은 분들이 성경을 직접 붓글씨로 펜으로 연필로 여러 도구를 이용해 성경을 모두 그대로 적어놓은 작품을 보았을 때만 해도 복사나 옮긴다라는 문구에 그리 신경이 쓰이지 않았다. 그러나 디지털의 시대에서는 크롤링을 통한 콘텐츠 복제와 저렴한 유통비용으로 인해 디지털 저작물의 무한 크롤링과 불법 사용이 일상다반사의 일이 되어 가고 있다. 저렴한 저장 공간과 높아지는 CPU의 성능 그리고 빠른 네트워크는 이러한 환경을 더욱 부추기고 있다.

## 1. 크롤링, 문제의 시작

사람이 수동으로 복사하여 이용하는 저작물과 프로그램을 이용하여 자동으로 대량의 저작물을 복사하여 사용하는 것이 프로그램을 모르는 일반 사용자들도 가능한 최근의 상황에서 우리는 자동복사 프로그램을 이용한 복사와 이용에 대한 주의를 가져 볼 필요가 있다. 프로그램을 이용한 대량복사에 사용되는 프로그램은 크롤러 또는 스크래퍼라고 부른다.

일반적으로 크롤링을 하는 크롤러는 사이트의 구조를 파악하고 링크를 쫓아 인덱스를 만드는 작업이라고 할 수 있고 스크래퍼는 의도한 특정 데이터를 정하고 그 데이터를 스크래핑하는 프로그램을 이야기한다. 이런 스크래핑을 이용해서 대량으로 특정 사이트의 데이터를 복사하여 이용하는 것은 어떤 문제를 가지고 있을까?

본 글에서는 스크래핑과 크롤링을 일반적 이해가 쉬운 크롤링으로 통합 지칭한다. 크롤링은 어떻게 정의해야 할까? 사람이 마우스를 이용하여 Ctrl+c와 Ctrl+v로 옮기는 행위와 최근에 쉽게 프로그램으로 복사와 저장을 하는 크롤링은 어떤 차이가 있는 것일까? 크롤링은 정해진 웹페이지를 자동적으로 프로그램이 정하는 방법으로 다운로드 받는 프로그램이다. 하나 이상의 URL에서 다운로드하고 그와 관련된 연결 하이퍼 링크를 추출한 후 하이퍼링크를 따라 웹페이지를 계속하여 다운로드한다. 처음 지정한 URL을 시작으로 최초 웹페이지와 연결되어 있는 하이퍼링크를 쫓아 모든 웹페이지를 다운로드하는 것이다. 웹페이지에 존재하는 하이퍼 링크를 찾아 연결된 또 다른 하이퍼링크의 웹페이지 다운로드하는 과정을 반복적으로 수행하면서 웹페이지에 있는 콘텐츠를 수집한다.

최초 크롤링은 검색엔진 또는 웹페이지에 있는 콘텐츠나 데이터의 수집을 기반으로 인덱스를 만들고 서비스를 강화하는 수단으로 시작되었다. 온라인서비스가 점점 방대해지고 디지털로 콘텐츠를 생산하고 저장하는 방식이 일반화 되면서 온라인 서비스는 점점 거대해져 갔다. 손으로 직접 집필했던 글은 키보드를 통해 입력되고 글은 디지털로 저장되고 손으로 그렸던 그림은 디지털 화면에서 생성되고 저장되고 있으며 영상 또한 필름에서 파일로 저장되면서 웹서비스는 세상의 모든 콘텐츠를 서비스하는 디지털 공간으로 확대되었다. 수많은 콘텐츠를 저장하고 서비스할 수 있도록 디지털 저장공간은 점점 더 저렴해졌고 이러한 콘텐츠를 가공하고 분석하여 서비스하기 충분할 정도로 CPU는 고도화되었으며 많은 콘텐츠의 전송을 저렴한 비용으로 전달하기에 충분할 만큼 네트워크는 빨라졌다. 하나의 웹사이트에서 서비스되는 콘텐츠의 수가 점점 더 방대해지고 이러한 웹사이트 역시 기하급수적으로 늘어나고 있다. 여기서 크롤링의 필요성이 나타나게 된다. 많은 정보를 쉽게 찾기 위해 관리자가 만든 인덱스는 짧은 시간 많은 콘텐츠가 생산되는 디지털에서 버전 관리가 어렵게 되고, 이런 문제를 해결하기 위해 자동화된 인덱스의 생성이 필요한 시점이 된 것이고 결국, 크롤링으로 수집한 데이터를 DB화하여 서비스하는 검색 사업자가 나타나게 된다.

크롤링으로 수집한 데이터는 결국 타인의 생각과 주관적 견해가 들어 있는 창작물인 경우가 많다. 사람의 창작물에는 저작권이 부여된다. 그렇다면 웹 서비스사이트에 접속한 사용자가 무료로 오픈된 콘텐츠를 무단으로 수집하여 그 정보를 이용하는 것은 저작권에서 자유로울 수 있을까? 일반적으로 누구나 서비스 받을 수 있는 오픈된 정보는 그 오픈된 취지에 크게 벗어나지 않는 정도에서 개인적 사용이 가능하다. 웹사이트를 서비스하는 주체가 자신의 웹사이트에 오픈하는 콘텐츠는 자신의 웹사이트에서 서비스되기를 바라는 서비스 주체의 취지가 있을 것이다. 아무리 오픈된 콘텐츠라 하더라고 웹사이트 서비스 주체의 취지와 관계없이 수집하여 이용하는 것은 옳지 않다. 다만 웹사이트의 정보를 크롤링하여 웹사이트의 콘텐츠 정보를 분석하고 검색서비스로 제공하여 정보를 찾는 사용자들에게 해당 사이트를 알려주는 검색의 경우는 웹사이트 서비스 주체의 서비스 취지에 크게 벗어난 행위는 아닐 것이다.

최근 대법원은 피고인들인 숙박정보제공업체의 직원이 경쟁업체의 모바일 애플리케이션 서버에 접속해 자신의 크롤링

프로그램을 통해 숙박시설 목록 등 데이터베이스를 복사한 사건에서 피고인들을 무죄로 판단한 바 있다. 그 판결은 ①서비스 제공자가 네트워크에 대한 접근권한을 제한하는지 여부는 보호조치나 이용약관 등의 대상에 의하여 결정되어야 하며, ②데이터베이스의 상당 부분은 양과 질 모두를 기초로 판단하여야 한다고 하였다. 이 같은 법리에 기초하여 정보통신망 침입, 데이터베이스 도용으로 인한 저작권법 위반, 업무방해 혐의 모두에 대해서는 무죄를 선고하였다. 이것은 크롤링을 통한 데이터 수집에 대한 최초의 대법원 판결이다.[1]

웹사이트의 서비스 정책을 위반하지 않았고 정보통신망 침입, 데이터베이스 도용으로 인한 저작권법 위반, 형법상 업무방해에 대해 구체적인 위반사항이 없다면 크롤링 자체를 위법으로 단정 짓기는 어렵다.

## 2. 크롤링, 기술의 접근

최근 크롤링을 할 수 있는 기술은 더 간단해지고 강력해지고 있다. 로그인을 우회하는 기술 그리고 몇 번의 클릭을 통해 원하는 콘텐츠에 접근하는 기술 등 오픈된 콘텐츠를 볼 수 있다면 크롤링은 어렵지 않다. 최근에 오픈된 Chat Gpt를 이용하면 누구나 어렵지 않게 원하는 사이트의 콘텐츠를 크롤링할 수 있다.

간단한 예로 Chat Gpt에서 2번의 프롬프트로 크롤링에 필요한 간단한 지식과 코드를 얻어 보았다. 처음의 프롬프트는 "귀하는 파이썬 개발자이며 웹 크롤링을 전문으로 합니다. 특정 사이트에서 콘텐츠를 크롤링하고 싶습니다. 방법을 쉽고 간단하게 설명해 주세요." 라는 프롬프트 입력으로 beautifulsoup4의 설치와 파이썬 스크립트 작성, 웹페이지 가져오기, BeautifulSoup으로 HTML 구문 분석, 추출 정보, 스크립트 실행까지의 간단한 코드와 설명을 얻을 수 있었고 더불어 웹사이트의 규칙을 위반하지 않도록 웹사이트의 robots.txt 파일과 서비스 약관을 확인해 달라는 친절한 설명도 얻을 수 있었다. 두 번째의 프롬프트에는 "위의 파이썬 라이브러리를 이용하여 https://www.chosun.com/ 사이트의 콘텐츠를 크롤링하고 하는 코드를 작성해 주세요." 라고 입력했고 정확하진 않지만 간단한 코드를 얻을 수 있었다. (그림 1)

**그림 1. Chat Gpt를 이용한 파이썬 크롤링 코드**

python — Copy code

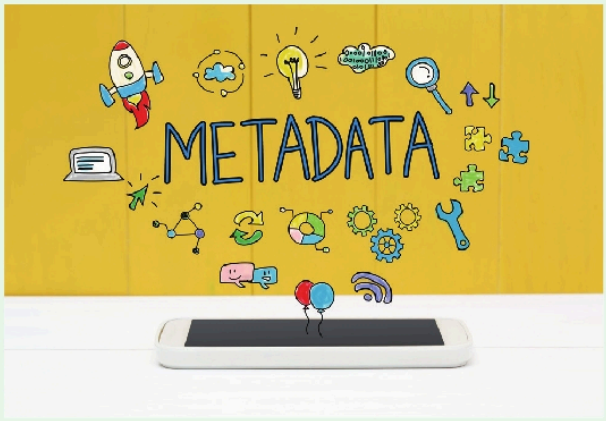
```python
import requests
from bs4 import BeautifulSoup

def crawl_chosun():
    # URL of the website to crawl
    url = "https://www.chosun.com/"

    # Send a GET request to the URL
    response = requests.get(url)

    # Check if the request was successful (status code 200)
    if response.status_code == 200:
        # Get the HTML content of the page
        html_content = response.content

        # Parse the HTML content with BeautifulSoup
        soup = BeautifulSoup(html_content, 'html.parser')

        # Example: Extract and print the titles of articles
        article_titles = soup.find_all('h3', class_='tit-news')
        for title in article_titles:
            print(title.text.strip())

    else:
        print(f"Failed to fetch the page. Status code: {response.status

if __name__ == "__main__":
    crawl_chosun()
```

1) 김태균 서울대학교 융합과학기술대학원 '웹 크롤링(Web Crawling)에 관한 연구 – 그 원리와 법적 책임에 관하여' 국문초록 참조

설명으로 'tit-news 클래스가 있는 <h3> 태그 내에 있다고 가정하고 조선일보 웹사이트의 헤드라인을 타겟팅했습니다. 그러나 관심 있는 특정 콘텐츠를 추출하려면 웹사이트의 HTML 구조를 검사하고 이에 따라 코드를 조정해야 할 수도 있습니다.'라는 설명을 덧붙여 수정이 필요한 부분을 알려주고 '웹사이트의 서비스 약관을 준수하고 차단을 방지하려면 짧은 기간에 너무 많은 요청을 보내지 않도록 하세요.'라는 친절한 설명도 덧붙여 주었다. HTML의 구조를 이해한다면 그리고 파이썬을 조금이라고 알고 있다면 어렵지 않게 크롤링을 할 수 있을 것이다. Chat Gpt와 더 많은 연결을 통해 좀 더 디테일한 크롤링의 조건들을 추가할 수도 있을 것이다. 더 많은 크롤링 기술과 자동화 프로그램의 기술들이 있으나 본 글에서는 간단한 흐름만을 서술한다. 보이는 콘텐츠는 HTML이라는 웹 표준 태그를 이용하여 우리가 사용하는 웹 브라우저에 표시된다. HTML 태그에서 H3 요소가 본문 텍스트, tit-news를 클래스명이라는 가정으로 Chat Gpt는 크롤링 파이썬 코드를 알려주었다. 얻고자 하는 웹사이트의 URL 규칙을 조금만 파악하면 콘텐츠의 크롤링은 어렵지 않다.

하루 콘텐츠 생산량이 가장 많다고 할 수 있는 언론사 기사의 경우는 정제되고 완결된 구조의 콘텐츠로 활용도가 높다. 대부분의 기사는 사람이 작성하고 사실의 적시와 함께 기자의 철학과 주관적 견해도 포함된 글이 많다. 이런 글들은 많은 사람이 볼 수 있도록 대부분의 기사가 무료 오픈이기에 로그인 없이 볼 수 있어 크롤링은 취약하다.

언론사의 기사를 크롤링하여 DB화하고 서비스한다면 법적 문제는 없는 것일까? 서비스 게시자의 오픈 취지에 맞는 사용인지 아니면 다른 용도의 사용인지가 중요한 판단이 될 것이다. 법적으로 검토할 사항들도 적지 않다. 정보통신망 이용촉진 및 정보보호 등에 관한 법률상 정보통신망 침입,




저작권법상 데이터베이스제작자의 권리 침해, 형법상 컴퓨터등장애업무방해 그리고 부정경쟁방지 및 영업비밀보호에 관한 법률상 부정경쟁행위가 되는지와 독점규제 및 공정거래에 관한 법률상 시장지배적 사업자의 남용행위 또는 불공정거래행위에 해당하는지 등 논의되고 고민해야 하는 사항들은 많다.

하지만 이런 고려사항들이 많음에도 불구하고 쉬운 개발적 접근과 저렴한 저장공간 그리고 빠른 네트워크 환경은 크롤링으로 콘텐츠를 수집하는 행위를 가속화 시키고 있다. 더불어 높은 CPU나 GPU성능은 이런 크롤링 데이터의 분석을 가능하게 하여 서비스가 아닌 데이터 분석용으로 데이터를 수집하는 개인이나 회사들의 크롤링을 가속화 시키고 있다.

### 3. 크롤링, 방지할 수 있는가?

누구나 쉽게 저비용으로 비용을 들여 생산하고 서비스하는 콘텐츠를 크롤링할 수 있게 되면서 오픈된 콘텐츠의 무분별한 크롤링이 사회적 문제가 되고 있다. 하지만, 현실적으로 오픈된 콘텐츠는 크롤러를 이용한 콘텐츠의 크롤링에서 보호받기 어려운 것이 현실이다.

최근에는 LLM(Large Language Models)의 학습을 이용한

서비스가 최신 서비스로 인정받고 있으면서 데이터의 크롤링이 더욱 가속화되고 있다.

더구나 완결된 구조의 언론사 기사는 크롤링을 통한 콘텐츠 수집이 더욱 심각한 상황이다.

저작물의 경우 온라인 게시와 함께 저작권을 표시하는 것은 이제 웹사이트 서비스의 기본이 되었다. 문서에서 많이 보이는 CCL[2](Creative Common License) 표기와 공공사이트에서 많이 볼 수 있는 4가지 유형의 공공누리가[3] 있고 대부분의 언론사에는 'Copyright, 회사이름 All rights reserved. 무단 전재 및 재배포 금지'라는 문구를 사이트 하단에서 볼 수 있다. 모두 저작권을 표시한 것으로 각각의 의미가 다르다. 오픈된 서비스에 게시된 저작물이라고 해서 아무렇게나 이용할 수 없다. 상업용도만 아니면 되는 것 아닌가 하는 생각으로 저작물을 크롤링하면 안되는 이유이다. 앞에서도 서술했지만 많은 법적문제도 고려해야만 하는 사안이다.

오픈된 콘텐츠의 크롤링을 근본적으로 막기는 어려울 것이나 사이트에 저작권표시로 저작권이 있음을 공표하는 것이 필요하다. 가능하다면 크롤링에 대한 정책을 만들어 오픈하는 것도 필요하다.

웹사이트 운영자는 자신의 웹사이트에 대한 웹 크롤링을 제어할 수 있다. 이를 위해 "robot.txt"파일을 업로드하게 되면 '로봇'은 이 파일의 내용에 따라 정보를 수집하지 않거나 허용된 범위 내에서만 수집하게 된다. 이를 로봇배제표준("robots exclusion standard" 또는 "-protocol")이라 하며, 웹 크롤러에 의한 자동적인 정보 수집을 거부하거나 제어하기 위하여 마련된 규약을 의미한다(IEFT, 2020). 다만, 로봇배제표준은 구글의 주도하에 운영되는 권고안에 불과하며 공식적인 인터넷 표준은 아니므로 "robot.txt"는 웹 크롤러를 법적으로 구속할 수 없고, 이를 준수하지 않는 웹 크롤러도 자주 발견된다.[4] 하지만 대부분의 웹 크롤러는 로봇배제표준을 준수하고 있어 적용이 필요하다.

검색 봇을 방지하는 방법으로 HTML 태그를 이용한 방법도 고려해 보자. HTML 페이지에서 메타태그는 주로 <head> 태그와 </head> 태그 사이에 입력되는데 <meta name="robots" content="noindex" />같은 방식으로 head태그 사이에 넣어 name에 차단하고자 하는 로봇의 종류를 설정하고, content에는 표시를 제한하고자 하는 항목에 대한 설정을 한다. 만일 특정 자동검색 봇에 대해 차단을 원한다면 name 부분에 해당하는 봇의 이름을 사용하면 된다.

또 다른 방법으로는 웹방화벽 및 웹프로그램으로 임계치 설정도 고려해 볼 만하다. 사람이 정보를 얻기 위해 웹사이트를 서핑하는 수준을 정하고, 사이트 내의 기계적 클릭을 방지하기 위한 자동클릭 방지와, 세션의 연결을 모니터링하며 초당, 분당 이벤트 임계치를 설정하여 관리하는 방법도 필요할 것이다. 검색 서비스에서 서비스하는 웹사이트의 정보검색과 URL 링크 연결이 필요하기에 URL 규칙을 랜덤하게 변경하는 것은 권장하지 않는다. 외부에 웹사이트의 정보가 오픈되어 사용자가 웹사이트에 쉽게 접근하기를 원하는 웹서비스 사업자의 경우 내부 서비스의 URL 규칙을 자주 변경하는 것은 외부 검색 연결을 어렵게 하기 때문에 권장하지 않는다.

---

2) CCL(Creative Common License)'이란 저작권자가 자신의 창작물에 대해 몇 가지 이용 방법과 조건을 붙여 자유롭게 이용할 수 있도록 한 일종의 표준약관이자 이용 허락표시, CCL 라이센스 유형은 크게 저작자 표시, 비영리, 변경 금지, 동일 조건 변경 허락 등 네 가지가 있다.

3) 공공누리는 국가, 지방자치단체, 공공기관이 4가지 공공누리 유형마크를 통해 개방한 공공저작물 정보를 통합 제공하는 서비스이다. https://www.kogl.or.kr/info/introduce.do

4) 권세진, 이정훈, 이창무 .'데이터 경제 시대에 있어서 웹 크롤링(crawling)의 법적 인식에 관한 연구' 韓國産業保安硏究 – 第11卷 第3號 (2021)

다만 외부 오픈이 필요한 콘텐츠와 로그인한 후 접근해야 하는 유료 콘텐츠의 구분을 명확히 하고 외부에서의 웹사이트 연결 시 프로그램 개발로 권한 있는 정당한 사용자만이 정당한 콘텐츠에 접근할 수 있도록 개발하는 것이 필요하다. 더불어 콘텐츠 접근의 임계치 설정으로 크롤러를 이용한 크롤링을 방지하는 것 또한 필요하다. 설정된 임계치는 넘게 되면 CAPTCHA를 적용하여 무분별한 크롤링을 방지하는 것도 방법이 될 수 있다.

이미지 속 텍스트도 추출이 가능한 시대인 만큼 내부 정책을 정하고 정책을 위반한 콘텐츠 사용자에게 경고하여 정책을 위반하는 콘텐츠 접근을 막아야 한다.

### 4. 창과 방패, 기술이 필요하다

과거 게임을 자동으로 플레이하여 간단한 게임상의 화폐를 모으는 '게임 보탤'이나 '게임 오토 플레이' 같은 프로그램이 유행하던 시절이 있었다. 게임 개발사들은 이런 불법 프로그램을 막기 위해 게임플랫폼에 많은 감시기능을 탑재하였지만 창과 방패의 싸움에서 항상 승리한 것은 아니었다. 누군가의 노력으로 만들어진 저작물을 동의 없이 수집·이용하는 것이 불법이라는 온라인 사용자들의 의식개혁이 필요하고 더불어 오픈된 데이터라도 그 취지 맞지 않는 무단 크롤링은 좀 더 세심한 법적 조건으로 관리를 해야 함과 동시에 빠르게 발전하는 크롤링 기술에 대응하기 위한 방지 기술의 개발이 필요하다. 오픈된 크롤링 프로그램은 계속 성능이 좋아지고 있지만 이를 막을 수 있는 방지 기술은 기초적인 수준에 머물러 있어 더 많은 불법 크롤링을 만들어 내고 있다. 선의의 피해자는 저작권을 가지고 서비스하는 웹서비스 업체와 창작자들일 것이다. 창작자의 권리를 보장하고 더 많은 창작물의 저작권이 불법으로부터 보호받기 위해 창을 막을 방패의 디지털 기술이 필요하다.