

챗GPT를 활용한 기록관리 메타데이터 추출 사례연구

A Case Study on Metadata Extraction for Records Management Using ChatGPT

김민지(Minji Kim)¹, 강성희(Sunghee Kang)², 이해영(Hae-young Rieh)³

Email: kmj990317@naver.com, kangsh@mju.ac.kr, hyrieh@mju.ac.kr



1 제1저자 명지대학교 기록정보과학전문대학원 데이터기록전공 석사
2 교신저자 명지대학교 기록정보과학전문대학원 데이터기록전공 교수
3 공동저자 명지대학교 기록정보과학전문대학원 기록관리전공 교수

논문접수 2024.4.16
최초심사 2024.4.18
게재확정 2024.5.10

ORCID

Minji Kim
https://orcid.org/0009-0006-8258-5314

Sunghee Kang
https://orcid.org/0009-0005-0610-8869

Hae-young Rieh
https://orcid.org/0000-0002-6300-2543

초 록

기록관리에서 메타데이터는 기록을 구성하는 필수 요소 중 하나로 기록물을 적절하게 관리하고 이해하도록 하는데 매우 중요한 역할을 한다. 기록관리 업무에서 메타데이터 요소들의 자동 부여가 불가능할 경우에는 기록전문가가 메타데이터 값을 직접 입력해야 한다. 이러한 업무의 불편함을 개선하기 위해 본 연구에서는 신기술인 챗GPT를 활용하여 기록관리 메타데이터 요소의 추출 방안을 제시하고자 하였다. 챗GPT 기술을 활용하기 위해 파이썬 프로그램과 랭체인 라이브러리를 이용하여 PDF 문서를 제시하고 질문을 통해 기록물의 메타데이터를 추출해보았고, 챗GPT 온라인 서비스를 통해 여러 건의 PDF 문서를 첨부하여 기록물의 메타데이터 요소를 추출해보았다. 그 결과 챗GPT-3.5 turbo를 사용한 랭체인에서는 보안상으로는 안전한 추출 방법이지만 하나 메타데이터의 정확한 요소를 얻기에는 다소 한계가 있었고, 챗GPT-4 온라인 서비스에서는 보안상 중요 문서를 첨부할 수 없지만 비교적 정확한 결과를 추출하였다. 이를 통해 기록관리에서의 메타데이터 추출을 위한 챗GPT 기술 활용의 가능성을 타진할 수 있었고, 챗GPT 관련 기술의 발달에 따라 좀 더 안전하고 정확한 결과 추출이 가능해질 것이다. 이러한 챗GPT의 장점을 활용함으로써 기록관에서 기록 및 메타데이터의 관리적 측면에서 업무의 효율성 및 생산성을 증대시키는데 도움을 줄 수 있을 것이라 기대한다.

ABSTRACT

Metadata is a crucial component of record management, playing a vital role in properly managing and understanding the record. In cases where automatic metadata assignment is not feasible, manual input by records professionals becomes necessary. This study aims to alleviate the challenges associated with manual entry by proposing a method that harnesses ChatGPT technology for extracting records management metadata elements. To employ ChatGPT technology, a Python program utilizing the LangChain library was developed. This program was designed to analyze PDF documents and extract metadata from records through questions, both with a locally installed instance of ChatGPT and the ChatGPT online service. Multiple PDF documents were subjected to this process to test the effectiveness of metadata extraction. The results revealed that while using LangChain with ChatGPT-3.5 turbo provided a secure environment, it exhibited some limitations in accurately retrieving metadata elements. Conversely, the ChatGPT-4 online service yielded relatively accurate results despite being unable to handle sensitive documents for security reasons. This exploration underscores the potential of utilizing ChatGPT technology to extract metadata in records management. With advancements in ChatGPT-related technologies, safer and more accurate results are expected to be achieved. Leveraging these advantages can significantly enhance the efficiency and productivity of tasks associated with managing records and metadata in archives.

Keywords: 메타데이터, 기록관리 메타데이터, 챗GPT, 랭체인, 기록정보서비스
Metadata, Records Management Metadata, ChatGPT, Langchain,
Archival Information Service

© 한국기록관리학회

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

■ 본 논문은 김민지의 석사학위논문 「기록 관리 메타데이터 추출을 위한 챗GPT 활용 방안」(2024)을 요약·수정된 것임.

1. 서론

1.1 연구의 필요성과 목적

기록관리 메타데이터 표준에서 정의하는 기록관리 메타데이터는 기록물의 맥락, 내용, 구조 및 장기간에 걸친 관리사항을 기술하며, 기록물의 생산·등록·분류·접근·보존·처분을 가능하게 하는, 구조화된 정보 또는 반구조화된 정보이다. 이는 기록물과 기록물을 생산·관리·유지·이용하는 사람, 과정과 시스템, 그리고 기록물을 관장하는 정책 등을 확인하고 그러한 활동의 맥락을 파악하는 데에 사용될 수 있다(기록관리 메타데이터 표준(v2.3). NAK 8:2022). 이러한 이유로 기록관리에서는 검색 및 보존을 위해 메타데이터 요소에 대한 설정이 제대로 되어야 하는데, 그렇지 않은 경우가 많다.

기록관리 메타데이터 표준의 작성방법에 대해 살펴보면 메타데이터 요소의 입력 시, 자동 입력 또는 직접 입력을 통해 작성한다고 제시하고 있다. 그러나 요소들의 자동 부여가 불가능할 경우에는 기록전문가가 메타데이터 값을 직접 입력해야 한다. 또한 시스템에 의해 초기값이 자동으로 부여되는 몇몇 하위·세부 요소는 초기값의 변화 시 선택값에서 데이터를 선택하여 직접 입력해야 한다. 이 외 시스템에 의한 자동 입력이 되지 않는 기본적인 필수 요소들은 기록전문가가 일일이 기록물을 파악한 후 메타데이터를 직접 입력해야 한다(기록관리 메타데이터 표준(v2.3). NAK 8:2022).

기록관리에서 중요한 요소인 메타데이터 입력 업무에 대한 불편한 점을 조금이라도 개선하기 위한 기술 및 방안이 도입되어야 할 필요가 있으나, 기존에는 특별한 기술을 활용하기보다는 기록전문가가 기록을 살펴보고 직접 메타데이터를 추출하여 작성하였고, 기록관리 분야에서 기록관리 메타데이터 요소의 자동 추출을 시도한 사례는 찾아볼 수 없었다. 본 연구에서는 이러한 업무의 불편함을 개선하기 위해 인공지능의 발전으로 등장한 챗GPT라는 기술을 활용하여 자동으로 기록관리 메타데이터를 추출하는 것이 가능한지 확인해 보고, 가능한 방안을 제안하고자 한다.

과거부터 현재까지 다양한 기술들이 등장하며 사회 발전에 큰 영향을 주어왔다. 최근에는 인공지능 분야가 급속도로 발전하고 있는데, 특히 '챗GPT(ChatGPT)'라는 대화 전문 인공지능(AI) 챗봇 서비스가 등장하여 전 세계가 여기에 주목하고 있다. 챗GPT는 기존 챗봇과는 달리, 인간 수준 이상의 다양한 인공지능 개발을 목표로 연구를 진행하기 위해 만들어진 언어 모델이다. 이는 인터넷의 방대한 텍스트 데이터셋을 학습하여 질문에 대한 답변을 생성하는 물론 인간과의 대화, 에세이 또는 이메일 작성, 텍스트 요약, 언어 번역, 코딩 생성 및 리뷰 등 언어와 관련된 다양한 작업을 수행할 수 있다(김수민, 백선환, 2023).

현재 챗GPT의 기능을 더욱 활성화하기 위해 많은 연구와 활용이 이루어지고 있다. 그중 오픈AI에서 제공하는 챗GPT의 개인 API key를 파이썬이나 자바 등의 프로그래밍 언어에서 활용하는 방안이 제시되고 있다. 이를 통해 다양한 기능을 결합할 수 있는 코드를 작성하여 챗GPT가 가진 기능을 더 효율적으로 사용할 수 있다. 또한 언어 모델을 기반으로 하여 애플리케이션 개발을 할 수 있는 프레임워크인 랭체인 기술을 이용해 챗GPT의 기능을 활용할 수 있다. 이를 이용하면 파이썬 환경에서 PDF 문서를 학습시켜 해당 문서에 대한 정보를 제공하거나 웹 검색이 가능한 코드를 작성해 웹에서 정보를 찾아 제공하는 서비스를 개발할 수 있으며 사용자의 데이터를 안전하게 보호할 수 있다(후루카와 히데카즈, 2023).

본 연구에서는 챗GPT 기술을 활용하여 기록물의 메타데이터 자동 추출이 가능한지 알아보고, 성능 확인에 도움이 될 수 있도록 파이썬 코드 및 랭체인을 연결하여 구현함으로써 기록전문가의 기록관리 업무를 개선해보고자 하였다. 먼저, 글자가 포함되어 있고 OCR 기술이 적용된 PDF/A 형식의 문서를 파이썬 프로그래밍을 통해 챗GPT에 입력한 후, 학습한 해당 문서의 내용에 대한 질문 및 답변 등의 작업을 실행하여 기록관리 메타데이터를

추출해보았다. 최근 업그레이드되어 유료 버전으로 이용할 수 있는 챗GPT-4를 활용하여 기록에 대한 정확한 분석을 할 수 있으나, 이는 부분공개나 비공개 기록의 경우, 문서에 포함된 기밀 또는 개인정보 등이 유출될 위험이 커진다. 따라서 오픈AI로부터 직접 챗GPT의 API key를 발급받아 파이썬을 활용하는 랭체인을 통해 기록의 메타데이터를 추출하는 것이 보안상 안전하다고 볼 수 있다. 앞으로 기록관리에서 메타데이터 추출에 챗GPT를 활용함으로써 기록전문가의 기록관리 업무에 도움이 될 것이며, 메타데이터 추출뿐만 아니라 기록물의 내용 요약의 기술에의 활용, 이미지 인식 등 다양한 업무에도 활용할 수 있을 것으로 기대한다.

1.2 선행연구

본 연구는 기록관리 메타데이터와 챗GPT에 관한 내용을 다루기 때문에 이와 관련된 연구를 살펴보고자 하였다. 메타데이터의 획득에 관한 직접적인 연구가 없기에 기록 메타데이터 요소, 설계 및 기술 등에 대한 연구를 참고하였다. 또한 챗GPT는 신기술이기 때문에 현재 진행된 연구가 많이 없어 발간된 단행본 및 학술 논문 등을 참고하여 연구하였다.

최상미, 이상용(2007)은 국내 및 호주, 영국의 국가 메타데이터 표준과 우리나라 공공기록물의 메타데이터 요소를 분석하여 현용전자기록물의 메타데이터 요소를 제안하고자 하였다. 현용전자기록물 단계에서 획득되는 19가지 요소 추출 및 미국의 메타데이터 참고모델에서 제시한 설계 내용 중 생산과 관리를 돕기 위한 필수 요소 7가지, 선택 요소 10가지를 추출하였다. 현재의 기록물등록대장과 기록물철등록부에서 메타데이터 요소가 생산되지만 생산 이유, 첨부문서 존재 여부 및 내용, 기록의 변경 및 접근 시도 등의 내용 파악이 불가능하며 자동 생성할 수 있는 메타데이터 요소들이 자동으로 생성되지 않거나 필수사항으로 지정되지 않아 추출하지 못하는 것을 확인할 수 있었다. 논문에서 제안한 메타데이터 요소를 추출한 결과, 현재 전자기록물 생산 환경에서 거의 모든 메타데이터 요소들의 생성이 충분히 가능하였으며 전자문서시스템의 문서 생산 과정에서 자동 생성의 기반을 마련하면 도움이 될 수 있을 것이라는 내용을 확인하여 기록관리에서의 메타데이터 자동 추출 기술의 필요성을 인식하였다.

최주호, 이재영(2012)은 영구기록관리시스템으로 이관되는 전자기록물에서 메타데이터를 추출하여 항목별로 분류한 후 이관되는 메타데이터 항목과 이를 비교 및 검증하기 위한 기술을 개발하고자 하여 다양한 메타데이터 추출 도구 및 도구별 추출 데이터 등을 비교하였다. 이를 바탕으로 인수 전자기록물의 원문으로부터 메타데이터를 추출하여 문서의 데이터 항목에 저장된 기존 메타데이터와 추출 메타데이터 간의 비교 검증을 수행하는 시스템을 개발하였다. 해당 연구에서 시스템 적용 가능성을 모색한 결과를 통해 기록관리에서 메타데이터 검증 자동화 프로세스 및 메타데이터 추출 기술이 요구되는 것을 확인하였다.

남용우(2013)는 메타데이터 추출도구의 필요성을 강조하며 전자기록관리시스템에 입력하기 위한 디지털 컴포넌트의 메타데이터 추출 도구가 존재하지 않는 문제점을 개선하기 위하여 추출 도구를 위한 프로세스를 설계하고자 하였다. 국가기록원의 표준운영절차를 참조하여 메타데이터 추출 절차에 추가할 부분을 확인하였고 국가기록원의 메타데이터 표준을 분석하여 자동 추출 정보와 메타데이터 항목의 매핑 여부를 살펴보았다. 직접 디지털 컴포넌트의 비트스트림 정보를 확인하여 자주 사용하는 디지털 컴포넌트의 속성정보를 분석하였으며 메타데이터 추출 도구를 위한 프로세스를 설계하였다. 이를 통해 자동으로 메타데이터를 추출하는 기술은 물론 다양한 포맷으로부터 메타데이터를 추출하여 전자기록의 접근성을 강화하는 방안을 제시할 필요가 있음을 살펴볼 수 있었다.

김인아, 강영선, 이규철(2020)은 국내의 지방자치단체에서 제공하는 사진 기록물의 표준 메타데이터 부재와 사진 정보 미활용의 문제점을 개선하기 위하여 사진 기록물의 메타데이터 설계 및 기계학습 기반 자동 인덱싱 기술을 제안하였다. 제안한 방안을 통해 OCR 기반 지자체 사진 기록물에 포함된 텍스트 및 개체를 자동 추출하고, 추출한 내용과 기본 정보를 해당 논문에서 설계한 사진 기록물 메타데이터 항목으로 구성된 데이터베이스 테이블

에 저장하는 프로그램을 개발하였다. 이를 통해 추출한 메타데이터를 저장 및 인덱싱하여 기록물 검색의 정확도를 향상할 수 있는 방안을 제시하였으며 기록관리에서 전자기록의 메타데이터 자동 추출 기술의 필요성뿐만 아니라 사진 기록 유형과 같은 포맷의 메타데이터를 추출하기 위한 방안의 부재를 인식하여 앞으로 기술의 발달이 필요함을 알 수 있었다.

서지인, 노지현(2022)은 디지털화 기록의 체계적인 관리와 품질 유지를 위해 필요한 메타데이터 요소(안)를 제안하고자 하였다. 이를 위하여 메타데이터의 설계 방향을 원천 기록과 디지털화 기록의 관계 정립, 자원 유형별 특성에 따른 자동 추출 메타데이터 요소 확대, 업무과정 및 프로젝트에 대한 메타데이터 작성으로 정리하였다. 디지털화 기록에서 자동으로 획득 가능한 요소와 메타데이터 작성자가 직접 입력해야 할 요소를 구분하여 제시함으로써 시스템으로부터 자동으로 획득되어야 하는 요소를 추출할 수 있는 기술의 개발 필요성을 확인하였다.

한편 Lund & Wang(2023)은 챗GPT가 학계와 도서관에 어떤 영향을 미칠지에 대해 면접관의 질문에 챗GPT가 답변한 내용을 정리하였다. 챗GPT는 학계에서 연구 및 학업 개선을 위한 챗GPT 사용법, 교육 환경에서 윤리 및 개인정보 보호에 미치는 영향, 도서관에서 자연어 처리를 위한 챗GPT 사용 의미, 학술 도서관에서 챗GPT 사용 시 고려해야 할 윤리적 문제에 대하여 답변하였다. 이를 통해 챗GPT가 불안하지만 학계와 도서관의 발전에 도움이 될 수 있으며, 전문가로서 함께 업무를 개선할 방법을 찾는 것이 중요하다고 제시하였다. 따라서 도서관뿐만 아니라 기록관리 분야에서도 챗GPT 기술을 활용함으로써 기록에 대한 질의응답이나 검색 등 업무 효율성을 높이는 다양한 방안을 제안할 수 있음을 살펴볼 수 있었다.

장운아, 오효정(2023)은 대체로 1인 기록관 체제인 국내 기록관리 기관의 전자기록물 관리 인력 및 자원이 부족한 상황에서 생성형 AI 기술을 활용할 수 있는 방안을 제안하였다. 기록관리 분야의 활용 방안으로 전자기록 관리 업무지원에서는 기록물 제목·메타데이터·요약문·기록물 생산현황 통보파일 자동 생성, 업무용 대화형 챗봇을 제시하였다. 기록정보서비스에서는 서비스용 대화형 챗봇 및 음성봇, 전시 콘텐츠 자동 생성, 지능형 큐레이션 로봇을 제안하여 각 업무 분야에 맞는 활용 방안을 정리하였다. 이를 통해 챗GPT가 기록물의 메타데이터 추출뿐만 아니라 기록관리의 여러 업무에서 다양하게 활용될 수 있는 기술임을 인식할 수 있었다.

1.3 연구범위 및 방법

본 연구는 챗GPT 기술을 프로그래밍 언어 및 프레임워크 등과 결합하여 기록전문가가 기록관리 메타데이터를 수월하게 추출할 수 있는 방안을 제안하기 위한 것으로 다음과 같이 연구를 수행하였다.

첫째, 국가기록원의 기록관리 메타데이터 표준인 「NAK 8:2022(v2.3)」에서 제시하는 기록관리 메타데이터 정의 및 중요성과 필수 요소들을 살펴보았다. 이를 챗GPT 구현에 사용하기 위해 표준에서 정의한 기준에 따라 메타데이터 요소를 표로 정리하여 제시하였다. 또한 챗GPT 및 랭체인 관련 문헌 연구를 통해 해당 기술에 대한 개념 및 정의 등을 살펴보았다. 최신 동향 및 활용 방안 등을 파악하기 위해 출판된 단행본 및 학술 논문을 참고하였다.

둘째, 기록 계층 중 기록물의 최소 단위인 기록건을 기준으로 하여, 국가기록원에서 문서보존포맷으로 제시한 PDF/A 포맷을 선정하여 구현에 사용하고자 하였다. 챗GPT 학습 대상 기록건은 공개된 결재문서를 이용할 수 있는 서울정보소통광장 홈페이지에서 제공하고 있는 PDF/A 포맷 형태의 문서를 사용하였다. 챗GPT가 다양한 내용 구조의 결재문서로부터 메타데이터 요소를 추출할 수 있는지 확인하기 위해 문화·관광, 행정·기타 분야의 기록물 시리즈에서 각 5건씩 저장하여 총 10건의 문서를 파이썬 환경에서 챗GPT-3.5 터보를 활용한 질문 랭체인, 챗GPT-4를 사용할 수 있는 온라인 서비스에서의 구현 및 버전별 결과 비교분석에 사용하였다.

셋째, 챗GPT를 유용하게 활용하기 위해 프로그래밍 언어인 파이썬과 프레임워크인 랭체인을 결합하여 다양하게 구현하였다. OpenAI로부터 개인 API Key를 발급받은 후, 챗GPT를 연결할 수 있는 파이썬 코드 및 랭체인으로 OCR 기술이 적용된 PDF 문서를 불러와 학습을 진행하였다. 파이썬 프로그래밍은 마이크로소프트사에서 만든 비주얼 스튜디오 코드(Visual Studio Code)에서 실행하였다. 국가기록원의 메타데이터 표준에서 제시한 필수 요소들을 챗GPT가 정확하게 추출할 수 있는지 판단하기 위해 챗GPT-3.5 터보와 챗GPT-4 버전을 사용하였다. 파이썬 환경에서는 챗GPT-3.5 터보 버전을 사용하였는데, 이는 챗GPT-3.5의 기능보다 발전하여 기존 모델에 비해 더 정확한 답을 제시할 수 있으며 기업의 보안 문서 등을 사용할 경우 기밀 유출을 방지할 수 있기 때문이다. 또한 온라인 서비스 환경에서 도출하는 결과를 확인하기 위해 유료 버전인 챗GPT-4를 사용하였다.

2. 이론적 배경

2.1 기록관리 메타데이터

국가기록원의 「기록관리 메타데이터 표준(v2.3)」은 장기간에 걸쳐 기록물의 진본성, 무결성, 신뢰성 및 이용 가능성을 보장하기 위해 공공기관이 생산 또는 접수하는 기록물에 대한 맥락과 내용, 구조 및 기록생애주기 동안 기록물 관리 기관이 관리할 사항을 기술하기 위해 제정되었다. 이는 정보객체가 아날로그이건 디지털이건 간에 기록물의 진본성·신뢰성·이용가능성·무결성을 보장하며, 기록물을 적절하게 관리하고 이해하는데 매우 중요한 역할을 한다고 정의하고 있다. 또한 결재과정에서 생산된 문서류 위주의 기록물에 대한 메타데이터 요소를 제시하고 있으며 시청각기록물, 웹 기록물, 행정정보데이터세트 등과 같은 기록물의 공통 항목에 대해서도 이 표준을 적용할 수 있다고 말하고 있다.

이 표준에 수록된 기록관리 메타데이터는 23개 상위 요소, 66개 하위 요소, 58개 세부 요소의 3계층으로 구성되며, 하위 요소 중 세트로 작성되어야 할 요소들은 컨테이너로 묶어 하위 요소 - 세부 요소로 계층화하여 정의한다. 상위 요소 중 필수 요소는 총 12개로 생산자, 기록계층, 기록식별자, 기록물명, 전자기록물 여부, 유형, 크기, 분류, 일시, 보존기간, 보존장소, 권한으로 제시하고 있다. 요소 설명은 정의, 적용기록계층, 컨테이너 여부, 필수 여부, 반복 여부, 선택값, 작성 방법, 초기값 순으로 기술되어 있고 적용기록계층, 선택값, 초기값은 요소의 계층에 따라 필요시 기술된다. 전자적으로 생산된 기록물의 진본성을 보장하기 위해 기록이 생산 혹은 획득되는 시점에서 생산자, 기록식별자, 기록물명, 일시, 생산이력 요소는 한 번 등록된 후 변경될 수 없다고 정의하고 있다.

아래 <표 1>은 표준에서 정의하고 있는 기록관리 메타데이터 요소표에서 메타데이터 작성 시 필수로 추출해야 하는 값들이다.

<표 1> 「기록관리 메타데이터 표준(v2.3)」 필수 요소

상위 요소	하위 요소	세부 요소
1 생산자	1.1 생산자 유형	
	1.2 기관명	
2 기록계층		
3 기록식별자	3.1 기본식별자	3.2.1 시스템식별자 유형
		3.2.2 시스템식별자
		3.3.1 보조식별자 유형
		3.3.2 보조식별자
4 기록물명	4.1 제목	
		4.2.1 기타제목 유형

		4.2.2 기타제목명
5 기술	5.1 기술 유형	
	5.2 기술 내용	
6 주제	6.1 주제 유형	
	6.2 주제명	
7 전자기록물 여부		
8 유형	8.1 기록 유형	
	8.4 컴포넌트 유형	
9 포맷	9.1 포맷명	
10 매체		
11 크기	11.1 용량	
	11.2 단위	
12 분류	12.1 분류체계 유형	
	12.2 분류값	12.2.2 분류명
13 일시	13.1 생산일시	
	13.2 종료일시	
14 생산이력		14.2.1 기관명
		14.2.3 부서명
		14.4.1 비치사유
		14.4.2 비치기간
		14.6.1 생산자
		14.6.5 처리상태
		14.6.7 처리순번
15 보존기간	15.1 보존기간	
16 보존장소		
17 권한		17.1.1 비밀분류
		17.1.2 비밀분류 근거
		17.1.3 보호기간
	17.3 공개	17.3.1 공개구분
		17.4.1 공공저작물여부
18 위치	18.1 소장처	
	18.2 소장위치	
19 관리이력	19.1 관리 유형	
	19.3 관리일시	
	19.4 관리행위자	19.4.1 기관명
		19.4.5 개인명
		19.5.1 변경요소명
	19.5.2 변경이전값	
20 이용이력	20.1 이용 유형	
	20.3 이용일시	
	20.4 이용자	20.4.3 부서명
		20.4.5 개인명
21 보존이력	21.1 보존처리 유형	
	21.3 보존처리일시	
	21.4 보존행위자	21.4.1 기관명
	22 관계	22.1 관계 유형
	22.2 관계대상식별자	
23 무결성체크	23.1 무결성체크법	
	23.2 무결성체크값	

2.2 챗GPT(ChatGPT)의 현황

챗GPT(ChatGPT)는 미국의 인공지능 연구 기업인 오픈AI(OpenAI)가 개발한 대화 생성 인공지능 언어 모델이다. GPT는 “Generative Pre-trained Transformer”의 줄임말로, 광범위한 데이터를 미리 학습하고 이를 토대로 질문을 이해하여 답변을 생성하고 대화의 상대가 이해할 수 있게 답을 변환하여 응답하는 시스템이다(안종배, 2023). 챗GPT는 ‘대규모’ 언어 모델(Large Language Model, 이하 LLM)로 빅데이터를 사용하여 언어 모델의 정확도나 활용도를 높아지도록 한다. LLM은 방대한 양의 문장 데이터를 학습하여 자연어 처리 작업을 수행하기 위해 풍부한 지식을 획득한 딥러닝 모델이다. 임의의 문장 뒤에 오는 다음 문장을 예측하는 능력과 자연스러운 문장을 만드는 능력을 학습하고 있으나, 수학처럼 논리적인 사고를 조합하는 데에는 서투르다(후루카와 히데카즈, 2023). 또한 자연어를 이해하고 생성하며 대화 내용 및 맥락을 기억하여 처리하는 등의 다양한 기술을 사용하여 사용자의 질문이나 요청에 대해 적극적으로 최적의 답안을 대응하는 서비스이다. 이에 사람들은 챗GPT를 활용하여 인공지능과의 대화, 글쓰기 창작 활동, 자료 검색, 엑셀 함수 활용, 프로그래밍 등 다양한 분야에서 도움을 받고 있다.

2023년 3월, 오픈AI는 챗GPT 모델을 외부에서 이용할 수 있도록 웹 API를 제공하기 시작하였다. 이 API를 통해 기업에서 자연어 처리 작업에 대한 입력을 전송하고, 해당 작업에 대한 결과를 받아올 수 있으며 여기에는 GPT-3.5 중 가장 성능이 뛰어난 GPT-3.5 터보(GPT-3.5-Turbo)라는 모델이 적용되었다(최은석, 2023). GPT-3.5 터보 모델은 기존의 GPT-3.5에 비해 업그레이드되어 빠른 일처리가 가능하며 기존 모델보다 대화가 자연스럽고 답변 속도가 빠르며 토큰을 적게 사용한다(이성용, 2023). 또한 챗GPT 온라인 서비스에 입력한 데이터가 오픈AI로 전송되어 기업 내부의 입력 내용 및 자료, 정보 등의 유출 우려 문제가 발생할 수 있는데, API를 사용하면 이러한 문제의 발생 가능성이 줄어든다(최은석, 2023).

이어 오픈AI에서 GPT-3.5의 업그레이드 버전인 GPT-4를 공개하였으며 학습 데이터 기간은 GPT-3.5와 달리 2023년 4월까지로 비교적 최신 정보를 갖고 있다. 기존의 GPT-3.5 모델이 한 번에 3,000개의 단어를 만들 수 있었다면, GPT-4에서는 약 25,000개로 늘어났다(엔미디어, 2023). 매개변수의 양 또한 5,000억 개로 기존의 챗GPT보다 약 5배 정도 증가하였다. 또한 GPT-4는 기존 모델의 학습 데이터와 모델 구조를 개선하여 보다 높은 수준의 자연어 처리 능력을 확보하여 GPT-3.5보다 더 정확한 정보를 제공할 수 있게 되었다(후루카와 히데카즈, 2023).

2.3 랭체인(LangChain)의 개념

랭체인(LangChain)이란 GPT-4, GPT-3.5와 같은 대규모 언어 모델을 이용하여 서비스를 개발하기 위해 사용할 수 있는 라이브러리이며 LLM을 활용한 애플리케이션 개발을 지원하는 오픈 소스 라이브러리이다. 기업에서 개인정보나 기밀이 포함된 문서에 대한 작업을 할 때, 챗GPT의 API만 이용하는 경우에는 간단한 대화를 할 수 있는 챗봇처럼 사용할 수 있다. 이에 랭체인을 함께 사용하면 여러 가지 기술을 연결하여 챗GPT API의 활용성이 더욱 높아지며 사용자의 데이터가 안전하게 보호될 수 있는 장점이 있다. 이는 다양한 모듈을 조합하여 복잡한 작업을 수행하는 애플리케이션을 개발할 수 있으며 대개 학습을 기반으로 대화하는 LLM에 책이나 프로그램을 전달하여 외부의 지식이나 계산 능력을 활용하게 하고 싶을 때 사용한다(후루카와 히데카즈, 2023).

또한, ‘에이전트’ 모듈을 통해 사용자의 입력으로부터 답변으로 요구되는 것을 추론하고, 필요한 도구를 선택하여 실행함으로써 최적의 답변을 도출한다. 에이전트는 내부적으로 답변이 불충분하다고 판단하면 추론과 행동을 반복하여 복잡한 작업을 높은 정확도로 구현 가능하다. 랭체인을 통해 개인 또는 회사의 자체 데이터를 연결하여

이에 기반한 질문에 답변하는 기능, 웹 서비스 연결 및 질문에 대한 답변 기능, 데이터베이스를 연결하여 자연어로 원하는 데이터를 원하는 형식으로 쿼리하는 기능, 프로그래밍 실행 환경과 연결하여 명령어에 따라 프로그램을 작성하고 실행하는 기능 등을 이용할 수 있다(후루카와 히데카즈, 2023). 이러한 기능들은 파이썬 환경에서 챗 GPT 모델을 불러와 랭체인에 연결하여 활용할 수 있다. 즉 메타데이터를 추출하는 데에 입력 데이터가 노출될 수 있는 챗GPT를 직접 활용하는 대신, 기관의 고유 기록물의 메타데이터를 내부적으로 추출하도록 활용할 수 있는 것이다.

3. 챗GPT를 활용한 기록관리 메타데이터 추출 구현 및 결과 분석

3.1 「기록관리 메타데이터 표준(v2.3)」의 필수 요소 추출 준비 과정

3.1.1 PDF/A 포맷 구현 대상 문서 준비

본 연구에서 실제 기록물을 대상으로 챗GPT를 활용하여 메타데이터를 추출하는 과정을 제시하고자 하였다. 기록관리 메타데이터의 자동 추출을 위한 학습 대상 포맷으로 PDF 문서를 선정하였고, 챗GPT에 여러 개의 PDF 문서를 학습시켜 해당 문서의 내용을 추출하는 작업을 시행하였다. PDF를 학습 대상 포맷으로 선정한 이유는 국가기록원에서 문서보존포맷은 문서의 내용 정보를 영구보존하기 위한 포맷 중 가장 만족도가 높은 것을 PDF/A로 제시하고 있으며, 대부분의 기관에서 디지털 기록을 PDF 문서로 제공하고 있기 때문이다.

또한 PDF는 온라인 챗GPT 서비스 환경뿐만 아니라 파이썬 환경에서 GPT 모델을 불러와 랭체인으로 문서를 학습시킬 때도 활용할 수 있다. 다만 챗GPT는 문서에 포함된 이미지는 읽을 수 없고, OCR을 통해 추출된 텍스트를 인식할 수 있기 때문에 OCR 기능이 수행된 PDF/A 포맷의 문서를 학습에 사용해야 한다. 이러한 문제점은 추후 딥러닝 기반의 AI-OCR(인공지능 광학식 문자 판독) 기술을 통해 인식이 낮았던 인쇄체, 필기체(손글씨), 타자 글씨, 사진, 팩스, 스캔문서 등 저품질·저화질 문서로부터 문서 추출이 가능해지고, 이를 PDF/A 포맷 문서에 적용하면 해결될 것으로 보인다.(안세진, 황현호, 임진희, 2022).

본 연구에서 구현에 사용한 PDF 문서는 서울정보소통광장에서 제공하고 있는 결재문서로 대부분이 PDF/A 포맷이며 공개된 상태로 이용 가능하여 학습 예시로 활용하기에 적합하다고 판단하였다. 또한 전자 결재문서이므로 대부분 스캔문서가 아닌 전자문서 형태로 생산 및 결재되었고, 문서에 첨부된 이미지를 제외한 내용의 대부분이 타자글씨로 작성되어 문자 인식이 가능한 문서이므로 챗GPT의 문서 내용 인식 및 학습에 수월할 것으로 판단하였기 때문이다. 구현에 사용한 문서는 서울정보소통광장의 결재문서 중 문화·관광, 행정·기타 분야의 기록물 시리즈에서 각 5건씩 선택하여 저장하였으며, <그림 1>과 같이 학습 편의상 폴더 내 결재문서의 제목을 수정하였다.

3.1.2 질문 대상 요소 선정

「기록관리 메타데이터 표준(v2.3)」에서는 필수적으로 작성되어야 하는 요소들을 제시하고 있다. 그중 서울정보소통광장의 문서들로부터 추출할 수 있다고 판단한 메타데이터 요소들을 <표 2>로 간단히 정리한 것이다. 질문에 사용할 필수 요소들은 <표 2>와 같이 회색 배경 셀의 '생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시'로 정하였다. 생산자, 기관명, 기록물명(제목), 생산일시와 같이 문서에 작성되어 있어 챗GPT가 인식하여 추출할 수 있는 요소들을 기준으로 선정하였다. 기록식별자 역시 문서의 생산 또는 접수 등록번호이므로 문서에 포함되어 있어 기준으로 정하였다. 다만 상위 요소 중 5 기술은 필수 요소가 아니기 때문에 구분을 위하여 기울임과 밑줄로 표시하였다.

- 1-1. 문화관광1_돈의문 박물관마을 소공원 조성공사 휴일작업 승인 알림
- 1-2. 문화관광2_2023년도 제8차 유물수집실무위원회 개최
- 1-3. 문화관광3_시-우리카드 우리WON 프로배구단 '서울시민 데이' 무료 초청 협조요청
- 1-4. 문화관광4_한성백제박물관 청소년 대상 자율관람 프로그램 수강신청 안내
- 1-5. 문화관광5_서울시립 미술아카이브 <라스트 제너레이션에게, 김용익>어린이·청소년 프로그램 참가 모집 안내
- 2-1. 행정기타1_소상공인 지원 관련 현안업무 검토
- 2-2. 행정기타2_2023년 제4기 국정 현안 법제의 이해과정(공통) 교육 신청 안내
- 2-3. 행정기타3_산림인접지역 쓰레기 불법소각 및 불피움에 따른 행정처리 절차 알림
- 2-4. 행정기타4_서서울미술관 설계 관련 간담회 비용 지급
- 2-5. 행정기타5_서울시립미술관 대관(1층 회의실) 협조 요청

<그림 1> 서울정보소통광장 문서 목록

<표 2> 「기록관리 메타데이터 표준(v2.3)」 필수 상·하위 요소

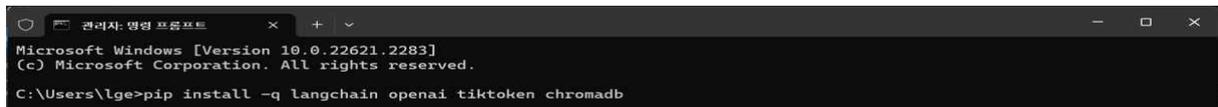
「기록관리 메타데이터 표준(v2.3)」 필수 상위 요소	「기록관리 메타데이터 표준(v2.3)」 필수 하위 요소
1 생산자	1.2 기관명
2 기록계층	
3 기록식별자	
4 기록물명	4.1 제목
5 기술	5.2 기술 내용
7 전자기록물 여부	
8 유형	
11 크기	11.1 용량
12 분류	
13 일시	13.1 생산일시
14 생산이력	
15 보존기간	
16 보존장소	
17 권한	

3.1.3 파이썬 코드 준비

파이썬으로 구현하기 위해 비주얼 스튜디오 코드를 실행하여 파이썬 파일을 생성한 후 <그림 2>와 같이 챗 GPT를 활용하기 위한 코드를 작성하였다. <그림 3>의 코드는 <그림 2>와 같이 비주얼 스튜디오에서도 실행 가능하다. 그러나 간혹 윈도우 및 파이썬 버전 등의 환경에 따라 실행 과정에서 오류가 생기는 경우도 있다. 이 때문에 본 구현 과정에서는 윈도우 환경에서 터미널을 관리자 모드로 열어 파이썬 패키지의 설치를 실행하였다.

```
# OpenAI API 키
!pip install -q langchain openai tiktoken chromadb
```

<그림 2> 비주얼 스튜디오 파이썬 패키지 설치 코드



<그림 3> 명령 프롬프트 파이썬 패키지 설치

PDF 문서를 불러와 질문하는 작업을 하기 위해 먼저 <그림 4>와 <그림 5>에서 작성한 경로에 있는 PDF

문서를 파이썬으로 불러온다. <그림 4>는 OpenAI에서 발급받은 개인 키를 사용하기 위해 openai 모듈을 가져오고, PyPDF2와 Langchain 라이브러리에서 다양한 클래스를 불러와 사용하기 위한 코드이다. 여기서 'OPENAI_API_KEY' 코드의 뒷부분인 ""에는 홈페이지에서 발급받은 개인 키(Key)를 사용해야 한다. 이는 개인별로 부여되고 사용량에 따라 요금이 부과되기 때문에 <그림 4>에서는 이를 제외하고 공백으로 표시하였다. 다음으로 PDF 문서를 불러오는 작업을 하기 위해 <그림 5>에서 작성한 경로에 있는 PDF 문서를 파이썬으로 불러온다. 여러 개의 PDF 문서를 한 번에 불러오기 위한 코드로 한 폴더 내에 문서를 저장하여 해당 폴더를 경로로 설정하였고, 이 경로에서 PDF 파일을 찾아오는 것을 의미한다.

```
# OpenAI API 키
!pip install -q grobid-client langchain openai faiss-cpu PyPDF2 tiktoken

import openai
from PyPDF2 import PdfReader
from langchain.embeddings.openai import OpenAIEmbeddings
from langchain.text_splitter import CharacterTextSplitter
from langchain.vectorstores import ElasticVectorSearch, Pinecone, Weaviate, FAISS
import os

os.environ["OPENAI_API_KEY"] = ""
```

<그림 4> API Key 및 랭체인 패키지 설치 코드

```
# PDF 파일이 있는 폴더 경로 설정
pdf_folder_path = "c:\\Users\\lge\\Desktop\\서울정보소통광장"

# 폴더 내의 모든 PDF 파일 목록 가져오기
pdf_files = [f for f in os.listdir(pdf_folder_path) if f.endswith('.pdf')]

# 추출된 텍스트를 저장할 리스트 초기화
texts = []
```

<그림 5> 구현 대상 PDF 문서의 폴더 경로 및 리스트 설정

폴더 내의 PDF 파일을 하나하나 처리하기 위한 반복문을 <그림 6>과 같이 설정한다. 'for'로 시작하는 코드는 반복문을 나타내며 주어진 조건을 여러 번 실행하는 것이다. 여기서 사용한 반복문의 형식은 'for 변수 in 반복을 지원하는 자료구조 : 변수를 사용하는 코드 문장' 구조이다(조용주, 임좌상, 2023). 이 코드는 반복문의 시작부터 끝까지 모두 드래그하여 한 번에 실행해야 한다. 이후 'print(len(texts))' 또는 'print(len(PDF_files))' 코드를 사용하여 texts 리스트에 저장된 PDF 파일의 개수를 출력하여 확인해야 한다.

```
# 각 PDF 파일을 처리하기 위한 반복문
for pdf_file in pdf_files:
    pdf_path = os.path.join(pdf_folder_path, pdf_file)
    pdf_reader = PdfReader(pdf_path)

    # 텍스트 추출
    pdf_text = ""
    for page in pdf_reader.pages:
        pdf_text += page.extract_text()

    # 추출된 텍스트를 리스트에 추가
    texts.append(pdf_text)

print(len(texts))
print(len(pdf_files))
```

<그림 6> 각 PDF 파일의 텍스트 추출 반복문

이후 불러온 랭체인 코드를 이용하여 <그림 7>과 같이 더 세세하게 질의응답 코드를 설정한다. 본 연구에서는

챗GPT-3.5 터보를 활용하므로 gpt-3.5-turbo 모델을 구현에 사용하였다.

```
# 질문 답변 체인 설정
from langchain import OpenAI
from langchain.chains import AnalyzeDocumentChain
from langchain.chains.question_answering import load_qa_chain
from langchain.chat_models import ChatOpenAI
model = ChatOpenAI(model="gpt-3.5-turbo")

qa_chain = load_qa_chain(model, chain_type="map_reduce")
qa_document_chain = AnalyzeDocumentChain(combine_docs_chain=qa_chain)
```

<그림 7> 질의응답 랭체인 설정

구현을 위한 문서의 선택 및 랭체인 활용을 위해 <그림 8>과 같이 코드를 작성하였다. 먼저 서울정보소통광장 폴더에서 문서를 하나하나 선택하는 코드를 작성하였다. 파이썬에서 첫 번째 위치가 부여받는 숫자는 1이 아닌 0이다. 따라서 0번째 인덱스는 폴더의 목록상 첫 번째 문서인 문화관광1 문서이다. 이어서 선택한 문서의 질의응답을 실행하기 위해 앞서 지정한 ‘생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시’에 대해 질문하는 코드를 작성하였다.

```
# 문서 선택
raw_text = texts[0]
raw_text = texts[1]
raw_text = texts[2]
raw_text = texts[3]
raw_text = texts[4]
raw_text = texts[5]
raw_text = texts[6]
raw_text = texts[7]
raw_text = texts[8]
raw_text = texts[9]

qa_document_chain.run(
    input_document=raw_text,
    question="생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시를 이 문서에서 추출해줘")
```

<그림 8> 리스트 내 문서 선택 및 질문 실행 코드

3.2 파이썬 환경에서 랭체인을 활용한 「기록관리 메타데이터 표준(v2.3)」의 필수 요소 추출

먼저 파이썬에서 랭체인을 이용하여 메타데이터 요소를 추출하는 방법이 보안상 안전하여 본 연구에서는 챗GPT-3.5 터보 모델을 불러와 메타데이터 추출을 시도해보고자 하였다. 서울정보소통광장에서 분야별로 각 5건씩 저장한 문화·관광, 행정·기타 분야의 문서를 순서대로 구현함으로써 결과를 확인하였다.

문서 내용과 실행 결과를 비교하기 위해 첨부한 문서 사진은 편의상 추출한 메타데이터가 포함된 부분만 캡처하였으며 문화관광 분야의 문서 2건의 결과는 <그림 9>, <그림 11>과 같다. 먼저 문화·관광 분야의 문서들에 대해, 챗GPT-3.5 터보로 질문한 결과, 문화관광3 문서는 요소를 제시하지 못하였고, 이를 제외한 4건에 대한 결과만 제공하였다. 좀 더 상세히 살펴보면, <그림 9>와 같이 문화관광1에 대해서는 기록물명, 기술 내용만 정확히 뽑아내었다. 기록물명은 제목이므로 그대로 추출하였고, 기술 내용은 결재문서의 내용 중 중요한 내용의 문장들만 뽑아내어 문서의 기술 내용에 적합하다고 보았다. 반면, 생산자는 수신자로, 기관명은 수신자의 기관명으로, 기록식별자는 관련 문서의 문서번호 일부로, 생산일시는 관련 문서의 생산일시를 찾아 답으로 제공하여 틀린 메타데이터 추출 결과를 제공하였다.

```
>>> raw_text = texts[1]
>>> question = "생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시를 이 문서에서 추출해줘"
>>> qa_document_chain.run(input_document=raw_text, question=question)
'생산자: 서울시설공단이사장공사감독처장\n기관명: 서울시설공단\n기록식별자: 3-8168 (2023.10.5)\n기록물명(제목): 돈의문 박물관마을 소공원 조성공사 휴일작업 승인 알림\n기술 내용: 귀처의 돈의문 박물관마을 소공원 조성공사 휴일작업 토월을 승인함을 알림. 현장안전관리 및 품질관리에 보다 철저를 기하여 주시기 바램. 휴일작업 이후 외부인들의 무단출입에 따른 차 사고 발생 또는 시설물의 훼손 도난 등이 발생하지 않도록 현장관리에 철저를 기하시기 바램.\n생산일시: 2023.10.5'
```

<그림 9> 파이썬에서 챗GPT-3.5 터보를 활용한 문화관광1 문서 질문 랭체인 실행 결과



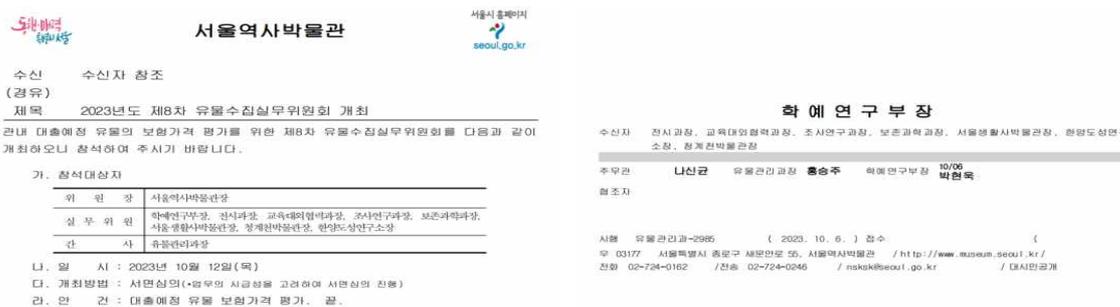
<그림 10> 서울정보소통광장 문화관광1 문서

출처 : 서울정보소통광장 - 결재문서(공개여부 : 공개, 기관 : 전체, 일반문서 : 포함) - 분야 : 문화·관광

문화관광2의 추출 결과에서는 <그림 11>과 같이 기관명, 기록물명만 정확히 뽑아내었다. 기록물명은 숫자가 뒤로 빠져 따로 추출된 것 외에는 문제가 없으므로 다른 요소들에 비하여 정확한 것으로 볼 수 있었다. 생산자는 문서 내용의 참석대상자 중 위원장인 서울역사박물관장, 기록식별자는 문화관광1과 동일하게 없었으며, 기술 내용은 문서의 내용 요약이 아닌 첫 문장을 추출하였고 생산일시는 유물수집실무위원회 개최 일시를 추출하였다.

```
>>> raw_text = texts[2]
>>> question = "생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시를 이 문서에서 추출해줘"
>>> qa_document_chain.run(input_document=raw_text, question=question)
'생산자: 서울역사박물관\n기관명: 서울역사박물관\n기록식별자: 없음\n기록물명(제목): 년도 제차 유물수집실무위원회 개최 2023.8\n기술 내용: 관내 대출예정 유물의 보험가격 평가를 위한 제 차 유물수집실무위원회를 다음과 같이 8개최하오니 참석하여 주시기 바랍니다.\n생산일시: 시일 목. : 2023 10 12'
```

<그림 11> 파이썬에서 챗GPT-3.5 터보를 활용한 문화관광2 문서 질문 랭체인 실행 결과



<그림 12> 서울정보소통광장 문화관광2 문서

출처 : 서울정보소통광장 - 결재문서(공개여부 : 공개, 기관 : 전체, 일반문서 : 포함) - 분야 : 문화·관광

나머지 문서 중 문화관광4의 추출 결과에서는 생산자, 기관명, 기록물명, 생산일시를 정확히 뽑아내었다. 생산자와 기관명은 기관명을 추출하였고, 기록물명은 제목을 추출하였으며 생산일시도 접수일시를 정확히 추출하였다. 반면, 기록식별자는 문서의 붙임에 있는 안내문을 답하였으 기술 내용은 문서의 내용을 그대로 서술하였다. 문화관광5의 추출 결과로 생산자, 기록식별자, 생산일시만 정확히 뽑아내었다. 생산자는 기관명으로 추출하였으며 기록식별자는 시행 문서번호를, 생산일시는 접수일자로 정확히 추출하였다. 기관명은 제목의 서울시립 미술아카

이브를 추출하였고 기록물명은 답변이 끊겼으며 기술 내용은 문서의 내용 요약이 아닌 첫 문장을 추출하였다. 이외 문화관광3 문서에서는 앞서 언급했듯이, 해당 요소를 찾을 수 없다는 답변을 하였다.

다음으로 행정·기타 분야의 문서 5건을 이전과 동일하게 챗GPT-3.5 터보로 질문한 결과, 이번에는 행정기타2, 5 문서 2건에 대해서는 결과를 제시하지 못하였으며, 행정기타1, 3, 4 문서 3건의 결과만 제공하였다. 먼저 행정기타1의 메타데이터 추출 결과에서는 기관명, 기록물명을 정확히 뽑아내었다. 생산자의 경우, 내부 결재이기 때문에 기관명 대신 결재자의 직급과 성명을 함께 추출하였으나 문서상 직급과 성명을 모두 추출하지는 않았다. 이를 통해 생산자 및 결재자를 각각 하나씩 빠트린 점만 제외하면 다른 요소에 비해 정확하지만 결과를 제시하지 못한 것으로 간주하였다. 기록식별자는 우편번호와 주소를, 기술 내용은 문화관광2와 같이 문서의 첫 문장을 요약하였고, 생산일시 또한 소상공인 지원 관련 현안업무 검토 간담회의 개최 일시를 추출하였다. 행정기타3의 추출 결과에서는 생산일시만 정확히 뽑아내었다. 생산자 및 기관명은 노원구청으로 추출하였으며 기록식별자와 기술내용은 없음, 기록물명은 답변이 끊겼다. 행정기타4의 추출 결과는 기록식별자, 기록물명, 생산일시를 정확하게 제공하였고, 기관명은 서울특별시로 답하였고, 기술 내용은 앞서와 같이 문서의 첫 문장을 요약하였다. 생산자는 행정기타1과 같이 내부 결재로 결재자의 직급과 성명 중 각각 하나씩 빠트렸으나 다른 요소에 비하면 알맞은 결과를 보였으며 이 또한 정확한 결과를 제시하지 못한 것으로 간주하였다.

각 5건씩 저장한 문화·관광, 행정·기타 분야의 문서에 질문 랭체인을 통해 필수 메타데이터 요소를 지정해서 질문한 결과, 챗GPT의 판단에 따라 해당 요소를 문서에서 추출하였으나 그 결과가 대부분 정확한 것은 아니었다. 기록물명이 가장 정확도가 높았으며 다음으로 생산자, 생산일시, 기관명 순으로 정확하였고, 기록식별자, 기술 내용이 가장 정확도가 낮았다. 이는 챗GPT가 기록물의 메타데이터에 관해 학습한 학습데이터의 양이 다른 분야의 학습데이터에 비해 상대적으로 적기 때문에 생산자나 기관명 등의 요소들이 명확히 제시되어 있음에도 불구하고 문서의 내용을 제대로 읽지 못 하거나 해당 요소에 대한 판단이 어려워 다른 요소를 대답한 것으로 추측할 수 있었다. 또한 질문에 대한 답변을 아예 하지 못한 경우도 있어 결과를 파악하기 어려운 한계점이 있었다. 행정기타 1, 4의 생산자의 경우, 다른 요소에 비하면 알맞은 결과이나 이는 정확한 결과가 아니므로 틀린 것으로 표시하였다. 10건의 문서로부터 추출해야 하는 메타데이터 요소의 개수는 총 60개로 기술 내용 요소를 제외하면 총 50개이며 이 중 16개만 정확히 추출한 것을 <표 3>을 통해 알 수 있다.

<표 3> 랭체인에서의 메타데이터 요소 추출 결과 비교

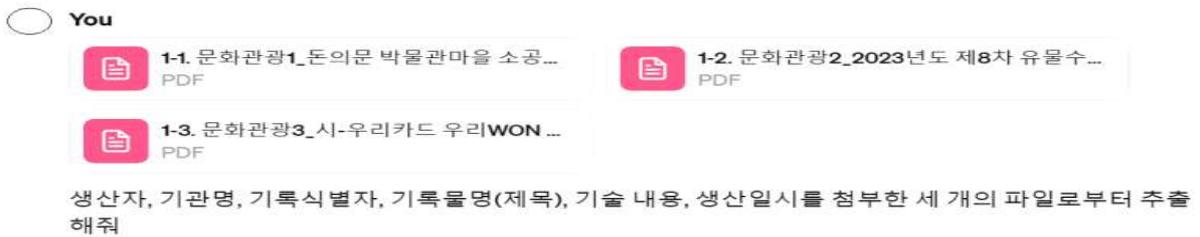
메타데이터 요소	문화1	문화2	문화3	문화4	문화5	행정1	행정2	행정3	행정4	행정5
생산자	X	X		O	O	X		X	X	
기관명	X	O		O	X	O		X	X	
기록식별자	X	X	X	X	O	X	X	X	O	X
기록물명(제목)	O	O		O	X	O		X	O	
생산일시	X	X		O	O	X		O	O	
기술 내용										

3.3 챗GPT-4에서의 「기록관리 메타데이터 표준(v2.3)」의 필수 요소 추출

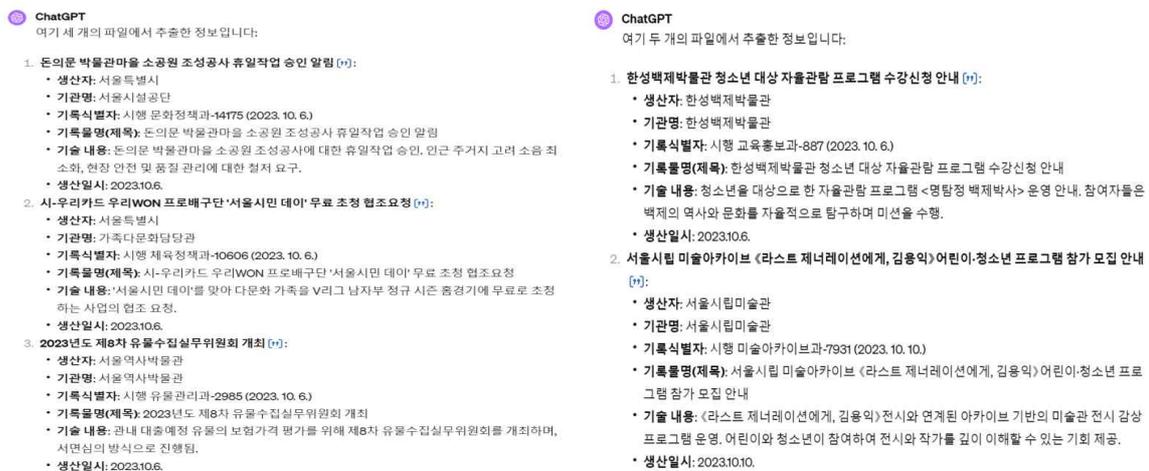
온라인 유료 서비스인 챗GPT-4는 파일 첨부 기능이 탑재되어 사용자가 여러 개의 텍스트 파일을 첨부하면 그것을 읽고 검색하여 해당 파일의 내용에 기반한 질문에 대답할 수 있다. 따라서 챗GPT-4를 사용하여 서울정보소통광장의 문서들로부터 메타데이터 요소를 정확히 뽑아내는지 시도해보았다. 챗GPT-4의 학습에 사용한 파일은 앞서 파이썬 구현에서 사용한 파일과 동일한 문서 10건이다. 챗GPT-4 온라인 서비스는 파이썬과 달리 여러 문서를 한 번에 첨부할 수 있는 기능으로 최대 10건까지 학습할 수 있다. 다만 파일의 용량 및 내용의 양에 따라

실행할 수 있는 문서의 양도 달라져, 각 분야의 문서 5건을 3건, 2건으로 나눠 첨부하여 질문하고자 하였다. 질문은 파이썬에서 실행한 것과 같이 “생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시”를 첨부한 문서에서 추출해줘”라고 하였다.

먼저 <그림 13>와 같이 챗GPT-4 온라인 서비스에 문화·관광 분야의 문서를 첨부하여 질문하였고 <그림 14>와 같이 메타데이터 요소를 추출하였다. 첨부한 문서들로부터 메타데이터 요소를 성공적으로 추출할 수 있었으며 몇몇 답안을 제외하고 대부분 제시한 요소들에 대해 정확한 답을 제공하였다. 생산자와 기관명은 문서를 작성한 기관으로 답하였으며, 주관적으로 작성될 수 있는 기술 내용 요소 또한 문서의 내용을 잘 요약하였다. 이를 통해 챗GPT-4에 문서에 작성된 내용을 정확히 읽고 찾아내어 추출하는 기능이 있는 것을 알 수 있었다.



<그림 13> 챗GPT-4 문서 첨부 및 지정 메타데이터 요소 추출 요청



<그림 14> 챗GPT-4의 문화관광 5개 문서 메타데이터 요소 추출 결과

다음으로 <그림 15>는 행정·기타 분야의 5개 문서를 첨부하여 질문한 결과이다. <표 4>를 통해 이번 구형 결과에서도 기관명을 제외한 5개의 요소들에 대해 알맞은 답변을 제공한 것을 알 수 있다. 특히 행정·기타 분야의 문서들은 동일한 구조의 내부결재 형식의 문서로 기관명이 정확히 명시되어 있음에도 불구하고 이를 인식하지 못하여 5건의 결과 모두 문서를 작성한 부서 또는 문서시행번호의 부서명을 추출한 것으로 추측할 수 있었다. 생산일시 또한 5개의 답변 중 1개를 제외한 나머지는 정확한 것을 알 수 있었다. 이로써 추출해야 할 메타데이터 총 60개 중 기술 내용 요소 10개를 제외하고 42개의 정확한 요소를 추출한 챗GPT-4는 랭체인에 비해 질문에 대한 정확한 결과를 제공하는 것을 확인할 수 있었다.

ChatGPT
여기 세 개의 파일에서 추출한 정보입니다.

- 소상공인 지원 관련 현안업무 검토 [7]:**
 - 생산자: 서울특별시
 - 기관명: 노동정책담당관
 - 기록식별자: 시행 노동정책담당관-12067 (2023. 10. 10.)
 - 기록물명(제목): 소상공인 지원 관련 현안업무 검토
 - 기술 내용: 소상공인 지원 관련 현안업무 검토 간담회 개최 및 관련 비용 지급 안내.
 - 생산일시: 2023.10.10.
- 2023년 제4기 국정 현안 법제의 이해과정(공통) 교육 신청 안내 [7]:**
 - 생산자: 상수도사업본부
 - 기관명: 교육협력과
 - 기록식별자: 시행 교육협력과-9949 (2023. 10. 10.)
 - 기록물명(제목): 2023년 제4기 국정 현안 법제의 이해과정(공통) 교육 신청 안내
 - 기술 내용: 중앙행정기관, 지방의회 공무원, 공공기관 직원 대상 교육 신청 안내 및 세부 사항 제공.
 - 생산일시: 2023.10.10.
- 산림인접지역 쓰레기 불법소각 및 불피움에 따른 행정처리 절차 알림 [7]:**
 - 생산자: 노원소방서
 - 기관명: 재난관리과
 - 기록식별자: 시행 재난관리과-5830 (2023. 10. 10.)
 - 기록물명(제목): 산림인접지역 쓰레기 불법소각 및 불피움에 따른 행정처리 절차 알림
 - 기술 내용: 산림인접 지역에서의 쓰레기 불법소각 및 불피움 방지를 위한 행정처리 절차 안내.
 - 생산일시: 2023.10.10.

ChatGPT
여기 두 개의 파일에서 추출한 정보입니다.

- 서울시미술관 설계 관련 간담회 비용 지급 [7]:**
 - 생산자: 서울특별시
 - 기관명: 박물관과
 - 기록식별자: 시행 박물관과-10483 (2023. 10. 10.)
 - 기록물명(제목): 서울시미술관 설계 관련 간담회 비용 지급
 - 기술 내용: 서울시미술관 설계 관련 간담회 개최 및 관련 비용 지급 안내. 비용은 53,000원으로, 이스트빌리지 일도씨탈갈비에서 사용되었으며, 박물관과장 배회정 포함 4명이 참석.
 - 생산일시: 2023.10.6.
- 서울시립미술관 대관(1층 회의실) 협조 요청 [7]:**
 - 생산자: 서울특별시의회
 - 기관명: 기획경제전문위원회
 - 기록식별자: 시행 기획경제전문위원회-2834 (2023. 10. 10.)
 - 기록물명(제목): 서울시립미술관 대관(1층 회의실) 협조 요청
 - 기술 내용: 도시농업 활성화 방안 연구를 위한 입법정책 연구용역 개최 목적으로 서울시립미술관 1층 회의실 대관 요청. 대관일시는 2023.10.12, 참석인원은 15여명.
 - 생산일시: 2023.10.10.

<그림 15> 챗GPT-4의 행정기타 5개 문서 메타데이터 요소 추출 결과

<표 4> 챗GPT-4에서의 메타데이터 요소 추출 결과 비교

메타데이터 요소	문화1	문화2	문화3	문화4	문화5	행정1	행정2	행정3	행정4	행정5
생산자	O	O	O	O	O	O	O	O	O	O
기관명	X	O	X	O	O	X	X	X	X	X
기록식별자	O	O	O	O	O	O	O	O	O	O
기록물명(제목)	O	O	O	O	O	O	O	O	O	O
생산일시	O	O	O	O	O	O	O	O	X	O
기술 내용										

3.4 기록관리 메타데이터 추출 결과 분석

지금까지 랭체인과 온라인 서비스를 각각 활용하여 「기록관리 메타데이터 표준(v2.3)」의 필수 요소들을 지정하여 챗GPT로부터 추출을 시도하였다. 랭체인과 온라인 서비스의 메타데이터 추출 결과를 쉽게 비교·분석하기 위해 각 문서별로 표를 작성하였고, 추출 결과의 정확도를 파악하기 위해 해당 문서를 보고 직접 판단한 요소들을 추가하여 비교하였다. 이 중 기술 내용 요소는 사람과 기계 모두 요약하는 방식이 다르고 각자의 주관이 들어가기 때문에 표로는 작성했지만 비교 대상에서는 제외하였다. 또한 편의를 위하여 결과 비교에는 챗GPT-3.5 터보를 활용하여 질문한 문화관광1, 2 문서와 행정기타1, 4 문서의 추출 결과를 활용하였다.

<표 5>는 문화관광1 문서의 결과 비교표이다. 여기에서 챗GPT 학습 환경의 표기는 직접 추출하여 작성한 요소는 대상 PDF 파일명으로, 파이썬에서 사용한 챗GPT-3.5 터보 버전은 랭체인 (챗GPT-3.5)로, 온라인 서비스는 온라인 (챗GPT-4)로 표시하였다. 이 중 가장 정확한 답변을 제시한 요소는 기록물명으로 3개의 답변 모두 일치하였다. 기록식별자, 생산일시는 직접 작성한 요소와 GPT-4의 답변이 동일하므로 온라인 서비스에서 랭체인에 비해 정확히 추출했음을 알 수 있었다. 랭체인에서의 결과는 기록물명을 제외하고 모든 답변이 일치하지 않았다.

<표 5> 랭체인과 온라인 서비스의 문화관광1 문서 메타데이터 추출 내용 비교

메타데이터 요소	문화관광1	랭체인 (챗GPT-3.5)	온라인 (챗GPT-4)
생산자	서울특별시장 (또는 주무관 이재익, 문화공간운영팀장 김은경, 문화정책과장 전재명)	서울시설공단 이사장공사감독처장	서울특별시
기관명	서울특별시	서울시설공단	서울시설공단
기록식별자	문화정책과-14175	3-8168 (2023.10.5.)	시행 문화정책과-14175 (2023. 10. 6.)
기록물명(제목)	돈의문 박물관마을 소공원 조성공사 휴일작업 승인 알림	돈의문 박물관마을 소공원 조성공사 휴일작업 승인 알림	돈의문 박물관마을 소공원 조성공사 휴일작업 승인 알림
기술 내용	2023년 10월 7일부터 9일까지 돈의문 박물관마을 소공원 조성공사의 휴일 작업 안내	귀처의 돈의문 박물관마을 소공원 조성공사 휴일작업 토월을 승인함을 알림. 현장안전관리...	돈의문 박물관마을 소공원 조성공사에 대한 휴일작업 승인. 인근 주거지 고려 소음 최소화 ...
생산일시	2023.10.06	2023.10.05.	2023.10.6.

<표 6>은 문화관광2 문서의 결과 비교로 기관명 요소의 3개의 답변 모두 일치하였다. 기록물명 요소의 결과 중, 랭체인과 온라인 서비스의 답변에서 숫자가 따로 추출되어 뒤로 빠져있으나, 그 점만 제외하면 3개의 답변이 똑같다고 볼 수 있다. 생산자는 랭체인과 GPT-4가 서울역사박물관장 및 서울역사박물관으로 답하였으나 이는 문서의 내용에서 추출한 유물수집실사무위원회의 위원장 및 기관명으로, 문서의 생산자가 아니기 때문에 두 버전 모두 틀린 요소를 답변한 것을 알 수 있었다. 기록식별자, 생산일시는 직접 입력한 요소와 GPT-4의 답변이 동일하여 정확히 추출했음을 보여주었다. 랭체인의 경우 기록식별자는 찾지 못하였으며, 생산일시는 해당 유물수집실사무위원회의 행사 진행 일시를 찾아 답변하였다.

<표 6> 랭체인과 온라인 서비스의 문화관광2 문서 메타데이터 추출 내용 비교

메타데이터 요소	문화관광2	랭체인 (챗GPT-3.5)	온라인 (챗GPT-4)
생산자	학예연구부장 (또는 주무관 나신균, 유물관리과장 홍승주, 학예연구부장 박현욱)	서울역사박물관장	서울역사박물관
기관명	서울역사박물관	서울역사박물관	서울역사박물관
기록식별자	유물관리과-2985	없음	시행 유물관리과-2985 (2023. 10. 6.)
기록물명(제목)	2023년도 제8차 유물수집실사무위원회 개최	년도 제차 유물수집실사무위원회 개최 2023 8	2023년도 제8차 유물수집실사무위원회 개최
기술 내용	2023년 10월 12일 목요일에 서면심의로 제8차 유물수집실사무위원회 개최 안내	관내 대출예정 유물의 보험가격 평가를 위한 제 차 유물수집실사무위원회를 다음과 같이 8개최하오니...	관내 대출예정 유물의 보험가격 평가를 위해 제8차 유물수집실사무위원회를 개최...
생산일시	2023.10.06	2023.10.12	2023.10.6.

<표 7>은 행정기타1 문서의 메타데이터 추출 결과로 기록물명이 3개의 답변 모두 일치하는 요소였다. 생산자, 기록식별자, 생산일시는 앞서와 같이 직접 작성한 요소와 GPT-4의 답변이 동일하여 GPT-4의 정확한 추출 기능을 보여주었다. 랭체인에서 추출한 생산일시의 경우에는 2023년 10월 10일로 동일해보이지만, 이는 결재문서의

내용 중 소상공인 지원 관련 현안업무 검토 간담회의 개최 일시로 문서의 생산일시를 추출한 것이 아니기 때문에 GPT-4의 추출 결과만 정확하다고 볼 수 있다. 랭체인이 추출한 기록식별자의 경우, 해당 기관의 주소를 찾아 답변하였으며 생산자의 경우, 문서 결재자의 직급 및 성명을 대답하였으나 일부만 추출하였다. 또한 기관명은 랭체인이 답변이 정확하고 GPT-4의 답변은 결재자 또는 문서시행번호로부터 추출한 것을 알 수 있다.

<표 7> 랭체인과 온라인 서비스의 행정기타1 문서 메타데이터 추출 내용 비교

메타데이터 요소	행정기타1	랭체인 (챗GPT-3.5)	온라인 (챗GPT-4)
생산자	서울특별시 (또는 주무관 박윤희, 노동정책팀장 장선경, 노동정책담당관 조완석)	박윤희 노동정책팀장 장선경 노동정책담당관	서울특별시
기관명	서울특별시	서울특별시	노동정책담당관
기록식별자	노동정책담당관-12067	우04514서울특별시 중구...	시행 노동정책담당관-12067 (2023. 10. 10.)
기록물명(제목)	소상공인 지원 관련 현안업무 검토	소상공인 지원 관련 현안업무 검토	소상공인 지원 관련 현안업무 검토
기술 내용	2023년 10월 10일에 유럽분식집에서 소상공인 지원 관련 현안업무 검토 간담회 개최	소상공인 지원 관련 현안업무 검토 간담회 개최 및 비용 지급	소상공인 지원 관련 현안업무 검토 간담회 개최 및 관련 비용 지급 안내.
생산일시	2023.10.10	2023.10.10. 11:42	2023.10.10.

마지막으로 <표 8>은 행정기타4 문서의 메타데이터 추출 결과 비교표이다. 기록식별자, 기록물명 요소에 대한 3개의 답변이 모두 일치하였고, 특히 기록식별자는 비교한 4건의 문서 중 처음으로 정확하게 추출한 결과를 보여 주었다. 생산자는 직접 작성한 메타데이터 요소와 챗GPT-4의 답변이 일치하였으며 기관명은 세 답변 모두 일치 하지 않았다. 이번 결과에서도 랭체인이 생산자 답변은 결재자의 직급과 성명을 추출하긴 하였으나 일부 직급 및 성명을 제외하여 제공하였다. 또한 GPT-4의 생산일시 답변은 문서 내용에 포함된 일시를 추출하였다.

<표 8> 랭체인과 온라인 서비스의 행정기타4 문서 메타데이터 추출 내용 비교

메타데이터 요소	행정기타4	랭체인 (챗GPT-3.5)	온라인 (챗GPT-4)
생산자	서울특별시 (또는 주무관 조서영, 박물관정책팀장 노은영, 박물관과장 배희정)	조서영, 박물관정책팀장 노은영, 박물관과장	서울특별시
기관명	서울특별시	서서울미술관	박물관과
기록식별자	박물관과-10483	박물관과-10483	시행 박물관과-10483 (2023. 10. 10.)
기록물명(제목)	서서울미술관 설계 관련 간담회 비용 지급	서서울미술관 설계 관련 간담회 비용 지급	서서울미술관 설계 관련 간담회 비용 지급
기술 내용	2023년 10월 6일 서서울미술관 설계 관련 간담회를 개최한 후, 사용한 비용 지급	간담회 비용 지급 관련 사항	서서울미술관 설계 관련 간담회 개최 및 관련 비용 지급 안내. ...
생산일시	2023.10.10	2023.10.10	2023.10.6.

다음으로 구현한 10개 문서의 모든 추출 결과를 한 눈에 비교하기 위한 표를 작성하였다. 분야별 문서의 추출 결과를 비교하기 위해 <표 9>와 <표 10>로 작성하였으며, 문서의 내용과 결과가 일치하는지 추출 결과만 비교 하기 위해 직접 작성한 요소는 제외하여 작성하였다. 또한 앞서 작성한 표에서도 기술 내용 요소의 비교를 제외하였기 때문에 이번 표에서도 제외하였다. 따라서 분야별 문서 5개로부터 추출한 요소의 전체 개수는 총 30개로

기술 내용 요소를 제외한 25개의 요소를 기준으로 비교하였다.

앞서 제시된 <표 3>과 <표 4>는 전체 10개 문서에 대한 랭체인에서의 챗GPT-3.5 터보의 구현 결과와 챗GPT-4 결과를 각각 보여주었는데, 이를 종합하여 제시하면 다음 <표 9>, <표 10>과 같다. 먼저 <표 9>는 문화관광 문서로부터 추출한 결과로 랭체인 (챗GPT-3.5)은 25개 중 10개 요소를 정확히 추출하였으며 챗GPT-4는 23개의 요소를 정확하게 추출한 것을 알 수 있다. <표 10>은 행정기타 문서로부터 추출한 결과로 랭체인 (챗GPT-3.5)은 25개 중 6개 요소를 정확히 추출하였으며, 챗GPT-4는 19개의 요소를 정확하게 추출한 것을 알 수 있다. 따라서 총 60개 요소 중 기술 내용 요소를 제외한 50개를 기준으로 랭체인 (챗GPT-3.5)은 16개, 챗GPT-4는 42개의 요소를 정확히 제시한 결과를 확인할 수 있다. 따라서 <표 9>와 <표 10>을 통해 챗GPT-4 온라인 서비스에서 추출한 메타데이터 요소는 비교적 정확하며 GPT-4 기술의 기능에 대해 알 수 있었고, 랭체인에서의 챗GPT-3.5 터보의 구현 결과로 추출된 요소는 온라인 서비스에 비해 정확도가 낮은 점을 살펴볼 수 있었다.

<표 9> 랭체인과 온라인 서비스의 문화관광 문서 메타데이터 추출 결과 비교

메타데이터 요소	문화관광1		문화관광2		문화관광3		문화관광4		문화관광5	
	랭체인 (챗GPT-3.5)	온라인 (챗GPT-4)								
생산자	X	O	X	O	X	O	O	O	O	O
기관명	X	X	O	O		X	O	O	X	O
기록식별자	X	O	X	O		O	X	O	O	O
기록물명 (제목)	O	O	O	O		O	O	O	X	O
생산일시	X	O	X	O		O	O	O	O	O
기술 내용	-									

<표 10> 랭체인과 온라인 서비스의 행정기타 문서 메타데이터 추출 결과 비교

메타데이터 요소	행정기타1		행정기타2		행정기타3		행정기타4		행정기타5	
	랭체인 (챗GPT-3.5)	온라인 (챗GPT-4)								
생산자	X	O	X	O	X	O	X	O	X	O
기관명	O	X		X	X	X	X	X		X
기록식별자	X	O		O	X	O	O	O		O
기록물명 (제목)	O	O		O	X	O	O	O		O
생산일시	X	O		O	O	O	O	X		O
기술 내용	-									

4. 기록관리 메타데이터 추출을 위한 챗GPT 활용 시 고려사항 및 시사점

4.1 챗GPT 사용 시 고려사항

본 연구에서는 챗GPT-3.5 터보 버전의 질문 랭체인 및 챗GPT-4 온라인 서비스에서 필수 메타데이터 요소를 지정하여 추출해달라는 질문을 하여 구현을 진행하였으며, 이를 통해 두 버전의 기능 차이가 확연히 구분되는 것을 확인할 수 있었다. 챗GPT-3.5 터보를 활용한 질문 랭체인의 결과에서는 대부분 틀린 답변을 제공하였고,

챗GPT-4 온라인 서비스의 결과가 챗GPT-3.5 터보에 비하면 정확도가 높은 것을 파악할 수 있었다.

챗GPT에서는 프롬프트의 중요성이 매우 강조되기 때문에 챗GPT에게 질문할 때, 모든 버전을 통틀어 명확한 조건을 제시할수록 답변의 정확도가 향상한다. 본 연구에서는 정확한 질문을 위해 메타데이터 표준에서 제시하는 필수 메타데이터 요소 중 챗GPT가 문서로부터 인식하고 추출할 수 있다고 판단한 요소들을 명확히 지정하였다. 파이썬에서는 분야별 문서를 한 건씩 순차적으로 불러와 랭체인을 통해 챗GPT-3.5 터보로 “생산자, 기관명, 기록 식별자, 기록물명(제목), 기술 내용, 생산일시를 이 문서에서 추출해줘”라고 질문하였으며, 챗GPT-4 온라인 서비스에서는 여러 건의 문서를 한 번에 첨부하여 동일한 질문을 실행하였다. 그 결과, 랭체인 및 온라인 서비스 모두 지정한 항목에 해당하는 요소들을 추출하였으나, 간혹 추출한 메타데이터 요소에 대한 답변과 해당 요소가 일치하지 않은 경우를 보였다. 이는 챗GPT가 학습한 데이터 중 기록물의 메타데이터 요소 추출과 관련된 학습데이터의 양이 상대적으로 적어서 문서로부터 해당 요소를 인식하지 못하여 발생하는 문제점으로 추측할 수 있다. 메타데이터 추출 결과로 보아 질문에 대한 이해도는 높은 것을 알 수 있어 기록전문가가 수동적으로 추출한 요소를 작성하는 메타데이터 업무에는 충분히 도움이 될 수 있다. 챗GPT를 사용한 환경에 관계없이 챗GPT에 적절하고 정확한 의도를 담은 프롬프트를 작성하여 질문하면, 질문자가 원하는 방향으로 더 명확하고 유용한 답변을 제공받을 수 있을 것이다. 또한 챗GPT가 추후 기록관리에서의 필수 메타데이터 요소에 대해 더 많고 정확한 정보를 학습한 모델로 발전한다면, 챗GPT를 활용하였을 때 메타데이터 요소의 추출 정확도를 향상시킬 수 있을 것이라 예상된다.

보안 위험이 적은 문서로부터 메타데이터를 추출할 경우, 온라인 서비스에서 여러 건의 문서를 첨부하여 한 번에 다량의 정확한 답변을 얻을 수 있으므로 업무 효율이 높아질 것이다. 다만 온라인 서비스에서는 최대 10건의 문서를 첨부할 수 있으나 이는 문서의 용량 또는 내용의 양에 따라 한 번에 답변할 수 있는 문서의 양과 답변의 정확도가 달라진다. 따라서 한 번에 다량의 문서를 첨부하기보다 문서의 양을 2-3건 정도로 적게 올리고, 질문 또한 정확한 기준을 제시하면 답변의 질이 상승한다는 것을 확인할 수 있었다. 반대로 보안상 중요한 기록의 메타데이터 추출 작업을 해야 할 경우, 사용자의 데이터를 안전하게 보호하는 랭체인을 활용하면 챗GPT가 인식한 문서의 내용이 국내의 챗GPT 서버에 저장되지 않아 보안 노출의 위험을 방지할 수 있다. 이는 온라인 서비스에 비해 한 번에 한 문서만 인식할 수 있는 단점이 있어 업무의 효율성은 비교적 낮아지지만, 개인정보 및 보안 내용의 유출은 방지하고 메타데이터 업무에 도움이 될 수 있는 요소를 추출하는 데 도움이 될 것이다.

앞으로 기록물 관리 기관에서 챗GPT 기술을 활용할 경우, 문서의 보안 수준에 따라 다양한 접근 방식을 고려할 수 있다. 보안상 중요한 문서의 경우, 보안성이 높은 파이썬 환경에서의 랭체인을 통해, 그렇지 않을 경우 편의성 및 접근성, 정확성이 비교적 높은 온라인 서비스의 활용을 권장한다. 챗GPT-3.5 터보를 활용한 랭체인의 구현 결과와 챗GPT-4 온라인 서비스의 결과를 비교하였을 때, 랭체인에서 챗GPT 모델을 불러와 질문하는 것보다 기술적으로 더 발전하고 온라인과 연결된 서비스에서의 결과가 비교적 더욱 유의미하고 정확한 메타데이터 요소를 추출했기 때문이다. 추후 기술의 발전에 따라 활용할 수 있는 랭체인 및 챗GPT 기능이 다양해져 현재 랭체인에서의 기술의 성능과 보안을 더욱 강화할 수 있다면 랭체인에서 챗GPT를 활용하여 기록물로부터 메타데이터를 추출하는 것이 더욱 의미 있을 것이다. 또한 온라인 서비스에서도 보안 유출의 위험을 줄이고 다양한 유형의 기록 및 첨부 파일의 수량 증가 등의 기능이 보완 및 발전하면 기록관리에 큰 영향을 끼칠 수 있을 것이다.

4.2 챗GPT 활용의 시사점

본 연구에서는 기록관리 메타데이터의 중요성을 기반으로, 기록전문가가 메타데이터를 수동으로 추출해온 불편함에 대응하여, 챗GPT를 활용한 기록물에서의 메타데이터 자동 추출의 가능성을 타진하였다. 여러 건의 기록에

대한 메타데이터를 한 번에 처리하여 이에 대한 정보를 입력 및 제공하는 것은 지금까지 기록관에서 상상할 수 없는 일이었다. 게다가 일반 사람들은 구글과 같은 인터넷 검색 엔진에서 모든 정보 단위에 접근하는 것을 선호하는데, 현재 우리나라 대부분의 기록물 관리 기관에서는 이러한 접근조차 거의 이루어지지 않고 있다. 또한 기록에 대한 정보는 기록 자체에 대한 메타데이터나 기술 정보 등이 제한적으로 제공되고 있으며, 극히 일부 원문 기록만이 제공되고 있는 것이 현실이다.

디지털화된 기록의 경우, AI OCR 등의 기술을 사용하여 기록 자체에 포함된 텍스트가 확인되면 챗GPT 기술을 활용해 필수 메타데이터 요소들을 추출할 수 있다. 본 연구를 통해 챗GPT를 활용하여 기록관리 메타데이터의 추출 가능성을 확인해 봄으로써 기록에 대한 접근성 향상을 통해 기록물 관리 기관에서의 고질적인 문제였던 검색도구의 한계와 불편함 등을 완화할 수 있는 가능성도 기대할 수 있을 것이다. 앞으로 챗GPT와 같은 생성형 AI 기술이 발전함으로써 기록관리의 오랜 과제였던 메타데이터, 검색도구 등에 이를 활용할 수 있을 것이라 예상된다. 또한 여러 최첨단 기술을 활용하여 기록관리에서는 이전보다 기록전문가의 노고를 줄여주고 업무의 효율성을 높여줄 많은 연구 및 방안이 제시될 수 있을 것이다. 나아가 다양한 유형의 기록으로부터 풍부한 메타데이터 요소를 사람보다 빠르고 정확하게 추출한 정보를 제공할 수 있다면, 이는 기록관리의 패러다임을 변화시키는 데에 기여할 수 있을 것이다. 챗GPT가 더 발전함에 따라 메타데이터 업무는 물론 다양한 기록관리 업무에서 더욱 유의미한 결과를 도출해낼 수 있을 것이라 예상된다. 따라서 이를 통해 앞으로 기록관리에서 업무의 효율성 및 생산성 증대 효과를 가져올 수 있을 것으로 전망한다.

5. 결론

본 연구에서는 기존의 수동적인 메타데이터 업무를 효율적으로 처리할 수 있도록 최근 빠르게 발전하고 있는 신기술인 챗GPT를 이용하여 기록관리 분야에서 메타데이터를 추출하는 방안에 대한 연구를 진행하였다. 기록관리에서 메타데이터는 기록의 맥락, 내용, 구조 등을 파악하는 데 사용할 수 있으며, 기록물을 적절하게 관리하고 이해하는데 매우 중요한 역할을 한다. 「기록관리 메타데이터 표준(v2.3)」에서 제시한 메타데이터의 개념을 종합적으로 분석 및 정리한 후, 필수 메타데이터 요소를 선정하여 챗GPT의 프롬프트에 사용하여 챗GPT의 기록관리에의 활용 가능성을 확인하고자 하였다.

챗GPT는 최근 새롭게 등장한 AI 신기술로 현재 챗GPT-3.5 터보와 챗GPT-4 버전이 출시된 상태이다. 챗GPT-3.5 터보는 대화형 인공지능 애플리케이션을 구축하기 위한 API에 적용된 모델로, 기존 모델인 GPT-3.5보다 자연스러운 대화가 가능하며 API 사용 시 기업 내부에서 입력한 데이터 등의 유출 문제를 방지할 수 있다. 챗GPT-4는 유료로 사용할 수 있고 기존 모델에 비해 높은 자연어 처리 능력을 확보하여 더욱 정확한 정보를 제공한다. 본 연구에서는 구현을 위해 챗GPT를 활용하기 위한 기술인 랭체인 라이브러리를 사용하였는데, 이는 대규모 언어 모델을 통해 사용할 수 있는 오픈 소스 라이브러리로 챗GPT에 다양한 기술을 연결할 수 있고 사용자의 데이터를 보호하는 기능이 있다는 장점을 갖는다.

구현에 사용한 문서는 국가기록원에서 문서보존포맷으로 제시하고 있는 PDF/A 포맷 형태의 문서로, 예시로는 서울정보소통광장에서 제공하는 결재문서 중 기록건에 해당하는 문서로 선정하였다. 파이썬 환경에서 챗GPT-3.5 터보를 활용한 랭체인과 챗GPT-4 온라인 서비스 환경에서 동일 PDF 문서의 메타데이터를 추출해 보았다. 랭체인에서의 메타데이터 추출 구현에는 문서를 하나씩 사용할 수밖에 없는 단점이 있으나 보안상 중요한 기록으로부터 메타데이터를 추출할 경우, 기록의 내용이 서버에 저장되지 않으므로 내용 유출을 방지할 수 있는 장점이 있다. 반면 보안 내용을 포함한 기록이 아닐 경우, 온라인 서비스를 통해 메타데이터를 추출하면 여러

건의 기록으로부터 필요한 메타데이터 요소를 한 번에 추출할 수 있어 메타데이터 업무의 효율성이 높아지는 장점이 있다.

구현 결과, 온라인 서비스의 학습 결과가 비교적 더욱 정확한 사실을 확인할 수 있었다. 챗GPT를 활용하여 랭체인으로 질문했을 때는 한 번에 하나의 문서를 불러와 실행하였으며, 질문에 상세한 조건을 추가함에 따라 답변의 정확도가 향상되었으나, 질문에 대한 요소를 문서에서 찾아내지 못하거나 해당 요소를 잘못 답변한 경우가 대다수로 추출해야 할 총 60개의 요소 중 기술 내용 요소 10개를 제외하고 정확하게 뽑아낸 요소는 17개에 불과하였다. 온라인 서비스에서는 한 번에 여러 건의 문서를 첨부하여 메타데이터를 추출할 수 있었고, 총 60개의 요소 중 8개의 요소만 잘못된 답변으로 랭체인에 비해 비교적 정확한 요소를 추출했음을 알 수 있었다. 이러한 결과를 통해 기록관리에서의 메타데이터 추출을 위한 챗GPT 기술 활용의 가능성을 타진해볼 수 있었다.

지금까지는 기록전문가의 기록관리 메타데이터에 대한 접근이 수동적으로 이루어져왔으나, 본 연구에서 제시한 메타데이터 추출 방안을 업무에 활용한다면 기록관리 업무의 효율성을 증대시킬 수 있을 것으로 보인다. 기록관에서 기록전문가가 하나씩 처리해야 했던 기록물을 챗GPT-4의 다수 문서 첨부 기능을 통해 일괄적으로 처리가 가능해지고, GPT가 발전함에 따라 더 많은 양의 문서 첨부가 가능해지면 챗GPT의 장점을 활용하여 기록관리 업무의 생산성 및 효율성을 더욱 향상시킬 수 있을 것이다. 보안 문서를 다룰 경우, 피어션 환경에서 랭체인으로 문서를 하나씩 불러와 정해진 질문을 하고 답을 얻음으로써 이전에 비해 기록전문가의 업무 복잡성을 상당히 줄여줄 수 있을 것이다. 앞으로 성능이 향상된 챗GPT 버전을 활용하여 랭체인 환경으로 구현하여 메타데이터 추출을 한다면 보안이 확보된 정확한 메타데이터 추출 결과를 얻을 수 있을 것이다.

다만, 현재 챗GPT가 전자문서로부터 인식할 수 있는 메타데이터 요소가 한정되어 있어 본 연구에서는 ‘생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시’ 총 6개 요소의 추출을 시도하였으며, 많은 양의 데이터를 처리할 수 없는 기술의 한계로 인해 10건의 문서를 추출에 사용하여 더 많은 양의 문서를 사용할 수 없었던 점이 가장 큰 한계점으로 남았다. 또한 랭체인의 기술적 한계로 피어션 환경에서는 학습할 구현 문서의 양을 1개씩만 인식할 수밖에 없었으며, 온라인 서비스에서는 답변의 정확도를 위해 최대 첨부 개수 10개 중 2-3개로 제한해 첨부하여 더 많은 개수의 문서를 구현에 사용하지 못한 점에서 한계가 있었다. 챗GPT가 문서에서 인식할 수 없는 필수 메타데이터 요소가 많아 생산자, 기관명, 기록식별자, 기록물명(제목), 기술 내용, 생산일시 요소만 추출을 시도했고 이 결과 또한 완벽하지 못한 점도 한계로 남는다. 그러나 이러한 점들은 챗GPT와 관련 기술의 발전에 따라 해결될 수 있을 것이라 본다. 챗GPT의 진보에 따라 챗GPT를 활용한 메타데이터 추출 작업은 향후 사람보다 더욱 빠르고 정확해질 것이며, 다양한 매체 및 형태의 기록으로부터 메타데이터 요소를 추출하는 것이 가능할 것으로 예상된다. 또한 기록에 대한 접근성이 보다 쉬워져 메타데이터 추출은 물론 기록 내용 요약의 기술에의 활용, 이미지 인식, 검색까지 가능해져 챗GPT의 활용이 기록관리 발전에 의미가 있을 것으로 전망한다.

앞으로 기록관리에서 챗GPT를 결합하여 사용할 수 있는 기술 및 플랫폼이 더 많이 출시되어 기록관리 분야에 도움이 되길 바란다. 아직 기록관리 분야에서 자동으로 메타데이터 요소를 추출하기 위해 챗GPT를 활용하는 방안은 찾아보기 어렵다. 본 연구를 시작으로 추후 발전한 챗GPT 기술을 통해 훨씬 많은 양과 다양한 유형의 문서 활용 방안이 제시되어 기록전문가의 업무 효율성 및 생산성과 기록의 접근성을 높여주는 데 기여할 수 있기를 기대한다. 또한 메타데이터 업무뿐만 아니라 기록관리 전반의 다양한 업무에서 챗GPT를 활용할 수 있는 방안이 제시될 수 있기를 희망한다.

참고문헌

- 강윤아, 오효정 (2023). 전자기록관리 업무 및 기록정보서비스에서의 생성형 AI 기술 활용. 한국기록관리학회지, 23(4), 179-200. <https://doi.org/10.14404/JKSARM.2023.23.4.179>
- 국가기록원 [발행년불명]. 전자기록관리 장기보존 포맷. 출처: <https://www.archives.go.kr/next/manager/electronicPreservation.do>
- 기록관리 메타데이터 표준(v2.3). NAK 8:2022(v2.3).
- 김수민, 백선환 (2023). 챗GPT 거대한 전환. 서울: 알에이치코리아.
- 김인아, 강영선, 이규철 (2020). 국내 지자체 사진 기록물의 효율적 관리를 위한 메타데이터 설계 및 기계학습 기반 자동 인덱싱 방법 연구. 한국기록관리학회지, 20(2), 67-83. <https://doi.org/10.14404/JKSARM.2020.20.2.067>
- 남용우 (2013). 기록 디지털 컴포넌트의 메타데이터 추출도구 설계. 석사학위논문, 명지대학교 기록정보과학전문대학원 기록관리전공.
- 서울정보소통광장 (2023. 10. 10.). 2023년 제4기 국정 현안 법제의 이해과정(공통) 교육 신청 안내. 출처: <https://opengov.seoul.go.kr/sanction/29414764>
- 서울정보소통광장 (2023. 10. 06.). 2023년도 제8차 유물수집실무위원회 개최. 출처: <https://opengov.seoul.go.kr/sanction/29400803>
- 서울정보소통광장 (2023. 10. 06.). 돈의문 박물관마을 소공원 조성공사 휴일작업 승인 알림. 출처: <https://opengov.seoul.go.kr/sanction/29398817>
- 서울정보소통광장 (2023. 10. 10.). 산림인접지역 쓰레기 불법소각 및 불피움에 따른 행정처리 절차 알림. 출처: <https://opengov.seoul.go.kr/sanction/29415812>
- 서울정보소통광장 (2023. 10. 10.). 서서울미술관 설계 관련 간담회 비용 지급. 출처: <https://opengov.seoul.go.kr/sanction/29415446>
- 서울정보소통광장 (2023. 10. 10.). 서울시립 미술아카이브 《라스트 제너레이션에게, 김용익》 어린이·청소년 프로그램 참가 모집 안내. 출처: <https://opengov.seoul.go.kr/sanction/29412009>
- 서울정보소통광장 (2023. 10. 10.). 서울시립미술관 대관(1층 회의실) 협조 요청. 출처: <https://opengov.seoul.go.kr/sanction/29408812>
- 서울정보소통광장 (2023. 10. 10.). 소상공인 지원 관련 현안업무 검토. 출처: <https://opengov.seoul.go.kr/sanction/29416535>
- 서울정보소통광장 (2023. 10. 06.). 시-우리카드 우리WON 프로배구단 '서울시민 데이' 무료 초청 협조요청. 출처: <https://opengov.seoul.go.kr/sanction/29398327>
- 서울정보소통광장 (2023. 10. 06.). 한성백제박물관 청소년 대상 자율관람 프로그램 수강신청 안내. 출처: <https://opengov.seoul.go.kr/sanction/29397511>
- 서지인, 노지현 (2022). 디지털화 기록관리를 위한 메타데이터 요소(안) 설계. 한국기록관리학회지, 22(4), 1-24. <https://doi.org/10.14404/JKSARM.2022.22.4.001>
- 안세진, 황현호, 임진희 (2022). 종이기록 데이터화를 위한 AI-OCR 적용 사례연구. 정보관리학회지, 39(3), 165-193. <https://doi.org/10.3743/KOSIM.2022.39.3.165>
- 안중배 (2023). 챗GPT-4 인공지능 미래세상. 파주: 광문각출판미디어.
- 앤미디어 (2023). 챗GPT & AI 활용법. 서울: 성안당.
- 이성용 (2023). 챗GPT&파이썬으로 AI직원 만들기. 서울: 이지스퍼블리싱.
- 조용주, 임좌상 (2023). 프로그래밍 입문을 위한 파이썬의 정석. 서울: 도서출판 길벗.

- 최상미, 이상용 (2007). 현용전자기록물의 메타데이터 요소에 관한 연구. 한국기록관리학회지, 7(1), 39-60. <https://doi.org/10.14404/JKSARM.2007.7.1.039>
- 최은석 (2023). 만들면서 배우는 나만의 인공지능 서비스. 파주: 위키북스.
- 최주호, 이재영 (2012). 전자기록물의 메타데이터 추출 및 비교 검증 기술 연구. 한국기록관리학회지, 12(1), 7-32. <https://doi.org/10.14404/JKSARM.2012.12.1.007>
- 후루카와 히데카즈 (2023). GPT-4, ChatGPT, 라마인덱스, 랭체인을 활용한 인공지능 프로그래밍. (트랜스메이트 옮김). 파주: 위키북스.
- Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. Library Hi Tech News, 40(3), 26-29.

• 국문 참고자료의 영어 표기

(English translation / romanization of references originally written in Korean)

- Ahn, Jong-Bae. (2023). ChatGPT-4: The Future World of Artificial Intelligence. Paju: KwangmoonGak Publishing Media.
- Ahn, Sejin, Hwang, Hyunho, & Yim, Junhee (2022). A Case Study on the Application of AI-OCR for Data Transformation of Paper Records. Journal of the Korean Society for Information Management, 39(3), 165-193. <https://doi.org/10.3743/KOSIM.2022.39.3.165>
- Ann Media (2023). ChatGPT & AI Usage. Seoul: Seongandang.
- Choi, Eunseok (2023). Learning by Creating Your Own AI Service. Paju: Wikibooks.
- Choi, Joo ho & Lee, Jae Young (2012). Extracting and Validating Metadata in Electronic Records. Journal of Korean Society of Archives and Records Management, 12(1), 7-32. <https://doi.org/10.14404/JKSARM.2012.12.1.007>
- Choi, Sang-Mi & Lee, Sang-Yong (2007). A Study on the Elements of Current Electronic Records. Journal of Korean Society of Archives and Records Management, 7(1), 39-60. <https://doi.org/10.14404/JKSARM.2007.7.1.039>
- Hurukawa, Hidekazu (2023). Using GPT-4, ChatGPT, LlamaIndex, and Langchain for AI Programming. Paju: Wikibooks.
- Jo, Yong-ju & Im, Jwa-sang (2023). The Definitive Guide to Python Programming for Beginners. Seoul: Gilbut Publishing Company.
- Kang, Yoona & Oh, Hyo-Jung (2023). The Use of Generative AI Technologies in Electronic Records Management and Archival Information Service. Journal of Korean Society of Archives and Records Management, 23(4), 179-200. <https://doi.org/10.14404/JKSARM.2023.23.4.179>
- Kim Sumin & Baek Sunhwan (2023). ChatGPT: The Great Transition. Seoul: RH Korea.
- Kim, InA, Kang, Young-Sun, & Lee, Kyu-chul (2020). Metadata Design and Machine Learning-Based Automatic Indexing for Efficient Data Management of Image Archives of Local Governments in South Korea. Journal of Korean Society of Archives and Records Management, 20(2), 67-83. <https://doi.org/10.14404/JKSARM.2020.20.2.067>
- Lee, Seongyong (2023). Creating AI Employees with ChatGPT & Python. Seoul: EJIS Publishing.
- Metadata Standard for Records and Archives Management Version 2.3. NAK 8:2022(v2.3).
- Nam, Yongwoo (2013). A Study on Design of Archival Digital Components Metadata Extraction Tool. Master's thesis, Myongji University Graduate School of Records, Archives & Information Science.

- National Archives of Korea [n.d.] National Archives of Korea's Electronic Records Management Long-Term Preservation Format. Available: <https://www.archives.go.kr/next/manager/electronicPreservation.do>
- Seo, Jiin, Rho & Jee-Hyun (2022). A Study on Designing Metadata Elements for the Management of Digitized Records. *Journal of Korean Society of Archives and Records Management*, 22(4), 1-24. <https://doi.org/10.14404/JKSAR M.2022.22.4.001>
- Seoul Information Communication Plaza (2023. 10. 10.). 2023 4th National Policy Issue Legislative Understanding Course (Common) Registration Guide. Available: <https://opengov.seoul.go.kr/sanction/29414765>
- Seoul Information Communication Plaza (2023. 10. 06.). 8th Artifact Collection Practical Affairs Committee Meeting Held in 2023. Available: <https://opengov.seoul.go.kr/sanction/29400804>
- Seoul Information Communication Plaza (2023. 10. 10.). Notification of Administrative Procedures for Illegal Burning of Trash and Lighting Fires in Areas Adjacent to Forests. Available: <https://opengov.seoul.go.kr/sanction/29415813>
- Seoul Information Communication Plaza (2023. 10. 06.). Notification of Approval for Weekend Work on the Small Park Construction Project at Donuimun Museum Village. Available: <https://opengov.seoul.go.kr/sanction/29398818>
- Seoul Information Communication Plaza (2023. 10. 10.). Payment for Costs Related to the Briefing Session on the Design of the West Seoul Art Museum. Available: <https://opengov.seoul.go.kr/sanction/29415447>
- Seoul Information Communication Plaza (2023. 10. 06.). Registration Guide for the Self-Guided Tour Program for Youths at the Hanseong Baekje Museum. Available: <https://opengov.seoul.go.kr/sanction/29397512>
- Seoul Information Communication Plaza (2023. 10. 06.). Request for Cooperation for the Free Invitation to 'Seoul Citizens Day' by the City-Woori Card WooriWON Professional Volleyball Team. Available: <https://opengov.seoul.go.kr/sanction/29398328>
- Seoul Information Communication Plaza (2023. 10. 10.). Request for Cooperation in Renting the Meeting Room on the First Floor of Seoul Museum of Art. Available: <https://opengov.seoul.go.kr/sanction/29408813>
- Seoul Information Communication Plaza (2023. 10. 10.). Review of Current Affairs Related to Support for Small Business Owners. Available: <https://opengov.seoul.go.kr/sanction/29416536>
- Seoul Information Communication Plaza (2023. 10. 10.). Seoul Metropolitan Art Archive 《To the Last Generation, Kim Yong-Ik》 Children and Youth Program Participation Recruitment Notice. Available: <https://opengov.seoul.go.kr/sanction/29412010>