논문

지능형 문서처리 도입과 기록관리 변화에 관한 연구

A Study on the Introduction of Intelligent Document Processing and Change of Record Management

류한조(Ryu, Hanjo)* · 이경남(Lee, Kyungnam)**
황진현(Hwang, Jinhyun)*** · 임진희(Yim, Jinhee)****

* 명지대학교 기록정보과학전문대학원 겸임교수(kalistland@gmail.com)(제1저자).
** 한신대학교 대학원 기록관리학과 강사(coarchivist@gmail.com)(공동저자).
*** 강릉원주대학교 일반대학원 기록관리협동과정 강사(jinhyunii@gmail.com)(공동저자).
**** 명지대학교 기록정보과학전문대학원 조교수(yimjhkr@mju.ac.kr)(교신저자).
■투고일: 2021년 03월 31일 ■최초심사일: 2021년 04월 09일 ■최종확정일: 2021년 04월 20일.
■기록학연구 68, 41-72, 2021, https://doi.org/10.20923/kjas.2021.68.041

〈초록〉

빅데이터 분석을 위해서는 기계가독성을 높이는 개방형 문서 포맷으로의 변화와 자연어 처리 기술 도구가 필요하다.

본 연구는 지능형 문서 처리의 도입 배경과 연구 현황을 공공부문 중심으로 살펴보고, 지능형 문서 처리가 가져올 업무의 변화를 예측해 보았다. 나아가 지능형 문서 처리가 기록관리 업무에 가져올 변화를 전망해보고, 기록관리 전문가의 역할의 변화와 요구되는 역량 등을 고찰해 보았다. 기록관 업무 단계와 아카이브 업무 단계의 광범위한 영역에 걸쳐 기록관리 업무의 변화를 전망하였고, 특히 반복적인 기록관리 업무의 자동화나 기록물의 기술 및 활용 업무에 영향을 미칠만한 변화들을 서술하였다. 이러한 업무 수행의 변화에 맞추어 기록관리계는 새로운 업무 절차와 방법, 그리고 필요한 역량을 준비해야 할 것이다.

주제어 : 지능형 문서 처리, 자연어 처리, 빅데이터, 텍스트 분석, 개방형 표준 포맷, 기계가독성

〈Abstract〉

In order to analyze big data, documents should be converted to a open standard format to increase machine readability. It also need natural language processing tools.

This study focused on the background of intelligent document processing and the status of research in the public sector, and predicted the changes in work that intelligent document processing would bring. This study noted the changes that intelligent document processing would bring to the archival work, and also considered changes in the role of archivist and their required competencies. Changes in archival work could be anticipated across a wide range of Records Management work and Archives Management work. In particular, it was expected to

have a significant impact on the automation of repetitive archival tasks or the description and utilization of records. This study proposed the need to prepare new archival work procedures, methods, and necessary competencies in response to these change in archival work.

Keywords : Intelligent Document Processing(IDP), Natural Language Processing(NLP), big data, text analysis, open standard format, machine readability

1. 서론

아래아 한글로 대표되는 문서편집기와 HWP 파일포맷은 문서를 이용한 업무처리에 오랫동안 활용되어 왔다. 특히 공문서의 경우 “행정 효율과 협업 촉진에 관한 규정”에 의해 전자적으로 처리하는 것을 원칙으로 하여 대다수가 전자문서로 생산되고 있다. 종이기반의 생산, 결재 등의 행위를 소프트웨어 편집기와 전자문서시스템을 통해 대신하는 형태로 자리잡았다.

전자문서는 문서의 생산과 유통에서 큰 장점을 보이지만 몇 가지 문제들이 지적되었다. 대표적 문제로 특정기업에 종속된 포맷의 독점성이 거론되었는데 이러한 문제를 해결하기 위해 ODF나 OOXML과 같은 오픈포맷이 등장하였다(정미리, 오세라, 임진희, 2016). 오픈포맷의 도입은 전자문서에 대한 기록관리 이용가능성을 높여줄 수 있어 긍정적으로 평가할 수 있으나 공공기관의 공문서 표준포맷으로 완전히 정착되지는 못했다.

전자문서는 대표적인 비정형데이터로 활용가능성이 높지 않다. 전자문서의 내용은 자연어 비중이 높으며, 수치화된 데이터가 포함되어 있어도 활용하기 쉬운 형태로 추출하기 어렵다. 따라서 접근점이 될 수

있는 생산자와 제목 등의 정보가 상대적으로 중요하게 여겨졌으며, 문서 활용을 위한 메타데이터에 대한 연구가 발전하였다.

그러나 최근에는 XML 기반포맷의 태그정보 입력, 비정형 데이터 처리 솔루션과 딥러닝 기술이 등장하면서 기계가독형 전환 시도가 일어나고 있다. Amazon이나 구글, IBM과 같은 글로벌 기업의 비정형 데이터 처리를 위한 딥러닝 오픈소스 라이브러리 개발도 활발히 이루어지고 있다. 예를 들어, OMEKA는 Ngram 텍스트 분석 플러그인을 활용하여 자연어 처리(Natural Language Processing, NLP) 기반의 텍스트 분석을 시도하고 있다(Omeka, 2017). 또한 아파치 OpenNLP, 구글 클라우드 Natural Language API, GATE, Carrot2, Gensim 등의 다양한 오픈소스 소프트웨어를 기반으로 한 자연어 처리 툴킷이 개발되고 있다.

자연어 처리 알고리즘을 활용하여 구조화되지 않은 방대한 텍스트를 분석하고 유의미한 결과값을 도출하는 작업은 기업의 시장 분석이나 비즈니스 수익 목적뿐만 아니라, 공공부문의 정책적 의사결정이나 새로운 과제 발굴에 활용할 수 있다.

이들 기술이 적용된 텍스트 분석을 위해서는 문서 내 데이터를 기계가독형(Machine-readable)으로 변환해야 하며 이를 위해 공공데이터의 개방 · 공유 · 활용을 위한 정보 처리 정책이 마련되고 있다. 미국 연방정부에서는 공공데이터를 기계가독이 가능한 형태로 공개하는 것을 의무화한 “Open, Public, Electronic and Necessary (OPEN) Government Data Act”를 2019년 제정하기도 하였다. 이 법은 2013년 버락 오바마 행정부의 오픈 데이터 정책을 기반으로 하며, 정부가 공개하는 데이터는 기계가독형이고 오픈 포맷을 사용하도록 하는 것을 골자로 한다(Data Coalition, 2021).

국내에서도 공공데이터를 기계가독이 가능한 형태로 제공하는 내용을 담은 “공공데이터의 제공 및 이용 활성화에 관한 법률”이 2013년 제정되었고, 공공데이터 포털을 통해 5만여 건의[1] 공공데이터가 공개되

었다. 공공데이터 포털은 주로 CSV와 같은 스프레트시트 포맷 파일 혹은 json형태로 제공하거나, 데이터에 접근할 수 있는 API를 공개하는 형태로 서비스하고 있다. 그러나 여전히 개방된 데이터의 일부는 표준화된 포맷이 아니어서 기계가 읽을 수 없다는 문제점이 지적되고 있다. 세입, 세출, 예산, 결산 자료 등을 CSV포맷이 아닌 pdf포맷으로 변환하여 공개하면 기계가독이 불가능하기 때문에 이를 표준포맷으로 변환하기 위한 번거로운 과정이 필요하다.

그럼에도 공공부문에서의 변화가 감지되고 있다. 기계가독형인 개방형 문서 포맷(Open Document Format for Office Application, ODF)의 필요성을 인지하고 개방형 문서 포맷에 대한 연구를 시작하고 있기 때문이다. 일례로 행정안전부에서는 공공서식의 기계가독형을 위한 공공서식 설계기준안과 디자인 표준안을 제안한 정책연구 “공공서식 디자인 재설계 방안 연구”를 2020년 진행하였다(행정안전부, 2020).

이와 같은 문서의 기계가독성 증진을 위한 환경 변화는 문서 생산 환경 변화를 의미하므로 기록관리 관점에서 주도적 자세로 개선 방안을 제시하는 연구가 필요하다. 국가기록원에서는 “문서 파일 포맷과 서식 개선 방안” 연구세미나를 통해 공문서 서식의 개방형 포맷의 필요성과 표준화된 메타데이터 태그의 필요성에 대해 서울시 사례를 통해 주의를 환기시켰다(국가기록원, 2020). 이 외에도 공문서 파일 포맷 현황을 분석하고, 정부기관 클라우드 업무관리시스템 전환 과정에서 파일 포맷을 새롭게 선택할 수 있게 됨에 따라 개방형 표준 포맷을 유지하는 전략에 대한 연구도 진행되었다(임진희, 2020). 또한 공문서 빅데이터 분석의 필요성과 기계가독형 문서 형식을 위한 자기기술(self-descriptive) 메타데이터의 확보, 문서 텍스트 태깅 등이 제안되기도 하였다(임진희, 2021).

---

1) 공공데이터 포털(data.go.kr)에 파일데이터 38,241개, 오픈API 6,777개, 표준데이터셋 122개(9,969개) 등 54,987건의 데이터를 공개하고 있다. (2021.3.25. 기준)

지능형 기술의 기록관리 분야 적용에 관한 선행연구는 주로 신기술의 적용 가능성을 탐색하는 연구가 진행되었다. 인공지능에 대한 배경정보를 제공하고 인공지능을 텍스트분석이나 영상인식, 음성인식 등 기록관리에 적용 가능한 분야와 도입 방안을 제시한 연구(김인택, 안대진, 이해영, 2017)를 필두로 기록관, 도서관, 박물관의 지능형 서비스 관련 사례를 소개하고 기록관리 분야의 적용을 위한 선결조건에 대해 분석한 연구(김태영 외, 2018)를 참고할 수 있다. 기관에서 사용하는 다양한 유형의 정보시스템과 기록관리시스템을 통합하여 하나의 시스템으로 구현한 사례 연구(주현우, 2019)를 통해서는 지능형 서비스를 적용한 기능 설계 사례를 볼 수 있다. 아울러 국가기록원(2019)에서는 지능형 전자기록관리 자동화 기술 개발을 추진하기 위해 중장기 추진 전략을 수립하기도 하였다. 이 연구를 통해 기록관리 업무에서 자동화 기술 지원이 가능한 기능을 도출하고 단계별 추진 계획과 목표, 성과 등을 제안하였다.

세부 기록관리 업무에 적용하기 위한 연구로 방재현(2018)은 아카이브 시스템에서 기계학습 알고리즘이 적용 가능한 기능을 도출하고, 기록물 평가 및 공개재분류, 이미지 대상으로 태깅 정보를 제공하는 업무에 지능형 기술 적용을 실험하였다. 의사결정 지원 도구로 지능형 기술을 실무에 적용하기 위한 연구로서 의의를 갖는다. 지능형 기술을 기록정보서비스에 활용하기 위한 연구도 진행되었다. 이창희(2018)는 대학기록관을 대상으로 이용자 요구 및 질의 분석을 기반으로 챗봇을 개발한 사례를 제시하였으며, 백지연과 오효정(2019)은 국가기록원의 웹포털서비스를 통해 장기간 누적된 대량의 질의 로그를 입수하여 빅데이터 분석을 통해 이용자의 정보요구를 분석하여 정보서비스에 활용하였다.

빅데이터 분석을 위해서는 문서 텍스트 태깅이 필요하다. 따라서 기계가독형 문서 형식에서 생산자, 생산활동, 문서간의 관계, 소속된 문

서 집합과의 관계 등을 메타데이터로 지정하고 메타데이터 항목간의 관계 설정이 필요하다. 이러한 관점에서 ICA Records in Context(RiC) 개념모델은 각 개체(Entity)의 맥락을 기술하고 상호운용성을 확보하는 기술표준으로서 의의를 갖는다. RiC은 박지영(2017)의 연구에서 기본 개념과 특징이 상세히 다루어졌고, RiC 개념모델을 이용해 실제 관계형 데이터 모델링을 구현한 사례 연구(신미라, 김익한, 2019)와 국가기록원의 영구기록물 기술에 RiC을 연계하여 기술하기 위한 기술요소를 도출한 연구(김수현, 이성숙, 2020)도 진행되었다. ICA에서는 2019년 RiC 개념모형을 개정하여 RiC-CM v0.2를 발표하였고, 이 버전 0.2의 특성을 다룬 연구(전예지, 이혜원, 2020)도 이어졌다.

LOD는 다양한 정보 자원을 연결함으로써 시멘틱 웹을 구현하기 위한 기술적인 방법이자 접근법이다(하승록, 안대진, 임진희, 2017). LOD에 관한 연구로는 공공데이터의 개방과 활용을 높이기 위해 LOD 형태로 개방할 필요성을 다룬 연구(이현정, 남영준, 2014)가 있다. 이후 보다 구체적인 실행 방안 연구가 이어졌다. 오픈소스 소프트웨어를 이용하여 LOD를 구축하는 절차를 5단계의 프레임워크로 제안한 연구에서는 기록관리 분야에 LOD를 구축하기 위한 필요요건들을 제시함으로써 기록정보의 연결과 확장을 위한 시사점을 제공하였다는데 의의가 있다(하승록, 임진희, 이해영, 2017). 또한 빅데이터, 인공지능 등의 환경에서 기록정보자원이 LOD 형태로 다양한 데이터와 상호연결하기 위한 방안을 제시한 연구에서는 국가기록원의 데이터 샘플을 대상으로 LOD를 구축하고 인물정보 데이터베이스와 상호연결을 테스트하여 시사점을 도출하기도 하였다(하승록, 안대진, 임진희, 2017).

문서 파일의 빅데이터 분석은 디지털 시대의 미룰 수 없는 핵심 과제임이 분명하다. 이러한 변화에 조응하여 공문서 형식도 HWP에서 공개표준인 ODF, OOXML, OWPML 등으로 변화하고 있다. 이와 같은 변화

는 문서의 생산 형식이 바뀌는 것을 의미한다. 기계가독성을 향상시키고 빅데이터 분석, 텍스트 분석이 가능해짐에 따라 기록관리 업무 내용과 절차, 방식 등도 변화할 것이다. 기록관리계는 이러한 새로운 전환에 대해 예측하고 준비해야 한다. 특히 아카이브의 역할에 대한 모색, 기록전문가의 역할과 역량에 대해 생각해볼 필요가 있다.

본 연구는 이러한 전환적 관점에서 지능화가 기록관리 업무에 미칠 영향을 분석하고 새로운 역할 수행을 준비해보는데 목적이 있다. 2장에서는 개방형 문서 포맷의 도입 배경과 자연어 처리 기술을 활용한 지능형 문서 처리의 현황에 대해 정리하여 기술적 부분의 이해를 높이고자 하였다. 3장에서는 지능형 문서처리 도입에 필요한 기술적인 요구들을 단계별로 살펴보고 이를 위해 필요한 사항들이 무엇인지 살펴보았다. 그리고 이러한 지능형 문서 처리가 업무수행과정과 기록관리에 어떤 변화를 가져올지, 기록관리계에서는 어떤 준비를 해야 하는지 4장에서 논의하였다.

2. 지능형 문서처리 도입 배경과 과정

1) 개방형 문서포맷 보급

개방형 문서포맷 도입은 2000년대 초반부터 이루어졌다. 지배적 시장 지위를 가진 마이크로소프트에 대한 종속에서 벗어나기 위해 썬마이크로시스템즈가 ODF를 지원하는 오픈오피스 버전 1.0을 출시하였으며, 이후 ODF는 OASIS(Organization for the Advancement of Structured Information Standards)의 주도 하에 2006년 국제표준화기구(ISO)와 국제전기표준회의(IEC)의 인증을 받아 국제 표준(ISO/IEC 26300)으로 제정되었다. 우리

나라에서도 2007년 ODF를 국가기술표준(KS)로 인증, 공공영역에 도입하자는 논의가 있었으나 바로 수용되지 않았다(오세라, 정미리, 임진희 2016. 30). 2016년에야 행정자치부 “정부지식 공유활용기반 고도화” 2차 사업에 공문서에 개방형 문서표준포맷인 ODT와 PDF 등을 적용할 것을 명시(오세라, 정미리, 임진희 2016. 81)하였고 현재 온나라 문서2.0에 ODF와 호환가능한 폴라리스오피스의 웹기안기를 사용하고 있다. 그러나 일부 지방자치단체에서는 2019년 온나라2.1을 구축하면서 한글과컴퓨터 웹기안기를 채택하여 개방형 포맷으로의 완전한 전환으로 보기는 어렵다(임진희, 2020, 28).

개방형 문서포맷은 문서 내용과 형식 등을 XML에 담아 기술하고 문서 내 이미지 등 다른 파일들이 포함된 경우 함께 묶어 압축하여 유통하는 구조를 지닌다. 이는 ①문서파일의 상호호환성을 높일 수 있고, ②내용에 해당하는 텍스트가 별도의 XML 문서로 만들어져 있어 내용검색 및 데이터 분석이 용이해지며, ③기존 문서들의 재활용에 유리한 장점을 지닌다(임진희, 2020, 31-32).

개방형 문서포맷 도입은 단순히 특정 포맷의 독점적 지위로 인한 접근성 및 상호호환성, 활용가능성 문제를 개선하는 것 이상의 의미가 있다. 대표적으로 HWP포맷 등의 문서를 대체하면서 사용되는 ODF 기반의 ODT 포맷은 XML으로 구성되어 있어 범용성이 높아져 활용가능성을 증대시킬 수 있다. 다음 그림 1[2]과 같이 XML기반의 개방형 문서포맷이 content.xml과 meta.xml로 분리되어 있을 경우 본문 내용 텍스트가 content.xml에 포함되어 텍스트 분석 전처리과정이 단순해질 수 있다.

2) 본문의 모든 그림은 연구내용 설명을 위해 직접 작성함.

〈그림 1〉 공문서 ODT Content.xml에서
추출 가능한 메타데이터

ODT 포맷은 문서파일의 텍스트 중 원하는 단어에 태깅 할 수 있어 이를 이용해 이용형태를 개선시킬 수 있다. 본문에 등장하는 주요 정책, 중요인물, 장소, 사건 등의 이름에 태깅하여 전거정보로 활용할 수 있다. 서식에 대한 태깅도 가능하다. 공문서의 경우 정해진 서식을 사용하는 경우가 많은데, 표준서식이나 정해진 양식에 의거하여 반복적으로 만들어지는 문서들은 서식 자체에 식별자를 부여하여 관리하고 태깅하여 활용할 수 있다(임진희, 2021, 129-131).

문서를 빅데이터를 효율적으로 처리하기 위해서는 문서의 기계가독

성을 높이는 것이 핵심이며, 기계가독성을 높이기 위한 가장 중요한 요건은 문서의 포맷이 개방형이어야 한다는 것이다. ODF와 같이 개방형 포맷을 사용할 경우 본문 내용의 텍스트 분석 전처리가 단순해지기 때문에 문서분석의 효율성이 높아진다. 예로 문서파일 단위로 키워드 검색, 공개 유형의 구분, 주제분류, 중복파일 체크, 무결성 검증, 가장 많이 참조한 파일 찾아내기(임진희, 2020, 34) 등 다양한 지능형 문서처리가 가능해질 수 있다. 개방형 문서포맷 도입의 확대는 지능형 문서처리 도입의 계기이자 선행조건으로 볼 수 있다.

2) 인공지능을 활용한 문서처리

자연어처리(Natural Language Processing, NLP)는 사람이 말하는 언어를 기계언어로 분석하여 컴퓨터가 읽어 들여 작동할 수 있는 형태로 만드는 연산과정이다. 자연어의 이해나 연산결과로 나온 기계어 형태를 반대로 인간이 이해할 수 있는 자연어로 표현하는 기술을 의미하는 것으로, 형태소 분석, 품사 태깅, 구문 분석 등의 다양한 기술을 기반(현영근 외, 2020, 138)으로 하는 인공지능의 핵심 기능 중 하나이다. 1950년대부터 기계 번역과 같은 자연어 처리 기술이 연구되기 시작하였으며, 1990년대 이후에는 대량의 코퍼스(corpus) 데이터를 활용하는 기계학습 기반 및 통계적 자연어 처리 기법이 주류를 이뤘다. 하지만 최근에는 딥러닝 기반의 자연어처리를 통해 방대한 텍스트로부터 의미 있는 정보를 추출하고 활용하기 위한 언어처리 연구 개발이 전 세계적으로 활발히 진행되고 있다(인공지능신문, 2021). 이러한 딥러닝 기반의 자연어처리 기술은 기계번역, 텍스트 인식을 기반으로 한 챗봇, 정보검색, 코퍼스 구축, 시맨틱웹, 딥러닝, 그리고 빅데이터 분석 분야뿐만 아니라 인간의 언어정보처리 원리와 이해를 위한 언어학과 뇌인지 언어정보처리 분야까지 확대되고 있다.

빅데이터 분석 시 자연어처리 기반이 되는 데이터인 코퍼스(corpus)는 언어 연구를 위해 텍스트를 컴퓨터가 읽을 수 있는 형태로 모아 놓은 언어 자료로(국립국어원 표준국어대사전) 학습을 위한 데이터로 사용된다. 한국어로는 '말뭉치' 또는 '말모둠' 이라고도 불리며 언어데이터의 집합이라 할 수 있다. 코퍼스 분석은 실제로 사용된 언어데이터를 그 대상으로 하고, 컴퓨터로 활용가능한 상태여야 하며 이들을 분석할 수 있는 컴퓨터 프로그램이 준비되어 있어야 한다.

자연어처리 기술 중 하나인 버트(Bidirectional Encoder Representations from Transformers, BERT)는 2018년 11월 구글이 공개한 인공지능(AI) 언어모델이다. 한국어를 포함해 다양한 언어처리가 가능한 모델로 대량의 코퍼스를 버트 언어모델에 적용할 수 있다. 기존의 자연어처리 모델이 사전훈련 없이 특정 태스크에 대해 학습시켜야 했던 것에 비해 대량의 코퍼스를 버트를 통해 다양한 태스크에 적용하고 미세조정할 수 있다. 또한 트랜스포머 구조를 사용하고 양방향적 특성(Ashish Vaswani et al, 2017, 3)을 가지고 있어 다음 단어의 예측 능력은 덜하지만 이후의 문맥을 고려한 자연스러운 결과를 획득할 수 있는 등 문장 의미 추출에 강점을 가진다.

이러한 변화들은 인공지능과 결합한 지능형 문서처리로 이루어진 것이다. 문서처리는 업무를 위해 담당자의 문서관리를 지원하는 개념으로 문서의 분류, 검색을 위한 정보나 지식의 추출 등의 기능이 결합된 형태이다. 이 기능에 인공지능 기술이 결합된 형태를 지능형 문서처리로 정의할 수 있다(Daniel E Cooke, 1994, 103). 따라서 지능형 문서처리는 자연어 처리 등을 이용한 데이터 추출과 이를 범주화하고 분류하는 과정이 필수적이다.

국내에서도 지능형 문서처리에 대한 다양한 고민들이 시작되었다. 무엇보다 업무수행 과정에서 생산되는 전자문서가 연간 약 6억건[3]에 달하는 등 수량이 많아 검색 및 활용에 어려움을 겪고 있으며, 민원 서

식과 같은 표준화된 문서파일은 정보를 정형화하여 활용할 수 있는 가능성이 존재함에도 확인 · 검토 · 검수 · 종합 · 분석 등에 많은 자원과 시간이 소모되고 있다. 특히 디지털 뉴딜정책과 관련하여 동일한 유형의 설문지가 전자파일로 확보되었음에도 자동화된 형태로 설문을 종합하는데 많은 노력과 시간이 소요된다(임진희, 2021. 107)는 점은 시사하는 바가 크다.

오픈포맷(Open Format) 문서 파일 사용의 증가는 것은 데이터에 대한 접근성과 활용성을 높여준다. 이는 지능형 문서처리 측면에서 긍정적 변화이다. 한글 및 MS워드프로세스는 OOXML, ODF, OWPML 등 오픈포맷으로 저장하거나 변환할 수 있는 기능을 제공하고 있으며, 오픈포맷을 사용할 경우 기술적으로 파일 내 메타데이터의 저장 및 유통이 가능하다. 또한 공공기관의 업무시스템에 임베딩(embedding)된 문서편집기의 경우에도 오픈포맷으로 저장되는 경우가 늘어나고 있으며 메타데이터를 활용한 문서파일 이용가능성이 높아지고 있다.

이에 지난 1월 과학기술부에서는 2021년 지정공모형으로 “문서 데이터의 정보화 및 협업을 위한 지능형 문서 처리 플랫폼 기술” 과제가 R&D연구로 발주되었다. 이 과제는 문서 내용에 해당하는 데이터를 토대로 메타데이터를 정보화 하고, 메타데이터를 이용한 의미기반 활용 유형을 제시하며, 정보 공유를 통하여 조직 내 협업을 증진하는 지능형 문서 처리 플랫폼 기술을 개발하는 것이다(정보통신기획평가원, 2020). 연구 결과에 따라 공공기관 등의 문서 내용이 데이터화되어 문서를 이용한 구성원 간 협업이 증대될 수 있고, 문서에 대한 의미 기반 검색, 추천, 분석 성능이 향상되어 문서를 이용한 빅데이터 분석에 긍정적 영

3) 국가기록원의 2020년 생산현황 분석 결과보고(19년 생산분)에 의하면 전자문서는 총 599,338,730건으로 전년 대비 약 5% 증가하였다. 1개 전자문서는 본문과 첨부파일을 고려하면 1개 이상의 문서파일을 담고 있으므로 총 문서파일 수는 더욱 많은 것으로 예상된다.

향을 줄 수 있을 것이다.

3. 지능형 문서처리의 단계별 기술

1) 메타데이터 추출 및 태깅

문서의 기계가독성을 높이고 효율적으로 처리하기 위해서는 문서 파일 단위로 데이터를 인식하고 활용할 수 있는 인프라를 구축해야 한다. 이러한 인프라는 문서의 내부 정보를 분석하기 위한 메타데이터 사용자 태깅 및 자동태깅 기능을 구현할 필요가 있다. 문서의 내용에서 의미단위를 분석하고 메타데이터를 지정하며 메타데이터 관계를 형성하는 기능은 데이터 객체와 데이터 객체 집합과의 관계, 데이터 객체와 관련 정보와의 관계 등의 메타데이터 의미관계를 형성하는 데 활용된다. 이로써 문서의 내용 분석을 실행할 수 있는 기반이 마련될 수 있다.

문서 내의 단어, 문단, 표, 그림 등의 의미단위에 태깅된 메타데이터는 문서의 검색이나 특정한 목적을 위한 문서의 분석, 문서와 문서 관련 맥락 정보의 연계 등을 실행하는데 활용되며 지능형 문서처리를 가능하게 한다. 문서 내의 데이터가 메타데이터 태깅을 통해 자동 추출되고, 이 데이터는 조직 내의 구성원들에게 제공될 수 있다. 또한 메타데이터 태깅은 구성원들의 참여에 의해 지정되고 공동편집이 가능하도록 지능형 문서처리를 실행할 수 있는 시스템에 포함되도록 설계되어야 효율성을 향상시킬 수 있다.

경우에 따라 상용 문서편집 소프트웨어를 이용하여 문서작성을 끝내고 생산자가 직접 메타데이터를 입력할 수 있는 부속프로그램(Add-in)이 필요할 수 있다. 이를 위해서는 ODT와 같은 오픈포맷을 토대로 문

서를 작성하고 유통하는 것을 전제로 하며, 메타데이터는 Content.xml에 포함될 수 있는 형태로 개발되어야 한다.

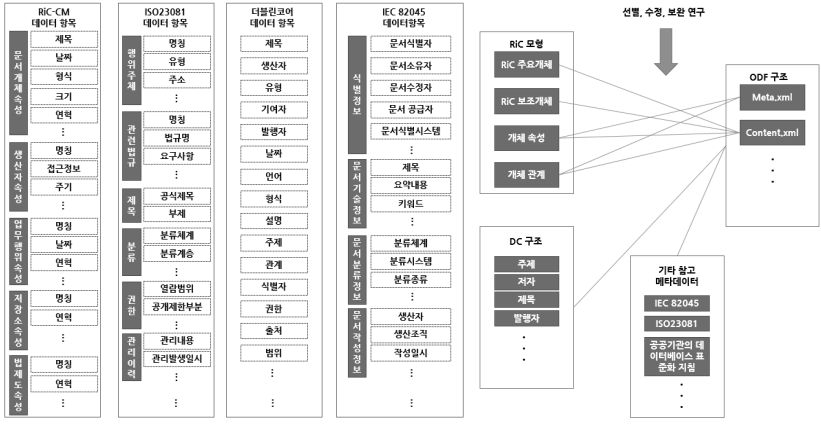
문서파일에서 추출되거나 입력된 메타데이터는 다시 그래프 형태의 데이터베이스를 활용하여 지식그래프의 시각화 방식으로 제공될 수 있다. 구성원들은 이를 공동으로 활용하거나 서로 다른 데이터와 연결함으로써 지식 정보를 확장하는데 활용이 가능하다.

메타데이터 태깅과 관련하여 몇 가지 요소들을 주목해야 한다. 먼저 메타데이터 요소 세트는 범용적으로 사용될 수 있도록 형태와 규격을 정의하여 표준화해야 한다. 표준화된 메타데이터를 활용해야 데이터의 공유와 활용의 효용성이 향상될 수 있기 때문이다.

또한 다양한 분야의 전문 메타데이터 표준을 포괄해야 한다. 이러한 기능 구현을 위해서는 메타데이터 세트의 업로드, 다운로드 기능과 함께 메타데이터 버전관리 기능이 필수적으로 요구된다. 또한 메타데이터 태깅을 공동 편집 기능을 활용해 수행하도록 하여 메타데이터의 지정과 변경, 수정, 검토, 확인 등의 실행이 적합하게 개발되어야 한다. 이 기능은 권한 관리 기능과 함께 구현되어야 한다. 메타데이터 태깅 업무의 수행 범위별 사용자 그룹을 구분하여 권한 관리 및 접근 통제 기능이 요구된다.

지능형 문서처리의 작동은 본문 내 메타데이터의 추출과 태깅으로만 가능한 것은 아니다. 메타데이터는 국제표준 등을 참고하여 상호운용성을 확보한 형태로 메타데이터 세트에 반영되어 관리되어야 한다. 그림 2와 같이 메타데이터 국제표준과 각 분야의 전문 메타데이터를 업로드하여 필요한 메타데이터를 차용하여 사용할 수 있도록 하며, 다양한 맥락정보의 표현을 위해 RiC 개념을 적용하여 활용할 수 있도록 구현될 필요가 있다.

〈그림 2〉 국제표준 기반의 메타데이터 활용 개념도



파일단위 활용을 적절하게 지원하는 형태의 메타데이터 세트를 구현하기 위해서는 문서파일과 맥락정보가 충분히 연결되어야 한다. 이를 위해 문서의 생산과 관련된 다양한 개체(entity) 정보를 문서파일에 바로 연결하는 방식을 떠올릴 수 있는데, 이는 RiC의 구조에서 차용할 수 있다. RiC 메타데이터의 구조 방향성이 객체 자체에 대한 정보 뿐 아니라 관리기관 정보, 생산자 정보, 기능정보 등 다양한 맥락정보를 시멘틱웹 환경에 적합한 통합을 지향했기 때문이다.

적절한 메타데이터 세트 설계 후 메타데이터가 다양한 영역에서 호환되는 형태로 활용되기 위해서는 메타데이터 레지스트리(MDR)에서 관리할 수 있는 형태여야 범용적 사용이 가능하다. 또한 각각의 문서의 내용을 충분히 담을 수 있는 맞춤화된 메타데이터가 필요하며, 각각의 메타데이터 항목에 담긴 정보 품질이 유지되어야 한다.

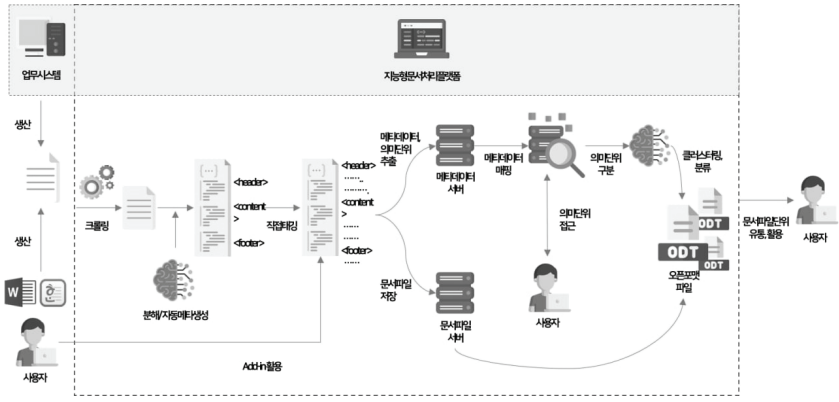
2) 의미단위 인식과 처리

자주 활용되는 기안문의 경우 특정 템플릿 등 (반)정형정보가 포함된

문서파일 등을 중심으로 메타데이터를 추출하고, 문서파일과 생산자, 생산기관과의 관계, 문서의 선후관계 등 메타데이터 항목간의 관계를 정의할 수 있다. 이렇게 정의된 항목간의 관계는 문서의 이용을 원활하게 한다.

메타데이터 항목 및 관계 정의는 문서 내 정보를 인식하여 메타데이터 항목으로 자동으로 추출시켜 줄 수 있는 기능이 함께 구현되어야 지능형 문서처리에 활용될 수 있다. 이를 위해서는 지능형 문서처리 플랫폼 내에 문서의 내용을 인식하여 처리할 수 있는 인공지능기술이 요구된다.

〈그림 3〉 인공지능을 활용한 지능형 문서처리 플랫폼 기본 흐름



인공지능이 효율적으로 작동하기 위해서는 앞서 설명한 자연어 처리 기술과 함께 언어데이터 집합 형태인 코퍼스 등이 필요하다. 이와 더불어 딥러닝 기반의 인공지능의 성능이 향상되기 위해서는 높은 품질의 학습데이터가 대량으로 필요하다. 일상생활에서 주로 사용하는 언어가 아닌 특정 산업군이나 영역 내에서 전문용어가 포함된 문서를 처리하기 위해서는 특성상 학습데이터를 적절히 생산해내는 것이 중요한 과제가 될 수 있다.

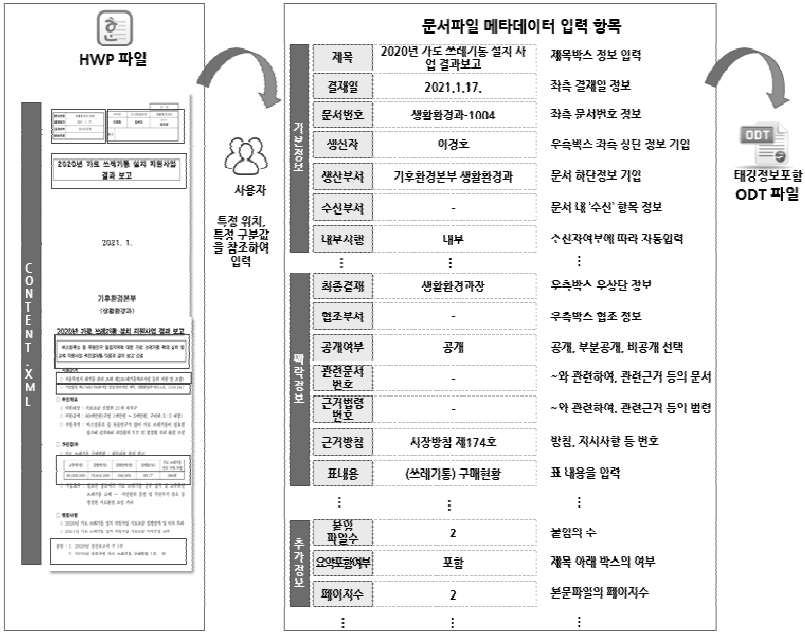
그림 3과 같이 지능형 문서처리 플랫폼은 생산된 문서파일을 시스템에서 인식하고 인공지능이 형식과 내용을 인식하여 자동으로 메타데이터를 생산할 수 있게 구현될 수 있다. 또한 사용자의 질의와 명령에 따라 저장된 메타데이터를 이용하여 다시 문서를 분류하거나 분석하여 제공하게 된다. 지능형 문서처리 플랫폼의 작동 흐름은 사용자가 편집기 혹은 업무시스템을 이용하여 생산한 문서가 크롤링 과정을 거쳐 시스템 내에서 메타데이터를 관리하게 된다. 기계가독된 본문이 영역별로 구분되어 메타데이터가 생성되며, 시스템 내 사용자가 직접 태깅하는 방식으로도 메타데이터가 보충되기도 한다. 시스템이 이러한 메타데이터를 통해 문서의 접근점을 향상시키고 사용자가 의미단위로 검색하는 것을 지원하게 되는데, 이 과정에서 유사한 문서를 클러스터링하거나 분류하는 기능도 수행할 수 있다. 이용자는 지능형 문서처리 플랫폼을 통해 기록 건단위가 아닌 문서파일 단위로 접근하여 이용할 수 있으며 의미단위로 쉽게 접근하여 업무에 더욱 쉽게 참조할 수 있는 이점을 가진다.

지능형 문서처리 플랫폼은 인공지능을 활용하여 메타데이터를 태깅하고 문서를 분류하기에 높은 수준의 업무를 위해서는 인공지능의 수준을 향상시킬 수 있는 높은 품질의 학습데이터가 충분히 제공되어야 한다. 또한 새로운 유형의 문서가 발견되거나 인공지능이 쉽게 인식하지 못하는 내용의 문서는 별도의 처리가 필요할 수 있으며, 이에 대한 개선을 위한 학습데이터도 필요하다.

나아가 특정 업무에 대한 안내, 취합, 계획수립 등 문서 전체 혹은 영역별로 범주화시킬 수 있는 공문서 본문 파일을 토대로 메타데이터 맵핑로직을 설계하여 이를 토대로 문서파일을 대량으로 분석한 후 적용가능한 학습데이터를 구축할 필요가 있다. 아래의 그림 4와 같이 기안문의 특정 템플릿 등 문서파일에서 의미단위별로 메타데이터 항목을 설정하고 메타데이터 항목간의 관계 정의를 통해 메타데이터를 추출하

여 활용할 수 있어야 한다.

〈그림 5〉 학습데이터를 위한 공문서 본문 파일의 메타데이터 입력



이 과정에서 지능형 문서 처리 플랫폼은 시스템 내 문서의 분석을 통해 의미분석 학습 데이터의 코퍼스(corpus)를 생성하고 데이터 전처리 과정을 거치게 된다. 전처리 과정을 거치면 의미요소 단위로 자연어가 기계가 이해할 수 있는 형태인 숫자로 나열되는 벡터(vector)로 변환된다. 사전학습(Pre-trained) 모델을 통한 전이학습(transfer learning)을 적용하여 파인 튜닝(Fine-tuning)과정을 거친다. 구현된 학습 모델을 통해 키워드, 주요문장, 문서요약 등의 기능을 수행하여 메타데이터를 자동생성하고 이를 시각화하는 기술을 구현하여 제공한다.

이를 위해서는 다시 특정한 도메인에 맞는 용어사전과 같은 도구가 필요할 수 있다. 도메인 용어사전은 문서파일의 내용을 분석하여 정책명, 기관명, 관련 주체나 단체, 사건 등의 주요 용어를 추출하여 의미단위로 지정하고 연결하여 목록화한 도구를 지칭한다. 또한 도메인 용어사전의 구축을 위해서는 의미단위간의 관계에 관한 정보를 수집하여 조직화하는 작업이 선행되어야 한다.

4. 문서 및 기록 업무의 변화 전망

1) 문서 수행 업무의 지능화

공공기관의 모든 업무는 문서를 중심으로 이루어진다. 특히 특정 목적을 위한 문서의 취합, 다양한 조사의 통계, 특정 서식에 대한 반복처리, 접수 문서의 배부, 업무담당자의 요청에 따른 관련 문서 검색 등 단순 반복 업무 수행에 많은 시간을 할애한다. 이러한 반복 업무는 조직 내 인력의 효율적 운영을 어렵게 만들며, 업무의 피로도를 증가시킨다.

인공지능과 개방형 포맷을 이용한 지능형 문서처리 기술은 이러한 문제를 해결할 수 있는 하나의 방안이 될 수 있다. 지능형 문서처리 기술은 문서의 형태와 서식, 키워드 등을 인식하고 여러 문서를 처리하는데 강점을 보이기 때문이다. 그간 문서를 접수하거나 인식한 후 본문과 첨부파일을 하나씩 열어 확인하고, 분석 등을 위해 새로운 파일에 옮긴 후 분석 결과를 위한 문서를 다시 생산하는 등 사람이 수행하던 업무를 일정 부분 대체할 수 있다. 지능형 문서처리 도구가 도입된다면 인공지능과 개방형 포맷을 활용하여 문서에서 정보를 추출·인식하고 이를 문서파일 단위로 검색하는 것이 가능해진다.

이 지능형 문서 처리 도구는 기관에서 문서를 접수하여 담당부서로

문서를 배부하는 업무의 자동화를 지원할 수 있다. 문서 내에 태깅된 메타데이터를 통해 업무 기능에 관련된 키워드를 확인하고 문서 요약을 통해 관련 업무 기능을 파악하여 해당 업무를 수행하는 담당부서로 문서 배부하는 업무에 활용이 가능하다.

의미기반의 문서검색 및 문서추천, 문서내용 요약 및 키워드 추출, 메타데이터 시각화 등의 활용도 가능해진다. 예를 들어 특정 업무나 정책, 인물과 관련하여 검토된 내용의 문서를 찾아달라는 요청을 받았을 경우, 현재는 업무명, 정책명, 인물명을 검색어로 입력하거나 이와 관련 키워드로 문서를 검색하여 해결한다. 지능형 문서처리가 가능해질 경우 각 문서파일에서 자동 추출되거나 태깅한 메타데이터를 활용하여 동일하거나 유사한 메타데이터가 포함된 문서들을 검색하고, 의미기반 검색·추천에 따라 관련된 내용의 문서를 추천해줄 수 있다.

대량의 설문조사와 같은 특정서식에 의한 통계가 필요할 시에도 설문조사 문서의 형식, 설문의 서식, 표 등에서 추출하여 자동 태깅된 메타데이터를 기반으로 통계분석 및 시각화가 가능해진다. 특히 자주 사용되는 공공서식의 경우 문서의 표기능을 이용하여 구분되어 있는 경우가 많은데, 완전하게 항목명과 기입란을 구분시키는 형태로 개선된다면 정보위치를 태깅하여 다수의 문서의 정보를 테이블 형태의 데이터베이스로 구성하여 활용할 수 있게 된다.

문서 내 데이터가 메타데이터 태깅으로 접근점이 늘어날 경우 이를 통해 업무와 관련된 문서 및 맥락 정보가 부서의 경계를 넘어 확장되어 확인되고 연결되며, 데이터는 넓어지고 깊어지게 된다. 이에 따라 데이터를 활용한 협업이 보다 쉬워지고 빨라지게 될 전망이다. 또한 문서 데이터의 의미 기반 검색과 추천, 분석 기능이 비약적으로 향상되며 새로운 인사이트가 도출될 수 있다.

이러한 과정은 ①기록 건이 아닌 기록 컴포넌트 단위인 문서 파일별로 메타데이터가 태깅되고, ②의미 분석에 의해 추출된 메타데이터

가 문서 파일에 임베디드되어 함께 유통됨으로써 ③관련 데이터를 활용한 지식이 축적되도록 할 것이다. 그 결과 업무참조를 위한 검색과 활용이 유용해짐은 물론, 조직의 업무 지식과 노하우를 보유한 구성원들이 메타데이터를 부가하는 지적 작업을 통해 협업을 촉진하는 흐름으로 이어질 것이라 기대할 수 있다.

지능형 문서처리의 데이터 분석을 통해 조직 운영 및 관리의 변화를 수반할 수도 있을 것이다. 데이터 분석 결과를 활용하여 조직 내 인력 배치 및 업무 효율을 점검할 수 있고, 지식관리의 관점에서도 조직의 핵심 지식과 활용 가치가 높은 지식을 발굴하는데 유용하게 활용(임진희, 2021)될 수 있을 것이다.

지능형 문서처리는 문서기반 업무의 효율성을 개선시키기 때문에, 빠른 도입이 예상된다. 인공지능 기술도입에 대한 범정부적 관심과 오픈포맷의 공공부문 도입 등과 함께 일어나고 있는 이러한 변화는 업무수행 방식에 있어 다방면의 변화를 예고하고 있다.

2) 대규모 기록관리 업무수행 방식 변화

지능형 문서처리는 기록의 '관리' 차원에서 도입되었다고 보기 어렵다. 기록관리 차원에서는 문서의 생산을 강요하거나 생산형태를 정의하는 등 업무수행에 직접적인 영향을 주지 않고, 제한된 역할을 수행할 뿐이다. 데이터의 추출이나 분류·범주화·수치화·검색에도 주도적 역할을 수행하지 않는다. 지능형 문서처리를 적용하는데 있어서는 기존과 다르게 기록관리를 업무수행 결과물에 대한 관리 차원으로 이해하기보다 고유한 업무자체로 이해하고 지능형 업무처리에 접근할 필요가 있다. 기록관리 업무를 수행하기 위한 지능형 문서처리 과정에서는 기록관리 업무분야를 구분하는 현용기록관리(RM단계), 영구기록관리(AM단계), 매뉴스크립트 관리 등의 단계적 구분에만 머무르지 않는다.

오히려 기관고유의 특성에 따른 문서의 형태, 단어의 사용 방식, 특별한 문서 생산 방식 등에 대한 관찰과 선제적 대응이 더욱 중요해진다.

일반적으로 지능형 업무처리는 대량 반복 업무를 처리하는데 강점을 가진다. 기록관리 업무에서 이에 해당되는 분야로는 기록의 공개여부 재분류와 평가심의가 대표적이다.

먼저 기록관 단계의 공개여부 재분류의 과정을 보면 원칙적으로 기록물의 비공개 재분류로부터 5년 주기, 영구기록물관리기관으로 이관 시기에 업무를 수행하게 된다. 현재의 기록관에서의 공개여부 재분류 과정은 ①재분류 기록 추출 ②공개재분류 계획 수립 ③기록물 확인 ④공개여부 내용 검토 ⑤처리과 의견조회 ⑥공개여부재분류심의 순으로 진행된다(국가기록원, 2009, 18).

공개여부 재분류는 기관이 생산한 비공개기록의 수량이 많을수록 늘어나며, 많은 기관에서 용역수행을 통해 “기록물 확인”과 “공개여부 내용검토”를 진행한다. 이 경우 수립된 원칙을 다수의 인력 투입을 통해 내용검토를 이행하므로 해당 원칙을 통일성 있게 적용하는 것이 현실적으로 어려우며, 기록의 수량에 따라 업무규모가 늘어나게 된다. 또한 기관에 따라서 공개재분류 업무의 품질 수준과 편차에 문제가 발생할 수 있으며, 범정부적으로 통일된 관리 기준이 적용되는 것은 현재로서는 사실상 불가능에 가깝다.

이러한 환경에서 인공지능을 이용한 지능형 업무처리를 적용하면 효율적인 업무 환경으로 변할 수 있다. 인공지능을 기반으로 한 지능형 업무처리를 공개재분류 업무에 적용한다면 자연어를 처리하여 유사문서를 구분하고 동일한 기준을 적용할 수 있게 된다. 지능형 문서처리를 기반으로 한 공개 재분류가 실행된다면 업무수행의 품질, 일관성, 속도 등을 개선하여 생산성을 높일 수 있을 것으로 기대된다.

예를 들어, 3장에서 설명한 지능형 문서 처리 과정을 통해 문서 내에서 의미단위를 식별하고 태깅을 통해 특정 주제어, 정책명 등을 도출해

낼 수 있다. 이 의미단위별로 비공개 대상 정보를 지정하여 공개재분류 업무에 활용할 수 있게 된다. 다른 법률, 명령에 따른 비공개 정보(비공개 대상 정보 1호)나 개인정보보호법에 따른 개인 정보(비공개 대상 정보 6호) 뿐만 아니라, 현재 진행중인 재판에 관한 내용(비공개 대상 정보 4호), 의사결정 과정 또는 내부검토 과정에 대한 내용(비공개 대상 정보 5호)도 의미단위 태깅을 통해 식별가능하기 때문에 공개재분류 과정의 자동화를 지원할 수 있다.

이러한 변화는 기록전문가의 업무수행 방식의 변화도 야기할 수 있다. 기록전문가가 계획을 수립하고 문서 목록을 추출하여 내용을 검토하거나 용역을 발주하는 대신 모범사례를 제시할 수 있는 학습데이터를 생산하고 이를 통제하는 업무를 수행하는 것이 더욱 효과적일 수 있다. 인공지능이 더욱 정교하게 업무를 수행하기 위해서는 더 많은 학습데이터와 예외사례 발굴 등의 과정이 필요하게 된다. 구체적으로는 범정부적으로 일관성 있는 업무수행과 자원의 효율적 활용을 위해 국가기록원을 중심으로 표준모형을 개발하고 기관의 특성에 맞는 고도화 모형을 개발하는 방식도 고려해 볼 수 있다. 이 과정에서 기록전문가는 기관의 특성을 드러내는 문서 유형을 학습 데이터로 포함시키고 학습데이터 구축 현황을 점검하고 수정하는 업무들을 수행할 수 있어야 할 것이다.

기록관 단위의 평가심의 업무 역시 지능형 문서처리를 이용한 업무수행이 적용될 수 있는 분야이다. 평가심의회 운영과정에서 유사한 기록을 범주화하고 일관성 있는 기준을 적용할 필요가 있기 때문이다. 지능형 문서처리를 이용하여 내부의 기준을 적용하여 평가심의회를 수행한다면 다수의 문서를 빠르게 인식하고 심의안을 만들어낼 수 있다.

이 경우에도 역시 높은 품질의 학습데이터가 중요하며 기관의 고유업무와 관련된 내용이 학습데이터에 반영되도록 해야 한다. 학습데이터 구축 과정에서 유사기능의 업무수행과 관련된 기록을 식별할 수 있

도록 정책명과 색인어 추출, 근거기록 연결, 업무과정에 대한 의미단위 인식 등이 가능할 수 있도록 시소러스와 같은 보조도구가 필요하게 된다. 특히 특정 기관의 고유업무는 대체로 해당 영역에서 한정적으로 사용하는 경우가 많기 때문에 영역별 학습데이터에 대한 별도의 생성이 요구된다. 특히 유사한 대상을 다루는 업무에서도 계획보고 · 결과보고 · 집행 · 요구 · 제출 등 실제 수행된 업무의 성격에 따라 보존기간이 상이한 경우가 많기 때문에 높은 수준의 인식과 구분이 필요하다.

평가심의회 운영 과정에서도 지능형 문서처리를 이용하여 수행한다면 품질과 통일성을 확보할 뿐 아니라 빠른 속도와 인력투입을 감소시켜 업무효율을 높일 수 있다. 특히 업무관리시스템을 사용하는 기관의 경우 개별 문서가 과제관리카드를 적절하게 선택하지 못하여 폐기대상이 된 경우를 쉽게 발견하기 어려울 수 있다. 평가심의 분야에 지능형 문서처리를 이용해서 더욱 정교하게 업무를 수행할 수 있다.

지능형 업무처리 방식을 이용한다면 기록의 정리와 기술업무 도 달라질 수 있다. 영구기록물관리기관이나 매뉴스크립트 보존소의 경우 정리와 기술(Description)에 관한 업무가 상당한 비중을 차지하고 있다.

우선 기술업무에 있어서는 지능형 문서처리를 통해 문서의 요약, 키워드 추출을 통해 메타데이터로 내보내는 기능이 활용될 수 있다. 텍스트 기반의 문서 내용을 토대로 메타데이터를 추출한다면 기록이 생산된 배경과 출처에 관한 정보를 자동화하는 것은 제한적이나 표준 기술항목에 존재하는 범위와 내용, 색인어 등을 입력하는 업무는 상대적으로 자동화시킬 수 있을 것으로 보인다. 현재 한국어 문서의 요약과 관련된 인공지능 기술은 상용화를 준비하거나 뉴스요약과 같은 특정 영역에서는 버트(BERT)를 이용한 솔루션을 구축해 적극적으로 활용되고 있다(연합뉴스, 2021). 공공부분 역시 한국정보화진흥원에서 국가과제로 진행하고 있는 등 빠르게 발전하고 있는 분야이다.

지능형 문서처리 도구 활용은 '건(item)' 단위의 기술에 우선적으로

적용될 수 있다. 또한 OCR(optical character reader) 기술을 이용하여 종이로 생산된 비전자문서를 전자화시켜 인식한 후 다시 인공지능을 이용하여 요약하고 색인어를 추출하는 방안도 생각해볼 수 있다.

향후에는 정부조직, 역사적 사건이나 인물정보가 저장된 DB와 연결시켜 생산자의 연혁 등과 관련된 기술도 자동화할 수 있을 것으로 예상된다. 이에 대한 정보는 범정부적으로 관리하고 있기 때문에 정보의 연계를 통해 충분히 활용할 수 있을 것이다.

장기적으로는 집합기술에 대한 부분도 지능형 문서처리를 적용할 수 있을 것으로 보인다. 특정한 기능이나 생산자로 인해 구성된 집합에 대한 기술도 다수의 학습데이터가 입력된 인공지능을 이용할 수 있을 것으로 예상된다.

완성도와 품질을 위해서는 기록관리 영역에서 합의를 통해 기록 기술이 가져야 할 요건과 요소들을 충분히 정의하여 통일성 있는 세부 기술 형식이 전제되어야 할 것이다. 기록전문가는 다양한 형태에 적합한 학습데이터를 충분히 생산하며, 특정 영역에서만 사용하는 용어를 정의하고 관련된 용어들을 연결하는 역할도 함께 수행하는 형태로 업무가 확장될 수 있다.

기록의 정리 역시 지능형 문서처리를 활용할 수 있는 분야이다. 기록관 단계에서 기록의 정리는 생산시스템에서 단위과제의 미지정 혹은 오지정을 해소하거나 미등록된 비전자기록을 확인하여 등록하는 과정을 지칭하며(국가기록원, 2020a, 30), 영구기록물관리기관에서 정리는 수집된 기록을 분석하여 물리적·논리적으로 분류하는 과정(국가기록원, 2020b, 22)을 포함한다.

지능형 문서처리 도구를 이용한다면 기록관 단계에서는 업무관리시스템에서 과제관리카드에 대한 재지정을 권고하거나, 경우에 따라 유사한 과제관리카드의 통폐합을 권고하는 방식도 가능할 수 있다. 또한 영구기록물관리기관에서는 각기 다른 수집처에서 수집하였으나 동일

한 사건을 증거하는 기록이나, 생산출처가 동일한 기록을 연결하는 도구로 활용할 수 있다. 이 경우 기록전문가는 지능형 문서처리 도구가 충분히 인식하지 못하는 문서들에 대한 사례를 추출하여 코퍼스 구축이나 학습데이터 생산을 통해 업데이트하는 업무를 수행할 수 있다.

3) 이용자 검색 및 활용 지원

지능형 문서처리를 도입하게 되면 기록관리 차원의 검색 서비스와 생산단계에서의 검색 서비스의 구분이 모호해질 수 있다. 기록관 단계의 기록관리시스템에서의 검색서비스와 생산단계인 업무관리시스템 차원에서의 검색 서비스 구분에 대한 실익이 사라질 수 있기 때문이다.

지능형 문서처리는 기록건이 아닌 문서파일 단위로 일어나며 기록의 활용 역시 문서를 최소단위로 할 수 있도록 한다. 문서로부터 검색을 위한 키워드 추출, 유관문서와의 연결, 업무관련자의 메타데이터 태깅 등이 일어난다면 기록에 대한 접근점이 늘어나 검색 편의성이 높아질 수 있다. 이러한 변화는 의미기반 검색을 포함하여 생산단계에서 활용에 대한 편의성이 상승하며, 기록관리시스템은 기록전문가를 통한 접근통제와 열람권한 및 이력 관리 등 몇 가지 기능을 제외하면 차별성을 가지기 어렵다. 클라우드를 시스템을 활용하여 생산시스템과 기록관리시스템이 통합운영되는 형태로 전환되는 흐름은(이진룡, 주현미, 임진희, 2018, 275) 이러한 변화를 현실화시키고 있다.

때문에 이용자에 대한 검색과 활용을 위해 기록전문가의 새로운 역할이 요구될 수 있다. 기록의 적극적인 제공을 위해 지능형 문서처리를 활용한 기록의 기획형 서비스가 그 예이다. 기록 간의 연결성이 높아지고 풍부한 메타데이터가 제공됨에 따라 자동화된 기록 큐레이팅이 가능해졌으며, 기록의 이용행태 등을 분석할 수 있는 데이터도 확보할 수 있다는 점도 긍정적인 부분이다. 데이터 분석 결과를 다양한 방식으로

시각화하여 제시할 수 있다는 장점도 갖는다.

예를 들어 기관의 외부 이용자를 위해서는 주제별, 유형별, 특정사업별 기록 모음 서비스를 제공할 수 있다. 정보공개 청구나 주요한 요구자료, 국민적 관심사 등을 토대로 기획형 기록모음 서비스를 준비할 수 있으며, 추출된 메타데이터와 기능분류체계를 혼합하여 구축하는 것도 가능하다.

또한 기관 내부 이용자를 위해서는 기록 검색 행태를 분석하여 자주 검색한 기록을 제시하는 방식으로 이용자들의 편의를 높일 수 있다. 검색결과는 다시 이용자들에게 태깅을 유도하여 문서간의 연결을 강화하는데 사용할 수 있다.

이러한 기획 서비스는 프레임워크가 충분히 구현되어 배포된다면 기관에 적합한 방식으로 조정하여 운영할 수 있다. 이는 기록전문가가 기관의 기록 이용자들과 호흡하며, 기관의 기록자원으로 기관의 서비스를 강화하는 역할을 수행한다는 점에서 의미를 지닌다.

5. 결론

본 연구는 공공영역에서 도입이 논의되고 있는 지능형 문서처리 도구의 배경과 구조를 분석해보고, 그에 따른 기록관리 차원의 변화를 예측해 보았다. 이와 더불어 변화에 대응할 필요성을 논의하였다.

전자기록 환경에서 지능형 문서 처리는 기록관리 업무에 광범위한 변화를 초래할 수 있다. 학습데이터를 활용하여 반복되는 업무의 처리를 자동화할 수 있으며, 기록물의 기술 업무에도 문서 요약 엔진 활용 및 색인어 추출 자동화 등을 적용할 수 있다. 또한 전거정보의 활용 및 기록물 분류 업무에도 지능화 기술의 적용 가능성이 높다. 특히 기록의 검색 활용이나 소장 기록물의 분석을 통한 콘텐츠 기획 업무에 인공지

능의 도입으로 획기적인 업무 수행이 가능할 것으로 예측된다.

이러한 기록관리 업무 변화에 맞추어 기록전문가는 수행해야 하는 업무 내용을 확인하고 이에 걸맞는 역량을 준비할 필요가 있다. 무엇보다 지능화 문서 처리 과정에서 기록전문가에게 부여되는 새로운 역할인 학습데이터 생산과 재생산, 모델의 통제 업무를 수행할 수 있어야 할 것이다. 문서 포맷의 전환과 문서 내의 데이터화를 통한 활용은 기록의 생산환경이 변화함을 의미하며 기록관리 업무의 자동화를 위해서는 학습데이터 구축 및 데이터 전처리 기술 전개 과정에 개입해야 하기 때문이다.

지능형 문서처리 기술 수준의 변화나 도입속도와 방식에 따라 기록관리 변화는 예상과 차이가 있을 수 있다. 하지만 '전자문서' 중심의 업무처리와 기록관리가 지속되는 한 지능형 문서처리가 기록관리 방식에 변화를 가져올 것이라는 예상은 타당할 것이다.

지금까지 인공지능이나 지능형 문서처리에 관한 연구는 소프트웨어 개발이나 자연어처리와 관련해서 주로 연구되었고, 기록관리와 직접적으로 관련한 논의는 상대적으로 드물었다. 이제 업무수행 과정에서 새로운 기술들은 어떠한 변화를 일으킬 것이며 기록관리에 어떠한 변화를 가져올 것인지 전망함과 동시에, 전통적인 기록관리 업무 자체를 어떻게 개선할 것인가에 대한 논의가 필요한 시점이다.

기록의 원문공개 혹은 데이터 공개 확대 관련된 논의 혹은 국가의 정보자원을 통합적으로 활용한 논의 등으로 인한 공공부문의 변화(머니투데이 2021)는 현용－준현용－비현용 단계의 구분을 근본적으로 무력화시킬 수 있다[4]. 공공데이터로서의 기록은 현용여부를 구분하지 않

4) 소위 '디지털 집현전'으로 국가지식정보 통합 플랫폼 구축을 위한 법안이 국회 21년 3월 과학기술정보방송통신위원회 소위원회를 통과하였으며, 국가기록원을 포함한 다양한 논문, 도서, 영상을 관리하는 기관의 국가지식정보를 통합하는 시도를 담고 있다.

고 활용될 수 있으며, 맥락정보는 기록이 폐기되지 않는 한 지속적으로 변경될 수 있다. 실제로 인공지능의 발전을 위해 다양한 데이터를 모아 대량으로 처리하는 시도는 이미 국가적 과제라 볼 수 있다.

이에 따라 기록관리 차원에서의 대응이 요구된다. 방향성만으로 본다면 먼저 대량의 반복 업무를 감소시키고 데이터에 대한 주도권을 적극적으로 가져와 '스마트하게' 업무방식을 바꾸는 시도가 필요하다. 또한 데이터를 중심으로 한 변화에 대응하면서도 기록관리의 기본원칙을 적용하는 지혜가 필요한 시점이다. 이를 위해 새로운 기록관리 표준절차 구축과 메타데이터 도출, 시스템의 재설계 등이 연구되어야 할 것이다.

향후 기록관, 영구기록물관리기관, 매뉴스크립트 보존소에서 일하는 기록전문가 역시 변화에 맞춘 역량이 요구된다. 구체적으로는 이를 위해 기록관리 커리큘럼의 변화에서부터 현장담당자들에 대한 재교육도 필요할 수 있다. 특히 학습데이터와 인공지능 발달, 자연어 처리 등에 대한 이해도와 소양을 높일 필요가 있다. 근본적으로는 지능형 문서처리의 도입은 종이문서를 기반으로 설계된 기록의 생산에서부터 관리체계를 근본적으로 변화시키며, 기록전문가가 생산단계에 더욱 깊숙이 개입할 수 있을 것이다.

본 연구가 앞으로 지능형 문서처리로 대표되는 인공지능의 도입에 대해 학계의 주목을 끌어내어 충분히 논의할 수 있는 계기가 되길 기대한다.

〈참고문헌〉

〈논문〉

김수현, 이성숙 (2020). RiC-CM을 적용한 영구기록물 기술방안 연구. 한국기록관리학회지, 20(1), 115-137.

김인택, 안대진, 이해영 (2017). 인공지능을 활용한 지능형 기록관리 방안. 한국기록관리학회지, 17(4), 225-250.

김태영, 강주연, 김건, 오효정 (2018). 지능형 기록정보서비스를 위한 선진 기술 현황 분석 및 적용 방안. 한국기록관리학회지, 18(4), 149-182.
박지영 (2017). RiC에 대한 기록공동체의 리뷰를 통해 본 기록물 기술표준 개선을 위한 제안. 기록학연구, 54, 81-109.
백지연, 오효정 (2019). 국가기록원 질의로그 빅데이터 기반 이용자 정보요구 유형 분석. 정보관리학회지, 36(4), 183-205.
신미라, 김익한 (2019). RiC을 적용한 아카이브 시스템 데이터 모델링 연구. 한국기록관리학회지, 19(1), 23-67.
오세라, 정미리, 임진희 (2016). 공개포맷에 기반한 전자기록 보존 포맷 재설계 방향 연구. 한국기록관리학회지, 16(4), 79-120.
이진룡, 주현미, 임진희 (2018). 차세대 기록관리를 위한 법체계 개선방안 연구. 기록학연구, 55, 275-305.
이창희, 이해영, 김인택 (2018). 기록정보서비스를 위한 메신저 기반의 챗봇 프로토타입 개발 연구: 명지대학교 대학사료실을 중심으로. 정보관리학회지, 35(3), 215-244.
이현정, 남영준 (2014). 우리나라 공공데이터의 이용활성화 방안에 관한 연구: 링크드 오픈 데이터와 전략을 중심으로. 정보관리학회지, 31(4), 249-266.
임진희 (2020). 클라우드 환경에서 공문서 파일포맷의 선택 전략. 기록학연구, 66, 5-35.
임진희 (2021). 공문서의 기계가독형(Machine Readable) 전환 방법 제언. 기록학연구, 67, 99-138.
전예지, 이혜원 (2020). RiC-CM v0.2 분석을 통한 온톨로지 모델링에 관한 연구. 한국기록관리학회지, 20(1), 139-158.
정미리, 오세라, 임진희 (2016). 공문서 컴포넌트 오픈포맷 채택이 기록관리에 미치는 영향 분석. 한국기록관리학회지, 16(2), 29-55.
주현우 (2019). 현용기록의 활용성 증진을 위한 지능형 기록관리시스템 구축: 한국중부발전 사례중심으로. 한국기록관리학회지, 19(4), 221-230.
하승록, 안대진, 임진희 (2017). 기록정보 LOD 구축을 위한 의미 상호연결 자동화 실험 연구. 한국기록관리학회지, 17(4), 177-200.
하승록, 임진희, 이해영 (2017). 오픈소스 도구를 이용한 기록정보 링크드 오픈 데이터 구축 절차 연구. 정보관리학회지, 34(1), 341-371.
현영근, 한정현, 채우리, 이기현, 이주현 (2020). 국내외 특허데이터 분석을 통한 자연어처리의 의미분석 관련 기술동향 분석에 대한 연구. 디지털융복합연구, 18(1), 137-146.

〈학위논문〉

방재현 (2018). 지능형 아카이브 시스템을 위한 기계학습 기술 적용 방안 연구: 심층신경망 적용을 중심으로. 한국외국어대학교 대학원 정보·기록학과 박사학위논문.

〈보고서 등〉

국가기록원 (2009). 잠자는 기록물에 날개를 달자: 기록물 공개재분류 실용매뉴얼 : 기록관 편.

국가기록원 (2019). 지능형 전자기록관리 기술연구 개발 기획연구. 연구결과보고서.

국가기록원a (2020). 기록물 관리지침(공통매뉴얼).

국가기록원b (2020). NAK 9:2020(v2.1)

국립국어원 표준대사전 https://stdict.korean.go.kr/search/searchView.do?word_no=517854&searchKeywordTo=3 [cited 2021.3.24.]

국가기록원 (2020). 영구기록물관리기관 표준모델: 기능 및 업무절차.

정보통신기획평가원 (2020). 2021년도 제1차 정보통신·방송 기술개발사업 및 표준개발지원사업 신규지원 대상과제 공고.

행정안전부 (2020). 공공서식 디자인 재설계 방안.

Ashish Vaswani, Noam Shazeer, Niki Parmar, etl (2017). Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017) https://arxiv.org/abs/1706.03762 [cited 2021.3.24.]

Daniel E Cooke (1994). The Impact Of Case Technology On Software Processes. World Scientific.

Data Coalition (2021). OPEN Government Data Act. [https://www.datacoalition.org/policy-issues/open-data/open-government-data-act/] [cited 2021.3.25.]

Omeka (2017). Text Analysis. Omeka Classic User Manual https://omeka.org/classic/docs/Plugins/TextAnalysis/ [cited 2021.3.24.]

〈언론기사〉

머니투데이 (2021.3.23.). '디지털 집현전' 법적 근거 생긴다… 과방위, 소위 '의결'. 출처: https://news.mt.co.kr/mtview.php?no=2021032318312470717&VNC_T

연합뉴스 (2021.3.27.). 연합뉴스, 인공지능 기사요약 서비스 첫선. 출처: https://www.yna.co.kr/view/AKR20210111095300527?input=1195m