논문

공문서의 기계가독형(Machine Readable) 전환 방법 제언

Suggestions on how to convert official documents to Machine Readable

임진희(Yim, Jin Hee)*

* 명지대학교 기록정보과학전문대학원 조교수(yimjhkr@mju.ac.kr)
■투고일: 2020년 12월 31일 ■최초심사일: 2021년 01월 05일 ■최종 확정일: 2021년 01월 13일.
■기록학연구 67, 99-138, 2021, https://doi.org/10.20923/kjas.2021.67.099

〈초록〉

빅데이터 시대에 정형데이터 뿐만 아니라 비정형데이터를 분석하는 것이 중요한 과제로 대두되고 있다. 정부기관이 생산하는 공문서도 텍스트 기반의 대형 비정형데이터로 빅데이터 분석의 대상이 된다. 기관 내부의 업무효율, 지식관리, 기록관리 등의 관점에서 공문서 빅데이터를 분석하여 유용한 시사점을 도출해 나가야 할 것이다. 그러나, 현재 공공기관이 보유 중인 공문서의 상당수가 개방포맷이 아니어서 빅데이터 분석을 하려면 비트스트림에서 텍스트를 추출하는 전처리 과정이 요구된다. 또한, 문서파일 내에 맥락 메타데이터가 충분히 저장되어 있지 못하여 품질 높은 분석을 하려면 별도의 메타데이터 확보 노력이 필요하다. 결론적으로 현재의 공문서는 기계가독(machine readable) 수준이 낮아 빅데이터 분석에 비용이 많이 들게 된다.

이 연구에서는 향후 공문서가 기계가독 수준을 높이기 위해서는 공문서의 개방포맷화, 기안문 서식의 표준태그화, 자기 기술(self-descriptive) 메타데이터 확보, 문서 텍스트 태깅 등이 선행될 필요가 있다는 점을 제안한다. 첫째, 문서가 스스로를 설명하기 위해 추가되어야 하는 메타데이터 항목들을 제시하고 이 메타데이터들이 기계가독형이 되도록 문서파일에 저장하는 방법을 제안한다. 둘째, 문서 내용 분석 시 자연어 처리에만 의존하지 않고 행정 맥락에 따라 중요한 키워드를 미리 국제표준 태그로 마킹하여 기계가독형이 되도록 하는 방안을 제안한다.

주제어 : 빅데이터, 텍스트 분석, 자기기술 메타데이터, UUID, 해시값, 국제표준 태그

〈Abstract〉

In the era of big data, analyzing not only structured data but also unstructured data is emerging as an important task. Official documents produced by government agencies are also subject to big data analysis as large text-based unstructured data. From the perspective of internal

work efficiency, knowledge management, records management, etc, it is necessary to analyze big data of public documents to derive useful implications. However, since many of the public documents currently held by public institutions are not in open format, a pre-processing process of extracting text from a bitstream is required for big data analysis. In addition, since contextual metadata is not sufficiently stored in the document file, separate efforts to secure metadata are required for high-quality analysis. In conclusion, the current official documents have a low level of machine readability, so big data analysis becomes expensive.

Keywords : bigdata, text analysis, self-descriptive metadata, UUID, hash value, international standard tag

1. 머리말

최근 공공영역에서 문서를 기계가독형으로 전환하는 것에 대해 관심을 둔 연구가 시작되고 있다. 먼저, 2020년 7월 국가기록원 연구세미나에서 “문서 파일 포맷과 서식 개선 방안”에 관한 발표가 있었다. 이 발표에서는 공문서 서식이 아날로그 방식으로 설계되어 있고 서울시의 경우 저장 포맷 또한 개방형 포맷이 아니어서 빅데이터 시대에 조응하지 못한다는 문제점을 제시하였다. 해결 방안으로는 첫째, 공문서의 저장포맷은 ODF, OOXML, OWPML 등의 개방형 포맷으로 전환해야 하며, 둘째, 공문서의 서식에 등장하는 구별된 항목들은 국제 표준 태그를 활용하여 정의하고 이 태그들이 문서에 저장되도록 하여 해당 항목의 값들이 기계가독성(Machine Readability)을 확보하도록 해야 한다는 제안이 있었다.(국가기록원 2020) 다음으로 행정안전부에서는 2020년 9월~11월 “공공서식 디자인 재설계 방안 연구”를 수행한 바 있다. 이 연구에서는 국내외 공공서식의 이용현황을 조사하고 분

석한 후 공공서식을 기계가독형으로 설계하기 위한 기준안과 디자인 표준안을 도출하였으며, 공공서식 개선을 위한 실증모델을 구현하였다.(행정안전부 2020) 다음으로 과기부에서는 2021년 지정공모형으로 "문서 데이터의 정보화 및 협업을 위한 지능형 문서 처리 플랫폼 기술"이라는 과제명으로 R&D연구를 발주하였다. 이 과제의 목표는 문서 내의 데이터(단어, 문단, 단락, 표 등)에 대한 메타데이터를 지정하여 정보화하고, 이를 토대로 검색 · 분석 · 검증 등 문서 정보의 활용성을 높이고, 정보 공유를 통하여 조직 내 협업을 증진하는 지능형 문서 처리 플랫폼 기술을 개발하고자 하는 것이다.(정보통신기획평가원 2020)

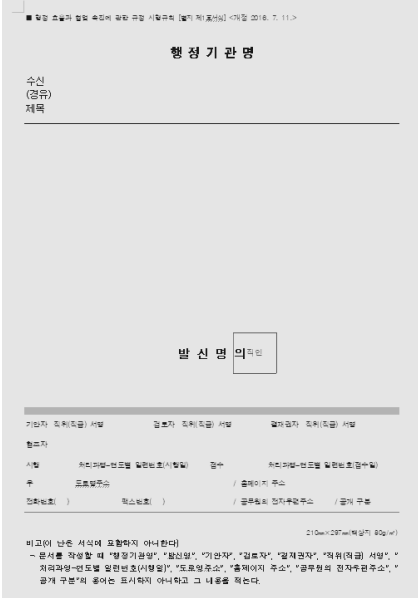
클라우드 기반의 업무관리시스템(온나라문서2.0)이 사용되면서 결재문서 기안문이 XML기반의 개방형 포맷인 odt로 저장되기 시작했고, 이는 문서가 기계가독형으로 전환되는 출발이라 할 수 있다.(정미리 외 2016) 그러나 붙임 문서파일은 여전히 hwp와 같은 독자포맷으로 만들어져서 기계가독 수준이 낮아지거나, xlsx나 pptx같은 OOXML 개방형 포맷 파일들은 업무관련 메타데이터가 충분히 저장되지 않은 상태로 만들어져서 기계가독 수준이 떨어져 있는 상황이다. 기록관리 관점에서 결재문서를 구성하는 모든 문서파일이 개방형포맷으로 전환되어야 한다는 것을 전제로 하고, 다양한 관점에서 문서를 분석하여 활용할 수 있도록 문서파일의 내부 구조를 개선해가야 할 때이다.(임진희 2020)

이 논문의 목적은 공문서 서식을 아날로그 기반에서 디지털 방식으로 전환해야 할 필요성을 업무적 관점과 지식관리 관점, 기록관리 관점으로 나누어 제시하고, 문서를 기계가독형으로 전환하기 위한 디지털 서식의 정의 방식과 풍부한 메타데이터를 포함하기 위한 방안을 제시하는 것이다. 이 논문은 문헌과 인터넷기사에서 얻은 정보 그리고 필자가 행정기관에 재직하면서 파악한 공문서의 생산 및 관리의 정량적, 정성적 현황을 토대로 작성되었다. '문서'라는 범주에는 다양한 유형과 출처의 문서가 포함될 수 있으나 이 논문에서는 '행정효율과 협업촉진에 관한 규정'(이하 '행정효율 규정')

에 따른 전자문서시스템과 업무관리시스템에서 생산하는 공문서로 범위를 제한하여 논의하고자 한다.

문서의 기계가독성이나 자기 기술성을 논의하기 위해서는 먼저 이 논문의 대상인 공문서의 기안문 서식과 문서관리카드에 대해 이해해야 한다. 우리나라 공문서 기안문에는 일정한 규격이 있다. 행정효율 규정에 기안문과 대장, 보고서, 신청서 등의 서식이 정해져 있고, 이를 전자문서시스템과 업무관리시스템에서 그대로 템플릿으로 탑재하였다. 따라서, 법정 서식과 다른 형태의 전자문서를 공문서라고 주장할 경우 진본여부에 대해 의심받게 된다. 행정효율 규정 시행령 시행규칙의 [별지 제1호서식] 기안문에 정의된 서식은 〈그림 1〉과 같다.

〈그림 1〉 행정효율 규정 시행령 시행규칙
[별지 제1호서식] 기안문
(국가법령정보센터 2017)


■ 행정 효율과 협업 촉진에 관한 규정 시행규칙 [별지 제1호서식] <개정 2016. 7. 11.>

행 정 기 관 명

수신
(경유)
제목

발 신 명 의 직인

기안자 직위(직급) 서명　　검토자 직위(직급) 서명　　결재권자 직위(직급) 서명

협조자

시행　　처리과명-연도별 일련번호(시행일)　　접수　　처리과명-연도별 일련번호(접수일)

우　　도로명주소　　/ 출퇴이지 주소

전화번호)　　팩스번호)　　/ 공무원의 전자우편주소　　/ 공개 구분

210mm×297mm(백상지 80g/㎡)

비고(이 난은 서식에 포함하지 아니한다)
~ 문서를 작성할 때 "행정기관명", "발신명", "기안자", "검토자", "결재권자", "직위(직급) 서명", "처리과명-연도별 일련번호(시행일)", "도로명주소", "출제이지 주소", "공무원의 전자우편주소", "공개 구분"의 용어는 표시하지 아니하고 그 내용을 적는다.

현재 사용 중인 공문서의 기안문 서식은 일제강점기 시절에 총독부 문서 서식에서 부터 연원한다.(이상훈 2009, 182) 국가기록원 대한민국정부 수립 이후 공문서의 항목명을 일본어에서 한글로 대체하고 항목의 일부를 수정하였다. 이후 사무관리 규정에 의거하여 공문서의 서식을 정하여 정부기관의 문서 양식을 표준화하고자 하였으며, 문서의 서식은 해당 문서가 원본인지 여부를 판단하는 기초정보가 되었다. 이후 전자정부가 추진되면서 전자결재시스템이 만들어졌는데 공문서의 서식이 hwp 문서파일로 만들어져 시스템에 탑재되어 활용되었다. 전자문서시스템을 이용하여 기안문을 작성할 경우, 기안 작성 메뉴에 들어가 기안문 서식파일을 선택하면 〈그림 1〉과 같이 특정 편집기의 문서가 열리고, 비어있는 서식에 데이터를 입력하여 기안문을 작성한다.

시스템에 탑재된 기안문 서식파일은 종이문서 시절 사무관리 규정에서 정의했던 문서의 모양을 그대로 옮긴 것이다. 행정효율 규정 시행령 시행규칙에는 [별지 제1호서식] ~ [별지 제12호서식]까지 총 15종의 서식파일이 제시되어 있으며, [별표 4]에는 공문서 서식을 설계할 때의 기준을 제시하고 있다. 서식의 설계기준에는 문서의 외양(Look&Feel)에 관한 내용이 대부분이다. 문서를 종이 A4 크기로 전제하고 여백을 얼마를 두고 내용 작성을 할 것인지, 표와 선, 칸, 글자 등을 넣을 때 간격과 여백, 서체 등을 어떻게 할 것인지를 정의하고 있는데 이는 모두 출력했을 때의 문서 모양을 중심으로 설계 기준을 제시한 것이다.

공문서 생산시스템이 전자문서시스템에서 업무관리시스템으로 전환 및 확장될 때, 서식에 문서관리카드가 새롭게 추가되었다. 행정효율 규정 시행령 시행규칙에서는 〈그림 2〉와 같이 문서관리카드의 서식을 제시하고 있다. 그러나, 문서관리카드는 업무관리시스템에 서식파일로 탑재되어 있지 않고 입력 화면을 제공하고 데이터베이스에 저장하는 형태로 구현되어 있다. 따라서, 문서관리카드 서식을 설명하는 '비고'에서는 문서관리카드의 외양에 대해서는 언급하는 내용을 담지 않고, 문서관리카드의 각 항목에 어떤

값을 어떻게 넣어야 하는지에 관해서만 지시하고 있다. 주목할 것은 '제목', '수신(경유)', '기안자', '검토자', '결재권자', '공개여부' 등 문서관리카드의 항목 중 일부는 기안문 서식의 항목과 중복된다는 점이다. 이는 애초에 문서관리카드를 설계할 당시 기안문 서식의 항목을 문서관리카드에도 모두 반영하고, 그 밖에도 더 관리가 필요한 정보를 추가하여 문서관리카드에 반영한 결과이다.

〈그림 2〉 행정효율 규정 시행령 시행규칙 [별지 제6호서식]
문서관리카드 (국가법령정보센터 2017)

□ 문서관리카드
○ 문서정보
[제목]
[과제카드명]
[관련정보]
[문서요지]
[본문]
[붙임]
○ 보고경로
[구분] [직위/성명] [의견/지시] [서명] [처리결과] [이력]
○ 시행정보
[발신기관명] [발신명의]
[생산등록번호]
[공개여부]
[수신]
[(경유)]
○ 관리정보
[열람범위] [열람제한]
[온-나라 지식나라]

문서관리카드는 참여정부 대통령비서실에서 개발하여 사용했던 e지원시스템에서 유래한다.(행정자치부 2007) 기존 기안문 서식의 항목이 단순하여 정보가 불충분하다고 판단하여 문서의 맥락과 출처, 결재과정에 대해 더욱

상세한 내역을 담도록 문서관리카드를 설계한 것이다. e지원시스템 안에서는 전자문서를 활용할 때 늘 문서관리카드를 함께 참조하도록 설계되어 있었다. 문서관리카드는 기안문 서식처럼 A4 규격에 맞춰 물리적인 모양을 설계한 것이 아니었고 문서파일이 아닌 데이터베이스에 저장되도록 설계되었다.

e지원시스템의 기능을 행정기관용 버전으로 재설계하여 만든 것이 업무관리시스템이다.(남서진 외 2017) 업무관리시스템으로 설계 시 기존 기안문 서식을 없애고 문서관리카드로 대체할 것에 대해 논의가 있었으나 결론은 기안문 서식은 남게 되었고 문서관리카드도 병행하여 사용하는 것으로 정해졌다. 기안문 서식이 전통적인 종이 공문서의 틀을 고수한 상태에서 몇 가지 중요한 메타데이터를 담는 구조라면, 문서관리카드는 A4라는 크기에 얽매이지 않고 전자적 문서 생산 및 처리과정에 관한 메타데이터를 폭넓게 입력하고 저장하기 위한 구조라 할 수 있다. 이러한 결정을 내리게 된 시점은 공공기록물법을 전자기록을 중심으로 전면 개정하던 시기이다. 디지털 시대에 맞는 공문서의 모습이 어떻게 진화해야 할 것인가에 대해 좀 더 전향적인 고민을 했더라면 이 시기에 공문서의 서식을 눈에 보이는 외양 뿐만 아니라 디지털 객체의 구조를 정의하는 것으로 확장했어야 했다. 그러나, 당시에는 그런 논의를 심화시킬 기회를 미처 갖지 못했고 결과적으로 정부기관이 만들고 유통하는 공문서는 디지털이면서도 디지털 데이터로서의 활용성이 저조한 형태로 남게 되었다.

업무관리시스템은 2014년부터 클라우드 기반으로 전환되기 시작하면서 새로운 전환점을 맞이한다. 기안문 파일이 hwp포맷에서 xml기반의 odt로 바뀌게 된 것이다. odt 파일은 세부적으로 텍스트 본문을 담은 content.xml과 메타데이터를 담은 meta.xml로 구분된다. 기안문 서식에 나타나는 제목, 수신처, 결재란 등의 항목에 데이터가 들어가게 되면 odt의 content.xml에는 데이터들이 각각의 태그로 구분되어 저장되는 것이다. odt파일의 공문서를 빅데이터 분석하는 경우, content.xml과 meta.xml 각각에 어떤 정보가

구분되어 들어가 있는지 직관적으로 알 수 있으며, content.xml에 태깅된 데이터들은 태그를 통해 의미를 파악할 수 있으므로 아무런 구조정보를 제공하지 못하는 hwp포맷보다 훨씬 유리하다.(임진희 2020) 대량의 문서 정보를 적시에 분석하기 위해서는 문서 자체가 기계가독형 서식을 갖추고, 기계가 이해할 수 있는 메타데이터가 풍부할 수록 분석의 효율이 높아지기 때문이다. 이상에서 문서의 서식, 문서관리카드, 포맷의 변화 등에 관해 살펴보았다. 2장에서는 공문서를 빅데이터로서 분석하고자 한다면 어떤 관점에서 어떤 목표를 가질 수 있을지에 대해 논의하고자 한다. 또한, 그런 분석이 현재의 문서 서식, 메타데이터, 포맷으로는 왜 어려운지를 알아보고자 한다. 3장과 4장에서는 공문서가 빅데이터로 잘 활용되기 위해서는 문서파일 단위에서 어떻게 변화해야 하며, 편집기에 어떤 기능이 추가되어야 하는지에 대해 논의해보고자 한다.

2. 공문서 분석의 필요성과 서식의 문제점

1) 빅데이터 분석이 필요한 영역

앞서 소개한 행안부 “공공서식 디자인 재설계 방안 연구”의 발주 배경은 중요한 시사점을 준다. 대통령실에 파견되어 일하던 행정관료가 어느 회의 자리에서 목격한 일인데, 코로나19로 일자리 제공이 새로운 뉴딜정책으로 논의되면서 요구가 많은 일자리 종류 분석 결과를 제출하라는 지시가 대통령실로부터 내려왔다고 한다. 그런데, 설문지 수십 만 장이 hwp파일로 쌓여있기는 한데 이를 분석하려면 두 달쯤 걸린다는 정부 기관 측의 답변이 나왔고, 대통령실에서는 전자정부에서 최고라는 정부가 빅데이터 시대에 이미 쌓여있는 데이터를 분석하는데 두 달이나 걸린다는게 말이 되느냐며 크게 질책을 했다는 것이다. 이를 목격했던 행정 관료가 행안부로

돌아와 시급히 발주한 것이 바로 이 연구 용역이었던 것이다. 이처럼 빅데이터 시대에 정형데이터 뿐만 아니라 비정형데이터를 분석하는 것이 중요한 과제로 대두되고 있다.(안대진 외 2017) 앞의 사례에서는 민원문서의 하나인 설문조사서에 대한 분석이 이슈였지만, 정부기관이 생산하는 공문서도 텍스트 기반의 대형 비정형데이터로 중요한 빅데이터 분석의 대상이 된다.

필자는 기관에 근무하면서 기관 내부의 공문서를 대상으로 빅데이터 분석을 하면 여러 영역에서 유용한 시사점을 얻을 수 있을 것이라는 확신을 갖게 되었다. 빅데이터 분석의 방법론에는 통계, 데이터 마이닝, 텍스트 마이닝, 인공지능 및 기계학습, 딥러닝 등 다양한 방법이 존재한다. 물론, 2020년 현재까지도 많은 공공기관이 생산하고 보유하는 공문서의 상당수가 개방포맷이 아니어서 빅데이터 분석을 하려면 문서파일에서 먼저 텍스트를 추출하는 전처리 과정이 요구된다. 또한, 문서파일 비트스트림 내에 맥락 메타데이터가 충분히 저장되어 있지 못하기 때문에 품질 높은 분석을 하기 위해서는 별도의 메타데이터 확보 노력이 필요하다. 문서의 기계가독성이 높다는 것은 인위적인 전처리, 즉 사람이 개입하여 의미를 정해주는 과정이 없이도 순수하게 소프트웨어적으로 문서의 구조를 이해하고, 정보의 항목별 의미를 이해할 수 있다는 것을 의미한다. 그러나, 현재의 공문서는 기계가독성 수준이 낮아 빅데이터 분석에 비용이 많이 들게 된다.(최주호 외 2012)

필자가 공공기관 근무 시 공문서에 대한 빅데이터 분석을 통해 얻을 수 있는 효과에 대해 확신했던 영역은 조직관리, 인사관리, 지식관리, 기록관리, 정보관리 등 다섯 분야이다. 먼저 업무효율의 관점에서 다음과 같은 시나리오를 가정해 보았다. 첫 번째로 조직관리의 관점에서 부서별로 적정 인원을 배치해 주기 위한 근거 데이터로 문서 분석결과를 활용하는 것이다. 전체 인원이 정해진 조직 내에서 부서별 인력 배치를 얼마나 합리적으로 하느냐는 조직 전체의 업무 효율을 크게 좌우하게 된다. 조직담당은 각 부

서의 업무 양과 질에 대한 평가를 통해 중요한 일을 많이 하는 부서에 우선적으로 인력을 배치해 줄 필요가 있다. 이 때, 각 부서가 생산하고 보유하는 공문서의 통계와 내용 분석을 통해 일의 경중을 가늠해 볼 수 있다. 필자가 근무했던 기관의 경우 직원들이 작성하는 문서의 서식이 약 180종이었다. 이 중 기관장 결재 방침서 서식이 사용된 문서가 중요 시책을 담고 있는 것이어서 가장 심혈을 기울여 내용을 작성하게 되고 개수는 적지만 중요도가 매우 높을 것이다. 반면 매일 같이 현장을 순찰하면서 현황을 적어 내는 일지류는 반복적으로 수행하는 업무를 반영한 것으로 문서의 개수는 많지만 상대적 중요도는 높지 않을 것이다. 방침서와 일지는 생산 주기가 1회성인가 반복적인가, 내용에 중요 시책의 계획을 담고 있는가 단순 측정 데이터를 담고 있는가, 결재 경로가 최종 기관장까지 가는가 하위 관리자 위임전결인가, 여러 부서의 협조를 필요로 하는가 하나의 과 조직에서 수행하는가 등 여러 측면에서 차이가 있다. 결국 부서별로 문서 유형별 통계를 내어보면 부서별로 수행하는 업무의 특징을 확실히 드러낼 수 있을 것이다. 이를 기반으로 조직의 업무량이나 업무강도, 필요한 전문역량 등을 연계하여 관리한다면 조직별 적정 인원을 산출하는데 도움이 될 것이다. 문서를 통해 업무량을 추정할 수 있는 기본 데이터를 뽑아볼 수 있음에도 아직까지는 어느 기관의 조직담당도 데이터에 기반한 과학적 근거를 가지고 인력배치를 하고 있다는 발표를 접한 바가 없다. 문서를 이런 목적으로 분석하려는 시도조차 해본 일이 없었을 것이며, 현재 문서의 존재 형태가 이런 분석이 용이하지도 않기 때문일 것이다. 앞으로는 문서의 서식이나 메타데이터 관리 등을 개선하여 이런 목적의 문서 분석도 가능하도록 하는 게 필요하다.

한편, 조직에 대한 평가 시 각 조직의 업무 효율에 대해 살펴볼 필요가 있다. 행정기관에서 문서 처리는 업무의 가장 기본이다. 따라서, 클라우드 기반의 전자결재시스템 나아가 모바일결재시스템이 구축되면서 문서의 결재 처리는 시공간의 제약을 많이 벗어날 수 있는 환경으로 변화하고 있다.

문서가 결재 경로를 거치면서 걸리는 시간을 분석한다면 조직별 문서 처리의 효율성을 파악할 수 있을 것이다. 특정 관리자가 자주 결재를 지연하여 업무의 추진을 늦추고 있는 것은 아닌지 모니터링 할 수 있고, 상대적으로 긴 시간 결재 대기를 하는 부서에서는 문제가 되는 핵심 경로(critical path)를 파악하여 문제를 해소해 나가야 한다.

두 번째로 인사관리의 관점에서 근무평정을 하거나 직무역량 개발을 할 때 근거 데이터로 개인별로 작성한 문서의 분석결과를 활용하는 것이다. 필자가 근무했던 기관의 경우 일 년에 두 번 직원들에 대한 근무평정을 수행한다. 인사관리시스템에 각자 자신의 업무 내용을 입력하고 팀장이 먼저 팀원들에 대해 평가를 하면 과장이 모든 팀원들에 대한 평가를 종합하여 확정한다. 승진대상자들은 자신의 업무 성과를 한 두 쪽으로 간략히 정리하여 제출하고 인사과에서는 이를 공시하여 직원들이 모두 볼 수 있게 한 후 최종 승진자를 확정한다. 기관에서는 직원들의 사기진작을 위해 특별한 과제를 수행하느라 고생한 개인이나 팀을 선정하여 인센티브를 주는데, 역시 한 두 쪽으로 간략히 성과를 정리하여 제출하도록 하고 이를 공시한 후 선별하게 된다. 그러나, 이에 대한 다양한 내부 비판적 의견들이 나오곤 한다. 기관 내부 익명 게시판에는 근무평정과 승진 시기가 되면 “정말 뼈빠지게 일하는 사람은 제외되고 눈도장 찍기 바쁜 사람만 좋은 점수받고 승진한다”고, “진짜 성과를 보였다면 증거를 대야하는 것 아니냐”고, “남이 한 일까지 자기가 했다고 업적 기술서에 써서 올리고 승진하는 사람을 걸러야 하는 것 아니냐”고 불평불만이 쏟아지곤 한다. 단순히 불만만 토로하는게 아니라 한걸음 더 나아가 대안을 제시하는 글도 간혹 올라왔었다. 예를 들어, “공무원은 결국 문서로 일하는 것” 아니냐며, “승진후보자들은 자기가 쓴 업적기술서 외에도 직접 기안한 문서 목록과 내용을 볼 수 있게 제공하라”는 요청이 있었다. 일하는 태도와 협업 능력, 의사소통의 원활함 등은 평판에 대한 정성적 평가로 취합하고, 처리한 문서 목록을 통해 업무의 양과 질을 판단하자는 것이다. 문서 목록은 업무

관리시스템에서 간단한 기능만 덧대면 바로 취합이 가능하므로 실행 가능성이 높다.

한편, 2018년 11월 인사혁신처에서는 딥러닝 기술을 이용하여 개방형직위에 공무원을 추천하는 시스템을 개발하고자 모색 중이었다. 특정 개방형 직위에 필요한 직무능력을 도출하고, 공무원들의 현 직무능력을 측정하여 매핑함으로써 후보자들을 추려보고자 한 것이다. 나아가 공무원들이 입직한 이래 현재까지 어떤 직무능력을 확보해왔는지를 추적하여 미래의 직무능력 개발 경로를 설계해주고자 계획하였다. 이를 위해, 활용할 수 있는 유일한 데이터가 문서라는 점에 착안하여 업무관리시스템에 누적된 문서들을 분석하고자 하였다. 먼저 NCS(National Competency Standards)와 BRM 단위과제를 매핑해 두고, 공무원별로 작성 혹은 결재한 문서를 찾은 후 문서의 단위과제를 이용하여 NCS의 역량을 연결하여 파악하고자 하였다. 문서에서의 역할이 작성자, 검토자, 결재자 등 어떤 것이었는가에 따라 가중치를 달리 주고, 기획서와 집행문서를 구분하여 가중치를 주고자 하였다. 나아가 문서의 내용을 딥러닝 기반으로 텍스트 분석하여 내용적 중요도를 측정하여 가중치를 부여하고자 하였다. 공무원들의 직무 역량 관리와 경력 관리에 문서를 기본 데이터로 활용하고자 착안하였다는 점에서 필자의 관심과 맥락을 같이 하는 탐색이었다.

세 번째로 지식관리의 관점에서 조직 내 문서는 지식의 기본 원천이므로 문서 내용이 조직 내에서 잘 공유되고 재활용되어야 하는데, 자주 참조되는 문서나 문서 공유의 중심 역할을 하는 직원을 추적하는데 문서의 분석결과를 활용하는 것이다. 서울시의 경우는 과장급 이상의 결재문서를 결재 다음 날 정보소통광장(opengov.seoul.go.kr)을 통해 시민들에게 원문공개하고 있다. 평균적으로 하루에 약 1만 건의 원문이 추가로 제공되고 있다. 서울시의 정보공개와 원문공개를 벤치마킹하고자 방문하는 기관의 업무담당자의 진술에 의하면, 자기 기관의 많은 업무담당자들이 정보소통광장의 문서를 검색하여 업무지식을 습득하거나 선진사례를 학습한다는 것이다. 서울

시는 정보소통광장에 구글 애널리틱스를 탑재하여 수 년 간에 걸쳐 이용 행태를 분석하고 있다. 사이트에 접속한 도메인을 보고 추정하여 분석한 결과 약 20~30%의 이용자가 타 행정기관의 업무담당자이며, 이용자의 상당수는 서울시 직원임을 알 수 있다. 서울시의 직원들조차도 문서 생산시스템인 업무관리시스템에서 문서 검색을 하기보다는 정보소통광장에서 검색하는 것이 유용하다는 반증으로 보인다. 서울시 정보소통광장의 경우 조직 내부뿐만 아니라 외부까지 포함하여 이미 문서 정보의 지식관리시스템 역할을 하고 있는 것이다.

하나의 결재문서는 본문파일과 붙임문서 파일들로 구성된다. 본문파일은 다른 결재문서의 붙임문서로 첨부되는 경우가 자주 있다. 다수의 결재문서의 구성에 공통적으로 첨부되는 파일이 있다면 이 결재문서들은 내용적으로 동일 사안을 다루고 있거나, 하나의 트랜잭션(transaction)을 처리하는 과정에서 일정한 순서로 엮여 있거나, 관련된 의사결정에 맥락을 제공해주는 관계로 얽혀있다고 볼 수 있다. 이렇게 연관되어 있는 결재문서의 집합을 찾아낼 수 있다면 단위과제를 넘어서는 업무적 연관을 파악할 수 있고, 나아가 조직을 넘어서는 정보의 흐름을 파악할 수도 있다. 또한, 문서 파일의 참조관계에 기반한 문서 클러스터를 찾아내면 이것은 지식의 귀납적 주제 영역으로 볼 수 있다. 문서 요약(document summary) 기능을 이용해 문서별, 문서 클러스터별 요약 정보를 생성해두고, 문서 텍스트에서 주요 키워드를 추출해 두면 지식의 검색 및 활용에 도움이 될 것이다.

지식관리에서 중요한 것은 어떤 정보가 가장 많이 참조되고 재활용되는 정보인가, 그리고 어떤 부서의 어떤 정보가 가장 많이 공유되는가를 추적하는 것이다. 조직이 업무를 원활히 수행하기 위해서 필수적으로 공유해야 할 중요 문서가 무엇인지 파악하고 있어야 선제적인 문서 공유가 가능하다. 시민에게는 비공개하는 문서라 해도 조직 내부에서는 적시에 공유될 필요가 있는 문서를 찾아 제대로 접근통제를 하면서 필요한 업무담당자들 간에 공유할 수 있도록 해야 한다. 타 부서에서도 중복적으로 참조되

는 문서를 가장 많이 보유한 부서는 조직 내에서 핵심적인 역할을 하는 부서라 볼 수 있고, 해당 문서의 작성자가 수행하는 직무가 조직 내에서는 핵심적인 직무라 볼 수 있다. 지식, 즉 문서의 흐름을 추적하면 부서 간의 업무적 연관을 파악할 수 있고 이를 네트워크 모형으로 분석하면 중심 부서와 중심 직무를 파악할 수 있다. 이러한 분석 결과는 지식 공유에 중심적 역할을 하는 부서와 직원에게 적절히 인센티브를 제공하는 근거로 사용할 수 있다.

네 번째로 기록관리의 관점에서는 문서의 단위과제 오분류를 해결하고, 공개여부 설정의 오류를 해결하며, 주제분류를 생성하고, 전거정보를 추출하는 등의 현안 과제에 문서의 분석결과를 활용할 수 있다. 2017년 국가기록원이 수행한 '차세대 기록관리 재설계 연구'에서 서울시 문서 2만 건을 대상으로 딥러닝을 하여 BRM 대기능에 분류하는 파일럿 실행 결과 97%이상의 정확도를 구현한 바 있다.(김인택 등 2017, 238) 단위과제에 대한 업무설명이 잘 기술되도록 하고, 학습된 데이터를 기반으로 문서의 생산자와 제목 및 내용 키워드를 이용하여 해당 문서가 속해야 할 단위과제를 찾아내도록 할 수 있다. 업무관리시스템에서 문서가 기안될 시점에 소속될 단위과제 추천값을 보여주고 기안자가 확정하도록 하는 프로세스로 설계된다면 오분류를 사전에 걸러낼 수 있을 것으로 기대한다. 문서의 공개여부 값 역시 학습된 데이터를 이용하여 업무관리시스템에서 문서의 기안 시점에 공개/비공개/부분공개 추천값을 보여주고 기안자가 확정하도록 할 수 있다. 문서 내용 중 일부의 정보만 가리면 되는 경우에는 해당 비공개 대상 정보를 발췌하여 기안자에게 보여주고 확인받도록 하며, 비공개 문서의 경우 정보공개법 제9조제1항의 1호~8호 중 적합한 사유를 선택하도록 할 수 있다. 아카이브에서는 장기보존 문서들을 텍스트 분석하여 기본적으로 조직 출처에 따라 분류 및 저장되는 기록들을 새롭게 주제분류하여 서비스를 제공할 수 있다. 딥러닝을 이용하여 문서의 특성(feature)를 도출한 후 특성이 유사한 것들을 클러스터로 묶은 후 적절한 주제명을 붙이는 방식으로 분류가

가능하다. 또한, 아카이브에서는 전거정보에 의거하여 문서를 탐색해 볼 수 있도록 서비스를 제공하는 것이 필요한데, 메타데이터로 도출된 전거정보 외에도 텍스트에 등장하는 키워드 중에 전거정보를 찾아내어 탐색이 가능하도록 할 수 있다. 문서 텍스트 중에 중요한 정책명, 중요 인물명, 중요 장소명 등에 해당하는 단어를 찾아 전거정보로 활용한다면 유용할 것으로 예상된다.

다섯 번째로 정보관리의 관점에서는 문서 분석을 통해 조직 내에서 보유하는 문서 중 중복된 문서를 찾아 제거함으로써 스토리지의 이용효율성과 검색의 효율성을 높일 수 있다. 필자가 근무했던 기관의 경우, 업무관리시스템에서 동일한 해시값을 가진 파일이 전체의 약 30% 존재하고, 복사본 개수가 2개부터 최대 600여개까지 다양한 양상을 보여주고 있다. 기록관리시스템에서는 동일한 해시값을 가진 파일이 전체의 약 70%를 차지하고 있다. 이처럼 해시값이 동일한 사본들은 간단한 정보기술로 찾아낼 수 있다. 하지만, hwp 본문파일을 pdf로 변환하여 재활용하는 경우에는 해시값이 변경되기 때문에 동일성을 확인할 수 없다. 업무담당자의 개인적 선택으로 pdf 변환 버전을 사용하는 경우가 상당수 존재하는 것을 확인했고, 때로는 외부 시행 문서의 경우 전자문서유통시스템에 파일 용량 제한을 피하기 위해 pdf 포맷으로 변환하여 첨부하는 것을 확인한 바 있다. 혹은 hwp 본문파일에서 직인을 제거하거나 결재란의 결재자 정보만을 삭제한 버전을 만들어 첨부하는 경우도 확인한 바 있다. 이렇게 만들어진 중복적 사본이 얼마나 되는지는 확인된 바 없다. 문서 처리 관행을 개선하여 불필요한 사본이 생성되지 않도록 하는 것과 더불어 중복 사본을 제거해야 하며, 이를 위해서는 파일의 포맷은 다르지만 내용이 동일한 문서, 거의 유사한 문서를 찾아내는게 필요하다. 문서 텍스트의 유사성(similarity) 분석을 통해 하나의 문서 파일을 출처로 하여 변형된 문서들을 찾아내어 처분할 수 있다.

2) 기계가독형 문서의 요건

앞에서 다섯 분야의 관점에서 문서를 빅데이터 분석한 결과를 활용하여 얻을 수 있는 혜택에 대해 살펴보았다. 문서의 빅데이터 분석을 용이하게 하기 위해서는 문서의 기계가독성을 높여야 한다. 문서 기계가독성의 첫 번째 요건은 파일의 포맷이 개방형이어야 한다는 것, 두 번째 요건은 문서 내에 자기 기술(self-descriptive) 메타데이터를 충분히 보유해야 하는 것, 세 번째 요건은 문서의 서식이나 텍스트에 국제표준 태그를 정의할 수 있게 하는 것이다.

첫 번째 요건인 공문서의 포맷이 개방형으로 전환되어야 할 필요성과 XML기반 개방형포맷의 구조에 대한 설명은 임진희(2020)의 연구에서 밝힌 바 있다. 현재 공공기관이 보유한 공문서는 저장포맷이 개방형이 아닌 문서파일도 다수 존재하며, 클라우드 기반의 업무관리시스템으로 전환된 기관에서는 결재문서의 본문파일만 개방형인 odt 포맷으로 저장하고 있는 상태이다. 이 상태에서는 문서의 기계가독성이 떨어지기 때문에 앞에서 제시한 다양한 목적의 분석을 하기 위해 많은 비용이 소요된다. 이 논문에서는 향후 새롭게 생성하는 문서파일들을 XML기반의 개방형 포맷으로 전환하고, 기존의 문서들도 최대한 개방형 포맷으로 변환하는 것을 전제로 하여 논의를 전개하고자 한다. XML기반의 개방형 문서 포맷에서는 content.xml과 meta.xml이 분리되어 있으며, 본문 내용 텍스트가 이미 content.xml에 갈무리되어 있어 텍스트 분석의 전처리과정이 단순해질 수 있기 때문이다.

두 번째와 세 번째 요건에 대해 상세하게 다루고자 한다. 먼저, 자기기술 메타데이터란 문서 파일에 자기를 설명해주는 메타데이터를 삽입(embed) 시켜둔 것을 말한다. 공문서 파일들이 업무관리시스템, 기록관리시스템, 정보소통광장과 같은 시스템 내에만 머물 때에는 각 파일의 명칭, 용도, 생산자, 생산일시 등에 관한 정보를 부가적으로 함께 조회해 볼 수 있다. 하지

만 파일들이 시스템 밖에서 별도로 유통될 때는 문서의 내용만 돌아다니게 되고 출처를 확인하기 어렵다. 기안문이나 보고서의 본문파일의 경우는 머리말에서 살펴본 바와 같은 서식에 의해 문서번호, 제목, 결재선 등의 정보가 적혀있어 출처를 식별해낼 가능성이 높다. 그러나, 기안문 서식에 저장된 정보가 문서의 실체를 특정하는데 충분한 정보가 아니라는 점과 이 정보들이 국제표준 태그로 저장되어 있지 않아 기계가 읽고 이해할 수 없다는 점에서 한계를 갖는다.

〈그림 3〉 표준기안문.odt 파일의 content.xml 중 제목 부분

```
- <table:table-cell table:style-name="PO_Table1.A6" table:protected="true">
    - <text:p text:style-name="PO_P5">
          <text:span text:style-name="PO_T4">제목</text:span>
      </text:p>
  </table:table-cell>
- <table:table-cell table:style-name="PO_Table1.B6" table:number-columns-spanned="4">
    - <text:p text:style-name="PO_P5">
          <text:bookmark text:name="CellField7i제목iT"/>
      </text:p>
  </table:table-cell>
```

〈그림 4〉 작성된 기안문.odt 파일의 content.xml 중 제목 부분

```
- <table:table-cell table:style-name="PO_Table1.A6" table:protected="true">
    - <text:p text:style-name="PO_P5">
          <text:span text:style-name="PO_T4">제목</text:span>
      </text:p>
  </table:table-cell>
- <table:table-cell table:style-name="PO_Table1.B6" table:number-columns-spanned="4">
    - <text:p text:style-name="PO_P5">
          <text:span text:style-name="PO_T4">중소기업 생산자금 지원을 위한 공공구매론 안내 및 업무 협조 요청</text:span>
          <text:bookmark text:name="CellField7i제목iT"/>
      </text:p>
  </table:table-cell>
```

기안문 서식이 hwp 포맷으로 사용될 때는 〈그림 1〉의 '제목' 란에 값이 저장되어 있어도 그것이 제목이라는 사실을 컴퓨터가 알아챌 수 없다. 제목이 그 자리에 적힌다는 사실을 사람이 인지한 후에 기안문과 동일한 문서파일만 모아서 그 영역에 적힌 값을 제목으로 추출하도록 프로그래밍을 해줘야 한다. 하지만, 기안문 서식이 odt 포맷으로 사용될 때는 content.xml

안에 제목에 대해 〈그림 3〉과 같이 태그를 달아두어 태그의 명칭만 미리 알면 쉽게 데이터를 추출하여 활용할 수 있다. 〈그림 3〉은 표준기안문 서식의 content.xml에서 제목 부분이 어떻게 태깅되어 있는지 보여주고 있고 〈그림 4〉는 기안문 샘플의 content.xml에서 제목부분을 발췌하여 보여주고 있다. 비어있는 서식인 〈그림 3〉과 기안한 문서인 〈그림 4〉에서 제목부분의 차이는 한 줄뿐이다. 〈그림 3〉과 〈그림 4〉의 세 번째 줄에 "제목"이라고 적힌 텍스트가 〈그림 1〉에서 보이는 "제목"이라는 글자로 디스플레이된 것이다. 〈그림 4〉에만 추가된 "중소기업 생산자금 지원을 위한 공공구매론 안내 및 업무 협조 요청"은 〈그림 1〉에서 제목 옆에 기안자가 입력한 텍스트가 저장된 것으로 기안기프로그램과 약속한 바대로 〈그림 3〉의 여덟 번째 줄에 삽입되어 〈그림 4〉와 같은 결과로 변경되었다. 〈그림 4〉의 아홉 번째 줄에 "CellField7i제목iT/"라는 값이 "text:bookmark"라는 태그의 값으로 지정되어 있는데, 이는 "중소기업 생산자금 지원을 위한 공공구매론 안내 및 업무 협조 요청"이라는 텍스트에 "CellField7i제목iT/"라는 태그를 붙였다는 의미이다. 즉, 표준기안문.odt파일을 만들 때 기안문 서식 모양대로 먼저 표를 그린 후 odt로 저장하면 기본으로 생겨나는 태그들 외에 서식의 각 셀에 대해 용도와 의미를 정하여 〈그림 3〉 여덟 번째 줄처럼 특정 셀에 북마크로 "CellField7i제목iT/"와 같은 태그를 추가하여 odt파일을 생성한 것이다. 따라서, 클라우드 업무관리시스템에서 표준기안문.odt를 열어서 작성한 모든 기안문에서는 문서의 제목에 모두 "CellField7i제목iT/"라는 태그가 붙게 되고, 기안문 본문을 텍스트 분석하고자 할 때는 "CellField7i제목iT/"태그가 붙은 텍스트를 문서의 제목으로 해석하면 되는 것이다. 기안문 서식 외에도 odt포맷의 파일에서는 북마크 태그를 이용하여 문서의 텍스트 중에서 중요한 키워드를 찾아 표시해줄 수 있다.

〈그림 5〉 더블린코어 메타데이터 이니셔티브 Section 2의 'title'

Term Name: title	More details
URI	http://purl.org/dc/terms/title
Label	Title
Definition	A name given to the resource.
Type of Term	Property
Has Range	http://www.w3.org/2000/01/rdf-schema#Literal
Subproperty of	• Title (http://purl.org/dc/elements/1.1/title)

〈그림 6〉 더블린코어 메타데이터 이니셔티브 Section 3의 'title'

Term Name: title	More details
URI	http://purl.org/dc/elements/1.1/title
Label	Title
Definition	A name given to the resource.
Type of Term	Property
Note	A second property with the same name as this property has been declared in the dcterms: namespace. See the Introduction to the document DCMI Metadata Terms for an explanation.

그런데, "CellField7i제목iT/"라는 태그는 클라우드 업무관리시스템을 구축하면서 공문서 기안문 서식을 odt로 만드는 과정에서 정의한 이름이라는 한계를 갖는다. '제목'이 클라우드 업무관리시스템이라는 시스템 내에서만이 아니라 세상 모든 곳에서 통용되려면 태그의 명칭이 국제 표준을 따르도록 해야 한다. 제목 뿐만 아니라 서식에서 필드로 지정한 항목들은 모두 국제 표준 태그나 온톨로지 태그로 정의해야 기계가독성의 수준이 높아진다. 예를 들어, 문서 '제목'의 경우 ISO15836표준인 더블린코어(Dublin Core)에서 'Title'이라는 메타데이터 요소와 동일한 것으로 볼 수 있다. 'DCMI metadata Terms'(https://dublincore.org/specifications/dublin-core/dcmi-terms/#)에서는 'Title'이라는 용어를 정의하고 있는 네임스페이스가 2개 이다. 먼저 'http://purl.org/dc/terms/' 네임스페이스에서는 'Title' 용어에 대해 〈그림 5〉와 같이 정의하고 있고, 다음으로 'http://purl.org/dc/elements/1.1/' 네임스페이스에

서는 'Title' 용어에 대해 〈그림 6〉과 같이 정의하고 있다. 기안문 서식의 '제목'이 의미적으로 더블린코어의 'Title'과 동일하다면 태그를 〈text:bookmark text:name="CellField7i제목iT/〉" 대신에 "〈dc:title〉"로 붙여주어야 한다는 것이다.

odt포맷의 문서 파일에서는 meta.xml에 메타데이터를 보유하고 있다. 클라우드 업무관리시스템의 기안문 서식으로 작성한 공문서 odt의 경우에는 〈그림 7〉에서 보는 바와 같이 문자세트, 태그들의 네임스페이스 정의문이 나오고, 메타데이터로는 "PolarisOffice"라는 기안기를 사용하여 만든 문서라는 점과 문서 작성자가 "cyj"라는 것이 기본 값으로 저장되어 있다. 문서 작성자를 더블린코어의 메타데이터와 태그를 이용하여 "〈dc:creator〉"로 표기한 것을 확인할 수 있다.

〈그림 7〉 표준기안문.odt의 meta.xml 내용 (임진희 2020, 23 〈그림 6〉 재인용)

```
<?xml version="1.0" encoding="UTF-8"?>
- <office:document-meta xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xhtml="http://www.w3.org/1999/xhtml" xmlns:xforms="http://www.w3.org/2002/xforms"
  xmlns:text="urn:oasis:names:tc:opendocument:xmlns:text:1.0" xmlns:tableooo="http://openoffice.org/2009/table"
  xmlns:table="urn:oasis:names:tc:opendocument:xmlns:table:1.0" xmlns:svg="urn:oasis:names:tc:opendocument:xmlns:svg-compatible:1.0"
  xmlns:style="urn:oasis:names:tc:opendocument:xmlns:style:1.0" xmlns:smil="urn:oasis:names:tc:opendocument:xmlns:smil-compatible:1.0"
  xmlns:script="urn:oasis:names:tc:opendocument:xmlns:script:1.0" xmlns:rpt="http://openoffice.org/2005/report"
  xmlns:presentation="urn:oasis:names:tc:opendocument:xmlns:presentation:1.0" xmlns:ooow="http://openoffice.org/2004/writer"
  xmlns:oooc="http://openoffice.org/2004/calc" xmlns:ooo="http://openoffice.org/2004/office" xmlns:officeooo="http://openoffice.org/2009/office"
  xmlns:office="urn:oasis:names:tc:opendocument:xmlns:office:1.0" xmlns:of="urn:oasis:names:tc:opendocument:xmlns:of:1.2"
  xmlns:number="urn:oasis:names:tc:opendocument:xmlns:datastyle:1.0" xmlns:meta="urn:oasis:names:tc:opendocument:xmlns:meta:1.0"
  xmlns:math="http://www.w3.org/1998/Math/MathML" xmlns:manifest="urn:oasis:names:tc:opendocument:xmlns:manifest:1.0"
  xmlns:loext="urn:org:documentfoundation:names:experimental:office:xmlns:loext:1.0" xmlns:grddl="http://www.w3.org/2003/g/data-view#"
  xmlns:formx="urn:openoffice:names:experimental:ooxml-odf-interop:xmlns:form:1.0" xmlns:form="urn:oasis:names:tc:opendocument:xmlns:form:1.0"
  xmlns:fo="urn:oasis:names:tc:opendocument:xmlns:xsl-fo-compatible:1.0" xmlns:field="urn:openoffice:names:experimental:ooo-ms-interop:xmlns:field:1.0"
  xmlns:drawooo="http://openoffice.org/2010/draw" xmlns:draw="urn:oasis:names:tc:opendocument:xmlns:drawing:1.0"
  xmlns:dr3d="urn:oasis:names:tc:opendocument:xmlns:dr3d:1.0" xmlns:dom="http://www.w3.org/2001/xml-events" xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:css3t="http://www.w3.org/TR/css3-text/" xmlns:config="urn:oasis:names:tc:opendocument:xmlns:config:1.0"
  xmlns:chart="urn:oasis:names:tc:opendocument:xmlns:chart:1.0" xmlns:calcext="urn:org:documentfoundation:names:experimental:calc:xmlns:calcext:1.0"
  xmlns:anim="urn:oasis:names:tc:opendocument:xmlns:animation:1.0">
  - <office:meta>
      <meta:generator>PolarisOffice</meta:generator>
      <meta:initial-creator>cyj</meta:initial-creator>
      <dc:creator>cyj</dc:creator>
      <meta:document-statistic meta:word-count="0" meta:paragraph-count="0" meta:page-count="1" meta:non-whitespace-character-count="0" meta:character-count="0"/>
    </office:meta>
</office:document-meta>
```

현재 meta.xml에 저장된 메타데이터만으로는 해당 문서파일이 어딘가에 유통되고 있을 때 출처를 확인하기 어렵다. content.xml에 적힌 서식의 북마크 정보를 통해 출처를 식별해야 한다. 필자는 문서의 식별정보와 맥락정보, 내용정보를 최대한 meta.xml에 표기하여 저장하는 것이 필요하다고 본다. 즉, 문서관리카드의 항목들의 대부분이 저장될 필요가 있다. 물론, 앞

에서 살펴본 바와 같이 더블린코어를 포함하여 국제표준 태그를 사용하여 저장해야 한다.

다음 3장에서는 문서의 자기기술을 충족시키기 위해 어떤 메타데이터들이 확보되어야 할지에 대해 상술하고자 한다. 또한 다음 4장에서는 문서의 텍스트에 태킹을 하기 위한 편집기 기능의 요건을 제시하고자 한다.

3. 문서의 자기 기술(self-descriptive) 메타데이터 제안

1) UUID를 포함하는 식별 메타데이터

필자가 범용 고유식별자인 UUID에 관심을 두게 된 것은 MoReq2010을 보게 되면서이다. 500여쪽에 달하는 요건 문서에 65회에 걸쳐 UUID를 언급하고 있다. UUID는 'Universally unique identifier'의 줄임말로, MoReq2010에서는 기록시스템을 구성하는 모든 엔티티와 서비스에 대해 UUID를 부여하여 관리하는 것을 필수조건으로 하고 있다. 이는 기록관리의 특성을 고려한 것으로, 장기보존 기록의 경우 해당 기록을 보존하고 있는 정보시스템의 수명보다 더 길게 남겨져야 하고, 그 과정에서 여러 기록시스템에 이관되면서 보존되어야 한다. 하나의 시스템에서 다른 시스템으로 옮겨지더라도 계속해서 고유하게 식별되도록 하기 위해 UUID를 사용하도록 요건화한 것이다.(DLM Forum Foundation 2011, 237)

MoReq2010에서는 여러 시스템의 기록이 모였을 때도 유일식별자 역할을 할 수 있도록 UUID를 부여함으로써 식별자를 변경하는 일이 없도록 하라는 요건을 제시하고 있다. UUID는 기록이 생산되는 서버에서 각자 발행하여 부여해도 서로 간에 중복되지 않도록 하는 식별자 체계이다. MAC 주소와 같이 서버의 고유한 속성값을 포함하도록 설계되어 분산적으로 발행하더라도 하나로 합쳤을 때 중복되지 않도록 설계된 것이다. 하나의 기계

에서 1초에 천만 개의 UUID를 발행해주는 알고리즘이 RFC4211:2005.2를 통해 제공되고 있다.(DLM Forum Foundation 2011, 238) 물론 서로 다른 서버에서 발행된 UUID값이 중복될 가능성이 0은 아니라는 점과 UUID값의 범위가 공공기록의 전체 건수를 포괄할 수 있는가는 추가 논의가 필요하다.

공문서의 경우 무엇으로 식별되는지 검토해볼 필요가 있다. 하나의 시스템 내에서 공문서를 구별하기 위해 부여된 식별자는 공문서가 해당 시스템을 떠나는 순간 식별자로서의 역할을 못하게 된다. 특히, 여러 기관의 문서들이 합쳐지는 아카이브에서는 각 기관에서 부여한 식별자들 간에 중복된 값이 발생할 수 있기 때문에 별도의 식별자를 다시 발행하여 관리할 수 밖에 없다. 그러므로, 업무관리시스템에서는 공문서에 대한 나름의 식별자가 부여되고, 기록관리시스템으로 이관되면 또 하나의 식별자가 부여되며, 국가기록원으로 이관되면 또 하나의 식별자가 부여되는 방식으로 관리된다. 식별자에는 시스템 내에서 유일한 번호를 발행하기 위해 인위적으로 만들어주는 식별자가 있고, 서로가 구별될 수 밖에 없는 고유한 특성을 조합하여 만드는 식별자가 있을 수 있다. 후자의 예로는, 국가기록원에 최종 이관된 공문서들의 경우 기관코드, 연도, 처리과코드, 일련번호 등 4개의 메타데이터를 합쳐봐야만 각 공문서를 고유하게 식별할 수 있게 된다. 복수의 메타데이터 조합을 식별자로 사용하는 것은 시스템 관점에서 비효율적이다. 결재문서 단위, 즉 기록 건 단위로 생성단계부터 UUID를 부여하여 생애주기 전반에 걸쳐 변함없이 사용하도록 하는 것이 시스템적으로 효율적일 뿐만 아니라 유용하다.

그렇다면, 문서파일들은 무엇으로 식별될까? 공문서가 전자문서유통시스템을 거쳐 다른 기관에 시행될 때는 본문파일과 붙임파일들이 한 덩어리로 움직인다. 기록관리시스템으로 이관될 때도 동일하다. 하지만, 업무관리시스템에서 문서를 재활용하거나 정보공개포털에서 원문공개가 될 때는 문서파일 단위로 취급된다. 2020년 말 클라우드 업무관리시스템 기안기에서는

붙임파일 각각에 대해 공개여부를 설정하도록 되어있다. 공개로 설정된 붙임파일만 원문공개가 될 수 있는 것이다. 그런데, 일부 문서파일만 원문공개가 되어 유통이 될 때는 이용자가 기억하지 않고는 해당 문서파일 자체만으로는 출처를 확인할 길이 없게 된다. 앞에서 지적한 바대로 이미 각 기관에서는 문서파일을 재활용하여 만들어진 공문서들이 많다. 중복을 방지하고 어떤 문서파일이 재활용이 많이 되고 있는지를 파악하기 위해 문서파일별로 UUID를 부여하여 관리하는 것이 필요하다.

공문서 결재 단위에 부여하는 UUID는 기록 건의 UUID가 되어 기록관리시스템이나 아카이브시스템에 이관되어도 유일식별자 역할을 할 수 있어 새롭게 식별자를 부여하는 일을 줄여준다. 또한, UUID 등록시스템을 운영하게 된다면 기록 건이 생산됨과 동시에 UUID를 등록해두어 생산현황을 바로 취합해 볼 수 있다. 문서파일에 부여하는 UUID는 문서파일이 재사용되는 곳에 참조번호로 사용될 수 있고, UUID 등록시스템에 문서파일의 해시값을 함께 등록해 두면 해당 문서의 존재여부와 무결성 검증에 활용할 수 있다. 따로 돌아다니는 문서파일을 취득한 이용자가 해당 문서파일의 출처를 확인하고자 할 때 UUID 등록시스템의 API를 호출하면서 UUID와 해시값을 보내면, UUID 등록시스템에서 해당 UUID의 문서파일이 존재하는지 확인하고 보내준 해시값과 보관 중인 해시값이 동일한지 판단하여 결과를 반환해주도록 하는 것이다.

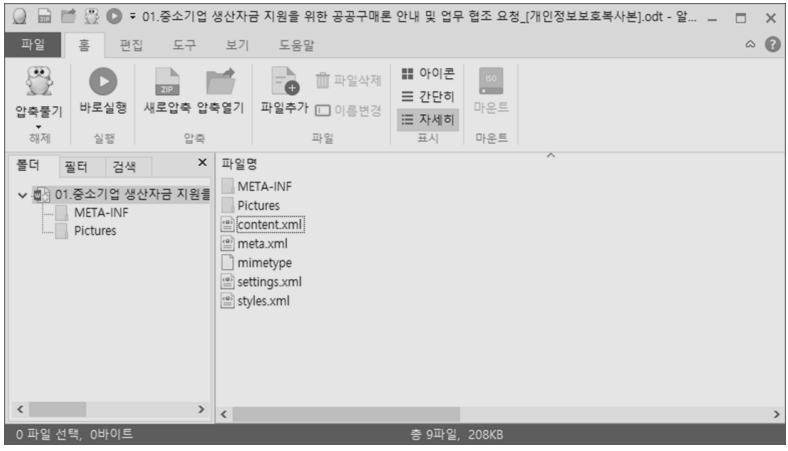
이와 같은 UUID 활용 범위는 최소한 업무관리시스템－기록관리시스템－영구기록관리시스템의 영역을 포괄할 수 있고, 나아가 클라우드 저장소에 공유하는 문서파일들에도 확장하여 적용할 수 있을 것이다. 또한, 기록 건과 문서파일 뿐만 아니라 문서의 서식, 업무담당자, 부서, 기관, 시스템 등 주요 메타데이터나 전거정보에 대해서도 UUID를 부여하여 관리하게 되면 유용할 것이다. 중앙기록물관리기관인 국가기록원은 DFR(Digital Format Registry)에 등록하는 각각의 포맷에 대해서도 UUID를 부여하여 관리하고, 공공기관에서는 보유 중인 문서파일의 포맷 정보를 기술할 때

이 UUID값을 참조하도록 한다면 유용할 것이다. 더 나아가 국가기록원이 공문서의 장기보존에 필요한 공통 RI(Representation Information)(CCSDS 2012 page4-22~4-25)들을 수집하고 등록하여 UUID를 부여하여 관리한다면, 영구기록물관리기관에서는 모두 이 정보를 참조하여 공문서에 대해 기술하고 활용할 수 있을 것이다.

2) 해시값을 포함하는 무결성 메타데이터

공공기관들은 전자기록 중심의 기록관리를 하고 있음에도 여전히 전자기록의 무결성 검증 체계를 완벽하게 구축하지 못한 상태이다. 최소한 공문서 기록에 대해서라도 무결성 확인을 위한 대안이 모색되어야 한다. 이 논문에서는 문서파일 단위의 해시값을 이용한 파일단위 무결성 검증 방법과 메타데이터의 해시값을 이용한 방법을 제안하고자 한다.

〈그림 8〉 표준기안문 서식으로 작성한 공문서의 본문파일 구조
(임진희 2020, 22 〈그림 4〉 재인용)



공문서는 본문파일과 붙임파일들로 구성된다. 이 논문에서 전제한 바와 같이 문서파일들이 모두 XML기반의 개방형 포맷으로 작성되고 저장되는 것으로 가정한다면, 문서파일들은 〈그림 8〉의 구조와 유사한 형태를 갖게 된다. odt로 저장되는 경우, 기안문 본문의 내용은 content.xml에 저장되고 메타데이터는 meta.xml에 저장된다. odt 파일 전체에 대한 해시값은 기록 건 단위에서 컴포넌트의 메타데이터 값으로 저장하여 관리하고, content.xml에 대한 해시값을 생성하여 meta.xml에 저장하도록 한다면 문서파일의 무결성 검증 체계는 더 확실해 질 것이다. 예를 들어, 어떤 문서파일이 입수되었을 때 해당 문서파일이 공문서의 일부인지를 확인하고자 할 때는 문서파일의 해시값을 생성한 후 1절에서 언급한 UUID 등록시스템의 API를 이용하여 동일한 해시값을 가진 문서파일에 관한 정보가 있는지 조회하도록 한다. 만약 동일한 해시값을 가진 문서파일에 관한 정보가 있다면 해당 문서파일을 이용하여 만들어진 공문서도 찾을 수 있게 되며, 만약 문서파일에 관한 정보가 없다면 공문서의 일부가 아니라는 것이 확인될 것이다. 나아가 1절에서 제안한 UUID를 문서파일 내의 meta.xml에 저장하도록 한다면 UUID 등록시스템을 통해 문서파일의 존재 여부를 UUID를 이용해서 확인할 수 있어 유용할 것이다. 다음은 문서파일에 발행된 UUID가 "123e4567-e89b-12d3-a456-426614174000"라 했을 때 meta.xml에 어떻게 기술될 수 있는지에 대한 예시이다. 새로운 메타데이터 항목으로 'uuid'를 명명해 태그를 달아줄 수도 있고, 앞에서 살펴본 국제표준 더블린코어의 'Identifier' 항목을 이용하여 태그를 달아줄 수도 있다. UUID 등록시스템에는 아래와 같이 문서파일에 UUID값이 저장된 후 해시값을 생성하여 등록하도록 한다.

〈meta:uuid〉123e4567-e89b-12d3-a456-426614174000〈/meta:uuid〉
〈dc:Identifier〉123e4567-e89b-12d3-a456-426614174000〈/dc:Identifier〉

만약 UUID 등록시스템 서비스가 부재한 경우라면 문서파일의 무결성 검증은 어떻게 할 수 있을까? 문서파일 안에 자기 자신에 대한 정보를 최대한 많이 넣어두도록 함으로써 무결성 검증도 용이해질 수 있다. 공문서에서는 문서 본문의 내용이 변경되는 것이 가장 큰 훼손이며 해시값이 변경되는 원인이 될 것이다. 즉, content.xml의 내용이 변경되었느냐 여부가 무결성 여부를 판단하는 기준이 될 것이다. 그렇다면, 문서파일의 content.xml에 대한 해시값을 생성하여 meta.xml에 함께 저장해두면 이 해시값을 이용하여 내용의 변경이 있었는지 검증할 수 있게 된다. 아래는 그 예시이다.

〈meta:content-hashval〉e0d123e5f316bef78bfdf5a008837577〈/meta:content-hashval〉

3) 문서관리카드 메타데이터

결재문서를 기안하는 과정에서 문서관리카드가 만들어진다. 〈그림 2〉와 같은 항목들로 설계된 문서관리카드는 〈그림 9〉와 유사한 화면을 통해 입출력된다. 앞에서 문서관리카드의 유래를 설명했듯이, 문서관리카드의 정보들은 결재문서를 설명하는 중요한 정보들로 이후 문서를 관리하는 데에도 꼭 필요한 정보들이다. 그러나, 문서관리카드의 정보는 데이터베이스에 저장되었다가 〈그림 9〉와 같이 업무관리시스템에서 조회하여 볼 수 있을 뿐이며, 문서의 본문파일 및 붙임파일과는 따로 유통되게 된다. 기록관리시스템에도 문서관리카드의 일부 정보만이 이관된다.

'문서정보', '결재경로', '시행정보'의 대부분의 항목은 의미 있는 메타데이터이다. 따라서, 이 데이터들을 문서파일에 저장하여 함께 다니도록 하여 자기 기술성을 높여 주는 게 좋다. 다만, 문서관리카드의 정보 중에는 공문서 전체에 관한 메타데이터와 문서파일 각각에 대한 메타데이터가 섞여 있으므로 이를 구분하여 판단하고 저장할 곳을 찾아 주는게 필요하다. 현재

〈그림 9〉 업무관리시스템의 문서관리카드 예시 화면
(임진희 2020, 9 〈그림 1〉 재인용)

까지 모든 버전의 업무관리시스템에서는 결재문서의 본문파일과 붙임 문서파일 단위의 관리 속성값을 명시적으로 충분히 관리하고 있지 않다. 특히 본문파일은 공문서 그 자체로 인식하는 경향이 있어 메타데이터 관리에 혼선이 빚어지고 있다.

예를 들어보자. '문서정보'의 '제목'은 본문파일 서식의 제목란의 값과 동일하게 유지되는 항목이다. 그렇다면 이 제목은 본문파일에만 해당하는 메타데이터인가 아니면 붙임 문서파일 전체에 해당하는 메타데이터인가? '문서번호'는 결재가 완료되는 시점에 채번되어 기안문 본문파일에도 적히고 문서관리카드에도 저장되는 값이다. 그렇다면 이 문서번호는 본문파일의

번호인가, 아니면 붙임 문서파일까지 포함한 결재문서 전체의 번호인가? '공개 여부'의 값은 결재문서 전체에 대한 설정값이다. 그런데, 〈그림 9〉에서는 상세히 표시되지 않았지만 업무관리시스템에서는 붙임 문서파일별로 공개여부를 설정하게 되어 있고, 본문파일에 대한 공개여부 설정 기능만 없다. '문서정보'의 '공개 여부'에는 문서파일별 설정값에다 본문파일에 대한 설정값까지 포함한 전체 설정값을 넣게 되어 있을 뿐이다. 업무담당자들이 자주 혼란에 빠지게 되는 부분이다. 향후 모든 전자결재시스템에서는 본문과 붙임 문서파일 각각에 대해 공개 여부 설정을 하도록 기능이 마련되어야 하며, 결재문서 전체에 대한 공개여부는 각 문서파일들에 대한 설정값을 합하여 산출되도록 해야 한다.

비공개 문서의 경우 조직내 열람 제한을 두기 위해 '열람 범위', '열람 제한'의 항목을 두고 있다. 이 설정값은 본문파일에 대한 것인가, 아니면 모든 파일에 대한 것인가? 이 역시 문서파일별로 따로 설정될 필요가 있는 항목이다. 비공개해야 하는 문서파일에 경우에만 열람 범위와 열람 제한을 기술해야 하기 때문이다. 그러나, 현재로서는 어느 문서파일에 대해 제한을 두는 것인지 표시가 되지 않는다. 앞에서 기술한 바처럼, 공문서의 본문파일은 재활용되어 다른 공문서에 붙임문서로 사용되는 경우가 많다. 애초에 공개로 설정되었던 본문파일이 다른 문서에서 붙임문서로 사용되면서 전체가 비공개로 설정되는 오류들이 벌어지는 이유는 문서파일 단위로 정밀하게 공개 여부와 열람 제한을 설정할 수 없도록 되어 있기 때문이다.

'결재 경로' 영역에는 기안자, 검토자, 결재자, 기안일시, 검토일시, 결재일시, 여러 버전의 본문파일들이 기술되어 있으며 이들 모두 중요한 메타데이터이다. 기안자, 검토자, 결재자 등에 대해 화면 상에는 직위(직급), 이름만 나와 있지만 고유식별자 정보도 함께 포착해야 한다. '시행 정보' 영역에서도 화면 상에는 조직명칭만 나오지만 고유식별자 정보도 함께 포착해야 한다. 그런데, 문제는 이 고유식별자가 업무관리시스템 내에서만 통용되는 값이라는 것이다. 장기보존 기록의 경우 조직 전거, 인물 전거 등을 관리해

야 하므로 전거정보에 해당하는 속성값들은 1절에서 설명한 UUID로 식별되도록 하는 것이 바람직할 것이다.

이상에서 살펴본 바와 같이 문서관리카드의 내용 중에는 결재문서 전체의 내용과 본문파일에 국한된 내용이 섞여있다. 또한, 붙임 문서파일들에 해당하는 메타데이터도 숨겨져 있다. 그렇다면, 문서관리카드의 내용을 문서에 자기기술 메타데이터로 포함하기 위해서는 어떤 방법이 있을까? 두 가지 해결 방안이 있다. 첫 번째는 결재문서 전체에 해당하는 메타데이터는 따로 doc_mata.xml(가칭)로 만들어서 본문 문서파일에 추가하고, 본문이나 붙임 문서파일 단위의 메타데이터는 각 문서파일의 meta.xml에 저장하도록 하는 방안이다.(오세라 외 2016) 이 방안에서는 본문 문서파일이 나머지 붙임 문서파일들 보다는 더 많은 메타데이터를 저장하고 중요한 역할을 하게 된다. 본문 문서파일이 다른 결재문서의 붙임 문서로 재활용될 때에도 원출처를 분명히 하면서 첨부될 수 있다는 점에서 장점을 갖게된다. 두번째는 결재문서 전체에 해당하는 메타데이터를 별도의 RDF(Resource Description Framework)로 만들어 발행하고,(하승록 외 2017) 분문이나 붙임 문서파일 단위의 메타데이터는 각 문서파일의 meta.xml에 저장하도록 하는 방안이다.

문서관리카드의 정보 외에도 문서파일에는 작성 환경에 관한 정보도 포함되는 것이 필요하다. 예를 들어, 기안기에서 작성되는 경우 업무관리시스템 명칭과 버전, 기안기 어플리케이션 이름과 버전 등을 추가하고, PC설치용 편집기에서 작성되는 경우 PC 정보, 운영체계 정보, 편집기 정보 등을 추가하는게 유용하다. PC에서 작성된 문서가 업무관리시스템에 업로드될 때는 업무관리시스템의 환경 정보가 메타데이터로 추가되도록 구성되어야 한다. 업무관리시스템도 기관마다 인스턴스가 다르므로 각 인스턴스마다 고유의 UUID를 부여하여 메타데이터에 포함시킬 필요가 있다.

4. 문서 텍스트 태깅을 위한 기능 요건

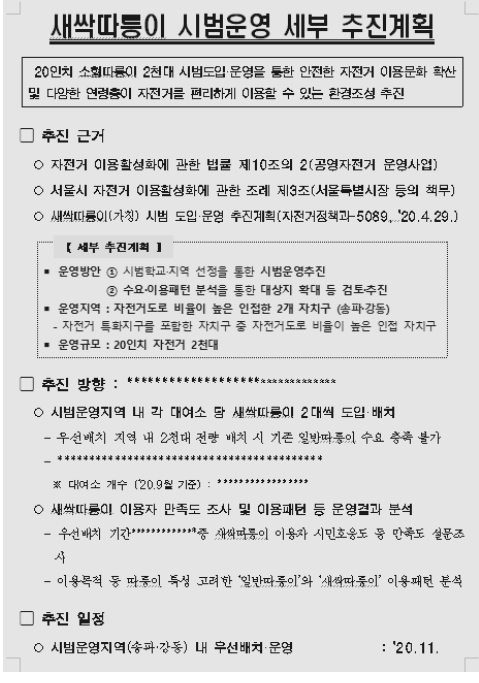
1) 전거정보 태깅 프로세스

XML기반의 개방형포맷으로 문서파일이 만들어지는 경우에는 문서파일의 텍스트에서 원하는 단어에 태깅을 할 수 있다. 지식관리 관점에서나 기록관리 관점에서는 문서에서 중요한 키워드를 표시해두는 것이 활용성을 높인다. 현재는 결재문서 작성 과정에서 키워드를 태깅하도록 하는 업무적 절차가 없고 따라서 편집기에서도 태깅 기능을 제공하고 있지 않다. 예외적으로 서울시 업무관리시스템에서는 한컴 기안기에 부가기능을 넣어서 기안문 본문의 텍스트 일부를 마킹하도록 하고 이 정보를 이용하여 해당 텍스트만 별표(*)로 치환하여 pdf로 만들어 정보소통광장에서 원문공개하고 있다.(오진관 등 2017) 그러나, 이 기능은 표준화된 태그 방식이 아니고 기계가독형 문서관리카드에 문서의 제목에는 나타나지 않는 핵심 키워드를 적도록 하는 란이 마련되어 있지만 내용을 기입하는 경우가 드물고, 또한 이런 방식은 본문 텍스트의 어느 위치에 해당 키워드가 출현하는지를 알 수 없기에 문서를 검색하는 키워드로 활용도가 제한된다. 반면 본문 텍스트에서 키워드가 나오는 그 자리에 태깅을 해둔다면 검색의 키워드로도 활용될 수 있을 뿐만 아니라, 본문의 해당 키워드에 하이퍼링크가 만들어져서 클릭을 통해 키워드에 대한 상세 정보를 사전에서 확인하거나 동일한 키워드가 들어있는 타 문서의 목록으로 연결되도록 서비스할 수 있게 된다.

예를 들어, 공문서의 경우에는 텍스트에 등장하는 주요 정책명이나 중요한 인물명, 장소명, 사건명 등에 태깅을 해두어 전거정보로 활용하면 유용할 것이다. 정책명의 경우 해마다 역점 시정사업이 정해지므로 이를 누적적으로 사전(dictionary)으로 만들어 두고, 문서 본문 텍스트에서 해당 사업명을 찾아 태깅하도록 할 수 있다. 조직/인물명, 장소명, 사건명의 경우 향후 검색 키워드로 사용될 가능성이 있는 것들을 대상으로 기안자가 판단하

여 태깅하고, 태깅되는 명칭을 사전에서 찾아 연계하거나 새로운 명칭은 등록하도록 할 수 있다. 이 때, 사전에 등재되는 명칭들은 UUID를 부여하여 관리하고, 동명이의어를 걸러내도록 한다. 또한, 문서에 등장하는 타 문서 번호도 태깅하면 문서 간의 연결 관계를 추적할 수 있어 유용할 것이다.

〈그림 10〉 원문공개된 추진계획서 예시
(정보소통광장 2020)



새싹따릉이 시범운영 세부 추진계획

20인치 소형따릉이 2천대 시범도입·운영을 통한 안전한 자전거 이용문화 확산 및 다양한 연령층이 자전거를 편리하게 이용할 수 있는 환경조성 추진

□ 추진 근거

○ 자전거 이용활성화에 관한 법률 제10조의 2(공영자전거 운영사업)

○ 서울시 자전거 이용활성화에 관한 조례 제3조(서울특별시장 등의 책무)

○ 새싹따릉이(가칭) 시범 도입·운영 추진계획(자전거정책과-5089, '20.4.29.)

【 세부 추진계획 】

▪ 운영방안 ① 시범학교지역 선정을 통한 시범운영추진
② 수요·이용패턴 분석을 통한 대상지 확대 등 검토추진

▪ 운영지역 : 자전거도로 비율이 높은 인접한 2개 자치구 (송파·강동)
- 자전거 특화지구를 포함한 자치구 중 자전거도로 비율이 높은 인접 자치구

▪ 운영규모 : 20인치 자전거 2천대

□ 추진 방향 : **********

○ 시범운영지역 내 각 대여소 당 새싹따릉이 2대씩 도입·배치

- 우선배치 지역 내 2천대 전량 배치 시 기존 일반따릉이 수요 충족 불가

- **********

※ 대여소 개수 (20.9월 기준) : **********

○ 새싹따릉이 이용자 만족도 조사 및 이용패턴 등 운영결과 분석

- 우선배치 기간**********중 새싹따릉이 이용자 시민호응도 등 만족도 설문조사

- 이용목적 등 따릉이 특성 고려한 '일반따릉이'와 '새싹따릉이' 이용패턴 분석

□ 추진 일정

○ 시범운영지역(송파·강동) 내 우선배치·운영 : '20.11.

〈그림 10〉에 보이는 결재문서를 예시로 살펴보자. 이 문서에서는 핵심 사업명이 '새싹따릉이'이며, 제목과 본문 안에 '새싹따릉이'가 여섯 번 출현하고 있다. 현재의 문서관리 체계에서는 제목에 등장하는 '새싹따릉이'는 검색엔진에 의해 키워드로 색인이 되지만 본문의 '새싹따릉이'까지 색인하지

못하는 경우가 많다. 여섯 개의 '새싹따릉이' 단어에 모두 동일한 명칭으로 태깅을 해둘 수 있다. '추진 근거'의 세 번째 내용이 odt포맷으로 만들어진다면 content.xml은 〈표 1〉와 같은 형태가 될 것이다.

〈표 34〉 태깅 전의 content.xml 예시

```
<text:span text:style-name="PO_T5">
    O 새싹따릉이(가칭) 시범 도입·운영 추진계획(자전거정책과-5089. '20.4.29.)
</text:span>
```

'새싹따릉이'가 중요 시책에 관련된 사업이라면 〈표 2〉과 같이 태깅을 해줄 수 있다. 중요 시책을 표시하는 북마크의 이름을 "authority"로 명명한다고 가정해 보았다. 여섯 군데에 동일한 태깅을 해놓고, 가상의 동일 UUID 값을 주도록 한다.

〈표 2〉 태깅 후의 content.xml 예시

```
<text:span text:style-name="PO_T5">
    O 새싹따릉이</text:span><text:bookmark text:name="authority">
    <meta:uuid>550e8400-e29b-41d4-a716-446655440000</meta:uuid>
<text:span text:style-name="PO_T5">
    (가칭) 시범 도입·운영 추진계획(자전거정책과-5089. '20.4.29.)
</text:span>
```

공문서를 작성하는 과정에서 태깅하는 시점은 먼저 기안자가 기안기에서 표준기안문 파일을 열어 본문 작성을 마치고 결재를 상신하는 시점에서 이루어지도록 하는게 좋다. 그 다음으로 검토자와 결재자가 검토와 결재하는 시점에 현재의 태그 정보를 확인하고 필요시 추가 태깅을 하도록 한다. 태깅할 대상을 찾는 방법은 첫째, 기안자가 의식적으로 키워드를 찾아내어 태

깅을 하는 경우와 둘째, 키워드 추천도구를 작동하여 추천된 키워드 중에 선택하여 확정하는 방식이 가능하다. 태깅하는 시점에 해당 키워드가 사전에 이미 등재된 항목인지 확인해보고 기존 항목이면 이미 할당된 UUID값을 가져와 태그에 포함시킨다. 만약 새로운 항목이라면 UUID발행기를 호출하여 식별자를 부여받고 사전에 항목을 등록하도록 한다.

전거정보를 태깅하려면 전거 사전(dictionary)이 만들어져 관리되어야 한다. 사전에는 전거정보의 종류를 시책, 인물, 조직, 장소, 사건 등으로 구분해주고, 용어별 UUID와 설명정보를 저장해 둔다. 그리고, 해당 용어가 태깅된 문서파일의 UUID들도 함께 관리하여, 하나의 전거정보를 클릭했을 때 관련 문서파일들을 조회할 수 있게 해준다.

2) 서식 태깅 프로세스

공문서는 용도에 따라 정해진 서식을 사용하는 경우가 많다. 서식은 반복적으로 작성해야 하는 문서의 경우 기입해야 하는 항목을 표준화하여 빠짐없이 정보를 기입하도록 하여 충실한 문서가 되게 하는 효과가 있다. 공공 서식의 경우, 표준화된 양식에 따라 기입하지 않으면 제출하는 서류가 무효화될 수 있기에 서식을 준수하는 것은 법적 효력을 좌우하는 중요 요소가 되기도 한다. 필자가 근무했던 기관의 경우 180종이 넘는 서식이 공문서에 사용되고 있는 것으로 파악되었고, 그 외에도 공식적인 서식을 정해두지는 않았지만 정기적으로 동일한 양식에 정보를 채우는 일지류도 상당히 많았다. 표준서식이나 정해진 양식에 의거하여 반복적으로 만들어지는 문서들은 빅데이터 분석을 할 때 동일한 종류의 데이터로 파악할 수 있어야 한다. 그러기 위해서는 서식 자체에 UUID를 부여하여 관리하도록 하고, 해당 서식으로 만들어지는 문서파일의 meta.xml에는 서식 UUID도 저장하도록 하여 문서의 식별성을 높이는게 필요하다.

대표적인 공문서 서식으로 〈그림 1〉의 표준기안문을 보면, 〈그림 3〉과

〈그림 4〉에서 보듯이 표의 각 셀에 대해 “제목”이라는 의미를 부여하고 있지만 클라우드 업무관리시스템에서만 통용되는 태그로 표시함으로써 그 의미가 협소해지고 기계가독성의 수준이 낮아졌다. 표준기안문 서식은 미리 odt파일로 만들어져서 업무관리시스템에 탑재되어 있고, 기안문 작성 시 이를 복사하여 사용하도록 되어 있다. 기안기에서는 미리 만들어 둔 서식의 태그들이 잘 보존되어 그대로 odt에 저장하도록 만들어졌다. 하지만, 기안기 외의 타 편집기에서 표준기안문 odt 파일을 열어 서식에 값을 넣고 저장하게 되면 그 odt 파일에서는 서식의 태그들이 사라지게 된다. 왜냐하면 태그들이 표준으로 널리 공표되어 모든 odt 편집기에서 지켜줄 수 있는 규격의 범주에 들어가 있지 않기 때문이다. 현재 기안문 서식의 셀 태그명들은 문서관리카드에 적히는 내용과 기안문의 내용을 일치시켜야 하는 업무관리시스템 기안문 만의 요구를 충족시키기 위해 만들어진 것이다.

기계가독형 문서 서식을 만들기 위해서는 문서 편집기에 추가 기능이 필요하다. 서식의 특정 셀에 저장되는 값에 의미를 지정하기 위해 국내외 표준 메타데이터를 활용하여 태깅을 하도록 하는 것이다. 물론, 이렇게 만들어진 서식의 태그들은 포맷의 표준 규격에 포함되도록 하고, 문서가 저장되거나 포맷변환이 되어도 최대한 사라지지 않도록 보장되어야 한다.

한편으로는 공문서의 서식 등록소를 두어 서식의 명칭과 목적을 입력하고 UUID를 발행하도록 해야 한다. 발행된 UUID를 서식파일의 meta.xml에 넣어두도록 하면, 해당 서식파일을 이용해서 만들어진 문서파일들은 동일한 서식 UUID에 의해 같은 구조를 가진 파일임을 바로 식별할 수 있게 된다. 서식, 문서파일, 전거정보 등 모든 주요 정보들이 UUID로 식별되고, 연결되며, 추적되도록 하는 것이다.

3) 기계가독형 문서의 편집기 기능

기계가독형 문서가 되기 위한 요건은 파일 포맷이 XML기반의 개방형이

라는 전제하에 문서 파일 내에 자기 기술 메타데이터를 충분히 보유하도록 하고, 서식이나 텍스트에 국제표준 태그를 정의할 수 있도록 하는 것임을 살펴보았다. 그런데, 이런 제안들은 문서를 생성하는 편집기에서도 관련 기능을 제공해야 실현 가능해진다. 문서파일에 결재환경과 과정에 대한 메타데이터를 충분히 저장한다는 것은 결재문서의 본문파일뿐만 아니라 붙임문서파일에도 공통적으로 적용되어야 의미가 있다. 따라서, 본문파일을 편집하는 기안기뿐만 아니라 붙임 문서파일을 작성할 때 사용하게 되는 PC설치형 오피스제품과 클라우드용 웹오피스제품 모두에게 필요한 기능이다. 또한, 서식에 태깅하거나 전거정보를 태깅해야 하는 문서는 본문파일만이 아니라 붙임문서도 대상이 될 수 있다. 본문파일은 기안기를 이용하여 작성하지만 붙임 문서파일들은 대부분 PC에서 작업한 후 업로드하는 방식이므로 기안기가 아니라 PC 설치용 편집기를 이용하여 태깅을 해야 한다. 만약 클라우드에 문서를 올려놓고 편집을 하는 경우라면 웹오피스 제품을 이용하여 태깅을 해야할 수도 있다. 결국, 기안기, PC 설치용 오피스, 웹오피스 등 3 종류의 오피스제품군 모두가 자기 기술 메타데이터를 저장하고 서식이나 전거정보를 태깅하는 기능을 갖고 있어야 한다.

현재 우리나라 공공기관에서 사용하는 전자문서시스템, 업무관리시스템 등에 탑재된 기안기는 한컴, 폴라리스오피스 두 종류이다. 두 제품 모두 클라우드 업무관리시스템에 탑재된 기안기 버전인 경우에 XML기반 개방형포맷인 odt로 기안문을 작성하면서 공문서의 기안문 서식을 지원하고 있다. (임진희 2020) PC설치형 오피스는 대부분의 기관에서 한컴과 마이크로소프트 오피스 제품을 겸용하고 있는 것으로 알고 있다. PC에서 만드는 hwp는 개방형 포맷이 아니므로 이 논문의 범위에서 제외된다. 마이크로소프트 오피스는 XML기반의 개방형포맷인 OOXML파일인 docx, xlsx, pptx를 생성하고 있으므로 이 논문의 범위에 포함된다. 클라우드용 웹오피스는 공공기관에서 아직 본격적으로 도입하지 않고 있으므로 이 논문에서는 논의에서 제외한다.

기안기와 PC설치형 오피스 제품군에서 자기 기술 메타데이터를 확보하여 넣도록 하고, 서식이나 전거정보에 태깅할 수 있도록 하기 위해서는 해당 기능에 대한 추가가 각각 필요하다. 대부분의 편집기에는 기본 기능 외에도 플러그-인 도구를 탑재하여 사용할 수 있도록 API를 제공한다. 앞에서 논의했던 요건들을 소프트웨어적으로 구현하여 플러그-인 도구로 제공하는 벤더가 있다면 편집기에서 기능을 확장하여 사용할 수 있다는 것이다. 이 논문에서 요구하는 기계가독형의 문서를 작성하려면 다음과 같은 플러그-인으로 도구들이 필요하다. 첫째, 업무관리시스템에서 문서관리카드와 시스템 정보를 획득하여 doc_meta.xml(가칭)을 만들어 주는 도구, 둘째, 업무관리시스템에서 문서파일의 자기 기술 메타데이터를 획득하여 meta.xml에 저장해 주는 도구, 셋째, PC환경에서 문서파일의 자기 기술 메타데이터를 획득하여 meta.xml에 저장해 주는 도구, 넷째, 문서 본문의 서식을 만들 때 항목별 태그를 임의로 정할 수 있고 이 정보가 content.xml에 잘 유지되어 저장되도록 해주는 도구, 다섯째, 문서 작성 중 본문의 텍스트에 태그를 추가할 수 있고 이 정보가 content.xml에 잘 유지되어 저장되도록 해주는 도구 등이다. 태그를 정할 때 기존의 활용가능한 국제표준 태그들 중에 선택할 수 있도록 해주는 기능이 도구에 포함되는게 필요하다.

편집기에 추가되어야 하는 도구들 외에도 자기기술 수준을 높이고 기계가독성을 높이기 위해 부가되어야 할 서비스들은 다음과 같다. 첫째, UUID 발행기 서비스, 둘째, UUID 등록소 서비스, 셋째, 사전(dictionary) 서비스, 넷째, RDF 발행기 등이다. 편집기 플러그-인 도구들이 선택적으로 이 서비스들을 이용하도록 하여 기계가독성을 더욱 높일 수 있다. 예를 들어, 자기기술을 위해 추가하는 메타데이터들 중에서 UUID를 새로 발행하여 함께 저장해야 하는 경우 먼저 UUID 발행기 서비스를 호출하고, 발행과 동시에 UUID 등록소 서비스를 호출하여 UUID와 객체의 명칭, 역할, 주요 속성정보를 저장하도록 한다. 전거정보를 태깅을 할 때는 사전 서비스를 호출하여 기존의 용어인지 확인한다. 새로운 전거정보인 경우 UUID 발행서비스

를 호출하여 UUID를 받아 사용하고 사전 서비스를 호출하여 UUID와 용어와 설명정보, 문서의 UUID를 저장하도록 한다. 결재문서 단위나 문서파일 단위로 메타데이터들을 모아 온톨로지 문서인 RDF로 발행하고자 할 때는 RDF 발행기를 호출하도록 한다.

5. 맺음말

필자는 기록관리 분야에서 십오 년이 넘게 연구자로 지내오면서 전자기록관리에 관한 여러 정책이나 법령 개정 사항에 직간접적으로 관여해 왔다. 기관에 재직할 때는 매월 70만 건 가량의 결재문서가 만들어지는 것을 보면서 이 많은 문서들을 단순히 보존기간에 맞춰 폐기하는 일에 집중하는 것이 기록관리의 핵심은 아니지 않은가 고민하게 되었다. 많은 비용을 들여 보유하고 있는 문서들이 의미있는 자원으로서 기관 내에서 유용하게 활용되었으면 좋겠다는 바람과 함께 아카이브에 쌓이는 공문서들이 갖는 가치를 실증적으로 분석해봐야 하지 않나 하는 과제를 품고 있었다. 한편으로는 빅데이터 분야가 정형데이터 분석뿐 아니라 반정형, 비정형 데이터들을 분석하여 통찰을 얻는 방향으로 진전함에 따라 공문서야 말로 대량의 반정형, 비정형 데이터로서 분석의 중요한 원자료가 될 것이라는 생각을 하게 되었다. 2장에서 조직관리, 인사관리, 지식관리, 기록관리, 정보관리 등의 관점에서 문서 분석의 유용성을 제시한 것은 근무했던 기관의 지식관리 현황이나 인사관리나 조직관리의 아쉬움을 보면서 고민한 결과이다. 하지만, 아직까지는 가상의 시나리오일 뿐 실제로 구현하여 실증을 거치려면 연구개발이 선행되어야 한다.

서울시는 업무관리시스템을 클라우드 기반으로 재구축하는 기회를 만들면서 서울시가 보유하는 문서를 최대한 개방형포맷으로 전환하고, 나아가 자기 기술 메타데이터를 문서파일에 충분히 저장하여 빅데이터 분석이 용

이하게 만들기 위한 준비도 진행한 바 있다. 그 과정에서 한컴오피스나 티맥스오피스 등 국내 오피스제품 벤더사들이 4장에서 제시한 본문 텍스트 태깅이 가능하도록 편집기에 기능을 추가할 의사가 있다는 점도 확인되었다. 오피스제품 벤더사들은 공공영역뿐 아니라 기업영역에서도 기계가독형 문서서식에 대한 요구가 있어 기술 개발을 적극적으로 해나가겠다는 입장이었다.(국가기록원 2020)

문서파일의 기계가독성을 높이는 것은 한국지능정보사회진흥원장의 기고문에서도 언급된 것처럼 이제는 디지털 시대, 빅데이터 시대에 조응하기 위해 시급히 국가적으로 추진해야할 과제이다.(동아일보 2020) 문서의 생산방식에서 부터 근본적인 변화가 수반되어야 하는 일이기 때문에 기록관리계에서도 이 과제를 소극적으로 대해서는 안될 것이다. 기록관리계에서도 문서파일의 기계가독성을 높일 수 있는 방안을 적극적으로 모색하고, 문서분석을 기반으로 지능적인 기록관리를 할 수 있도록 투자하고 관심을 가질 필요가 있다.

〈참고문헌〉

국가기록원. 2020.7.8. 문서 파일포맷과 서식 개선 방안－서울시 사례를 중심으로. https://www.youtube.com/watch?v=UN8fqGAl7vM

국가법령정보센터. 2017.10.17. 행정 효율과 협업 촉진에 관한 규정 시행령 및 시행규칙. https://law.go.kr/lsSc.do?section=&menuId=1&subMenuId=15&tabMenuId=81&eventGubun=060101&query=%ED%96%89%EC%A0%95%ED%9A%A8%EC%9C%A8%EA%B3%BC+%ED%98%91%EC%97%85#undefined (2020년 12월 30일 접근)

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

남서진, 임진희. 2017. 정부기능분류체계(BRM)의 재정비를 위한 사례연구－'문화재' 정책영역을 중심으로. 『한국기록관리학회지』, 17(2), 129-163.

동아일보. 2020.12.17. 인공지능 시대 정부 문서, 근본을 바꿀 때다. https://www.donga.com/news/Opinion/article/all/20201216/104481261/1 (2020년 12월 30일 접근)

안대진, 임진희. 2017. 제4차 산업혁명 기술의 기록관리 적용 방안. 『기록학연구』, 54, 211-248.
오세라, 정미리, 임진희. 2016. 공개포맷에 기반한 전자 기록 보존 포맷 재설계 방향 연구. 『한국기록관리학회지』, 16(4), 79-120.
오진관, 오세라, 최광훈, 임진희. 원문정보공개 지원을 위한 민감정보 필터링 요건에 관한 연구. 『한국기록관리학회지』, 17(1), 51-71.
이상훈. 2009.7. 한국정부 수립 이후 행정체제의 변동과 국가기록관리체제의 개편(1948년~64년). 『기록학연구』, 21, 169-246.
이화여자대학교 산학협력단. 2010.4 온톨로지 기반 태그를 기반으로한 정보 검색 기법. 교육과학기술부 한국연구재단 https://scienceon.kisti.re.kr/srch/selectPORSrchReport.do?cn=TRKO201000014135 (2020년 12월 30일 접근)
임진희. 2020. 클라우드 환경에서 공문서 파일포맷의 선택 전략. 『기록학연구』, 66, 5-35.
정미리, 오세라, 임진희. 2016. 공문서 컴포넌트 오픈포맷 채택이 기록관리에 미치는 영향 분석. 『한국기록관리학회지』, 16(2), 29-55.
정보소통광장. 2020.11.4. 새싹따릉이 시범운영 세부 추진계획. https://opengov.seoul.go.kr/sanction/21539219 (2020년 12월 30일 접근)
정보통신기획평가원. 2020.12. 2021년도 제1차 정보통신·방송 기술개발사업 및 표준개발지원사업 신규지원 대상과제 공고. https://ezone.iitp.kr/common/anno/02/form.tab?PMS_TSK_PBNC_ID=PBD202000000098 (2020년 12월 30일 접근)
최주호, 이재영. 2012. 전자기록물의 메타데이터 추출 및 비교 검증 기술 연구. 『한국기록관리학회지』, 12(1), 7-32.
하승록, 안대진, 임진희. 2017. 기록정보 LOD 구축을 위한 의미 상호연결 자동화 실험 연구. 『한국기록관리학회지』, 72(4), 177-200.
행정자치부. 2007. 청와대비서실의 보고서작성법. https://copy.or.kr/884 (2020년 12월 30일 접근)
행정안전부. 2020.12. 공공서식 디자인 재설계 방안 연구. http://www.prism.go.kr/homepage/entire/retrieveEntireDetail.do?pageIndex=1&research_id=1741000-202000076&leftMenuLevel=160&cond_research_name=%EA%B3%B5%EA%B3%B5%EC%84%9C%EC%8B%9D&cond_research_start_date=&pageUnit=10&cond_order=3
CCDSD. 2012. REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS). https://public.ccsds.org/Pubs/650x0m2.pdf (2020년 12월 30일 접근)
DLM Forum Foundation. 2011. MoReq2010. https://www.moreq.info/files/moreq2010_vol1_v1_1_en.pdf (2020년 12월 30일 접근)