발 간 등 록 번 호
11-1741050-000032-14

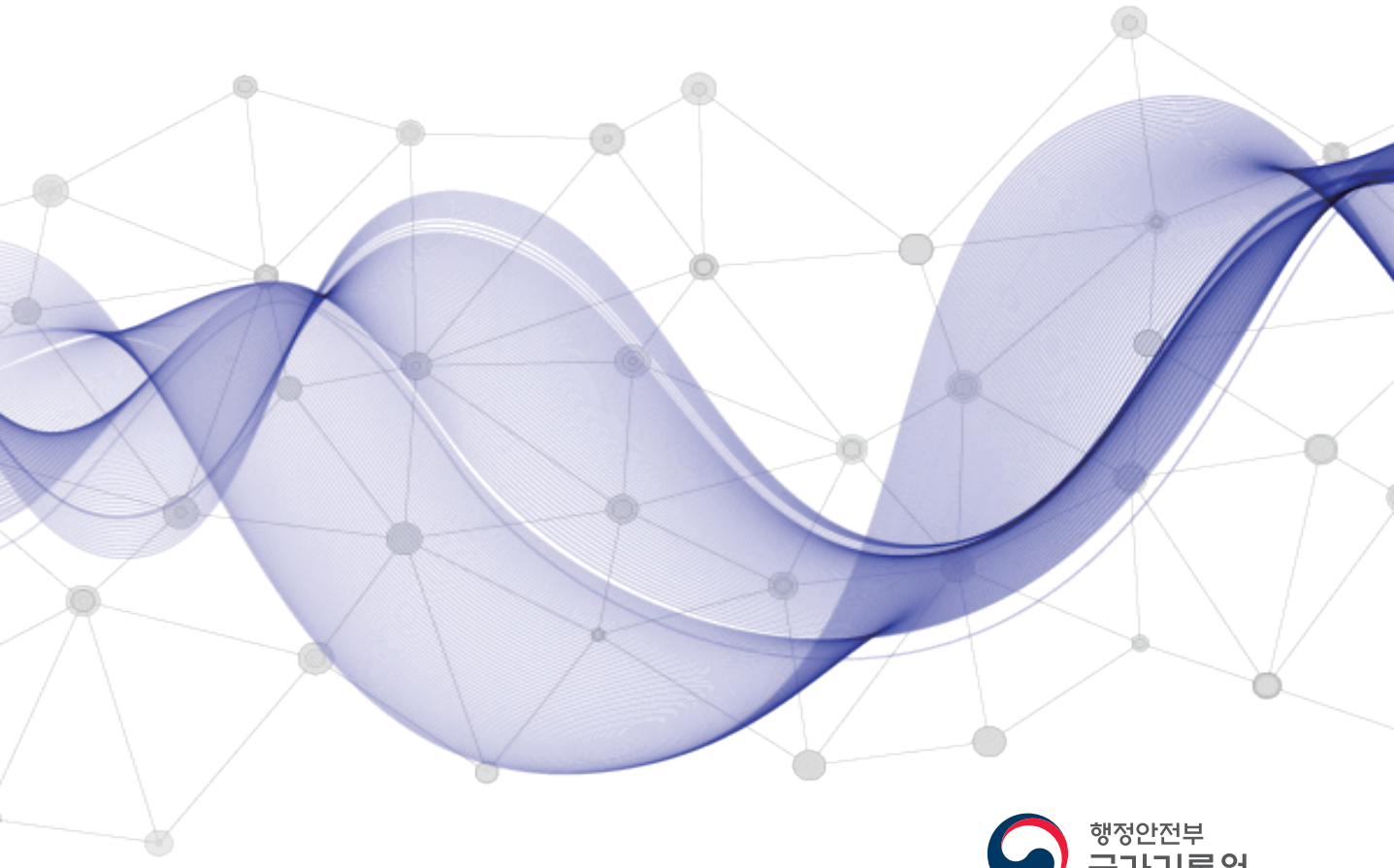
발행처 : 국가기록원 www.archives.go.kr **발행인** : 이소연 **발행일** : 2020. 5. 13.

기록관리 이슈페이퍼 vol.19

2020

디지털화 기록의 문자인식

– OCR 적용 사례 및 테스트 결과를 중심으로 –



행정안전부
국가기록원

디지털화 기록의 문자인식

– OCR 적용 사례 및 테스트 결과를 중심으로 –

박지혜 공업연구관(기록보존서비스부 복원관리과)

목 차

Ⅰ. 들어가며
Ⅱ. 디지털화 기록의 OCR 적용 사례
Ⅲ. 소장기록 OCR 솔루션 적용 테스트 결과
Ⅳ. 소장기록 OCR 인식성능 개선방안 연구
Ⅴ. 맺음말

요 약

국가기록원은 기록에 대한 접근을 개선하고, 기록의 내용에서 유의미한 정보를 추출하여 재분류, 개인정보 식별 등 업무에 활용할 수 있도록 디지털화 기록의 문자인식을 추진하고 있다. 본 글은 영국, 독일, 미국 등 다양한 해외 아카이브, 국내 도서관의 OCR 적용 사례와 국가기록원에서 소장하고 있는 실제 디지털화 기록에 상용 OCR 솔루션을 적용하였을 때 인식률이 어느 정도 수준인지 알기 위해 추진한 문자인식 테스트 결과를 정리한 내용이다.

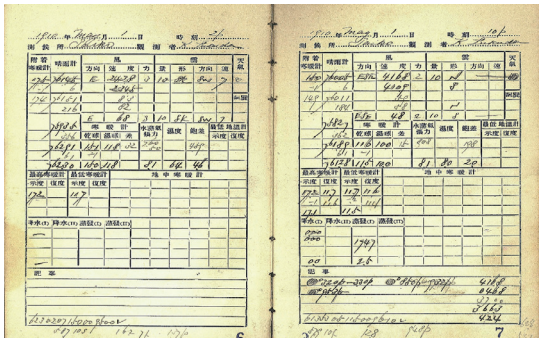
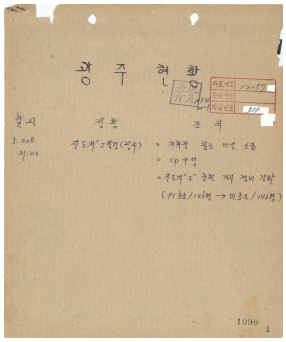
국내·외 아카이브와 도서관의 OCR 적용 사례 조사를 통해 OCR 플랫폼 등 적용 기술, 적용 대상 기록의 선정시 고려사항, 학습데이터량에 따른 인식률 개선 효과 등 다양한 시사점을 확인할 수 있었다. 문자인식 테스트 결과를 통해서는 기록의 작성 방식(출력, 타자, 수기), 언어(한글, 영어, 한자, 일본어), 유형(문서, 카드, 대장) 등 효과적인 OCR 기술의 적용을 위해 고려해야할 기록의 특성이 매우 다양하며, 자동화가 가능한 기록과 그렇지 않은 기록에 대한 접근 방법을 달리할 필요가 있다는 점을 확인하였다.

이러한 결과를 바탕으로 2020년부터 디지털화 사업 대상 중 일부의 자동 인식이 가능한 기록에 상용 OCR 솔루션 적용을 추진하고 있으며, 상대적 인식률이 저조한 타자 기록을 중심으로 OCR 인식 성능을 개선하는 연구를 병행하고 있다.

디지털화 기록의 문자인식과 관련한 노력은 이제 시작 단계로, 텍스트적 요소(수기 기록 또는 타자 기록), 비텍스트적 요소(마크, 체크박스, 구분기호), 레이아웃(페이지 구조, 양식, 표) 등에 대한 인식 연구 및 실무 적용을 단계적으로 진행해야 실효성 있는 성과를 거둘 수 있다. 앞으로 기존 OCR 기술의 성능 개선, 타 언어 대비 부족한 한글 학습 데이터세트 확보, OCR 데이터의 검색시스템 연계 등 단기적인 성과와 함께 중장기적으로 디지털화 기록의 검색·활용 개선 및 기록관리 업무 자동화의 기반 마련 등의 성과를 거둘 수 있도록 노력할 예정이다.

Ⅰ. 들어가며

국가기록원은「공공기록물 관리에 관한 법률」제6조(기록물의 전자적 생산 · 관리), 제21조(중요 기록물의 이중보존) 및 동법 시행령 제49조(영구기록물관리기관의 보존매체 수록)에 근거하여 기록의 디지털화를 추진하고 있다. 2019년에는 매체수록 심의회를 통해 선정된 5.18 민주화운동 관련 기록, 평창동계올림픽 기록, 과거사위원회 기록, '50~60년대에 생산된 법무부 수용자신분장 등 1.4만여권을 디지털화하여 중앙기록관리시스템에 등록하였고, 단계적인 원문서비스 검토를 통해 국가기록원 홈페이지에 공개하고 있다.

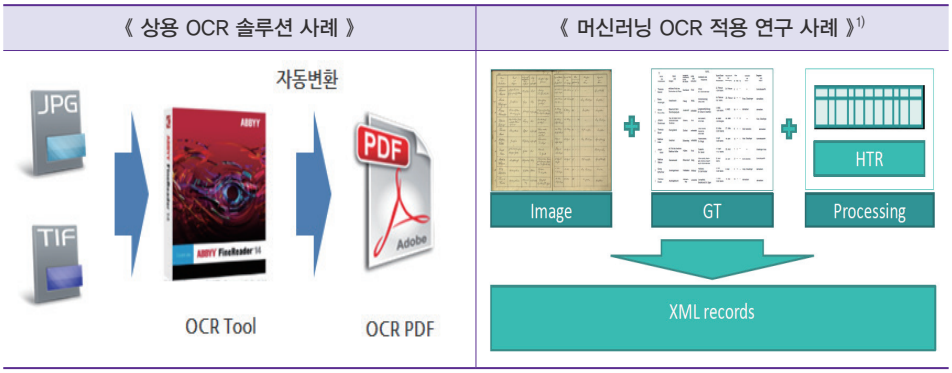
《 1910년 대구 측후소 관측야장 》	《 1980년 5.18 민주화운동 관련 기록 》
	

〈그림 1〉 2019년 종이 기록의 디지털화 사례

현재, 종이 기록의 디지털화 파일은 TIFF, JPG 등 이미지 형식으로 저장 · 관리되고 있으며, 기록의 열람시 원문의 텍스트 검색이 어렵고, 공개 재분류, 개인정보 식별, 마스킹 등 기록관리 업무를 사람이 일일이 수작업으로 해야 하는 한계가 있다. 이미지 파일에서 텍스트를 추출하는 광학문자인식(Optical Character Recognition, 이하 OCR) 기술이 있으나, 다양한 방식으로 생산된 소장 기록에 적용했을 때 효과성을 담보할 수 있느냐의 문제로 본격적인 검토가 이루어지지는 않았다.

OCR이란 사람이 쓰거나 기계로 인쇄한 문자의 이미지를 인식하여 기계가 읽을 수 있는 문자로 변환하는 패턴인식 분야의 기술로, 크게 기계로 작성된 글자를 인식하는 인쇄체 인식 방법과 필기체로 작성된 글자를 인식하는 필기체 인식 방법으로 구분된다. 인쇄체는 글자가 정형화되어 있어 OCR 인식률이 높으며 많은 기업에서 상용화된 OCR 솔루션을 출시하고 있다. 필기체는 사람마다 다양한 필체가 존재하고 문서의 형태가 다양하여 인식이 쉽지 않아 머신러닝 적용 탐색을 위한 연구가 진행 중이다

(이규철, 2017). OCR은 비교적 최근 기술은 아니며, 우편, 의료, 금융 분야 등에서 문자의 자동인식을 위하여 다양하게 활용되어온 기술이다. 다만, 기록관리 분야에서는 인식률 문제로 활발히 적용되어오지 못하고 있으나, 최근 해외 아카이브에서 기록의 텍스트 추출과 검색 연계 등에 OCR을 접목하기 위한 노력이 시작되고 있는 추세를 확인한 바 있다(국가기록원, 2019).



〈그림 2〉 상용 OCR 솔루션 및 머신러닝 OCR 적용 연구 사례

국가기록원은 공공기록관리 혁신과제의 일환으로 매체수록 전략 수립을 추진하면서, 기록의 디지털화 확대 및 활용 강화를 세부과제로 설정하고 디지털화 기록의 문자인식을 위한 OCR 기술 적용을 검토하고 있다(국가기록원, 2019).

이번 글에서는 국내 · 외 아카이브와 라이브러리의 OCR 기술 적용 사례와 함께 2019년에 국가기록원이 디지털화 기록에 OCR 기술을 본격적으로 적용하기 전 사전 검토가 필요한 부분을 확인하기 위해서 수행한 테스트 결과를 중심으로 디지털화 기록의 문자인식에 대해 말하고자 한다.

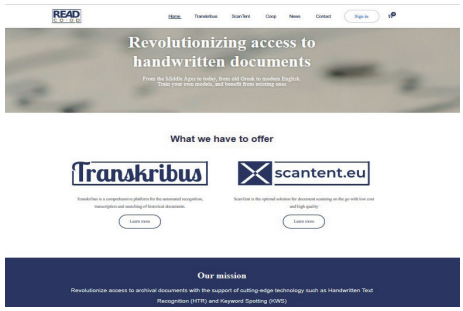
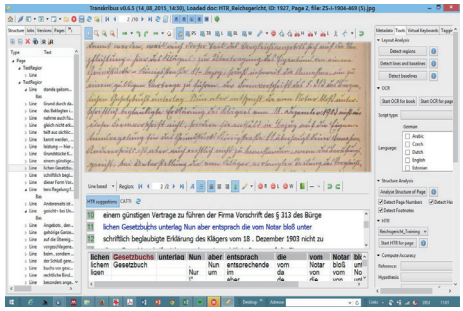
II. 디지털화 기록의 OCR 적용 사례

1. 유럽 연합 READ 프로젝트[2)]

유럽 연합에서는 최신 기술을 사용하여 보존기록에의 접근성을 높이기 위한 영구 보존문서 인식 및 강화(Recognition and Enrichment Archival Documents,

1) https://readcoop.eu/wp-content/upload/2018/11/LANG-PASSAU.pdf
2) https://readcoop.eu/

이하 READ) 프로젝트를 수행하고 있다. 이 프로젝트의 핵심적인 목표는 기록의 자동 인식, 전사 및 검색을 위한 서비스 플랫폼을 제공하는 것이다. 프로젝트를 통해 필기 기록의 텍스트 인식(Handwriting Text Recognition, 이하 HTR), 레이아웃 분석(layout analysis), 키워드 스폿팅(Keyword Spotting, 이하 KWS) 분야에서 성과를 내고 있으며, Transkribus라는 필기 기록의 문자인식 플랫폼을 개발하여 아카이브, 도서관, 인문학자들이 기술을 활용하도록 하였다. 프로젝트의 주요 성과는 필기 기록의 텍스트 인식오류율 저하, 레이아웃 분석의 개선, 대규모 학습 데이터 생성 및 머신러닝 방법을 활용한 기술의 향상 등이다. 영국, 독일 등의 아카이브에서 필기 기록의 텍스트 추출을 위한 프로젝트를 수행하면서 Transkribus 플랫폼을 적용한 사례가 최근 공개되고 있다.

《 READ 프로젝트 홈페이지 》[3]	《 Transkribus 화면 》[4]
	

〈그림 3〉 READ 프로젝트 및 문자인식 플랫폼(Transkribus) 화면

2. 영국 국가기록원(TNA) 필기체 인식 파일럿 프로젝트[5]

TNA는 필기체의 OCR 적용 가능성 검토를 위하여 앞서 언급한 Transkribus 플랫폼을 적용하여 필기체 텍스트 인식(HTR) 파일럿 프로젝트를 진행하고 있다. HTR 프로젝트에서는 유언장(PROB 11)을 대상으로 하고 있는데 법원의 서기가 작성한 유언장 사본이 포함되어 있어 필기체가 균일하며, 합법적인 문서이고, 구조화된 언어 패턴이 있으며, 사람, 장소, 상품, 사회 및 경제 네트워크와 시간과 공간에 걸친 기타 요소에 대한 세부 정보가 포함되어 있는 문서이기 때문이다.

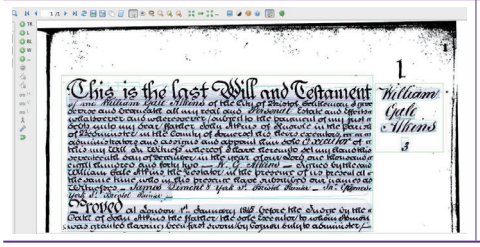
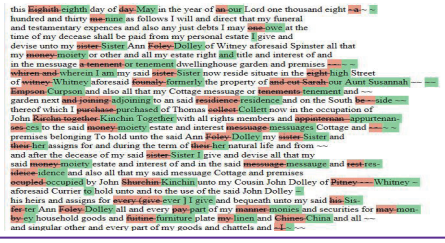
3) https://readcoop.eu/

4) http://cordis.europa.eu/project/id/674943

5) https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/

필기체 인식을 위한 첫 번째 단계는 기록의 이미지 파일을 플랫폼에 업로드하고 텍스트 영역과 기준선을 정의하는 분할 작업이다. 이는 소프트웨어가 텍스트를 찾을 위치를 확인하는 것으로 대부분 자동화되어 있지만, 결과를 확인하고 수정해야 하는 경우도 있다. 분할 작업이 완료되면 실측 자료(Ground Truth, 이하 GT)를 업로드하거나, 모델이 있다면 HTR 소프트웨어를 실행하여 자동 전사를 생성하는 과정으로 진행된다.

TNA의 프로젝트에서 얻을 수 있는 시사점 중의 하나는 OCR 인식률 측정방식이다. OCR 및 HTR 전사의 정확도는 단어 에러율(Word Error Rate, 이하 WER)과 문자 에러율(Character Error Rate, 이하 CER)로 측정하는데, 15,000 단어 학습시 WER 39%, CER 21%, 37,000 단어 학습시 WER 28%, CER 14%로 정확도가 향상된다. 이러한 데이터에서 살펴볼 수 있는 바와 같이 학습 데이터량이 많아질수록 문자 인식의 정확도가 향상될 수 있기 때문에, TNA는 온라인 자원봉사자 커뮤니티와 함께 60,000 단어의 추가 전사를 축적하고 테스트하는 작업을 계속하며 프로젝트를 진행하고 있다.

《 텍스트 영역 및 기준선 추가 》	《 필기체 인식 결과 》
	

〈그림 4〉 TNA의 필기체 인식 파일럿 프로젝트 사례[6]

3. 독일 법원기록 디지털화 및 필기체 인식 프로젝트[7]

독일에서는 Greifswald 대학, Wismar 아카이브, READ 프로젝트팀 등이 협력하여 과학기금을 지원받아 법원기록의 디지털화와 필기체 인식을 추진하고 있다. 여기에서도 문자인식 플랫폼으로 READ 프로젝트의 Transkribus를 활용하고 있다.

법원기록을 프로젝트의 대상으로 정한 이유는, 형사법의 역사와 범죄 연구의 기본이 되는 기록으로, 일상생활, 사고 및 성별 연구 등과 관련이 있으며, 특히 법적

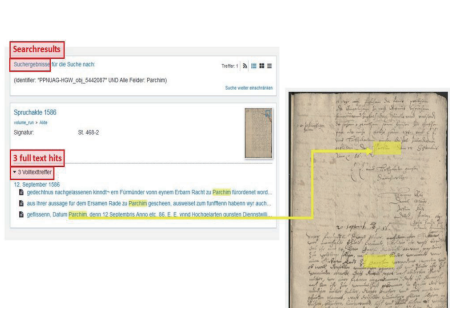
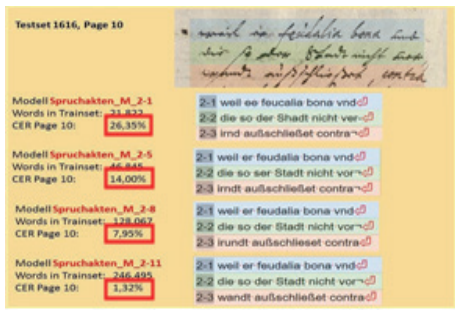
6) https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/

7) https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/

실무에 중점을 두는 기록이기 때문이다. 프로젝트를 통해 1580~1800년에 생산된 Greifswald 대학 법학부의 법적 지시사항 102,000페이지, 1746~1845년에 생산된 Wismar 재판소 판사 의견 130,000페이지, 1701~1879년에 생산된 Wismar 협의회 법원 판사 의견 25,000페이지의 3가지 법원기록을 디지털화하고 HTR과 KWS를 사용하여 기록의 이미지 파일에서 텍스트를 검색하도록 한다.

프로젝트의 워크플로우는 파일 및 프로세스 준비, 스캐닝, 문서 구조 및 메타데이터 인식 강화, Transkribus용 파일 제공, Transkribus를 사용한 필기 텍스트 인식(HTR) 등이다.

이 프로젝트에서는 필기체 인식을 위한 학습데이터 구성과 관련한 시사점을 확인할 수 있었다. 필기체 인식의 핵심요소는 GT로 설명하고 있으며, 모델이 없는 경우 1~2페이지에 1시간이 소요되며, 모델이 개선될수록 단축되어 오류율 8% 미만 모델에서는 시간당 약 6페이지의 GT를 생성한다는 데이터를 제시한다. 테스트 결과로 GT의 양을 두 배로 늘리면 모델의 오류율이 절반으로 줄어드는 결과를 얻었고, 적어도 5만 단어의 GT로 모델을 훈련해야 하며, 10만 단어의 훈련을 통해 우수한 HTR 모델을 만들 수 있을 것으로 예측하고 있다.

《 디지털 이미지에 검색어 표기 》	《 OCR 적용 인식률 》
	

〈그림 5〉 독일의 법원기록 디지털화 및 필기체 인식 프로젝트 사례[8]

4. 미국 국가기록관리청(NARA) 카탈로그 OCR 적용 사례[9]

NARA는 2019년 9월 뉴스레터를 통해 「내셔널 아카이브즈 카탈로그」에 OCR 검색 기능을 추가한다고 발표했다. 지금까지 카탈로그에서는 제목 및 디스크립션 등 일부

8) https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/category/transkribus-in-practice/ground-truth/

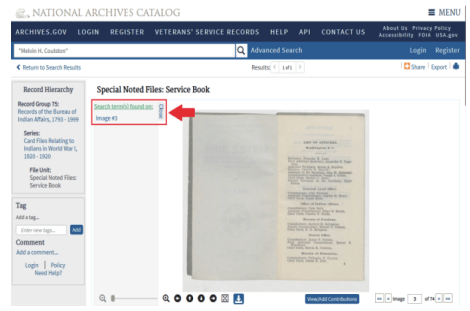
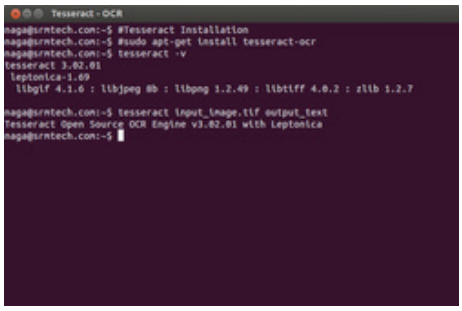
9) https://narations.blogs.archives.gov/2019/09/09/new-search-feature-optical-character-recognition-ocr/

메타데이터를 활용하여 검색이 가능하였으나, OCR 기능의 적용을 통해 수백만 페이지에서 검색을 개선하고 이미지의 일부 텍스트를 찾을 수 있게 한다. 카탈로그에 검색어를 입력하면 OCR 데이터와 기록이 보여지고, 검색어는 굵게 표시된다. 파란색 제목을 클릭하면 검색어가 있는 페이지를 볼 수 있고, 목록의 페이지 또는 강조 표시된 페이지 축소판을 클릭하면 해당 페이지로 이동한다.

NARA는 OCR 엔진으로 구글에서 개발을 지원하고 있는 오픈소스 소프트웨어인 Tesseract를 사용하고 있다. 이는, 윈도우, 리눅스 등의 OS에서 활용이 가능하며, 영어, 프랑스어, 중국어, 일본어 및 한국어 등 100개 이상의 언어를 지원한다는 장점이 있으나, 한글 인식에서는 높은 정확도를 보여주지는 못한다. 그 이유는 영문의 경우 총 26개의 알파벳으로 문장을 구성하지만, 한글은 총 40개의 자모음이 있어 약 11,000가지가 넘는 조합 가능성이 있고 이를 가지고 문장을 구성하는 특성으로 인해 한글 OCR 인식에는 어려움이 있기 때문이다(이승훈, 2017).

NARA의 Tesseract 활용 결과를 살펴볼 때 영문 인식에 있어서도 현재까지는 기록의 검색에 도움이 되긴 하지만 인식이 완벽하지 않고 여전히 수작업으로 하는 전사가 OCR 보다 정확한 상황이라 NARA는 설명한다.

해당 프로젝트에서 확인할 수 있었던 시사점 중 하나는 그간 스캐닝된 이미지 파일의 처리이다. NARA는 현재까지는 2019년 6월 이후 추가되는 JPG 또는 PDF 형식의 기록에 적용하고 있지만, 이전에 디지털화된 기록에 소급하여 OCR을 적용하기 위한 방법도 모색 중이라 말한다. 향후 이에 대한 방법론 등의 고민이 이어져야 할 부분이다.

《 디지털 이미지의 텍스트 검색 》[9]	《 Tesseract 화면 》[10]
	

〈그림 6〉 미국 NARA OCR 카탈로그 적용 사례

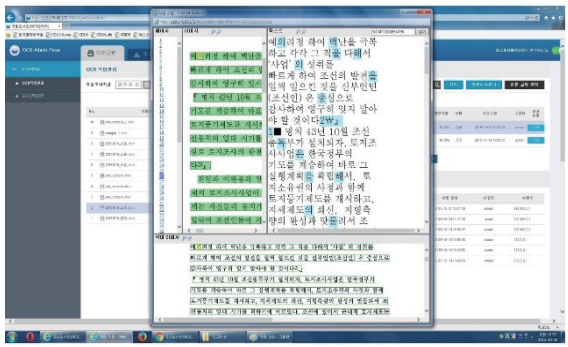
10) https://narations.blogs.archives.gov/2019/09/09/new-search-feature-optical-character-recognition-ocr/

11) https://en.wikipedia.org/wiki/Tesseract_(software)?utm_source=newsletter&utm_medium=email&utm_campaign=ocr-sept2019

5. 국내 도서관 OCR 적용 사례 및 네이버, 카카오 등 기술 사례

국립중앙도서관은 1995년부터 전자도서관 구축의 일환으로 디지털화에 착수하였으며, 디지털화 전략연구, 중장기계획 수립 및 관련 예산 확보를 거쳐, 2018년 본 예산 기준 100억원의 예산으로 디지털화 사업을 추진하고 있다. 디지털화 절차 내에 상용 OCR 솔루션을 적용한 텍스트 추출을 수행하고 있으며(서혜란, 2018), 일부 고서를 제외하고 인식률이 약 80% 수준이라 알려져 있다.

국회도서관은 1998년 전자도서관 구축사업을 통해 디지털화를 시작하였으며, 2018년 기준 일반도서, 고서, 석·박사 학위논문, 학술지, 공공정책정보 등 481만 건의 디지털화를 추진하였고 디지털화 예산은 20억원 규모이다. 국회도서관은 대량의 자료를 전자도서관을 통해 제공하고 있으며, 텍스트 및 음성 서비스 구현을 위해 OCR 처리, XML 파일 생성, 오류 수정, 로그관리, 통계관리 기능 등으로 구성된 웹 기반 OCR 시스템을 구축하였다. 이는 OCR 서버로 XML 파일을 생성하고 웹 환경에서 오탈자를 검증한 후 국회도서관의 DB에 업로드하는 절차로 진행된다.

《 국회도서관 OCR XML 파일 》	《 국회도서관 OCR 시스템 화면 》
	

〈그림 7〉 국회도서관 OCR 적용 사례[12)]

네이버는 번역, 명함, 영수증, 물류 송장의 문자 인식 및 업무 자동화 분야의 도입 확대 등을 추진하며 OCR 인식모델을 개발하고 있다. 네이버의 OCR 기술은 OCR 분야의 국제 문서 분석·인식 학회(International Conference on Document Analysis and Recognition, 이하 ICDAR)의 4개 분야에서 1등을 하는 등 사물에 대한 정보를 얻는 인식 분야 테스트에서 높은 수준을 보이고 있다.[13)]

12) http://www.retia.co.kr/cnt/info/info_case_read.html?uid=62

13) https://clova.ai/m/techdemo?lang=ko

카카오는 카카오 드라이버에서의 자동차 번호판 인식, 카카오뱅크에서 신분증 글자의 추출 등 다양한 카카오서비스를 위하여 핵심기술 중 하나로 이미지내 글자의 자동인식 기술을 개발 중으로 최근 OCR 베타 서비스를 시작하고 있다. 이는, 이제 시작하는 단계로 앞으로 지속적으로 성능을 개선하고 지원하는 언어의 수를 늘려가 서비스 품질을 개선해나갈 필요가 있다고 언급하고 있다.[14)]

6. 국제 문서 분석 및 인식 학회(ICDAR)[15)]

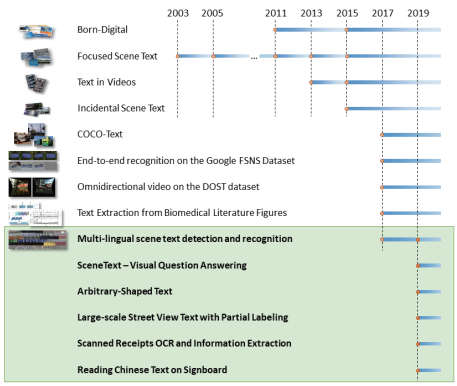
OCR 분야의 기술수준 탐색을 위하여 해당 분야 국제 학술대회의 정보를 확인하였다. ICDAR은 국제 패턴인식협회(IAPR)에서 시작한 국제 경진대회로 OCR 분야의 경진대회 중 권위가 높은 학회로 알려져 있다. ICDAR에서는 다양한 챌린지가 진행되는데, 기록의 문자인식과 관련 있는 내용은 다음과 같다.

ICDAR 2019에서는 기준선 탐지 대회(Competition on Baseline Detection, 이하 cBAD)가 있었다. 대회의 목표는 7개의 유럽 아카이브에서 수집된 문서 이미지 데이터세트 3,021건에서 기준선을 자동으로 찾는 것이다. ICDAR 2020에서는 문서의 시각적 질문 응답 대회(Document Visual Question Answering, 이하 DoV/QA)가 있다. 여기에서는 문서 이미지에 대한 정보를 시각적으로 이해하고, 텍스트 요소(손으로 작성한 내용 또는 타이핑한 내용), 비텍스트 요소(마크, 체크 표시 상자, 구분 기호, 도표), 레이아웃(페이지 구조, 양식, 표) 및 스타일(글꼴, 색상, 강조 표시) 정보 등을 이해하고 질문에 응답하는 과제이다. 〈그림 8〉의 ICDAR DoV/QA 예제를 살펴보면, 영수증과 관련하여 손으로 작성된 내용과 표 형식 등의 정보를 인식하고 지불 총액과 송장 번호 등 질문에 대한 답을 인식된 기록정보를 이용하여 응답하는 과제이다.

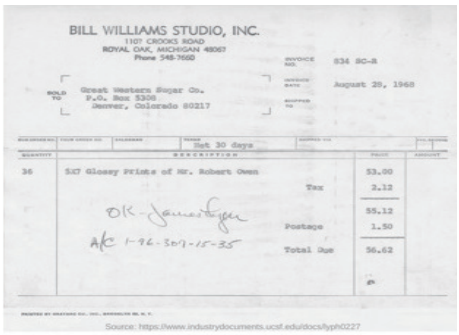
14) https://brunch.co.kr/@kakao-it/318

15) https://rrc.cvc.uab.es/

《 ICDAR 2019 과제 》



《 ICDAR DoV/QA의 예제 》



지불해야 할 총액은 얼마인가?56.62
송장 번호는 무엇입니까?834 SC-R

〈그림 8〉 국제 문서 분석 · 인식 학회(ICDAR) 과제 및 예제[16]

Ⅲ. 소장기록 OCR 솔루션 적용 테스트 결과

국가기록원은 2019년 디지털화 사업을 통해 OCR 솔루션을 적용하기 전 사전 단계로 일부 테스트를 추진하였다. 10개 기관의 디지털화 대상 기록 25권(62면)에 대해 상용 OCR 솔루션을 적용하고, 출력, 타자, 수기 등 기록 방식과 한글, 영어, 한자 등 기록 언어 및 스캔 품질, 솔루션 종류 등 다양한 인자가 OCR 인식률에 미치는 영향을 살펴보았다. 이번 테스트에서 OCR 인식률은 원문의 총 글자 수와 OCR 결과로 인식된 글자 수를 비교하는 방법으로 측정하였으나, 앞으로 인식된 문자의 검색·활용을 고려하여 문자 에러율(WER)을 검토할 예정이다.

〈표 1〉 OCR 관련 테스트 항목

구분		주요내용
검토 항목별 테스트	기록 방식	출력, 타자, 수기
	기록 언어	한글, 영어, 한자, 일본어
	기록 유형	문서, 카드, 대장
	스캔 품질	해상도(400dpi, 200dpi), 흐린 문서, 기울어져 스캔된 문서, 다언어 혼용 등
	솔루션 종류	국외 A사, 국외 B사, 국내 C사
생산기관별 OCR 테스트	인식률	테스트 페이지 OCR 인식률
	적용가능성	적용 적합, 부적합, 선별적 적용 가능

16) https://rrc.cvc.uab.es/?ch=17

1. 검토 항목별 인식률 테스트 결과

1) 기록 방식

출력된 기록의 OCR 인식률은 80~90% 이상으로 확인되었다. 그러나 타자와 수기 기록의 경우에는 상용 솔루션 적용시 인식이 어려운 것으로 확인되었는데 이 중 타자 기록은 한글의 인식률을 기준으로 한 결과로 영문 타자의 경우 90% 이상의 사례도 확인되었다. 이 결과를 통해 한글 타자 기록의 인식성능 개선 연구를 우선 고려할 필요가 있음이 확인되었다.

〈표 2〉 기록 방식별 OCR 인식률 테스트 결과

구분	출력	타자	수기
인식률 (문서+한글 기준)	90% 이상	40~60%	인식불가

2) 기록 언어

같은 조건에서 한글, 영어, 한자 등 기록 언어별 인식률의 차이는 크지 않으며, 인쇄된 경우 모든 언어에서 80% 이상의 인식률을 확인할 수 있었다. 다만, 여러 언어가 혼재된 경우 인식률이 저하되는 경향도 확인되었다.

〈표 3〉 기록 언어별 OCR 인식률 테스트 결과

구분	한글	영어	한자	일어
인식률 (문서+출력 기준)	90% 이상	90% 이상	90% 이상	80% 이상

3) 기록 유형

출력 문서는 80% 이상의 인식률이 확인되었고, 카드, 대장의 경우에는 표 형식에 여러 언어가 사용되고 수기로 작성된 경우가 많아 인식률이 저조함을 확인하였다.

〈표 4〉 기록 유형별 OCR 인식률 테스트 결과

구분	카드	대장
인식률 (한글+수기 기준)	인식저조	인식저조

4) 스캔 품질

해상도의 차이에 따른 인식률을 비교한 결과, 관련한 유의미한 차이는 확인하지 못했다. 이는 국가기록원이 보유하고 있는 디지털화 파일의 해상도가 문자 인식에 영향을 미칠 정도가 아니라고 일차적으로 해석될 수 있다. 다만, 오래되고 상태가 좋지 않은 기록이 많아 변색, 파손, 오염 등 기록의 상태나, 기울어짐, 노이즈 등 복사 상태의 영향이 있을 것으로 보이며, 이에 대한 세부적인 검토가 필요하다.

〈표 5〉 스캔 품질별 OCR 인식률 테스트 결과

구분	400dpi	200dpi	흐린문서	기울어진스캔
인식률 (문서+한글+출력 기준)	97%	97%	80%	82%

5) 솔루션 종류

문서, 한글, 출력 기준의 테스트 결과 도서관 등의 적용사례가 있는 국외 A사 솔루션의 인식률이 상대적으로 높음을 확인하였다.

〈표 6〉 솔루션 종류별 OCR 인식률 테스트 결과

구분	국외 A사	국외 B사	국내 C사
인식률 (문서+한글+출력 기준)	83%	69%	74%

2. 생산기관별 OCR 테스트 결과

국내 도서관의 사례에서와 같이 디지털화 프로세스에 상용 OCR 솔루션을 적용할 수 있는지 가능성을 확인하기 위하여, 실제 디지털화 기록의 이미지 파일에 OCR 솔루션을 적용하는 테스트를 진행하였다. 검토 항목별 OCR 테스트 결과와 동일하게 인식률은 주로 출력, 타자, 수기 등 기록 방식, 스캔 품질 등에 영향 받음을 알 수 있었으며, 적용이 가능한 유형의 기록과 그렇지 않은 기록의 일부를 확인할 수 있었다.

〈표 7〉 생산기관별 OCR 테스트 결과 및 적용 특징

생산 기관	생산 년도	기록 특징	OCR 테스트 결과			적용 가능성
			전체 문자	인식 문자	인식률	
A	'65~'76	문서+한글+타자	153	83	50% 수준	적용 부적합
		문서+한글 · 한자+수기	32	0	인식 불가	

생산 기관	생산 년도	기록 특징	OCR 테스트 결과			적용 가능성
			전체 문자	인식 문자	인식률	
B	'92~'02	문서+한글+타자+복사	487	364	70% 수준	선별적 적용 검토
		문서+한글+출력	249	146	50% 수준	
C	'76	문서+한글+타자	391	13	5% 이하	적용 부적합
		문서+한글+타자+복사	458	52	10% 수준	
D	'78~'92	문서+한글+수기	225	0	인식불가	적용 부적합
		허가증+한글+수기	245	38	15% 수준	
E	'79	문서+한글+타자	106	18	20% 수준	적용 부적합
		문서+한글+수기 · 등사+복사	432	41	10% 수준	
F	'10~'11	문서+한글+출판물	1,400	1,398	90% 이상	선별적 적용 검토
		문서+한글+출력+복사	279	64	20% 수준	
G	'06~'10	문서+한글+출력	436	434	90% 이상	선별적 적용 검토
		문서+한글+출력	595	65	10% 수준	
H	'46~'08	문서+한글+출력	563	542	90% 이상	선별적 적용 검토
		대장+한글+수기	197	9	5% 수준	
I	'53~'00	카드+한자+수기 · 등사	248	137	50% 수준	적용 부적합
		문서+한글+출력+복사	345	99	30% 수준	
J	'90	문서+한글+타자	386	212	50% 수준	선별적 적용 검토
		문서+한글+출력	299	273	90% 이상	

국가기록원에서 소장하고 있는 기록의 특성은 매우 다양하기 때문에 같은 기록물철 안에도 여러 기록 방식과 언어가 혼재되어 있는 경우가 많아 이를 고려하지 않을 경우 문자인식의 효과성을 담보할 수 없다. 테스트 결과를 통해 효과적인 문자인식을 위해서는 기록의 특성을 고려할 필요가 있다는 결론을 얻었기 때문에, 2020년 OCR 적용시 이를 고려하게 되었다. 2020년 디지털화 사업에서 전체 디지털화 대상의 10%에 OCR 상용 솔루션을 적용할 예정으로, 사업 대상 기록을 디지털화 특성별로 구분한 주요 유형 중 유형-1에 해당하는 기록에 문자인식 적용을 추진할 예정이다.

〈표 8〉 기록 특성별 디지털화 주요 유형

구분	유형	주요 내용	OCR 적용
유형-1	문서	규격+재질(백상지, 중질지)+자동스캔+인쇄	○
유형-2	문서	규격+재질(미농지, 갱지)+수동스캔+타자/필기	×

구분	유형	주요 내용	OCR 적용
유형–3	문서	비규격+특수지(감열/감광/감압)+수동스캔	×
유형–4	간행물	출판물+수동스캔+인쇄	×
유형–5	간행물	제본물+수동스캔+타자/필기	×

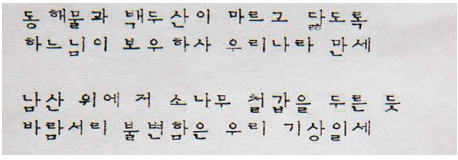
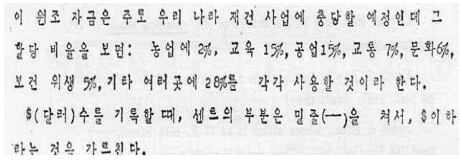
Ⅳ. 소장기록 OCR 인식성능 개선 연구

앞서 살펴본 바와 같이, 수기나 타자로 작성된 기록의 경우 상용 OCR 솔루션을 직접 적용하기 어려운 상황으로 문자인식에 대한 효과성을 높이기 위해서는 인식이 어려운 기록 유형에 대한 인식성능 개선이 필수적이다. 해당 부분은 2020년 국가기록관리 · 활용기술 연구개발사업(R&D)의 연구과제 중 하나로 추진하고 있으며 주요 내용은 아래 표와 같다.

〈표 9〉 OCR 인식성능 개선 연구 주요내용

구분	주요내용
연구과제명	• 소장기록물 특성을 고려한 OCR 성능 개선방안 연구
기간/예산/주관기관	• 2020. 4. ~ 11. (8개월) / 135백 / ㈜로민
연구목표	• 타자 인식 평가용 데이터셋의 문자탐지 및 기술수준 달성 • 이미지 형태 데이터의 OCR 적용 및 디지털화 프로세스 연구 • 문자인식 성능 평가 테스트베드 구축 및 OCR 기술수준 검증 • 문자인식 딥러닝 모델 학습 및 평가를 위한 데이터셋 구축
주요 연구내용	• 소장기록물 OCR 적용 관련 활용 및 관리방안 마련 – OCR 파일 관리방안 도출 및 검색기능(CAMS) 연계 방안 – CAMS 파일 OCR 적용 방안 ⇨ **디지털화 프로세스 개선방안 도출** • 소장기록물 특징을 고려한 OCR 연구 및 기술수준 검증 – 문자인식 알고리즘 기술 동향 탐색 – 소장기록물 특성을 고려한 OCR 플랫폼 성능 비교 분석 ⇨ **OCR 인식 성능 목표 달성** • 문자인식 성능 개선을 위한 학습데이터 추출 · 적용 방안 연구 – 학습데이터 관련 연구개발 동향 조사 및 프레임워크 설계 – 테스트베드 및 실측 데이터 생성 – 타자기록물 성능 개선방안 제시 및 정보서비스 방안 제시 ⇨ **타자기록물 데이터셋 구축 및 OCR 테스트베드 구축**
기대효과	• 기존 OCR 성능 부족으로 적용되지 못한 OCR 적용 기술 확보 • 타언어 대비 부족한 한글데이터셋 확보 • OCR 데이터의 검색시스템 연계로 접근성 및 활용성 개선

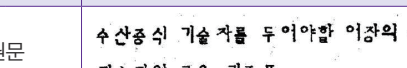
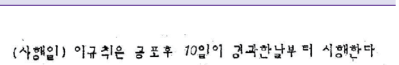
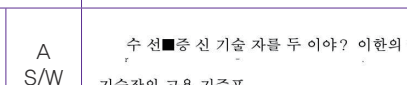
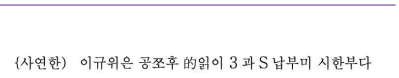
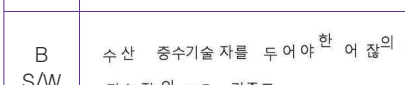
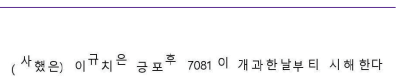
출력 기록이나 영문 타자 기록과는 달리 한글로 작성된 타자 기록은 초성, 중성, 종성 간의 간격이 불규칙하고 띄어쓰기나 자간이 컴퓨터 폰트와는 다르며 타자기의 종류 등에 따라 글자체가 다양하다는 특징 때문에 상대적 인식률이 저조하였다. 올해 연구는 타자 기록을 중심으로 진행할 예정으로 연구를 진행하면서 수기로 작성된 기록에까지 범위를 확대해야 할지에 대한 고민도 이어져야 할 것으로 보인다.

《 세벌식 타자기 글자체 》	《 다섯벌식 타자기 글자체 》
	

〈그림 9〉 타자 기록 유형(예시)

〈표 10〉은 연구가 시작 전 사전 테스트로 진행한 결과로, 상용, 오픈소스 등 다양한 OCR 솔루션을 적용했을 때 타자 기록의 인식 결과이다. 현재 솔루션을 적용한 문자 인식 결과물은 직접 활용할 수 없는 상황임을 확인할 수 있다.

〈표 10〉 타자 기록의 OCR 솔루션 인식 성능 사전 비교

구분		기록 1	기록 2
원문			
인식 결과	A S/W		
	B S/W		
	C S/W		

문자인식을 위한 주요 연구내용은 다음과 같다. 이미지에서 문자 영역의 위치를 검출하는 Text Localization과 검출된 문자 영역에서 글자를 인식하는 Text Recognition 단계에 Mask-RCNN 기반 Two-stage Detection 모델과 Segmentation 기반 모델인 CRAFT을 개발하여 비교 실험을 진행할 예정이다. 또한 인식 성능 개선을 위한 타자

기록 10만 단어, 2,000장 규모의 데이터셋을 생성하고 데이터 라벨링을 수행할 예정이며, 이 외에 합성 이미지 데이터셋도 추가로 생성하여 구축된 테스트베드를 활용한 성능 개선 및 평가에 활용할 예정이다.

〈그림 10〉 문자 인식 프로세스

Ⅴ. 맺음말

2019년 하반기에 국가기록원은 기록에 대한 접근성을 개선하고 기록의 내용에서 추출할 수 있는 정보를 기록관리 업무에 활용하기 위하여, OCR 기술 관련 사례 조사와 기초적인 테스트를 수행하였다. 그 결과로 비용 대비 효과적인 OCR 기술의 적용을 고려하여 자동 인식이 가능한 유형과 불가능한 유형의 기록에 대한 접근 방식을 디지털화 사업과 R&D로 달리 적용하고자 한다.

자동 인식이 가능한 유형의 기록은 2020년 디지털화 사업을 통해 한글, 한자 등으로 생산된 출력 문서의 일부에 상용 OCR 솔루션을 시범 적용할 계획이다. 동일 유형의 기록에도 기록 방식, 기록 언어 등이 혼재되어 있기 때문에 자동 인식의 효과적인 적용을 위해서는 대상 기록의 특성에 대한 사전조사가 필수적으로 출력, 타자, 수기, 언어 등을 고려하여 단계적인 적용을 추진하고자 한다.

자동 인식이 불가능한 유형의 기록은 2020년 연구개발 과제로 소장 기록의 특성을 고려한 OCR 인식성능 개선방안 연구를 추진할 예정이다. 주요 연구의 내용은 OCR 적용 파일의 유형별 관리방안, 검색 기능 연계 활용 방안 마련, 소장 기록의 특징을 고려한 OCR 기술 수준의 검증, 문자인식 성능 개선을 위한 학습데이터 추출 및 적용 방안 연구 등이다.

이를 통해 디지털화 기록의 검색과 활용을 위한 문자인식 기술 기반을 구축하고, 소장 기록의 접근성 및 활용성을 높일 수 있도록 검색시스템을 연계하며, 향후, 기록의 개인정보 등 자동 태깅과 마스킹 기능의 연계 활용 등을 장기적으로 추진할 계획이다.

〈참 고 문 헌〉

국가기록원 2019. 기록물 매체수록 해외 동향 및 향후 과제.

이규철, 유지상, 「한글 음식 메뉴 인식을 위한 OCR 기반 어플리케이션 개발」, 『한국정보통신학회논문지』 21(5), 2017.

이승훈, 전진호, 홍해성, 강동혁, 박미화, 「OCR 기술을 이용한 한글 처방전 문자 인식 시스템」, 『한국정보과학회 학술발표논문집』, 2017.

서혜란, 「도서관 장서 디지털화 사업의 현황과 과제」, 『한국문헌정보학회 학술발표논문집』, 2018.

https://readcoop.eu/wp-content/upload/2018/11/LANG-PASSAU.pdf

https://readcoop.eu/

http://cordis.europa.eu/project/id/674943

https://blog.nationalarchives.gov.uk/machines-reading-the-archive-handwritten-text-recognition-software/

https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/

https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/category/transkribus-in-practice/ground-truth/

https://narations.blogs.archives.gov/2019/09/09/new-search-feature-optical-character-recognition-ocr/

https://en.wikipedia.org/wiki/Tesseract_(software)?utm_source=newsletter&utm_medium=email&utm_campaign=ocr-sept2019

http://www.retia.co.kr/cnt/info/info_case_read.html?uid=62

https://clova.ai/m/techdemo?lang=ko

https://brunch.co.kr/@kakao-it/318

https://rrc.cvc.uab.es/

https://rrc.cvc.uab.es/?ch=17

‖ 기록관리 이슈페이퍼 발간 목록 ‖

발간호	제목	작성자	발간일
vol. 1	기관 심층인터뷰를 통한 BRM 단위과제 운영 개선 방안 수립	황정원 기록연구사	2019. 10. 8.
vol. 2	「공공기록물법」 상의 기록의 개념 검토 ① 기록의 개념과 성립요건 - 정보와 증거로서의 기록의 함의를 기록물법에 적용하기 - ② 기록이란 무엇인가? -「공공기록물법」에 따른 기록관리 대상의 범위와 관련하여-	이젬마 서기관 임신영 기록연구사	2019. 10. 22.
vol. 3	대통령기록물 평가체계 개선 방안	윤정훈 행정사무관	2019. 10. 31.
vol. 4	"도전! 기록관리 명강사되기" 기록관리 강사양성제도 도입	김명옥 사서사무관	2019. 11. 15.
vol. 5	국가기록원 블록체인 기록관리 플랫폼 구축사업의 의미와 전망	왕호성 기록연구사	2019. 11. 22.
vol. 6	전자기록 장기보존정책의 방향	이지영 공업연구사	2019. 12. 5.
vol. 7	디지털기반 대통령기록관리체계 모델 재설계	김현숙 공업연구사	2019. 12. 12.
vol. 8	건축아카이브의 해외 동향 및 향후 과제 - ICAA BRAGA 2019 참관기 -	김수연 전문임기제 허인영 전문임기제	2019. 12. 13.
vol. 9	전자기록 장기보존패키지 모델 시험과 새로운 모델 제안	신동혁 공업연구사 김상국 전산사무관 나미선 학예연구관	2019. 12. 17.
vol. 10	기록물 매체수록 해외 동향 및 향후 과제	박지혜 공업연구관	2019. 12. 24.
vol. 11	기록물 생산현황 분석 결과(2018년 생산분)	김현애 기록연구사	2020. 1. 15.
vol. 12	기록물관리시스템을 통한 생산현황 통보 자동화 방안	하정하 기록연구관	2020. 1. 29.
vol. 13	기록물관리 전문요원 양성제도 현황과 전망 - 교육원 과정을 중심으로 -	성주영 기록연구사	2020. 2. 12.
vol. 14	정기실태점검을 통해서 본 기록관리 개선방안	박지태 학예연구관 송혜현 사서사무관	2020. 2. 27.
vol. 15	공공기록물법과 전자정부법과의 관계	임신영 기록연구사	2020. 3. 11.
vol. 16	해외 공공기록 평가선별제도 관련 사례 및 시사점	조영주 사서주사	2020. 3. 20.
vol. 17	속기록 의무생산회의에 대해 묻고 답하다	박이준 학예연구관 이주현 기록연구사	2020. 4. 3.
vol. 18	기록관리 현장 지원을 위한 기관방문 컨설팅의 추진	나창호 기록연구관 정경택 공업연구사	2020. 4. 21.
vol. 19	디지털화 기록의 문자인식 - OCR 적용 사례 및 테스트 결과를 중심으로 -	박지혜 공업연구관	2020. 5. 13.
vol. 20	통계로 알아보는 국가기록포털의 현재	서경란 전산주사보	2020. 5. 27. 발간예정

발간 예정 목록

- 특수지 기록물과 보존
- 비공개 기록물 공개재분류 업무절차 개선 방안
- 국가 중요기록물 보존성 향상을 위한 〈맞춤형 복원 복제 지원사업〉, 10년을 돌아보며
- 손상파일 검사 복구 프로토타입 적용 방안
- 행정안전부 기록관리기준표 개선 추진
- 비밀기록물 생산현황 통보 서식
- 국외 소재 한국 병풍 〈오륜 행실도〉 복원처리 지원을 위한 영국박물관 방문기
- 전자기록 이관에 따른 데이터 적합성 문제
- 2013~2019년 기록관리시스템 컨설팅 현황

※ 향후 이슈페이퍼의 주제 및 발간 일정은 원내 사정에 의해 일부 변경될 수 있습니다.

「기록관리 이슈페이퍼」는 기록관리 현장의 다양한 현안 논의와 기록인 여러분의 귀중한 연구성과 공유를 기다립니다.

국가기록원 연구협력과 ☎ (042) 481-6353 ✉ issuepaper@korea.kr

기록관리
이슈페이퍼 vol.19

' 신뢰받는 기록관리로 정부는 투명하게, 국민은 행복하게 '

행정안전부
국가기록원

35208 대전광역시 서구 청사로 189 정부대전청사 2동
Tel 042-481-6353 **Fax** 042-481-6234