2020년 제3차 기록관리 정책포럼

기록서비스 확대를 위한 개인정보 문제 공동 해결 방안

| **일 시** | 2020. 11. 12. (목) 14:00 ~ 17:00

| **장 소** | 행정기록관 2층 세미나실

| **주 최** | 행정안전부 국가기록원

행정안전부
국가기록원

정책포럼 진행 일정

시 간		주 요 내 용	비 고	
13:30~14:00	30’	등 록	발열체크 및 등록	
14:00~14:05	5’	개 회	사 회 자	
14:05~14:10	5’	인사말씀	이소연 국가기록원장	
14:10~14:25	15’	주제발표 1 디지털문서의 개인정보 필터링 및 마스킹 기술	임석종(KISTI)	
14:25~14:40	15’	주제발표 2 전자기록물 공개재분류를 위한 비공개정보 필터링 및 마스킹 기술	김진아(국가기록원)	
14:40~14:55	15’	주제발표 3 국립중앙도서관의 온라인 서비스와 개인정보보호	최윤경(국립중앙도서관)	
14:55~15:10	15’	주제발표 4 개인정보 가명처리 정책동향	박윤식(한국인터넷진흥원)	
15:10~15:25	15’	휴 식	장내 정리	
15:25~17:00	95’	종합토론	사회	조이형(국가기록원)
			토론	이영도(국가기록원) 오용석(한국인터넷진흥원) 김순석(한라대학교)
17:00	-	폐 회	사 회 자	

Contents

‖ 주제발표

‖ 토론요지문

주제발표 1

디지털 문서의 개인정보 필터링 및 마스킹 기술

임 석 종(KISTI)

디지털 문서의 개인정보 필터링 및 마스킹 기술

임 석 종(KISTI)

차 례

1. 개요

1) 개인정보처리의 개념

○ 개인정보처리의 정의

- 개인정보 보호법 [시행 2017. 10. 19.] [법률 제14839호, 2017. 7. 26., 타법개정] 제2조에 따른면 1. "개인정보"란 살아 있는 개인에 관한 정보로서 **성명, 주민등록번호 및 영상** 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)를 말한다. 2. "처리"란 **개인정보의 수집, 생성, 연계, 연동, 기록, 저장, 보유, 가공, 편집, 검색, 출력, 정정(訂正), 복구, 이용, 제공, 공개, 파기(破棄), 그 밖에 이와 유사한 행위**

2) 개인정보처리의 기술 동향

○ 인터넷 사용자 ID 관리 및 서비스

- 인터넷의 사용자 ID의 분산·중복 저장 및 사용자 ID의 증가와 관리의 불편함, 개인정보 유출로 인한 프라이버시 문제 발생으로 ID 통합관리시스템 개발을 통한 통합 ID 서비스 제공

○ 인적자원관리 및 개인정보보호

- 'SW 개발보안 의무화 및 정보관리범위 확대' 등으로 HRM(인적자원관리) 시스템 내 개인정보보호기술 적용

○ 빅데이터 환경의 개인정보처리

- 빅데이터에서 텍스트, DB 등 다양한 형태로 표현되는 개인정보 데이터 자동화 고속 식별, 개인정보 익명처리를 위한 비식별화 정보 생성, 빅데이터 개인정보보호 기술 응용 서비스 개발

○ 비식별개인정보의 온라인 맞춤형 서비스

- 비식별개인정보를 활용한 온라인 맞춤형 서비스를 위한 비식별개인정보의 수집 및 활용 기술

3) 개인정보처리 관련 기술 현황

○ 개인정보 탐지 및 모니터링 기술

- 네트워크 환경에서 모바일 사용자의 개인정보 접속 모니터링 및 탐지하는 기술

○ 빅데이터 분석 기술

- 빅데이터를 처리하는 데 필요한 원본 데이터의 저장과 처리 그리고 데이터 마이닝 기술을 활용한 개인정보처리 기반 기술

○ 정보손실 방지를 위한 비식별화 기술

- 변환 레코드 생성, 암호화, 삭제, 재식별 불가능화 기술

2. KISTI 개인정보처리 기술 현황

1) KISTI 개인정보처리 기술 개요

○ 디지털 문서 파일내에 포함된 개인정보의 자동 검출 및 제거 기능

- 다양한 포맷의 텍스트 전자문서(Excel, Hwp, PDF, PPT, Word, txt 등)에 포함된 개인정보를 검사하고 개인정보를 설정한 패턴(마스킹처리)대로 처리한다.
- 개인정보(전화번호, 이메일, 주소, 주민번호, 계좌번호, 카드번호, 이력정보, 인명, 등)에 대한 개인정보 검사 패턴 등록 및 관리를 통하여 업데이트한다.

○ 연구보고서의 제출 형식인 PDF 파일 뿐만 아니라 엑셀, 한글, 파워포인트, 워드, 텍스트 등 다양한 파일에 포함된 개인정보를 추출하는 기술

○ 연구보고서에 포함된 다양한 개인정보의 유형별로 패턴을 분석하여 관리하고 업데이트하는 기술

○ 전자문서에 포함된 개인정보의 민감도와 수준을 환경 설정에서 사용자 선택에 따라 조정하여 사용할 수 있는 기술

○ 추출된 개인정보의 비식별 사용을 위한 데이터 변환 및 마스킹 처리 하는 기술

2) KISTI 개인정보처리 연구개발 내용

○ 개인정보처리기 주요 기능

① 사용자 PC의 폴더 및 파일 선택기능

② excel, hwp, pdf, ppt word, text 파일의 개인정보 검증 기능 및 개인정보 치환기능

③ 검사결과 저장기능

④ 특정 개인정보 패턴의 조건에 따른 마스킹처리(*) 기능 제공

⑤ 개인정보 치환 파일명 설정

⑥ 개인정보 검사 대상 및 개인정보 치환 환경설정

○ 개인정보처리기 사용 방법

① 개인정보검사

가. 개인정보 검사 대상 선택

나. 개인정보 검사 선택

다. 선택한 문서타입(excel, hwp, pdf, ppt word, text)을 구분해 해당 텍스트 정보추출

라. 추출된 텍스트 정보를 이용해 환경설정에서 등록한 정규식 패턴을 통해 개인정보와 해당 페이지 정보 반환

마. 개인정보검출 목록 정보를 엑셀로 다운로드

② 개인정보제거

가. 검출된 문서 페이지와 검출단어 치환 패턴정보를 이용해 문서의 개인정보를 치환

나. 해당 문서를 모두 치환처리가 완료되면 화면에 치환되면 문서명, 검출항목, 삭제개수, 정보를 반환

마. 치환 목록 정보를 엑셀로 다운로드

○ 개인정보처리기 사용 환경

가. 사용 OS : Window, Linux

나. 사용 언어 : Java 1.8 이상

○ 개인정보자동검출기 사용법

〈개인정보자동검출기 메인화면〉

① 초기화면에서 환경설정을 선택하면 파일명 지정 버튼을 선택함

② 개인정보가 삭제된 파일 저장 시 저장을 원하는 파일명 입력.

③ 삭제옵션을 선택/해제함

④ 저장/취소 버튼을 클릭하여 작업을 완료함.

① 초기화면에서 환경설정을 선택하고 시스템 검사항목을 다시 선택함

② 시스템 검사항목의 대항목과 상세항목을 확인하고 검사할 항목을 선택/해제함

③ 시스템 검사항목 설정 후, 저장/취소 버튼을 클릭하여 작업을 완료함.

① 초기화면에서 환경설정을 선택하고 사용자 검사항목을 다시 선택함

② 사용자 검사항목의 대항목과 상세항목을 확인하고 검사할 항목을 선택/해제함

③ 신규 사용자 검사항목 등록을 위해 Row를 생성하거나, 선택된 검사항목의 Row를 삭제함

④ 정규표현식 교육자료를 다운로드함

⑤ 사용자 검사항목 설정 후, 저장/취소 버튼을 클릭하여 작업을 완료함.

① 개인정보검사 대상 파일 선택화면을 호출함

② 파일을 검사 대상 목록에 추가함

③ 개인정보검사 버튼을 클릭하여 검사함.

① 개인정보검사버튼을 클릭함.

② 출력된 결과에 대해 개인정보를 제거함.

③ 개인정보 검사 결과 목록을 저장함.

		검사대상 전자문서 목록		2016-11-8
순번	**파일명**	**파일사이즈**	**검출갯수**	**제거갯수**
1	의료방사선 측정표준 확립(1).hwp	18.58MB	0	0
2	20130308054713610.pdf	6.4MB	5	0
3	20130308060849765.pdf	3.37MB	4	0
4	220-2011-1-E00027.doc	9.66MB	3	0
5	220-2011-1-D00098.doc	26.29MB	3	0
6	14162예방안071.pdf	2.95MB	8	1
7	20130308055528112.pdf	2.18MB	3	0
8	10기가급 DDoS 방어 및 컨텐츠 보안	19.06MB	0	0
9	차세대 전자파 측정표준 개발.hwp	13.78MB	0	0
10	고 안정성을 가지는 초저가, 무독성	11.49MB	6	0
11	20130308060847832.pdf	18.08MB	34	0
12	동물용 패혈증 신속 진단 키트 개발	16.2MB	0	0
13	KOLAS(교정) 및 측정분석 기술지원	18.52MB	0	0
14	지속가능한 담수화 기술을 위한 차세	17.07MB	4	0
15	고성능 고온 고분자 전해질 연료전지	1.53MB	4	0
16	원자 및 나노 수준의 균일한 다공성을	8.52MB	4	0
17	나노바이오융합 및 나노재료 안전성	15.07MB	0	0

① 검사 및 제거 결과가 저장된 엑셀 파일을 선택하여 처리 내역을 확인함.

3. KISTI 개인정보처리 기술 개발 방향

□ 개인정보처리기 연구개발 현안 및 방향

○ 디지털 네트워크 환경의 정보보호에서 전자문서의 개인정보 보호기술 이슈

- 개인정보 보호시스템은 방화벽, 네트워크 보안장비와는 별도로 웹과 전자문서에서 개인정보를 검색 및 처리하는 기술로서 전자문서의 파일의 범용성과 개인정보 유형에 대한 패턴을 확장하는 기술

○ 디지털 이미지 형식의 전자문서에 포함된 개인정보의 식별 및 변환 이슈

- 개인정보가 텍스트가 아닌 이미지 형식으로 저장된 경우 개인정보의 탐지 및 제거를 위한 접근으로 OCR 기술 및 기계학습 모델 적용

4. 향후 과제(전자문서 개인정보처리 기술의 개선 및 확장)

○ 이미지 개인정보에 대한 기계학습 및 개인정보처리 모델 개발

○ 한글 문서 OCR 적용을 개인정보 텍스트 변환 및 개인정보처리

○ 데이터 전처리 및 큐레이션 기술 적용

- 과학기술정보에 포함된 텍스트 개인정보 뿐만 아니라 특수문자, 수식, 그래프에 대한 데이터 전처리와 큐레이션 기술 개발

○ 개인정보 비식별화 데이터 활용 및 기술정보의 연계

- 개인정보는 보호하되 비식별화된 개인정보를 활용할 수 있는 기술정보와 연계하여 활용될 수 있도록 개발

○ 국가 R&D 연구성과 관리기관 및 연구수행기관과의 협력

참고문헌

[국내논문]

· 「개인정보보호기술에 관한 연구(A Study on Privacy Enhancing Technologies)」

· 「HRM 시스템의 개인정보보호기술 적용방안(Applied Method of Personal Information Protection Technology in HRM System)」
강민영(가천대학교 일반대학원 모바일소프트웨어학과 UU0000008), 박석천(가천대학교 컴퓨터공학과 UU0000008), 홍석우(화이트정보통신 사업총괄 CC0001798), 한국정보처리학회 2013년도 제39회 춘계학술발표대회 2013 May 10, pp. 691-694, 2013.

[학위논문]

· 장성수, 「영상정보처리장치(CCTV)와 개인정보보호에 관한 연구」, 2013, xi, 93p., 동아대학교 법무대학원, 국내석사

[국가R&D보고서]

· (주)이지서티, 「빅데이터 환경에서 비식별화 기법을 이용한 개인정보보호 기술 개발」, 2019. 7.

· 서울대학교, 「비식별개인정보의 보호 및 활용에 관한 연구」, 2010. 8.

주제발표 2

전자기록물 공개재분류를 위한 비공개정보 필터링 및 마스킹 기술

김 진 아(국가기록원)

전자기록물 공개재분류를 위한 비공개정보 필터링 및 마스킹 기술

김 진 아(국가기록원)

차 례

1. 머리말

국가기록원이 각급 기관으로부터 **이관 받는 전자기록물은 매년 그 수량이 증가**하고 있다. 2019년의 경우 총 159개 기관에서 83,688철(3,765,563건)을 인수하였으며, 이는 ’18년 대비 27.22%(’18년 약 296만 건→ ’19년 약 377만 건, ↑약 81만 건)[1]가 증가한 수치이다. 10년 전에 생산되어 이관된 상황을 감안한다면, 앞으로 매년 이관 및 관리 대상 전자기록물이 지속적으로 증가할 것이라는 것은 충분히 예측할 수 있는 일이다. 늘어난 전자기록물의 양만큼이나 **국가기록원에서는 기록관리 각각의 프로세스에서 전자기록물 처리에 대한 방안을 검토**하여 관련 체계를 구축하고자 노력하고 있다.

국가기록원은 소장하고 있는 비공개기록물을 적극 공개하고 활용하기 위해 「공공기록물

1) 국가기록원 보존인수과, 「2019년 전자기록물 인수결과보고」, 2020.2월, 2쪽.

관리에 관한 법률(이하 '「기록물관리법」')」에 따라 공개재분류를 추진해왔다. 특히 전자기록이 이관된 지 5년이 도래됨에 따라 **전자기록물의 특징을 고려한 공개재분류 방안 마련의 필요성**이 검토되었다.

비공개기록물에 대한 공개 여부 재분류는 소장 기록물의 적극적인 공개 및 대국민 서비스를 강화하기 위한 기록 관리의 필수 과정으로, 국가기록원은 2007년 공공기록물법을 전면 개정하여 '비공개기록물 중 생산년도 30년이 경과한 기록물은 공개 한다', '공개재분류 후에도 비공개되는 기록물은 매 5년마다 재분류 한다'는 조항을 마련하여 **비공개기록물의 공개재분류**를 적극적으로 추진해 오고 있다. 그러나 국가기록원 소장 기록물 중 공재재분류 시기(생산 후 30년경과)가 도래한 비공개기록물과 생산 · 이관 당시 공개 여부가 누락된 미분류 기록물의 수량은 **매년 200만 권**에 달하는 실정이다. 올해의 경우 법정 도래 대상량 총 1,931,397권 중 **인력과 예산을 고려하여 약 7.9%**에 해당하는 153,059권[2]만 대상으로 공개재분류를 추진하고 있으며, 이는 업무 담당자와 건별 입력 위탁사업을 통해 소화하고 있다.

또한 기록물의 적극적 공개, 기록물에 대한 국민의 관심 증가 및 이관 기록물의 활용을 위한 **기관 및 민원인의 열람도 지속적으로 증가**하고 있는 실정이다. 2019년의 경우 29,340건의 청구건에 대해 총 1,281,898건[3]이 처리되었으며, 이 중 재산관계(토지조사부, 지적원도, 농지상환대장, 분배농지부 등)가 551,219건, 기타(독립운동자료, 관보, 총독부 문서, 국무회의록 등)이 489,733건, 인사관계(강제연행기록물, 인사기록카드, 세대별 주민등록표, 학적부 등)이 132,331건, 행형관계(판결문, 약식명령문, 형사사건부, 수용자 신분장 등)이 108,614건의 순으로 열람이 이루어졌다. 이는 2017년 총 776,423건, 2018년 총 1,050,246건에 비해 상당히 증가된 수치이다.

공개재분류에서는 매년 누적되어 증가되는 기록물에 대한 비공개대상정보 확인시, 열람실에서는 기록물에 대한 비공개정보 처리 시, **그간 그 많은 기록물들을 육안으로 검수**(공개재분류의 경우 비전자기록물 대상)**하여 처리**를 해 왔다. 기록물에 담겨 있는 비공개정보 유형을 확인하고 공개여부가 확정되면 중앙영구기록물관리시스템(CAMS)에 「공공기관의 정보공개에 관한 법률(이하 '「정보공개법」')」에 따라 공개여부(공개, 비공개, 부분공개)와 해당 호수(정보공개법 재9조 제1항각호), 비공개정보유형이 탑재된다. 열람실에서는 CAMS에서 해당 내용 확인 및 열람대상 기록물과 공개재분류 현황을 파악하여 비식별화 처리를 하여 제공하는 식이다.

이는 **국가기록원 뿐 아니라 각급 기관의 기록관에서 동일한 방식으로 운영됨에 따라 겪고 있는 어려움**이기도 하다. 특히 1인 기록관 체계에서 기록관리전문요원 1인이 매년 도래되는 공개재분류 처리와 기관의 정보공개업무까지 담당하고 있는 상황에서는 추진하기 어려운

2) 국가기록원 공개서비스과, 「2020년 비공개 기록물 공개재분류 추진 계획」, 2020.1월, 3쪽.
3) 국가기록원, 「2020년도 국가기록원 주요통계연보」, 2020.6월, 16.국가기록원 열람.

업무라고 할 수 있다. 기록관 단위의 사례를 통해 1인 기록관 체제에서 5년 주기 재분류의 어려움과 문제점을 검토하여 대상기록물을 생산 보존기간 30년 이상으로 한정하는 등의 제안[4]이 있기는 하였으나, 법령 개정이 되지 않는 이상 기록관의 공개재분류 제도 운영상의 실질적인 문제에 대한 검토 없이 이상적인 제도 운영에 대한 제안을 담고 있어 공허함만 남을 뿐[5]이다.

이런 **실무에서의 어려움**은 대량의 기록물에 대한 처리, 해당 기록물이 비공개정보를 포함하고 있는지의 여부를 파악해야 한다는 것, 비공개정보대상이 어디 있는지 일일이 찾아내야 한다는 것이다. 또한 예측하지 못한 대상(공개기록물일 경우 등)에서 비공개정보가 포함되어 있는 경우와 찾아낸 비공개정보를 일일이 비식별화 처리를 해야 하다는 것이다.

본 연구는 올해 전자기록물을 대상으로 공개재분류를 처음 실시('15년부터 전자기록물 본격 이관('05년 이전 생산된 전자기록) → 5년 재분류에 따른 법적 도래)하게 되면서, **이러한 어려움에 해소해 보고자 하는 고민을 담아** 시작하게 되었다. 이미 국가기록원에서는 전자기록의 문제점 및 요구사항 등을 개선하기 위한 지능형 전자기록관리 R&D 사업을 기획[6]하였으며, 현 업무·IT 환경변화를 고려한 전자기록관리 프로세스 전과정*에 대한 새로운 개념의 전자기록 관리체계 도출 필요가 있음을 제시하였다.

* **수집, 인수, 보존포맷, 기술분류, 공개재분류, 보존기간 재평가, 검색/활용, 폐기, 기준관리, 공통/이력관리 등**

전자기록 환경의 변화와 실무의 고민으로 시작된 **『전자기록물 공개재분류를 위한 비공개정보 필터링 및 마스킹 기술 적용방안 연구』**에 대해 소개해 보고자 한다.

2. 연구개발 내용 및 방법

연구를 추진하기 위해서 우선 국가기록원 소장기록물에 대한 비공개정보대상 유형 및 자료에 대한 분석을 추진하였다. **첫째,** 기존에 추진되어 온 공개재분류 기준서(2007년 이후 약 10년치)를 분석하여 정보공개법 제9조제1항제6호(개인정보) 중심에 해당하는 정보를 추출해 보기로 하였다. 이는 기록물의 비공개정보대상 중 가장 많은 부분을 차지(2019년 공개재분류 추진 결과 비공개 사유별 현황 중 제6호가 전체의 약 55.1%를 차지[7])하기도 하며, 기록물 공개 여부시 「개인정보 보호법」에 의해 가장 민감하게 반응되는 항목이기 때문이다. **둘째,** 분석을 통해 필터링을 위한

4) 임희연, 「기록관에서의 공개재분류 제도 개선 방안-서울특별시교육청 사례 중심」, 『기록학연구』 제49호, 한국기록학회, 2016. 7.
5) 권미현, 「비공개기록물 공개재분류 업무절차 개선방안」, 제8차 기록관리 연구세미나 발표자료(2019.10.16.)
6) 국가기록원, 「지능형 전자기록관리 기술연구 개발 기획연구」, 2019.8월
7) 국가기록원, 「2019년 비공개기록물 공개재분류 추진 결과보고」, 2019.12월. 23쪽.

기준을 도출하고 적용 방안을 찾아보도록 하였다. 관련 유사사례 및 개인정보 탐지 상용 S/W를 분석하고, 전자기록물 공개재분류에 적용시킬 수 있는 적용방안 연구하는 것이다. **셋째,** 분석을 통해 도출된 필터링 모델 적용 및 학습데이터를 실험을 위한 테스트베드 구축하여 기술을 적용해 보는 것이다.

1) 소장기록물 분석 및 사례연구

(1) 비공개정보 대상 유형 및 분석

국가기록원에서 그 동안 수행해온 약 **10년간의 공개재분류 기준서**(기록물 유형, 생산기관, 생산년도, 기록물철, 기록물개요, 상세내용, 공개재분류 결과, 검토의견, 비공개정보대상 등)를 바탕으로 제6호(개인정보) 중심으로 분석을 실시하였다.

- 분석대상 : 총 5,418개(전체 9,567개 중 중앙행정기관의 제6호 포함사항)
- 분석내용 : **업무분류체계 및 개인정보유형별**로 구분하여 **주요 단어 및 내용 분석** 추출
  - 업무분류체계 : 외교, 감사, 소청·소원·소송 등 약 23개 유형[8]과 개인식별정보 및 개인증빙기록으로 구분
  - 개인정보유형 : 가족정보, 교육 및 훈련정보, 병역정보, 부동산 정보 등 약 15개[9]

8) 국가기록원 기록관리 공공표준, NAK 16-2 2013(v1.0)「기록물 공개관리 업무-제2부:영구기록물관리기관(v1.0)」, 2013
9) 개인정보보호포털, www.privacy.go.kr

업무분류체계 기록물유형	비공개정보대상 제6호	비고
공통되는 개인정보	〈개인식별정보〉 • 이름, 성별, 본과, 본적, 주소, 출생지, 생활근거지, 전화번호, 생년월일, 주민등록번호, 연령, 직업, 소속, 직위, 직업, 최종학력, 출신학교, 학력, 경력, 재산, 혈액형, 호주와의 관계, 취미, 특기, 가족사항, 가족관계, 신체사항, 재산사항, 전자우편주소 등 • 계좌번호, 은행명, 차량번호 • 여권번호, 발급일, 여권 유효기간, 여행목적지, 여행목적, 국적 • 군번 및 계급 〈개인증빙기록〉 • 주민등록표, 등기부등본, 인감증명원·서, 인감신고서, 인감증명신청, 주민등록초본, 주민등록등본, 세대별주민등록표, 호적초본, 호적등본, 자동차등록증, 호적조회서, • 인사기록카드, 인사발령통지서, 인사기록부, 개인신상조사서, 신상카드, 신상관계서, 사망진단서, 신분증 사본, 동일인증명서, 법최사실증명서 • 신원증명원·서, 신원조사회보서, 이력서, 가족관계증명서, 병적증명서, 신원진술서, 신원조사서, 신원보증서, 인감증명서, 재직증명서, 재산증명원, 재산증명서, 위임장, 건강진단서, 병적증명서, 제적등본, 인우보증서 • 공무여행심사기록, 영어성적증명서, 자격증 사본, 시험합격증서, 여권 사본, 외국인등록표 • 대학 졸업증명서, 석사학위증명서, 박사학위증명서, 성적증명서, 소득증명서 • 취임승낙서, 결근계, 연명부, 각종명부, 근무상황카드, 사직서, 출근부, 월평균보수지급액, 출근표, 급여지급명세서, 기무승락서, 취임동의서 (이하 생략)	

개인정보 항목	비공개정보대상	비고
가족정보	• 가족구성원들의 이름, 성별, 연령, 관계, 출생지, 생년월일, 주민등록번호, 소속, 주소, 학력, 종교, 직업, 직장, 직위, 소속, 직급, 전화번호, 여권번호, 경력, 생활환경, 동거여부, 부양여부 • 부양가족사실증명	
교육 및 훈련정보	• 학교출석현황, 최종학력, 학교성적, 기술 자격증 및 전문 면허증, 이수한 훈련 프로그램, 동아리활동, 상벌사항 • 교육훈련(기간 계급, 종류, 성적, 실시기관) • 교육정소, 임교수로일자, 성적평가(과목별 점수, 총점, 평균, 석차), 교육근태(특이사항) • 학적부, 경찰교육성적대장, 교육원부 • 심사점수, 출결결과가 포함된 평단(전문조사요원 심사지로, 점수산정, 최종선발자평단, 선발대상자 평단, 교육근태, 조사요원가산점수환산 등)	
병역정보	• 군번 및 계급, 제대유형, 필/미필, 사유, 제적등위, 복무기관주특기, 근무부대, 역종, 군별, 병과, 입대년월일, 제대년월일, 복무상황, 훈기번호 • 병적증명(조회)원서, 병역수첩, 병적확인증, 병역필확인증, 병역복무기간 정정내역, 병역전역증 • 병역/소집 면제자 명부	
부동산 정보	• 소유주택, 토지, 자동차, 기타소유차량, 상점 및 건물 등 • 소유 부동산 혹은 동산 금액, 재산총액, 가옥, 부업, 생활정도	
소득정보	• 월수입, 현재 봉급액, 봉급경력, 보너스 및 수수료, 기타소득의 원천, 이자 소득, 사업소득, 적금 • 직책수당, 근속수당, 법인소득금액 • 봉급(본봉, 정근수당, 직무수당, 장근수당)	
기타 수익정보	• 보험 (건강, 생명 등) 가입현황, 회사의 판공비	
신용정보	• 대부잔액 및 지불상황, 저당, 신용카드, 지불연기 및 미납의 수, 임금압류 통보에 대한 기록 • 신용카드이용대금명세서, 압금증, 긴이영수증, 무통장입금증, 신용카드번호, 계좌번호 등	
고용정보	• 현재의 고용주, 고용인(이름, 주민등록번호, 주소, 전화번호 등 개인정보), 회사주소, 상급자의 이름, 직무수행평가기록, 훈련기록, 출석기록, 상벌기록, 성격 테스트결과 직무태도 • 비정규직고용계약서	
법적정보	• 학직위반자·징계대상자·자퇴자·익사자 등의 이름, 학과, 학번, 주민등록번호, 생년월일, 성적, 출신고, 본적 및 주소의 읍면동 이하, 가족관계, 가정환경 등 개인정보 • 전과기록, 범죄내용, 위법적용 조항 및 1, 2심 판결내용, 자동차 교통 위반기록, 파산 및 담보기록, 구속기록, 이혼기록, 납세기록, 소송 내용 • 징계형벌(년월일, 계급, 종류, 사유, 발령기관)	
의료정보	• 가족병력기록, 과거 의료기록, 개인 의료기록, 정신질환기록, 신체장애, 혈액형, IQ, 약물테스트 등 각종 신체테스트 정보 • 채용신체검사서, 건강검진결과, 건강진단카드, 건강진단서 • 의료보험번호, 의료기록, 질환기록 • 사체해부기록, 진단서부기록서, 병인증서, 사망진단서	
조직정보	• 노조가입, 종교단체가입, 정당·사회단체 가입, 클럽회원, 가입단체 (이하 생략)	

(2) 선행 연구 및 사례 분석

본 연구의 선행연구로는 **기계학습 기반의 기록관리 모델 연구**인, 서울시 정보소통광장 기록물(전자결재) 대상 기계학습 기반의 문서 자동분류 연구(서울시 BRM체계 활용)[10]로 기록 텍스트 자동 분류를 위한 기계학습 플랫폼을 구축하고 수행한 사례와, 대용량의 텍스트로부터 다양한 자연어처리(텍스트 분석)의 작업들을(개체명인식 모델, 문서분류, 대화모델 등) 사전학습과(pre-trainning) 후 처리(fine-tuning)작업을 통해 산출해 낸 연구[11] 등이 있다.

또한 **공공기관의 유사사례**를 분석하여 연구 동향에 대해 파악하였다. 관세청의 경우 정책연구로 공공데이터 개인정보에 대한 비식별화 연구를 수행하여 비식별화 적용대상과 활용 가능성, 비식별화 가이드라인, 비식별화 방안, 비식별화 처리 및 프라이버시 보호모델 S/W 개발 가이드라인에 대한 연구[12]를 수행하였고, 법원행정처의 경우 판결문에 대한 비식별화 작업을 위하여 자연어 처리, 기계학습 모델 기능, 개체명, 패턴 추출, 비식별화 처리 및 비식별화 검증 관리 기능 등의 솔루션 사업을 계속 사업[13]으로 진행하고 있다.

기록물에 대한 유사사례를 분석하기 위해 **시중 상용화 되어 있는 개인정보 탐지용** S/W를 분석한 결과, 평균 약 7~12개의 개인정보(숫자 등으로 구성된 일정 패턴)을 검출하는 것으로 나타났으나, **이름에 대한 검출은 없는 것**으로 조사되었다.

10) ㈜아카이브랩, 「국가기록원 차세대 기록관리 모델 재설계 연구 개발」, 정보관리학회지, 34(4), 321-344
11) Google, 「Pre-training of Deep Bidirectional Transformers for Language Understanding」, 2018.10.11., arXiv.org
12) 관세청, 「공공데이터 개방범위와 개인정보 등 비공개 정보 비식별화 방안 연구」, 2019년, www.prism.go.kr
13) 법원행정처, 「2019년 지능형 비식별환 솔루션 도입(2단계) 사업」, www.g2b.go.kr

제조사	제품명	수용 개인정보 유형	비고
지란지교	SERVERFILTER for RMS	주민등록번호, 유사주민번호, 여권번호, 운전면허번호, 신용카드번호, 외국인등록번호, 건강보험번호, 이메일주소, 핸드폰번호, 계좌번호, 법인번호, 사업자번호	12개
이지서티	U-PRIVACY SAFER	주민번호, 외국인등록번호, 여권번호, 운전면허번호, 핸드폰번호, 일반전화번호, 이메일주소, 건강보험번호, 계좌번호, 기관번호	10개
글로벌다윈	SmartXFilter	핸드폰번호, 이메일주소, 계좌번호, 주민등록번호, 신용카드번호, 보수개인정보, 고유식별정보	7개
센티널 테크놀로지	CoolFilter	주민등록번호, 외국인번호, 여권번호, 운전면허번호, 유사주민번호, 신용카드번호, 휴대전화번호, 계좌번호, 이메일주소, 법인번호, 사업자번호, 건강보험번호	12개
나노디엠에스	SPK Server Filter	주민번호, 외국인등록번호, 여권번호, 운전면허번호, 핸드폰번호, 일반전화번호, 이메일주소, 건강보험번호, 계좌번호, 기관번호	10개
컴투루 테크놀러지	PrivacyCenter	주민등록번호, 외국인번호, 여권번호, 운전면허번호, 유사주민번호, 신용카드번호, 휴대전화번호, 계좌번호, 이메일주소, 법인번호, 사업자번호, 건강보험번호	12개

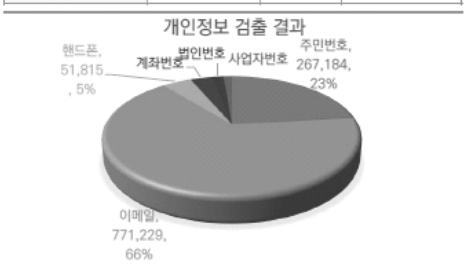
각급 기관에서 기록관에서 사용 중인 표준기록관리시스템 내에 개인정보 필터를 도입한 기관을 파악하여 사용 현황을 조사한 결과, 현재 사용 중인 한 개 기관의 샘플 데이터를 분석할 수 있었다. 2018년에 실시한 결과(2004년 생산 기록물 대상)로, 분석 결과 총 **9개 유형**(주민등록번호, 여권번호, 운전면허, 신용카드, 이메일, 핸드폰, 계좌번호, 법인번호, 사업자번호)이 검출이 가능하며, 이메일 〉 주민등록번호 〉 핸드폰번호 순으로 검출된 걸 확인할 수 있었다. 또한 공개재분류 대상인 부분공개, 비공개 기록물 뿐 아니라 **공개기록물에서도 개인정보에 해당**하는 비공개정보대상이 전체 검출 대상 중 약 88.8%를 차지한 것을 알 수 있었다. 이는 공개기록물 대상으로도 비공개정보에 필터링 대책이 필요하다는 것을 알 수 있는 사례이다.

개인정보 검출 결과

구분	검출 기록물 건수	%	비고
주민등록번호	267,184	22.88%	총 : 962,955개 기록물
여권번호	3,509	0.30%	
운전면허	63	0.01%	
신용카드	58	0.01%	
이메일	771,229	66.03%	
핸드폰	51,815	4.44%	
계좌번호	39,153	3.35%	
법인번호	20,067	1.72%	
사업자번호	14,815	1.27%	
총합계	1,167,893	100%	중복된 검출 포함

개인정보 포함 현황

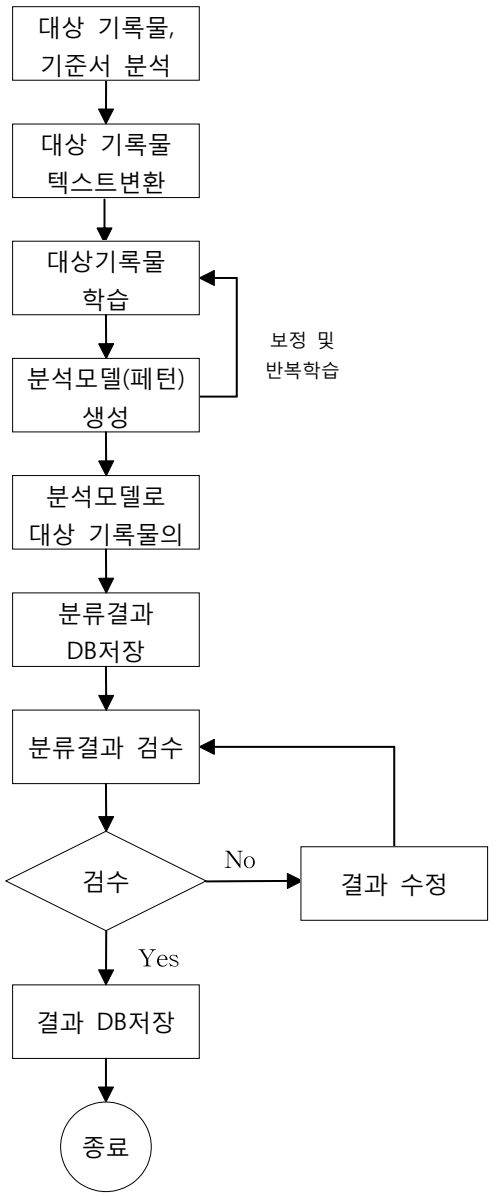
문서 공개여부별 구분 (공개 값)	개인정보포함 기록물 건수	개인정보 포함%	비고
공개기록물(1)	855,057	88.8%	공개 재검토 필요
부분공개기록물(2)	25,015	2.5%	법정 공개재분류 대상 기록물
비공개기록물(3)	82,885	9.7%	
총합계	962,957	100%	



2) 테스트베드 설계/구축

전자기록물 대상 **학습기/분석기**에 입력할 수 있는 **텍스트파일 변환**과 변환된 기록물의 **기계학습 수행**을 위한 과정의 연구, 최종 분석모델로 기록물을 분류하고 DB에 저장, DB 결과를 받아 비공개정보대상을 마스킹하는 기능으로 절차를 정의하고 다음과 같이 설계하였다.

실행 절차(안)



구축 사항

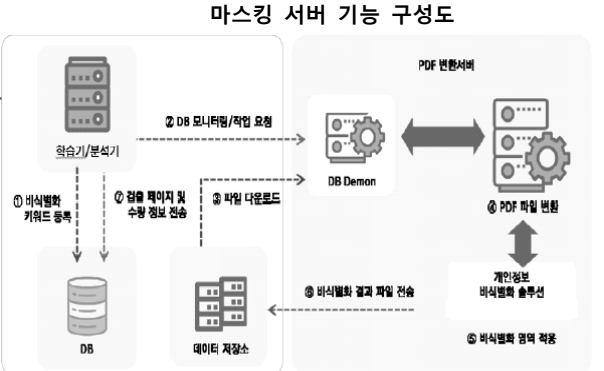
학습기/분석기(하드웨어)

구분	상세 규격
CPU	AMD Quad Core 4.2Ghz
GPU	GeFORCE 1.7GHz, 6GB Memory
MEMORY	32GB
DISK	1,000GB
Network Adapter	1,000Mbps
Monitor	17" LED

학습기/분석기(소프트웨어)

구분	제품명
OS	마이크로소프트사 Window10
DBMS	오라클사 ORCLE
검색엔진(문서필터)	다이퀘스트사 Mariner R
학습 S/W	이투온사 PyTorch
분석 S/W	이투온사 PyTorch

마스킹 서버 기능 구성도



※ 성능과 안전성, 기술지원 용이성, 연구용이성 등을 감안하여 해당 H/W, S/W 등을 선정

또한 향후 **국가기록원 CAMS와의 연계**를 위해 해당 팀과 업무협의를 하였으며, CAMS의 보안성 확보와 운영의 용이성을 감안하여 별도의 연계 API(Application Programming Interface)를 연계하는 방안으로 정리하였다.

CAMS 연계 검토	협의사항	비고
CAMS DB 사용/동기화	CAMS DB 연계 : CAMS 서버와 API 통신을 통한 연결방식	연계 API 개발 필요
공개재분류 서버 파일 전송방법	XTORM(EDMS) 솔루션을 통해 스토리와 연결, XTORM 직접 Access	
공개재분류 서버 전송 파일 형태	전송 파일 형태 : neo	
CAMS 공개재분류 메뉴 연동	AP 서버와 API 통신을 통한 연결방식	
마스킹 처리 파일 전달방식	부분공개 열람요청시 마스킹 서비스 파일은 API를 통해 CAMS에 전송	

3) 분석모델 및 기계학습

(1) 텍스트파일 변환 분석

학습기/분석기 학습을 실행하기 위해서는 아래아 한글과 같은 원본파일 또는 문서보존포맷인 PDF는 **반드시 텍스트로 변환**하여야 한다. **초기에 문서보존포맷인 PDF파일**을 대상으로 연구설계 하였으나, 변환시 정합성 오류가 발생하여 이후 **원본파일을 대상으로 추진**하였으며 파일 변환 분석 결과는 아래와 같다.

텍스트 파일 변환 분석 결과

▶ 학습기/분석기는 분석 대상의 기록물을 반드시 텍스트로 변환하여야 처리 가능
  * CAMS 내 다양한 문서 포맷을 텍스트로 변환하는 변환기(문서필터)가 필요하며, 문서 내 이미지(스캔이 된 전자화 기록물)을 제외한 모든 문서 변환 제공 가능

▶ 연구 초기에는 CAMS 문서보존포맷인 PDF파일을 텍스트변환 파일로 변환하여 학습기/분석기에 로딩하는 방안으로 연구 설계를 수행

▶ PDF 변환시, 정합성 등 여러 종류의 오류사항이 발생, 변환 엔진에 따라 텍스트 변환 오류가 조금씩 다르게 발생
  * PDF 변환 엔진에서 인지하지 못하는 폰트나 특수폰트의 경우 이미지 변환 PDF 파일로 변환됨

<PDF 변환시 오류의 예시>
- 괄호안의 숫자가 누락되거나 영문이 누락되는 경우(괄호안의 숫자와 영문의 폰트를 인식하지 못해 이미지로 변환되어 PDF 파일이 생성되는 것으로 추정)
- 숫자+문자의 형태의 경우 순서가 바뀌어 변환되는 경우
- 표가 있는 경우 표안의 내용이 누락되어 변환(폰트 인식 및 특수폰트 사용으로 인한 이미지 변환으로 PDF 파일이 생성되는 것으로 추정)

▶ 실험 연구에 따라 오류케이스를 분석, PDF 혹은 원본문서의 텍스트 변환을 비교 확인하여, 공개재분류 기록물의 포맷 대상(원본기록물, PDF)결정을 위한 검증 추진

▶ CAMS 경우, XML Parser를 통하여 원본문서, PDF 파일을 추출 가능

▶ RMS의 경우는 원본문서만 보관되어 있는 경우가 대부분이어서 원본문서를 대상으로 테스트 변환 수행가능

학습기/분석기를 통한 **텍스트 변환의 일반적인 예시**이다. 아래아 한글파일로 PDF와 한글파일에서 변환된 사례이다. **동일하게 변환됨**을 확인할 수 있다.

변환 전 원문파일	PDF를 변환시	한글파일을 변환시
감 사 결 과 요 약 대한주택공사는 1962. 7. 1. 주택 등을 건설·공급 및 관리하기 위한 목적으로 설립된 이후 2003년 말까지 총 144만여 호의 주택(분양 80만여 호, 임대 64만여 호)을 건설·공급하였다. 위 공사는 2003년도에 부동산 경기의 호조로 2,033억여 원의 당기순이익을 실현하였으나 정부의 국민임대주택 100만 호 건설계획(2003~2012년)에 따라 수익성이 낮은 임대주택건설물량을 늘리고 수익성이 높은 분양주택의 공급을 축소하는 등 사업구조가 임대주택사업 위주로 변화되어 재무건전성을 유지하는데 큰 어려움이 예상되었다. 이에 따라 감사원에서는 경영수지개선 및 재무자원의 효율적 운용을 도모하는데 감사목적을 두고 2004. 3. 29.부터 같은 해 4. 14.까지 대한주택공사 본사 및 2개 지역본부를 대상으로 재무감사를 실시하였던 바 감사결과 나타난 재무자원 현황, 결산확인 및 분야별 문제점은 다음과 같다. <재무자원 현황 및 결산확인> 대한주택공사의 재무자원은 2003년 말 현재 자산 16조 7,901억여 원, 부채 10조 1,285억여 원, 자본 6조 6,616억여 원으로 2001년 이후 최근 3년간 자산, 부채 및 자본이 지속적으로 증가하였다. 이는 주로 국민임대주택의 건설물량 증가에 따른 국민주택기금의 증가 등으로 부채가 3,621억여 원 증가하고, 정부출자금과 당기순이익이 큰 폭으로 증가하여 자기자본이 1조 248억여 원 증가하였기 때문이다. 최근 3개년간(2001~2003년)의 결산내용을 분석해 본 결과 매출액이 감소하고 임대사업에서 손실(805억여 원)이 발생하였는데도 부동산경기의 활성화로 분양	.PAGE:3 -1- 감 사 결 과 요 약 대한주택공사는 1962. 7. 1. 주택 등을 건설공급 및 관리하기 위한 목적으로 설립된 이후 2003년 말까지 총 144만여 호의 주택(분양 80만여 호, 임대 64만여 호)을 건설공급하였다. 위 공사는 2003년도에 부동산 경기의 호조로 2,033억여 원의 당기순이익을 실현하였으나 정부의 국민임대주택 100만 호 건설계획(2003~2012년)에 따라 수익성이 낮은 임대주택건설물량을 늘리고 수익성이 높은 분양주택의 공급을 축소하는 등 사업구조가 임대주택사업 위주로 변화되어 재무건전성을 유지하는데 큰 어려움이 예상되었다. 이에 따라 감사원에서는 경영수지개선 및 재무자원의 효율적 운용을 도모하는데 감사목적을 두고 2004. 3. 29.부터 같은 해 4. 14.까지 대한주택공사 본사 및 2개 지역본부를 대상으로 재무감사를 실시하였던 바 감사결과 나타난 재무자원 현황, 결산확인 및 분야별 문제점은 다음과 같다. <재무자원 현황 및 결산확인> 대한주택공사의 재무자원은 2003년 말 현재 자산 16조 7,901억여 원, 부채 10조 1,285억여 원, 자본 6조 6,616억여 원으로 2001년 이후 최근 3년간 자산, 부채 및 자본이 지속적으로 증가하였다. 이는 주로 국민임대주택의 건설물량 증가에 따른 국민주택기금의 증가 등으로 부채가 3,621억여 원 증가하고, 정부출자금과 당기순이익이 큰 폭으로 증가하여 자기자본이 1조 248억여 원 증가하였기 때문이다. 최근 3개년간(2001~2003년)의 결산내용을 분석해 본 결과 매출액이 감소하고 임대사업에서 손실(805억여 원)이 발생하였는데도 부동산경기의 활성화로 분양	감 사 결 과 요 약 대한주택공사는 1962. 7. 1. 주택 등을 건설공급 및 관리하기 위한 목적으로 설립된 이후 2003년 말까지 총 144만여 호의 주택(분양 80만여 호, 임대 64만여 호)을 건설공급하였다. 위 공사는 2003년도에 부동산 경기의 호조로 2,033억여 원의 당기순이익을 실현하였으나 정부의 국민임대주택 100만 호 건설계획(2003~2012년)에 따라 수익성이 낮은 임대주택건설물량을 늘리고 수익성이 높은 분양주택의 공급을 축소하는 등 사업구조가 임대주택사업 위주로 변화되어 재무건전성을 유지하는데 큰 어려움이 예상되었다. 이에 따라 감사원에서는 경영수지개선 및 재무자원의 효율적 운용을 도모하는데 감사목적을 두고 2004. 3. 29.부터 같은 해 4. 14.까지 대한주택공사 본사 및 2개 지역본부를 대상으로 재무감사를 실시하였던 바 감사결과 나타난 재무자원 현황, 결산확인 및 분야별 문제점은 다음과 같다. <재무자원 현황 및 결산확인> 대한주택공사의 재무자원은 2003년 말 현재 자산 16조 7,901억여 원, 부채 10조 1,285억여 원, 자본 6조 6,616억여 원으로 2001년 이후 최근 3년간 자산, 부채 및 자본이 지속적으로 증가하였다. 이는 주로 국민임대주택의 건설물량 증가에 따른 국민주택기금의 증가 등으로 부채가 3,621억여 원 증가하고, 정부출자금과 당기순이익이 큰 폭으로 증가하여 자기자본이 1조 248억여 원 증가하였기 때문이다. 최근 3개년간(2001~2003년)의 결산내용을 분석해 본 결과 매출액이 감소하고 임대사업에서 손실(805억여 원)이 발생하였는데도 부동산경기의 활성화로 분양사업에서 1조 2,134억여 원의 매출이익이 (이하 생략)

(2) 모델설계

기계학습을 위한 텍스트파일의 **비공개정보 분석 결과, 비공개정보의 종류 및 특성, 패턴을 고려 정규표현식(패턴)과 기계학습을 결합**하는 것으로 모델 설계를 하였다. 기계학습은 정규표현식(패턴)으로는 해결하지 못하는 인물명 중심으로 수행이 가능하며, 단어사전을 활용해 비공개 추출을 하는 것이 효율적인 것으로 판단하였기 때문이다.

모델 설계 방안
▶ 제6호(개인정보)와 제2호(국가안전보장·국방·통일·외교) 비공개단어 분석 결과, 비공개정보 종류 및 특성, 또는 패턴을 고려 패턴 분석과 기계학습을 결합하는 것으로 판단
▶ 기계학습을 통해 텍스트에서 바로 인식할 수 있는 개체명은 인명, 성별, 장소, 시간 등에 해당하나, 제6호 및 제2호 비공개정보 특성을 분석하고, 원본파일 또는 PDF 파일을 변환한 텍스트파일 샘플 테스트 결과를 종합적으로 고려하여 인물명 중심의 기계학습 수행
▶ 개체명 인식이 가능한 장소의 경우, 주소 정규식 패턴과 주소DB를 이용할 경우 기계학습에 비해 상대적으로 상세한 주소에 대해 분석 또는 인식이 가능
▶ 정규식 형태의 비공개정보(이메일, 여권번호, 주민등록번호, 외국인등록번호, 사업자등록번호 등 10개)는 정규식 패턴인식을 통해 비공개정보 추출 또는 검색할 경우 기계학습에 비해 상대적으로 높은 인식률 달성이 가능
▶ 제6호 및 제2호 비공개정보대상의 경우, 단어사전과 글자수 패턴을 적용

이메일, 여권번호, 주민등록번호, 외국인등록번호, 사업자등록번호, 군번, 운전면허번호, 카드번호, 주소, 소송번호 총 10개의 **정규식 패턴**을 설정하고 주소 정규식에서 **제외할 단어**(주소와 붙어서 같이 쓰이는 단어 등)를 선택 후 패턴 설정하였다.

주소 정규식 패턴에서 제외한 단어 목록
'공개','분리대','구간','공사','개선','위치','지점','지역','점용','전용','앞','옆','좌측','우측','맞은편','근처','재난','변경','설계','설치','매설','주변','면적','분할','합병','통합','계획','관광지','사고','검토','폐쇄','준공','신청','허가','명칭','시설용지','시설부지','현장','포교당','당사는','귀청과','승인','발생','침수','민원','노력','합계','총계','경기도계','강원도계','충북도계','충남도계','전남도계','전북도계','경북도계','경남도계','제주도계',' 계',' 지내',' 일원',' 일대',' 소재',' 근처',' 부근',' 제방',' 교량',' 대교', ' 정문',' 인근', '외 ','대교 ', '면 적'

제6호 (개인정보)의 **단어사전을 구축하여 글자수 패턴**을 적용하였다.

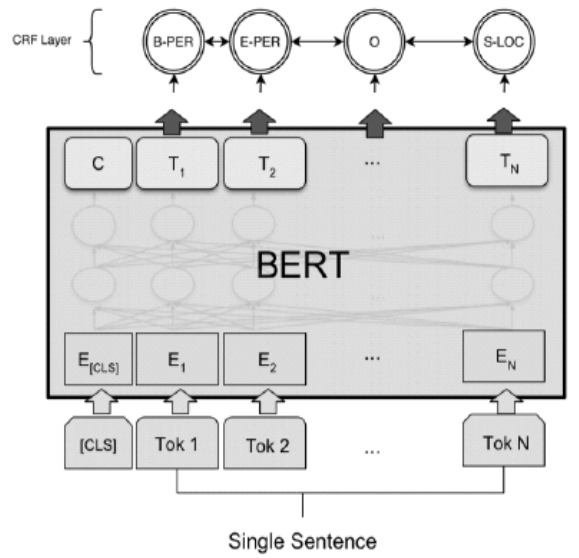
- 총 157개의 개인식별정보와 186개의 개인증빙기록총, 343개의 비공개단어를 개인정보 유형별과 업무분류체계별 2가지 기준으로 유형 분류하여 비공개정보사전 구축
- 기본 비공개단어에서 글자간의 띄어쓰기를 추가 후 사전으로 구축
- 텍스트 변환시 띄어쓰기 오류가 발생한 단어까지 검출 가능

제2호 (국가안전보장·국방·통일·외교관계)도 86개의 비공개정보를 단어 사전을 구축하였으며, 기타 직업과 관련한 단어도 사전으로 구축하여 적용하였다.

기계학습 사전학습모델로 **Kobert 모델**[14]**을 선정**하였다. 국가기록원 기록물에 포함되어 있는 사람 이름을 추출하여, Kobert 사전학습 데이터 및 기계학습 베스트 모델에 해당하는 것을 pytorch-bert-crf-ner 폴더 내에 별도 저장하였다. 총 44,,832개의 전자기록물 파일을 대상으로 학습을 수행, **반복학습**을 통하여, 예측에 벗어난 케이스를 연구하여 정확도를 향상시켰다.

14) Google에서 발표한 BERT 알고리즘 기반으로 BERT+CRF 기반의 한국어 NER Target을 활용한 Kobert 모델을 적용하였으며, 기본 구분 태그는 인명을 포함하여 10개로 구성

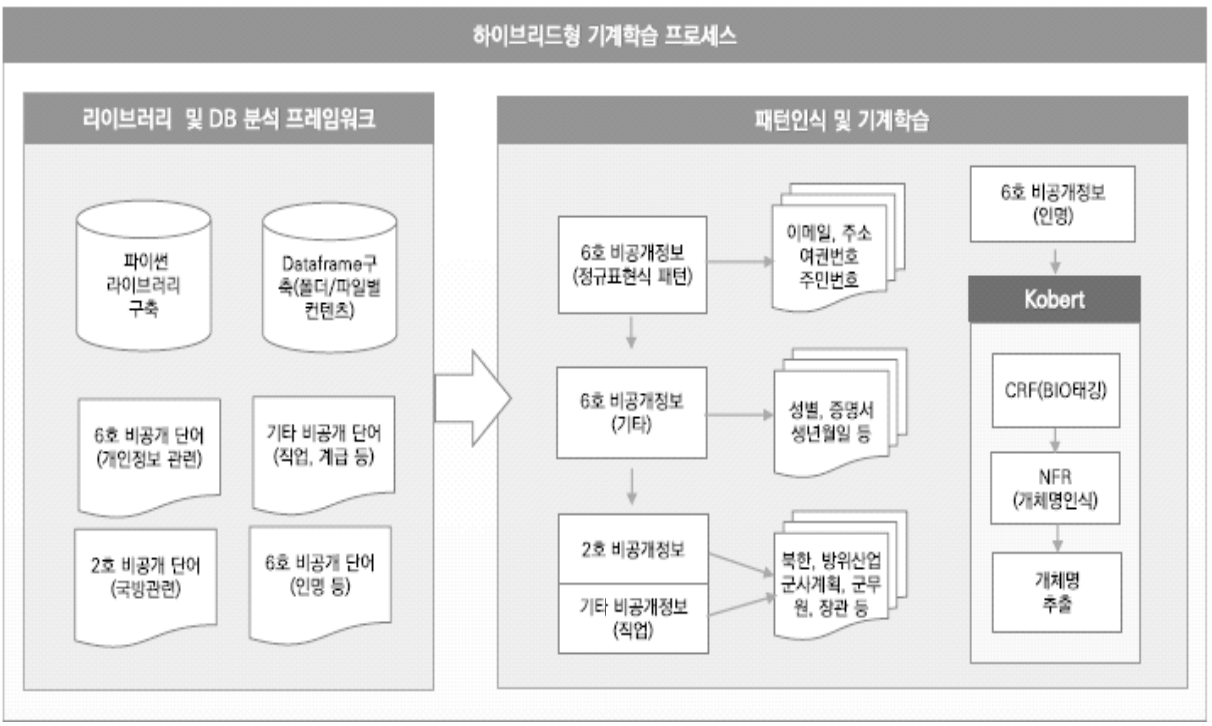
BERT의 논리적 구성도



분석모델 업그레이드 반복

구 분	수행 업무
데이터 분석	• 텍스트 문서 분석 • 비공개 기준 설정
분석모델	• 비공개정보 유형별 패턴분석 • 단어사전 및 분석모델, 패턴분석 업그레이드
기계학습모델	• 기계학습모델 분석 • 기계학습모델 성능 파악 및 반복학습
사용자 인터페이스	• 국가기록원 요구 분석 • 사용자 요구사항 반영한 프로토타입 구현

데이터 분석, 패턴분석, 단어 사전 및 분석모델, 패턴분석 업그레이드, 기계학습을 반복하여 기계학습과 패턴인식의 조합인 '하이브리드형 기계학습 프로세스'를 추진하였으며, 기계학습의 장점과 정규식 분석의 장점을 각각 취합하여 구성하였다.

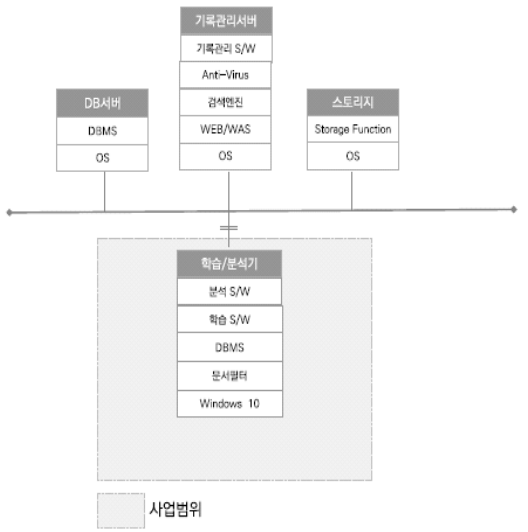


3. 연구개발 결과

1) 테스트베드 구축

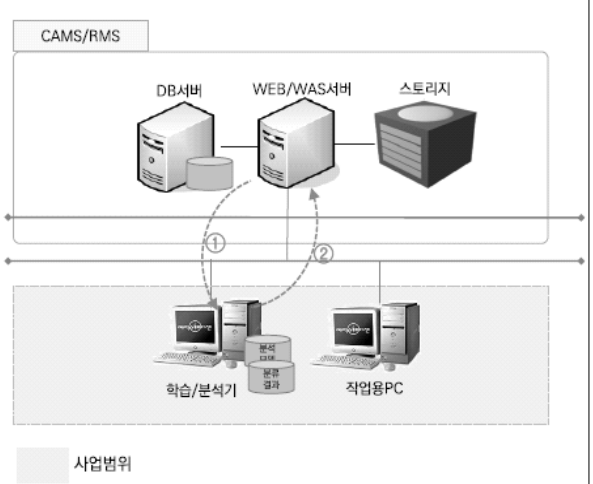
테스트베드를 구축하기 위해 적합성 검토를 통해 H/W를 구축하였고, 학습/분석 소프트웨어 탑재 및 운영을 위한 하드웨어와, Kobert 학습 모델을 위해 GPU를 설치하였다. 구축된 테스트베드의 H/W 및 S/W는 다음과 같다.

학습기/분석기 하드웨어

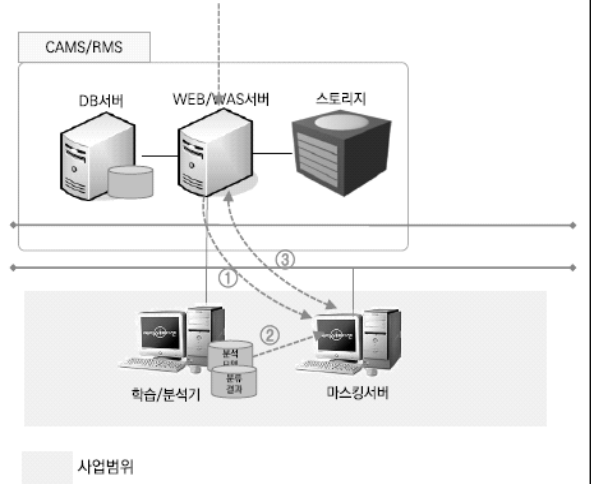


구분	제품 명	상세 기능 및 선정사유	비고
OS	Microsoft Window10	• 오픈, 개방형 운영체계 • 학습기/분석기 설치에 가장 많이 사용 • 기술지원의 용이성이 높아 선정	테스트베드 적용
DBMS	오라클 ORACLE	• DB 구성 및 시험의 편의성 및 효율성 제공 • 대용량 DB에 성능 및 안정성 고려 선정	테스트베드 적용
검색엔진 (문서필터)	다이퀘스트 Mariner R	• 다양한 기록물의 텍스트 변환 용이 • PDF 변환은 이미지로 인식하여 변환이 되는 상황이 많아 원본 기록물을 대상으로 검출하는 것이 바람직 함. • 시중 가장 많은 다양한 포맷과 버전의 기록물과 한글을 변환할 수 있는 상용 소프트웨어 선정	테스트베드 적용
학습S/W	이루온 PyTorch	• 국가기록원의 기록물과 공개재분류기준서를 기반으로 학습을 통해 분석 모델(패턴) 제작 • 본 연구용역의 원활한 학습 진행을 위해 제안사의 학습기 선정	테스트베드 적용
분석 S/W	이루온 PyTorch	• 학습기의 분석기 모델(패턴)을 받아 분류 태그값과 매치되는 기록물을 분류하기 위한 모듈 • 본 연구용역의 원활한 기록물 분류 진행을 위해 제안사의 분석S/W 선정	테스트베드 적용
마스킹S/W	유니닥스 Black Marking	• 정보공개서버의 요청을 기록관리시스템에서 받아 부분공개 대상 기록물을 추출하여 마스킹 처리 업무 수행 • 등록된 PDF전문사의 개발의 도움을 받아 DB연동을 하기위해 상용소프트웨어 선정	마스킹 소프트웨어 연구와, 연계를 위한 추가적인 개발이 진행되어야 함으로 테스트베드에는 적용하지 않음.

학습기/분석기 소프트웨어



마스킹 서버(설계)

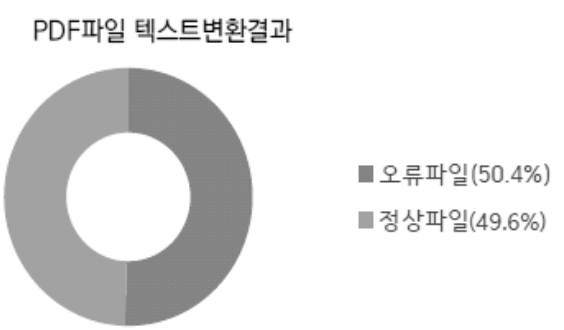
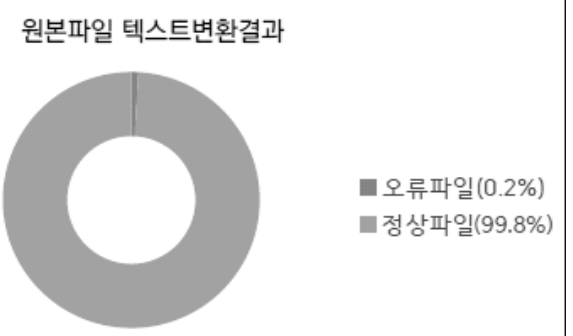


비공개정보 필터링 및 마스킹을 수행을 위하여, 국가기록원 CAMS 및 RMS와의 연계 설계를 진행하였다. CAMS/RMS 연계, 공개재분류 서버, 마스킹 서버로 구분하였고 연계를 위한 기능 요구 도출 사항은 아래의 표와 같다.

요구사항명	연계 기능요구사항	비고
CAMS/RMS 연계 Agent	○ 공개재분류 서버 연계 Agent - CAMS와 RMS의 공개재분류 기능과 API로 연계되어 공개재분류대상 기록물을 공개재분류 서버에 전송 - 공개재분류 결과 DB를 CAMS와 RMS의 공개재분류 DB테이블에 API를 활용 전송 저장 ○ 마스킹 서버 연계 Agent - CAMS와 RMS의 부분공개 기록물을 API를 통하여 마스킹 서버로 추출 전송 - 학습/분석기의 공개재분류 결과 DB를 마스킹 서버에 API를 통하여 전달 - 마스킹서버로부터 마스킹 처리된 기록물을 API를 활용 전송받아 DB와 기록물을 저장	
공개재분류 서버	○ CAMS/RMS 연계 공개재분류 Agent - CAMS와 RMS의 공개재분류 기능과 연계되어 공개재분류대상 기록물을 공개재분류 서버에 API를 통하여 전송된 기록물을 수신 저장 - 공개재분류서버 DB를 CAMS와 RMS의 공개재분류 DB테이블에 API를 통하여 전송 ○ 텍스트 변환 모듈 - CAMS/RMS의 API로부터 전송된 기록물을 텍스트로 변환 저장 - 텍스트로 변환된 파일을 학습기/분석기에 전송 ○ 학습/분석기 - 텍스트 변환 모듈로부터 전송받은 텍스트 기록물을 반복 학습을 통하여 학습모델 제작 - 학습모델을 통하여 공개재분류대상 기록물을 분류 - 분류 결과와 비공개 사유를 학습/분석기 DB에 저장 - 공개재분류 결과 DB를 CAMS와 RMS의 API연계 모듈을 통하여 전송	
마스킹 서버	○ CAMS/RMS 연계 마스킹 Agent - CAMS/RMS의 부분 공개 열람요청을 받아 해당 기록물을 API를 통하여 마스킹 서버에 기록물 전송 - 마스킹 처리된 기록물을 연계 API를 통하여 CAM/RMS에 전송하여 DB에 등록 ○ 마스킹 모듈 - PDF파일과 공개재분류 DB 결과값을 수신 - 결과값에 해당하는 키워드를 검색하여 검출 - 키워드 검색 검출 결과 값을 DB에 저장 - 1차로 검색된 결과를 육안으로 검수 - 수정할 사항은 PDF편집기로 편집 - 편집 완료된 PDF파일의 비공개 대상 키워드를 마스킹 처리 - 마스킹 처리된 파일을 마스킹 API를 통하여 CAMS/RMS로 전송	

2) 문서필터 텍스트 변환 검증 결과

국가기록원 중앙영구기록물관리시스템(CAMS) 내 PDF파일과 원문파일을 텍스트파일 변환하여 비교한 결과, PDF **파일은 49.6%** 정확도, **원문파일은 99.2%**의 정확도로 변환이 되었다.

PDF 파일 텍스트변환 결과(정상 49.6%)	원본파일 텍스트변환 결과(정상 99.8%)
	

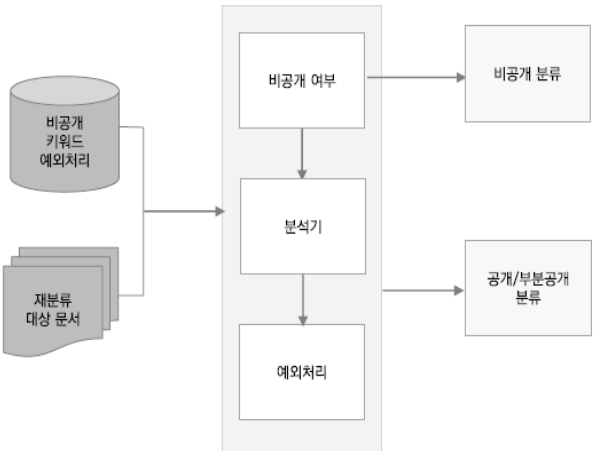
3) 학습기/분석기 구축 결과

각 비공개정보 유형별로 가장 효율적이면서 정확도가 높은 방식을 찾아 연구하였으며, 연구 결과 **기계학습과 정규식 표현 방식, 사전 등을 복합적으로 활용**하여 **적용**하는 것이 가장 정확도가 높은 방법임을 확인하였다. 학습기/분석기 구성은 여러 가지 분류모델(기계학습, 정규표현식, 사전 등) 적용 및 유기적인 결합을 위해 파이썬을 활용 개발 통합하였다. 또한 기록물에 비공개 정보 출현시, 형태로는 비공개정보이지만, 내용상 공개정보인 대상 처리를 위하여 **비공개로 분류하는 선조치 작업 후 정보 공개 예외처리를 위한 후처리 작업**을 진행하였다.

예외처리 예시(정보공개법 제9조제1항제6호)

순번	내용
1	법령에서 정하는 바에 따라 열람할 수 있는 정보
2	사생활의 비밀 또는 자유를 부당하게 침해하지 아니하는 정보
3	공개하는 것이 공익이나 개인의 권리 구제를 위하여 필요하다고 인정되는 정보
4	직무를 수행한 공무원의 성명·직위
5	국가 또는 지방자치단체가 업무의 일부를 위탁 또는 위촉한 개인의 성명·직업

예외처리 패턴화 관리 절차



지속적인 학습/분석을 위해 프로토타입 메뉴를 제작하였다. 이 메뉴는 문서별 비공개정보 확인, 관리자모드에 정규식 추가, 비공개 제6호 단어, 제2호 단어, 사전 단어 추가 및 인명 검색 기능 등을 제공할 수 있도록 구성하였다.

프로토 타입 메뉴	프로토 타입 메뉴 기능
========== 국가기록원문서 비공개정보 검색 ========== 1. 문서별 비공개정보 2. 관리자 모드 0. 종료 *관계자 외 관리자 모드 사용 금지 원하는 번호를 선택하세요(ex. 1): 1 [문서별 비공개정보] ========== 체크할 폴더 (ex. EA0000216): 체크할 파일 (ex. 127.0.0.1_0005):	▶ 연구 일정을 고려 성능 체크를 위한 파이썬으로 작성되었으며, 연구과제 결과에 따라 추후 유저 인터페이스 업그레이드 예정 ▶ 문서별 비공개정보 : 문서내 비공개정보 여부 파악을 위한 메뉴이며, 해당 메뉴 선택시 제6호 비공개 단어에 대해 정규표현식, 단어사전과 패턴검색, 그리고 Kobert 기반 인명검색 등으로 순차적으로 수행 ▶ 관리자 모드 : 관리자 모드로, 문서 내 비공개정보를 분석 처리하여 DB에 저장하기 위한 메뉴로, DB 규모에 따라 처리시간이 달라짐. DB 관리자 외 사용제한

이번 연구의 **정확도 측정**을 위하여 **연구의 결과값을 기록물 육안검수를 통해 확인**하였다. 제6호 개인정보의 경우는 업무분류체계 및 개인정보유형별로 구분하여 해당 유형명과 개수를 표시하도록 하였으며, 제6호의 정규식(이메일, 여권번호 등 10개)는 각각 세분화하여 확인하였다. 인명의 경우 별도의 유형으로 구분하여 검토 확인하였다.

기계학습과 정규식 표현 방식, 사전 등을 **하이브리드 방식으로 진행한 최종 결과값은 약 97.5%**로 나타났다. **정규표현식은 100%, 주소 DB와 인명 기계학습은 각각 97%, 70%**로 나타났다. 제6호(개인정보) 중 정규식 패턴과 단어사전+글자패턴을 이용한 검증은 연구 범위내에서 비교적 정확하게 필터링 되는 것을 확인하였다. 그러나 인명의 경우 다양한 경우의 수와 추가적인 반복학습, 알고리즘 개발과 연구가 필요한 대상으로 나타났다.

- 비공개정보를 비공개라고 인식한 정도(Recall, 재현율)을 기반으로 체크, 육안검수와 비교
- 대상 기록물 파일수 : 44,832개, 검수 파일수 : 2,730개(약 6% 샘플 검수)

비공개정보 유형		패턴	단어사전	DB	사전학습	적용	정확도	비고
6호 비공개 단어 패턴	이메일	O				정규표현식	100%	343개 비공개 정보 유형
	여권번호	O				정규표현식	100%	
	주민등록번호	O				정규표현식	100%	
	외국인등록번호	O				정규표현식	100%	
	사업자등록번호	O				정규표현식	100%	
	군번	O				정규표현식	100%	
	운전면허번호	O				정규표현식	100%	
	카드번호	O				정규표현식	100%	
	주소	O		O		정규표현식 + 주소DB	97%	
	소송번호	O				정규표현식	100%	
6호 비공개 단어 – 인명					O	기계학습 사전학습모델	70%	
6호 비공개 단어 – 단어		△△	O			단어사전 + 글자패턴	99%	
2호 비공개정보 – 단어		△△	O			단어사전 + 글자패턴	99%	87개 비공개 정보 유형
기타 비공개정보			O			단어사전	100%	
총 평균							97.5%	

4. 맺음말

이상 본 연구는 국가기록원에 소장하고 있는 전자기록물 공개재분류 추진을 위해 비공개정보 필터링 및 마스킹 연구를 추진하였고 그에 대한 내용을 소개하였다.

연구 결과를 정리해보면, 원본파일을 대상으로 텍스트 변환 후 학습데이터로 활용하는 것이 적합한 것으로 나타났다. 이는 변환 오류율이 낮았고(0.2%), 왜곡되지 않은 학습데이터는 정확도를 높이는 필수 필요 조건이기 때문이다. 기계학습 및 재분류 방안으로는, 비공개 단어들의 특성에 따라 가장 적합한 방안을 선택하여 정확도를 높일 수 있었다. 본 연구에서는 인명의 경우 기계학습 모델인 Ko-bert를 사용하였다. 왜냐하면, 비공개단어의 특성상 패턴이나 단어사전 등의 방식으로는 분류할 수 없는 방법이기 때문이다. 나머지 개인정보 제6호와 국가안전보장·국방·통일·외교의 제2호에 해당하는 비공개단어는 패턴과 단어사전을 활용하여 정확도를 높였으며, 주소의 경우 주소DB를 통해 검출하였다. 그리고 이메일, 여권번호, 주민등록번호 등은 패턴 검출 방식을 사용하여 활용하였다. 특히 소송번호의 경우 기존 상용 개인정보 필터링에 포함되지 않은 항목으로 기록원에서 새롭게 패턴방식을 적용한 사례라고 할 수 있다. 연구의 한계점도 있었다. 비공개기록물을 대상으로 하다 보니 연구 개발임에도 불구하고 장소의 제약(국가기록원 內 사업장, 인터넷 사용 불가)이 있었으며, 불편한 연구 환경에도 불구하고 연구를 진행하였다.

연구 결과에서 나타났듯이, 인명의 경우 앞으로도 다양한 경우의 수와 추가적인 반복학습이 필요하다. 샘플링 검수에서 정확성이 높게 나타나기는 했지만, 더 많은 전자기록물을 대상으로 추가적인 내용과 비공개정보에 대한 유형 분석이 필요하다. 또한 제외처리에 대한 반복학습 및

알고리즘 개발과 개인정보 처리를 위한 구체적인 연구 개발도 필요하다. 또한 마스킹 처리 구현과 실험도 향후 추가적으로 제고되어야 할 부분이다. 이러한 부분은 앞으로 지속적인 연구를 통해 해결되리라는 바램을 여지로 남겨놓고자 한다.

짧은 기간(7개월) 자료 분석과 텍스트 변환 검증, 유형에 따른 모델 설계 및 샘플링 검수를 진행하면서 미시적으로는 전자기록물의 특징을 반영한 대량 기록물 대상 공개재분류의 효율적 방안을 찾고 싶었고, 거시적으로는 이 연구가 국가기록원의 지능형 전자기록관리 발전을 위한 한 걸음이 되기를 기대하였다. 또한 향후 현장에서 동일한 어려움을 겪고 있는 기록관에 도움이 될 수 있을 거라는 희망을 가져본다.

앞으로 국가기록원이 소장하고 있는 전자기록물(빅데이터), 공개재분류 이력과 기준서(딥러닝), 지속적인 기계학습으로 공개재분류 업무의 향상은 물론 디지털 혁신 기반의 전자기록관리의 실현이 될 수 있기를 기대해 본다.

참고문헌

권미현, 「비공개기록물 공개재분류 업무절차 개선방안」, 제8차 기록관리 연구세미나 발표자료, 2019. 10. 16.

임희연, 「기록관에서의 공개재분류 제도 개선 방안-서울특별시교육청 사례 중심」, 『기록학연구』 제49호, 한국기록학회, 2016. 7.

국가기록원, 「2019년 전자기록물 인수결과보고」, 2020. 2.

국가기록원, 「2020년 비공개 기록물 공개재분류 추진 계획」, 2020. 1., 3쪽.

국가기록원, 『2020년도 국가기록원 주요통계연보』, 2020. 6., 16.국가기록원 열람.

국가기록원, 「지능형 전자기록관리 기술연구 개발 기획연구」, 2019. 8.

국가기록원, 「2019년 비공개기록물 공개재분류 추진 결과보고」, 2019. 12., 23쪽.

국가기록원, 기록관리 공공표준, NAK 16-2 2013(v1.0) 「기록물 공개관리 업무-제2부:영구기록물관리기관(v1.0)」, 2013.

㈜아카이브랩, 「국가기록원 차세대 기록관리 모델 재설계 연구 개발」, 『정보관리학회지』, 34(4), 321-344쪽.

관세청, 「공공데이터 개방범위와 개인정보 등 비공개 정보 비식별화 방안 연구」, 2019., www.prism.go.kr

법원행정처, 「2019년 지능형 비식별환 솔루션 도입(2단계) 사업」, www.g2b.go.kr

개인정보보호포털, www.privacy.go.kr

Google, 「Pre-training of Deep Bidirectional Transformers for Language Understanding」, 2018. 10. 11., arXiv.org

주제발표 3

국립중앙도서관의 온라인 서비스와 개인정보보호

최 윤 경(국립중앙도서관)

국립중앙도서관의 온라인 서비스와 개인정보보호

최 윤 경(국립중앙도서관)

차 례

1. 도서관과 개인정보보호

도서관에서 개인정보보호는 이용자의 지적자유 보장과 밀접하게 관련되어 있다. 1939년 ALA(미국도서관협회, American Library Associaton) Council의 '도서관 권리선언(Library Bill of Rights)'[15] 이후 1999년 국제도서관협회연맹(IFLA)이 '도서관과 지적자유에 관한 성명'을 발표하였다. 여기에서는 도서관들은 지적 자유를 수호하고 개인의 표현과 사상의 자유를 비롯한 국민의 알권리를 보장해야 하며, 이를 위해 도서관에 대한 검열 반대는 물론 이용자의 개인정보 보호가 전제되어야 한다고 표명하였다(노영희, 2012)[16].

우리나라도 1980년대부터 개인정보보호에 대한 논의들이 진행되었으며, 1994년 「공공기관의 개인정보보호에 관한 법률」(이하 '개인정보보호법')이 제정되면서 국내 도서관계에서도 개인정보보호의 중요성이 부각되었다. 1997년 한국도서관협회에서도 '도서관인 윤리선언'을 제정

15) 도서관 권리선언은 1944년, 1948년, 1961년, 1967년, 1980년에 차례로 개정되었으며, 최신 개정은 2016년에 이뤄졌다.
16) 노영희. 2012. 도서관의 개인정보보호정책 개발 및 제안에 관한 경우. 한국문헌정보학회지, 46(4): 204-242.

하였으며, 2019년 개정한 바 있다. 최근 개정된 선언문에서는 도서관인이 국민의 알권리를 보장하는 사회적 책무를 갖고 있으며, “도서관인은 도서관서비스 과정에서 수집되는 이용자의 프라이버시와 개인정보를 적극 보호한다.”는 윤리 지표를 제시하였다.

이외에 「도서관법」에서도 개인정보보호에 대한 조항들이 마련되어 있다. 제1조에서는 도서관의 사회적 책임으로 국민의 정보 접근권과 알권리에 대한 보장, 제8조(이용자의 개인정보보호)에서는 도서관이 이용자의 정보보호에 대한 의무와 책임이 명시되어 있다. 또한 제20조의2(온라인 자료의 수집)와 동법 시행령 제13조5(개인정보의 정정 및 삭제 청구)를 통해 국립중앙 도서관에서 수집한 온라인 자료에 개인정보가 포함된 사실을 알게 된 경우 국립중앙 도서관장 에게 해당 정보의 정정 또는 삭제 등을 청구할 수 있고, 국립중앙도서관은 10일 이내 필요한 조치를 취한다는 내용이 기술되어 있다.

도서관의 개인정보보호정책을 다룬 최근 연구로는 노영희(2012), 노영희와 김태경(2015)[17]의 연구가 대표적이었다. 2012년 연구는 국내 도서관의 개인정보보호 정책의 필요성과 가이드라인 개발 방향을 제시하였고, 2015년에는 도서관 관종과 관계없이 공통적으로 적용 가능한 개인정보 가이드라인을 제안하였다. 여기에서는 공공기관으로서의 일반적인 개인정보보호 뿐만 아니라 도서관 업무와 서비스 수행에서 필요한 지침들도 함께 제시했다는 점에 의의가 있었다.

한편, 최근 4차산업혁명, 빅데이터, 인공지능(AI) 등 변화하는 기술과 정보환경에서 개인정보보호의 중요성이 더욱 커지고 있다. 그와 더불어 공공과 민간 영역에서 혁신적인 고객 서비스를 개발하고 고안하기 위해 개인 데이터를 활용하고 있으며, 개인정보보호정책도 개인정보의 활용과 보호를 모두 강화하는 방향으로 변화하고 있다. 이러한 개인정보보호 정책 환경의 변화에 따라 국립중앙도서관은 공공기관이자 국가를 대표하는 도서관으로서 각종 서비스에 개인정보보호정책을 반영해왔다.

특히 코로나19로 인한 도서관의 임시휴관 및 제한적 운영이 계속되면서 온라인 서비스를 더욱 강화할 계획이며, 서비스의 개선과 새로운 서비스 개발을 위해 이용자 데이터에 대한 활용을 적극적으로 검토하고 있다. 대표적인 서비스로는 국립중앙도서관의 대표 누리집을 통한 서비스와 온라인 자료에 대한 열람 서비스를 꼽을 수 있다. 그리고 이용자의 독서성향과 도서관 이용행태를 기반으로 하는 개인별 맞춤 서비스도 계획하고 있다.

본 고에서는 국립중앙도서관을 비롯한 외국 국가도서관의 개인정보보호 정책과 온라인 서비스에 대한 적용 현황을 살펴보았다. 그리고 온라인 서비스의 발전 방향과 향후 개인정보보호정책에 대한 개선 방향에 대해 제안하였다.

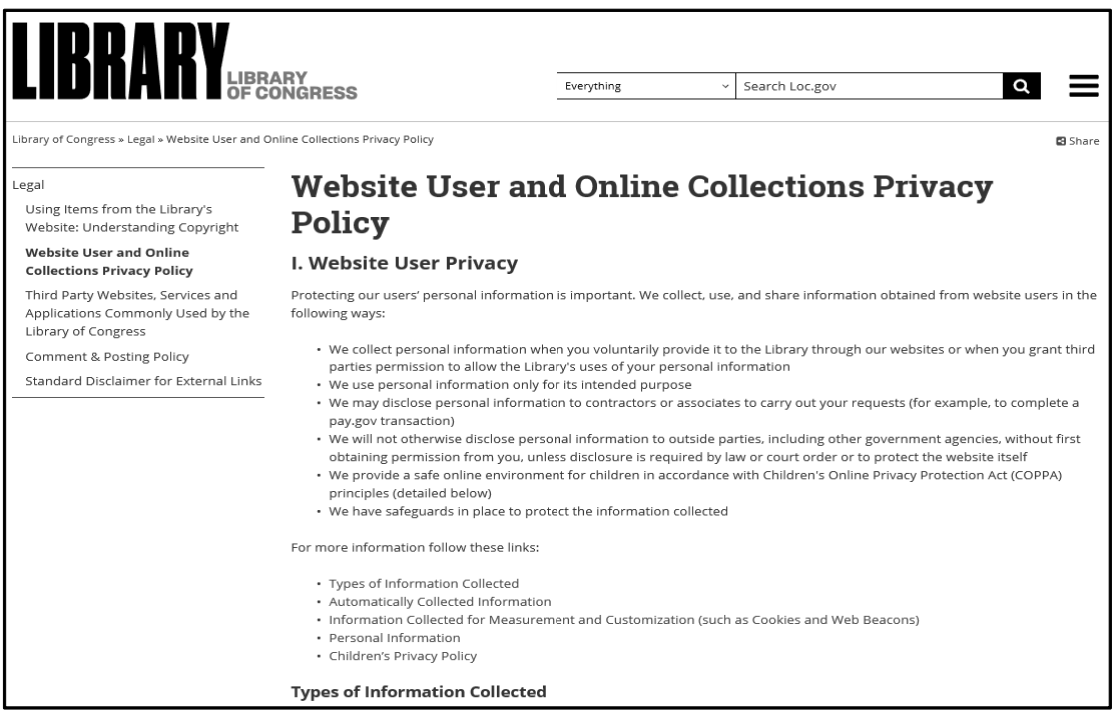
17) 노영희, 김태경. 2015. 도서관 개인정보보호 가이드라인 개발에 관한 연구. 정보관리학회지, 32(2) : 25-61.

2. 외국 국가도서관 사례

1) 미국의회도서관

미국의회도서관의 개인정보정책은 웹사이트 이용자와 온라인 컬렉션으로 구분하여 제시하고 있다. 웹사이트 이용자에 대한 개인정보정책에는 수집 정보의 유형, 자동 수집 정보, 서비스 평가나 개인화를 위해 수집된 정보, 개인정보 취급, 어린이를 위한 개인정보정책 등이 포함되어 있다. 여기에서는 이용자 정보에 대한 관리가 도서관에서 매우 중요한 역할임을 선언하면서, 개인화 서비스를 목적으로 자동화 도구들을 통해 수집되는 정보들을 구체적으로 나열함으로써 개인정보의 활용 측면도 강조하고 있다.

또한 수집된 정보들은 개인을 식별하기 위한 용도로는 사용하지 않되 이용자 분석이나 웹사이트 개선을 위한 용도로 활용할 수 있다는 점을 명시하였다. 이외에도 미국의회도서관에서 제공하는 방대한 온라인 컬렉션 일부에 개인정보가 포함될 수 있지만, 이용자에게 개인정보 보호와 저작권 등의 법적 책임을 준수하는 것을 전제로 자유로운 정보접근권을 제공한다고 하였다(Library of Congress, 'Website User and Online Collections Privacy Policy', 〈그림 1〉 참조)[18].


LIBRARY LIBRARY OF CONGRESS

Everything | Search Loc.gov

Library of Congress » Legal » Website User and Online Collections Privacy Policy

Share

Legal
- Using Items from the Library's Website: Understanding Copyright
- **Website User and Online Collections Privacy Policy**
- Third Party Websites, Services and Applications Commonly Used by the Library of Congress
- Comment & Posting Policy
- Standard Disclaimer for External Links

Website User and Online Collections Privacy Policy

I. Website User Privacy

Protecting our users' personal information is important. We collect, use, and share information obtained from website users in the following ways:

- We collect personal information when you voluntarily provide it to the Library through our websites or when you grant third parties permission to allow the Library's uses of your personal information
- We use personal information only for its intended purpose
- We may disclose personal information to contractors or associates to carry out your requests (for example, to complete a pay.gov transaction)
- We will not otherwise disclose personal information to outside parties, including other government agencies, without first obtaining permission from you, unless disclosure is required by law or court order or to protect the website itself
- We provide a safe online environment for children in accordance with Children's Online Privacy Protection Act (COPPA) principles (detailed below)
- We have safeguards in place to protect the information collected

For more information follow these links:

- Types of Information Collected
- Automatically Collected Information
- Information Collected for Measurement and Customization (such as Cookies and Web Beacons)
- Personal Information
- Children's Privacy Policy

Types of Information Collected

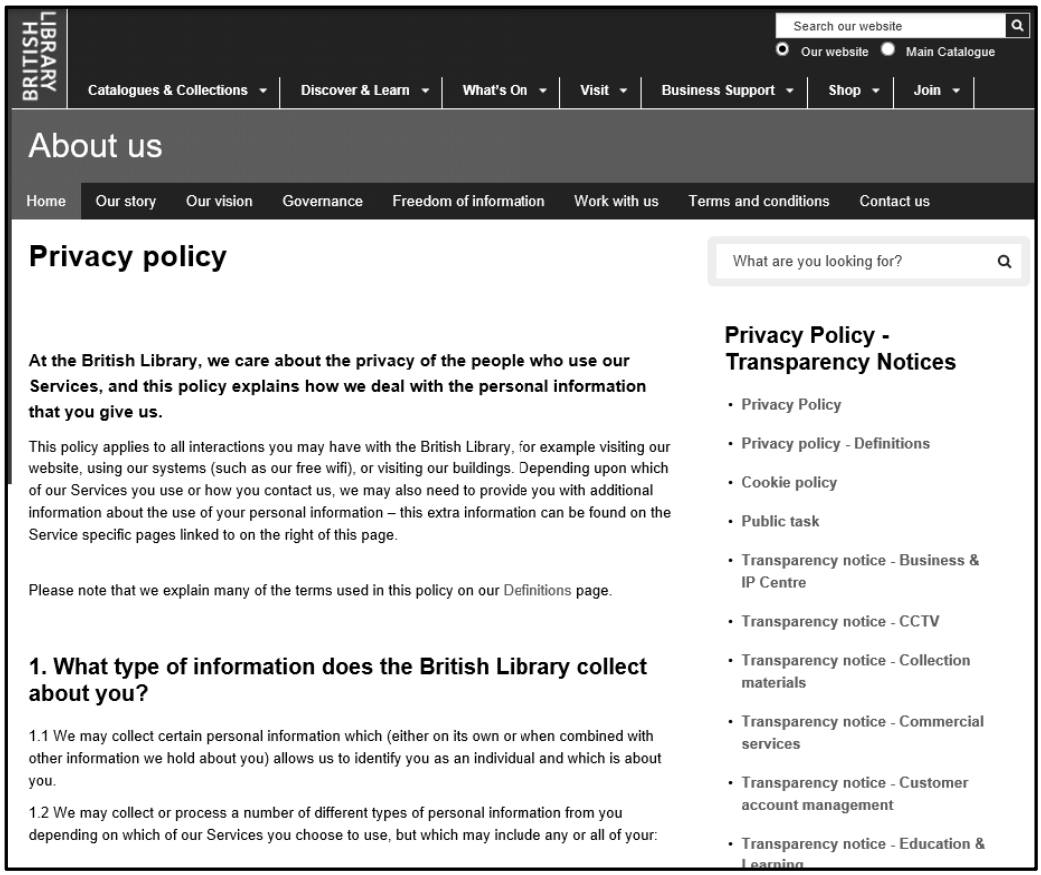
〈그림 1〉 미국의회도서관의 개인정보정책 웹사이트

18) Library of Congress. Website User and Online Collections Privacy Policy. https://www.loc.gov/legal/privacy-policy/

2) 영국국가도서관

영국국가도서관은 '정보접근의 자유'와 더불어 미국보다 더 상세하게 개인정보보호정책이 기술되어 있었다. 또한 일반적인 규정이나 규칙에서 나타난 표현보다는 질의응답 방식으로 이용자에게 친절하고 친숙하게 안내하고 있다는 점이 특징적이었다. 구성 내용은 수집 정보의 유형, 수집 방법, 개인정보 활용, 비개인화 및 익명 처리 방식, 제3자와의 공유, 개인정보의 관리, 이용자의 권리 등이었다(British Library, 'Privacy policy'[19], 〈그림 2〉 참조).

영국국가도서관에서 수집하는 개인정보의 유형은 매우 다양했는데, 일반적인 이름이나 주소, 이메일, 연령, 생일, 성별 이외에도 사회관계망서비스 ID, 사용자 의견이나 피드백, 제안, 도서관 누리집에서의 브라우징 행태(이력). 도서관 방문 이력 및 행태 등이 포함되었다. 또한 비개인화 및 익명처리를 통해 이용자 데이터의 심층적인 분석과 다른 정보와의 결합이 가능하도록 명시하고 있다.



〈그림2〉 영국국립도서관의 개인정보보호정책 사이트

19) British Library. Privacy policy. https://www.bl.uk/about-us/privacy-policy

이외에도 투명성 안내(Transparency notices)라는 메뉴를 통해 CCTV, 이용자 계정 및 쿠키 관리, 상용 서비스 등 도서관 서비스에서 개인정보의 취급이나 관리, 활용과 세부 지침을 국민에게 공개하고 있다. 일례로 '도서관 장서에 관한 투명성 안내'에서는 도서관 소장자료에는 다양한 유형의 개인정보가 포함될 수 있다는 점을 안내하였다. 만약 민감 정보가 발견되면 공익을 위한 보존의 목적으로 처리되며, 정보 공개 시 정보 주체에 피해를 줄 가능성이 있다면 공개 접근을 불허하되 미래 세대를 위해 영구적인 보존은 유지한다고 명시되어 있다[20].

3) 일본국립국회도서관

일본국립국회도서관도 누리집을 통해 기관 및 누리집의 서비스 운영을 위한 개인정보보호정책을 공개하고 있다[21]. 내용와 항목 구성은 다른 국가도서관들과 매우 유사하였지만, 도서관 전체 운영에 관한 정책은 다소 차이가 있다.

특징적인 점은 서지 서비스에 대한 개인정보보호 취급 규정이 별도로 마련되어 있다는 것이다[22]. 자료에 대한 목록정보(서지데이터)를 작성할 때 저자명, 출생년 등 개인정보가 입력되는데 서지 작성 대상 자료나 발간된 인명사전, 공공기관이나 다른 국립도서관에서 제공하는 데이터베이스에서 참조한 정보에 대해서는 개인정보 활용 동의가 필요하지 않지만, 이외의 방식으로 수집한 개인정보를 서지데이터에 활용하는 경우는 본인의 동의를 받도록 규정하고 있다. 그러나 온라인 자료나 도서관 장서에 대한 개인정보보호정책은 별도로 제시하지 않았다.

3.국립중앙도서관의 개인정보보호 적용 현황

1) 도서관 이용자에 대한 개인정보보호

국립중앙도서관의 이용자에 대한 개인정보보호 정책은 대표누리집 내 '개인정보처리방침'[23]과 '영상운영처리기기 운영·관리방침'[24]에서 확인할 수 있다. 개인정보처리방침을 중심으로 살펴

20) British Library. Transparency notice - Collection materials. https://www.bl.uk/about-us/privacy-policy/transparency-notice-collection-materials
21) 國立國會図書館. プライバシーポリシー. https://www.ndl.go.jp/jp/privacypolicy/index.html, 國立國會図書館. の個人情報の取扱いについて. https://www.ndl.go.jp/jp/privacy/index.html
22) 國立國會図書館. 当館が作成する書誌データとその修正について. https://www.ndl.go.jp/jp/data/basic_policy/policy/personal.html
23) 국립중앙도서관 개인정보방침. https://www.nl.go.kr/NL/contents/N70101000000.do
24) 국립중앙도서관 영상정보처리기기 운영·관리방침. https://www.nl.go.kr/NL/contents/N70102000000.do

보면 총 13개 부문으로 개인정보의 처리목적, 처리 및 보유기간, 제3자 제공, 개인정보처리의 위탁, 처리하는 개인정보의 항목 등으로 구성된다. 국회도서관이나 서울도서관 등 국내 도서관들도 유사한 형식을 사용하는데 이는 대부분 공공 기관에서 개인정보보호위원회의 '개인정보 처리방침 작성 예시'를 준용하면서 개별 도서관의 상황을 반영하여 작성하기 때문이다.

개인정보보호방침에서는 국립중앙도서관의 현황을 기초로 작성되었는데, 현재 도서관에서 제공하는 20개의 서비스에 대해 각각의 운영근거와 처리목적, 처리 및 보유기간, 처리하는 개인항목을 구체적으로 나열하고 있다. 위의 미국과 영국 사례와 비교해볼 때 기본적인 조항들과 내용들은 유사하였다. 그러나 기술된 내용이 상대적으로 설명적이기보단 간략하였고, 웹사이트 중심의 개인정보정책 위주로 작성되었다. 도서관 장서나 어린이 서비스 등 도서관만이 가진 고유한 임무나 특성에 기반한 내용은 없었지만, 개인정보 수집 범위에서는 외국 사례와는 달리 필수와 선택 사항을 명확하게 구분하여 제시하고 있었다.

아래 〈표 1〉은 국립중앙도서관 회원정보와 관련된 통합회원DB에서 정의한 개인정보 수집 항목인데, '도서관 이용자 관리 및 맞춤서비스 제공'을 목적으로 마지막 이용일로부터 5년까지 수집·활용된다. 그런데 앞서 기술한 바와 같이 공공기관을 위한 공통 형식을 적용하다 보니 내용적인 측면에서 도서관의 고유성이 나타나기보다는 공공기관의 특성 위주로 나타나고 있었다.

〈표 36〉 국립중앙도서관의 개인정보 수집 범위(통합회원DB)

구분	수집 항목
필수	ID, 비밀번호, 이름, 이메일, 전화번호(또는 핸드폰번호), 생년월일, 주소, 법정대리인 이름 (만 14세 미만 회원의 경우)
선택	성별, SMS통보여부, 소속기관, 이용목적, 관심분야, 메일링서비스 수신여부, 학교정보, 책바다 소속도서관, 책바다 소속도서관 이용증 번호 또는 이용자 ID

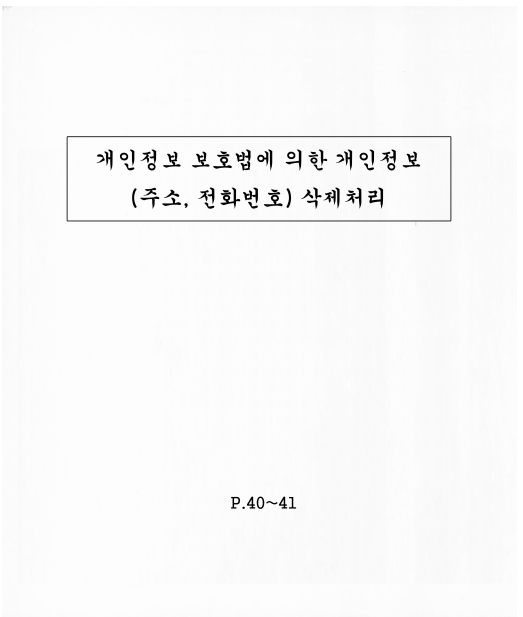
2) 온라인 콘텐츠에서의 개인정보보호

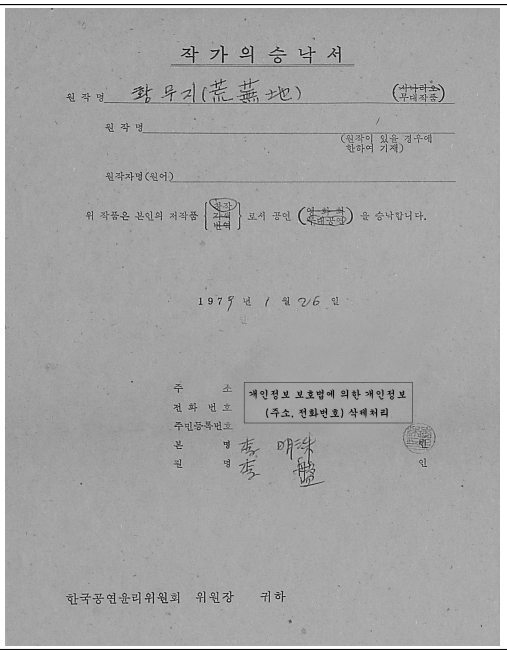
국립중앙도서관은 국가대표도서관으로서 국내에서 발행된 온·오프라인 자료를 수집할 책무를 가진다. 온·오프라인 자료에 개인정보가 포함되어 비공개 요청이 들어오면 해당 자료에 대한 검색 및 열람을 제한시키고 있다. 수집된 자료는 이용자에게 자유롭게 이용되는 것이 원칙이지만, 개인정보에 포함된 자료에 대해서는 정보공개법(「공공기관의 정보공개에 관한 법률」) 제9조에 해당하거나 저자와 발행자가 비공개를 요청한다면 「국립중앙도서관 자료

이용제한 처리지침(규정 제530호)」 제3조(이용제한 자료의 종류)와 제5조(이용제한의 범위)에 따라 자료의 열람이나 검색을 제한할 수 있다. 먼저 서비스를 하고 사후에 이용 제한을 하는 이유는 현실적으로 자료량이 방대해 일일이 개인정보의 포함 여부를 확인하기 어려운 점도 있지만, 국민의 정보 접근성과 알권리를 보장해야 한다는 측면에서 특별한 요청이 없다면 최대한 자료 이용을 보장하는 것이 도서관의 임무이기 때문이다.

한편, 국립중앙도서관에서 디지털화로 구축한 온라인 자료에 대해서는 개인정보보호를 다소 다르게 처리하고 있다. '회원명부'와 같이 개인정보가 전체 또는 다수 포함된 자료로 예상이 가능한 경우는 아예 디지털화를 하지 않거나, 하더라도 자료는 보존하되 열람은 제한하고 있다. 그러나 대다수의 자료는 일부분만 개인정보가 포함된 경우가 많아 해당 부분만 개인정보 보호 처리를 한 후 온라인으로 원문을 제공하고 있다. 이는 기계적인 방식으로 특정 부분에 대한 개인정보 보호 처리가 더 쉽고 용이한 온라인 자료의 강점을 살려 이용자의 정보 접근성을 최대화하고자 한 것이다.

아래는 국립중앙도서관에서 온라인 자료에 대한 개인정보보호 취급 사례이다.

<table>
<tr>
<td> 
</td>
<td>· 연속적으로 여러 페이지에 개인정보가 나오는 경우 '개인정보 보호법에 의한 개인정보(주소, 전화번호) 삭제처리' 문구를 전체 면에 삽입</td>
</tr>
</table>



작 가 의 승 낙 서

원 작 명 황무지(荒蕪地) (생산학술)

원 작 명 (원작이 있을 경우에 한하여 기재)

원작자명(원어)

위 작품은 본인의 저작물 [원작 / 번역] 로서 공인 (생산학술) 을 승낙합니다.

1975 년 1 월 26 일

주 소 개인정보 보호법에 의한 개인정보 (주소, 전화번호) 삭제처리

전 화 번 호

주민등록번호

본 명 李 明 [illegible]

필 명 李 [illegible] 인

한국공연윤리위원회 위원장 귀하

- 개인자택 주소, 전화번호, 주민등록번호에 해당하는 개인정보가 본문 내용에 일부 표기된 경우 해당하는 개인정보 부분만 선택하여 처리

표준작성 공헌자

표준 번호 : TTAK.KO-01.0145

이 표준의 제·개정 및 발간을 위해 아래와 같이 여러분들이 공헌하였습니다.

구분	성명	위원회 및 직위	연락처 (Tel, E-mail)	소속사
과제 제안	이종화	VoIP 프로젝트그룹 의장	042-860-5278 jhyiee@etri.re.kr	한국전자통신연구원
표준 초안 제출	이종화	VoIP 프로젝트그룹 의장	042-860-5278 jhyiee@etri.re.kr	한국전자통신연구원
	정옥조	VoIP 프로젝트그룹 위원	042-860-1587 lij1024@etri.re.kr	한국전자통신연구원
	장용선	-	02-2131-0221 jangys@nia.or.kr	한국정보사회진흥원
	최창호	-	02-2131-0782 chchoi@nia.or.kr	한국정보사회진흥원
표준안 심의	민경선	전송통신 기술위원회 의장	042-524-0100 개인정보 보호법에 의한 개인정보 (주소, 전화번호) 삭제처리	KT
		외 기술위원회 위원		
사무국 담당	박정식	팀장	031-724-0080 jspark@tta.or.kr	TTA
	김영재	선임	032-724-0195 yjkim@tta.or.kr	TTA

- 기관주소, 회사메일 등 개인의 직접적인 정보가 아닌 경우 보호처리 제외
- 포털에서 제공하는 개인메일 또는 개인 홈페이지는 개인정보 처리

이 력 서

성 명 : 진 근 찬 (陳 根 贊)

생년월일 : 1946년 8월 21일

출 생 지 : 경기도

본 적 : 개인정보 보호법에 의한 개인정보(주소, 전화번호) 삭제처리

학 력

1964 ~ 1971 한양대학교 공과대학 기계공학과 (B.S)

1975 ~ 1977 한국과학원 재료공학과 (M.S)

1979 ~ 1982 한국과학기술원 재료공학과 (Ph.D)

경 력

1971.1 ~ 1972.5 이천전기공업 (주)

1972.5 ~ 1982 현재 홍능기계공업회사 (선임연구원)

- 개인정보가 본문내용에 포함된 경우, 해당하는 개인정보 부분만 보호처리
- 생년월일, 특정 지역명은 개인정보로 보지 않고 보호처리 제외

○ 옥타브 컨소시엄

KT		
1	호처리연동장치 정보	- (I-CSCF) 224.114.39.227/5060 --> 타 사업자가 접속해야 할 서버 주소 - (S-CSCF) 222.114.39.227/5070
2	미디어서버 정보	- ip address : SDP 협상을 통해 오픈됨 - 처리 가능한 미디어 : 음성, 영상 - 지원코덱 : G.711, H.263 QCIF, CIF, jpeg
3	연동 시험용 단말 정보	- 참여단말 : SoIP 영상 단말 * LG Sandwich LN101-806D * 삼성 Ginny 101-806DW - 지원미디어 : 영상, 음성, 텍스트, 이미지 * 영상통화 : H.264, G.722 우선 협상 협상순서:(H.264, H.263, MPEG-4), * 음성통화 : G.722 우선협상 협상순서 : (G.722, G.711a, G.711u, G.729) - 번호 : 07077006000 ~ 07077006002 * sip:07077006000@octave.com, tel:07077006000 - domain : octave.com
4	컨퍼런스 URI	- 07077006100 ~ 07077006101 * sip:07077006100@octave.com, tel:07077006100
5	연동 시험 담당자 정보	- KT: 이성희, 042-870-8042, 개인정보 보호법에 의한 개인정보 (주소, 전화번호) 삭제처리 이운영, 042-870-8317

○ 케이블 컨소시엄

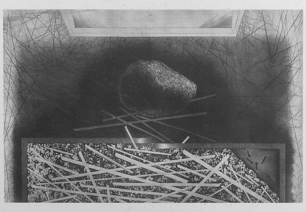
케이블 컨소시엄	
연동 시험용 단말 정보	- Capability : INVITE, CANCEL, ACK, BYE, OPTIONS - Audio Codec : G.711(a-law, m-law), G.723.1 - Video Codec : H.263 CIF (352 x 288 pixels, 15~20 fps) H.263 QCIF (176 x 144 pixels, 30 fps)
연동 시험 담당자 정보	- KLabs 한재천 선임 (02-300-3482, 개인정보 보호법에 의한 개인정보 (주소, 전화번호) 삭제처리, jupiter@klabs.re.kr) - 티브로드 김희정 과장 (okay@tbroad.com)
기타사항	- 타 컨소시엄과 영상전화 연동시험을 위해 소프트스위치 주소 및 단말 정보 제공 필요 - 영상통화 시험은 정상적인 호 설정, 해지, BYE 의 정상적인 처리 등을 시험하기로 함

· 기관 또는 회사의 전화(내선번호)에 해당하지 않고 개인 또는 자택전화번호가 노출된 경우 부분적으로 보호처리

· 원칙적으로 개인정보가 노출된 위치에 처리

한 성 희
HAN, Soung-Hee

1950년생
홍익대학교 미술대학 및 동대학원 졸업.
부산미술대전 대상, 공간국제판화공모전 가작상 및 우수상 수상,
부산현대판화전, 한국현대판화가협회전, 가나기와 국제판화앙당팡당전 등에 참여
동판화 개인전 3회, 현재 부산대, 동아대, 부산여대 출강중
개인정보 보호법에 의한 개인정보(주소, 전화번호) 삭제처리
출품작 〈Image of Form" 87-Ⅰ, Drypoint, Etching, Aquatint, 80×55㎝, 1987〉



· 개인정보가 노출된 위치와 본문 내 글자 크기를 기준으로 보호처리

ncia NIA

구분	내용	업체	담당자	연락처
서버	IBM P550	트라이얼 정보통신	김재길 부장	개인정보 보호법에 의한 개인정보 (주소, 전화번호) 삭제처리
WEB/WAS	TMAX JEUS TMAX WebToB	TMAX	박현수 과장	
보안솔루션	Key# BIZ	루멘소프트	전형석 실장	

다. 각 IPTV사 및 개발 협력사

구분	업체	담당자	연락처
KT	KT	이성재 책임	개인정보 보호법에 의한 개인정보 (주소, 전화번호) 삭제처리
	람즈	최현우 과장	
SKB	SKB	유원택 부장	
	유라클	김용찬 책임	
LGT	LGT	정유식 책임	
	유라클	김근호 과장	

· 개인정보 내용과 개인정보가 아닌 내용이 동시에 노출되고 개인정보로 분류되는 부분이 다수인 경우 일괄적으로 보호처리

4. 서비스 활성화를 위한 발전 방향

1) 개인화 서비스 개발을 위한 개인정보 수집 확대

국립중앙도서관은 코로나19로 인한 비대면 사회에 이용자의 창작·학술·창업 활동을 안정적으로 지원하기 위해 온라인 서비스를 강화하고 있다. 특히 인공지능이나 빅데이터 등 지능형 기술을 접목하여 누리집이나 모바일을 통해 도서관 자료나 프로그램에 대한 맞춤형 추천 서비스를 제공할 계획이다(〈그림 3〉 참조).

〈그림 32〉 이용자 맞춤형 자료 추천 서비스(안)

지금까지 도서관에서는 이용자가 미리 등록한 관심 키워드나 주제어를 바탕으로 신착자료나 학술지 메일링 리스트를 제공하고 있다. 그러나 이미 이용자가 선택하기 전에 볼만한 콘텐츠나 서비스를 먼저 추천하는 민간 서비스와 비교하자면 추천 결과에 대한 만족도나 실효성이 현저히 떨어질 수밖에 없다. 따라서 맞춤형 추천 서비스의 만족도와 정확성을 높이려면 이용자의 취향이나 독서성향이나 도서관 시설이나 자료에 대한 이용 이력 등 다양하고 방대한 이용자 데이터가 수집·분석되어야만 한다.

그러나 현행 개인정보처리방침으로는 이용자 데이터를 수집하는 데에 한계가 있다. 따라서 이용자의 자료나 시설 이용 이력, 위치 데이터, 교육이나 전시 프로그램 등 서비스 활용 이력 등을 수집할 수 있도록 개인정보 수집 범위를 확대하는 방안을 논의하고 있다. 공공기관에서 개인정보를 추가적으로 수집하는 것은 가능하지만, 변경 사항에 대해서는 개인정보보호위원회에 개인정보 침해요인 평가를 요청할 필요가 있다.

2) 개인정보처리방침에 도서관 특수성 반영

국립중앙도서관을 비롯한 대다수 도서관들이 공개하는 개인정보처리방침은 형식과 내용이 유사하고, 공공기관을 위한 표준화된 서식을 적용하다 보니 도서관의 임무나 서비스에 대한 특색이 드러나지 않았다. 외국 국가도서관의 경우는 국가 차원의 개인정보처리정책을 준용하고는 있으나, 도서관 업무와 서비스뿐만 아니라 도서관으로서의 정체성이 정책에서도 나타나고 있다.

도서관이 소장하고 있는 자료 중에는 개인정보가 포함된 자료도 많을 것이다. 그러나 도서관은 이용자에게 자유로운 정보 접근권을 제공하는 차원에서 정보주체의 특별한 요청이 없는 한 소장자료를 제공하는 것을 원칙으로 한다. 따라서 도서관이 개인의 정보를 보호하면서 국민의 지적 자유권을 보장한다는 근본적인 임무와 책임을 명확하게 안내하는 방안을 검토하고 있다.

3) 온라인 자료 제공 확대를 위한 기술 개발

현재 국립중앙도서관에서는 개인정보가 노출되는 부분에 수작업으로 이미지를 삽입하여 온라인 자료를 처리하고 있다. 수집된 자료와 앞으로 수집될 막대한 양의 온라인 자료에 대해 개인정보를 일일이 확인하는 것은 불가능하다. 이를 보완하기 위해서는 온라인 자료에 노출될 수 있는 개인정보의 패턴을 인식하여 자동으로 마스킹할 수 있는 기술도 도입할 필요가 있다. 그러나 현대간행물이 이외에 고문헌이나 근대자료는 본문이 세로로 쓰여있거나 한자가 많아 기계적으로 패턴을 읽기가 어려워 개인정보의 노출 위험은 여전히 남아있다.

4) 도서관과 기록관 등 유관기관과의 협력 강화

2020년 2월 IFLA와 ICA(세계기록관리협의회, International Council on Archives)는 한층 강화된 개인정보보호법을 지지하면서도 도서관과 기록관에서의 자료 수집과 보존 활동을 비롯하여 이용자의 정보접근성을 보장해야한다는 내용으로 공동 성명[25]을 발표하였다. 성명서에서는 개인정보보호법에 대해 1) 도서관과 기록관 같은 전문기관들이 개인식별정보를 담은 자료를 수집하고 보존할 수 있도록 예외를 허용하고, 2) 법적으로 자료의 파기나 삭제가 불허하다는 점을 보장해야한다는 점과 3) 개인사생활이나 기밀성, 문화적 민감 요소를 보호해야하는 경우에만 자유로운 접근을 제한하고, 4) 도서관과 기록관은 개인정보가 수록된 자료를 보존하는 것에 대한 법적 면책 특권이 있어야 한다고 하였다.

우라나라도 올해 데이터 3법[26] 법제화를 통해 가명정보 정의 및 처리·활용에 대한 법적

25) IFLA-ICA Statement on Privacy Legislation and Archiving. https://www.ifla.org/publications/node/92939

근거를 마련하고, 개인정보처리자의 안정성 확보조치 및 배상책임 의무화 등 책임성을 강화하였다. 그러나 「개인정보보호법」에서도 도서관과 기록관과 같이 방대한 자료와 기록을 보유한 기관에 대한 예외 조항은 없다. ALA의 '도서관 권리선언(2019)' 제4조를 통해 '도서관은 표현의 자유와 아이디어에 대한 자유로운 접근을 거부하는 데 관련된 모든 개인 및 그룹과 협력해야한다'고 표명한 것처럼, 국립중앙도서관과 국가기록원도 공공기관이지만 자료를 수집·보존하고 국민의 지적 자유와 알권리를 보장해야 하는 기관으로서, 개인정보보호 정책으로 인해 인류의 지적 문화 유산의 수집과 보존이 제한되거나 폐기를 요청받는 상황을 방지할 수 있도록 양 기관이 협력하여 대응할 필요가 있다.

26) 개인정보보호법, 정보통신망 이용촉진 및 정보보호 등에 관한 법률, 신용정보의 이용 및 보호에 관한 법률

주제발표 4

개인정보 가명처리 정책동향

박 윤 식(한국인터넷진흥원)

개인정보 가명처리 정책동향

박 윤 식(한국인터넷진흥원)

차 례

1. 머리말

가명정보 또는 개인정보의 가명처리 관련 정책동향을 거론하기 위해서는 개인정보에 대해 이해하는 것이 우선적으로 필요하다. 「개인정보 보호법」에서는 “개인정보”란 “살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보”라고 정의하고 있다. 여기에 더 나아가서 “해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보. 이 경우 쉽게 결합할 수 있는지 여부는 다른 정보의 입수 가능성 등 개인을 알아보는 데 소요되는 시간, 비용, 기술 등을 합리적으로 고려하여야 한다”고도 정의하고 있다.

통상의 개인정보는 그 자체가 보호의 대상이거나, 해당 주체의 동의에 의해서만 수집·이용이 가능한 정보였다. 하지만, 언론보도를 통해 많이 접하듯이 데이터3법(개인정보 보호법, 정보통신망 이용촉진 및 정보보호 등에 관한 법률, 신용 정보의 이용 및 보호에 관한 법률)의 개정으로 개인정보의 새로운 개념인 “가명정보”가 탄생하게 되었고, 개인정보에 대한 정의에 “가명정보”가 추가되었다. “가명정보”란 “개인정보의 일부를 삭제하거나 일부 또는 전부를

대체하는 등의 방법으로 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보"라는 정의가 개인정보의 정의에 추가되었다. 개인정보를 보호하고자 하는 기존의 입장은 유지하면서도, 법에서 허용한 목적(통계작성, 과학적 연구, 공익적 기록보존 등)을 수행하기 위해서는 정보주체의 동의 없이 가명처리하여 활용할 수 있는 길이 열린 것이다.

〈 「개인정보 보호법」 상 허용된 가명처리의 목적 〉

[개인정보보호법(2020.2.4. 개정, 2020.8.5. 시행)]

제28조의2(가명정보의 처리 등) ① 개인정보처리자는 통계작성, 과학적 연구, 공익적 기록보존 등을 위하여 정보주체의 동의 없이 가명정보를 처리할 수 있다.

[가명정보 처리 가이드라인(2020.9.24.)]

개인정보처리자는 정당한 처리 범위 내에서 통계작성, 과학적 연구, 공익적 기록보존 등의 목적으로 정보주체의 동의 없이 가명정보를 처리할 수 있음

가. 통계작성 : 통계란 특정 집단이나 대상 등에 관하여 작성한 수량적인 정보를 의미

나. 과학적 연구 : 과학적 연구는 기술의 개발과 실증, 기초연구, 응용연구 및 민간 투자 연구 등 과학적 방법을 적용하는 연구를 의미

다. 공익적 기록보존 : 공공의 이익을 위하여 지속적으로 열람할 가치가 있는 정보를 기록하여 보존하는 것을 의미

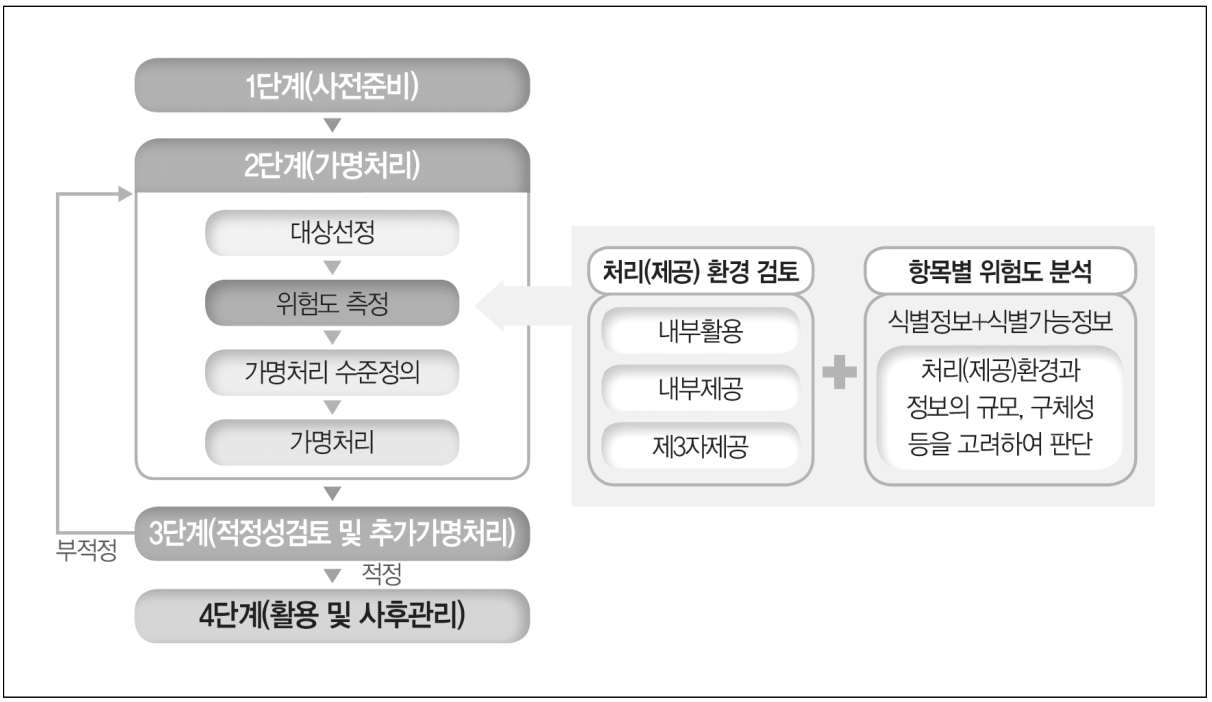
이미 언급한 바와 같이 "가명정보"는 "원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보"이며, 이와 별개로 "익명정보"라는 개념도 있다. 익명정보는 "시간, 비용, 기술 등을 합리적으로 고려할 때 다른 정보를 사용 하여도 더 이상 개인을 알아볼 수 없는 정보"로 정의할 수 있으며, 개인 식별성이 없기 때문에 개인정보에는 포함되지 않는다. 가명정보와 익명정보의 가장 큰 차이점은 개인의 식별가능성 여부에 있다고 할 수 있다.

가명정보라는 정의를 탄생하게 한 "데이터3법"에서는 또 하나의 중요한 개념이 포함되어 있다. 바로 "가명정보의 결합"이라는 개념인데, 개인정보를 가명처리한 결과물인 가명정보를 다른 개인정보처리자의 가명정보와 결합함으로써 공익과 사회 발전을 위한 새로운 분석을 시도할 수 있다. 예를 들면, 각 기관별로 별도로 가지고 있던 복지혜택 제공 정보를 결합하여, 복지의 사각지대를 찾는다던지, 급여 수준에 따른 고위험 질병에 대한 치료방안 연구 등이 가명정보의 결합을 통해 얻을 수 있는 새로운 정보들이다.

2. 가명처리

우선 가명정보 생성을 위한 가명처리에 대해 소개하고자 한다. 가명처리를 하기 위한 단계별 절차는 다음과 같다.

〈 가명처리 단계별 절차도 〉



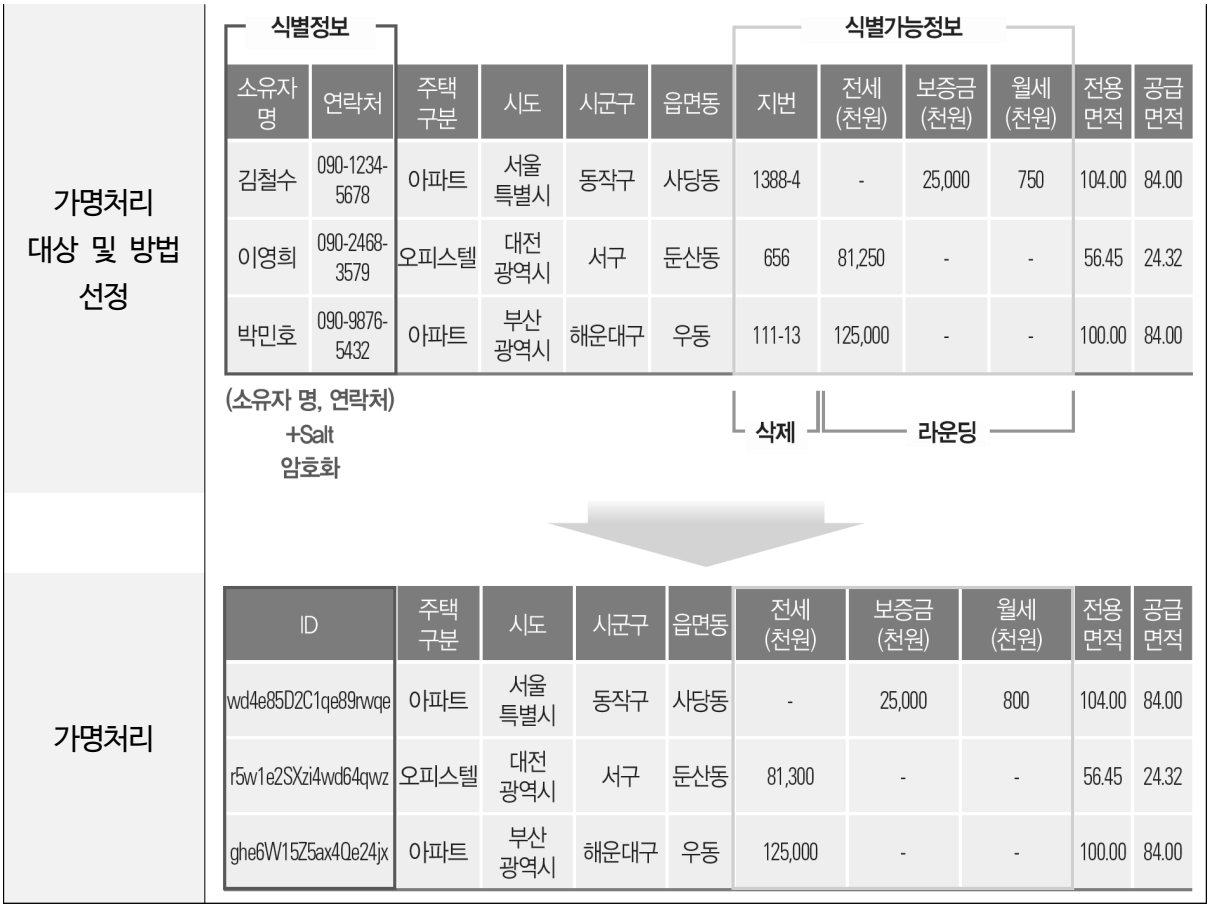
먼저 1단계는 사전준비 단계로, 가명처리의 목적을 명확히 하고, 가명처리 대상 항목 및 처리수준을 정의하는 단계이다. 이 때, 처리 목적이 법에서 허용한 목적(통계작성, 과학적 연구, 공익적 기록보존 등)에 적합한지 여부에 대한 확인이 필요하다. 또한, 가명정보를 제3자에게 제공하는 경우에는 이용목적 및 방법, 재식별 위험관리 등 가명정보의 안전성 확보를 위하여 필요한 조치를 마련토록 해야 한다. 가명정보 처리에 관한 내부관리계획이 별도로 수립되어 있지 않은 경우에는 가명정보의 안전한 관리를 위해 내부관리계획을 수립해야 한다.

다음으로 2단계는 가명처리 단계로, 가명처리의 목적에 필요한 최소한의 항목만을 가명처리 대상으로 선정하는 등 개인정보의 최소처리원칙을 준수하여야 한다. 그리고, 가명처리의 방법을 정할 때에는 처리 목적, 처리환경, 정보의 특성 등을 종합적으로 고려하여야 한다. 가명처리 단계는 세부적으로 ①대상선정, ②위험도 측정, ③가명처리 수준정의, ④가명처리로 나눌 수 있으며, 가명처리 단계에서 생성되는 추가정보(암호알고리즘, Salt 등)는 가명정보와 분리하여 별도로 저장하거나, 삭제할 수 있다. 추가정보를 분리 보관할 때에는 재식별에 악용되지 않도록 접근 권한을 최소화하고, 접근통제를 강화하는 등 필요한 조치를 적용해야 한다.

「가명정보 처리 가이드라인」에서 안내하는 가명처리 절차는 아래와 같다.

〈 가명처리 절차(예시) 〉

<table>
<tr><th>절차</th><th>처리방법(예시)</th></tr>
<tr><td>원본정보</td><td>
<table>
<tr><th>소유자명</th><th>연락처</th><th>주택구분</th><th>법정동코드</th><th>시도</th><th>시군구</th><th>읍면동</th><th>지번</th><th>건물명</th><th>전세(천원)</th><th>보증금(천원)</th><th>월세(천원)</th><th>전용면적</th><th>공급면적</th></tr>
<tr><td>김철수</td><td>090-1234-5678</td><td>아파트</td><td>2635010700</td><td>서울특별시</td><td>동작구</td><td>사당동</td><td>1388-4</td><td>대우마리나</td><td>-</td><td>25,000</td><td>750</td><td>104.00</td><td>84.00</td></tr>
<tr><td>이영희</td><td>090-2468-3579</td><td>오피스텔</td><td>3611011000</td><td>대전광역시</td><td>서구</td><td>둔산동</td><td>656</td><td>푸르지오시티</td><td>81,250</td><td>-</td><td>-</td><td>56.45</td><td>24.32</td></tr>
<tr><td>박민호</td><td>090-9876-5432</td><td>아파트</td><td>4311410100</td><td>부산광역시</td><td>해운대구</td><td>우동</td><td>111-13</td><td>평화</td><td>125,000</td><td>-</td><td>-</td><td>100.00</td><td>84.00</td></tr>
</table>
</td></tr>
<tr><td>가명처리 대상선정</td><td>
<table>
<tr><th>소유자명</th><th>연락처</th><th>주택구분</th><th>시도</th><th>시군구</th><th>읍면동</th><th>지번</th><th>전세(천원)</th><th>보증금(천원)</th><th>월세(천원)</th><th>전용면적</th><th>공급면적</th></tr>
<tr><td>김철수</td><td>090-1234-5678</td><td>아파트</td><td>서울특별시</td><td>동작구</td><td>사당동</td><td>1388-4</td><td>-</td><td>25,000</td><td>750</td><td>104.00</td><td>84.00</td></tr>
<tr><td>이영희</td><td>090-2468-3579</td><td>오피스텔</td><td>대전광역시</td><td>서구</td><td>둔산동</td><td>656</td><td>81,250</td><td>-</td><td>-</td><td>56.45</td><td>24.32</td></tr>
<tr><td>박민호</td><td>090-9876-5432</td><td>아파트</td><td>부산광역시</td><td>해운대구</td><td>우동</td><td>111-13</td><td>125,000</td><td>-</td><td>-</td><td>100.00</td><td>84.00</td></tr>
</table>
</td></tr>
<tr><td>위험도 측정</td><td>
처리환경
A사의 부동산 시세정보를 B기관에 제공(계약)
제공되는 항목의 '지번'의 경우 등기부열람을 통해 특정개인식별 가능성 존재
+
분류
소유자명, 연락처는 개인정보로 분류하고 가명처리(암호화)
구체적인 지번은 분석목적에 관계 없어 삭제 조치 및 시세정보는 분석에 필요한 단위(만원)로 가명처리
</td></tr>
</table>

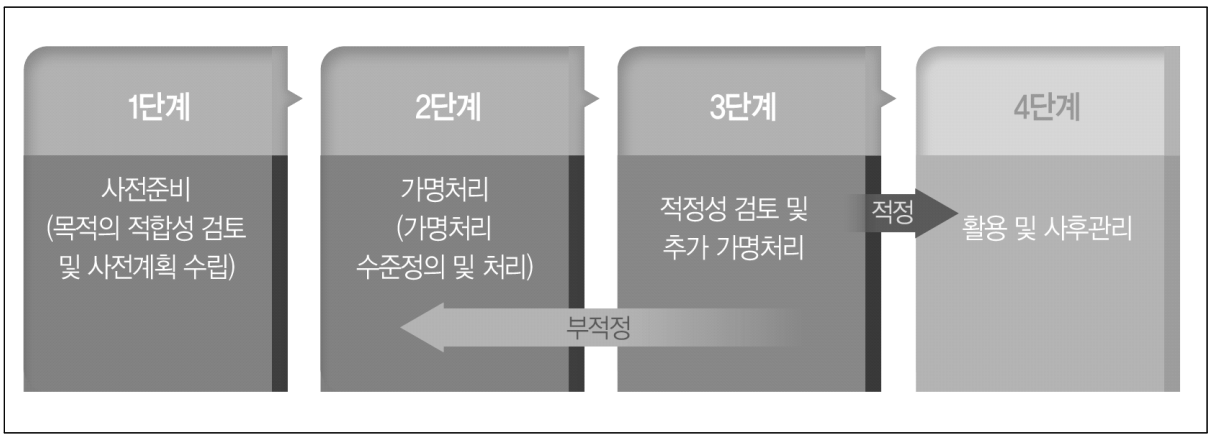


다음 단계인 3단계는 적정성 검토 및 추가 가명처리 단계로, 목적달성을 위해 적절한 수준으로 가명처리가 이루어졌는지, 재식별 가능성은 없는지 등에 대한 최종적인 판단절차를 수행한다. 적정성 검토는 개인정보처리자의 판단에 따라 외부전문가로 구성된 적정성 평가단을 구성하여 검토할 수 있다. 검토 결과, 목적 달성이 어렵거나 재식별 가능성이 있다고 판단되는 경우에는 2단계(가명처리)를 반복하거나, 부분적으로 추가 가명처리를 할 수 있다. 특히, 항목별 위험도를 바탕으로 가명처리한 경우에도 특이정보를 통해 개인식별이 가능할 수 있으므로, 검토를 통해 추가 가명처리를 할 수 있다. 예를 들면, 국회의원과 같이 특정 지역에 소수의 인원이 존재하는 직업의 경우, 지역구 국회의원 명단 등을 통해 개인이 식별될 수 있으므로, 인접 지역과 병합하거나, 직업을 일반화(예: 정치인)하는 등의 추가 가명처리를 통해 처리할 수 있다.

마지막 단계인 4단계는 활용 및 사후관리 단계로, 3단계에서의 적정성 검토 결과 가명처리가 적정하다고 판단되면 가명정보를 본래 활용 목적을 위해서 처리할 수 있다. 활용 시에는 누구든지 특정 개인을 알아보기 위한 목적으로 가명정보를 처리해서는 안되며, 가명정보 처리 과정에서 개인식별 가능성이 증가하는지 여부를 지속적으로 모니터링 하여 안전하게 처리하여야 한다. 특정 개인이 식별되는 경우에는 즉시 처리중지, 회수, 파기 등을 통해 개인식별의 위험을 제거하기 위한 조치를 취해야 한다.

여기까지 소개한 가명처리 절차를 요약하면 아래 그림과 같다.

〈 가명처리 단계별 절차도(요약) 〉



3. 가명정보 결합 및 반출

다음으로, 가명정보의 결합 및 반출에 대한 절차를 소개하고자 한다. 이번에는 가명정보 및 가명처리를 주요 주제로 다루고 있으므로, 요약하여 소개하고자 하며, 향후 기회가 되면 가명정보의 결합 및 반출에 대해 상세히 소개하고자 한다.

서로 다른 개인정보처리자 간의 가명정보의 결합 및 반출은 「개인정보 보호법」 제28조의3(가명정보의 결합 제한)에 따라, 개인정보보호위원회 또는 관계 중앙행정기관이 지정하는 전문기관이 수행할 수 있다. 현재 개인정보보호위원회를 포함한 여러 정부 기관에서 결합전문기관 지정을 위한 절차를 진행하고 있으며, 일부 지정절차를 마친 기관도 있다.

〈 가명정보의 결합 및 반출 관련 규정 〉

[개인정보보호법(2020.2.4. 개정, 2020.8.5. 시행)]

제28조의3(가명정보의 결합 제한) ① 제28조의2에도 불구하고 통계작성, 과학적 연구, 공익적 기록보존 등을 위한 서로 다른 개인정보처리자 간의 가명정보의 결합은 보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관이 수행한다.

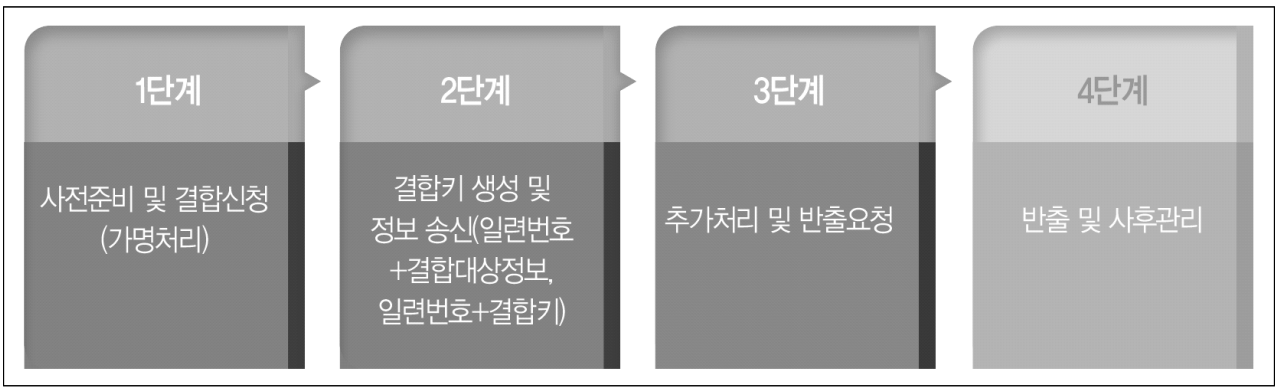
② 결합을 수행한 기관 외부로 결합된 정보를 반출하려는 개인정보처리자는 가명정보 또는 제58조의2에 해당하는 정보로 처리한 뒤 전문기관의 장의 승인을 받아야 한다.

[가명정보의 결합 및 반출 등에 관한 고시(2020.9.1. 제정·시행)]

제1조(목적) 이 고시는 「개인정보 보호법」(이하 "법"이라 한다) 제28조의3과 「개인정보 보호법 시행령」(이하 "영"이라 한다) 제29조의2부터 제29조의4까지의 규정에 따른 결합전문기관 지정 및 가명정보의 결합·반출에 관한 기준·절차 등을 정함을 목적으로 한다.

가명정보의 결합은 서로 다른 개인정보처리자가 보유한 개인정보를 가명처리하여 결합하는 것이므로, 결합하고자 하는 정보가 많을수록, 결합 대상 기관이 많을수록 개인이 식별될 수 있는 가능성이 높아진다. 그렇기 때문에, 공인된 결합전문기관에서 결합하고, 반출심사를 통과한 경우에만 반출할 수 있도록 법령에서 정의함으로써 개인이 식별되지 않도록 철저히 관리될 수 있도록 하였다.

〈 결합신청자 기준 가명정보 결합·반출 절차도(요약) 〉



위 그림은 결합신청자를 기준으로 한 가명정보 결합 및 반출 절차이다.

1단계에서 결합신청자는 사전에 서로 다른 결합신청자 간의 협의를 통해 결합신청에 필요한 가명처리를 수행하거나 결합신청서를 작성하는 등 가명정보 결합에 대한 사전 준비를 수행하고, 결합전문기관에 결합 신청한다.

다음으로, 2단계에서 결합신청자는 결합전문기관과 결합 일정, 전송 방법 등을 협의하고, 결합키관리기관으로부터 결합키 생성에 이용되는 정보를 수신하여 결합키를 생성한 후 결합 관련 정보를 송신한다. 결합전문기관에는 일련번호와 결합대상정보를, 결합키관리기관에는 일련번호와 결합키를 송신하여 결합을 준비한다. 이 때, 결합신청자는 결합키관리기관에게 사전결합률에 대한 확인을 요청할 수 있으며, 사전결합률에 따라 결합을 계속 진행할지를 결정할 수 있다.

앞서 언급한 바와 같이, 결합신청자를 기준으로 결합 절차를 설명하고 있으므로, 2단계와 3단계 사이에 진행되는 절차(결합키관리기관은 결합키연계정보 생성, 결합전문기관은 결합 수행)가 생략되어 있다. 실제로는 3단계가 시작되기 전에 결합전문기관에는 결합이 완료되어 결합된 정보가 생성되어 있는 상태이다.

다음으로, 3단계에서 결합신청자는 결합된 정보를 반출하기 전 결합전문기관 내에 설치된 별도의 공간(결합전문기관에서 제공)에서 추가 가명·익명처리를 할 수 있으며, 결합전문기관에 반출을 요청하고자 하는 경우 반출심사를 위한 자료를 제출하여야 한다.

마지막 4단계에서 결합신청자는 반출심사가 완료된 후 반출된 가명정보를 제공받아, 가명정보 처리에 대한 안전조치를 준수하여 당초 결합을 신청한 목적에 따라 처리할 수 있다. 결합신청자는 특정 개인을 알아보기 위한 목적으로 결합데이터의 재식별 또는 재결합 시도 및 결합 목적 외의 활용을 하여서는 안된다.

4. 맺음말

개인정보는 가장 우선하여 보호해야 할 대상이며, 유출될 경우의 피해는 어떠한 종류의 피해보다 크다고 할 수 있다. 하지만, 현재의 암호화 기법 등 개인정보를 보호하기 위한 기술 또한 상당 수준으로 발전하였고, 개인정보를 다른 형태로 가공하여 활용할 수 있는 상황에 이르렀다. 개인정보 보호와 활용이라는 전혀 어울리지 않는 주제를 개인정보 보호 기술이 서로 이어주고 있는 것이다. 다양한 분야에서 가명정보의 활용에 대해 논의하고 있는 만큼, 국가기록의 관리 분야에서도 가명정보에 대한 활발한 논의와 함께, 새로운 사용처를 발굴할 수 있도록 모두 함께 노력할 필요가 있다. 오늘 이 시간은 가명정보 제도에 대한 소개만으로도 부족한 시간이지만, 다음에는 국가기록에 대한 가명처리, 그리고 새로운 활용처 발굴 등을 위해 심도깊은 논의를 하는 시간이 되길 바란다.

토론요지문 1

‘기록서비스 확대를 위한 개인정보 문제 공동 해결 방안’에 대한 토론문

이 영 도(국가기록원)

‘기록서비스 확대를 위한 개인정보 문제 공동 해결 방안’에 대한 토론문

이 영 도(국가기록원)

○ 이번 정책포럼이 개인정보의 필터링이나 마스킹 기술, 가명처리 등에 한정된 것이라 아쉬움이 있음. 다만, ‘국립중앙도서관의 온라인 서비스와 개인정보보호’ 발표는 국가기록원의 긍정적 검토와 상호 협조가 필요함.

○ 개인정보보호법의 ‘개인정보’와 정보공개법의 비공개대상정보 6호 ‘개인정보’에 대한 법적인 검토와 논의가 차후 더 필요함.

- 개인정보보호법의 개인정보 : 살아 있는 개인에 관한 정보, 그러면 ‘살아 있음’은 누가 증명을 해야 하는 지? 사자(死者)의 경우 어떻게 확인할 것인지?
- 개인정보보호법의 개인정보(개인을 알아볼 수 있는 정보), 민감정보(사상, 신념, 건강, 유전정보, 범죄, 인종, 민족 등), 고유식별정보(주민등록번호, 여권번호, 면허번호, 외국인등록번호)의 범주와 처리
- 정보공개법 제9조제6호의 공개가능한 개인정보의 범주를 어떻게 정할 것인지?

(주제1) 디지털 문서의 개인정보 필터링 및 마스킹 기술

1. Privacy Free(전자문서 내 개인정보 자동검출, 제거하는 프로그램)에 대한 질의
   - 원본파일 삭제 기능의 의미는?
   - 개인정보 포함된 암호화된 파일도 가능한지?
   - 처리 속도는 어느 정도인지?
2. 비공개 개인정보와 공개 가능한 개인정보가 혼합되어 있는 전자문서의 경우
   - 개인정보 검사 결과에 대해 개인정보 제거시 공개 가능한 개인정보*의 경우는 어떻게 되는지?

*** 상훈법(훈포장 수상자 개인정보), 국적법(국적 취득 및 상실자의 성명, 생년월일, 등록기준지), 직무를 수행한 공무원(성명, 업무용 전화번호, 전자우편 주소, 소속, 직위 등) 등**

(주제2) 전자기록물 공개재분류를 위한 비공개정보 필터링 및 마스킹 기술

1. (주제1)의 토론 2와 연계하여, 비공개대상 개인정보의 범주는 어디까지 인지? 또한, 사업자번호나 법인번호는 공개가능한 정보가 아닌지?
2. 발표문 2-(2) 선행연구 및 사례분석(p.21)의 '표준기록관리시스템 내에 개인정보 필터를 도입한 기관 중 한 개 기관의 샘플 데이터 분석'은 정확도 제고를 위해 실제 확인 필요함.

〈개인정보 검출 결과〉

- 총 대상 기록물 962,955개 기록물? 962,955건?
- 주민등록번호 267,184건 : 진짜 주민등록번호인지? 앞자리 6자리+뒷자리 7자리의 조합을 모두 주민등록번호로 인식한 것은 아닌지?
- 이메일 771,229건 : 시행문·접수문의 담당자 메일주소가 아닌지?

〈개인정보 포함 현황〉

- 전체 기록물 중 공개/부분공개/비공개기록물 건수 파악이 우선임.
- 개인정보 포함 기록물 건수 962,957건 중 공개기록물에서 855,057건(88.8%)이나 검출된 것은 이해할 수 없음. 사실 관계 확인이 필요.
- 대부분의 공개기록물에 포함되어 있는 공개가능한 개인정보(직무수행자 이메일)가 아닌지?

3. 2-3)-(2)모델설계(p.25)에서 제6호(개인정보)이외에도 제2호(국가안전보장·국방·통일·외교)를 대상으로 하여 학습·분석한 이유는?
4. 원문파일이나 PDF파일을 텍스트파일로 변환하여 비공개정보를 필터링할 경우, 마스킹은 어느 파일에서 하는지?

(주제3) 국립중앙도서관의 온라인 서비스와 개인정보보호

○ 2020년 2월 IFLA(국제도서관협회연맹)와 ICA(세계기록관리협의회)의 공동성명(개인정보보호법을 지지하면서도 도서관과 기록관에서의 자료 수집과 보존 활동을 비롯하여 이용자의 정보접근성을 보장해야한다는 내용) 동의, 특히 개인사생활이나 기밀성, 문화적 민감 요소를 보호해야하는 경우에만 자유로운 접근을 제한하는 부분에 공감함

○ 국립중앙도서관과 국가기록원의 협력·대응에 동의함

3.1. 도서관 이용자에 대한 개인정보보호

- 수집하는 개인정보의 유형*과 개인정보보호법에 의해 보호해야 할 개인정보 구분 여부

*** 영국국가도서관의 경우 수집하는 개인정보의 유형 중 사용자 의견이나 피드백, 제안, 브라우징 이력, 도서관 방문 이력 및 행태 등이 포함**

3.2. 온라인 콘텐츠에서의 개인정보보호

○ 온·오프라인 자료에 포함된 개인정보의 경우 비공개 요청이 들어오면 검색 및 열람을 제한

- 당사자의 요청이 있어야 하는지? 아니며 제3자가 요청을 할 수 있는지?
- 오프라인 자료 중 특정 페이지에만 개인정보가 포함되어 있을 시 어떻게 하는지?
- 개인정보 비공개 요청에 대한 도서관의 최종적인 결정 프로세스가 있는지?

○ 디지털화로 구축된 온라인 자료에 대한 개인정보보호

- 개인정보 보호 처리 후 온라인 원문을 제공하는 비율은?
- 온라인 자료에 대한 개인정보보호 취급 사례에서 개인정보의 범주*는?

*** 사례에서 이력서의 경우 성명, 생년월일, 학력, 경력 등을 공개하고 있음**

- '기계적인 방식으로 특정 부분에 대한 개인정보 보호 처리'에서 기계적인 방식에 대해 좀 더 구체적으로 설명을 부탁드림

(주제4) 개인정보 가명처리 정책동향

1. 개인정보보호법 제28조의2(가명정보의 처리 등) '통계작성, 과학적 연구, 공익적 기록보존 등'을 목적으로 가명정보를 처리할 수 있는데, 법에 언급된 구체적인 3개의 목적 외에 '등'의 의미는? '공익적 기록보존'의 구체적 의미 또는 사례는?
2. '가명정보와 익명정보의 가장 큰 차이점은 개인의 식별가능성 여부에 있다' 부연 설명을 부탁드립니다.
3. 가명정보 처리의 범주에 '개인정보(개인을 알아볼 수 있는 정보)', '민감정보(사상, 신념, 건강, 유전정보, 범죄, 인종, 민족 등)', '고유식별정보(주민등록번호, 여권번호, 면허번호, 외국인등록번호)'를 모두 포함하고 있는지?

   ※ 개인정보보보호법 제23조(민감정보의 처리 제한), 제24조(고유식별정보의 처리 제한), 제24조의2 (주민등록번호 처리의 제한)

4. 가명정보 결합전문기관 사례 : 공공기관, 법인, 단체, 기업 등

토론요지문 2

‘기록서비스 확대를 위한 개인정보 문제 공동 해결 방안’에 대한 토론문

오 용 석(한국인터넷진흥원)

‘기록서비스 확대를 위한 개인정보 문제 공동 해결 방안’에 대한 토론문

오 용 석(한국인터넷진흥원)

토론에 앞서

AI와 빅데이터 등의 신기술과 4차산업혁명, 그리고 올해 코로나-19로 인해 모든 일상생활 속에서 디지털 전환이 가속화되고 있으며, 각종 온라인 서비스 이용을 위해 개인정보의 이용이 확대되고 있습니다. 민간 뿐만 아니라 공공·금융·의료 등 사회 전 분야에서 개인정보의 활용 요구를 반영하여 공공데이터의 적극적 개방 및 마이데이터 사업을 추진하고 있으며, 개인정보보호법에서는 가명정보라는 개념을 신설하여 특정 목적하에 정보주체의 동의없이 추가 이용할 수 있는 기반도 마련되어 시행 중입니다. 그렇지만 각종 국가기록물에는 개인정보의 포함 여부에 따라 「정보공개법」 제9조(비공개 대상정보)에 의해 상당 자료가 비공개로 설정되어 활용이 제한되고 있는 현실입니다. 국민의 알권리 충족과 공익목적의 기록보존 및 데이터개방·공유와 동시에 개인정보자기결정권을 슬기롭게 조화시키는 노력이 필요할 때입니다.

1. 「디지털 문서의 개인정보 필터링 및 마스킹 기술」에 대한 토론

먼저 “개인정보”와 “처리”에 대한 법률적 개념에 대해 말씀 주신 사항에 대한 간단한 설명 드립니다. 이에 대한 사항은 KISA의 박윤식 팀장님의 발표 자료에 더욱 상세하게 기술되어 있습니다. 개인정보보호법이 올해 2월에 개정된 사항 중 “개인정보”의 정의가 아래와 같이 일부 수정되었습니다. “결합가능성”에 대해 이전보다 구체화하였으며, 신설된 제58조의2와 같이 “익명정보”에 대한 정의를 신설하여, “익명정보”는 「개인정보보호법」의 적용을 제외했습니다. 또한 개인정보의 범주에 포함된 “가명정보”에 대한 정의를 신설하였습니다.

개인정보 보호법
[시행 2020. 8. 5] [법률 제16930호, 2020. 2. 4, 일부개정]

제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다. 〈개정 2014. 3. 24., 2020. 2. 4.〉
1. "개인정보"란 살아 있는 개인에 관한 정보로서 다음 각 목의 어느 하나에 해당하는 정보를 말한다.
가. 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보
나. 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보. **이 경우 쉽게 결합할 수 있는지 여부는 다른 정보의 입수 가능성 등 개인을 알아보는 데 소요되는 시간, 비용, 기술 등을 합리적으로 고려하여야 한다.**
다. 가목 또는 나목을 제1호의2에 따라 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보(이하 "가명정보"라 한다)
1의2. "가명처리"란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 말한다.
2. "처리"란 개인정보의 수집, 생성, 연계, 연동, 기록, 저장, 보유, 가공, 편집, 검색, 출력, 정정(訂正), 복구, 이용, 제공, 공개, 파기(破棄), 그 밖에 이와 유사한 행위를 말한다.

제58조의2(적용제외) 이 법은 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보에는 적용하지 아니한다.
[본조신설 2020. 2. 4.]

디지털 문서 파일 내에 포함된 개인정보의 자동검출을 위해 많은 솔루션이 패턴 기반으로 개인정보를 검출하고 있습니다. 주민등록번호나 이메일과 같이 정규식으로 표현할 수 있는 패턴에 대한 자동검출은 거의 정확히 검출하고 있으나, 이력정보/인명과 같이 정규식으로 표현이 제한되는 정보는 정보를 DB화해서 검출하는 것이 일반적인 방법입니다. RDBMS와 같이 정형데이터에 대해서는 각 필드에 대한 속성값(예: 주민등록번호, 이름, 전화번호)에 대한 메타정보를 알 수 있어서 검출이 용이한 반면, 일반 텍스트·웹게시글과 같은 비정형 데이터에 대해서는 이력정보/인명과 같은 정보를 검출하기 어려운 것이 현실입니다.

KISTI의 개인정보처리 기술현황에서도 정형데이터에 대해 정규식으로 표현 가능한 개인정보에 대한 검출 기능만 있는 것으로 보여집니다.

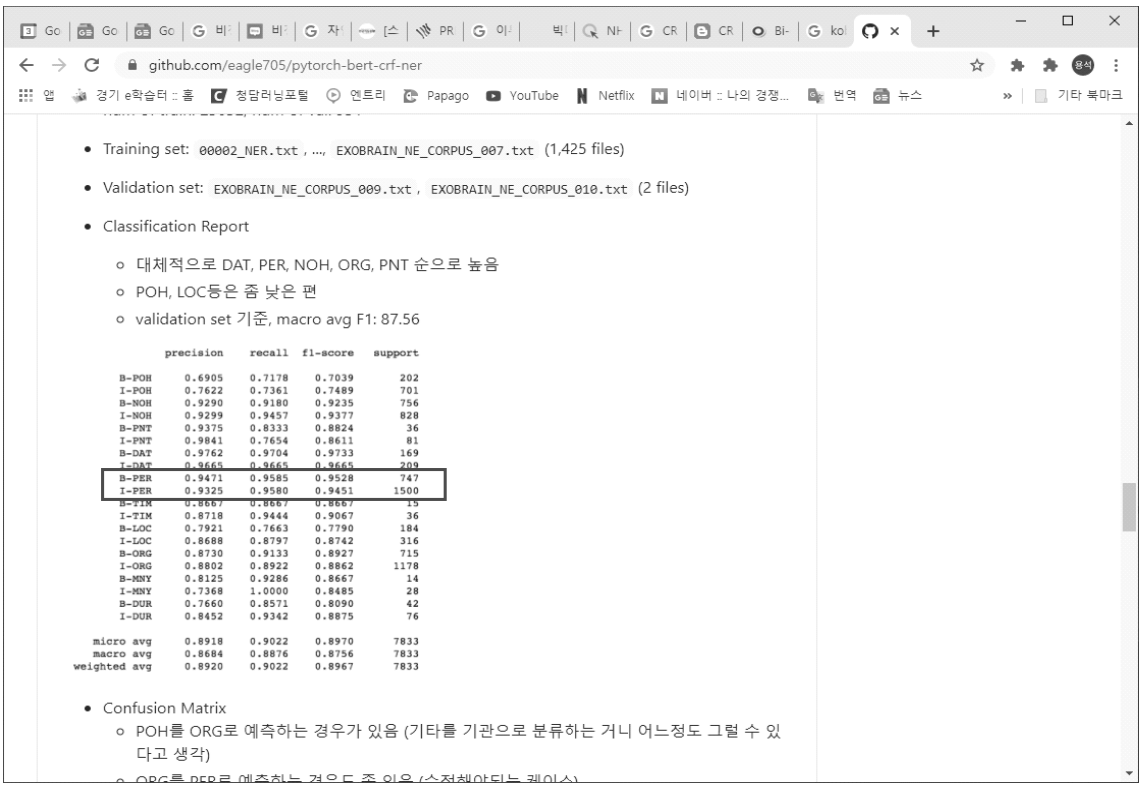
이에 자연어처리(NLP, Natural Language Processing)에서 문맥을 고려한 개인정보 검출하는 연구가 아직도 진행하고 있는 것으로 알고 있습니다. 이에 관한 연구가 계획되어 있으신지 궁금합니다.

국가기록원에서 관리하고 계시는 다양한 문서도 공문, 보고서, 이미지 등 다양한 형태의 문서를 보관하시고 계실 텐데, 향후 텍스트/이미지/동영상 등 비정형데이터에 대한 개인정보 검출에 관한 연구가 필요할 것으로 생각됩니다. KISTI에서 고민하고 계신 이미지 등 비정형 데이터에 대한 개인정보 검출에 대한 연구 방향이 어떤지 궁금합니다.

2. 「비공개정보 필터링 및 마스킹 기술」에 대한 토론

우선 공공데이터 공개 및 개방과 국민의 알권리 차원에서 국가기록원이 소장한 기록물에 대한 비공개자료의 적극적인 공개정책 및 이를 시스템화하는 연구를 추진하는 것이 바람직한 정책방향이라고 생각합니다.

정규표현식으로 해결하지 못하는 개인정보 중 인물명 중심으로 기계학습을 통해 개인정보 검출하는 모델을 설계하셨다고 하는데, 기계학습 후 탐색률(정확도)이 사전기반의 탐색 등 기계학습 전 탐색률(정오탐)에 비해 어느 정도 향상되었는지 궁금합니다. 참고로 Pytorch-BERT-CRF-NER 관련 GitHub에는 인명(PER)에 대한 정확도가 약 0.93~0.94으로 나오는데, 국가기록원 데이터로는 0.7(70%)이 나오는 결과에 대해 비교분석을 해보셨는지 궁금합니다. 물론 깃허브에서 사용한 데이터와 국가기록원 데이터간 차이에 의해서 발생했을 것으로 생각됩니다.



또한 하이브리드형 기계학습 프로세스에서 생년월일은 정규표현식으로 기술이 가능할텐데, 무슨 이유로 정규표현식 패턴이 아닌 기타로 분류하였는지 궁금합니다.

정보공개법 제9조제1항제6호에 해당하는 예외처리 사항에 대한 의사결정은 수동으로 하는지 자동화된 룰이 있는지 궁금합니다.

3. 「개인정보 가명처리 정책동향」에 대한 토론

개인정보를 가명처리하는데 법에서 허용한 목적 중 “공익적 기록보존”이 있습니다. 국가기록원의 소장 자료 중 상당수가 “공익”을 위한 목적이라고 생각이 드는데요, 이 경우 가명처리를 할 수 있는지요? 예를 들어 앞서 「비공개정보 필터링 및 마스킹 기술」에서 비공개정보에서의 개인정보를 검출하여 마스킹한 후 공개를 추진하는데 있어서, 마스킹 기술이 가명처리인지 혹은 익명처리인지와 이렇게 마스킹을 한 후 공개를 해도 무방한지 궁금합니다.

가명처리된 정보에 대한 적정성 검토가 중요할 것으로 생각됩니다. 가명처리된 정보에 대한 적정성 검토는“개인정보처리자의 판단에 따라 외부전문가로 구성된 적정성 평가단을 구성하여 검토할 수 있다.”고 하셨는데, 기관이 자의적으로 외부전문가를 구성하면 되는지 혹은 개인정보보호위원회(KISA) 차원에서 Pool을 운영할 계획인지요?

정형데이터에 대한 적정성 검증은 여러 가지 활용할 수 있는 도구가 있는 것으로 알고 있습니다. 그런데 보고서와 같은 텍스트 데이터와 이미지와 같은 비정형데이터에 대한 가명·익명처리의 적절성에 대한 검토 기준이 있는지 궁금합니다.

4. 「국립중앙도서관의 온라인 서비스와 개인정보보호」에 대한 토론

개인화 서비스 개발을 위한 개인정보 수집 확대를 위해 “공공기관에서 개인정보를 추가적으로 수집하는 것은 가능하지만, 변경 사항에 대해서는 개인정보보호위원회에 개인정보 침해요인 평가를 요청할 필요가 있다.”고 하셨는데, 개인정보처리방침을 변경해서 개인의 동의를 받는 방법은 고려하지 않으셨는지 궁금합니다. 참고로 “개인정보 침해요인 평가”는 「개인정보보호법」 제8조의2에서 “소관 법령의 제정 또는 개정을 통하여 개인정보 처리를 수반하는 정책이나 제도를 도입·변경하는 경우”에 해당합니다. 다른 방법으로는 동법 제18조제2항 제5호 “개인정보를 목적 외의 용도로 이용하거나 이를 제3자에게 제공하지 아니하면 다른 법률에서 정하는 소관 업무를 수행할 수 없는 경우로서 보호위원회의 심의 · 의결을 거친 경우”를 고려할 수 있는데, 이 경우는 공공기관만 해당하며 이때는 위치정보 등의 개인정보 이용이 반드시 소관 업무를 수행하는데 입증을 해야 합니다.

개인정보처리방침에 도서관 특수성을 반영하는 것에 대해서 제안해주신 것에 대해 깊이 공감합니다. 개인정보보호법 시행 안착을 위해 개인정보처리방침 예시를 보급한 것을 그대로 차용하여 기관별 특수성이 없이 사용되고 있는 것이 현실입니다. 특히 도서관 중 미성년자가 많이 이용하는 어린이도서관의 경우는 미성년자의 특수성을 반영하는 것도 검토되어야 할 것으로 생각합니다.

[참고]

GDPR 전문 제4조

(4) 개인정보처리는 인류에 봉사할 수 있도록 설계되어야 한다. **개인정보보호권은 절대적 권리가 아니며, 개인정보보호권은 사회에서의 개인정보보호 기능과 관련하여 고려되어야 하며 비례의 원칙에 입각하여 다른 기본권과 균형을 이루어야 한다.** 본 규정은 모든 기본권을 존중하고 여러 협약에 구현된 헌장의 자유와 원칙을 준수한다. 그러한 협약에는 특히 사생활 및 가족생활, 가정과 통신을 존중할 권리, 개인정보보호, 사상과 양심 및 종교의 자유, 표현 및 정보의 자유, 기업 활동의 자유, 효과적인 구제 권리와 공정한 재판을 받을 권리, 그리고 문화적, 종교적, 언어적 다양성 등이 포함된다.

GDPR 제85조 개인정보 처리 및 표현과 정보의 자유

1. 회원국은 법률로써 본 규정에 의거한 **개인정보 보호권과 언론 목적 및 학술, 예술 또는 문학적 표현 목적의 개인정보 처리 등 표현과 정보의 자유권 사이의 균형을 유지**시켜야 한다.
2. 언론 목적이나 학술, 예술 또는 문학적 표현의 목적으로 시행되는 개인정보 처리에 대하여 회원국이 개인정보 보호권과 표현 및 정보의 자유권 사이의 균형을 유지시켜야 할 필요가 있는 경우. 제2장(원칙), 제3장(정보주체의 권리), 제4장(정보처리자 및 수탁처리자), 제5장(제3국 또는 국제기구로의 개인정보 이전), 제6장(독립적 감독기관), 제7장(협력 및 일관성), 제9장(특정 정보처리 상황)의 면제 또는 적용 일부 제외를 규정해야 한다.
3. 각 회원국은 2호에 따라 채택한 자국법의 조문과 이에 영향을 미치는 차후의 개정법 또는 개정안을 지체 없이 집행위원회에 통보해야 한다.

GDPR 제86조 개인정보 처리 및 공식 문서 공개

공공기관, 공공기구 또는 민간기구가 공익을 위해 실시하는 업무의 수행을 위해 보유하고 있는 개인정보는 본 규정에 따른 **공식 문서의 일반 공개와 개인정보 보호권 사이의 균형을 유지시키기 위해 유럽연합 법률 또는 해당 공공기관이나 기구에 적용되는 회원국 법률에 의거하여 해당 기관이나 기구가 공개할 수 있다.**

GDPR 제89조 공익을 위한 유지보존의 목적, 과학이나 역사 연구의 목적 또는 통계 목적에서의 개인 정보 처리에 적용되는 안전조치 및 적용의 일부 제외

1. 공익을 위한 유지보존의 목적, 과학이나 역사적 연구의 목적 또는 통계 목적에서의 개인정보 처리는 본 규정에 따른 정보주체의 권리와 자유를 위해 적절한 안전조치의 적용을 받아야 한다. 그 같은 안전조치를 통해 특히 데이터 최소화의 원칙을 보장하기 위한 기술·관리적 조치가 구비되어 있어야 한다. 상기 목적들이 이 같은 방식으로 충족될 수 있다면 기술·관리적 조치에 가명처리가 포함될 수 있다. 정보주체를 식별할 수 없거나 더 이상 식별할 수 없는 개인정보의 추가 처리를 통해 상기 목적들이 충족될 수 있는 경우, 그 목적들은 이 같은 방식으로 충족되어야 한다.
2. 개인정보가 과학이나 역사적 연구 목적 또는 통계 목적으로 처리되는 경우, 유럽연합 또는 회원국 법률은 제15(열람권), 16(수정권), 18조(처리제한) 및 제21조(반대한권리)에 명시되고 본 조문 1호의 조건 및 안전조치에 따른 권리로 인해 특정 목적의 달성이 불가능하거나 심각하게 저해될 가능성이 있고 적용의 일부 제외가 그 같은 목적의 충족에 요구되는 한, 해당 권리의 적용을 일부 제외하도록 규정할 수 있다.

※ 17조 삭제권, 19조 고지의무, 20조 이전권은 없음

토론요지문 3

‘기록서비스 확대를 위한 개인정보 문제 공동 해결 방안’에 대한 토론문

김 순 석(한라대학교)

'기록서비스 확대를 위한 개인정보 문제 공동 해결 방안'에 대한 토론문

김 순 석(한라대학교)

1. 디지털 문서의 개인정보 필터링 및 마스킹 기술 발표문에 대한 토론

개인정보의 처리 기술과 관련하여 KISTI의 사례를 자세히 설명해 주셔서 우선 감사드립니다. 관련하여 몇 가지 의문 사항이 있어 질문 드리고자 합니다.

첫째, 디지털 문서 파일내에 포함된 개인정보의 자동 검출 및 제거 기능과 관련하여 다양한 포맷의 텍스트 전자문서(Excel, Hwp, PDF, PPT, Word, txt 등)에 포함된 개인정보를 검사하고 개인정보를 설정한 패턴(마스킹처리)대로 처리한다고 설명해 주셨습니다. 이에 KISTI에서 자체적으로 설정하고 계신 개인정보의 분류나 유형이 있으신지요? 그리고 설정하고 계신 패턴이 구체적으로 어떤 것인지 궁금합니다. 예컨대 Regular expression의 경우 간단한 예시도 좋습니다.

둘째, 전자문서에 포함된 개인정보의 민감도와 수준을 환경 설정에서 사용자 선택에 따라 조정하여 사용할 수 있는 기술에 대해 말씀주셨습니다. 저는 개인적으로 이 부분에 매우 공감이 갑니다. 해서 말씀주신 개인정보의 민감도와 수준을 어떠한 방법으로 분류하고 계신지 궁금합니다.

셋째, 추출된 개인정보의 비식별 사용을 위한 데이터 변환 및 마스킹 처리하는 기술에 대한 질문입니다. 대개 마스킹 기술의 경우 예컨대 김순석을 김＊＊로 처리하거나 하는 방식을 떠올리게 됩니다. 이와 관련하여 KISTI에서 처리하고 계신 마스킹 기술(보다 자세히는 이름이나 주소, 날짜 등에 있어 마스킹의 범위)과 그 외 비식별 처리 기술들에 대해 몇 가지만 소개해 주실 수 있으신지요?

넷째, 최근 개인정보보호법 등 데이터3법 개정을 통해 개인정보의 가명처리를 통한 데이터 활용 수요가 늘어날 것으로 전망됩니다. 이와 관련하여 혹시 KISTI에서 준비하고 계신 사항이나

활용 사례가 있다면 소개해 주시면 감사하겠습니다.

끝으로 향후 일반 텍스트 데이터 뿐만아니라 이미지 등 비정형 멀티미디어 기술 개발 및 개선을 통해 보다 다양한 형식의 데이터 처리가 이루어질 수 있기를 기대합니다. 수고하셨습니다.

2. 전자기록물 공개재분류를 위한 비공개정보 필터링 및 마스킹 기술 발표문에 대한 토론

우선 실무에서의 고민과 어려움을 듣고나니 안타깝다는 생각이 듭니다. 비록 짧은 기간이기 했지만 같은 연구자의 한 사람으로서 이번 연구가 좋은 결실을 맺은 것 같아 다행스럽습니다. 다만 아직 초기단계이고 인명의 경우 앞으로도 다양한 경우의 수와 추가적인 반복학습이 필요한 향후 과제도 있지만 매우 의미있는 시도였다고 생각됩니다. 또한 앞으로도 이러한 시도와 연구가 지속적으로 이루어질 수 있도록 정부차원의 행재정적 지원이 더욱 필요하다고 생각됩니다.

연구결과의 정확도에 있어 기록물 육안검수와 비교시 정규표현식은 100%, 주소 DB와 인명 기계학습은 각각 97%, 70%로 나타난 것을 알 수 있었습니다. 좀 아쉬운 부분이기 하지만 70%의 경우 안전성을 100% 담보할 수 없기 때문에 여전히 이 부분에서는 육안검수가 필요할 수 밖에 없는 상황으로 보입니다.

향후 지속적인 학습/분석을 위한 프로타입 메뉴에 있어서는 후속 과제를 통해 보다 다양화했으면 하는 바램입니다. 왜냐하면 이번 연구 수행을 위해 사용된 기록물이 샘플데이터로서 모든 유형과 상황을 반영하기에는 한계가 있을 수 있다는 생각을 해보았습니다. 또한 PDF 파일의 텍스트 변환 부분에 있어서의 정확도 향상, 인명 부분에 있어서의 기계 학습을 위한 알고리즘 등도 개선이 필요할 것으로 사료됩니다.

이번 포럼을 통해 다시한번 국가기록물의 공개재분류에 있어 어려움을 알게되었고 향후 후속 연구를 위한 정부차원이 지원이 시급하게 필요하다는 인식을 제고하는 계기가 된 것 같아 좋은 자리였다고 생각합니다.

끝으로 이번 연구 이후 향후 국가기록원이 가지고 계신 향후 후속 연구 진행 상황이나 계획이 있으시면 좀 들어보고 싶습니다.

3. 국립중앙도서관의 온라인 서비스와 개인정보보호 발표문에 대한 토론

우선 최윤경 사무관님께서 발표해주신 내용 잘 들었습니다. 이번 계기로 국립중앙도서관 등 도서관의 개인정보보호의 어려움을 알게 된 좋은 계기가 된 것 같습니다. 특히 도서관인으로서 국민의 알권리를 보장하면서도 동시에 서비스 과정에서 수집되는 이용자의 프라이버시를 함께 보호해야하는 어려움이 그것입니다. 발표내용을 들으면서 몇가지 궁금한 사항이 생겨 여쭙고자

합니다.

첫째, 외국 도서관의 사례를 들으면서 이에 따른 우리나라의 현황은 어떨까에 대한 질문입니다. 미국의 경우 '수집된 정보들은 개인을 식별하기 위한 용도로는 사용하지 않되 이용자 분석이나 웹사이트 개선을 위한 용도로 활용할 수 있다는 점을 명시'하고 있는데 우리나라도 명시되고 있거나 아니면 활용하고 계신지 궁금합니다. 만일 그러하다면 이는 활용측면에서는 좋은 사례라 여겨집니다. 그리고 영국의 경우 '비개인화 및 익명처리를 통해 이용자 데이터의 심층적인 분석과 다른 정보와의 결합이 가능하도록 명시'하고 있는데 이는 미국의 경우와 유사한 맥락이나 한걸음 더 나아가 정보의 결합까지도 허용하는 것으로 이해됩니다. 우리나라도 그러한지 궁금합니다. 이는 개인정보의 활용 측면에서 볼 때 영국보다 오히려 우리나라가 좀더 범위가 넓은 것으로 해석됩니다. 왜냐하면 이번에 개정된 우리나라의 개인정보보호법에 의하면 익명이 아닌 한국인터넷진흥원의 박팀장님께서 말씀해 주신 것처럼 가명수준으로도 합법적인 활용이 가능하기 때문입니다. 다시 말씀드려서 공익을 위한 기록 보존의 목적으로 민감정보를 포함한 개인정보를 익명수준이 아닌 가명수준으로 처리하여 활용할 수 있기 때문입니다.

둘째, 발표 내용을 들어보면 국립중앙도서관에서도 앞서 국가기록원의 공개재분류 사례처럼 온라인 자료에 대해서는 기계적인 방식의 개인정보처리가 필요한 것으로 사료됩니다. 이와 관련하여 향후 자동화된 패턴인식 및 마스킹 처리 SW 등 솔루션 도입이나 도입 방안에 대한 연구 등 어떠한 계획이 있으신지 궁금합니다.

끝으로, 말미에 말씀해주신 '국립중앙도서관과 국가기록원도 공공기관이지만 자료를 수집·보존하고 국민의 지적 자유와 알권리를 보장해야 하는 기관으로서, 개인정보보호 정책으로 인해 인류의 지적 문화 유산의 수집과 보존이 제한되거나 폐기를 요청받는 상황을 방지할 수 있도록 양 기관이 협력하여 대응할 필요가 있다.'는 부분에 저로서도 매우 공감하는 바입니다. 이와 관련하여 국가 정책입안자들이나 개인정보보호위원회 등 관련 기관에 정책적으로 제한하고자 하시는 말씀이 있으실 것 같습니다. 부탁드립니다.

4. 개인정보 가명처리 정책동향 발표문에 대한 토론

우선 발표자님께서 해주신 발표를 잘 들었습니다. 아마도 제가 파악한 핵심은 맨 마지막에 말씀해주신 “국가기록의 관리 분야에서도 가명정보에 대한 활발한 논의와 함께, 새로운 사용처를 발굴할 수 있도록 모두 함께 노력할 필요가 있다.”이 부분이 아닐까 합니다. 모두가 다들 잘 알고 계시듯 지난 8월 5일 개인정보보호법을 비롯한 이른바 데이터3법의 개정을 통해 개인정보를 가명처리함으로서 개인정보를 활용할 수 있는 길이 열리게 되었습니다. 물론 활용에는 ‘안전한’이라는 접두어가 수식어로 붙습니다. 이는 가명정보라도 개인정보이기에 안전성은 필수적으로 담보되어야 한다는 의미이기도 합니다.

한편 국가기록 분야의 경우 앞서 발표자님께서 말씀해 주신대로 공공의 이익을 위하여 지속적으로 열람할 가치가 있는 정보를 기록하여 보존하는 것을 의미하는 “공익적 기록보존”의 목적으로 개인정보를 가명처리하여 활용할 수 있게 되었습니다. 이를 활용하기 위해서는 다음과 같은 조건들이 필요할 것입니다. 첫째, 공공의 이익을 위할 것. 둘째, 지속적으로 열람할 가치가 있는 정보일 것. 셋째, 기록하여 보존하는 일과 관계될 것. 넷째, 활용을 위해 정보주체로 부터 사전 동의를 받지 않는 대신 개인정보를 안전하게 가명처리할 것. 그리고 끝으로 가명 처리된 정보일지라도 여전히 개인정보인 만큼 내부관리계획 등 관리적, 기술적, 물리적인 안전조치를 취할 것입니다. 따라서 활용을 위해 위와 같은 사항에 대한 담보가 필요하다는 의미이기도 합니다.

현재 일부 시민단체에서는 비록 가명처리가 된 정보이긴 하지만 여전히 정보 주체로부터 동의없이 활용하는데 있어 불안감이 없지 않은 것이 사실입니다. 법이 시행 초기인 만큼 이러한 시민들의 불안감과 우려를 불식시키기 위한 방법 중 하나가 바로 발표자님께서도 말씀해 주신 좋은 모범사례의 발굴이 아닐까 합니다. 이러한 모범 사례 발굴을 통해 시민들께 안전에 대한 우려를 불식시킴과 동시에 활용을 위한 물꼬가 비로소 트이게 되지 않을까 전망해 봅니다.

이러한 관점에서 볼 때 국가기록원에서도 가명정보의 활용을 위한 여러 모범 사례들을 발굴할 수 있지 않을까 기대해 봅니다. 이것이 곧 이 법을 개정하게된 취지이기도 합니다만 바꿔말해 이것이 곧 Regulator의 역할이자 공공기관의 역할이기도 한 것 같습니다.

끝으로 박윤식 팀장님께 한가지 질문을 드리는 것으로 발제를 마칠까 합니다. 개인정보보호법 개정 이후 개인정보보호위원회를 비롯하여 향후 정부에서 공익적 기록보존 등 가명정보의 활용과 관련한 정책과 행재정적 지원 방향에 대해 여쭙고 싶습니다.

MEMO

MEMO

MEMO

MEMO