

발 간 등 록 번 호

11-1371029-000238-01



정책연구용역 최종보고서

‘코로나19 아카이브’ 구축방안 연구

2022.9.



문화체육관광부
국립중앙도서관

코로나19 아카이브 구축방안 연구

2022. 09

- 연구 기관 : 중앙대학교 산학협력단
- 연구책임자 : 김학래 (중앙대학교 문헌정보학과 교수)
- 연구 원 : 양지선 (스텔스랩(주) 대표)
- 연구보조원 : 김현지 (중앙대학교 문헌정보학과 석사과정)
박하람 (중앙대학교 문헌정보학과 석사과정)
선은택 (중앙대학교 문헌정보학과 석사과정)
송채은 (중앙대학교 문헌정보학과 석사과정)



문화체육관광부
국립중앙도서관

제 출 문

국립중앙도서관 귀하

본 보고서를 『코로나19 아카이브 구축방안 연구』 과제의 최종보고서로 제출하며, 연구결과 공개에 동의합니다.

2022. 09

- 연구 기관 : 중앙대학교 산학협력단
- 연구책임자 : 김학래 (중앙대학교 문헌정보학과 교수)
- 연구 원 : 양지선 (스텔스랩(주) 대표)
- 연구보조원 : 김현지 (중앙대학교 문헌정보학과 석사과정)
박하람 (중앙대학교 문헌정보학과 석사과정)
선은택 (중앙대학교 문헌정보학과 석사과정)
송채은 (중앙대학교 문헌정보학과 석사과정)

[목 차]

제 1 장 서론	1
1.1 연구 필요성	1
1.1.1 배경	1
1.1.2 필요성	1
1.1.3 기대효과	2
1.2 연구 내용	2
1.2.1 ‘코로나19 아카이브’ 구축 방향 정립	2
1.2.2 ‘코로나19 아카이브’ 서비스 방안 제안	2
1.2.3 ‘코로나19 아카이브’ 활용 및 홍보 방안 수립	3
1.3 연구 방법	3
1.3.1 코로나19 관련 자원과 디지털 아카이브 사례 조사	3
1.3.2 디지털 아카이브의 평가	3
1.3.3 ‘코로나19 아카이브’ 구축과 서비스 방안 수립	3
 제 2 장 국내외 코로나19 아카이브 사례 분석	4
2.1 국내	4
2.1.1 국가기관	4
2.1.2 공공기관	9
2.1.3 민간기관	15
2.2. 국외	18
2.2.1 국가기관	18
2.2.2 공공기관	19
2.2.3 민간기관	23
2.3 코로나19 웹 아카이브 프로젝트	32
2.3.1 수집 정책 분석	32
2.3.2 주요 국가별 분석	34
2.4 ‘코로나19: 우리의 기억’	4
2.4.1 개요	41
2.4.2 주요 특징	42
2.4.3 데이터 수집과 분석	45
2.4.4 데이터 시각화	48
 제 3 장 ‘코로나19 아카이브’의 구축 방향	55
3.1 개요	55

3.1.1 비전	55
3.1.2 연차별 단계적 접근	57
3.2 핵심가치와 목표	59
3.2.1 수집된 자료의 끊임 없는 재현	59
3.2.2 아카이브의 의미적 연결	60
3.2.3 공동활용을 위한 공유	61
3.2.4 새로운 가치 발굴을 위한 활용	61
3.3 아카이브 구축을 위한 표준	62
3.3.1 오아시스(OASIS) 재난아카이브의 연계와 확대	62
3.3.2 국내외 표준 기반 자원 연계 방안	66
제 4 장 ‘코로나19 아카이브’의 서비스 방안	86
4.1 데이터의 수집·관리와 연계	68
4.1.1 재사용이 가능한 코로나19 관련 자료의 수집과 관리	68
4.1.2 코로나19 관련 자료의 연계	70
4.2 기술적 고려사항	72
4.2.1 코로나19 자료의 통합 검색 서비스	72
4.2.2 지능형·소통형 서비스를 위한 인공지능 서비스 모델 분석	75
4.2.3 인공지능 서비스의 도입을 위한 우선순위	77
4.2.4 UI/UX를 포함한 시각화	83
4.3 아카이브의 활용 활성화	88
4.3.1 민간 아카이브의 연계	88
4.3.2 국민 참여형 서비스	92
4.3.3 코로나19 아카이브 활용과 홍보 방안	96
제 5 장 결론	99
5.1 연구의 요약	99
5.1.1 국내외 코로나19 아카이브 사례 분석	99
5.1.2 ‘코로나19: 우리의 기억’ 사례 분석	101
5.1.3 ‘코로나19 아카이브’의 구축 방향과 서비스 방안	111
5.2 시사점	102
5.2.1 국가 차원의 코로나19에 대한 기록	102
5.2.2 수량 중심에서 문제해결 중심으로	103
[붙임]	105
[참고문헌]	115

[표 목차]

[표 2-1] 해외 코로나19의 수집 정책 내 정책 요소 분석	33
[표 2-2] WARCnet의 Working Group2에서 조사한 웹 아카이브 목록	35
[표 2-3] WARCnet에서 진행한 주요 국가별 코로나19 아카이브 분석	36
[표 2-4] ‘코로나19: 우리의 기억’의 원칙	24
[표 2-5] 주제별 수집 데이터 수량	46
[표 2-6] 코로나19 관련 수집 데이터	47
[표 3-1] 분야별 디지털 아카이브의 주요 목표	56
[표 3-2] ‘코로나19 아카이브’ 구축을 위한 로드맵	75
[표 3-3] DCAT 클래스와 설명	64
[표 3-4] 재난안전법의 재난유형	67
[표 4-1] FAIR 원칙의 세부 지표	69
[표 4-2] 디지털 아카이브가 보유한 자원의 총 수량 질의문	74
[표 4-3] SKOS 어휘를 사용한 분류체계 질의문	74
[표 4-4] 지능형·소통형 서비스를 위한 인공지능 서비스 모델 목록	76
[표 4-5] 단기·중기·장기 과제의 개발 소요 시간과 서비스 모델	77
[표 4-6] ARCH 디자인 프로세스에서 진행한 5단계 UX 테스트	85
[표 4-7] 오아시스 시스템의 웹사이트 수집을 위한 권장사항	89
[표 4-8] 오아시스의 자료 이용 범위 안내	91
[표 4-9] 공공누리 유형안내	91
[표 4-10] 이용자의 참여 중심의 디지털 아카이브 목록	93
[표 4-11] 이용자의 참여 방식에 따른 아카이브 유형	94
[표 4-12] 이용자 참여와 관련된 기능의 비교 결과	95

[그림 목차]

<그림 1-1> 연구 범위와 방법	3
<그림 2-1> 보건복지부 코로나바이러스감염증-19 대시보드	5
<그림 2-2> 통계청 코로나19 상황판	7
<그림 2-3> SGIS 코로나19 발생현황	8
<그림 2-4> 코로나19 특허정보 내비게이션	9
<그림 2-5> 코로나19 연구정보포털	10
<그림 2-6> covid19board 소비 주제의 데이터 시각화	11
<그림 2-7> 코로나19 A to Z의 코로나19 소식보기	12
<그림 2-8> 과학으로 본 코로나19	13
<그림 2-9> OASIS의 코로나바이러스감염증-19 웹 아카이브	14
<그림 2-10> 아시아지역정보센터의 코로나19 데이터	15
<그림 2-11> 코로나나우 데이터센터	16
<그림 2-12> Kaggle에서 제공하는 DS4C 데이터	17
<그림 2-13> 영국 정부 홈페이지의 데이터 요약 정보	18
<그림 2-14> Coronavirus Web Archive 컬렉션의 데이터	19
<그림 2-15> CDC의 Sciences 주제에서 제공하는 데이터 시각화	20
<그림 2-16> UKRI의 코로나19 관련 연구 메타데이터 Topic Map	21
<그림 2-17> WHO의 지도 시각화	22
<그림 2-18> Documenting 코로나19에서 루이지애나 주 검색 결과	24
<그림 2-19> IIPC Novel Coronavirus	25
<그림 2-20> Our World in Data 코로나19 확진자수 지도 시각화	26
<그림 2-21> Oxford COVID-19 Government Response Tracker의 데이터 시각화	27
<그림 2-22> NEW YORK TIMES에서 제공하는 코로나19 확진 heatmap 시각화	28
<그림 2-23> Bloomberg의 US Cases 미국 지도 시각화	29
<그림 2-24> 성별 분리된 데이터를 제공하는 국가의 지도 시각화	30
<그림 2-25> Journal of the Plague Year에서 제공하는 지도 시각화	31
<그림 2-26> WARCnet 보고서	34
<그림 2-27> ‘코로나19: 우리의 기억’ 중앙대 아카이브 기증식	24
<그림 2-28> ‘코로나19: 우리의 기억’ 의 깃허브 저장소	44
<그림 2-29> ‘코로나19: 우리의 기억’ 의 데이터 공유 체계	54
<그림 2-30> Google Colab을 활용한 데이터 분석	48
<그림 2-31> ‘코로나19: 우리의 기억’ 타임라인	94
<그림 2-32> ‘코로나19: 우리의 기억’ 스토리맵	94
<그림 2-33> 서울의 등교수업 조정 현황	50

<그림 2-34> 공연예술 산업의 코로나19 전후 매출액 비교	50
<그림 2-35> 청와대 청원 홈페이지 키워드 분석	51
<그림 2-36> 2020년 상반기 고용률과 취업자 증감	52
<그림 2-37> 코로나 이슈 타임라인	53
<그림 2-38> 분류기준별 집단감염 누적 확진자 수	53
<그림 2-39> 국가별 백신 확보량	54
<그림 2-40> 한국의 월별 백신 공급 현황	54
<그림 3-1> ‘코로나19 아카이브’의 비전과 핵심가치	55
<그림 3-2> ‘코로나19 아카이브’ 구축을 위한 주요 과제	75
<그림 3-3> 데이터맵 어휘 클래스의 다이어그램	63
<그림 3-4> DCAT 주요 클래스 다이어그램	64
<그림 3-5> SKOS 주요 클래스 다이어그램	65
<그림 3-6> 디지털 아카이브 연계를 위한 지식 모델	66
<그림 3-7> SKOS 어휘를 적용한 오아시스 재난아카이브 분류체계(안)	67
<그림 4-1> 코로나19 관련 자원의 연계를 위한 지식 모델	72
<그림 4-2> 디지털 아카이브 통합 검색 서비스 아키텍처	73
<그림 4-3> ‘코로나19 아카이브’ 검색 그래프 시각화	57
<그림 4-4> 인공지능 서비스 모델의 도입 난이도 대비 성과 현황	77
<그림 4-5> 국회·지방의회 의정정보시스템의 검색어 자동완성 기능	78
<그림 4-6> Japan Disasters Digital Archive의 재난 관련 소셜미디어 해시태그 분석	79
<그림 4-7> 한 문장 단위의 개체명 인식	81
<그림 4-8> 링크드 오픈 데이터로 구축된 WarShapo의 인물 정보	83
<그림 4-9> ARCH의 분석 서비스	84
<그림 4-10> ARCH 디자인 프로세스 4단계	84
<그림 4-11> 3단계 외부 UX 테스트에서 실시한 만족도 평가의 정량적 결과	85
<그림 4-12> Blacklight의 핵심 기능만 구현한 데모	86
<그림 4-13> Blacklight의 적용 사례	87
<그림 4-14> OpenSeadragon을 사용하여 이미지 콜로주를 생성한 사례	88
<그림 4-15> 오아시스 시스템의 웹사이트 수집절차	89
<그림 4-16> 팀 버너스-리의 Five-Star 오픈 데이터 단계	90
<그림 4-17> 이용자 참여 관점의 아카이브 현황	96
<그림 4-18> Archives Unleashed Toolkit을 이용한 웹 아카이브의 분석	97
<그림 4-19> Archives Unleashed Toolkit을 이용한 네트워크 분석	98

요 약

- 본 연구는 ‘코로나19 아카이브’ 구축을 위한 목표, 방향성, 서비스 방안과 아카이브의 활용 활성화를 위한 홍보 방안을 제안하고 있음
- 코로나19 감염병으로 인해 정치·경제·사회·문화 전반에 걸쳐 급격한 변화가 진행 중이고, 감염병 확산의 방지와 종식을 위한 대응 과정에서 다양한 기억과 기록이 생성되고 있으므로, 시의성 있게 자료를 수집·보존해야 할 필요성이 제기됨
- 코로나19와 관련된 개인과 국가의 스토리를 체계적으로 수집·보존·활용하기 위한 범국가 차원의 ‘코로나19 아카이브’ 구축이 필요하고, 지능형·소통형 서비스 구현을 위한 최적화 방안을 검토해야 함
- 특히, 국립중앙도서관은 오아시스(웹 자원 아카이빙)를 통해 코로나19 관련 재난아카이브를 구축하고 있으며, 이를 기반으로 코로나19 관련 영상·기록·구술자료 등 다양한 자료를 포함하는 서비스를 제공하기 위해 오아시스의 확대·개편 방안을 모색하고 있음
- 2장은 국내외 코로나19 아카이브의 현황과 디지털 자료의 현황을 조사하고, 팀.케일리에서 구축한 ‘코로나19:우리의 기억’ 아카이브도 국내 사례로 소개하고 있음. 국내에서 코로나19를 전문적으로 다루고 있는 아카이브는 부족한 현실임. 반면, 해외는 대학의 도서관, 지방자치단체 등 다양한 기관이 코로나19에 대한 자료를 수집·보존하고 있음. 기존의 디지털 아카이브는 디지털 자료의 수집·보존에 핵심 가치가 있지만, 아카이브의 디지털 자료의 활용에 제한이 있음
 - WARCnet 프로젝트는 해외 주요 도서관에서 코로나19 아카이브를 운영하는 주요 아카이브 담당자를 인터뷰하고 있음
 - ‘코로나19:우리의 기억’ 아카이브는 코로나19와 관련된 텍스트, 영상, 오디오, 문서 등 다양한 형식의 매체를 수집하고, 주제별로 데이터 분석·시각화하고 있음. 특히, 분석에 활용된 데이터와 코드를 함께 제공하여 재사용할 수 있는 기회를 제공함
- 3장은 국가 차원의 ‘코로나19 아카이브’ 구축과 서비스를 위한 목표와 비전을 제안하고 있음. 수집한 자료의 연계 체계를 마련하고, 국민의 참여가 가능한 개방형 아카이브를 지향함

- ‘코로나19 아카이브’는 디지털 자원의 수집·보존을 넘어 국가 차원에서 공공·민간의 연계와 융합을 지향하는 아카이브 플랫폼을 지향함
 - ‘코로나19 아카이브’는 핵심 가치는 ‘재현’, ‘연결’, ‘공유’, ‘활용’ 등 국민이 참여하고 협업할 수 있는 개방형 아카이브를 지향함
- 4장은 3장에서 수립했던 목표와 비전을 토대로, 코로나19 디지털 자원의 수집과 연계, 지능형·소통형 서비스를 구현하기 위한 기술적 고려사항, 아카이브 활용 활성화를 위한 방안을 제안하고 있음
- 핵심 가치를 실현하기 위해 코로나19 자료의 수집·보존, 수집한 웹사이트의 검색·재현, 수집 자료의 의미적 연결, 공공·민간 아카이브의 연계, 재난안전 자료의 공동활용체계, 시민 참여 기반의 생태계 채널, 재난안전 자료의 다차원 분석, 고가치 데이터의 재생산 체계 등 8개의 핵심 목표를 수립함
 - ‘코로나19 아카이브’는 구축기반 확보, 연계체계 확보, 활용 체계 확보의 3단계로 추진하고, 이후 오아시스 재난 아카이브와 확대 개편 또는 고도화를 추진함
- 국가 차원의 코로나19 아카이브의 효과적인 구축과 운영을 위해 체계적인 로드맵 수립과 조직·기술·정책 차원의 적극적 대응이 필요함
- 특히, 수집 자원의 수량적 접근이 아닌 문제해결을 위한 방식으로 코로나19 아카이브의 수집운영이 추진되는 것이 바람직함
 - 국민 참여 중심의 개방형 아카이브를 구축하기 위해 전문가, 커뮤니티, 시민의 참여를 통해 아카이브가 성장할 수 있는 기반이 마련되는 것이 중요함
 - 수집된 데이터를 데이터 과학자, 주제 분야 전문가가 활용할 수 있는 방안 제공 필요

제 1 장 서론

1.1 연구 필요성

1.1.1 배경

- 코로나19 감염병의 출현, 확산과 소멸에 이르는 과정이 수년에 걸쳐 진행됨에 따라 국가적·사회적 재난에 관한 기록의 수집·보존에 대한 필요성이 대두됨
 - 국가적 재난에 대한 축적된 기록은 각종 재난의 예방과 대응을 위한 정책 수립과 연구 자료로 활용될 수 있음
- 국립중앙도서관은 오아시스(OASIS, www.oasis.go.kr^{*})의 재난아카이브에서 코로나19의 발생부터 감염 확산 방지를 위한 노력과 의학·과학·사회·경제적 현상에 대한 웹사이트, 웹문서, 동영상, 이미지와 같은 웹 자료를 지속적으로 수집하고 있음
 - * 오아시스(OASIS, Online Archiving & Searching Internet Sources)는 국립중앙도서관이 2004년부터 온라인 디지털 정보자원을 수집·보존하고 있는 사업임. 특히 ‘재난 아카이브’는 오아시스 프로젝트의 일환으로 2014년 세월호 침몰사고, 2015년 메르스 유행, 2019년 강원도 산불 등 국가적 재난에 대한 기록을 담고 있음
- 국립중앙도서관은 2008년부터 회원 기관으로 참여하고 있는 국제인터넷보존컨소시엄(IIPC, International Internet Preservation Consortium)의 코로나바이러스감염증-19 국가별 웹 아카이브 컬렉션 구축 협력 프로젝트에 참여하고 있음
- 코로나19와 관련된 다양한 자료를 효과적으로 보존하고, 향후 활용하기 위해 국가 차원에서 디지털 자료를 보존하기 위한 관리체계를 수립하고, 체계적인 수집과 관리, 활용하기 위한 디지털 아카이브를 구축하는 것이 필요함
 - 코로나19와 관련된 정보는 중앙부처, 지방자치단체, 공공기관, 민간에서 관리하고 있고, 디지털 아카이브, 데이터포털 등 서로 다른 정보 시스템을 활용하고 있음
 - 코로나19 자료를 체계적으로 관리하기 위해 국가 차원의 노력과 개별적으로 관리되는 자료를 연계하기 위한 방안을 검토해야 함

1.1.2 필요성

- 코로나19 감염병으로 정치·경제·사회·문화 전반에 걸쳐 급격한 변화가 발생함에 따라 코로나19를 경험한 국가와 국민들의 다양한 기억과 기록을 수집·보존하고 향후 활용할 수 있는 방안을 검토해야 함

- 코로나19를 통해 영향 받은 개인과 국가의 다양한 사건과 자료를 체계적으로 수집·보존·활용하기 위한 범국가 차원의 ‘코로나19 아카이브’ 구축이 필요함
- 오아시스(웹 자원 아카이빙)에서 코로나19 관련 재난아카이브를 구축하고 있으며, 이를 기반으로 코로나19 관련 영상·기록·구술자료 등까지 수집·등록 후 종합적인 서비스 제공하기 위한 시스템으로 확대·개편 방안을 모색해야 함
- 코로나19 관련 포괄적 아카이빙, 대국민 활용 서비스 개발, 지능형·소통형 서비스 구현을 위한 ‘코로나19 아카이브’ 구축 방안을 수립해야 함

1.1.3 기대효과

- 코로나19 극복 과정에서 국민과 국가가 경험한 다양한 자원을 체계적으로 관리할 수 있는 기반을 확보할 수 있음
- 지속 가능한 디지털 자료의 관리를 통해 향후 유사 사회적 재난 발생 시, 정부 정책의 기민한 수립과 위기 대응의 기본 자료로 활용할 수 있음
- 코로나19 데이터의 기록·보존과 활용·분석 기반을 확보하고, 국립중앙도서관의 웹 자원 아카이빙(오아시스)과 연계하여 코로나19 관련 영상·기록·구술 자료를 종합적으로 서비스할 수 있는 체계 확보

1.2 연구 내용

1.2.1 ‘코로나19 아카이브’ 구축 방향 정립

- ‘코로나19 아카이브’ 목표, 가치, 방향성, 단계별 추진전략 수립
- ‘코로나19 아카이브’ 구축 대상 및 범위, 방법 제시
- 국내외 코로나19 아카이브 사례 조사 및 벤치마킹
- 코로나19 재난아카이브 확대 개편을 통한 최적 구현 방안 제시

1.2.2 ‘코로나19 아카이브’ 서비스 방안 제안

- 웹·영상·기록·구술 채록 등 코로나19 관련 자원 연계 구축 및 서비스 방안
- 코로나19 자원 통합검색 시스템 구현 방안
- 신기술 적용 지능형·소통형 서비스 구현 방안
- 민간 아카이브 연계 통합 서비스 방안
- ‘코로나19 아카이브’ 프레임워크 설계 및 UI/UX 제안
- 개인정보, 저작권 등 현안 검토 및 진단, 해결 방안 제시

1.2.3 ‘코로나19 아카이브’ 활용 및 홍보 방안 수립

- 대국민 서비스 및 활용 방안 수립
- 국내외 성과 공유 및 확산 등 홍보 방안 수립

1.3 연구 방법

1.3.1 코로나19 관련 자원과 디지털 아카이브 사례 조사

- 국내외 코로나19 아카이브의 현황과 프로젝트 검토
- 코로나19 관련 디지털 자원, 연구보고서, 오픈소스 프로젝트의 현황 분석
- 웹사이트, 문헌조사를 통한 현황 검토와 분석 진행

1.3.2 디지털 아카이브의 평가

- 국내외 디지털 아카이브 평가 방법론 조사
- FAIR 데이터 프레임워크 기반 디지털 아카이브의 평가와 분석

1.3.3 ‘코로나19 아카이브’ 구축과 서비스 방안 수립

- 국내외 웹 아카이브의 기능과 서비스 목표에 대한 조사와 분석
- 코로나19 관련 자원의 특징을 기반으로 ‘코로나19 아카이브’의 목표와 방향성 정의
- 인공지능, 통합검색 등 코로나19 서비스 구현 방안 조사와 분석



<그림 1-1> 연구 범위와 방법(김학래, 2022)

제 2 장 국내외 코로나19 아카이브

사례 분석

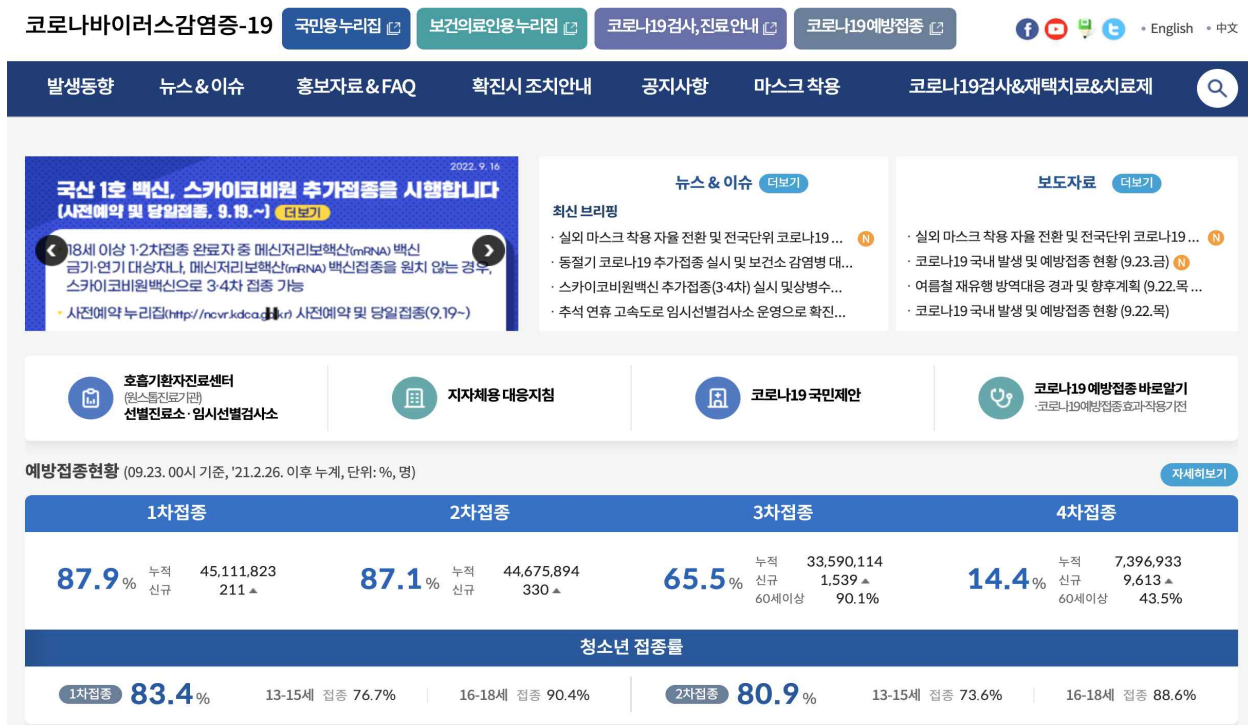
2.1 국내

2.1.1 국가기관

□ 보건복지부 코로나바이러스감염증-19(질병관리청)¹⁾

- 국내 코로나19 확진자 현황과 예방접종 현황에 대한 시각화, 코로나 관련 뉴스를 제공하는 보건복지부 질병관리청의 공식 웹사이트임
 - 코로나19 확진 현황은 국내 발생 현황, 국외 발생 현황, 시도별 발생동향 3가지로 구분하여 세부적으로 정보를 제공함
 - 코로나19 관련 통계성 정보는 공공데이터포털을 통해 전체 데이터세트를 제공하고 있음
 - 코로나19 관련 뉴스는 보도자료, 화제이슈, 일별 브리핑 3가지로 구분하여 정보를 제공함
- 정례브리핑, 지자체별 대응 지침, 정책 등 정부 공식 발표와 행동 지침 정보를 확인할 수 있는 기능을 제공함
 - 행동 지침 정보로 확진 시 조치안내를 제공하고, 공지사항으로 입국자 및 해외여행객, 자가 격리자, 집단 및 다중이용시설 사용 방법을 제공하여 필요에 따라 원하는 정보를 얻을 수 있음
 - 일별 00시 기준으로 확진 현황을 업데이트하고, 공공데이터포털에 관련 데이터세트를 제공하고 있음
 - 정례브리핑 등 정부 대응과 관련된 자료는 문서 형식으로 제공되며, 일별로 생성된 자료는 코로나19 감염증과 관련된 현황 정보와 정부의 대응이 상세하게 포함되어 보존 가치가 매우 높음
- 관련 사이트로 국민용 누리집, 보건의료인용 누리집, 코로나19 검사/진료 안내, 코로나19 예방접종 링크를 제공함으로써, 코로나19와 관련된 모든 정보를 하나의 웹사이트에서 열람 가능함
 - 코로나19 검사/진료 안내 사이트를 통해 전체 지역의 호흡기환자진료센터, 선별진료소, 임시선별검사소, 자동차 이동형 선별진료소 정보를 제공함
 - 코로나19 예방접종 사이트를 통해 예방접종 실적현황과 뉴스, 접종 후 행동지침 정보를 제공하고 있음

1) <http://ncov.kdca.go.kr/>



<그림 2-1> 보건복지부 코로나바이러스감염증-19 대시보드(보건복지부, 2022)

□ 코로나19 데이터 타임캡슐(4차산업혁명위원회, 2021)

- 코로나19로 인한 국가사회 전반의 영향과 변화를 기록·저장하기 위한 보존과 활용체계의 수립을 목적으로 추진된 프로젝트임
 - 향후 코로나 유사 위기가 발생하는 경우 대응하기 위해 관련 데이터를 축적하는 프로젝트로, 2021년 5월부터 10월까지 약 5개월 동안 327개 기관에서 총 8,298건의 코로나19 관련 데이터 목록을 식별하는 작업을 진행함
 - 2022년 보건복지부에서 코로나19 데이터 타임캡슐 구축을 위한 빅데이터 플랫폼 사업을 추진하고 있음
- 범정부 차원에서 코로나19 관련 데이터를 보유하거나 관리하는 현황을 파악하기 위해 공공분야 조사를 진행함
 - 공공분야 조사에 참여한 기관은 총 307개의 기관으로, 중앙부처 46개, 광역지자체 17개, 소속·산하 공공기관 244개 기관이 포함되었음. 177개 기관이 코로나19 관련 데이터를 보유하고 있지만, 130개 기관은 코로나19 관련 데이터를 전혀 보유하고 있지 않음
 - 중앙부처는 기관 업무, 특성과 관련된 데이터를 보유하고 있고, 광역지자체는 지자체 코로나19 현황과 대응 중심의 데이터 보유하고 있음
- 코로나19 관련 데이터의 활용과 필요한 데이터를 발굴하는 방법에 대해 고찰하고

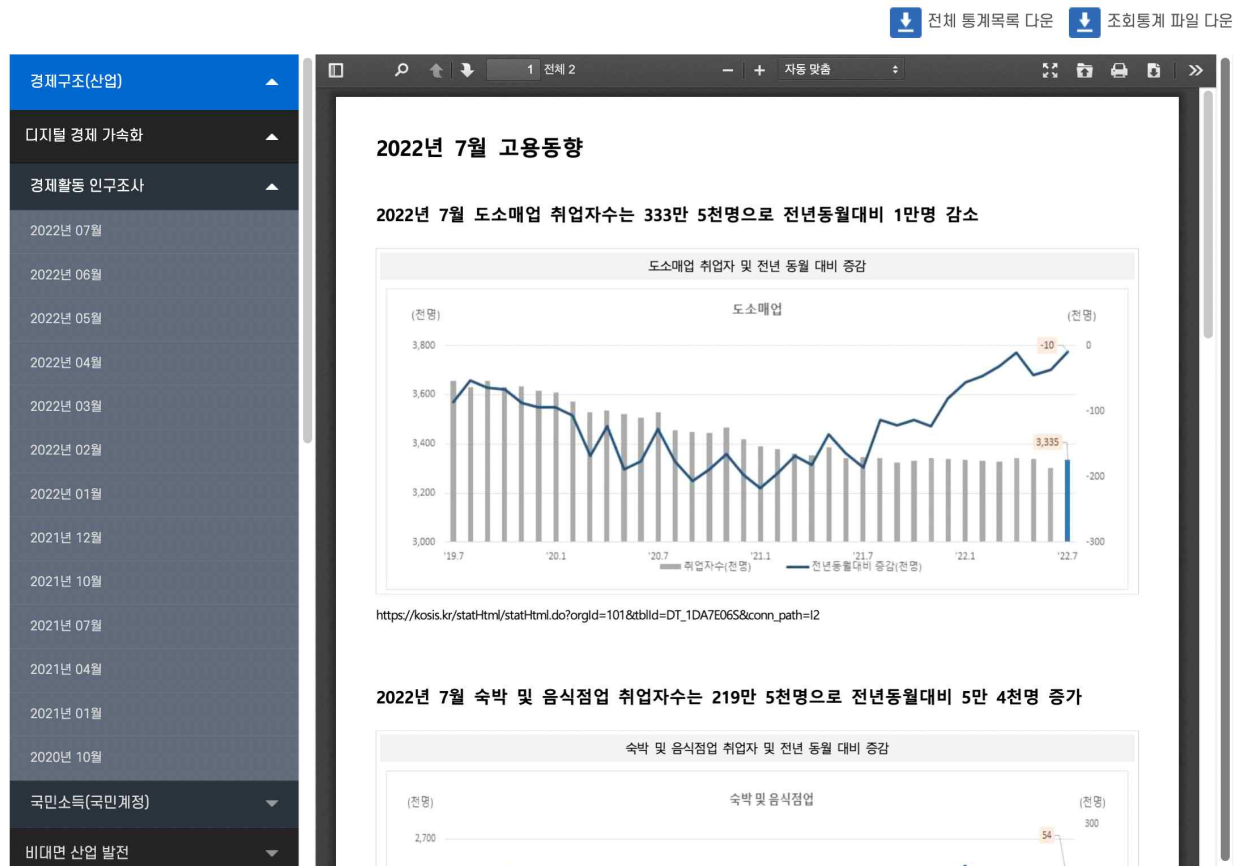
자 민간분야 조사를 진행함

- 국내 15개 분야별 빅데이터 플랫폼에서 코로나19 데이터를 조사하여 총 245건의 관련 데이터를 선정함 (유통 분야가 142건으로 가장 많은 데이터를 보유함)
- 연구자 대상의 코로나19 데이터의 활용 현황을 설문조사함. 코로나19 확진자 데이터는 설문에 참여한 대부분의 연구자가 기초 데이터로 활용함. 백신과 정부 정책에 대한 상세한 데이터에 대한 요구가 매우 높음

□ 코로나바이러스 감염증-19 상황판(통계청)²⁾

- 통계청이 보유하고 있는 코로나19 관련 통계성 데이터를 제공하고 있음
 - 확진자와 사망자, 백신접종완료 수를 제공하고, 팬데믹이 생활 전반적으로 많은 영향을 끼친 만큼, WHO 팬데믹 선언 시기를 기점으로 월별 물가동향, 고용동향, 산업동향과 가계 동향 데이터를 코로나 확진자수와 함께 꺾은선 그래프로 제공함
 - 전반적인 확진 현황을 국내와 국외로 구분하여 세부적으로 통계데이터를 제공함
 - 국내 데이터는 전국 월별, 일별, 누적, 발생경로별 확진자수, 사망자수, 백신접종완료자 수를 제공하며, SGIS통계지리정보서비스를 연결하여 지도 시각화를 제공하고, 각 지역별 세부적인 데이터를 제공함
 - 국외 데이터는 지도 시각화 형태로 대륙별 누적확진과 누적사망을 표현하고, 각 대륙에 속한 나라의 세부 데이터를 제공함
- 통계청의 주요통계(물가, 고용, 산업 활동, 온라인쇼핑, 가계, 인구, 초과사망, 모바일 인구이동)를 월별 동향으로 분석하여 제공
 - 각 주제별로 간단하게 현황을 텍스트로 정리하여 제공하고, 관련 보도자료 링크와 통계표 링크를 제공하며 조회통계 파일도 다운로드할 수 있음
- 코로나 상황판을 통해 분야 구분(경제구조, 사회구조, 노동시장, 에너지·환경), 주요관심사항(디지털 경제 가속화, 비대면 산업 발전, 대면 서비스업 변화, 집합공연, 레저시설 위축, 자동·무인화 촉진, 홈코노미 트렌드 확산 등) 정보를 제공함
 - 각 분석 보고서별로 간단하게 현황을 텍스트로 정리하고, 통계표 링크, 전체 통계목록과 조회통계 파일을 제공함

2) https://kosis.kr/covid/covid_index.do



〈그림 2-2〉 통계청 코로나19 상황판(통계청, 2022)

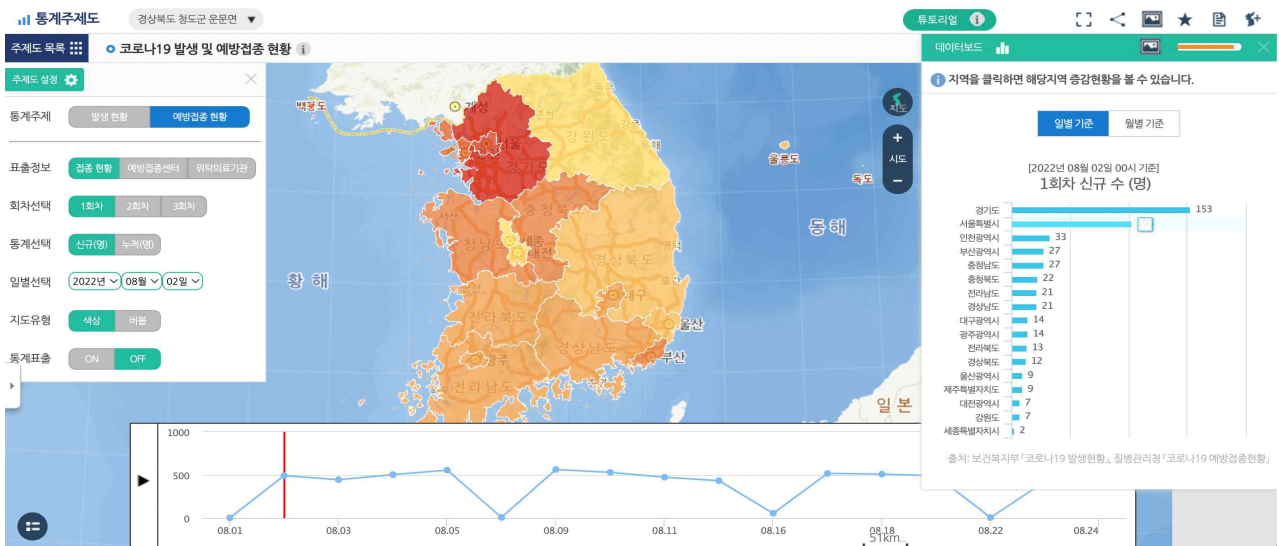
□ SGIS 코로나19 발생현황(통계청)³⁾

- 통계청에서 주관하는 통계지리정보서비스로, ‘건강과 안전’ 주제에서 코로나19 발생현황에 대한 데이터보드를 제공함
 - 일별 00시를 기준으로 국내발생과 해외유입으로 구분하여 일일 확진자와 사망자, 확진자 성별과 연령별 분포의 통계 정보, 지역별 증감정보를 제공함
 - 월별 신규확진자 정보를 그래프 시각화 형태로 제공함
- 코로나19 관련 발생 현황과 예방접종 현황을 지도 시각화 형태로 제공함
 - 발생 현황은 표출정보, 세부선택, 일별선택, 지도유형, 통계 표출을 설정하여 지도상에서 데이터 시각화가 가능함
 - 예방접종 현황은 표출정보, 회차선택, 통계 선택, 일별선택, 지도유형, 통계 표출을 설정하여 지도상에서 데이터 시각화가 가능함

3)

https://sgis.kostat.go.kr/view/thematicMap/thematicMapMain?stat_thema_map_id=sAXkcVzk5V202007141335257355ued9032uw&theme=CTGR_005&mapType=05

- 2가지 주제 모두 꺾은선 그래프 형태의 시각화를 제공하여 각 지역 뿐만 아니라, 전 지역의 현황률을 합친 수치도 확인할 수 있음



<그림 2-3> SGIS 코로나19 발생현황(통계청, 2022)

□ 코로나19 특허정보 내비게이션(특허청)⁴⁾

- 통계청에서 운영하는 코로나19 특허정보 안내 홈페이지로, 코로나19와 관련된 주요 특허기술 동향, 특허기술 정보, 특허분석·동향보고서, 코로나19 연구포털, 진단키트/K-워크스루 제품안내 등의 정보를 제공함
 - 주요 특허기술 동향, 특허기술 정보, 특허분석·동향보고서는 방호/방역, 진단/검사, 치료제/백신, 폐기물 처리 등의 주제로 세분화하고, 코로나19와 관련된 연구개발을 위해 각 주제별로 국내외 특허 기술정보를 테이블 형태로 제공함
 - 출원/공개/등록번호를 클릭하면 특허정보검색서비스인 키프리스로 이동하여 자세한 특허 신청정보와 법적상태, 심사진행상태 등 자세한 정보를 열람할 수 있음
- 진단키트/K-워크스루는 우리나라 진단키트 제품을 소개하고, 도별로 통과하면서 검체를 채취하는 진료방식인 K-워크스루에 대해 소개함
 - 진단키트 탭에서 국내외 체외 진단기기 제품 정보를 진단시약, PCR장비, 핵산추출시약, 핵산추출장비, 검체채취키트로 구분하여 각 제품명과 제조사, 제품특징, 인증 정보를 얻을 수 있음
 - K-워크스루는 채취방식을 기준으로 음압식, 양압식, 혼합식으로 구분하여 채취방식에 대한 설명, 제조사, 제품명과 제품특징 및 생산능력에 대한 정보를 제공함

4) <https://www.kipo.go.kr/ncov/>



〈그림 2-4〉 코로나19 특허정보 내비게이션(특허청, 2022)

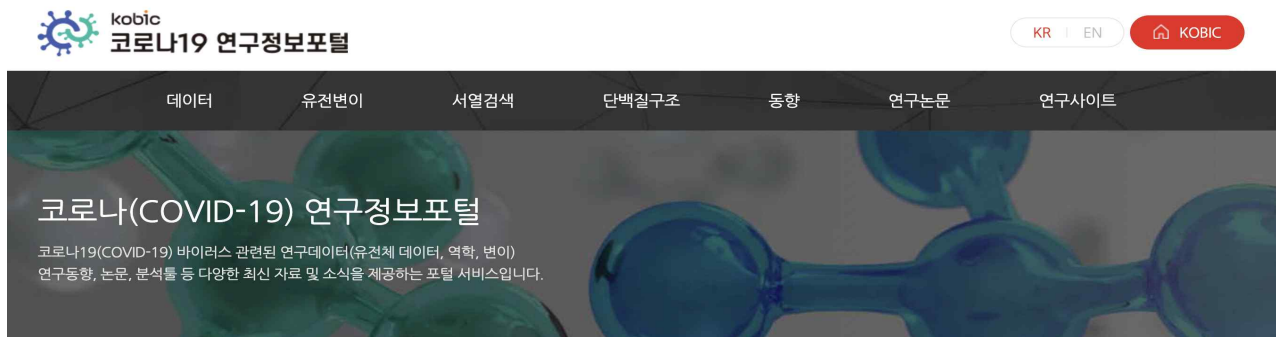
2.1.2 공공기관

□ 코로나 연구정보포털(국가생명연구자원정보센터)⁵⁾

- 코로나19 관련 논문, 연구사이트, 유전변이와 서열검색, 단백질 구조 등 연구자에게 도움이 될 만한 코로나19 연구정보를 제공하는 서비스임
 - COVID-19 관련 연구논문의 리스트와 논문 검색 기능을 제공하여, 메타데이터로 제목, 저자, DOI 번호, 출판 날짜 PMID 번호를 제공함. 미국 NCBI의 PubMed 사이트의 해당 논문으로 이동 가능한 링크를 제공하여 원문의 열람도 가능함
 - 단백질 정보를 제공하는 총 5개의 데이터베이스 사이트 링크를 제공하며, 각 사이트의 특징과 중점적으로 수집하는 정보를 정리하여 연구자들이 원하는 데이터를 편리하게 검색할 수 있도록 도움을 줌
- 국내외 코로나19 연구와 관련된 뉴스를 수집하여 제공함
 - 국내외 연구 현황을 신속하게 전달하기 위해 국내외 보도 자료를 6개 분야로 구분하여 제공하며, 2022년 9월을 기준으로 약 3,832건의 보도 자료를 제공함
 - * 6개의 분류 기준: 백신, 변이, 연구, 진단, 치료, 항체
- 전 세계에서 발표된 코로나19 유전체와 단백질 데이터를 수집하여 국가별로 구분하여 제공함
 - 유전체 서열, 단백질 서열, 전사체로 나눠 각 요소의 등록번호와 수집일자, 종명, 수집지역 등의 정보를 제공함

5) <https://www.kobic.re.kr/covid19/>

- 전 세계에서 알려진 모든 코로나19 RNA 염기 변이와 단백질 변이 정보를 수집하여 제공함
 - 서열 시작과 끝 번호, 참조서열과 변이서열, 유전자와 분류 정보를 제공함



코로나 바이러스 데이터 현황

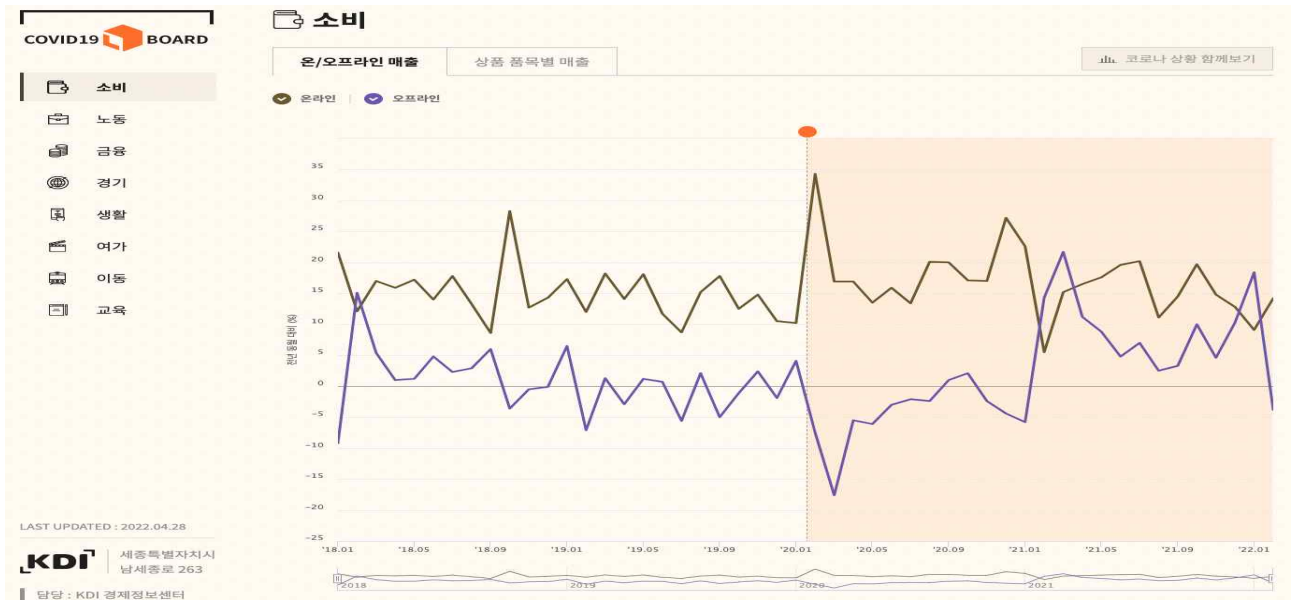
10,352 국내 유전체 서열	399 국내 단백질 서열	3 국내 전사체	7,801,493 국외 유전체 서열	26,785,075 국외 단백질 서열
---------------------	------------------	-------------	------------------------	-------------------------

<그림 2-5> 코로나19 연구정보포털(연구정보포털, 2022)

□ covid19board(기획재정부, 한국개발연구원 경제정보센터)⁶⁾

- 지역별 누적 확진자 현황 시각화와 함께, 코로나19 이후 변화된 삶이라는 주제로 소비, 노동, 금융, 경기, 생활, 문화, 교통, 교육 총 8개의 측면에서 코로나19 전후 상황을 비교하는 시각화를 총 21건 제공함
 - 각 주제별로 차트 하단의 내비게이터를 통해 기간을 조절할 수 있으며, 데이터에 대한 설명도 차트 하단에서 확인할 수 있음
 - 코로나19 전후를 효과적으로 알 수 있는 인터랙티브 그래프 형태이며, 코로나19의 전염성이 심각했던 시기의 색을 진하게 표현하여 함께 비교가 가능함
 - 차트에서 월 기준으로 데이터가 구분됨
- 시각화에 사용한 데이터의 출처는 공유하지만, 데이터는 제공하지 않음
 - 원본 데이터 출처는 산업자원통상부, 한국고용정보원, 코리아크레딧뷰로(KCB), 모바일인텍스, 한국개발연구원(KDI) 등으로 홈페이지 하단에서 확인이 가능함
 - 현재 마지막 업데이트는 2022년 4월 28일임

6) <http://www.covid19board.kr/>




<그림 2-6> covid19board 소비 주제의 데이터 시각화(covid19board, 2022)

□ 코로나19 A to Z(과학기술정보통신부, 한국생명공학연구원)⁷⁾

- 코로나19에 대한 과학 지식, 국내·외 최근 이슈, 연구동향, 진단·백신·치료제 개발 관련 연구현황 등 관련 정보를 한눈에 볼 수 있도록 제공함
- 코로나19 바로알기, 코로나19 소식보기, 코로나19 깊이알기, 코로나19 극복하기, 코로나19 연구개발 애로사항 접수 총 5가지 주제로 정보를 구분하여 제공함
 - 코로나19 바로알기 주제는 코로나19에 대한 과학 지식이나 이슈를 영상이나 글, 카드뉴스, 외부 링크 형식으로 제공하며, 제공 출처는 한국생명공학연구원, KISTI 등 매우 다양함
 - 코로나19 소식보기 주제는 코로나19에 대한 국내·외 뉴스, 정책동향, 산업동향, 제도동향 등 전반적인 최신 소식을 담은 외부 링크를 제공함
 - 코로나19 깊이알기 주제는 코로나19의 과학적 정보와 지식을 다루는 4개의 웹사이트 링크를 제공함
 - * 4개 웹사이트: 코로나19 연구정보 포털, 코로나19 특허정보 내비게이션, ScienceON 코로나19 과학기술 정보·데이터, BRIC 과학으로 본 코로나19(코로나19)
 - 코로나19 극복하기 주제는 코로나19 극복을 위한 연구현황, 국민 행동수칙 등을 담은 외부 영상이나 카드뉴스를 제공함
 - 코로나19 연구개발 애로사항 문의는 국내의 과학연구를 지원하는 기관과 연락처를 제공함

7) https://www.kribb.re.kr/kor/sub05/sub05_03_01_01.jsp


 코로나19(COVID-19)에 대한 국내·외 뉴스, 연구동향 등 다양한 소식을 알려 드립니다.

전체 검색 전체보기

번호	구분	제목	작성일	첨부
4481	뉴스	감염병 대응 위해 "보건의료데이터 중요성 더 커졌다" (약사공론)	2022-08-31	
4480	뉴스	정부, 하반기 백신접종 계획 오늘 발표...입국 전 검사도 조정 (MBN)	2022-08-31	
4479	뉴스	접종 여부 상관없이 입국 전 코로나 검사 폐지 (서울신문)	2022-08-31	
4478	뉴스	코로나19 감소세, 감염재생산지수 9주만에 1이하로 (파이낸셜뉴스)	2022-08-31	
4477	뉴스	오미크론 변이 대응 코로나19 백신 나오나 (청년외사)	2022-08-30	
4476	뉴스	"팬데믹 대응 위해 백신개발 패러다임 변해야" (의학신문)	2022-08-30	
4475	뉴스	[미래기술25]①11개월 만에 백신 출시...mRNA 세상을 구원하다 (팜이데일리)	2022-08-30	
4474	뉴스	[미래기술25]②"백스트 팬데믹 대비"...mRNA 국산화 잔걸음 (팜이데일리)	2022-08-30	
4473	뉴스	호주는 이미 유행...독감에 코로나 겹친 '트윈데믹' 발생할까 (아시아경제)	2022-08-30	
4472	뉴스	mRNA 백신 3차접종, 중증화 예방효과 87%...최대 6개월 지속 (뉴스1)	2022-08-30	

〈그림 2-7〉 코로나19 A to Z의 코로나19 소식보기(한국생명공학연구원, 2022)

□ BRIC(생물학연구정보센터) - 과학으로 본 코로나19⁸⁾

- 코로나19에 대한 과학적 지식정보를 제공하기 위한 웹사이트로, BRIC에서 보유하고 있는 코로나19 관련 자료를 한 곳에 모아 서비스하고 있음
 - 토픽, 기획특집, 웨비나(웹+세미나), 논문, 리포트, 특허, 토론, 제품, BOOK, 관련링크로 구분하여 다양한 유형의 정보를 제공함
 - BRIC 뉴스 정보를 링크로 연결하여 코로나19 관련 최신 토픽 기사와 기획특집을 제공하며, 논문과 리포트, 책도 BRIC 링크로 연결됨
 - 진행했던 웨비나의 주제와 내용, 일시와 연사 정보를 간단히 텍스트로 전달하고, 영상 링크를 제공함
 - 코로나19 관련 정보를 다루는 외부 링크를 제공함
- 코로나19 관련 학술적 토론방을 마련하여, 많은 연구자들이 질문이나 도움을 청할 수 있는 서비스를 제공함
 - 일방향적으로 정보를 제공하는 것이 아니라, 이용자들이 서로 소통하면서 정보를 생성할 수 있는 공간임

8) <https://www.ibric.org/labox/index.php?labox=5>



<그림 2-8> 과학으로 본 코로나19(생물학연구정보센터, 2022)

□ OASIS(Online Archiving & Searching Internet Sources, 국립중앙도서관)⁹⁾

- 국립중앙도서관에서 주관하는 디지털 지적 문화유산을 국가적 차원에서 수집·보존하는 온라인 디지털자원의 수집·보존 프로젝트임
 - 국립중앙도서관은 2004년부터 오아시스 프로젝트를 진행해오고 있으며, 인터넷 정보자원을 수집·보존하고 급변하는 정보환경에 대응하기 위해 공통의 도구와 기술 및 표준을 개발하는 국제인터넷보존컨소시엄(IIPC)에 2008년 가입하여 웹 아카이빙을 연구하고 집중적으로 기록을 수집함
 - 그 결과 2021년 12월을 기준으로 128,438개의 웹사이트와 103,201개의 웹 자료를 수집했으며, 현재는 총 156만 건 이상의 수집된 인터넷 자원이 서비스되고 있음
 - 헤리트릭스 3.4버전을 적용하고 있으며, 웹사이트 수집 후 저작권자의 이용동의를 받는 절차를 진행함
 - 자료 유형이 웹사이트인 경우, 출처 사이트가 자주 변경되거나 삭제될 수 있기 때문에 시간 순으로 변경내역을 모두 기록하여 달력 형식으로 제공함
 - 각 자료 유형별로 출처 URL, 표제, 발행사항 등 텍스트 형식의 메타데이터와 MODS형식의 메타데이터를 제공하며, 수집한 웹 아카이브를 활용한 웹 키워드 태그 클라우드, 웹 트렌드 등 여러 시각화 서비스도 제공함

9) <https://www.nl.go.kr/oasis/>

- 자료유행

웹사이트 (14,448)

뉴스 (7,186)

보도자료 (1,387)

웹페이지 (126)

블로그 (37)

웹사이트 (35)

카드뉴스 (28)

컬렉션

코로나바이러스감염증.. (14,448)

메르스 환자 발생 (1)

이천 물류창고 화재 (1)

기후변화 (1)

이용범위

국립중앙도서관 (14,433)

모두이용 (15)

발행연도

2021 ~ (6,885)

2016 ~ 2020 (7,531)

2011 ~ 2015 (5)

2006 ~ 2010 (2)

+ 더보기(2)

수집년도

2022 (1,977)

2021 (5,148)

2020 (7,323)

주제

기술과학 (14,393)

“이슈:코로나바이러스감염증-19(COVID-19)” (전체 14,448 건)

도움말

웹사이트

14,448건

검색순

URL

발행처

수집일

covid19.artnuri.or.kr/

변경기록

웹페이지

(아트누리)나만의 맞춤지원 예술도움

발행처: 한국문화예술위원회 | 발행연도: 2021

수집일: 2022.07.03

이용범위: 국립중앙도서관

온라인보기

http://www.cdc.go.kr/npt/

변경기록

웹사이트

감염병포털

발행처: 질병관리청 | 발행연도: 불명(不明)

수집일: 2020.07.10

이용범위: 국립중앙도서관

온라인보기

https://www.mfds.go.kr/mfds/pop/microPage_covid19...

변경기록

웹사이트

자가감사카드 관련 정보를 정확하고 신속하게 제 공하겠습니다

발행처: 식품의약품안전처 | 발행연도: 2022

수집일: 2022.07.12

이용범위: 국립중앙도서관

온라인보기

http://covid19seoulmind.org/

변경기록

웹사이트

서울시 COVID19 심리지원단

발행처: 서울시 | 발행연도: 2020

수집일: 2020.08.12

이용범위: 국립중앙도서관

온라인보기


– 14 –

2.1.3 민간기관

□ 아시아지역정보센터(Asia Regional Information Center, ARIC)¹⁰⁾

- 아시아 지역의 연구를 위한 데이터베이스 개발과 Country Report를 통한 지역 전문가들의 지식 체계화, 코로나19, SDGs, 글로벌 불평등, 환경 문제 등 글로벌, 지역 문제 해결을 위한 개방형 플랫폼 제공을 목적으로 설립된 단체임
 - 서울대학교와 Futureearth, Asia Development Institute, Asian Journal of Political Science, Korea International Cooperation Agency가 파트너십을 체결함
- 질병관리청(KCDC)의 보도자료와 질병관리청 통계 정정 발표를 반영한 코로나19 Korea Data와 ECDC, JHU, OXFORD Government Response Tracker, Our World in Data를 반영한 코로나19 World Data를 제공함
 - 수집한 원본 데이터와 활용하기 편리하도록 전처리한 데이터, 경제적 영향, 인구이동, 공공의료와 관련된 주제별로 분석한 데이터를 제공함
 - 데이터는 CSV, SAS7BDAT, DAT 형식으로 제공함
 - 각국이 코로나19에 대응하는 방법과 직면해 있는 문제들에 대해 설명하는 국가 보고서를 출판함
 - 데이터는 일별로 업데이트가 되었지만, 2022년 5월을 기점으로 데이터의 업데이트는 중지됨

Daily update of COVID-19 data More Information



SNU ARIC (Asia Regional Information Center)

- ▼ Home
- ▼ Data
- ▼ Research & Projects
- ▼ Notice

KOREA RAW DATA

DATE	CONFIRM	DEATH	RELEASE	QUARANT	TOTAL_TEST	UNDER_TEST	NEGATIVE	seriously or critically ill case	일일 신규
2022-06-16	18256457	24407	98
2022-06-15	18248479	24399	93
2022-06-14	18239056	24390	98
2022-06-13	18229288	24388	95
2022-06-12	18225460	24371	98
2022-06-11	18218078	24351	101
2022-06-10	18209650	24341	107
2022-06-09	18200346	24323	106
2022-06-08	18188200	24305	114

[Cases in Korea_Original](#)
[영양제보](#)
[보상](#)
[Number of COVID-19 vaccination by age \(1st\)](#)
[Number of COVID-19 vaccination by age \(Fully\)](#)
[Number of COVID-19 vaccination by age \(Booster\)](#)
[Number c](#)

Source : Korea Centers for Disease Control & Prevention

Q

1) History

ARIC constructs the dataset based on the press releases of KCDC. ARIC provides additional datasets that reflect statistical corrections by KCDC

2) Variables

(1) Sheet 1 : Cases in Korea

- CONFIRM : Cumulative confirmed cases
- RELEASE : Cumulative number of people released from quarantine
- QUARANT : Number of people quarantined
- DEATH : Cumulative number of people deceased
- TOTAL_TEST : Total tests as the sum of tests in progress, positive and negative tests
- UNDER_TEST : Number of tests in progress

<그림 2-10> 아시아지역정보센터의 코로나19 데이터(아시아지역정보센터, 2022)

10) <https://sites.google.com/view/snuaric/data/covid-19-data>

□ 코로나나우 데이터센터¹¹⁾

- 코로나나우 가넷은 대구대학교봉사(DU봉사단)을 주축으로 비영리적으로 운영되고 코로나19 대응현황, 뉴스, 심층 분석 등 코로나19와 관련된 다양한 자료를 제공함
 - 국내사례 중 민간에서 주도하여 코로나19 데이터 센터를 구축한 대표적인 사례임
- 현재 코로나나우 가넷 웹 페이지는 운영이 중단되었으나 사용된 데이터는 ‘코로나나우 데이터센터’ 플랫폼을 통해 요청 시 데이터를 제공하고, 일부 데이터는 공공 데이터포털 링크를 제공함
 - 제공되는 데이터의 주제는 총 6개로, 확진 현황을 나타내는 코로나19 데이터, 각종 이슈나 뉴스 보도 내용을 담은 대응관련 데이터, 코로나나우 접촉자수/페이지뷰데이터와 통계 보고서를 담은 기타 데이터, 기타 플랫폼에서 제공하는 민간 데이터, 질병관리청에서 제공하는 데이터, 코로나퓨처나우 일별 데이터를 제공함
- ‘코로나나우 데이터센터’ 플랫폼을 구축한 팀 코로나나우는 코로나 발병 초기, 데이터가 잘 정리되어 있지 않던 시기부터 정보를 정리하고 가공하여 제공하는 서비스를 시작함
 - 코로나19 초기, 팬데믹 상황에서 데이터의 주기적 관리가 필요함을 알려주는 사례로 작용하여 다양한 뉴스를 통해 이슈가 되었음
 - 현재 데이터의 업데이트는 중지되었음



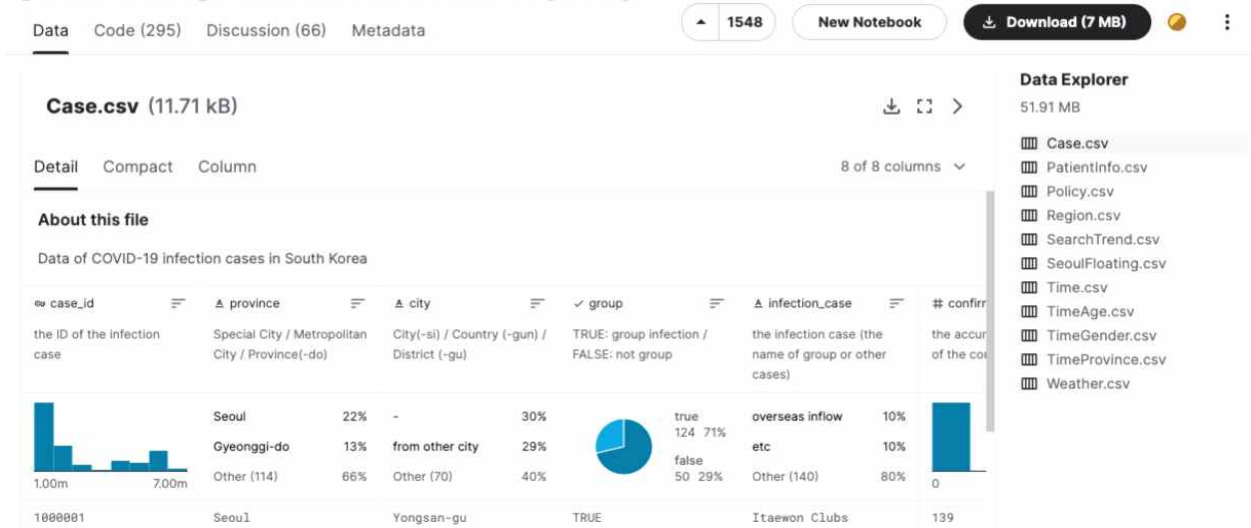
<그림 2-11> 코로나나우 데이터센터(코로나나우 데이터센터, 2022)

11) https://www.coronanow.kr/dev-datacenter/dev_data.html

□ DS4C(Data Science for 코로나19 in South Korea)¹²⁾

- 한국 질병관리청과 지방자치단체에서 공개하는 확진자 동선, 확진까지의 시간 등 코로나19 확진자의 상세한 정보를 대량으로 수집해 데이터세트를 구축하여 공개함
 - Kaggle과 GitHub를 통해 데이터와 시각화 자료를 공유함
 - 각 데이터의 컬럼별로 값의 종류와 비율을 간단히 시각화하여 제공함
 - 제공 데이터는 확진 사례, 환자 정보와 그 외 추가적으로 지역, 정책 데이터 등 총 4개의 주제에서 12개의 데이터를 공개함
 - 각 데이터세트들의 구조와 연결 관계를 다이어그램 형식으로 제공하고, 각 데이터의 컬럼 값과 의미를 설명하여 이용자들이 데이터에 대한 이해도를 높일 수 있도록 함
 - Kaggle에서 데이터를 이용한 간단한 EDA와 시각화 코드를 제공하여 활용방안을 제시함
 - 주로 월별로 데이터의 업데이트가 진행되었으나, 현재는 업데이트가 중지됨
- Kaggle 코로나 데이터세트 랭킹 3위를 차지한 이력이 있으며, AI 학회 뉴립스(NeurIPS) 워크숍에서 발표된 바 있음
 - 2022.9.26. 기준으로 이용자들이 Kaggle에서 115,643번 다운로드 했으며, 575,294건 조회를 기록하여 이용률이 높다는 사실을 알 수 있음
 - Kaggle의 Discussion 탭을 통한 이용자와의 커뮤니케이션이 가능하여 데이터에 대한 질문을 직접 할 수 있다는 장점이 있음

[NeurIPS 2020] Data Science for COVID-19 (DS4C)



<그림 2-12> Kaggle에서 제공하는 DS4C 데이터(DS4C, 2022)

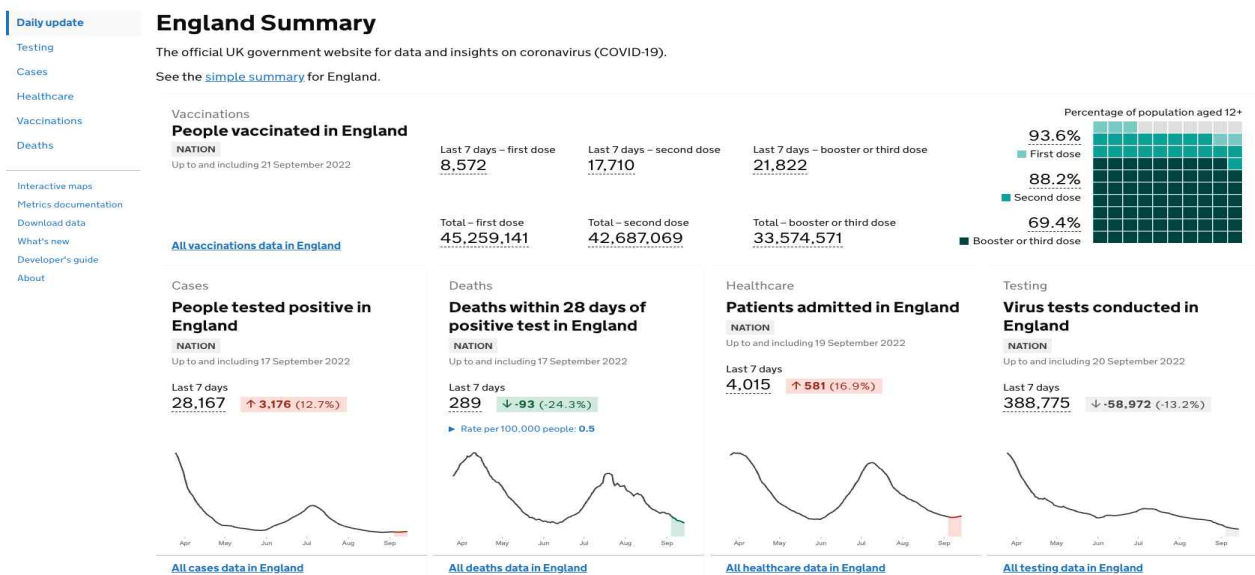
12) <https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset>

2.2. 국외

2.2.1 국가기관

□ 영국 정부 홈페이지¹³⁾

- 코로나19와 관련된 뉴스 및 커뮤니케이션, 연구 및 통계, 정책 문서 및 상담, 정보 공개의 투명성과 자료 등 크게 6개의 콘텐츠 타입으로 구분하여 자료를 공개함
 - 코로나19 주제로 진행되었던 기자 간담회 영상 링크를 제공하며, 기자회견에 사용되었던 웹 슬라이드 자료와 데이터세트를 PDF와 EXCEL 형식으로 제공함
 - 지역을 기준으로 정부에서 권고하는 가장 최신의 안내 사항이나 시행 지침 정보를 해외 여행이나 일상생활 등 주제별로 얻을 수 있음
- PCR 테스트 현황, 확진, 입원, 백신 접종과 사망 현황이 지도 시각화 형태로 제공되며 사용된 데이터는 다운로드 가능함
 - 메인화면은 지난 7일을 기준으로 수치의 증감여부를 간단하게 나타낸 데이터를 보여주고, 각 기준을 클릭하면 자세한 데이터를 확인할 수 있음
 - 일별로 업데이트되었으며, 현재는 2022년 9월 22일을 마지막으로 업데이트가 중지됨
- 지도 시각화 형태로 확진과 백신 접종 현황을 제공함
 - 백신 접종 현황의 시각화는 1차와 2차, 부스터 접종 현황을 구분하여 비교할 수 있도록 제공함



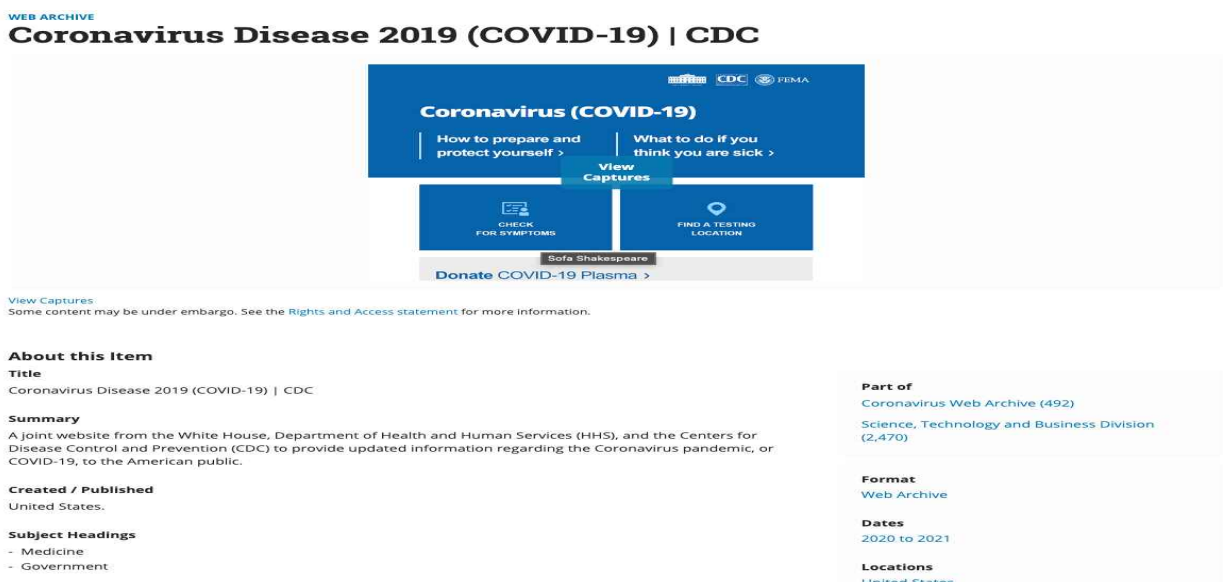
<그림 2-13> 영국 정부 홈페이지의 데이터 요약 정보(영국 정부 홈페이지, 2022)

13) <https://www.gov.uk/coronavirus>

2.2.2 공공기관

□ Coronavirus Web Archive(Library of Congress)¹⁴⁾

- 미국의회도서관에서 운영하는 웹 아카이브로, 코로나19 대유행이 진행되는 동안 다양한 분야의 웹 문서를 수집하고 일반 이용자들에게 공개함
 - 정부, 과학, 비즈니스, 정책 등 450개의 웹 아카이브가 포함됨
 - 다양한 기술, 관점과 주제 전문성을 고려한 전문가를 선정하여 팀을 구성함
 - * 다루는 주제: 정부(연방, 주, 지방), 과학 기술, 경제와 노동, 예술과 문화, 공공과 사회정책, 교육, 심리학, 스포츠와 레크리에이션, 종교
 - 대부분 월 단위와 주 단위로 데이터가 수집됨
 - 약 23개의 다양한 언어로 수집된 아카이브를 보유함
 - * 대상 23개 언어: 영어, 버마어, 프랑스어, 독일어, 그리스어, 힌디어, 인도네시아어, 이탈리아어, 일본어, 크메르어, 한국어, 라오스어, 말레이어, 말라얄람어, 마라티어어, 페르시아어, 포르투갈어, 러시아어, 싱할라어, 스페인어, 타갈로그어, 태국어, 터키어
- 제목, 요약, 제작자, 출판된 지역, 주제, 자료유형, 원본 URL 정보를 메타데이터로 제공함
 - MODSXML 형식으로 된 메타데이터 파일을 제공하고, 선택한 자료의 인용문구도 제공하여 이용자의 편의성을 높임
 - 해당 웹 아카이브가 포함된 컬렉션, 형식, 날짜, 위치, 언어, 주제의 메타데이터에 대한 값을 링크 형식으로 구축하고, 같은 메타데이터 항목값을 갖는 다른 자료나 컬렉션을 추천하여 이용자의 지속적인 탐색에 도움을 줌

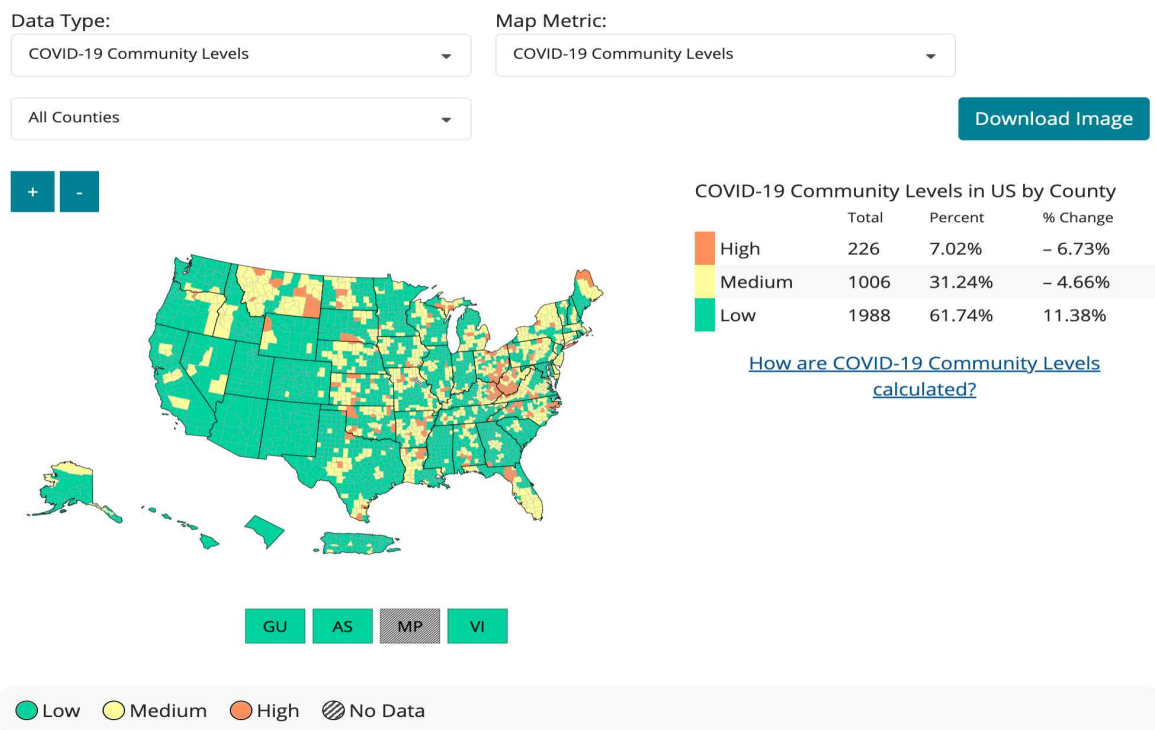


<그림 2-14> Coronavirus Web Archive 컬렉션의 데이터(미국의회도서관, 2022)

14) <https://www.loc.gov/collections/coronavirus-web-archive/about-this-collection/>

□ 미국 질병통제예방센터(CDC)¹⁵⁾

- 미국 질병통제예방센터(이하 CDC)의 공식 코로나 정보 제공 홈페이지로, 아래의 8 가지 주제로 구분하여 안내사항을 제시하고 관련 데이터를 제공함
 - Your Health 주제는 코로나19에 대한 기본 사항, 자주 묻는 질문, 외출 주의사항, 여행 가이드라인 등 일상에서 주의해야하는 정보를 제공함
 - Vaccines 주제는 예방 접종 현황, 예방 접종 가이드, 예방 접종 후 주의사항 정보를 제공하며, COVID Data Tracker 페이지로 이동하는 링크를 연결하여 회차별, 백신의 종류별, 주별 접종 현황을 시각화한 정보를 제공함
 - Cases&Data 주제는 확진자, 사망자 등 통계, 커뮤니티, 예방접종, 사례/사망 및 테스트, 인구통계학적 동향, 건강관리 설정, 계층 감시, 혈청 유병률, 고위험군, 예방 조치 및 사회적 영향에 대한 정보를 기술하고 있음
 - Specific Settings 주제는 세척/소독 및 환기, 학교와 아이들 돌보기, 노후 및 공유주택, 노숙자, 구금 시설 등 특정한 장소나 사람을 대할 때 주의사항을 기술하고 있음
 - Healthcare Workers 주제는 의료 종사자를 위한 코로나19 정보를 제공함
 - Health Depts 주제는 감염관리, 데이터 분석, 사례분석, 커뮤니티를 위한 지침 등의 정보를 제공함
 - Science 주제는 CDC의 데이터를 기반으로 매주 분석 보고서와 COVID Data Tracker 탭을 통해 기간 및 주제별 데이터 시각화를 제공함

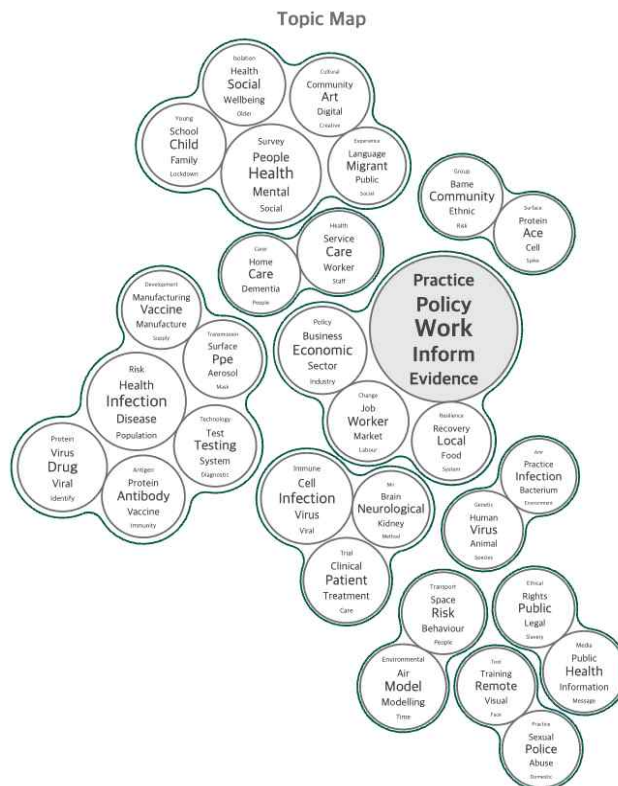


<그림 2-15> CDC의 Sciences 주제에서 제공하는 데이터 시각화(미국 질병통제예방센터, 2022)

15) <https://www.coronavirus.gov>

□ UKRI(영국 연구혁신기구)¹⁶⁾

- UKRI는 영국의 비정부 공공기관으로 7개 연구회(NERC, BBSRC, MRC, EPSRC, STFC, ESRC, AHRC)와 경제 분야를 담당하는 산하기관인 Innovate UK, 대학 연구 지원 기능을 수행하는 조직인 Research England를 통합하여 2018년에 설립됨
 - 연구자금지원과 조달역할을 하며 통합 운영을 통해 연구활성화를 추구함
- 코로나바이러스 프로젝트 섹션을 제공하며, 관련 프로젝트의 제목과 초록을 기반으로 코로나 19 유행병에 대응하여 수여되는 UKRI 연구 보조금을 시각화함
 - Topic Map을 통해 코로나19 연구들의 주제 비율을 한 번에 볼 수 있고, 각 주제마다 가중치를 알 수 있으며, 가장 많은 가중치를 차지한 주제는 21.79%를 차지한 정책 부분임
 - Top Documents를 통해 원하는 주제에 속하는 연구들의 ID, 제목, 수행기관, 관리자, 예산을 명시함
 - Topic Labels은 선택한 Topic의 주제를 워드 클라우드 형태로 나타냄
 - Document는 Top Documents에서 원하는 문서를 클릭하면, 연구의 기본적인 정보(예: 문서의 ID, 초록, 제목)를 제공함



<그림 2-16> UKRI의 코로나19 관련 연구 메타데이터 Topic Map(UKRI, 2022)

16) https://strategicfutures.org/TopicMaps/UKRI/research_map.html

□ WHO¹⁷⁾

- 국제보건기구(WHO)에서 서비스하는 웹사이트로, 전 세계의 코로나19 확진과 사망, 백신 접종 현황을 대륙과 국가별로 인터랙티브 그래프 형식으로 제공함
 - 지도 시각화를 이용하여 전 세계의 확진자와 사망자, 백신접종 현황을 누적치, 7일, 24시간을 기준으로 제공함
 - 막대그래프를 이용하여 대륙별 일일과 주간을 기준으로 확진자와 사망자의 수와 누적치를 시각화하여 제공함
 - Table View 주제는 그래프 시각화가 아니라, 표 형식으로 데이터를 제공하며 나라와 지역을 기준으로 검색 서비스도 제공함
- Measures 주제에서는 코로나19의 확산 방지를 위한 각 나라의 지침을 마스크, 학교, 일, 모임, 국내 이동, 해외여행을 기준으로 구분하여 PHSM(공중 보건 및 사회적 조치)를 측정하고, 각 기준을 통합하여 전체 PHSM 심각도를 측정함
 - 세계 지도 시각화를 이용하여 각 나라의 통합 PHSM 심각도를 시각화하여 제공하고, 클릭 시 선택한 나라의 상세한 각 기준의 측정 내용을 볼 수 있음
- 대시보드와 시각화에 사용된 국가별 확진 현황, 예방접종 현황 데이터를 일별로 업데이트하고, 데이터에 대한 설명을 상세하게 제공함
 - 데이터 업데이트의 주별로 진행되며, CSV 형식으로 제공됨



<그림 2-17> WHO의 지도 시각화(WHO, 2022)

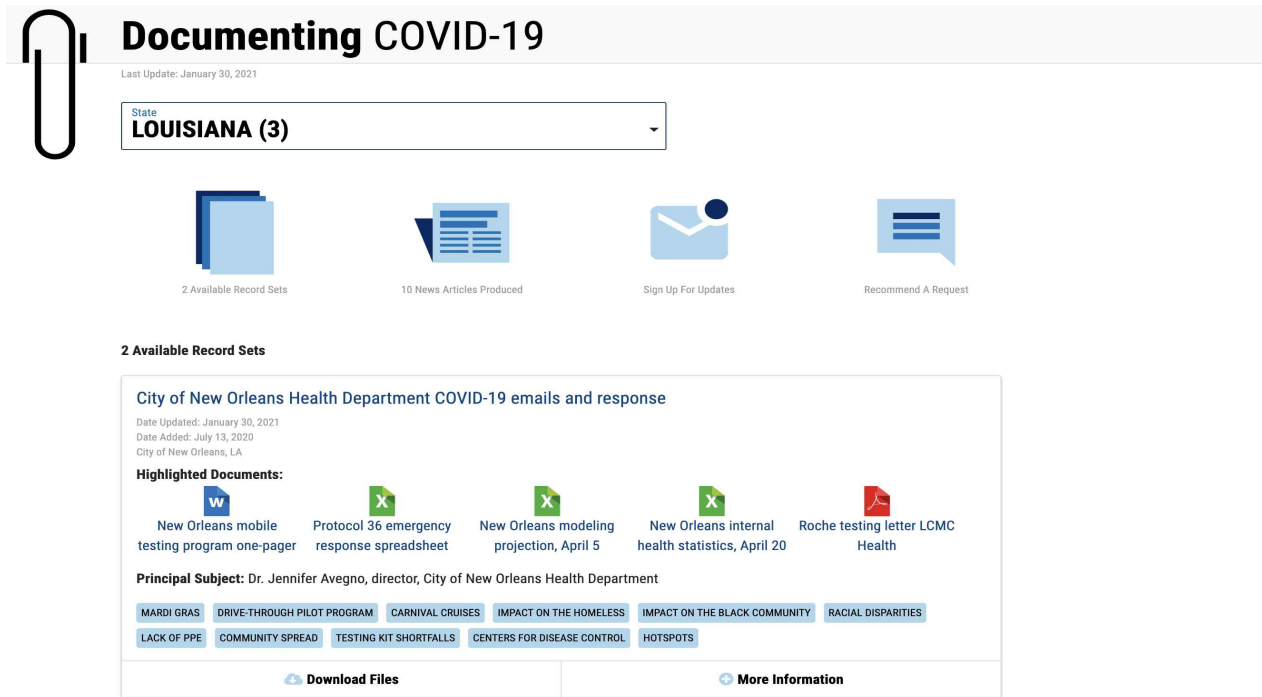
17) https://covid19.who.int/?gclid=EAlaIqobChMzLKghvjf6QIVO8EWBR1MZQIUEAAYASAAEgKQefD_BwE

2.2.3 민간기관

□ “Documenting COVID-19” 프로젝트¹⁸⁾

- 세계 각국의 중앙정부, 지방자치단체, 대학 도서관, 커뮤니티에서 코로나19 관련 자료를 기록하고 있음
 - 2022년 4월 기준의 현황으로, 주정부 공개기록법·정보 자유법(Freedom of Information Act)을 근거로 수집된 291개 레코드세트를 보유하고 있음
 - 구글에서 “Documenting COVID-19”로 구문 검색 결과, 37,800개 검색 결과가 나타남
- 기관에 따라 운영 방식에 차이가 있지만, “Documenting COVID-19” 유형의 웹사이트는 기부, 참여 방식으로 운영됨
 - 사진·이미지, 일기·저널·블로그·에세이, 비디오·오디오, 구술자료, 코멘트, 인터뷰, 노래, 예술작품, 사인, 시 등 다양한 유형의 자료를 수집하고 있음
 - 개인, 전문가, 산학연 수준의 다양한 참여가 있고, 주로 이메일, 구글폼을 이용해서 데이터를 제출하고 있음
- “Documenting COVID-19”는 코로나19 관련 검색이 가능한 문서의 저장소로, Columbia University의 Brown Institute for Media Innovation에서 시작된 협력적 오픈 레코드 프로젝트임
 - 미국 탐사 다큐멘터리 프로그램인 FRONTLINE PBS, Columbia Journalism School’s Stable Center for Investigative Journalism과 the Fund for Investigative Journalism, National Geographic에서 지원하는 프로젝트임
 - 코로나19 팬데믹 상황에서 알고리즘 및 기타 예측 도구 사용에 대한 정보를 얻는 프로젝트인 The Algorithms 프로젝트와 미국 전역의 지역 검시관의 코로나19 사망에 대한 자세한 기록을 수집하는 The Examiners Project 프로젝트에서 활용됨
- 탐색 방법은 크게 미국의 주별 탐색과 태그별 탐색 방법이 있음
 - 주별 탐색의 결과로, 각 주의 레코드세트와 뉴스 보도자료를 함께 제공함

18) <https://documentingcovid19.io/>




<그림 2-18> Documenting 코로나19에서 루이지애나 주 검색 결과(Documenting COVID-19, 2022)

□ IIPC Novel Coronavirus(국제 인터넷 보존 컨소시엄)¹⁹⁾

- 국제 인터넷 보존 컨소시엄(IIPC)에서 인터넷 콘텐츠를 보존하기 위해 Archive-It에 공개한 웹 아카이브임
 - IIPC는 전 세계의 국립 도서관과 대학 도서관, 연구소 같은 여러 기관이 회원 기관으로 참여하고 있으며, 다양한 프로젝트를 후원하고 협력함
 - 2020년 2월 21일 최초 크롤링을 시작하여 컬렉션 구성을 시작함
 - 30개 이상의 IIPC 회원사와 100명 이상의 개인과 기관 담당자들이 참여함
- 현재 코로나19와 관련하여 약 15,956건의 아카이빙 자료를 보유하고 있으며, 국가, 웹사이트 유형, 도메인, 언어를 기준으로 웹 아카이브 자료를 탐색할 수 있음
 - 감염확산, 의학적·과학적 정보, 백신개발과 예방접종, 사회경제적 이벤트로 주제를 나눠서 아카이브 자료를 구축함
 - 총 127개의 국가에서 45개의 언어로 된 자료를 포함함
 - 약 8,000개 이상의 시드 (웹사이트)를 대상으로 크롤링을 수행함
 - 웹 아카이브의 메타데이터로 웹사이트 URL, 사용되는 언어, 웹사이트 제공 국가, 웹사이트 유형과 도메인을 제공함

19) <https://netpreserve.org/projects/collaborative-collections/>



Novel Coronavirus (COVID-19)

Collected by: [International Internet Preservation Consortium](#)

Archived since: 2월, 2020

Description: A collection created by the Content Development Group of the International Internet Preservation Consortium in collaboration with Archive-It to preserve web content related to the ongoing Novel Coronavirus (Covid-19) outbreak. Identification of seed websites and initial web crawling began in February 2020, and the collection will continue to add new content as needed during the course of the pandemic and its containment. High priority subtopics include: coronavirus origins; information about the spread of infection; regional or local containment efforts, including vaccine development and implementation of vaccinations; medical and scientific aspects; social aspects; economic aspects; and political aspects. Websites from anywhere in the world and in any language are in scope.

Subject: [Science & Health](#), [Spontaneous Events](#), [COVID-19 \(Disease\)](#), [Epidemics](#), [Coronavirus infections](#), [COVID-19 Pandemic, 2020](#)

Creator: [International Internet Preservation Consortium](#)

Collector: [International Internet Preservation Consortium](#)

Narrow Your Results

Language Sort By: Count (A-Z)

- Spanish (6755)
- English (3313)
- Portuguese (2040)
- French (827)
- German (776)

[More ▼](#)

Website Type Sort By: Count (A-Z)

- Media article (6024)
- Government agencies (National) (4374)
- Government agencies (State/Provincial/Local) (1772)
- Media topic page (921)
- Non-profit/Mutual aid/Volunteer (666)
- Commercial (458)
- Personal/Social media (442)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

[Sites](#) [Search Page Text](#)

Page 1 of 160 (15,956 Total Results)

[Next Page ►](#)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Title: 외교부 해외안전여행

URL: <http://0404.go.kr/dev/main.mofa>

Description: Ministry of Foreign Affairs safe travel overseas

Captured once on September 07, 2020

Language: Korean

Country of publication: Korea (South)

Top-Level Domain: .kr

Website Type: Government agencies (National)

<그림 2-19> IIPC Novel Coronavirus(Archive-It, 2022)

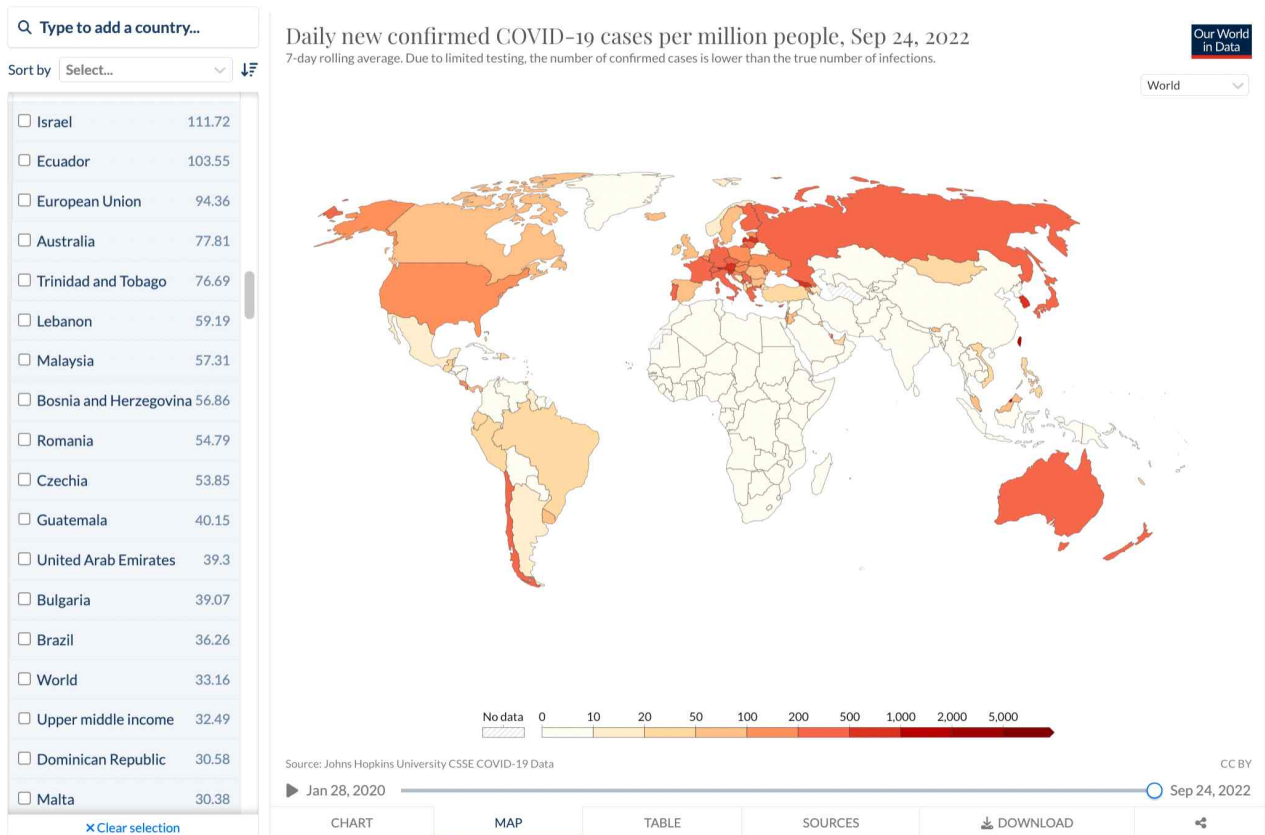
□ Our World in Data²⁰⁾

- 옥스퍼드 대학교의 연구자들과 Global Change Data Lab 비영리단체와의 파트너십을 체결하여 진행된 프로젝트임
- 차트, 맵, 테이블 형태의 인터랙티브 시각화를 제공하며, 사용자가 주제, 기간, 인구수 등 원하는 조건을 설정해 시각화의 조작이 가능함
 - 2020년 01월 28일부터 현재까지의 데이터를 기반으로 시각화를 제공함
 - 대륙이나 특정 나라를 선택할 수 있으며, 주기도 선택할 수 있음
- Country Profiles, Vaccinations, US Vaccinations, Cases, Deaths, Testing,

20) <https://ourworldindata.org/coronavirus-source-data>

Hospitalizations, Policy Responses, Mortality Risk, Excess Mortality 등 주제별로 데이터를 제공하며, 데이터는 일별로 업데이트됨

- 모든 데이터와 시각화 자료는 개방되어 있으며, GitHub를 통해 소스 코드와 데이터세트, 데이터세트에 대한 오버뷰를 제공함
- 데이터는 XLSX, CSV, JSON, PNG, SVG 형식으로 제공됨



<그림 2-20> Our World in Data 코로나19 확진자수 지도 시각화(Our World in Data, 2022)

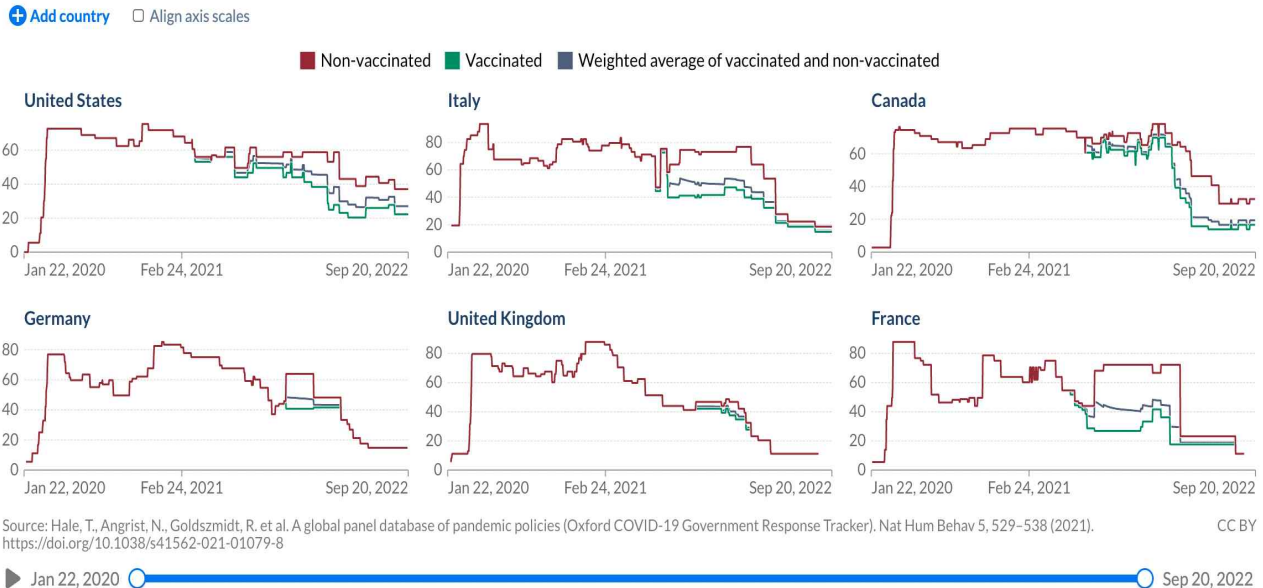
□ Oxford COVID-19 Government Response Tracker²¹⁾

- 지속적으로 정부 대응 정책을 관찰하고 국가별 정책을 비교하기 위해 시작된 프로젝트로, Oxford University team of academics와 Blavatnik School of Government가 주도하는 팀이 데이터를 수집함
 - 코로나와 정부 대응의 관계를 중심으로 데이터와 시각화를 제공함
 - 정부 정책을 세 개의 주제로 나누어 총 20개의 지표를 선정함
 - * (c1-c8): 봉쇄 및 폐쇄 정책, (e1-e4): 경제 정책, (h1-h8): 의료 시스템 정책
- 시계열 형식의 데이터를 제공하며, 국가별 데이터, 히트맵을 활용해 기간과 국가를

21) <https://www.bsg.ox.ac.uk/research/research-projects/코로나19-government-response-tracker>

조작할 수 있는 인터랙티브 시각화를 제공함

- 국가별 데이터 시각화를 통해 기간에 따라 정부 정책과 확진자 수의 관계를 비교할 수 있음
- GitHub를 통해 소스 코드와 데이터 지표를 설명함
- 데이터는 JSON, CSV, SVG, PNG 형식으로 제공함



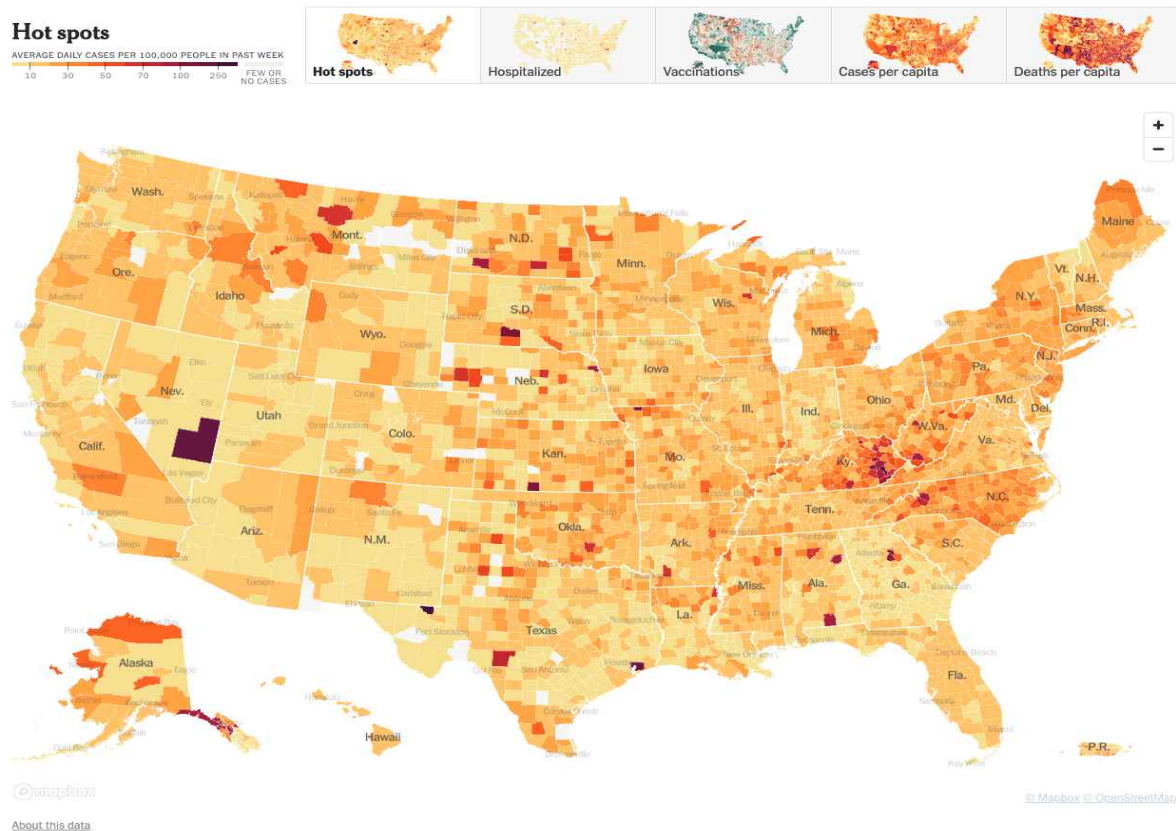
<그림 2-21> Oxford COVID-19 Government Response Tracker의 데이터 시각화(Oxford대학교, 2022)

□ NEW YORK TIMES²²⁾

- 데이터 시각화에 강한 뉴욕 타임즈는 미국 전역의 코로나19 상황을 인터랙티브 그 래프 형태로 다양한 시각화를 제공함
 - 미국 전역의 코로나19 확진자 수와 사망자 수, 연령대별 확진자 비율, 백신 접종 현황, 병 원 입원 사례 등의 8종의 시각화를 제공하며, 직관적이고 컬러풀한 시각화로 보다 쉽게 코로나19 상황을 이해할 수 있게 함
 - 미국의 주를 기준으로 지역을 구분하여 확진자 수와 사망자 수, 신규 입원자 수와 지도 시각화, 전체적인 트렌드 시각화를 제공함
 - 그 외 호주, 독일 등 13개국에 대한 데이터와 시각화 자료도 제공함
- * 데이터와 시각화를 제공하는 13개국: 호주, 브라질, 캐나다, 중국, 프랑스, 독일, 인도, 이탈리아, 일본, 멕시코, 남아프리카, 스페인, 영국
- ‘Track Coronavirus Cases in Place Important to You’ 메뉴를 통해 구독자의 지역 에 따라 코로나19의 상황을 파악할 수 있는 집단 감염 경로 검색 서비스를 제공함

22) <https://www.nytimes.com/interactive/2021/us/covid-cases.html>

- 해당되는 지역의 평균 신규 확진자 수와 새롭게 보고된 사망자의 수에 대한 정보를 텍스트 형태로 시각화와 함께 제공함
- 시각화에 사용한 데이터는 GitHub을 통해 오픈 데이터로 공개되며, 데이터의 메타 데이터에 대한 정의도 기술되어 있음



<그림 2-22> NEW YORK TIMES에서 제공하는 코로나19 확진 heatmap 시각화(NEW YORK TIMES, 2022)

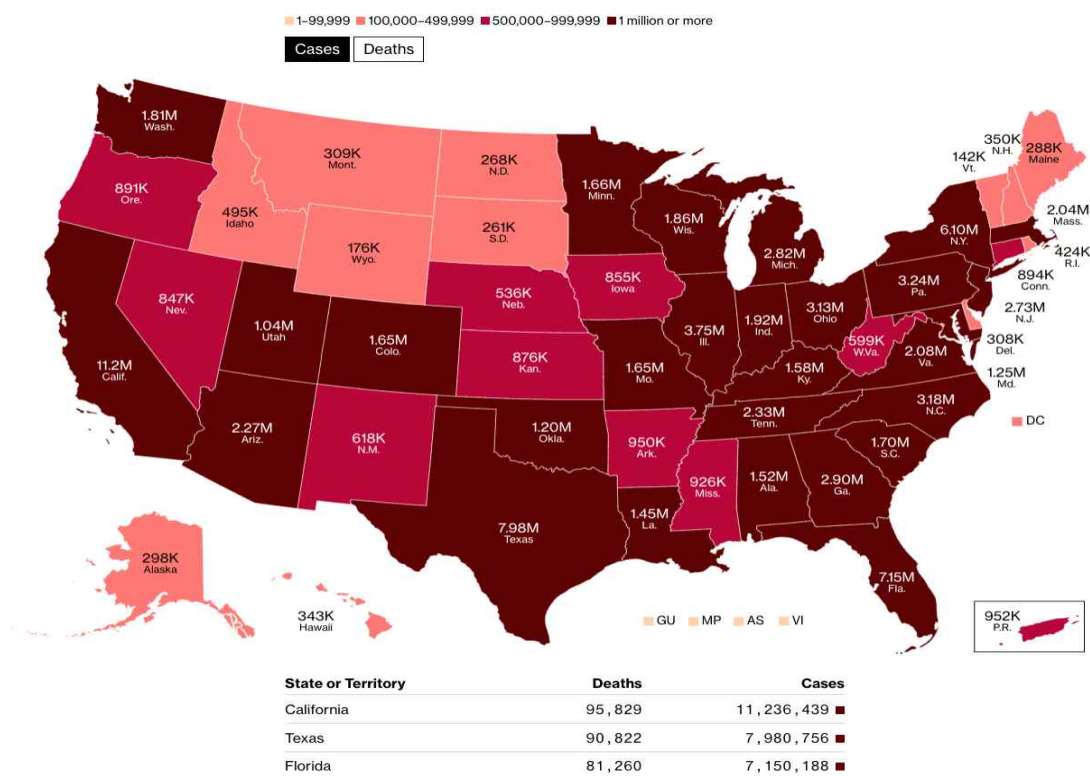
□ Mapping the Coronavirus Outbreak Across the World(Bloomberg)²³⁾

- Bloomberg는 미국의 경제 및 데이터 기업으로 사회 뉴스와 데이터를 제공하며, 국제 코로나19 관련 데이터를 기반으로 주제 분류에 따라 다양한 시각화와 분석결과를 보여줌
 - 사용된 코로나19 데이터의 출처는 Johns Hopkins University Center for Systems Science and Engineering임
 - 데이터의 업데이트는 일별로 진행됨
- Vaccine Tracker, Global Cases, US Cases, US Regions 총 4개의 주제로 데이터가

23) <https://www.bloomberg.com/graphics/2020-coronavirus-cases-world-map/>

구성되어 있음

- Vaccine Tracker 주제는 백신 투여가 코로나19의 확산 감소에 끼친 영향을 타임라인으로 정리하고, 지도 시각화를 이용해 전 세계의 백신 보유 및 계약 현황, 국가 간 백신 격차를 나타내며, 백신 투여 현황과 미국의 주별 백신 투여 현황 내용을 제공함
- Global Cases 주제는 주요 국가의 코로나19 확진과 사망 현황을 테이블 형태의 데이터와 그래프와 지도 시각화를 이용하여 표현함
- US Cases 주제는 미국 각 주의 확진과 사망 사례에 대한 누적 수치를 시각화하고 제공하며 일일 확진과 사망 사례에 대한 수치도 제공함
- US Regions 주제는 미국 각 주에서도 지역을 더 세분화하여 사망자 수를 지도 시각화로 나타내는데, 우편번호나 장소 이름으로도 검색이 가능하다는 특징이 있음



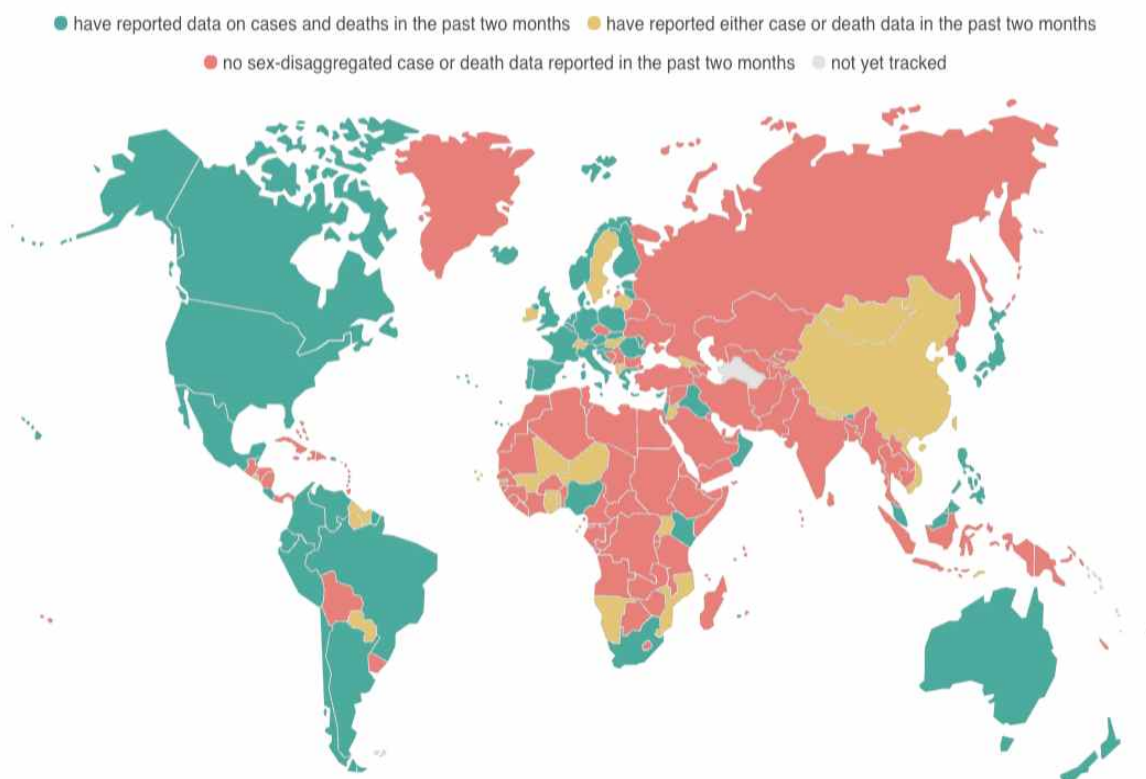
<그림 2-23> Bloomberg의 US Cases 미국 지도 시각화(Bloomberg, 2022)

□ The sex, gender and COVID-19 Project²⁴⁾

- Global Health 50/50, APHRC(African Population and Health Research Center), ICRW(International Center for Research on Women)이 파트너십을 체결하여 성별 분리된(Sex-Disaggregated) 코로나19 데이터베이스 구축을 목표로 함
- 빌 & 멜린다 게이트 재단(Bill & Melinda Gates Foundation)의 후원으로 운영됨

24) <https://globalhealth5050.org/the-sex-gender-and-코로나19-project/#>

- 젠더 차이에 초점을 두어 세계 194국의 코로나19 데이터를 분석한 내용을 바탕으로 기술통계와 성별 확진자 비율, 성별 확진자 진료 양상, 시간에 따른 성별 확진자 양상 등을 인터랙티브 그래프로 시각화하여 제공함
 - 성별 분리된 데이터를 기반으로 월별 및 지역별 보고서를 제공하는데, 2021년 11월 보고서를 마지막으로 업데이트는 중지됨
 - 국가를 검색하여 각 국가의 성별 분리된 데이터의 시각화 내용을 확인할 수 있으며 데이터의 출처는 링크로 연결하여 제공함
 - 시각화에 사용된 데이터는 CSV, XLSX, PDF 형식으로 다운로드가 가능함
- 해당 프로젝트는 전 세계 국가 코로나19 보건 정책에서 성별을 고려했는지 근거 자료를 찾기 위해 Policy Portal이라는 개별 주제로 각 국가의 정책 내용을 분석함
 - 백신, 공중 보건 메시지, 임상 지침, 의료 종사자 보호, 감시체제, 필수 의료 서비스 정책에서 젠더를 고려한 정책을 시행했는지에 대해 지도 시각화를 이용하여 표현함
 - 국가별, 정책별 검색을 할 수 있으며 테이블 데이터 형식으로 각 국가에서 시행하는 정책을 정리함



〈그림 2-24〉 성별 분리된 데이터를 제공하는 국가의 지도 시각화
(The sex, gender and COVID-19 Project, 2022)

□ A Journal of the Plague Year: Share your story²⁵⁾

- Arizona State University의 공공역사기금을 기반으로 코로나19 팬데믹 상황의 개인 기록을 수집, 공유, 보존하기 위해 전 세계의 교수진, 대학원생이 참여하는 시민 참여형 아카이브임
 - 웹사이트에서 시민들이 직접 자신이 보유한 오디오, 비디오, 이미지, 텍스트 자료에 대해 직접 제목과 설명, 해시태그, 개체와 연결된 URL, 관련 지역 등을 입력하여 등록할 수 있으며, 익명 제출도 가능함
 - 입력한 메타데이터는 Dublin Core와 OMEKA CLASSIC을 따라 기술됨
- 아카이브에서 원하는 주제 또는 내용의 자료를 효과적으로 탐색할 수 있도록 상세 검색 기능을 제공하며, 일종의 전시처럼 특정한 주제를 중심으로 관련 자료만 선별하고 큐레이팅하여 제공함
 - 제목이나 설명, 태그, 날짜, 형식, 창작자 등으로 검색이 가능하며, Dublin Core나 OMEKA CLASSIC 등 사용되는 메타데이터 스키마의 포함여부로도 검색이 가능함
- Global Pandemic Map 메뉴는 지도 시각화를 이용하여 전 세계적으로 코로나19와 관련된 사건이 일어났었던 지역의 기록을 열람하는 기능을 제공함
 - 지도의 확대와 축소가 가능하며, 자세하게 축소할수록 관련 지역과 사건에 대한 설명 정보를 얻을 수 있음



<그림 2-25> Journal of the Plague Year에서 제공하는 지도 시각화
(Journal of the Plague Year, 2022)

25) <https://covid-19archive.org/s/archive/page/Share>

2.3 코로나19 웹 아카이브 프로젝트

2.3.1 수집 정책 분석

- 아카이브 구축 기관에게 수집 정책은 법적인 근거와 기록의 일관성을 보장하는 수단이므로 체계적으로 마련될 필요가 있음
 - 법적 고려 사항과 다양한 유형에 대한 수집 기준, 보존 계획을 포함하는 수집 정책 등의 내용을 포함해야 함
 - 체계적인 수집 정책은 다양한 주제의 아카이브를 구축할 때 참고할 수 있는 기준점이 될 수 있음
- 현재 구축되는 코로나19 아카이브는 기존의 아카이브와는 다른 특성을 갖기 때문에 코로나19 아카이브만의 수집 정책이 필요함
- 코로나19의 영향력이 크기 때문에 전 세계적으로 이 주제에 공감하고 영향을 받는 사람들이 많고, 이러한 사람들이 만들어내는 관련 자료의 양이 매우 방대함
- 코로나19는 현재까지도 진행되고 있다는 특징에서 실시간으로 생산되는 기록을 수집하는 과정에 주목할 필요가 있음
 - 국내·외 사례를 보면, 기술이 발전함에 따라 도큐먼트 형식이 아닌 SNS 기록물, 웹 기록물, 데이터 형식 기록물, 동영상 기록물 등 다양한 형식의 기록을 수집하는 사례가 증가함
 - 기록의 수집 과정에서 기존의 도큐먼트의 기증과 이관 방식이 아닌, 적극적인 조사와 이용자 참여를 통한 기록 수집의 사례가 증가하였고 이에 따라 즉각적인 메타데이터 작성 이 필요해짐
- [표 2-1]은 해외 코로나19 아카이브를 구축하는 기관 중, 수집 정책 관련 정보를 제공하는 기관을 대상으로 수집 정책 내용을 분석한 내용임
 - 해외 코로나19 관련으로 구축된 웹 아카이브 중, 성문화된 수집 정책을 보유하고 있는 기관은 약 78.5%로 비교적 높음

[표 2-1] 해외 코로나19의 수집 정책 내 정책 요소 분석(심지연 & 김지현, 2021)

유형	도서관						기록관				
이름	텍사스 알링턴 대학	포덤 대학	프린 스턴 대학	페이 필드 대학	샌프란 시스코 시립 도서관	사스 카툼 주립 도서관	태즈메 니아 기록관	네브래 스카 대학	오레곤 주립 대학	스프링 필드 대학	라이 어슨 대학
아카이브 목적		○	○	○	○		○		○	○	○
이용자											
기록물 범주		○	○	○				○	○		○
수집 참여자	○	○	○			○	○	○		○	
수집방법	○	○	○	○		○	○	○	○	○	
수집절차	○	○	○	○	○	○	○	○		○	
수집매체	○	○	○	○	○	○		○		○	○
수집 우선순위	○										○
개인정보 보호			○	○			○		○	○	○
저작권	○	○	○	○	○	○	○		○	○	○
수집 제외항목	○						○				○

- 가장 보편적으로 포함되어 있는 요소는 수집 매체에 대한 정의와 저작권 관련 사항에 대한 요소로 총 10개의 기관에서 포함하고 있으며, 순서대로 수집 방법(9개), 아카이브 목적(8), 수집 참여자(8), 수집 절차(8), 기록물 범주(7), 개인정보보호(7)로 높았고, 상대적으로 적은 기관이 채택하고 있는 요소는 수집 제외 항목(3), 수집 우선순위 항목(2)임
- 국내에서 코로나19 관련 아카이브를 진행하는 곳은 많지만, 코로나19 아카이브의 특성을 고려하여 수집 정책을 마련하고 명시하는 곳은 거의 없음
- 수집 목적, 수집 범위와 매체, 수집 과정과 절차, 저작권 관련 사항의 범위를 명시한 성문화된 수집 정책을 마련한다면 법적 근거가 될 수 있으며 일관성 있는 수집이 가능함
- 그 외 다른 요소들은 부가적 사항으로 수집 정책에 포함시킨다면 체계적인 수집 정책의 마련이 가능함

2.3.2 주요 국가별 분석

- WARCnet은 웹 아카이빙 기관과 연구자의 네트워크로, 웹 아카이브의 디지털 문화 유산의 활용과 이해를 돕기 위한 목적으로 설립됨
 - 2020년에 활동을 시작하여 2023년까지 활동 계획을 갖고 있으며, 매해 정기미팅을 통해 각자의 결과를 공유함
 - 연구주제를 기준으로 구성원을 Working Group으로 구분하여 총 6개의 Working Group이 있으며, 정기미팅을 통해 서로 연구 성과를 공유하고 그 내용을 논문으로 출판함
 - Working Group2의 주요 연구 주제는 ‘초국가적 사건 분석’으로 세계적으로 많은 영향을 끼친 사건을 중심으로 연구를 진행하고, 그 중 코로나19와 관련된 연구도 포함됨
 - 영국도서관, 프랑스 국립도서관, 프랑스 국립방송아카이브(INA), 룩셈부르크국립도서관, 인터넷보존컨소시엄(IIPC)이 서로 협업하여 국가별 주요 코로나19 아카이브의 구축 방안, 운영 현황에 대한 분석을 진행함
 - 코로나19 관련 아카이브를 구축한 8개국의 9개 기관을 대상으로 구술 인터뷰를 추진하여, 컬렉션의 선택, 큐레이션 과정과 범위에 대해 조사를 진행함([표 2-2] 참고)
 - 결과는 컬렉션의 현황, 접근성과 검색가능성, 파트너십과 활용으로 구분하여 분석내용을 정리함
- * 최종 분석표는 [표 2-3]을 참고

Holownia & Geeraert: Exploring special web archives collections related to COVID-19: The case of the Library of Congress



This WARCnet paper is part of a series of interviews with European web archivists who have been involved in special collections related to COVID19. This is an interview with the Library of Congress.

Kurzmeier, Finegan and Ryan: Exploring special web archives collections related to COVID-19: The National Library of Ireland



This WARCnet paper is part of a series of interviews with European web archivists who have been involved in special collections related to COVID19. The interview is made with The National Library of Ireland

Aasman, Bingham, Brügger, De Wild, Gebeil & Schafer: Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections



This report is the first in a short series of WARCnet papers which aim to provide feedback on an internal datathon conducted by Working Group 2 of the WARCnet project.

De Wild, Kyritsis, Teszelszky & De Bode: Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)



This WARCnet paper is part of a series of interviews with European web archivists who have been involved in special collections related to COVID19. The interview is made with the Dutch Web Archive (KB).

Brügger: The WARCnet network: The first year



This WARCnet Paper provides an overview of the WARCnet network's organisation and of the network's different activities in 2020, including dissemination of results.

Gebeil & Schafer: Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)



This WARCnet paper is part of a series of interviews with European web archivists who have been involved in special collections related to COVID19. The interview is made with the French National Library (BnF).

<그림 2-26> WARCnet 보고서(Warcnet, 2022)

[표 2-2] WARCnet의 Working Group2에서 조사한 웹 아카이브 목록(WARCnet, 2022)

번호	국가	기관
1	미국	미국의회도서관(Library of Congress)
2	프랑스	프랑스 국립방송아카이브(INA)
3		프랑스 국립도서관(Bnf)
4	덴마크	덴마크 웹 아카이브(Netarkivet)
5	헝가리	국립 세체니 도서관(National Szecheny Library)
6	영국	영국 도서관의 웹 아카이브 (British Library - UKWA)
7	스위스	스위스 국립도서관(Swiss National Library)
8	네덜란드	네덜란드 국립도서관의 웹 아카이브 (National Library of Netherlands - Dutch Web Archive)
9	아이슬란드	아이슬란드 웹 아카이브(Icelandic Web Archive)

[표 2-3] WARCnet에서 진행한 주요 국가별 코로나19 아카이브 분석(WARCnet, 2022)

구분	Library of Congress (미국)	INA (프랑스)	Netarkivet (덴마크)	National Szecheny Library (헝가리)	UK Web Archive (영국)	Swiss National Library (스위스)	French National Library (프랑스)	Dutch Web Archive (네덜란드)	Icelandic Web Archive (아이슬란드)
인터뷰 날짜	2022-1-5	2020-7-13	2020-9-14	2020-8-13	2020-9-3	2020-9-14	2020-10-15		
인력	. 10명 (현재 8명)			. 2명 (Full-time) . 다른 프로젝트 동시 진행	. 9명 (Full time, British Library) . 수집은 자발적인 참여 기반		. 52 명 (BC Web App 참여) . 전체 약 80명 (아카이브 수집)		. 1명 (주 2-4시간)
범위	. 2,000여개의 웹사이트에서 코로나19으로 태그된 곳 수집 . 뉴스 수집 사이트는 대부분 포함되어 있음	. 시청각 콘텐츠에 집중해서 수집 . 유튜브 포함, 트위터 중심 . 페이스북, 인스타그램은 수집하지 않음 . Trends24를 사용해서 해시태그 분석과 모니터링 . 실시간 API가 제한되어 있어 모든 데이터를 수집하지 못함 . 데이터 수집을 위해 여러 개 계정을 사용하고 있음	. 코로나19 이벤트 크롤러: Royal Danish Library와 Danish cultural institutions 의 협업으로 진행 . 덴마크 사회에 영향을 주는 주제를 포함하는 모든 웹사이트를 대상으로 수집 . 이벤트 크롤러 방식으로 수집	. 웹사이트, 해시태그 등 수집 . 헝가리와 관련된 콘텐츠	. 페이스북 제한적으로 수집 . 트위터 해시태그 수집 . 영어를 중심으로 수집 . 영어 이외의 언어로 된 콘텐츠 수집/큐레이션은 큐레이터에 따라 수준이 다름 . STEM 분야는 웹 아카이브에서 제대로 표현되지 않는 분야이며, 관련 전문성이 부족함	. 전체 웹사이트를 대상으로 수집 . 법적 이유로 소셜미디어는 수집하지 않음 . 코로나19을 위한 웹사이트 . 스위스에서 출판되거나, 스위스 관련 콘텐츠 . 기수집된 아카이브서 코로나19 관련 콘텐츠 . 뉴스는 핵심 사이트 중심으로 수집	. .fr 도메인 한정 프랑스 웹 수집 . 소셜미디어 . 기관 웹사이트 . 지방자치단체 웹사이트 . 블로그와 유튜브 . 해시태그 . 새롭게 생성된 웹사이트 . 미디어 웹사이트는 2단계로 진행 (수집 빈도를 높임)	. 네덜란드의 웹사이트에서 네덜란드어로 출판된 자료 한정 . nl 도메인에 한정하지 않음 . 네덜란드 시민에 의해 출판된 모든 웹사이트 포함 . 온라인 신문은 수집하지만, 유평사이트는 제한적으로 수행 . 매일 수집 . 소셜미디어 관련 프로젝트가 있지만, 소셜미디어 전문가는 부재	. 뉴스 미디어 웹사이트 . 아이슬란드 웹사이트 . 영어, 노르웨이어 언어 일부 포함 (핀란드어 제외) . 검색 키워드: “iceland“와 COVID 관련 용어
수집 내용	. 우선순위: 연방, 주, 지방 및 토착 법률, 공공 정책,	. 1.2억건 이상의 트윗이 수집되어 있음	. 2,300개 이상의 덴마크 웹사이트 사이트 포함	. 텍스트 중심의 콘텐츠에 집중 . 국가, 지역,		. 문화, 스위스의 . 주제: 일상, 뉴스 협회,	. 페이스북, . 트위터, 비디오 수집을 위해	. 코로나19 관련 디지털북 . 주제: 여행,	

	과학, 비즈니스 및 문화 콘텐츠	. 트위터에 언급된 250,000개 이상의 비디오	. 17개 이상의 팟캐스트 . 250개 이상의 페이스북 프로필 . 2,200개 이상의 트위터 계정 . 400개 이상의 인스타그램 계정 . 1,800개 이상의 Reddit 포스트 . 300개 이상의 영상, 채널 . 5,000개 이상의 Twitch 채널	국제 뉴스 포털에서 수집 . 헝가리, 주변 국가 (슬로바키아, 루마니아)		자원봉사, 의료 기관, 직원 조직, 행정, 전문가, 스포츠, 교육, 전문 조직, 정치, 농업 등	Heritrix 이용 (페이스북은 10~15% 성공률로 비교적 낮음)	정부, 경제, 의학, 네덜란드 스포츠, 평론, 교육, 보험 사회, 종교 등	
소셜미 디어 수집	. 검토 중 . Asian Division에서 소셜미디어 콘텐츠를 추가하고 있음		. 수집하고 있지만, 기술적 문제로 제대로 수집하지 못하고 있음 . 유튜브, 페이스북	. 수집하지 않음 (소셜미디어 수집 도구가 없음) . 인스타그램을 수집했으나 관련 없는 내용임 . 페이스북 수집을 위해 Webrecorder를 테스트했지만, 오류 발생	. 페이스북, 인스타그램의 일부 웹사이트를 Webrecorder 또는 Conifer 툴을 이용해서 수집 . 코로나 컬렉션을 위한 인력이 없음			. 수집하지 않음 (전문가 없음)	. 수집하지 않음 (기술적 문제)
국제적 데이터 수집		. 해시태그 기준으로 수집, 언어 종속성 없음 . 프랑스어가 아닌 콘텐츠도 포함	. 덴마크의 코로나19에 집중 . 덴마크 관련 국제 뉴스 수집을 일부 진행 . Legal Deposit 참고 (이미지)	. 언어 기준으로 수집 . 헝가리어로 된 국제 이벤트 . 헝가리의 일반 생활에 영향을 미치는 이벤트	. UK 중심, 기관과 개인에게 주는 영향에 집중 . 영국 웹사이트에 접근해서 수집 . 도메인: .uk, .scot, .wales . 자동으로	. 글로벌한 특징 . 스위스 관련 콘텐츠 집중 수집 . 스위스 도메인에 한정하지 않음 . 독일어, 이탈리아어, 프랑스어	. 글로벌한 특징 . 프랑스 언어권이 포함될 수 있음 . 프랑스에 집중된 관점을 갖고 있음	. 국제적 관점은 적음 . Frisian 언어를 사용하는 도메인 선정해서 수집 . KB 아카이브는 큐레이션을 기반으로 하기 때문에 중요한	

					Geo-IP 테스트를 수행하고 수집 범위 결정	콘텐츠도 수집		웹사이트를 선정해서 수집	
수집주 기	. 매일 수집X . 일별, 월별, 분기별로 진행 . 대부분은 주별로 수집		. 뉴스 미디어는 일별 . 업데이트 주기는 웹사이트에 따라 다름	. 중요한 정보는 주별 . 컬렉션마다 상이	. 컬렉션마다 상이	. 1년		. 1회 / 1년 (필요시 수집) . 코로나19 컬렉션은 주기적으로 수집	
소프트 웨어								. Heritrix crawler	
URL 포함			. URL, 비디오 URL, 날짜, 출판자			. URL, 웹사이트의 제목, 언어, DDC, 기관 등 포함			. 일부 자원에 URL 포함
수집사 이트									. 83개 사이트 (일주일, 5일기준)
시작일	. 2020년 7월 (공식적으로 개시)	. 2020년 3월 13일	. 2020년 1월 . IIPC 이벤트 컬렉션: 2020년 2월 중순 시작	. 2020년 2월	. 2020년 3월	. 2020년 3월	. 2020년 1월	. 코로나19 발병과 함께	2020년 3월
종료일			. 정해지지 않음	. 현재 없음	. 현재 없음	. 현재 없음		. 2020년 까지로 설정, 그러나 진행 중	. 종료시까지
품질관 리	. 1년 단위로 평가	. 모니터링 도구로 매일 체크 . 인터페이스를 통해 시각화	. 샘플을 추출해서 관리 . 큐레이터가 크롤링 로그와 수집된 콘텐츠의 샘플 체크	. 인력이 부족해서 진행하지 못함	. 서비스 모니터링 . Alert으로 수집 실패한 사이트 정보를 얻음 . Wayback을 이용해 개별 웹 콘텐츠 확인 가능	. 수작업으로 진행 . 품질이 좋지 않으면 재수집	. 컬렉션에 따라 품질이 다름 . 감염병 온라인에 대해 Wayback Machine으로 품질 체크 수행	. 수집 전에 web archivability 확인 . 스페셜 컬렉션에 대한 최초 수집의 내용 확인, 이후 정기적으로 품질 체크 진행	. 진행하지 않음 (수집문제, 웹 페이지 접근이 불가능한 일부 경우에 확인)
컬렉션 규모				. 120개의 seed URL	. 6,000개 이상의 웹사이트 (600개 트위터 계정 포함)	. 약 500개 웹사이트를 대상으로 수집 . 100개	. 5,000개 URL로 수집 시작 . 비디오, 온라인 뉴스를 포함할	. IIPC와 KB에 의해 수집된 345개의 웹사이트 . IIPC가 크롤링한	

					<div><div><div><div><div></div><div>기술적 문제로 비디오는 명시적 링크로 제공되는 경우 외 수집X</div><div>이미지, HTML 태그, 웹 문서</div><div>구술 자료 포함</div></div></div><div><div><div><div></div><div>웹사이트는 코로나19 컬렉션을 위해 신규로 생성</div><div>컬렉션 규모: 약 1.75 TB</div></div></div></div></div></div>	<div><div><div><div></div><div>경우, 15TB 데이터 또는 2.75억 URLs</div><div>1,014 컬렉션 태스크</div><div>15,504 WARC 파일</div></div></div></div>	<div><div><div><div></div><div>96개의 웹사이트</div><div>KB가 크롤링한 41개의 웹사이트</div></div></div></div>		
이슈			<div><div><div><div></div><div>수집한 데이터는 메인 웹 아카이브에 포함되지 않아서, 인덱스 되지 않고 검색이 불가능함</div></div></div></div>				<div><div><div><div></div><div>소셜미디어는 동의없이 수집 불가</div><div>코로나 관련 앱 수집 불가 (메타데이터만 수집)</div><div>js, 임베딩 콘텐츠 등이 포함된 웹 페이지는 수집 제한</div><div>웹사이트 소유자가 수집을 거절하는 경우</div></div></div></div>	<div><div><div><div></div><div>수집 대상을 선정하는 데 기준 없음</div><div>인력 없음</div></div></div></div>	
모니터링								<div><div><div><div></div><div>결과만 확인</div></div></div></div>	
접근 권한	<div><div><div><div></div><div>최초 450개 웹 아카이브 제공 (1년 제한 정책)</div><div>수집 레코드는 월 기준 정책 적용</div></div></div></div>	<div><div><div><div></div><div>INA reading room에서 접근 가능</div></div></div></div>	<div><div><div><div></div><div>Netarkivet 콘텐츠와 동일한 규칙</div><div>덴마크 연구소와 관련 있어야 함</div><div>SolrWayback 설치 준비하고 있음</div></div></div></div>	<div><div><div><div></div><div>저작권 이유로 공개 접근 불가</div><div>서비스 환경이 충분하지 않음</div><div>제한적으로 도서관에서 접근 가능</div></div></div></div>	<div><div><div><div></div><div>reading room에서 URL 접속시 열람 가능</div><div>품질보증 작업을 진행할 예정,</div><div>웹사이트로 공개하지 않음</div><div>메타데이터 표준화 작업 진행</div></div></div></div>	<div><div><div><div></div><div>파트너에 한정해서 reading room에서 URL을 통해 접근 가능</div><div>웹사이트는 전문 검색 지원</div><div>재생산은 불가 (법적 이유)</div></div></div></div>	<div><div><div><div></div><div>BnF interactive terminal을 통해 접근</div></div></div></div>	<div><div><div><div></div><div>KB 아카이브: 도서관 네트워크로 접근 가능</div><div>개인 연구 목적의 활용 가능</div><div>재생산은 조건부 허용</div><div>IIPC 코로나19 : 자유롭게 이용</div></div></div></div>	<div><div><div><div></div><div>URL을 통해 접근 가능 (https://vefsafn.is/)</div><div>기본 URL 검색 제공</div><div>접근 제한 없음</div><div>모두에게 공개</div></div></div></div>

					예정			가능	
데이터 분석과 활용		. 향후 계획이 있음	. 2개 프로젝트에서 접근 권한을 확보	. 없음	. 없음	. 없음	. 관심은 많지만 구체적인 프로젝트는 없음	. 프로젝트 2개 진행	. 없음
홍보	. 언론보도 . 소셜미디어 . 캠페인		. 언론 홍보 . 웹사이트에 뉴스 홍보 . Royal Danish Library 페이스북 계정 활용	. 언론보도 . Hungarian Library Association과 온라인 행사 개최	. Digital Preservation Coalition 블로그 . UK JISC mail web archiving list, Welcome Trust 활용 . British Library에 홍보 . UK Web Archive 트위터, 블로그에 홍보	. 외부 홍보 없음	. BnF 블로그, Web Corpora에 홍보	. 팟캐스트 인터뷰 진행: 수집방법과 내용 설명 . 언론보도	. 없음
파트너 십		. IIPC와 협력할 계획 있음	. Archive-It		. Royal College of Nursing in the UK. Royal Collages of Health and Medicine의 the Welcome Trust Network . Public Record Office in Northern Ireland Trinity College Dublin Public Record Office of Northern Ireland	. IIPC . cantonal libraries . Web Archive Switzerland 멤버	. IIPC, Archive-It	. IIPC . the Digital Heritage Network in the Netherlands	. IIPC

- 웹 아카이브의 범위와 수집 절차를 중심으로 8개국의 아카이브를 조사한 결과, 도큐먼트 형식과 함께 대규모 소셜미디어 데이터를 수집하고자 노력하는 모습을 알 수 있음
 - 코로나19는 전 세계 사람들의 생활에 많은 변화를 이끌었고 공통의 공감대를 형성하였기 때문에 소셜미디어에서 많이 다뤄지며 대량의 데이터를 생산함
 - 그 동안의 아카이브 방식은 소셜미디어보다 도큐먼트 형식, 웹 페이지 수집에 집중해왔기 때문에 법률적, 기술적으로 소셜미디어의 수집 기반이 아직 갖춰지지 않았고, 이 문제는 많은 아카이빙 기관이 갖는 공통의 문제임
 - 대중들이 소셜미디어를 통해 생산하는 데이터의 양은 시간이 흐를수록 더 증가할 것이고, 향후 아카이빙 작업의 필요성도 지속될 것이기 때문에 기반을 다지는 작업이 필요함
- 인터넷보존컨소시엄(IIPC)과 WARCnet처럼 아카이빙 필요성을 느끼는 기관과 사람들 사이의 네트워크를 형성하여 소통하고 있음
 - 웹 아카이브 구축에 참여하는 종사자와 연구자들이 모여서 서로의 문제점과 결과물을 공유하면서 완성도 있는 웹 아카이브를 구축하고자 노력하고 있음

2.4 ‘코로나19: 우리의 기억’

2.4.1 개요

□ 코로나19 디지털 아카이브 프로젝트

- ‘코로나19: 우리의 기억’은 코로나19 팬데믹으로 인한 우리 사회의 변화를 가치중립적으로 기록하고, 미래의 참고를 위해 코로나19 관련 데이터를 보존함
 - ‘코로나19: 우리의 기억’은 2020년 5월부터 진행한 디지털 아카이브 프로젝트로, 중앙대학교 문헌정보학과 김학래 교수와 팀.케일리(중앙대학교 사회과학대학 재학생으로 구성)가 참여함
 - 주요 활동은 감염병에 대한 사회과학적 문제 정의와 이를 해결하기 위한 공학적 접근, 구성원의 집단지성을 통해 문제의 해결 과정을 포함하고 있음. 주요 기록은 코로나19의 감염과 확산 현황을 분석하고, 사회 전반에 커다란 영향을 미치거나 일상생활의 변화를 이끌었던 사건을 포함함
- ‘코로나19: 우리의 기억’은 민간이 구축한 아카이브를 국가도서관에서 영구보존하는 국내 최초의 관학협력 사례임
 - 팀.케일리는 행정안전부가 2020년 7월 30일에 주최한 정부혁신 해커톤(개발끝장대회)에서 ‘코로나19: 우리의 기억’ 프로젝트로 커뮤니티상을 수상하였고, 2021년 3월 6일에 개최된 오픈데이터데이(Open Data Day)에서 ‘코로나19 데이터 분석 A to Z’를 주제로 발표를 진

행함

- 국립중앙도서관에 기증된 ‘코로나19: 우리의 기억’ 디지털 아카이브는 2019년 코로나19의 첫 확진자 발생 이후부터 2020년 2월까지 코로나19와 관련된 국내외 데이터를 포함하고 있음. 국립중앙도서관은 ‘코로나19: 우리의 기억’을 포함해 범국가 차원의 코로나19 아카이브 구축을 추진할 예정임



<그림 2-27> ‘코로나19: 우리의 기억’ 중앙대 아카이브 기증식

2.4.2 주요 특징

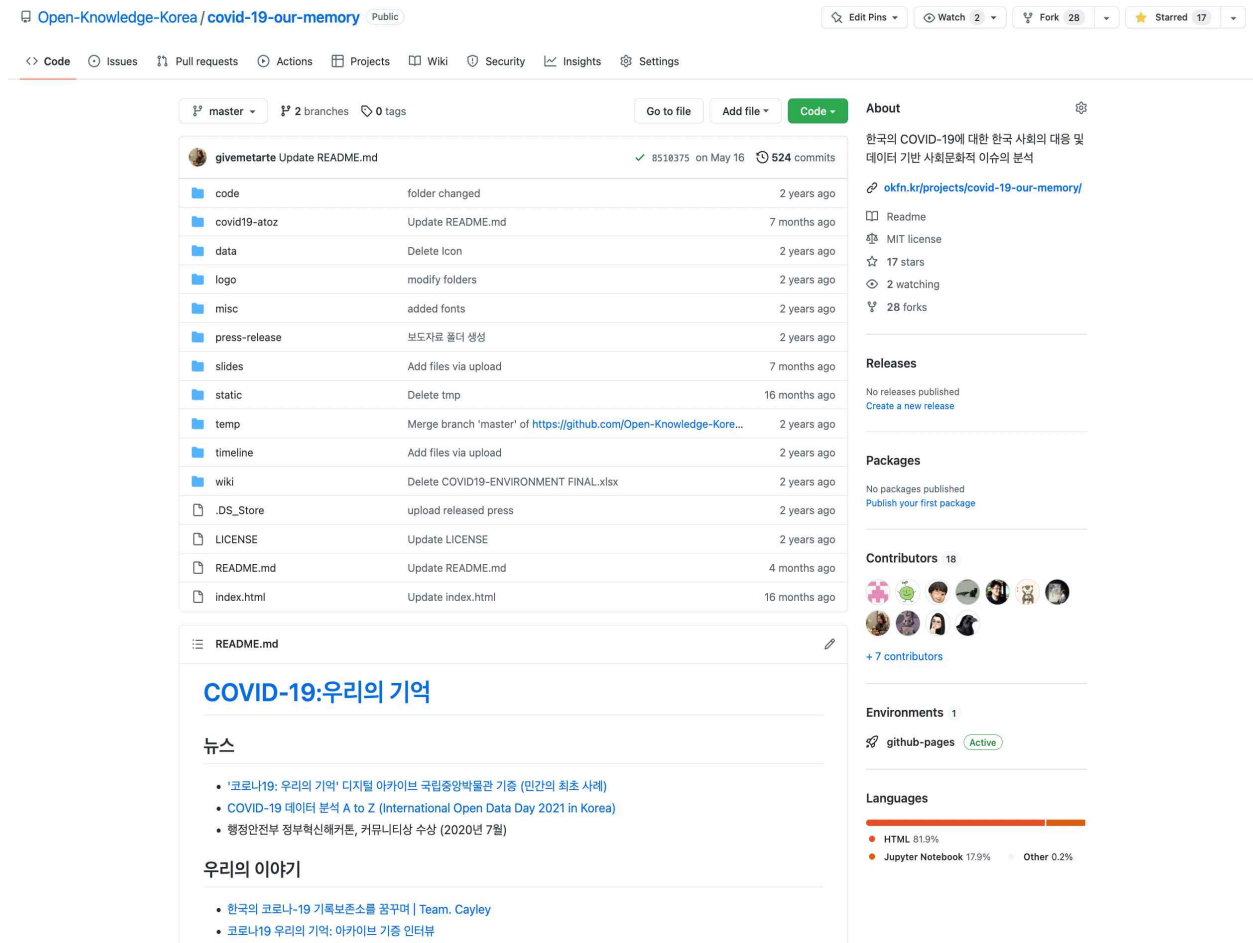
□ ‘코로나19: 우리의 기억’의 세 가지 원칙

- ‘코로나19: 우리의 기억’은 디지털 아카이브로서 사실 정보를 기록하고, 기록의 검증과 재현을 위해 세 가지 원칙을 정의하고 있음

[표 2-4] ‘코로나19: 우리의 기억’의 원칙(코로나19: 우리의 기억, 2022)

- ① 가치중립적 디지털 기록(Value-neutral Archiving)
- ② 지속 가능한 협업 모델(Zero-contacted, but Collaborative)
- ③ 오픈 데이터 원칙(Open Data First)

- **(가치중립적 디지털 기록)** 모든 기록은 구성원의 논의를 거쳐 협업적이고 중립적인 관점에서 선정되고 기술됨
 - 코로나 바이러스 감염증은 사회 전반에 큰 영향을 미치고 있고, 역사적 사건이 아닌 현재 진행형인 사건임. 역사적 사건의 기록을 해석하는 것과 달리, 현재 진행 중인 사건은 가치중립적이고 객관적으로 기록되어야 함
 - 디지털 아카이브의 주요 주제는 정부 정책과 대응, 경제, 교육, 문화, 사회, 환경으로 구분함. 세분화할 수 있는 주제(예: 의료, 정치, 노동, 인권, 여성)가 있지만, 개별 이슈의 지속성과 빈도를 고려해 세분화하지 않고 주요 주제에 포함함
 - 주요 주제별로 관련된 이벤트(사건 또는 이슈)를 수집하고, 핵심 이벤트는 수집된 이벤트에서 사회적 영향과 파급력을 고려하여 선정함. 선정 과정은 각 주요 주제팀의 1차 논의와 전체 구성원이 참여한 2차 검증으로 진행되고, 주제별로 균형 있는 분포를 포함하도록 주제별 필수 이벤트를 5-6개로 제한함
- **(지속 가능한 협업 모델)** ‘코로나19: 우리의 기억’은 비대면 환경에서 지속 가능한 협업 방식으로 진행됨
 - ‘코로나19: 우리의 기억’은 사회적 거리두기가 엄격히 지속되는 기간에 진행되었음. 개별 팀과 전체 구성원의 정기회의는 온라인 환경에서 비대면으로 진행됨
 - 슬랙(Slack)은 구성원 사이의 정보 공유를 위한 목적으로 사용하고, 구글 드라이브(Google Drive)는 아카이브에 활용되는 모든 자료를 공유하는데 사용됨. 매주 진행되는 정기회의는 줌(Zoom)을 사용함
- **(오픈 데이터 원칙)** ‘코로나19: 우리의 기억’에서 수집 또는 구축한 데이터는 오픈 데이터 원칙을 적용하여 모두 공개하는 원칙을 적용함
 - ‘코로나19: 우리의 기억’의 목표는 단편적이고 일시적인 프로젝트를 넘어 지속가능한 협력을 이끌어내는 것을 지향함. 수집된 원시 데이터, 기록된 모든 데이터, 데이터 분석을 위해 개발된 모든 소스 코드는 대중에게 공개함
 - 모든 자료는 깃허브(GitHub)를 통해 공유하고, 특수한 사례를 제외하고 크리에이티브 커먼즈-저작권 표시 대한민국 라이선스(CC-BY 2.0 KR)를 적용함

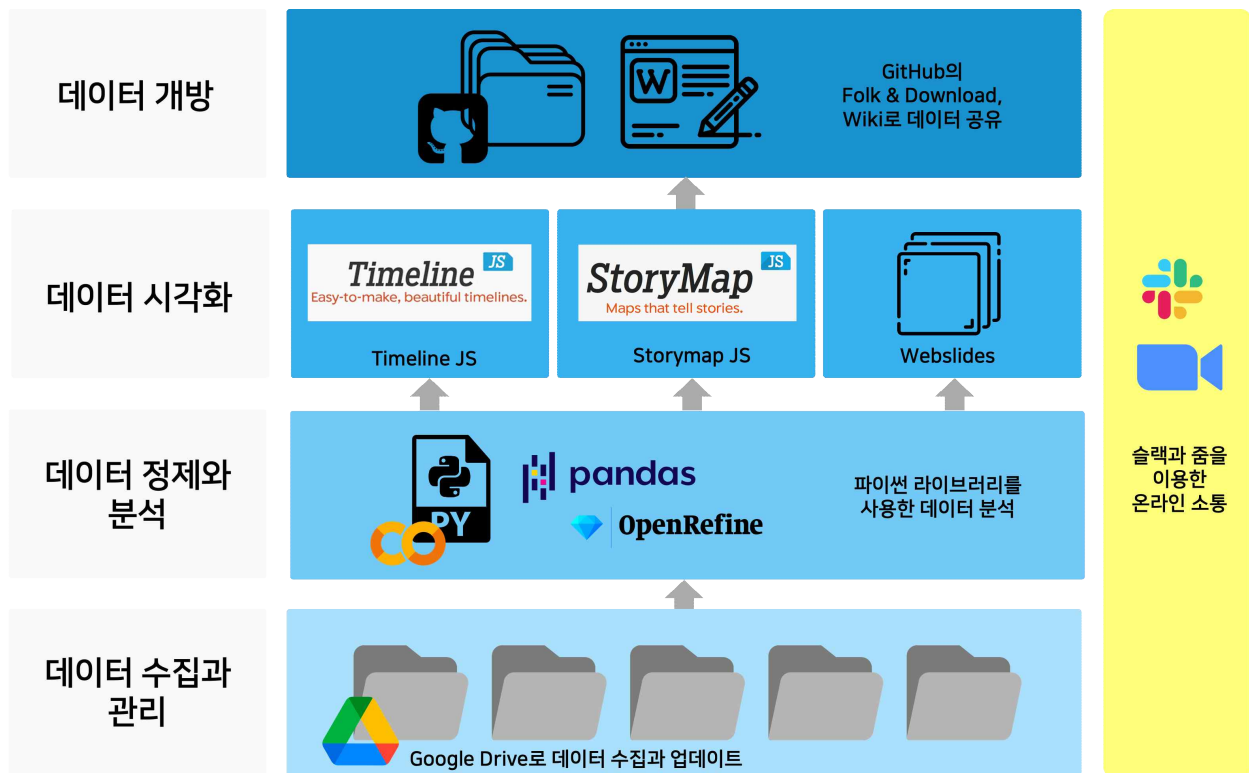


<그림 2-28> ‘코로나19: 우리의 기억’의 깃허브 저장소(‘코로나19: 우리의 기억’의 깃허브, 2020)

□ 데이터의 재사용을 고려한 공유 체계

- ‘코로나19: 우리의 기억’은 데이터의 공유와 재사용이 가능한 온라인 환경을 구축하고, 데이터의 수집과 관리, 분석과 시각화, 개방의 모든 단계에서 오픈소스를 활용함
- (데이터 수집과 관리) 다양한 유형의 데이터는 구글 드라이브에 저장되고, 변경되거나 업데이트된 데이터도 지속적으로 구글 드라이브에서 관리됨
- (데이터 정제와 분석) 파이썬 라이브러리 또는 오픈소스를 사용하여 데이터 분석을 함. 구글 드라이브에 내장된 구글 콜랩(Google Colab)으로 코드를 작성함. 데이터는 파이썬의 판다스(Pandas)를 사용하거나 오픈리파인(OpenRefine)을 사용하여 정제함
- (데이터 시각화) 데이터를 분석한 결과는 오픈소스를 활용하여 시각화함. 코로나19와 관련된 뉴스 타임라인은 Timeline JS를 사용했고, 첫 확진자부터 코로나19의 확산 동선을 추적한 시각화는 Storymap Js를 활용함. 그 외의 다양한 데이터 분석은 파이썬의 라이브러리(예: Matplotlib, Plotly)를 활용하고, 웹 슬라이드(WebSlides)로 데이터 분석 결과를 공유함

- (데이터 개방) 구축된 모든 데이터와 소스 코드는 깃허브 레포지토리로 개방되고, 포크(Folk)와 다운로드를 통해 모든 자원을 재사용할 수 있음. 깃허브 위키는 이용자에게 데이터를 사용할 수 있는 방법에 대한 가이드라인을 제공함
- (온라인 소통) 슬랙과 줌을 통해 활발히 소통함. 슬랙은 협업 도구로 만들어진 메신저 어플리케이션으로, 공지와 작업 진행 상황이 모두 슬랙을 통해 이루어짐. 줌은 화상회의 도구로, 매주 진행 상황을 공유하는 회의가 줌으로 진행됨



〈그림 2-29〉 ‘코로나19: 우리의 기억’의 데이터 공유 체계(김학래, 2022)

2.4.3 데이터 수집과 분석

□ 데이터 수집

- 코로나19가 사회 전반에 미친 영향을 분석하기 위해 정부 정책 및 대응, 교육, 문화, 사회, 산업, 경제, 환경으로 구분하여 총 1,511건*의 데이터세트를 수집함

* [표 2-5]는 주제 분야별로 수집한 데이터세트, 멀티미디어, 문서 등의 수량을 나타낸 것임

- 정부 정책 및 대응 데이터는 정부 정책브리핑과 확진자 데이터, 코로나19 관련 기사를 수집하여 정책의 변화를 분석함
- 교육 데이터는 전국 지자체의 교육 정책을 분석하기 위해 각 학교의 등교 조정 실행 기간을 조사하고 교육부의 등교수업 현황 보도자료를 수집함
- 문화 데이터는 영화, 공연, 여행 산업의 변화를 분석하기 위해 영화관입장권통합전산망, 공연예술통합전산망, 한국관광공사, 한국공항공사의 데이터를 수집하고, 비대면 문화 활동

을 분석하기 위해 서울특별시의 온라인과 오프라인 대출 데이터를 수집함

- 사회 데이터는 재난 불평등을 분석하기 위해 산업체별 확진 현황 데이터와 청와대 청원 홈페이지에서 코로나19 관련 데이터와 사회 분야의 기사를 수집함
 - 산업과 경제 데이터는 고용, 소비, 주식을 중심으로 경제 분야를 분석하기 위해 통계청, 한국거래소, 한국은행, 뉴스·기사 데이터를 수집함
 - 환경 데이터는 미세먼지 데이터, 폐기물 배출량 등의 데이터를 수집하여 코로나19가 환경에 준 영향을 다각적으로 분석함
- 정부와 보건복지부에서 개방하는 코로나19 관련 공공데이터의 활용성을 높이기 위해 데이터 수집과 분석방법을 공개함. 코로나19 관련 데이터 API, 집단감염 정례브리핑, 국외 백신 현황을 중심으로 분석함
- 글로벌 오픈데이터데이의 ‘COVID19 데이터 분석 A to Z’에서 3개의 세션을 통해 코로나19 관련 국내외 데이터의 분석 결과와 소스 코드를 공유함
 - 세션 1은 공공데이터 포털에서 공개하는 오픈API의 정확도를 개선하고 다양한 시각화 분석을 제공함. 세션 2는 집단감염을 주제로 정례브리핑에서 공개하는 집단감염 데이터(hwp, pdf)를 기계가 처리할 수 있는 형태(csv)로 변환하여 데이터 분석을 진행함. 세션 3은 국내외 백신 수급 현황과 정책을 다각도에서 분석해 우리나라 백신 정책의 방향성을 제시함

[표 2-5] 주제별 수집 데이터 수량(코로나19: 우리의 기억, 2022)

대분류	소분류	데이터수량	대분류	소분류	데이터수량
문화	도서관	24	산업	숙박	21
	영화	5		업종별 통계	54
	OTT	2		업종별 매출	16
	공연	10		통계청	7
	정책	3	환경	대기오염	8
경제	가계동향	14		미세먼지	11
	고용	29		석유 소비	2
	경제 통계	25		온실가스	2
	재난지원금	2		항공	9
	주식	30		지역별 교통량	19
	인플레이션	7		재활용품	2
	배달업	5	보건 정책	정례브리핑	178
	자영업	46		확진자	258
교육	대중교통	2		주요 뉴스	309
	대면수업	82		공공 마스크	66
사회	청원	14		정부 보도자료	184
	기타	20		공공 마스크 통계	45

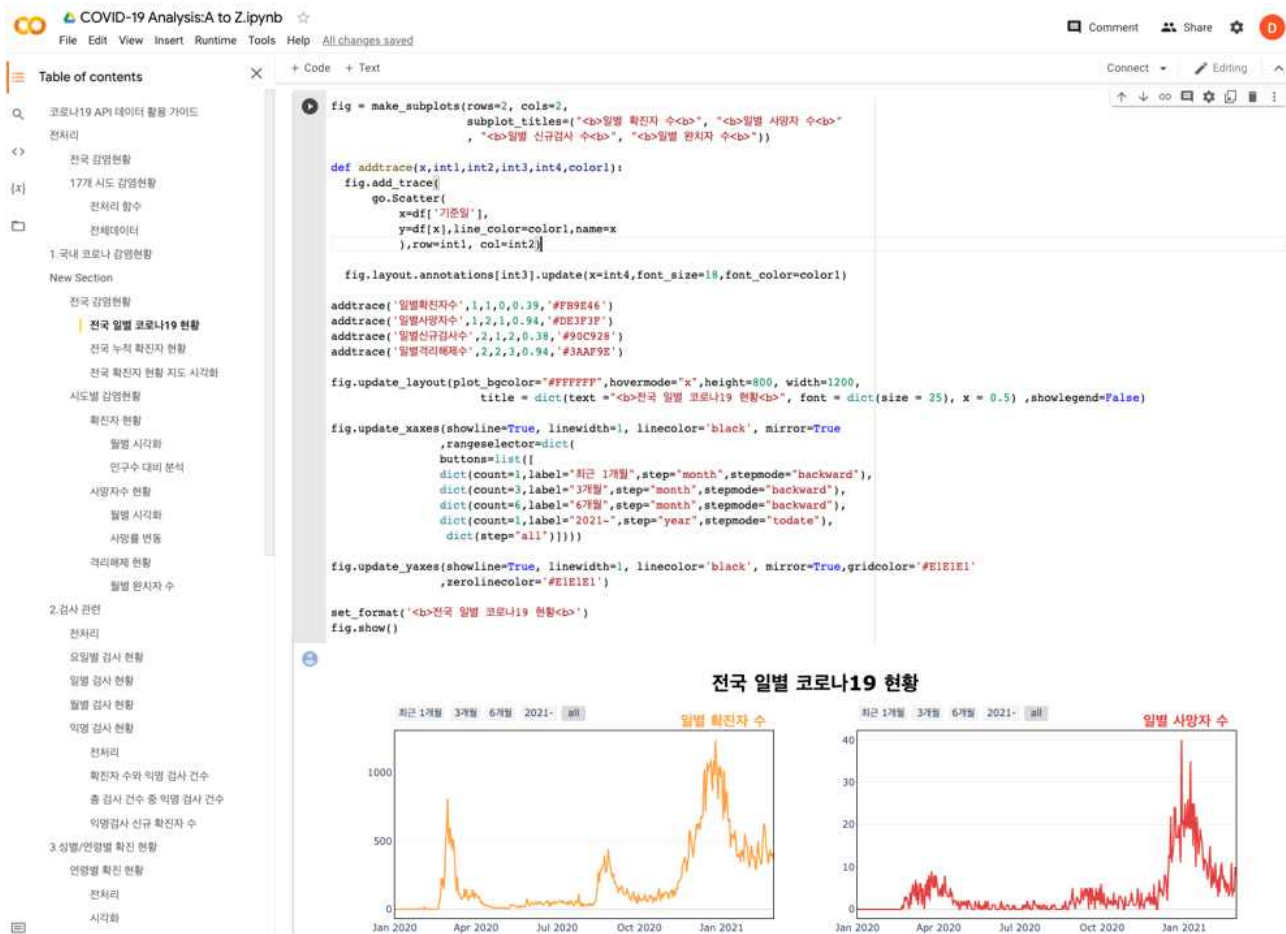
□ 데이터 분석 방법

- 수집한 코로나19 데이터는 [표 2-6]와 같이 뉴스·기사, 데이터세트, 데이터 분석, 데이터 시각화 형태로 활용됨
 - 뉴스·기사는 웹 페이지 형태로 URL과 내용을 요약한 데이터를 수집하여 코로나19의 주요 이슈를 분석하는 데 사용됨
 - 확진자 현황 데이터를 비롯한 주제별 통계 데이터는 데이터 분석과 시각화의 원천데이터로 사용되고, 분석 도구는 구글 코랩에서 파이썬 라이브러리를 활용함
 - 데이터 시각화는 코로나19 이슈를 시계열로 분석한 타임라인과 위치 기반의 스토리맵으로 구성함

[표 2-6] 코로나19 관련 수집 데이터(코로나19: 우리의 기억, 2022)

구분	내용
뉴스·기사	주요 주제별로 수집한 뉴스 데이터
데이터세트	감염 현황 및 통계, 주요 이슈별로 수집한 모든 데이터
데이터 분석	주제별 분석한 소스 코드
데이터 시각화	타임라인, 스토리맵을 포함하는 데이터 시각화

- (Google Colab) 구글 코랩은 웹 브라우저에서 파이썬을 실행할 수 있는 서비스로, 클라우드 기반의 주피터 노트북(Jupyter Notebook) 개발 환경을 제공함.
 - 별도의 설치 없이 데이터 분석에 사용되는 파이썬 라이브러리(Pandas, Numpy, Matplotlib 등)를 기본적으로 제공함
 - 온라인에서 공유 가능한 환경으로 여러 사람의 협업이 가능하고, 구글 드라이브와 연동하여 외부 링크로 공유할 때는 편집 권한을 제한하고 공유할 수 있음
 - ‘코로나19: 우리의 기억’ 아카이브의 데이터 시각화는 [그림 2-30]와 같이 구글 코랩을 사용해 분석하고, 깃허브에 소스 코드를 공유함



<그림 2-30> Google Colab을 활용한 데이터 분석(코로나19: 우리의 기억, 2020)

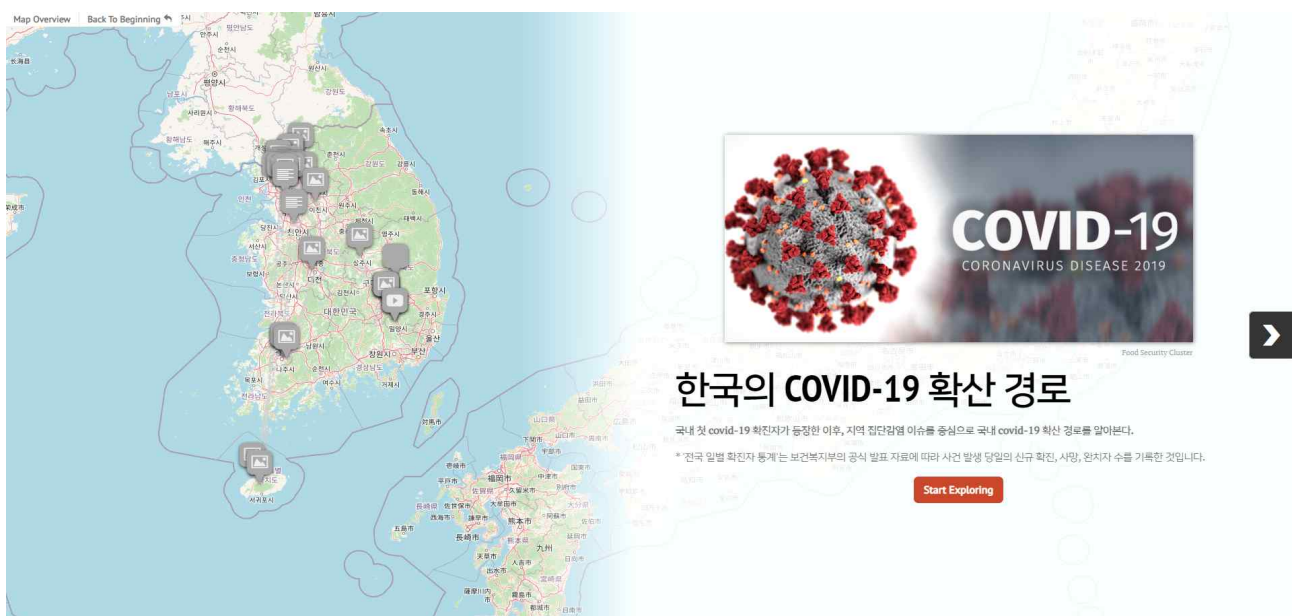
2.4.4 데이터 시각화

□ 주제별 코로나19 영향 분석

- (정부 정책 및 대응방안) 정부발표, 방역, 확진자와 관련된 이슈를 수집하여 미디어 자료와 함께 정부의 정책과 대응방안을 시간의 흐름에 따라 살펴볼 수 있음
- ‘코로나19: 우리의 기억’ 타임라인은 범국가적 재난 상황에 대응하는 정부의 다양한 정책과 이슈를 문화, 경제, 환경, 교육 등의 주제별 이슈에 해당하는 자료와 함께 시간순으로 시각화함
- ‘코로나19: 우리의 기억’ 스토리맵은 지역 집단감염 이슈를 중심으로 국내 코로나19의 확산 경로 위치 기반으로 시각화하고, 관련 뉴스·기사를 제공함

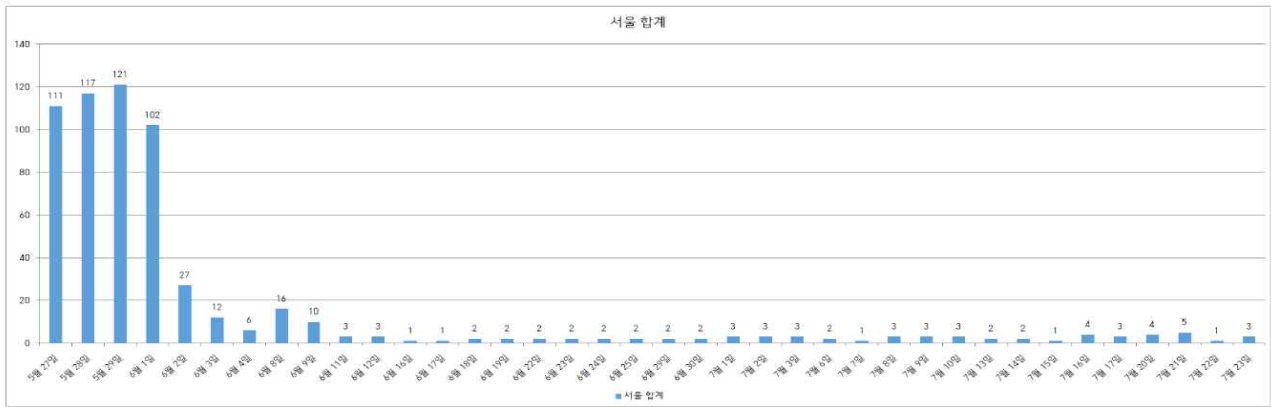


<그림 2-31> ‘코로나19: 우리의 기억’ 타임라인(코로나19: 우리의 기억, 2020)



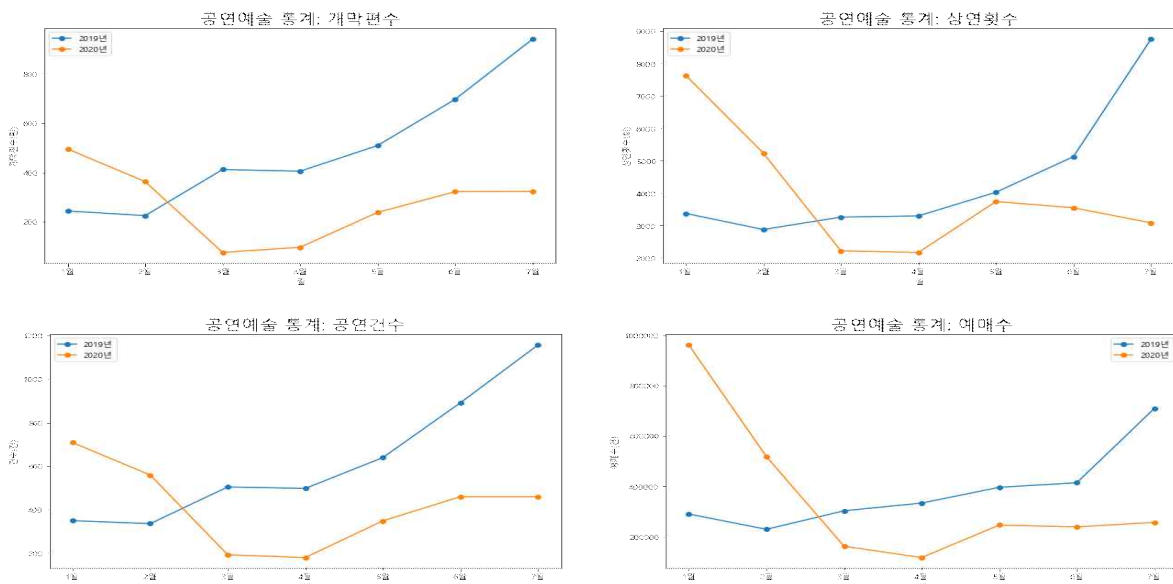
<그림 2-32> ‘코로나19: 우리의 기억’ 스토리맵(코로나19: 우리의 기억, 2020)

- (교육) 코로나19의 영향을 받은 교육 분야의 변화를 정리하고, 교육부의 보도자료를 통해 전국 학교의 대처 방안을 수집하고 분석함
 - 감염 방지를 위해 대면 수업의 일부 혹은 전체 일시 중지를 시행한 전국 17개의 지자체 학교의 등교 조정 실행기간을 조사함
 - 전국 학교의 등교조정 현황과 집단감염 사례를 분석하여 등교수업 조정 기준의 필요성을 제안함



〈그림 2-33〉 서울의 등교수업 조정 현황(코로나19: 우리의 기억, 2020)

- (문화) 코로나19의 확산 방지를 위해 ‘사회적 거리두기’의 시행의 영향으로 변화한 문화예술공연 산업 현황을 기록함
 - 영화 산업, 공연-전시업, 관광-항공업 데이터를 분석하여 코로나19 전후를 비교함
 - ‘영화관입장권통합전산망(KOBIS)’, ‘공연예술통합전산망(KOPIS)’, 한국관광공사와 한국공항공사 데이터를 수집하고 분석함
- 비대면 환경의 언택트 산업(도서-출판업, OTT 서비스업)의 현황과 관련된 뉴스와 데이터를 수집함
 - 서울특별시의 공공도서관 데이터를 수집해 온라인과 오프라인 대출 추이를 조사하여 사회적 거리두기 이후 오프라인 대출이 증가함을 분석함
 - 인터넷을 통해 콘텐츠를 제공하는 OTT(Over The Top) 서비스업 매출액의 추세를 조사함



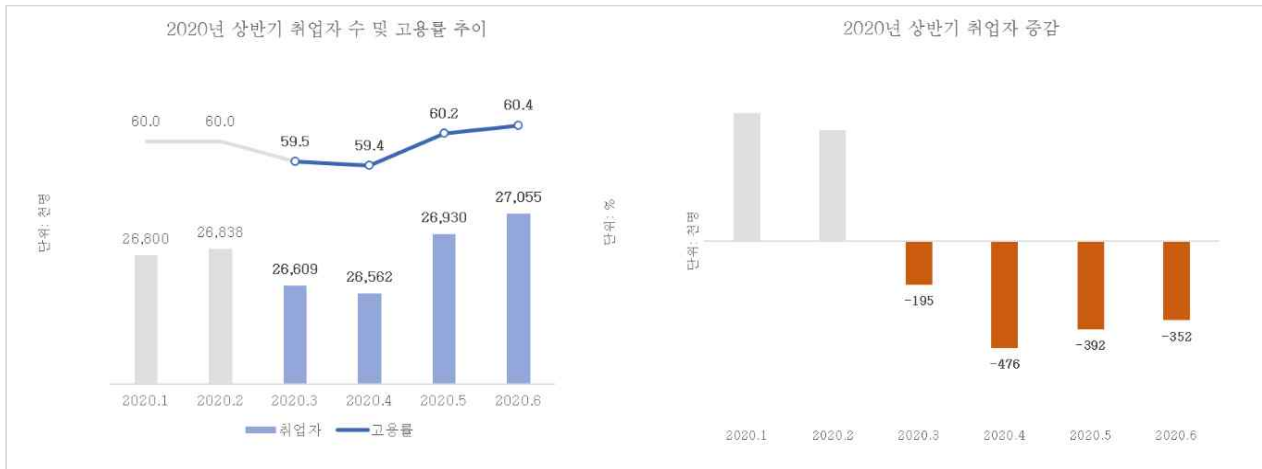
〈그림 2-34〉 공연예술 산업의 코로나19 전후 매출액 비교(코로나19: 우리의 기억, 2020)

- (사회) 코로나19로 인한 재난 불평등을 주제로 취약계층과 감염병의 관계를 분석함
 - 특정 산업체와 확진현황의 상관관계 분석을 위해 KOSIS의 시군구별 산업체 데이터를 수집함. 집단감염이 발생한 산업체 종사자와 확진자 수를 분석하여 제조업, 보건업, 사회복지 서비스업, 교육 서비스업 등이 유의미한 상관관계가 있음을 나타냄
 - 청와대 국민청원 홈페이지에서 코로나19 관련 게시글을 수집하고 자연어 처리한 결과를 기간별로 시각화*하여 국민들이 집중하는 이슈가 어떤 것인지 분석함
- * <그림2-35>는 월별로 코로나19 관련 키워드의 비중을 나타낸 워드 클라우드임
- 사회와 관련된 키워드 중 여성, 노동, 장애인, 성소수자, 종교를 선정하여 기사를 수집해 다양한 계층의 이슈를 조사함



<그림 2-35> 청와대 청원 홈페이지 키워드 분석(코로나19: 우리의 기억, 2020)

- (경제) 코로나19 확산 이후 침체된 경제 흐름을 분석하기 위해 전년대비 취업율, 실업률, 업종별 매출 동향, 주식 동향을 다각적으로 조사하여 경제에 미친 영향을 분석함
 - 통계청 데이터의 고용률과 실업률을 조사한 결과, 운수창고업은 사회적 거리두기로 택배 주문이 증가하면서 다른 업종에 비해 취업자 수가 유지되었고, 대면 접촉이 많은 도소매업과 숙박음식점업은 매장 방문율이 줄어들어 고용률이 대폭 감소함
 - 서비스업, 여행/교통 서비스, 문화/레저 서비스, 음식료품, 의류/신발, 의료/보건, 교육 업종별 매출 소비 데이터를 수집하고 증감률을 분석함



<그림 2-36> 2020년 상반기 고용률과 취업자 증감(코로나19: 우리의 기억, 2020)

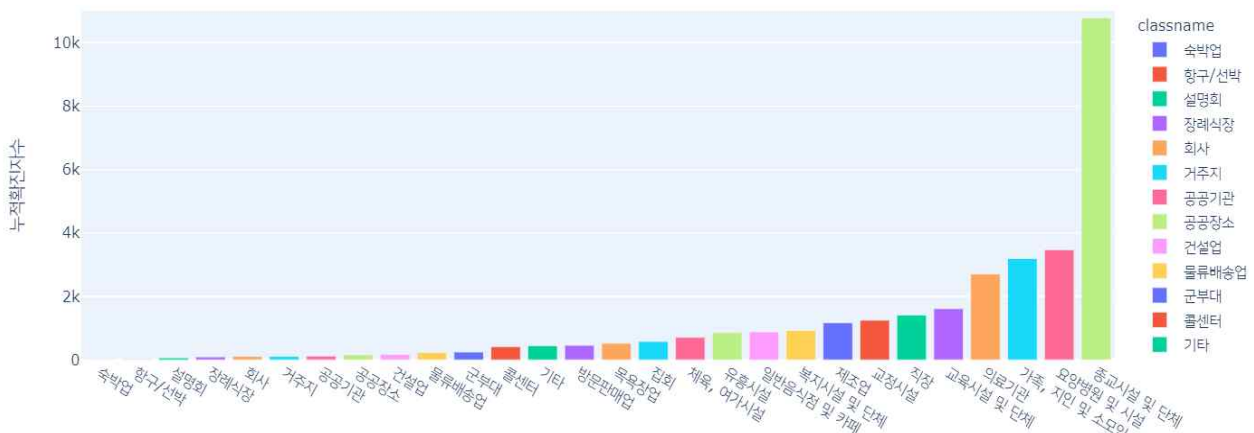
- **(환경)** 코로나19가 환경에 미친 영향을 분석하기 위해 미세먼지 데이터와 폐기물 배출량, 의료 폐기물 처리 과정, 일회용품 사용량을 탐색함
 - 국내 대기질이 개선된 원인을 조사하기 위해 석유 소비량, 전국 교통량 변화, 국내 항공량 변화 데이터를 수집하고 분석함
 - 생활 폐기물과 의료 폐기물 처리 현황을 분석하기 위해 온라인 쇼핑 거래액, 배달 거래액 데이터를 수집하고, 일회용품 사용량으로 인한 영향을 알아보기 위해 다각적으로 조사함

□ 코로나19 분석 A to Z

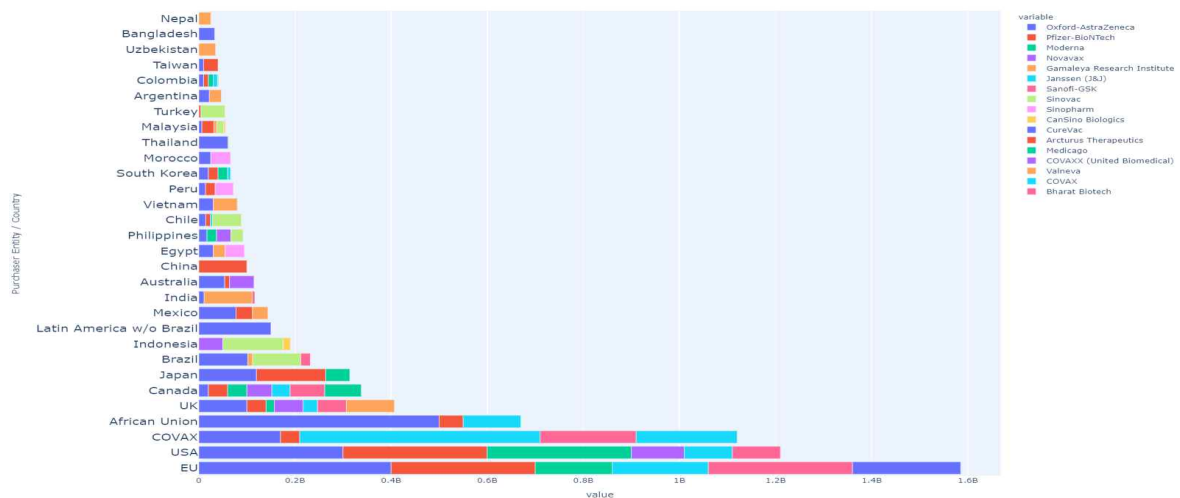
- **(코로나19 공공데이터 API 활용 가이드)** 보건복지부에서 공개하는 코로나19 감염현황, 코로나19 시도발생 현황, 코로나19 연령/성별 감염현황, 코로나19 해외발생 API를 수집하고 국내 코로나19 감염 추이와 확진자 집단별 차이를 분석함
 - 기존 코로나 정보 제공은 누적 확진 현황을 표 형식으로만 제공하고, 데이터 간 연결성이 미흡하다는 단점을 해결하기 위해 API 데이터를 전처리하고 시각화함
 - API 데이터의 정확성을 높이기 위해 컬럼과 실제 값이 불일치한 값을 수정하고 누적 기준의 값을 일별 기준으로 변환함
 - 국내 검사현황 분석을 위해 익명 검사 데이터와 요일별 검사 데이터를 수집하여 시각화하고, 선별진료소의 수량과 위치를 분석하기 위해 시군구별 선별진료소 수량과 인구 수 대비 선별진료소 수량을 분석함
 - 정부의 방역 정책 중 사회적 거리두기의 단계 변화에 따른 확진자 수 추이를 분석하기 위해 코로나 이슈 타임라인을 <그림 2-37>과 같이 시각화함



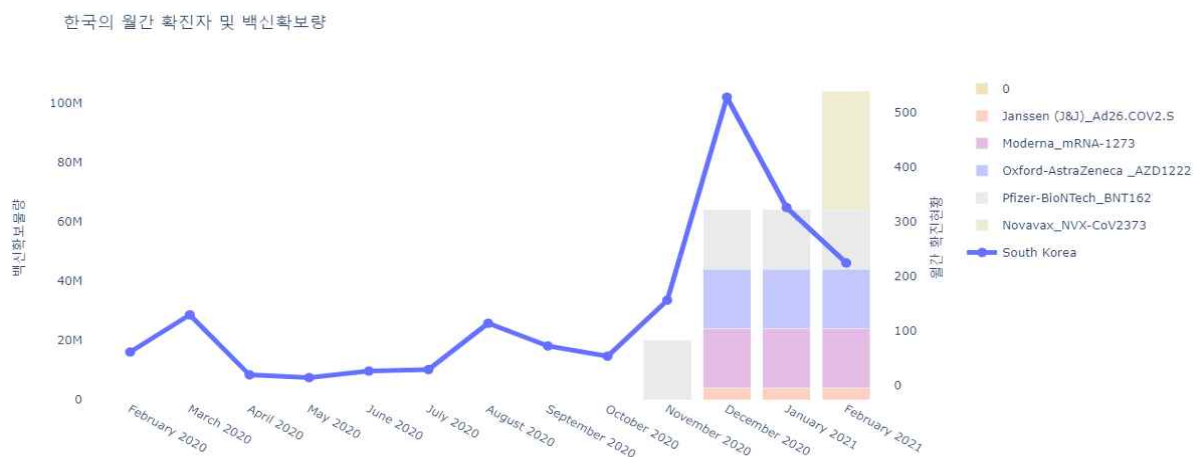
- **(집단감염 데이터 분석)** 방역 정책을 위해 개방하는 코로나19 관련 정례브리핑에 포함된 집단감염 정보를 이용자 접근성, 오픈데이터, 가공 용이성, 지속성에 맞는 기준 수집하고 분석함
 - 집단감염 데이터는 학교·회사·병원·공장·종교시설 등 한정된 지역이나 집단에서 발생하는 감염 사례에 대한 데이터로, 감염이 발생하는 집단에 대한 정보를 파악할 수 있음
 - 중앙재난안전대책본부(중대본)에서 발표하는 정례브리핑 보도자료(HWP, PDF)를 수집하고, 기계가 읽을 수 있는 형태인 CSV 형태로 변환함
 - ‘코로나19: 우리의 기억’에서 구축한 집단감염 데이터는 지역, 최초 감염 인식 날짜, 발생지, 확진자 수, 누적 확진자 수, 브리핑 날짜, 발생지 업종명을 수집함
 - 발생지를 업종별로 분류하고 분류기준별 집단감염 확진자와 지역별 집단감염 확진자를 분석하고 시각화함



- (백신으로 보는 국가별 코로나 대응 전략) 코로나19 백신 보급 초기 시기의 국가별 대응 전략을 분석함. 코로나19 백신 공급을 시각화하고, 한국의 백신 대응 현황을 파악함
- 전 세계 확진 현황 데이터는 ‘Our World In Data’, 백신 확보 현황과 확보일자 데이터는 듀크 대학교에서 제공하는 ‘launch and speedometer’의 데이터를 사용함
- <그림 2-39>은 국가별 백신 확보량을 시각화한 것으로, 미국·유럽·영국·캐나다와 같은 선진국이 다량의 백신을 보유함을 알 수 있음
- 1인당 GDP를 기준으로 국가들을 그룹화하고 백신 현황을 분석한 결과, 백신 수량을 많이 확보한 국가는 백신의 종류 또한 다양하게 확보함
- <그림 2-40>는 한국의 백신 확보 현황을 월별로 분석한 것으로, 백신 확보 계약 체결 시기가 다른 선진국에 비해 비교적 늦음



<그림 2-39> 국가별 백신 확보량(코로나19: 우리의 기억, 2020)



<그림 2-40> 한국의 월별 백신 공급 현황(코로나19: 우리의 기억, 2020)

제 3 장 ‘코로나19 아카이브’의 구축 방향

3.1 개요

3.1.1 비전

- 기존 웹 아카이브의 수량 중심의 수집·보존을 넘어 디지털 자료를 활용하고 재생산할 수 있는 체계를 통해 전 세계적으로 경쟁력 있는 디지털 아카이브를 지향함
 - 현재 구축·운영되고 있는 디지털 아카이브는 디지털 자료의 수집과 보존, 기록물의 재현과 제공을 핵심 가치로 정의하고 있음. 일부 아카이브는 디지털 자료의 공유를 목표로 갖고 있지만, 새로운 가치를 발굴하기 위한 활용이나 연계를 핵심 가치로 정의한 아카이브는 많지 않음 ([표 3-1] 참고)
 - 웹 아카이브의 기본적인 가치인 수집·보존·재현을 넘어 수집된 자료의 의미적 연계, 개별 데이터의 분석을 통한 활용 활성화를 핵심 가치와 목표로 정의해야 함
- 비전선언문은 ‘코로나19 아카이브’가 도전적으로 추구해야 할 목표를 정의함
 - 코로나19와 관련된 주제, 국가의 재난안전과 관련된 주제를 포괄하고, 디지털 자료의 활용과 분석을 지향하는 플랫폼을 지향함
- 코로나19 아카이브는 코로나19의 핵심 자료를 수집·보존하고, 동시에 재난 분야의 핵심 아카이브로 확대할 수 있는 지향점을 갖고 있음
 - 국가 수준의 재난안전, 감염병에 대해 다양한 매체에서 생산되는 자료를 종합적으로 수집·보존하고, 디지털 자료의 재사용과 분석을 지원하는 플랫폼을 지향함
- 핵심가치는 ‘재현’, ‘연결’, ‘공유’, ‘활용’ 등 네 가지의 핵심 가치를 정의하고, 개별 핵심 가치를 현실화하기 위해 핵심 목표를 구체적으로 수립함

비전	국가 재난안전에 대한 지식을 융합하는 아카이브 플랫폼			
핵심가치	재현	연결	공유	활용
핵심목표	코로나19 자료의 수집·보존	수집 자료의 의미적 연결	재난안전 자료의 공동활용체계	재난안전 자료의 다차원 분석
	수집한 웹사이트의 검색·재현	공공·민간 아카이브의 연계	시민 참여 기반의 생태계 채널	고가치 데이터의 재생산 체계

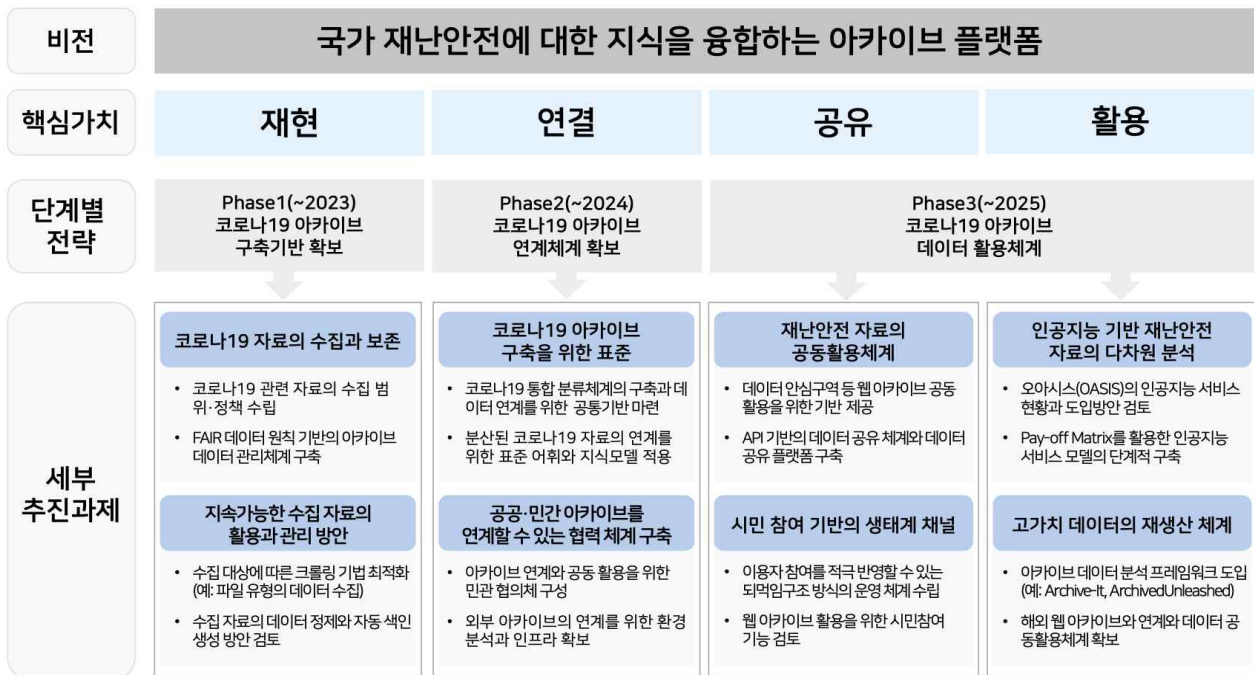
<그림 3-1> ‘코로나19 아카이브’의 비전과 핵심가치(김학래, 2022)

[표 3-1] 분야별 디지털 아카이브의 주요 목표 (개별 아카이브의 핵심 단어를 추출)(김학래, 2022)

아카이브	주요 목표와 내용
오아시스 (국립중앙도서관)	국립중앙도서관은 국가대표도서관으로서 디지털 지적 문화유산을 국가적 차원에서 <u>수집·보존</u> 하여 후대에 <u>제공</u> 할 목적으로 2004년부터 온라인 디지털자원의 수집·보존 프로젝트인 OASIS(Online Archiving & Searching Internet Sources) 사업을 추진해오고 있습니다.
코리안 메모리 (국립중앙도서관)	국립중앙도서관에서는 ‘코리안 메모리-지역 문화유산 디지털아카이브’ 구축 기반을 마련하고 향토자료, 희귀자료 등 공공도서관 주요 장서의 <u>디지털화 지원</u> 을 통한 국가문헌의 <u>확충</u> 및 <u>보존</u> , <u>제공</u> 에 기여하고자 합니다.
경기도 메모리 (경기도)	경기도메모리는 개인·단체·기관과 함께 경기도의 기억과 기록을 모아 서비스하는 디지털 아카이브입니다. 사람과 마을과 지역을 <u>잇고</u> 공유하여 경기도를 <u>재현</u> 합니다.
재난 디지털 아카이브 (조선대학교)	디지털아카이브는 동아시아 각종 자료에 수록된 재난 관련 기록과 기억 정보의 데이터베이스를 <u>축적</u> 하여, 사회적 상호 연관성 등을 <u>탐색</u> 할 수 있는 디지털 기반의 연구 환경을 <u>구축</u> 한다. 디지털 인문학의 연구 방법론에 입각하여 고대부터 현재까지 동아시아에서 발생한 자연, 환경, 기후, 재해, 질병 등 자연재해, 민란, 분쟁, 전쟁 등 인위적 사회적 재난과 그 기록의 <u>아카이빙</u> 을 시도한다.
근현대사 아카이브 (대한민국역사박물관)	대한민국역사박물관은 대한민국의 근현대사를 증명하는 다양한 사료의 <u>수집·관리</u> 및 <u>서비스</u> 를 통하여 역사 기록 자원을 기반으로 근현대사 연구 및 기반 문화산업의 <u>활성화</u> 에 이바지 하고자 근현대사 디지털아카이브를 <u>구축·운영</u> 하고 있습니다.
한국예술디지털아카이브	한국예술디지털아카이브(Korea Digital Archives for the Arts : DA-Arts)는 2011년 예술자료를 시간·공간의 제약을 넘어 <u>활용</u> 할 수 있도록 하고자 소장 예술자료 중 일부를 <u>디지털화</u> 하여 <u>제공</u> 하면서 시작되었습니다.
예술자료 디지털 정보저장소	예술단체와 예술인이 예술자료를 저장하고 활용할 수 있는 디지털 공간을 <u>제공</u> 하고, 그것을 통합적으로 <u>검색</u> 하고 <u>활용</u> 하는 것을 목표로 하고 있습니다. 현장에서 생산된 예술자료를 보다 체계적이고 효율적으로 <u>관리</u> 하고자 하는 예술단체와 예술인의 많은 <u>동참</u> 을 기다립니다.
무형유산 디지털 아카이브 (국립무형유산원)	국립무형유산원은 여러 세대에 걸쳐 전승되어 온 무형문화제에 대한 다양한 자료를 생산 및 <u>수집</u> 하여 <u>보존</u> , <u>관리</u> 하고 있습니다. 무형유산 디지털 아카이브에서는 수집한 자료를 <u>공개</u> 하여 국민들이 무형유산에 대한 정보 및 자료를 한 곳에서 보고 이용하도록 <u>서비스</u> 하고 있습니다.

3.1.2 연차별 단계적 접근

- 핵심 가치를 실현하기 위해 우선순위가 높은 핵심 목표를 선제적으로 추진하고, 시민 참여와 데이터 활용 활성화를 위해 단계적으로 접근
- **(기반확보)** 대규모 디지털 자료를 효과적으로 수집보존하기 위한 환경을 구축하고, 수집한 자료의 검색과 재현을 지원할 수 있는 기반 마련
- **(연계체계 확보)** 수집된 자료 사이의 의미적 연결을 위한 가이드라인의 수립과 적용, 공공·민간 아카이브를 연계할 수 있는 협력 체계 구축
- **(활용체계)** 코로나19 아카이브를 공동활용하기 위한 체계의 수립, 국민이 참여하고 협업할 수 있는 개방형 플랫폼의 구축
- **(확대개편)** 코로나19 아카이브와 오아시스 재난아카이브의 연계 또는 확대 개편



<그림 3-2> ‘코로나19 아카이브’ 구축을 위한 주요 과제

[표 3-2] ‘코로나19 아카이브’ 구축을 위한 로드맵

세부 추진과제	1단계 (~2023년)	2단계 (~2024년)	3단계 (~2025년)
코로나19 아카이브의 지도·감독체계 확립	<ul style="list-style-type: none"> 코로나19 아카이브에 대한 모니터링 체계 수립 	<ul style="list-style-type: none"> 개별 아카이브의 특성과 품질에 따른 모니터링 최적화 방안 수립 	-

코로나19 관련 자료 수집정책 개발	<ul style="list-style-type: none"> 수집대상 아카이브의 범위와 유형 정의 아카이브 특성에 따른 수집정책 마련 	<ul style="list-style-type: none"> 수집대상 아카이브의 수집과 업데이트 자동화 도구 개발 	-
조직 및 인력 재설계	<ul style="list-style-type: none"> 코로나19 아카이브를 위한 전담인력 확보 	<ul style="list-style-type: none"> 전문 교육 프로그램 설계 직원 재교육 프로그램 시행 	<ul style="list-style-type: none"> 오아시스 확대 개편을 위한 업무 와 조직 개편
코로나19 자료의 수집과 보존	<ul style="list-style-type: none"> 코로나19 관련 자료의 보존 형식과 방법에 대한 기술 검토 코로나19 데이터의 재사용을 위한 소프트웨어 지원 방안 수립 	<ul style="list-style-type: none"> FAIR 데이터 원칙 기반의 아카이브 데이터 관리체계 구축 	<ul style="list-style-type: none"> 장기보존정책 수립
지속가능한 수집 자료의 활용과 관리 방안	<ul style="list-style-type: none"> 수집 대상에 따른 크롤링 기법 최적화 (예: 파일 유형의 데이터 수집) 	<ul style="list-style-type: none"> 수집 자료의 데이터 정제와 자동 색인생성 방안 마련 	<ul style="list-style-type: none"> 전문가/사용자에게 공유할 수 있는 데이터 선정 클라우드 기반 데이터 분석 기능 제공
코로나19 아카이브 구축을 위한 표준	<ul style="list-style-type: none"> 코로나19 통합 분류체계의 구축과 데이터 연계를 위한 표준 체계 마련 	<ul style="list-style-type: none"> 분산된 코로나19 자료의 연계를 위한 표준 어휘와 지식모델 적용 	<ul style="list-style-type: none"> 국내외 표준 기반 코로나19 자료의 지식그래프 구축
공공·민간 아카이브를 연계할 수 있는 협력 체계 구축	<ul style="list-style-type: none"> 아카이브 연계와 공동 활용을 위한 민간 협의체 구성 	<ul style="list-style-type: none"> 외부 아카이브의 연계를 위한 환경 분석과 인프라 확보 	<ul style="list-style-type: none"> 공공·민간 아카이브와 오아시스 재난 아카이브의 확대 개편
재난안전 자료의 공동활용체계	<ul style="list-style-type: none"> 공동활용 가능한 코로나19 아카이브 데이터 선정 선정데이터의 제공 방안과 활용 결과를 반영할 수 있는 체계 수립 	<ul style="list-style-type: none"> API 기반의 데이터 공유 체계와 데이터 공유 플랫폼 구축 	<ul style="list-style-type: none"> 데이터 안심구역 등 웹 아카이브 공동활용을 위한 기반 제공

시민 참여 기반의 생태계 채널	<ul style="list-style-type: none"> ▪ 이용자 참여를 적극 반영할 수 있는 되먹임구조 방식의 운영 체계 수립 ▪ 웹 아카이브 활용을 위한 시민참여 기능 검토 	<ul style="list-style-type: none"> ▪ 시민과 담당자가 지속적으로 의사소통할 수 있는 기능 제공 (데이터 기증, 데이터 분석공유 기능 등) ▪ 시민 참여를 통한 데이터 가공·분석 기능 제공 	<ul style="list-style-type: none"> ▪ 해커톤 등 아카이브 데이터를 활용할 수 있는 프로그램 마련 ▪ 해외 아카이브 커뮤니티와 공동으로 데이터 활용 프로그램 기획
인공지능 기반 재난안전 자료의 다차원 분석	<ul style="list-style-type: none"> ▪ 오아시스(OASIS)에 적용 가능한 인공지능 서비스 선정과 도입방안 검토 ▪ 오픈소스 기반 지능형 솔루션 검토와 테스트 ▪ 오픈소스 기반 아카이브 시각화 도구 테스트 	<ul style="list-style-type: none"> ▪ 오픈소스 기반 코로나19 아카이브 시각화 도구 구현 ▪ 지식그래프 기반 통합 검색서비스 아키텍처 설계와 테스트 ▪ FAIR 데이터 기반 아카이브 데이터 접근과 재사용 검증 	<ul style="list-style-type: none"> ▪ 오아시스 재난 아카이브와 연계한 통합 검색 서비스 제공
고가치 데이터의 재생산 체계	-	<ul style="list-style-type: none"> ▪ 아카이브 데이터 분석 프레임워크 도입(예: Archive-It, ArchivedUnleashed) 	<ul style="list-style-type: none"> ▪ 해외 웹 아카이브와 연계와 데이터 공동활용체계 확보 ▪ 아카이브 서비스의 기능과 전체 라이프사이클을 고려한 사용자 친화적 UI/UX 구축

3.2 핵심가치와 목표

3.2.1 수집된 자료의 끊임 없는 재현

- (코로나19 관련 자료의 특성) 코로나19 관련 자료는 감염병의 출현과 대응 방안을 포괄하며, 사회의 기민한 변화를 반영하는 특성이 있음
- (주제의 다양성) 코로나19 감염병은 비대면 사회의 출현과 같이 전 인류사회에 광범위한 영향을 만들었고, 감염병 확산 방지를 위한 정부의 정책과 대응, 코로나19로 인한 사회·경제·정치적 영향 등 관련 범위가 매우 포괄적임

- **(자원의 다양성)** 코로나19 자료는 텍스트, 동영상, 이미지 등 다양한 디지털 형식을 포함하고 있고, 온라인 공간에 산재되어 있음
- **(자원의 휘발성)** 사회적 거리두기, 방역 정책 등 코로나19로 인한 사회변화가 급속히 진행됨에 따라 정보자료의 가변성이 매우 높은 특징이 있고, 적시에 수집·보존하지 않은 자료가 소실될 가능성이 매우 큰 특징을 갖고 있음
- **(코로나19 자료의 수집과 보존)** 코로나19의 출현과 감염병 확산 방지를 위한 사회적 노력, 산·학·연·관·민의 일련의 활동과 관련된 자료를 효과적으로 수집하여 코로나19 관련 자료를 체계화시킴
 - ‘코로나19’의 수집범위와 수집대상, 수집 우선순위, 수집과정과 절차, 유관기관의 협력체계, 수집 자료의 라이선스와 이용조건, 활용 방안 등 수집을 위해 합리적인 정책과 가이드라인 수립이 요구됨
 - 수집 대상과 우선순위는 관련 분야 전문가 협의회를 구성하여 객관적으로 선정하고, 대표성, 정확성, 시급성, 효용성 등 우선순위를 판단할 수 있는 정량적인 지표를 설계하고 평가체계를 수립해야 함
- **(코로나19 자료의 재현)** 웹 환경의 기록화는 원천 웹사이트와 자료의 맥락을 유지하고, 원천 자료의 시각에서 ‘재현’하는 것이 중요함
 - 정적 웹 페이지, 자바스크립트가 포함된 동적 페이지, 파일 유형의 데이터, 소셜미디어 API 등 수집 대상에 따라 웹 크롤링 기법을 최적화하고 색인생성을 위한 방안을 검토해야 함
 - 원천 자료의 수집을 위해 웹 크롤링, 색인생성과 관리, 수집 자료의 검색과 재현을 위한 기반 기술과 환경을 구축해야 함

3.2.2 아카이브의 의미적 연결

- **(기록의 연계)** 코로나19와 관련된 방대한 디지털 자료는 의미적 관계를 기반으로 서로 연결되어 데이터 탐색·검색 또는 사용자의 요청에 효과적으로 대응해야 함
 - 디지털 자료의 연결은 주제별 컬렉션, 개별 자료 수준에서 검토할 수 있고, 통합 분류체계를 구축하여 공통으로 적용할 수 있는 기반 마련이 필요함. 국가 재난과 코로나19 감염병을 포함하는 분류체계는 표준화되어 있지 않고, 기계가 읽을 수 있는 형식으로 제공되고 있지 않음
- **(아카이브 간 연계)** 코로나19 아카이브는 독자적으로 관련 자료를 수집하고, 동시에 공공·민간의 아카이브에 연계할 수 있는 연계체계를 수립해야 함
 - 코로나19 아카이브는 디지털 자료의 통합적 접근보다 산재 또는 독립적으로 운영되는 데

이더 포털과 디지털 아카이브를 연결하는 방식이 바람직함. 코로나19 관련 자료는 주기적으로 생산된 자료, 규모가 큰 디지털 자료를 포함하고 있기 때문에 웹 아카이브에서 통합하는 데 제한이 있음

- 공공·민간이 보유하고 있는 디지털 자료는 코로나19 아카이브에서 연결할 수 있는 방안을 제공하고, 아카이브·개별 자료에 적용할 수 있는 공통 가이드라인을 수립하고 협력하는 것이 필요함
- 예를 들어, ‘코로나19 데이터 카탈로그’와 같이 물리적 공간과 관계없이 유사한 자료에 접근할 수 있는 방안을 검토하고, 유관기관들과 협력체계를 구축하는 것이 필요함

3.2.3 공동활용을 위한 공유

- **(공동활용체계)** 주제 전문성을 높이기 위해 공동활용 체계를 수립하고, 국민이 참여하고 협업할 수 있는 개방형 아카이브를 지향함
 - 코로나19 등 감염병 관련 자료는 광범위한 범위를 포함하기 때문에, 유관기관이 보유한 데이터를 연계·활용하기 위한 방안이 필요함
 - 웹 아카이브의 제약 사항인 라이선스와 이용조건을 해결하기 위해 데이터 프리존과 같은 방식으로 수집 자료를 공유할 수 있는 방안을 검토해야 함
 - 수집된 디지털 자료는 원천데이터의 형식으로 제공하고, 동시에 주제별 컬렉션과 같이 가공된 데이터는 이력을 함께 제공하는 것이 필요함
- **(참여 기반 생태계)** 웹 아카이브의 자료를 이해하고 활용하기 위해 적극적으로 시민 참여를 유도하고, 웹 아카이브를 활용할 수 있는 다양한 방법을 제공해야 함
 - 시민은 코로나19 아카이브의 사용자이며, 동시에 관련된 자료를 기증 또는 제공할 수 있는 주체이기 때문에, 시민 참여는 자료의 수집과 활용에 이르는 전 단계에서 핵심적으로 고려해야 함
 - 웨이백 머신(Wayback machine)으로 수집된 자료를 탐색하는 방식은 사용자에게 편리하지만, 재현이 어려운 자료가 다수 발생할 수 있기 때문에 주제별 컬렉션을 구축하기 위한 데이터 관리체계를 검토해야 함

3.2.4 새로운 가치 발굴을 위한 활용

- **(아카이브 활용 활성화)** 축적된 디지털 자료를 효과적으로 활용하기 위한 시스템을 구축하고, 자료의 활용·재생산할 수 있는 선순환체계를 지원함
 - 웹 아카이브로 축적된 자료를 주제 분야에 따라 분류체계를 적용하여 쉽게 접근·검색하기 위한 기능을 제공하고, 연관 있는 데이터를 분석할 수 있는 환경을 제공해야 함
 - 코로나19 관련 자료는 텍스트, 통계 등 데이터 분석이 가능하고, 공공·민간 아카이브의 다양한 자료를 융합하여 특정한 문제를 해결하는 데 활용될 수 있음

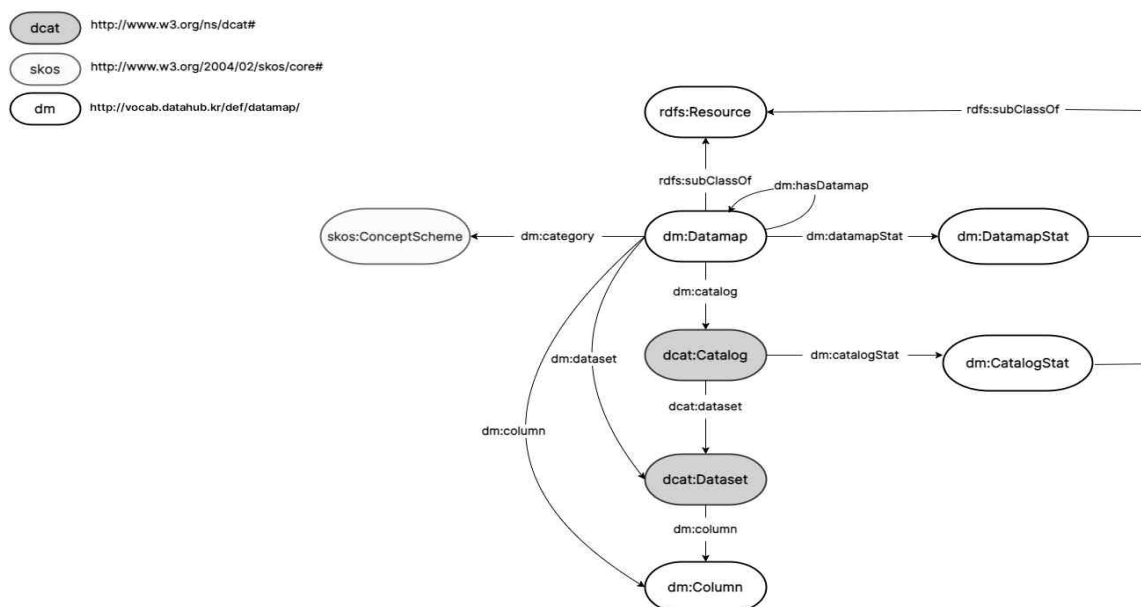
- 데이터 과학자가 웹 아카이브의 자료에 쉽게 접근하고 분석할 수 있는 개발·분석·시각화 환경을 구축하고, 관련 연구자, 개발자 커뮤니티와 협력할 수 있는 채널을 확보해야 함
- **(고가치 데이터)** 수집된 자료의 분석과 가공을 통해 특정 사회 현안의 해결, 정부 정책의 수립에 활용할 수 있는 고품질의 데이터를 확보할 수 있음
 - 수집된 자료의 데이터 분석을 지원하기 위해 웹 아카이브에 있는 자료를 선별적으로 추출할 수 있는 기능을 제공함. 수집된 디지털 자료를 일관성 있고 확장 가능한 분류체계를 적용하고, 재생산되는 데이터에 동일하게 적용하여 관리체계의 확장성을 높임
 - 디지털 자료의 효과적인 탐색과 활용을 위해 지능형 검색 등 인공지능 기술을 적극적으로 적용함. 텍스트 중심의 자료는 자연어 처리를 위해 기계학습 환경을 구축할 필요가 있음

3.3 아카이브 구축을 위한 표준

3.3.1 오아시스(OASIS) 재난아카이브의 연계와 확대

- **(오아시스 재난아카이브)** 오아시스(OASIS)는 국립중앙도서관에서 진행하는 국가적 차원의 온라인 디지털 자원의 수집과 보존 프로젝트로, 재난아카이브를 별도로 구성함
 - 재난안전법에 따라 특별 재난으로 선포된 재난과 재해, 정치·사회·문화적 파급력이 큰 자연재난과 사회재난을 대상으로 함
 - 재난유형별, 발생기간, 지역별로 자료를 구분하고 있고, 총 40건의 재난 사건을 기록한 웹 자원 35,974건을 수집함
 - * 자세한 재난아카이브의 내용은 불임2을 참고
- **(연계 방안)** 재난아카이브와 코로나19 아카이브의 모든 정보는 검색과 관리를 위해 상호 접근할 수 있도록 연계되는 것이 바람직하며, 디지털 아카이브의 자원을 데이터 목록의 형식으로 구축하여 의미적으로 연계하고 확장 가능한 방식을 검토해야 함
 - 디지털 아카이브의 웹 자원을 데이터 관점에서 관리하기 위해 DCAT, SKOS, Datamap 등 국제·국내 표준을 활용하여 데이터 목록과 이에 대한 상세 정보를 기계가 읽을 수 있는 형식으로 기술할 수 있음
- **(데이터맵 어휘)** 데이터 서비스(예: 데이터 포털, 디지털 아카이브)가 보유하거나 제공하는 메타데이터를 기술하고, 서로 다른 데이터 서비스를 연계하기 위한 표준임
 - 한국정보통신기술협회의 정보통신단체표준(TTAS)으로 제정됨(데이터맵 어휘, TTAK.KO-10.1291, 2021.6.)

- 데이터 서비스의 주요한 정보 개체와 개체 사이의 관계를 정의하고, 데이터맵 사이의 연계 방안을 RDF기반의 어휘로 제공함. 데이터 서비스가 제공하는 데이터세트, 정보 자원에 대한 수치 정보를 표현하고 서로 다른 데이터맵의 현황을 통합하고 탐색하는 데 활용할 수 있음
- 데이터 서비스는 데이터 포털, 디지털 아카이브 등이 보유한 데이터세트 또는 디지털 자원과 이에 대한 메타데이터를 제공하고, 동시에 서비스의 운영과 관리를 위한 정보를 함께 갖고 있음. 각 정보자원은 웹상에서 관리되기 때문에 고유한 URI 체계로 식별할 수 있어야 함
- 데이터맵 어휘의 핵심 클래스는 dm:Datamap으로 데이터 서비스를 운영하고 관리하기 위해 필요한 메타데이터 집합을 표현함. 개별 데이터세트의 메타데이터는 DCAT이나 Schema.org의 클래스로 데이터세트를 표현할 수 있고, 컬럼에 대한 메타데이터는 dm:Column으로 표현함
- 데이터 서비스에서 제공하는 분류체계는 dm:category 속성을 SKOS 어휘와 연결하여 해당 정보를 표현함. SKOS에서 분류체계 요소의 집합은 skos:ConceptScheme으로 표현하고 개별 항목은 skos:Concept으로 표현함. skos:Concept은 skos:related, skos:broader, skos:narrower 등을 통해 서로 다른 분류체계 간의 관계를 의미적으로 나타낼 수 있음
- 데이터맵 어휘는 메타데이터 수준에서 수치정보를 표현하기 위한 속성을 정의하고 있음. dm:numberOfOrganization, dm:numberOfDataset, dm:numberOfCatalog는 각각 데이터맵에 연계된 기관, 데이터세트, 데이터목록의 전체 개수를 표현함



<그림 3-3> 데이터맵 어휘 클래스의 다이어그램(김학래, 2022)

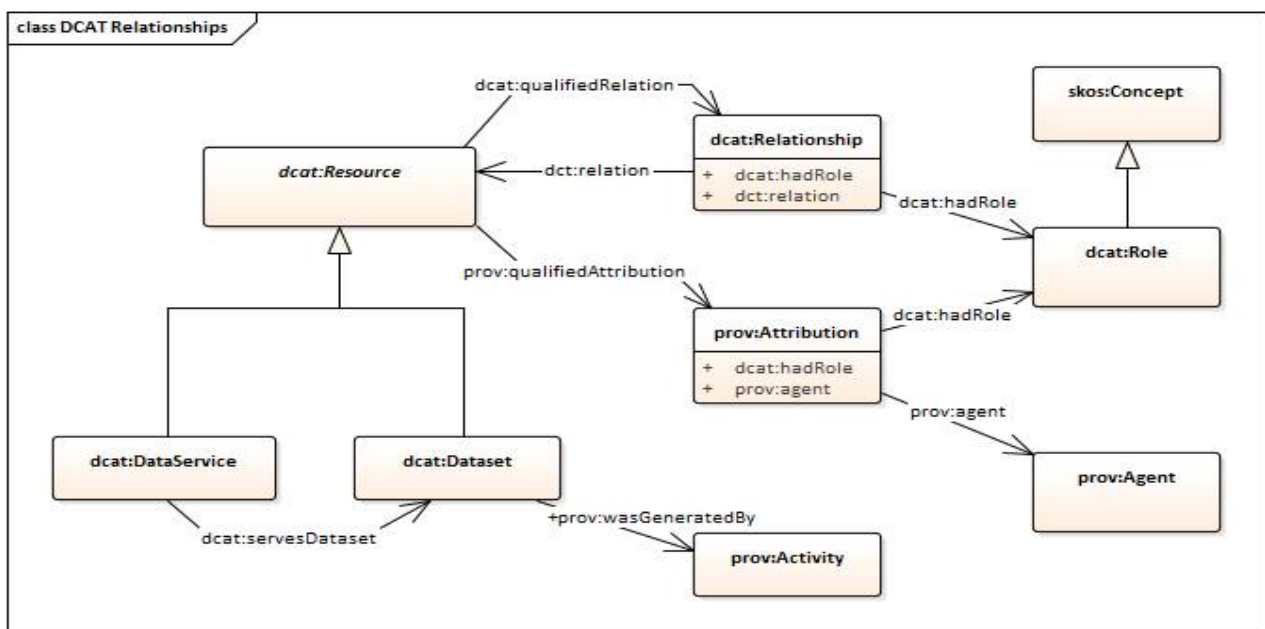
- (DCAT) W3C 권고안으로 웹에 게시된 데이터 카탈로그와 데이터세트 사이의 상호 운용성이 용이하도록 설계된 RDF 어휘이며, 메타데이터를 체계적으로 구조화하고

의미적으로 연결하고 접근할 수 있는 기능을 제공함

- DCAT은 dcat:Catalog, dcat:Dataset, dcat:Distribution, dcat:Resource 클래스를 사용하여 용어의 재사용과 모듈화를 강화시켰음. dcat:Resource는 dcat:DataService를 하위클래스로 정의하여 느슨하게 구조화된 카탈로그의 관계를 표현할 수 있음. 현재 권고하는 버전은 DCAT2이지만, 업데이트 버전인 DCAT3은 공간, 시간적 개념을 추가하고 dcat:CatalogedResource를 통해 데이터세트의 버전 관리를 위한 속성을 추가함
- DCAT은 카탈로그, 데이터세트를 표현하기 위한 핵심 어휘로 구성되어 있음([표 3-3] 참고)
- DCAT은 RDFS, OWL2를 기반의 표현력을 갖고 있고, 더블린코어(Dublin Core 1.1), 더블린 코어 메타데이터(DCTERMS), FOAF(Friend of A Friend), SKOS(Simple Knowledge Organization System) 등 기존 어휘와 의미적 연계가 가능함
- 구글 중심으로 개발하고 있는 Schema.org는 DCAT에서 표현하는 데이터 카탈로그, 데이터 세트에 대한 유사한 클래스를 정의하고 있음. Schema.org는 비교적 DCAT보다 느슨한 구조를 가지고 있어 다양한 자원 형태를 표현할 수 있음

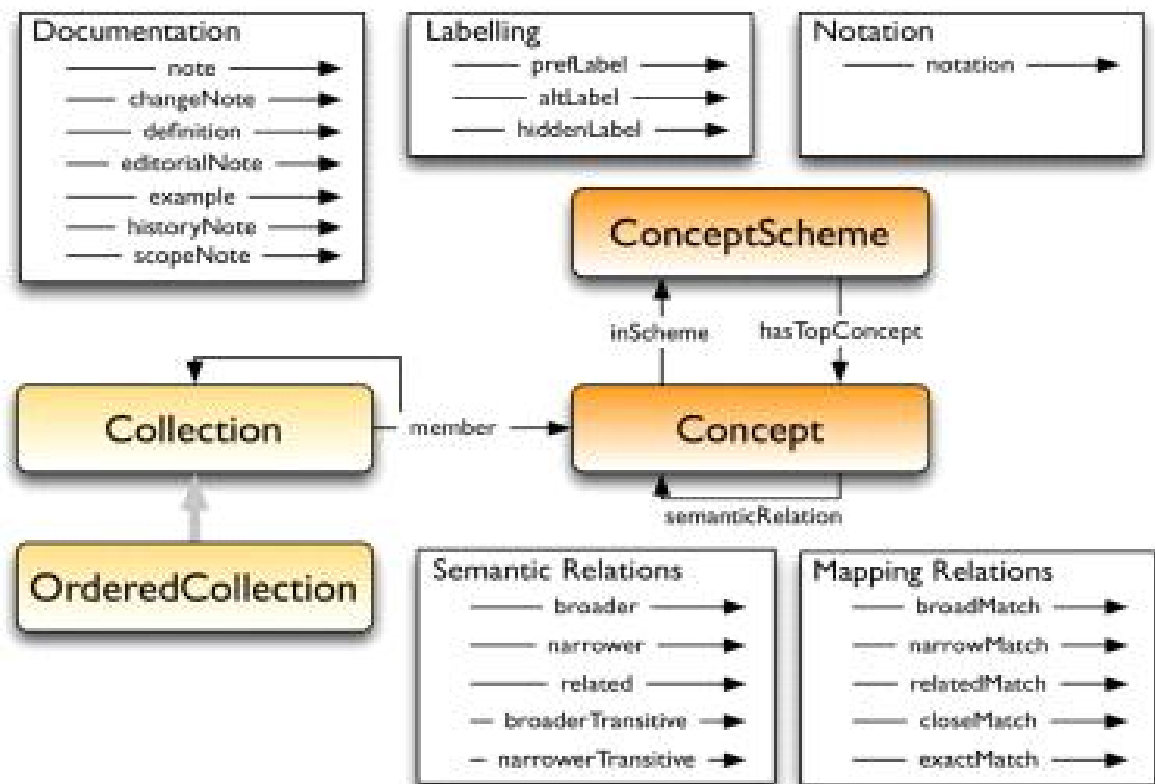
[표 3-3] DCAT 클래스와 설명

DCAT 클래스	설명
dcat:Catalog	데이터세트 또는 데이터 서비스의 메타데이터를 표현하기 위한 클래스
dcat:Resource	카탈로그의 메타데이터를 기술하기 위한 클래스
dcat:Dataset	데이터세트(데이터의 집합)을 표현하기 위한 클래스
dcat:Distribution	데이터세트의 접근 형식을 표현하기 위한 클래스
dcat:DataService	데이터를 제공하는 서비스를 표현하기 위한 클래스



<그림 3-4> DCAT 주요 클래스 다이어그램(W3C, 2020)

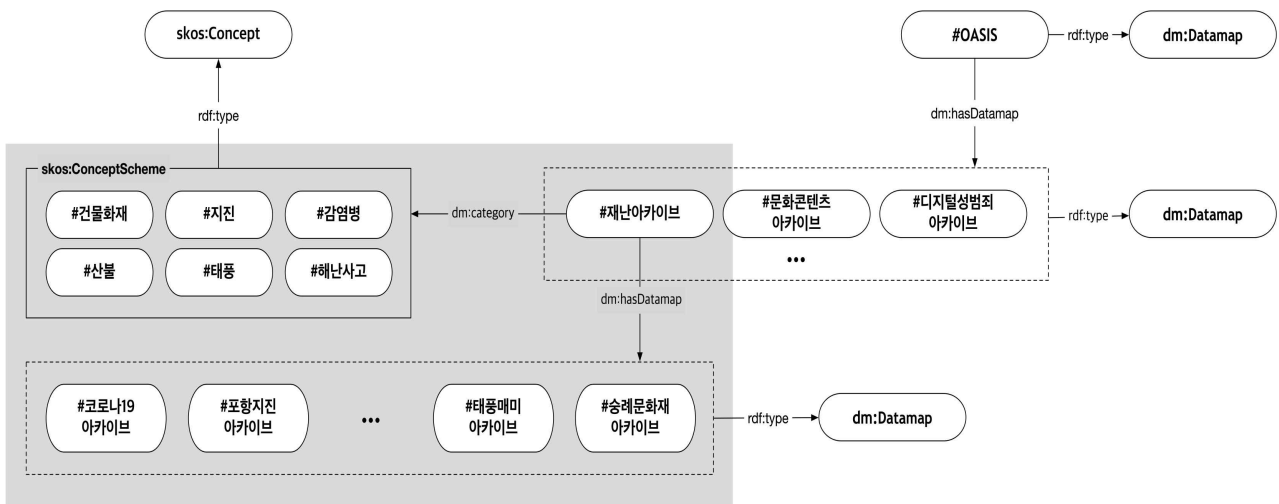
- (SKOS) W3C 권고안으로 분산되고 연결 가능한 방식으로 지식을 조직하는 체계(통제 어휘, 시소러스, 텍소노미 등)를 표현하는 방법을 제공함
- 지식조직시스템(knowledge organization system)은 통제 어휘의 구조와 사용 방식을 구조적으로 표현하기 위한 어휘 체계이며, 데이터의 표현·공유·교환할 수 있는 표준이 필수임
- 개념 스키마의 콘텐츠를 표현하기 위한 기본 구조와 모델을 제공하며, 기계가 읽을 수 있는 형식의 데이터로 표현할 수 있고 웹에 데이터를 게시·공유·교환하기 위한 방법을 제공함
- 시소러스를 표준적으로 표현하기 위한 클래스와 속성으로 구성되어 있고, 개념(concept) 중심으로 어휘를 기술할 수 있음(skos:Concept)
- 시소러스의 상위어(Broader Term), 하위어(Narrower Term), 관련어(Related Term)을 표현할 수 있는 의미적 구조를 표현하고 있고, 개념에 대한 어휘를 유니코드 문자열로 작성할 수 있음(예: 선호 - skos:prefLabel, 대체 - skos:altLabel, 숨겨진 - skos:hiddenLabel 등)
- <그림 3-5>는 SKOS의 클래스 다이어그램을 나타낸 것으로 각 클래스에 쓰이는 속성과 skos:ConceptScheme, skos:Concept, skos:Collection, skos:OrderedCollection 클래스 간의 관계를 표현함



<그림 3-5> SKOS 주요 클래스 다이어그램(Baker, 2013)

3.3.2 국내외 표준 기반 자원 연계 방안

- Datamap, DCAT, SKOS 어휘를 적용한 디지털 아카이브 지식그래프 연계 모델은 데이터를 의미적으로 탐색할 수 있는 구조로 표현되고, 분류체계를 기준으로 데이터의 검색이 가능함
- 오아시스는 재난아카이브를 비롯한 여러 디지털 아카이브 컬렉션을 관리하고 있음. 각 컬렉션은 관련 웹 자원을 웹사이트, 문서, 이미지, 동영상으로 구분하여 보존함



〈그림 3-6〉 디지털 아카이브 연계를 위한 지식 모델(김학래, 2022)

- (데이터맵 어휘 기반 디지털 아카이브) 오아시스의 웹 아카이브 자원 현황과 관련된 정보는 데이터맵 어휘를 통해 기계가 읽을 수 있는 형식으로 표현되어 이종의 디지털 아카이브를 연계할 수 있음
- 데이터맵은 DCAT 또는 Schema.org로 생성한 데이터 목록과 데이터세트를 재사용할 수 있기 때문에 정적으로 관리되는 기존의 디지털 아카이브에 비해 연계와 확장이 용이함
- 정보자원의 규모(dm:numberOfDataset), 관리기관(dm:numberOfOrgagnization), 서비스 상태(dm:availableStatus) 속성을 적용하여 상세 정보를 기술함
- 주제 분류는 오아시스에서 관리하는 분류체계 항목을 SKOS의 개념(skos:Concept) 클래스의 인스턴스로 정의하고, 개별 분류 사이의 관계는 skos:related 속성을 이용해 분류체계 사이의 관계를 표현함
- 데이터맵 어휘를 사용한 디지털 아카이브는 RDF 형식의 기계가 읽을 수 있는 형식으로 변환되며, 데이터를 처리할 수 있는 환경을 구성하면 질의와 탐색이 가능함
- (분류체계 모델) 오아시스 재난아카이브의 분류체계*를 기준으로 기계가 읽을 수 있는 형식으로 표현하고, SKOS 어휘를 적용하여 모델을 구성함

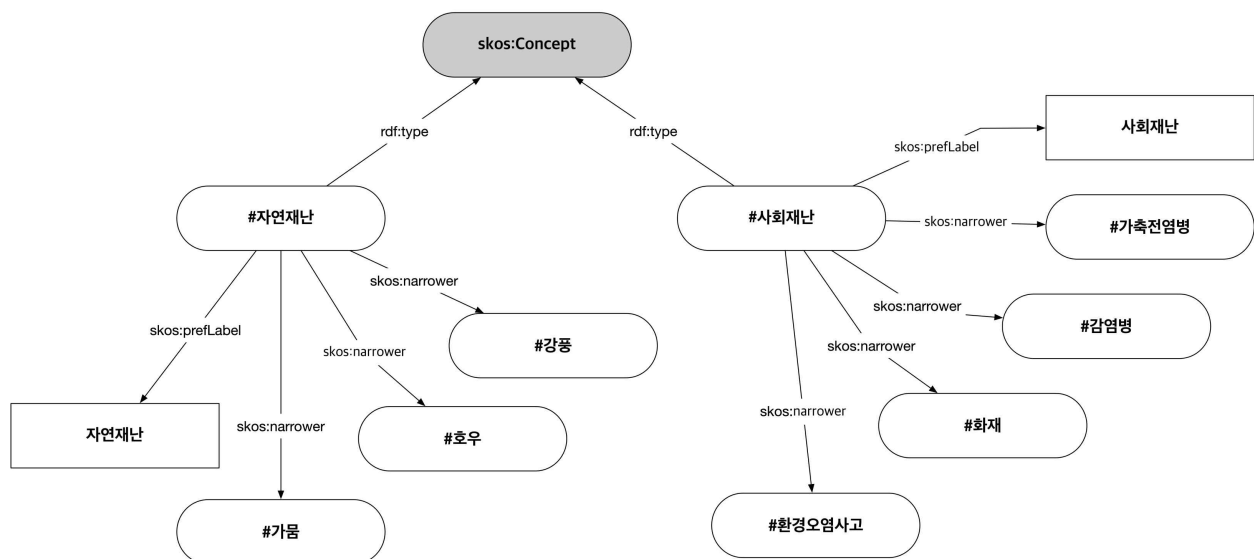
* 분류체계는 재난안전법의 재난유형을 사용함([표 3-4] 참고)

- 오아시스에서 재난아카이브 외의 디지털 아카이브 컬렉션은 KDC를 활용한 주제분류로 사용하고, 재난아카이브는 재난안전법의 재난유형을 사용함
- 서로 다른 분류체계를 사용하는 디지털 아카이브를 연계하기 위해 모든 분류체계의 항목을 skos:Concept으로 표현하고, 항목 간의 관계는 상위어(broader)-하위어(narrower) 관계 또는 관계성(related)만 표현하는 방식을 적용함(<그림 3-7> 참고)

[표 3-4] 재난안전법의 재난유형(재난안전법, 2022)

대분류	소분류
자연재난	가뭄, 강풍, 낙뢰, 대설, 자연우주물체의 추락·충돌, 조류대 발생, 조수, 지진, 태풍, 풍랑, 해일, 호우, 홍수, 화산활동, 황사, 기타
사회재난	가축전염병, 감염병, 교통사고, 항공사고, 해상사고, 국가기반체계의 마비, 붕괴, 폭발, 화생방사고, 화재, 환경오염사고, 해외재난

skos <http://www.w3.org/2004/02/skos/core#>
 분류체계안 <http://example.com/category/def/>



<그림 3-7> SKOS 어휘를 적용한 오아시스 재난아카이브 분류체계(안) (김학래, 2022)

제 4 장 ‘코로나19 아카이브’의 서비스 방안

4.1 데이터의 수집·관리와 연계

4.1.1 재사용이 가능한 코로나19 관련 자료의 수집과 관리

- **(필요성)** 대부분의 아카이브는 수집한 자료의 활용과 연계에 대한 계획이 미흡하므로, 수집된 자료의 활용을 위해 수집 과정에서부터 FAIR 데이터 원칙(FAIR Data Principle)을 도입하는 것이 필요함
 - 대부분의 웹 아카이브는 수집한 자료의 활용이 어렵고, 이는 웹 아카이브의 공통적인 문제로 작동함. 특히, 웹 아카이브가 수집한 자료 또는 콘텐츠를 활용한 사례가 매우 적음
 - FAIR 데이터 원칙은 사람의 개입을 최소화하며 웹 아카이브의 콘텐츠를 사용하기 위한 원칙으로 사용할 수 있음
- **(정의)** FAIR 데이터 원칙은 데이터를 찾을 수 있고(Findable), 데이터에 접근 가능하며(Accessible), 데이터의 상호운용이 가능하고(Interoperable), 데이터를 재사용(Reusable)할 수 있는 일련의 가이드라인임
 - 2016년 G20 항저우 정상회담에서 연구 분야에 FAIR 원칙의 적용에 서명했고, FAIR 원칙의 구체적인 실행 방안에 대한 국제적 논의가 지속적으로 진행되고 있음
 - FAIR 원칙은 오픈 사이언스(Open Science)의 맥락에서 연구 데이터의 재사용성을 강화하기 위한 도구로 인식되었으나, 현재는 장기적으로 데이터를 보존하고 관리하기 위한 보편적인 프레임워크로 인식됨
 - FAIR 데이터 원칙은 디지털 아카이브의 데이터 관리 원칙으로 적용할 수 있고, 데이터 관리체계의 기준으로 검토가 가능함
- **(FAIR 데이터 원칙의 요소)** FAIR 원칙은 탐색성(Findability), 접근성(Accessibility), 상호운용성(Interoperability), 재사용성(Reusability) 요소로 구분되고, 각 요소는 평가할 수 있는 세부 지표*로 구성됨
 - * 세부 지표는 [표 4-1]를 참고
 - **(탐색성)** 데이터를 재사용하는 첫 번째 단계는 데이터를 찾는 것임. 기계가 읽을 수 있는(machine-readable) 메타데이터는 데이터의 자동 검색에 필수적이므로, 데이터와 메타데이터는 사람과 기계가 모두 찾기 쉬운 방식으로 기술되어야 함
 - **(접근성)** 사용자는 인증(authentication)과 권한(authorization)을 포함해 필요한 데이터에 접근

근할 수 있어야 함. 특히, 데이터는 개방형 표준 프로토콜(예: HTTP)을 사용해 누구나 쉽게 접근할 수 있어야 함

- **(상호운용성)** 데이터는 일반적으로 다른 데이터와 통합되어야 함. 데이터는 데이터의 분석과 저장, 처리를 위해 어플리케이션이나 워크플로우에 관계없이 상호운용이 될 수 있어야 함
- **(재사용성)** FAIR의 궁극적인 목표는 데이터를 최대한 재사용할 수 있도록 하는 것임. 데이터와 메타데이터는 서로 다른 환경에서 복제나 통합이 가능하도록 잘 기술되어야 함

[표 4-1] FAIR 원칙의 세부 지표(GO FAIR, 2021)

FAIR 구분		내용
탐색성	F1	(메타)데이터는 전역적으로 고유하고 영구적인 식별자가 할당되어야 한다.
	F2	데이터는 풍부한 메타데이터로 기술되어야 한다.
	F3	메타데이터는 기술하는 데이터에 대한 명확하고 명시적인 식별자를 포함해야 한다.
	F4	(메타)데이터는 검색 가능한 자원에 등록 또는 색인화되어야 한다.
접근성	A1	(메타)데이터는 표준화된 통신 프로토콜을 사용하여 식별자로 검색된다.
	A1.1	프로토콜은 개방형이며, 무료이고 보편적인 방법으로 구현 가능해야 한다.
	A1.2	프로토콜은 (필요한 경우) 인증과 권한 부여 절차를 허용해야 한다.
	A2	(메타)데이터는 더 이상 사용할 수 없는 경우에도, 계속 접근할 수 있어야 한다.
상호운용성	I1	(메타)데이터는 지식 표현을 위해 공식적이고 접근 가능하며 공유되고 광범위하게 적용 가능한 언어를 사용해야 한다.
	I2	(메타)데이터는 FAIR 원칙을 따르는 어휘를 사용해야 한다.
	I3	(메타)데이터는 다른 메타데이터에 대한 정규화된 참조가 포함되어야 한다.
재사용성	R1	(메타)데이터는 정확하고 관련성 높은 속성이 여러 개 있어야 한다.
	R1.1	(메타)데이터는 명확하고 접근 가능한 라이선스로 공개되어야 한다.
	R1.2	(메타)데이터는 상세화된 출처와 관련이 있다.
	R1.3	(메타)데이터는 도메인과 관련된 커뮤니티 표준을 충족해야 한다.

- **(아카이브의 FAIR 원칙 적용)** 아카이브 관점에서 FAIR 원칙은 디지털 자원의 탐색과 상호 연결, 지속 가능성을 위한 보존·관리 프레임워크를 제공할 수 있음
 - 대부분의 국내 디지털 아카이브는 디지털 자원의 탐색성과 상호운용성, 재사용성이 낮은 형태로 관리되고 있음. 대부분의 디지털 자원은 탐색에 활용할 수 있는 메타데이터 정보가 미흡하고, 표준 어휘를 사용하지 않아 데이터의 연계가 어려운 형태로 제공됨
 - FAIR 데이터 원칙은 디지털 자원의 기계가용성(machine-actionable)과 재사용성을 중심으로 평가함. 디지털 자원이 수집 단계에서부터 FAIR 원칙을 적용해 관리·보존된다면, 아카이브가 수집한 디지털 자원을 최대한 활용할 수 있게 됨
- * 불임3은 디지털 아카이브를 위한 FAIR 데이터 평가 프레임워크를 제공함

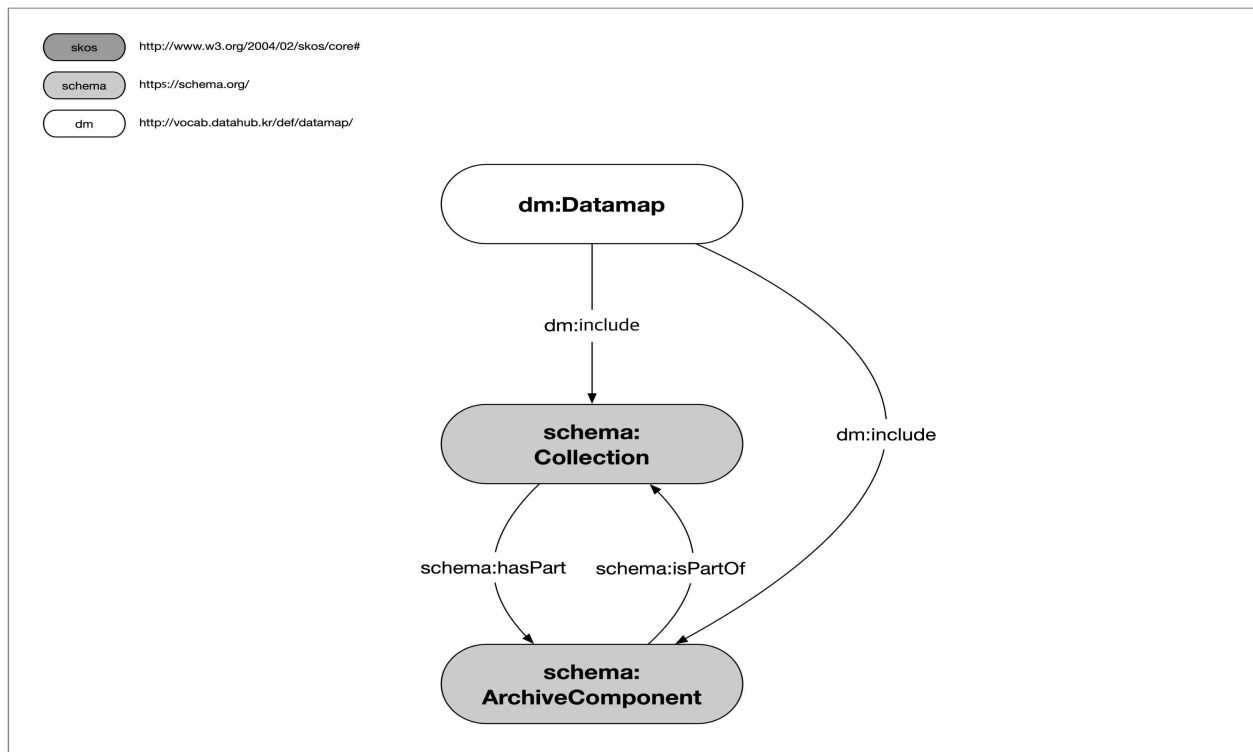
4.1.2 코로나19 관련 자원의 연계

- **(필요성)** 수집된 자원은 연계와 활용이 어려운 형태로, FAIR 원칙을 준수하고 국제·국내 표준을 적용한 자원을 구축해야 함
 - FAIR 원칙은 상호운용성을 강조하며, 기계가 읽을 수 있는 형태의 자원을 구축하기 위해 표준 어휘를 사용하여 자원을 표현할 것을 제시함. FAIR 원칙은 주제 분야의 특성을 고려해 기존 도메인의 표준 어휘를 사용할 것을 권장함
 - 코로나19 관련 자원의 형태와 주제 분야는 다양할 수 있지만, 디지털 자원의 관점에서 국제·국내 표준인 Schema.org와 데이터맵을 활용해 기계가 읽을 수 있는 형식으로 자원을 기술할 수 있음
 - 기계가 읽을 수 있는 형식의 디지털 자원은 RDF 형식으로 변환되고, 데이터가 처리될 수 있는 환경이 구축되면 연계된 디지털 자원의 질의와 탐색이 가능해짐
- **(Schema.org)** 구글, 마이크로소프트, 야후와 애플에서 공동 개발한 어휘로, 광범위한 주제(예: 사람, 장소, 사건)를 의미적으로 표현하기 위해 만들어진 RDF 어휘임
 - Schema.org는 광범위한 지식을 표현하는 797개의 유형과 1,453개의 속성을 갖고 있고, 확장(Extension)을 통해 특정 도메인(예: 도서관과 서지, 아카이브)의 어휘를 추가할 수 있음
 - Schema.org는 다차원적인 기록 체계를 표현하는데 한계가 있지만, 디지털 아카이브를 링크드 데이터(linked data)로 쉽게 발행할 수 있고, 일반적인 검색 엔진(예: 구글)에서 디지털 자원을 보다 쉽게 검색할 수 있음
 - Schema.org의 아카이브 커뮤니티 그룹은 디지털/물리적 아카이브(digital and physical archive)와 내용(contents)을 표현하기 위한 스키마를 개발하고 있음. 대표적으로 schema:ArchiveComponent는 디지털 아카이브에 포함된 디지털 객체를 표현하는 클래스로, 기존의 클래스(예: Collection, CreativeWork)와 결합하여 사용할 수 있음

- (코로나19 관련 자원 연계를 위한 지식 모델) 코로나19 관련 아카이브에 담긴 자원과 메타데이터는 Schema.org와 데이터맵 어휘를 적용하여 모델을 구성하고, 구축된 데이터는 FAIR 원칙을 준수하며 기계가 읽을 수 있는 형식으로 표현됨
- 단일한 웹 아카이브 또는 디지털 아카이브는 dm:Datamap으로 표현되고, 하나 이상의 컬렉션(schema:Collection) 또는 하나 이상의 자원(schema:ArchiveComponent)이 하나의 dm:Datamap에 포함(dm:include)될 수 있음. 즉, 데이터맵은 컬렉션 또는 자원이 목록화된 집합으로 표현됨
- 단일한 아카이브가 포함하는 하나 이상의 컬렉션은 schema:Collection²⁶⁾으로 표현됨. schema:Collection은 하나 이상의 자원(schema:ArchiveComponent)으로 구성되고, 하나의 자원은 컬렉션의 일부(schema:hasPart)로 표현됨
- 컬렉션의 상세 정보는 속성으로 표현이 가능함. 예를 들어, 컬렉션이 포함하는 자원의 개수는 schema:collectionSize 속성으로, 컬렉션의 출처 정보는 schema:archiveAt 속성으로 표현할 수 있음
- 아카이브에 보존된 자원은 schema:ArchiveComponent²⁷⁾로 정의할 수 있음. 컬렉션에 포함된 자원은 컬렉션의 일부(schema:isPartOf)로 연결되고, schema:hasPart와 역관계로 구성됨. 웹 아카이브에 포함되거나 컬렉션에 포함되지 않는 자원은 웹 아카이브(dm:Datamap)와 직접 포함관계(dm:include)를 가짐
- 아카이브에 보존된 물리적 자원 또는 디지털 자원의 상세정보는 속성으로 표현할 수 있음. 속성은 물리적 자원이 보존된 위치(schema:itemLocation), 아카이브를 보유하는 기관(schema:holdingArchive) 등을 표현할 수 있음

26) <https://schema.org/Collection>

27) <https://schema.org/ArchiveComponent>

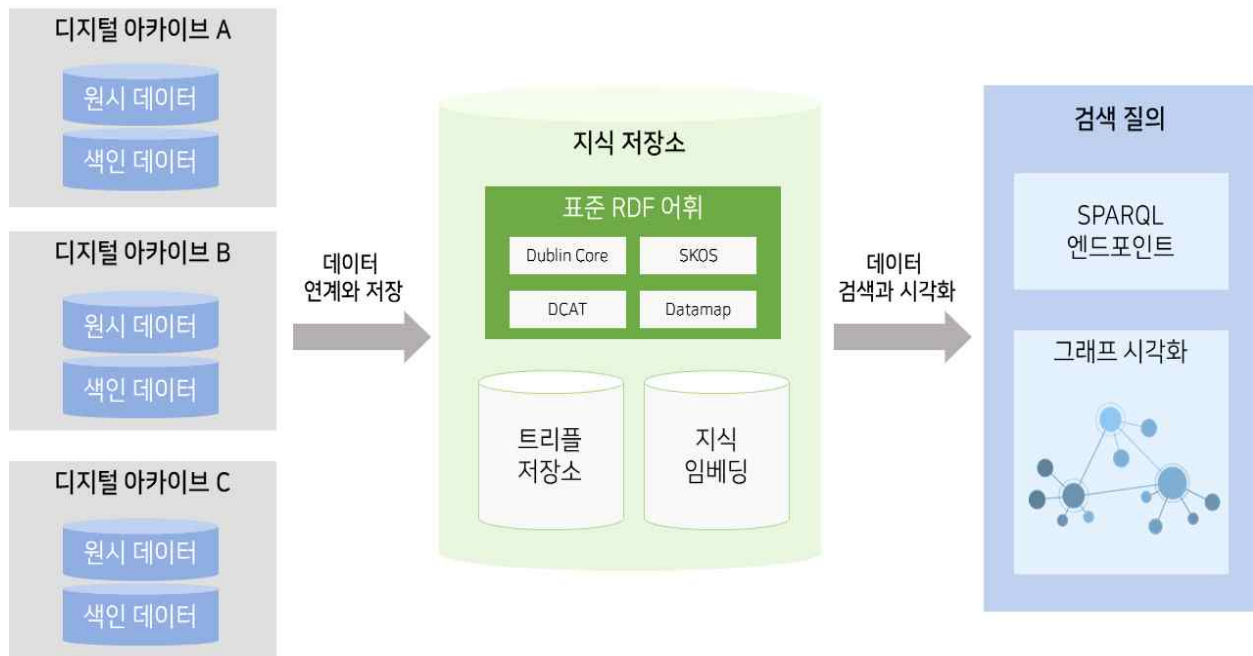


〈그림 4-1〉 코로나19 관련 자원의 연계를 위한 지식 모델(김학래, 2022)

4.2 기술적 고려사항

4.2.1 코로나19 자원의 통합 검색 서비스

- 코로나19 아카이브와 OASIS의 통합 검색 서비스는 서로 다른 유형의 자원을 검색하기 위해 메타데이터를 의미적으로 연계하고 통합할 수 있는 기술과 환경을 고려해야함
- 자원 사이의 의미적인 관계는 RDF 어휘를 적용한 표준 어휘(Dublin Core, SKOS, DCAT, Datamap)를 통해 모델링할 수 있고, RDF 형식으로 변환된 데이터를 처리하기 위한 트리플 저장소가 필요함
- 특히 다양한 자원의 메타데이터 정보 제공, 자원 사이의 관계 식별, 암묵적인 관계를 추론할 수 있는 지식그래프 기술(지식그래프 저장소, 지식그래프 모델링, 지식그래프 URI 설계 등)이 필요함
- 〈그림 4-2〉은 복수의 디지털 아카이브에서 자원을 통합 검색하기 위한 아키텍처를 표현한 것으로, 디지털 아카이브의 메타데이터를 RDF 어휘로 변환하고 연계한 후 검색할 수 있는 시스템 구조를 나타냄



〈그림 4-2〉 디지털 아카이브 통합 검색 서비스 아키텍처(김학래, 2022)

- (그래프 데이터베이스) RDF 어휘를 사용한 데이터는 트리플 구조 (Subject-Property-Object)를 표현할 수 있는 RDF/XML, Turtle 등의 형식으로 표현되고, 트리플 데이터를 저장하는 그래프 데이터베이스(Graph Database)가 존재함
- (SPARQL 엔드포인트와 질의) SPARQL(Simple Protocol and RDF Query Language)은 RDF 기반의 온톨로지 질의(Query) 언어이고, PREFIX, SELECT, ASK, DESCRIBE, CONSTRUCT, WHERE 등의 명령어로 데이터를 탐색함
 - 디지털 아카이브를 RDF 어휘로 변환하고 그래프 데이터를 트리플 저장소에 저장하고 관리하면 SPARQL 질의문을 통해 의미적인 검색을 할 수 있음
 - [표 4-2]는 데이터맵 어휘로 표현한 디지털 아카이브가 보유한 정보 자원의 총 수량을 고유한 URI 기준으로 집계하는 질의문으로, 데이터맵(dm:Databmap)은 복수의 데이터맵을 통합할 수 있기 때문에 다양한 디지털 아카이브의 관리 현황을 손쉽게 파악할 수 있음
 - [표 4-3]는 정보 자원의 분류체계를 SKOS(skos:Concept)로 표현하여 분류 항목 사이의 관계(skos:related, skos:narrower, skos:broader)를 연계한 데이터를 구축하고, 복수의 데이터맵에서 ‘감염병’으로 분류하는 자원을 검색하는 질의문임

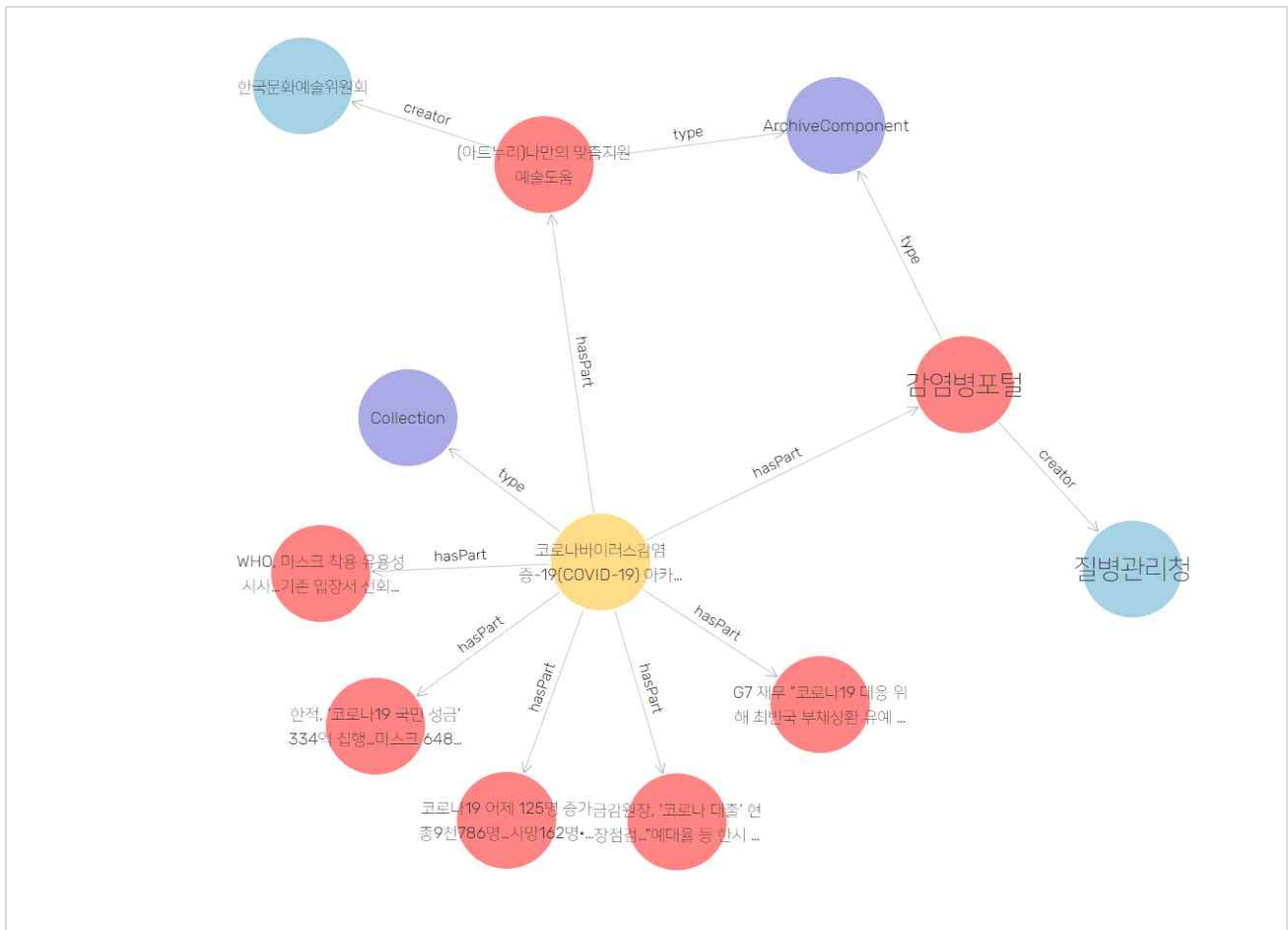
[표 4-2] 디지털 아카이브가 보유한 자원의 총 수량 질의문

질의 내용	질의문
다수의 디지털 아카이브가 통합된 자원의 총 수량	<pre> PREFIX dm: <http://vocab.datahub.kr/def/datamap/> PREFIX dct: <http://purl.org/dc/terms/> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX dcat: <http://www.w3.org/ns/dcat#> SELECT (SUM(?contentsCount) AS ?sum) WHERE { :Archive-1 a dm:Datamap ; dm:datamapStat ?stat1 ; dm:hasDatamap ?subArchive. ?subArchive a dm:Datamap ; dm:datamapStat ?subStat . ?subStat a dm:DatamapStat ; dm:numberOfDataset ?contentsCount .}</pre>

[표 4-3] SKOS 어휘를 사용한 분류체계 질의문

질의 내용	질의문
주제분류에서 ‘감염병’ 또는 관련 주제로 분류된 정보 자원을 보유한 데이터맵, 분류체계명, 수량을 탐색하는 질의문	<pre> PREFIX dm: <http://vocab.datahub.kr/def/datamap/def/> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX dcat: <http://www.w3.org/ns/dcat#> PREFIX skos: <http://www.w3.org/2004/02/skos/core#> PREFIX dct: <http://purl.org/dc/terms/> SELECT distinct ?DatamapName ?relatedName (count(distinct ?dataset) as ?dsCount) where { ?Datamap dm:dataset ?dataset ; rdfs:label ?DatamapName . {?dataset a dcat:Dataset ; dcat:theme :감염병 ; dcat:theme/skos:prefLabel ?relatedName .} union {?dataset a dcat:Dataset ; dcat:theme/skos:relatedMatch :감염병 ; dcat:theme/skos:prefLabel ?relatedName} filter(lang(?relatedName)='ko') } group by ?DatamapName ?relatedName</pre>

- (그래프 시각화) RDF 어휘로 표현한 자원을 이해하기 쉽도록 그래프 형태의 시각화를 제공할 수 있음. 웹 서버 또는 로컬 환경에서 구축된 트리플 저장소의 SPARQL 엔드포인트에 접근하여 특정 자원을 검색하고 연계된 데이터를 시각화함
- <그림 4-3>은 코로나19 관련 자원 연계를 위한 지식 모델을 적용한 그래프 데이터에서 코로나19 아카이브와 개별 자원의 관계를 나타냄



<그림 4-3> ‘코로나19 아카이브’ 검색 그래프 시각화

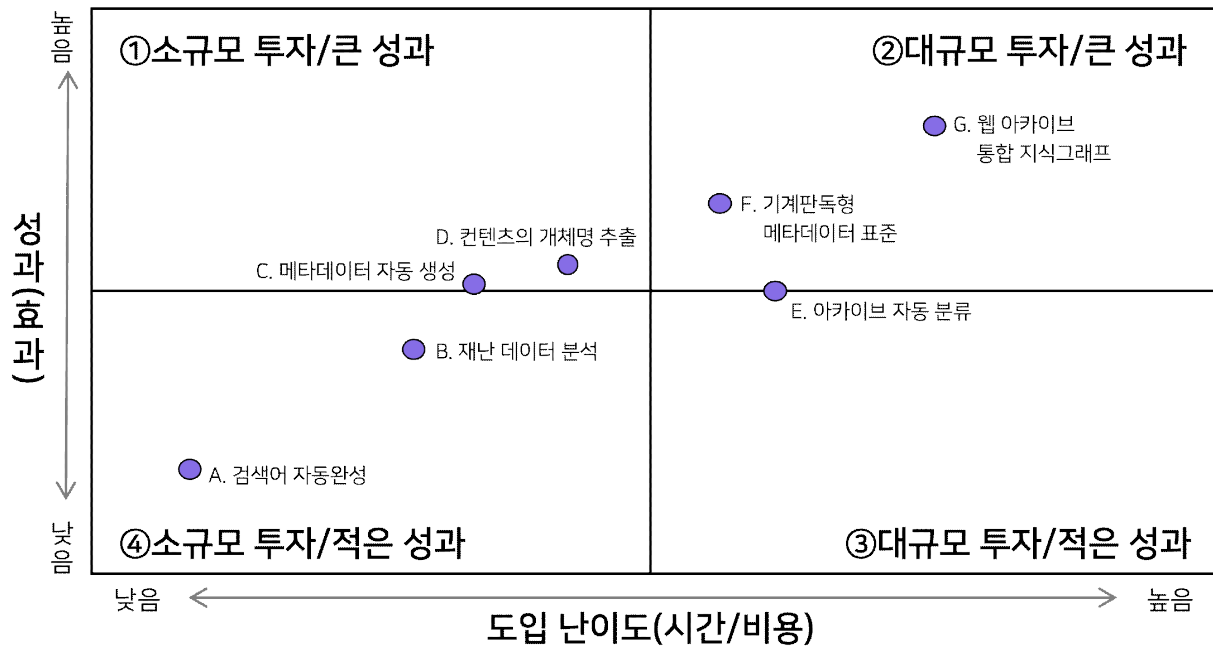
4.2.2 지능형·소통형 서비스를 위한 인공지능 서비스 모델 분석

- 페이오프 매트릭스(Pay-OFF Matrix)를 기준으로 ‘코로나19 아카이브’에 적용 가능한 인공지능 서비스 모델의 도입 효과를 분석함
- 페이오프 매트릭스는 팀이나 기업의 의사결정 과정에 사용하며, 제한적 합리성(bounded rationality)을 기반으로 의사결정에 효과적임
- 딥러닝, 머신러닝과 자연어 처리 기술의 도입으로 예상되는 효과, 도입 난이도(시간/비용)를 기준으로 분석함

[표 4-4] 지능형·소통형 서비스를 위한 인공지능 서비스 모델 목록

항목	서비스 이름	서비스 설명
A	검색어 자동완성	검색창에 키워드를 입력하면 예상되는 검색어를 추천해주는 기능
B	재난 데이터 분석	아카이브가 보유한 여러 자료를 기반으로 데이터를 분석하여 인사이트를 제공하는 서비스
C	메타데이터 자동생성	디지털 자료의 메타데이터를 자동으로 생성해주는 기능
D	콘텐츠의 개체명 인식·추출	텍스트 기반 콘텐츠에서 유의미한 개체명을 인식하고 추출하는 기능
E	아카이브 자동 분류	유사도를 기반으로 아카이브 콘텐츠를 자동으로 분류하는 기능
F	기계판독형 메타데이터 표준	메타데이터를 일관적으로 기술하는 체계를 구축하고, 기계가 처리할 수 있는 형태로 메타데이터를 구축하는 서비스
G	웹 아카이브 통합 지식그래프	아카이브의 자료를 의미적으로 연계하고, 인공지능이 처리 가능한 형태로 지식베이스를 구축하는 서비스

- 페이오프 매트릭스의 사분면은 도입 난이도 대비 성과에 따라 구분되고, 각 사분면은 다음과 같이 해석됨
 - (1. 소규모 투자/큰 성과) 비교적 쉽게 구현이 가능하나, 실제 오아시스에서 서비스되지 않은 영역에 해당됨. C와 D는 오아시스에 쉽게 적용이 가능하나, 오아시스에서 제공하지 않는 서비스 모델임
 - (2. 대규모 투자/큰 성과) 고수준의 기술과 인프라가 기반 되어야 하는 서비스 영역에 해당됨. E와 G는 현재 오아시스가 도입하지 않는 서비스 모델이고, F는 오아시스에서 확대 적용할 수 있음
 - (3. 대규모 투자/작은 성과) 유지비용이 높지만, 혁신성이 낮은 영역에 해당됨. 오아시스에 도입하기 쉽지 않고, 비교적 큰 성과를 얻기 어려움
 - (4. 소규모 투자/작은 성과) 중요도나 활용성은 비교적 낮지만, 시장에서 보편적으로 관련 기능을 제공하는 영역임. 오아시스에서 부분적으로 제공하는 서비스 중 확대 적용이 가능한 모델은 A과 B임



〈그림 4-4〉 인공지능 서비스 모델의 도입 난이도 대비 성과 현황

4.2.3 인공지능 서비스의 도입을 위한 우선순위

- 웹 아카이브에 적용하기 위해 인공지능 서비스 모델을 단기·중기·장기 과제로 구분함
 - (단기 과제) 6개월 이내에 도입할 수 있는 서비스 또는 기능에 해당됨. 오아시스의 일부 서비스에 확대 적용이 가능한 서비스 모델 A와 B가 단기 과제에 포함됨
 - (중기 과제) 1~3년의 기술 개발과 적용이 필요한 서비스 또는 기능에 해당됨. 서비스 모델과 관련된 SDK가 있어 비교적 개발이 쉬우나, 오아시스에서 제공하지 않는 서비스 모델 C와 D, E가 해당됨
 - (장기 과제) 3~5년의 기술 개발과 적용이 필요한 서비스 또는 기능에 해당됨. 단기 또는 중기 과제가 선행되어야 도입할 수 있는 서비스 모델이거나 관련 SDK를 직접 개발해야 하는 서비스 모델 F와 G가 해당됨

[표 4-5] 단기·중기·장기 과제의 개발 소요 시간과 서비스 모델

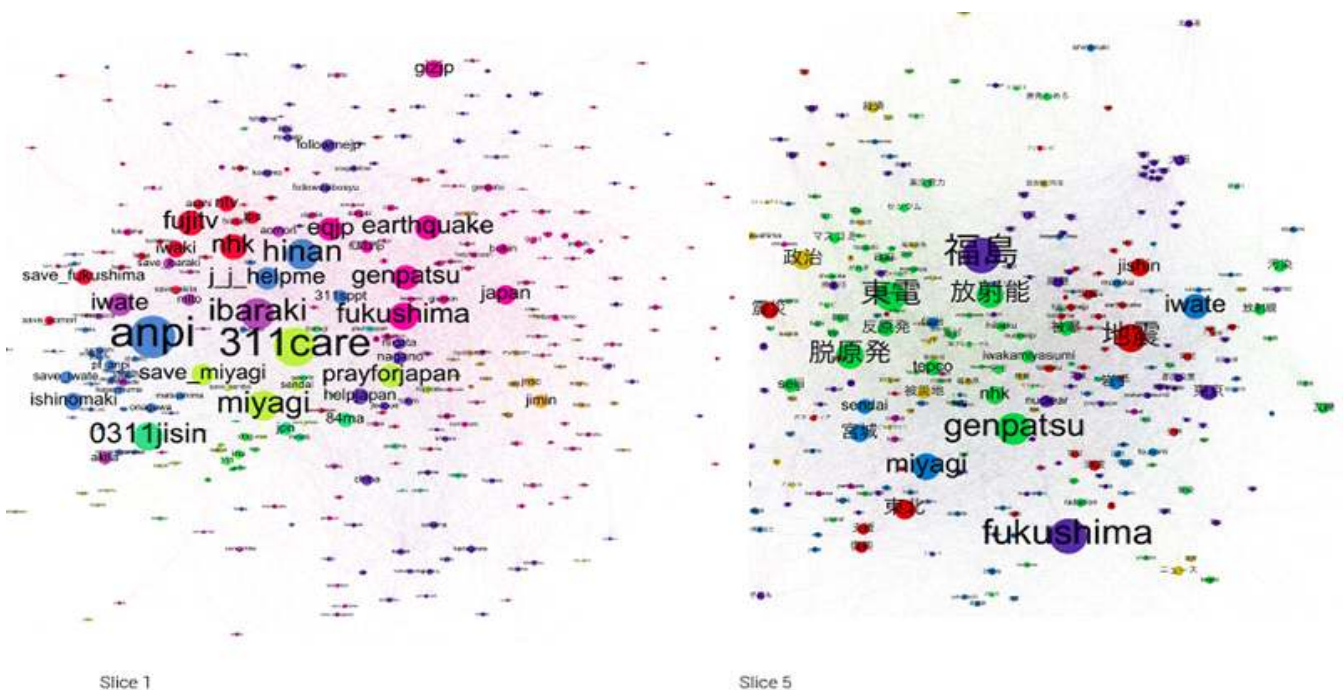
분류	단기	중기	장기
개발 소요 시간	6개월 이내	1~3년	3~5년
서비스 모델	A. 검색어 자동완성 B. 재난 데이터 분석	C. 메타데이터 자동 생성 D. 콘텐츠의 개체명 추출 E. 아카이브 자동 분류	F. 기계판독형 메타데이터 표준 G. 웹 아카이브 통합 지식그래프

- (A. 검색어 자동완성) 검색창에 키워드를 입력할 때 자동으로 예상 키워드를 추천하여 검색의 효율성을 높이는 서비스
 - (세부 내용) 검색어로 입력한 키워드는 자소 단위로 분석해 예상되는 키워드를 자동으로 완성해주는 기능임. 예를 들어, ‘코로나19’를 입력하면 코로나19와 관련된 ‘코로나19 영향’, ‘코로나19 대학’ 등의 키워드를 자동으로 제공함(<그림 4-5>). 다양한 SDK가 공개되고 있어 단기간에 도입할 수 있는 기능에 해당됨
 - (관련 기술 또는 사례) 검색어 자동완성과 관련된 다양한 언어의 오픈소스 라이브러리(예: algolia의 autocomplete, JQuery의 autocomplete)이 공개되어 있어 쉽게 구현이 가능함. 검색을 지원하는 대부분의 포털이나 디지털 아카이브는 이용자 편의를 위해 검색어 자동완성 기능을 제공함
 - (도입 효과) 오아시스는 2004년부터 수집한 방대한 양의 웹 자원을 보유하고 있지만, 웹 자원을 효과적으로 검색할 수 있는 기능을 제공하지 않고 있음. 오아시스에 검색어 자동완성 기능이 적용되면 이용자가 원하는 키워드를 가진 웹 자원이 있는지 쉽게 파악할 수 있고, 검색어와 관련된 맥락 정보를 얻을 수 있음



<그림 4-5> 국회·지방의회 의정정보시스템의 검색어 자동완성 기능(국회·지방의회 의정정보시스템, 2022)

- (B. 재난 데이터 분석) 재난과 관련된 데이터 분석과 시각화를 제공하여 이용자가 효과적으로 정보를 탐색할 수 있는 서비스
 - (세부 내용) 오아시스에 수집된 코로나19 관련 웹 자원의 분석을 통해 코로나19가 한국 사회에 미친 영향을 다각도로 분석하는 서비스를 제공할 수 있음. 데이터 분석에 필요한 기술적 장벽이 비교적 낮고, 오아시스의 재난아카이브가 보유하는 웹 자원을 기반으로 단 기간에 수행할 수 있음
 - (관련 기술 또는 사례) Japan Disasters Digital Archive²⁸⁾는 2011년 3월에 발생한 쓰나미와 지진 피해를 기록하는 디지털 아카이브로, 관련 트윗이나 증언, 영어뉴스·기사 전문 등의 디지털 자원을 보존하고 있음
 - 대표적인 데이터 분석으로 쓰나미와 관련된 소셜미디어의 해시태그를 분석하여 재난이 사회에 미친 영향을 분석함(<그림 4-6>). Gephi를 활용하여 해시태그를 클러스터링하고, 해시태그의 출현 빈도 또는 중심성을 분석함
 - 분석 결과는 재난 당시에 재난 피해자를 위한 도움 요청이 활발했고, 재난 이후는 재난으로 인한 사회적, 정치적 영향이 논의되었음
 - (도입 효과) 오아시스는 웹트렌드 분석을 통해 특정 키워드의 출현 빈도를 제공하고 있음. 그러나 출현 빈도는 통계 수치만 드러낼 뿐, 오아시스가 보유하는 웹 자원에 내재된 의미나 인사이트를 도출하지 않음
 - 오아시스는 코로나19와 관련된 다양한 웹 자원을 보존하고 있고, 이를 효과적으로 분석한다면 코로나19 재난에 대해 의미 있는 인사이트를 제공할 수 있을 것으로 기대됨



<그림 4-6> Japan Disasters Digital Archive의 재난 관련 소셜미디어 해시태그 분석(Japan Disasters Digital Archive, 2022)

28) <https://jdarchive.org/en>

- (C. 메타데이터 자동 생성) 자동으로 디지털 자원의 메타데이터를 생성하는 기능
 - (세부 내용) 메타데이터는 디지털 자원의 검색에 핵심적인 역할을 수행하고, 디지털 자원을 이해하는데 도움이 되는 맥락 정보를 제공함
 - 그러나 사람이 광범위한 디지털 자원의 모든 메타데이터를 기술하는데 한계가 있음. 인공지능 기반의 메타데이터 생성 기능은 디지털 자원을 관리하는 사자에게 훌륭한 보조 도구가 될 수 있음
 - 한편, 웹 자원의 특성을 파악하여 적합한 메타데이터 항목과 값을 설계해야 하므로 중기적인 계획을 수립해야 함
 - (관련 기술 또는 사례) Demosaurus²⁹⁾는 네덜란드 국립도서관 연구소에서 개발한 인공지능 기반의 메타데이터 자동 생성 라이브러리임. Demosaurus는 각 작품의 저자와 출판사 형식과 같은 메타데이터를 자동으로 추가할 수 있는 기능을 제공함
 - 머신러닝을 활용해 해당 작품의 저자일 가능성이 큰 인물을 추천하고, 저자의 상세정보는 네덜란드 NTA(저자 이름 시소러스)와 연결하여 상세 정보를 제공함
 - (도입 효과) 오아시스의 디지털 자원은 MODS를 사용하여 메타데이터가 기술됨. 메타데이터 자동생성 도구는 MODS로 기술되는 항목의 값을 추천하여 효율적으로 메타데이터를 생성하도록 도울 수 있음

- (D. 콘텐츠의 개체명 인식·추출) 웹 자원의 콘텐츠에서 이름을 가진 개체(named entity)를 인식하고, 개체의 유형을 파악하는 기능
 - (세부 내용) 웹 자원의 콘텐츠를 파악할 수 있는 기술로, 개체명 인식을 통해 재난과 관련된 특정 인물 또는 사건명을 추출할 수 있는 기능을 제공함
 - 개체명 인식을 위한 다양한 오픈소스 라이브러리(예: NLTK, SpaCy, Flair)가 존재하지만, 한국어로 표현된 재난아카이브 관련 개체명 인식을 지원해야 하므로 중기적 관점에서 준비해야 함
 - (관련 기술 또는 사례) 한국어 개체명 인식 관련 데이터세트는 국립국어원에서 제공하는 개체명 사전(2021)³⁰⁾과 네이버에서 제공하는 개체명 인식 데이터세트³¹⁾가 있음
 - 대표적으로 포르투갈의 기록 보관소는 국가의 역사를 담고 있는 디지털 자원을 개방하지만, 디지털 자원에 대한 주석이 제공되지 않아서 디지털 자원에 포함된 내용의 개체를 인식하고 검색 결과에 활용될 수 있도록 함
 - (도입 효과) 웹 자원의 활용성을 강화할 수 있는 기술이며, 웹 자원의 콘텐츠에서 인물이나 사건명을 추출할 수 있음. 오아시스의 재난아카이브는 키워드를 메타데이터로 표현하지 않으므로, 관련 인물 또는 사건을 메타데이터로 표현하면 아카이브의 검색 성능이 향상될 수 있고 웹 자원을 파악할 수 있는 정보를 제공할 수 있음

29) <https://github.com/KBNLresearch/Demosaurus>

30) https://rlkujwkk7.toastcdn.net/71/NIKL_DX_2021_v1.0.pdf

31) http://air.changwon.ac.kr/?page_id=10

Ousted **WeWork** founder **Adam Neumann** lists his **Manhattan** penthouse for **\$37.5 million**
 [organization] [person] [location] [monetary value]

<그림 4-7> 한 문장 단위의 개체명 인식(Analytics Vidhya, 2022)

- (E. 아카이브 자동 분류) 내용의 유사도를 기반으로 디지털 자원을 자동으로 분류하는 기능
 - (세부 내용) 오아시스의 재난아카이브는 여러 유형의 재난으로 구분됨. 인공지능을 활용해 수집한 웹 자원 중 재난에 해당되는 자원을 구분하고, 각 웹 자원의 재난 유형을 자동으로 분류하는 기능을 제공할 수 있음
 - 한편, 웹 자원 중 재난에 관련된 자원을 구분하고, 웹 자원의 재난 유형을 분류하기 위한 학습 데이터와 분류체계를 마련해야 하므로 중기적인 계획이 필요함
 - (관련 기술 또는 사례) NDC Predictor는 일본 국립국회도서관에서 개발한 서지 자동 분류 서비스임. 머신러닝을 사용해 도서의 서지정보로부터 일본의 십진분류법(NDC 9판)을 자동으로 분류하는 서비스를 제공함
 - 네덜란드 국립도서관은 머신러닝을 활용하여 자동으로 듀이 십진분류법을 생성하는 서비스를 개발함. 다층 퍼셉트론(MLP), 합성곱신경망(CNN) 등의 인공지능 기술을 활용해 DDC 분류 성능을 높임
 - (도입 효과) 오아시스가 수집하는 웹 자원은 광범위함. 사람이 웹 자원 중에서 재난아카이브에 포함될 자원을 선별하고, 재난유형을 구분하는 것은 한계가 있음. 아카이브 자동 분류는 사람이 웹 자원의 분류를 보조할 수 있는 수단이 됨
- (F. 기계판독형 메타데이터 표준) 웹 자원의 메타데이터의 항목과 값이 일관적으로 기술되고, 모든 항목의 메타데이터가 기계가 읽고 이해할 수 있는 형식(machine-readable)으로 표현함
 - (세부 내용) 오아시스의 웹 자원은 MODS를 사용하여 메타데이터를 기술하고, XML 형태로 메타데이터를 배포하고 있음. 그러나 MODS와 같이 XML로 표현되는 메타데이터는 구조와 형식은 통일할 수 있으나, 다른 메타데이터 스키마와 상호운용성이 낮음
 - 인공지능이 자동으로 읽고 이해하는 RDF 방식의 메타데이터 체계를 검토할 필요가 있음. 기계판독형 메타데이터 구축은 MODS로 표현된 메타데이터를 RDF 형식으로 변환하기 위해 광범위한 작업이 요구되므로 장기적인 계획이 필요함
 - (관련 기술 또는 사례) 미국 의회도서관은 XML로 표현된 MODS 메타데이터를 RDF 형식으로 변환하거나 BIBFRAME으로 변환하는 작업을 진행하고 있음³²⁾
 - BIBFRAME은 미국 의회도서관에서 서지 표현을 위해 개발한 RDF 어휘로, MARC 21를 대체하기 위해 개발되었음

32) <https://www.loc.gov/standards/mods/modsrdf/>

- **(도입 효과)** RDF로 표현된 메타데이터는 기계가 읽고 이해할 수 있는 형식으로 표현되고, 웹 자원의 상호운용과 재사용이 가능해짐
- **(G. 웹 아카이브 통합 지식그래프)** 다양한 웹 아카이브의 정보를 의미적으로 연결하여 관계를 만들고, 인공지능이 읽고 이해할 수 있는 형태로 지식베이스를 형성함
 - **(세부 내용)** 인공지능이 읽고 이해할 수 있는 형태의 표준 어휘를 사용하여 웹 자원 사이의 관계를 의미적으로 표현하고, 텍스트로 표현된 메타데이터 값을 엔티티(entity)로 표현하여 웹 자원의 풍부한 맥락 정보를 제공할 수 있음
 - 지식그래프의 구축은 재난아카이브의 웹 자원을 파악·정제하고, 그 관계를 지식 모델로 표현해야 하는 절차가 필요하므로 장기적 관점에서 통합 지식그래프를 구축하는 것이 필요함
 - **(관련 기술 또는 사례)** WarSampo는 핀란드에서 구축한 제2차 세계대전에 대한 시맨틱 웹 디지털 아카이브로, 제2차 세계대전에서 핀란드와 관련된 이종의 분산된 데이터셋을 링크드 오픈 데이터(Linked Open Data)로 통합함(<그림 4-8> 참고)³³⁾
 - 일본 국립국회도서관은 다양한 분야(예: 도서/간행물, 문화재, 미디어 예술)의 디지털 아카이브를 연계해 광범위한 디지털 자원의 검색과 활용 서비스를 제공하는 국가 플랫폼임³⁴⁾. 83개의 기관에서 운영하는 디지털 아카이브를 통합하기 위해 RDF 기반의 JSP-RDF 스키마로 메타데이터를 표현하고, Web API와 SPARQL API로 통합된 디지털 자원의 검색을 지원함
 - **(도입 효과)** 웹 자원의 메타데이터는 이종의 시스템에서 상호운용이 가능하고, 데이터의 공유와 교환이 용이한 형태로 구축될 수 있음. 텍스트로 제공되는 메타데이터의 값은 외부의 데이터(예: 위키데이터)와 연계되어 웹 자원을 이해하는데 유용한 정보를 제공할 수 있음

33) <https://www.sotasampo.fi/en/>

34) <https://jpsearch.go.jp/>

Persons

Search for known persons from the past Finnish wars by writing their name in the text input below and/or selecting a person from the list below. Information regarding the person and recommended links will appear on the right. **If you cannot find the person you are looking for, and know in which military unit they have served, you can take a look at the unit's timeline.**

Search by person name

Aakala, Aarne Eetu (Sergeant)

Aakko, Juho Vilhelm (Jääkäri)

Aakkonen, Antti (Sergeant)

Aakkula, Antero (Major)

Aakkula, Eino Rikhard (Corporal)

Aaku, Eero (Captain)

Aalto, Aulis Antero (Private)

Aaitto, Erkki

Aalto, Aarne Ensio (Private)

Aalto, Aarne Erkki (Private)

Aalto, Alli Valfrid (Private)

Aalto, Arvo Aadolf (Lance Corporal)

Aalto, Arvo Anselm (Private)

Aalto, Eero August (Lance Corporal)

Aalto, Einar Johannes (Jääkäri)

Aalto, Eino Kaarlo (Corporal)

Aalto, Erkki (Seaman)

Aalto, Frans Jalmar (Private)

Aalto, Ilmari (Private)

Aalto, Jaakko Johannes (Private)

Aalto, Juho Kustaa Armas (Corporal)

Aalto, Kaario Albert (Military Engineer)

Aalto, Kaario Antero (Lance Corporal)

Aalto, Kaario Emil (Private)

Aalto, Kaale Edvard (Lance Corporal)

Aalto, Kosti Johannes (Private)

Aalto, Lahja Johannes (Lieutenant Colonel)

Aalto, Lauri (Private)

Aalto, Lauri (Private)

Aalto, Lauri (Lance Corporal)

Aalto, Lauri Edvard (Private)

Paavo Juho Talvela

InformationTimelinePhotographs

Paavo Juho Talvela (January 19, 1897 in Vantaa – September 30, 1973, Helsinki) was a Finnish soldier and a Knight of the Mannerheim Cross. He was one of the volunteers who served in the Finnish Jaeger battalion in Germany in 1916 to 1917. He was a battalion commander in the Finnish Civil War. In 1919 he took part in the Aunus expedition as Commander in Chief.

During the Winter War (1939 - 1940), Talvela commanded "Group Talvela" which took part in the Battle of Tolvajärvi. For this success he was promoted to Major General in December 1939, the first promotion to general's rank during the war. In February 1940 Talvela took the command of the Finnish III Corps in the Karelian Isthmus. When the war ended on 13 March 1940, Talvela returned to civilian life. However, once the Finnish-German relations warmed, he was used in semi-official missions to Germany in late 1940.

During the early Continuation War Talvela commanded the Finnish IV Corps in Karelia. From January 1942, when he was promoted to Lieutenant General, until February 1944 Talvela was the Finnish representative at the German High Command. Once back in Finland, Talvela commanded first the Finnish II Corps in northern Karelia until June 1944 when he took over the command of the Aunus Group. In July 1944 Talvela was sent back to Germany, where he remained until Finland made peace with the Soviet Union in early September 1944. When he was about to depart for Finland, Himmler reportedly asked Talvela to become the head of a pro-German faction in Finland. Talvela refused out of hand.

After the war Talvela spent some years in South America as a representative of Finnish paper industry, until returning to Finland. He was promoted to General of Infantry in retirement in 1966.

Talvela was very able and aggressive commander in offense, but he was less well suited to defensive warfare. He was prone to vanity and temper tantrums and his stubbornness made Talvela a very difficult subordinate. He performed best when given independent commands. Talvela was awarded the Mannerheim Cross in 1941. (Wikipedia)

URI: http://tdf.fi/warsa/actors/person_50

Personal Details

Family name	Talvela
Given names	Paavo Juho
Born	19.01.1897
Municipality of birth	Helsingin mlk ^[4]
Rank	Jalkaväenkenraali
Military Unit	VI armeijakunta (Continuation War) Aunuksen ryhmä (Continuation War)

Death Details

Date of death	30.09.1973
Place of death	Helsinki ^[4]

Sources

[1] Suomi Sodassa. ISBN 9519078940, Valitut Palat 1983.
[2] Knights of the Mannerheim Cross. http://www.mannerheim-ristinitarit.fi/ritarit
[3] Suomalaisen Kirjallisuuden Seura, Biografiakeskus. www.kansallisbiografia.fi
[4] Wikipedia

Talvela, Paavo

Born on 1897 in Helsinki. Died on 1973 in Helsinki.

armeijakunnankomentaja

Kenraali Paavo Talvelaa on sanottu Suomen korkeaa-arvoisimmaksi reservinupseeriksi. Tälle leikkimieliselle sanonnalle on katetta sikäli, että Talvela erosi neljä kertaa armeijan vakinaisesta palveluksesta joko osallistuaakseen vapaaehtoisena heimosotiin tai toimiakseen liike-elämän palveluksessa. Talvelalla oli kuitenkin keskeinen tehtävä talvi- ja jatkosotien aikaisena sotatoimiyhtymän komentajana sekä ylipäällikön edustajana Saksan sotavoimien pääesikunnassa. Hän liittyi monin tavoin myös Suomen itsenäisyyden ajan historiaan: hän oli lapuanliikkeen organisaaattori, Alkoholiilikeen johtotehtävissä ja Petsamon liikenteen järjestäjä.

Source: [Semanttinen Kansallisbiografia](#) \ SKS:n Kansallisbiografia

<그림 4-8> 링크드 오픈 데이터로 구축된 WarShapo의 인물 정보(WarShapo, 2022)

4.2.4 UI/UX를 포함한 시각화

- (사용자 친화적 접근) UI/UX는 코로나19 아카이브의 사용자와 관리자 측면에서 현실적이고 실용적인 방안으로 검토하고, 단편적인 기능과 기술의 적용이 아닌 서비스의 전체 라이프사이클을 고려하여 접근해야 함
 - 코로나19 아카이브는 웹 아카이브와 개별 디지털 아카이브의 특징이 결합된 하이브리드 방식으로 사용자에게 따라 차별화된 UI를 제공해야 함
 - 예컨대, ARCH의 분석 서비스를 위한 UI 설계는 기획에서 테스트의 4단계 모델을 적용하고, 각 단계별로 실무자, 사용자의 검증을 수행하여 서비스에 구현됨
 - 웹 아카이브의 다양한 자료를 탐색, 시각화 기능은 관련 오픈소스를 충분히 검토하여 적용할 필요가 있음
- (ARCH) 웹 아카이브 분석을 위한 새로운 인터페이스로, 대규모 웹 아카이브 연구를 지원하기 위해 Archives Unleashed Cloud의 분석 도구와 Archive-It의 웹 아카

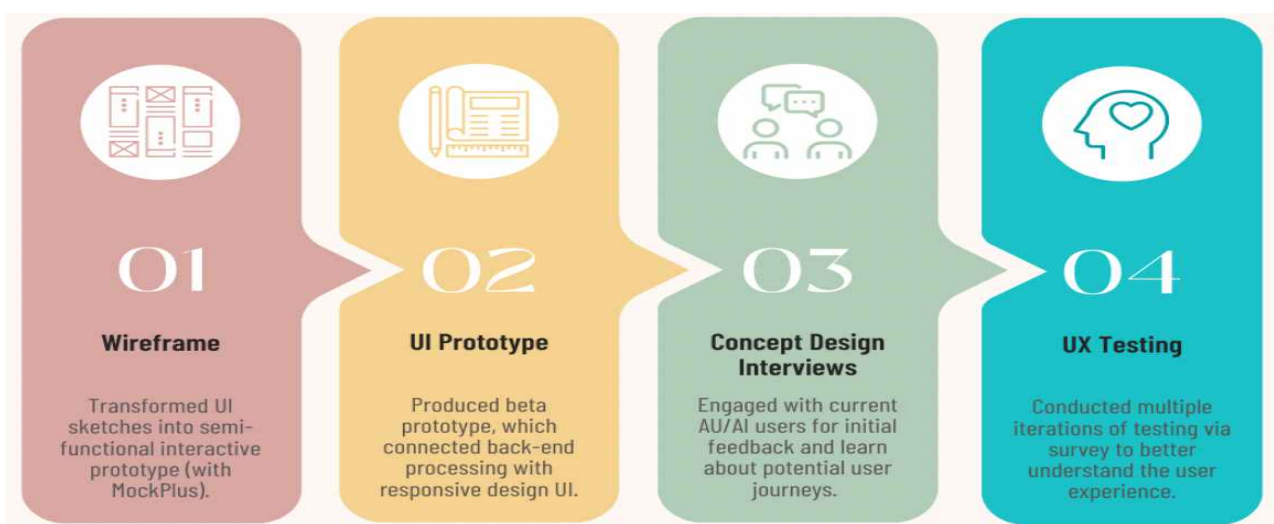
이브 플랫폼을 통합하는 작업을 진행하여 Archive-It 컬렉션에 대한 확장된 분석 기능을 제공함

- 도메인 빈도 통계, 하이퍼링크 네트워크 그래프 및 추출된 전체 텍스트를 포함하여 12개 이상의 데이터셋을 생성할 수 있는 기능을 제공하며, 자료를 탐색하기 위한 브라우저 내 시각화를 제공함

The screenshot displays the ARCH Web Archive Analysis interface. The left sidebar shows the 'Archive Research Compute Hub' with a 'Collections' section. The main content area is titled '#blacklivesmatter Web Archive Analysis' and includes a 'Collection overview' section showing a 6.6 TB public collection. Below this, the 'Jobs in Process' section shows two jobs: 'Extract text files (html, text, css, js, json, xml) information (Example)' and 'Extract video information (Example)', both in the 'Processing' stage with a 'Running' queue. A 'Start new job' button is visible at the bottom right.

<그림 4-9> ARCH의 분석 서비스(Archive-it 지원 센터, 2022)

- (ARCH 디자인 프로세스) (1) 스케치를 프로토타입으로 변환하는 와이어프레임, (2) 백엔드 처리 과정과 연결한 초기 UI 프로토타입, (3) 인터페이스에 대한 인터뷰, (4) UX 테스트 순서로 사용자 경험을 측정하며 서비스의 개선을 진행하였고, 각 단계는 상호 연결되어 있음



<그림 4-10> ARCH 디자인 프로세스 4단계(ARCH UX Testing, 2022)

- 특히 (4) UX 테스트를 통해, 사용자의 이용 행태와 사용하는 과정에서 발생한 문제, 충족되지 않은 이용자의 요구는 무엇인지 파악이 가능함
- 총 5단계의 UX 테스트를 진행함 ([표 4-6] 참고)
- 테스트 진행 방식의 한 예시로, 3단계의 외부 UX 테스트는 먼저 시험 대상자들의 배경 지식과 경험 정보를 얻기 위한 인터뷰를 진행했으며, UX를 사용한 후의 만족도를 정량적으로 측정하고, 따로 피드백을 얻는 과정을 진행했음 (<그림 4-11> 참고)

[표 4-6] ARCH 디자인 프로세스에서 진행한 5단계 UX 테스트(ARCH UX Testing, 2022)

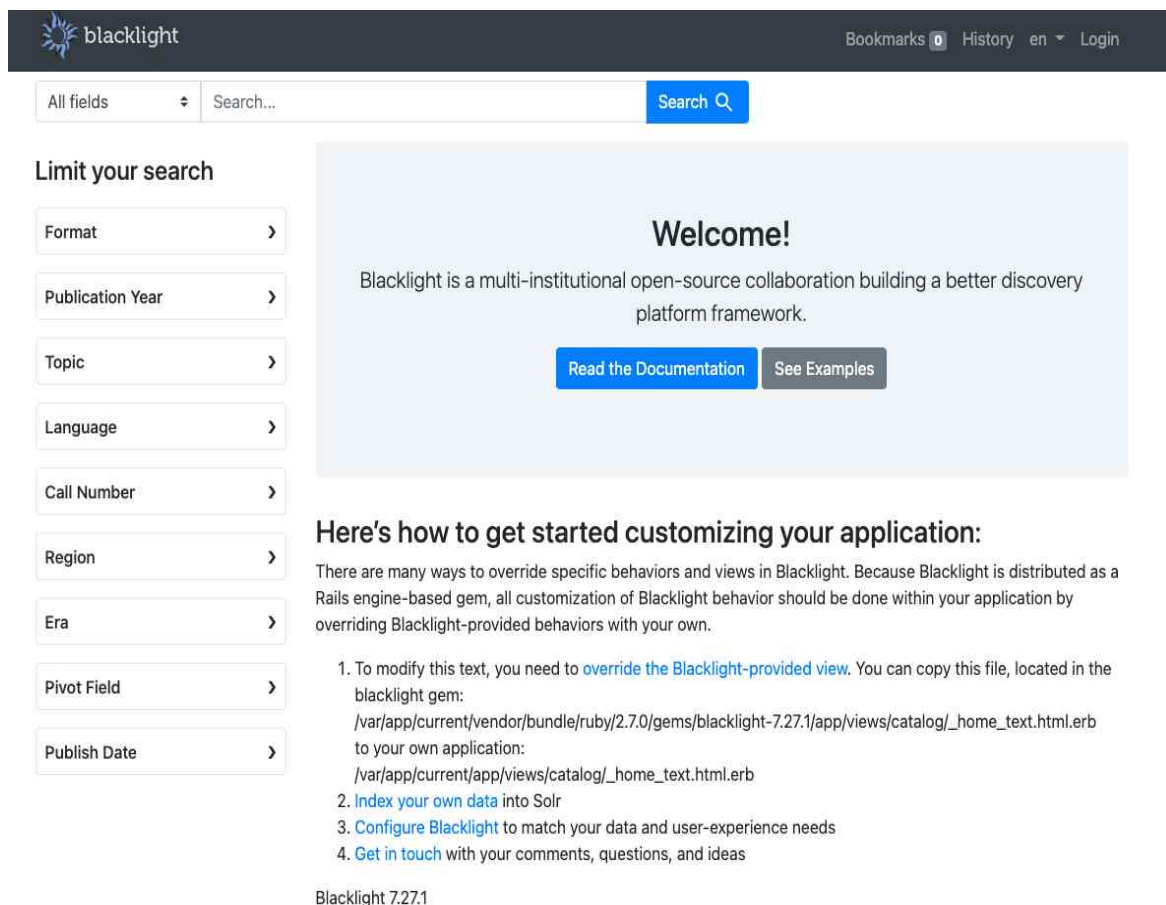
순서	단계	설명
1단계	컨셉 디자인 인터뷰 (화상 통화/화면 공유)	인터페이스를 시각적으로 구현한 와이프레이미를 검토하고, 앞으로의 워크플로우에 대한 논의를 진행함
2단계	내부 UX 테스트 (설문조사)	Archive-it 팀 구성원과 함께 프로토타입 초안을 토대로 내부 UX 테스트를 진행함
3단계	외부 UX 테스트 (설문조사)	자문 위원회 의원과 컨셉 디자인 인터뷰에 참여했던 사람들을 대상으로, 더 많은 인원에게 UX 테스트 후 피드백을 요청함
4단계	얼리어답터 UX 테스트 (설문조사)	업그레이드된 탐색, 접근성 개선 등 보완된 시스템 기능을 바탕으로, Archive-it의 사용자들에게 테스트 후 피드백을 얻음
5단계	코호트 UX 테스트 (화상 통화/화면 공유)	몇 달 동안 ARCH를 사용해본 코호트 팀의 경험을 토대로, 사용 경험에 대한 인터뷰를 진행함

Statement	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	Avg
I found navigating the system easy and straightforward	Disagree	Agree	Strongly Agree	Agree	Agree	Agree	Agree	Agree	Agree	Agree	Strongly Agree	4.0
The workflow of the system was clear and easy to follow	Neutral	Neutral	Strongly Agree	Agree	Neutral	Agree	Agree	Agree	Neutral	Agree	Strongly Agree	3.8
The terminology used to explain functions and features was clear	Agree	Agree	Neutral	Agree	Agree	Agree	Agree	Agree	N/A	Agree	Agree	3.5
The visualizations/graphs were helpful for understanding content of datasets	Agree	Agree	Agree	Neutral	Disagree	Agree	Neutral	Agree	Strongly Agree	Agree	Agree	3.6
I found the timing for processing datasets acceptable	Agree	Agree	Agree	Strongly Agree	Strongly Agree	Neutral	Strongly Agree	Disagree	Disagree	Agree	Strongly Agree	3.7
Generating datasets through the Cloud will improve my current research methods	Neutral	Neutral	Strongly Agree	Agree	Strongly Agree	Agree	Agree	Agree	Agree	Agree	Neutral	3.9

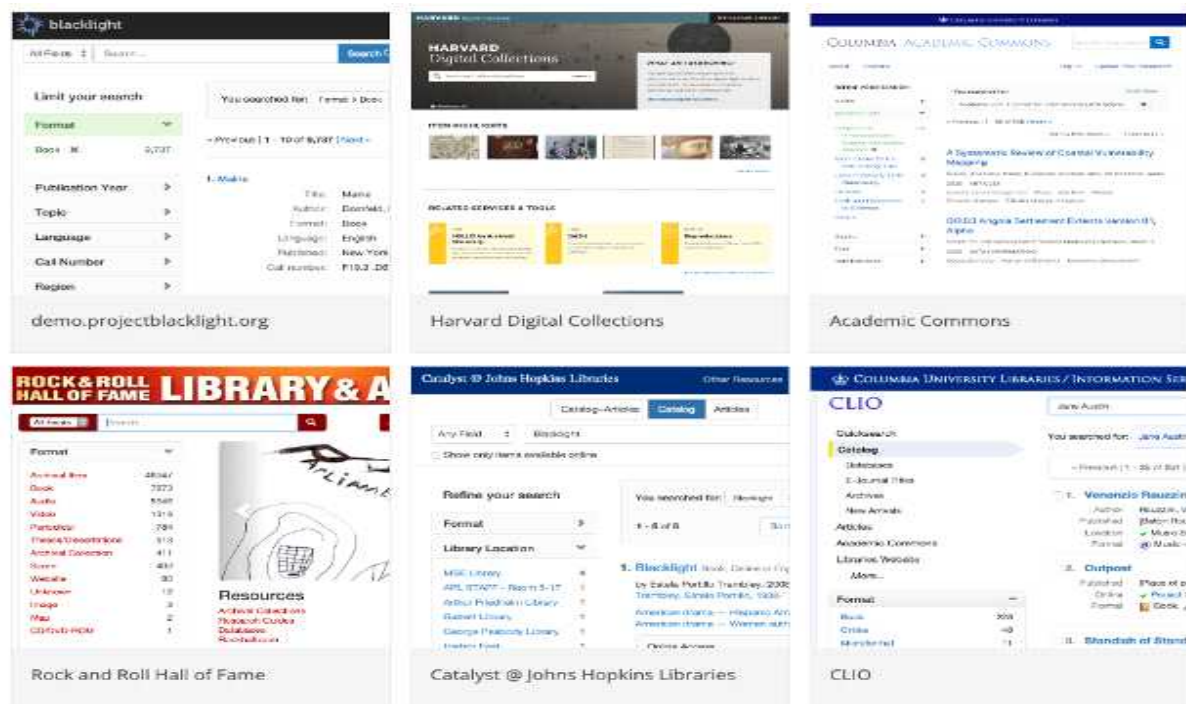
<그림 4-11> 3단계 외부 UX 테스트에서 실시한 만족도 평가의 정량적 결과(ARCH UX Testing, 2022)

- (Blacklight) 자바로 개발된 오픈소스 검색 엔진인 Apache Solr를 사용하여 텍스트 또는 메타데이터 검색 기능을 제공하는 인터페이스 검색 플랫폼임
- Apache Solr은 전문 검색, 검색 결과의 하이라이팅, 실시간 색인과 데이터베이스 통합 등의 기능을 가능하게 하며 전 세계 많은 대규모 인터넷 사이트들의 검색과 내비게이션 기능을 지원함
- Blacklight는 웹 프레임워크인 Ruby on Rails로 구성된 프론트엔드를 제공하고, 모든 소스 코드는 오픈 소스로 GitHub을 통해 공개하며 설치와 활용하는 방법에 대해 자세한 안내 사항을 제공함

- **(제공하는 기능)** 필드 검색, 패시 검색, 검색결과에 대한 정렬, 페이지네이션, 검색 쿼리의 저장과 공유, 레코드 검색을 위한 JSON API 기능, 적합성 기반 검색, 자료의 북마크, 모든 자료의 영구 URL, 자료의 이용자 태깅 등 여러 기능을 지원함
 - 서지 데이터와 관련 연구 자료를 관리하기 위한 서지관리 소프트웨어 Zotero 같은 플러그인이 검색 결과 페이지에서 데이터를 쉽게 추출할 수 있도록 구성되었으며, OpenSearch와 Schema.org를 사용하여 검색 결과를 쉽게 공유할 수 있음
 - **(적용 사례)** 하버드 대학교 도서관의 디지털 컬렉션 검색 사이트처럼 전 세계적으로 다양한 대학교, 도서관이나 박물관 등에서 Blacklight를 사용하여 웹사이트를 구축하여 서비스하고 있음
 - 핵심적인 기능을 담은 기본 데모는 제공되되, CSS처럼 이용자들이 직접 소스 코드를 수정하거나 추가하여 기능을 확장할 수 있기 때문에 다양하게 활용할 수 있음
- * Blacklight를 사용하여 웹사이트를 구축한 사례목록은 불임4를 참고



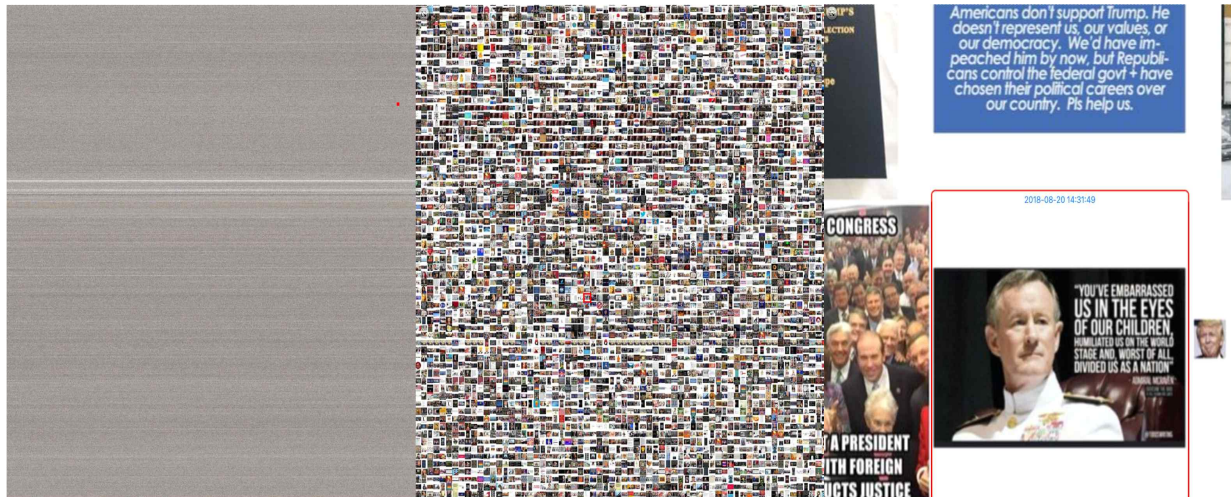
<그림 4-12> Blacklight의 핵심 기능만 구현한 데모(Blacklight, 2022)



<그림 4-13> Blacklight의 적용 사례(Blacklight, 2022)

- (OpenSeadragon) 자바스크립트 언어로 구성된 고해상도 이미지를 확대·축소할 수 있는 오픈 소스의 웹 기반 뷰어로, 대규모 이미지의 콜로주(collage) 생성이 가능함
- OpenSeadragon을 활용할 수 있는 방법으로 GitHub에 공개된 소스 코드를 사용하는 방법과 공식 홈페이지에서 ZIP, TAR 폴더를 다운로드받는 방법, npm이나 bower, CDN 형식으로 소스 코드에서 사용하는 방법이 있으며, 데스크탑과 모바일 환경에 적용 가능함
- (제공하는 기능) 이미지 데이터를 분석하여 웹 페이지에 맞게 이미지의 확대와 축소가 가능하며, 기준이 되는 좌표계로 이미지의 실제 픽셀, 웹 페이지의 픽셀과 뷰포트 3개 중에서 선택하여 활용할 수 있음
- OpenSeadragon을 적용하기 위해 특정 이미지의 확대·축소된 이미지들이 필요함. 사용하려는 이미지가 큰 경우 사전에 변환작업을 진행해야 하며 필요에 따라 이미지를 변환할 수 있는 여러 옵션을 선택하여 사용할 수 있음
- * 사용가능한 변환 옵션: DZI, IIF, TMS, Zoomify
- 다양한 이미지 서비스 프로토콜을 지원하며 필요한 프로토콜을 지원하지 않는 경우, 사용자 지정 타일 소스를 사용하여 추가할 수 있고 각 프로토콜 예시 소스 코드를 공식 홈페이지에서 제공함
- 이미지 외에도 OSM(Open Street Maps) 방식으로 지도 시각화의 활용이 가능함
- (구현 사례) <그림 4-14>은 OpenSeadragon을 활용하여 이미지 콜로주를 생성한 모습임
- 약 9,300만개의 0.25 Mpixel 이미지 타일을 활용하여 약 4 TeraPixel의 이미지 콜로주를 생성하였으며, 차례대로 전체 콜로주 사진, 이미지 확대를 6단계로 실행한 사진, 이미지 확대를 끝까지 다 실행한 사진임

- (적용 사례) 전 세계적으로 대학교, 박물관의 웹사이트에서 사진과 역사 자료를 시각화할 때 많이 사용되고 있으며, 산업분야나 미디어, 지도 시각화가 필요한 곳에서 사용됨
- (기타 이미지 시각화 방법) Apache Spark를 사용하여 웹 아카이브를 분석하는 오픈 소스 플랫폼인 Archives Unleashed Toolkit과 딥러닝을 쉽게 구동할 수 있도록 돕는 라이브러리인 Tensorflow를 이용한 이미지 데이터 분석과 시각화가 가능함



<그림 4-14> OpenSeadragon을 사용하여 이미지 콜로주를 생성한 사례(OpenSeadragon, 2022)

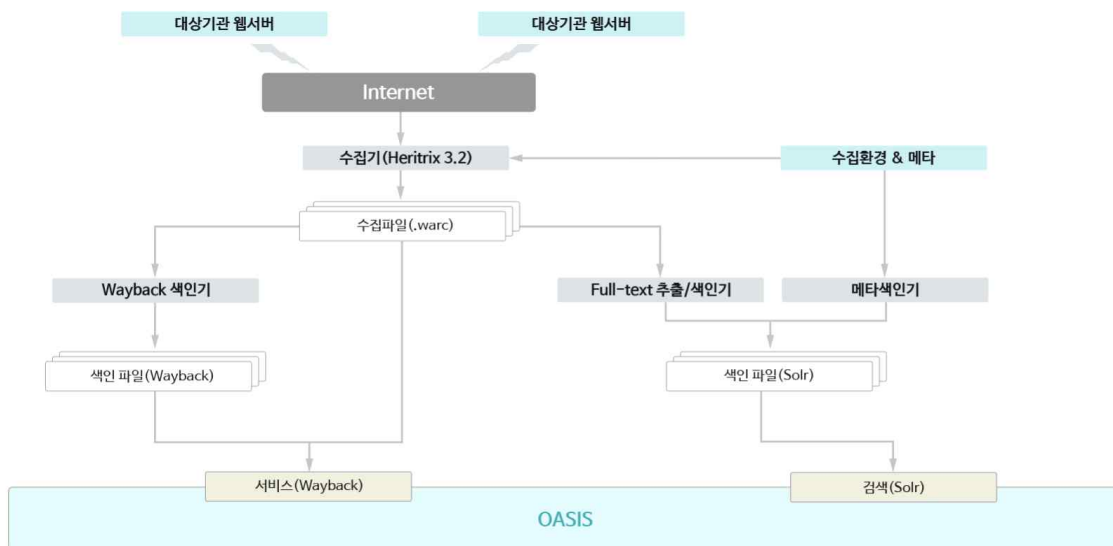
4.3 아카이브의 활용 활성화

4.3.1 민간 아카이브의 연계

- 민간 아카이브는 목적에 따라 설립과 관리 주체가 상이하고, 기록물의 주제와 유형, 대상, 수집 방법, 기록관리의 수집과 관리 정책이 중요함
 - 민간 아카이브의 종류는 민간 형태의 모기관에서 생산하거나 접수하여 장기보존의 가치를 인정받은 보존기록을 관리하는 형태의 단독 이관형 민간 아카이브, 지역이나 경험 공동체를 포함한 고통의 가치나 경험을 공유하는 민간 집단의 기록을 수집·관리하는 공동체 기록 수집형 민간 아카이브가 있음
 - 데이터의 관리 차원에서 공공 아카이브에 비해 정기적 이관이 드물고, 기록에 대한 수집을 능동적으로 진행함. 개별 아카이브마다 수집의 기준이 모호하고 유통과정에서 손실되거나 변경되기 때문에 유지보수가 어려움
- (시스템 연계) 오아시스는 민간에 생성되는 웹사이트 자원을 수집하거나 기증받아 관리하고 있음. 웹사이트 수집은 정보제공동의서를 요청 후 자동 수집 로봇을 활용해 정부기관, 학술단체, 연구기관, 기타 이슈사이트의 자료를 수집함
 - <그림 4-15>은 웹사이트 수집 로봇(Heritrix 3.4)은 직접 웹 문서, HTML 소스, 웹 이미지

등을 수집한 데이터를 저장하고 관리하는 과정을 나타냄

- [표 4-7]은 외부에서 생성되는 웹사이트를 오아시스 시스템의 웹사이트 수집 로봇이 정확하게 수집하고 보존하기 위해 권장하는 웹사이트 구축 방법임
- 오아시스는 웹 자원을 수집하는 절차와 시스템 권장사항을 제공하고 있지만, 아카이브 차원에서 시스템적으로 연계할 수 있는 방안은 미흡함. 민간 아카이브는 목적에 따라 형태가 다양해서 웹사이트 수집 로봇만으로 수집하는 방안은 한계가 존재함

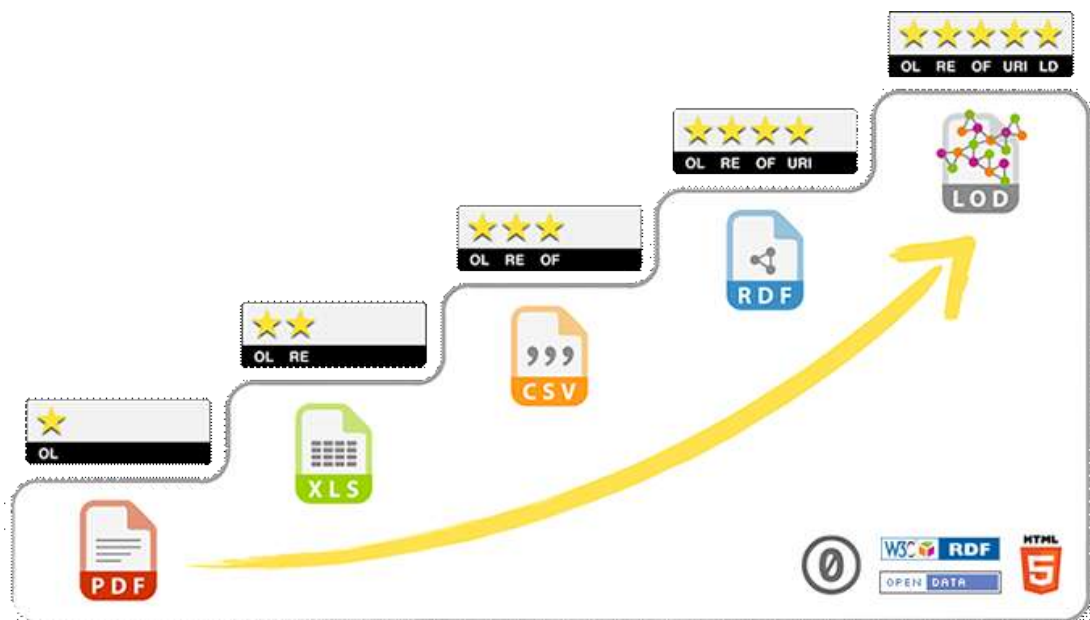


<그림 4-15> 오아시스 시스템의 웹사이트 수집절차(국립중앙도서관, 2022)

[표 4-7] 오아시스 시스템의 웹사이트 수집을 위한 권장사항(국립중앙도서관, 2022)

구분	설명
웹사이트 구축	웹 서버 혹은 방화벽에서 오아시스 수집로봇의 접근을 허용함
	robots.txt 파일은 반드시 root에 위치해야 함
	robots.txt 파일에 수집로봇 접근을 허용함
웹 페이지 작성	URL정보를 변경할 때에는 HTML 헤더에 메타 태그를 사용함
	모든 페이지의 제목을 동일하게 표현하지 않음
	css파일이 위치한 상위 경로까지 seedURL을 정확히 작성함
	마우스 오버 동작 이미지는 두 개 이미지를 전부 작성함
	페이지에 플래시를 사용하지 않음
	https로 강제 접속을 사용하지 않음
	화면에서 iFrame을 사용하지 않음
	activeX를 사용하지 않음
	콘텐츠 문자셋을 한 가지로 일치함

- **(API 서비스 제공)** 데이터를 제공하는 API(Application Programming Interface)는 데이터베이스에 저장된 데이터를 요청하는 클라이언트에게 전달하는 것으로, 오픈 API는 누구나 사용할 수 있도록 공개하는 형식임
 - 민간 아카이브의 자원은 웹사이트, 이미지, 동영상, 문서, 음성 등 여러 형태로 제공하기 때문에 하나의 시스템으로 연계하기 어려움. 시스템 연계가 어려운 경우, 데이터 단위에서 연계하기 위해 API 형태로 서로 다른 아카이브의 데이터를 공유할 수 있는 체계를 마련해야함
 - 오아시스의 코로나19 아카이브의 웹사이트 자료는 14,448건 중 15건만 모두 이용 가능함. 반면 국립중앙도서관의 장서 자료는 Open API³⁵⁾ 형태로 이용자가 데이터를 직접 활용할 수 있도록 제공함
 - 즉, 오아시스의 웹 자원을 민간 아카이브와 연계하기 위해 API 형태로 공유하고 연계하는 방안이 필요함
- **(오픈 데이터)** 코로나19 아카이브의 웹 자원은 저작권의 대부분의 자원 접근이 제한되어 이용자 열람에 한계가 있음. 데이터 저작권의 활용범위를 확장하고, 오픈 데이터 원칙을 도입하여 활용성을 높여야 함
 - <그림 4-16>는 팀 버너스-리가 제안한 오픈 데이터의 다섯 단계를 도식화 한 것으로, 데이터의 라이선스에 제한을 두지 않는 오픈 라이선스를 제안해 이용자가 활용하기 쉬운 데이터 제공 환경의 중요성을 표현함. 특히, 높은 단계의 오픈 데이터는 웹상의 구조화된 방법으로 제공하여 기계가 읽을 수 있는 형식으로 이용자의 접근성을 높이고 상호이용성을 보장함



<그림 4-16> 팀 버너스-리의 Five-Star 오픈 데이터 단계(5-star Open Data, 2012)

35) <https://www.nl.go.kr/NL/contents/N31101010000.do>

- 코로나19 아카이브를 비롯한 오아시스의 웹 자원 이용범위는 [표 4-8]와 같이 ‘관외 이용’, ‘국립중앙도서관’, ‘비공개’로 제한함. 이용자는 이용범위의 제한으로 대부분 자료에 접근할 수 없고, 데이터 활용과 분석이 어려움
- [표 4-9]의 공공누리(Korea Open Government License, KOGL) 제도는 공공기관에서 생성된 공공저작물의 저작권을 표기하여 표준화된 약관으로 개방을 통한 활용성을 높임
- 오아시스의 웹 자원도 국립중앙도서관에서 수집하고 관리하는 공공저작물의 기준으로 판단할 수 있으므로, 원본 저작권자의 이익을 부당하게 해치지 않는 조건에서 이용범위 확대를 고려해야함

[표 4-8] 오아시스의 자료 이용 범위 안내(국립중앙도서관, 2022)

구분	조건
관외 이용	웹사이트 저작권자의 ‘이용 동의’가 있을 경우, 국립중앙도서관 및 인터넷에서 이용 가능
국립중앙도서관	웹사이트 저작권자의 ‘이용 동의’가 없을 경우, 국립중앙도서관 내에서만 이용 가능
비공개	웹사이트 저작권자가 비공개 요청을 하거나 국립중앙도서관 정책상 공개가 어려운 경우, 이용할 수 없음

[표 4-9] 공공누리 유형안내(공공누리, 2022)

구분	이미지	조건
제1유형		<ul style="list-style-type: none"> • 출처표시 • 상업적, 비상업적 이용 가능 • 변형 등 2차적 저작물 작성 가능
제2유형		<ul style="list-style-type: none"> • 출처표시 • 비상업적 이용만 가능 • 변형 등 2차적 저작물 작성 가능
제3유형		<ul style="list-style-type: none"> • 출처표시 • 상업적, 비상업적 이용가능 • 변형 등 2차적 저작물 작성 금지
제4유형		<ul style="list-style-type: none"> • 출처표시 • 비상업적 이용만 가능 • 변형 등 2차적 저작물 작성 금지

- **(데이터 공유 플랫폼)** 민간 아카이브와 서로 다른 데이터를 연계하기 위한 데이터 표준을 마련하고, 공유할 수 있는 플랫폼 환경을 제공해 공공과 민간 사이의 정보의 재사용성을 높여야 함
 - 수집한 자료를 통합적으로 제공할 수 있는 데이터 공유 플랫폼을 구축하여 이용자의 접근성을 개선하고 데이터 생성과 관리 현황을 파악할 수 있는 채널을 확보해야 함
 - 디지털 아카이브의 데이터는 보존과 동시에 재사용 가능한 형태로 관리해야 함. ‘코로나19: 우리의 기억’은 깃허브를 통해 디지털 아카이브를 구축한 사례로, 원본 데이터와 데이터 분석 소스 코드를 모두 오픈 소스로 제공한 사례임. ‘공공데이터 포털’은 행정안전부에서 관리하는 공공데이터를 제공하는 데이터 포털로, 분산적으로 존재하는 공공데이터를 통합적으로 제공함
 - 민간 아카이브와 데이터를 공유하기 위해선 깃허브와 같이 외부에서 데이터 제공에 기여할 수 있는 구조 또는 공공데이터 포털처럼 분산된 자원을 통합해서 제공할 수 있는 데이터 창구가 필요함
 - 출처가 서로 다른 데이터를 관리하기 위해선 메타데이터 수준의 연계가 필요함. 코로나19 아카이브의 주제분류는 재난안전법의 분류체계와 KDC(한국십진분류법)을 사용하고 있지만 두 분류체계 간의 관계는 정의하지 않음. 코로나19 데이터의 분류체계와 용어사전 같은 주요 표준을 마련하여 데이터의 체계적 관리와 접근 용이성을 확보할 수 있음

4.3.2 국민 참여형 서비스

- **(사례 선정)** 국민 참여형 서비스의 도입을 위해 이용자의 참여가 두드러진 디지털 아카이브 사례를 분석함
 - 대상은 (1) 코로나19 또는 재난과 관련된 자료를 수집하고, (2) 이용자의 참여가 활발한 디지털 아카이브를 선정함
 - 국내 2개와 국외 5개의 아카이브가 선정되었고, [표 4-10]은 선정된 아카이브의 목록임
- * 선정된 디지털 아카이브에 대한 자세한 설명은 2장 국내외 코로나19 아카이브 사례 분석을 참고

[표 4-10] 이용자의 참여 중심의 디지털 아카이브 목록

이름	구분	설명
4.16 기억저장소	국내	2014년 4월 16일의 세월호 참사를 기억하고 기록하기 위한 아카이브로, 시민이 세월호와 관련된 기록을 업데이트하는 참여형 아카이브임
코로나19: 우리의 기억	국내	한국의 코로나19와 관련된 다양한 사회이슈를 기록한 디지털 아카이브임. 디지털 자원의 수집과 보존을 넘어 디지털 자원을 공유하고 재사용할 수 있는 환경을 구축함
The September 11 Digital Archive (이하 911DA)	국외	2001년 9월 11일에 뉴욕, 버지니아, 펜실베이니아에서 발생한 테러의 기억과 기록을 수집·보존하기 위한 참여형 아카이브임
HINAGIKU (이하 HINA)	국외	일본의 국립국회도서관에서 구축한 디지털 아카이브로, 1995년에 발생한 동일본 대지진의 디지털 데이터를 검색·활용할 수 있음
Documenting COVID-19 (이하 DC19)	국외	세계 각국의 중앙정부, 지방자치단체, 대학 도서관과 커뮤니티의 코로나19 관련 자료를 기록하고, 기부와 참여 방식으로 다양한 유형의 자료를 수집하고 있음
Coronavirus Web Archive (이하 CWA)	국외	미국의회도서관에서 코로나19가 진행되는 동안 다양한 분야와 언어의 웹 문서를 수집하고 일반에 공개하고 있음
A Journal of the Plague Year: Share your story (이하 SYS)	국외	코로나19 상황의 개인기록을 수집하고 보존하기 위해 전 세계의 교수와 대학원생이 참여하는 시민 참여형 아카이브임

- **(분석 프레임워크)** 이용자 참여와 관련된 기능을 추출하여 오아시스와 비교·분석하고, 이용자의 참여 허용도에 따라 아카이브의 유형을 구분함
 - 이용자 참여와 관련된 기능은 박진경과 김유승(2017) 연구의 비교 기준을 확대하였음. 이용자가 활발히 디지털 자원을 이용하거나 분석할 수 있는 기능(예: 콘텐츠와 메타데이터 다운로드, 데이터 분석)을 추가해 총 10개의 기능이 선정됨
 - [표 4-12]는 오아시스와 이용자 참여 중심의 디지털 아카이브의 기능을 비교한 결과임
 - 아카이브 유형은 아카이브를 관리하는 기관의 통제력과 이용자의 참여 허용도를 기준으로 ‘적극적 참여’, ‘허브방식’과 ‘소극적 참여’으로 구분됨. 각 유형의 특징과 설명은 [표 4-11]에서 확인할 수 있음

[표 4-11] 이용자의 참여 방식에 따른 아카이브 유형(박진경 & 김유승, 2017)

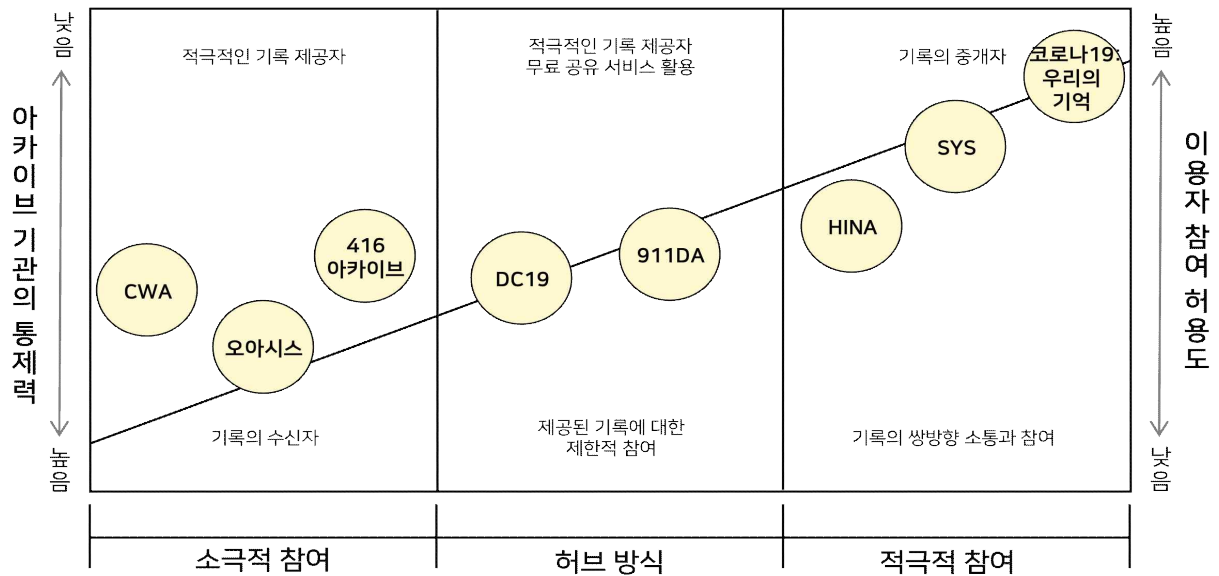
유형	기관의 통제력	이용자 참여 허용도	설명
적극적 참여	서비스의 중개자 · 조력자 역할	기록의 생산자 · 소비자 역할	이용자의 참여가 극대화된 유형. 이용자가 직접 기록의 생산과 관리 과정에 참여할 수 있음
허브 방식	콘텐츠의 주도권 보유	제한적 참여	무료 공유서비스(예: 온라인 플랫폼, 온라인 허브)를 활용해 기록을 제공하는 유형. 주요 콘텐츠 생산자는 전문가 또는 기관이며, 이용자는 기록의 생산과 관리에 제한적으로 접근 가능함
소극적 참여	강력한 통제력 행사	정보의 수신자 역할	아카이브 관리기관 또는 전문가가 기록의 생산과 관리 과정에 강한 통제력을 행사하는 유형. 이용자는 제공되는 기록의 활용만 가능함

- (아카이브별 기능 비교) 이용자 중심의 디지털 아카이브는 디지털 자원을 직접 보고, 활용이 가능한 형태로 다운로드하는 기능과 데이터 분석을 대표적으로 제공하고 있음
- 저작권 문제로 대부분의 웹 자원을 온라인으로 이용할 수 없는 오아시스와 달리, 모든 아카이브는 디지털 콘텐츠를 온라인으로 볼 수 있는 기능을 제공함. 미국의회도서관의 Coronavirus Web Archive는 오아시스와 같이 웹 자원을 수집하지만, 시간별로 웹 자원의 캡처를 제공하여 이용자가 자원을 파악할 수 있도록 도움
- 디지털 자원을 공유하거나 개방하는 기능은 이용자가 아카이브의 자원을 적극적으로 활용할 수 있게 함. 다수의 아카이브에서 콘텐츠를 다운로드받거나 디지털 자원의 데이터 분석 서비스를 제공하고 있음

[표 4-12] 이용자 참여와 관련된 기능의 비교 결과

기능	오아시스	416 아카이브	코로나19: 우리의기억	911DA	HINA	DC19	CWA	SYS
콘텐츠 온라인 보기		O	O	O	O	O	O	O
콘텐츠 다운로드			O	O	O	O		O
데이터 분석	O		O	O		O		O
기증	O	O			O			O
메타데이터 다운로드				O	O		O	
콘텐츠 수정·가공 반영		O	O					O
즐거찾기/스크랩			O		O			
콘텐츠 업로드			O					O
댓글			O					
API 서비스					O			

- **(아카이브 유형)** ‘적극적 참여’ 유형의 아카이브는 이용자가 적극적으로 기록의 생산부터 관리 과정에 참여할 수 있는 반면, ‘소극적 참여’ 유형의 아카이브는 관련 기관 또는 전문가의 기록관리에 대한 통제력이 두드러짐
- **(적극적 참여)** ‘코로나19: 우리의 기억’, ‘A Journal of the Plague Year: Share your story’, ‘HINAGIKU’가 포함됨. 이용자가 자유롭게 관련 기록을 업로드하고, 수정·편집할 수 있고, 직접 기록을 공유하고 재사용할 수 있는 방안을 마련함
- **(허브 방식)** ‘The September 11 Digital Archive’와 ‘Documenting 코로나19’가 해당됨. 아카이브 플랫폼을 중심으로 콘텐츠를 제공하고 있지만, 콘텐츠를 게시하고 관리하는 권한은 기관 또는 전문가가 주도하고 있음. 이용자는 콘텐츠를 보거나 다운로드하는 일부 기능만 이용할 수 있음
- **(소극적 참여)** ‘4.16 기억저장소’, ‘오아시스’, ‘Coronavirus Web Archive’가 포함됨. 이 아카이브의 공통점은 기증을 통해서 이용자의 콘텐츠가 아카이브에 포함될 수 있고, 기록의 수집과 관리는 오직 기관 또는 전문가에 의해 수행됨. 콘텐츠의 다운로드와 같이 기록을 활용할 수 있는 기능이 극히 제한되어 있음



〈그림 4-17〉 이용자 참여 관점의 아카이브 현황

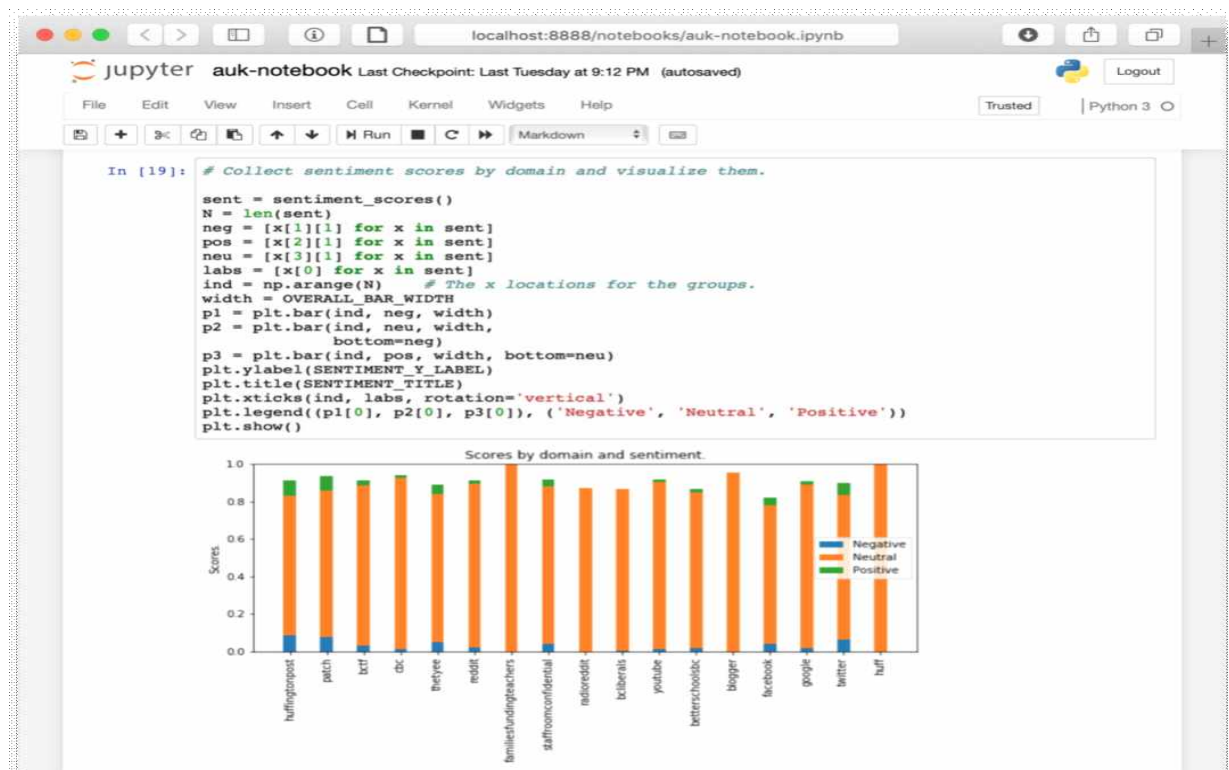
- (오아시스 적용 방안) ‘적극적 참여’ 유형의 아카이브와 같이 적극적으로 기록을 공유하고 재사용할 수 있는 환경을 제공하고, 이용자가 적극적인 기록의 생산자 또는 기여자가 될 수 있는 기반을 마련하는 것이 필요함
 - 오아시스는 수집된 웹 자원의 메타데이터만 확인할 수 있고, 웹 자원의 직접적인 온라인 이용은 제한되어 있음
 - 미국의회도서관의 ‘Coronavirus Web Archive’는 오아시스와 같은 웹 아카이브이지만, 시간별 웹 자원의 캡처를 통해 수집된 웹 자원을 확인할 수 있음. 오아시스의 적극적인 활용을 위해 수집된 웹 자원부터 온라인으로 확인할 수 있는 환경부터 구축되어야 함
 - 오아시스는 수집된 웹 자원의 목록과 메타데이터만 공개하고 있음. 극히 일부의 웹 자원만 볼 수 있고, 콘텐츠와 메타데이터의 다운로드도 제공하지 않음
 - 반면, 이용자의 참여 정도가 높은 아카이브(예: ‘코로나19: 우리의 기억’, ‘HINAGIKU,’ ‘A Journal of the Plague Year: Share your story’)는 기록과 메타데이터의 벌크 다운로드 또는 API를 제공하거나 직접 아카이브에 기록을 업로드할 수 있는 기능을 제공함
 - 오아시스가 국민 참여형 서비스를 도입하기 위해서 오아시스가 보유한 웹 자원과 메타데이터를 이용할 수 있는 환경을 만들어주는 것이 선행되어야 함

4.3.3 코로나19 아카이브 활용과 홍보 방안

- (데이터 프리존) 웹 아카이브는 수집한 자료의 라이선스와 이용조건으로 특정한 공간과 시설에서 자료를 활용하는 제약이 있음
 - 데이터 프리존은 물리적 보안환경이 갖추어진 별도의 특정장소에서 제공받은 자료를 활

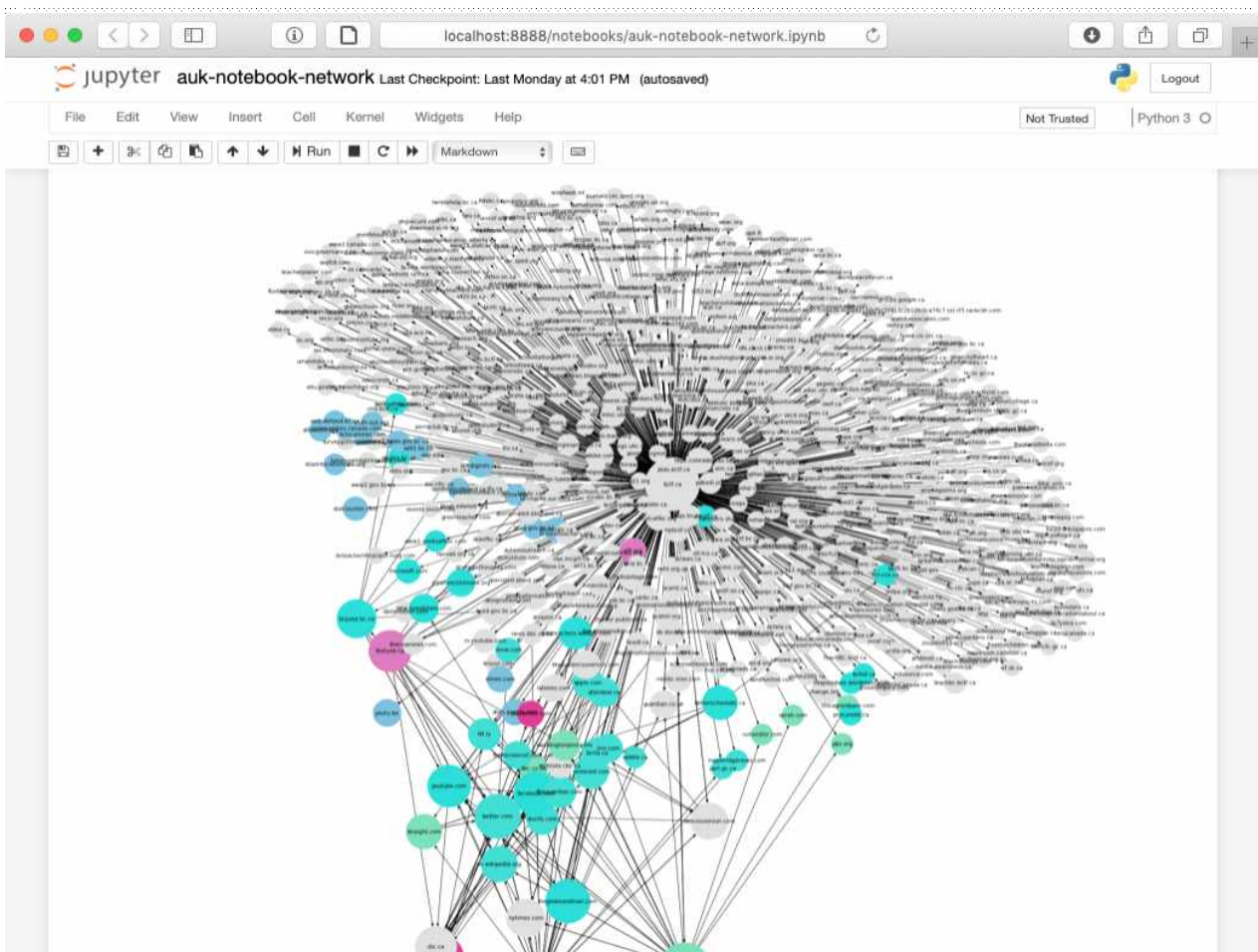
용, 분석 작업을 수행한 후 제공기관이 승인 하에 제공하는 방식임

- 국립중앙도서관이라는 장소적 한계를 넘어, 전국의 주요 도서관에 데이터 프리존을 설치하여 사용자의 접근성을 높일 수 있음
 - 웹 아카이브 자료의 이용 조건과 활용 등급을 명시적으로 규정하고, 사용자의 데이터 활용, 분석, 분석 결과의 외부 반출을 위한 가이드라인을 제공하여 참여를 확대시키는 방안 검토가 필요함
- (데이터 분석) 웹 아카이브에서 기존의 웹 사이트를 재현하고, 디지털 자료의 상호연계를 통해 데이터를 구체적으로 활용할 수 있는 방안을 제공해야 함
- 대부분의 웹 아카이브는 웨이백 머신 등 디지털 자료의 재현에 초점이 있고, 사용자는 과거의 자료를 탐색할 수 있음
 - 그러나, 웹 아카이브에 보존된 자료를 분석하고 연계하기 위한 방법이 제공되지 않기 때문에, 자료의 활용이 제한적임
 - 사용자의 데이터 분석과 활용을 위해 웹 아카이브에서 필요한 데이터를 자동으로 추출하고 분석 환경에서 활용할 수 있는 환경 제공이 필수적임.
 - 예를 들어, ArchiveUnleashed 프로젝트는 웹 아카이브의 디지털 자료를 분석하기 위한 도구 (toolkit)를 제공하고, 파이썬 환경에서 디지털 자료를 분석할 수 있는 오픈소스를 제공하고 있음



<그림 4-18> Archives Unleashed Toolkit을 이용한 웹 아카이브의 분석(ArchivesUnleashed, 2022)

- (참여 중심의 홍보) 웹 아카이브에 대한 연구 차원의 관심과 주제 분야 전문가의 적극적 참여를 위한 홍보 방안을 검토해야 함
 - 웹 아카이브는 문헌정보, 기록관리, 웹 표준, 웹 크롤링, 데이터베이스 등 다학문적 융합과 협력을 통해 전문성을 확보해야 함
 - 디지털 자료의 수집 정책에서 탈피해서 효과적인 활용을 위해 국립중앙도서관 중심으로 연구협력체계를 구축하고, 기본계획을 수립하여 장기적으로 대응할 필요가 있음
 - 웹 아카이브에서 수집하는 대상과 범위와 관련한 연구자, 기관, 커뮤니티에 타겟형 홍보가 필요하고, 단편적인 실적의 홍보가 아닌 참여할 수 있는 방안을 구체적으로 제안해야 함
 - 데이터 해커톤(hackathon), 밋업(meetup) 등 전문가 커뮤니티가 자유롭게 참여할 수 있는 이벤트를 통해 웹 아카이브의 현황을 공유하고, 실질적으로 수집된 자료를 활용하고 개선할 수 있는 기회를 제공해야 함



<그림 4-19> Archives Unleashed Toolkit을 이용한 네트워크 분석(ArchivesUnleashed, 2022)

제 5 장 결론

5.1 연구의 요약

5.1.1 국내외 코로나19 아카이브 사례 분석

- **(국내)** 국내에서 코로나19 관련 데이터나 도큐먼트를 수집하거나 데이터를 분석하여 시각화 자료를 제공하는 아카이브 사례를 조사함
 - 총 13개의 아카이브 사례를 조사하고, 제공하는 데이터와 시각화 자료, 특징에 대해 분석하여 정리함. 질병관리청, 통계청 같은 국가기관과 공공기관뿐만 아니라 코로나나우 데이터센터와 DS4C처럼 민간수준에서도 공익을 위하여 여러 아카이브를 구축하고 있음
 - 연구, 특허, 경제 등 각 단체의 주제나 특성과 연관된 코로나19의 자료를 수집하고 시사점을 제시하고 있으며, 따라서 여러 분야에서 코로나19에 관심을 기울고 있다는 사실을 알 수 있음
 - 디지털 지적 문화유산을 국가적 차원에서 수집·보존하는 프로젝트인 국립중앙도서관의 오아시스는 재난아카이브의 한 테마로 ‘코로나바이러스감염증-19(코로나19)’ 아카이브를 구축하고 있음
- **(국외)** 국외에서 코로나19 관련 데이터나 도큐먼트를 수집하거나 데이터를 분석하여 시각화 자료를 제공하는 아카이브 사례를 조사함
 - 총 13개의 아카이브 사례를 조사하여 수집 데이터와 이용자에게 제공하는 형식, 시각화 형태 등 특징에 대해 분석하여 정리함
 - 대부분 미국에서 구축한 아카이브 사례들이며, 정부, 대학, 사기업 등 다양한 단체에서 아카이브를 구축하고 있음
 - 13개의 아카이브 사례 중, 6개의 아카이브 사례는 각 아카이브를 구축하는 나라에 해당되는 데이터뿐만 아니라 전 세계의 데이터를 분석하여 지도시각화 형태로 제공하고 있음
 - * 전 세계의 데이터를 분석하여 시각화 자료를 제공하는 6개의 아카이브 사례: Our World in Data, Oxford 코로나19 Government Response Tracker, New York Times, Bloomberg, The sex, gender and 코로나19 Project, WHO
 - Journal of the Plague Year: Share your story 아카이브는 자료를 일방향적으로 제공하는 것이 아니라, 시민 참여형으로 직접 사람들이 자신의 실제 생활 모습을 담아내기 때문에 다양한 범위와 자료를 수집할 수 있다는 장점이 있음
- **(코로나19 웹 아카이브 프로젝트)** 현재 구축된 코로나19 아카이브의 수집 정책과 현황, 검색가능성과 접근성, 파트너십을 분석한 프로젝트 내용을 바탕으로 개선된 아카이브를 구축하기 위해서 보완되어야 할 부분을 정리함

- **(코로나19 아카이브 수집 정책 분석)** 해외 코로나19 아카이브를 구축하는 기관 중, 수집 정책 관련 정보를 제공하는 기관 11개를 대상으로 정책 요소를 분석함
- 아카이브 구축 과정에서 정책은 법적 근거와 기록의 일관성을 보장하는 수단임
- 각 기관의 정책을 11개의 세분화된 요소로 분석한 결과, 가장 많이 적용된 요소는 수집 매체에 대한 정의와 저작권 관련 사항으로 총 10개의 기관에서 적용하고 있음. 수집 방법, 아카이브 목적, 수집 참여자, 수집 절차, 기록물 범주, 개인정보보호 요소도 많이 포함되어 있음. 반면 수집 제외 항목과 수집 우선순위 항목은 각각 3곳, 2곳으로 가장 적게 포함되어 있음
- **(주요 국가별 코로나19 아카이브 분석)** 웹 아카이빙 기관과 연구자의 네트워크 기관인 WARCnet은 코로나19 아카이브를 구축한 8개국 9기관을 대상으로 컬렉션 현황, 접근성과 검색가능성, 파트너십을 중심으로 아카이브 담당자를 인터뷰한 내용을 표로 정리함
- 각 기관은 도큐먼트 형식의 자료를 수집할 뿐만 아니라, 소셜미디어 데이터를 수집하고자 노력하고 있지만, 법률적, 기술적 기반이 아직 완벽히 갖춰지지 않았음

5.1.2 ‘코로나19: 우리의 기억’ 사례 분석

- **(추진배경)** ‘코로나19: 우리의 기억’은 코로나19 범유행 상황에서 일어난 한국 사회의 변화를 가치중립적으로 기록한 디지털 아카이브 프로젝트로, 2019년에 첫 확진자 발생 이후부터 2020년 2월까지 코로나19와 관련된 국내외 데이터를 포함함
 - ‘코로나19: 우리의 기억’은 민간이 구축한 디지털 아카이브를 국가도서관에서 보존하는 첫 사례임
- **(특징과 공유 체계)** 기록의 검증과 재현을 위해 세 가지 원칙에 따라 아카이브를 구축하고, 데이터의 공유와 재사용이 활발한 체계를 구축함
 - 첫째, 가치중립적으로 디지털 기록을 수집·생산·보존하고, 둘째, 지속 가능한 협업 모델을 지향하며, 셋째, 오픈 데이터로 구축한다는 원칙을 따름
 - 데이터의 수집과 관리, 분석과 시각화, 개방의 모든 단계에서 오픈소스를 활용하고, 누구나 데이터를 활용할 수 있는 온라인 공유 환경을 구축함
- **(데이터 수집과 분석 방법)** 코로나19 관련 데이터세트 총 1,511건과 공공데이터 포털 API, 정부의 보도자료와 정례브리핑을 수집하고, 파이썬 환경에서 데이터 분석 후 시각화함
 - 주제 분야별로 데이터를 구분하고 공공 데이터 포털에서 제공하는 코로나19 관련 API 데이터를 수집함. 정례브리핑의 자료를 기계가 읽을 수 있는 형태로 변환하여 데이터를 직접 구축함
 - 수집한 데이터는 구글 코랩 환경에서 파이썬 라이브러리를 활용해 분석을 진행하고, 분석

결과는 깃허브를 통해 공유함

- **(데이터 분석과 시각화 사례)** ‘코로나19: 우리의 기억’ 아카이브는 정부 정책 및 대응, 교육, 문화, 사회, 경제, 환경으로 주제를 구분해 코로나19 관련 데이터를 분석한 시각화 사례와 확진 현황 공공데이터, 집단감염, 국가별 백신 현황 데이터를 분석한 사례를 제공함
 - 코로나19 타임라인과 스토리맵을 제공해 시간의 흐름에 따라 변화하는 코로나19 관련 주요 이슈를 시각화함
 - ‘COVID19 데이터 분석 A to Z’ 세미나에서 공공데이터 API, 집단감염 데이터, 국가별 백신 현황을 시각화하고 정리한 웹 슬라이드와 데이터 분석 소스 코드를 함께 공유함

5.1.3 ‘코로나19 아카이브’의 구축 방향과 서비스 방안

- **(목표와 비전)** 전 세계적으로 경쟁력 있는 ‘코로나19 디지털 아카이브’의 구축을 위해 ‘재현’, ‘연결’, ‘공유’, ‘활용’의 4가지 핵심 가치를 정의하고, 구체적인 실현 방안을 제공함
 - 코로나19 관련 자료가 사회의 기민한 변화를 반영하는 특성을 고려하여 이를 효과적으로 수집·관리할 수 있는 정책과 가이드라인이 필요하고, 수집된 자료가 자원이 내포하는 맥락을 고려하여 ‘재현’될 수 있어야 함
 - ‘코로나19 아카이브’는 산재하고 있는 코로나19 관련 자료를 연계하여 일종의 허브로 기능하는 것이 바람직하며, 이를 위해 독자적으로 자료를 수집하고 공공과 민간의 아카이브를 연계할 수 있는 방안이 필요함
 - 공동활용체계를 구축해 협업이 가능한 개방형 아카이브를 지향해야 하며, 국민이 웹 아카이브를 활용할 수 있는 다양한 서비스를 제공해야 함
- **(구축방안)** 코로나19 아카이브의 연계와 확장을 위해 국내외 표준 기반을 적용한 데이터를 구축하고, 오아시스 재난아카이브와 연계할 수 있는 프레임워크를 제안함
 - RDF 기반의 표준 어휘인 DCAT, SKOS, Datamap은 디지털 자원을 표현하기 위한 어휘로, 표준 어휘로 구축한 코로나19 아카이브는 메타데이터와 정보 자원을 의미적으로 탐색할 수 있음
 - 국내외 표준 어휘를 사용해 디지털 아카이브를 기술하고, 오아시스의 재난아카이브와 코로나19 아카이브의 연계를 위한 지식 모델을 설계함
- **(데이터 수집·관리와 연계)** 디지털 자원의 탐색, 상호 연결을 위해 FAIR 데이터 원칙을 준수한 데이터 수집과 관리 방안이 도입되어야 하고, 표준 어휘를 활용해

코로나19 관련 자료를 연계해야 함

- 웹 아카이브가 수집한 자원이나 콘텐츠를 활용하는 사례가 매우 적으므로, 자원의 수집 단계부터 FAIR 원칙을 도입하여 자원의 공유와 재사용이 가능한 체계를 구축해야 함
- FAIR 원칙을 준수한 자원은 표준 어휘로 메타데이터가 기술될 뿐만 아니라, 기계가 읽고 이해할 수 있는 형식으로 구축되어 코로나19 관련 자원 사이의 연계가 가능해짐
- **(기술적 고려사항)** 코로나19 아카이브는 자원의 재사용을 위해 기계가 읽을 수 있는 형식으로 표현하고, 통합적으로 검색할 수 있는 프레임워크를 구축해야 함. 이용자가 활발히 아카이브를 이용할 수 있는 인공지능 기반의 서비스 모델과 UI/UX 디자인을 검토할 필요가 있음
 - 디지털 아카이브의 통합 검색 서비스를 위해 표준 어휘로 구축한 아키텍처를 설계함. 트리플 저장소와 SPARQL 엔드포인트에서 의미적으로 연계된 자원을 질의할 수 있고, 이용자의 이해를 돕기 위한 그래프 형태의 시각화를 제안함
 - 지능형·소통형 서비스를 위해 7개의 인공지능 서비스 모델을 제안하였고, 각 서비스 모델은 개발 소요 시간에 따라 2개의 단기 과제와 3개의 중기 과제, 장기 과제 2개로 구분됨
 - 아카이브의 효과적인 시각화를 위해 활용할 수 있는 오픈 소스와 도구를 조사하고 각 기능과 특징, 실제 적용 사례를 정리함
- **(아카이브의 활용 활성화)** 코로나19 자원의 확대와 활용 활성화를 위해 민간 아카이브와 연계할 수 있는 방안을 제시하고, 국민 참여형 서비스를 위해 도입할 수 있는 기능을 검토함
 - 민간 아카이브는 수집 대상과 관리 주체가 다양하므로 오아시스의 웹사이트 수집 로봇에 한계가 있음. 데이터 수준의 연계를 위한 API 서비스 제공과 오픈 데이터 정책을 도입한 공유 체계 마련이 필요함
 - 국민 참여형 서비스를 도입하기 위해 이용자의 참여가 활발한 디지털 아카이브 사례 7개를 오아시스와 비교한 결과, 오아시스는 적극적으로 기록을 공유할 수 있는 환경을 제공하고, 이용자가 적극적인 기록의 생산자 또는 기여자가 될 수 있는 기반을 마련하는 것이 필요함

5.2 시사점

5.2.1 국가 차원의 코로나19에 대한 기록

- 코로나19 감염병은 전 세계적으로 유례없는 재난으로, 이에 대한 국가 차원의 자료를 수집, 보존하는 것은 매우 중요한 의미가 있음

- 현재 코로나19 자료는 중앙부처, 지방자치단체, 공공과 민간 기관에서 자발적이고 산발적으로 추진되고 있음
- 코로나19 아카이브 구축을 통해 산재하고 있는 자료를 수집하고, 서로 다른 기관이 보유하고 있는 자료를 연계하기 위한 관리체계를 수립하고 적용하는 것이 필요함
 - 광범위한 주제를 포괄하는 코로나19 관련 자료를 체계적으로 수집·보존하기 위해 수집 정책, 보존 가이드라인, 자료의 연계를 위한 규정을 명확히 정의하고, 유관 기관과 긴밀하게 협력해야 함
 - 코로나19 자료는 연관 있는 주제와 범위가 매우 넓지만, 이에 대한 자료가 충분하지 않기 때문에 활용에 제한이 있음. 따라서, 수집 대상과 범위를 명확히 정의해야 함
- 향후 오아시스 재난아카이브와 통합 또는 확대 개편할 수 있도록 기술·데이터·정책 차원에서 종합적인 로드맵을 수립해서 대응하고, 관련 예산과 조직 구성을 위해 선제적으로 대응해야 함

5.2.2 수량 중심에서 문제해결 중심으로

- 웹 자료의 수집과 보존은 중요하지만, 웹 아카이브의 활용이 매우 저조하기 때문에 활용을 개선하기 위한 방안 검토가 시급함
- 수집 자료의 수량적 접근이 아닌 문제 해결을 위한 방식으로 웹 아카이브의 수집과 운영 정책의 변화가 필요함
 - 수집한 자료가 활용되지 않는 상황에서 수집 중심으로 아카이브가 운영되면 저장 공간의 물리적 제약, 예산 확보의 어려움이 있고, 궁극적으로 아카이브 지속 가능성에 문제가 발생함
 - 사회 현안 해결을 위해 웹 아카이브의 적극적 활용 방안을 모색하고, 수집된 자료에서 고가치 데이터를 추출하여 활용할 수 있는 기능을 제공해야 함
- 웹 아카이브와 데이터 포털의 자원을 연계하고, 적극적으로 데이터 재생산을 지원하는 정책적 검토가 필요함
 - 코로나19 관련 데이터는 기관에 따라 관리 형식이 다르고, 다수의 데이터는 데이터포털 또는 정보 공개 방식으로 이미 공개되고 있음
 - 웹 아카이브에 수집된 자료와 공공데이터로 개방된 자료를 연계·융합할 수 있는 기술적 방안을 검토하고, 라이선스 또는 이용조건에 대한 법률적 검토를 통해 활용을 저해하는 요소를 제거해야 함
 - 웹 아카이브의 라이선스·이용조건으로 장소/공간/시간의 제약이 있기 때문에 데이터 프리존으로 접근할 수 있는 환경이 조성되어야함. 이외에도 관련 데이터의 연계와 융합을 지원할 수 있는 방안을 모색해야함

□ 국민 참여 중심의 개방형 아카이브 플랫폼

- 전문가, 커뮤니티, 시민의 참여를 통해 아카이브가 성장할 수 있는 되먹임 구조 방식의 운영 체계를 수립하고, 아카이브를 통해 새로운 가치와 협력을 이끌어 낼 수 있는 플랫폼 생태계 구축이 필요함
 - 현재 운영 중인 대부분의 디지털 아카이브는 디지털 자료의 보존과 제공에 한정되어 있고, 참여 기반이 매우 취약한 구조임. 자료의 수집과 활용에 이르는 전 단계에서 사용자의 피드백을 적극적으로 수용하고, 사용자가 제공하는 자료를 공유할 수 있는 기능을 제공해야 함
- 웹 아카이브는 자료의 수집과 보존 측면에서 문헌정보, 기록관리의 이론과 실무적 지식이 필요하지만, 웹 크롤링, 색인, 자료의 추출과 검색은 소프트웨어 공학적 지식이 상당 부분 필요함
 - ArchivedUnleashed, Archive-It 에서 추진하고 있는 클라우드 기반 데이터 추출과 분석은 수집된 자료의 활용을 위해 참고할 수 있는 사례임
 - 코로나19 아카이브를 포함한 웹 아카이브는 복합 학문을 융합하여 대응하는 것이 효과적임. 이를 위해 국립중앙도서관을 중심으로 문헌정보, 기록관리, 소프트웨어공학의 전문가들이 협업할 수 있는 체계를 마련하는 것이 필요함

붙임 1. 코로나19 국내외 데이터 사례 목록

번호	국내외 분류	사이트명	URL
1	국내	보건복지부 코로나바이러스감염증-19(질병관리청)	http://ncov.mohw.go.kr/
2	국내	코로나19 연구정보 포털(국가생명연구자원정보센터)	https://www.kobic.re.kr/covid19/
3	국내	코로나19 데이터 타임캡슐(4차산업혁명위원회, 2021)	-
4	국내	코로나바이러스 감염증-19 상황판(통계청)	https://kosis.kr/covid/covid_index.do
5	국내	SGIS 코로나19 발생 및 예방접종 현황(통계청)	https://sgis.kostat.go.kr/view/thematicMap/thematicMapMain?stat_thema_map_id=sAXkcVzk5V202007141335257355ued9032uw&theme=CTGR_005&mapType=05
6	국내	코로나19 특허정보 내비게이션(특허청)	https://www.kipo.go.kr/ncov/
7	국내	아시아지역정보센터 (Asia Regional Information Center, ARIC)	https://sites.google.com/view/snuaric/data/covid-19-data
8	국내	covid19board (기획재정부, 한국개발연구원(KDI) 경제정보센터)	http://www.covid19board.kr/
9	국내	코로나나우 데이터센터	https://www.coronanow.kr/dev-datacenter/dev_data.html
10	국내	DS4C	https://www.kaggle.com/kimjihoo/coronavirusdataset/activity
11	국내	한국생명공학연구원 - 코로나19 A to Z	https://www.kribb.re.kr/kor/sub05/sub05_03_01_01.jsp
12	국내	생물학연구정보센터 - 과학으로 본 코로나19	https://www.ibric.org/labox/index.php?labox=5
13	국내	OASIS (Online Archiving & Searching Internet	http://www.oasis.go.kr/collection/mainDisa.do

		Sources)	
14	국외	Documenting COVID-19	https://documentingcovid19.io/
15	국외	Coronavirus Web Archive (Library of Congress)	https://www.loc.gov/collections/coronavirus-web-archive/about-this-collection/
16	국외	IIPC Novel Coronavirus (국제 인터넷 보존 컨소시엄)	https://netpreserve.org/projects/collaborative-collections/
17	국외	CDC(미국 질병통제예방센터)	https://www.coronavirus.gov
18	국외	Our World in Data	https://ourworldindata.org/coronavirus-source-data
19	국외	Oxford COVID-19 Government Response Tracker	https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker
20	국외	UKRI(영국 연구혁신기구)	https://strategicfutures.org/TopicMaps/UKRI/research_map.html
21	국외	NEW YORK TIMES	https://www.nytimes.com/interactive/2021/us/covid-cases.html
22	국외	Bloomberg	https://www.bloomberg.com/graphics/2020-coronavirus-cases-world-map/
23	국외	The sex, gender and COVID-19 Project	https://globalhealth5050.org/the-sex-gender-and-covid-19-project/#
24	국외	WHO	https://covid19.who.int/?gclid=EAlaQobChMIzLKghvjf6QIVO8EWBR1MZQIUEAAYASAAEgKQefD_BwE
25	국외	영국 정부 홈페이지	https://www.gov.uk/coronavirus
26	국외	A Journal of the Plague Year: Share your story	https://covid-19archive.org/s/archive/page/Share

붙임 2. 오아시스 재난아카이브

재난유형	제목	발생기간	총 수량	웹사이트 수량	문서 수량	동영상 수량	이미지 수량
호우	2020년 여름 집중호우	2020.06.24~진행중	373	351	16	0	6
건물화재	이천 물류창고 화재	2020.04.29~진행중	207	178	2	0	27
전염병	코로나바이러스감염증-19(코로나19)	2020.01.20~진행중	16,103	14,448	234	17	1,404
생물재난	아프리카돼지열병	2019.09.17~2019.12.31	108	108	0	0	0
해난사고	헝가리 유람선 사고	2019.05.29~2019.07.30	262	250	4	3	5
산불	2019 강원도 산불	2019.04.04~2019.04.04	178	71	51	25	31
전염병	메르스 환자 발생	2018.09.08~2018.10.16	649	601	15	7	26
호우	한반도 집중호우 강타	2018.08.26~2018.09.11	395	328	30	4	33
폭염	전국폭염비상	2018.08.01~2018.08.23	860	620	81	60	99
해난사고	세월호 4주기	2018.04.16~2018.04.23	187	158	3	10	16
건물화재	제천 스포츠센터 화재	2017.12.21~2018.01.11	184	155	5	10	14
지진	2017년 포항지진	2017.11.15~2017.11.15	143	143	0	0	0
호우	2017년 중부지방 폭우 피해	2017.07.16~2017.07.16	94	84	3	7	0
생물재난	2017년 구제역	2017.02.01~진행중	234	165	37	31	1
생물재난	2017년 조류인플루엔자(AD)	2017.01.01~2017.06.28	329	240	66	14	9

화재	2016년 대구 서문시장화재 및 2017년 소래포구어시장화재	2016.11.30~2017.03.18	93	72	10	11	0
지진	2016년 경주지진	2016.09.12~진행중	491	272	186	31	2
해난사고	2015년 추자도 돌고래호 전복 사고	2015.09.05~2015.09.05	73	17	12	44	0
전염병	2015년 대한민국 중동호흡기증후군(메르스, MERS) 유행	2015.05.20~2015.12.23	1,417	723	672	6	16
건물붕괴	2014년 판교 공연장 환풍구 붕괴 사고	2014.10.17~2014.10.17	10	0	9	0	1
건물화재	2014년 장성 요양병원 화재 사고	2014.05.28~2014.05.28	56	31	4	20	1
해난사고	2014년 세월호 침몰사고	2014.04.16~진행중	8,694	7,425	893	355	21
건물붕괴	2014년 경주 마우나오션리조트 체육관 붕괴 사고	2014.02.17~2014.02.17	71	27	1	42	1
유도선사고	우이산호 충돌 유류 오염사고	2014.01.31~2014.01.31	220	167	53	0	0
유해화학물질 유출사고	2012년 구미 불산 누출 사고	2012.09.27~2012.09.27	409	55	97	119	138
태풍	2012년 태풍 산바	2012.09.15~2012.09.17	75	40	0	35	0
태풍	2011년 집중호우와 제9호 태풍	2011.08.06~2011.08.10	41	41	0	0	0

	「무이파」						
산사태	2011년 우면산 집중호우 산사태	2011.07.25~2011.07.28	73	41	6	25	1
해난사고	2010년 천안함 침몰 사고	2010.03.26~2010.03.26	1,767	1,354	412	0	1
문화재재난	2008년 숭례문 화재 사건	2008.02.10~2008.02.11	371	299	72	0	0
해난사고	2007년 허베이스피리트호 원유 유출 사고	2007.12.07~2007.12.07	568	280	1	128	159
해난사고	2007년 골든로즈호 충돌 침몰 사고	2007.05.12~2007.05.12	38	37	0	0	1
건물화재	2007년 여수출입국관리사무소 화재 사고	2007.02.11~2007.02.11	237	236	0	1	0
산불	2005년 강원 양양군 화재	2005.04.04~2005.04.06	31	24	0	6	1
태풍	2003년 태풍 매미	2003.09.12~2003.09.13	60	31	7	15	7
화재	2003년 대구지하철 화재 사고	2003.02.18~2003.02.18	352	177	6	39	130
산불	2000년 동해안 산불	2000.04.07~2000.04.15	56	44	2	9	1
건물화재	1996년 씨랜드 청소년수련원 화재 사고	1999.06.30~1999.06.30	139	109	30	0	0
항공재난	1997년 대한항공 801편 추락 사고	1997.08.06~1997.08.06	19	0	0	18	1
건물붕괴	1995년 삼풍백화점 붕괴	1995.06.29~1995.06.29	307	196	73	37	1

붙임 3. 디지털 아카이브를 위한 FAIR 데이터 평가 프레임워크

원칙	평가대상	점수	평가지표	설명	관련표준
F	식별자	1	No PID	<ul style="list-style-type: none"> 웹 표준(W3C), 디지털 아카이브 관련 표준 기반의 식별자 제공 링크드 데이터의 URI 식별 체계 적용(전역적인 식별자 생성) 	doi, Handle, Archival Resource Key(ARK), ISO 9834-8 UUIDs, PURLs
		2	내부 시스템 ID (예: 데이터베이스 ID)		
		3	URL		
		4	전역 URI		
		5	전역, 영구적 PID		
	메타데이터	1	메타데이터 없음	<ul style="list-style-type: none"> 디지털 기록의 포괄적 메타데이터 (더블린코어, Schema.org) 디지털 기록의 수집, 생성, 기술, 출처에 대한 메타데이터 Readme 파일, 버전관리 파일 검색엔진이 디지털 아카이브의 메인페이지를 통해 데이터 형식을 처리할 수 있는 메타데이터(schema: Collection 등) 	RDA, AMREMM, BIBFRAME, MARC 21, MODS, VRA Core, DCMI, Schema.org
		2	메타데이터가 충분하지 않음		
		3	제한된 메타데이터 제공		
		4	충분한 메타데이터 제공		
		5	확장된 메타데이터 제공과 부가적인 문서 제공		

A	메타 데이터	1	데이터 또는 메타데이터 접근 불가	<ul style="list-style-type: none"> • 개방형 프로토콜을 통해 디지털 아카이브에 접근(예: HTTP, HTTPS, FTP, TFTP, SFTP) • 명시적으로 데이터 접근을 기술한 문서 제공 	OAI-PMH, Search/Retrieve via URL(SRU)
		2	메타데이터에 한정해 접근		
		3	제한된 방식으로 접근 (상용목적의 접근, 엠바고 정책 적용)		
		4	일반적인 접근 (public access) - 등록 필요		
		5	제한 없는 전면 접근		
I	데이터형식	1	독점적인 형식	<ul style="list-style-type: none"> • 특정 소프트웨어에 제한되지 않는 개방형 파일 형식 • 기계가 읽고 이해할 수 있는 메타데이터 형식으로 제공 • 디지털 자원의 내용협상(content negotiation) 제공 • SPARQL 엔드포인트를 통해 디지털 자원에 접근, 검색 기능 제공 	X3D, FLAC, MP3, DPX, MPEG 4, TIFF, Jpeg2000(JP2), ASCII Text, CSV
		2	높은 비율의 독점적인 형식 제공		
		3	개방형 포맷		
		4	개방형, 구조적인 형식		
		5	개방형, 구조적, 기계가 처리할 수 있는 형식		
	어휘	1	표준어휘를 사용하지 않음	<ul style="list-style-type: none"> • 디지털 아카이브와 관련된 통제 어휘, 텍소노미, 온톨로지의 적용 • 기록관리 표준(국내) • 기록관리 메타데이터 표준: EAD, ISAD(G), OAIS, PREMIS, 	<ul style="list-style-type: none"> • 아카이브 메타데이터 구조: ISAD(G), RAD • 메타데이터 구조: ISO 15836, ISO 23081 • XML DTD
		2	자체적으로 정의한 어휘 사용		

		3	제한적으로 표준 어휘 사용	AGRkMS • 온톨로지 어휘: RiC-O, Schema.org	/스키마: EAD, EAG, MODS,RDF • 디지털 보존: ISO/TR 18492, PREMIS • 레퍼런스모델 /프레임워크: ISO 15489, ISO 14721, PAIMAS(ISO 20652), BSI PAS 197
		4	표준어휘 사용		
		5	의미기반 표준 어휘 사용		
R	라이선스	1	라이선스 없음	• 데이터 라이선스에 대한 설명, 링크 정보 제공 • 기계가 처리할 수 있는 형식의 라이선스 정보 제공	Creative Commons(CC0, CC BY, CC BY NC, CC BY NC ND), BSD, MIT, GNU GPL, Open Database license
		2	이용 약관(Terms of Use)		
		3	라이선스가 있으나 구체적인 정보를 제공하지 않음		
		4	라이선스 있음(기계가 읽을 수 있는 방식 X)		
		5	기계가 읽을 수 있는 라이선스		

붙임 4. Blacklight를 사용하여 웹사이트를 구축한 사례

번호	사이트명	URL
1	Havard Digital Collections	https://library.harvard.edu/digital-collections
2	Academic Commons	https://academiccommons.columbia.edu/
3	Rock and Roll Hall of Fame	http://catalog.rockhall.com/
4	Catalyst @ Johns Hopkins Libraries	https://catalyst.library.jhu.edu/
5	CLIO	https://clio.columbia.edu/
6	Yale Collection of Musical Instruments	http://collection-media.yale.edu/
7	US National Library of Medicine Digital Collections	https://collections.nlm.nih.gov/
8	United States Holocaust Memorial Museum	https://minerva.ushmm.org/
9	NCSU Libraries' Digital Collections	https://d.lib.ncsu.edu/collections
10	European Collections	https://www.europeana.eu/en
11	Stanford' s Online Exhibits	https://exhibits.stanford.edu/
12	DTU Findit - Your digital library	https://findit.dtu.dk/
9	French Revolution Digital Archive	http://frda.stanford.edu/
10	Cornell University Library Catalog	https://newcatalog.library.cornell.edu/

11	Revs Digital Library	https://exhibits.stanford.edu/automobility
12	Penn State ScholarSphere	https://scholarsphere.psu.edu/
13	University of Virginia Library Catalog	https://search.lib.virginia.edu/
14	Stanford University Library Catalog	https://searchworks.stanford.edu/
15	Prints & Photographs Online Catalog	http://wallachprintsandphotos.nypl.org/
16	Catholic Parish Registers at the NLI	https://registers.nli.ie/
17	Brown University Library	https://library.brown.edu/
18	Indiana University IUCAT	https://iucat.iu.edu/
19	Linn-Benton Community College	http://libfind.linnbenton.edu/
20	Vera & Donald Blinken Open Society Archives	https://catalog.osaarchivum.org/
21	Royal Danish Library – Digital collection	http://www5.kb.dk/editions/any/2009/jul/editions/en
22	Archive for Danish literature	https://tekster.kb.dk/
23	Temple University Library	https://librarysearch.temple.edu/
24	Phoebe A. Hearst Museum of Anthropology	https://portal.hearstmuseum.berkeley.edu/
25	Pennsylvania state University Libraries’ Catalog	https://catalog.libraries.psu.edu/
26	Chorus: Towards an open source stack for e-commerce search	http://chorus.dev.o19s.com:4000/

[참고문헌]

○ 논문

- 김학래. (2020). “코로나-19: 우리의 기억” : 코로나바이러스 감염증과 사회변화에 대한 디지털 아카이브. 한국기록관리학회지 20(4), 229-236.
<http://dx.doi.org/10.14404/JKSARM.2020.20.4.229>
- 김학래. (2021). FAIR 원칙: 데이터 관점의 디지털 아카이브 구현을 위한 고려사항. 한국기록관리학회지, 21(2), 155-172. <https://doi.org/10.14404/JKSARM.2021.21.2.155>
- 김현정. (2017). 재난기록의 웹 아카이빙에 대한 연구 - OASIS 재난아카이브의 역할과 기능을 중심으로 -. 한국기록관리학회지, 17(2), 27-56.
- 김희정. (2010). 웹 아카이빙 인터페이스 유형 및 특성 분석 - 국가주도 웹 아카이빙 사업을 중심으로 -. 한국기록관리학회지, 10(2), 147-170.
- 남재우, 이수영. (2021). 웹 아카이브 OASIS 개선방안에 관한 연구. 융합정보논문지, 11(2), 1-9. <https://doi.org/10.22156/CS4SMB.2021.11.02.001>
- 박진경, 김유승. (2017) 참여형 디지털 아카이브 활성화 방안. 한국비블리아학회지, 28(2), 219-243. <https://doi.org/10.14699/kbiblia.2017.28.2.219>
- 배양희, 김유승. (2017). 민간 디지털 아카이브 수집정책 연구. 기록학연구, 53, 119-159. <https://doi.org/10.20923/kjas.2017.53.119>
- 손동유. (2021). 민간 아카이브 활성화를 위한 법·제도 모색. 기록학연구, 69, 7-33. <https://doi.org/10.20923/kjas.2021.69.007>
- 심지연 & 김지현. (2021). 코로나19 아카이브를 위한 기록 수집 방안 연구. 기록학연구, 70, 189-243.
- 유신성 & 오효정. (2021). 재난백서 구축을 위한 온라인 기록물 활용 방안. 한국기록관리학회지, 21(2), 135-154.
- 유호선 & 오효정. (2018). 이슈 생존 주기 모형 기반 재난 아카이브 수집 방안. 정보관리학회지, 35(2), 115-139.
- 윤정옥. (2013). 차세대 도서관 목록의 제반 기능에 관한 분석- SearchWorks를 중심으로 -. 한국문헌정보학회지, 47(4), 5-23.
- 윤정옥. (2014). 웹 아카이브 OASIS 수집 콘텐츠의 분석. 한국도서관 정보학회지, 45(4), 45-65. <https://doi.org/10.16981/kliss.45.201412.45>
- 한희정, 강주연, 김용, & 오효정. (2017). 자동수집 기반 재난안전 기록정보자원

아카이브 설계 및 구축전략. 한국기록관리학회지, 17(4), 127-154.
<https://doi.org/10.14404/JKSARM.2017.17.4.127>

현문수. (2013). 민간·공동체 아카이브 네트워크 구축을 위한 기술정보 공유 방향 연구. 한국기록관리학회지, 13(3), 127-150.

Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013). Key Choices in the Design of Simple Knowledge Organization System (SKOS). Journal of Web Semantics, 20, 35-49. <https://doi.org/10.1016/j.websem.2013.05.001>

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021). Named Entity Recognition and Classification on Historical Documents: A Survey. arXiv.2109.11406. <https://doi.org/10.48550/arXiv.2109.11406>

Guha, R.V., Brickly, D. & Macbeth, S. (2016). Schema.org: Evolution of Structured Data on the Web. Communications of the acm, 59(2), 44-51.
<https://doi.org/10.1145/2844544>

Hawkins, A. (2022). Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. Archival Science, 22, 319-344. <https://doi.org/10.1007/s10502-021-09381-0>

Lampron, P., Mixter, J., & Han M. K. (2016). Challenges of Mapping Digital Collections Metadata to Schema.org: Working with CONTENTdm. In Metadata and Semantics Research. New York City: Springer, 181-186.

Matienzo, M. A., Roke, E. R., & Carlson, S. (2017). Creating a Linked Data-Friendly Metadata Application Profile for ArchivalDescription. Proc. Int' l Conf. on Dublin Core and Metadata Applications 2017, 112-116.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

○ 단행본

딘 알레망, 제임스 핸들러, 파이엔 갠던. (2021). 온톨로지 과학자를 위한 지식그래프. 한빛아카데미

○ 웹 페이지

5 Star 오픈 데이터. (2022). <https://5stardata.info/ko/>

국립중앙도서관 웹 자원 아카이브 OASIS(Online Archiving & Searching Internet Sources). (2022). <https://nl.go.kr/oasis/>

국립중앙도서관. (2022).

<https://www.nl.go.kr/NL/contents/N50606000000.do?schM=view&id=36566&schBcid=photo0104>

데이터맵 어휘 표준 (TTAK.KO-10.1291). (2022).

<https://www.tta.or.kr/tta/ttaSearchView.do?key=77&rep=1&searchStandardNo=TTAK.KO-10.1291&searchCate=TTAS>

보건복지부 코로나바이러스감염증-19(질병관리청). (2022). <http://ncov.mohw.go.kr/>

생물학연구정보센터 - 과학으로 본 코로나19. (2022).

<https://www.ibric.org/labox/index.php?labox=5>

아시아지역정보센터(Asia Regional Information Center, ARIC). (2022).

<https://sites.google.com/view/snuaric/data/covid-19-data>

정유철. (2022. 02. 12). 중앙대 문헌정보학과 ‘코로나-19: 우리의 기억’ 디지털 아카이브 영구 보존. K스피릿.

<http://www.ikoreanspirit.com/news/articleView.html?idxno=66646>

코로나 연구정보포털(국가생명연구자원정보센터). (2022).

<https://www.kobic.re.kr/covid19/>

코로나19 상황판(통계청). (2022). https://kosis.kr/covid/covid_index.do

코로나19 특허정보 내비게이션(특허청). (2022). <https://www.kipo.go.kr/ncov/>

코로나나우 데이터센터. (2022). https://www.coronanow.kr/dev-datacenter/dev_data.html

한국생명공학연구원 - 코로나19 A to Z. (2022).

https://www.kribb.re.kr/kor/sub05/sub05_03_01_01.jsp

Analytics Vidhya. (2022).

<https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/>

ARCH UX Testing. (2022).

<https://news.archivesunleashed.org/arch-ux-testing-designing-for-the-users-f977923af7e>

Archive-it Blog. (2022). <https://archive-it.org/blog/post/arch-program-update/>

Archive-It Help Center. (2022). <https://support.archive-it.org/hc/en-us>

Blacklight GitHub. (2022). <https://github.com/projectblacklight/blacklight>

Blacklight. (2022). <http://projectblacklight.org/>

Bloomberg. (2022).
<https://www.bloomberg.com/graphics/2020-coronavirus-xwcases-world-map>

CDC(미국 질병통제예방센터). (2022). <https://www.coronavirus.gov>

Coronavirus Web Archive(Library of Congress). (2022).
<https://www.loc.gov/collections/coronavirus-web-archive/about-this-collection/>

covid19board.kr(기획재정부, 한국개발연구원(KDI) 경제정보센터). (2022).
<http://www.covid19board.kr/>

Data Catalog Vocabulary (DCAT) - Version 2. (2022).
<https://www.w3.org/TR/vocab-dcat-2/>

DCAT Application Profile for data portals in Europe. (2022).
<http://data.europa.eu/w21/32d70b6e-0d27-40d9-9230-017e4cd00bcc>

DCMI Metadata Terms. (2022). <http://purl.org/dc/elements/1.1/>

Dixon, Jennifer. (2020. 04. 28). Documenting the Pandemic: Libraries Launch COVID-19 Archival Projects.
<https://www.libraryjournal.com/story/Documenting-Pandemic-Libraries-Launch-COVID-19-Archival-Projects-archives>

Documenting COVID-19. (2022). <https://documentingcovid19.io/>

DS4C. (2022). Kaggle. <https://www.kaggle.com/kimjihoo/coronavirusdataset/activity>

GO FAIR. (2021). FAIR Principles. <https://www.go-fair.org/fair-principles/>

GOV.UK (2022). <https://www.gov.uk/coronavirus>

NEW YORK TIMES. (2022).
<https://www.nytimes.com/interactive/2021/us/covid-cases.html>

Nick Ruest Official Blog. (2022). <https://ruebot.net/45-images/>

OpenSeadragon GitHub. (2022). <https://github.com/openseadragon/openseadragon>

OpenSeadragon. (2022). <https://openseadragon.github.io/#examples-and-features>

Our World in Data. (2022). <https://ourworldindata.org/coronavirus-source-data>

Oxford COVID-19 Government Response Tracker. (2022).
<https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker>

RISS Linked Open Data 활용서비스. (2022). <http://data.riss.kr/serviceHome.do>

Schema.org. (2022). <https://schema.org/>

SGIS 코로나19 발생현황(통계청). (2022).
https://sgis.kostat.go.kr/view/thematicMap/thematicMapMain?stat_thema_map_id=sAXkcVzk5V202007141335257355ued9032uw&theme=CTGR_005&mapType=05

SKOS Simple Knowledge Organization System. (2022).
<http://www.w3.org/2004/02/skos/core#>

The sex, gender and COVID-19 Project. (2022).
<https://globalhealth5050.org/the-sex-gender-and-covid-19-project/#>

UKRI. (2022). https://strategicfutures.org/TopicMaps/UKRI/research_map.html

WARCnet. (2022). <https://cc.au.dk/en/warcnet>

WHO. (2022).
https://covid19.who.int/?gclid=EAlaIQobChMzLKghvjf6QIVO8EWBR1MZQIUEAAYASAAEgKQefD_BwE



문화체육관광부
국립중앙도서관



ISBN 979-11-6513-298-9