

미래 도서관 웹 콘텐츠 수집 정책

세미나 I 웹툰 수집 현황과 과제

세미나 II 미래형 타임캡슐 : 웹 아카이빙(OASIS) 진화와 발전방안

일시 | 2022년 10월 13일(목) 13:40~16:40

장소 | 창원컨벤션센터 601호

주최 | 국립중앙도서관 온라인자료과



제59회 전국도서관대회 세미나

미래 도서관 웹 콘텐츠 수집 정책

세미나 I 웹툰 수집 현황과 과제

세미나 II 미래형 타임캡슐 : 웹 아카이빙(OASIS) 진화와 발전방안

- 일시 : 2022년 10월 13일(목) 13:40~16:40
- 장소 : 창원컨벤션센터 601호
- 주최 : 국립중앙도서관 온라인자료과



국립중앙도서관

National Library of Korea

프로그램 순서

● 세미나 I

- (주 제) 웹툰 수집 현황과 과제
- (일 시) '22. 10. 13.(목) 13:40~15:00(80분)
- (주요내용) 국가지식자원으로서 웹툰 수집 확대의 발전 방향 모색

시 간	구 분	내 용	비 고
13:40-13:50 (10분)	사회	■ 세미나 안내(발표자 등 소개)	김경철 사무관 (국립중앙도서관 온라인자료과)
13:50-14:15 (25분)	발표 1	■ 웹툰산업의 현황 & 주요과제	강태진 대표 (주)코니스트
14:15-14:25 (10분)	발표 2	■ 국립중앙도서관 웹툰 수집 현황 및 과제	배진경 주무관 (국립중앙도서관 온라인자료과)
14:25-14:50 (25분)	발표 3	■ 웹툰 수집·보존의 필요성 및 방향성	한상정 교수 (인천대학교/만화연구가)
14:50-15:00 (10분)	질의응답	■ 질의응답	전체

● 세미나 II

- (주 제) 미래형 타임캡슐 : 웹 아카이빙(OASIS) 진화와 발전방안
- (일 시) '22. 10. 13.(목) 15:20~16:40(80분)
- (주요내용) 웹 아카이빙 현재와 미래, 활용방안 고찰

시 간	구 분	내 용	비 고
15:20-15:30 (10분)	사회	<ul style="list-style-type: none"> ■ 세미나 안내(발표자 등 소개) 	김경철 사무관 (국립중앙도서관 온라인자료과)
15:30-15:40 (10분)	발표 1	<ul style="list-style-type: none"> ■ 국립중앙도서관 웹자원 아카이빙 (OASIS) 현황 및 사례 	안경자 주무관 (국립중앙도서관 온라인자료과)
15:40-16:05 (25분)	발표 2	<ul style="list-style-type: none"> ■ 국가 디지털 아카이브 구축방안 '코로나19 아카이브' 	김학래 교수 (중앙대학교 문헌정보학과)
16:05-16:30 (25분)	발표 3	<ul style="list-style-type: none"> ■ 도서관의 미래, 오아시스(OASIS) 발전 방안 	이수상 교수 (부산대학교 문헌정보학과)
16:30-16:40 (10분)	질의응답	<ul style="list-style-type: none"> ■ 질의응답 	전체

· 목 차 ·

세미나 I 웹툰 수집 현황과 과제

웹툰산업의 현황 & 주요과제	3
강태진 ((주)코니스트 대표)	
국립중앙도서관 웹툰 수집 현황 및 과제	17
배진경 (국립중앙도서관 온라인자료과 주무관)	
웹툰 수집·보존의 필요성 및 방향성	25
한상정 (인천대학교 교수 / 만화연구가)	

세미나 II 미래형 타임캡슐 : 웹 아카이빙(OASIS) 진화와 발전방안

국립중앙도서관 웹자원 아카이빙(OASIS) 현황 및 사례	35
안경자 (국립중앙도서관 온라인자료과 주무관)	
국가 디지털 아카이브 구축방안 '코로나19 아카이브'	53
김학래 (중앙대학교 문헌정보학과 교수)	
도서관의 미래, 오아시스(OASIS) 발전 방안	79
이수상 (부산대학교 문헌정보학과 교수)	

부록 전자책 납본 가이드	93
오아시스(OASIS) 홍보 리플릿	97

웹툰 수집 현황과 과제



세미나 I

발표 1

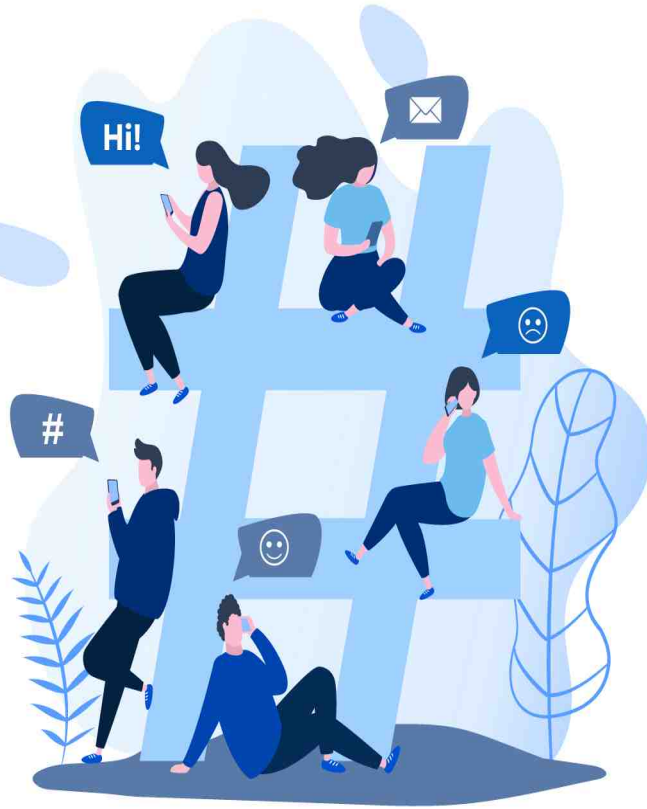
웹툰산업의 현황 & 주요과제

강태진

((주)코니스트 대표)

웹툰산업의 현황 & 주요과제

(주) 코니스트
강태진 대표

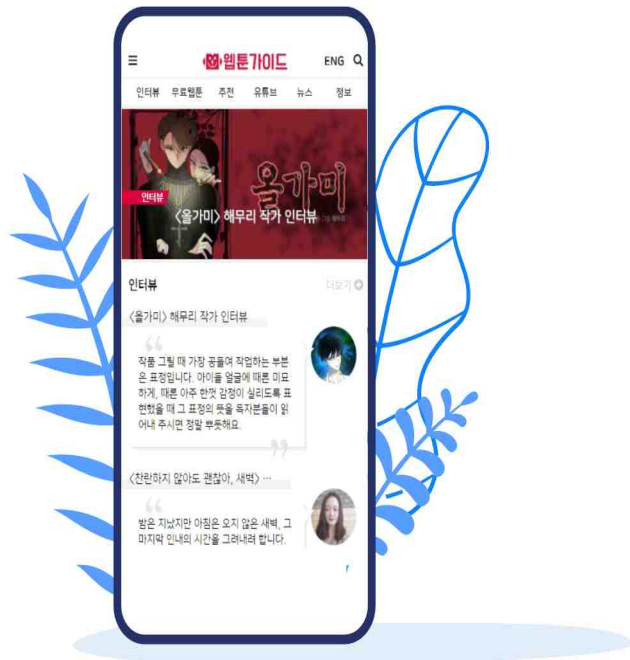


CONIST

CONTENTS

I 웹툰 생태계 및 현황

II 이슈 & 과제



I. 웹툰 생태계 & 현황

(Webtoon Eco-system & Summary)

0. Intro

BTS

0. Intro



<https://youtu.be/x5Nim22WaCs>

0. Intro

문화
(Culture)

0. Intro

문화 비즈니스 인구 비즈니스 헐리웃 영화의 파워

0. Intro

미국: 3.28억
캐나다: 0.35억
영국: 0.66억
호주: 0.25억
인도: 13.53억
필리핀: 1.07억

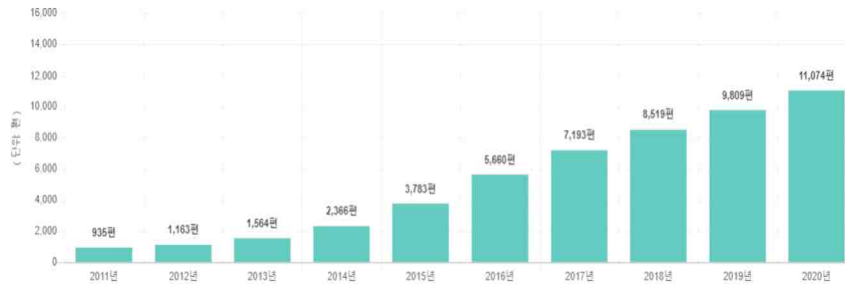
겨울왕국 매출(Revenue)

1년에 3조

→ 19.14억 X 10% X 10,000원 = ???

2. 웹툰 생산 규모

10년간 누적 작품 변화추이



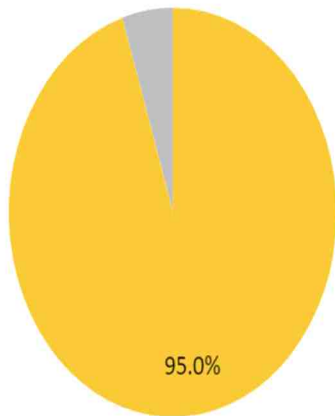
10년간 신규 작품 변화추이



2. 웹툰 생산 규모 (노블코믹스)

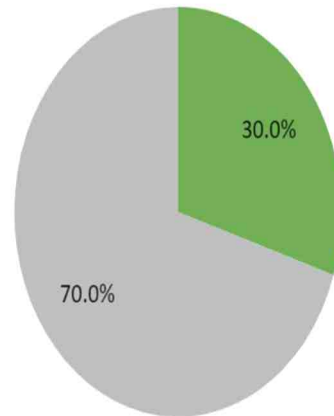
카카오페이지 TOP 200내 노블코믹스 비율

- 노블코믹스 점유율
- 그외의 작품

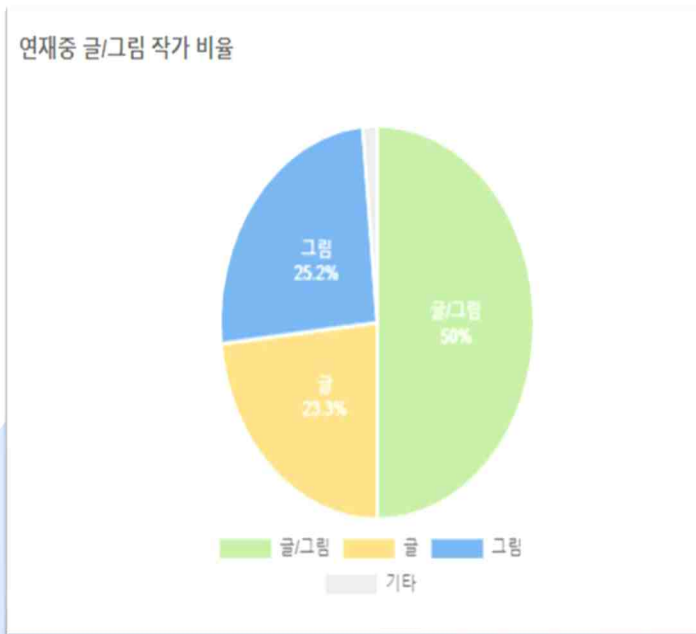


네이버웹툰 TOP 200 내 노블코믹스 비율

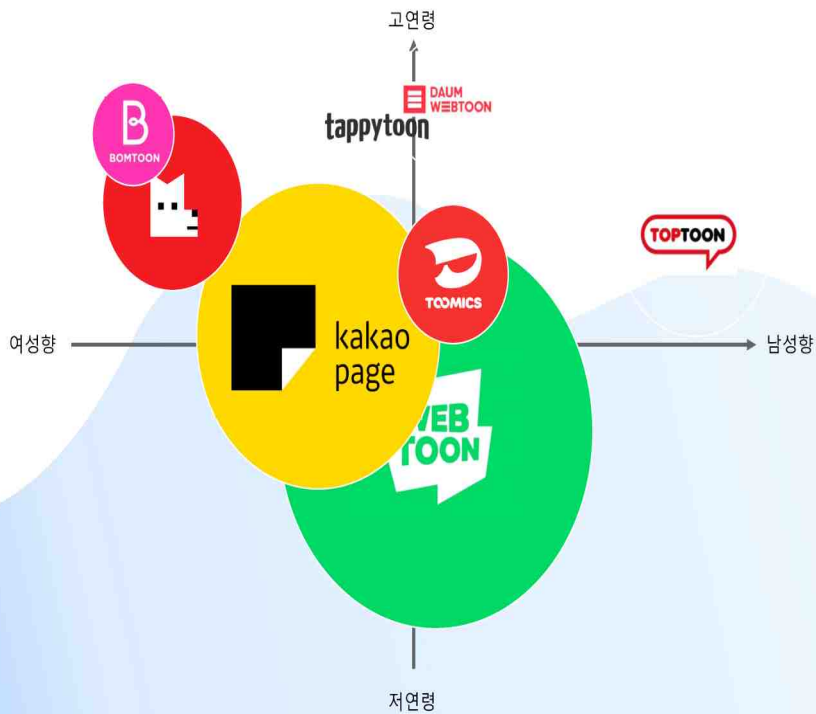
- 노블코믹스 점유율
- 그외의 작품



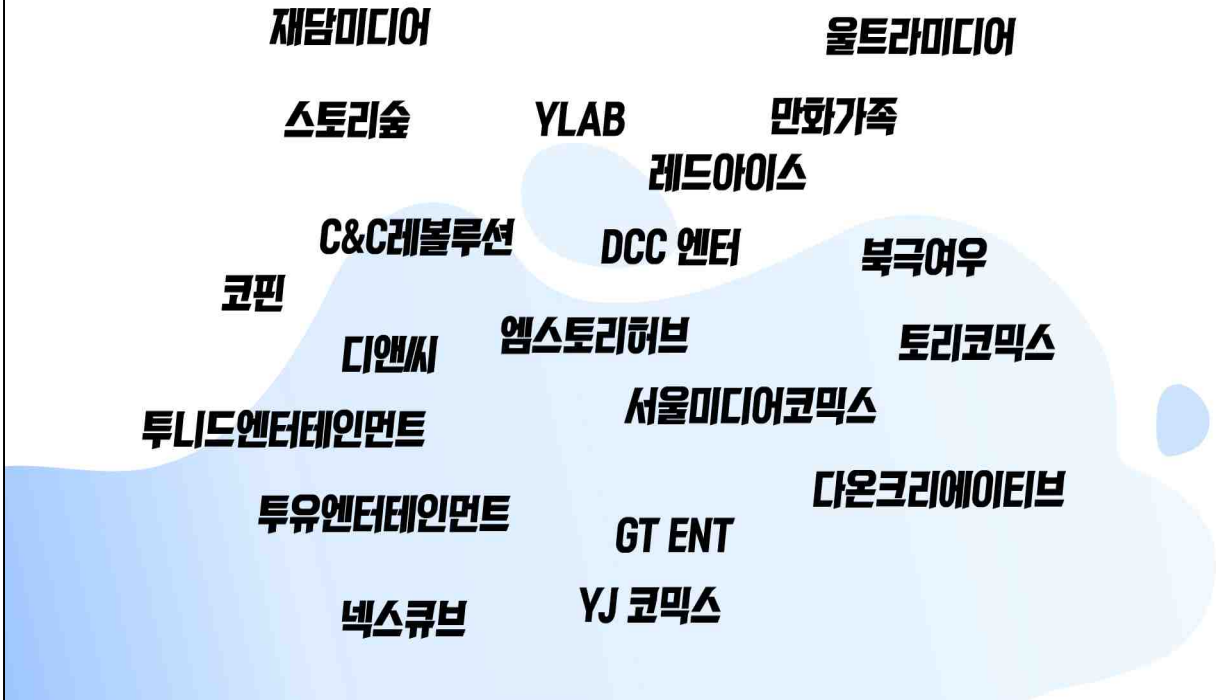
3. 작가 & 작품



4. 국내 주요 웹툰 플랫폼



5. 국내 주요 에이전시 & 제작 스튜디오



5. 국내 주요 에이전시 & 제작 스튜디오 (매출현황)



II. 이슈 & 과제 (Issues)

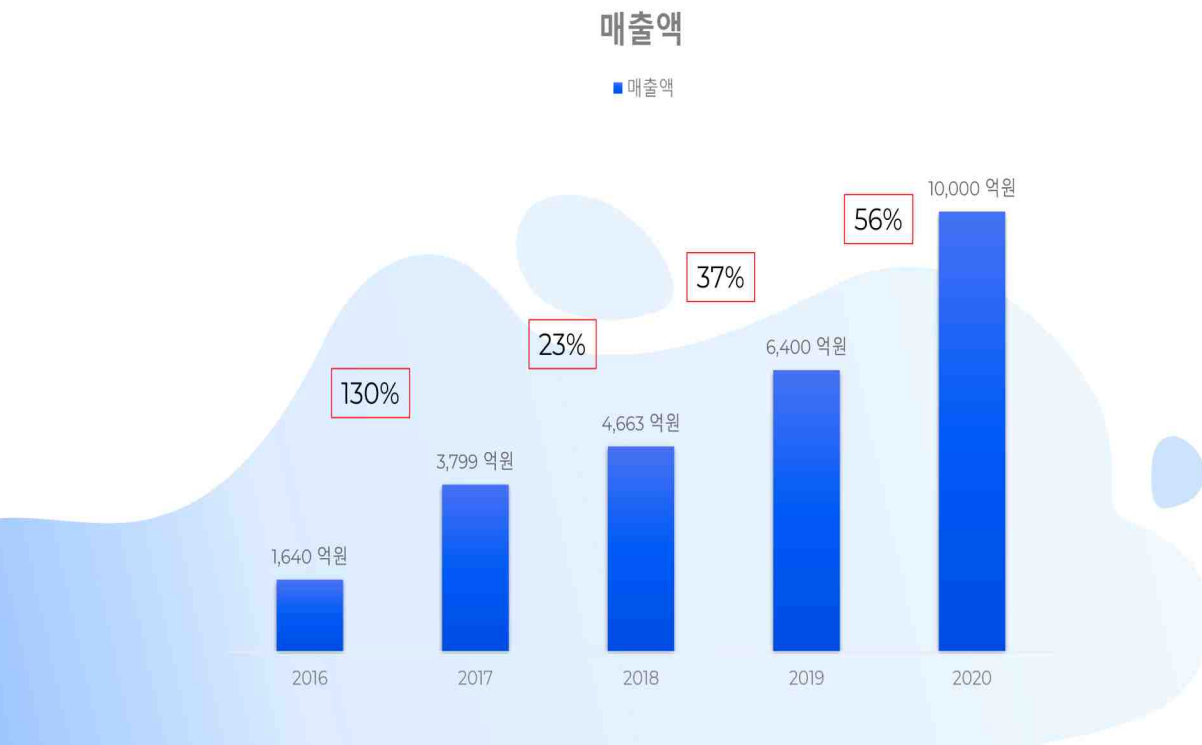
1. 주요 이슈

- ◆ 온라인 콘텐츠 수요 폭증
- ◆ 네이버웹툰과 카카오엔터테인먼트의 경쟁 격화
- ◆ 글로벌 콘텐츠 수급 전쟁
- ◆ 웹툰산업계에 대한 M&A 활성화
- ◆ 웹툰 제작의 산업화 (대형화 & 스튜디오)
- ◆ IP Transmedia 성공 (사내맞선, 내일, 어게인마이라이프, 옷소매끝동)
- ◆ 글로벌 진출 일상화
- ◆ 전자책/웹소설/애니/게임 업체들의 웹툰 진출

1. 주요 이슈

- ◆ 네이버의 왓패드 인수, 카카오페이지의 래디쉬 인수
- ◆ 픽코마의 성장, 6천억 투자, 10조 상장
- ◆ 노란색 줄, 초록색 줄 선택의 기로
- ◆ 직접 해외 진출의 좁은 문
- ◆ 경쟁력 있는 작품의 제작 공식 & 노하우
- ◆ 인력 확보 전쟁 (PD, 제작인력)
- ◆ 규모의 대형화 (레드아이스, 코핀, 스토리술)

2. 웹툰산업 규모와 성장 추이



2. 웹툰산업 규모와 성장 추이

게임 산업 규모 17조 93억

2. 웹툰산업 규모와 성장 추이

■ 웹툰 산업 매출 10년 안에 10배 이상으로 성장

- 언어적 장벽이 큰 문제가 되지 않는 현 상황
- 온라인 생산과 소비가 가능한 완벽한 언택트 비즈니스
- 글로벌 진출의 가속화

SM엔터테인먼트 매출 추이 (단위: 억원)

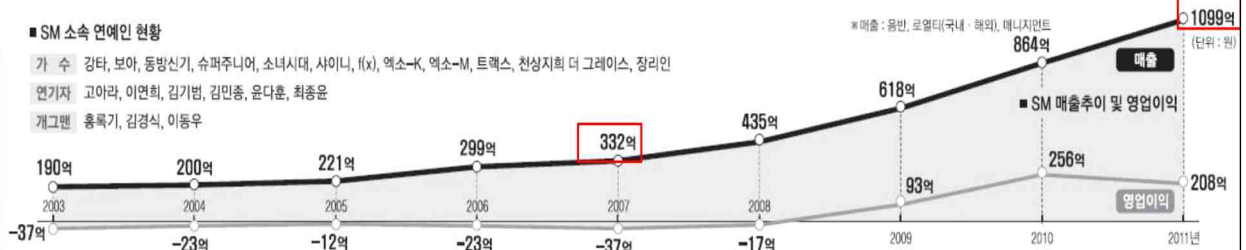
	매출액	국외 매출
2007년	332	75
2008년	435	120
2009년	618	190
2010년	826(예상)	393(~3분기)

*자료: 대우증권

→ 2017년 매출 3,654억원

■ SM 소속 연예인 현황

가 수 강타, 보아, 동방신기, 슈퍼주니어, 소녀시대, 샤이니, f(x), 엑소-K, 엑소-M, 트랙스, 천성지희 더 그레이스, 장리인
연기자 고아라, 이연희, 김기범, 김민중, 윤다훈, 최중운
개그맨 홍록기, 김경식, 이동우



2. 웹툰산업 규모와 성장 추이



Thank You



세미나 I

발표 2

국립중앙도서관 웹툰 수집 현황 및 과제

배진경

(국립중앙도서관 온라인자료과 주무관)

국립중앙도서관 웹툰 수집 현황 및 과제

국립중앙도서관 온라인자료과 배진경 주무관

- 목 차 -

- 01 온라인 자료 수집 업무
- 02 웹툰 수집 현황
- 03 고민과 과제

01

온라인 자료 수집 업무

1. 온라인 자료 수집 업무

✓ 온라인 자료의 정의

도서관법 제2조(정의)

온라인 자료란 정보통신망(「정보통신망 이용촉진 및 정보보호 등에 관한 법률」 제2조 제1항 제1호의 정보통신망을 말한다)을 통하여 공중송신(「저작권법」 제1조 제7호의 공중송신을 말한다) 되는 자료를 말한다.

✓ 온라인 자료 수집의 필요성

- 자료 출판, 생산 환경의 변화
- 높은 휘발성

1. 온라인 자료 수집 업무

✓ 온라인 자료 수집 구분



"누구든지 도서관자료를 발행 또는 제작한 경우 그 발행일 또는 제작일부터 30일 이내에 그 도서관자료를 국립중앙도서관에 납본하여야 한다."



"국립중앙도서관은 대한민국에서 서비스되는 온라인 자료 중에서 **보존가치가 높은 온라인 자료**를 선정하여 수집·보존하여야 한다"

1. 온라인 자료 수집 업무

✓ 관련 법령

도서관법 제20조의2(온라인 자료의 수집) 제1항

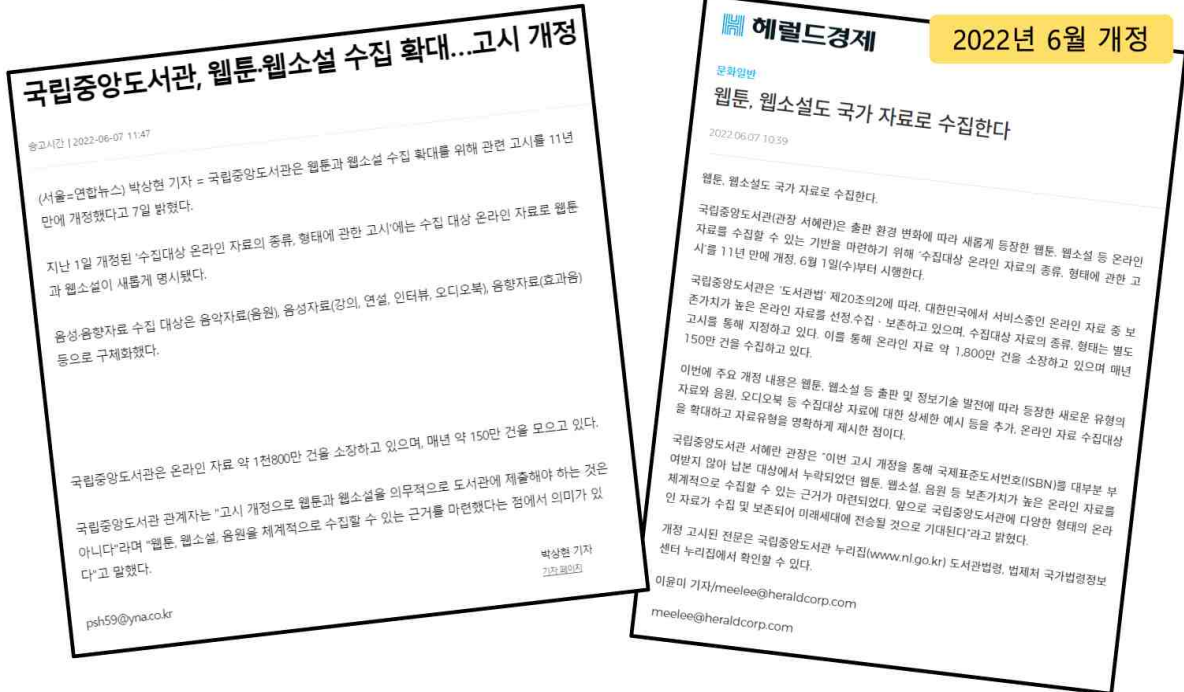
"국립중앙도서관은 대한민국에서 서비스되는 온라인 자료 중에서 **보존가치가 높은 온라인 자료**를 선정하여 수집·보존하여야 한다"

도서관법 시행령 제 13조의2(온라인 자료의 수집) 제1항

법 제20조의2제1항에 따라 국립중앙도서관이 수집하는 온라인 자료는 전자적 형태로 작성된 웹사이트, 웹자료 등으로써 국립중앙도서관장이 시행령 제13조의3에 따른 도서관자료심의위원회의 심의를 거쳐 선정하여 **고시**하는 자료로 한다.

1. 온라인 자료 수집 업무

✓ 고시 수집대상 온라인 자료의 종류 형태에 관한 고시 (2022)



1. 온라인 자료 수집 업무

✓ 수집대상 온라인 자료의 종류

「수집대상 온라인 자료의 종류, 형태에 관한 고시」 국립중앙도서관 고시 제2022-1호

문자자료	• 전자책, 전자저널, 학위논문, 보고서, 전자신문, 웹툰, 웹소설 등
음성음향자료	• 음악자료(음원), 음성자료(강의, 연설, 인터뷰, 오디오북), 음향자료(효과음)
영상자료	• 방송, 영화, 공연, 이러닝자료 등
이미지자료	• 사진, 회화 등

02

웹툰 수집 현황

2. 웹툰 수집 현황

✓ 국립중앙도서관 웹툰 수집 현황

연도		2020	2021	2022	합계
수집건수	작품단위	2,316	321	144	2,781
	화단위	133,116	11,948	15,139	160,203

* 한국만화영상진흥원을 통한 수집

✓ 수집 구분

ISBN 발급

납본



수집

ISBN 발급 여부 무관

03

고민과 과제

3. 고민과 과제

✓ 웹툰 수집의 고민과 과제



수집 대상 선정의 어려움



협력 강화 필요



보상체계 부재

세미나 I

발표 3

웹툰 수집·보존의 필요성 및 방향성

한상정

(인천대학교 교수 / 만화연구가)

웹툰 수집·보존의 필요성 및 방향성

인천대학교
불어불문학과/문화대학원
한상정

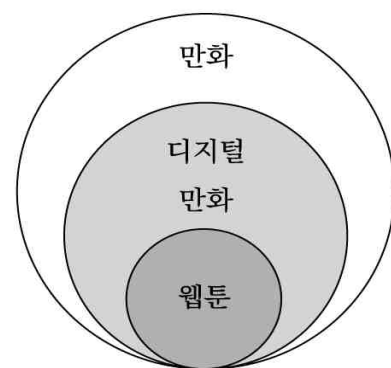
0. 전제

만화의 (형태) 분류

- 한간만화(cartoon)
- 다칸만화(comics)

만화의 (지지대) 분류

- 종이(출판)만화 : 단행본, 시리즈, 잡지, 신문...
- 디지털(출판)만화 : 스캔본, PDF, 웹툰...



웹툰(webtoon)

- 세로스크롤 형식을 주로 하는 디지털출판만화의 한 형식을 지칭했으나
관습상 디지털만화 전체를 대표하기도 함
- 해외의 경우, 한국의 디지털코믹스(digital comics)의 형태로 인지함

1. 웹툰 수집·보존의 필요성

1-1. 보존가치가 높음

- ‘사회’를 읽어내기에 가장 유력한 형식 중의 하나 : 모든 걸 ‘그려낼’ 수 있음
 - 2020년대 한국의 학교생활은 어떤 모습인가
 - 한국사회는 어떻게 그려지고 있는가, 구어는 어떤 식으로 쓰고 있는가
 - 여성과 남성의 신체묘사는 어떻게 재현(그려지고) 있는가
- 다양한 계급 계층의 정서 파악 가능
 - 초등학생부터 성인까지, 여성부터 남성, 사회적 약자에 이르기까지
 - 다양한 계급, 계층의 정서를 가장 잘 읽어낼 수 있는 표현형식 중의 하나
- 동시대 사회적 이슈에 대한 생생한 반응의 집적체
 - 댓글

1. 웹툰 수집·보존의 필요성

1-1. 보존가치가 높음

- 1960년대 유신, 모든 대중예술 영역을 사전심의, 검열. 유독 만화만, 1968년 ‘밀수, 도박, 탈세, 폭력, 마약’과 함께 ‘6대 악’으로 규정
 - 왜 이렇게까지 ?
- 일본만화의 영향력에도 불구하고, 오늘날까지 자국만화가 활발하게 창작, 향유되고 있음

컬럼비아대학 도서관 만화자료 사이트

COMICS @ COLUMBIA

New titles

- Aquaman and the war against oceans: comics activism and allegory in the Anthropocene [CF](#) [i](#) Sep 17, 2022
- Button pusher [CF](#) [i](#) Sep 17, 2022
- Disciples [CF](#) [i](#) Sep 17, 2022
- Free speech handbook: a practical framework for understanding our free speech protections [CF](#) [i](#) Sep 17, 2022
- Things we create [CF](#) [i](#) Sep 17, 2022
- The graphic novel classroom: powerful teaching and learning with images [CF](#) [i](#) Sep 17, 2022
- Saltando matones [CF](#) [i](#) Sep 17, 2022
- Smartguy: the life and times of Barney Frank [CF](#) [i](#) Sep 17, 2022
- Black Panther and philosophy: what can Wakanda offer the world? [CF](#) [i](#) Sep 17, 2022
- Captain Erik: vittoria [CF](#) [i](#) Sep 17, 2022
- Erotic comics in Japan: an introduction to eromanga [CF](#) [i](#) Sep 17, 2022
- Howard Chaykin [electronic resource]: conversations [CF](#) [i](#) Sep 17, 2022

Comics in the Columbia Libraries

The Columbia University Libraries collect both graphic novels for the circulating collection in the 4th-floor West reading rooms and also creator archives in the Rare Book & Manuscript Library. The circulating collection launched in 2005, when the libraries held three titles, and by the end of 2020 the collection featured more than 19,000 titles in over two dozen languages. The archival collections, which already contained disparate comics holdings, launched in earnest in 2011, with the acquisition of writer Chris Claremont's papers.

The circulating holdings contain a diverse collection, with mainstream and alternative titles, archival reprints, independent comics, Kickstarter projects, and other content. These materials have been used in courses from East Asian Languages and Cultures, to English and Comparative Literature, to Narrative Medicine, and have been featured in the American Studies course "The American Graphic Novel." Students have used the collection for term papers, senior theses, and M.A. essays, while faculty have found curricular and research inspiration there.

To suggest a title for the circulating collection, or to donate to the archival collections, please contact Karen Green, the Curator for Comics and Cartoons (see contact info on the right).

Other sites of interest (see also Materials Elsewhere)

- The German Graphic Novel: A site for German language educators who would like to incorporate authentic graphic novels into their courses—and also for those generally interested in the German comics scene.
- Comic Strip Library: High-res scans of a few early 20th-century comic strips
- WESSWeb European Comics and Graphic Novels page: A guide to locating, understanding, and appreciating European comics, from the West European Studies Section of the American Library Association

Curator for Comics and Cartoons

Contact:
 Hg19@columbia.edu
 (212) 653-0429
 6M50 Butler Library
 Rare Book & Manuscript Library

Rare Book and Manuscript Library

<https://guides.library.columbia.edu/graphic-novels>

The circulating holdings contain a diverse collection, with mainstream and alternative titles, archival reprints, independent comics, Kickstarter projects, and other content. These materials have been used in courses from East Asian Languages and Cultures, to English and Comparative Literature, to Narrative Medicine, and have been featured in the American Studies course "The American Graphic Novel." Students have used the collection for term papers, senior theses, and M.A. essays, while faculty have found curricular and research inspiration there.

- 주류 만화와 대안 만화, 메인스트림, 독립만화, 아카이브용 재판, 독립만화, 킥스타터 프로젝트(텀블벅), 그리고 다른 콘텐츠들의 컬렉션이 있음
- <극동아시아 언어와 문화>, <영어학>, <비교문학>, <문학치료> 등의 수업 자료로 활용하고 있으며, <미국 연구> 중 특히 '미국 그래픽 노블' 수업에도 사용하고 있음
- 학생들은 이 컬렉션들을 기말 페이지, 소논문, 석사학위논문 등의 소재로 사용하고 있으며, 교수자들은 수업의 영감과 커리큘럼을 발견하고 있음

: 우리나라는?

1. 웹툰 수집·보존의 필요성

1-2. 휘발성이 강한 디지털 자원

- 2007년 국내최초 유료웹진 플랫폼 <만끽>사이트가 폐쇄. <이끼>등 연재 작품 관련 정보 유실
- 2012년 야후 코리아, 파란의 서비스 종료로 <Yahoo 카툰 세상>, <Paran 카툰> 사라짐
- 2015년 작가들의 컴퓨터가 랜섬웨어 등 악성코드로 웹툰 원고 유실
- 창작자/에이전시/플랫폼 사업자들의 불안감 : 문화적으로, 산업적으로 필요

1-3. 수집·보존 주체의 한계점

- 경기도 부천시 출연기관인 한국만화영상진흥원으로는 수집·보존의 한계가 뚜렷
- 문체부 산하기관에서 운영하던 '문화콘텐츠닷컴' 폐쇄 : 2002년부터 2010년까지 30만건의 문화원형디지털콘텐츠 - 사업예산 지원으로는 한계가 있음
- 그럼 어디에서 ? : 작가들이 신뢰할 수 있는 기관. 원칙이 변화하지 않는 공간

2. 방향성 : 고민의 지점들

2-1. 저작권 관리 문제와 분리

- 저작권 위탁관리와 분리할 필요. 저작권 관련 정보 제공 정도에 그쳐야

2-2. 제도적 수집방안 : 만화진흥법 또는 납본법 개정

- 웹툰식별번호가 있는 고화질 작품
 - ISBN이 아닌 새로운 식별번호가 필요(작가, 제목, 장르, 주제, 연재기간, 연재처...)
 - 한국전자출판협회의 UCI(국가콘텐츠식별체계)의 적절성 ?
 - 대상 : 연재 / 완결 (미완결로 끝날 경우), 원고료를 받는 작품(자기 연재?)
 - 식별번호(작가, 연재처, 연재기간 등) 배부처 : 작가협회 / 플랫폼(인스타 ? 중복?)
 - OCR을 통해 텍스트 정보들을 읽어낼 수 있도록(인물, 장소, 내용 등 파악에 도움)
- 웹툰연재 플랫폼들 : 오아시스
 - 웹툰 작품과 댓글, 광고 등
 - 작품보다는 작품과 독자가 만나는 상태, 그 컨텍스트를 확인할 수 있도록

2. 방향성 : 고민의 지점들

2-3. 적극적 수집

- 대표적인 문제작부터 목록화한 후 순차적으로
- 광범위한 관련 자료들
 - 작가/작가군에 대한 자료(출생, 데뷔, 작품들, 평가, 평론, 인터뷰, 수상...)
 - 작품에 대한 자료(설정 자료들, 시놉시스의 변화, 시나리오, 콘티...)
 - 작품에 대한 2차 자료(연구 및 평론, 매체 이전 ...)

2-4. 활용방안에 대한 최소한의 합의 도출 필요

- 완결 후 유료화한 웹툰 작품들을 무작위로 공개할 수 없음
- 국립중앙도서관의 한정된 기기에서만 작품을 볼 수 있도록(공개 선택한 작품에 한해)
- 각 작품에 대한 정보, 텍스트 정보들은 공개 - 각종 연구들을 위해
- 이 부분이 정리가 되어야, 작가들의 적극적인 참여 보장

미래형 타임캡슐 : 웹 아카이빙(OASIS) 진화와 발전방안



세미나 II

발표 1

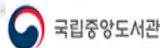
국립중앙도서관 웹자원 아카이빙(OASIS) 현황 및 사례

안경자

(국립중앙도서관 온라인자료과 주무관)

국립중앙도서관 웹자원 아카이빙(OASIS) 현황 및 사례

제59회 전국도서관대회 2022.10.13.(목)



목차

CONTENTS

- 01** 오아시스 개요 : 추진경과, 국제활동, 해외동향
- 02** 오아시스 현황 : 수집, 컬렉션, 시스템
- 03** 오아시스 사례 : 구축사례, 누리집 소개
- 04** 향후 계획

웹자원 아카이빙(OASIS) 개요



OASIS : Online Archiving & Searching Internet Sources

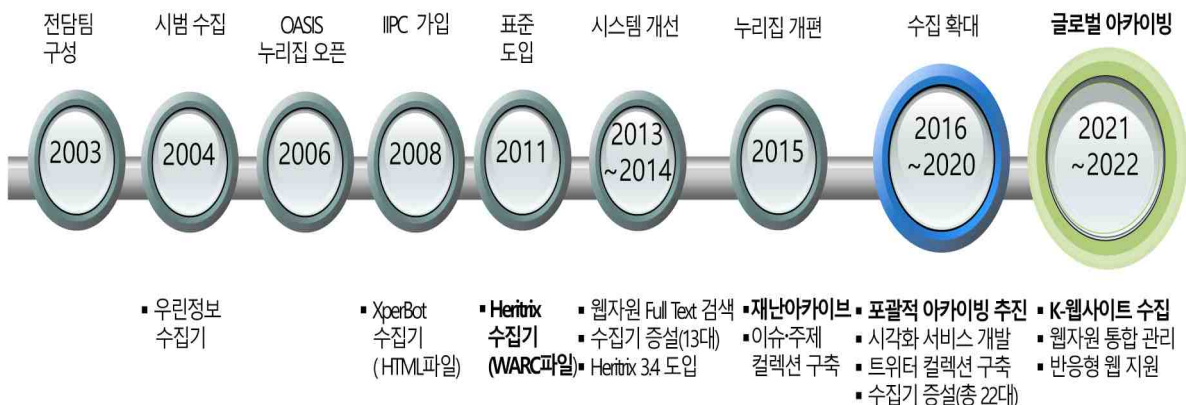
■ 소멸되기 쉬운 공개된 인터넷자원(웹사이트/웹자료)을 국가 디지털 지식문화유산으로 수집·보존하여 미래세대에 제공하기 위한 프로젝트('04년~)

* '03년 10월 유네스코 "디지털유산 보존에 관한 헌장" 발표

관련법규

- 도서관법 제20조 2(온라인 자료의 수집)
 - ① 국립중앙도서관은 대한민국에서 서비스되는 온라인 자료 중에서 보존가치가 높은 온라인 자료를 선정하여 수집·보존하여야 한다.
- 수집대상 온라인자료의 종류(국립중앙도서관 고시 제2011-1호, 제3조)
 1. 웹사이트 2. 웹자료 (문자자료, 음성음향자료, 영상자료, 이미지자료)

추진 경과



국제인터넷보존컨소시엄



INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

- **설 립** : 2003년 7월 / **국립중앙도서관 가입** : 2008년 10월
- **목 적** : 전 세계에서 생산되는 인터넷 정보자원을 수집·보존, 급변하는 정보환경에 대응, 공통의 도구(tool), 기술 및 표준 개발
- **조 직** : 위원장, 부위원장, 프로그램 및 통신책임자, 운영위원회(15개 기관), 워킹그룹, 포트폴리오
- **회 원** : 35개국 53개 기관(국립중앙도서관, LC, BL, BnF, 인터넷 아카이브 등)
* 국가도서관(34), 지역도서관(2), 대학 및 연구도서관(7), 국립문서보관소(2), 비영리재단 등 기타(7)
- **주요 업무**
 - 웹 아카이빙 관련 지식 공유 및 포럼 제공
 - 인터넷 콘텐츠 수집, 보존, 장기적 접근성 제공을 위한 표준 개발
 - 소프트웨어와 도구 개발 및 지원
 - 인터넷 보존 이니셔티브 활동
 - 보존 콘텐츠 활용을 위한 연구 네트워크 조성

국제 협업 컬렉션



INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

- **IIPC 「COVID-19」 등 공개 웹 아카이브 컬렉션 참여, 국내 코로나 관련 웹사이트 정보 제공**

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 outbreak and its containment. High priority subtopics include: coronavirus origins, information about the spread of infection, regional or local containment efforts, 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, Epidemics, Coronavirus infections, COVID-19 Epidemic, COVID-19 (Disease)

Creator: International Internet Preservation Consortium

Collector: International Internet Preservation Consortium

Title: 외교부 해외안전여행 URL: http://0404.go.kr/dev/main.mofa Description: Ministry of Foreign Affairs safe travel overseas Captured once on 9월 7, 2020 Videos: 6 Videos Captured Language: Korean Country of publication: Korea (South) Top-Level Domain: .kr Website Type: Government agencies (National)
Title: 코로나바이러스감염증-19(COVID-19) URL: http://ncov.mohw.go.kr/ Description: Government health website Captured 34 times between 2월 22, 2020 and 10월 26, 2020 Language: Korean Country of publication: Korea (South) Top-Level Domain: .kr Website Type: Government agencies (National)
Title: Coronavirus Infection-19 (COVID-19) URL: http://ncov.mohw.go.kr/index_main.jsp Description: Political aspects Captured 10 times between 2월 21, 2020 and 8월 6, 2020 Language: Korean Country of publication: Korea (South) Top-Level Domain: .kr Website Type: Government agencies (National)
Title: South Korean Ministry of Foreign Affairs - Latest safety news URL: http://www.0404.go.kr/dev/newest_list.mofa Description: Information about travel restrictions affecting South Koreans Captured 6 times between 3월 19, 2020 and 10월 27, 2020 Language: Korean Country of publication: Korea (South) Top-Level Domain: .kr Website Type: Government agencies (National)

해외 동향

OASIS



오아시스 현황

OASIS

- 선택적 수집과 포괄적 수집 병행
글로벌 K-웹사이트(2022~)

- 국가 재난 아카이브, 주제·이슈 컬렉션 구축
「코로나-19」컬렉션 구축(2020~)



- 수집 대상 및 컬렉션 선정, 가이드라인 자문
웹사이트 수집 지침 개정(2020)

- 태그 클라우드, 웹트렌드 분석 서비스 개발

수집 현황



■ 웹사이트 886,088건 / 용량 770TB (2021. 12월 기준)

웹사이트 수집

구분	'04~'10	'11~'15	2016	2017	2018	2019	2020	2021	계
선택적 수집	21,054	128,422	23,164	10,357	16,383	16,058	5,001	8,263	236,874
포괄적 수집	-	-	52,830	142,028	110,000	104,000	120,181	120,175	649,214
계	40,372	20,464	75,994	152,385	126,383	120,058	125,182	128,438	886,088

컬렉션 구축

재난·행사·주제	웹사이트	웹문서	이미지	동영상, 사운드	계
「코로나19」 등 154건	78,812	9,297	44,206	16,809	149,124

수집 지침



기본 방향

대한민국 국가도메인 웹사이트 전체 대상으로 포괄적 수집,
개별 웹사이트, 국가재난·행사, 주제 및 이슈 관련 웹자원 선택적으로 수집·보존

수집 대상

- 대한민국 국가도메인 웹사이트 및 한국 관련 해외 웹사이트
- 웹자료(납본대상이 아닌 문서자료, 동영상, 이미지, 음향 등 멀티미디어 자료)
- 정부부처, 지방자치단체, 공공기관 블로그 및 트위터 등 주요 소셜미디어
- 데이터세트, 신규 웹자원 및 소셜플랫폼 등

수집 기준

신뢰성, 유일성, 적합성, 유용성, 가치성, 대중성, 언어, 시급성, 접근 용이성 고려

수집 제외

- 원본 콘텐츠가 아닌 단순 편집본으로 보존 가치가 없다고 판단되는 웹자원
- 기술적으로 수집 또는 보존할 수 없는 웹자원

공개 및 보존

- 수집 자원은 영구 보존되며, 국립중앙도서관내 이용 자료로 제공
- 저작권자의 이용 동의 절차를 거쳐 관외 이용 자료로 제공

수집 방법



구분	포괄적 수집('16년~)	선택적 수집('04년~)
목적	▲당대 지식문화유산의 누락 없이 전승	▲국가 이슈나 테마에 대해서 체계적 수집
방법	▲'.kr', '.한국' 등 국가 도메인 대규모 수집	▲수집 기준에 의한 선별 수집
	▲연1회, 깊이 3depth 수집	▲사이트별 수집 주기(분기별, 월별, 일별 등), 깊이 (1~5depth) 지정
	▲메타데이터 구축 안함, 자동색인	▲검색용 메타데이터 구축
서비스	▲관내 서비스	▲관내 서비스 또는 공개(이용동의서 받은 웹사이트)

선택적 수집



개별 사이트

- 국가기관, 공공기관, 학회, 교육 · 연구기관
- 민간단체, 시민사회단체 및 비영리조직
- 상업기관, 언론기관
- 개인 홈페이지, 소셜 플랫폼
- 한국 관련 해외 웹사이트
- 방문수 많은 웹사이트

수집 제외

편집 사이트나 원본 콘텐츠가 아닌 웹사이트
기술적으로 수집 및 보존이 불가능한 웹사이트
(예시 : Javascript Code 인식 오류 등)

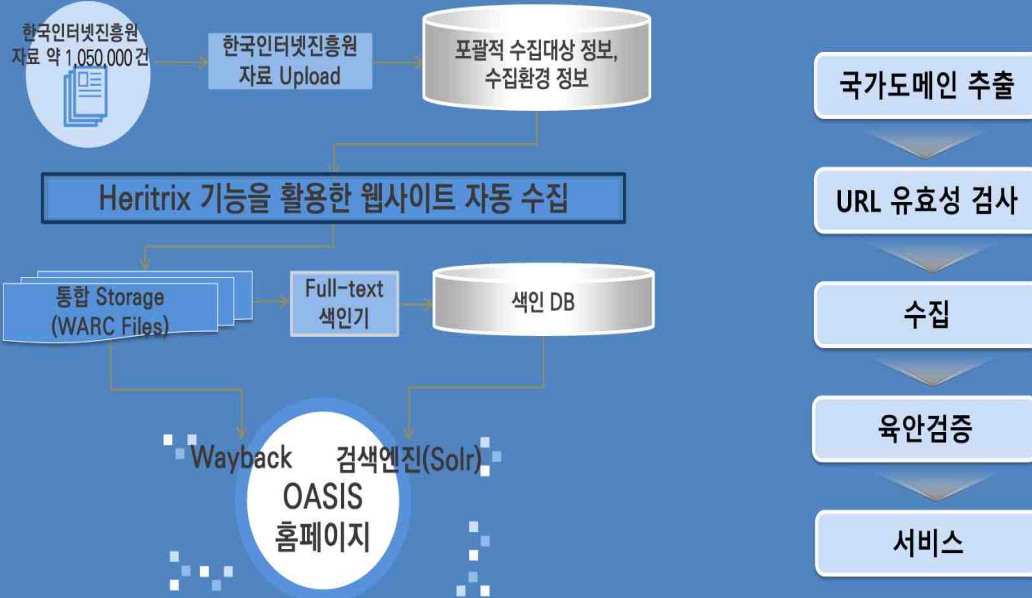
국가적 재난·행사·주제 컬렉션

- 국가 재난 관련 웹사이트
- 선거, 올림픽, 국제행사 등 사회문화적으로 중요한 이슈 관련 웹사이트
- 인문·사회·자연과학 및 기술 주제 분야 관련 웹사이트
- 전문가 그룹 추천 웹사이트

수집 유형

웹사이트, 웹페이지, 뉴스기사
카드뉴스, 보도자료 등 웹자료

포괄적 수집



국가재난아카이브

- 「코로나-19 컬렉션」 등 국내 주요 40건 재난 사건에 대한 8만건의 재난 기록 보유
- 국민의 생명·신체·재산과 국가에 피해를 주거나 줄 수 있는 것으로서 **자연재난과 사회재난** 웹자원 수집
- 「재난 및 안전관리 기본법(재난안전법)」에 따라 **특별재난**으로 선포된 재난과 재해, 이 밖에 **정치·사회·문화적 파급력이 큰 재난 및 재해** 관련 웹자원 수집



1995-06-29



2003-02-18



2014-04-16



2015-05-20



2020-01-20

주제·이슈 컬렉션

• 국가적으로 중요한 행사

- 선거, 국제회의, 남북정상회담, 올림픽, 월드컵 등

• 정치·사회·문화적으로 큰 반향을 일으킨 사건

- 한·일 위안부 협상 타결, 가슴기 살균제 사건 등

• 인문·사회·자연과학 및 기술 분야 테마 컬렉션

- 만화에서 웹툰까지, 미세먼지, 4차 산업혁명 등



2021/22컬렉션



시스템 현황



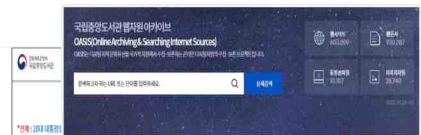
DB구축시스템

- 웹사이트 수집
(사이트명, URL, 수집주기, 수집깊이)
- 자동 수집(Heritrix 22대)
- 상세 메타데이터 입력
(KCR 4판, MODS 3.5, KDC 6판)

번호	URL	상태
1	www.kci.go.kr	정상
2	www.kci.go.kr	정상
3	www.kci.go.kr	정상
4	www.kci.go.kr	정상
5	www.kci.go.kr	정상
6	www.kci.go.kr	정상
7	www.kci.go.kr	정상
8	www.kci.go.kr	정상
9	www.kci.go.kr	정상
10	www.kci.go.kr	정상

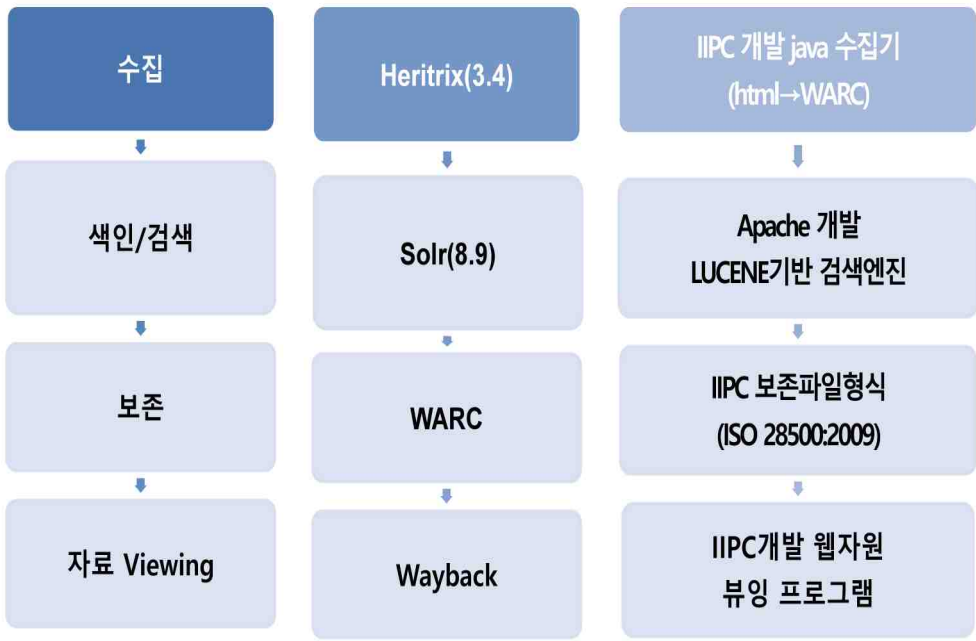
KOLIS III

- 등록
- 검증
- 색인 생성
- 서비스 이관

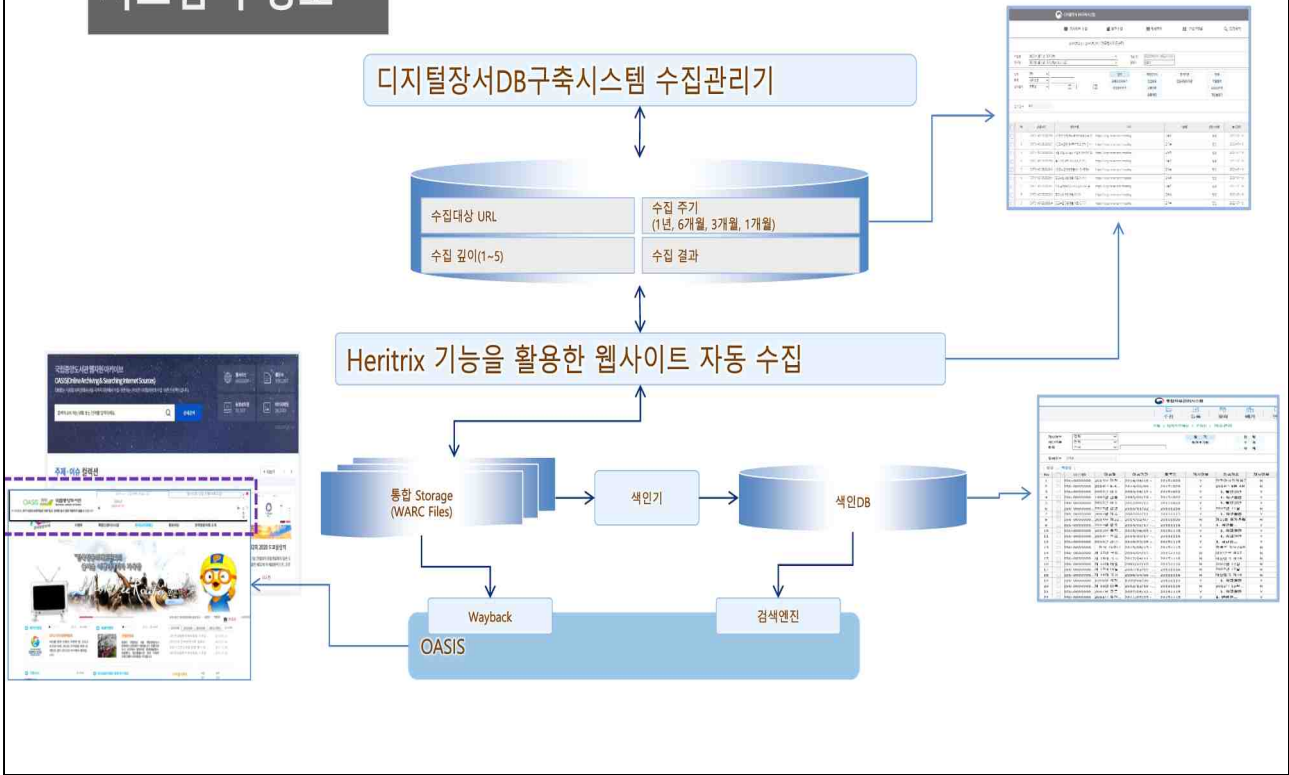


국립중앙도서관 &
오아시스
누리집 서비스

시스템 도구



시스템 구성도



수집 사례



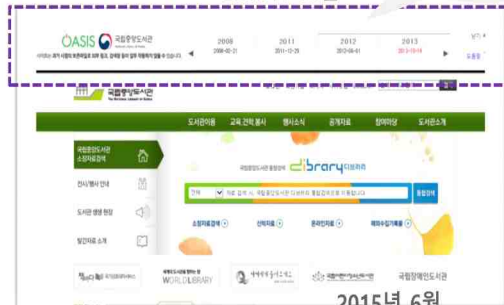
수집 사례

WAYBACK

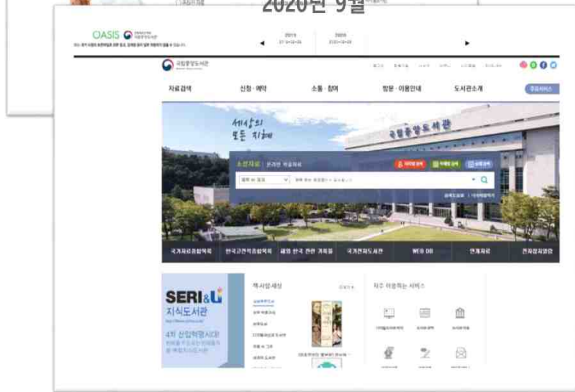
2011년 12월



2013년 10월



2020년 9월



2015년 6월

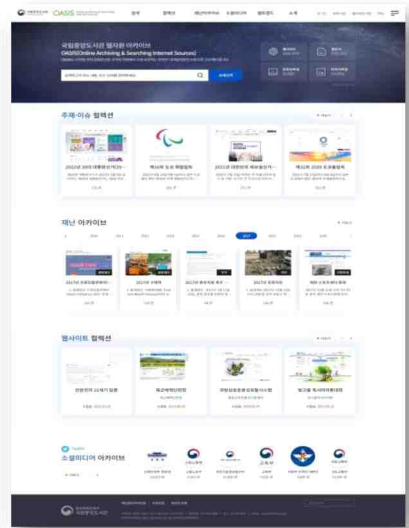


오아시스 누리집 nl.go.kr/oasis

2006 - 2015

2016 - 2021

2022 -



오아시스 누리집 nl.go.kr/oasis

검색

- 통합검색, KDC주제 별보기, 초성보기
- 본문검색(Full-Text)

주제·이슈 컬렉션

재난아카이브

웹사이트 컬렉션

소셜미디어 아카이브

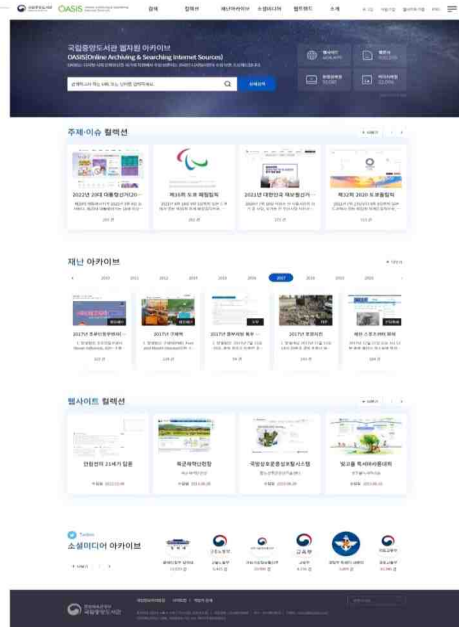
- 트위터

웹트렌드

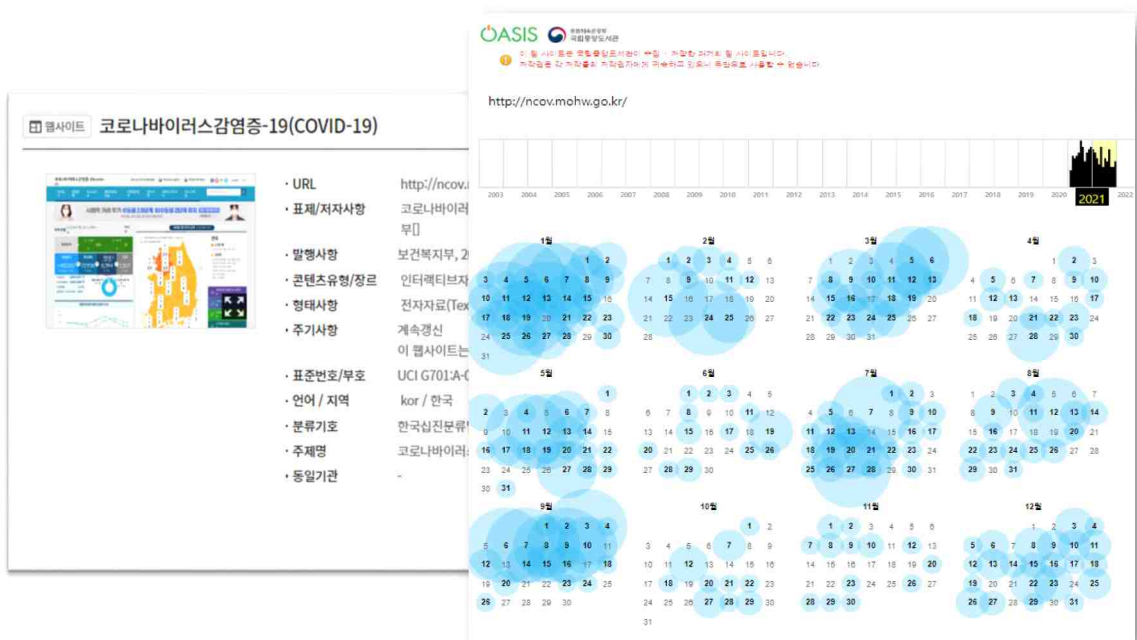
- 웹트렌드, 태그클라우드

소개

- 개요, 통계, 자료실, 공지사항, 자주묻는질문, 웹사이트 기증



웹사이트 변경기록 MEMENTO



웹트렌드 / 태그클라우드



웹사이트 기증 / 이용 동의

웹사이트 기증안내

기관이나 개인이 운영하는 웹사이트를 국가 지식문화유산으로 영구 보존하고 싶다면 오아시스에 기증 신청할 수 있습니다.

- 웹사이트 기증은 소유권 제공이 아니라 국립중앙도서관에서 수집 보존
- 신청해주시는 웹사이트는 수집 로봇에 의해 자동 수집 후, 해당 정보를 구
- 단, 기술적으로 수집이 불가능한 경우, 별도 안내함

기증대상 | 국내 기관 또는 개인이 운영하고 있는 웹사이트

신청방법 | 온라인신청

당면 과제



- (진본성) 원본과 편집본, 상업용 웹자원에 대한 가치 판단
- (보 안) 악성코드 감염 우려, 개인정보 등 공개 제한
- (품 질) 진화하는 웹사이트 수집, 무결성 한계
- (시스템) 스토리지, 서버 과부하, 웨이백 서비스 불안정
- (협 업) 정책과 기술의 합심 필요

“현 시대의 가장 복잡한 웹 정보자원”

향후 계획



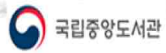
- (아카이빙 확장) 선택적 수집과 포괄적 수집 병행, 국내외 K-웹자원 수집
- (웹자원 통합) 효율적인 웹자원 수집 · 정리 · 서비스 체계 구축
- (컬렉션 확대) 기관별 · 주제별 웹 컬렉션 확대
- (시스템 확충) 수집부터 서비스 개선까지 예산 확보를 통한 시스템 고도화
- (빅데이터 제공) 방대한 웹자원 데이터 활용 서비스 개발 및 제공

“국가 웹 아카이브 서비스 구현”



감사합니다.

국립중앙도서관 온라인자료과



세미나 II

발표 2

국가 디지털 아카이브 구축방안 '코로나19 아카이브'

김학래

(중앙대학교 문헌정보학과 교수)

국가 디지털 아카이브 구축방안 '코로나19 아카이브'

김학래 | 중앙대학교


목차

1. 웹 아카이브·재난 아카이브의 개요
2. 코로나19 관련 디지털 아카이브 현황
3. 사례 연구 - 공공데이터 | 코로나19:우리의 기억
4. 코로나19 아카이브 구축 방안
5. 정리하며

소개


HIKE · Team.Cayley | 중앙대학교

2020




무엇을 지정하는가
COVID-19 관련 데이터의 수집과 분석을 위한
주요 과제를 지정한다.

팀, 케일리 구성
코로나19 관련 데이터 수집과 공유




정부혁신해커톤 커뮤니티상 수상
행정안전부

2021



COVID19 데이터 분석 A to Z
OPEN DATA DAY IN KOREA 2021
2021.3.6 (월) 14:00-17:00

International Open Data Day 개최
행정안전부 후원




코로나 데이터, 타임랩스에 담는다...유사 위기 대비
144억 코로나19 데이터를 시간여행에 담은 승리
주최: 행정안전부, 한국인터넷진흥원, 한국과학기술정보연구원

코로나19 데이터 타임랩스에 담는다...유사 위기 대비
주최: 행정안전부, 한국인터넷진흥원, 한국과학기술정보연구원


코로나19 공공기관 데이터 보유 전수조사
4차산업혁명위원회

2022




covid19-cluster-infection

코로나19 집단감염 데이터 구축·공유
KT·빌&벨린다 재단



‘코로나19: 우리의 기억’ 아카이브 기증
국립중앙도서관



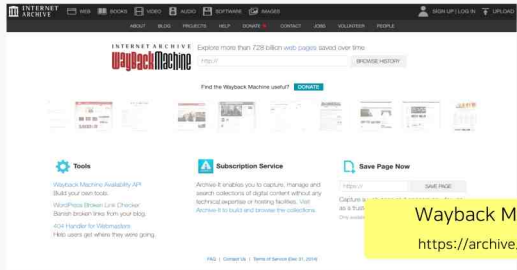
Hyper Interlinked Knowledge Engineering Lab.

3

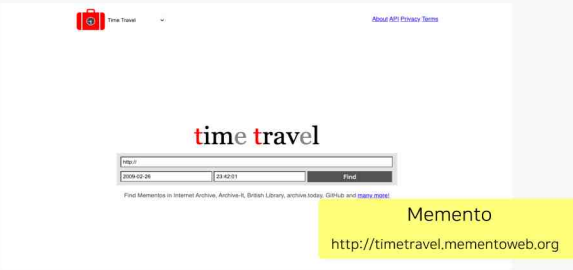
개요

웹 아카이브

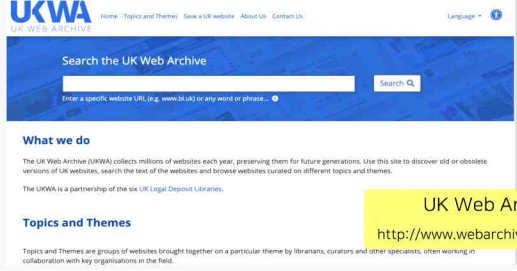
특정 웹사이트를 영구히 저장하여 재현하는 서비스



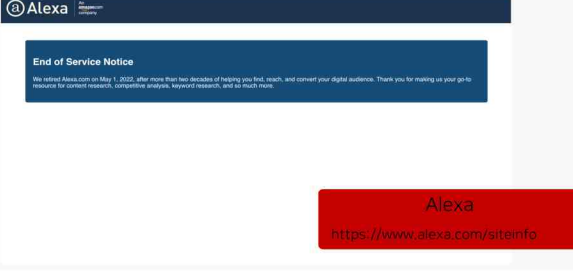
Wayback Machine
<https://archive.org/web>




Memento
<http://timetravel.mementoweb.org>



UK Web Archive
<http://www.webarchive.org.uk/ukwa>



Alexa
<https://www.alexa.com/siteinfo>



Hyper Interlinked Knowledge Engineering Lab.

4

개요

재난 아카이브 - 해외 사례



The September 11 Digital Archive

<https://911digitalarchive.org/>

- 2001년 9월 11일, 뉴욕, 버지니아, 펜실베이니아에서 발생한 테러의 기억과 기록을 수집, 보존하기 위한 참여형 아카이브

- City University of New York Graduate Center, Roy Rosenzweig Center for History & New Media, Alfred Foundation 등 다수 기관 참여

- 이야기, 사진, 디지털 아트 이미지, 이메일, 음성 메일, 비디오, 애니메이션, 공식 문서, 구전 역사 등 약 15만 개의 디지털 파일 축적



HINAGIKU

<https://kn.ndl.go.jp/#/>

- 동일본 대지진 (1995년)과 관련된 디지털 데이터를 검색·활용할 수 있는 디지털 아카이브

- 정부, 지자체, 대학, 도서관, 연구기관이 보유·관리하는 동일본대지진의 기록을 국가 차원에서 수집·보존·공개

- 동일본 대지진과 관련된 음성, 동영상, 사진, 웹정보를 포괄적으로 수집하고, 학술연구와 교육에 활용할 수 있도록 제공

HINAGIKU: Hybrid Infrastructure for National Archive of the Great East Japan Earthquake and Innovative Knowledge Utilization

Hyper Interlinked Knowledge Engineering Lab.

코로나19

코로나19 데이터 타임캡슐 - 4차산업혁명위원회 (2021)

코로나19로 인한 국가사회 전반의 영향과 변화를 기록·저장하기 위한 보존·활용체계 수립

- 327개 기관에서 총8,298건의 데이터 목록 식별 (21년 5월 ~ 10월, HIKE 연구실)
- 코로나19 데이터 타임캡슐 구축을 위한 빅데이터 플랫폼 사업 추진 (22년, 보건복지부)



Source: <https://tinyurl.com/6xzprw8>

Hyper Interlinked Knowledge Engineering Lab.

코로나19

오아시스 (OASIS) - Online Archiving & Searching Internet Sources

<http://www.oasis.go.kr>

국립중앙도서관이 2004년부터 온라인 디지털 정보자원을 수집·보존하고 있는 프로젝트

- 포괄적 수집과 선택적 수집을 병행하며 개별 사이트와 국가적으로 중요한 이슈·재난 분야의 디지털 기록을 수집 . 웹사이트 886,088건, 파일 3,351,150개, 351TB 규모의 데이터 수집 (2021년 12월)
- 웹 아카이브를 활용한 웹 키워드 분석, 웹 트렌드 서비스 제공
- 국제인터넷보존컨소시엄 (IIPC, International Internet Preservation Consortium) 가입 (2008년)
- 재난 아카이브는 세월호 침몰사고 (2014), 메르스 유행 (2015), 강원도 산불 (2019)에 대한 재난 기록을 보존하고 있음
- 코로나19 관련 자료: 2020년 3월 이후 현재 359건의 자료를 수집하여 제공하고 있음



"코로나19 발생부터 감염 확산, 확산 방지를 위한 노력, 의학·과학·사회·경제적 양상 등을 다룬 정부 기관과 관련 기관·단체의 인터넷 웹사이트와 웹 문서, 동영상, 이미지와 같은 웹 자료 등이 수집된다.
코로나19는 현재 진행 중이므로 웹 정보자원은 16일부터 종료 시까지 지속해서 업로드한다. 또 국립중앙도서관은 2008년부터 회원기관으로 참여하는 국제인터넷보존컨소시엄(IIPC)의 '코로나바이러스 감염증-19' 국가별 웹 아카이브 컬렉션 구축 협력 프로젝트에도 참여한다. 국립중앙도서관은 IIPC 동참 요청으로 오아시스 재난 아카이브에 구축될 코로나 감염 확산과 대응에 관한 우리나라 웹사이트 정보를 제공한다."

코로나19 발생부터 소멸까지 모든 것 기록으로 남긴다, 연합뉴스, 2020년 3월 10일

Source: https://img5.yna.co.kr/etc/inner/KR/2020/03/10/AKR20200310052000005_02_i_P4.jpg

Hyper Interlinked Knowledge Engineering Lab.

7

코로나19

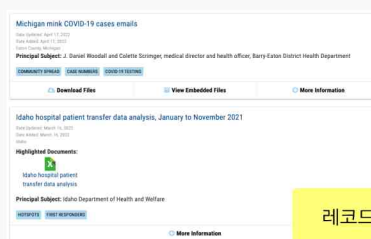
Documenting COVID-19 (<http://documentingcovid19.io>)

COVID-19과 관련된 검색이 가능한 문서의 저장소

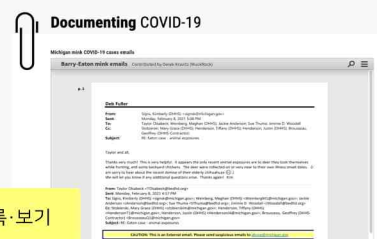
- Columbia University의 Brown Institute for Media Innovation에서 시작된 협력적 오픈 레코드 프로젝트
- FRONTLINE PBS, Columbia Journalism School's Stabile Center for Investigative Journalism, the Fund for Investigative Journalism, National Geographic에서 지원하고 있음
- 주정부 공개기록법·정보 자유법 (Freedom of Information Act)를 근거로 수집된 291개 레코드세트 보유 (2022년 4월)
- The Algorithms 프로젝트, The Examiners Project 프로젝트에서 활용 (John S. Knight Journalism Fellowships (Stanford 대학)에서 지원)



수집 데이터 현황



레코드목록·보기



Source: <https://tinyurl.com/6x7zprw8>

Hyper Interlinked Knowledge Engineering Lab.

8

코로나19

“Documenting COVID-19 at” 프로젝트

세계 각국의 중앙정부, 지방자치단체, 대학 도서관, 커뮤니티에서 코로나-19 관련 자료를 기록하고 있음

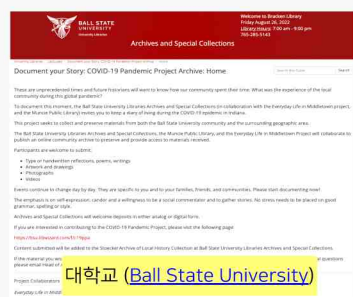
- 구글에서 “Documenting COVID-19”로 구글 검색: 37,800개 검색 결과가 나타남
- 기관에 따라 운영 방식에 차이가 있지만, “Documenting COVID-19” 유형의 웹사이트는 기부, 참여 방식으로 운영
- 사진·이미지, 일기·저널·블로그·에세이, 비디오·오디오, 구술자료, 코멘트, 인터뷰, 노래, 예술작품, 사인, 시 (poetry) 등 다양한 유형의 자료를 수집하고 있음
- 개인, 전문가, 산학연 수준의 다양한 참여가 있고, 주로 이메일, 구글폼을 이용해서 데이터를 제출하고 있음



중앙정부 (Canada)



지방정부 (Missoula)



대학교 (Ball State University)

Source: <https://tinyurl.com/6x7zprw8>

Hyper Interlinked Knowledge Engineering Lab.

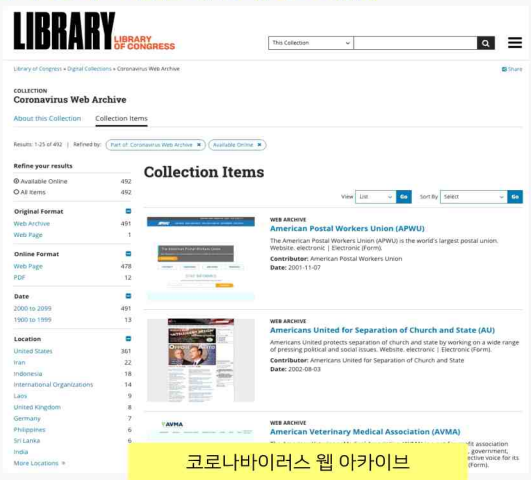
코로나19

Coronavirus Web Archive - Library of Congress

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

코로나-19 대유행이 진행되는 동안 다양한 분야의 웹 문서를 수집하고 일반에 공개하고 있음

- 정부, 과학, 비즈니스, 정책 등 450개의 웹 아카이브가 포함됨
- 2020년 6월에 공식적으로 수집 계획 수립 (현재 지속되고 있음)
- 도서관 전체에서 다양한 기술, 관점, 주제 전문성을 고려하여 10명의 담당자를 선정하여 팀을 구성 (현재 8명)
- 컬렉션의 수집은 대부분 weekly 로 진행 (매일 업데이트 없음)
- 다양한 아카이브를 포함하고 있음: Business in America Web Archive, State Government Websites of the United States Web Archive, General News on the Internet Web Archive, Professional Organizations for Performing Arts Web Archive, Public Policy Topics Web Archive.
- 수집 언어: English, with Burmese, French, German, Greek, Hindi, Indonesian, Italian, Japanese, Khmer, Korean, Lao, Malay, Malayalam, Marathi, Persian, Portuguese, Russian, Sinhala/Sinhalese, Spanish, Tagalog, Thai, Turkish



코로나바이러스 웹 아카이브

Hyper Interlinked Knowledge Engineering Lab.

코로나19

IIPC CDG Collection (<https://netpreserve.org/projects/collaborative-collections/>)

IIPC 회원과 공동으로 주요 주제에 대한 컬렉션 구축

- 공개 웹 아카이브 컬렉션을 구축하고 Archive-It을 통해 서비스
- IIPC 회원들의 높은 관심, 특정 구성원에 책임 부가 배제, 여러 기관의 폭넓은 관점, 국가를 초월하는 공동의 주제
- [Research Working Group](#)과 아카이브 활용을 위해 협력

주요 컬렉션

컬렉션 제목	최종수집일자	웹사이트	데이터 규모
National Olympic and Paralympic Committees	2019년 5월 30일	279	1020.2GB
Online News Around the World	2019년 11월 12일	1,493	1.8TB
Climate Change	2019년 12월 2일	769	1.2TB
Intergovernmental Organizations	2020년 2월 5일	394	2.7TB
Artificial Intelligence	2020년 5월 1일	388	643.7GB
Novel Coronavirus (COVID-19)	2020년 6월 10일	7,822	2.9 TB

Source: <https://netpreserveblog.wordpress.com/2020/07/01/iipc-content-development-groups-activities-2019-2020/>
Hyper Interlinked Knowledge Engineering Lab.

The screenshot shows the IIPC website interface. At the top, there are navigation links: HOME, EXPLORE, LEARN MORE, CONTACT US. Below that, the IIPC logo and name are displayed. A search bar is visible with the text 'Enter search terms here'. The main content area shows search results for '2010 Winter Olympics' and '2012 Summer Olympics', including their descriptions and collection dates.

코로나19

IIPC Novel Coronavirus (COVID-19) (<https://archive-it.org/collections/13529>)

COVID-19에 대한 웹 아카이브를 Archive-It에 공개

- 2020년 2월 초에 컬렉션 구성 (2월 21일 최초 크롤링 시작함)
- 30개 이상의 IIPC 회원사, 100명 이상의 개인/기관 담당자 참여
 - . 주제: 감염확산, 의학적·과학적 정보, 백신개발과 예방접종, 사회경제적 이벤트
 - . 약 8,000개 이상의 시드 (웹사이트)를 대상으로 크롤링 수행
 - . 수집된 데이터는 약 2.9 TB 규모 (2020년 6월 11일 기준)
- 웹사이트 유형
 - 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), University (383), Medical research (212), Cultural (210), Pandemic data tracking/mapping (127), Intergovernmental (110), Religious (74), Wiki (66), Military (32)
- 국가별 현황: 127개 국가, 45개 언어를 포함
 - United States (1986), Brazil (1318), Argentina (1253), Peru (1243), Uruguay (1110), Chile (1017), Bolivia (601), France (591), Paraguay (547), Portugal (521), Spain (432), Netherlands (422), Singapore (399), Germany (357), Serbia (305)

The screenshot shows the IIPC website interface for the 'Novel Coronavirus (COVID-19)' collection. It features a search bar and a list of search results. One result is highlighted, showing the title 'Title: 외교부 해외안전여행', URL 'http://0404.go.kr/dev/main.mofa', and description 'Description: Ministry of Foreign Affairs safe travel overseas'. The page also includes navigation links and a 'Narrow Your Results' section.

코로나19

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

웹 아카이빙 기관과 연구자의 네트워크

- 웹 아카이브의 디지털 문화유산의 활용과 이해를 돕기 위한 목적
- 2020년-2023년까지 활동, 매해 정기미팅을 통해 결과 공유
 - . Independent Research Fund Denmark | Humanities (grant no 9055-00005B)
 - . 주요 연구자: Niels Brügger, Valérie Schafer, Jane Winters

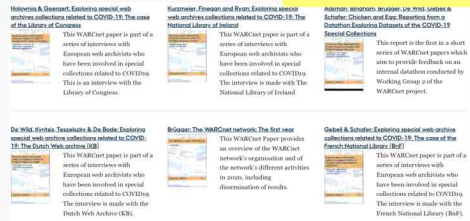
- WG2: Analysing transnational events

- COVID-19 과 관련된 웹 아카이브의 구축과 테스트베드 구축
- British Library, BnF (National Library of France), INA (the French National Audiovisual Institute) in France, the BnL (National Library of Luxembourg), IIPC가 서로 협업해서 진행
- 특별 컬렉션의 선택과 큐레이션 프로세스와 범위에 대한 기술 인터뷰 추진
 - INA, the BnF, the BnL, the IIPC, Netarkivet (Denmark), the National Széchényi Library in Hungary, the UK Web Archive, the Swiss National Library, the National Library of the Netherlands (KB) and the Icelandic Web Archive

Working Groups 현황

- Working Group 1: Comparing entire web domains
- Working Group 2: Analysing transnational events
- Working Group 3: Digital research methods and tools
- Working Group 4: Research Data Management across borders
- Working Group 5: The WARCnet Code Book of web archive data formats
- Working Group 6: Post WARCnet

WARCnet 보고서



WARCnet 프로젝트

주요 국가별 COVID-19 아카이브

WARCnet 보고서의 국가별 주요 COVID-19 아카이브의 구축 방안, 운영 현황에 대한 분석

- 스페셜 컬렉션의 현황, 접근성과 검색가능성, 파트너쉽과 활용으로 구분하고 있음
- Scope
 - . What did you collect? Websites? social networks?
- Frame
 - . When did you start? When did/do you plan to stop? What was the frequency?
 - . How was quality control done on the collection?
 - . Can you give me a quantitative overview of the data?
- Accessibility and Searchability
 - . Can we talk about the accessibility and searchability of these data?
- Partnerships and Uses
 - . Are researchers already asking you about the COVID-19 collection and wanting to analyse it?
 - . How did you communicate about this special collection? Did you have some partnerships? (with local stakeholders, with Archive-It? With the IIPC?
 - . What were issues, challenges or limits of this collection activity?
 - . How do you archive nationally something which is fundamentally global?

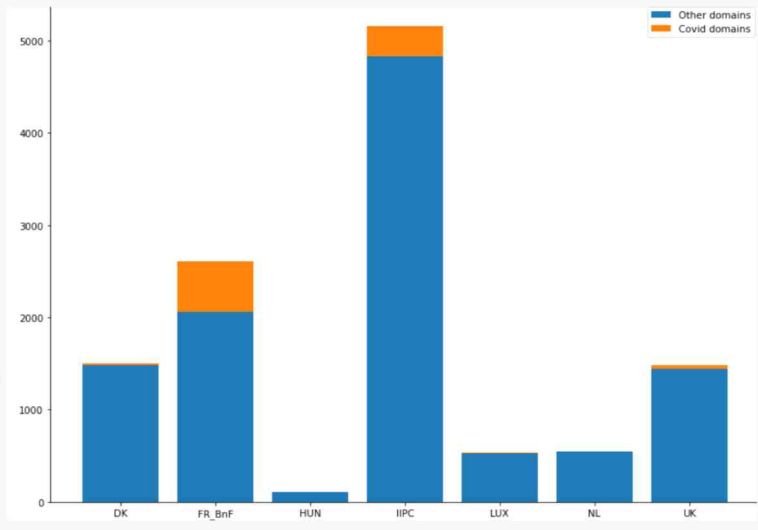
아카이브 특징 분석과 요약

WARCnet 프로젝트

국가별 웹 아카이브 - 100% COVID-19 웹사이트

코로나19 관련 웹 아카이브의 유형

- 'corona', 'korona', 'covid'를 포함하는 아카이브 자료에서 유일한 도메인명 검토
- 100% COVID-19 웹사이트: COVID-19 때문에 새롭게 만든 만든 웹사이트
- 이미 존재하는 웹사이트: COVID-19 관련 콘텐츠를 웹사이트에서 포함
- 국가별 아카이브에서 100% COVID-19 웹사이트를 구축하는 사례가 많지 않음
- 프랑스 BnF, 영국 British Library는 100% COVID-19 웹사이트에서 콘텐츠 수집
- 대부분의 COVID-19 자료는 기존 웹사이트를 통해 제공되고 있음



Source: <https://netpreserveblog.files.wordpress.com/2021/11/awac2-table2.png>

WARCnet 프로젝트

국가별 웹 아카이브 - 품질관리와 사용 관점

품질관리에 비해 수집 콘텐츠의 사용에 대한 계획 미흡

- 수집하는 웹사이트에 따라 품질관리 방법은 다름
- 모니터링 도구와 품질 관리를 위한 프로세스를 갖춘 아카이브의 품질이 상대적으로 높음 (LoC, INA 등)
- 샘플을 추출하고 큐레이터가 수작업을 통해 진행 (덴마크)
- 인력, 기술적 지원의 한계로 품질관리가 진행되기 어려움
- 특히, 소셜 미디어에 대한 수집과 이에 대한 품질 관리는 현존하는 소프트웨어로 해결하는 데 한계가 있음
- 수집 콘텐츠의 활용은 대부분의 아카이브에서 공통적인 문제로 지적하고 있음
- 수집된 콘텐츠를 활용하는 사례가 매우 적고, 특정한 연구그룹과 협력 또는 도서관 자체적으로 활용 방안을 검토하고 있음



Reality Check

아카이브 활용의 제한

대부분의 웹 아카이브는 저작권 이슈로 수집된 자료의 열람을 제한하고 있음

- 조사대상 중 Library of Congress만 일반 대중에게 자료 공개를 하고, 대부분 reading room에서 접근 허용

The screenshot shows a search result for 'Screen Beer: You'll never drink alone!'. The metadata window indicates that access to this publication is limited. A red dashed arrow points from the search results to the metadata window, highlighting the access restriction.

Covid-19: Documenting a crisis

접근제한으로 스냅샷을 열람하지 못함

Reality Check

아카이브 활용의 제한

수집된 콘텐츠에 접근하는 데 기술적 제한이 있음

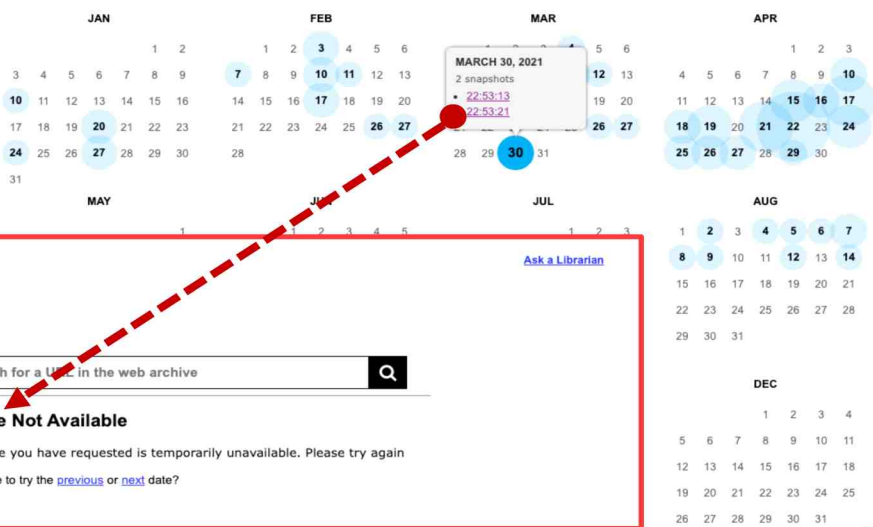
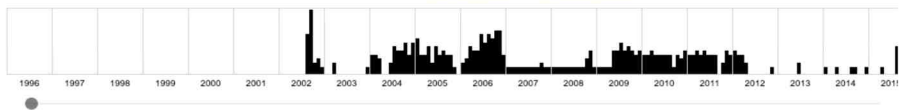


WEB ARCHIVES

Ask a Librarian

Search for a URL in the web archive

http://www.who.int/en
Saved 662 times between August 7, 2002 and August 14, 2021.



사례연구

지방자치단체의 데이터 관리실태

지방자치단체에서 제공하는 데이터의 관리·활용에 대한 분석

문제 정의

- 지방자치단체는 데이터를 효과적으로 관리하고 있는가?
- 사용자는 제공한 데이터를 활용하고 있는가?

연구 방법

- 공공데이터포털에 제공된 데이터세트 수집
- 데이터 관리 정책의 준수 여부 분석
- 데이터 활용 지수 계산
- 지방자치단체의 특징 분석

Considerations in Releasing Public Data: The Case of Local Governments in Korea

Chaeun Song¹ and Hakke Kim¹

Abstract
Local governments play a very important role in providing and disseminating public data, with the assistance of the national government. They also strive to promote effective policies for accessing this data. However, local governments face challenges in independently promoting public data policies owing to budget, manpower, and technical constraints. This study analyses public data provided by local governments in South Korea from the perspective of management and use, apart from suggesting considerations for the public data policies of local governments. Public data provided by local governments were also collected, and both the data management method and use of such data were evaluated. Data management measures the currentness in complying with the update policy for each dataset, and the data utilization measures the relationship between view and download as usefulness. A clustering analysis was conducted to analyse the common characteristics of individual local governments. According to the research results, most local governments do not systematically manage their datasets, and the use of the data provided is extremely low. It is therefore necessary to establish an effective data policy that considers the characteristics of local governments.

Keywords
Public data, Government data, Local government, Data use

Introduction

The disclosure of public data continues to attract the attention of governments across the globe as an effort to enhance transparency [1, 2], and is an important national resource for adding economic value [3, 4, 5]. The increasing importance of the data economy has resulted in a corresponding rise in the importance of public data in creating an ecosystem for data-related industries. Governments have recently established national data strategies to prepare for future industries, including artificial intelligence and autonomous vehicles [6, 7].

The Korean government has continuously provided public data since 2013 [8]. According to the 'Act Promoting the Provisioning and Use of Public Data' [9], the scope of such data is broadly defined as all data or information created by public institutions and public-sector information leading to communication and cooperation among all citizens [9]. The initial goal of the government was data security. However, as data are now recognised as a key resource for realising a data-driven economy, the South Korean government is strengthening its support to maximise the usage of public data in various industries. The government has proposed spending approximately 562 billion USD over the next five years on the Korean Digital New Deal, which aims to establish foundations for a digital economy [10]. In particular, the Digital New Deal will address data usability by creating a 'data deal' used to collect data from both public and private sources and then standardise the data such that it can be properly analysed [11]. A proposal to utilize 549.2 million USD to establish 15 new big data platforms has been also created by the South Korean government. In addition to facial recognition, voice, and carrier datasets related to artificial intelligence, the Digital New Deal will also unveil

datasets essential to our daily lives, including maps, weather, and traffic information as public data [12].

As entities in a public data policy, public institutions play a key role in the provisioning and management of data. The term 'public institutions' refers to 'any state agency, local government, or public institution', as defined in subparagraph 19 of Article 3 of the Framework Act on National Informationization [13]. Local governments can provide data to solve social problems based on regional characteristics and changing circumstances. In fact, local governments in South Korea provide approximately 60% of the public data through a public data portal [5, 14]. However, complaints from users regarding public data are on the rise because only fragmented or heavily redacted data have been released, rather than the raw data [15].

Many studies have been conducted to improve the data policy of local governments or to assess public data at both the local and national levels [16, 17, 18]. Such studies can be divided into two main categories. First, studies related to the public data of local governments tend to focus on the format and functional aspects of the data portals [19, 20, 21]. This approach is effective in comprehensively reviewing the data policies of local governments from the perspective of such portals. Second, the data policies of local governments tend

¹Department of Library and Information Science, Chung-Ang University
Corresponding author:
Hakke Kim, Chung-Ang University, 84 Heukseom-ro, Donggu-gu, Seoul, Korea
E-mail: hakke@cau.ac.kr
https://orcid.org/0000-0001-2023-0316/
10.1016/j.jis.2022.101111

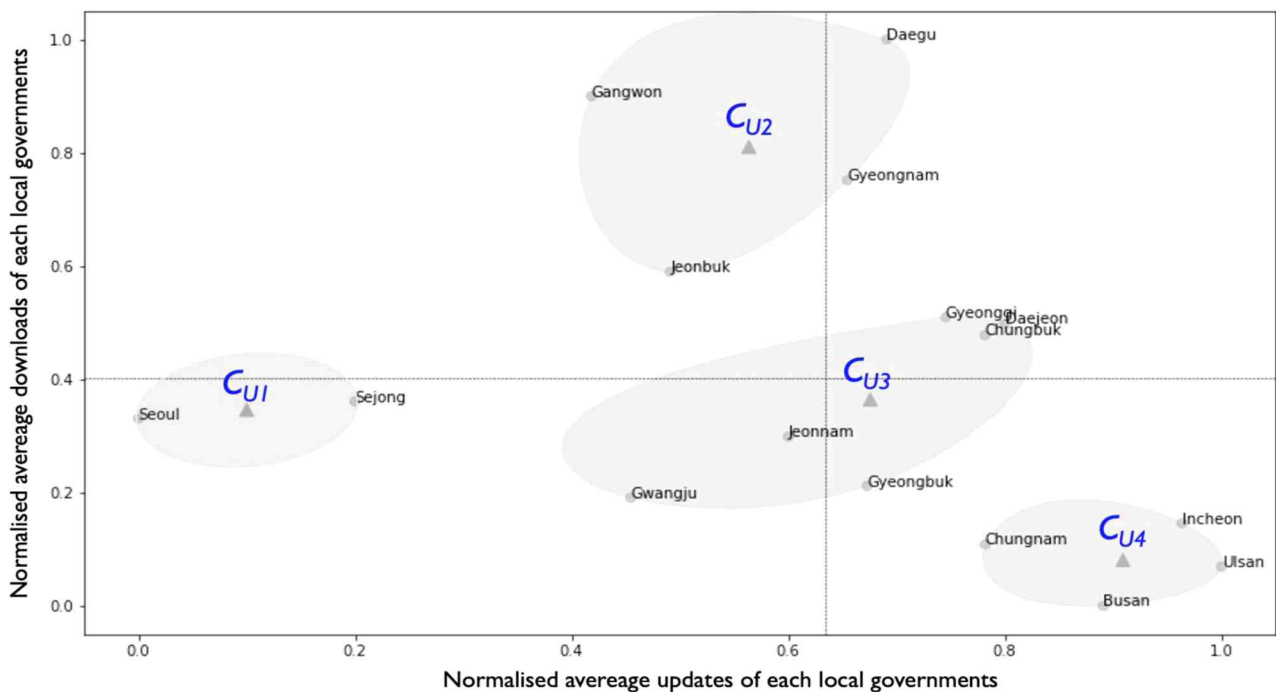
Source: Song & Kim, Considerations in Releasing Public Data: The Case of Local Governments in Korea, Journal of Information Science, 2022

Hyper Interlinked Knowledge Engineering Lab.

19

사례연구

지방자치단체 분류: 업데이트-다운로드



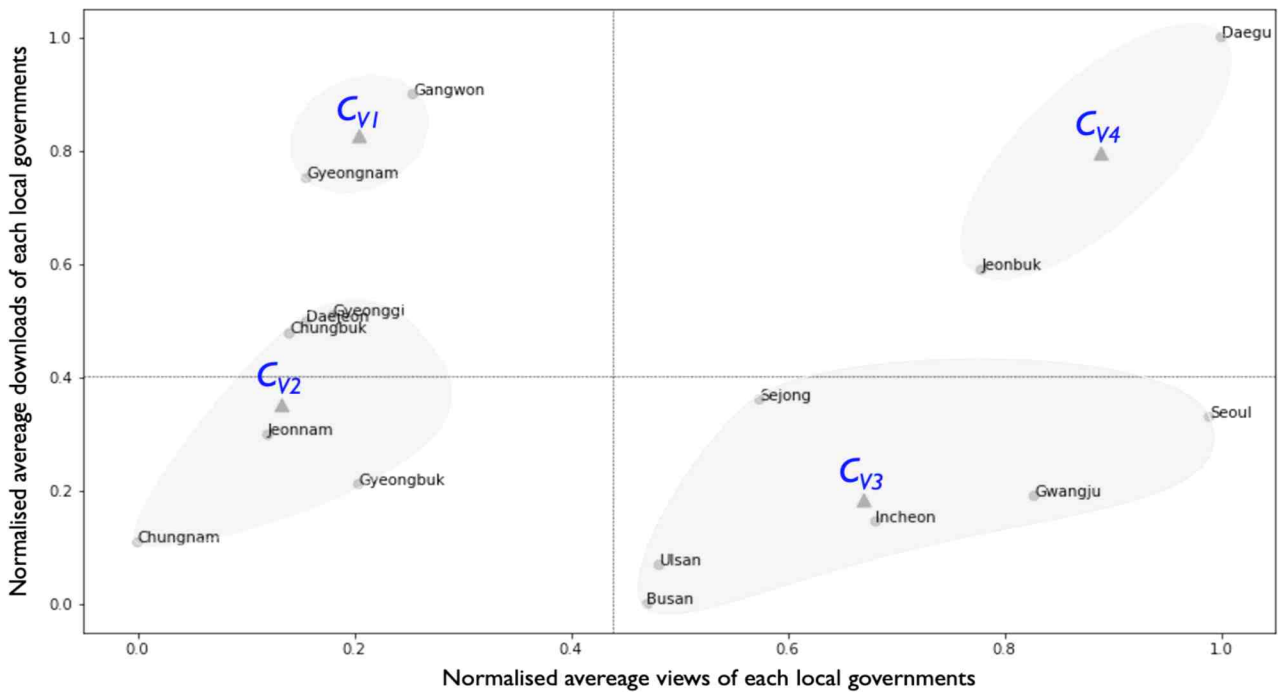
Source: Song & Kim, Considerations in Releasing Public Data: The Case of Local Governments in Korea, Journal of Information Science, 2022

Hyper Interlinked Knowledge Engineering Lab.

20

사례연구

지방자치단체 분류: 조회-다운로드



Source: Song & Kim, Considerations in Releasing Public Data: The Case of Local Governments in Korea, Journal of Information Science, 2022

Hyper Interlinked Knowledge Engineering Lab.

21

소결

양적인 개방 정책은 효과적이지 않다.

- 공공기관의 데이터 역량을 강화하기 위해 파괴적 혁신이 필요하지 않을까?
- 데이터 전문인력 확보와 지원을 위해 과도할 정도의 정책적 대응이 필요하지 않을까?
- 데이터 생산·개방·유통을 위한 기술과 소프트웨어 지원이 필요하지 않을까?

제언

공공기관의 포괄적인 데이터 역량강화를 위한 **현실적·제도적 지원**이 필요하다.

Hyper Interlinked Knowledge Engineering Lab.

22

수집

양적인 개방 정책은 효과적이지 않다.

- 공공기관의 데이터 역량을 강화하기 위해 파괴적 혁신이 필요하지 않을까?
웹 아카이브 기관
- 데이터 전문인력 확보와 지원을 위해 과도할 정도의 정책적 대응이 필요하지 않을까?
웹 아카이브
- 데이터 생산·개방·유통을 위한 기술과 소프트웨어 지원이 필요하지 않을까?
웹 아카이브 활용

제언

공공기관의 포괄적인 데이터 역량강화를 위한 **현실적·제도적 지원**이 필요하다.

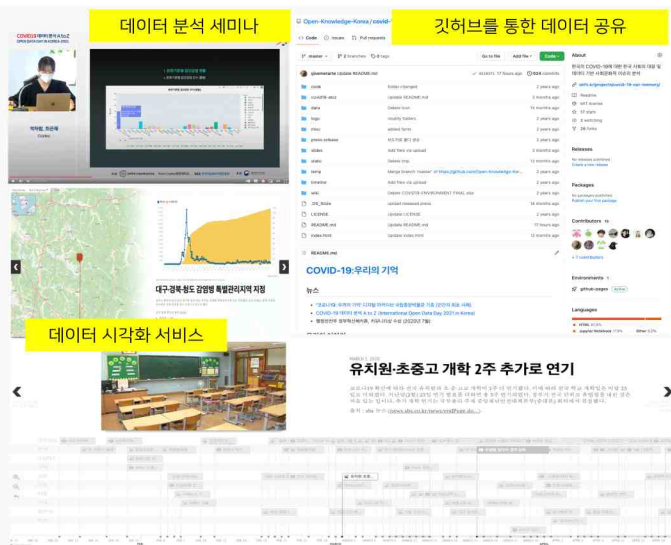
웹 아카이브 기관

사례연구

코로나19: 우리의 기억

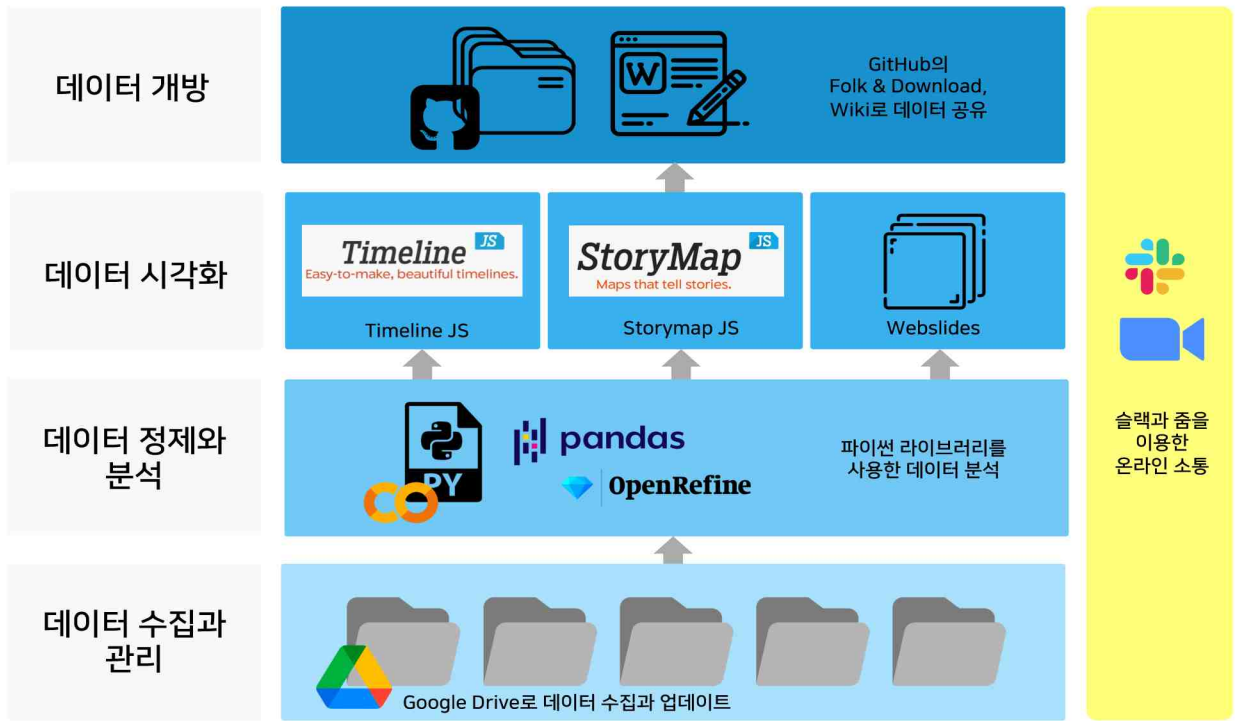
핵심 특징과 기능

- ✔ 2020년부터 현재까지 코로나19에 대한 기록
- ✔ 코로나19에 대한 가치중립적 기록
- ✔ 국내외 표준 기반 데이터 표현과 연계
- ✔ 모든 데이터, 자료의 공유 (CC-BY 라이선스)
- ✔ Interactive 디지털 아카이브 구현
 - 7개 주제의 1,511 데이터셋 포함
 - 데이터분석을 위한 28개 파이썬 코드
 - 3종의 데이터 시각화 서비스 제공
 - 80개 데이터 분석 차트 제공



사례연구

코로나19: 우리의 기억



코로나19 아카이브

비전과 핵심 가치

국가 차원의 재난 (감염병)에 대한 자료를 수집·보존하고 디지털 자료의 융합을 지향하는 아카이브 플랫폼

- ① 사회경제문화적 변화와 관련된 자료의 수집·보존과 끊임 없는 재현
- ② 공공·민간 아카이브의 연계와 수집된 자료의 의미적 연결
- ③ 국민이 참여하고 협업할 수 있도록 디지털 자료의 공유
- ④ 새로운 가치 발굴을 위한 디지털 자료의 융합과 활용

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

핵심가치: 재현

수집된 자료의 끊임 없는 재현

원천 웹사이트의 맥락에서 재현하고, 수집된 자료 사이의 통합 검색 지원

- 동적 웹페이지, 소셜미디어의 수집을 위한 웹 크롤링 기법의 최적화 검토 (오픈소스 활용 방안)
- 정적 웹페이지 외 데이터 파일, 멀티미디어의 추출과 저장 방안 검토 (메타데이터 적용과 확장 필요)
- FAIR 데이터 원칙을 적용하여 코로나19 아카이브의 관리체계를 수립하고 평가 모델로 적용
 . FAIR: 탐색성 (Findability), 접근성 (Accessibility), 상호운용성 (Interoperability), 재사용성 (Reusability)



오아시스의 수집절차 (국립중앙도서관, 2022)

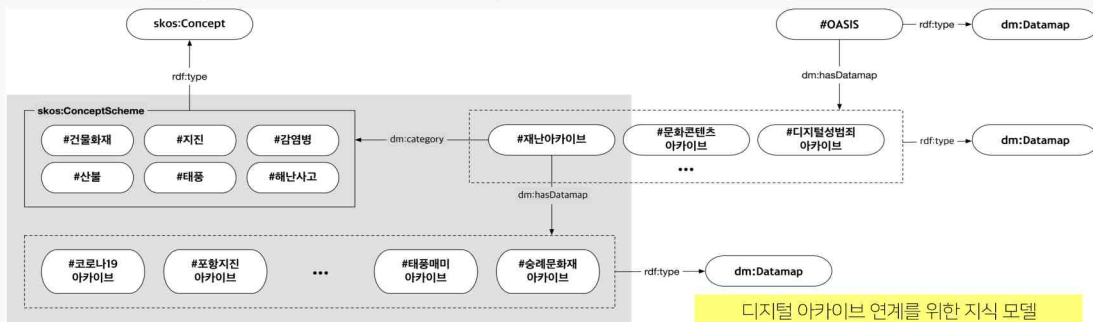
디지털 아카이브 통합 검색 서비스

핵심가치: 연결

아카이브와 기록의 연결

공공·민간의 디지털 아카이브에 연계와 기록자료의 연결을 위한 연계 체계 수립

- 관리기관에 따른 디지털 아카이브의 의미적 연결을 위해 지식그래프 기술 적용 필요
- Datamap, DCAT, SKOS 등 국내외 표준 어휘를 적용한 디지털 아카이브의 연계 모델 설계와 적용
- 오아시스 재난 아카이브의 분류체계를 기계가 읽을 수 있는 형식으로 표현 (예: SKOS)
 . 재난안전법의 재난유형을 기본으로 분류체계를 구축하고, 디지털 아카이브 연계를 위한 공통 표준으로 제정 필요



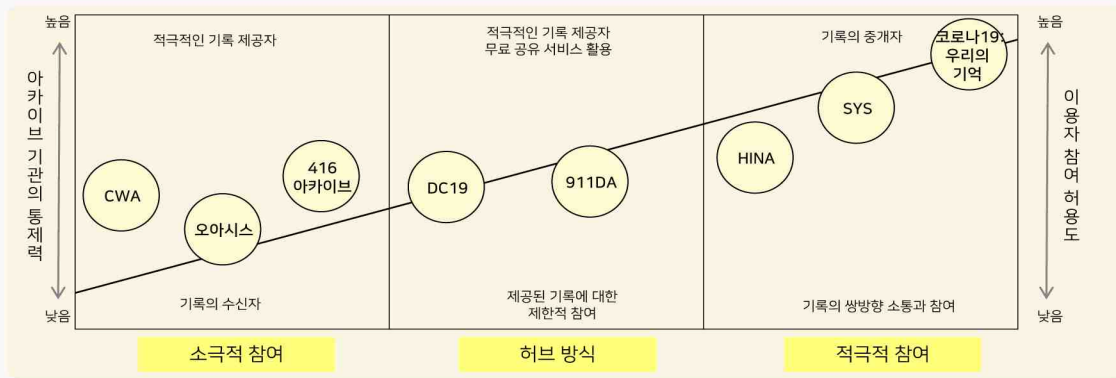
디지털 아카이브 연계를 위한 지식 모델

핵심가치: 공유

국민 참여형 서비스

기록을 공유하고 재사용할 수 있는 환경 제공, 동시에 이용자가 적극적인 기록 생산자로 기여하는 기회 부여

- 수집된 웹 자원을 온라인 환경 또는 사용자가 선호하는 방식으로 활용할 수 있는 공유 체계 마련
- 단방향 기록 제공이 아닌 사용자가 생산자가 될 수 있는 참여 기반 확대 필요

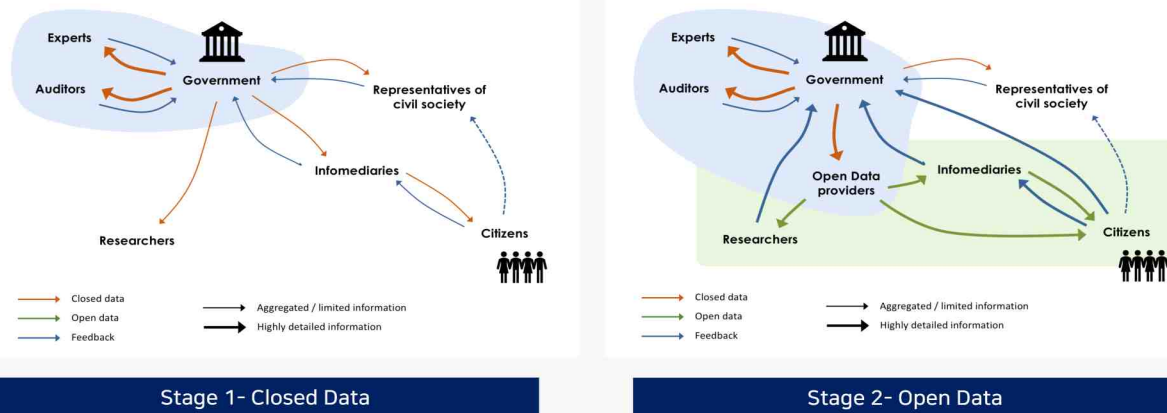


핵심가치: 활용

국민 참여형 서비스

기록을 공동활용하기 위한 체계 수립과 국민이 참여하고 협업할 수 있는 개방형 아카이브 플랫폼 지향

- 대부분의 아카이브는 단방향으로 Push 형식으로 연구자·이해관계자의 참여와 피드백을 반영하기 힘든 구조



자원연계

코로나19 아카이브에서 연계의 대상은?

목록	D-데이터허브		건축데이터 민간개방시스템	경기데이터드림		경남 빅데이터 허브 플랫폼	경남 빅데이터 허브 플랫폼	고속도로 공공데이터 포털	고속도로 공공데이터 포털	공공데이터포털	공공데이터
	파일	API	파일	파일	API	파일	API	파일	API	파일	API
메타데이터 개수	13	13	6	15	17	18	14	11	5	27	
데이터명	데이터명	데이터명	서비스명	서비스	서비스	데이터셋 명	데이터셋 명	서비스(파일명)	서비스	제목	제목
분류체계	카테고리	카테고리		분류(분류체계)	분류(분류체계)			자료 대구분		분류체계	분류체계
조회수	조회수	조회수		조회수	조회수	조회	조회			조회수	조회수
다운로드 건수	다운로드 횟수					다운로드 횟수				다운로드(바로그기)	
활용신청 건수							활용신청 건수				활용신청
제공기관	제공기관	제공기관	제공기관/제공부 서	제공 기관	제공 기관	제공기관	제공기관			제공기관	제공기관
관리부서명	제공부서	제공부서		제공 부서	제공 부서	관리부서	관리부서			관리부서명	관리부서명
관리부서(기관) 연락처						관리부서 전화번호	관리부서 전화번호			관리부서 전화번호	관리부서 전
관리부서(기관) 이메일											
관리부서(기관) 홈페이지											
보유근거										보유근거	
업데이트 주기	업데이트 주기			갱신 주기	갱신 주기	업데이트 주기		자료 제공주기		업데이트 주기	
차기 등록 예정일	차기등록 예정일					차기등록예정일				차기 등록 예정일	
매체유형										매체유형	
확장자	파일 대표 확장자	데이터 포맷				서비스 제공 형태	데이터포맷		검색결과포맷	확장자	데이터포맷
다운로드 URL	첨부파일		다운로드	파일변환저장	파일변환저장	다운로드		다운로드		다운로드	
키워드				태그	태그	키워드	키워드			키워드	키워드
등록일	등록일	등록일	데이터제공년월	등록일자(데이터 기	등록일자(데이터 기	데이터개방일	데이터개방일			등록	등록
수정일	수정일(데이터 최종 수정일)	수정일(데이터 최종 수정일)		최종 수정일자	최종 수정일자	수정일	수정일			수정	수정
접근 URL		참고 URL								URL	
제공형태						제공형태				제공형태	
설명	파일 데이터 설명	API 데이터 설명		내용	내용	설명				설명	설명
기타 유의사항										기타 유의사항	

자원연계

국내외 표준 기반 웹 아카이브의 연계

Level 0: 도메인 데이터 (인스턴스)

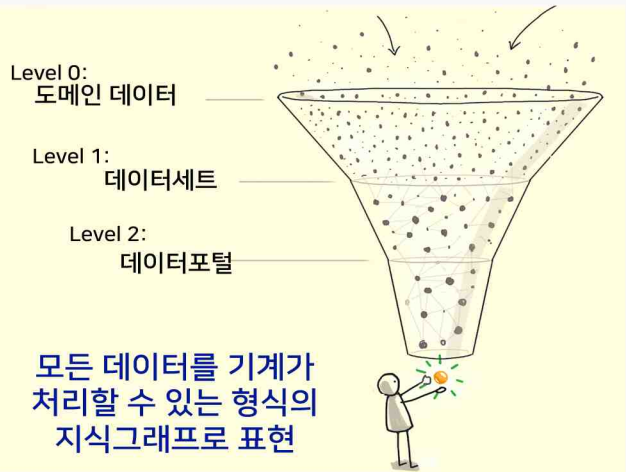
- 개별 주제, 분야에서 생성되는 가공되지 않은 데이터
- 다양한 형식으로 표현, 관리되고 있음

Level 1: 데이터셋 메타데이터

- 도메인 데이터의 특성을 반영한 메타데이터
- W3C DCAT, Schema.org 등 국제 표준의 적용

Level 2: 데이터 서비스 메타데이터

- 데이터 서비스 단위의 포괄적인 데이터의 연계
- 데이터맵 어휘 (TTA 표준)

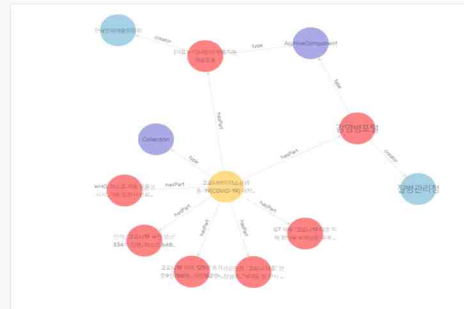
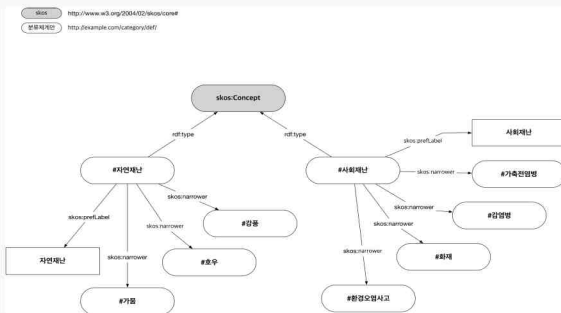


자원연계

데이터 목록의 연계 - 메타데이터와 개별 웹 아카이브

국내·국제 표준을 적용하여 FAIR 원칙을 준수하는 기계가 읽을 수 있는 아카이브 목록 구축

- DCAT: 권고안으로 웹에 게시된 데이터 목록 사이의 상호운용성을 위해 설계된 어휘
- SKOS: W3C 권고안으로 시소러스, 분류, 택소노미 등 구조화된 지식을 의미적으로 표현하기 위한 어휘
- Datamap: 데이터 포털의 메타데이터 상호운용성을 위해 설계된 어휘 (개별 웹 아카이브를 데이터맵 수준으로 표현)



Source: <https://tinyurl.com/6x7zprw8>

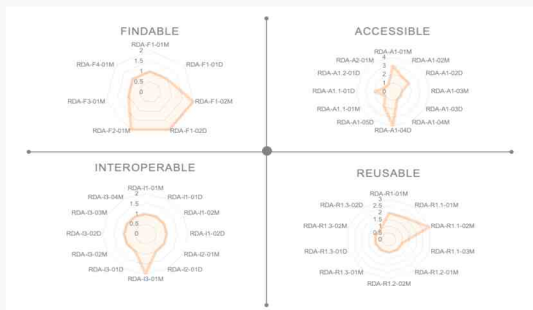
Hyper Interlinked Knowledge Engineering Lab.

지능형·소통형

데이터 보존의 지능화 - 의미적으로 처리 가능한 메타데이터

FAIR 원칙: 사람의 개입을 최소화하며 웹 아카이브의 콘텐츠를 사용하기 위한 원칙

- 연구 분야에 FAIR 원칙 적용에 서명 (G20 Hangzhou summit, 2016)
- 국제 조직 (GO FAIR) 설립과 지원 (독일, 네덜란드, 프랑스, 2017)
- FAIR 원칙의 구체적 실행 방안에 대한 국제적 논의가 지속적으로 진행되고 있음
- 디지털 아카이브의 데이터 관리 원칙으로 적용할 수 있고, 데이터 관리체계의 기준으로 검토 가능



국내 디지털 아카이브 사례

- F. 온라인 접근이 가능하지만, 데이터의 완결성이 매우 낮음
- A. 메타데이터 정보가 부정확하고 탐색에 활용하기 어려움
- I. 표준 어휘가 적용되지 않아 데이터의 연계가 쉽지 않음
- R. 기계 판독이 가능한 라이선스를 명시적으로 제공하지 않음

Source: <https://tinyurl.com/6x7zprw8>

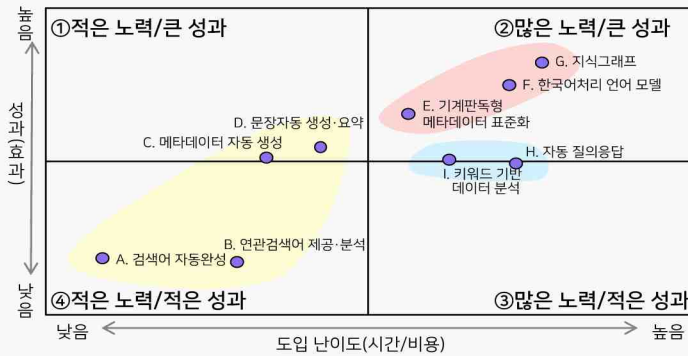
Hyper Interlinked Knowledge Engineering Lab.

지능형·소통형

인공지능 서비스 모델의 분석

페이오프 매트릭스 기준으로 인공지능 서비스 모델의 도입 효과를 분석

- 페이오프 매트릭스: 팀이나 기업의 의사결정 과정에 사용하며, 제한적 합리성을 기반으로 의사결정에 효과적임
- 딥러닝, 머신러닝 기술과 자연어 처리 기술의 도입으로 예상되는 효과, 도입 난이도 (시간/비용)를 기준으로 분석
- 페이오프 매트릭스의 사분면은 도입 난이도 대비 성과에 따라 구분됨



Pay-Off 매트릭스

- ① 비교적 쉽게 구현 가능, 실제 서비스되지 않는 영역
- ② 고수준의 기술·인프라 기반 서비스 영역
- ③ 유지비용이 높지만, 혁신성이 낮은 영역
- ④ 시장에서 보편적화된 기능을 제공하는 영역

Source: <https://tinyurl.com/6x7zprw8>

Hyper Interlinked Knowledge Engineering Lab.

35

지능형·소통형

인공지능 서비스의 도입을 위한 우선순위

인공지능 서비스 모델을 웹 아카이브에 적용을 단기·중기·장기 과제로 구분

- (단기) 6개월 이내에 도입할 수 있는 서비스 또는 기능
- (중기) 1년 ~ 3년의 기술 개발과 적용이 필요한 서비스 또는 기능
- (장기) 3~5년의 기술 개발과 적용이 필요한 서비스 또는 기능

서비스 모델에 따른 적용 시기

분류	단기	중기	장기
개발소요시간	6개월 이내	1년 ~ 3년	3년 ~ 5년
서비스모델	A. 검색어 자동완성 B. 연관 검색어 제공·분석	C. 메타데이터 자동생성 D. 문장 요약·생성 H. 아카이브 자동분류	E. 기계판독형 메타데이터 표준 F. 웹아카이브 한국어 언어모델 G. 웹 아카이브 통합 지식그래프
주요기술	한국어 형태소 처리 기술 자연어 분석·처리 기술	CNN, BERT, GPT Ko-BERT, KoGPT	온톨로지 표준 연계·매핑 대규모 지식그래프 처리·추론

Source: <https://tinyurl.com/6x7zprw8>

Hyper Interlinked Knowledge Engineering Lab.

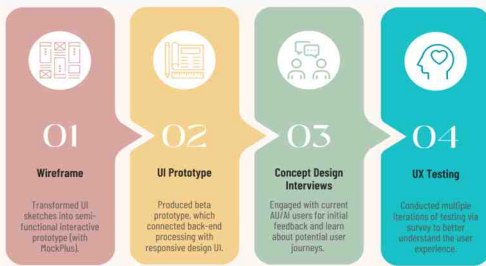
36

UX/UI

ARCH 디자인 프로세스 - Archives Research Compute Hub

12개 이상의 데이터셋을 생성할 수 있는 기능을 제공 (도메인 빈도 통계, 하이퍼링크 네트워크 그래프, 추출 텍스트 분석)

- Archives Unleashed Cloud와 Archive-It의 분석 기능을 통합하기 위한 협력 프로그램 운영
- ARCH 디자인 프로세스: 와이어프레임, 초기 UI 프로토타입, 인터페이스에 대한 인터뷰, UX 테스트로 구분해서 진행됨 . Concept Design Interview (1월-2월, 2021) → 내부 UX 테스트 (5월) → 외부 UX 테스트 (8월-9월) → 초기 사용자 UX 테스트 (1월, 2022) → Cohorts UX 테스트 (1월, 2022) → 최종 UX 테스트 (Spring, 2022)



ARCH Design Process

ARCH 서비스

Source: <https://archive-it.org/blog/files/2021/11/arch-process.png>
 Source: <https://news.archivesunleashed.org/arch-ux-testing-designing-for-the-users-f977923af7e>
 Source: <https://support.archive-it.org/hc/en-us/articles/360061122492-Introduction-to-ARCH>

데모 시연

UX/UI

Blacklight - Discovery platform framework (<http://projectblacklight.org/>)

Ruby on Rails 탐색 기능을 제공하는 오픈소스

- 필드 검색, 패시 검색, 검색결과에 대한 정렬과 페이지네이션 지원
- Apache Solr 인덱스에서 검색을 지원하는 기능 제공

프로젝트 웹사이트

적용 사례

Source: <https://tinyurl.com/6x7zprw8>

UX/UI

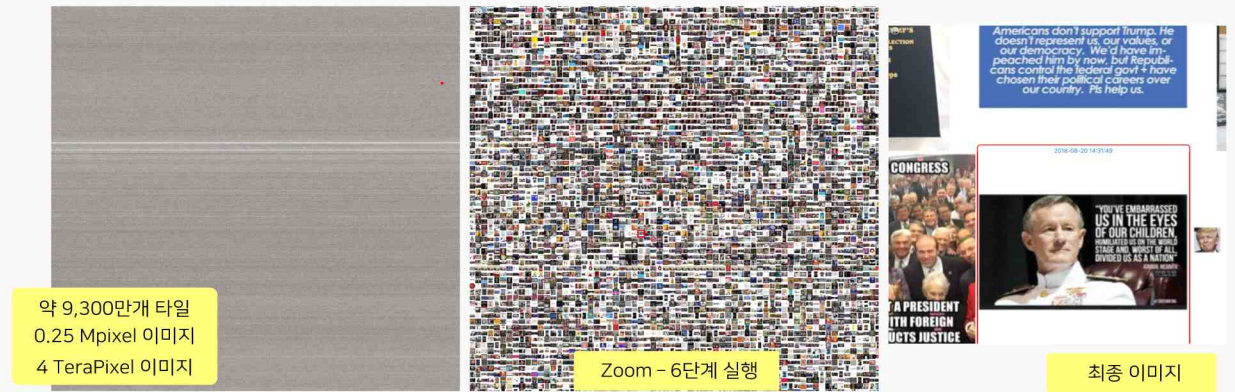
웹 아카이브의 이미지 분석과 시각화

OpenSeadragon: <https://openseadragon.github.io/>

juxta: <https://github.com/tokee/juxta>

OpenSeadragon을 사용하여 대규모 이미지의 콜로주 (collage) 생성

- 자바스크립트로 구현된 고해상도 이미지 확대/축소 기능을 제공하며, 오픈소스로 데스크탑·모바일 환경에 적용 가능
- Archives Unleashed Toolkit, Tensorflow를 이용한 이미지 데이터 분석과 시각화



Source: <https://ruebot.net/45-images/>
 Source: <https://ruebot.net/geocities-icdl2019/bird/>
 Hyper Interlinked Knowledge Engineering Lab.

활용

Archives Unleashed 프로젝트 (<https://archivesunleashed.org/>)

페타바이트 규모의 인터넷 콘텐츠를 연구자에게 제공

- 웹 아카이브 검색, 데이터 분석 도구 개발
 - . [Archives Unleashed Toolkit](#): Apache Spark 기반 웹 아카이브 분석 플랫폼
 - . [Archives Unleashed Notebooks](#): Toolkit으로 만든 데이터의 동적 시각화
 - . [Wardlight](#): WARC, ARC 형식의 웹 아카이브의 탐색 지원
 - . [Archives Unleashed Cloud](#) (Archive-It Collection과 통합)
- 다양한 기관과 데이터톤 (datathon) 개최
 - . LoC, Internet Archive, British Library, University of Toronto 등
- Archive-It 과 웹 아카이브의 사용성 향상을 위한 협력
- Archives Unleashed Cohorts (2022-2023)
 - . Latin American Women's Rights Movements: Tracing Online Presence through Language, Time and Space
 - . Historicizing Aughts-Era Mormon Mommy Blogging Media Landscapes
 - . Web Archiving and Saskatchewan COVID Archive: Expanding Coverage to Capture Social Media, Medical Misinformation, and Radicalization
 - . Querying Queer Web Archives
 - . Using Web Archives for Mapping the Use of Cultural Practices in Postconflict Societies and During Reconciliation Process



The Project
 The Archives Unleashed project aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Our team develops web archive search and data analysis tools to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the World Wide Web.

Supported by The Andrew W. Mellon Foundation, Archives Unleashed is partnering with our colleagues at the Internet Archive. We will be integrating the Cloud with the Archive-It service to ensure the project's long-term sustainability, as well as enhance the usability and accessibility of web archives.

[Read more about the next stage of the project here.](#)

Acknowledgments

The work of this project is made possible thanks to the generous support of financial and in-kind support from the following institutions:



활용

Archives Unleashed Notebooks – Analysis on Cloud

Archives Unleashed Toolkit에서 생성한 데이터를 작업하기 위한 개발 환경

- 툴킷에서 제공하는 도메인 규모, 추출된 전체 텍스트, 네트워크 시각화 데이터를 Interactive 환경으로 탐색
- 파이썬, 판다스 (Pandas) 에 대한 경험이 있으면, Jupyter 또는 Colab으로 분석 가능

The image shows two side-by-side screenshots. The left screenshot is the 'Archives Unleashed Cloud' interface, featuring a table of collections with columns for Title, Status, Date Analyzed, Public, Files, and Size. The right screenshot is a Jupyter Notebook interface titled 'auk-notebook-network', displaying a complex network graph visualization with numerous nodes and connecting lines.

Title	Status	Date Analyzed	Public	Files	Size
Leonard Cohen Collection			No	159	34.3 GB
University of Toronto Libraries Digital Collections			Yes	125	73.2 GB
Toronto 2016 Pan Am & Parapan American Games	Completed	April 16, 2020	Yes	294	90.4 GB
Canadian Political Parties and Political Interest Groups			Yes	6137	691 GB
Federal Election Candidate Sites 2015			Yes	310	206 GB
Toronto Mayoral Election 2014	Completed	April 16, 2020	Yes	292	292 GB
Canadian Government Information			Yes	14356	4.66 TB
Canadian Labour Unions			Yes	7757	1.03 TB
Ontario Provincial Election 2011	Completed	August 8, 2020	No	108	7.91 GB
Sweden Archive	Completed	April 16, 2020	Yes	42	7.16 GB
Canadian Political Interest Groups			Yes	100	8.75 GB
Ontario Provincial Election 2018	Completed	April 16, 2020	Yes	939	113 GB
University of Toronto Archives Web Collection			Yes	10624	1.35 TB
University of Toronto Scarborough			No	3	27.7 MB

Source: <https://archivesunleashed.org/images/cloud-collections.png>

Source: https://archivesunleashed.org/images/AUK_Notebook_Network.png

41

활용

아카이브 데이터를 직접 활용·참여할 수 있는 기회 제공

해커톤 (hackaton), 미트업 (Meetup) 형식의 참여형 이벤트에 대한 검토 필요

- Archives Unleashed 프로젝트는 주기적으로 '데이터톤'을 개최
- 커뮤니티의 참여를 유도하는 직접적인 방안이며 아카이브 데이터, 소프트웨어의 품질 검증이 가능



Archives Unleashed Datathon, Gelman Library, George Washington University

Source: <https://netpreserveblog.files.wordpress.com/2021/03/au-datathon.png>

Hyper Interlinked Knowledge Engineering Lab.

42

정리하며 시사점

국가 차원의 코로나19 관련 재난에 대한 자료를 수집·보존은 매우 중요함

- 코로나19 아카이브를 통해 산재되어 있는 자료를 수집하고, 국가 차원에서 자료 연계를 위한 관리체계 수립이 필요
- 향후 오아시스 재난 아카이브와 통합 또는 확대 개편할 수 있도록 조직·기술·정책 차원의 종합적인 체계 마련 필요

수량 중심에서 문제 해결 중심으로

- 수집 자료의 수량적 접근이 아닌 문제 해결을 위한 방식으로 웹 아카이브의 수집과 운영 정책의 변화 필요
- 웹 아카이브의 물리적·공간적 제약을 해결하기 위한 방안 모색 필요 (예: 데이터 프리존 적용)

국민 참여 중심의 개방형 아카이브 플랫폼

- 전문가, 커뮤니티, 시민의 참여를 통해 아카이브가 성장할 수 있는 되먹임구조 방식의 운영 체계 수립
- 단방향의 자료 제공이 아닌 시민이 디지털 자료를 활용·개선한 결과를 반영할 수 있는 개방형 구조 도입 필요
- Archives Unleashed 프로젝트와 같이 수집된 데이터를 데이터 과학자가 직접 활용할 수 있는 방안 제공

정리하며 다학문적 협력의 필요성

한국기록학회논문지에서 '아카이브', '아카이빙', '오아시스' 키워드를 포함한 논문 검색 (118건)

- 문헌정보, 기록관리와 웹·컴퓨터 공학의 이론과 실무를 융합할 수 있는 협력 체계 구성이 반드시 필요함



논문 제목에 '활용'이 포함된 목록

- 한상은 (2022) 디지털 아카이브의 위키데이터 활용방안 연구
- 송주형 (2021) 지능형 아카이브 솔루션을 활용한 공개재분류 연구: 2020년 국가기록원 공개재분류 사업
- 이봉환 외 (2019) 클라우드 컴퓨팅 기반 에뮬레이션 전략을 활용한 전자기록 장기보존 방안 연구
- 이재능 외 (2019) 클라우드소싱을 통한 참여형 기상기록정보의 수집과 활용에 관한 연구
- 이상현 외 (2018) ICT기술을 활용한 동영상 기록관리 및 이용의 활성화 방안 연구
- 김은실 외 (2017) 대통령기록물을 활용한 다문화가정 자녀의 자아존중감 향상을 위한 교육 프로그램 개발
- 김인택 (2017) 인공지능을 활용한 지능형 기록관리 방안
- 최태석 (2016) 아구기록 활용에 기반한 한국아구명예의전당 운영 방안 연구
- 이혜원 (2016) 역사콘텐츠 활용을 위한 디지털 큐레이션 프레임워크 제안
- 강주연 외 (2015) 5.18 민주화 운동 기록물을 활용한 PBL모형의 교육 프로그램 개발 고등학교를 중심으로
- 설문원 (2014) 오픈소스를 활용한 사건 아카이브 구축에 관한 연구 밀양 송전탑 건설 갈등 사건을 사례로
- 최석현 (2013) 아카이브의 디지털 전시 활용효과 분석

감사합니다!

세미나 II

발표 3

도서관의 미래, 오아시스(OASIS) 발전 방안

이수상

(부산대학교 문헌정보학과 교수)

도서관의 미래, 오아시스(OASIS) 발전 방안

부산대학교 문헌정보학과
이수상 교수
sslee@pusan.ac.kr

1



NIS

목 차

1. 오아시스(OASIS) 개요
 1. 웹 아카이빙
 2. Internet Archive
 3. OASIS
 4. Virginia Tech Web Archive
 5. 웹 아카이빙의 주요 기술
2. 오아시스(OASIS)의 평가
 1. OASIS가 잘하고 있는 점
 2. 선행연구에 나타난 비판
 3. Internet Archive와 비교
3. 오아시스(OASIS)의 발전 방안
 - 5가지 제안

2



NIS

I. 오아시스(OASIS) 개요

1. 웹 아카이빙

■ 관련 용어 정의

- 아카이빙(Archiving)
 - ✓ 기록물의 보존작업
 - ✓ 대상에 따라, **디지털 아카이빙, 웹 아카이빙** 등
- 디지털 아카이빙(Digital Archiving)
 - ✓ 지속적으로 보존할 가치를 가진 디지털 객체 [대상]
 - ✓장기간 관리하여 이후의 이용을 보장하는 활동 [목적]
 - ✓관련 시스템: 디지털 아카이브/리포지토리
- [참고] 디지털 보존(digital preservation): 디지털 아카이빙의 한 과정
 - ✓장기 보존(long-term preservation)과 접근(access) 유지
 - ✓지속적 가치(enduring/ongoing values) 유지
 - ✓보존기술: 매체전환(Reformatting), 파일재생(Refreshing), 전환(Migration), 에뮬레이션(Emulation), 캡슐화(encapsulation) 등
- 출처: 위키백과 등

3



NIS

I. 오아시스(OASIS) 개요

■ 웹 아카이빙(Web Archiving)

- 웹자원의 아카이빙
 - ✓대상: 표면웹(surface web)과 심층웹(deep web), 웹페이지와 웹문서
- 웹(Web)에 기록된 데이터를 웹 아카이브(Web archive)로 수집, 저장, 보존, 활용(미래의 연구에 사용)할 수 있도록 하는 프로세스

■ 웹 아카이빙의 주요 사례

- **OASIS**(Online Archiving & Searching Internet Sources)
- **Internet Archive**
- Australia's Web Archive Pandora
- NLI(아일랜드 아카이브)
- Library of Congress Web Archive Collection
- Government of Canada Web Archive
- UK Web Archive
- **Virginia Tech Web Archive**
- University of North Texas' CyberCemetery

4



NIS

I. 오아시스(OASIS) 개요

2. Internet Archive

■ Internet Archive(www.archive.org)

- 1996년 설립된 비영리 단체
 - ✓핵심적인 임무는 모든 지식에 대한 보편적인(무료/공개) 액세스 제공
- 인터넷 매체의 보관 시작
 - ✓현재 세계 최대의 웹 아카이브: 현재 6,250억 개의 웹 페이지 서비스 제공
- 디지털 형식의 기타 문화 유물의 **디지털도서관 역할**까지 수행
 - ✓현재 전 세계 18개 지역에서 하루 4,000권의 책을 스캔 (현재 3,800만 권의 책과 텍스트 등 보유)
 - ✓1927년 이전에 출판된 책의 다운로드 가능
 - ✓수십만 권의 현대 책을 **Open Library 사이트**를 통해 대출 가능
- **Wayback Machine**을 통해 25년 이상의 웹 기록에 액세스 가능
- **Archive-it** S/W를 도서관 및 기타 파트너와 협력, 중요한 웹 페이지를 아카이빙
- **Heretrix**라는 웹크롤러 개발
- 출처: www.archive.org, https://newstapa.org/article/gZpYB

5



NIS

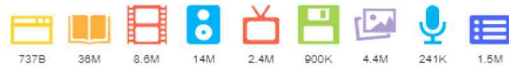
I. 오아시스(OASIS) 개요

■ Internet Archive의 검색(www.archive.org)



Wayback Machine 검색

Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.



Search **GO**

- Search metadata
- Search text contents
- Search TV news captions
- Search radio transcripts
- Search archived web sites

Advanced Search

"national library of korea" 검색 시 (2022-09-23)
-전체 50건
* Internet Archive의 검색결과와 차이 발생

Internet Archive 검색

"national library of korea" 검색 시 (2022-09-23)
-전체 28,397건
-web 유형: 82건

이것을 선택할 경우, Wayback Machine의 검색결과와 동일

6



NIS

I. 오아시스(OASIS) 개요

■ Wayback Machine

- Internet Archive의 웹 아카이브
- 공개 웹페이지를 저장 및 영구 보존하고 이를 공개하는 서비스
- 매일 10억 개 이상의 영구 저장된 URL을 저장해 공개
- 언론인, 연구자, 팩트체크 전문가, 활동가뿐만 아니라 일반 대중이 무료로 웨이백머신을 다양한 방법으로 이용 중
- 지금까지 이 서비스 또는 이 서비스 결과물을 출처로 작성된 기사는 수천 건

■ Archive-it (<https://archive-it.org/>)

- 2006년부터 시작, 현재 24개국 이상에 있는 1,100개 이상의 조직(도서관, 문화유산 및 연구 기관, 사회적 집단, 교육 및 오픈지식 프로젝트 등)에 웹 아카이빙 서비스(구독형) 제공
- Archive-it 이용자들이 현재 400억 개 이상의 웹자료를 보존 중
- 아카이빙된 자료는 Internet Archive에서 독립적으로 소유 및 운영하는 비영리 데이터센터에 저장되고 서비스됨

■ 출처

7



- <https://archive-it.org/explore>, www.archive.org, <https://newstapa.org/article/gZpYB>, <https://archive-it.org/blog/learn-more/> 등

NIS

I. 오아시스(OASIS) 개요

3. OASIS

■ OASIS

- 2004년 시범수집, 2006년 OASIS 홈페이지(www.oasis.go.kr; nl.go.kr/oasis) 오픈
- 목적
 - ✓ 소멸되기 쉬운 공개된 인터넷자원(웹사이트/웹자료)을 국가적인 차원에서 디지털 지식문화유산으로 수집·보존
 - ✓ 미래세대에 연구 자료로 제공

■ OASIS의 아카이빙 대상

- 선택적 수집: 주요 사이트를 선별하여 수집. 선별기준 필요
- 포괄적 수집: 국가도메인(.kr 등), 국내에서 운영되는 일반최상위도메인(.com, .net, .org)까지 확대 수집 가능

■ OASIS의 주요 프로세스

- 수집: IIPC에서 개발한 Heritrix 사용
- DB 구축: DB구축시스템 사용
- 검증, 색인작성: KOLIS II 사용
- 온라인 서비스: 국립중앙도서관/OASIS 홈페이지

8



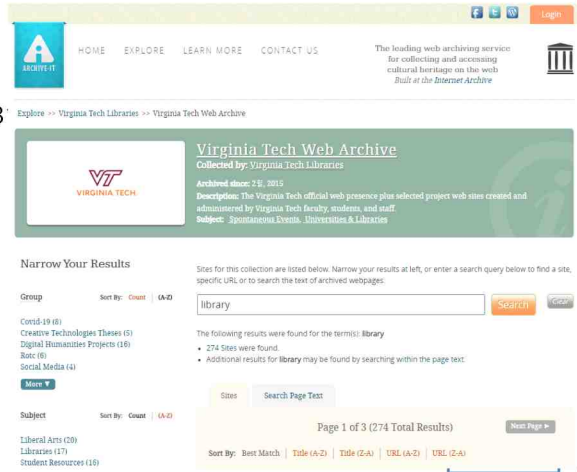
NIS

I. 오아시스(OASIS) 개요

4. VTWA(Virginia Tech Web Archive)

■ 주요 특징

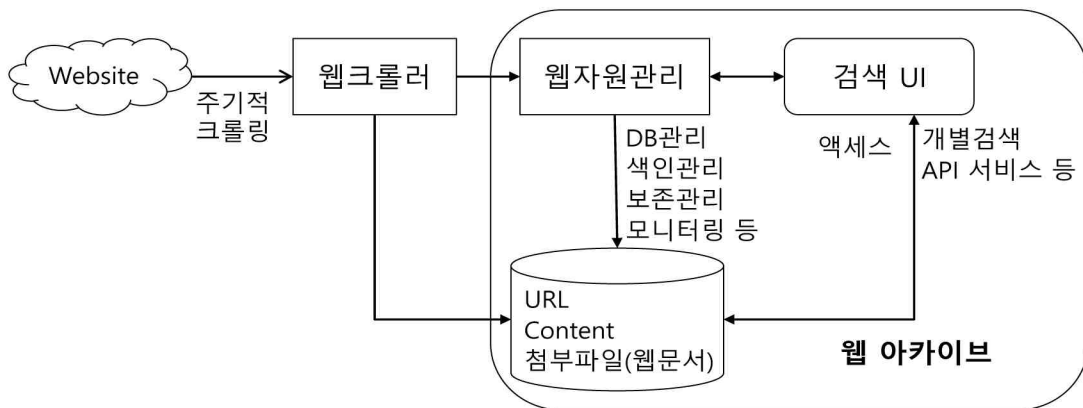
- 2015년부터 시작, Archive-it을 사용
- 수집범위
 - ✓ Virginia Tech 도메인(vt.edu)을 포함
 - ✓ 원본 시드에서 4단계 수준의 하이퍼링크를 검색
- 웹 아카이브는 공개적으로 사용 가능
- 관련 사이트
 - ✓ <https://www.archive-it.org/collections/53>
- 이외에도 Archive-it을 사용하는 대학들 많음
 - ✓ <https://archive-it.org/explore> 참조



I. 오아시스(OASIS) 개요

5. 웹 아카이빙의 주요 기술

■ 웹 아카이빙 시스템 주요 구성도



I. 오아시스(OASIS) 개요

■ 웹 아카이빙의 주요 과정

- 선택과 계획(Selection and planning)
 - ✓아카이브의 사명에 따라 대상 선택(범용, 지역별, 주제별 등으로 제한 가능)
 - ✓특정 기관, 기업, 국가정부에 서비스 제공 가능
- 권한(Permissions): 사이트의 수집권한 확인
- 수집(Harvesting)
 - ✓무료 또는 상용의 웹크롤러 사용
 - ✓시작해야 하는 URL(시드라고 함)의 링크를 통해 각 페이지와 관련 파일들을 저장
- 기술(Description): 파일이 수집되면 파일의 특성과 출처를 정확하게 설명하는 메타데이터와 함께 저장
- 장기 보존(Long-term preservation): 웹 아카이브는 자체 검사 및 수리를 위한 여러 중복 사본과 프로세스가 있는 공인된 보안 환경에 파일을 저장
- 액세스(Access): 수집되고 기술된 WARC 파일을 원래의 형식으로 재현
- 출처: <https://blogs.kent.ac.uk/isnews/how-to-archive-your-website/>

11



NIS

I. 오아시스(OASIS) 개요

■ 웹 아카이빙의 주요 개념과 기능

- 웹크롤링(Web Crawling; Web Harvesting)
 - ✓시드(seeds)라고 하는 URL들의 세트리스트(set list)를 방문, 해당 시드의 모든 하이퍼링크를 식별, 시드를 "크롤링"할 때 관련 정보를 복사 및 저장하는 프로세스
 - ✓이 정보는 WARC 파일포맷으로 된 스냅샷으로 저장되며, WARC 뷰어로 재생할 때 캡처된 그대로 재생됨
- 웹크롤링의 유형
 - ✓원격 수확(Remote harvesting): 웹 크롤러를 사용하여 웹페이지 탐색 및 캡처
 - ✓데이터베이스 수확(Database harvesting): 쿼리를 위해 데이터베이스에서 XML과 같은 표준 스키마로 데이터 내보내기
 - ✓트랜잭션 수확(Transactional harvesting): 페이지 보기와 같은 이벤트를 기반으로 웹 페이지를 캡처
- 웹크롤러(Web Crawler; Web Harvester): 웹크롤링 도구
 - ✓오픈소스 웹 크롤러: Heritrix, Scrapy, DeepArc 등

12



NIS

I. 오아시스(OASIS) 개요

- WARC(Web ARChive) 포맷
 - ✓복수의 디지털 자원들을 관련 정보와 함께 하나의 복합 아카이브 파일(an aggregate archival file)로 결합하는 방법을 제공
 - ✓WARC 형식 파일은 하나 이상의 WARC 레코드를 연결
 - ✓WARC 레코드는 레코드 헤더와 레코드 콘텐츠 블록으로 구성
 - ✓WARC 레코드의 8가지 유형: 'warcinfo', 'response', 'resource', 'request', 'metadata', 'revisit', 'conversion' 및 'continuation'
 - ✓레코드 헤더에는 필수 필드(레코드의 날짜, 유형, 길이 등)가 있어, 각 자원들의 편리한 검색을 지원
 - ✓WARC 파일의 콘텐츠 블록에는 모든 포맷의 자원들(예: HTML 페이지에 포함되거나 링크되는 이미지 파일, 시청각 파일)을 포함
 - ✓참조: OCLC에서 개발한 Web Archiving Metadata 스키마
- 출처: <https://guides.lib.vt.edu/webarchiving/home>. 웹 자원 아카이빙을 위한 웹 크롤러 연구 개발

13



NIS

II. 오아시스(OASIS)의 평가

1. OASIS가 잘하고 있는 점

■ 콘텐츠 측면

- 방대한 규모의 웹사이트 자료 아카이빙
 - ✓구축현황: 웹사이트 886,088건 / 파일 3,351,150개 / 용량 391TB (2021. 12월 기준)
- 특정 주제에 대한 웹 아카이빙
 - ✓국가 이슈·재난·주제 분야 등(현재 154건) 관련 웹 아카이빙(2014년 이후)
- 소셜미디어 수집
- 선택적 수집에서 포괄적 수집으로 전환(2016 이후): 수집량의 증가

■ 기술적인 측면

- Heritrix 수집기 사용
 - ✓오픈소스, Internet Archive에서 개발한 도구

14



NIS

II. 오아시스(OASIS)의 평가

■ 표준화 측면

- WARC 포맷 사용

■ 서비스 측면

- 컬렉션 서비스

✓전체 컬렉션 보기 (2022년 9월 현재 88건)

✓테마별 구성

❖ 상위 테마: 총류, 예술, 문학, 역사, 사회과학, 자연과학, 기술과학

- 웹트렌드 분석

✓시간의 흐름에 따라 웹사이트에서 나타난 특정 단어들의 관심도(웹페이지에 나타난 특정 단어의 출현율) 변화 추이와 단어들 간의 상관관계 파악

✓시각화: 키워드 비중(단어들 간 출현비중), 수집년도 구분(전체자료의 검색결과 비율), 도메인 구분

✓외부자료 혼합: 검색된 자료를 그래프로 분석한 자료에 개인의 지식, 또는 각종 리서치 자료, 공공데이터 등 다른 데이터들을 연계하면 더 풍부한 의미를 부여

✓태그클라우드 등

15



NIS

II. 오아시스(OASIS)의 평가

2. 선행연구에 나타난 비판

■ 선행연구 1

- 콘텐츠의 품질에 대한 의문
- 수집 웹사이트의 주제적 불균형
- 수집 웹사이트의 지속성 불확실
- 수집 웹사이트의 최신성 불분명
- 수집 웹사이트 통계의 부정확

■ 선행연구 2

- 웹 아카이빙과 OASIS 프로젝트에 대한 사람들의 인지도가 매우 낮음
- OASIS 웹 사이트는 첫 접속자가 많았고, 이탈률이 높음
- OASIS 이용률 제고를 위한 적극적인 홍보 활동 부족
- 서비스 개편의 부족

■ 선행연구

- 1.윤정옥. 2014. 수집 웹사이트의 지속성 불확실
- 2.남재우, 이수영. 2021. 웹 아카이브 OASIS 개선방안에 관한 연구

16



NIS

II. 오아시스(OASIS)의 평가

3. Internet Archive와 비교

■ 비교

	Internet Archive	OASIS
콘텐츠	약 6,300억 개의 웹페이지	약 900,000건 웹사이트
콘텐츠 유형	웹사이트, 도서, 영화, S/W, 음악 등	웹사이트, 문서, 동영상, 이미지 *웹으로 유통되는 콘텐츠
통합 검색	모든 유형의 콘텐츠 대상 웹사이트 대상 *홈페이지(www.archive.org)에서 검색창 분리	웹사이트 대상: nl.go.kr/oasis 국립중앙도서관(nl.go.kr)에서 통합검색
구축자	협업 (플랫폼 형태) *Wayback Machine(개인), Archive-it(조직체) 사용: 도서관 및 파트너(24개국 1,100개 이상의 조직체에서 사용)	단독(국립중앙도서관) *단독 시스템
개방성	보편적인(무료/공개) 접근 제공	국립중앙도서관 내에서만 접근 *아카이빙된 국립중앙도서관 웹사이트도 외부공개 불가
기술	다양한 오픈소스 S/W 제공 *Archive-it, Heretrix 등	국제적으로 통용되는 기술과 표준 사용 (Heretrix, WARC 포맷 등)

17



NIS

III. 오아시스(OASIS)의 발전 방안

5가지 제안

- **콘텐츠의 다양화: 국가 디지털 문화유산의 아카이빙**
 - 웹사이트 위주에서 다양한 콘텐츠들 확대 강화: 국내 모든 웹콘텐츠를 대상
- **협업 구축환경 제공: 다양한 파트너와 협업**
 - 지역 대표도서관, 지역 중앙도서관, 대학도서관, 전문도서관 등
 - 공공기관, 전문기관 및 단체, 개인 등
- **개방성의 확대: 저작권 문제의 해결**
 - 현재: 아카이빙 대상 웹사이트 기관들과 협약(이용 동의, 비공개 요청) 처리
 - 향후: 공정이용으로 인정 또는 옵트아웃(opt-out) 체제 제안
 - 공정이용: 보도·비평·교육·연구 등을 위한 인용일 경우, 허락 없이 이용 가능
 - 옵트아웃(opt-out) 체제 제안
 - ✓ 저작권자가 특별한 의사표시를 하지 않는 한 새로운 권리처리가 적용되는 체제
 - ✓ 정보가 인터넷에서 이용 가능한 것을 고려할 때, 옵트-아웃 방식을 채택하여 저작권자의 이익을 부당하게 해치지 않는 범위 내에서 저작물을 자유롭게 이용
 - ✓ 참조: 김창화. 2021. 데이터 마이닝과 저작권 면책의 범위 및 한계. 계간 저작권, 34(2): 5-31.

18



NIS

III. 오아시스(OASIS)의 발전 방안

■ 협업을 위한 서비스 플랫폼 구축

- 클라우드/플랫폼 형태의 S/W 사용 (무료 구독형으로 제공)
- 국가별, 지역별, 기관별, 개인별 웹 아카이빙 서비스 제공
- 참조: 개인의 Wayback Machine 이용, 조직체의 Archive-it 이용 등

■ 서비스 플랫폼으로서 요건 강화

- 기존의 목적(8쪽 참조)을 플랫폼으로 구현
- 제공자(구축자): 구축 대상 디지털 지식문화유산의 확대
- 소비자(연구자, 기자, 작가, 학생 등)
 - ✓ 활용목적(조사연구 등) 요구 충족
 - ✓ 활용을 위한 다양한 프로젝트 수행
- 편의 서비스의 확대
 - ✓ 다양한 활용사례들 공유
 - ✓ 대시보드 서비스 등

19



NIS

감사합니다.

질의 & 응답

20



NIS

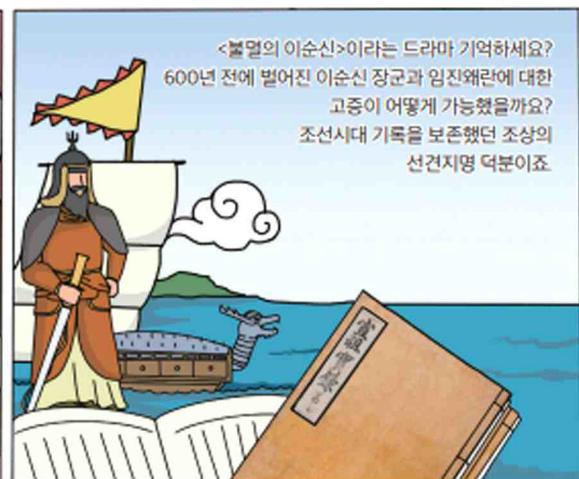
부 록

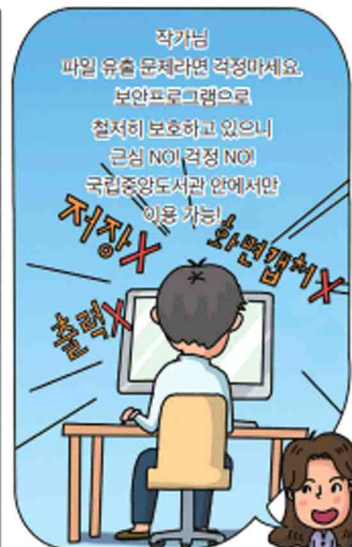
전자책 납본 가이드

- 오아시스(OASIS) 홍보 리플릿



당신의 전자책은 소중한합니다.





*자세한 '전자책 납본 방법'은 <전자책 납본 가이드> 페이지를 참조하세요.

GUIDE

전자책 납본 가이드

납본대상

국제표준도서번호(ISBN)를 부여받은 전자책

납본방법



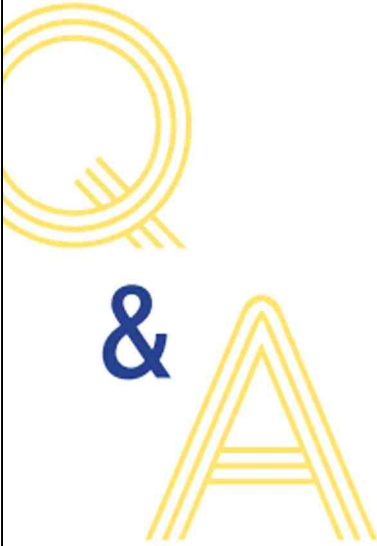
자료이용

<p>관내이용</p> <ul style="list-style-type: none"> · 국립중앙도서관 내에서만 이용 	<p>이용범위</p> <ul style="list-style-type: none"> · 열람만 가능 · 인쇄, 저장, 화면 캡처 불가 	<p>USER 수</p> <ul style="list-style-type: none"> · 동시 접속자 수 1명 	<p>보안</p> <ul style="list-style-type: none"> · 방화벽 구축 · 관내 전용 뷰어 개발 · DRM(디지털 저작권 관리 시스템) 적용
--	---	--	--

납본문의

전화 02-590-0700, 내선 3번 - 전자책 6번 **팩스** 02-590-0672
이메일 onsujip2020@korea.kr
주소 (06579) 서울시 서초구 반포대로 201 국립중앙도서관 온라인자료과 전자책 납본 담당자
홈페이지 국립중앙도서관 ISBN-ISSN-납본시스템 → 전자책 납본 → 온라인 자료 납본 안내

전자책 납본 초보자를 위한



Q1

**전자책을 출판하면
무조건 납본해야 하나요?**

네. 국제표준도서번호(ISBN)를 부여받은 전자책은 국립중앙도서관에 납본해야 합니다.

※ 근거 : 도서관법 제20조(도서관자료의 납본)

Q2

**전자책에 대한 납본 제외
기준이 있나요?**

전자책이지만 일반적인 책의 형태나 구성요소를 갖추지 않은 자료, 이미 공개된 내용을 편집한 자료, 단순 스캔자료 등은 납본 제외 자료로 반려될 수 있습니다. '납본 제외 자료'는 국립중앙도서관 ISBN-ISSN-납본시스템(<http://nl.go.kr/seoji>) → 온라인 자료 납본안내 에서 확인할 수 있습니다.

Q3

**전자책 납본 파일의
외부 유출 위험은 없나요?**

국립중앙도서관은 방화벽 구축과 함께 전자책 납본 파일의 외부 유출 및 불법 복제 등의 방지를 위해 디지털 저작권 관리(DRM) 기반의 기술적 보호조치를 통해 철저히 관리하고 있습니다. 특히 보존용 파일은 외부 접근이 불가능하도록 최신 보안시스템이 설치된 환경 속에서 철저히 관리, 보존하고 있습니다.

Q4

**전자책 납본 절차가
어렵지 않나요?**

전자책 납본은 국립중앙도서관 홈페이지를 통해 신청 가능합니다.

- ① 국립중앙도서관 ISBN-ISSN-납본시스템(<http://nl.go.kr/seoji>)에 전자책 파일을 업로드
- ② 대량 납본은 전자책 파일과 서지정보를 저장매체(USB, 외장하드 등)에 담아 송부

※ 판매용 전자책은 열람용 1부에 대한 정가를 납본 보상금으로 지급하오니, 관련 서류(납본서·보상청구서, 전자계산서)를 제출하시면 됩니다.

Q5

**납본한 전자책을
집에서도 볼 수 있나요?**

납본한 전자책은 외부에서 볼 수 없습니다. 국립중앙도서관을 직접 방문해야만 열람할 수 있고, 저작권법에 따라 동시에 열람할 수 있는 이용자도 한 명으로 제한하고 있습니다. 인쇄 및 저장, 화면 캡처는 모두 불가능합니다.

OASIS

국립중앙도서관 웹자원 아카이브

OASIS Online Archiving & Searching Internet Sources

오아시스(OASIS)는 공개된 인터넷자원을 국가 디지털 지적 문화유산으로 수집·보존하여 후대에 전승하기 위한 온라인 디지털자원 수집·보존 프로젝트입니다.

국립중앙도서관은 도서관법 제20조의2(온라인자료의 수집)에 의거, 2004년부터 공개 웹자원을 수집·보존하는 오아시스(OASIS) 사업을 추진하고 있습니다.

수집한 웹자원은 웹 아카이브로 구축하고 주제 큐레이션하여 국립중앙도서관 누리집 및 오아시스 누리집을 통해 대국민 서비스를 제공하고 있습니다.

아울러 국제인터넷보존컨소시엄(IPC : International Internet Preservation Consortium)에 가입하여 COVID-19 공동 컬렉션 구축에 참여하는 등 국제적인 기술·지원 공유 및 협력 활동을 수행하고 있습니다.



추진 경과

2004년	웹사이트 시범 수집
2006년	OASIS 누리집 오픈
2008년	국제인터넷보존컨소시엄(IPC) 정회원 가입
2011년	국제표준(Heritrix 수집기, WARC파일) 도입
2016년	포괄적 수집 개시
2017년	웹트렌드 시각화 서비스



수집 대상

- 우리나라에서 생산된 웹사이트 및 웹사이트 내 공개된 자료(웹자료)를 수집합니다.
- ▶ 대한민국 국가도메인 웹사이트 및 한국 관련 해외 웹사이트
(국가·공공기관, 교육연구기관, 민간·시민사회단체 및 비영리조직, 상업·언론기관 등)
 - ▶ 웹사이트를 통해 공유되는 웹자료
(문자자료, 음성·음향자료, 영상자료, 이미지자료 등)
 - ▶ 정부부처, 지방자치단체, 공공기관의 블로그 및 트위터 등 주요 소셜미디어
 - ▶ 데이터세트 등 기타 다양한 형식의 웹자원

수집 방법

대한민국 국가도메인 웹사이트 전체를 대상으로 한 포괄적 수집을 기본으로 하며, 국가 재난 및 주요 행사와 관련된 웹자원에 대해 선택적 수집을 병행하고 있습니다.



국립중앙도서관 웹자원 아카이브 오아시스를 통해,
현 시대의 웹기록물을 만나보실 수 있습니다.

웹자원 수집 현황(2004~2020)

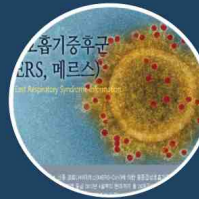
2004년부터 웹사이트 75만건, 웹자료 123만건을 수집하여, 전체 약 200만건의 웹자원을 보유하고 있습니다.



연도	웹사이트	웹자료	계(건)
2020년	133,354	75,809	209,163
2019년	120,058	75,314	195,372
2018년	126,383	93,864	220,247
2017년	152,385	90,238	242,623
2016년	75,994	54,830	130,824
2015년	50,100	100,678	150,778
2014년	38,540	191,491	230,031
2013년	20,464	83,484	103,948
2012년	15,242	65,280	80,522
2011년	4,076	158,821	162,897
2004~2010년	21,054	246,654	267,708
총계	757,650	1,236,463	1,994,113

국가재난아카이브

'코로나19' 컬렉션을 비롯하여 국내에서 발생하는 주요 재난에 대한 웹자원을 컬렉션 단위로 수집·보존합니다.



오아시스 컬렉션

국가적으로 중요한 행사와 정치·사회·문화적으로 큰 반향을 일으킨 사건, 인문·사회·자연과학 및 기술 분야별 중요한 주제 관련 웹자원을 수집·보존합니다.



공개 및 보존

수집한 웹자원은 영구 보존되며, 원칙적으로 국립중앙도서관내 이용 자료로 제공됩니다.
(단, 저작권자의 이용 동의가 있을 경우, 관외 이용 자료로 제공됩니다.)

서비스 이용 동의

오아시스 웹자원의 대국민 서비스 제공을 위해서 저작권자의 이용 동의가 필요합니다.

[이용 동의 방법]

- ▶ 온라인 신청 : 오아시스 누리집 oasis.go.kr > 웹사이트 기증 > 이용 동의서 등록
- ▶ 공개용 온라인자료 이용동의서 양식 작성 후 회신(메일, 팩스, 우편)

memo ●

A large, empty rectangular box with a thin black border, occupying most of the page below the 'memo ●' header. It is intended for the user to write their memo content.

memo ●

A large, empty rectangular box with a thin black border, occupying most of the page below the 'memo ●' header. It is intended for the user to write their memo content.

제59회 전국도서관대회 세미나

미래 도서관 웹 콘텐츠 수집 정책

인 쇄 | 2022년 10월 13일

발 행 | 2022년 10월 13일

발행처 | 국립중앙도서관 온라인자료과

주 소 | 서울특별시 서초구 반포대로 201

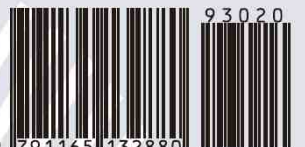
ISBN 979-11-6513-288-0 93020



국립중앙도서관
National Library of Korea

비매품/무료

93020



9 791165 132880

ISBN 979-11-6513-288-0