Submitted on: 19.06.2020

Extracting Online Publications Embedded in Websites: NDL Initiatives and Challenges

INOIE Nobuaki
Digital Library Division, Kansai-kan of the National Diet Library (NDL), Kyoto, Japan.

SHIBATA Masaki
Kansai-kan of the National Diet Library (NDL), Kyoto, Japan.

KUDO Tetsuro
Digital Library Division, Kansai-kan of the National Diet Library (NDL), Kyoto, Japan.

Abstract:

The National Diet Library (NDL) has been operating the Web ARchiving Project (WARP) since 2002 and steadily archiving Japanese websites. However, it is often difficult for users to find e-books, e-zines and other online publications embedded in websites, because they are stored as a part of websites and do not have sufficient metadata.

To help alleviate this problem, in 2010 the NDL started a project to extract those publications from the archived data of WARP and create metadata for each of them. In this paper, we will describe the criteria for selecting publications to extract and the workflow for the extraction and creation of metadata in the project. As of March 2020, we have extracted and created metadata for 570,000 publications. Those extracted publications are efficiently discoverable on the NDL Digital Collections with the metadata and linked from the online catalogue of the NDL.

In addition to the above, we will briefly refer to the challenges we face in the project: Improving the efficiency of the workflow, enriching metadata and archiving moving image files. We will continue to tackle these challenges.

Keywords: web archiving, online publications, metadata, moving image files, national libraries

1 Introduction

Since 2002, the National Diet Library (NDL) has been operating the Web ARchiving Project (WARP) and has been steadily archiving Japanese websites. WARP currently archives over 12,600 websites comprising 1.7 PB of data and is becoming one of the largest web archives in the world. That consequently means WARP preserves a tremendous amount of e-books, e-zines and other online publications that are embedded in websites.

However, it is often difficult for users to find and access those publications archived in WARP in spite of the fact that they have both social and cultural significance. It is because they lack sufficient metadata (e.g., title, publisher’s name, publication date) and for this reason, can neither be efficiently searched for nor easily listed.

To help alleviate this problem, the NDL has since 2010 been extracting and creating metadata for publications embedded in websites that have been archived in WARP. As of March 2020, 570,000 online publications have had metadata added and are discoverable on the NDL Digital Collections, which is the database for digitized materials and online publications collected by the NDL.

In this paper, we will introduce web archiving by WARP and describe our workflow for extracting embedded online publications from WARP. Also, we will briefly refer to the challenges we are facing in the project.[1]

2 Web archiving by WARP

WARP archives websites hosted both by Japanese public agencies and by highly-public private organizations, e.g. public interest corporations, academic societies, political parties, etc. Public agencies’ websites are now comprehensively archived as prescribed in the 2010 revision of the National Diet Library Law, whereas private organizations’ websites are selectively archived based on the permission of their webmasters.

Figure 1 shows the transition in data size and the number of archived websites of WARP. As of March 2020, WARP has archived 12,556 websites, of which about 5,800 are public agencies' websites and about 6,700 are private organizations' ones. The amount of data has reached 1,679 terabytes. Around 85% of the websites are available on the internet with permissions by the webmasters.

The proportion of file formats in the archived websites is shown in Figure 2. The number of archived files has reached 8.5 billion. The NDL extracts some files in .pdf, .doc(x), and .xls(x) formats and adds metadata to them.[2]

[1] For other projects by the NDL for collecting digital materials, such as the e-legal deposit system and acquisition of digital dissertations, which are not mentioned in this paper, see Akiyama (2014), Uwabo and Aihara (2019).

[2] For further information on the history, system and legal aspects of WARP, see Murakami (2015).

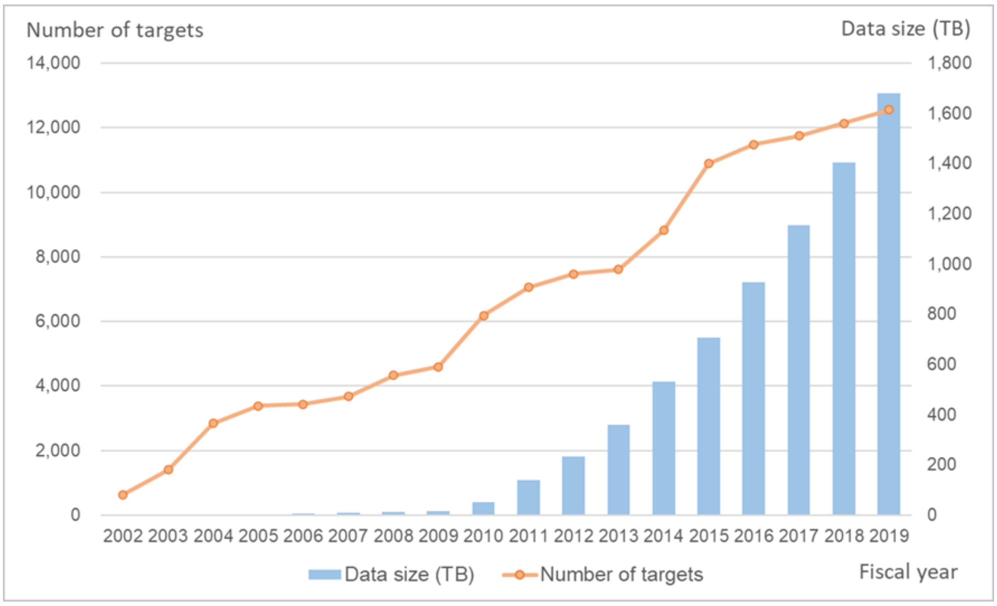


Figure 1. WARP's transition in data size and number of targets

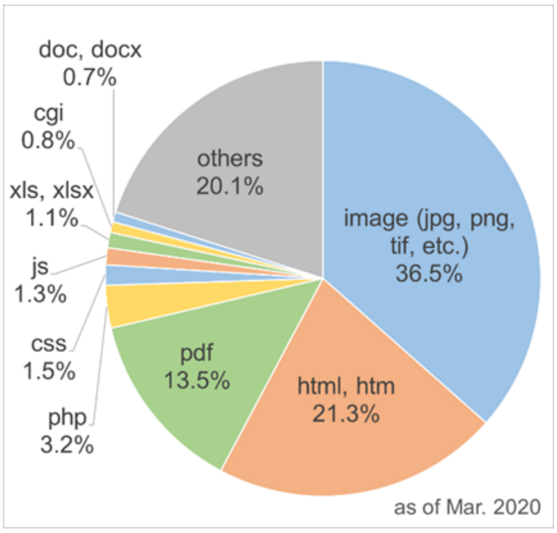


Figure 2. Proportion of formats in the archived files in WARP

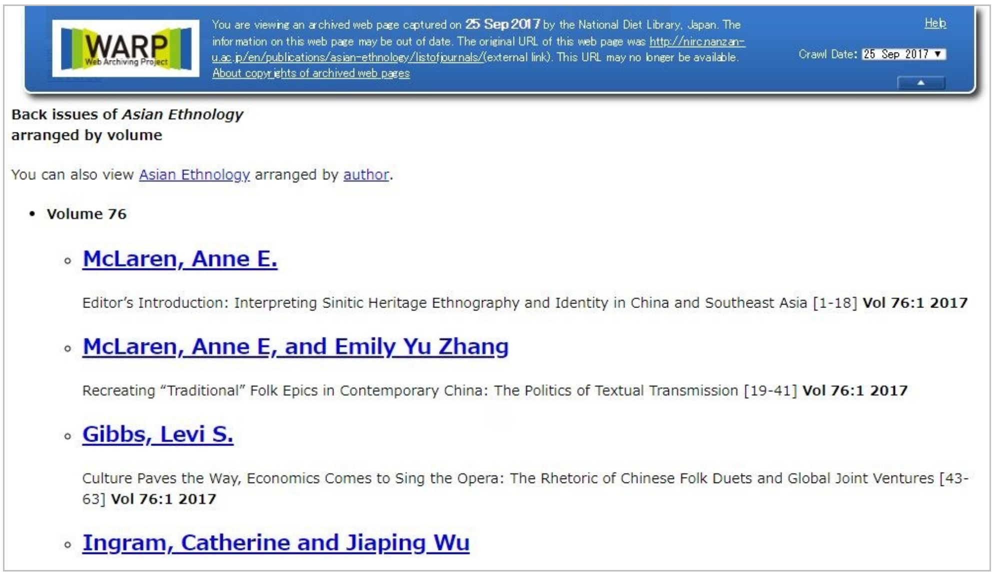
3 Extracting online publications embedded in websites

3.1 Criteria for Selection

It is impossible to extract all the embedded publications from WARP because of the huge amount. Thus, we select publications to extract by prioritization. Specifically, our main targets of extraction are white papers, annual reports, yearbooks, handbooks, official journals, public relations magazines, bulletins, academic journals, technical reports and research reports. Serial publications that were once published in printed form (and collected by the NDL) are especially highly prioritized. Publications related to the Great East Japan Earthquake also have high priority in our list. If there is a request to extract some publication from inside or outside of the NDL, we add them to the list after taking into account the need and workload. One thing to be noted is that publications that are available in repositories of universities and other research institutions are not targets for extraction. This is because WARP is not archiving online publications that are stored in institutional repositories, as those publications are already guaranteed to be accessible to the public for the long term by the institutions.

3.2 Workflow

The first step is specifying publications to extract. We find the webpages where the publications to extract are embedded from the archived websites of WARP (Figure 3). Then, we extract the information of the publications such as anchor texts and URLs, using a tool that utilizes Visual Basic for Applications (VBA) in MS Excel.



You are viewing an archived web page captured on 25 Sep 2017 by the National Diet Library, Japan. The information on this web page may be out of date. The original URL of this web page was http://nirc.nanzan-u.ac.jp/en/publications/asian-ethnology/listofjournals/(external link). This URL may no longer be available.
About copyrights of archived web pages

Back issues of *Asian Ethnology*
arranged by volume

You can also view Asian Ethnology arranged by author.

- Volume 76
  - McLaren, Anne E.

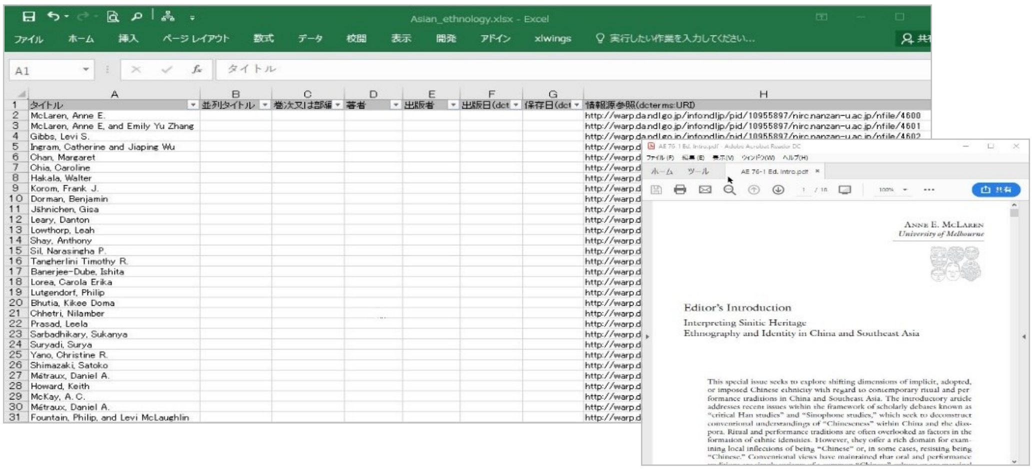
    Editor's Introduction: Interpreting Sinitic Heritage Ethnography and Identity in China and Southeast Asia [1-18] **Vol 76:1 2017**
  - McLaren, Anne E, and Emily Yu Zhang

    Recreating "Traditional" Folk Epics in Contemporary China: The Politics of Textual Transmission [19-41] **Vol 76:1 2017**
  - Gibbs, Levi S.

    Culture Paves the Way, Economics Comes to Sing the Opera: The Rhetoric of Chinese Folk Duets and Global Joint Ventures [43-63] **Vol 76:1 2017**
  - Ingram, Catherine and Jiaping Wu

Figure 3. Specifying a publication to extract on WARP

Secondly, we prepare the metadata of the publications. With the information extracted in the previous step, we access each URL of the publication and prepare metadata on a MS Excel sheet (Figure 4). We create metadata with a VBA tool following the National Diet Library Dublin Core Metadata Description (DC-NDL), which is standardized by the NDL based on the Dublin Core. Of the 66,000 metadata that we prepare per year, almost half of them are in-house production and the other half are prepared by an outside supplier that is specialized in cataloguing.



	A	B	C	D	E	F	G	H
1	タイトル	並列タイトル	巻次又は部編	著者	出版者	出版日(dct	保存日(dct	情報源参照(dcterms:URI)
2	McLaren, Anne E.							http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc.nanzan-u.ac.jp/nfile/4600
3	McLaren, Anne E, and Emily Yu Zhang							http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc.nanzan-u.ac.jp/nfile/4601
4	Gibbs, Levi S.							http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc.nanzan-u.ac.jp/nfile/4602
5	Ingram, Catherine and Jiaping Wu							http://warp.d
6	Chan, Margaret							http://warp.d
7	Chia, Caroline							http://warp.d
8	Hakala, Walter							http://warp.d
9	Korom, Frank J.							http://warp.d
10	Dorman, Benjamin							http://warp.d
11	Jähnichen, Gisa							http://warp.d
12	Leary, Danton							http://warp.d
13	Lowthorp, Leah							http://warp.d
14	Shay, Anthony							http://warp.d
15	Sil, Narasingha P.							http://warp.d
16	Tangherlini Timothy R.							http://warp.d
17	Banerjee-Dube, Ishita							http://warp.d
18	Lorea, Carola Erika							http://warp.d
19	Lutgendorf, Philip							http://warp.d
20	Bhutia, Kikee Doma							http://warp.d
21	Chhetri, Nilamber							http://warp.d
22	Prasad, Leela							http://warp.d
23	Sarbadhikary, Sukanya							http://warp.d
24	Suryadi, Surya							http://warp.d
25	Yano, Christine R.							http://warp.d
26	Shimazaki, Satoko							http://warp.d
27	Métraux, Daniel A.							http://warp.d
28	Howard, Keith							http://warp.d
29	McKay, A. C.							http://warp.d
30	Métraux, Daniel A.							http://warp.d
31	Fountain, Philip, and Levi McLaughlin							http://warp.d

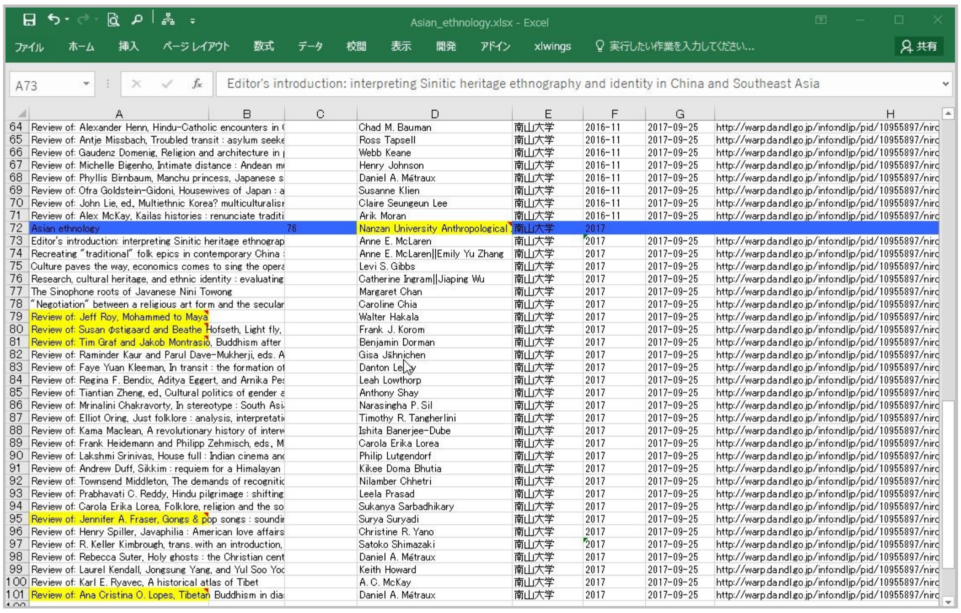
Anne E. McLaren
University of Melbourne

Editor's Introduction
Interpreting Sinitic Heritage
Ethnography and Identity in China and Southeast Asia

This special issue seeks to explore shifting dimensions of implicit, adopted, or imposed Chinese ethnicity with regard to contemporary ritual and performance traditions in China and Southeast Asia. The introductory article addresses recent issues within the framework of scholarly debates known as "critical Han studies" and "Sinophone studies," which seek to deconstruct conventional understandings of "Chineseness" within China and the diaspora. Ritual and performance traditions are often overlooked as factors in the formation of ethnic identities. However, they offer a rich domain for examining local inflections of being "Chinese" or, in some cases, resisting being "Chinese." Conventional views have maintained that oral and performance

Figure 4. Preparing metadata of a publication

In the third step, we check the metadata created both in house and by the outside supplier (Figure 5). VBA tools are used again in this step so that we can check the metadata efficiently.



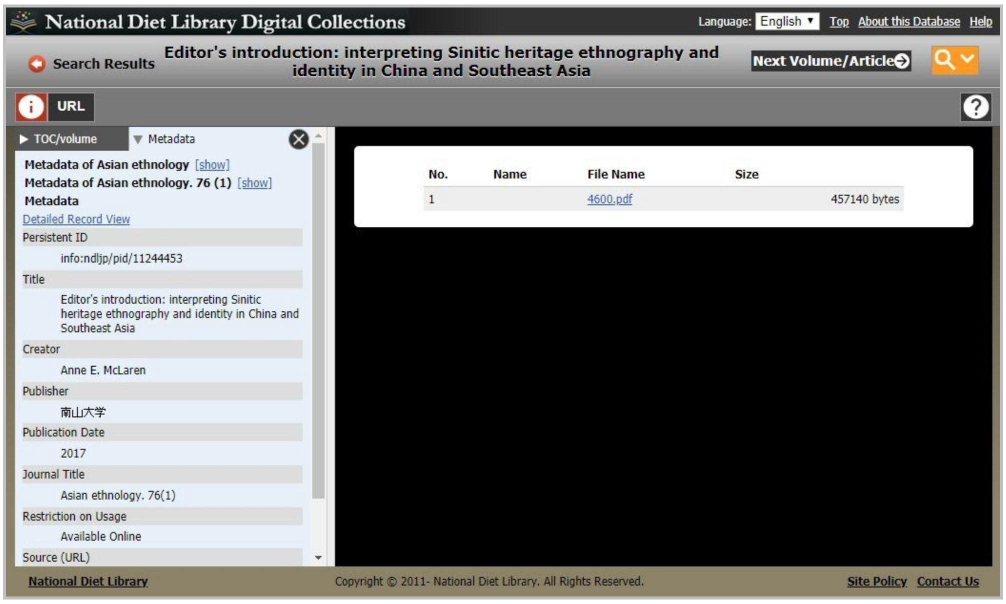
Asian_ethnology.xlsx - Excel
ファイル ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示 開発 アドイン xlwings 実行したい作業を入力してください… 共有
A73

Editor's introduction: interpreting Sinitic heritage ethnography and identity in China and Southeast Asia

	A	B	C	D	E	F	G	H
64	Review of: Alexander Henn, Hindu-Catholic encounters in (			Chad M. Bauman	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
65	Review of: Antje Missbach, Troubled transit : asylum seeke			Ross Tapsell	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
66	Review of: Gaudenz Domenig, Religion and architecture in			Webb Keane	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
67	Review of: Michelle Bigenho, Intimate distance : Andean m			Henry Johnson	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
68	Review of: Phyllis Birnbaum, Manchu princess, Japanese s			Daniel A. Métraux	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
69	Review of: Ofra Goldstein-Gidoni, Housewives of Japan : a			Susanne Klien	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
70	Review of: John Lie, ed., Multiethnic Korea? multiculturalisr			Claire Seungeun Lee	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
71	Review of: Alex McKay, Kailas histories : renunciate traditi			Arik Moran	南山大学	2016-11	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
72	Asian ethnology	76		Nanzan University Anthropological	南山大学	2017		
73	Editor's introduction: interpreting Sinitic heritage ethnograp			Anne E. McLaren	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
74	Recreating "traditional" folk epics in contemporary China :			Anne E. McLaren||Emily Yu Zhang	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
75	Culture paves the way, economics comes to sing the opera			Levi S. Gibbs	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
76	Research, cultural heritage, and ethnic identity : evaluating			Catherine Ingram||Jiaping Wu	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
77	The Sinophone roots of Javanese Nini Towong			Margaret Chan	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
78	"Negotiation" between a religious art form and the secular			Caroline Chia	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
79	Review of: Jeff Roy, Mohammed to Maya			Walter Hakala	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
80	Review of: Susan Østigaard and Beathe Hofseth, Light fly,			Frank J. Korom	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
81	Review of: Tim Graf and Jakob Montrasio, Buddhism after			Benjamin Dorman	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
82	Review of: Raminder Kaur and Parul Dave-Mukherji, eds. A			Gisa Jähnichen	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
83	Review of: Faye Yuan Kleeman, In transit : the formation of			Danton Le[illegible]y	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
84	Review of: Regina F. Bendix, Aditya Eggert, and Arnika Pes			Leah Lowthorp	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
85	Review of: Tiantian Zheng, ed., Cultural politics of gender a			Anthony Shay	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
86	Review of: Mrinalini Chakravorty, In stereotype : South Asi			Narasingha P. Sil	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
87	Review of: Elliot Oring, Just folklore : analysis, interpretati			Timothy R. Tangherlini	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
88	Review of: Kama Maclean, A revolutionary history of interw			Ishita Banerjee-Dube	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
89	Review of: Frank Heidemann and Philipp Zehmisch, eds., M			Carola Erika Lorea	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
90	Review of: Lakshmi Srinivas, House full : Indian cinema anc			Philip Lutgendorf	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
91	Review of: Andrew Duff, Sikkim : requiem for a Himalayan			Kikee Doma Bhutia	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
92	Review of: Townsend Middleton, The demands of recognitic			Nilamber Chhetri	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
93	Review of: Prabhavati C. Reddy, Hindu pilgrimage : shifting			Leela Prasad	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
94	Review of: Carola Erika Lorea, Folklore, religion and the so			Sukanya Sarbadhikary	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
95	Review of: Jennifer A. Fraser, Gongs & pop songs : soundir			Surya Suryadi	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
96	Review of: Henry Spiller, Javaphilia : American love affairs			Christine R. Yano	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
97	Review of: R. Keller Kimbrough, trans. with an introduction,			Satoko Shimazaki	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
98	Review of: Rebecca Suter, Holy ghosts : the Christian cent			Daniel A. Métraux	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
99	Review of: Laurel Kendall, Jongsung Yang, and Yul Soo Yoc			Keith Howard	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
100	Review of: Karl E. Ryavec, A historical atlas of Tibet			A. C. McKay	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc
101	Review of: Ana Cristina O. Lopes, Tibetan Buddhism in dia			Daniel A. Métraux	南山大学	2017	2017-09-25	http://warp.da.ndl.go.jp/info:ndljp/pid/10955897/nirc

Figure 5. Checking created metadata

Lastly, online publications are uploaded to the NDL Digital Collections with metadata and finally become accessible and discoverable for users (Figure 6). If the archived website in which a publication was originally embedded is available online on WARP, the publication is also available online on the NDL Digital Collections.



Editor's introduction: interpreting Sinitic heritage ethnography and identity in China and Southeast Asia

Metadata of Asian ethnology [show]
Metadata of Asian ethnology. 76 (1) [show]
Metadata
Detailed Record View

Persistent ID
info:ndljp/pid/11244453

Title
Editor's introduction: interpreting Sinitic heritage ethnography and identity in China and Southeast Asia

Creator
Anne E. McLaren

Publisher
南山大学

Publication Date
2017

Journal Title
Asian ethnology. 76(1)

Restriction on Usage
Available Online

Source (URL)

No.	Name	File Name	Size
1		4600.pdf	457140 bytes

National Diet Library   Copyright © 2011- National Diet Library. All Rights Reserved.   Site Policy Contact Us

Figure 6. Uplaoding an extracted publication with metadata

As a result of implementing the workflow in our day-to-day business, the number of extracted online publications has been growing as shown in Figure 7. In recent years we have extracted 66,000 publications per year, and 565,962 publications were discoverable on the NDL Digital Collections as of March 2020.

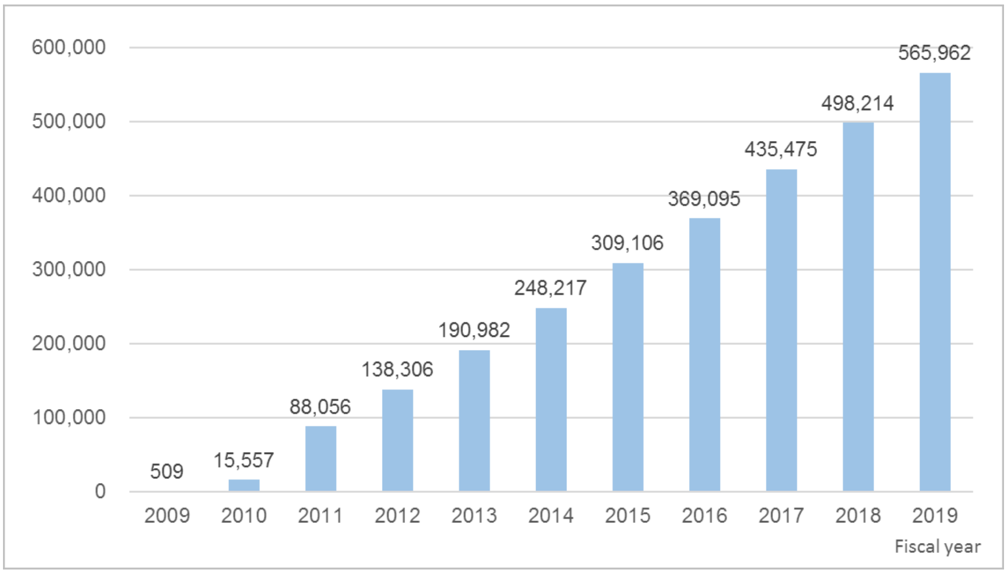
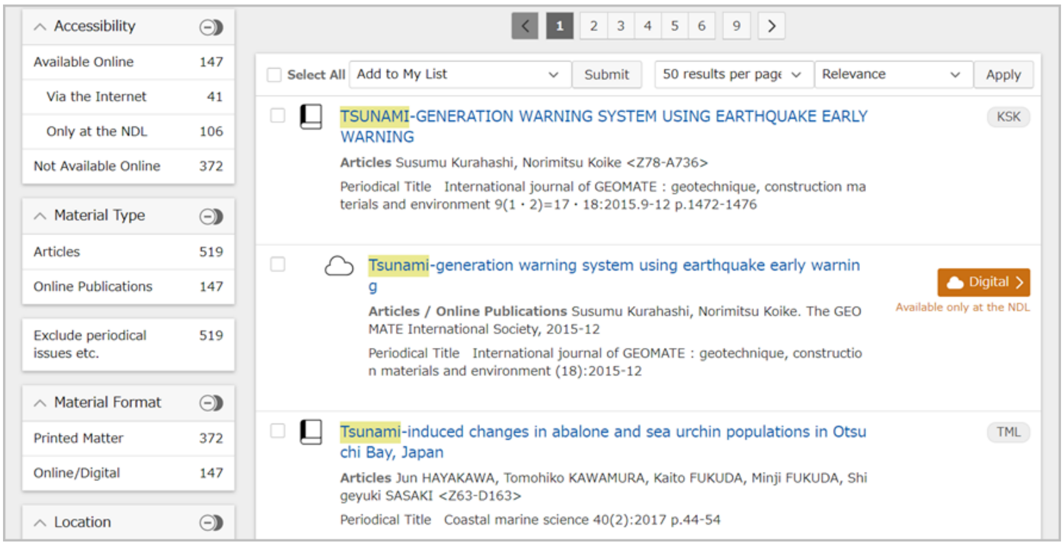


Figure 7. The number of online publications on the NDL Digital Collections

3.3 Metadata linking to the online catalogue

Metadata of online publications in the NDL Digital Collections is linked via an API to NDL Online, the online catalogue of the NDL. Users can find online publications on the online catalogue just as they would look for printed ones. We are still making efforts to enhance the accessibility to online publications on NDL Online, and one of the recent examples of our efforts is grouping indication. If the NDL has both printed and online versions of the same publication, both versions are displayed as one group on NDL Online (Figure 8).



Accessibility	
Available Online	147
Via the Internet	41
Only at the NDL	106
Not Available Online	372
Material Type	
Articles	519
Online Publications	147
Exclude periodical issues etc.	519
Material Format	
Printed Matter	372
Online/Digital	147
Location	

□ TSUNAMI-GENERATION WARNING SYSTEM USING EARTHQUAKE EARLY WARNING

Articles Susumu Kurahashi, Norimitsu Koike <Z78-A736>

Periodical Title International journal of GEOMATE : geotechnique, construction materials and environment 9(1・2)=17・18:2015.9-12 p.1472-1476

□ Tsunami-generation warning system using earthquake early warning

Articles / Online Publications Susumu Kurahashi, Norimitsu Koike. The GEOMATE International Society, 2015-12

Periodical Title International journal of GEOMATE : geotechnique, construction materials and environment (18):2015-12

□ Tsunami-induced changes in abalone and sea urchin populations in Otsuchi Bay, Japan

Articles Jun HAYAKAWA, Tomohiko KAWAMURA, Kaito FUKUDA, Minji FUKUDA, Shigeyuki SASAKI <Z63-D163>

Periodical Title Coastal marine science 40(2):2017 p.44-54

Figure 8. Search results on NDL Online

4 Challenges

As described above, we established a workflow for extracting publications embedded in websites and making them more easily accessible, but there is still much room for improvement. We will briefly refer to the three major challenges we are facing.

4.1 Improved efficiency

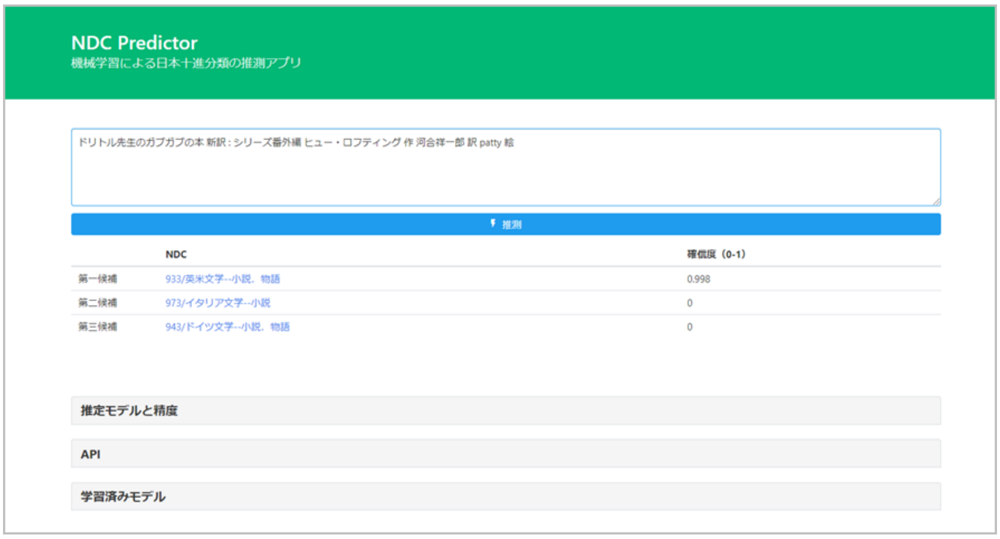
At the moment, we are extracting only a limited number of publications from WARP. One of the main challenges we face is overcoming this limitation through improved efficiency in our current workflow, in which most parts are performed manually.

Seeking a solution for this challenge, we have since 2019 been developing a tool to generate metadata for online publications. The tool analyses text on the front page of an article in the PDF format and extracts the title and author's name(s) of the article. Although the development of the tool is still at the experimental phase, we expect these kinds of tools and other technologies related to machine learning will be of great help in improving the efficiency of our workflow, and it is becoming more important for us to acquire programming knowledge and skills in order to cope with this kind of challenge.

4.2 Enrichment of metadata

The second challenge is the enrichment of metadata. The metadata we are currently adding to online publications have a limited number of elements. We are now considering adding elements such as subject, classification or keywords to the metadata.

One of the tools suggested as a solution to this challenge is the NDC Predictor (Figure 9). It is an application which was developed by the Research and Development for Next-Generation Systems Office of the NDL using machine learning technology in 2019. It predicts a class number in the Nippon Decimal Classification for publications from a bibliographic record. We are considering applying this kind of tool to enrich our metadata for online publications.



NDC Predictor
機械学習による日本十進分類の推測アプリ

ドリトル先生のガブガブの本 新訳：シリーズ番外編 ヒュー・ロフティング 作 河合祥一郎 訳 patty 絵

	NDC	確信度（0-1）
第一候補	933/英米文学--小説．物語	0.998
第二候補	973/イタリア文学--小説	0
第三候補	943/ドイツ文学--小説．物語	0

Figure 9. The NDC Predictor

4.3 Archiving moving image files

Although we currently archive only online publications that are in static and relatively simple formats, we recognize the challenge to archive more complex and varied formats, such as moving image files.

Our immediate concern with this challenge is archiving videos posted on YouTube by Japanese public agencies. The most feasible way for that is delivering data from the agencies to the NDL, but this method would impose an additional workload on both the agencies and the NDL. Thus, we have begun considering direct downloads of videos from YouTube. To establish a new framework for direct downloads, we will keep analysing precedent for video archiving by other institutions such as the National Archives, UK and keep considering legal aspects including YouTube’s terms of service, which does not allow users to download videos.

5 Conclusion

As we described above, we have established a workflow for extracting online publications embedded in Japanese websites and enhanced their discoverability. However, the amount and the variety of those publications are rapidly increasing. We will keep making efforts to improve the efficiency of our workflow and to establish new frameworks for preserving complex publications including moving image files in order to cope with this rapid increase.

References

AKIYAMA, Tsutomu (2014) Struggles of the National Diet Library in Collecting Online Publications in Japan. Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 87 - Information Technology with Preservation and Conservation and National Libraries. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. http://library.ifla.org/id/eprint/886, (accessed 2020-06-26).

MURAKAMI, Kosuke (2015) Lessons learned from twelve years' operation of the Web Archiving Project (WARP). Paper presented at: IFLA WLIC 2015 - Cape Town, South Africa in Session 90 - Preservation and Conservation with Information Technology. http://library.ifla.org/id/eprint/1089, (accessed 2020-06-26).

UWABO, Yoshie and AIHARA, Masaki (2019) Grey Literature in Japanese Academia—Efforts by the National Diet Library to Acquire and Provide Access to the Journals and Conference Proceedings of Academic Societies and Associations. Paper presented at: IFLA WLIC 2019 - Athens, Greece - Libraries: dialogue for change in Session S12 - Serials and Other Continuing Resources with National Libraries. In: Grey Literature: Scholarly Communication in a Digital World. http://library.ifla.org/id/eprint/2719, (accessed 2020-06-26).