

IFLA PRESENTS

Professional Unit Virtual Events



IFLA Section
Preservation and
Conservation



IFLA Section
Information Technology

IFLA PRESENTS

Extracting Online Publications Embedded in Websites: NDL Initiatives and Challenges

INOIE Nobuaki

SHIBATA Masaki

KUDO Tetsuro

National Diet Library (NDL), Japan

18/11/2020



privacy

This is a recorded presentation.

Recording, presentation slides and full paper will be posted on PCS & ITS publication page.

Questions or comments? Please type into the chat or Q&A box.

The talk is GDPR-compliant

IFLA and ZOOM privacy policies:

<https://www.ifla.org/data-protection-policy>

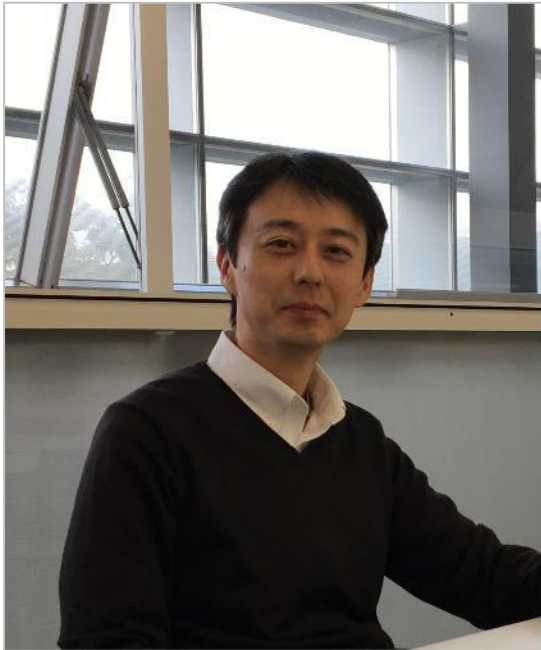
<https://zoom.us/privacy>

Questions regarding privacy:

professionalsupport@ifla.org



speaker



INOIE Nobuaki

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



Overview: Extracting Online Publications Project

- Background: Web Archiving Project (WARP) in the NDL
- Scope
- Workflow
- Metadata linking to the online catalog
- Challenges and future plans

Background (WARP) (1/4) - Scope

- “WARP” = Web ARchiving Project, since 2002
- Archiving websites of
 - Japanese public agencies comprehensively based on the National Diet Library Law
 - private organizations selectively based on the permissions of their webmasters (political parties, private universities, international events, etc.)



The screenshot shows the WARP website interface. At the top, there is a logo for WARP (Web Archiving Project) and the text "国立国会図書館 インターネット資料収集保存事業". Below the logo is a navigation menu with "Language: English", "FAQ", "Help", and "Site Map". The main content area is divided into several sections:

- Quick Search:** A search bar with a "Search" button and a link to "Advanced Search".
- Collection Search:** A search bar with a "Search" button and a link to "Advanced Search".
- Monthly Special:** A section titled "Where the WARP link is (in Japanese)" featuring two thumbnails. The first is for the "Prime Minister's Office of Japan 'Past Websites'" (2 Apr. 2020) and the second is for the "Japan Atomic Energy Agency 'Fukushima Nuclear Accident Archives'" (13 Jul. 2020).
- News:** A section with two news items: "Monthly access ranking (Sep 2020) has been updated." (12 Oct 2020) and "Monthly special 'Where the WARP link is' has been released." (1 Oct 2020).
- Recommended:** A section with a link to "Mechanism of Web Archiving".
- Web Archives of the World:** A section with a link to "Web Archives of the World".
- Featured Collections:** A section with a link to "Featured Collections".
- Use Cases (in Japanese):** A section with a link to "Use Cases (in Japanese)".
- Statistics:** A section with a link to "Statistics".
- Monthly Access Ranking:** A section with a table of rankings.

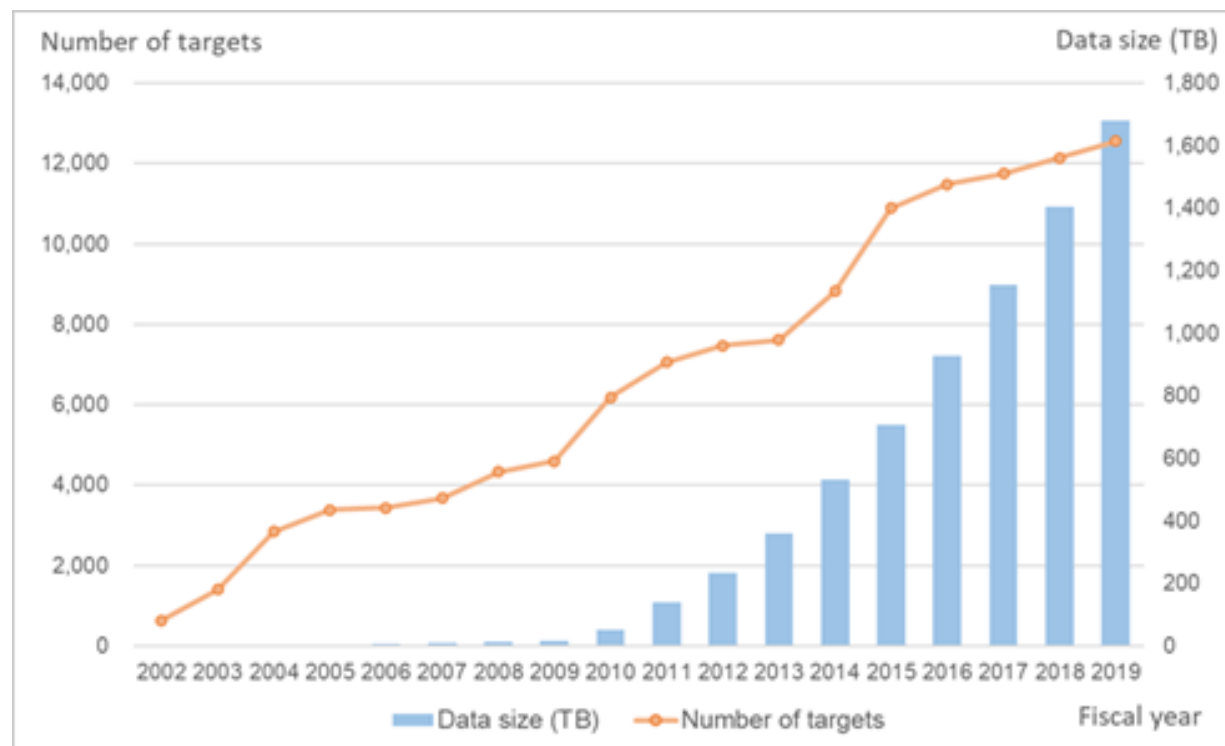
Rank	Organization	Access Count
1位	経済産業省 (2019年8月1日)	527,186
2位	文部科学省 (2019年8月1日)	524,907
3位	文部科学省 (2019年10月1日)	313,709
4位	日本貿易振興機構 (JETRO) (2016年1月1日)	274,799

WARP website
<http://warp.da.ndl.go.jp/>



Background (WARP) 2/4

- 1.7PB in data size
- 12,600 websites archived (as of Mar. 2020)
 - 5,800 public agencies and 6,700 private organizations
 - 85% available on the internet

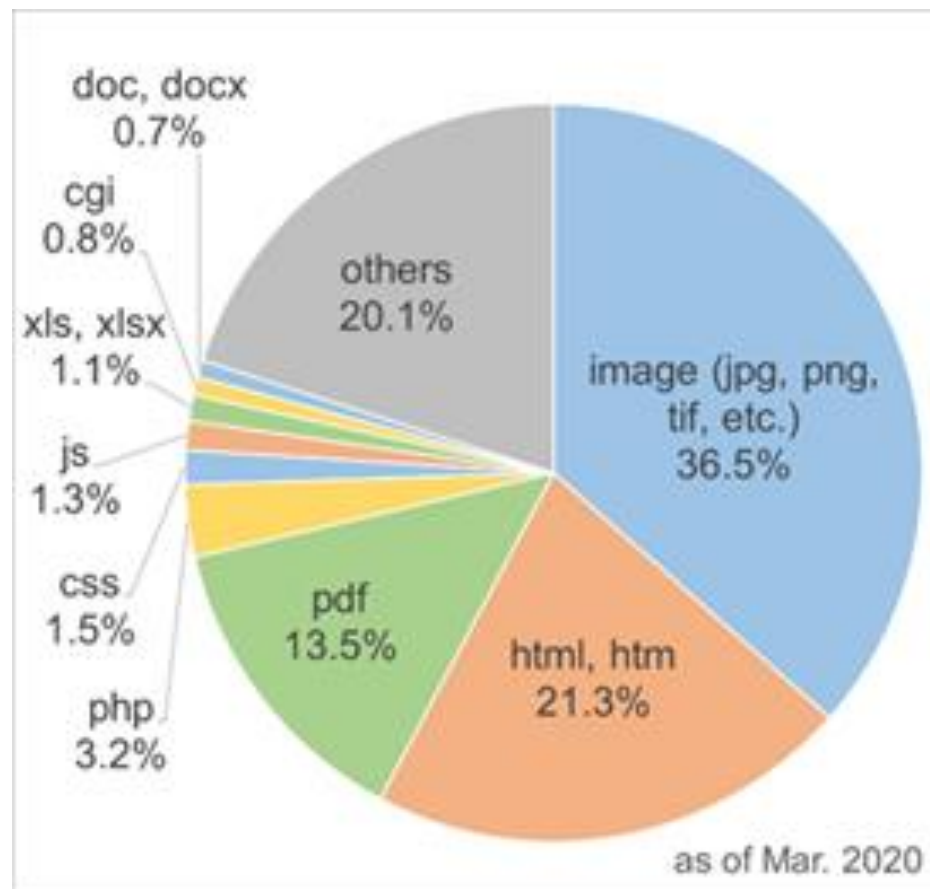


WARP's transition in data size and number of targets



Background (WARP) 3/4

- 8.5 billion files are archived
- Many of the .pdf, .doc(x) and .xls(x) files are e-books or e-zines



Proportion of formats
in WARP archived files

Background (WARP) 4/4

- Online publications in websites are
 - Embedded with other files
 - Lacking sufficient metadata



- Not efficiently searched for nor easily listed



- Need to extract online publications

Scope

- Main targets are
 - .pdf, .doc(x) and .xls(x) files on archived websites in WARP
 - Information-rich publications
 - White papers, annual reports, yearbooks, handbooks, official journals, public relations magazines, bulletins, academic journals, technical reports, research reports, etc.
 - Serial publications that were once published in printed form
 - Publications related to the Great East Japan Earthquake in 2011
- Publications in institutional repositories are out of scope

Workflow (2/6) – Specifying Publications

- Step 1: Specifying a publication to extract on archived websites
 - Specify a webpage with publications to extract
 - Extract anchor texts, URLs, etc. with VBA tools



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](#)

Workflow (3/6) – Creating Metadata

- Step 2: Creating metadata
 - Create metadata for a publication with VBA tools
 - Metadata follows the [National Diet Library Dublin Core Metadata Description \(DC-NDL\)](#)
 - Half of metadata by in-house production (The other half by an outside supplier)

The image displays three overlapping screenshots illustrating the metadata creation workflow. On the left is an Excel spreadsheet titled 'Asian_ethnology.xlsx' with columns for author names and URLs. The middle screenshot shows a web browser displaying the 'Editor's Introduction' page for the journal 'Asian Ethnology', featuring the title 'Interpreting Swiss Heritage Ethnography and Identity in China and Southeast Asia' and the author 'Anne E. McLaren, University of Melbourne'. On the right is a screenshot of a metadata entry form, likely for the National Diet Library, with fields for title, author, and URL, and a '保存' (Save) button.

Screenshots of creating metadata

Workflow (4/6) – Checking Metadata

- Step 3: Checking metadata
 - NDL staff check metadata created both by in-house production and by outside supplier
 - VBA tools are used in checking

	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.dnli.go.jp/infondljp/pid/10955897/nirc	
65	Review of: Antje Missbach, Troubled transit: asylum seek		Ross Tapsell	南山大学	2016-11	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
69	Review of: Ota Goldstein-Gidon, Housewives of Japan: a		Susanne Klien	南山大学	2016-11	2017-09-25	http://warp.dnli.go.jp/infondljp/pid/10955897/nirc	
70	Review of: John Lie, ed., Multiethnic Korea? multiculturaliz		Claire Seungeun Lee	南山大学	2016-11	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
72	Asian ethnology	75	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.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
75	Culture paves the way, economics comes to sing the oper		Levi S. Gibbs	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
77	The Sinophone roots of Javanese Nini Towong		Margaret Chan	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
79	Review of: Jeff Froy, Mohammed to Maya		Walter Hakala	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/pid/10955897/nirc	
80	Review of: Susan ostegaard and Beathe Holseth, Light fly,		Frank J. Korom	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
83	Review of: Faye Yuan Kleeman, In transit: the formation of		Danton Le	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/pid/10955897/nirc	
84	Review of: Regina F. Bendix, Aditya Eggert, and Annika Pei		Leah Lowthorp	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/pid/10955897/nirc	
85	Review of: Tiantian Zheng, ed., Cultural politics of gender c		Anthony Shay	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/pid/10955897/nirc	
86	Review of: Minalini Chakravorty, In stereotype: South Asi		Narasimha P. Sil	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
88	Review of: Kama Maclean, A revolutionary history of interv		Ishita Banerjee-Dube	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
90	Review of: Lakshmi Srinivas, House full: Indian cinema and		Philp Lutendorf	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
92	Review of: Townsend Middleton, The demands of recognitic		Nilamber Chhetri	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/pid/10955897/nirc	
93	Review of: Prabhavati C. Reddy, Hindu pilgrimage: shifting		Leela Prasad	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
95	Review of: Jennifer A. Fraser, Gongs & jpp songs: soundi		Surya Suryadi	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
99	Review of: Laurel Kendall, Jongsung Yang, and Yul Soo Yoo		Keith Howard	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/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.dnli.go.jp/infondljp/pid/10955897/nirc	
101	Review of: Ana Cristina O. Lopes, Tibetan Buddhism in dis		Daniel A. Métraux	南山大学	2017	2017-09-25	http://warp.dnli.go.jp/infondljp/pid/10955897/nirc	

Screenshot of an Excel sheet for checking metadata

Workflow (5/6) – Uploading to the NDL Collections

- Step 4: Uploading to the NDL Digital Collections
 - Publications uploaded with metadata
 - 85% available on the internet

The screenshot displays the National Diet Library Digital Collections interface. The main content area shows a search result for the article "Editor's introduction: interpreting Sinitic heritage ethnography and identity in China and Southeast Asia". A sidebar on the left provides detailed metadata for the document, including the title, creator (Anne E. McLaren), publisher (Nanshan University), and publication date (2017). A table in the main area lists the uploaded file:

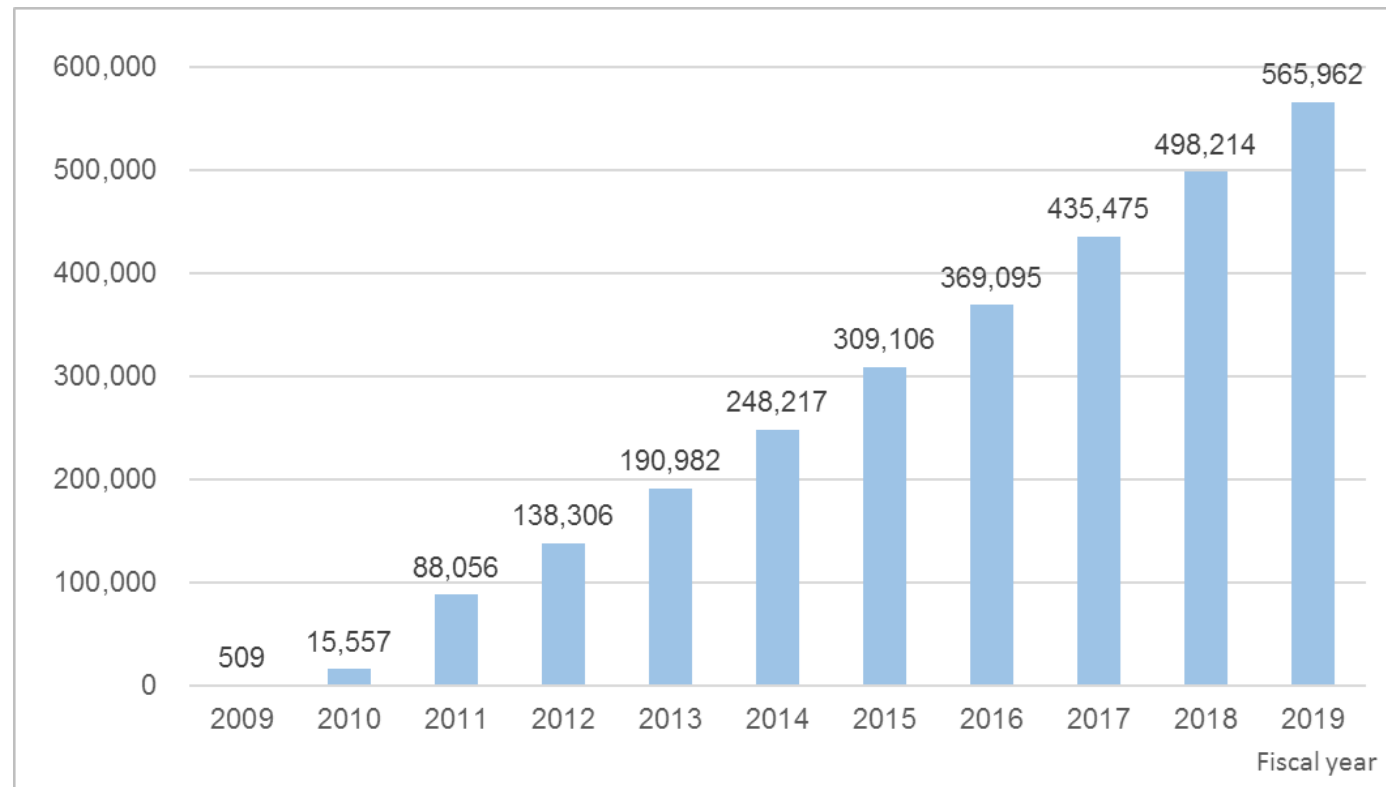
No.	Name	File Name	Size
1		4600.pdf	457140 bytes

The interface also includes a search bar, navigation links, and a footer with copyright information and site policies.

Screenshot of the NDL Digital Collections

Workflow (6/6) – Current Status

- 66,000 publications extracted per year
- 565,962 publications were discoverable on the NDL Digital Collections (as of Mar. 2020)



The number of online publications on the NDL Digital Collections

Metadata Linking to the Online Catalog

- Metadata of online publications linked to [NDL Online](#) via an API
- Grouping indication of printed and online versions

The screenshot displays a search results page on NDL Online. On the left, there are filter panels for Accessibility, Material Type, Material Format, and Location. The main content area shows three search results. The first result is a printed version of an article, indicated by a book icon and the label 'Printed ver.'. The second result is an online version, indicated by a cloud icon, a 'Digital' button, and the label 'Online ver.'. The third result is another printed version, indicated by a book icon and the label 'TML'. The search results include titles, authors, and periodical titles.

Material Type	Count
Available Online	147
Via the Internet	41
Only at the NDL	106
Not Available Online	372

Material Type	Count
Articles	519
Online Publications	147

Material Format	Count
Printed Matter	372
Online/Digital	147

Material Format	Count
Printed Matter	372
Online/Digital	147

Material Format	Count
Printed Matter	372
Online/Digital	147

Grouping indication in a search result on NDL Online

Challenges & Future Plans (1/3) - Improved Efficiency

- Most parts of the workflow done manually
 - = Limits number of publications extracted



- Metadata generation tool (experimental phase)
 - Being developed with Python from 2019
 - Extracts a title and name of author from the title page of a .pdf file
- Future plan: Semi-automated creation of basic metadata

Challenges & Future Plans (2/3) - Enrichment of Metadata

- Current metadata have only basic elements
 - Do not include subject, classification, keywords, etc.
- NDC Predictor (2019)
 - Developed with machine learning technology
 - Analyzes bibliographic records of publications
 - Predicts a classification based on Nippon Decimal Classification (NDC)

	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

NDC Predictor
<https://lab.ndl.go.jp/ndc/>



Challenges & Future Plans (3/3) - Archiving Moving Image Files

- Complex publications such as moving images are not archived
- Immediate concern: YouTube videos posted by Japanese public agencies



- Future plan: New framework for direct downloads from YouTube
- Need to keep analyzing precedent for video archiving
- Need to keep considering legal aspects

thank you

Check out our other IFLA events at
www.ifla.org/events/all

Visit our Unit's webpage to find out more about our work
at

<https://www.ifla.org/preservation-and-conservation>
<https://www.ifla.org/it>

