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:

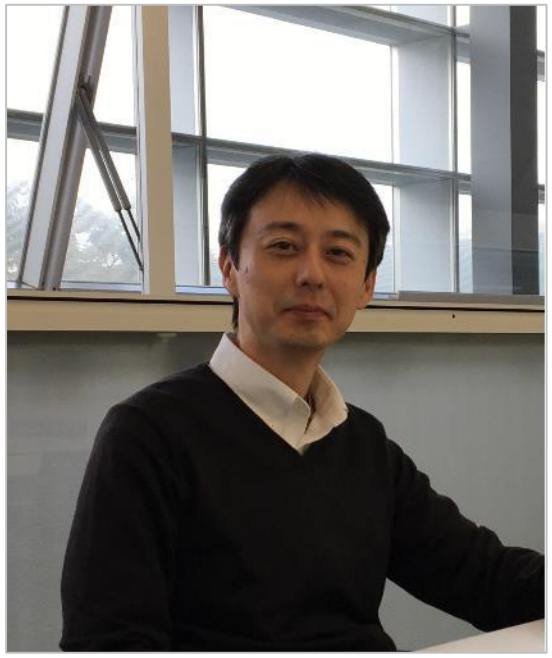
https://www.ifla.org/data-protection-policy

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

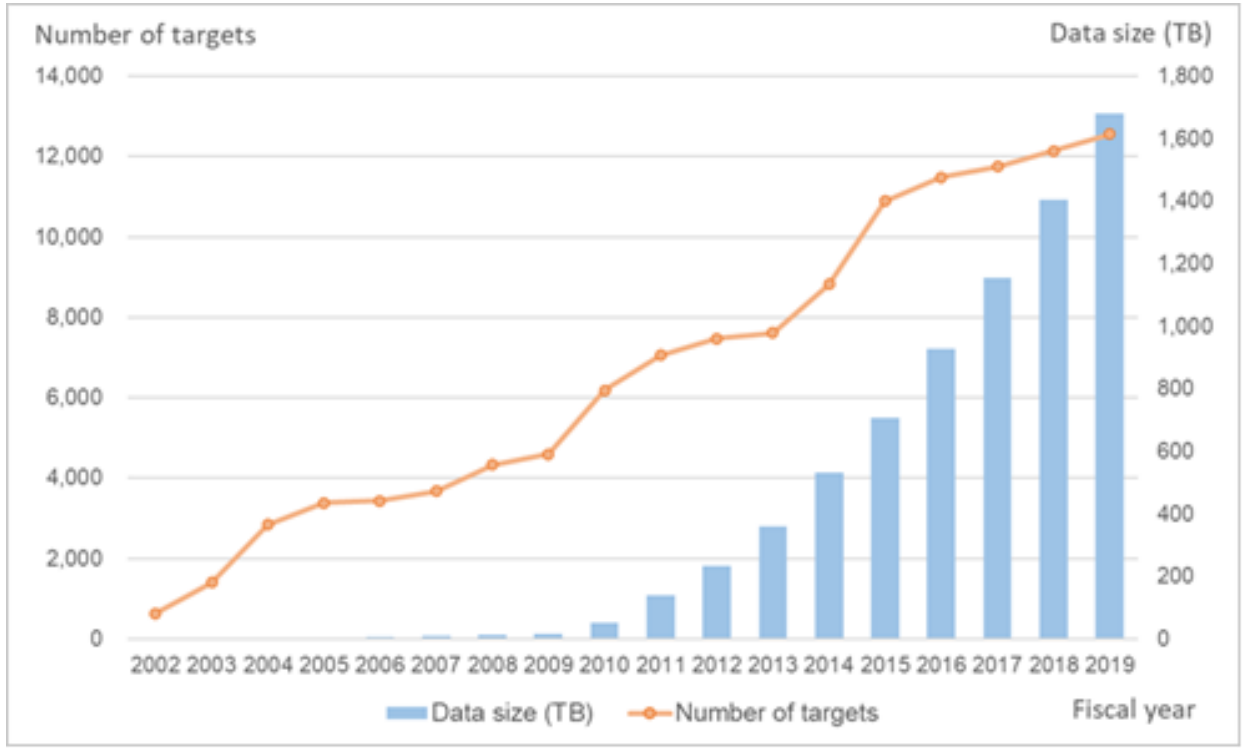
Background (WARP) (1/4) - Scope

- “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.)

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

Background (WARP) 2/4

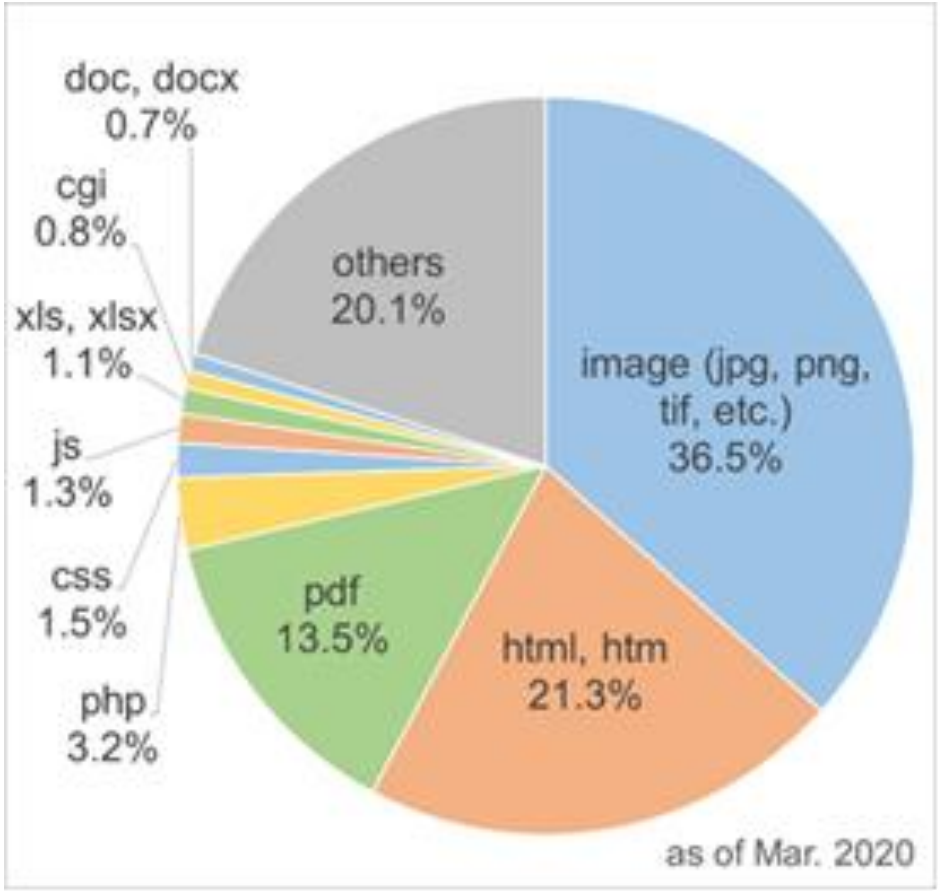
- 1.7PB in data size
- 12,600 websites archived (as of Mar. 2020)
  - 5,800 public agencies and 6,700 private organizations
  - 85% available on the internet



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

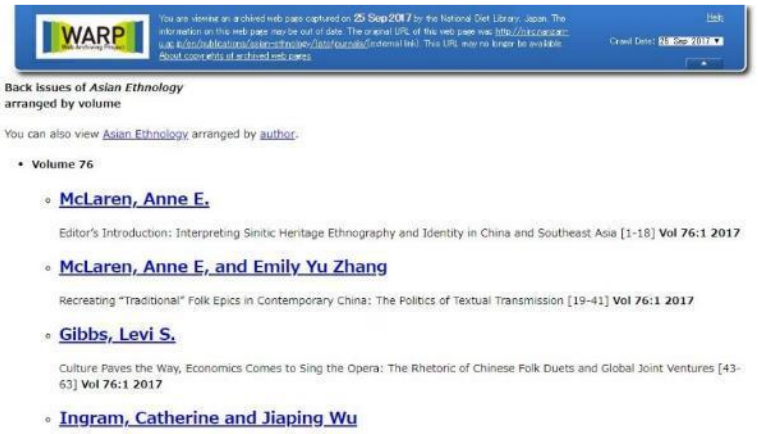
Scope

- 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

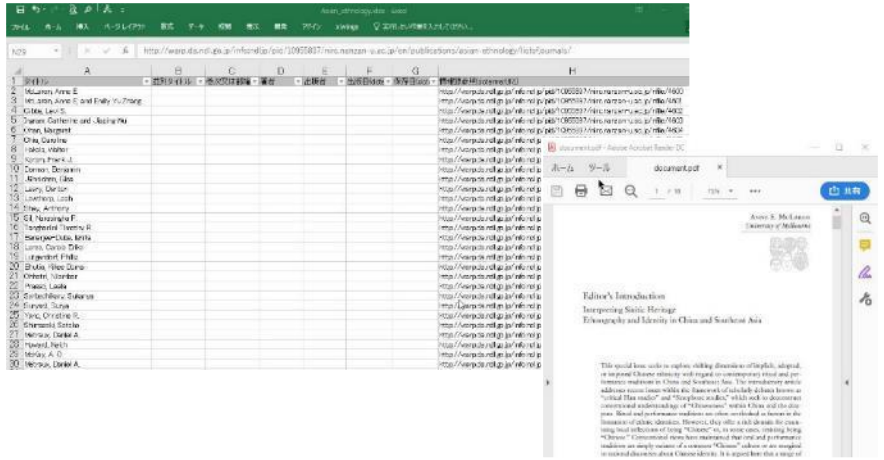
- Publications in institutional repositories are out of scope

Workflow (1/6) - Overview

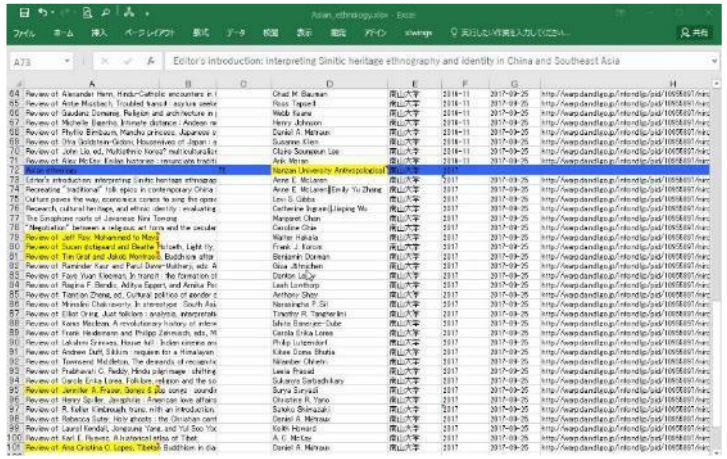
1. Specifying a publication to extract



2. Creating metadata



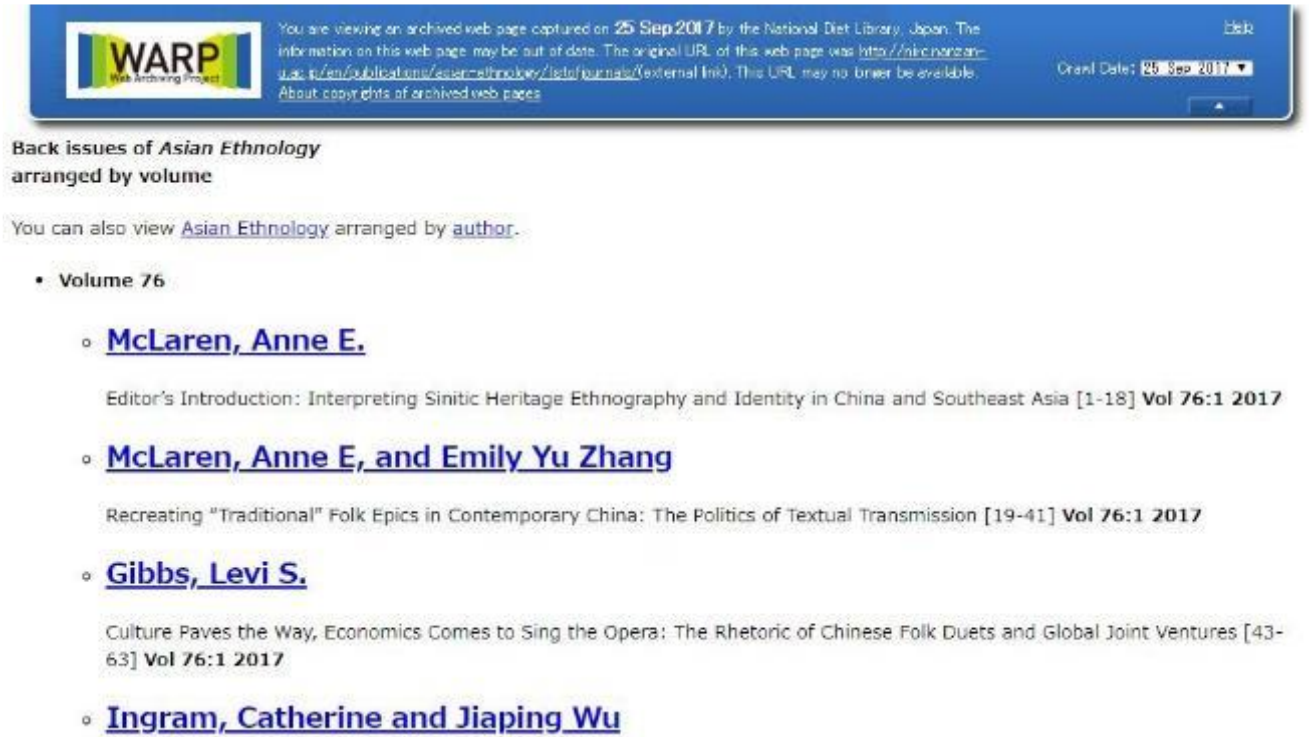
3. Checking metadata



4. Uploading to the NDL Digital Collections

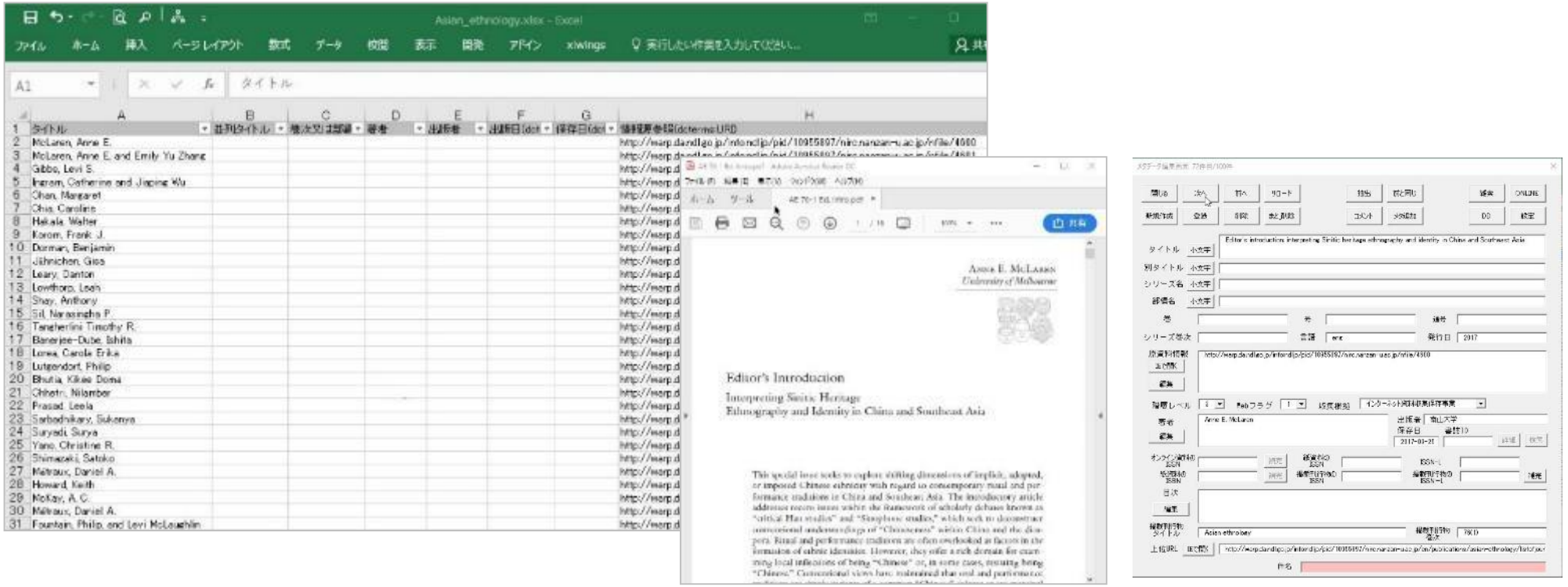
Workflow (2/6) – Specifying Publications

- 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



Workflow (3/6) – Creating Metadata

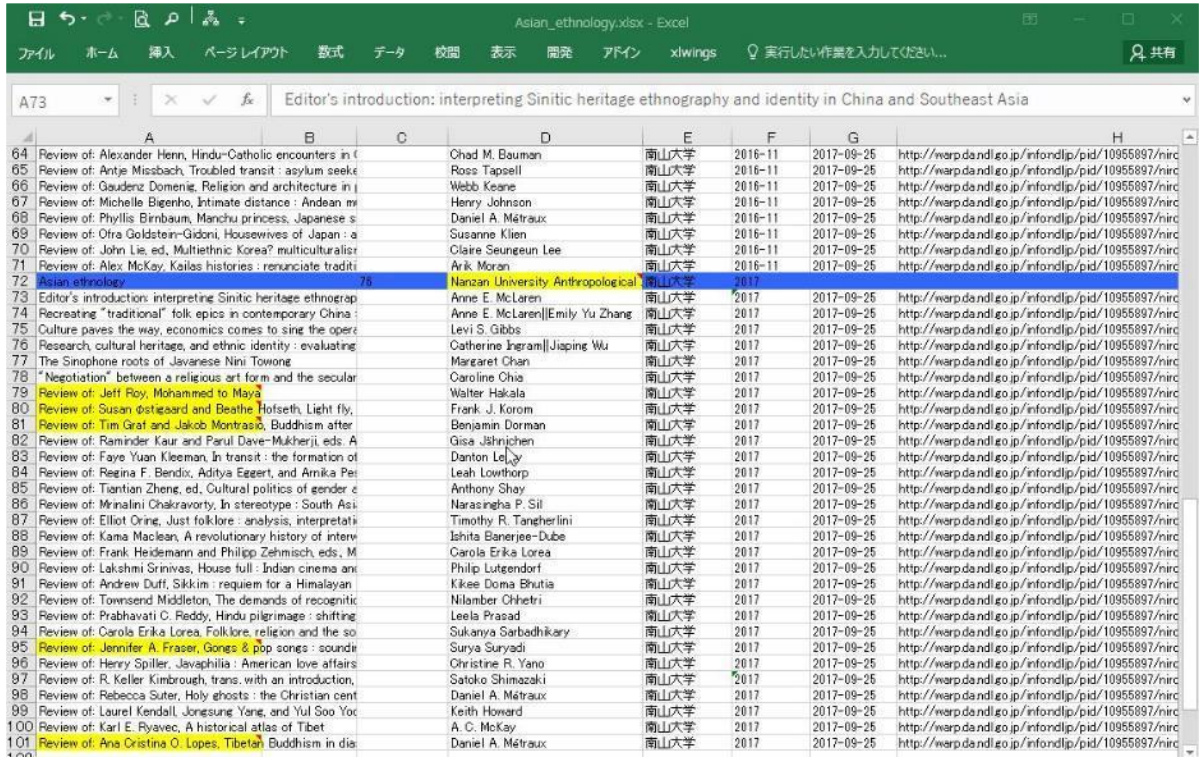
- Step 2: Creating metadata
  - Create metadata for a publication with VBA tools
  - Metadata follows the National Diet Library Dublin Core Metadata Description (DC-NDL)
  - Half of metadata by in-house production (The other half by an outside supplier)



Screenshots of creating metadata

Workflow (4/6) – Checking Metadata

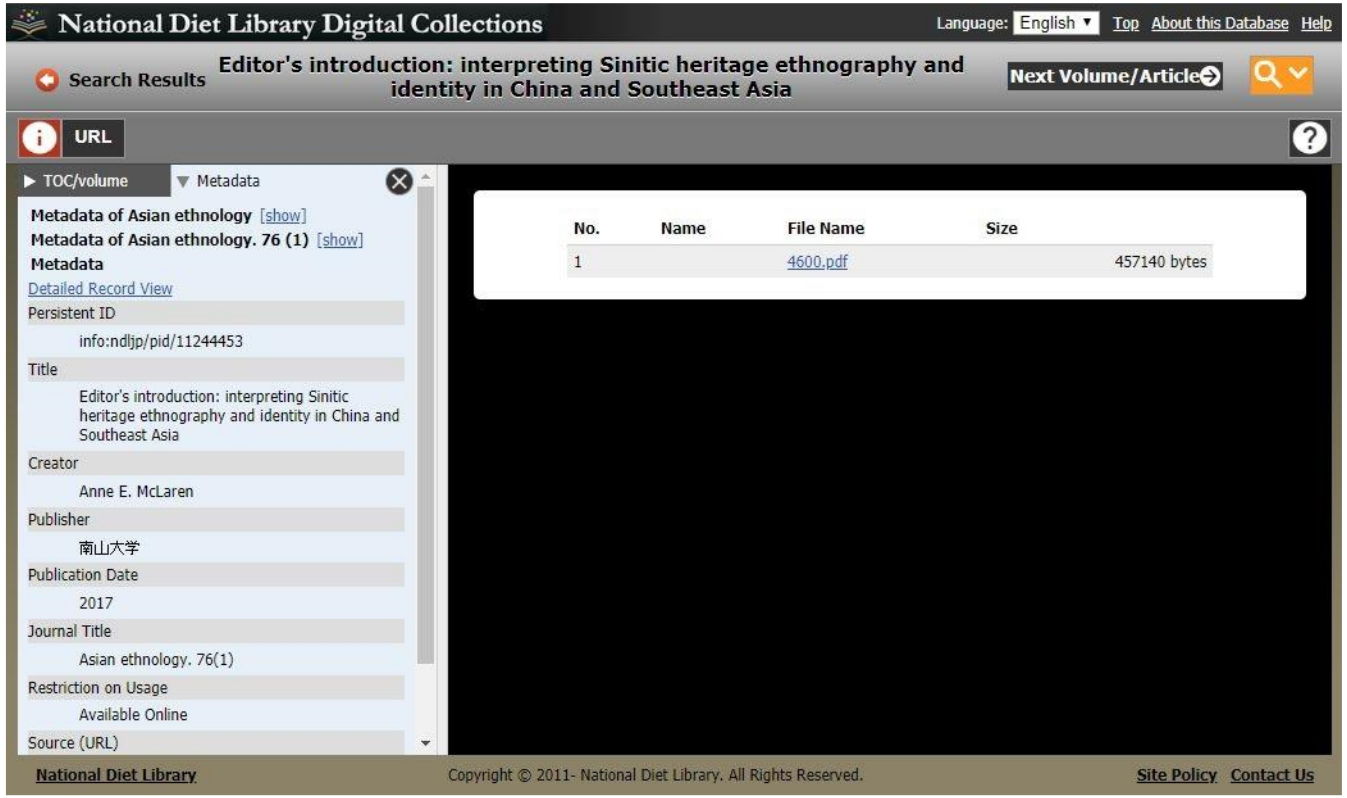
- Step 3: Checking metadata
  - NDL staff check metadata created both by in-house production and by outside supplier
  - VBA tools are used in checking



Screenshot of an Excel sheet for checking metadata

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

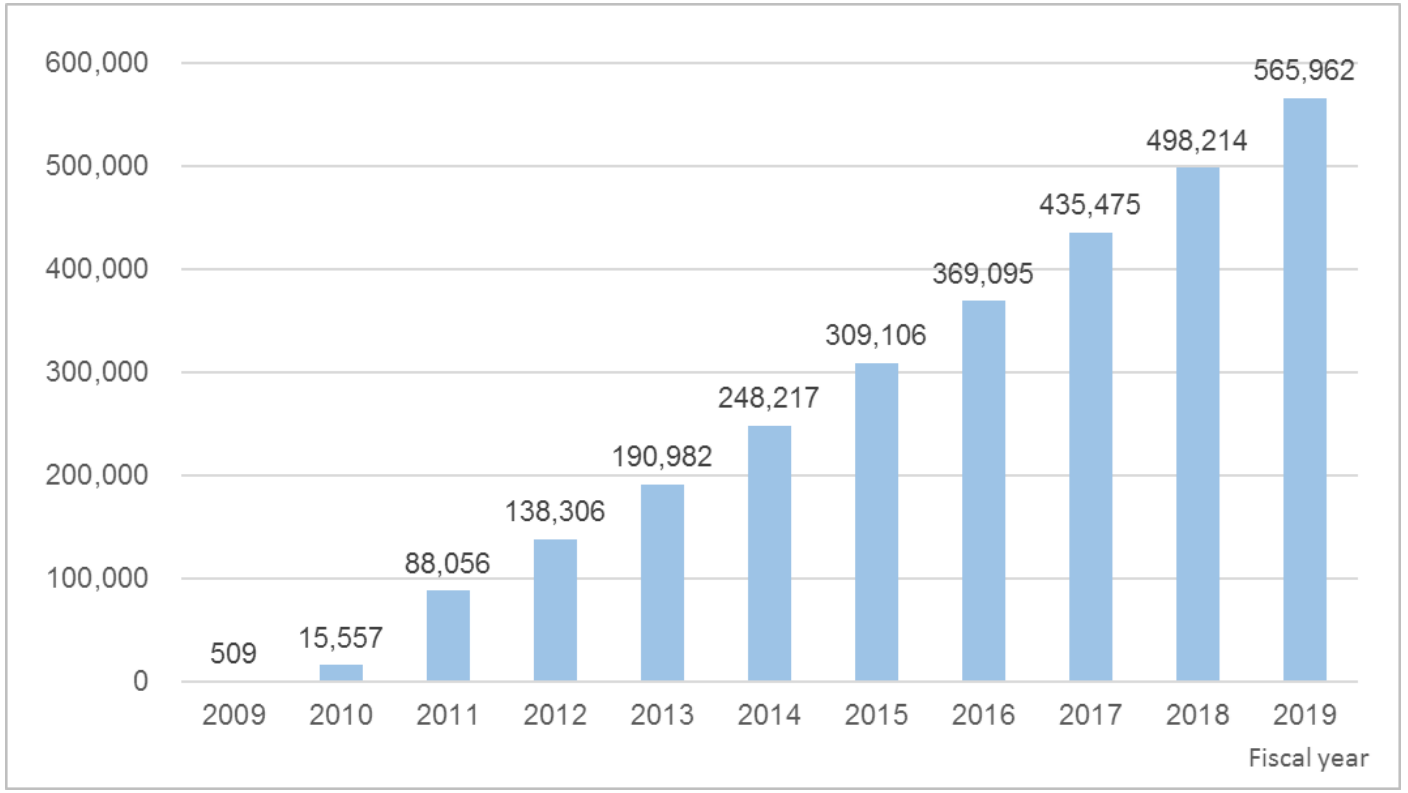
- Step 4: Uploading to the NDL Digital Collections
  - Publications uploaded with metadata
  - 85% available on the internet



Screenshot of the NDL Digital Collections

Workflow (6/6) – Current Status

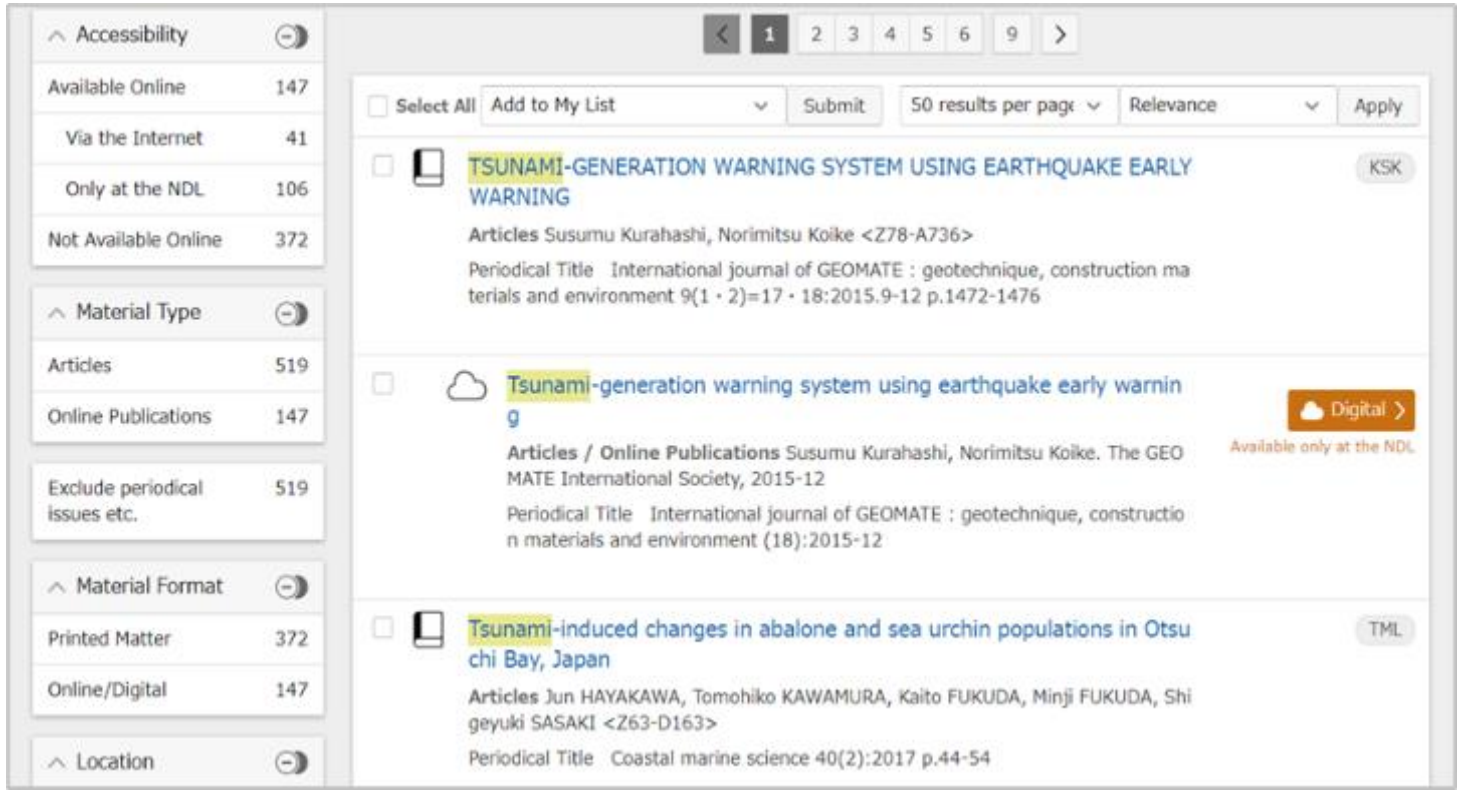
- 66,000 publications extracted per year
- 565,962 publications were discoverable on the NDL Digital Collections (as of Mar. 2020)



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



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 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