(Free) tools for data documentation and dissemination

(using DDI Codebook for microdata)

November 2022

Contacts:

Mehmood Asghar: masghar@worldbank.org

Olivier Dupriez: odupriez@worldbank.org

Aivin Solatorio: asolatorio@worldbank.org

Components

1. Guide on the use of metadata standards and schemas
   - Advocacy for the use of metadata standards/schemas
   - Technical guidelines
2. Cataloguing application: NADA (National Data Archive)
   - Open source
   - Multi-standard, including DDI Codebook
3. Metadata Editor
   - To replace the Nesstar Publisher; multi-standard including DDI Codebook
   - In testing phase
4. Research/exploratory work on improved data discoverability
   - Use of machine learning for semantic searchability, recommender system, metadata augmentation

1

A set of standards for multiple data types

DDI CODEBOOK 2.5 FOR MICRODATA

ISO19139/19115/19110 FOR GEOGRAPHIC DATASETS

DUBLIN CORE/ MARC21/BIBTEX FOR DOCUMENTS

DUBLIN CORE/IPTC FOR IMAGES

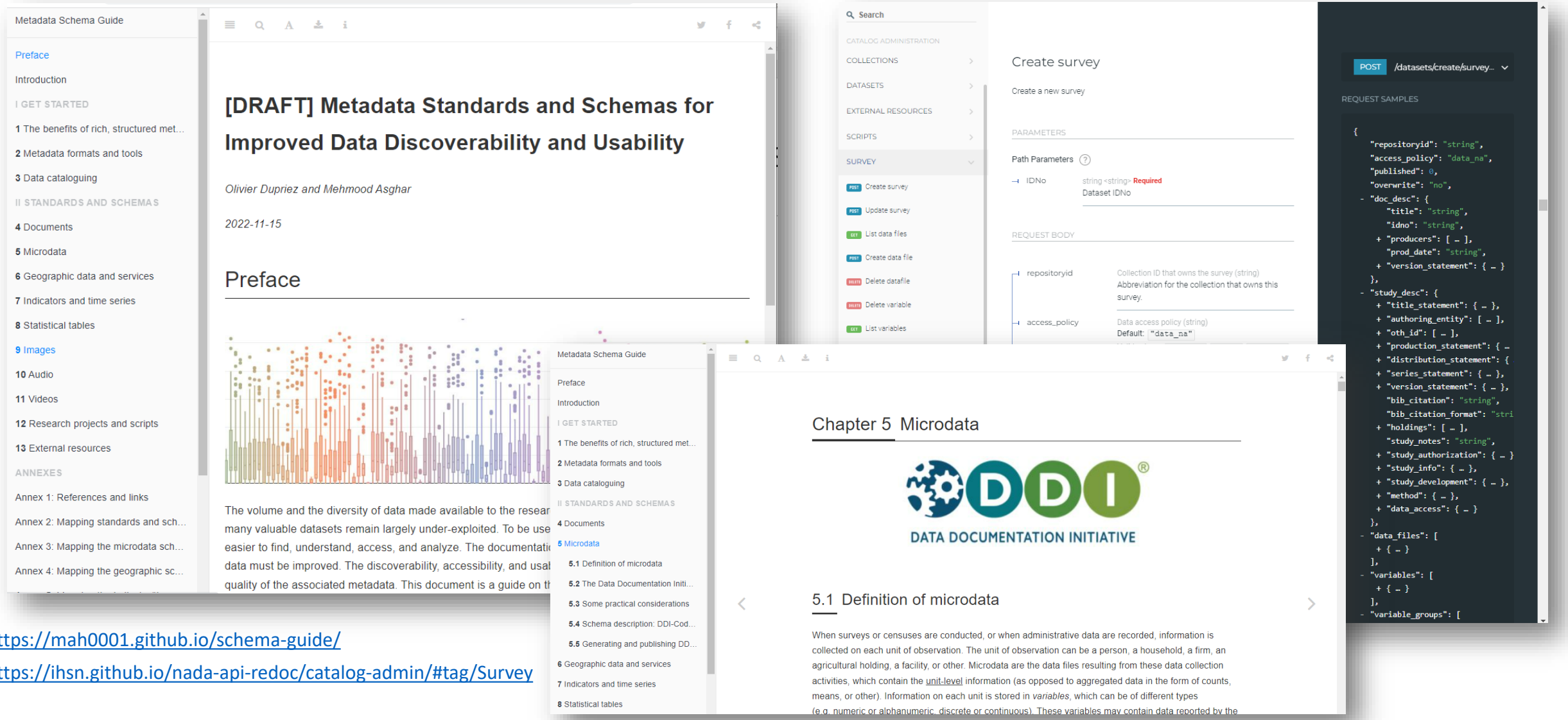
CUSTOM SCHEMAS FOR INDICATORS, TABLES, VIDEOS, REPRODUCIBLE SCRIPTS

ALL TYPES: MAPPING TO SCHEMA.ORG FOR SEARCH ENGINE OPTIMIZATION

DDI®
DATA DOCUMENTATION INITIATIVE

1

Guidelines on standards and schemas



https://mah0001.github.io/schema-guide/

https://ihsn.github.io/nada-api-redoc/catalog-admin/#tag/Survey

2

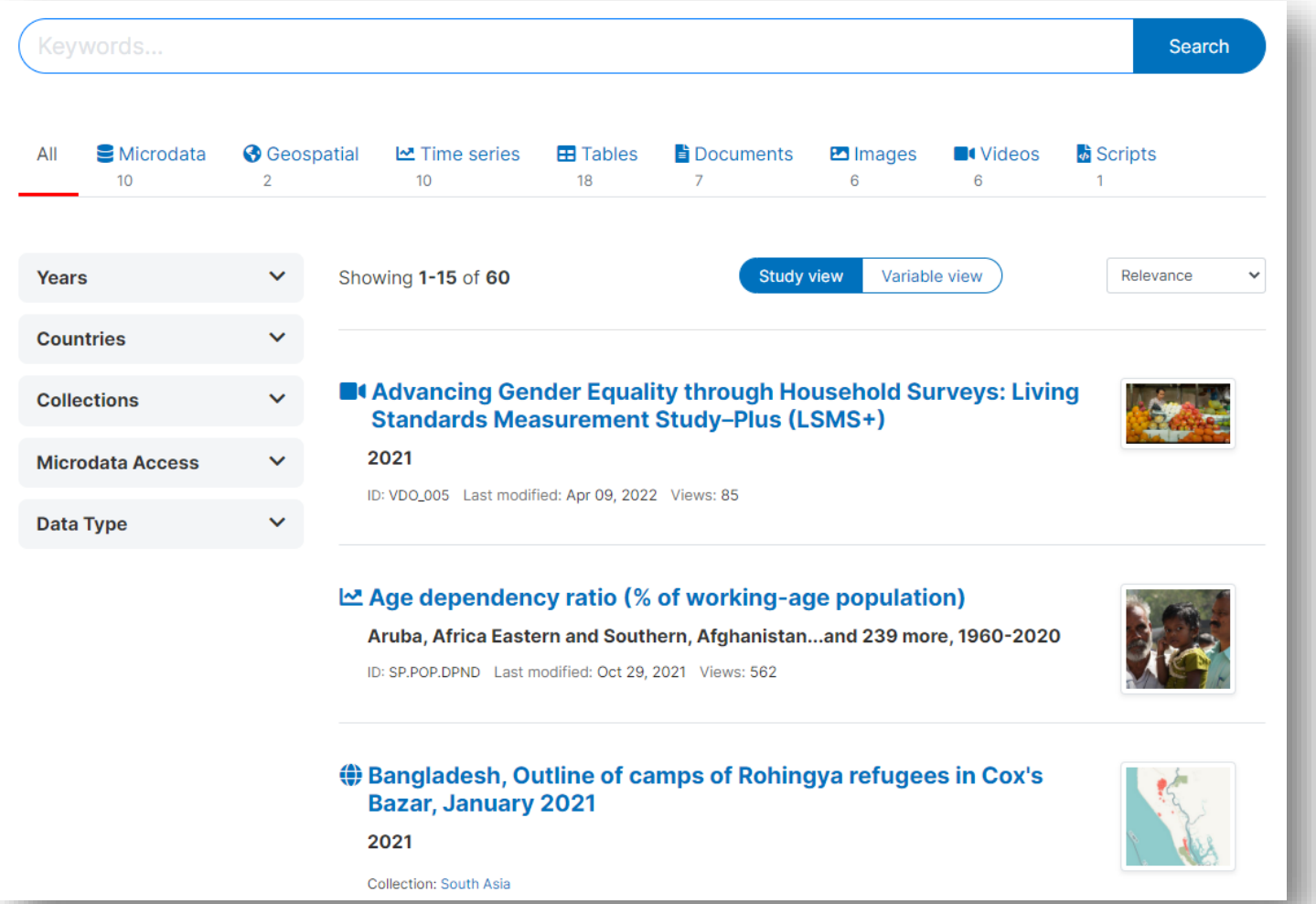
NADA: a multi-standard cataloguing tool

- Open source
  - https://github.com/ihsn/nada
  - https://ihsn.github.io/nada-documentation/
- Technologies: PHP; SQL (metadata); mongoDB (data); Solr (optional)
- Widgets for flexible additions (e.g., embed visualizations or data grids in catalog pages using JS libraries of your choice)
- R package and Python library for automation of tasks
  - To generate, harvest, publish, edit, augment, extract, import/export metadata
- Internationalization: embedded translation tool

2

NADA: multiple data types / standards

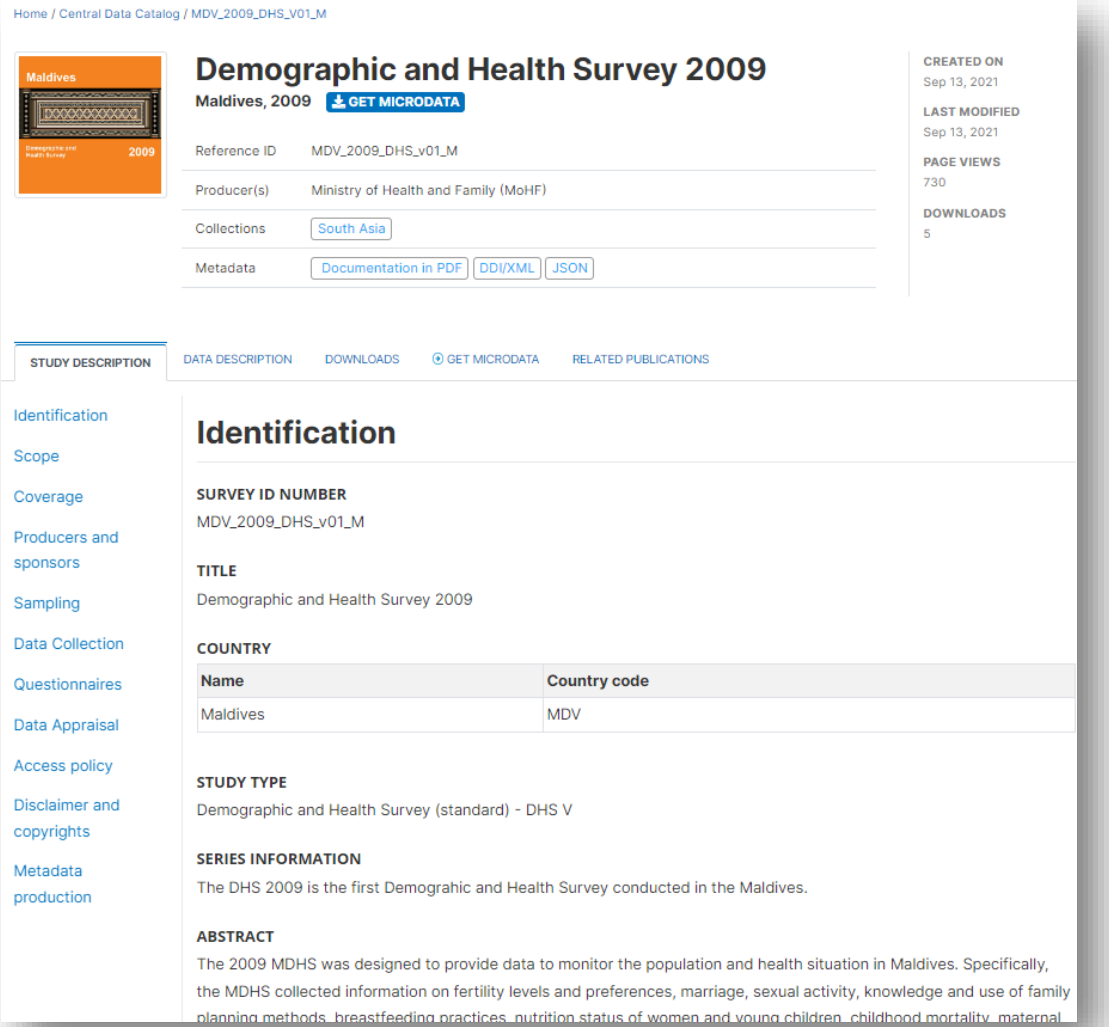
- Multi-standard (for multiple data types and reproducible scripts)
- Data organized by type and (optional) by collection
  - E.g., thematic collections, and/or collections by data producer



2

NADA: rich, structured metadata

- Rich metadata
- Flexible data access control
- Can federate the catalog administration
  - Sub collections
  - Fine-grained roles/permissions system
- Embedded SEO (Google data structure / schema.org)
- Metadata accessible via API
- R package for automation of tasks
- Responsive design (bootstrap4)

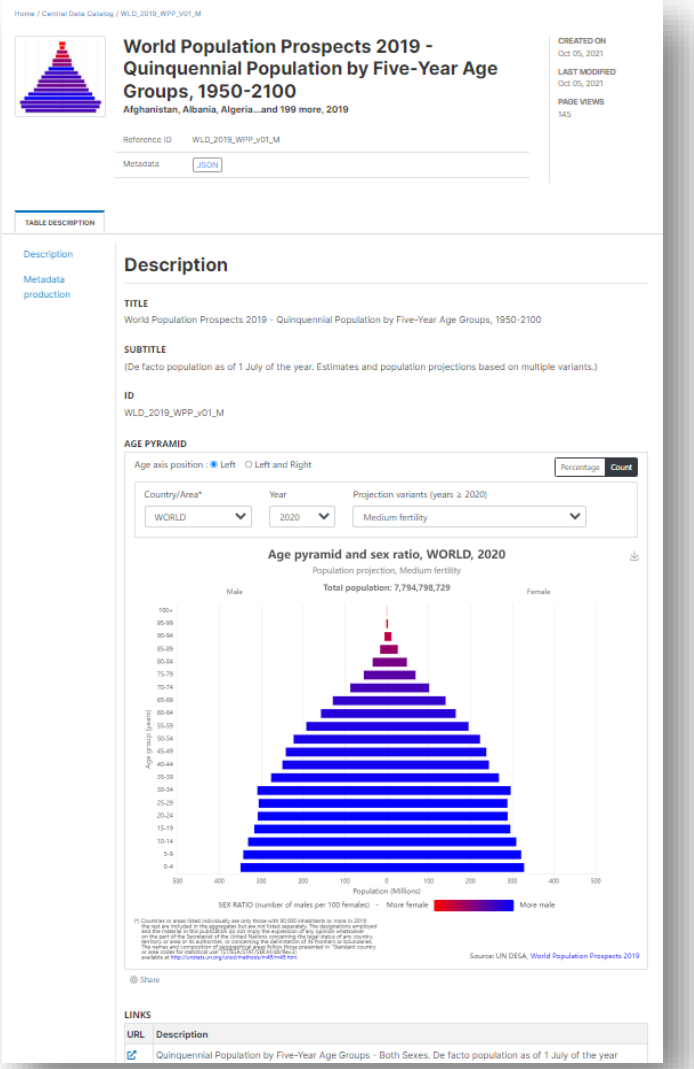


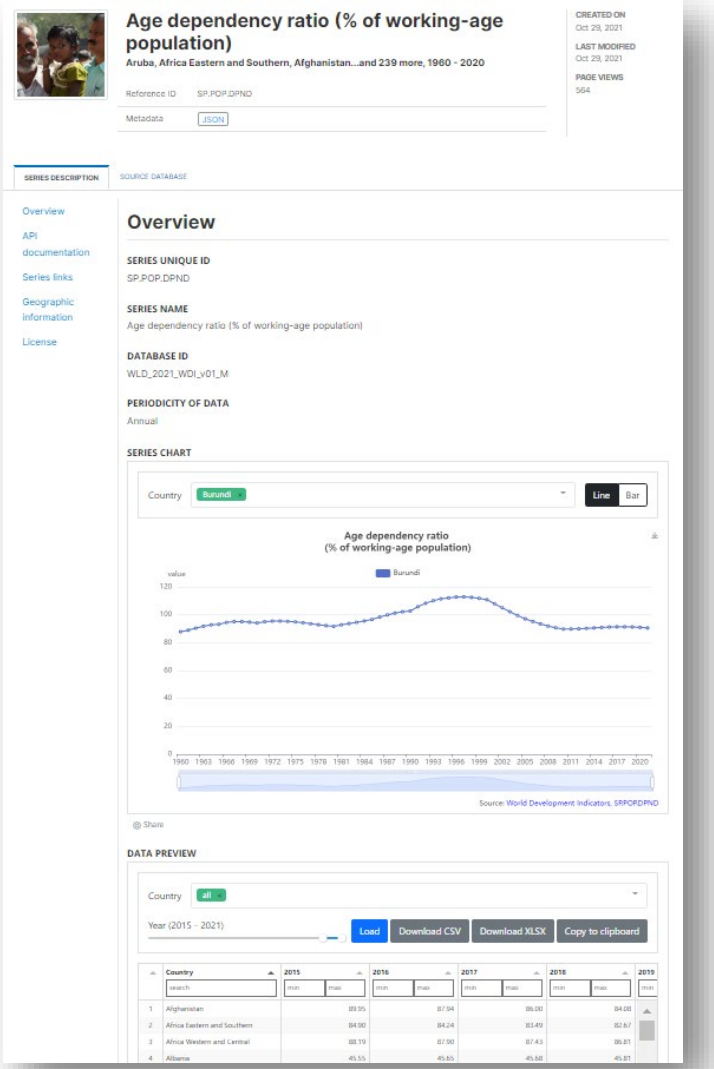
2

NADA: embed visualizations and more

Embed visualizations, data grids, maps, etc. created with other applications, e.g., eCharts

- Mostly for indicators
- Requires data accessible via API (NADA API or external API)
- Can build your own tools using NADA as back-end

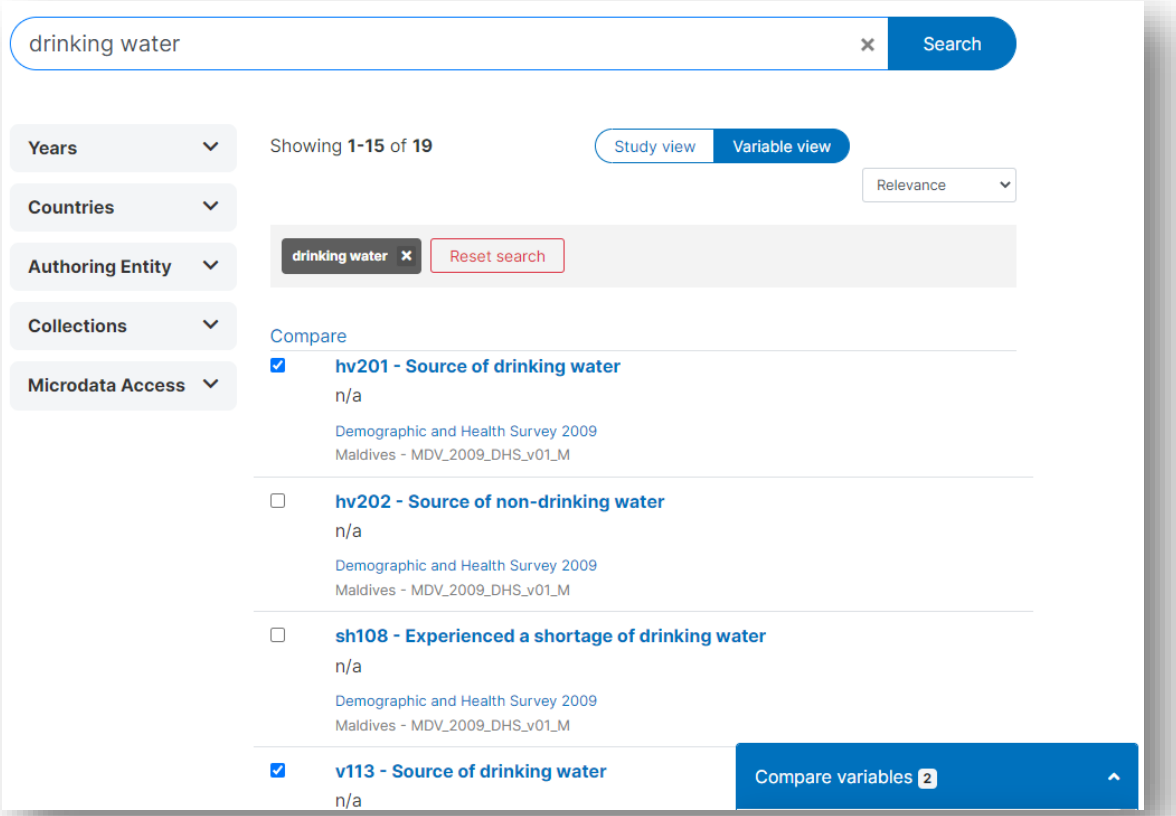


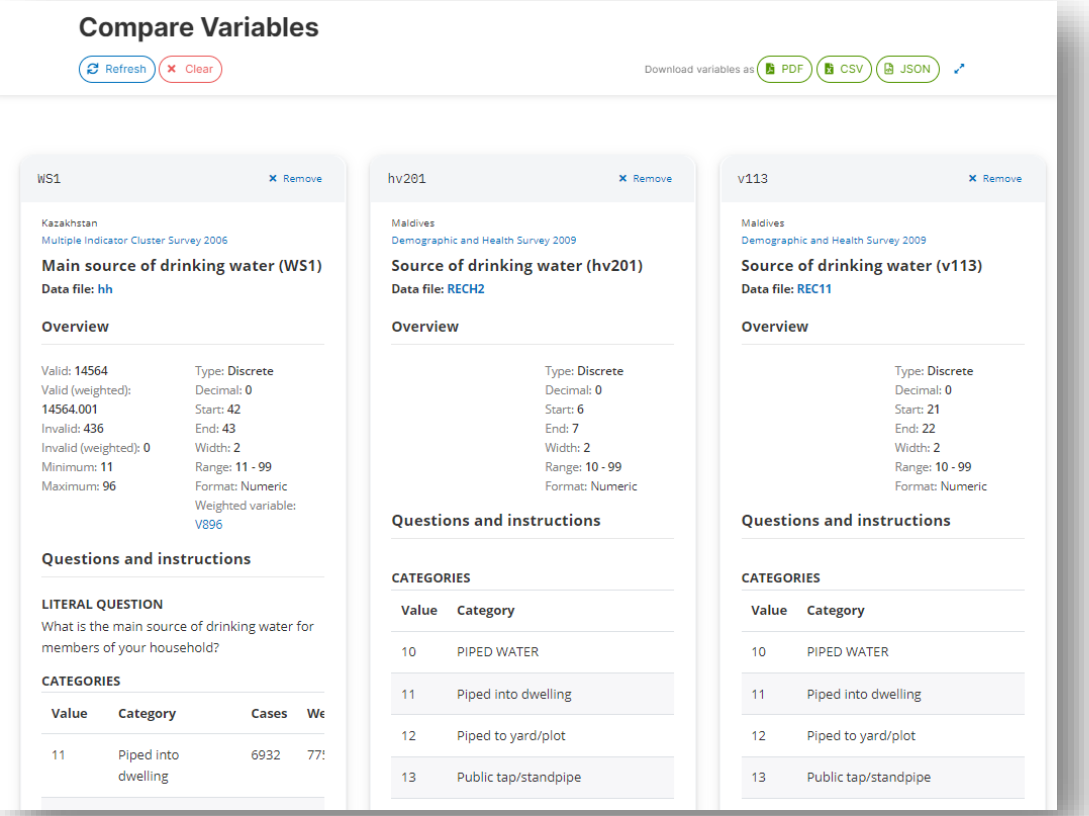


2

NADA: variable-level search and comparison

For microdata: variable-level search and comparison tools

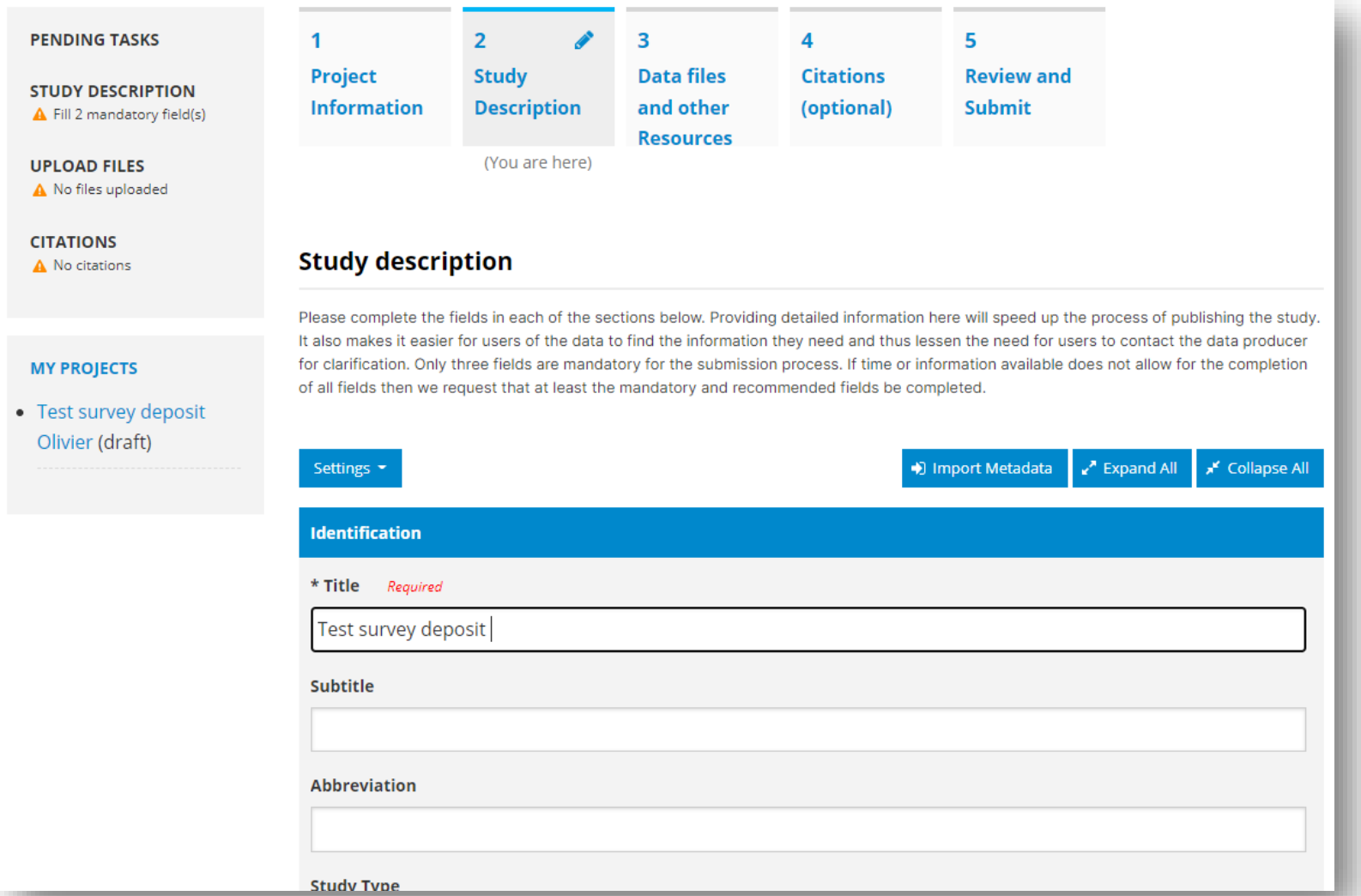




2

NADA: data deposit system

Data deposit system for controlled acquisition of data/metadata



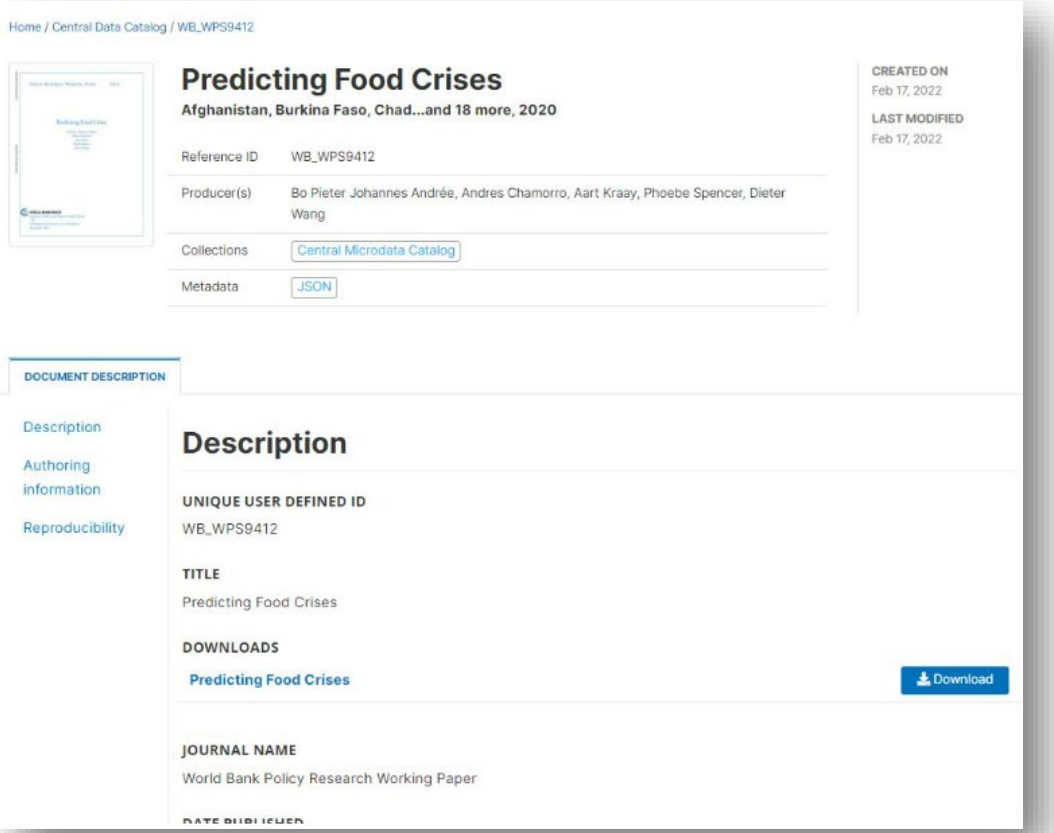
2

NADA: Management using R or Python (APIs)

```r
library(nadar)

# ------------------------------------------------------------------------------
my_keys <- read.csv("C:/confidential/my_API_keys.csv", header=F, stringsAsFactors=F)
set_api_key("my_keys[1,1")
set_api_url("https://.../index.php/api/")
set_api_verbose(FALSE)
# ------------------------------------------------------------------------------

setwd("C:/my_folder")
doc_file <- "WB_PRWP_9412_Food_Crises.pdf"

id <- "WB_WPS9412"

thumb_file <- gsub(".pdf", ".jpg", doc_file)
capture_pdf_cover(doc_file)  # Capture cover page for use as thumbnail

example_1 <- list(

  document_description = list(

    title_statement = list(idno = id, title = "Predicting Food Crises"),

    date_published = "2020-09",

    authors = list(
      list(last_name = "Andrée", first_name = "Bo Pieter Johannes",
           affiliation = "World Bank",
           author_id = list(list(type = "ORCID", id = "0000-0002-8007-5007"))),
      list(last_name = "Chamorro", first_name = "Andres",
           affiliation = "World Bank"),
      list(last_name = "Kraay", first_name = "Aart",
           affiliation = "World Bank"),
      list(last_name = "Spencer", first_name = "Phoebe",
           affiliation = "World Bank"),
      list(last_name = "Wang", first_name = "Dieter",
           affiliation = "World Bank",
           author_id = list(list(type = "ORCID", id = "0000-0003-1287-332X")))
    ),

    journal   = "World Bank Policy Research Working Paper",
    number    = "9412",
    publisher = "World Bank",

    ref_country =  list(
      list(name="Afghanistan",      code="AFG"),
      list(name="Burkina Faso",     code="BFA"),
      list(name="Chad"              code="TCD")
```

Generate metadata using R or Python and publish it in NADA.

Use of NADA API allows automation of many tasks (scraping or harvesting, transforming, checking, documenting, publishing)



Full examples available in https://mah0001.github.io/schema-guide/chapter04.html

3

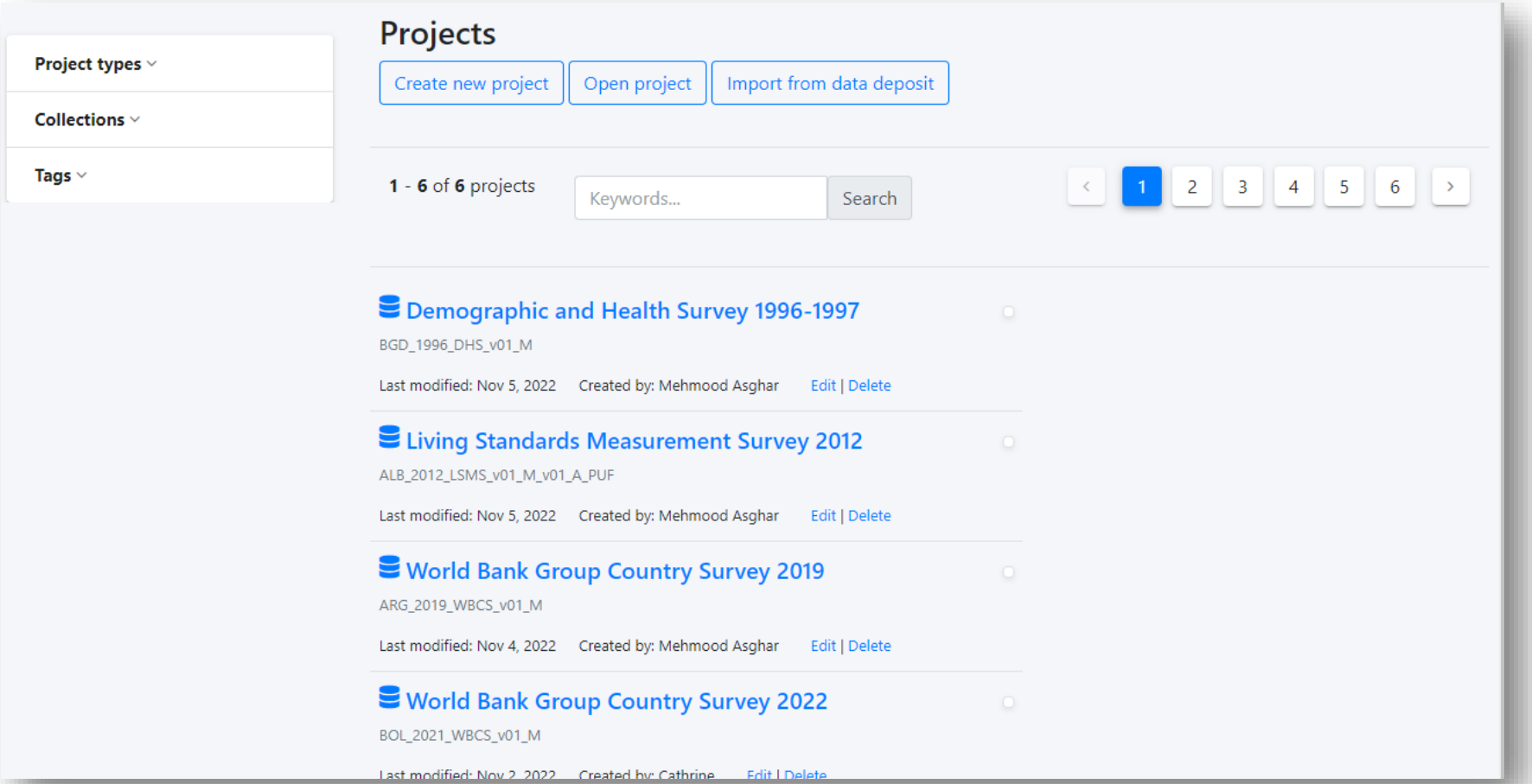
Metadata Editor (DDI Codebook for microdata)

- To replace the Nesstar Publisher (and more)
- Multi-standards; can accommodate user-defined standards/schemas
  - Easy to upgrade when new versions of standards like DDI Codebook are released
- Uses R (haven package) to import microdata files/generate summary statistics
  - Easy to upgrade when new data file formats are released
- Multi-platform; stand-alone or server application
- In final testing phase (as of Dec. 2022); public release around March 2023
- In future version: add metadata augmentation utilities using plugins
  - Connect to machine learning APIs for keyword/topic extraction, classification and tagging, word embeddings; image labeling; speech-to-text for videos, translation, …)

3

Metadata Editor – Home page

- “Project type” can be microdata, geospatial, image, document, table, indicator, etc.
- Access to projects controlled by user login
- Projects can be:
  - Private
  - Shared with a selected group
  - Shared with all logged-in curators

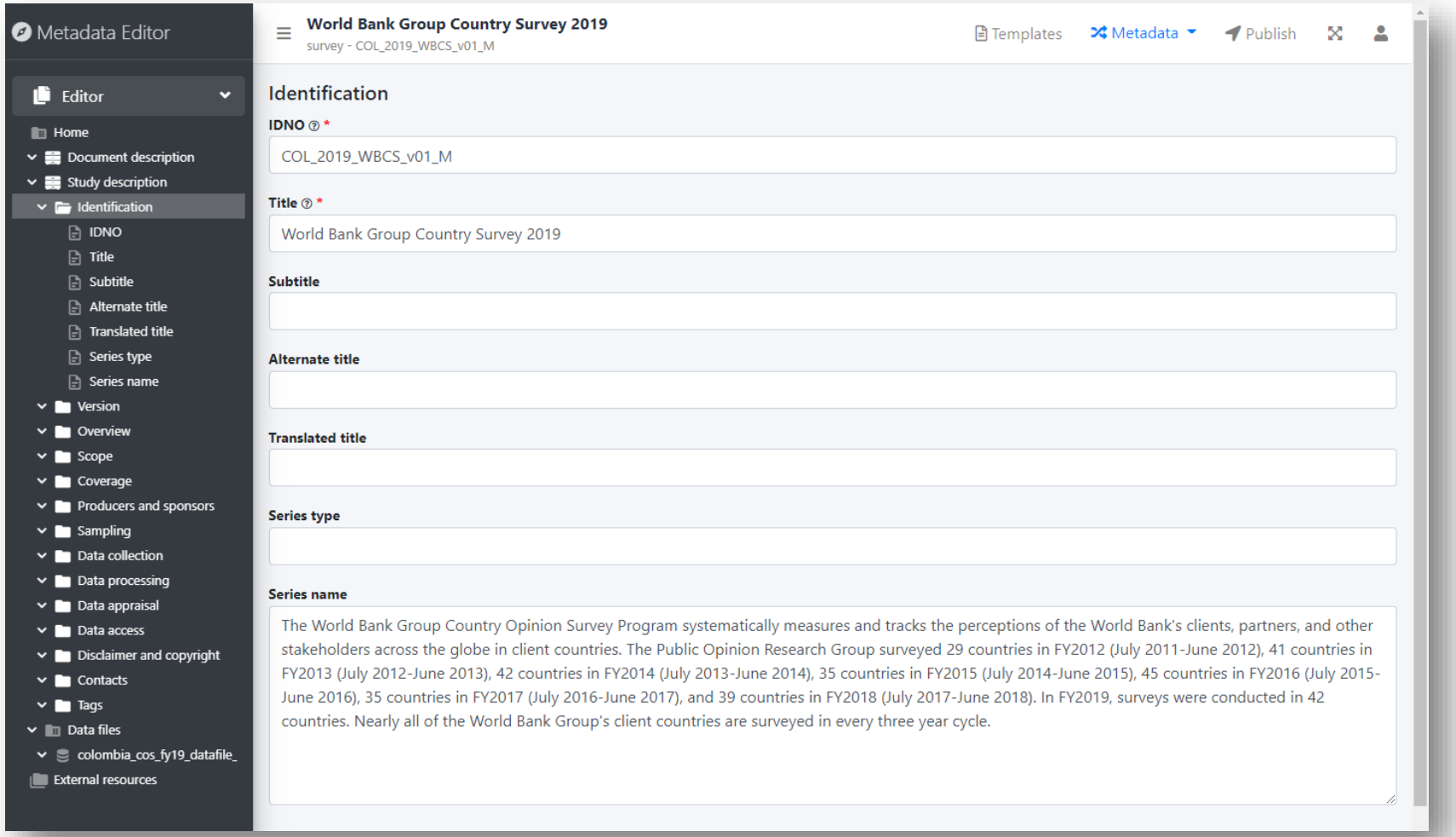


3

Metadata Editor – Study-level metadata page

Process: like in Nesstar Publisher

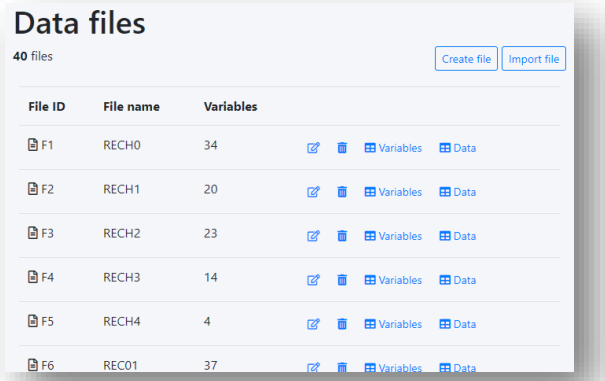
1. Select a template
2. Import data (for microdata)
3. Add metadata and external resources
4. Augment metadata (in future version)
5. Save/export/publish
   - Project saved as a ZIP file that includes the data, metadata, external resources, study thumbnail, and the template.

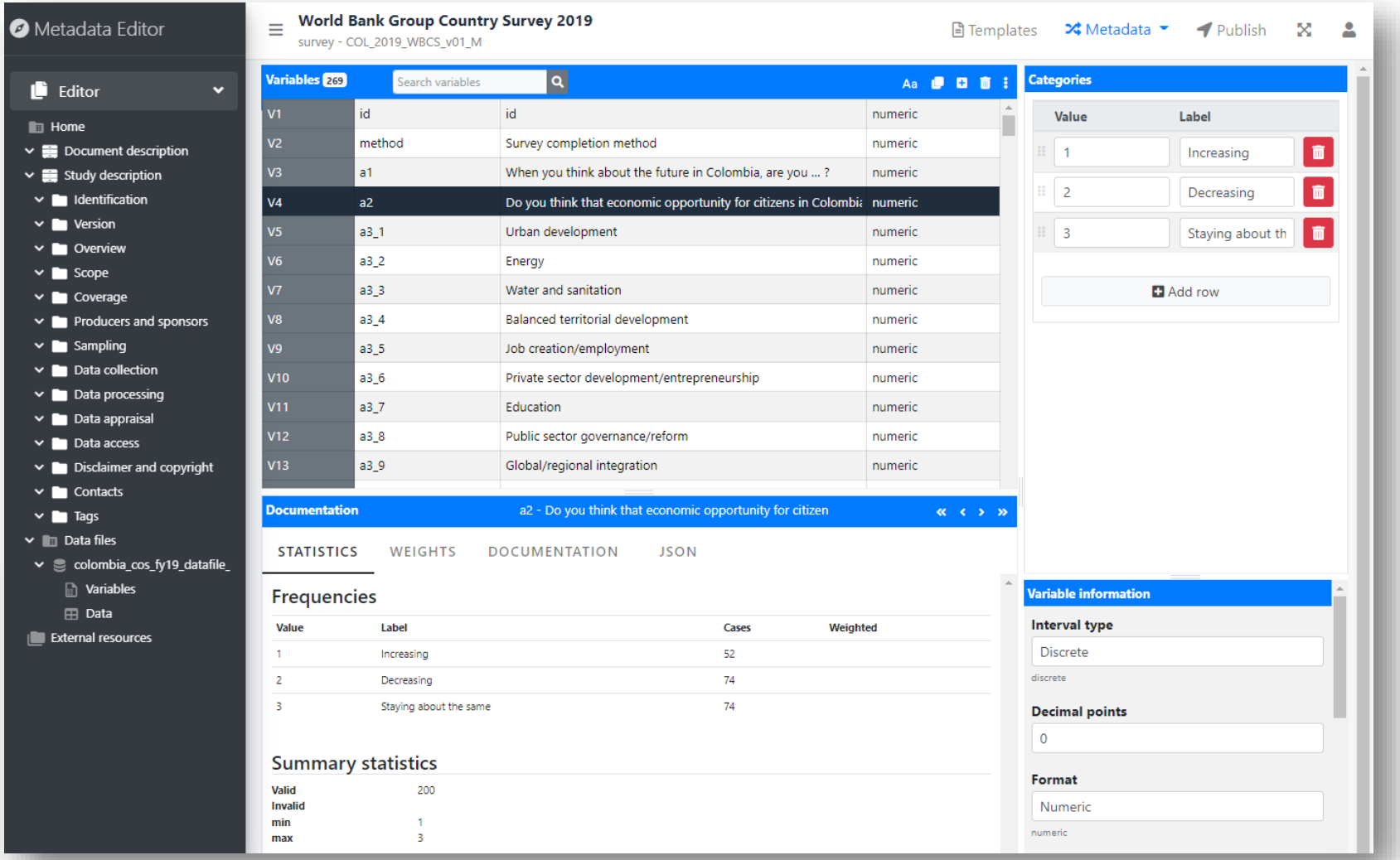


3

Metadata Editor – Variable-level metadata page

Imports/ exports microdata files and generates summary statistics using R (haven package)

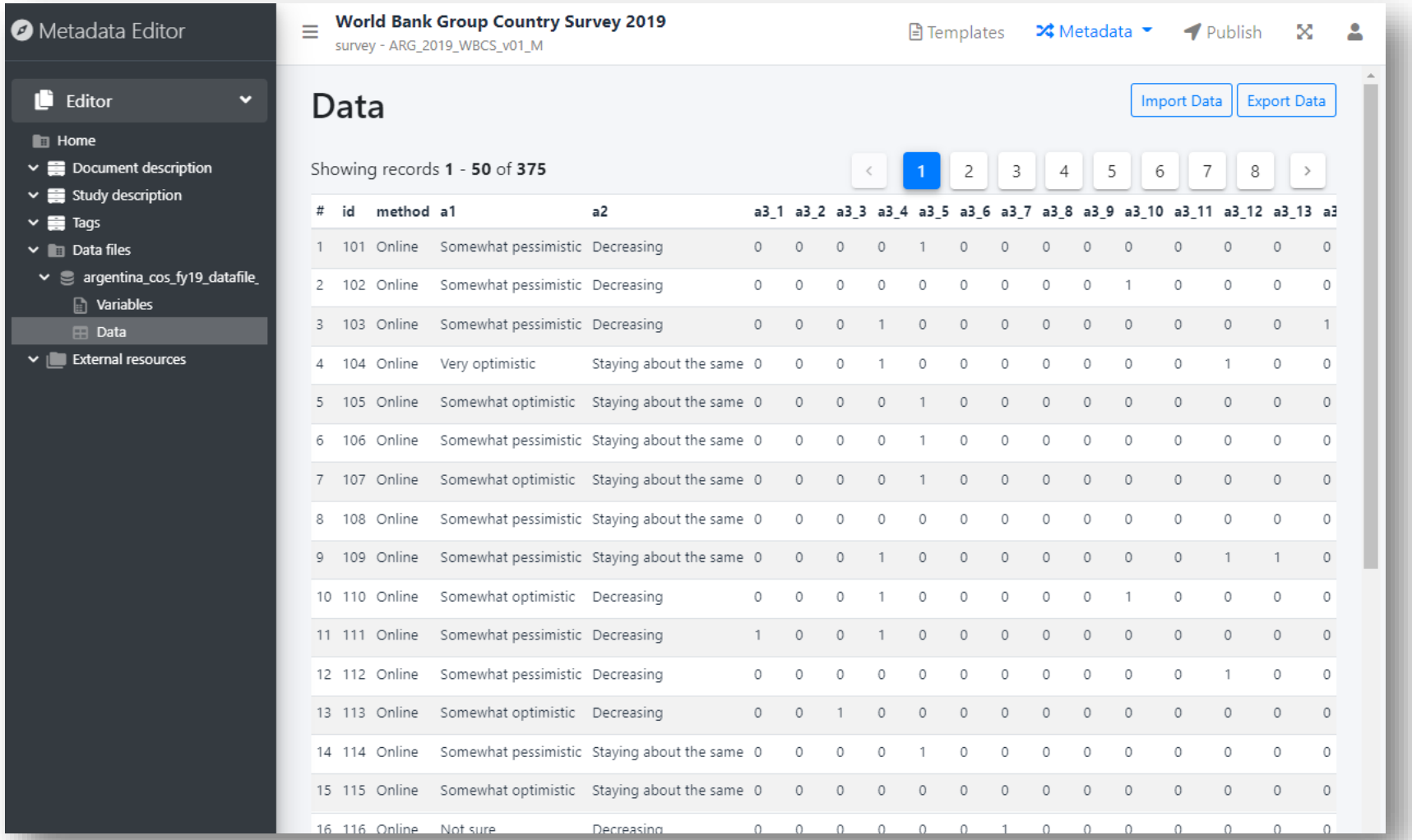




3

Metadata Editor – Data preview page

Imports/ exports data files and generates summary statistics using R (haven package)

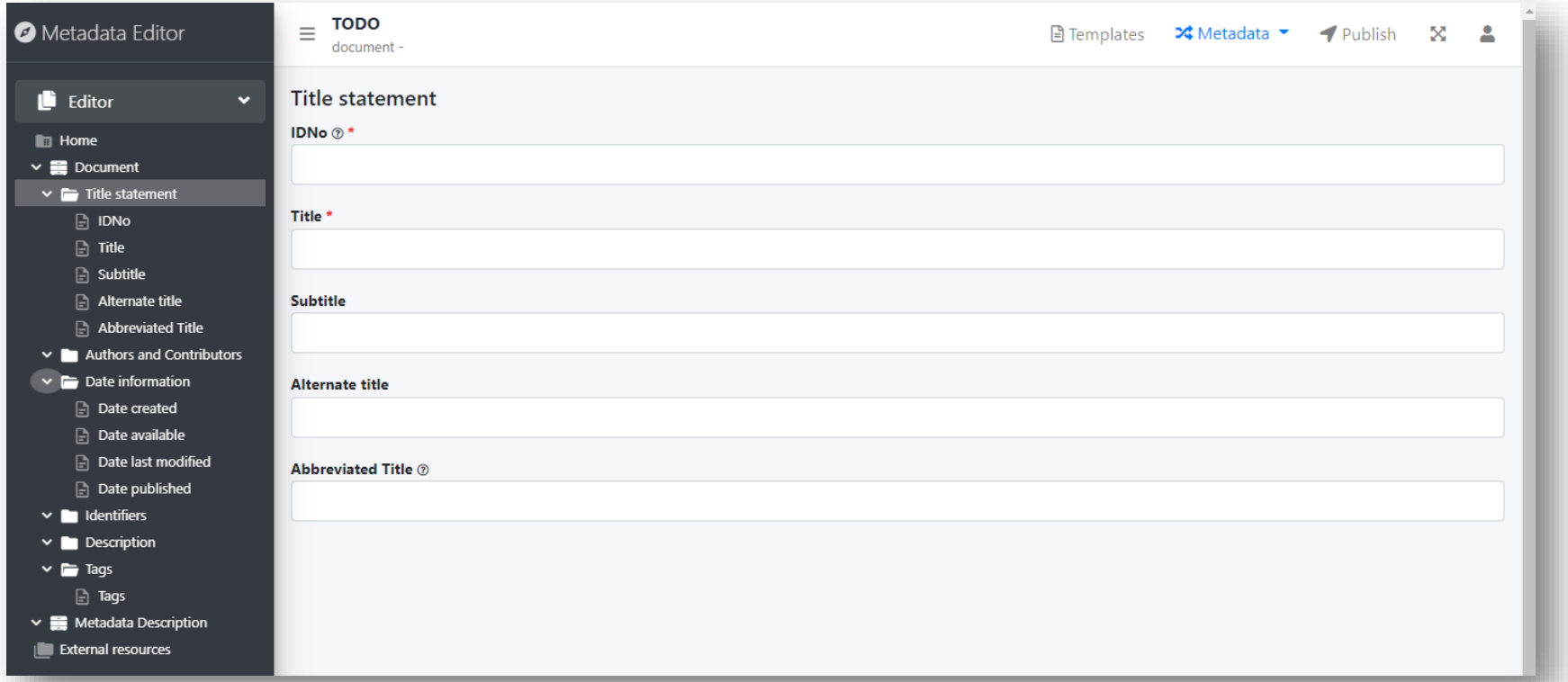


#	id	method	a1	a2	a3_1	a3_2	a3_3	a3_4	a3_5	a3_6	a3_7	a3_8	a3_9	a3_10	a3_11	a3_12	a3_13	a3
1	101	Online	Somewhat pessimistic	Decreasing	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	102	Online	Somewhat pessimistic	Decreasing	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3	103	Online	Somewhat pessimistic	Decreasing	0	0	0	1	0	0	0	0	0	0	0	0	0	1
4	104	Online	Very optimistic	Staying about the same	0	0	0	1	0	0	0	0	0	0	0	1	0	0
5	105	Online	Somewhat optimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
6	106	Online	Somewhat pessimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	107	Online	Somewhat optimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
8	108	Online	Somewhat pessimistic	Staying about the same	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	109	Online	Somewhat pessimistic	Staying about the same	0	0	0	1	0	0	0	0	0	0	0	1	1	0
10	110	Online	Somewhat optimistic	Decreasing	0	0	0	1	0	0	0	0	0	1	0	0	0	0
11	111	Online	Somewhat pessimistic	Decreasing	1	0	0	1	0	0	0	0	0	0	0	0	0	0
12	112	Online	Somewhat pessimistic	Decreasing	0	0	0	0	0	0	0	0	0	0	0	1	0	0
13	113	Online	Somewhat optimistic	Decreasing	0	0	1	0	0	0	0	0	0	0	0	0	0	0
14	114	Online	Somewhat pessimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
15	115	Online	Somewhat optimistic	Staying about the same	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	116	Online	Not sure	Decreasing	0	0	0	0	0	0	1	0	0	0	0	0	0	0

3

Metadata Editor – Document other data types

Works similarly for other data types, with their own standard or schema, e.g, using elements from DCMI and MARC21 to document a publication.



Research/exploratory work on NLP

- Objective: improved discoverability; exploit natural language processing (NLP) models to build a semantic search and recommender system for data cataloguing applications (NADA and others)
- Also: build tools for automated metadata enhancement to be embedded in metadata editors
- Models/tools already tested with satisfactory results, but still needs tuning and tools development (open APIs)
- Exploratory work available in nlp4dev platform (pwd = “nlpexplorer”) https://www.nlp4dev.org/

4

Research/exploratory work on NLP

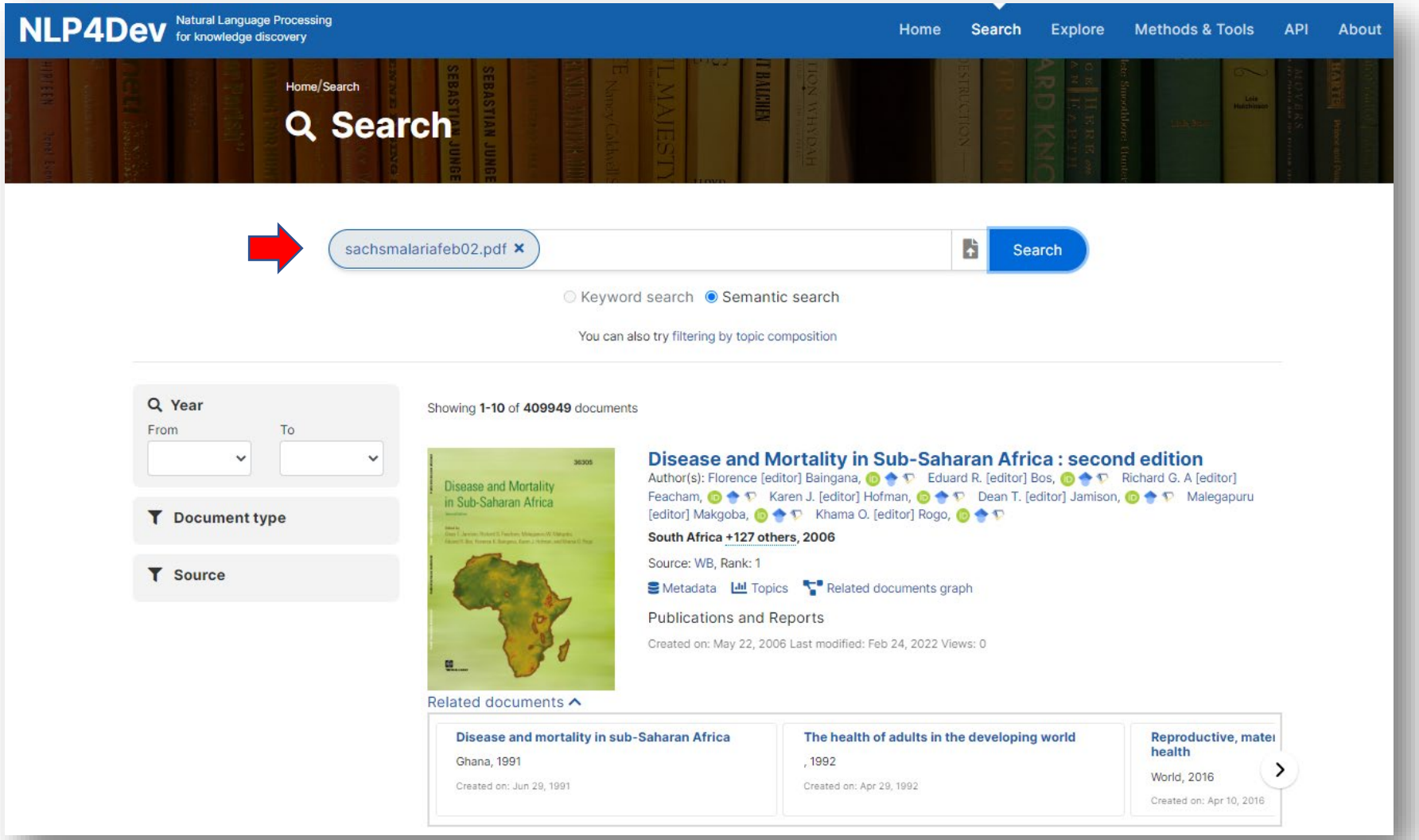
Testing a semantic search and recommender system

Example:

User can submit a PDF document as a query

→ Document is processed and analyzed

→ Closest matches are returned (ranked by semantic closeness)



4

Research/exploratory work on NLP

- → The closest datasets available in the data catalog are also returned (based on semantic closeness between the query/PDF document and the metadata available for each dataset
- → Could also use geographic coverage, year, and other criteria for ranking results)
- → Requires rich and augmented metadata to return relevant results

1 2 3 4

Ongoing/planned activities, and collaboration

- Addition of multiple new features in NADA and Metadata Editor
- Exploratory work on use of machine learning for better data discoverability
  - Semantic search; D3 indexing for geographic data; query parsers; improved search results ranking; and more.
  - Objective: develop smarter search algorithms and a recommender system for data catalogs, and metadata enhancement tools (with open-source software / open APIs)
- Collaboration welcome in multiple areas, e.g.:
  - Review/improvement of our custom metadata schemas
  - Development and/or testing of software (Metadata Editor, NADA, R package, PyNADA)
  - Production of technical documentation and training materials for Metadata Editor/NADA
  - Translation of Metadata Editor and related documentation
  - Design new features for Metadata Editor and NADA
  - Support to resource-constrained organizations (in low-income countries)