Exploring upgrade options for the CESSDA Data Catalogue

Authors:
Maja Dolinar, Slovenian Social Science Data Archives (ADP)
Katja Moilanen, Finnish Social Science Data Archive (FSD)
Markus Tuominen, Finnish Social Science Data Archive (FSD)
Benjamin Beuster, Norwegian Centre for Research Data (NSD)

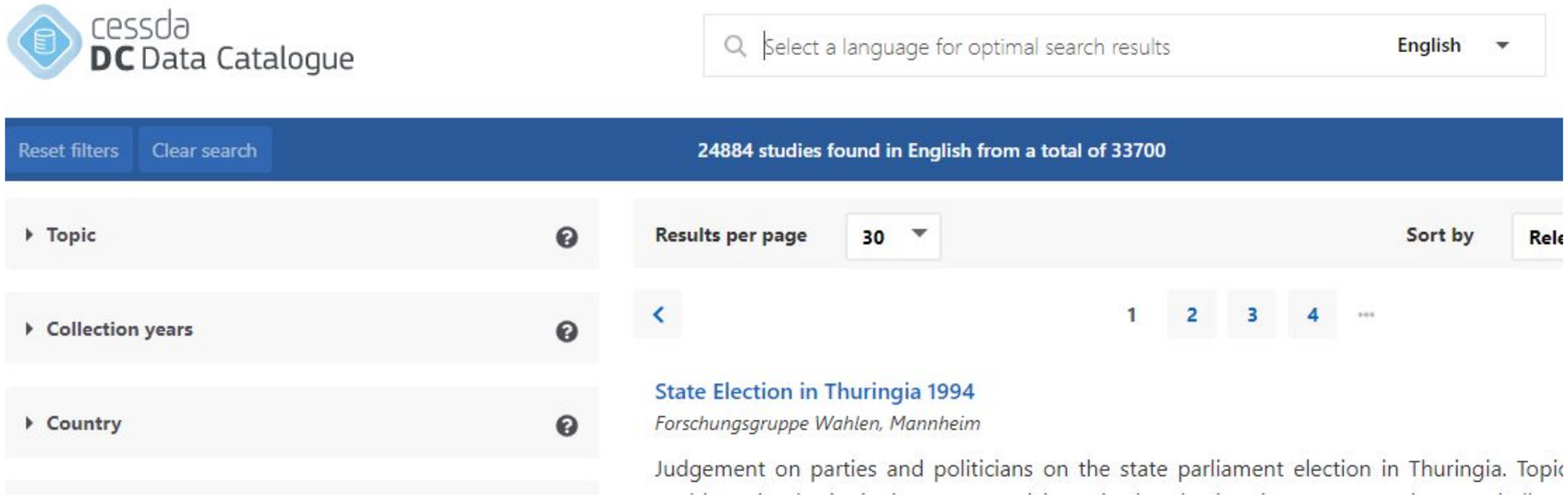
EDDI21, the 13th Annual European DDI User Conference, November 30 – December 1, 2021

cessda.eu @CESSDA_Data

CESSDA DC Data Catalogue

https://datacatalogue.cessda.eu/

- contains metadata of studies from CESSDA service providers (SPs) and serves as an entry point for search and discovery of European social science data



Aim of the presentation

- Present the findings of the CDC Upgrade Task group under the CESSDA Agenda 21-24
- Aim of the task: to develop a list of recommended features and updates for the next version of CDC
  - including implementing search and discovery of variable-level information in CDC and selecting other prioritised items from the CDC wish list
- Methodologies implemented:
  - Desk research of available resources (SPs existing catalogues, other similar catalogues)
  - Survey of SPs about needed improvements of the CDC
  - Analysis of SPs variable-level metadata examples

Keeping track with current developments

- **Established contact** with CDC User Group, Dataverse Basecamp, Metadata Office, CDC User Experience Project

- **Following news and developments** within CESSDA:
  - CESSDA Data Access Policy
  - CDC upgrades and additional features:
    - validation of harvested metadata,
    - PID validation,
    - mapping to OpenAIRE, B2find, schema.org and Dublin Core
  - Tools Open Hour on Dataverse developments

Insight into SPs' views on CDC development

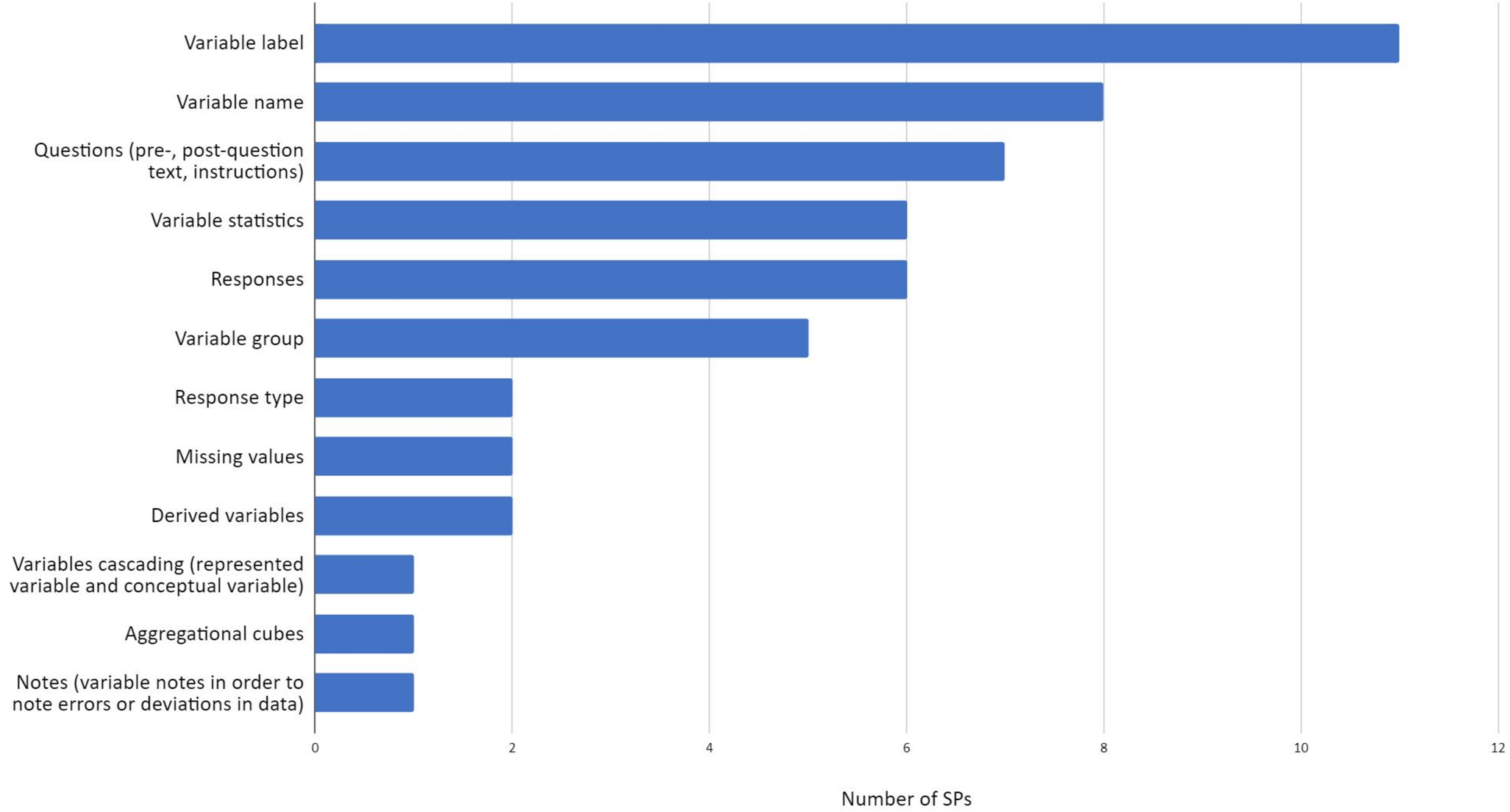
- 15 SPs responded to the survey and 9 provided variable metadata examples.
- **Additional filters that would be helpful:** data access, time method, dataset type
- **Richness of metadata in the current version:** enough metadata (6x), additional suggestions: universe, type of dataset, date when metadata published in CDC
- **Improvements of functionalities:** no improvements (9x), some ideas for improvements in terms of clarity and findability

Insight into SPs' views on CDC development

- **Improvements of user experience:**
  - language selection issues reported,
  - User Guide should be made more prominent,
  - adding a demonstration video on how to use the Catalogue,
  - similar studies list should include all studies in the catalogue and not only studies by the individual SP,
  - strong multilingual support is important

- **Additional information to be included in the CDC:** variable-level information, dataset type, citation, information about related publications etc.

Willingness to include variable-level information in CDC



Summary of the variable metadata examples

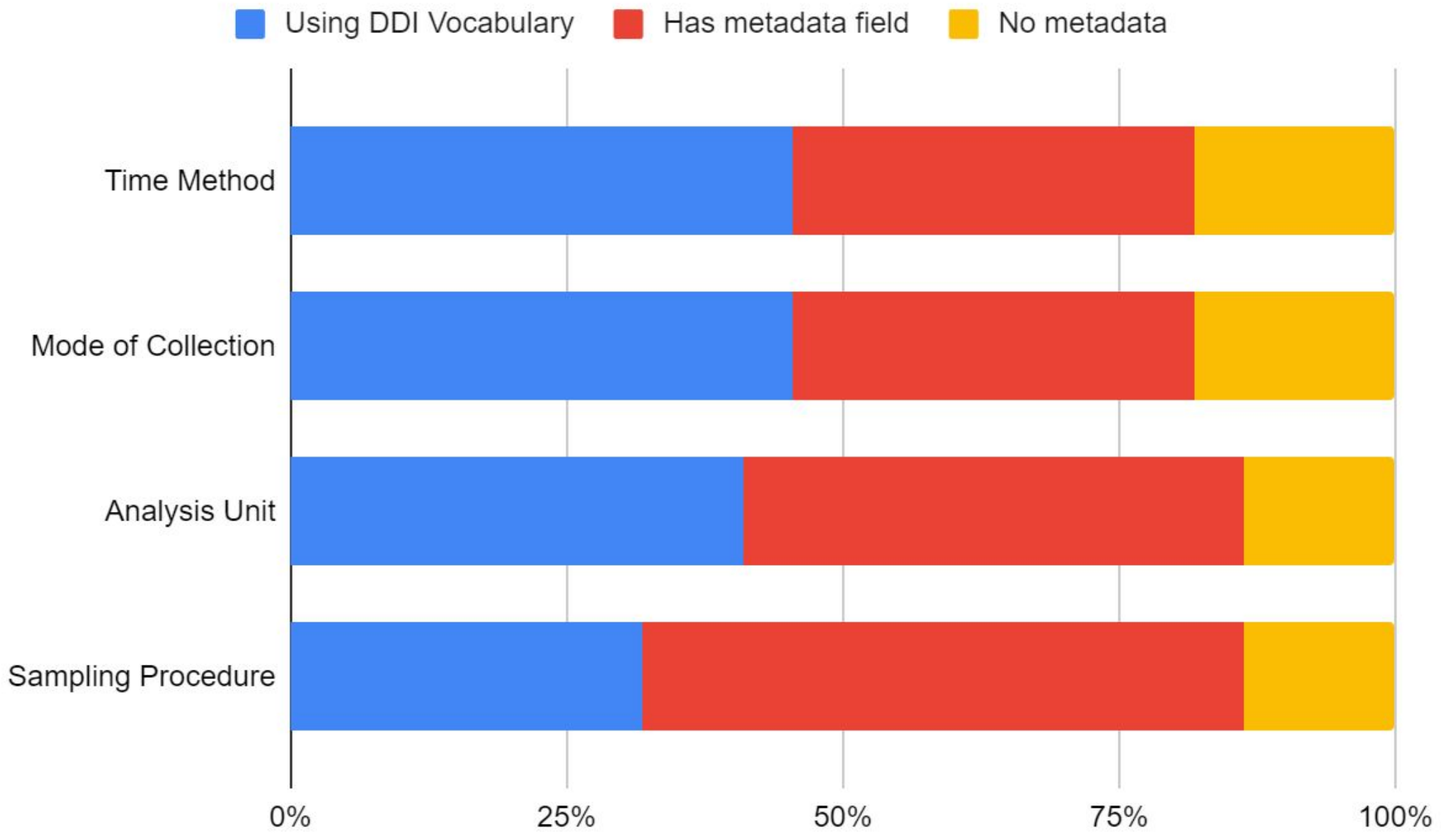
- 9 SPs sent us examples of variable information
- 8 DDI-C, 1 DDI-L

- **Variable information (9/9)**
  - All had ID, name and label
  - 8/9 if variable is continuous or discrete; 7/9 literal question text and 6/9 interviewer instructions
- **Summary Statistics 8/9**
- **Information on Categories and values**
  - All had label and value
  - 7/9 if category represents missing value and frequency of the category
- **Technical format (character or numeric) 7/9**
- **Variable groups (7/9)**
  - All 7 had reference to variables and label/text explaining variable group

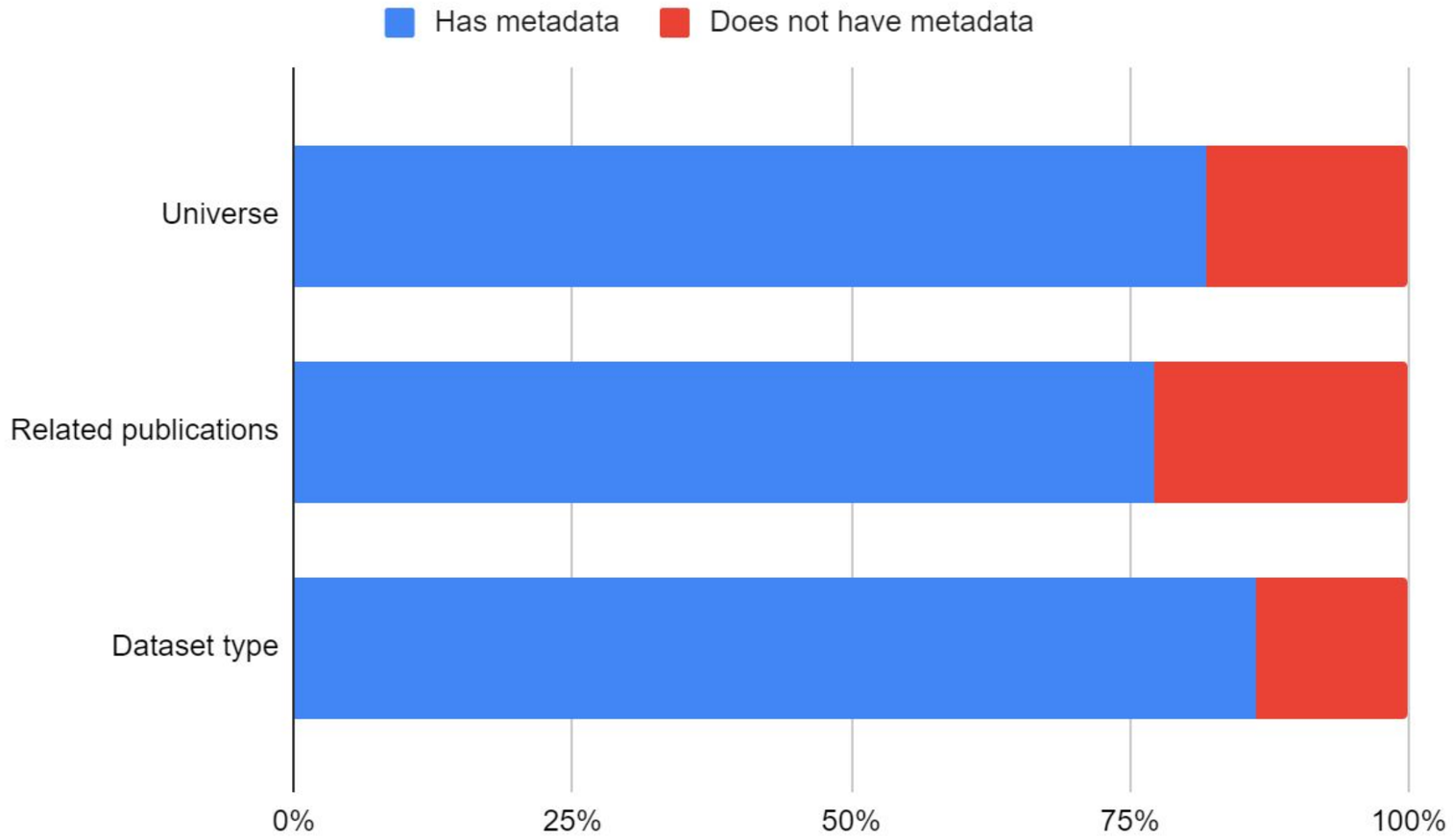
Desk research of SPs' metadata

- Data collected mostly in August 2021
- Source of the data: SPs' own catalogues
- Data include both SPs that already have their metadata in CESSDA catalogue and SPs not existing in catalogue yet
- "Research" questions included:
  - How many of the SPs are using DDI vocabularies or have at least some information about:
    - Analysis Unit
    - Mode of Collection
    - Time Method
    - Sampling Procedure
  - Is there information about related publications, universe and data type (quali/quanti or other)?
  - Which information SPs have about variables?

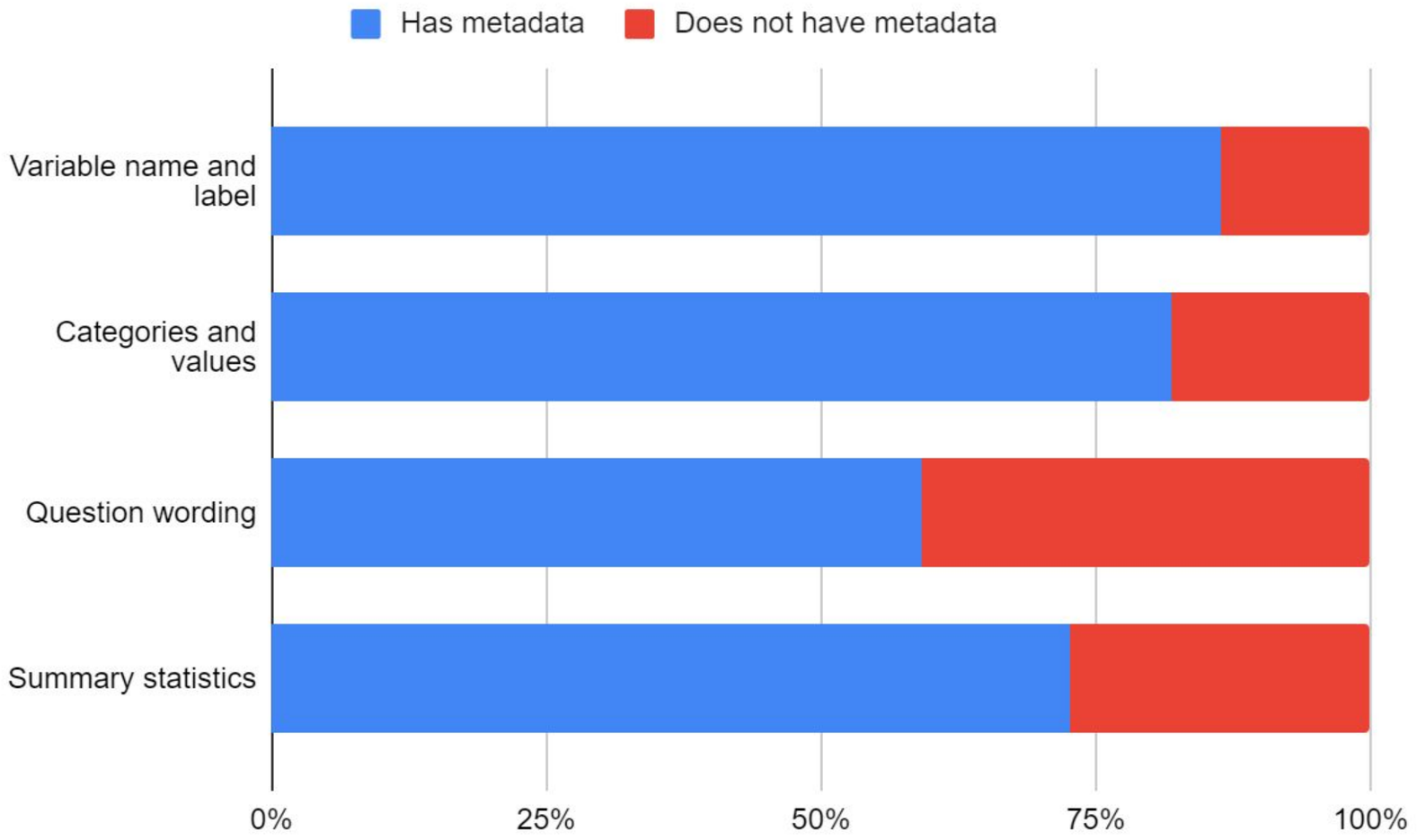
Usage of DDI vocabularies in specific fields



Presence of selected metadata fields



Variable-level metadata



Building the CDC on “ready made” applications

- Explored applications: Dataverse, Colectica

Use of Dataverse for CDC: good for presenting study metadata, however currently does not allow for variable-level discovery and search (some current developments from CESSDA SPs show possible solutions that CDC could use to display variable-level information).

- result: great for searching and filtering study metadata (customization available)
- result: potential multilinguality issues for metadata
- result: studies need to have a DOI
- result: could potentially be used in the future for CDC

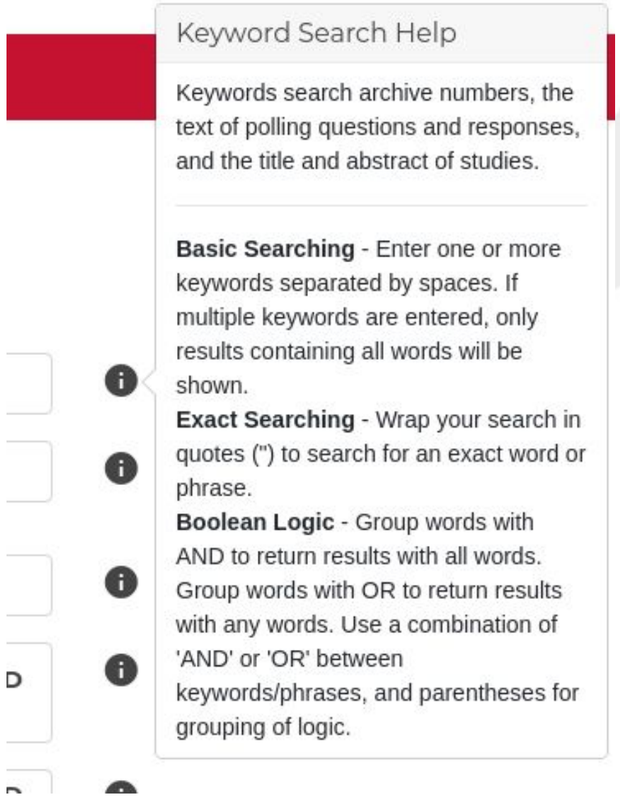
Use of Colectica for CDC: Covers almost the entire research lifecycle. Works well for longitudinal and repeat studies. Supports variable documentation including variable cascade and variable lineage structure. Multilingual metadata

- result: could potentially be used in the future for CDC
- result: web representation (Colectica Portal) is not designed for this purpose. Requires changes in configuration. New interface could be developed on top of the repository

A mix of both?

Comparing other data catalogues to CDC

- Data catalogues in general look similar and almost always have the same basic features
- Some useful features found on other catalogues that could be added to CDC:
  - Tooltips to include help about using the filter and the search in general (Roper iPoll)
  - Possibility to hide/show some information like the abstract in the result list (ICPSR)
  - Expanding some filters by default (Réseau Quetelet)
- Filters have some of the biggest differences
  - The amount of available filters varies a lot
    - CDC currently has 4 filters for the studies while ICPSR has 17
  - Other catalogues usually try to list all the options for filters in some way which ends up more complicated and takes too much space compared to CDC



Conclusions

- Challenge: build the CDC on the SPs' existing metadata (i.e., primarily DDI Codebook) or to advocate for a metadata upgrade (i.e., DDI Lifecycle, currently used by only a handful of SPs)
  - result could be limited functionality for end-users who want to search and find variables documented in different languages

- **Final conclusions of the task are work in progress**
  - Final report "*Requirements specification and plan for CDC upgrade / competitive evaluation*" due in March 2022

cessda DC Data Catalogue

Thank you!

Questions?

https://datacatalogue.cessda.eu/

cessda.eu

@CESSDA_Data