Linked Open Research Data for Social Science

A concept registry for granular data documentation

Pascal Siegers[1], Antonia May[1], Jana Nebelin[3], Dagmar Kern[1], Andreas Daniel[2], Ben Zapilko[1], Claudia Saalbach[3], Fakhri Momeni[1], Knut Wenzig[3], & Jan Goebel[3]

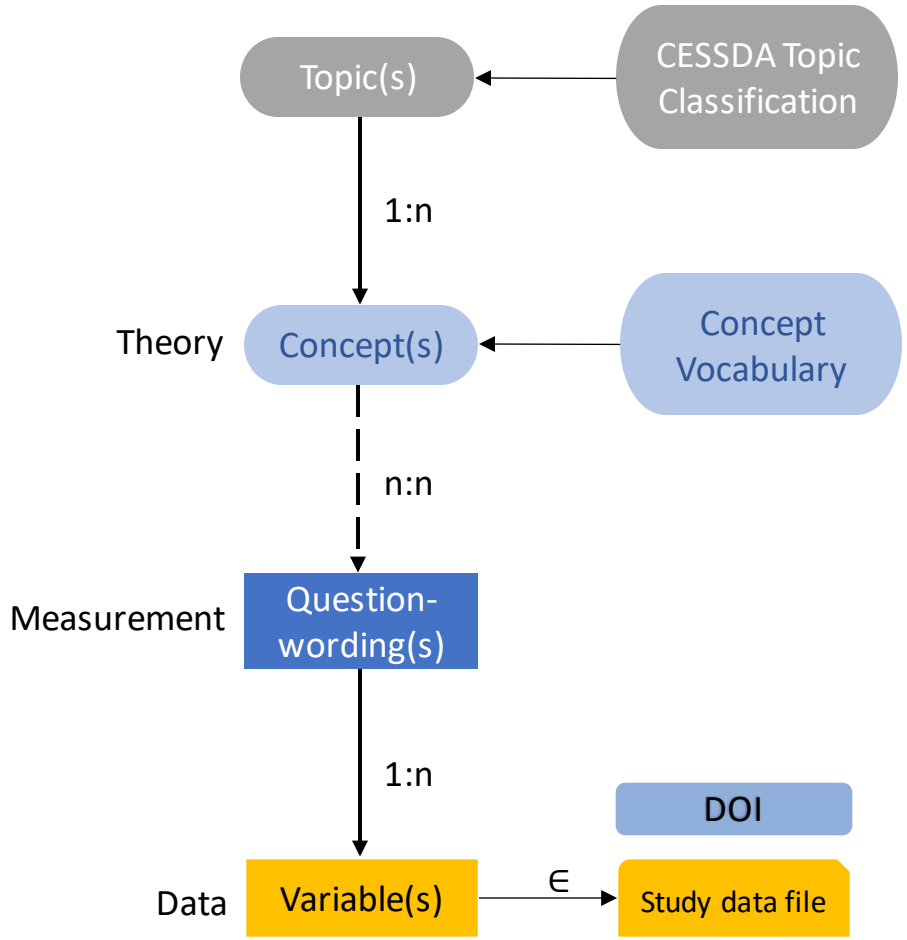
[1]GESIS Leibniz-Institute for the Social Sciences, [2]German Centre for Higher Education Research and Science Studies, [3]German Socio Economic Panel (SOEP) at the German Institute for Economic Research

1. Conference on Research Data Infrastructure [CoRDI]

12. – 14. September 2023 in Karlsruhe – doi:10.5281/zenodo.8420378

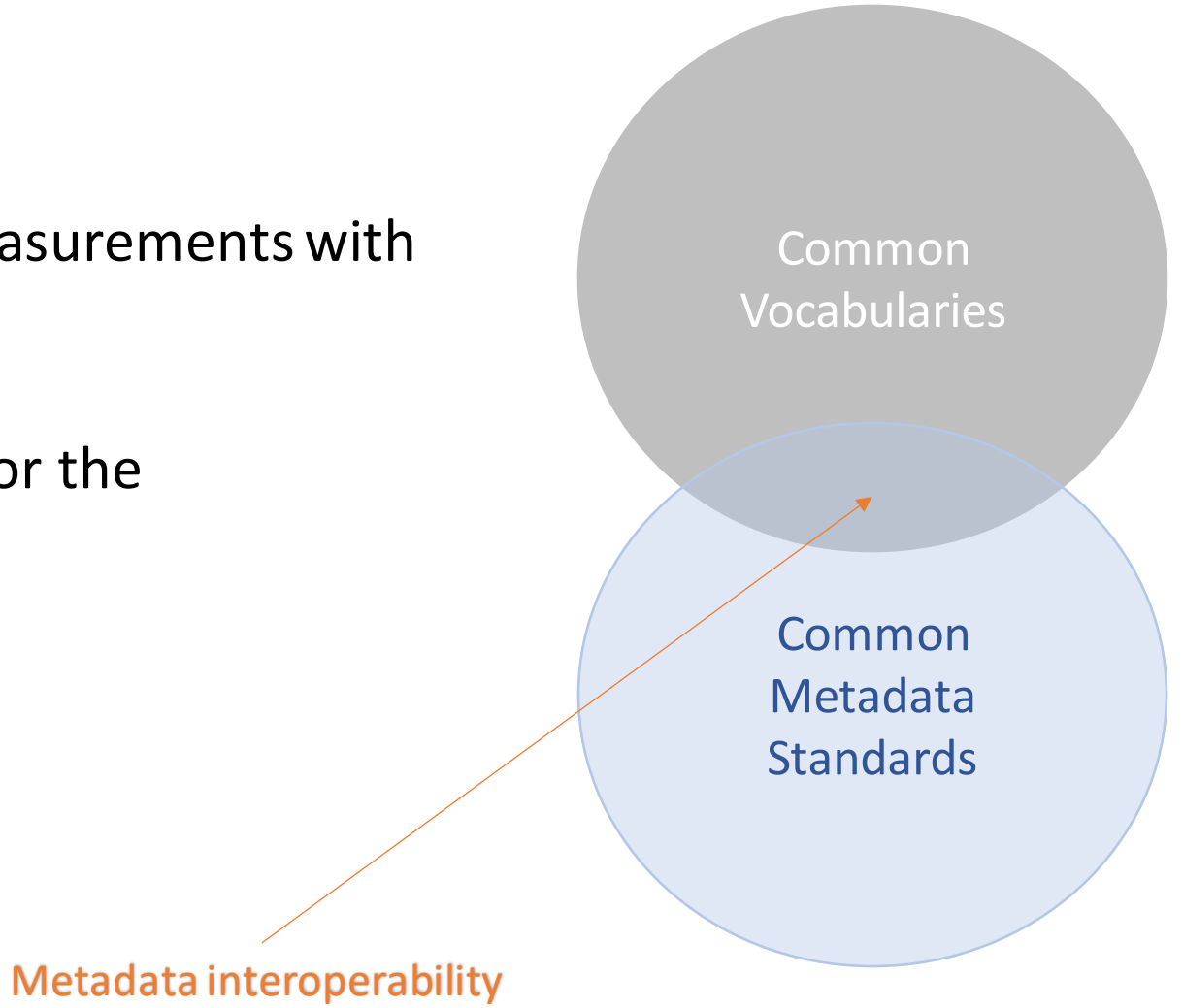
I. The missing „link“ in data documentation I

- General vocabularies for topics (e.g. ELSST, CESSDA Topic Classification)
- Extensive documentation of questions wordings (→ DDI)
- Extensive documentation of variables in data sets (labels, codes, code labels, missing values, etc.)
- Missing: often **no** information on theoretical concepts intended to measure
  → No concept vocabulary available for data documentation



I. The missing „link“ in data documentation II

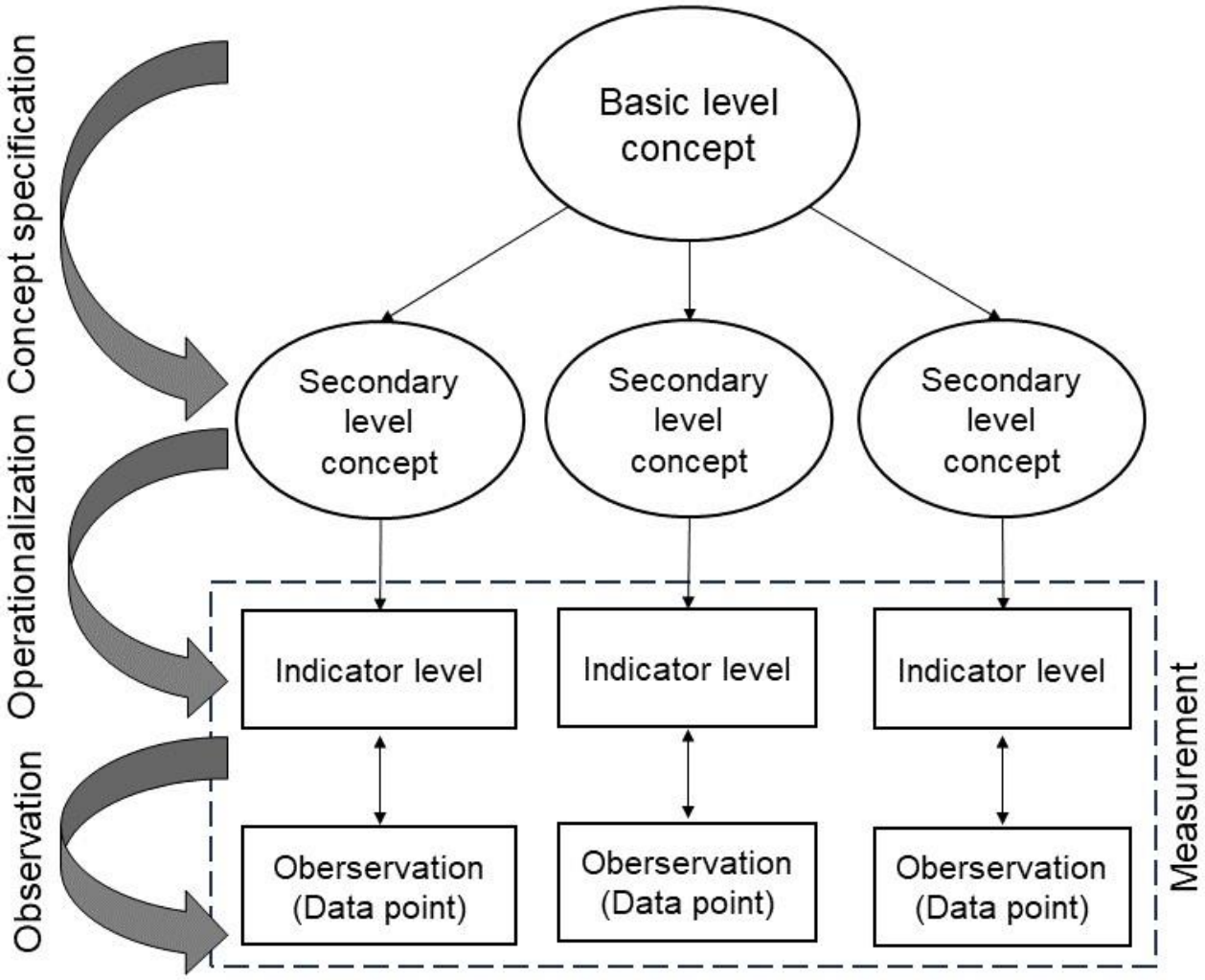
- Why concepts in documentation?
  - Supporting data search by linking measurements with concepts
  - Identifying different measurements for the same/similar concepts
  - FAIRification of research data



II. Concepts in Social Science research I

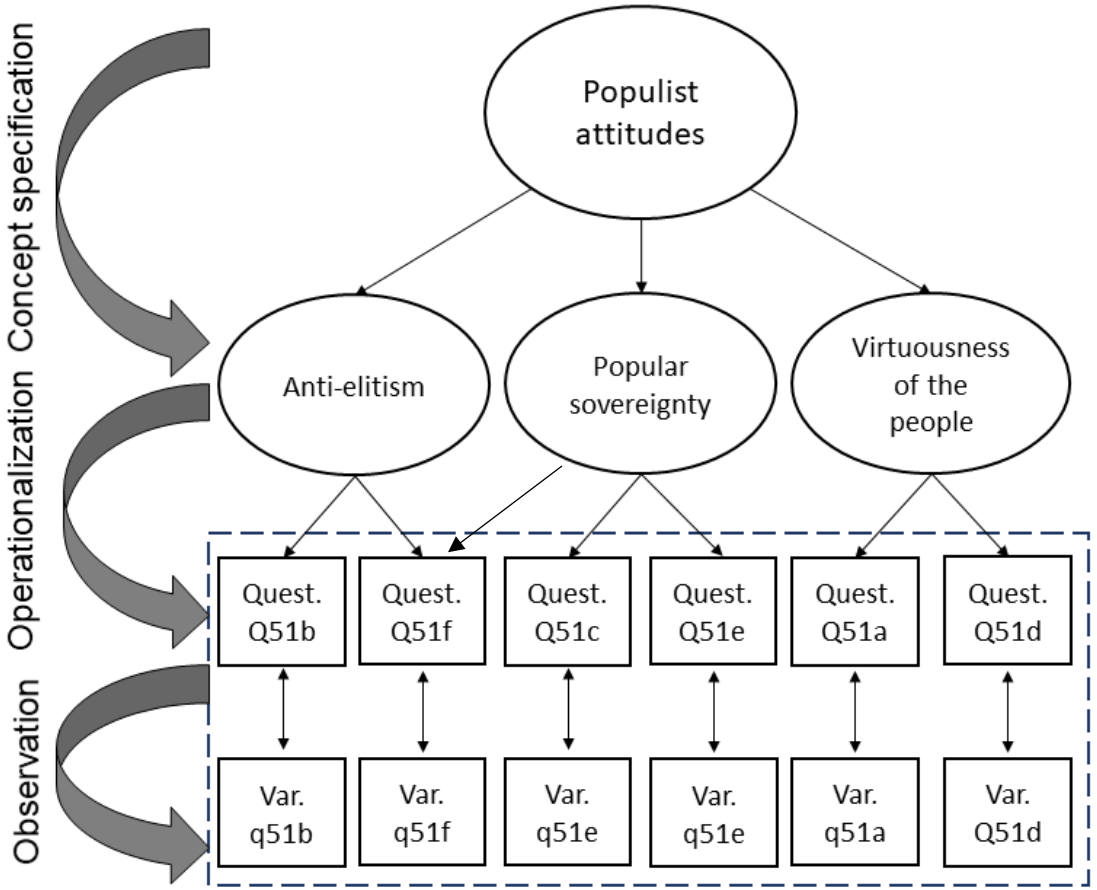
- Concepts are central elements of scientific language and knowledge representation

- Goertz (2006) distuinghishes three levels of social science concepts
  - **I. Basic level:** terminology used in theoretical propositions about reality
  - **II. Secondary level**: Components of basic level concepts (dimensions)
  - **III. Indicator level**: specifications for measurement



II. Concepts in Social Sciece research II

Example from GLES 2021



Example of indicator question (Q51d):
Question text:
Please say how much you agree or disagree with each of these statements.
Question Item
"Differences between the elite and the people are larger than the differences among the people."
Answer scale
(1) Strongly agree; (2) Agree; (3) Neither agree nor disagree; (4) Disagree; (5) Strongly disagree.

Different measures for populism used in research practice

Diversity is the rule rather than the exeption

n:n relationships between concepts and measurements

III. Conceptualizing a Concept Registry

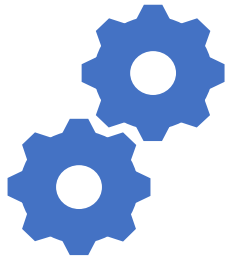
Construction principles:

I. Open and user driven development of the vocabulary

II. Theory language

III. Embedding existing vocabularies

IV. Open interface(s) for re-use



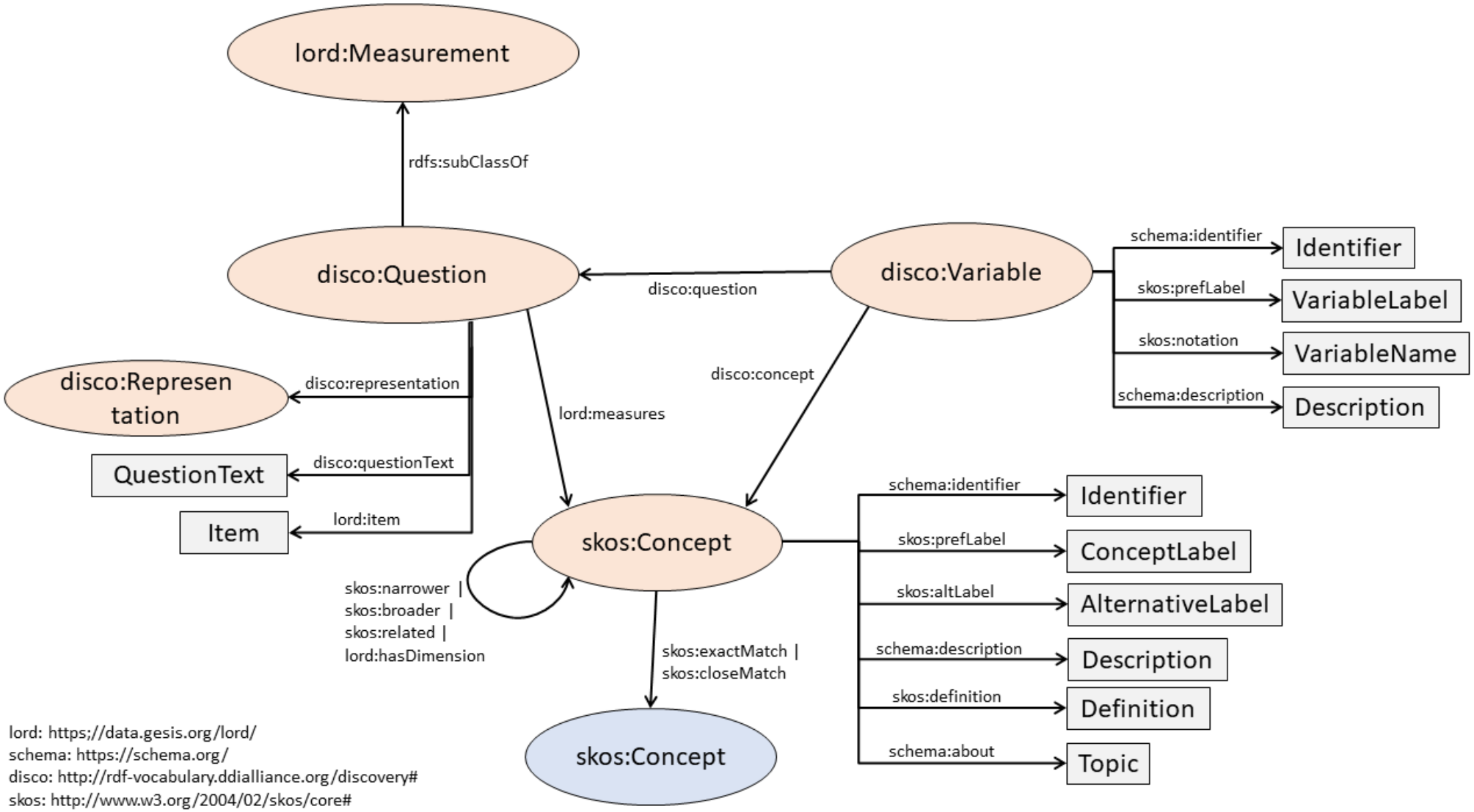
Components:

I. Data model for the concept registry

II. Annotation Tool (linking concepts to the measurement metadata)

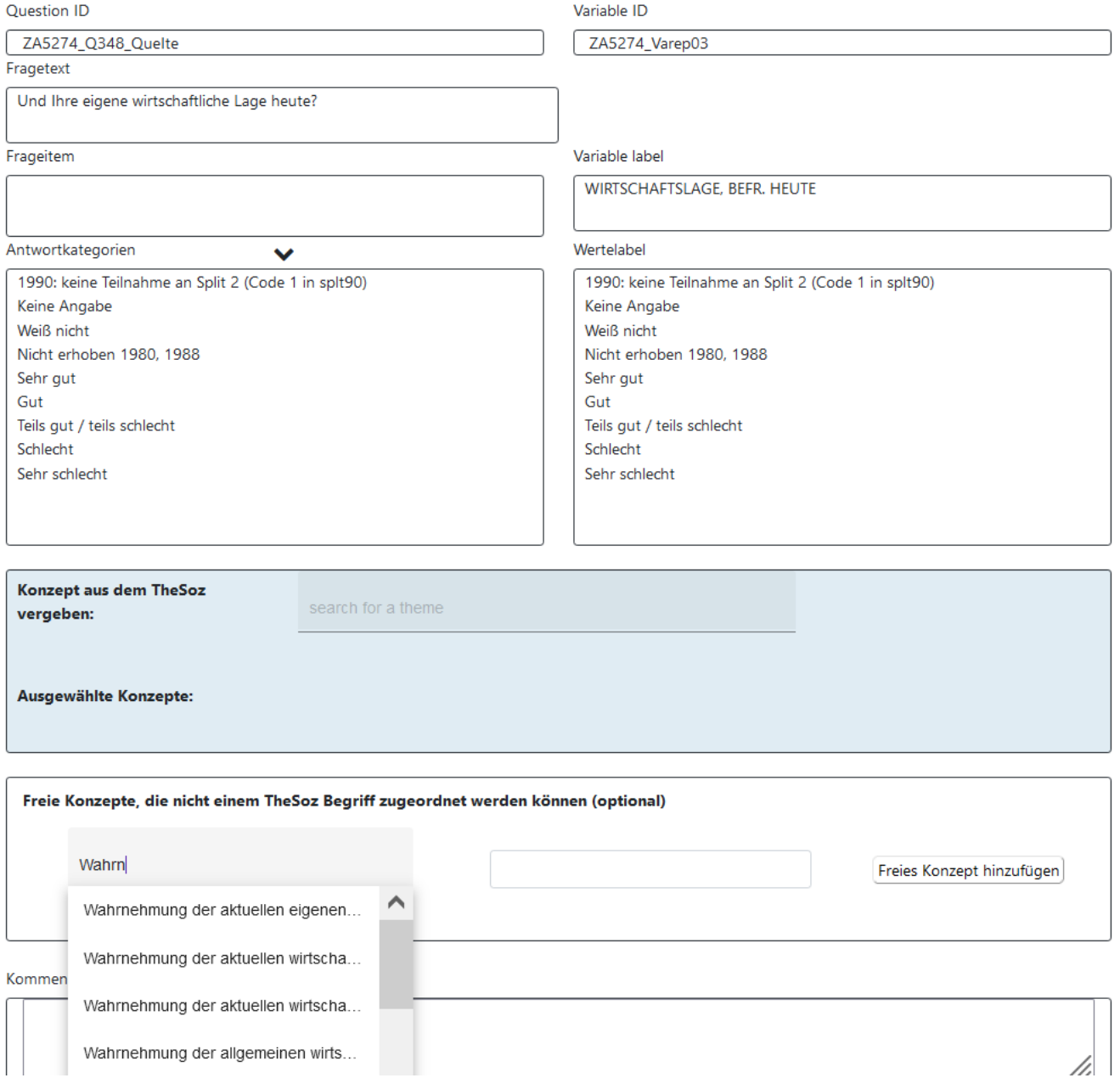
III. Triple Store

III. The LORD Data Model



III. The LORD Annotation Tool

- Displays question and variable metadata
- Allows to select a concept/topic from Thesaurus Social Sciences (TheSoz)
- New concepts are linked to the measurement and added to the concept registry
  - Several concepts can be linked to the metadata



IV. Lessons Learned from the pilot study I

- Test annotation
  - German Socio-economic panel (GSOEP), German National Academics Panel Study (nacaps), and German General Social Survey (GGSS)
  - Each project partner annotated selection of questions from the three surveys (topics: health, income, migration etc.)

- Core questions for test:
  - Do annotations „overlap“?
  - Is there a *between concepts structure* „emerging“ from the annotations?

SOEP

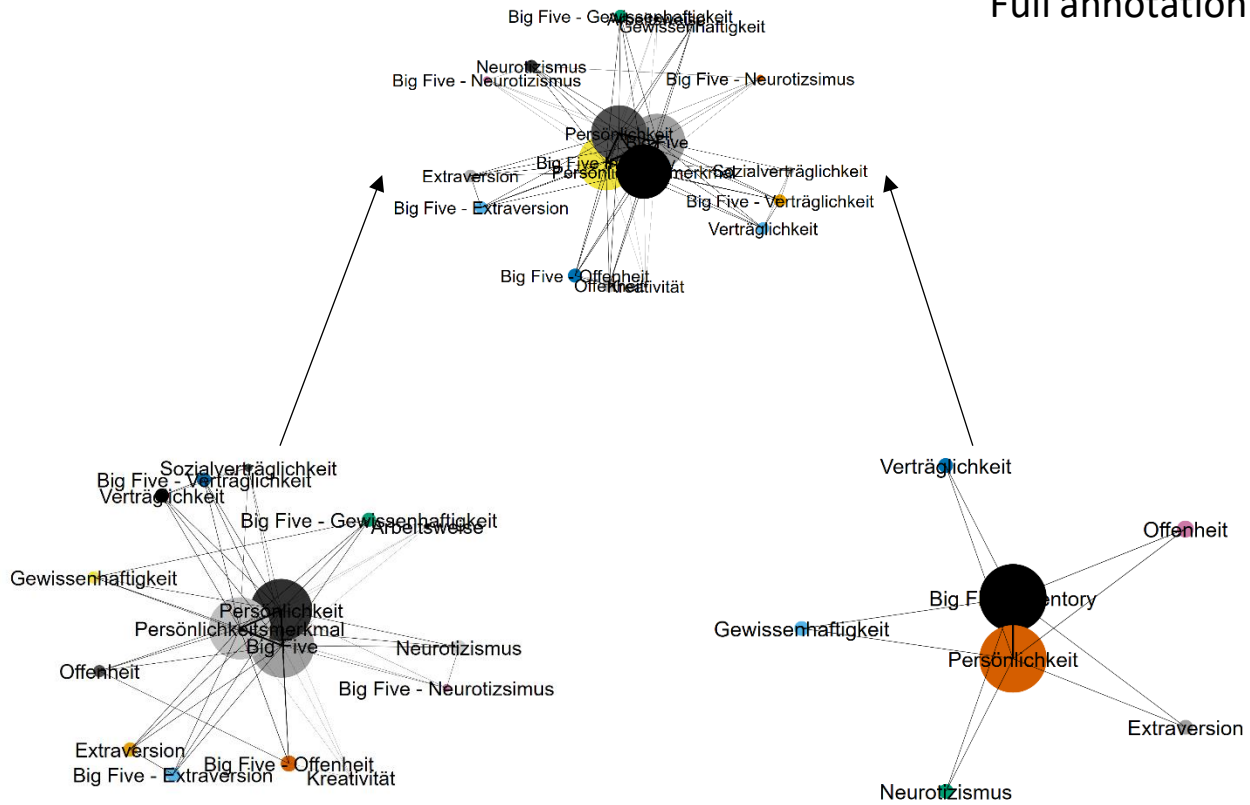
ALLBUS

nacaps
National Academics Panel Study

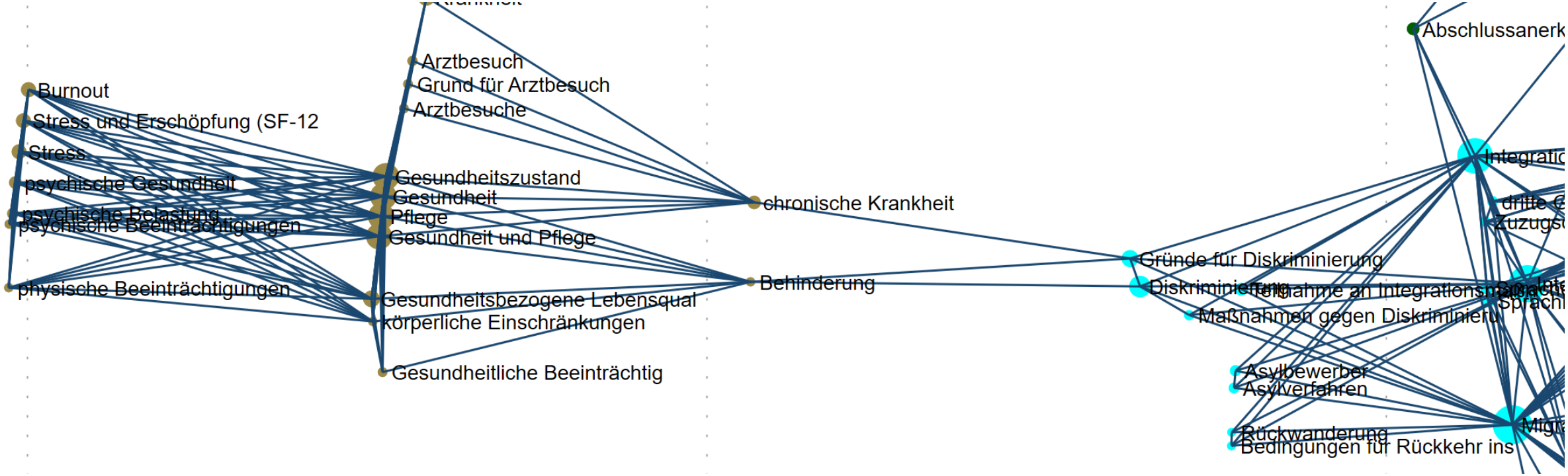
IV. Lessons Learned from the pilot study II

- Great diversity in individual annotation styles

- Results in large amount of different concept terms that cover very similar measurements
  - Non-substantive differences in concepts



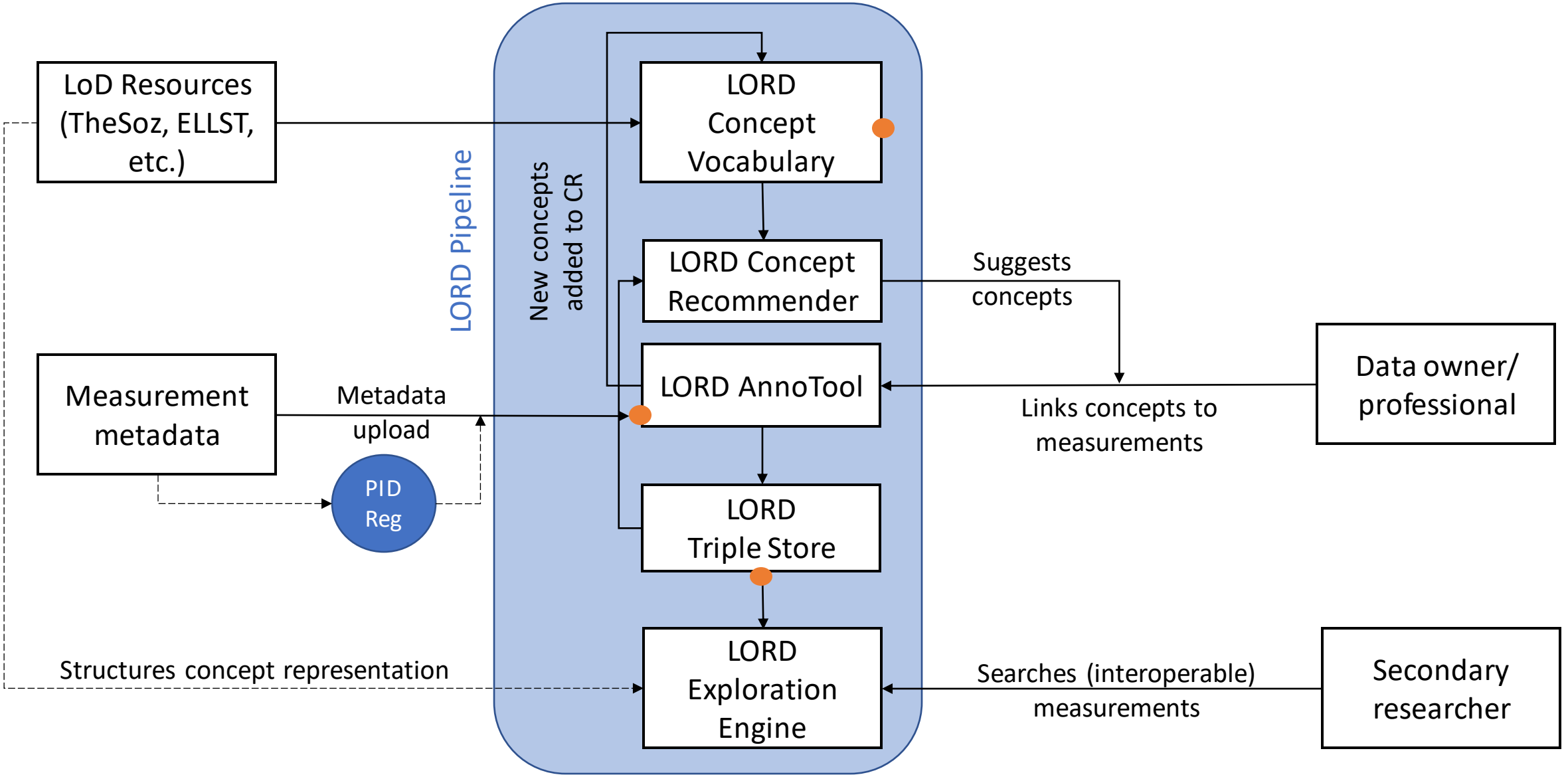
IV. Lessons Learned from the pilot study III



V. Outlook: developing the LORD „pipeline“

- Current project only covers the exploration phase
- A user driven concept registry will require additional functionalities
  - Performant recommendation systems for concept based on measurement metadata
  - The possibility to create links between concepts (not part of the current tools)
- Start phase: curated corpus of terms and relationships for the concept registy to improve recommender systems
- Graph based concept exploration engine

V. Outlook: developing the LORD „pipeline“



Thank you for your attention

https://www.diw.de/de/diw_01.c.862891.de/projekte/linked_open_research_data_for_social_science_pilot_study__lordpilot.html

LORDpilot received funding from the German Science Foundation (Grant Number: 464413245)

III. The LORD Data Model: Example (in German)