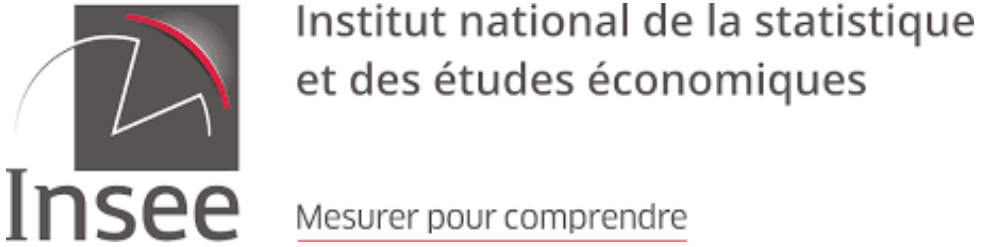
Managing repeated representation of variables in DDI Lifecycle

Christophe Dzikowski, INSEE, Jon Johnson, CLOSER, UCL,

Dr Sofiane Kab, INSERM

EDDI 2023, Ljubjlana



Collaborative Workshop

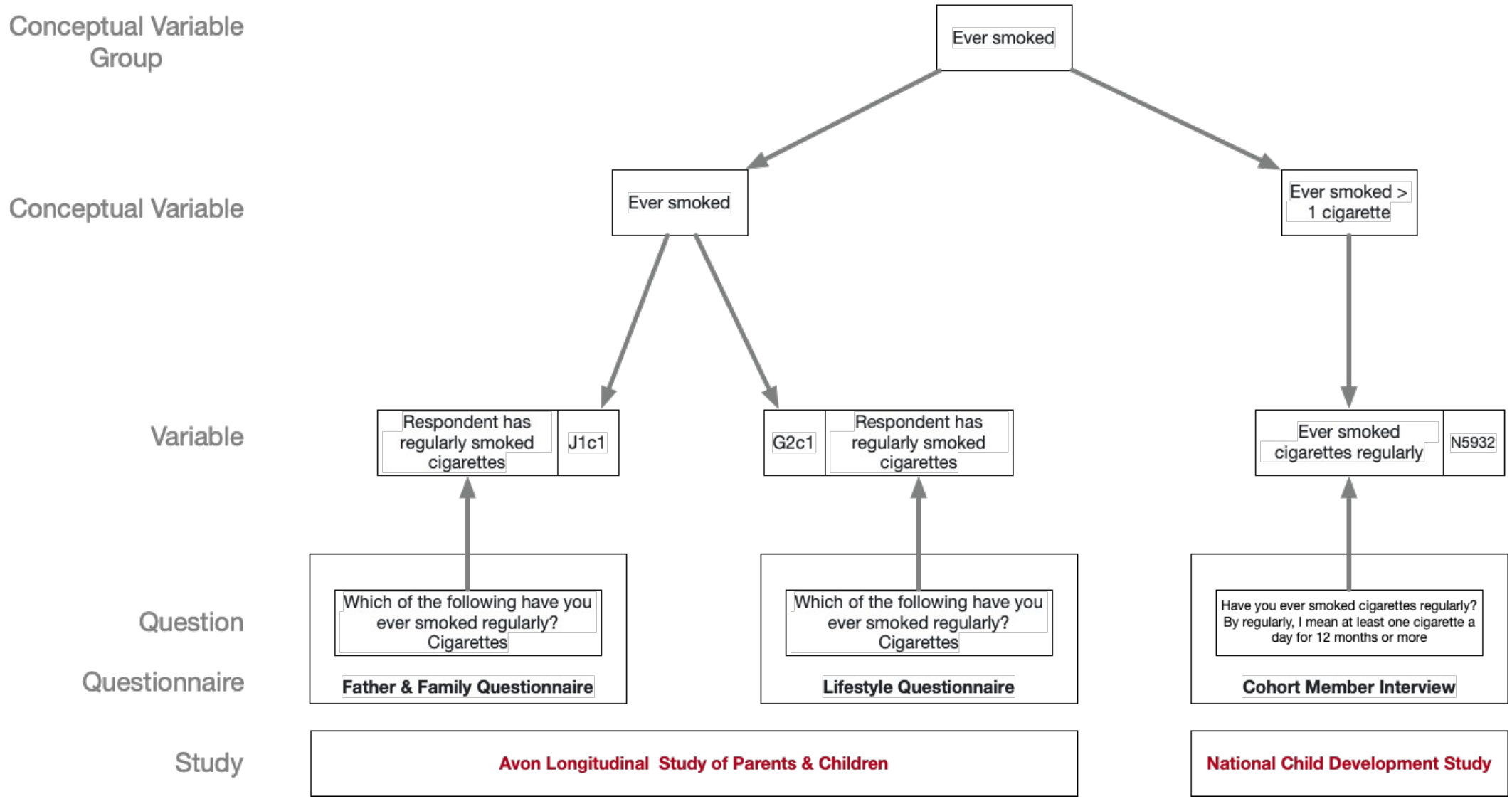
Two day workshop between:

- CLOSER: social science and biomedical longitudinal population studies
- Constances: National biobank, with surveys, linked health, admin data
- Insee: National Statistical Organisation

  - What are the difficulties encountered with repeated variables
  - Comparison of different use case
  - Which improvement of the metadata may be needed
  - What is sustainable to implement

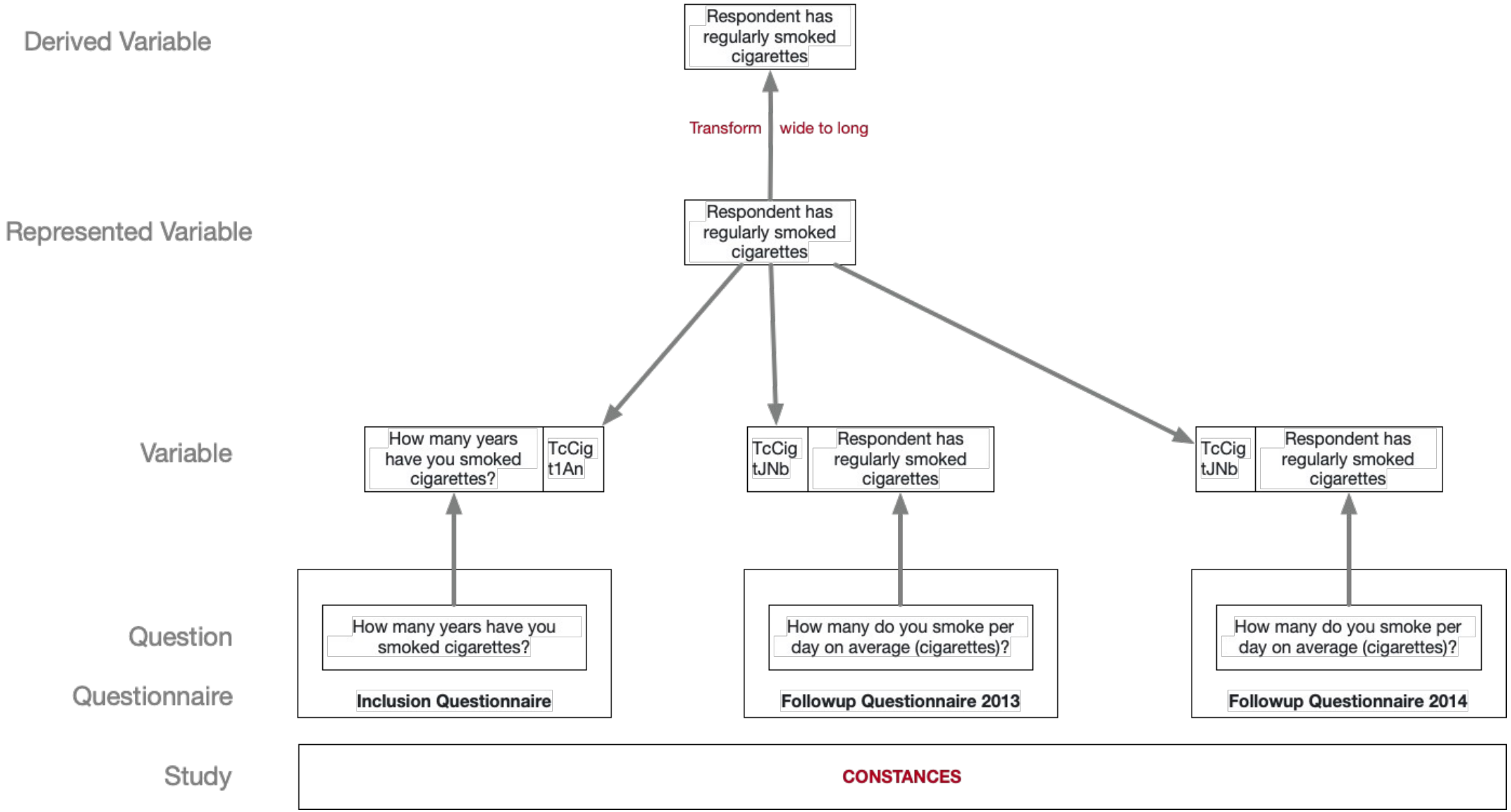
CLOSER- Use Case

- A collection of longitudinal population studies
- Survey questionnaires, mixture of paper and CAI data collections
- Metadata is retrospectively generated
- Data is disseminated for analysis primarily as collected (but data cleaning etc.)
- Comparability is within both a single study and across studies
- Code list and measurement comparison in concordance tables
- Researcher refines decisions based on alignment of high-level conceptual variables


Conceptual Variable Group
Conceptual Variable
Variable
Question
Questionnaire
Study
Ever smoked
Ever smoked
Ever smoked > 1 cigarette
Respondent has regularly smoked cigarettes
J1c1
G2c1
Respondent has regularly smoked cigarettes
Ever smoked cigarettes regularly
N5932
Which of the following have you ever smoked regularly? Cigarettes
Which of the following have you ever smoked regularly? Cigarettes
Have you ever smoked cigarettes regularly? By regularly, I mean at least one cigarette a day for 12 months or more
Father & Family Questionnaire
Lifestyle Questionnaire
Cohort Member Interview
Avon Longitudinal Study of Parents & Children
National Child Development Study

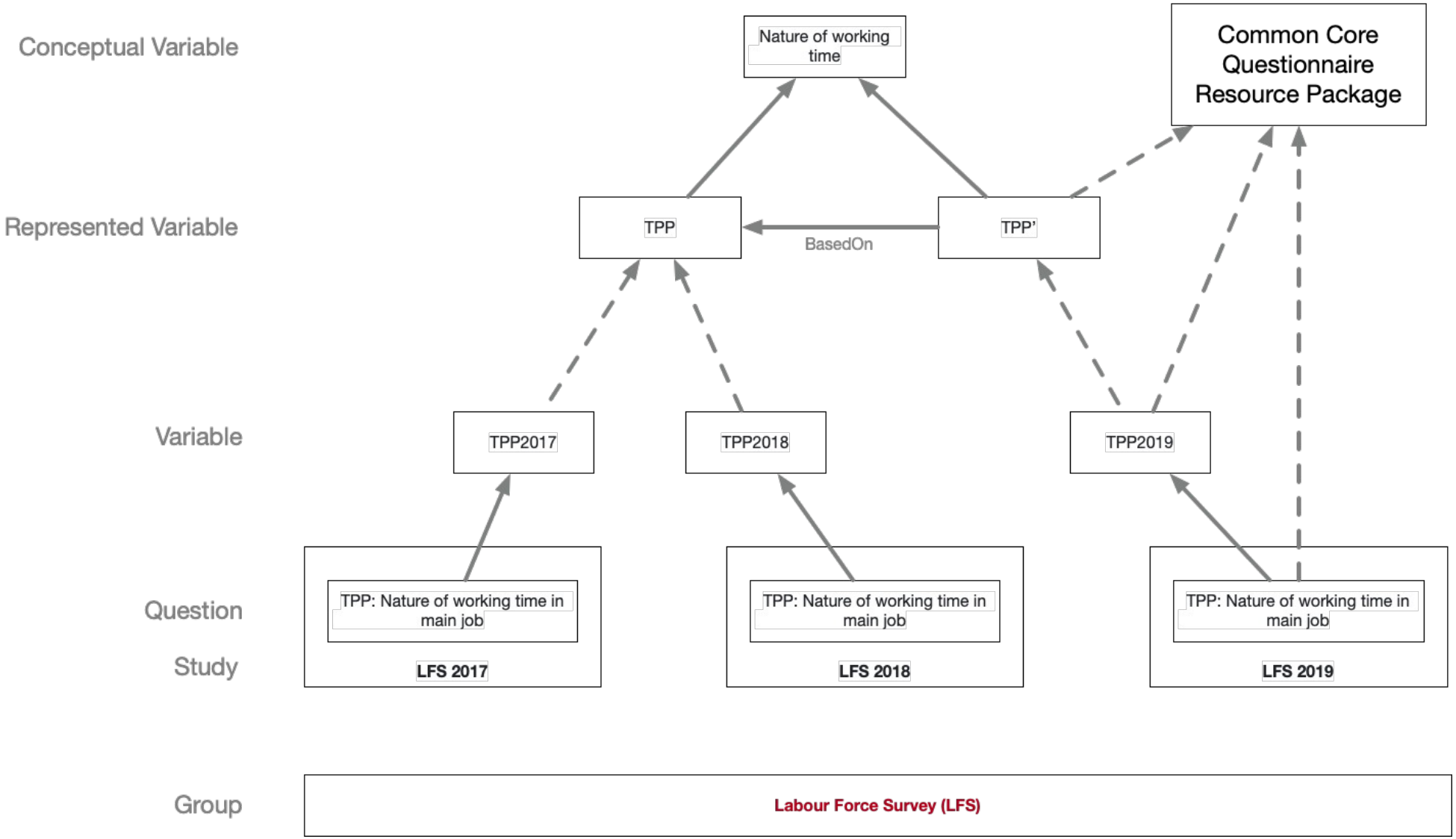
Constances- Use Case

- France's most extensive general cohort, a single longitudinal design with refresh cohorts
- Traditional questionnaires with administrative, biologic, and paraclinical data
- Metadata is currently retrospectively generated, but in the future transition to prospective metadata generation
- Data transformed to longitudinal/event formats for comprehensive analysis
  - By default this is harmonization to a represented variable
- Maintains comparability within the cohort
- Supports expansive data strategy for future research


Derived Variable
Respondent has regularly smoked cigarettes
Transform
wide to long
Represented Variable
Respondent has regularly smoked cigarettes
Variable
How many years have you smoked cigarettes?
TcCig t1An
TcCig tJNb
Respondent has regularly smoked cigarettes
TcCig tJNb
Respondent has regularly smoked cigarettes
Question
How many years have you smoked cigarettes?
How many do you smoke per day on average (cigarettes)?
How many do you smoke per day on average (cigarettes)?
Questionnaire
Inclusion Questionnaire
Followup Questionnaire 2013
Followup Questionnaire 2014
Study
CONSTANCES

INSEE- Use Case

- Cross-sectional surveys with reuse of classifications and represented variables
- Questionnaires, mixture of paper, CAI data collections, specification with DDI L
- Collection instruments creation is driven by the metadata
- The process between the collected data and the disseminated data is still to be build, this is a long term objective
- Thus, metadata **concerning the represented variable for documenting the disseminated data** is retrospectively generated, but in the future will be carried out prospectively , drawn from the instanced variables coming from the questionnaire


Conceptual Variable
Represented Variable
Variable
Question
Study
Group
Nature of working time
Common Core Questionnaire Resource Package
TPP
BasedOn
TPP'
TPP2017
TPP2018
TPP2019
TPP: Nature of working time in main job
TPP: Nature of working time in main job
TPP: Nature of working time in main job
LFS 2017
LFS 2018
LFS 2019
Labour Force Survey (LFS)

Comparability Criteria

- Data originates from the same question
  - This may be in a different questionnaire and asked in a different order
  - May take place at a different time
  - Comparable response domain and categories / measurement
- Universe
  - Longitudinal comparison WITHIN a study, (CLOSER and Constances) comparing universe is not a major concern as comparing the same individuals
  - Cross sectional (INSEE) and cross-study (CLOSER) understanding differences between universes is **very** important
- Other
  - Mode changes, and respondent may be important to some researchers

Variable cascade

- The variable cascade enables grouping of variables:
  - A conceptual variable that defines a common concept
  - A represented variable that additionally defines a common representation
  - Instance variables that may have the same representation.

- It does not explicitly enable provenance - however
  - Provenance of instance variables can be established though the Question Reference (QuestionItem, QuestionGrid, QuestionBlock)
  - QuestionConstruct within an instrument defines the universe and;
  - DataCollection the broad spatial and temporal coverage.

Controlled Vocabularies

- BasedOn
  - All item types support the BasedOn attribute
  - This supports the ability to provenance an item which is re-used in a different context (whether changed on not)
  - The GESIS CV for Variable relations proposes way of applying a CV to variables
    - https://vocabularies.cessda.eu/vocabulary/Variables-Relations?lang=en
  - This may be a way on placing provenance closer to the variable
  - This may help the researcher to understand and compare variables

Controlled Vocabularies

- Universes
  - Universes in some contexts are very important to ensure that there is a valid comparability
  - Using information held in QuestionConstructs can be used to define and classify universes
    - AGE > 16
    - Employees > 1000 < 5000
  - Universe can be expressed as the intersection of 2 (or more) universes
    - Variable references two universe's "Women" , "age > 40" , then we define the universe of women over 40.

Initial Conclusions

- Questions is important in assisting in comparison of repeated measures
- Variable cascade is not completely aligned between DDI-Lifecycle and DDI-CDI.
  - DDI-CDI also allows Universe on represented variables
- Different use cases in the use of the cascade and the universe
  - Longitudinal analysis / comparison, universe is implied by joining same persons
  - Cross-sectional analysis / comparison, universe is required before comparison
- BasedOn at a variable (instance or represented) level can assist in management of comparable items
  - Different use cases will benefit in different ways
  - Continue discussion on controlled vocabularies for BasedOn using the three use cases

Future Plans

- Continue the collaboration
  - Workshop in April 2024
- Developing Best Practice
  - https://github.com/Making-Sense-Info/Varese
- Potential to include further collaborators with new use cases