# Use Cases

Participants: Esra, Wolfgang, Sanda, Jon

## Data Collection

Simple questionnaire

Identify the question and connect it with external system (e.g. a maintainable ID?)

Document the question text and question label

Substitute he/she dependant on gender question previously answered

Document Interviewer instruction

Document introductory text to question

Document the sequence of questions, together (optionally) with numbers

Document the filters (input filters, output filters)

Document show card

Document language

Document answer categories or open ended or measurement

Show question block, consisting of more than one item

Relationship question - concept

Relationship question - (sub-)universe

Relationship question - mode of data collection

Relationship question - collection place

Relationship question - date of collection

Relationship question - publisher/provider

Relationship question - classification/keywords

Re-use of question in Data Collection

Identify the re-used question

Relationship to variable

Simple protocol

Define the observational object

Define the measure, together with the concept

Define the collection time and place

Relation to Variable(s), within dataset

Document instructions

Document choice of web or paper mode within an interview, break in interview - e.g. switches from web to face-to-face

Obtaining register data (official data)

Polling or update of same data structure to update records

Internet source (e.g. web scraping) - requires workflow and protocol for information captured about the web scraping event. Transformation path from source to dataset

Which type of source (HTML, specific social media platform, etc.)

Selection of sources: Universe, sampling (optional), filter by criteria?

Which data is being captured? Is it classified (procedure?)

Coverage by topics (e.g. hashtags, keyword filtering etc)

Geographic or IP-range coverage, date/time coverage

Specifics about the application (Twitter, Facebook etc) as a datasource, also specific APIs or tools/commercial services may be used

Connecting information within the network if a “snowball” approach is used

Legal and ethical considerations (informed consent, intellectual property, copyright)

Streaming data source (e.g. smart meter data, IOT) - what is limited due to requirement of qualitative

In addition to the “Internet source” information:

Update interval of the data and/or dataset

Versioning issues (identification, documentation)

Selections while capturing the stream

Think about re-use

## Data Description

File name, location, identifier, software type

Relationship to study

Data layout

Simple rectangular dataset

One record per case

Multiple records per case

Hierarchical/Multilevel data

Records relationships and linkage

Aggregate data

Describe cell content in relation to source variables (dimensions)

Statistical Classification

* (example already in process of being done in XML)

Variable cascade

* (example already in process of being done in XML)

Variable relationship to question/capture in instrument

Weight variable composed of two weight variables

Multiple weight variables with different uses for analysis

Logical description of variables

Instance variable description:

ID, Name, Label, Description, Categories, link to question, link to concept (?)

Relationship Instance/Represented/Conceptual variable

Represented Variable Description:

ID, (name?) (Label?) Description, Categories, link to question (?) link to concept (?)

Relationship to Conceptual variable

Conceptual variable description:

ID, (name?) (Label?) Description, Categories, link to question (?) link to concept

Think about re-use

## Data Transformation

Recode a variable into another (new) variable

Create new variable based on a procedure applied to one or more existing variables (e.g., compute, etc.)

Harmonise two variables using a represented variable

Verification processing during archive ingest (data checks)

* Data contains only values in a code list
* Data contains only values within a range

Confidentialization

* Is this just basically a recode (income > $100,000 do something)
* Check for combinations of variables that have fewer than x respondents in them

Imputation

* Identify sensitive variables (address, email, income, profession, geographic etc.)
* If a is missing a= b+10%

Build index by calculation of several variables, relation to concept

Think about re-use

## Study Information

Simple Study

Document name

Document sampling procedure

Document mode of collection

Document universe

Document coverage (spatial, temporal, topical)

Document persons/affiliations/institutions (creator, contributor, publisher, funding agency):

Document on contributor (name, identifier, type,...)

Document on funding (name, identifier, award, ...)

Document availability/rights

Document on study descriptions (abstract, table of content,...)

Document classifications/keywords

Identify studies (study number, study id)

Citation of study

Publications on study (structured/unstructured)

Connecting a dataset to an instrument inside a data collection

Connecting a dataset to a Study directly

Grouping two instruments in a study

Grouping two studies in a study series

Archiving a study (including description, provenance and ownership)

Think about re-use

## Data Description & Capture

Common Data Element example

* <https://cde.nlm.nih.gov/home> or <http://webarchive.nationalarchives.gov.uk/20160106185646/http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-and-questions/long-lasting-health-conditions-and-illnesses--impairments-and-disability.pdf>

Outputting household loop as pid/age vs age\_1 pid1 age\_2 pid2 etc