The Longitudinal Impossible Dataset: Helping Users Navigate the ONS Longitudinal Study

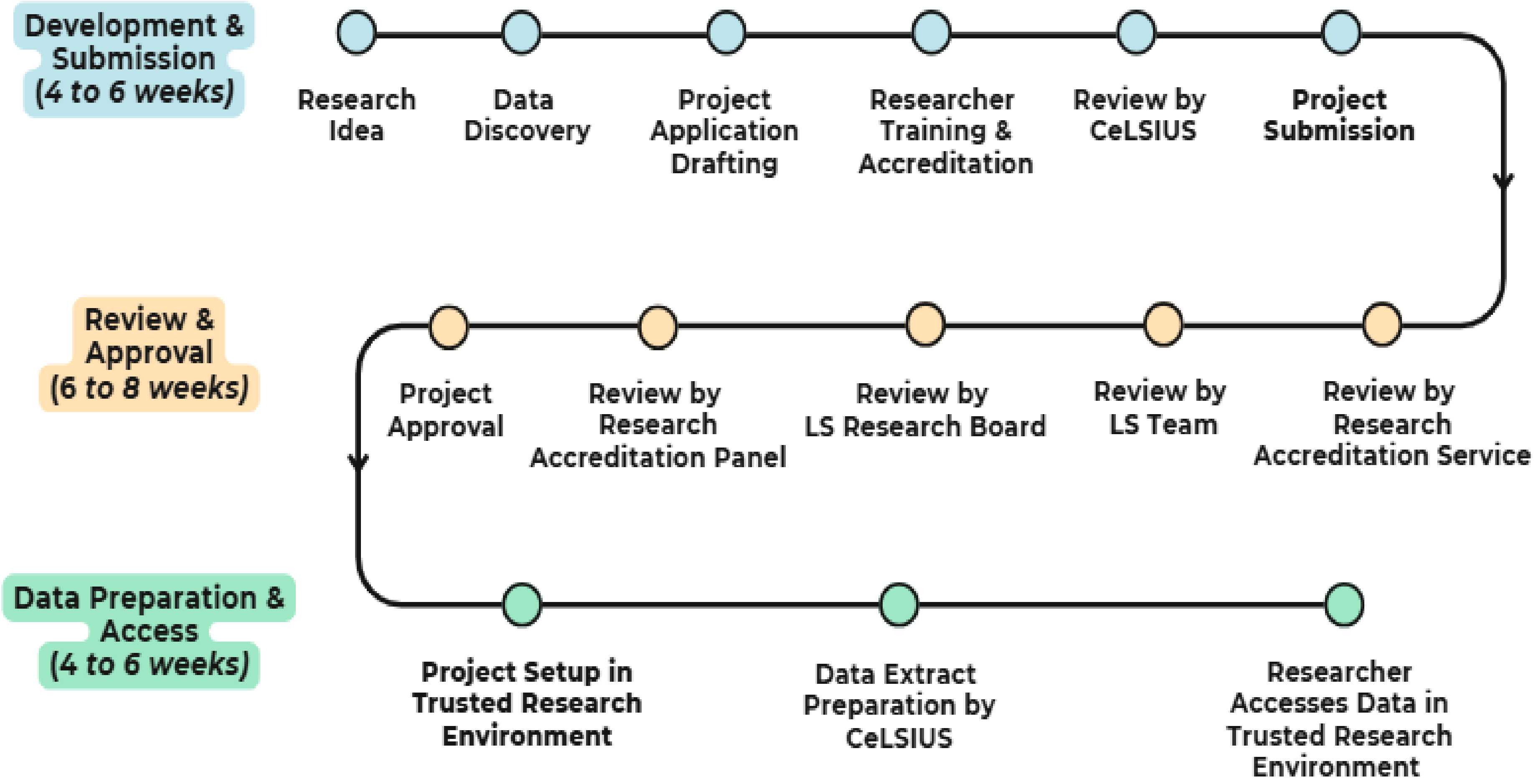
Supported by funding from: UKRI Economic and Social Research Council

Andreas Mastrosavvas

Nicola Shelton

Epidemiology and Public Health, University College London (UCL)

Accessing the ONS LS



Development & Submission (4 to 6 weeks)

Research Idea

Data Discovery

Project Application Drafting

Researcher Training & Accreditation

Review by CeLSIUS

Project Submission

Review & Approval (6 to 8 weeks)

Review by Research Accreditation Service

Review by LS Team

Review by LS Research Board

Review by Research Accreditation Panel

Project Approval

Data Preparation & Access (4 to 6 weeks)

Project Setup in Trusted Research Environment

Data Extract Preparation by CeLSIUS

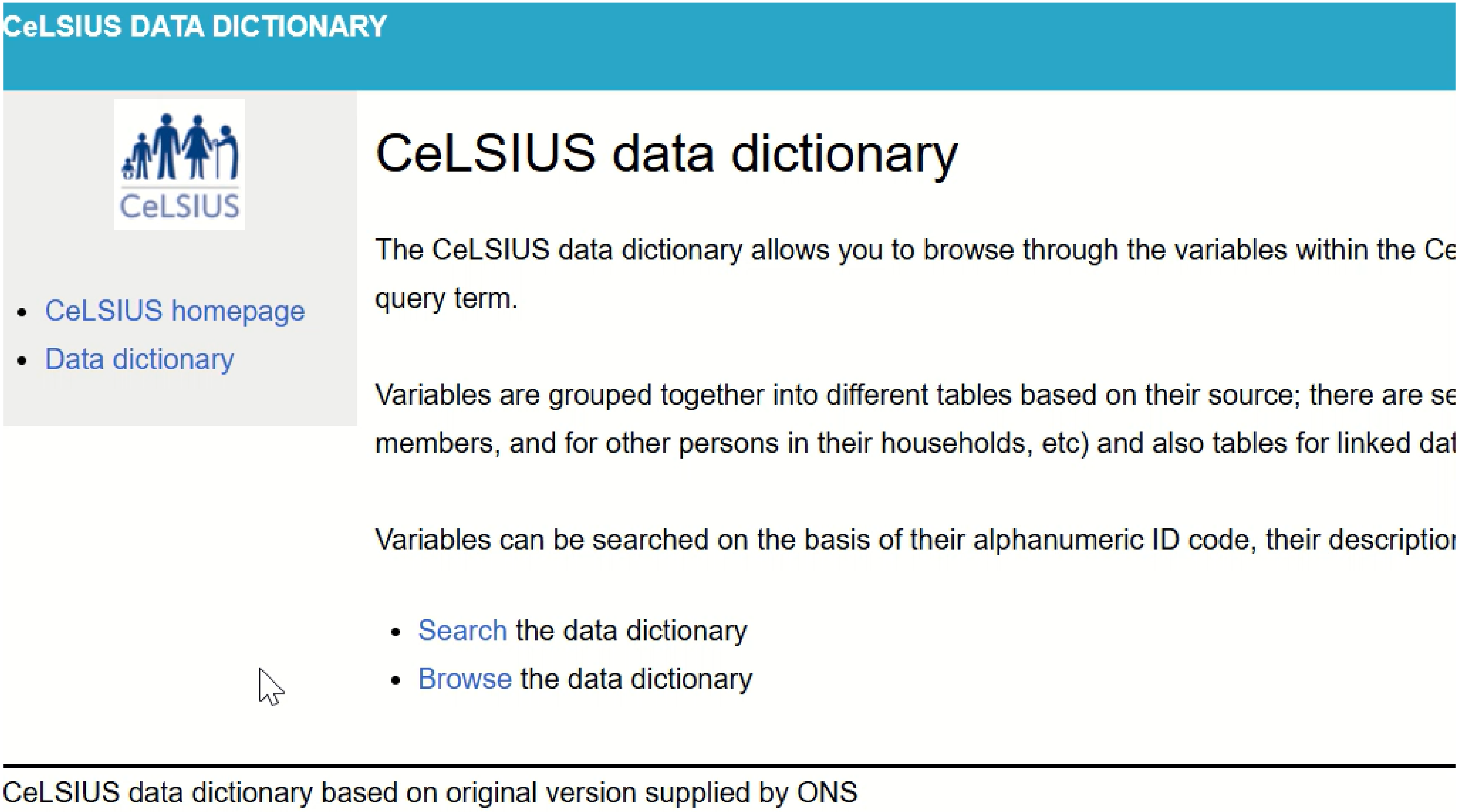
Researcher Accesses Data in Trusted Research Environment

Accessing the ONS LS – common sources of delay

- Data discovery is complex
  - Over 600 LS variables per Census
  - Researchers need to think carefully about variable selection pre-application
  - Researchers sometimes find they need to re-apply for additional variables

- Researchers unable to develop code from the start of their project

The CeLSIUS data dictionary



CeLSIUS DATA DICTIONARY

- CeLSIUS homepage
- Data dictionary

CeLSIUS data dictionary

The CeLSIUS data dictionary allows you to browse through the variables within the Ce
query term.

Variables are grouped together into different tables based on their source; there are se
members, and for other persons in their households, etc) and also tables for linked dat

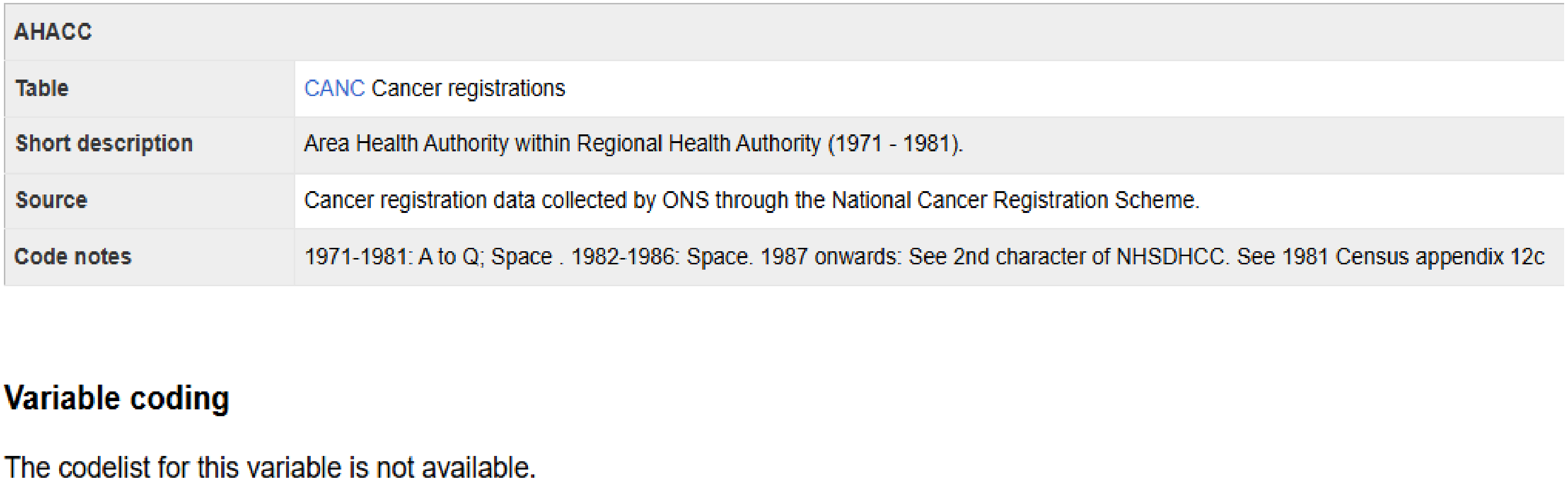
Variables can be searched on the basis of their alphanumeric ID code, their descriptio

- Search the data dictionary
- Browse the data dictionary

CeLSIUS data dictionary based on original version supplied by ONS

The CeLSIUS data dictionary

- Can be challenging to navigate, esp. for new users
- Variable metadata can be unstructured, esp. for linked data



AHACC	
Table	CANC Cancer registrations
Short description	Area Health Authority within Regional Health Authority (1971 - 1981).
Source	Cancer registration data collected by ONS through the National Cancer Registration Scheme.
Code notes	1971-1981: A to Q; Space . 1982-1986: Space. 1987 onwards: See 2nd character of NHSDHCC. See 1981 Census appendix 12c

Variable coding

The codelist for this variable is not available.

The Longitudinal Impossible Dataset

Objectives

- Enable users to better understand what variables they need to request
- Enable users to develop code prior to data access
- Avoid real or perceived disclosure of ONS LS data

Concept

- A fake, user-generated dataset that…

  … has the *same structure* as ONS LS data

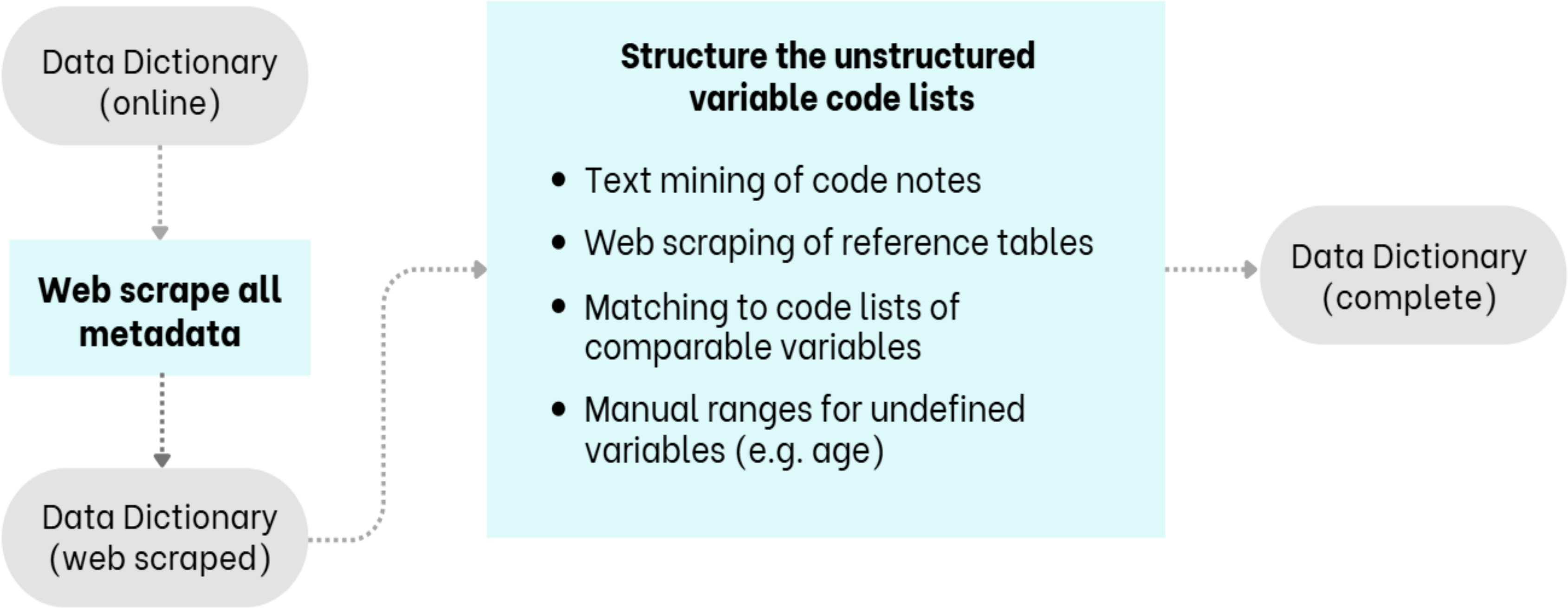
  … is purely based on and representative of *public metadata*

  … is *impossible* – observations are clearly not real

  … is *longitudinal* – adheres to some realistic rules e.g. on ageing

The Longitudinal Impossible Dataset

Development – Input Metadata



Data Dictionary (online)

Web scrape all metadata

Data Dictionary (web scraped)

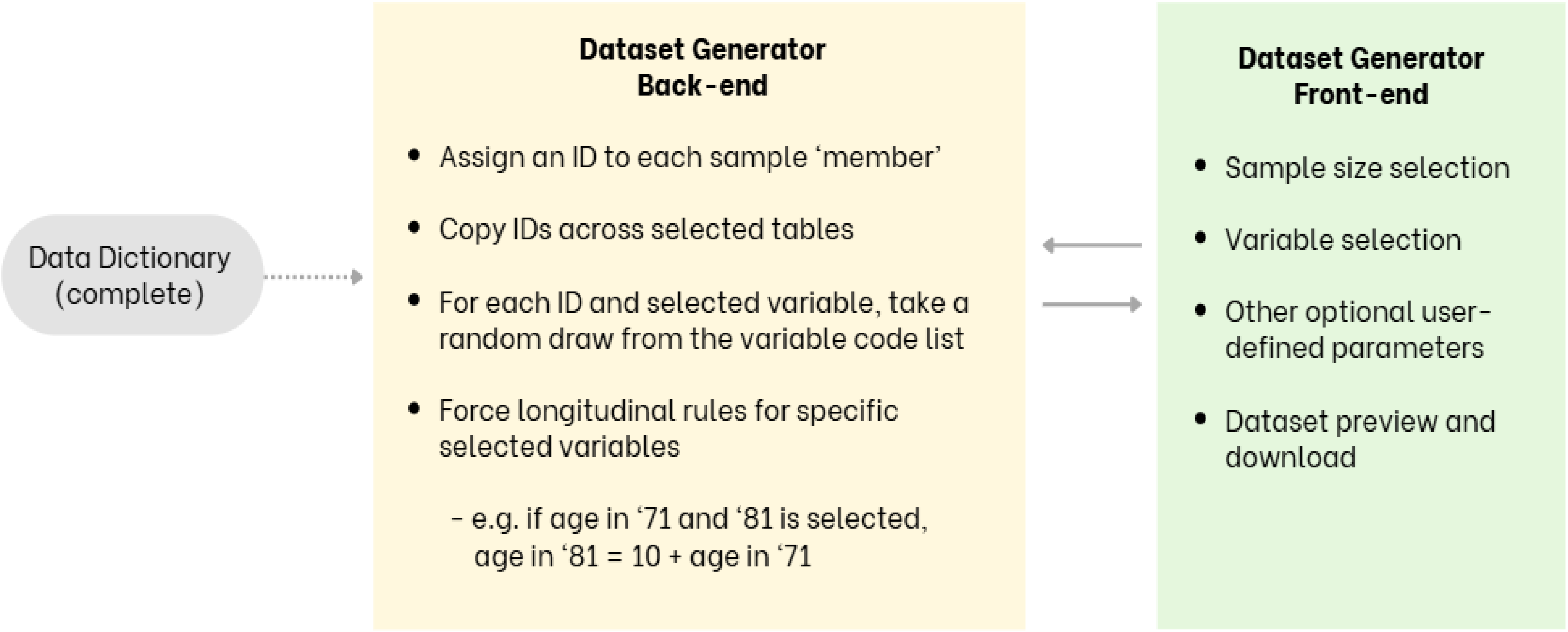
Structure the unstructured variable code lists

- Text mining of code notes
- Web scraping of reference tables
- Matching to code lists of comparable variables
- Manual ranges for undefined variables (e.g. age)

Data Dictionary (complete)

The Longitudinal Impossible Dataset

Development – Dataset Generator Application (RShiny)



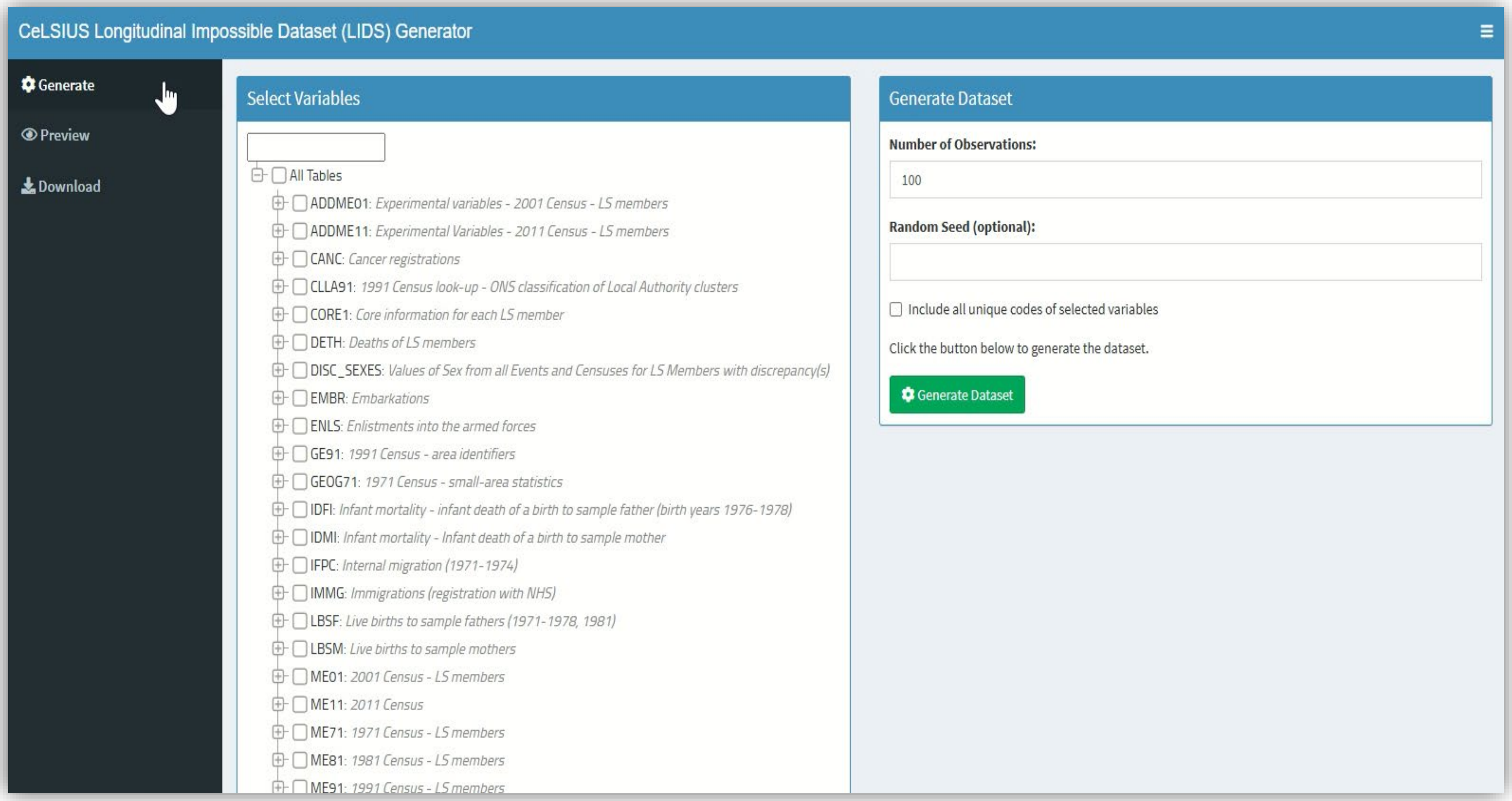
Data Dictionary (complete)

Dataset Generator Back-end

- Assign an ID to each sample 'member'
- Copy IDs across selected tables
- For each ID and selected variable, take a random draw from the variable code list
- Force longitudinal rules for specific selected variables
  - e.g. if age in '71 and '81 is selected, age in '81 = 10 + age in '71

Dataset Generator Front-end

- Sample size selection
- Variable selection
- Other optional user-defined parameters
- Dataset preview and download



CeLSIUS Longitudinal Impossible Dataset (LIDS) Generator

Select Variables

- ⊟ □ All Tables
  - ⊞ □ ADDME01: *Experimental variables - 2001 Census - LS members*
  - ⊞ □ ADDME11: *Experimental Variables - 2011 Census - LS members*
  - ⊞ □ CANC: *Cancer registrations*
  - ⊞ □ CLLA91: *1991 Census look-up - ONS classification of Local Authority clusters*
  - ⊞ □ CORE1: *Core information for each LS member*
  - ⊞ □ DETH: *Deaths of LS members*
  - ⊞ □ DISC_SEXES: *Values of Sex from all Events and Censuses for LS Members with discrepancy(s)*
  - ⊞ □ EMBR: *Embarkations*
  - ⊞ □ ENLS: *Enlistments into the armed forces*
  - ⊞ □ GE91: *1991 Census - area identifiers*
  - ⊞ □ GEOG71: *1971 Census - small-area statistics*
  - ⊞ □ IDFI: *Infant mortality - infant death of a birth to sample father (birth years 1976-1978)*
  - ⊞ □ IDMI: *Infant mortality - Infant death of a birth to sample mother*
  - ⊞ □ IFPC: *Internal migration (1971-1974)*
  - ⊞ □ IMMG: *Immigrations (registration with NHS)*
  - ⊞ □ LBSF: *Live births to sample fathers (1971-1978, 1981)*
  - ⊞ □ LBSM: *Live births to sample mothers*
  - ⊞ □ ME01: *2001 Census - LS members*
  - ⊞ □ ME11: *2011 Census*
  - ⊞ □ ME71: *1971 Census - LS members*
  - ⊞ □ ME81: *1981 Census - LS members*
  - ⊞ □ ME91: *1991 Census - LS members*

Generate Dataset

Number of Observations:

100

Random Seed (optional):

□ Include all unique codes of selected variables

Click the button below to generate the dataset.

The Longitudinal Impossible Dataset

Key features

- Purely based on public metadata – anyone could ‘make their own’
- Purely random relationships between variables – anything is possible
- Customisable – helps users think about what variables they need

What LIDS can be used for

- Discovering relevant variables and their coding
- Familiarising with LS data extract relational table structure
- Developing mock code before accessing real data

What LIDS cannot be used for

- Estimating size of sub-populations of interest
- Identifying relationships between variables/events

The Longitudinal Impossible Dataset

Next steps

- Sense-checking extracted code lists from unstructured metadata
- Explore additional longitudinal rules (e.g. births and deaths)
- Explore additional user-defined parameters
- Supporting documentation (interface & website)
- Launch – by end of 2025

Census Innovation at CeLSIUS: other initiatives

User support for secure Census flow data and microdata

- CeLSIUS-ONS Steering Group
- Developing supporting guides
  - Census 2021 and the COVID-19 pandemic
  - Data catalogue
  - Data access process and restrictions (SRS/IDS)
- Exemplary research projects
- Other support and guidance

Questions and comments

Thank you!

Andreas Mastrosavvas – amastrosavvas@ucl.ac.uk

Nicola Shelton – n.shelton@ucl.ac.uk

CeLSIUS – celsius@ucl.ac.uk