

Introduction to the DDI Metadata Standard



European DDI User Conference
“Discovering DDI” Track
26 November 2021



License: [Creative Commons Attribution 4.0 International \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)¹



Jane Fry, Data Services Librarian, Carleton University



Arofan Gregory, Data Management
Consultant



Jared Lyle, Executive Director, DDI Alliance and Archivist, ICPSR, University of Michigan



Barry Radler, Distinguished Researcher, University of Wisconsin-Madison

What is DDI?

- DDI = Data Documentation Initiative
- A free and open international metadata standard
 - Used primarily in the social and behavioural sciences, economics, health
 - An open standard designed for data sharing and reuse
- A structure for describing data and its related information
- Describes data from surveys and other observation-based data collection methods
 - Currently moving towards covering new data types and data from new domains.

DDI Products

- The DDI standard has developed over time
 - <https://ddialliance.org/products/overview-of-current-products>
- Currently has two main products:
 - DDI-Codebook
 - DDI-Lifecycle
- Each designed for a different purpose
- Other products:
 - Controlled vocabularies
 - RDF (XKOS)
 - Standard Data Transformation Language (SDTL)

Developing Products

- The DDI standard continues to develop based on what users want
 - <https://ddialliance.org/products/developing-products-of-the-alliance>
- Developing products:
 - DDI-Cross Domain Integration (DDI-CDI)
 - RDF (Disco)

DDI users include...

Agencies

- Norwegian Social Science Data Services
- Harvard University
- Statistics Canada
- Health Canada
- ICPSR
- U.S. Bureau of Labor Statistics
- ESRC Data Archive (UK)
- Zentralarchiv für Empirische Sozialforschung (GESIS)

Projects

- Australian Social Science Data Archive
- CESSDA Data Portal
- DAMES Project (UK)
- DataFirst (at University of Cape Town)
- Israel Social Science Data Center
- ICPSR Data Catalog
- ODESI (Canada)
- Statistics New Zealand

DDI is being used in over 80 countries around the world!



DDI Alliance

- Self-sustaining member organization (created in 2003)
- Members have a voice in DDI development
- Organizational structure
 - Executive Board - The policymaking and oversight body of the Alliance.
 - Scientific Board - Facilitates the scientific and technical work activities.
 - Technical Committee - Maintains and updates the various DDI products, in collaboration with the different working groups of the DDI Alliance.
 - Working Groups - Convened to work on different activities and topics within the work areas of the DDI Alliance.

Document, Discover and Interoperate

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI is a free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving. Documenting data with DDI facilitates understanding, interpretation, and use -- by people, software systems, and computer networks. Use DDI to **D**ocument, **D**iscover, and **I**nteroperate!

Why Use DDI?

- ✓ Generate interactive codebooks
- ✓ Implement data catalogs
- ✓ Build question banks
- ✓ Create concordance mappings
- ✓ Harmonize and compare data
- ✓ Manage longitudinal data sets

[Find Out More!](#)

Featured DDI Adopters



[CESSDA Data Catalogue](#)

The Consortium of European Social Science Data Archives (CESSDA) offers a data catalogue providing access to research data from archives across Europe. The catalogue includes nearly 30,000 studies (~22,000 with english descriptions) distributed by

News

[New DDI Alliance Privacy Policy](#)

Mar 17, 2021

[DDI Product Suite Logos](#)

Mar 5, 2021

[2021 DDI Annual Meetings](#)

Mar 1, 2021

Benefits of Membership

- **Vote** on Alliance products, including additions and modifications to specifications.
- Be **elected** to Alliance leadership, including the Executive Board and the Scientific Board.
- Have a **seat** at annual meetings, including the Annual Meeting of Members and the Annual Meeting of the Scientific Community.
- **Participate** on working groups, which develop DDI products and advise DDI leadership.
- **Display** the Alliance trademarks on promotional material and publicize participation in the Alliance.
- **Send participants** to Alliance-sponsored events at reduced or no cost.
- **Request access** to Member-only information.

DDI-Codebook

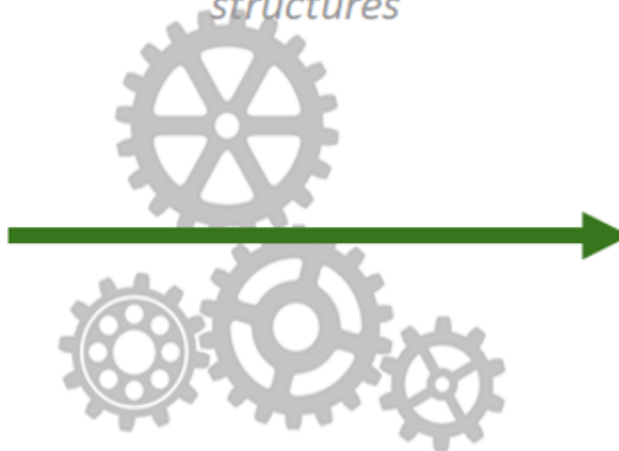
Jane Fry
Data Services Librarian, Carleton University

Metadata not structured



Lost Metadata
Manager

*DDI Codebook
structures*



Metadata structured by DDI Codebook standard



DDI Codebook

- A structure facilitating the production of machine-readable codebooks and data dictionaries.
- Built to emulate a physical codebook
 - that is, to catalog a dataset, to describe a single study
 - Expressed in an XML schema
- Different tools are used to interpret it
 - Including: Nesstar, Colectica, Dataverse, MTNA, Kuha 2, R
- The latest version of DDI-Codebook is version 2.5

DDI Codebook ...

- Relatively straight forward
- Sections
 - *Document Description*
 - *Study Description*
 - *Data Files Description*
 - *Variable Description*
 - *Other Study Related Materials*

DDI Codebook - Section 1

- Document Description
 - *Describes the actual document that you are putting together to describe the survey or study*
 - *For example*
 - *Title*
 - *Date*
 - *Author (of document)*
 - *...*

DDI Codebook - Section 2

- Study Description
 - *Describes the actual survey or study that was conducted*
 - *For example: purpose, how, why, where, how long, funding of the study, ...*
 - *Does not describe the data file*
 - *Many of the metadata tags in this section are the same as the ones in the first section - 'Document Description'*

DDI Codebook - Section 3

- Data Files Description
 - *Exactly what it says!*
 - *Describes the datafiles that were created by the study*
 - *For example*
 - *Number of variables*
 - *Number of cases*
 - *Structure of the datafile*
 - *Hierarchical, rectangular (the default) or relational*

DDI Codebook - Section 4

- Variable Description
 - *Any and all information there is about each and every variable in the datafile*
 - *For example*
 - *the question text (including any pre-questions)*
 - *the variable label*
 - *all the values and associated labels for that particular variable*
 - *the population*
 - *any additional notes*
 - *for example, instructions for the interviewer*

DDI Codebook - Section 5

- Other Study Related Materials
 - *These are documents that are separate from this particular file you are creating*
 - *A link is provided for separate download*
 - *For example*
 - *Questionnaires*
 - *User Guides*
 - *Codebooks*
 - *...*

Examples of DDI Tags

<titl>Canadian Community Health Survey, 2012: Annual Component </titl>

<labl>Questionnaire (.pdf)</labl>

<dataDscr><notes>The variables in this study are identical to earlier waves. </notes></dataDscr>

<titl>Canadian Gallup Poll, May 2000</titl>

<dataChck>Quality checks were performed by Carleton University Data Centre. </dataChck>

<titl>Survey of Household Spending, 2001 [Canada]</titl>

<varQnty>255</varQnty>

<titl>Canadian Gallup Poll, May 1949, #186</titl>

<copyright>Copyright Gallup Canada Inc., 1950</copyright>

Best Practices Document

- Started with the DDI Alliance technical document
 - Made it human readable
- Updated - 2019
 - Used when marking up in [ODESI](#)
 - Uses the Nesstar platform
- Has lots of examples for the tags
- Great for training purposes

Best Practices Document Based on DDI 2.x

Version 3.1



odesi.ca

January 2019

Jane Fry (Carleton University)
Alexandra Cooper (Queen's University)
Susan Mowers (University of Ottawa)
Carys Carrington (Carleton University)

1.1.6.3

<notes>

Notes and Comments

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, type, subject, level, resp, sdatrefs

Description: Used to indicate additional information regarding the version or the version responsibility statement for the marked-up document, in particular to indicate what makes a new version different from its predecessor. "Notes" sections appear in several places in the DTD. The attributes for notes permit a controlled vocabulary to be developed (type and subject), the level of the DTD to which the note refers to be identified (study, file, variable, etc.), and the author of the note to be indicated (responsibility).

Note 1:

Every time this document is changed, this tag should be used, with the most recent note being entered first, followed by the older notes.

Example 1:

```
<notes>Additional study information was added to this document.</notes>
```

Example 2:

```
<notes resp="Smith, Jane">Additional information on derived variables  
has been added to this marked-up version of the documentation.</notes>
```

Example 3:

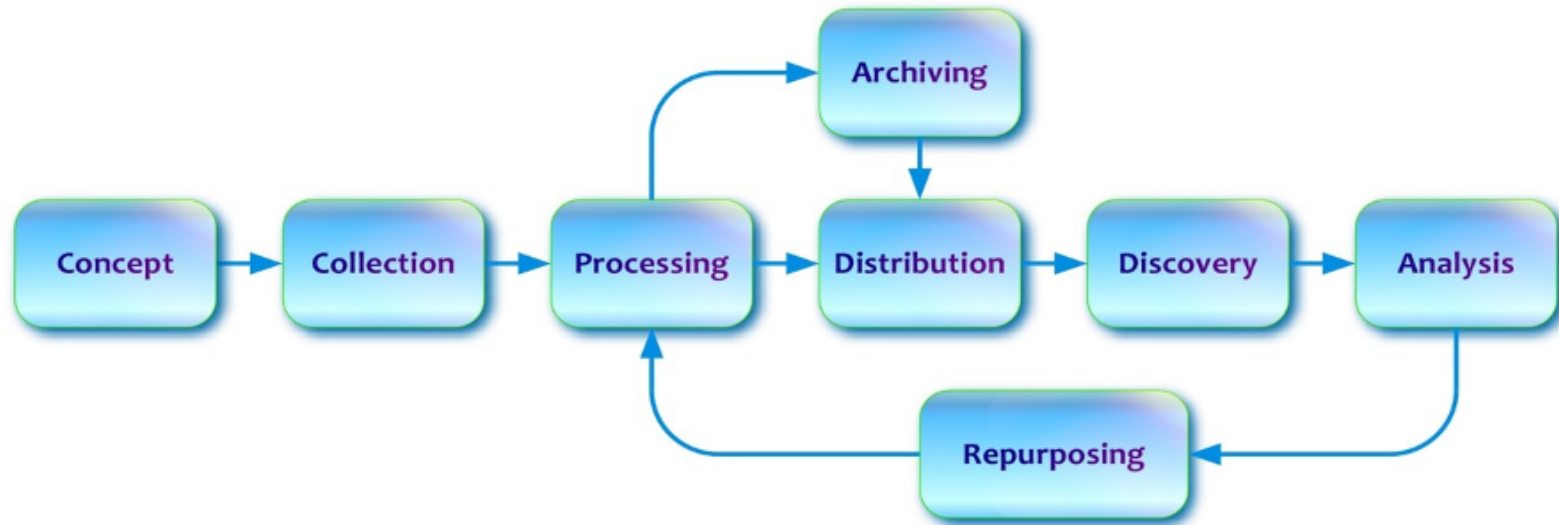
```
<notes> Version 2008-01-18 - made file compliant to <odesi> Best Practices Standards;  
added documentation for each variable.<br />  
Version 2007-11-10 - changed information in Document Description, and Other Materials.  
</notes>
```

DDI-Lifecycle

Barry Radler

Distinguished Researcher, University of Wisconsin-Madison

Research data lifecycle



DDI Lifecycle 3.3



- Website: ddialliance.org

The screenshot shows the website ddialliance.org/Specification/DDI-Lifecycle/3.3/#description. The page features the DDI logo and a navigation menu with options: Learn, Products, Membership, Events, Publications, and About. The 'Products' menu is open, showing a list of items: Overview of Current Products, DDI Lifecycle (highlighted), DDI-Codebook, Controlled Vocabularies, XKOS, SDTL, DDI Agency ID Registry, and Developing Products of the Alliance. Below the menu, there is a 'Download this product' button and an XML Schema entry point: <https://ddialliance.org/Specification/DDI-Lifecycle/3.3/XMLSchema/instance.xsd>. The 'Content' section includes links for [Description](#), [Applications](#), [Informational Documents](#), [Markup Examples](#), and [License](#).

DDI Lifecycle 3.3



- Github: ddialliance.github.io/ddimodel-web/DDI-L-3.3

DDI Data Documentation 3.3

Topics

- Study
- Classification
- Data Capture
- Data Description
- Foundational
- Agent
- Non-Packaging Items
- All Content Items

Items and Fields

- All Item Types
- All Composite Types

Use Case: Midlife in the US



Key characteristics of MIDUS:

- **Multiple longitudinal samples**
- **Multidisciplinary design**
 - Survey, Cognitive, Daily Diary, Bio, Neuro
- **Many linked data products**
 - 28 primary datasets, 37,000 variables, N~13,000
- **Wide secondary usage – Open Data philosophy**
 - Top data download at ICPSR
 - 140k downloads; 24k users; 1,500+ publications
- **MIDUS-Colectica Portal**



Use Case: Midlife in the US



Particular benefits of DDI Lifecycle (3.3) for MIDUS:

- Intelligent search function
- Harmonization (internal, post-hoc)
- Facilitates Custom Data Extracts



The MIDUS Colectica Portal

<http://midus.colectica.org>




MIDUS Search Explore Baskets 7 Admin Help Feedback bradler@wisc.edu



Home

Welcome

Welcome to the MIDUS Colectica Portal. This website provides users access to publicly available data and richly-structured documentation for the MIDUS study.



The Portal is based on the [DDI metadata standard](#) and is powered by [Colectica software](#).



Midlife in the United States

Midlife in the United States (MIDUS) is a national longitudinal study of health and well-being that was conceived by a multidisciplinary team of scholars interested in understanding aging as an integrated bio-psycho-social process. Since its inception in 1995 MIDUS has continued to grow, and now includes data from over 12,000 individuals, comprising thousands of variables in different scientific topics among multiple cohorts.

This website is a one-stop portal that provides MIDUS data and metadata (information about research data) for exploration and analysis. As of 2016 the portal is organized around four functions:

- **Home:** The home page still allows users to browse variables, read abstracts, download instruments and documentation, and search for variables within individual MIDUS projects. Also included for each project is a link to the full project dataset available from the [official MIDUS archive at ICPSR](#).

Intelligent Search



- Searches different fields: variable name, label, question text, assigned concepts
- Search results are arrayed
- Intelligent searches across *ALL* 30k MIDUS variables

Series >

Variables

Smoking Sort by: Alphabetical

[Syntax](#)

Item types: Variables
Query: Smoking

Results 1 to 50 of 155 (0.02 seconds)

[K1SB2](#)

of cigarettes/day during heaviest year (current smoker)
On average, about how many cigarettes did you **smoke** per day in the one year in your life when you **smoked** most heavily?

[A1PA42](#)

Age began to **smoke** regularly
History of **Smoking**

[CACA38](#)

Age began to **smoke** regularly
History of **Smoking**

Appears Within

- Studies
 - MIDUS 1 Project 1
- Data Files
 - M1P1
- Variable Groups
 - M1P1 VariableGroup
 - Phone
 - Phone A: Health

Age began to smoke regularly

MIDUS 1 Project 1 M1P1 M1P1 VariableGroup 3

< A1PA40

M1P1 (140 of 2098)

A1PA43 >

Variable Description

- Name A1PA42
- Label Age began to smoke regularly
- Question Text At what age did you begin to smoke REGULARLY?
- Backward Skip (IF A1PA41 = NEVER HAD A CIGARETTE; IF A1PA40 = NO OR DK)
- Dataset M1_P1_SURVEY_N7108_20190116

Value	Label	Frequency	% of valid	% of all
97	DONT KNOW	16		0.23%
99	INAPP	3,432		48.28%

Valid	Invalid	Min	Max	Mean	StdDev
3660	3448	5	56	18.76	4.76

Internal Harmonization



- Clarifies the related nature of versions of longitudinal and cross-cohort variables

- Survey
- Administration
- Recession Experience
- Health
 - Self-Evaluated Health
 - Physical/Mental Health Limits Work
 - Health at Age 16
 - History of Neurological Disorder
 - History of Heart Trouble
 - Cardiovascular Health
 - Angina
 - Chest Pain
 - Blood Pressure
 - Cancer
 - Hysterectomy/Ovary Removal
 - History of Smoking**
 - History of Drinking
 - Depression - Depressed Affect
 - Depression - Anhedonia
 - Depression - Depressed Affect and Anhedonia
 - Anxiety/Worry/Panic
- Education, Occupation, and Marital Status
- Household Roster and Children
- Caregiving
- Living Arrangements
- Race and Ethnicity
- Life Satisfaction
- Your Health
- Health Questions for Women
- Health Insurance
- Parent's Health
- Personal Beliefs
- Work
- Finances
- Community Involvement

	M1P1	M2P1	M3P1	MKE1	MKE2	MRP1	MKER
Age had first cigarette	A1PA41	B1PA37	C1PA37	BACA36	CACA36	RA1PA37	RAACA36
Ever smoked cigarettes regularly	A1PA40	B1PA38A	C1PA38A	BACA37	CACA37	RA1PA38A	RAACA37
Age began to smoke regularly	A1PA42	B1PA38B	C1PA38B	BACA38	CACA38	RA1PA38B	RAACA38
Now smoke cigarettes regularly	A1PA43	B1PA39	C1PA39	BACA39	CACA39	RA1PA39	RAACA39
Cigarettes per day during heaviest year - Current smoker	A1PA44	B1PA40	C1PA40	BACA40	CACA40	RA1PA40	RAACA40
Ever tried to quit smoking	A1PA45	B1PA41	C1PA41	BACA41	CACA41	RA1PA41	RAACA41
Age last smoked regularly	A1PA46	B1PA42	C1PA42	BACA42	CACA42	RA1PA42	RAACA42
Cigarettes per day during heaviest year - Past smoker	A1PA47	B1PA43	C1PA43	BACA43	CACA43	RA1PA43	RAACA43
Ever used pipe, cigars, snuff, chew		B1PA44	C1PA44	BACA44	CACA44	RA1PA44	RAACA44
Lived with smoker growing up - Father		B1PA45A		BACA45A		RA1PA45A	RAACA45
Lived with smoker growing up - Mother		B1PA45B		BACA45B		RA1PA45B	RAACA45
Lived with smoker growing up - Other		B1PA45C		BACA45C		RA1PA45C	RAACA45
In home anyone smoke or use tobacco current		B1PA46	C1PA46	BACA46	CACA46	RA1PA46	RAACA46
At job anyone smoke or use tobacco ever		B1PA47	C1PA48	BACA47	CACA47	RA1PA47	RAACA47
At job anyone smoke or use tobacco ever		B1PA48	C1PA47	BACA48	CACA48	RA1PA48	RAACA48

Details...

Conceptual Variable

Name P1_ConceptualVariable_267

Label Lived with smoker growing up - Mother

Comparability Notes Not in M1, M3, and MKE2; was introduced at M2, MKE1, MR, MKER as a new baseline item.

Comparability Class General Item Availability; Baseline Item

Statistics

Code Comparison

Correspondence Tree

% of valid % of total

	M2P1	MKE1	MRP1	MKER
	B1PA45B	BACA45B	RA1PA45B	RAACA45B
YES	31.59%	36.15%	34.77%	43.98%
NO	40.54%	33.11%	31.33%	29.19%
NO ONE	27.86%	30.74%	33.90%	26.82%
DON'T KNOW				
REFUSED				
INAPP				
















Dataset	Variable	Valid	Invalid	Min	First Quartile	Median	Third Quartile	Max	Mean	StdDev
M2P1	B1PA45B	4,960	3	1				4		
MKE1	BACA45B	592	0	1				4		
MRP1	RA1PA45B	3,575	2	1				4		
MKER	RAACA45B	507	1	1				4		

Custom Data Extract



- Researchers can focus on variables of interest
- Facilitate accurate merges across numerous datasets
- Ease data management burden

- Survey
- Daily Diary
- Cognitive
- Biomarker
- Neuroscience

		M1P1	M2P1	M3P1	MKE1	MKE2	MRP1	MKER
	Age had first cigarette	A1PA41	B1PA37	C1PA37	BACA36	CACA36	RA1PA37	RAACA36
	Ever smoked cigarettes regularly	A1PA40	B1PA38A	C1PA38A	BACA37	CACA37	RA1PA38A	RAACA37
	Age began to smoke regularly	A1PA42	B1PA38B	C1PA38B	BACA38	CACA38	RA1PA38B	RAACA38
	Now smoke cigarettes regularly	A1PA43	B1PA39	C1PA39	BACA39	CACA39	RA1PA39	RAACA39
	Cigarettes per day during heaviest year - Current smoker	A1PA44	B1PA40	C1PA40	BACA40	CACA40	RA1PA40	RAACA40
 	Ever tried to quit smoking	A1PA45	B1PA41	C1PA41	BACA41	CACA41	RA1PA41	RAACA41
	Age last smoked regularly	A1PA46	B1PA42	C1PA42	BACA42	CACA42	RA1PA42	RAACA42
	Cigarettes per day during heaviest year - Past smoker	A1PA47	B1PA43	C1PA43	BACA43	CACA43	RA1PA43	RAACA43
 	Ever used pipe, cigars, snuff, chew		B1PA44	C1PA44	BACA44	CACA44	RA1PA44	RAACA44
 	Lived with smoker growing up - Father		B1PA45A		BACA45A		RA1PA45A	RAACA45
 	Lived with smoker growing up - Mother		B1PA45B		BACA45B		RA1PA45B	RAACA45

ASIST DDI Lifecycle Demo

This is your active basket.

Variables (7)

[Download Data](#)[Download Metadata](#)

 	BACA36	Age had first cigarette
 	C1PA37	Age had first cigarette
 	RA1PA37	Age had first cigarette
 	RAACA36	Age had first cigarette
 	A1PA41	Age had first cigarette
 	CACA36	Age had first cigarette
 	B1PA37	Age had first cigarette

ASIST DDI Lifecycle Demo

Generated on Monday, April 19, 2021 11:37 AM

Title

ASIST DDI Lifecycle Demo

Data File

Variable Count	12
----------------	----

MIDUSID - Master ID created by concatenating M2ID, MRID, MIDJA_IDs

Type	Numeric (Integer)
Data File	MIDUSID_SampleID_20160525

Valid	Invalid	Minimum	Maximum
13225	0	10001	39999

SAMPLMAJ - Major sample identification (aka Sample)

Type	Code
Data File	MIDUSID_SampleID_20160525

Valid	1	2	3	4	13	14	20	21	30	Total	Frequency	% of total	% of valid
	MAIN RDD	SIBUNG	TWIN	QTY OVERSAMPLE	MILWAUKEE	BOSTON NEW	MIDUS REFRESHER	MILWAUKEE REFRESHER	MIDJA		3,487	26.4%	26.4%
											952	7.2%	7.2%
											1,996	15.1%	15.1%
											757	5.7%	5.7%
											592	4.5%	4.5%
											328	2.5%	2.5%
											3,577	27.0%	27.0%
											509	3.8%	3.8%
											1,027	7.8%	7.8%
											13,225	100.0%	100%

Valid	Invalid	Minimum	Maximum
13225	0	1	30

M2ID - MIDUS 2 ID number

Type	Numeric (Integer)
Data File	MIDUSID_SampleID_20160525

Valid	Invalid	Minimum	Maximum
-------	---------	---------	---------

8112	5113	10001	19193
------	------	-------	-------

MRID - MIDUS Refresher ID number

Type	Numeric (Integer)
Data File	MIDUSID_SampleID_20160525

Valid	Invalid	Minimum	Maximum
4086	9139	30000	39999

MIDJA_IDs - Midlife in Japan ID number

Type	Numeric (Integer)
Data File	MIDUSID_SampleID_20160525

Valid	Invalid	Minimum	Maximum
1027	12198	20008	29987

BACA36 - Age had first cigarette

Type	Code
Study	MIDUS 2 Milwaukee
Data File	MKE1
Variable Group	GAPI
Variable Group	Section 1: Health
Forward Skip	IF BACA36 = 96, GO TO BACA44 .
Interviewer Instructions	INTERVIEWER: IF R SAYS "I don't smoke", PROBE: "At what age did you have your very FIRST cigarette, if EVER?"
Question Text	At what age did you have your very first cigarette?
Pre-Question Text	The next questions are about smoking cigarettes.
ICPSR DOI for Project	http://dx.doi.org/10.3886/ICPSR22840

Valid	96	NEVER HAD A CIGARETTE	Frequency	% of total	% of valid
		Total	123	20.8%	21.0%
			587	99.2%	100%
Missing	97	DON'T KNOW	5	0.8%	
		Total	5	0.8%	

Valid	Invalid	Minimum	Maximum	Mean	StdDev
587	5	1	96	32.9	32.82

C1PA37 - Age had first cigarette

Type	Code
Study	MIDUS 3 Project 1
Data File	M3P1

DDI Cross-Domain Integration (DDI-CDI)

Arofan Gregory (Convenor)

DDI Modelling, Representation and Testing (MRT) Project Team

A New DDI Specification

- DDI-CDI is a new specification
 - Currently in revision after public review
 - Release in early 2022
- DDI-CDI is an implementation of the “DDI 4”/”DDI Moving Forward” model
 - Specific focus on cross-domain data integration
 - Model-based standard
 - XML and other syntax representations supported
 - Designed to be machine-actionable
- Complementary to other DDI specifications
 - Works with DDI Codebook and DDI Lifecycle
 - Extends metadata coverage to support integration with other domain data
 - Can work with other (non-DDI) domain metadata specifications

Why a New Specification?

- The volume of research data is increasing exponentially
 - New sources
 - New formats/structures
- The use of data across domain boundaries is increasing
 - “Grand challenges” (e.g., COVID-19, climate change)
 - New technologies and new approaches (e.g., AI, machine learning)
- Problems of scale demand machine-actionability
 - For metadata harvesting
 - For navigating data at all levels and across domain boundaries

DDI-CDI Functionality

- Structural description across diverse sources/types of data
 - Wide (rectangular) data
 - Tall (long), event data, sensors
 - Key-value data (big data, NoSQL data)
 - Multi-dimensional data cubes, indicators, time series
- Describes provenance of data between different structures/forms
 - Processing framework
 - Relies on other standards (PROV-O, SDTL, etc.)
- Describes data at an atomic level
 - Variables
 - Datums/cells

To learn more about DDI-CDI...

- Public Review documentation:
 - <https://ddialliance.org/announcement/public-review-ddi-cross-domain-integration-ddi-cdi>
 - Links to draft specification and documentation
 - Links to webinar recordings and other resources
- Report on EOSC and DDI-CDI
 - Available in May 2021
 - Lots of examples of specific uses
 - Shows connection to broader FAIR ecosystem

References

- Beuster, Benjamin, & Fry, Jane. (2020, November). What can DDI do for you? An introduction to the DDI. Presented at the 12th Annual European DDI User Conference (EDDI20), Zenodo. <http://doi.org/10.5281/zenodo.4298302>
- Orten, Hilde, Beuster, Benjamin, & Jääskeläinen, Taina. (2019, December). What can DDI do for you? An introduction to the DDI. Presented at the 11th Annual European DDI User Conference (EDDI19), Tampere, Finland: Zenodo. <http://doi.org/10.5281/zenodo.3597192>

Thank you!



secretariat@ddialliance.org