Who is our audience- will they already know about data. What is the vocabulary of DDI is using that the audience may know already (but not necessarily as developers use variable in a different way).

Explanatory text can include why and who collected the data.-

Add notes to slides for trainer

Do we need an intro slide?

Why are variables important- they are the fundamental unit of analysis.

These are the things we want to do with the data (benefits)….

**What is a variable in DDI?**

The whole purpose is to tie the variables together which are measuring a comparable topic.

Should we list what other domain use to describe a variable e.g. column in excel, relational database, attribute in object-oriented programing etc.

* Start with a table of data- a variable is a column of data.
* When people are approaching DDI they think of data as records of data
* Are we missing info unless we explain why we have a rectangular file? How do you say that each row is related to the same column? Are we making an assumption– show an example and explain that we asked each people there marital status each row is a person.
* We can visualise this as a table
* Variable = instance variable?
* Discuss as item type level in relation to other items rather than details about each.
* Repeated measures in a column (not associated these measures with anything yet)
* How do we describe the content of the variable? What is the data type? (this may be too much detail at first- but will be needed for the cascade. Four examples of different columns of data- text, numeric, code lists, date.
* We should introduce the value domain when introducing the variable. E.g. variables have different types of data and list them.
* Leave rows till later on in the slides.
* Code list are related to variable items. E.g. 1= Married, 2 = Single -in DDI the categories are also separate items (this may be too much detail).
* Variables point to a code list with the enumerated value.
* Or it could be a number, or text- it is how we describe the values in the set of data.
* What is the data in this column look like? E.g. characters or numbers (two kinds- qual and quant)

**What is variable cascade?**

* We explain how representative variable works with other data types.- start with the code list and the give examples of quantitative data types
  + e.g. marital status
  + e.g. how far did you walk today? In different units miles and km or height.
  + e.g. birth date – date format.
* Three obviously different types- codes, numbers and dates.
* Represented variable means that they are not directly comparable but can be made to be comparable.
* Show example of three different sets of data
* We measure martial status several times e.g. over time. Three Different (distinct) sets of data with a different variables all measuring marriage. Two datasets sharing same code list, one with different code list. E.g. 1= Married, 2=Single, 3=divorced.
* They are related in some way, but how do we relate these together- the variable cascade helps us do this.
* We should start with easy terms e.g. examples of variable names and then can switch to DDI terms later.
* All three variables are trying to measure the same thing which marriage which is a conceptual variable. We don’t tell you how. Each of the variables in the datasets is an actual measurement of marital status.
* How we actually did a measurement is the representation. If they are measured in the same way then they both point to the same representative variable.
* Represented variables link variables which are directly comparable.
* Value domain = Variable representation in DDI
* Same representation of the value domain. It doesn’t care about the question asked- only the value domain.
* The represented variable is a way to measure marriage and so links to the marriage conceptual variable.
* If the value domain changes, then it adds in another layer.
* The third dataset has a different code list (i.e. value domain) and so it needs a new representative variable which then links to the conceptual variable.
* Concept variable defines a measure without defining the measurement.

**Benefits**

Comparability and provenance.

* Lineage of the data/provenance
* Identify comparable data
* Compare how the response values have changed over time or over different sources
* Used to manage to change and comparability
* Supports people working on harmonisation (gives you the structure to describe the variables and decide whether you can harmonise them or not)

**Useful resources**

Daniel Gillman EDDI 2017: The DDI-4 Variable Cascade and Datum-Centered Approach

Colectica docs site - <https://docs.colectica.com/designer/manage-content/data/harmonize-variables/>

**Questions/answers**

What level do we publish? – it depends on how you want to display the datasets.

All the variables could have the same variable name, but in DDI they will have a different identifier. It is up to you which you see.

We should always have the conceptual and represented in place, even if there is no change in the code value.

DDI separates the conceptual variables and representative variable is so that someone can query the relationships without having to look at every variable and value domain which is time consuming and also not reliant on a machine- as it is shows that person has decided that these are comparable.

Should there always be a new variable for each data collection- can they be the same variable? It depends on how you describe the dataset- there is lots more info about a dataset then just its type e.g. who it’s measured about e.g. universe. If you have different datasets the people/household changes. If you have new dataset then it is usually a new instance variable. Specific to a datasets as the it contains other metadata e.g. time of collection. If you publish data every quarter the variable instance will be different each time.

If there is only one dataset- whey would you use represented and conceptual variables- If you want people to re-use it or to use with other study’s data then it is valuable to including the represented and conceptual variable.

Ideal world- create a concept then create a question for that. Which is linked to the represented and conceptual variable. – questions to be discussed elsewhere.