Authors: Neil Kaye, Hayley Mills and Jon Johnson

2 Structured metadata

2.1 Structured metadata

Structured metadata defines the relationship between data items to enable computer systems to understand the contextual meaning of the data – to display the relevant information on a website, for instance.

Structured metadata tells a computer what something is, how it relates to other objects and what to do with it. By standardising the content and structure, it makes it easier for computers to automatically extract information from the metadata.

This information can then be provided to researchers to help them discover and access data from many different sources. It facilitates data sharing and allows data collected in one study to be re-used in the future by other researchers.

2.1.1 How structured metadata helps: an example

The example below shows a dataset of some variables which we might guess as being related to apples and oranges, but without additional metadata like measurement units we are not sure what exactly the variables relate to and how we can interpret and compare the data.

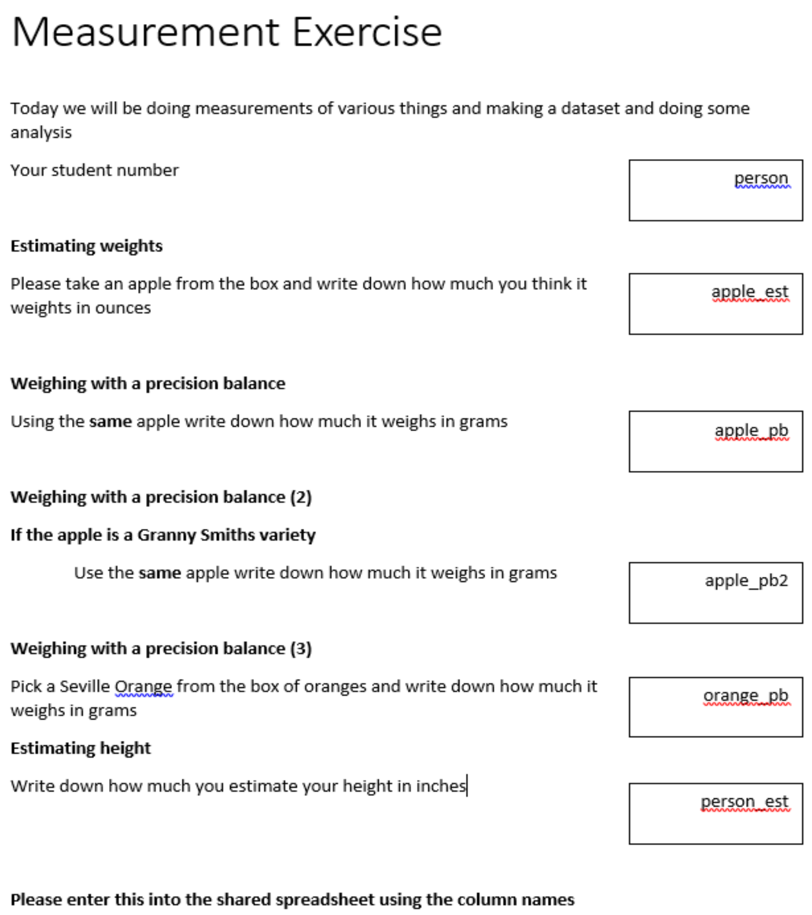
We might assume that we can compare all apples together, and that we don't want to compare apples with oranges.

person	apple_est	apple_pb	apple_pb2	orange_pb	person_est
1	6	170	165		140
2	7	200		250	180
3	18	250	270		370
4	5	125		190	100
5	6	115	140		275

Suggested citation: Kaye, N., Mills, H. & Johnson, J. (2020). *Understanding metadata*. CLOSER Learning Hub, London, UK: CLOSER

6	5	180	170	190	300

We can use the accompanying documentation to gather some more information. This tells us the data came from a student exercise and using this we now know the units of each variable and how the measurements were made.



Measurement Exercise

Today we will be doing measurements of various things and making a dataset and doing some analysis

Your student number — person

Estimating weights

Please take an apple from the box and write down how much you think it weights in ounces — apple_est

Weighing with a precision balance

Using the **same** apple write down how much it weighs in grams — apple_pb

Weighing with a precision balance (2)

If the apple is a Granny Smiths variety

Use the **same** apple write down how much it weighs in grams — apple_pb2

Weighing with a precision balance (3)

Pick a Seville Orange from the box of oranges and write down how much it weighs in grams — orange_pb

Estimating height

Write down how much you estimate your height in inches — person_est

Please enter this into the shared spreadsheet using the column names

This might be all the information we need and we can use it as it is, in a document form to determine what variables are useful to compare.

However, we could add some structure to this metadata to allow us to systematically assess the variables for similarities and differences.

Suggested citation: Kaye, N., Mills, H. & Johnson, J. (2020). *Understanding metadata*. CLOSER Learning Hub, London, UK: CLOSER

Label	Weight of apple (oz)	Weight of apple (g)	Weight of Granny Smith (g)	Weight of Seville (g)	Height of Person (inches)
Names	apple_est	apple_pb	apple_pb2	orange_pb	person_est
Concept	weight	weight	weight	weight	height
Unit type	apples	apples	apples	oranges	person
Method	Estimated	Precision balance	Precision balance	Precision balance	Estimated
Unit	Ounces	Grams	Grams	Grams	Inches
Population	All apples	All apples	Granny Smith apples	Seville oranges	All students

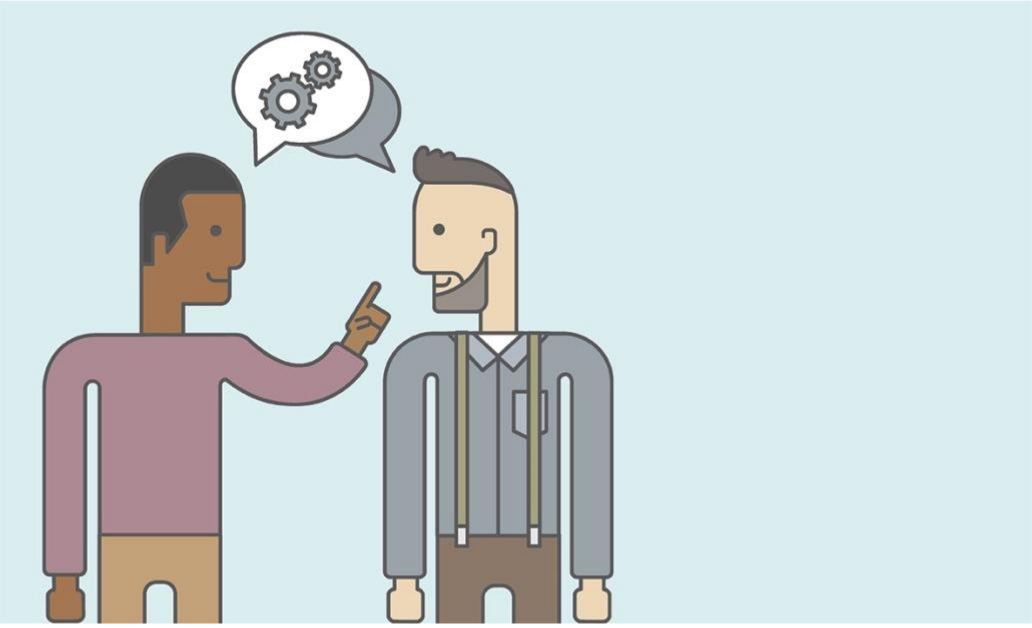
In the table above we have given each type of metadata in the document a label, for example Unit which is the measurement unit (e.g. ounces), or concept which is what we are measuring (e.g. weight). Doing this allows us to see clearly which variables are comparable or not, or which may need further transformation (e.g. converting from ounces into grams).

For example, if we want to compare the mean weights of two types of fruit, we might want to compare apples and oranges. The table will help us decide which apple variable to use. Comparing the mean height of a person vs mean weight of a fruit might not make sense as they are not comparable Unit *Types*, but they both have the same Method of measurement (Estimated) so you might want to look at how good students are at estimating small objects, such as an apple, compared to large objects such as themselves.

Documenting the dimensions of the data in the form of structured metadata is helpful for a human to understand the variables, but it is vital for the metadata to be read by a computer which is not possible from the word document.

2.2 Metadata standards

A metadata standard provides a framework to establish a common set of definitions for various characteristics or attributes of data. Standardising metadata, including language, spelling, format, variable coding, etc., allows different datasets to ‘speak’ to each other. If everyone uses a different standard, it can be very difficult to compare data from different sources. Because there is not one overall standard, it is necessary for ‘translation’ programmes to map between standards. In this way systems can read datasets using different standards.



Metadata standards allow:

- re-use of data
- data discovery
- data access
- interoperability of systems – that is, systems and machines can talk to each other and know they are referring to the same thing
- sharing of metadata between communities (e.g. data providers and data users)

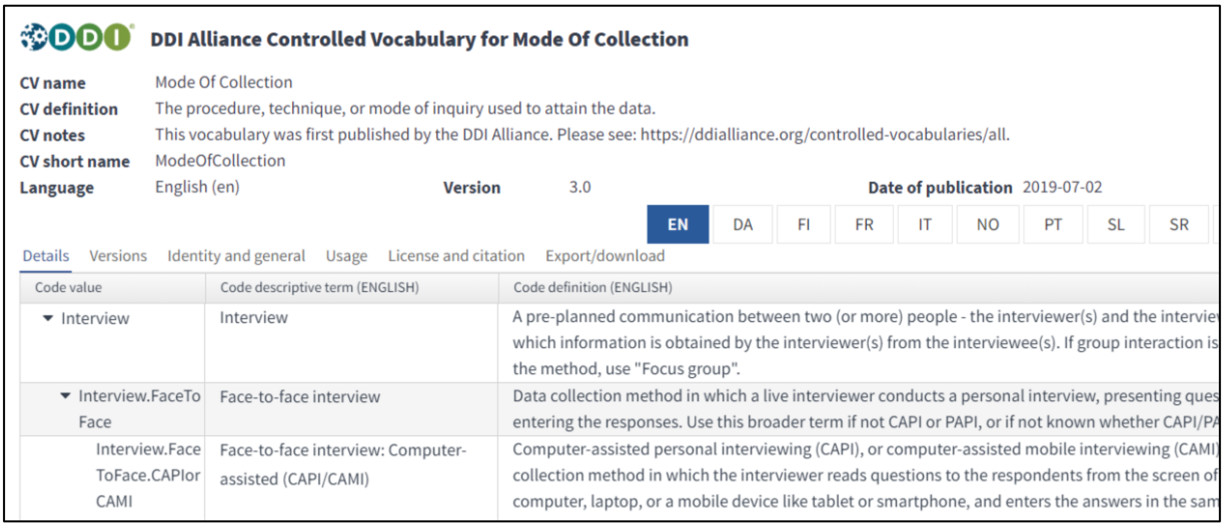
Suggested citation: Kaye, N., Mills, H. & Johnson, J. (2020). *Understanding metadata*. CLOSER Learning Hub, London, UK: CLOSER

2.3 Controlled Vocabularies

Using common vocabularies is a powerful way of describing related items that assists in data discovery. They are often referred to as *Controlled Vocabularies*, which are maintained within a community to describe commonly-used terms within that discipline. They will most often consist of a name, description and a definition. The description and definition will also in, some cases, be available in multiple languages so that a consistent way of describing something in maintained across different countries.

Having a standardised list facilitates the discovery of relevant data. In the example below, standardising 'face-to-face interview' as a controlled vocabulary term means that researchers do not have to include alternative terms – e.g. 'in-person interviews', 'personal interviews', 'at-home interviews', etc. – in their search.

E.g. DDI Alliance Controlled Vocabulary for Mode of Collection:



DDI Alliance Controlled Vocabulary for Mode Of Collection

CV name	Mode Of Collection
CV definition	The procedure, technique, or mode of inquiry used to attain the data.
CV notes	This vocabulary was first published by the DDI Alliance. Please see: https://ddialliance.org/controlled-vocabularies/all.
CV short name	ModeOfCollection
Language	English (en)
Version	3.0
Date of publication	2019-07-02

Code value	Code descriptive term (ENGLISH)	Code definition (ENGLISH)
▾ Interview	Interview	A pre-planned communication between two (or more) people - the interviewer(s) and the intervie which information is obtained by the interviewer(s) from the interviewee(s). If group interaction is the method, use "Focus group".
▾ Interview.FaceToFace	Face-to-face interview	Data collection method in which a live interviewer conducts a personal interview, presenting ques entering the responses. Use this broader term if not CAPI or PAPI, or if not known whether CAPI/PA
Interview.FaceToFace.CAPIorCAMI	Face-to-face interview: Computer-assisted (CAPI/CAMI)	Computer-assisted personal interviewing (CAPI), or computer-assisted mobile interviewing (CAMI) collection method in which the interviewer reads questions to the respondents from the screen of computer, laptop, or a mobile device like tablet or smartphone, and enters the answers in the sam

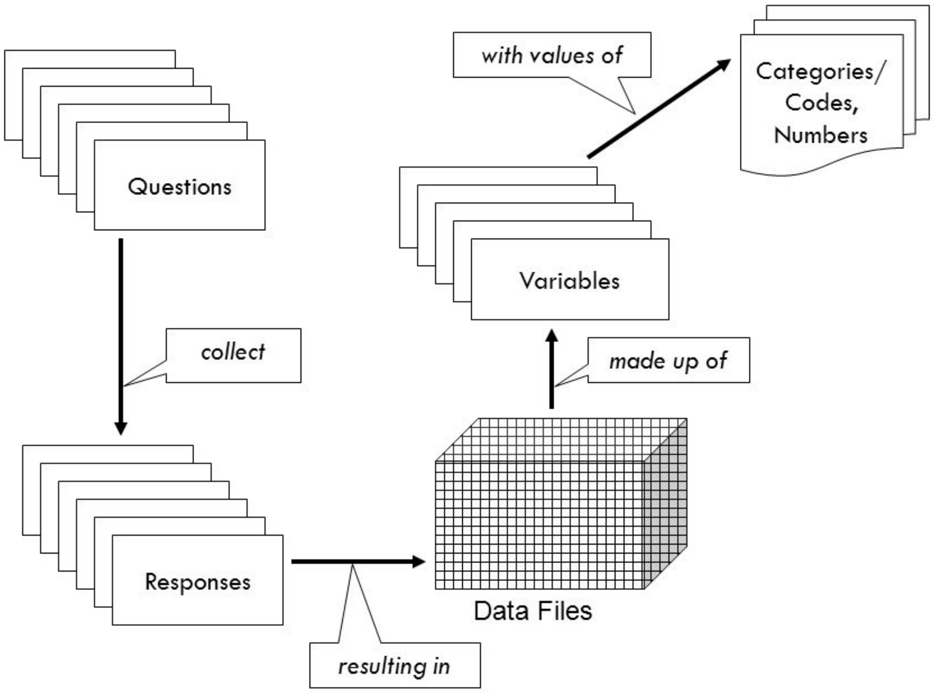
Source: *CESSDA, 2020, DDI Alliance Controlled Vocabularies*

Suggested citation: Kaye, N., Mills, H. & Johnson, J. (2020). *Understanding metadata*. CLOSER Learning Hub, London, UK: CLOSER

2.4 Variables, questions and measurements: how metadata helps to make sense of data

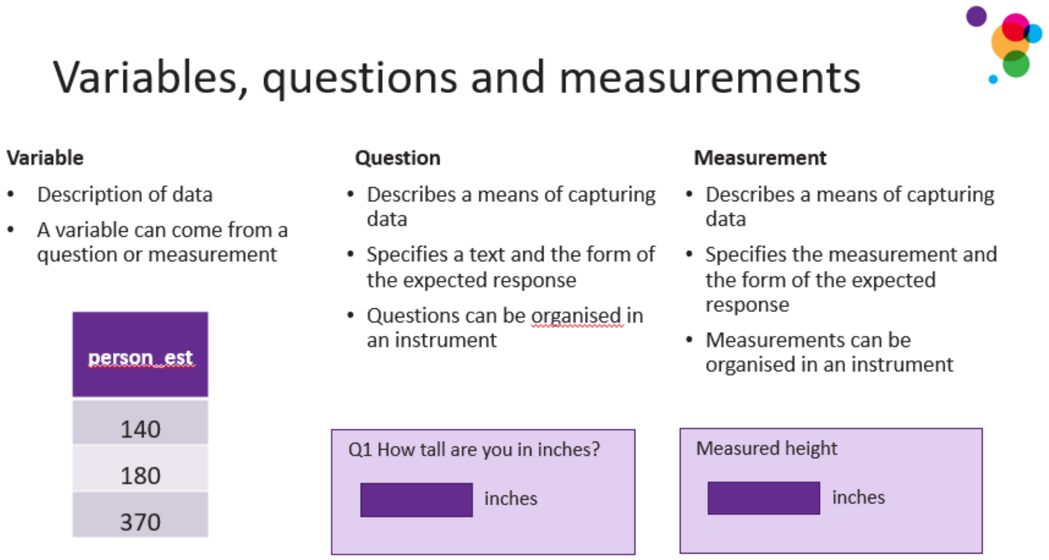
Metadata is collected at every stage of the research life cycle, from pre data collection to analysis and publication.

If we look at the creation of data from a survey instrument or questionnaire we can split this into different metadata elements starting with the questions and resulting in the variables. The figure below shows how the questions are used to collect responses, which results in data that is made up of variables containing values of numbers.



What is the difference between a variable, a question and a measurement?

Whilst variables refer to any data item that describes an attribute or characteristic of an object, questions and measurements refer to two different means of capturing these data items: a question provides text and a prescribed way to respond to the text; a measurement specifies what characteristic or element of a thing is to be measured, how and in what units this should be taken.



Variables, questions and measurements

Variable

- Description of data
- A variable can come from a question or measurement

person_est
140
180
370

Question

- Describes a means of capturing data
- Specifies a text and the form of the expected response
- Questions can be organised in an instrument

Q1 How tall are you in inches?

inches

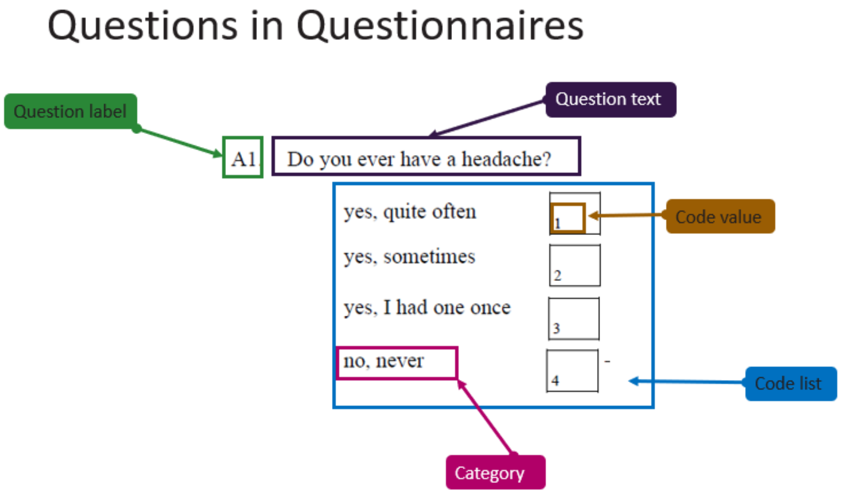
Measurement

- Describes a means of capturing data
- Specifies the measurement and the form of the expected response
- Measurements can be organised in an instrument

Measured height

inches

A question is formed of more than just the question text, but can be broken down into different elements; the question name or label, the question text, how the participant responds, and any instructions on how to answer the question.



Questions in Questionnaires

A1 Do you ever have a headache?

yes, quite often 1

yes, sometimes 2

yes, I had one once 3

no, never 4

In addition to questions, questionnaires contain other elements to help the participant navigate through the questionnaire including accompanying text or statements and routing (i.e. when to answer or skip a question). Each of these elements is a piece of metadata which helps us to understand how the data were collected.

2.5 Common standardised classifications

In the UK, and internationally, statistical authorities – e.g. Office for National Statistics; UK Statistical Authority; Eurostat; International Labour Organisation – have developed a number of standardised classifications with the aim of assisting data collection, presentation of statistics and evaluating policy effects.

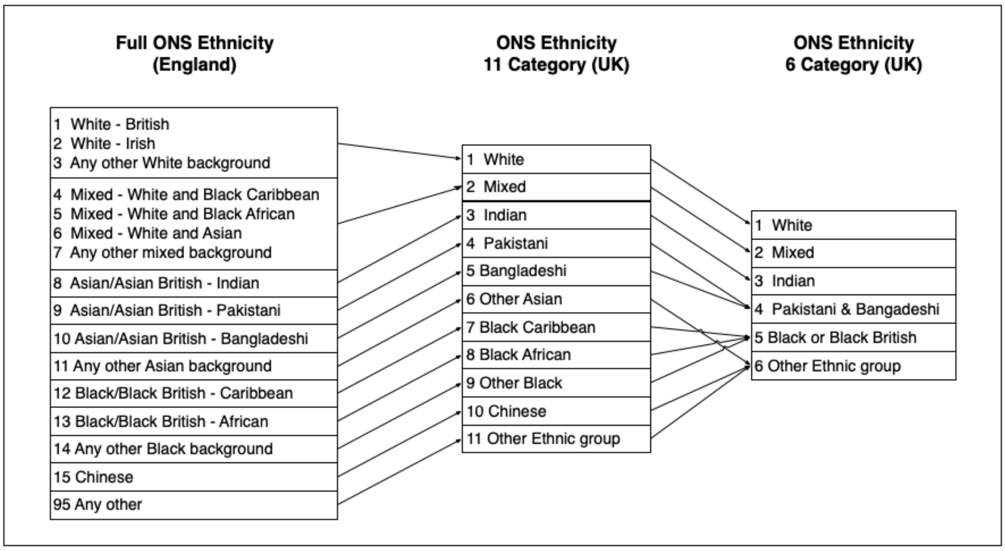
These classifications can assist in the collection of subjective and/or complex information in relation to an individual's identity (e.g. ethnic group) or economic activity (e.g. occupational classification), or to assist in the comparability of measures across different contexts (e.g. educational attainment in different national systems).

Academic researchers commonly use these centrally-standardised classifications when designing a study. This ensures that findings are consistent with the policy discourse and official statistical outputs. It also allows for comparability with other research (including previous and future studies) using the same standardised codes.

2.5.1 Ethnic group

Ethnic group membership is highly subjective, multifaceted and dynamic, which makes the collection of data on ethnicity complex. Whilst there is no consensus on what constitutes an ethnic group, the Office for National Statistics has constructed a standard classification based on categories according to groups with 'shared characteristics'.

There are 18 categories in England (including one for 'other'), which can be collapsed into 11 categories for UK-wide comparability, or into 6 broader 'top-level' categories:


Full ONS Ethnicity (England)
ONS Ethnicity 11 Category (UK)
ONS Ethnicity 6 Category (UK)
1 White - British
2 White - Irish
3 Any other White background
4 Mixed - White and Black Caribbean
5 Mixed - White and Black African
6 Mixed - White and Asian
7 Any other mixed background
8 Asian/Asian British - Indian
9 Asian/Asian British - Pakistani
10 Asian/Asian British - Bangladeshi
11 Any other Asian background
12 Black/Black British - Caribbean
13 Black/Black British - African
14 Any other Black background
15 Chinese
95 Any other
1 White
2 Mixed
3 Indian
4 Pakistani
5 Bangladeshi
6 Other Asian
7 Black Caribbean
8 Black African
9 Other Black
10 Chinese
11 Other Ethnic group
1 White
2 Mixed
3 Indian
4 Pakistani & Bangadeshi
5 Black or Black British
6 Other Ethnic group

2.5.2 Educational attainment

The UNESCO Institute for Statistics (UIS) has developed the International Standard Classification of Education (ISCED) to serve as an instrument to compile and present statistics both nationally and internationally.

This maps national educational qualifications onto internationally-comparable ISCED levels, which take into account the level (e.g. primary, lower secondary, etc.), orientation (e.g. academic, vocational, etc.) and type (e.g. access to higher education, etc.) of programme undertaken by students.

The following table provides ISCED categories for the main UK qualifications, allowing for international comparison of educational attainment:

Level		Category (orientation)	Sub-category (type)	UK qualification
Pre-primary	0	01	010	
		02	020	
Primary	1	10	100	
Lower secondary	2	24 General	243	Key Skills
Upper secondary	3	34 General	342	GCSE / Scottish Standard or Intermediate
			343	AS level / Scottish Higher
			344	A level / Scottish Advanced Higher
		35 Vocational	352	NVQ level 2
			354	NVQ level 3
Post-secondary	4	44 General	-	
		45 Vocational	-	
Short-cycle tertiary	5	55 Vocational	551	NVQ level 4, HNC
			554	Foundation degree, NVQ level 5, HND
Bachelor's or equivalent	6	66 Orientation unspecified	665 First degree	Bachelor's degree
Master's or equivalent	7	76 Orientation unspecified	767 Further degree	Master's degree
Doctor or equivalent	8	86 Orientation unspecified	864	PhD