Tutorial on DDI controlled vocabularies

Taina Jääskeläinen, FSD Finnish Social Science Data Archive
Anne Etheridge, UK Data Service

EDDI 2017, Lausanne

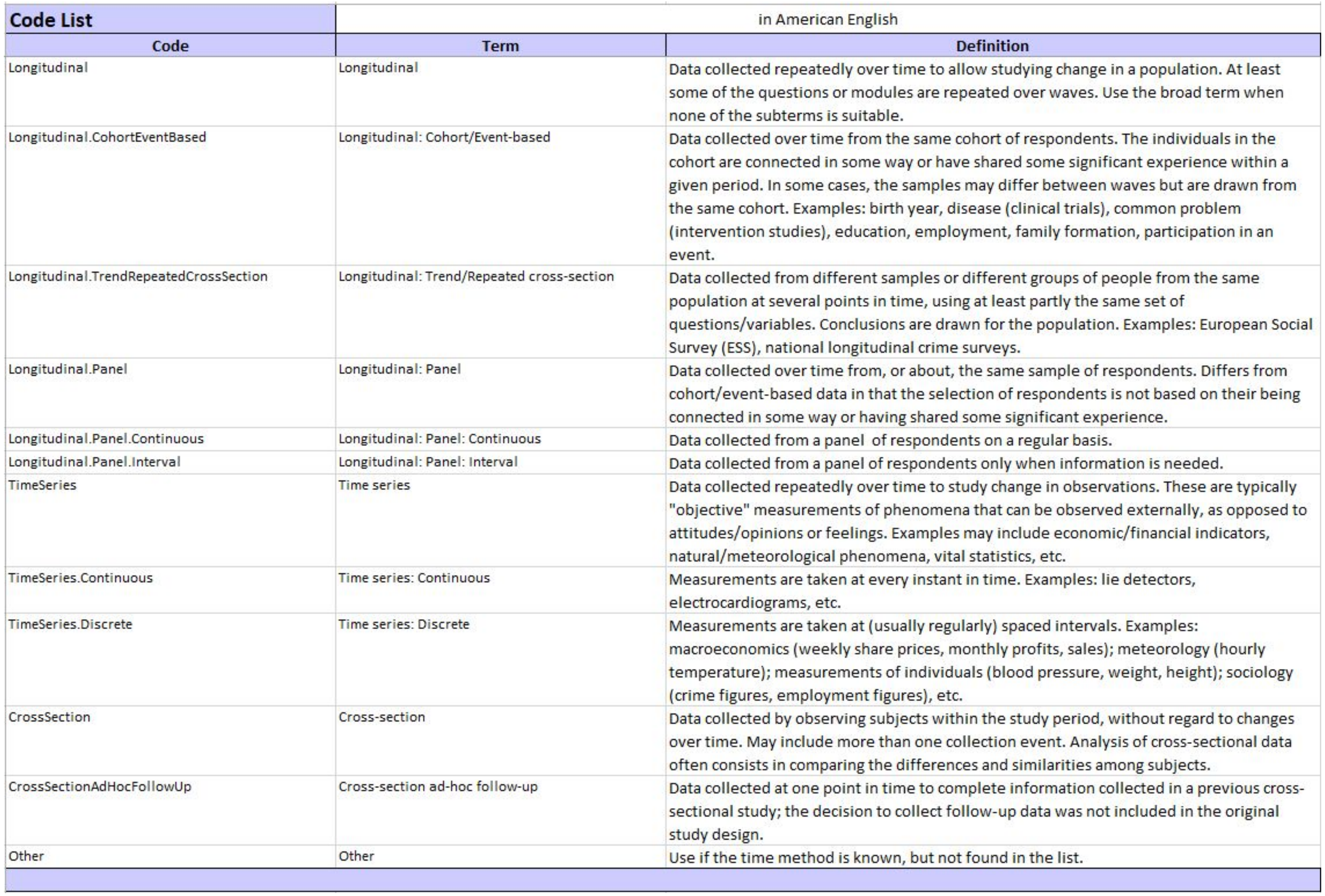
What is a controlled vocabulary?

- a list of terms that have been enumerated explicitly
- list is controlled by, and is available from, a controlled vocabulary registration authority
- terms exclusive and not overlapping
- terms should have an unambiguous, non-redundant definition

Why use a controlled vocabulary?

- enhances content consistency, predictability, and comparability of data
- supports machine-actionability
- increases precision of searches and retrieval
- promotes semantic and technical interoperability
- facilitates metadata creation, organising, and sharing

Example DDI CV - Time method



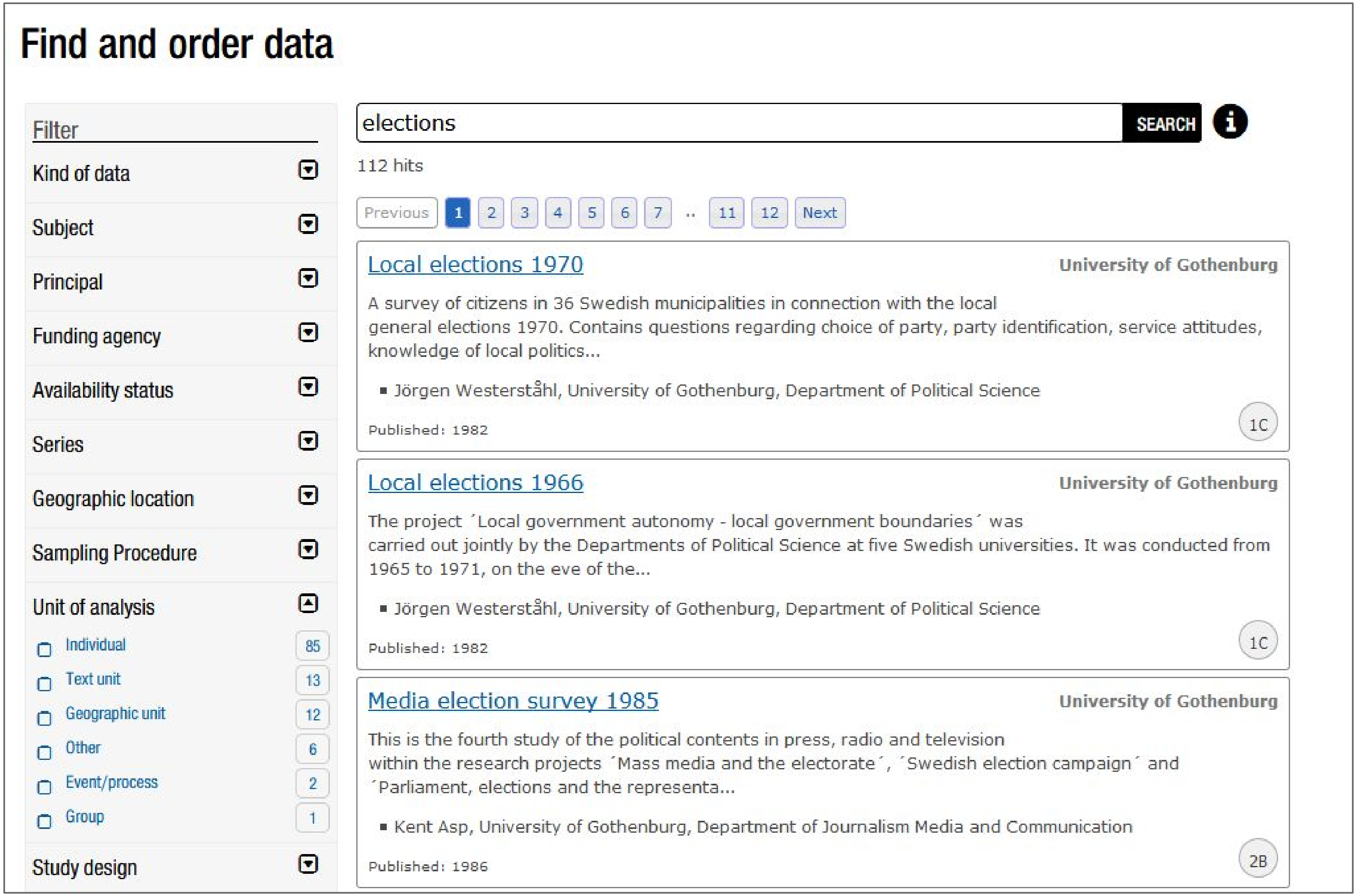
Code List	in American English	
Code	**Term**	**Definition**
Longitudinal	Longitudinal	Data collected repeatedly over time to allow studying change in a population. At least some of the questions or modules are repeated over waves. Use the broad term when none of the subterms is suitable.
Longitudinal.CohortEventBased	Longitudinal: Cohort/Event-based	Data collected over time from the same cohort of respondents. The individuals in the cohort are connected in some way or have shared some significant experience within a given period. In some cases, the samples may differ between waves but are drawn from the same cohort. Examples: birth year, disease (clinical trials), common problem (intervention studies), education, employment, family formation, participation in an event.
Longitudinal.TrendRepeatedCrossSection	Longitudinal: Trend/Repeated cross-section	Data collected from different samples or different groups of people from the same population at several points in time, using at least partly the same set of questions/variables. Conclusions are drawn for the population. Examples: European Social Survey (ESS), national longitudinal crime surveys.
Longitudinal.Panel	Longitudinal: Panel	Data collected over time from, or about, the same sample of respondents. Differs from cohort/event-based data in that the selection of respondents is not based on their being connected in some way or having shared some significant experience.
Longitudinal.Panel.Continuous	Longitudinal: Panel: Continuous	Data collected from a panel of respondents on a regular basis.
Longitudinal.Panel.Interval	Longitudinal: Panel: Interval	Data collected from a panel of respondents only when information is needed.
TimeSeries	Time series	Data collected repeatedly over time to study change in observations. These are typically "objective" measurements of phenomena that can be observed externally, as opposed to attitudes/opinions or feelings. Examples may include economic/financial indicators, natural/meteorological phenomena, vital statistics, etc.
TimeSeries.Continuous	Time series: Continuous	Measurements are taken at every instant in time. Examples: lie detectors, electrocardiograms, etc.
TimeSeries.Discrete	Time series: Discrete	Measurements are taken at (usually regularly) spaced intervals. Examples: macroeconomics (weekly share prices, monthly profits, sales); meteorology (hourly temperature); measurements of individuals (blood pressure, weight, height); sociology (crime figures, employment figures), etc.
CrossSection	Cross-section	Data collected by observing subjects within the study period, without regard to changes over time. May include more than one collection event. Analysis of cross-sectional data often consists in comparing the differences and similarities among subjects.
CrossSectionAdHocFollowUp	Cross-section ad-hoc follow-up	Data collected at one point in time to complete information collected in a previous cross-sectional study; the decision to collect follow-up data was not included in the original study design.
Other	Other	Use if the time method is known, but not found in the list.

Code:
Value of the code - the ID across languages, to be used in metadata for interoperability. Machine actionable. Good results if organisations have an Editor with drop-down lists for CVs - no spelling mistakes.

Term:
Descriptive, human-readable term, to be shown in published metadata

Definition:
Concise description of the meaning of a term within the context of the vocabulary. Should explain the difference between terms and have examples.

Example: Analysis unit CV terms used for a filter in a search interface (SND)



Find and order data

Filter

- Kind of data
- Subject
- Principal
- Funding agency
- Availability status
- Series
- Geographic location
- Sampling Procedure
- Unit of analysis

□ Individual	85
□ Text unit	13
□ Geographic unit	12
□ Other	6
□ Event/process	2
□ Group	1

- Study design

elections SEARCH

112 hits

Previous 1 2 3 4 5 6 7 .. 11 12 Next

Local elections 1970 — University of Gothenburg

A survey of citizens in 36 Swedish municipalities in connection with the local general elections 1970. Contains questions regarding choice of party, party identification, service attitudes, knowledge of local politics...

- Jörgen Westerståhl, University of Gothenburg, Department of Political Science

Published: 1982 — 1C

Local elections 1966 — University of Gothenburg

The project ´Local government autonomy - local government boundaries´ was carried out jointly by the Departments of Political Science at five Swedish universities. It was conducted from 1965 to 1971, on the eve of the...

- Jörgen Westerståhl, University of Gothenburg, Department of Political Science

Published: 1982 — 1C

Media election survey 1985 — University of Gothenburg

This is the fourth study of the political contents in press, radio and television within the research projects ´Mass media and the electorate´, ´Swedish election campaign´ and ´Parliament, elections and the representa...

- Kent Asp, University of Gothenburg, Department of Journalism Media and Communication

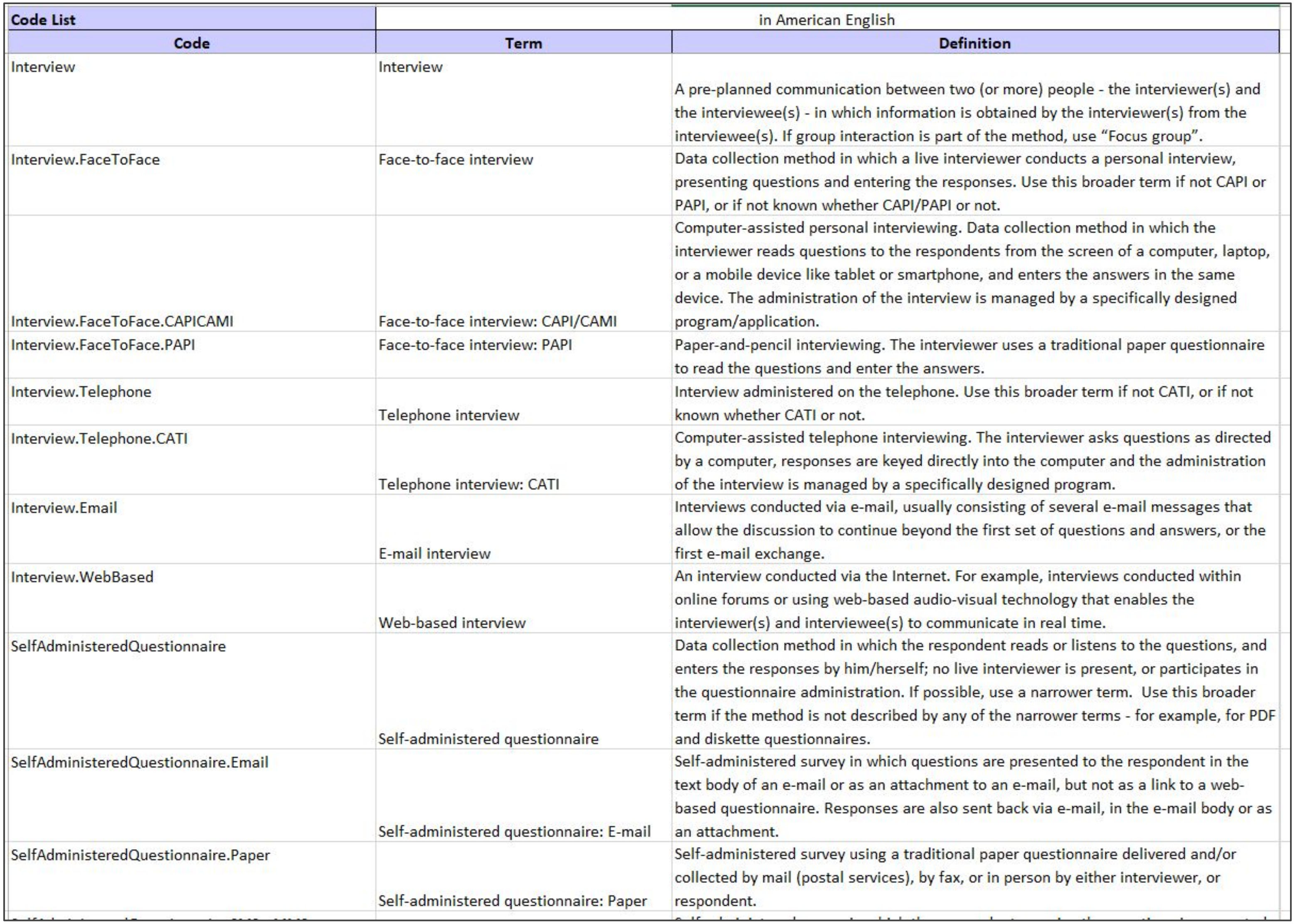
Published: 1986 — 2B

DDI CVs

- currently (December 2017) 23 DDI CVs
- continually under review by the DDI Controlled Vocabularies Group (DDI CVG)
- CVG currently creating a Contributor role CV
- values of CVs irrespective of DDI version
- many CVs are, or will be used by CESSDA, and are included in the CESSDA metadata model

Using CVs

Example DDI CV - mode of collection



Code List		in American English
Code	**Term**	**Definition**
Interview	Interview	A pre-planned communication between two (or more) people - the interviewer(s) and the interviewee(s) - in which information is obtained by the interviewer(s) from the interviewee(s). If group interaction is part of the method, use "Focus group".
Interview.FaceToFace	Face-to-face interview	Data collection method in which a live interviewer conducts a personal interview, presenting questions and entering the responses. Use this broader term if not CAPI or PAPI, or if not known whether CAPI/PAPI or not.
Interview.FaceToFace.CAPICAMI	Face-to-face interview: CAPI/CAMI	Computer-assisted personal interviewing. Data collection method in which the interviewer reads questions to the respondents from the screen of a computer, laptop, or a mobile device like tablet or smartphone, and enters the answers in the same device. The administration of the interview is managed by a specifically designed program/application.
Interview.FaceToFace.PAPI	Face-to-face interview: PAPI	Paper-and-pencil interviewing. The interviewer uses a traditional paper questionnaire to read the questions and enter the answers.
Interview.Telephone	Telephone interview	Interview administered on the telephone. Use this broader term if not CATI, or if not known whether CATI or not.
Interview.Telephone.CATI	Telephone interview: CATI	Computer-assisted telephone interviewing. The interviewer asks questions as directed by a computer, responses are keyed directly into the computer and the administration of the interview is managed by a specifically designed program.
Interview.Email	E-mail interview	Interviews conducted via e-mail, usually consisting of several e-mail messages that allow the discussion to continue beyond the first set of questions and answers, or the first e-mail exchange.
Interview.WebBased	Web-based interview	An interview conducted via the Internet. For example, interviews conducted within online forums or using web-based audio-visual technology that enables the interviewer(s) and interviewee(s) to communicate in real time.
SelfAdministeredQuestionnaire	Self-administered questionnaire	Data collection method in which the respondent reads or listens to the questions, and enters the responses by him/herself; no live interviewer is present, or participates in the questionnaire administration. If possible, use a narrower term. Use this broader term if the method is not described by any of the narrower terms - for example, for PDF and diskette questionnaires.
SelfAdministeredQuestionnaire.Email	Self-administered questionnaire: E-mail	Self-administered survey in which questions are presented to the respondent in the text body of an e-mail or as an attachment to an e-mail, but not as a link to a web-based questionnaire. Responses are also sent back via e-mail, in the e-mail body or as an attachment.
SelfAdministeredQuestionnaire.Paper	Self-administered questionnaire: Paper	Self-administered survey using a traditional paper questionnaire delivered and/or collected by mail (postal services), by fax, or in person by either interviewer, or respondent.

Using the CVs in metadata

- use the most specific term possible
- if unclear which narrower term is appropriate, use the broader term
- if organisation decides to use only broader terms, data may be excluded in filtered search results for systems that use narrower terms
- interoperability best if the value of the code is entered for the term; for the CV itself, the minimum is agency, machine-actionable CV ID, and version

Multilingual CVs

Example is part of the DDI Time method CV with terms and definitions translated into German by GESIS

	Definition	GESIS (German)	GESIS (German) definition
TIME METHOD	DDI: Describes the time dimension of the data collection.	Erhebungsdesign	Beschreibt die zeitliche Dimension der Datenerhebung.
Longitudinal	Data collected repeatedly over time to allow studying change in a population. At least some of the questions or modules are repeated over waves. Use the broad term when none of the subterms is suitable.	Längsschnitt	Daten, die an mehreren Zeitpunkten erhoben wurden, um Veränderungen innerhalb einer Population zu untersuchen. Mindestens einige der Fragen oder Module werden dabei in Wellen wiederholt abgefragt. Verwenden Sie diesen breit gefassten Begriff, wenn sich keiner der Unterterme eignet.
Longitudinal: Cohort/Event-based	Data collected over time from the same cohort of respondents. The individuals in the cohort are connected in some way or have shared some significant experience within a given period. In some cases, the samples may differ between waves but are drawn from the same cohort. Examples: birth year, disease (clinical trials), common problem (intervention studies), education, employment, family formation, participation in an event.	Längsschnitt: Kohorte/Ereignisbasierte Daten	Daten, die an mehreren Zeitpunkten einer gleichbleibenden Kohorte erhoben wurden. Die Individuen der Kohorte sind in einer spezifischen Art und Weise miteinander verbunden oder teilen besondere Erfahrungen innerhalb eines bestimmten Zeitraums. In manchen Fällen können die Stichproben zwischen den Wellen variieren, werden jedoch aus derselben Kohorte gezogen. Beispiele: Geburtsjahr, Krankheit (klinische Studien), gemeinsames Problem (Interventionsstudien), Bildung, Beschäftigung, Familiengründung, Teilnahme an einer Veranstaltung.
Longitudinal: Trend/Repeated cross-section	Data collected from different samples or different groups of people from the same population at several points in time, using at least partly the same set of questions/variables. Conclusions are drawn for the population. Examples: European Social Survey (ESS), national longitudinal crime surveys.	Längsschnitt: Wiederholter Querschnitt	Daten, die an mehreren Zeitpunkten aus unterschiedlichen Stichproben oder Personengruppen einer gleichen Population erhoben wurden, wobei zumindest teilweise die gleichen Fragen oder Variablen verwendet wurden. Die Schlussfolgerungen beziehen sich auf die Gesamtpopulation. Beispiele: European Social Survey (ESS), nationale Längsschnittstudien über Kriminalität.
Longitudinal: Panel	Data collected over time from, or about, the same sample of respondents. Differs from cohort/event-based data in that the selection of respondents is not based on their being connected in some way or having shared some significant experience.	Längsschnitt: Panel	Daten, die an mehreren Zeitpunkten von einer gleichbleibenden Stichprobe erhoben wurden. Der Unterschied zu Kohorten - bzw. ereignisbezogenen Daten liegt darin, dass die Befragten keine Verbindung aufweisen oder bedeutende Erfahrungen teilen.

Multilingual DDI CVs

- multilingual CVs work across countries and organisations
- CESSDA has provided translations of some DDI CVs
    - mandatory CVs from CESSDA metadata model
    - will be published in 2018
    - translated initially into Danish, German, Finnish, Norwegian, and Slovenian
- structure and hierarchy of a CV is determined by the source/master CV, cannot be changed in translated version
- value of the code is the ID that remains the same across languages
- only the descriptive (human readable) term and definitions are translated

Translation tips for CVs

Translating terms:

- in published metadata, CV terms appear alone without definitions
- therefore need to be translated into terms that are understandable in themselves and used in the local language
- for example, the term ‘Summary’ was translated into Finnish as three different words ‘Yhteenveto, lyhennelmä tai tiivistelmä’ as there is no one word that cover the whole meaning of the English term
- note that translated terms are also used for data discovery (see the filter example in search interface in slide 5), which emphasises need for easy understandability
- it is also good to provide access to term definitions for search interface users

Translation tips for CVs

Translating definitions:

- as closely as possible
- if literal translation means arbitrary, hard-to-understand definition, concentrate on translating the meaning, even using some other words or expressions if understandability requires this. But take care to convey the same meaning!
- definitions that are not understandable are not usable

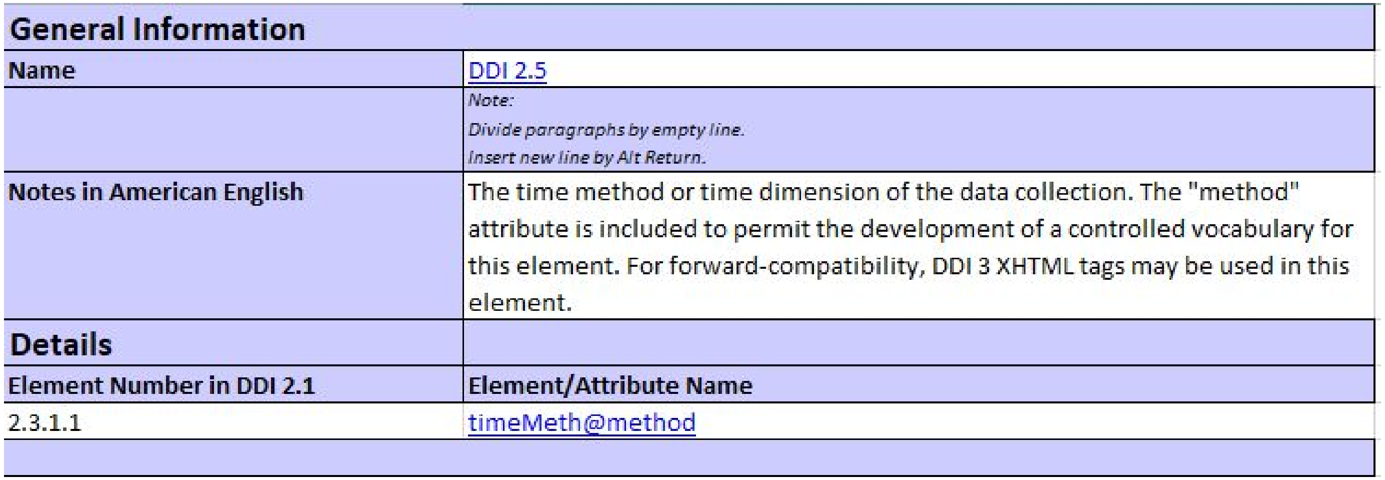
CV Manager tool

- ongoing project of CESSDA
- tool will be launched in autumn 2018
- will support both CESSDA and DDI CVs
- tool will manage CVs – create, edit, update
- will mark CVs ready to be published
- DDI CVs will be published both on DDI Alliance website and CV Manager User interface
- users have different roles and access rights

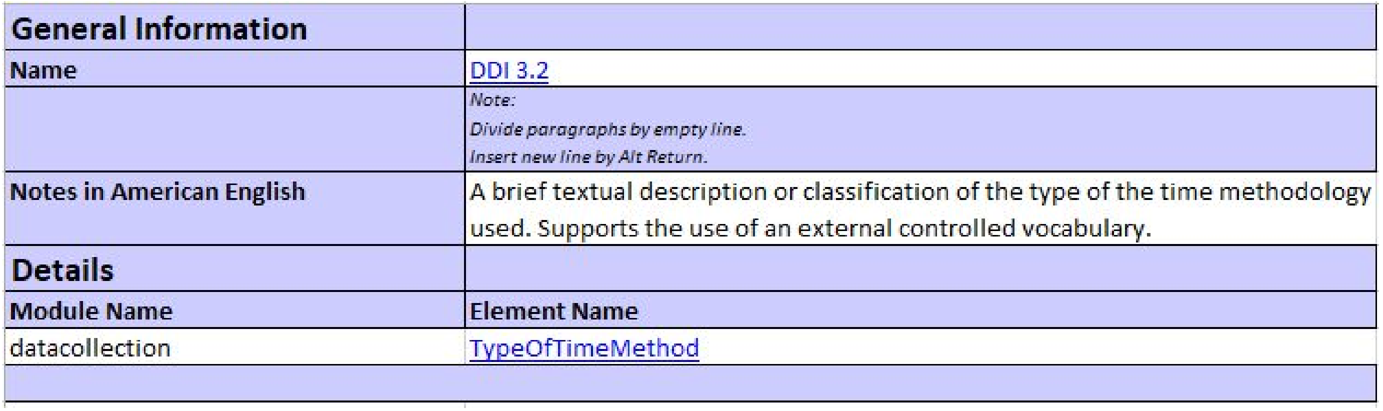
CV Manager tool

- online access
- persistent ids
- validation of content
- handles both source CVs and their language versions
- guaranteed continuity and access rights
- access to CV tool data for various purposes through exports or an API
- search interface

CV metadata: specifying the element for each DDI version



General Information	
Name	DDI 2.5
	Note: *Divide paragraphs by empty line.* *Insert new line by Alt Return.*
Notes in American English	The time method or time dimension of the data collection. The "method" attribute is included to permit the development of a controlled vocabulary for this element. For forward-compatibility, DDI 3 XHTML tags may be used in this element.
Details	
Element Number in DDI 2.1	**Element/Attribute Name**
2.3.1.1	timeMeth@method



General Information	
Name	DDI 3.2
	Note: *Divide paragraphs by empty line.* *Insert new line by Alt Return.*
Notes in American English	A brief textual description or classification of the type of the time methodology used. Supports the use of an external controlled vocabulary.
Details	
Module Name	**Element Name**
datacollection	TypeOfTimeMethod

DDI CVs in different DDI versions

- in metadata for a CV, there is always information in which element that particular CV can be used in different DDI versions
- items within the CV are valid for all versions of DDI
- DDI3: in elements that allow a CV item to be used, there are always:
  - a TypeOfxxx element for typology, i.e. for entering the value of the code
  - a free-text element for giving more detailed information

So, for Sampling Pocedure: the value of the code is entered into TypeOfSamplingProcedure whereas more detailed information on sample design and drawing the sample can be entered into Description element.

XML mark-up example in DDI 2.5 for a German CV term

```
<method>
  <dataColl>
    <timeMeth>
<concept xml:lang=”de” vocabURI="URI for urn:ddi-cv:TimeMethod:1.2"
vocab="TimeMethod">Longitudinal.Panel</concept>
    </timeMeth>
  </dataColl>
</method>


<!-- content of element "concept" is used for entering the value of the code of the controlled
vocabulary-->
<!-- In the published metadata in a German data catalogue, the German term Längesschnitt: Panel
would be shown instead of Longitudinal.Panel.-->
<!-- More information on the CV itself (version etc, long and short names etc) can be given in
<controlledVocabUsed> elements
```

XML mark-up example in DDI 3.2 for an English CV term

```
<r: TypeOfTimeMethod
codeListID=”TimeMethod” codeListName=”Time Method”
codelistAgencyName=”DDI Alliance” codeListVersionID=”1.2”
codeListURN=”urn:ddi-cv:TimeMethod:1.2”>Longitudinal.Panel
</r: TypeOfTimeMethod>



<!--Please note that in DDI 3.3  and DDI4,  all "CodeListXxx" attribute names
will be changed to "controlledVocabularyXxx", becoming therefore
"controlledVocabularyID", "controlledVocabularyAgencyName" etc.-->
```

DDI CVG

- members from several countries including USA, UK, Finland, Sweden, Australia, Germany and Norway

- Controlled Vocabularies Working Group (DDI CVG)

https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/39911435/Controlled+Vocabularies+Working+Group

- comments on vocabularies and more members are welcome
- webinars once a month, on a Tuesday at 13:00 UTC

Contact

Taina Jääskeläinen, FSD Finnish Social Science Data Archive
Taina.Jaaskelainen@staff.uta.fi

Anne Etheridge, UK Data Service
aether@essex.ac.uk