# **Getting Started with DDI**

You have identified DDI as a good standard for the task you want to accomplish, so that’s great. Now you need to move forward in using the standard. Below we provide some common use cases and tasks that DDI can support. Each task has a description, a set of related resources, and recommended tools to accomplish the task. If some of this language is unfamiliar to you, please see the [DDI Glossary](http://ddidev.web.itd.umich.edu/resources/ddi-glossary) for definitions.

## Create a Data Catalog for Discovery +/-

To create an online data catalog of study-level metadata records, you will need to “mark up” or tag those records using DDI. If you have existing records in some other form like Dublin Core or MARC, you can use the existing fields in those records as a foundation for the new DDI records. To do that, you need to map the fields to DDI. A [mapping](http://www.ddialliance.org/resources/tools/dc) from Dublin Core to DDI is available for this purpose. Since Dublin Core is very limited, you will most likely want to add additional DDI fields to the records you are creating.

### Content to Include

Study-level metadata available in data catalogs usually describe the program or project from which the data come (i.e., context). This includes key elements about a study, including purpose, the data collection process, and information describing aspects of the research design, such as the universe and sample being surveyed, time period, geographic location, etc.

Following is the set of fields that ICPSR includes in its catalog:

•Title, Alternate Title

•Study Number

•Principal Investigator

•Funding

•Bibliographic Citation

•Series Information

•Summary

•Subject Terms

•Geographic Coverage

•Time Period

•Date of Collection

•Unit of Observation

•Universe

•Data Type

•Sampling

•Weights

•Mode of Collection

•Response Rates

•Extent of Processing

•Restrictions

•Version History

### Use Cases and Examples

It can be helpful to examine which fields others are using. [DDI Lite](http://www.ddialliance.org/specification/ddi2.1/lite/index.html) provides a set of basic DDI fields for discovery. You may also want to consult the [CESSDA](http://cessda.net/Data-Catalogue), [<odesi>](http://search1.odesi.ca/#/), [IHSN](http://catalog.ihsn.org/index.php/catalog), Dataverse, and [ICPSR](http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp) catalogs. ICPSR provides its records in DDI Codebook (Version 2.1 and 2.5) and DDI Lifecycle (Version 3.1) XML format, available on each individual study homepage or in bulk for the [full collection of records](http://www.icpsr.umich.edu/icpsrweb/content/membership/or/metadata/index.html#obtain).

**Example of ICPSR Data Catalog Record**





**Example of ICPSR Data Catalog Record in DDI 3.1**



###

### Example of <odesi> Data Catalog Record



**Example of <odesi> Data Catalog Record in DDI 1x**



### Tools

There are several good tools to create DDI metadata for use in a data catalog(s), including Nesstar Publisher (used by CESSDA and <odesi> catalogs shown above) and NADA (used by the IHSN catalogs). These tools are efficient because they enable the creation of data catalogs with limited effort. And there are excellent training tools (both guides and videos) available to help get you started.

[Nesstar Publisher, Server, and Webview](http://www.nesstar.com/index.html)

-- [User Guide](http://www.nesstar.com/help/4.0/publisher/download_resources/Publisher_UserGuide_v4.0.pdf) to creating DDI in Nesstar Publisher

-- [How to Mark-up in Nesstar](http://www.library.carleton.ca/help/nesstar-how-mark-nesstar) -- Guides and short videos created by Carleton University

[Microdata Cataloging Tool (NADA)](http://www.ihsn.org/home/software/nada)

-- [User Guide](http://documentation.ihsn.org/nada/4.2/) to installing and managing NADA

[Colectica](http://www.colectica.com/)

-- [Document a Study](http://docs.colectica.com/designer/manage-content/studies/document-study/)

-- [Instructional videos](http://www.colectica.com/video) on using Colectica

[DDIEditor from the Danish Data Archive](http://www.samfund.dda.dk/dditools/default.htm) (open source)

### Metadata Resources

Dublin Core mapping to DDI

DDI Lite

[CESSDA catalog](http://www.cessda.net/catalogue/)

<[odesi> catalog](http://search1.odesi.ca/)

[IHSN-ADP catalogs](http://adp.ihsn.org/country_activities)

[ICPSR catalog](http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies?q=&searchSource=find-analyze-home&sortBy=)

[ICPSR records in DDI format](https://www.icpsr.umich.edu/icpsrweb/content/membership/or/metadata/xml/index.html)

DataCite mapping

Dataverse catalog

##

## [Create a Codebook](https://docs.google.com/document/d/1HP10ZcE7MLv6iiMdvZD3IzcLP6NprAGP-EgltNB-B2Q/edit) (Amber)

\*\*\*\*\* go to DDI Training Group > Getting started guides > create a codebook

## Design/Document Survey Questionnaires from Scratch (Jane)

Researchers reusing data often want information on how data were collected. This includes the actual questions that are used in the questionnaire. There is other information (listed below) that is also relevant and needs to be included when documenting the survey questionnaire. Basically, any information surrounding the actual questionnaire and why the different questions were included (or not) should be included in the documentation.

**Why document the questionnaire?**

It is important that any documentation (metadata) surrounding the questionnaire be recorded as the questionnaire is being developed. This is not only good research practice, it will also ensure that the informDation is preserved for the future. There are different methods for creating questionnaires, and documenting the information is necessary at all times. Some of the different circumstances that can occur are: a separate company is designing the actual questionnaire; a different company is administering the actual survey; there is a team of researchers working on different parts of the questionnaire; there is a research assistant working on the questionnaire who is on a contract and will not be there for the duration of the research project; and documenting which CAI (computer assisted interviewing) system was used, if one was used. The commonality among the previous situations is that there may not be anyone to answer any questions about the questionnaire once the research is complete if the information has not been recorded. Thus, it is important that all information be recorded.

DDI is useful for documenting questionnaires as it provides comprehensive information about the data collection phase of the research data lifecycle. Data producers can export DDI profiles from the CAI systems they use, thereby capturing full metadata and making it much easier to document and archive the data. For more information on this subject, check out “Survey metadata reusability and exchange: A call to action for questionnaire documentation”.

<http://www.ddialliance.org/survey-metadata-reusability-and-exchange>

### Content to Include (*Note: this is not an exhaustive list*)

* Actual questions
* Question order
* Skip patterns for the survey
* Introductions used (for the questionnaire and for the different sections)
* Any information that is not directly asked of the respondent but is determined by the interviewer (e.g., gender)
* Format of the questionnaire (e.g., paper, telephone, internet)
* Questionnaire designer
* Universe for each question
* Showcard information
* Interviewer instructions

### Tools

*Note: this is not an exhaustive list, it is merely intended to get you started.*

Data Collection, Colectica Designer

<http://docs.colectica.com/designer/manage-content/data-collection/>

Michigan Questionnaire Documentation System (MQDS)

(based on the Blaise data model)

<http://inventions.umich.edu/technologies/4975_michigan-questionnaire-documentation-system>

Onyx

(open source software for epidemiological data management and analysis)

<http://www.obiba.org/>

REDCap

(for building and managing online surveys and databases, and to support data capture)

<http://www.project-redcap.org/>

REDCap and DDI Exporting Data and Metadata with the API

<http://kuscholarworks.ku.edu/handle/1808/11047>

“The DASISH Questionnaire Design Documentation Tool: Keeping Track of the Questionnaire Design Process” Bakkmoen, H.V., Orten, H. (Norwegian Social Science Data Services), Prestage, Y. (City University London)

<http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/paper/viewFile/164/98>

Rogatus

Some basic features include,

* Create a structured survey instruments
* Create conditional branches in the survey instrument
* Export and Import instrument in DDI format
* HTML preview of instrument

<http://web.opit.hu/QBEE>

**Use Cases and Examples**

The figure below is from the “Canadian General Social Survey, Cycle 25, 2011”. It is an illustration of the DDI information that can be garnered about a variable when the relevant and available information concerning the actual question is documented.

(taken from odesi.ca)

 

The metadata in this figure includes:

* the variable name and label
* the prequestion text
* the literal question
* the postquestion text (note: the total text is cut off in this example)
* the value labels and frequencies
* the summary statistics
* the interviewer instructions (note: the total text is cut off in this example)
* the universe for this particular question

This variable has been well documented and would serve as an excellent example to someone doing similar research.

**Getting started**

Once you have designed your research question, conducted your literature review, determined your population of interest and your method of inquiry, it is time to figure out how you are going to collect your data. This is where the design of the questionnaire happens.

The type of research you are conducting (e.g., qualitative versus quantitative) will also help determine the types of questions that will be used. For example, if you are conducting qualitative research, you will want to use more open-ended and probing questions to help gather data. If you are conducting quantitative research, your literature review will help determine the groups and types of questions you will be using.

When you are designing a questionnaire for the first time, it is helpful to examine other questionnaires that have been used in similar studies. Many studies will publish their questionnaires so you will be able to examine them to get tips for your data collection.

**Question banks**

**Why a question bank?**

If you think about the number of surveys that have already been conducted, you realize that there are many questionnaires out there that could be used to help you design one. As well as question banks, there are also resources for question evaluation that will help you to improve your questionnaire design.

Some organizations show their question wording online as part of the metadata for the variable description. For example, the Minnesota Population Center has a series of integrated public use microdata files (<https://www.ipums.org/>). These series include the IPUMS (harmonized data from censuses around the world), the Integrated Health Interview Series, the American Time Use Survey, the Integrated Demographic and Health Series, and Terra Populus (integrated data on population and the environment).

Tip: Some of the tools are not called question banks, however they can be searched and will list the actual questions.

**Examples**

Social Science Variables Database: (located at ICPSR)

Search over 4 million variables.

Also able to compare questions across studies and series.

<http://www.icpsr.umich.edu/icpsrweb/ICPSR/ssvd/index.jsp>

Q-Bank:

Collaborative resource for question evaluation.

<http://wwwn.cdc.gov/qbank/home.aspx>

American National Election Studies (ANES):

Core utility - lists out concepts/topics by category and lists the actual question

<http://electionstudies.org/CoreUtility/all.htm>

Search Utility - searches ANES Time Series study codebooks to locate questions and variables by keyword.

<http://isr-anesweb.isr.umich.edu/_ANESweb/utilities/cbksearch/searchANES.htm>

UK Data Service Variable and Question Bank:

Search hundreds of surveys.

<http://discover.ukdataservice.ac.uk/variables>

Survey Data Netherlands:

Over 36,000 questions to search.

<http://surveydata.nl>

Tools

SQBL - The Simple Questionnaire Building Language:

Lightweight XML format. Compatible with DDI.

<http://sqbl.org/wiki/index.php/SQBL_-_The_Simple_Questionnaire_Building_Language>

The Canard Question Module Editor:

Cross-platform open-source tool for the design of structured questionnaires.

<http://www.iassistdata.org/downloads/2014/2014_4m_spencer.pdf>

Colectica - Manage a Question Bank:

Manage questions independent of the survey in which they may be used.

Create your own question banks.

<http://docs.colectica.com/designer/manage-content/data-collection/manage-questions/>

**Getting started**

Let’s say you are interested in looking at youth and smoking. You want to know the types of questions that have been asked in this area.

You start by going to the Social Science Variables Database at ICPSR and type in ‘youth and smoking’. There are 148 results returned.



Then you select the results in which you are interested and the variables will be compared for you. The information given includes: name of the survey; the variable name; the question; the responses; the time period; and the universe.

This particular database also has search tips that will help you to narrow your search down.

**Capture the Output of a CAI System (Jon)**

Computer Aided Interviewing is increasingly the preferred method of data collection in larger studies, through dedicated software such as Blaise [<http://blaise.com/>} , IBM Data Collection [http://www-01.ibm.com/software/uk/analytics/spss/products/data-collection/] or CASES [http://cases.berkeley.edu/], and increasingly through web based software such as REDCap [<https://catalyst.harvard.edu/services/redcap/>] or LimeSurvey [<https://www.limesurvey.org/en/>].

As well as capturing what was implemented in the survey collection (as opposed to what was specified), it is important in provenancing the output variables and reusing this. Without the need for manual entry this is obviously also a great time saver.

Unfortunately, none of the software packages support DDI-L natively, but much can be extracted from them that can be used to capture the majority of the questionnaire used.

The way in which CAI software is programmed can make a big difference to what is retrievable at the end of a survey. For instance, for convenience and display purposes, questions and interviewer instructions are often concatenated together. Splitting these out as separate items when programming will make extracting these metadata elements more straightforward.

A further structural problem with most CAI software is that it separates out the flow controls of the questionnaire from the questions and instructions, so software that converts to DDI-L where these are unified into a single model will need to combine them.

**Tools and Examples**

**Blaise** (using the Michigan Questionnaire Documentation System (MQDS) plugin). A version that supports Blaise 5 and DDI-L 3.2 is currently in development, although earlier versions of MDQS export out DDI-L 3.1, which can be transformed to DDI-L 3.2 using tools such as SledgeHammer and Colectica Designer. Colectica Designer also has a plugin that allows extraction of DDI-L 3.2 from a Blaise (bim) source file.

**IBM Data Collection** has its own proprietary XML format which contains most of the information you would want to extract to DDI-L. Colectica Designer also has a plugin that allows extraction of DDI-L 3.2 from a IBM Data Collection (mdd) source file.

**LimeSurvey** has its own internal XML format. This is not as rich as the most current version of DDI-L, but they have produced tools that will convert to DDI at [<https://surveys.acspri.org.au/quexmltools/>]. Colectica Designer also has a plugin that allows extraction of DDI-L 3.2 from a LimeSurvey (queXML) source file.

**REDCap** has arguably a less sophisticated export format than the previously mentioned CAI software, but again contains much which can be repurposed to DDI-L. There is an R package available from [<http://kuscholarworks.ku.edu/handle/1808/11047>] which converts a REDCap (csv) source file. Colectica Designer also has a plugin that allows extraction of DDI-L 3.2 from a REDCap (csv) source file.

**CASES** is currently being redeveloped, DDI files can be imported into CASES. Colectica Designer has a plugin that allows extraction of DDI-L from a CASES (.q;.m.FILES) source file.

## Document Data in Excel (Dan)

Excel files are notorious for being able to store many kinds of data, charts, formulas, and other content in any style and layout a user chooses. While it is difficult to to document all types of Excel files, DDI is well suited to document well formatted Excel files containing tabular data.

There are two types of data that a well formatted Excel file may contain.

* Measurement data
* Aggregate data

DDI describes both measurements data (microdata) and aggregate data. Measurement data is arranged in defined columns, with an optional first row containing the column headers or names. Each row of measurement data contains a set of measures or values. Aggregate data is normally arranged in a cube or pivot table, with a set of dimensions and aggregate measurements.

An Excel file can contain multiple sheets. A well formatted Excel file will have one dataset per sheet. This allows describing each sheet as a tabular or aggregate dataset, and to document relationships between them.

### Tools

An Excel file may be easily documented using the free Colectica for Excel addin. This tool integrates into Excel to allow documenting Excel workbooks and sheets using the DDI standard. It provides an extra panel to record more detailed descriptions of the columns and variables of a dataset, and stores the additional information directly within the Excel file. The tool will also create pdf data dictionaries from documented Excel files. The Professional version of Colectica for Excel can also import and document datasets saved in proprietary statistical package formats, such as SPSS and Stata.

Colectica for Excel
<http://docs.colectica.com/excel/>

<http://www.colectica.com/software/colecticaforexcel>

\*\*\*\*\*

## Document and Manage Longitudinal Data (Jon)

Hopefully you are using a database (of some description) to hold your data. Perhaps you might have some way of looking up in your database the provenance of each data item, i.e., which question (or questions) did this come from. Other ways of managing this problem are in a panel study where it is common to keep the name of the data item the same, but prepended with a wave identifier. In either case having a lookup of comparable variables is an essential starting point.

Normally, users of the data will not access your data directly, but in a statistical package such as SPSS or Stata and this is what you will want to document.

There are several tools available to document the dataset(s) in the current version of DDI-L, StatTransfer, SledgeHammer and Colectica. They do this by extracting the rich metadata that already exists in your statistical dataset, such as the variable labels, the code lists, and other information such as the number of cases, and optionally the statistics (mean, max, frequency etc) for the variables themselves and creating a DDI-L file in a standardized format that can be used across all of your data.

DDI-L provides several mechanisms for grouping variables that are comparable. Fuller discussion of this is in the DDI Best Practice series *Enabling Longitudinal Data Comparison Using DDI [*<http://dx.doi.org/10.3886/DDILongitudinal02>] although you should be aware this paper refers to a previous iteration of DDI-L (version 3.1).

The following diagram illustrates a simplified version of the connection between a source question, the question used in a set of surveys, the variables and the mapping between them, The Design/Document Survey Questionnaires from Scratch shows how you can document the questions (which are held in a QuestionScheme).

In this example, we have incorporated the name of the question and variable into the DDI URN which uniquely identifies the question or variable with DDI-L. This is composed of an Agency (the owner of the object) and identifier for the object and a version. The same question was asked at two different sweeps of two studies, the study is indicated by the DDI Agency uk.cls.mcs and uk.cls.ncds and ethe1, ethe2, ethe7 and ethe8 are the identifier for the questions,

A represented variable is a mechanism to allow use to create a ‘virtual variable’ that we can use to link variables together and allows us to document their similarity.

In this illustration, a represented variable has been created within each study uk.cls.mcs.ethe2000 and uk.cls.ncds.ethe2000. An additional represented variable has also been created that references them and potentially other occurrences of this represented variable.



### Tools

The above can be accomplished using traditional XML tools such as XSLT or using Colectica Designer and Repository.

\*\*\*\*\*

## Create a Question Bank (Jane - this is combined with questionnaires section further up, as discussed at last mtg)

\*\*\*\*\*\*

## Create Comparison Maps/Crosswalks/Concordances (Dan)

DDI Lifecycle allows you to make comparisons of variables across datasets. These comparisons can be recorded in a standardized manner, detailing semantically meaningful relationships between similar variables. This information can form the basis of exploratory tools for researchers.

### Tools

“Colectica allows you to harmonize variables that measure the same information at different times or among different populations. Colectica uses the DDI Lifecycle *represented variable* item type to describe this harmonization. Harmonized variables documented in DDI can also be explored and compared on the web using Colectica Portal.”

Colectica - [Manage Harmonized Variables](http://docs.colectica.com/designer/manage-content/data/harmonize-variables/)

*[Data Dissemination and reuse] Provide simple searches at the study and variable level*

* *is this similar to above? Or would this be like a discovery record?*
* *DDI Index tool*

*[Data Dissemination and reuse] Citing datasets*

 *-similar to above? List the necessary tags to properly identify and cite data; i.e. versioning, URI etc. title, year, format, producer, distributor etc.*

*-machine actionable*

*-”credit where credit is due” paper (Altman)*

 *-contributions to datasets*

 *-CASRAI site*

 *-Dagstuhl workshop and paper on data citation*

*Quick guide to extracting DDI from existing datasets using StatTransfer*

*-or Sledgehammer*

*-or Colectica*

 *-----------------------------------------------------------------------------------------------------------*