cessda eric
Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

Research Data Management

International Summer School
in Uganda

Dr. Anja Perry

Oliver Watteler

gesis Leibniz Institute
for the Social Sciences

cessda eric

Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

Documentation and Metadata

INTRODUCTION TO DATA DOCUMENTATION AND METADATA

What are metadata?

- **Metadata are…**
  - … data about data, i.e. a description of data
  - … "structured or semi-structured information which enables the creation, management, and use of records [i.e. data] through time and within and across domains" (Day, 2005)

$\Rightarrow$ **Metadata tell a story about the data, i.e. makes …**
- … sense of the data
- … the data detectable

Functions of metadata

Cite

Administrate and archive

Share and re-use

Metadata

Understand and interpret

Visible and findable

Replicate/Analyze

Images pixabay (CC-0)

CREDIBILITY

Levels of data documentation

Metadata on...

study level

variable level

Metadata can be…
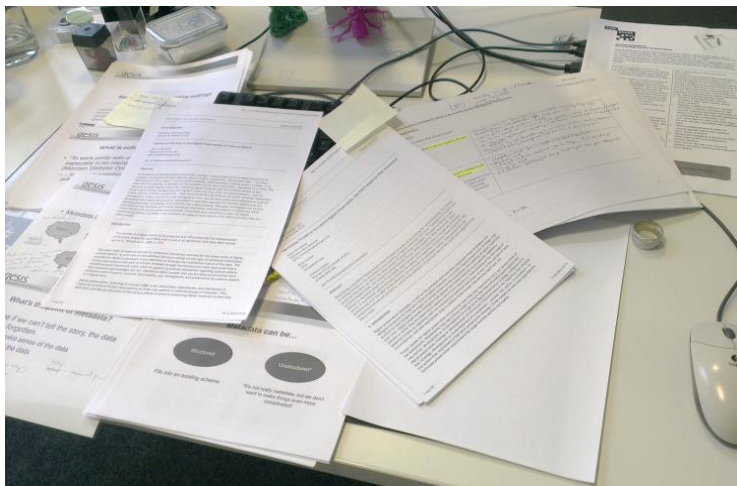
unstructured

Un-/semi-standardized documentation of the data, e.g.

- codebooks
- methodical reports etc.
- protocols

structured

structured information in a standardized manner

- ISCO (Classification)
- DDI (Metadata)

Standards for Structured Metadata

- Different standards exist, for example
  - Dublin Core
  - Data Documentation Initiative (DDI)
  - Statistical Data and Metadata eXchange (SDMX)
- Standard guidelines for documentation
- Only indirectly relevant for smaller research projects
- Directly relevant for data repositories and archives

Take-Home Message

- Documentation and registration are important
- Minimum requirements for re-use and replication (good scientific practice):
  - Data file
  - Variable documentation
  - Questionnaire
  - Methodological report
- Scope of documentation depends on project size, amount of data, personal and financial resources
- Create data management plan and apply for sufficient grants for data documentation and processing work

DOCUMENTATION AT THE STUDY LEVEL

Study Level

- **Study description**
  - aim of study
  - primary researchers
  - funders etc.
- **Study design**
  - population and sampling procedure
  - method of data collection etc.
- **Data processing**
  - data cleaning
  - anonymization etc.

Basic Elements

General context in which data were collected and the methodological steps in generating the data at hand:

- Methodological or technical report

- Questionnaire

- Any other materials (e.g. interviewer selection grid for households, respondent stimuli)

Methodological Reports

- (aka field reports or technical reports)
- Provide researchers with an overview of
  - survey design,
  - data collection, and
  - statistical processing of the data to be analyzed
- Should enable the researchers to assess the quality and the analytical potential of the data
- Related to possible "errors" occurring during the survey research process based on the total survey error (TSE) approach.

Total Survey Error Framework

Representation

Measurement

| Target Population | | Construct |

Coverage Error

Validity

| Sampling Frame | | Measurement |

Sampling Error

Measurement Error

| Sample | | Response |

Nonresponse Error

Processing Error

| Respondents | | Edited Data |

Adjustment Error

| Survey Statistic |
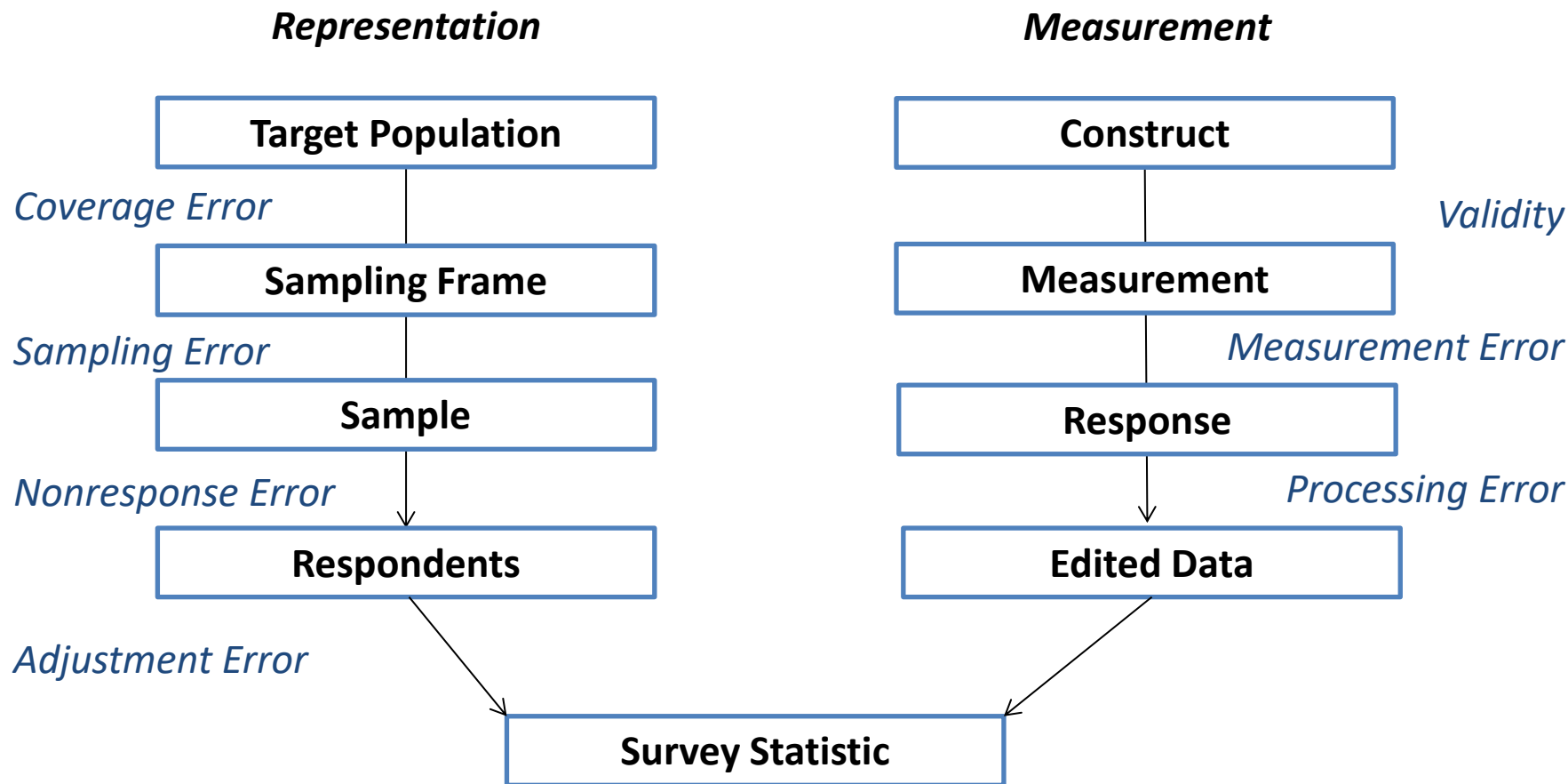
Source: Adapted from Groves et al. (2004).

Guidelines for Survey Documentation

Question	Report Section	Sources of Error
For what purpose were the data collected?	Objective and Design	
How were the respondents selected?	Target Population and Sampling	Coverage error Sampling error Unit nonresponse
How were the data collected?	Mode of Data Collection	Mode measurement error
What information was collected?	Survey Instrument	Respondent measurement error Item nonresponse
Who has collected the data when and where?	Fieldwork	Interviewer measurement error House effects
How were the data edited, coded, and weighted?	Data Processing	Processing errors Adjustment errors
Were provisions of data protection laws respected?	Data Protection and Ethical Issues	Ignoring legal issues

Guidelines for Survey Documentation

Methodological report front matter:

- Study title/series title

- Principal investigators/project team

- Funding

- Recommended citation

Guidelines for Survey Documentation

Objectives and Design: background, research objectives, research design (e.g. cross-section, trend, panel)

Example

The European Values Study (EVS) is a large-scale, cross-national, and longitudinal survey research program on how Europeans think about family, work, religion, politics and society. Repeated every nine years in an increasing number of countries, the survey provides insights into the ideas, beliefs, preferences, attitudes, values, and opinions of citizens all over Europe. Four waves of surveys were executed from 1981 to 2008. These surveys explore value differences, similarities, and value changes.

Source: EVS 2008 Method Report

Guidelines for Survey Documentation

Target Population and Sampling: the target population and eligibility criteria, sampling frame, respondent selection at each sampling stage, description of clustering and/or stratification

Example EVS Sampling Procedure Germany

Persons 18 years or older who are resident within private households, regardless of nationality and citizenship or language. [...] The basis for the study is a random sample drawn from resident registers of German municipalities (a national resident register does not exist). Therefore the sampling has to proceed in two steps: (1) a random sample of municipalities had to be drawn and (2) random samples of the municipalities' resident registers. The sample design is disproportional and takes the distinction of East and West Germany into account where the East and West of Berlin are attributed to the respective parts of Germany. In order to realize the oversampling of East German population, the sample of municipalities is stratified according to the federal states and to 7 size-classes of the municipalities (in order to reflect the population of the municipalities in the sampling probabilities).

Source: EVS 2008 Method Report

Guidelines for Survey Documentation

Mode of Data Collection: self-administered vs. interviewer-administered, computer-assisted vs. not computer-assisted, mixed-mode designs

Example EVS
In all countries, fieldwork was conducted on the basis of detailed and uniform instructions prepared by the EVS advisory groups. The EVS questionnaires were administered as face-to-face interviews in the appropriate national language(s). As far as the data capture is concerned, CAPI or PAPI was used in nearly all countries. Exceptions are Finland (internet panel) and Sweden (postal survey).

Source: EVS 2008 Method Report

Guidelines for Survey Documentation

Survey Instrument: topics of the questionnaire, construction of scales (psychometric quality), special instruments, and results of traditional or cognitive pretests, translation

Table 2:   Thematic categories and related variables in EVS 2008

Preceding character	Thematic category	New category	Number of variables	Number of *new* variables	Number of variables used in EVS 1999 *and* 2008
A	Perceptions of life		155	12	73
B	Environment		14	6	1
C	Work		63	5	25
D	Family		71	7	24
E	Politics and Society		160	9	63
F	Religion and Morale		164	8	46
G	National identity		30	20	4

Source: EVS 2008 Project and Data Management

Guidelines for Survey Documentation

Fieldwork: field dates, number of interviews, interview duration, interviewer training and monitoring

Example EVS Belgium
The fieldwork was preformed by a total of 104 interviewers. The majority (n = 87) of the interviewers, selected by the field agency, had previous experience with scientific survey research. Interviewers were classified as experienced if they had participated in at least one academic social survey in the past three years. The inexperienced interviewers (n = 17) were trained by the field agency and have already worked on non-academic market research. They also participated in the EVS-Briefing, as did all interviewers. The interviewers had an average age of 54 years, ranging from 25 years to 90 years. A crosstabulation of educational level by gender is provided in table 2.6.

Source: EVS 2008 Method Report

Guidelines for Survey Documentation

Response Rates: contact rates, cooperation rates, response rates and refusal rates (AAPOR)

Standard Definitions

Final Dispositions of Case Codes and Outcome Rates for Surveys

Revised 2016

RDD Telephone Surveys
In-Person Household Surveys
Mail Surveys of Specifically Named Persons
Mail Surveys of Unnamed Persons
Internet Surveys of Specifically Named Persons

Guidelines for Survey Documentation

Data Processing: how the data were edited and cleaned up in general, coding of open-ended responses, creation of different types of weights

Example EVS Weighting Procedure Germany
The variable "weight" was computed by the EVS for all national datasets on the basis of information and population statistics provided by the EVS countries. The weight is constructed on the basis of gender and age categories (-24; 25-34; ... ; 65-74; 75 and over). Value '0' implies that year of birth information was missing in the data. The weight adjusts the socio-structural characteristic in the samples to the distribution of gender and age of the universe-population. For German and Belgium data an additional country-specific weight variable (weight_c) is provided that includes a special weight factor for the regions of Germany (East- and West) and of Belgium (Brussels capital region, Flanders and Walloon region). This design weight corrects for the disproportional sample size of these regions in both countries. The reported population sizes refer to adult inhabitants, i.e. people of age 18 and older. For the computation of the weighing factors two calculation steps are necessary:
*(I) (inh. Region1 / inh. country) * total sample size = proportional share of interviews from Region1*
(II) proportional share of interviews / realized share of interviews = weighing factor

Source: EVS 2008 Method Report

Guidelines for Survey Documentation

Data Protection and Ethical Issues: documentation of informed consents, any measures to protect respondent privacy (e.g. aggregation of variables).

Documentation of Survey Instruments

- Include the original questionnaire and any other materials used during fieldwork processes (respondent- and interviewer-related)
- Document fieldwork questionnaire as used by the survey vendor
- Use screenshots from the survey software to detect differences between original questionnaire and the questionnaire in field
- Respondent materials: advance letters, show cards, consent forms, experimental stimuli etc.

Exercise: Data Documentation

💬 work in 5-6 groups

🕐 time: about 30 minutes

☺ afterwards, we will commonly discuss your suggestions

☞ see Exercise-Booklet for details on Exercise 7

Content and Language

- **Clear and unambiguous language**
  - "college" and "university" are no synonyms

- **Describe and don't evaluate**
  - current sample can be considered to represent the population of interest

- **Complete information**
  - who were the interviewers?
  - how had they been trained?

- **Consistent information**
  - e.g. no lectures at University of St. Gallen in July 2015

Structure of Field Report

1) **Research population and area**

2) **Sampling process**

   $\Rightarrow$ method of sampling and underlying population

3) **Realization of interviewing**

   $\Rightarrow$ who did what, when (field period) and how (method of interviewing)

4) **Realized sample**

   $\Rightarrow$ realized interviews, drop-outs and response rate

DOCUMENTATION AT THE VARIABLE LEVEL

Variable Level or …

- **Questionnaire**
  - original question wording
  - provided answer categories
  - explanations, interviewer instructions etc.

- **Variables**
  - labels and meanings of variables
  - labels and meanings of coding
  - variable notes etc.
  - scales

Image: pixabay (CC-0)

What is a Codebook?

"A codebook describes the contents, structure, and layout of a data collection. A well-documented codebook "contains information intended to be complete and self-explanatory for each variable in a data file"

Why provide a Codebook ?

- Documentation of the published dataset on variable level

- Description of a dataset's content, structure, and layout

- Provision of contextual information about the study the published data file is part of

- Connection of variables with questions, answer categories

- A report of findings of data checks and modifications during data cleaning/processing for specific variables

- An accompanying assisting handbook for dataset end-users

Content of a Codebook

- **A detailed codebook should contain**
  - information on variable level in naming, structure, format
  - explanations for deviations between field questionnaire and dataset after data cleaning that are not obvious
  - Information about used standards

- **It might also include methodological information like**
  - if and which standards were used (e.g. ISO codes for geographic units, International Standards Classification of Education)
  - reference to scales and instruments used (if they are not documented in the report)

Content of a Codebook

Codebook front matter section should contain basic (bibliographic and identificational) information about dataset and study/series for unambiguous assignment:

- study title, series information
- principal investigators of study/publisher of data
- information about the dataset file
    - persistent identifier of study if available
    - data set version and changes from previous changes
    - citation recommendation for dataset
- introducing part (purpose and format of a codebook)
- table of content of the CB
- An overview/list of all dataset variables

Content of a Codebook

Codebook main body contains information about the dataset that are self-explanatory for each variable in a data file

- Variable name, variable label, variable format
- Question text / question number in field questionnaire
- Meaning of values (variable values)
    - For scale variables: The variable's units of measurement
    - For categorical variables: if numeric codes and what they represent
- missing data: Values and labels of user and system missing data (e.g. question not asked, not applicable, don't know, no answer)
- Variable notes: filter, modifications, standards, variable functions, inconsistencies

Example of Variable Documentation

Variable: Educational level of respondent in the European Values Study 2008 ([doi:10.4232/1.11004](doi:10.4232/1.11004))

1. Online data documentation provided for retrieval purposes and data exploration in GESIS ZACAT

ZA4800: EVS 2008: Integrated Dataset

ZA4800 Datafiles and Documentation download (via data catalogue)

Variable v336_r: education level (Q110) (recoded)

LITERAL QUESTION
EDUCATIONAL LEVEL OF RESPONDENT - RECODED (3 CATEGORIES)

Source variable: v336_4
[What is the highest level you have completed in your education?]

-5 other missing
-4 question not asked
-3 not applicable
-2 no answer
-1 don't know
1 Lower
2 Middle
3 Upper

Comparability:
Trend question (EVS 2008=EVS 1999).

Show Card

Examples of Variable Documentation

2. Variable documentation (Codebook in PDF) provided in **GESIS Data Catalogue**

EVS 2008 - Germany

GESIS StudyNo 4753, v1.1.0, http://dx.doi.org/doi:10.4232/1.10151

v1 - how important in your life: work (Q1A)

WE START WITH SOME QUESTIONS ABOUT LIFE IN GENERAL, LEISURE TIME ACTIVITIES AND WORK

Q1

<SHOW CARD 1>

Please say, for each of the following, how important it is in your life.

Q1.A Work

-5 other missing
-4 question not asked
-3 nap
-2 na
-1 dk
1 very important
2 quite important
3 not important
4 not at all important

v1, weighted by weight_c

Value	Label	Missing	Count	Percent	Valid Percent
-2	na	M	4	0.2	
-1	dk	M	21	1.0	
1	very important		1000	48.2	48.8
2	quite important		696	33.5	34.0
3	not important		184	8.9	9.0
4	not at all important		170	8.2	8.3
	Sum		2075	100.0	100.0
	Valid Cases		2050		

Persistent identifiers

Persistent Identifiers

- **Ensures**
  - permanent findability
  - unambiguous identificati
  - citation

- **Example: Digital Object Identifier (DOI)**
  - a kind of PID
  - ca. 5000 registration agencies, e.g. DataCite or da|ra
  - „DOI-resolver"
    e.g. http://www.doi.org/index.html

DOI: 10.ORGANIZATION/ID
Example: 10.4232/1.11159
Prefix/Suffix

da|ra schema: main categories (rich metadata)

	General Resource Type		
	Resource Type		
0			
0.1			
	Title		
	Other Titles		
1	Title Type		
2	Collective Title		
2.1	Numbering		
3	Creator		
3.1			
4	Person		
4.1	Person ID		
4.1.1	Vocabulary of Person ID		
4.1.1.1	URI Name Authority Record[2]		
4.1.1.2	Affiliation		
4.1.2	Affiliation ID		
4.1.2.1	Vocabulary Affiliation ID		
4.1.2.1.1	URI Institution Authority Record		
4.1.2.1.2	Institution		
	Institution ID		
4.2	Vocabulary of Institution		
4.2.1	URI Institution Authority ID		
4.2.1.1	URI Institution Authority Record		
4.2.1.2	URL		
	DOI Proposal		
8	Version		
9	Keywords Internal		
10	Internal Identifier		

Sampled Universe Publication Date Keywords (controlled)

Availability (controlled) Relation

Title Other Titles Dataset Contributor

Version Creator Description Availability (free)

Temporal Coverage Alternative Identifier

Language Collection Mode (controlled)

Classification Internal Collection Mode (free) Notes

Geographic Coverage URL

Time Dimension Rights

Collective Title Sampling

Publications Publication Place

DOI Proposal

General Resource Type Keywords (free)

44

Cite the data you use!

https://datacite.org/cite-your-data.html

45

ZA6980: International Social Survey Programme: Social Networks and Social Resources - ISSP 2017

Bibliographic Citation	**Content**	**Methodology**	**Data & Documents**	**Errata & Versions**	**Further Remarks**
Publications	**Groups**				

Citation	*ISSP Research Group (2019): International Social Survey Programme: Social Networks and Social Resources - ISSP 2017. GESIS Data Archive, Cologne. ZA6980 Data file Version 1.0.0, doi:10.4232/1.13251*
Study No.	ZA6980
Title	International Social Survey Programme: Social Networks and Social Resources - ISSP 2017
Current Version	1.0.0, 2019-3-18, doi:10.4232/1.13251
Date of Collection	13.01.2017 - 07.03.2019
Principal Investigator/ Authoring Entity, Institution	1 2 3 4 5 6 • Muckenhuber, Johanna - Institut für Soziologie, Karl-Franzens-Universität Graz, Graz, Austria • Höllinger, Franz - Institut für Soziologie, Karl-Franzens-Universität Graz, Graz, Austria • Hadler, Markus - Institut für Soziologie, Karl-Franzens-Universität Graz, Graz, Austria

Afrobarometer – Data use policy

Copyright

Afrobarometer data are protected by copyright. Authors of any published work based on Afrobarometer data or papers are required to acknowledge the source, including, where applicable, citations to data sets posted on this website. **Please acknowledge the copyright holders in all publications resulting from its use** by means of bibliographic citation in this form:

Afrobarometer Data, [Country(ies)], [Round(s)], [Year(s)], available at http://www.afrobarometer.org.

We also request users to send/email copies of any publications, papers, or reports that employ Afrobarometer data to Brian Howard, Afrobarometer's publications manager, at bhoward@afrobarometer.org.

https://www.afrobarometer.org/data/data-use-policy

Exercise: Data Evaluation

🗨 work in 5-6 groups

🕐 time: about 10 minutes

☺ afterwards, we will commonly discuss your suggestions

☞ see Exercise-Booklet for details on Exercise 8

What to look for when assessing quality?

Image: CESSDA Training Working Group (2017) (CC-BY-SA)

Can I establish

- Why the data was created?
- What the dataset contains?
- How data was collected?
- Who collected the data and when?
- How was the data processed?
- Any manipulations done to the data?
- What quality assurance procedures were used?

Metadata Standard DDI

cessda eric
Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

What is DDI?

- Data Documentation Initiative

- *"An effort to develop a specification for documenting data files in XML. The DDI Alliance is the organization that created the specification, …"*

- More information can be found on the DDI website: http://www.ddialliance.org/

DDI as a standard

- DDI is a *standard* for metadata

- The standard structure means that all computers, even if they are using different applications, can work on the same data and related information (metadata/documentation)
    - The formats are not proprietary to any specific system
    - Uses generic XML technology as the basis for cross-platform use

http://www.ddialliance.org/community/join

Example XML

```xml
<StudyUnit xmlns="ddi:studyunit:3_1" id="StudyUnit36053" versionDate="2015-04-16">
    <Citation xmlns="ddi:reusable:3_1">
        <Title>Cognition and Aging in the USA (CogUSA) 2007-2009</Title>

                        <Creator xmlns="ddi:reusable:3_1" affiliation="University of S

                        <Creator xmlns="ddi:reusable:3_1" affiliation="University of M

                        <Creator xmlns="ddi:reusable:3_1" affiliation="University of M

            <Publisher>Inter-university Consortium for Political and Social Research</Publ
                    <Contributor role="distributor">ICPSR</Contributor>
                    <PublicationDate>
                    <SimpleDate>2015-04-16</SimpleDate>
                    </PublicationDate>
                    <InternationalIdentifier xmlns="ddi:reusable:3_1" type="ICPSR Number">
                    <InternationalIdentifier xmlns="ddi:reusable:3_1" type="DOI">doi://10.
    </Citation>

    <Abstract isIdentifiable="true" id="Abstract36053">
        <Content xmlns="ddi:reusable:3_1">
        <div xmlns="http://www.w3.org/1999/xhtml" id="Summary36053">Cognition and Aging in
```
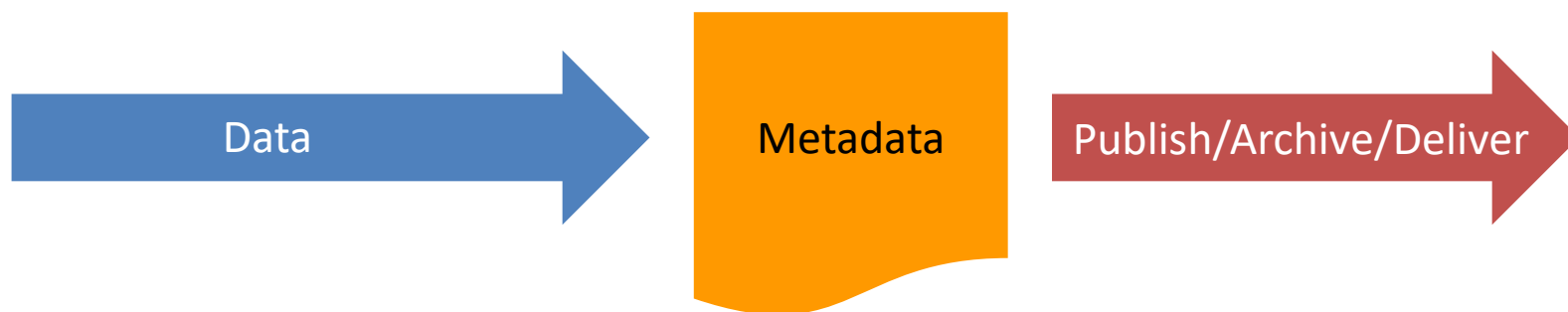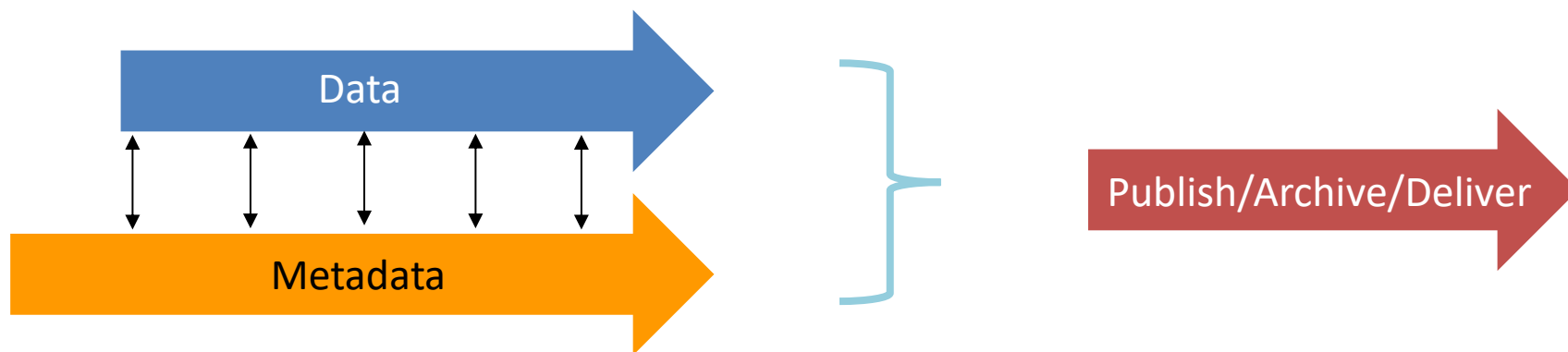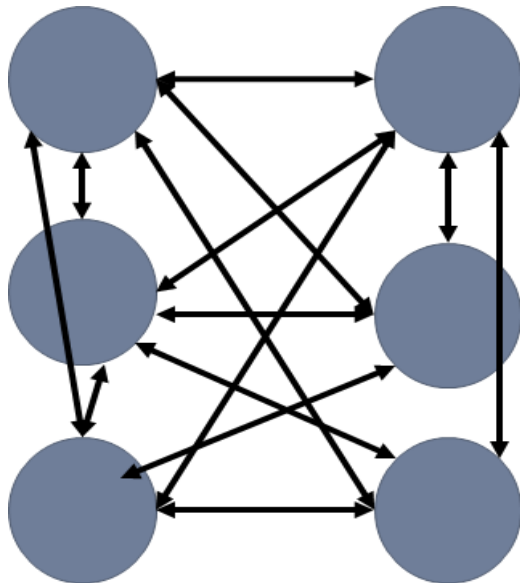
https://ddialliance.org/training/getting-started/data-catalog

"Upstream" vs. "Downstream" Metadata Capture

"Downstream" Metadata Capture

Data → Metadata → Publish/Archive/Deliver

"Upstream" Metadata Capture

Data

Metadata

Publish/Archive/Deliver

DATA DOCUMENTATION INITIATIVE

Management Patterns for Data and Metadata

Decentralized:

Difficult to manage

Centralized:

Easier to manage

DDI Research Lifecycle

http://www.ddialliance.org/training/why-use-ddi
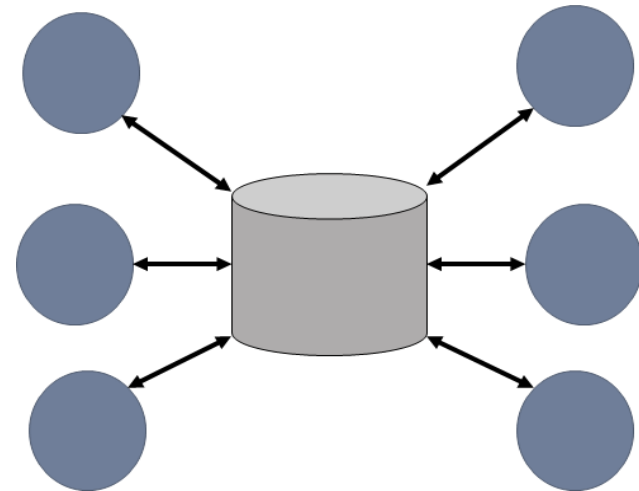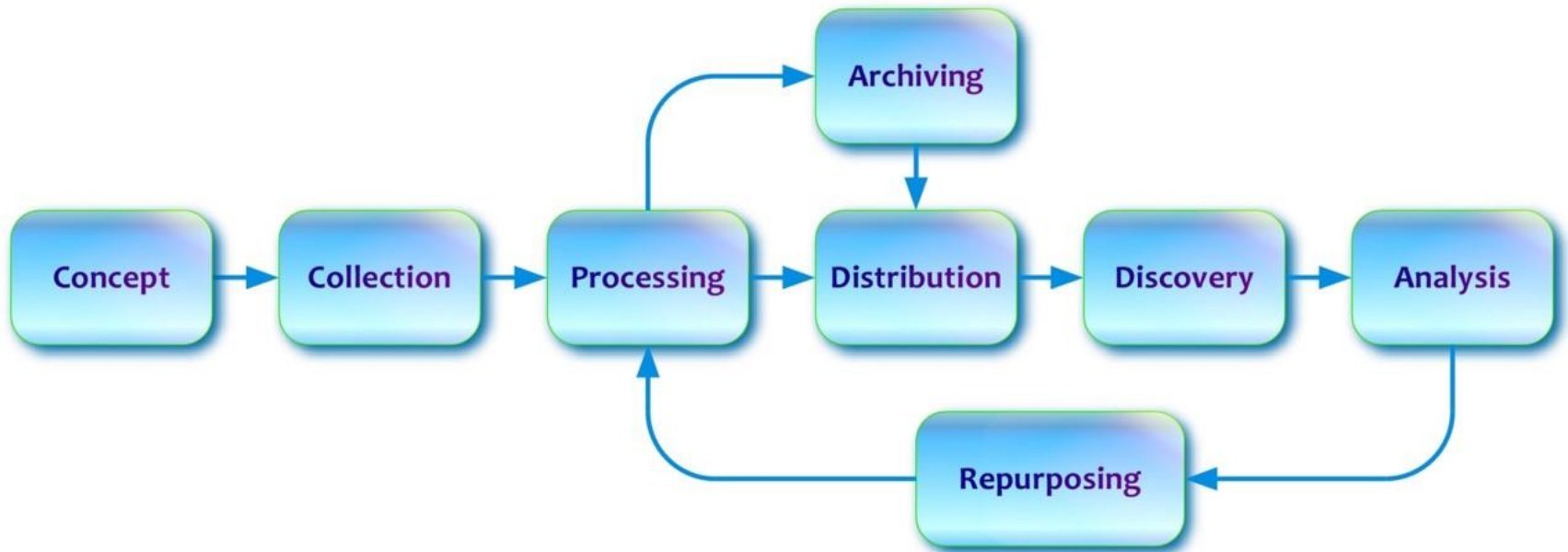
Getting started with DDI

- **Daunting at first**
  - Process is broken down into steps
- **Lots of help available**
  - DDI Alliance
  - http://www.ddialliance.org/training/getting-started
  - Colleagues
  - Other researchers
- **DDI List-serv**
- **DDI Best Practices**
  - Work in progress
  - *Feedback always welcome*

Tools to help you get started

Tool Purpose	DDI version(s) supported	Availability	Supported Operating Systems
- Any - ▼	- Any - ▼	- Any - ▼	- Any - ▼

Search Tools by Name

[ ] **Filter** **Reset**

Name	Version(s) supported	Availability	Description	Purpose

Thank you for your attention!

anja.perry@gesis.org
oliver.watteler@gesis.org

Adapt your DMP

- Adapt your Data Management Plan regarding Data Documentation and Metadata.

Your task

Exercise: Data documentation

Please complete exercise 9 in the booklet and answer the questions concerning three sample datasets.