What do you know about your data? Metadata and PIDs at the GESIS Data Archive

Meet the Experts – GESIS online talks

Data Services, Data Archiving, and Research Data Management

Wolfgang Zenk-Möltgen, Jan Schwalbach, Kokila Jamwal, 2023-10-26

gesis
Leibniz Institute
for the Social Sciences

Meet the Experts

Leibniz
Association

Speakers

Wolfgang Zenk-Möltgen

- Deputy head of team Metadata Standards and Interoperability
- Service: Data documentation workflow and software, metadata standards
- Research interests: DDI, Data sharing and citation, environmental attitudes

Dr. Jan Schwalbach

- Postdoctoral researcher in team Archiving
- Services: Work Package Manager Archiving & Provision of Digital Behavioral Data
- Research interests: Legislative politics, pol. behavior and pol. representation

Kokila Jamwal

- Doctoral researcher in team Metadata Standards and Interoperability
- Services: Metadata for Digital Behavioral Data
- Research interests: Metadata, data findability and data reusability

Data Services for the Social Sciences (DSS)

Data Services	Community Data Collection	Metadata & PIDs
RDM & RDM Training	DP-R|EX	Risk & Sensitive Data
	Certification	

▼ Archiving

▼ Data Acquisitions and Access

▼ Metadata Standards and Interoperability

Outline

- Introduction
- Selection of standards
- Conversions to exchange metadata
- Example applications
- Current state of workflows
- Challenges for metadata applications
- Metadata for Digital Behavioral Data
- Conclusions

Introduction

- 6592 research data sets in the GESIS Data Archive
- Mostly from Germany and EU Countries
- Predominantly from Sociology and Political Science and mostly survey data:

Mass political behaviour, attitudes/opinion (2,017)

Government, political systems and organisation (1,496)

Elections (929)

Economic systems and development (704)

International politics and organisation (695)

Employment (663)

Mass media (542)

Family life and marriage (494)

Social conditions and indicators (457)

Social behaviour and attitudes (449)

more...

Research data (6,592)

GESIS-DBK (6,266)

GESIS-SowiDataNet|datorium (234)

GESIS-MISSY (92)

Germany (3,122)

German Democratic Republic (1949-1990) (950)

France (787)

Italy (745)

Netherlands (745)

Spain (700)

Ireland (695)

Belgium (691)

Denmark (682)

Portugal (662)

more...

Meet the Experts

Introduction

- Tasks of the Data Archive include
  - Creating study descriptions
  - Offer data deposit and access
  - Help researchers to create documentation
  - Create complex documentation
  - Give advice for research data management
  - Long-term archiving
- User Services
  - Formerly done often by special purpose portals
  - Will be integrated as much as possible into general purpose website
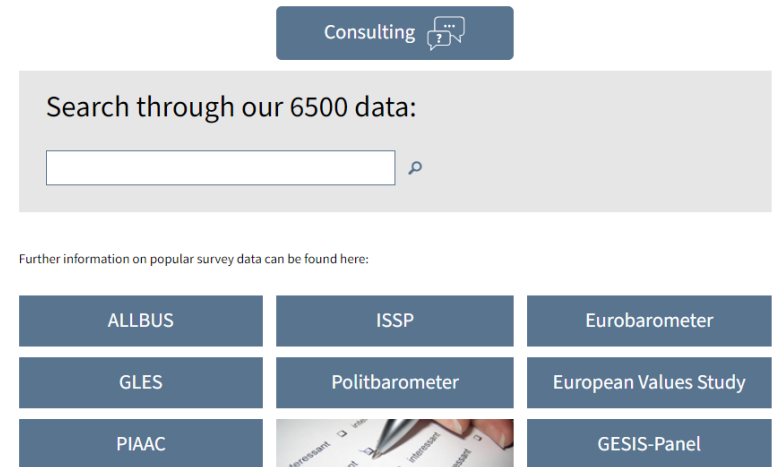
Selection of standards

- Why? Selection builds a ground for development of applications

- Well defined interfaces are needed

- Until now: DDI-Lifecycle as the basis

- Extended and combined with additional standards

- Technically managed with the DDIFlatDB approach[1]

[1] Hopt, Oliver, Claus-Peter Klas, and Alexander Mühlbauer. "Flexible DDI storage."
IASSIST Quarterly 42.2 (2018): 1-12.

Selection of standards

Additional standards used:

- ISO standards for date, time, and country
- CESSDA CMM metadata (DDI compliant)
- da|ra and DataCite metadata schema
- Controlled Vocabularies from all these

Need to do version management (metadata, CVs)

- E.g., ISO Countries do change over time
- Also, metadata schemata do change, especially versions of CVs may differ

Selection of standards

Additional relevant standards

- Semantic web, RDF, SKOS

- Schema.org, JSON-LD, FAIR

- OpenAire, DCAT

- Sometimes applied several times through chained workflows

Topic Classifications

- ELSST, TheSoz, ZA Categories, CESSDA Topic Classification, Continuity Guides, etc …

Selection of standards

Resources:

- DDI-Lifecycle at the Data Archive, http://nbn-resolving.de/urn:nbn:de:0168-ssoar-52487-3
- DDI-Codebook 2.5, https://ddialliance.org/Specification/DDI-Codebook/2.5/
- DDI-Lifecycle 3.2, https://ddialliance.org/Specification/DDI-Lifecycle/3.2/
- Controlled vocabularies, https://vocabularies.cessda.eu/
- Da|ra Metadata Schema 4.0, https://doi.org/10.4232/10.mdsdoc.4.0
- DataCite Metadata Schema 4.4, https://doi.org/10.14454/3w3z-sa82
- F-UJI Automated FAIR Data Asessment Tool, https://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool

Conversions to exchange metadata

Mappings or Cross-walks between different standards

- Efficiency
  - document once, export to many portals
  - Automatic procedure

- Problems
  - semantic meanings do not always support this
  - Higher or lower granularity may be needed
  - Context might change

Example applications

- Study Description
  - Created in DBKEdit, saved in MySQL
  - Exported to DDI-C XML for CESSDA CDC
  - (was exported to another DDI-C for ZACAT Nesstar)
  - Exported to DDI-L for DDIFlatDB
    - Used e.g. by ExploreData and CESSDA EQB
  - Exported to da|ra for DOI metadata
  - Indexed by Solr Search for GESIS Search

Example applications

DBKEdit 1.9.2 Menu

Studienbeschreibung 3950

Studie dt.	Studie en.	Entwurf dt.	Entwurf en.	Vorspann dt.	Vorspann en.	vorige (3949)	aktuelle (3950)	nächste (3951)	Logdatei	XML	Vorschau in DBKSearch	Vorschau (intern)	Druckansicht

3950: International Social Survey Programme: Citizenship - ISSP 2004

ZA-Studiennummer	3950	
Titel	International Social Survey Programme: Citizenship - ISSP 2004	Bearbeiten
Weitere Titel		Liste
Produktzuordnungen		Liste
Gruppenzuordnungen	▪ GN0001: International Social Survey Programme (ISSP) ▪ GN0136: ISSP - Module Topic: Citizenship	Liste
FDZ-Zuordnungen	▪ FDZ0003: Internationale Umfrageprogramme (FDZ Int. Umfrageprogramm bei GESIS)	Liste
letzte Version	▪ **1.3.0**, 26.06.2012, Update: correction of errors (see changes in this version) Update: correction of errors (see changes in this version) brien 10.4232/1.11372 (Version wurde bereits registriert) (Metadaten zu dieser Version sind aktuell.)	Liste
Erhebungszeitraum standardisiert/frei	▪ 03.2003 - 02.2006 Australia: 07. September 2005 to 29. December 2005 Austria: November to December 2004 Brasil: 05. January 2006 to 29. January 2006 Bulgaria: 15. July 2005 to 26. July 2005 Canada: 29. January 2004 to 31. March 2004	Liste

Example applications

International Social Survey Programme: Citizenship - ISSP 2004

Wilson, Shaun; Gibson, Rachel; Hadler, Markus [+]

GESIS Data Archive, Cologne. ZA3950 Data file Version 1.3.0, https://doi.org/10.4232/1.11372

Abstract: "Citizen and state. Topics: Qualities of a good citizen; attitude toward the right of public meetings for religious extremists, people who want to overthrow the government by force, or people prejudiced against any racial or ethnic group; social and political participation; memberships; importance of different people`s rights in democracy (scale); estimation of political influence possibilities (political efficacy); likeliness of counter-action against an unjust law and expected chance of serious attention to people`s demand; interest in politics; personal trust in politicians and people;..." more

Principal Investigator/ Authoring Entity, Institution: Wilson, Shaun - Macquarie University, Sydney, Australia| Gibson, Rachel - ACSPRI Centre for Social Research (ACSR) Research School of Social Sciences, The Australian National University, Canberra, Australia| Hadler, Markus - Institute of Sociology, University of Graz, Austria| Hoellinger, Franz - Institute of Sociology, University of Graz, Austria|... more

Publisher: GESIS Data Archive

Study number: ZA3950

Current Version: 1.3.0, 2012-06-26, https://doi.org/10.4232/1.11372

DOI: 10.4232/1.11372| 10.4232/1.10078

Publication year: 2012

Date of Collection: 03.2003 - 02.2006

Geographic coverage: Australia| Brazil| Chile| Bulgaria| Denmark| Germany| Finland| Vlaams Gewest| France| Great Britain| Ireland| Israel| Japan| Canada| Latvia| Mexico| Netherlands| New Zealand| Norway| Austria| Philippines| Poland| Portugal| Russian
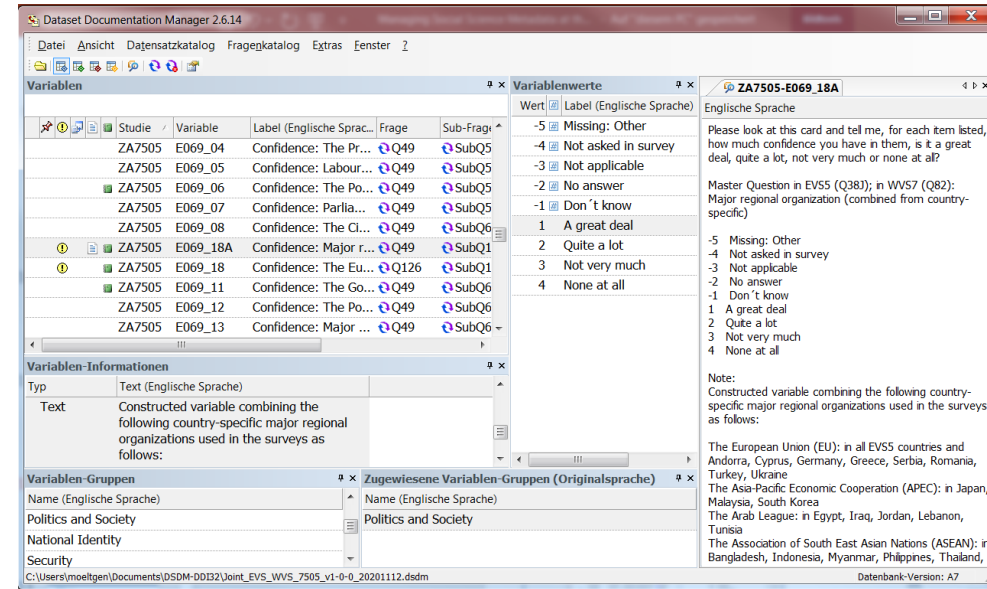
Materials

Datasets

Questionnaires
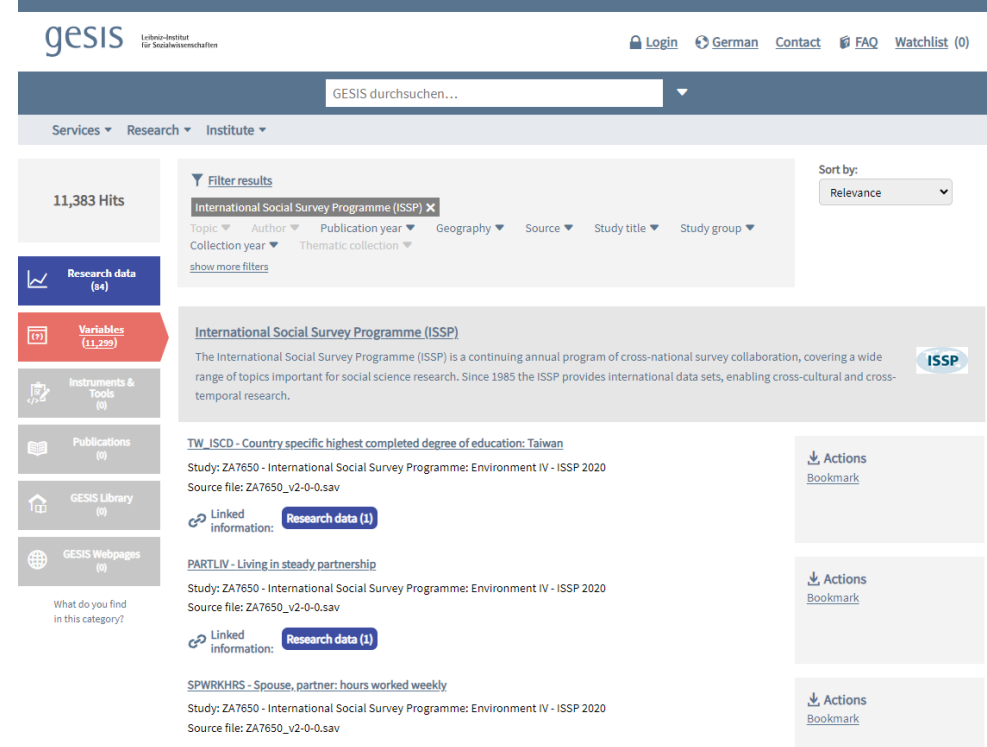
Codebook

Other documents

Actions

Cite

14

Example applications

- Variable and Question Documentation
  - Created in DSDM, saved in MS Access
  - Exported to CodebookExplorer for VariableReports
  - (was exported to DDI-C for ZACAT Nesstar)
  - Exported to DDI-L for DDIFlatDB
    - Used e.g. by ExploreData and CESSDA EQB
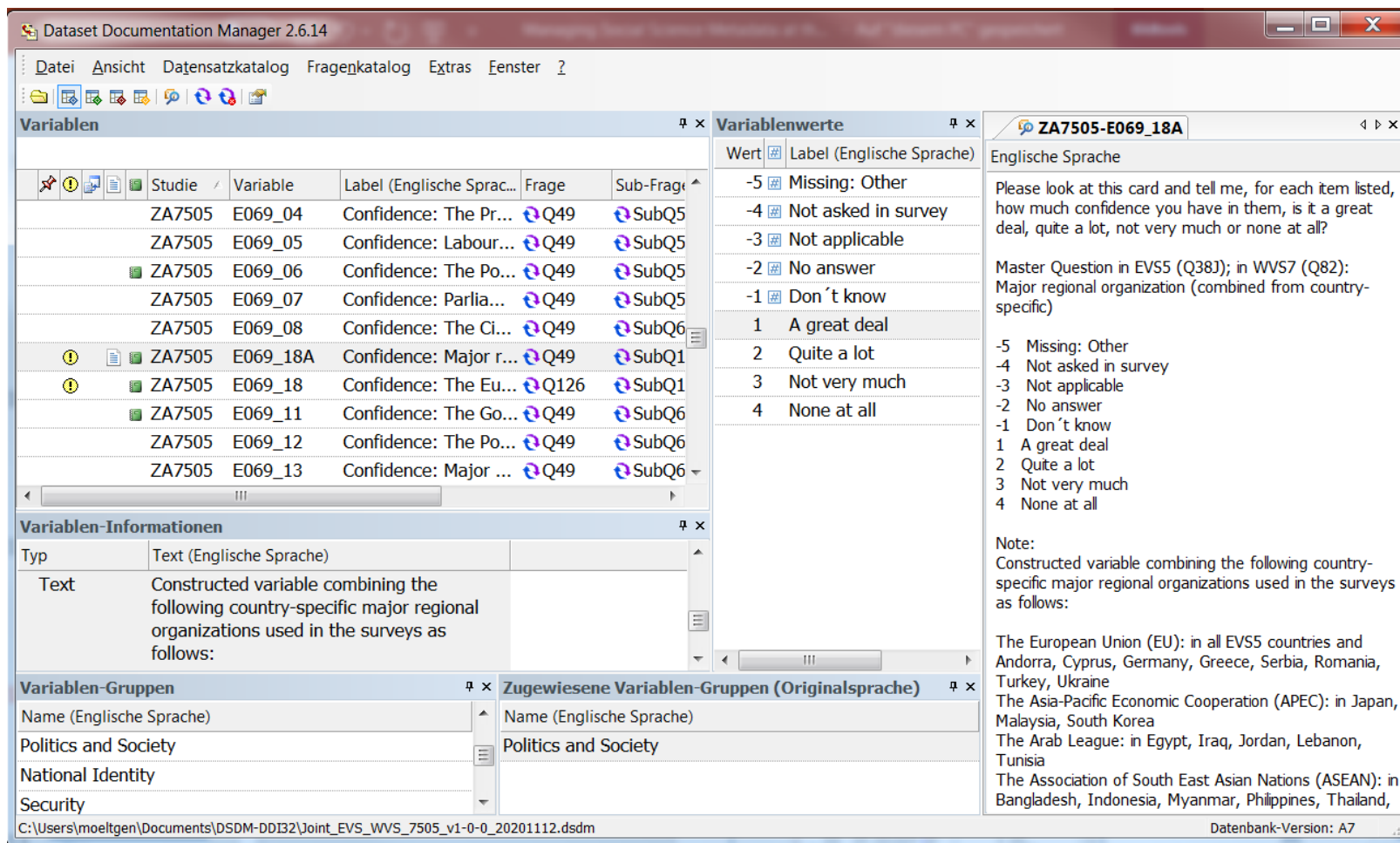   - Indexed by ElasticSearch for GESIS Search

Example applications

http://dbkapps.gesis.org/software/dsdm/

Example applications

https://search.gesis.org/

Current state of workflows

Portals for end-users

| da|ra & DataCite | GESIS Search | SowiDataNet |datorium | ZACAT & ExploreData | CESSDA CDC | Question naires |

common metadata schema (DDI-L with CVs) implemented in DDIFlatDB

| DBKEdit | DSpace | DSDM | QMD |

Studies   Datasets and Variables   Questions
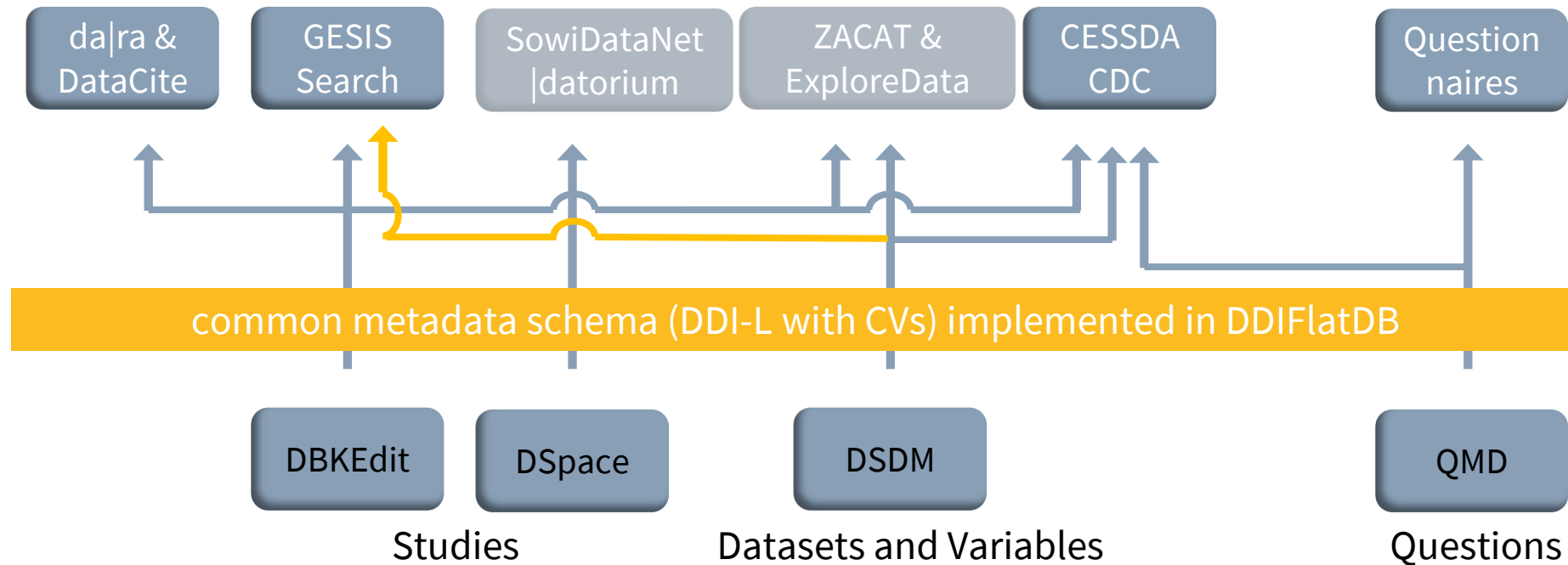
Metadata production tools

Current state of workflows

- Using a common metadata store to feed many applications
- Using legacy editor tools to fill this store, but enabling a stepwise migration of these
- Integrated search and access for end-users
- Special functions may still require extra portals
- Support for partner portals where necessary
- Limited support for different metadata versions
- We will introduce Colectica (DDI-L software) for integrated study and variable documentation

Challenges for metadata applications

- Higher amounts of data

- Wider variety of data

- Context of data collection more often inter-disciplinary

- More exchange of data and metadata

- Multilinguality of metadata

- Expectations of fast provision and access

- Need to comply with various standards

- Meet the F.A.I.R Criteria (https://www.go-fair.org/fair-principles/)

Digital Behavioral Data

- Digital Behavioral Data (DBD) encompasses digital observations of human and algorithmic behavior

- Examples:
  - Social media data
  - Web tracking data
  - Geolocation data

- Full collection here: https://www.gesis.org/en/services/finding-and-accessing-data/digital-behavioral-data-datasets

Challenges in describing DBD

- GESIS focus on survey data for decades

- Some metadata **not relevant** for DBD

- Adaptation to DBD often difficult: e.g., collection period ≠ content period

Challenges in describing DBD

- Some metadata that **is relevant** for DBD is **missing**

- E.g., data collection mode →Controlled Vocabulary like API and Web Scraping

Data collection mode

Select option   [+]

- ▼ Interview
  - ▸ Face-to-face interview
  - ▸ Telephone interview
    - E-mail interview
    - Web-based interview
  - ▸ Self-administered questionnaire
  - ▸ Focus group
  - ▸ Self-administered writings
  - ▸ Observation
  - ▸ Experiment
    - Recording
    - Content Analysis
    - Transcription
    - Compilation/Synthesis

Metadata for Digital Behavioral Data

Examples of metadata need to describe DBD:

- User ID/Identifier
- Location
- Device Information
- Platform information
- Data collection period
- Mode of data collection
- …

CVs for DBD

- ## Controlled Vocabularies:

A list of clearly defined, mutually exclusive, and exhaustive terms, which are the only choices for usage in a specific context.

- ## E.g. :

CV code name	Value codes
AggregationMethod	Sum ArithmeticMean Count Mode Median Maximum Minimum Percent CumulativePercent Other

Controlled Vocabularies for DBD

- New Controlled Vocabulary, e.g., for Platform:
    Youtube

    Wikipedia

    Telegram data

    ChatGPT data

    Reddit Data

    Twitter/X…

- Use latest version of DDI CV for ModesOfCollection
    Web Scraping

    API

Conclusions

We need to:

- Use the common metadata store to feed all the remaining applications

- Evaluate state-of-the-art editor tools to manage the metadata workflow

- Do even more integration for end-users

- Complete the concept for managing different versions of CVs

- Include the conversion functions into the concept, assuming that there will be ongoing change in the metadata schemata and CVs

Conclusions

You need to manage

Conversions · Content · Standards · Software · Workflows
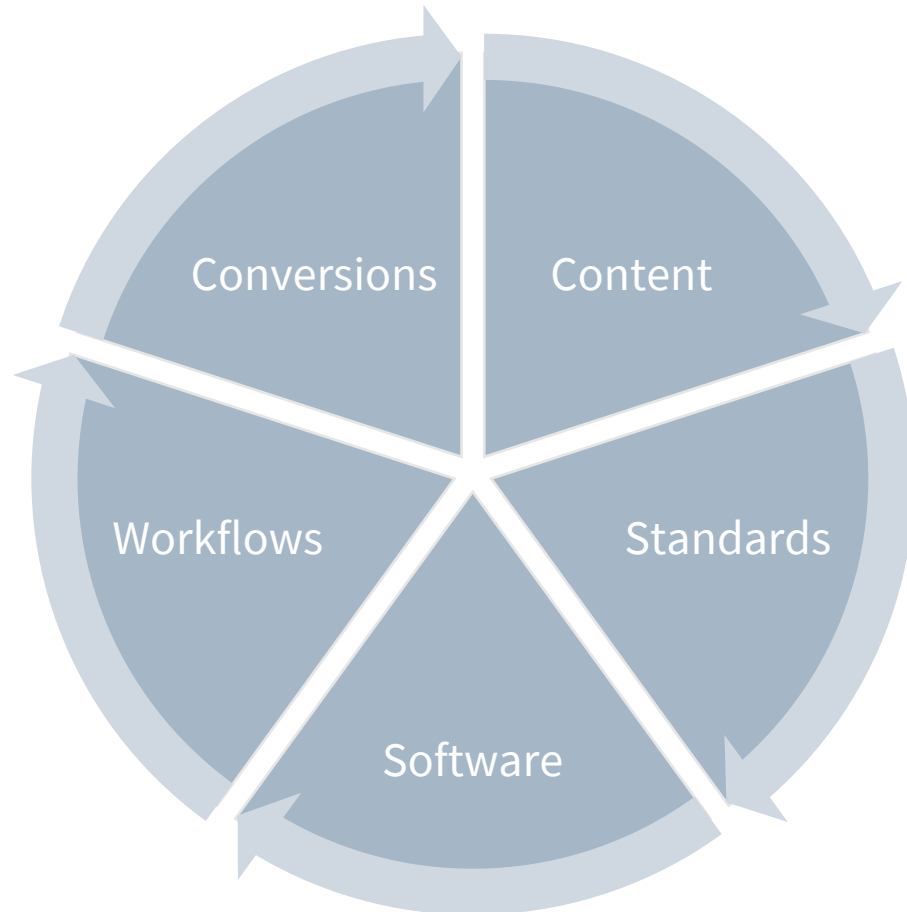
… to meet the challenges of more and heterogeneous data types

- Extended content
- Additional standards
- More exchange
- More languages
- Fast access

Expert contacts & GESIS consulting

Contacts: you can reach the speakers via e-mail:

Wolfgang.Zenk-Moeltgen@gesis.org

Jan.Schwalbach@gesis.org

Kokila.Jamwal@gesis.org

GESIS Consulting: GESIS offers individual consulting in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle.

Please visit our website www.gesis.org for more detailed information on available services and terms.

Data Services, Data Archiving, and Research Data Management

Certification	14.09.2023	**Certification for data archives and research data centers** *Jonas Recker, Kerstin Beck*
Data Services	28.09.2023	**Data Services: An Overview** *Oliver Watteler*
Community Data Collection	12.10.2023	**Community Data Collection** *Alexander Jedinger*
Metadata & PIDs	26.10.2023	**Metadata and PIDs** *Wolfgang Zenk-Möltgen, Jan Schwalbach & Kokila Jamwal*
DP-R|EX	09.11.2023	**An Introduction to Domain-Specific Data Infrastructures: DP-R|EX** *Alexander Jedinger, Marlene Hilgenstock & Pascal Kolkwitz-Anstötz*
RDM & RDM Training	23.11.2023	**Research Data Management & RDM-Training** *Anja Perry & Sebastian Netscher*
Risk & Sensitive Data	07.12.2023	**Secure Data Center: Secure Access to Sensitive Data** *Deborah Wiltshire & Jara Kampmann*

https://www.gesis.org/en/services/sharing-knowledge/consulting-and-guidelines/meet-the-experts

Thank you for participating!

Questions?