Managing Social Science Metadata at the Data Archive

EDDI 2020 – Virtual Conference, 1-2 December 2020

Wolfgang Zenk-Möltgen

GESIS – Leibniz Institute for the Social Sciences, Germany

Outline

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

Introduction

Research data (6,592)

- GESIS Data Archive has predominatly survey data: 6592 research data sets

- Mostly from Germany and other European Countries

- With typical topics from Sociology and Political Science

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)

GESIS-DBK (6,266)

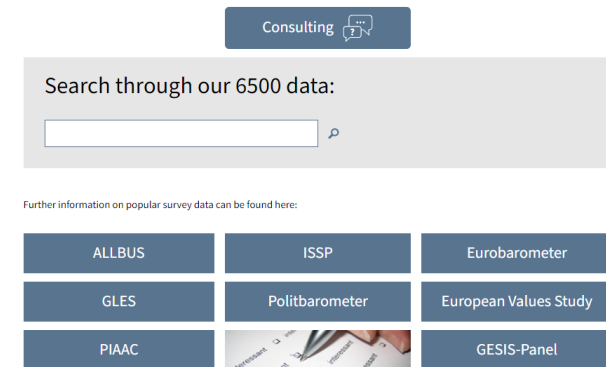GESIS-SowiDataNet|datorium (234)

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...

Introduction

- Data are coming more and more from other sources than surveys
- In very diverse data formats
- Examples are
  - Social media (e.g. Twitter, Facebook)
  - Geographic or spatial data
  - Free-text (from open ended questions, but also from party programs or other texts)
  - Code for data preparation or analysis
  - Link collections

Introduction

- **Tasks of the Data Archive include**
  - Creating study descriptions
  - Offer data deposit and access
  - Help researchers to create documentation
  - Create complex documentation
  - Give advise 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

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
- Multi-linguality of metadata
- Expectations of fast provision and access
- Need to comply with various standards

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
- OpenAire, DCAT
- Sometimes applied several times through chained workflows

Topic Classifications

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

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
  - ▶ 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
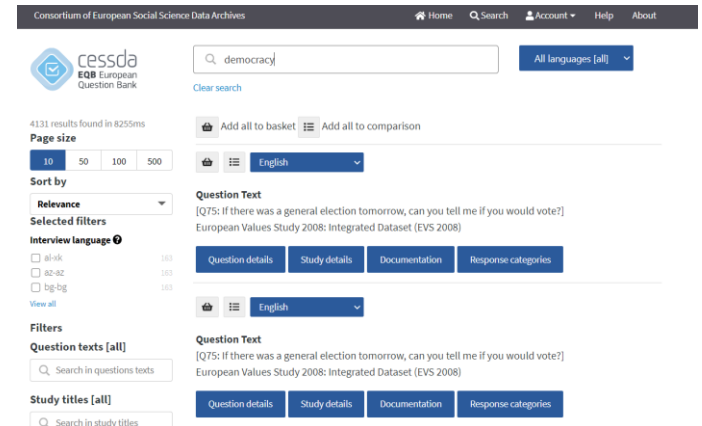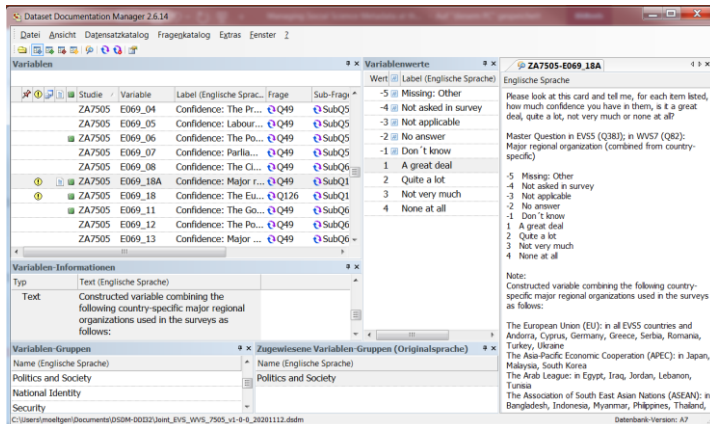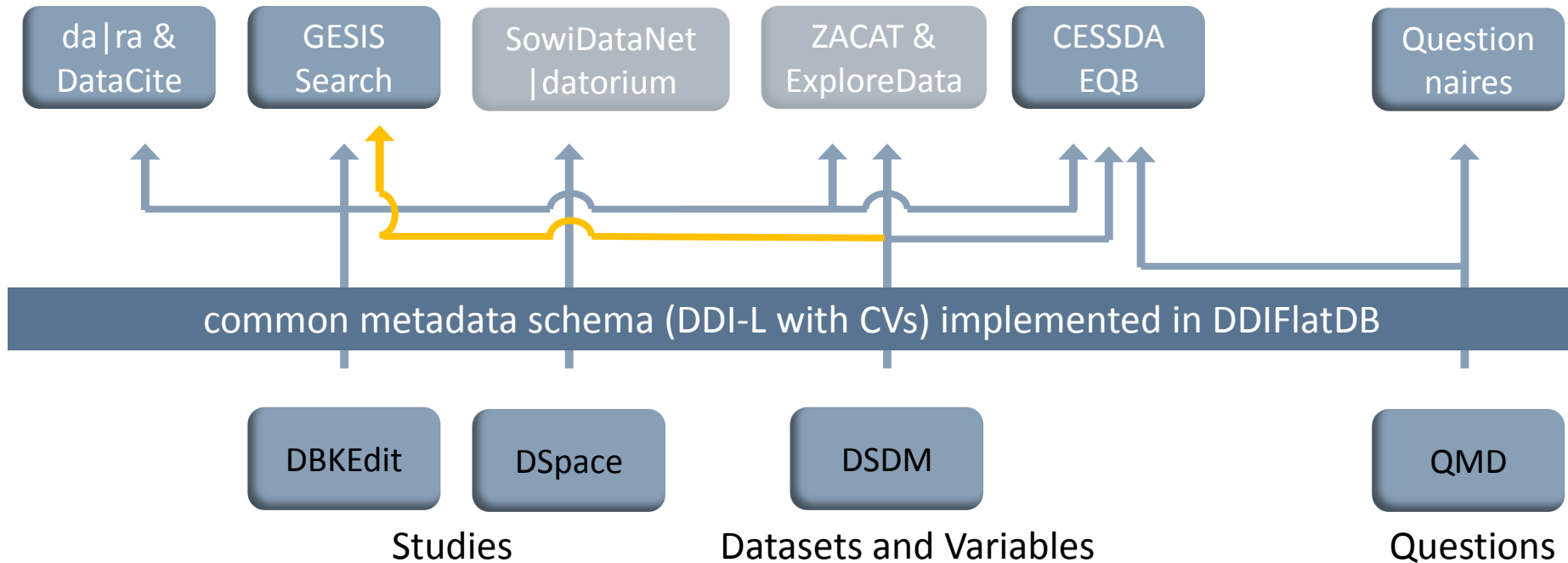
Example applications

- **Variable and Question Documentation**
  - ▶ Created in DSDM, saved in MS Access
  - ▶ Exported to CodebookExplorer for VariableReports
  - ▶ 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

Current state of workflows

Portals for end-users

da|ra & DataCite	GESIS Search	SowiDataNet |datorium	ZACAT & ExploreData	CESSDA EQB	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
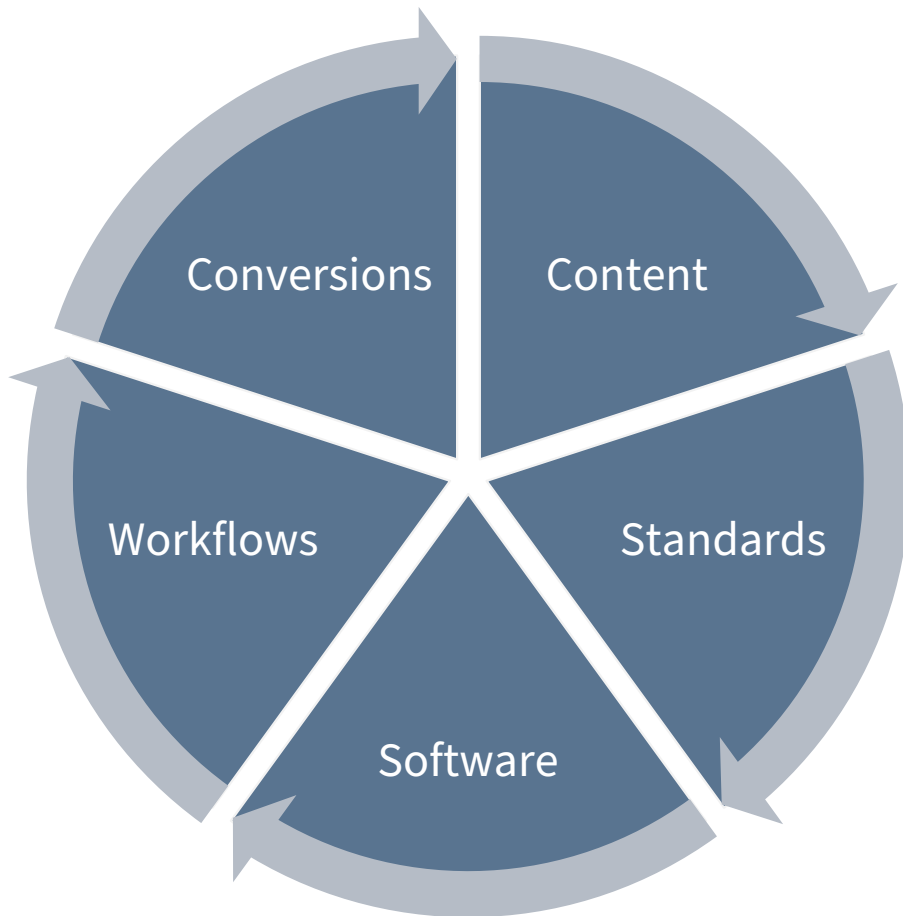- Limited support for different metadata versions

Conclusions

We need to do

- 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

Content
Standards
Software
Workflows
Conversions

… to meet the challenges of more and heterogenious data types

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

Thank you! Questions?

gesis   Leibniz Institute
         for the Social Sciences

Leibniz
Association

Contact: Wolfgang.Zenk-Moeltgen@gesis.org