Connecting GESIS research data and publication information systems

Katarina Boland
Department Knowledge Technologies for the Social Sciences
GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

OpenAIRE Interoperability Workshop, Braga, Portugal
08.02.2013
Outline

1. Introduction to GESIS
2. GESIS information systems
3. Linking publications to datasets
4. Connecting research data and publication information systems
GESIS

- largest infrastructure institution for the Social Sciences in Germany
- five scientific departments (Mannheim, Cologne, Berlin)
Connecting GESIS research data and publication information systems
GEISIS
Publications

- SSOAR - Social Science Open Access Repository
- SOWIPORT - the social science portal
Publications

SSOAR - Social Science Open Access Repository

electronic full texts (Social Sciences) for free access mainly pursues the Green Way of Open Access

http://www.ssoar.info/en.html

SOWIPORT - the social science portal
Publications

- SSOAR - Social Science Open Access Repository
- SOWIPORT - the social science portal

approximately 7 million references on publications and research projects from 18 databases

additional information on institutions and events

http://www.gesis.org/sowiport/en/
Research Data

detailed information, documentation on variable-level and beyond

- ZACAT - Online Study Catalogue
- MISSY - Microdata Information System

documentation on study-level

- da|ra - Registration agency for social and economic data
- DBK - Data Catalogue
Research Data

- da|ra - Registration agency for social and economic data
- DBK - Data Catalogue
Research Data

- da|ra - Registration agency for social and economic data

German DOI registration service for social science and economic data by GESIS and ZBW - German National Library of Economics, in cooperation with DataCite

http://www.da-ra.de/en/home/
Research Data

- da|ra - Registration agency for social and economic data
- DBK - Data Catalogue

**Data Catalogue - Simple Search**

The Data Catalogue comprises the study descriptions from all archived studies and empirical primary data from survey research, historical social research and texts for content analyses. The documentation of data from official statistics will be stepwise added.

- **Basic Search - Advanced Search**
- To highlight search terms and to use multiple selection, please activate JavaScript
- You can browse the complete list of studies available at the Data Archive here: Study List.
- You can browse a list of available products here: Product List.
- Order & Download is now possible directly from within the data catalogue.

**GESIS Quicklinks**

Choose page...

**New Data**

- Eurobarometer 76.4 15. January 2013 (2011) (ZA5596) available
- Demographic Differences in Life Course Dynamics in Eastern and Western Germany? (ZA5584) available
- GESIS Online Panel Pilot Family Life (survey 3) (ZA5584) available
- How innovative is Germany? (ZA5639) available
- Views on European Unification (ZA5637) available

**Terms of use**

- Charge regulations
- Usage regulations

study descriptions from survey research, historical social research and texts for content analyses
documentation of data from official statistics will be added successively

Outline

1. Introduction to GESIS
2. GESIS information systems
3. Linking publications to datasets
4. Connecting research data and publication information systems
Linking publications to datasets

the InFoLiS project:
Integration of research data and publications for the Social Sciences

InFoLiS is funded by the DFG (SU 647/2-1)
InFoLiS project goals
Presentation and discussion of the empirical findings. For this purpose, data from the Socio-Economic Panel (SOEP) of the years 1990 and 2003 are used and for both periods, the impact factors are estimated using linear regression models.

data from the <title> of the years <year> are used
References to datasets

Tabelle 1: Bevölkerungsvorausberechnung für Deutschland nach Altersgruppen - Anteile in Prozent

(Databasis: 10. Bevölkerungsvorausberechnung des Statistischen Bundesamtes, Variante 5)

Table 1: Population forecast for Germany depending on age cohorts - proportion in percent.

References to datasets

_Tabelle 1: Bevölkerungsvorausberechnung für Deutschland nach Altersgruppen - Anteile in Prozent_

_(Datenbasis: 10. Bevölkerungsvorausberechnung des Statistischen Bundesamtes, Variante 5)_

_(Data base: <number>.<title> of the <data collector>, version <version>)_
References to datasets


Consulted were furthermore ...
Consulted were furthermore <title1>, <title2>, <title3>, ..., <titleN>.

References to datasets

Table 3: Sample of the surveys conducted in the years 2003 and 2004 as well as size of the sample, with valid data from both surveys

(Source: Ditton et al. 2005a)
References to datasets

Tabelle 3: Stichprobe der Untersuchung in den Jahren 2003 und 2004 sowie Größe der Stichprobe, mit gültigen Daten aus beiden Erhebungen

(Quelle: Ditton u.a. 2005a)

(Source: <citation of descriptive publication>)
Grafik 7: Einschätzung der wirtschaftlichen Lage: Einschätzung der eigenen wirtschaftlichen Lage (in Prozent)

(Quellen: Allbus/Sozialstaatssurvey)

(Sources: Allbus/Sozialstaatssurvey)
Grafik 7: Einschätzung der wirtschaftlichen Lage: Einschätzung der eigenen wirtschaftlichen Lage (in Prozent)

(Quellen: Allbus/Sozialstaatssurvey)

(Sources: <title1>/<title2>)
Linking publications to datasets

References to datasets are not standardized!

see also...


Automatic identification of references

Why not simply search for study titles in publications?

- Studies are referenced using abbreviations, alternative names or literature
- Study titles may be common nouns - ambiguous!
- there is no complete list of all conducted studies
General idea

How do humans recognize study references?

_Source: Estimations based on SOEP, wave 2002._
General idea

How do humans recognize study references?

*Source: Estimations based on xyz, wave 2002.*
General idea

How do humans recognize study references?

Source: Estimations based on xyz, wave 2002.

→ Learn patterns: typical contexts for study references
General idea

How do humans recognize study references?

*Source: Estimations based on xyz, wave 2002.*

→ Learn patterns: typical contexts for study references
→ Sparse Data Problem: use iterative bootstrapping approach
Evaluation: Precision & Estimate of Recall

about 14% of the found references are not study names, but citations of publications → not counted as incorrect here
subset of SSOAR with keyword “empirisch-quantitativ” (empirical quantitative)
German, n = 259
conversion pdf → txt with automatic correction
Reference extraction

for details see...

Matching to da|ra records
Matching to da|ra records

**ALLBUS/GGSS 2000 PAPI (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2000 PAPI)**

- **DOI:** [10.4232/1.3452](10.4232/1.3452)
- **Version:** 1.0.0
- **Principal Investigator:** Allmendinger, Julita (Universität München)
- **Temporal Coverage:** 2000-01 - 2000-07
- **Publication Date:** 2003
- **Availability:** Download

**ALLBUS/GGSS 2000 CAPI (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2000 CAPI)**

- **DOI:** [10.4232/1.3451](10.4232/1.3451)
- **Version:** 1.0.0
- **Principal Investigator:** Allmendinger, Julita (Universität München)
- **Temporal Coverage:** 2000-01 - 2000-07
- **Publication Date:** 2003
- **Availability:** Download

**ALLBUS/GGSS 2000 CAPI-PAPI (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2000 CAPI-PAPI)**

- **DOI:** [10.4232/1.11365](10.4232/1.11365)
- **Version:** 2.0.0
- **Principal Investigator:** Allmendinger, Julita (Universität München)
- **Temporal Coverage:** 2000-01 - 2000-07
- **Publication Date:** 2003
- **Availability:** Download
Matching to da\textit{ra} records

\rightarrow Precise matching to DOI not always possible!
\rightarrow Instead: Matchings to \textit{relevant} sources
Matching to da|ra records

→ Precise matching to DOI not always possible!
→ Instead: Matchings to relevant sources
→ Definition of relevance depends on application
Matching to da|ra records

→ semantic web technologies
Links
Connecting information systems

Service I: InFoLiS

Services II, III & Architecture:
Dennis Wegener,
Daniel Hienert,
Dimitar Dimitrov
(SOWIPORT, da|ra)

Connecting GESIS research data and publication information systems
Connecting information systems

Demo: da|ra test system with SOWIPORT links
Connecting information systems

Demo: da|ra test system with SOWIPORT links (offline version)
Conclusion: our aim

- interlink our own repositories and information systems
- provide services for reference extraction and matching to all interested institutions (free access to webservices)
  - domain- and language-independent
- link to publications and data stored in external repositories
Thank you for your attention!

katarina.boland@gesis.org