Distributed Archiving of Social Science Research Data: On the Way to Best-Practice Guidelines

Reiner Mauer, Oliver Watteler
GESIS, Data Archive

Presentation prepared for IASSIST 2013 (Cologne)
Aim of this presentation

• Challenge well known:
  – Archive A has something,
  – Archive B has something.
  • Both holdings “belong together”!

• Organizational and technical solutions available.

• **BUT** intellectual input still necessary!
  Keep creation contexts coherent for third parties.

Work in progress: (use case) distributed archiving of qualitative and quantitative data.

• One aim of the project: create living example for further cooperations of the same type.
Distributed archiving - intro

• What is research data?
• What is distributed archiving?
• Two working examples:
  – sowiport (GESIS)
  – CESSDA catalogue
• Cooperation QualiService (Bremen) and GESIS
(A) Research data?

• Research data can be ...
  – "... **reinterpretable representation of information** in a formalized manner suitable for communication, interpretation, or processing.“ [OAIS]
  – "... **recorded factual material** commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This ‘recorded’ material excludes physical objects ... (OMB 1999)”
  – "... representation of a part of societal reality which can be **used to analyse social phenomena** (Watteler 2012:125).“
(A) Distributed archiving

- Archiving „... is what archives do“ *(based loosely on 'The Little Prince')*
- Archive [OAIS] = „organization, […] that has accepted the responsibility to preserve information and make it available for a Designated Community (CCSDS 2012:1-1).”
- Distributed archiving „... is what two or more archives do in a joint effort to reach this goal.“
- Joint effort = Archiving information packages in more than one place while preserving the context of origin.
(A) Making up for downsides of „distribution“

• „Tearing apart“ of provenance!
• Several data holding institutions involved!
  → Communication for coordination (e.g. rights management, documentation ...)
  → Possibly, differing access regimes!
  → ...

Solution

• „Conjoint access structure“
(A) Working example:

Type of information
- Publications (2632)
- Studies (63)
- Research projects (20)

Database
- GesisBib (1521)
- SOLIS (495)
- SSOAR (427)
- Soc.Abs. (74)
- FIS (65)

Person(s)
(A) Working example: sowiport

- Working example from a Social Science data infrastructure:

<table>
<thead>
<tr>
<th>PRO’s</th>
<th>CON’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows for search across multiple holdings.</td>
<td>„Classical“ database system: needs import.</td>
</tr>
<tr>
<td>Controls for variations using cross-concordances.</td>
<td>Does not group results.</td>
</tr>
<tr>
<td>Faceted search possible.</td>
<td></td>
</tr>
</tbody>
</table>
(A) Working example: CESSDA
(A) Working example: CESSDA

- Working example from the world of Social Science data archives

<table>
<thead>
<tr>
<th>PRO’s</th>
<th>CON’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows for search across multiple holdings.</td>
<td>No ex post filtering of results possible.</td>
</tr>
<tr>
<td>Uses European Language Social Science Thesaurus (ELSST).</td>
<td>No sorting of results.</td>
</tr>
<tr>
<td>Search on various levels (studies, variables ...).</td>
<td>No „connection“ of distributed holdings possible.</td>
</tr>
<tr>
<td>Links directly to data holdings.</td>
<td></td>
</tr>
</tbody>
</table>
(B) Special cases

- National data-sets that actually belong together
- Variations of data-sets in different places
- Qualitative and quantitative data units in one project context
(B) „National“ data-sets that actually belong together

- How do you best document international project contexts if data-sets are preserved in national archives?
- “Classic” problem
- Example: TIES project ("The Integration of European Second Generation")
  - 11 countries
  - Dutch data at DANS
  - German data at GESIS
  - Others?
- No ex ante planning of archiving.
- No technical solution yet.
(B) Varieties of data-sets in different places

- How do we learn about the varieties in data holdings?
- Example: German Micro Census (annual survey; 1% sample of all households in Germany)
  - Scientific Use File at RDC
  - Public Use File for Order
  - Campus File for free download
  - Remote Access e.g. at IZA, Bonn
  - Metadata e.g. to set up trend files at GESIS
(B) Varieties of data-sets ...
possible solution
(B) Qualitative and quantitative data units in one project context

- How do you connect qualitative and quantitative data from the same project which are archived in different locations?
- Project partners: QualiService and GESIS
- Challenge: Research projects including qualitative and quantitative data
- Working example: „Migration and societal integration“
(B) qualixservice and gesis

• Brief history of a cooperation
  – Foundation Archive for Live Course History (ALLF), Bremen, 2001
  – Feasibility study for full qualitative archive, 2001-2003
  – Transfer of ALLF holdings to QualiService, 2011

• Why this cooperation?
  – GESIS has strong commitment to quantitative social research.
  – Each data type needs expertise and tacit knowledge!
  – Archiving data often a matter of trust by peers!
  – Cooperation means optimal combination of forces, BUT also coordination of efforts!
(B) Challenge: Research projects including qualitative and quantitative data

- Combination of qualitative and quantitative research designs of growing importance (e.g. mixed-methods).
- Distinction:

<table>
<thead>
<tr>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory Building</td>
<td>Testing</td>
</tr>
<tr>
<td>Main Method</td>
<td>Mostly open forms of interviewing</td>
</tr>
<tr>
<td>(Individual Level)</td>
<td>Mostly closed forms of interviewing, but</td>
</tr>
<tr>
<td>Data</td>
<td>Mainly transcripts</td>
</tr>
</tbody>
</table>

- Example for combination: use of open interviews to come with theory which is later tested in quantitative fashion.
(B) Working example in the making

• Aim: societal integration of Jewish citizens from former S.U.
• Comparison with other migrant groups in new home countries.
• 5 projects, one „roof“
  – 3 quantitative
  – 2 qualitative
Footnote: Of parents and children

„Father“ project

Research Consortium

„Children“

Project a

Project b

Project c

Project d

Research Data

Quali a1

Quanti a2

Quali b1

Quanti b2

Quanti b3

Quanti c1

Quali d1
(C) The aim: best-practise guideline

Distributed archiving means ...

• Coordinate acquisition contacts!
• Coordinate rights management and access regimes!
• Define common metadata!
• Work out grouping function in each archival catalogue!
• Work out linking mechanism between groups!
• Jointly document project framework!
• Keep ALL the „stuff“ alive!
(C) The workflow so far

- New data available
- Who takes care of what?
- Ingest QualiService
- Ingest GESIS
- Documentation QualiService
- Documentation GESIS
- Linking of documentations
- RELEASE
(C) Put the pieces together (Step 1)

- First three section also needed in „regular“ archiving.
- E.g. ISSP:
  - 27 waves
  - Program description

\[\text{AIP} = \text{Archival Information Package [OAIS]}\]
(C) Put the pieces together (Step 2)
(C) Put the pieces together (Step 3)
Put the pieces together (No plan yet)

But wait! If you use DDI, why not go the LOD way?

< Distributed Holdings>

Release Metadata into LOD cloud ...
Thank you!
Any questions?

reiner.mauer@gesis.org
oliver.watteler@gesis.org
References