Data Management, DDI-based Documentation and Visualization of Business and Organizational Research Data at the DSZ-BO

Johanna Vompras

University Library Bielefeld

Datenservicezentrum Betriebs- und OrganisationsDATEN

Dec 4th, 2012
EDDI2012 – Bergen, Norway

Session B1: Infrastructure for Data Collection, Research, and Archiving
Contents

1 DSZ-BO
2 Data Infrastructure
3 Technical Solutions and Tools
4 Summary
You have never heard about 'Bielefeld'???

Population: 323,000

Location:
in the east of North Rhine-Westphalia (Ost-Westfalen)

City naturally divided by the Teutoburg Forest

Famous companies:
Dr. Oetker, Seidensticker, Schüco
Contents

1 DSZ-BO

2 Data Infrastructure

3 Technical Solutions and Tools

4 Summary
Task and Scope of the DSZ-BO

Collect, archive, distribute and maintaining a catalogue of Business and Organizational Data from the Social Sciences, like ...

- Surveys with multiple organizations, e.g. interviews with human resource managers of different firms,
- Qualitative case studies and mixed methods,
- Process generated numbers, e.g. average time of patients in different hospitals, business catalogues,
- Observations, e.g. informal processes in one local office,
- Linked employer employee data (LEE).
DSZ-BO Services

Data Catalogue
- General Information
- Detailed Metadata
- Discovery Services

Data Producers
- Data Registration
- Archiving
- Data Sharing

Data Access
- Secondary Use

Data service center for Business and Organizational DATA (DSZ-BO)

Data Users

Support & Training
- Data Management
- Documentation
- Data Protection, Anonymization
Contents

1 DSZ-BO
2 Data Infrastructure
3 Technical Solutions and Tools
4 Summary
Specific Requirements

- **Standardized Documentation:** Mapping of "complex" study structure into DDI:

  - Study Structure:
    - BEATA
    - Emloyer inquiry
    - Quant. Survey
    - Qual. Interview
    - Document analysis
    - Employee survey
    - Partner survey

- **Data Catalogue:**
  - Data with certain methods,
  - Data which contains variables that operationalize certain research questions,
  - Datasets with certain levels of analysis,
  - Examples for good practice in a certain field.
Co-Operation DSZ-BO and Library Services

- Data access
- Distribution
- Data sharing
- Long-term archiving
- Authentication
- Indexing
- Retrieval
- Privacy policies

**Functionalities**

**Internal Representation**
- Metadata-format: DDI 3.1
- XML schemes
- XML database(s)

**Study**
- Coding into DDI 3.1, semi-automatically
- Textual information, publications, surveys

**Data**
- Study: data collection questions, variables
- Data sets in SPSS/STATA, etc.

**Meta- and Microdata**
- Relation to business or organizational unit
  (e.g. linked employer-employee data)

**Services**
- Development of the technical infrastructure
- Data cleaning
- User support/help desk (editing, access, training, etc.)
- Content maintenance
- Data privacy policies

**Co-Operation**

**University Library**
- Support of the data lifecycle
  - Multiple-language support

**Faculty of Sociology**
- EDDI2012, December 3–4 2012, Bergen, Norway
Co-Operation DSZ-BO and Library Services

Data re-use, study comparison, support of the **data lifecycle**
multiple-language support

meta- und microdata
relation to business or organizational unit
(e.g. linked employer-employee data)

Data Infrastructure
Technical Solutions and Tools

Specific Requirements
DSZ-BO as a part of Institutional Infrastructures

University Library
Faculty of Sociology

Johanna Vompras
EDDI2012, December 3–4 2012, Bergen, Norway
Documentation or putting ”research” into a DDI ...

Study Investigation
- Expert Knowledge
- Systematic Literature Search and Selection
- Acquisition of Data from Data Producers

Study Classification
Standards: A, B, C
- A: Own Research / Detailed Documentation
- B: Normal Documentation
- C: Minimal Documentation for Cataloguing and Registration purposes

Additional parameters for classification:
Kind of Data, Data and Documentation Quality, Traceability of Research Work, Expert Knowledge, Processing Effort

Study Documentation
using DDI v3

Specific Data Processing and Documentation Workflows

Examples:
- Conversion of SPSS to DDI for variable level documentation
- DDI-XML-Templates for A-Standard documentation
- XML Forms for C-Standard documentation

Johanna Vompras  
EDDI2012, December 3–4 2012, Bergen, Norway
Form template (.xsn) provides a view on XML data → fill it out!

- Form is compressed,
- XML Schemas, default XML data,
- XSLT files for view in the form,
- Script files and form definition files.

# Tool 1: Editor for C-Standard Studies

## Form template (.xsn)

### Interne Infos

<table>
<thead>
<tr>
<th>Author:</th>
<th>md ◦ bitte auswählen!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datenbankname (BaseX-XML)</td>
<td>Ausf_s0063</td>
</tr>
<tr>
<td>Zeitstempel 1. Eintrag (*)</td>
<td>03.02.2012 ◦ wird automatisch ergänzt!</td>
</tr>
<tr>
<td>Zeitstempel letzte Änderung (*)</td>
<td>03.02.2012 ◦ wird automatisch ergänzt!</td>
</tr>
<tr>
<td>Datenursprung</td>
<td>AuF ◦ bitte auswählen!</td>
</tr>
<tr>
<td>Verwaltungs Stelle</td>
<td>Link auf Akquise DDI</td>
</tr>
<tr>
<td>Ablageort der Daten</td>
<td>Daten im DSZ-BO</td>
</tr>
</tbody>
</table>

### Studiendokumentation in DDI

#### 1. DDIInstance

<table>
<thead>
<tr>
<th>Version</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agency</td>
<td>Universität Bielefeld</td>
</tr>
<tr>
<td>Version Date</td>
<td>2012-02-03</td>
</tr>
<tr>
<td>s0063</td>
<td>DDI_Instance_s0063</td>
</tr>
</tbody>
</table>

#### 2. StudyUnit

| s0063 | z.B. s0001 für Albus |

### 2.1 Allgemeine Studieninformation

<table>
<thead>
<tr>
<th>Title</th>
<th>Telekom. Wie machen die das? Die Transformation der Beschäftigungsverhältnisse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurztitel/Abkürzung</td>
<td></td>
</tr>
<tr>
<td>Projektleitung</td>
<td></td>
</tr>
<tr>
<td>Finanzierendes Institut</td>
<td>DFG - Deutsche Forschungsgemeinschaft</td>
</tr>
</tbody>
</table>
Tool 2: Content Administration Backend

Inspired by researchers workflows for Content Creation and Publishing

Concentration of all processing steps needed to publish the 'DDI-Instance'

- Upload and Archival of the DDI XML file,
- Assignment of internal ID,
- Storage of Metadata and XML Database Generation,
- Preview and Publishing Options.

### Studie bearbeiten

<table>
<thead>
<tr>
<th>Attribut</th>
<th>Wert</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>2</td>
</tr>
<tr>
<td>StudyID</td>
<td>s0002</td>
</tr>
<tr>
<td>Titel der Studie</td>
<td>BEATA</td>
</tr>
<tr>
<td>Standard</td>
<td>A</td>
</tr>
<tr>
<td>Deaktiviert</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribut</th>
<th>Wert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datenbankname</td>
<td>beata</td>
</tr>
<tr>
<td>DDI-Datei</td>
<td>s0002_beata_v2.xml</td>
</tr>
</tbody>
</table>

DDI-Datei für Studie mit der ID 2: [Durchsuchen] [DDI hochladen]

[Im Studienportal publizieren]
Technical Infrastructure: 3 Layers

Requirements:

- Easy to maintain (contents),
- Easy to extend (DDI model and queries)
- General approach for mapping into a visualization
- Data Encapsulation

Information Portal
- Search
- Query Processing
- Visualization

Storage Layer
- XML/SQL Databases
- Other Resources (Tables, Statistical Files, etc.)

Administration Backend (Testing/Productive)
- Catalog Maintenance
- Editorial Work
- Content Management
Data Storage Layer

- Data storage:
  - DDI File: XML Database (BaseX)
  - Other Metadata: Relational Database
- Data Queries/Modification:
  - XQUERY/MySQL
- Data Visualization:
  - JavaScript Framework
  - PHP Script with XQUERY ⇒ Results ⇒ JSON ⇒ Input for Components (e.g. GridPanel or DataView)
Data Retrieval and Visualization

- Each 'Study' is stored in a single XML database
- Queried by XQuery language, e.g. 'select all publications related to the study ALLBUS'
- Results are returned as lists of items and transformed into JSON format

XQuery: find all publications related to ALLBUS study

```xml
FOR $node IN doc("allbus")//s:StudyUnit/r:OtherMaterial[@type='text']
RETURN
  ( CONCAT(
    data($node/r:Citation/r:Creator)," (",data($node/r:Citation/r:PublicationDate),"),",
    data($node/r:Citation/r:Title)),
    data($node/r:Citation/r:PublicationDate),
    data($node/r:ExternalURLReference)
  )
```
Visualization of structured JSON data

Components: Windows, (Tree)Panels, Tabs, (Grouping)Grid

Functions e.g. within GroupingGrid: Sorting and Grouping
Functionality of the Search

Search, Browse, Visualization

- Selection of studies, display of general study information
- Listings of data collections, questions, concepts, etc.
- Linking of data with questionnaires and publications, and other materials
- Search by keywords, or (thesaurus) concepts
- Filtering (e.g. by year, country, standard)
Contents

1 DSZ-BO

2 Data Infrastructure

3 Technical Solutions and Tools

4 Summary
DSZ-BO: Experiences from technical point of view

To find a collection of light-weighted documentation tools tailored for specific scope of studies and data, which are easy to learn and easy to operate,

- Special requirements for afterwards documentation → decision about classification
- Adjustment to researchers’ working workflows:
  - tools,
  - editors,
  - data processing, and
  - content authoring system.
- WYSIWYG-like information portal,
- Easy-to-use content management system, not only for technical staff.
Thank you! Questions?

Contact

Bielefeld University Library
Research Data Management Services and Infrastructure Projects
Johanna Vompras

johanna.vompras@uni-bielefeld.de

Data Service Center for Business and Organizational Data
dsz-bo@uni-bielefeld.de
Appendix: XQUERY Query Generation

Table *ddi_path*:

<table>
<thead>
<tr>
<th>id</th>
<th>attribute</th>
<th>label</th>
<th>path</th>
<th>root</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>title</td>
<td>Titel</td>
<td>data($node/r:Citation/r:Title)</td>
<td>s:StudyUnit</td>
</tr>
<tr>
<td>2</td>
<td>creator</td>
<td>Erstellt von</td>
<td>$node/r:Citation/r:Creator</td>
<td>s:StudyUnit</td>
</tr>
<tr>
<td>3</td>
<td>funding</td>
<td>Gefördert durch</td>
<td>for $knoten in doc($database) ...</td>
<td>s:StudyUnit</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>studyresults</td>
<td>Ergebnisse</td>
<td>NULL</td>
<td>d:DataCollection</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>realization.sampling_notes</td>
<td>Anmerkungen</td>
<td>$node/r:Note/r:Content</td>
<td></td>
</tr>
</tbody>
</table>

Example ’Funding’:

FOR $knoten in doc($database)//a:Archive/a:OrganizationScheme/a:Organization
WHERE $knoten/@id = $node/r:FundingInformation/r:AgencyOrganizationReference/r:ID
RETURN
concat(data($knoten/a:OrganizationName),utilities:ifexistsPar(data($knoten/a:Nickname)))

corresponds to the relational statement:

\[ \pi(a:OrganisationName) \]

\[ (a:Archive/a:OrganizationScheme/a:Organization \bowtie@id=r:ID \]

\[ r:FundingInformation/r:AgencyOrganizationReference ) \]