Structuring unstructured data for business, economic and related research

Anja Busch¹, Ulrich Krieger², Atif Latif¹, Fidan Limani¹, Irene Schumm², Ahmed Saleh¹

¹: ZBW – Leibniz Information Centre for Economics; ²: University Library, University of Mannheim

July 2021
Our understanding of individual and social behavior is currently significantly expanded due to the availability of new data types.
## Use Case: Unemployment Research

<table>
<thead>
<tr>
<th>1930’s</th>
<th>1980’s</th>
<th>Since 2010’s</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="1930’s Image" /></td>
<td><img src="image2.jpg" alt="1980’s Image" /></td>
<td><img src="image3.jpg" alt="Since 2010’s Image" /></td>
</tr>
</tbody>
</table>

### 1930’s
- **detailed**
  - observer error
  - small scale
  - no inference

### 1980’s
- **standardized**
  - large scale
  - inference
  - expensive
  - high burden
  - misreports

### Since 2010’s
- **standardized**
  - large scale
  - inference
  - cheap
  - low burden
  - complex post-processing
  - tools and infrastructure lacking

*Source: Archives for the History of Sociology in Austria (Graz), »Marienthal« Virtual Archives, ISR Archive, IAB SMART Study, Kreuter et al.*
Challenge of the New Analytical Paradigm

Survey and Experimental Design
*High effort prior to data collection*

Automated Data Recording
*Low effort prior to data collection*

Low-Dim. Measures

Survey Data

Text, Language

Images

IOT / High-Dimensional (Dirty) Digital Trace

Standard Processing
*Low effort post data collection*

Complex Processing and Analysis
*High effort post data collection*

Linkage

BERD@NFDI in a Nutshell
New Enriched Model of Social Science Research ...

TRADITIONAL MODEL OF EMPIRICAL RESEARCH

Structured and standard data

Empirical Methods for causal analysis and prediction

Standard (traditional) sources

Scientific progress from discovery and new knowledge generation
• Abundant complex data and data types: Huge potential for exciting discoveries and social gains
• "Methodological" costs much higher in analysis
• Interwoven with technical burden
• Risk of misleading and irreproducible results
BERD@NFDI Structure

- Einbindung der Community & wissenschaftlichen Bibliotheken
- Öffentlichkeitsarbeit
- Benutzerzentrierte Bedarfserhebung
- Impact-Messung

- Education & Training
- Support/Data Stewardship Services
- Feedback Loops
- EOSC-Integration

- FAIRes Datenmanagement
- Suche
- Remote Access & Sichere Datenspeicherung
- Portal
- Persistente Identifikatoren
- Migrationsstrategie

- Datenerfassungstools
- Semantische Anreicherung
- Verlinkung von Daten (z.B. Wissensgraph, NEL/NER)
- Speicher- und Rechenkapazitäten

- Daten-Qualität und Normalisierung
- Anonymisierung
- Prozessierung digitalisierter Dokumente
BERD is an Important Missing Piece of the NFDI

Paradigm shift
from individual analysis and data silos to data and ML on one integrated platform

• **Open**
  Linked unstructured and structured data

• **Fast and accessible computation**
  By cloud-based HPC solution

• **Best practices in ML**
  Platform provides guidance on methods

• **Reproducible and Transparent**
  Documented used data and methods

• **Management of the entire data life cycle**
The BERD Consortium - a Unique Combination

GoFAIR Office
EconBiz
da | ra Agency
GESIS Data Search
TweetsKB
ZBW Leibniz Information Centre for Economics
OCR-D
BERD-ontology
ZEW
Univ. Mannheim
IAB
LMU München
Univ. Hamburg
Univ. Cologne
MCML
Open ML
Cloud & HPC BW
datorium
LRZ Leibniz Supercomp. Centre

BERD@NFDI in a Nutshell
BERD@NFDI in a Nutshell

BERD Flywheel for NFDI

Structured & Unstructured Data

Artificial Intelligence Algorithms

Unique consortium of experienced, strong researchers and leading infrastructure experts
Backup
## Services Roll-out Schedule

<table>
<thead>
<tr>
<th>Task Area</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Area 1</td>
<td>Continuous feedback generation using methods of User-Centered Requirements Engineering; Web and social media presence; Scientific publication on BERD</td>
<td>Dissemination events</td>
<td>Focused crawlers and data harvesters</td>
<td>Upload feature</td>
<td>Evaluation report on impact and success</td>
</tr>
<tr>
<td>Task Area 2</td>
<td>Guidelines and legal consulting services</td>
<td></td>
<td></td>
<td></td>
<td>Ingestion services fully operable metadata normalization tool implemented</td>
</tr>
<tr>
<td>Task Area 3</td>
<td>Recommender service for OCR tools</td>
<td>Initial set of standards for data quality assessment and data normalization</td>
<td>Inventory of German Firm Data and Sources; Integrated OCR-D workflow</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Area 4</td>
<td>Storage and computing capacity set up</td>
<td>Algorithm repositories connected; Initial reporting standards for performance of data analysis tools available</td>
<td></td>
<td>Continuous assessment of algorithms established; Graphical User Interface for BERD ontology</td>
<td>Information extraction from unstructured resources based on BERD ontology</td>
</tr>
<tr>
<td>Task Area 5</td>
<td>Metadata Schema specification; Prototype of search infrastructure; PID service technically integrated; Information portal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Area 6</td>
<td>Self-learning modules for researchers; Training events for researchers and librarians; Consultancy service for research data management; Automated data stewardship services pilot</td>
<td>Support for BERD@NFDI infrastructure</td>
<td>Fully automated data stewardship services; Automated feedback loops</td>
<td></td>
<td>Export and exchange services</td>
</tr>
</tbody>
</table>
Apart from the implementation of new BERD@NFDI services, existing services (from BERD@BW, OpenML etc.) are continued and will be integrated into the BERD@NFDI information portal.
## Training and Teaching

Based on vast experience at Mannheim University and LMU

<table>
<thead>
<tr>
<th>Target groups</th>
<th>Forms of learning</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Researchers</td>
<td>• Self-learning modules</td>
<td>• Generic</td>
</tr>
<tr>
<td>• Data stewards</td>
<td>• On-site workshops</td>
<td>• Data type specific</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• From beginner to specialist level</td>
</tr>
</tbody>
</table>

BERD@NFDI in a Nutshell
BERD@NFDI takes the challenge

Data sources & types
Unstructured and non-standard data from primarily digital sources

Task Area 2
Harvesting
Sharing by users
Access to providers

Task Area 3
Quality assurance and normalization
Anonymization
Processing of digitized documents

Task Area 4
Data analysis for structuring (AI algorithms)
Semantic enrichment and linking

Task Area 5 + 6
Preserving and accessing data and services
Supporting users
STRUCTURED DATA
Characteristics:
- e.g. survey data, administrative data
- high standardization
- homogeneity of sources and formats
- standardized collecting, managing and analyzing
- standardized tools and methods
- sufficient computing and storage capacity
- no interconnected data infrastructure
- open legal and ethical issues

UNSTRUCTURED DATA
Characteristics:
- e.g. text, video, audio
- low standardization
- heterogeneity of sources and formats
- no standardized collecting, managing and analyzing
- no standardized tools and methods
- no sufficient computing and storage capacity
- no developed data infrastructure
- open legal and ethical issues

Common goal
Interconnected infrastructure for handling both structured and unstructured data
OpenML Technical Architecture

- WEB UI (REACT)
- PYTHON API
- C# API
- R API
- JAVA
- CLI

- SERVER (FLASK)
- CORE
- REST API

- SEARCH (ELASTICSEARCH)
- DATABASE (MYSQL)
- S3 OBJECT STORE (MIN.IO)

ML library integrations

bindings

service layer

data layer
Hidden Technical Depth of Machine Learning

- ML systems for unstructured and *dirty* data are hugely complex
- Plethora of different pipeline steps
- If not embedded in a proper infrastructure, users are lost and projects fail (late)

BERD Building on OpenML

- All objects linked and searchable: data, algorithms, scripts, results
- Many major ML toolkits integrated
- Programming language agnostic
- Fully reproducible

I shared a new data set

I found a better model!
BERD as an Open Platform for Analysis

**Domain Specialist**
- Define task in accordance with theory
- Refine theory based on results

**BERD**
- Interactive notebooks
- Similarity search on studies
- AutoML removes drudge work
- ...

**Data Scientist**
- Map task to analysis
- Refine and optimize analysis pipelines

BERD facilitates optimal collaboration between domain specialists and data scientists