Pulling the wool over users´eyes

Why is a German Research Data Center interested in Synthetic Data?

Stefan Bender
(Institute for Employment Research, Germany)

NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection 2009, Washington, D.C.

31. July 2009

Overview

- A very short History of German Data Access
- The Portfolio Approach to Confidentiality Protection Lane 2007)
- The RDC of the BA in the IAB
- Imputed data sets for research
- Conclusions/Future Work

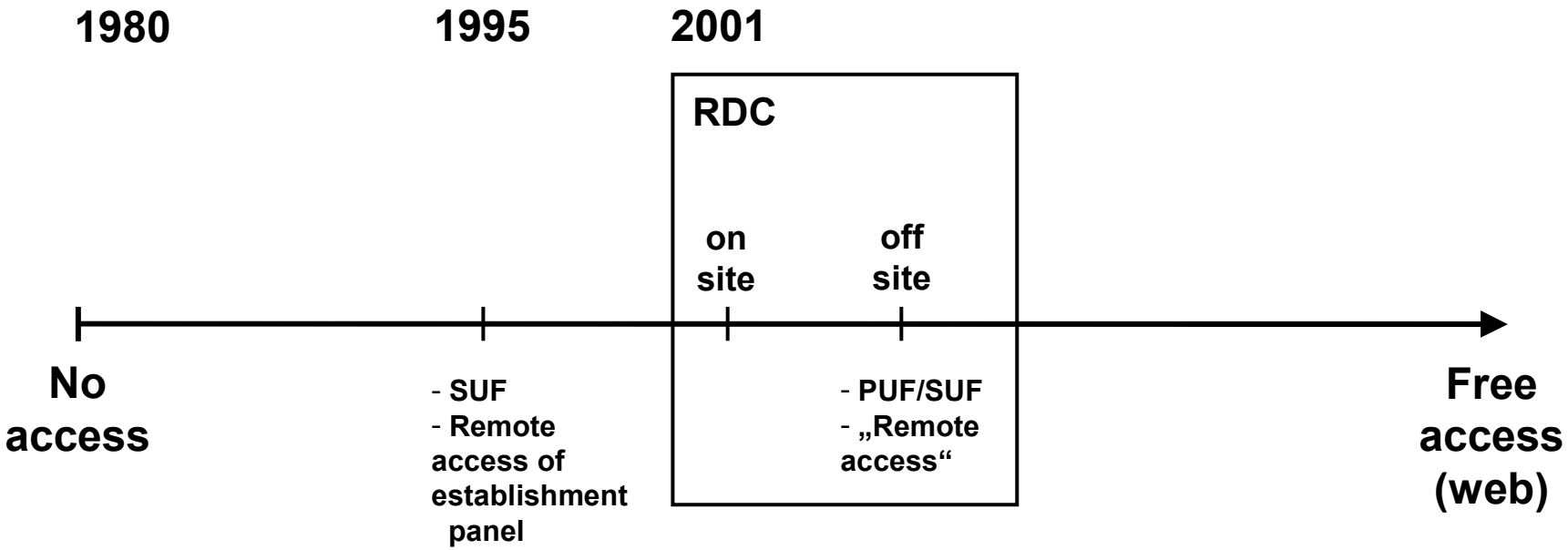
Data Access in Germany: The 80´s

- Law for the census (1987): privilege for research
- Federal Statistics Law: from absolute to factual anonymity
- First scientific use files were published

Data Access in Germany: The 90´s

- Constant pressure of the scientific community „Set the data free“
- The „Commission to Improve the Statistical Infrastructure in Cooperation with the Scientific Community and Official Statistics“ (*KVI 1999, report 2001)* established by the Federal Ministry of Education and Research (*BMBF).*
- German Council for Social and Economic Data (RatSWD).
  The Council’s main purpose is to advise in the development of the German data infrastructure for empirical research in the social and economic sciences.
- Establishment of Research Data Centers (RDC) by data producers and Data Service Centers (DSC). At the beginning all co-financed by the BMBF.

Development in Germany



RDCs provide researchers access to micro data for non-commercial empirical research in the fields of social security and employment

- Advisory service on data selection and data access
- Handling of remote execution
- Assistance and support for visiting researchers
- Online data documentation and documentation of methodological aspects of data
- Clarification of questions on data protection
- Updates of scientific use files and other research datasets
- Organization of workshops and user conferences

The Portfolio Approach to Confidentiality Protection (Lane 2007)

- How should data be protected at the disseminating institution?
  - Technological Protection
  - Statistical Protection
  - Operational Protection
  - Legal Protection

(RDC in RDC approach, Remote Data Access)

Statistical Protection

- Replacing all personal/organizational identifiers
- Drawing samples
- Standardised microdata (no individual data solutions)
- Generating scientific use files (deleting variables, aggregation, **multiple imputation**)
- Disclosure limitation review
- Different kinds of data access

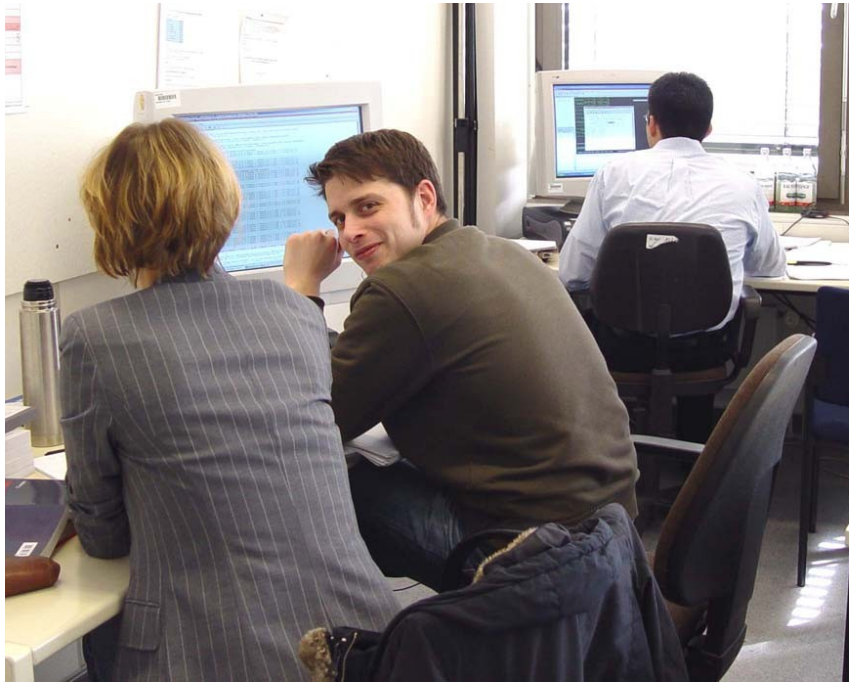
RDC of the BA in the IAB

- Started in April 2004; positive evaluation in April 2006, since December 2006 100% financed by the Federal Employment Agency
- 6 researchers, 3 non-researchers, office in Nuremberg
- Data based on:
  - the notification process of the social security system,
  - the internal procedures of the Federal Employment Agency
  - data from IAB-surveys (e.g. IAB Establishment Panel).
- Available data:
  - IAB Establishment Panel, Establishment History Panel
  - IAB Employment Sample, BA Employment Panel, Integrated Employment Biographies Sample of the IAB
  - Linked Employer Employee of the IAB

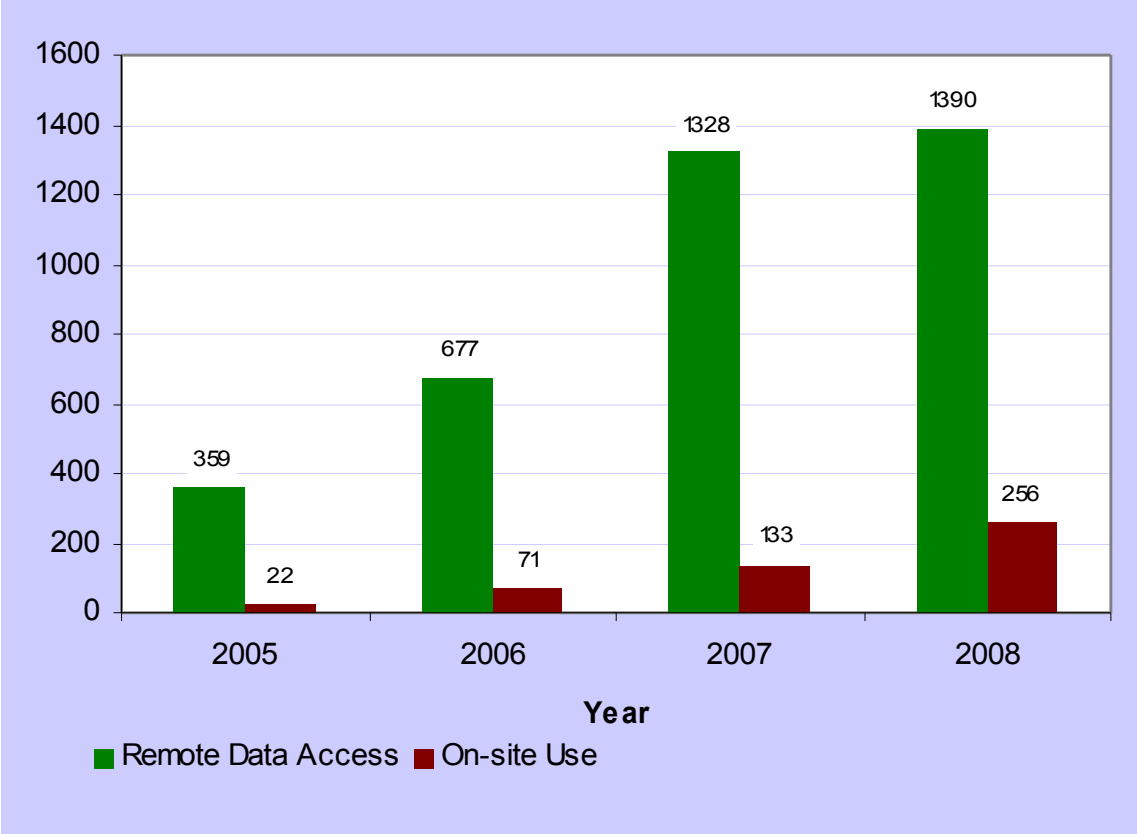
Data Access

- Restriction: access for non-commercial empirical research in the fields of social security and employment
  - On-Site Use
  - Remote Data Access
  - Scientific Use Files

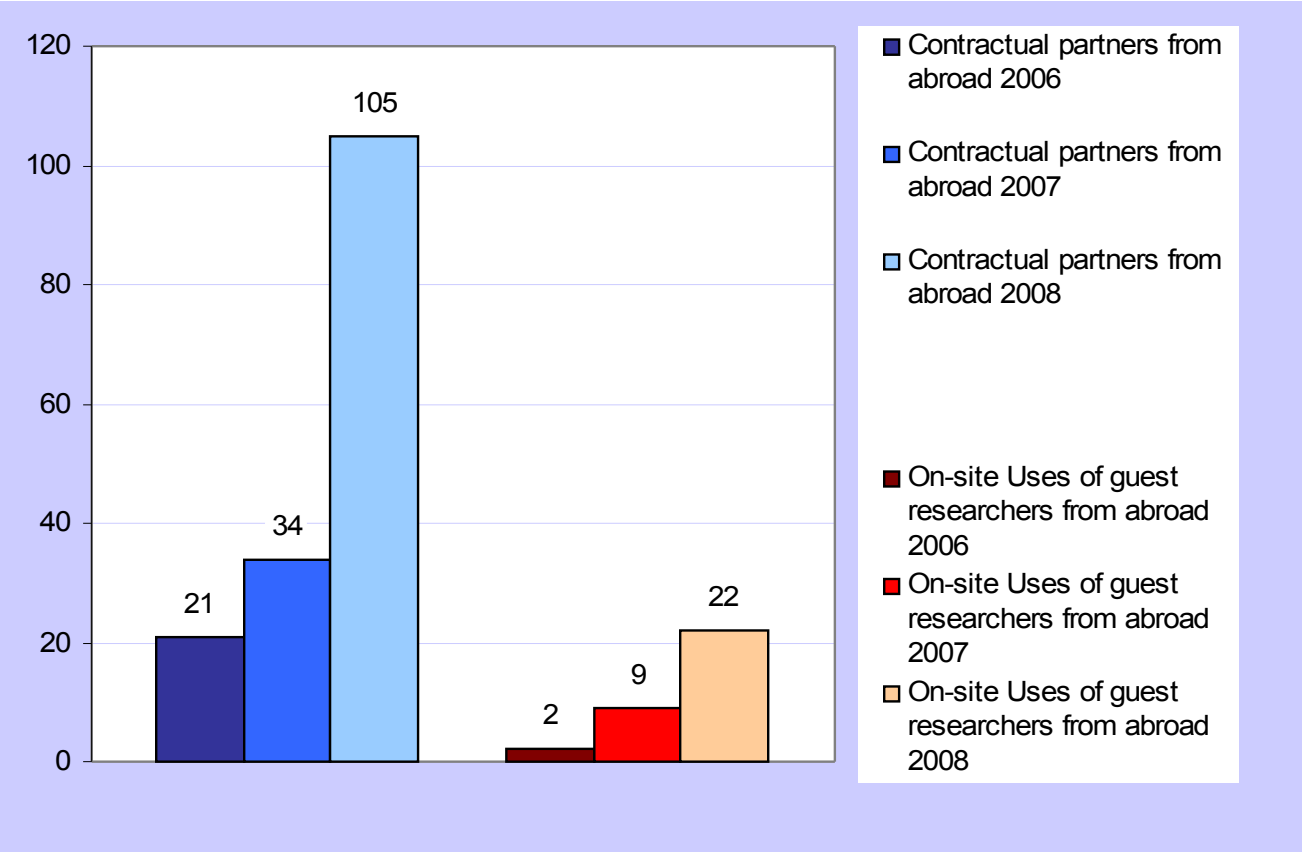
- **No costs**
- **Financial support for guest researchers from abroad**



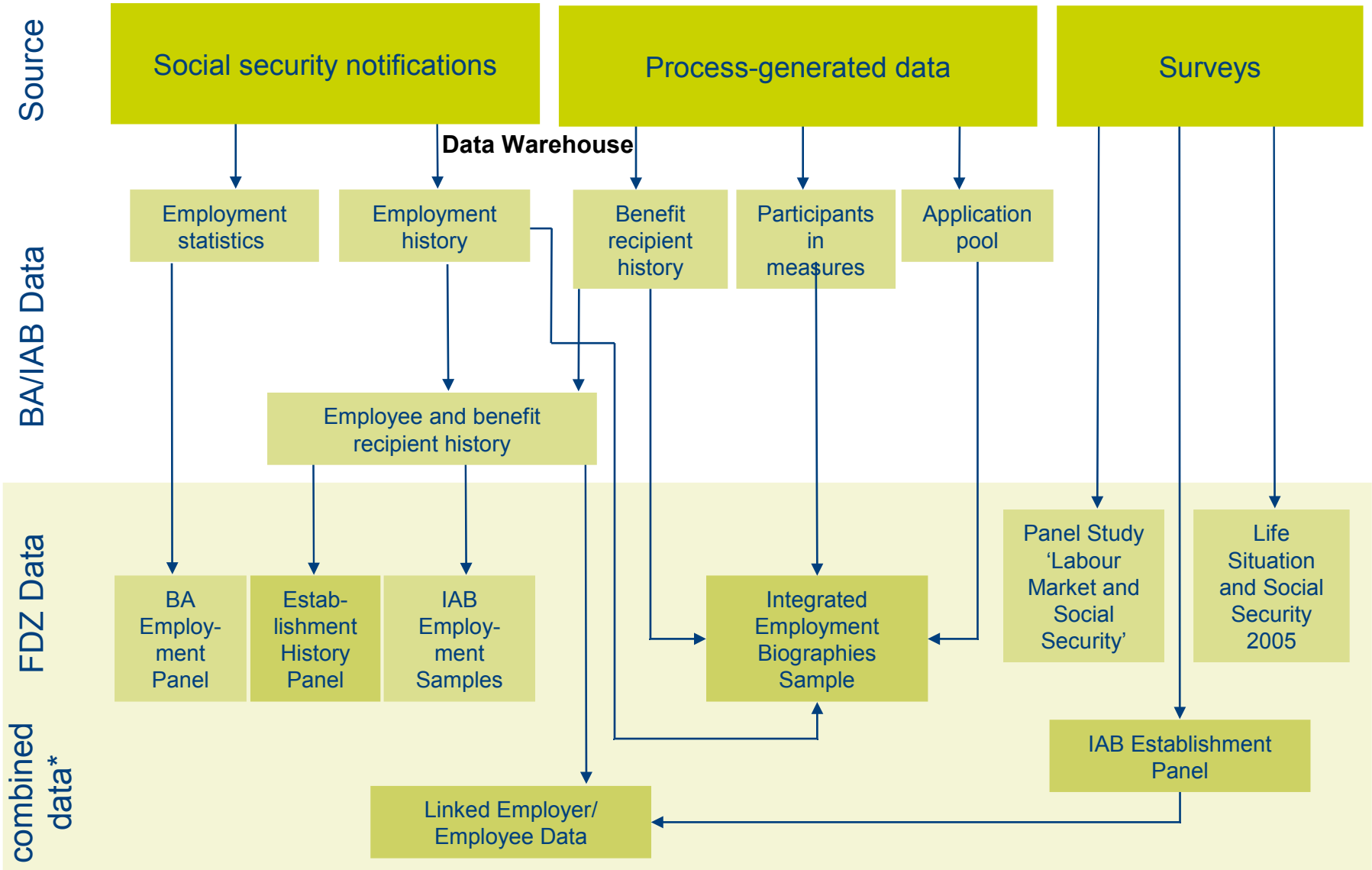
of remote data accesses and on-site uses



of users outside Germany (contractual partners, guests)



Data Sources and Paths



Data Access at the RDC

Datasets	Data Access		
	On-site Use	Remote Data Access	Scientific Use File
IAB Employment Samples	✓	✓	✓
BA Employment Panel	✓	✓	✓
Integrated Employment Biographies Sample	✓	✓	✓
IAB Establishment Panel	✓	✓	X
Linked-Employer/Employee Data	✓	✓	?
Establishment History Panel	✓	✓	?
Cross-sectional survey 'Life Situation and Social Security 2005 (LSS 2005)'			✓
Panel Study 'Labour Market and Social Security' (PASS)			✓

We have only test data for the Establishment Data.

Anonymising business micro data (FAWE) – Motivation I

- Public release of business data is often considered too risky

  - Skewed distributions make identification of single units easy
  - Number of units in total population is small
  - Information on businesses in the public domain
  - High benefits from identifying a single unit
  - High probability of inclusion for large establishments

- Standard perturbation methods have to be applied on a high level

- Release of high quality data is very difficult

FAWE - Motivation II

- „Democratic“ access to micro data (RatSWD guidelines for RDCs).
- Research on anomization techniques.
- Need to have scientific use files for establishment/firm data.

- European” movement towards a better data access (Essnet projects)

RDC-Motivation to join the Project

- Yearly conducted establishment survey (IAB Establishment Panel)
- Strong demand for access from external researchers
- Only on-site and remote access possible so far, only structural test data.
- High costs in terms of time and money.
- Project-goal: Generate synthetic datasets of the survey for public release.

- Project start: summer 2006
- Number of people working on this project: 1-1,3

FAWE – Project Information

- Former project (2001-2005) by Federal Statistical Office Germany and Statistical Offices of the Länder, Institute for Applied Economic Research (IAW), Centre for European Economic Research (ZEW).
- Result: Release of some cross sectional firm data as scientific use files from the Statistical Offices.
- New project with Institute for Employment Research (IAB) to release panel data and data of the IAB.
- Financed by the Federal Ministry for Education and Research (BMBF)

Results of our Partners

- Compared estimation results between anonymized real data and original data (German Turnover Tax Statistics, German Structure of Costs Survey).
- Anonymization Techniques: information suppression, categorization, micro aggregation and additive or multiplicative noise.
- Micro aggregation and adding noise lead to biased estimations. For example, Adding independent noise to the covariates leads to the error-in-variables problem.
- Corrections for some (mostly linear) estimator.

Brother Grimm: The Hare and the Hedgehog

A fairy tale, where the hedgehog tricks the rabbit in a race with his wife as a double, already waiting at the finish.

Our Results

- Results: Jörg´s presentation

- Real re-identifikation experiments:
  - Used data sets: IAB-Establishment Panel and a public available commercial data base (AMADEUS).
  - True matches between the two data sets by using business names and addresses.
  - Probabilistic record linkage with three blocking variables and 13 matching variables (EM algorithm frame work).
  - Against the literature the re-identification risk of firms in our experimental setting is comparable to individuals.

Conclusions: Imputation

- Synthetic datasets provide a high level of disclosure protection.
- Partially synthetic fulfill our needs for data protection.
- Synthetic datasets offer a high level of data utility.
- First dataset almost ready for release.

Future Work

- Provide metadata for the user.
- Weighting of data set.
- Long term goal: release complete longitudinal data.

Conclusion for the RDC

- Synthetic data is a promising way for generating public use files for sensitive data like business data.
- Generating synthetic datasets is a labour intensive task.
- Faster release of imputed data?
- Need to compare anonymization techniques (imputed data vs real data is only one dimension).
- Need to convince users to use imputed data (also convince referees).
- Trust in researchers´ activities (“culture of confidentiality”). Raise researchers knowledge and how their activities are related to data confidentiality (researcher training).
- Imputation is just one dimension (RDC in RDC approach)

Very Short Summary

Contact:
stefan.bender@iab.de

Web site of RDC:
http://fdz.iab.de/en

The IAB Establishment Panel

- Annually conducted establishment survey
- Since 1993 in Western Germany, since 1996 in Eastern Germany
- *Population*: All establishments with at least one employee covered by social security
- *Source*: Official Employment Statistics
- Response rate of repeatedly interviewed establishments more than 80%
- Sample of more than 16.000 establishments in the last wave
- *Contents*: employment structure, changes in employment, business policies, investment, training, remuneration, working hours, collective wage agreements, works councils

“RDC in RDC”

- Remote Access in Germany is far behind solutions like in Denmark or in the Netherlands.
- We have to fulfil requirements of data protection and data security.
- Problem: not every country has the same data protection standards/regulations/laws.
- Nearly the same standards in RDCs all over the world (or the standards can be established).
- Main problem for German data protection: how to control who is sitting at the PC. That is possible in other RDCs.

Model of the RDC in RDC approach