

**IAB-Datenbasis für Evaluationszwecke
und
Aufbau einer allgemeinen
Maßnahme-Teilnehmer-Grunddatei**

Elmar Kellner (IAB)
Jürgen Passenberger (IAB)

IAB Institut für
Arbeitsmarkt- und
Berufsforschung

Aufbau einer Datenbasis für Evaluationszwecke im IAB

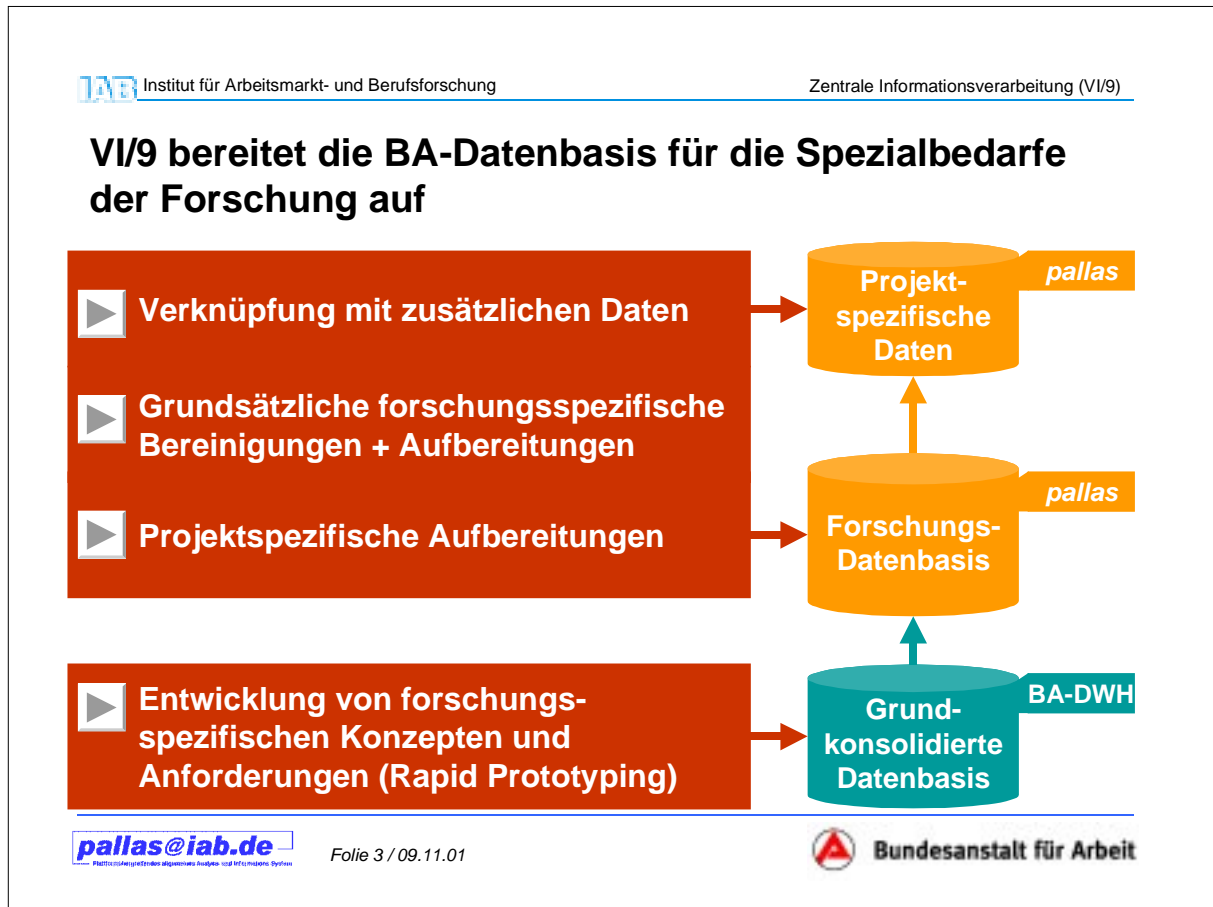
(Maßnahme-Teilnehmer-
Grunddatei)

Elmar Kellner
Team Forschungsdatenmanagement (FDM)
im Bereich VI/9

Dr. Jürgen Passenberger
Bereichsleiter VI/9

Inhalt

- ▶ **Forschungsdatenbasis für Evaluationszwecke**
- ▶ **ABM-/SAM-Datenbasis als erste Stufe**
- ▶ **Aufbau einer allgemeinen Maßnahme-Teilnehmer-Grunddatei (MTG) im IAB**

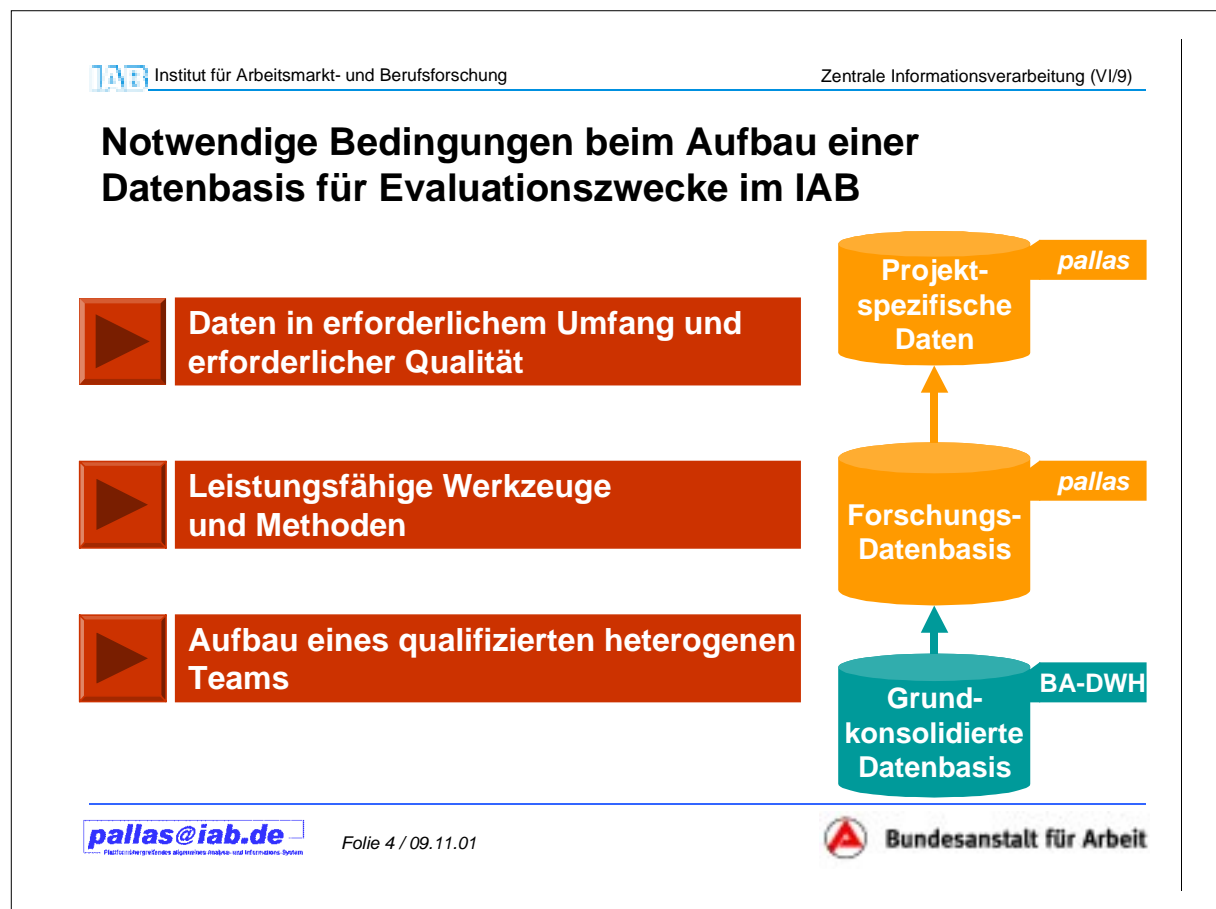


Eine zweckmäßig angelegte dezentralisierte Arbeitsmarktpolitik benötigt qualifizierte Informationen für eine effektive Steuerung der Prozesse am Arbeitsmarkt. Von der Forschung werden auf die Fragen der Politik differenzierte Antworten „just in time“ erwartet. Forschung im IAB benötigt daher eine leistungsfähige dedizierte Informationsverarbeitung (VI/9), die die BA-Datenbasis für forschungsspezifische Bedarfe aufbereitet. Zentrale Leistungen des Teams „Forschungsdaten-Management“ (FDM) im Informationsverarbeitungsbereich des IAB (VI/9) sind:

- 1) Erweiterung und forschungsspezifische Bereinigung der BA-Datenbasis (d.h. Data Warehouse- und BS2000-basierte Altsystem- Datenbasis der BA); insbesondere umfassen diese Arbeiten:
 - a) die Verknüpfung der BA-Daten mit zusätzlichen internen und externen Daten sowie eigen-erhobenen Daten des IAB,
 - b) grundsätzliche forschungsspezifische Bereinigungen und Datentransformationen, z.B. Gebietsstandsanpassungen zur Bildung langer Zeitreihen mit Regionsbezug, Aufbau personenbezogener Individualdatensätze von Erwerbspersonen zur Verfolgung von Erwerbskarrieren
 - c) projektspezifische Aufbereitungen der IAB-/BA-Datenbasis, z.B. Zusammenfassungen von Merkmalskategorien, Missing-Data-Ersetzungen, projektspezifische Generierung von Synthesemerkmalen etc.
- 2) Entwicklung von forschungsspezifischen Datenbereitstellungskonzepten und -anforderungen durch Systematisierung und Verallgemeinerung von Rapid-Prototyping-Umsetzungen: Datenererschließungs- und Datenbereitstellungsanforderungen seitens einzelner IAB-Forschungsprojekte werden mit Methoden und Werkzeugen des Datawarehousings sowie dem gesamten Funktionspektrum der Datenanalyse-Werkzeuge der Wirtschafts- und Sozialforschung prototypisch vom FDM-Team des IAB schnell umgesetzt („Rapid Prototyping“). Auf Basis dieses Prototyping-Prozesses können (verallgemeinerte) Datenbereitstellungskonzepte und -anforderungen an das BA-

DataWarehouse entwickelt, formuliert und nach Abstimmung mit den Fachbereichen des IAB als Change-Request-Anforderung an das Datawarehouse-Team der BA weitergeleitet werden. Diese Arbeiten im Vorfeld einer regulären, d.h. auf Dauer zu stellenden Übernahme und Aktualisierung von Daten in das BA-DataWarehouse sind notwendige Vorarbeiten für eine zügige qualitätsvorgeprüfte Aufnahme von Daten in das BA-DataWarehouse.

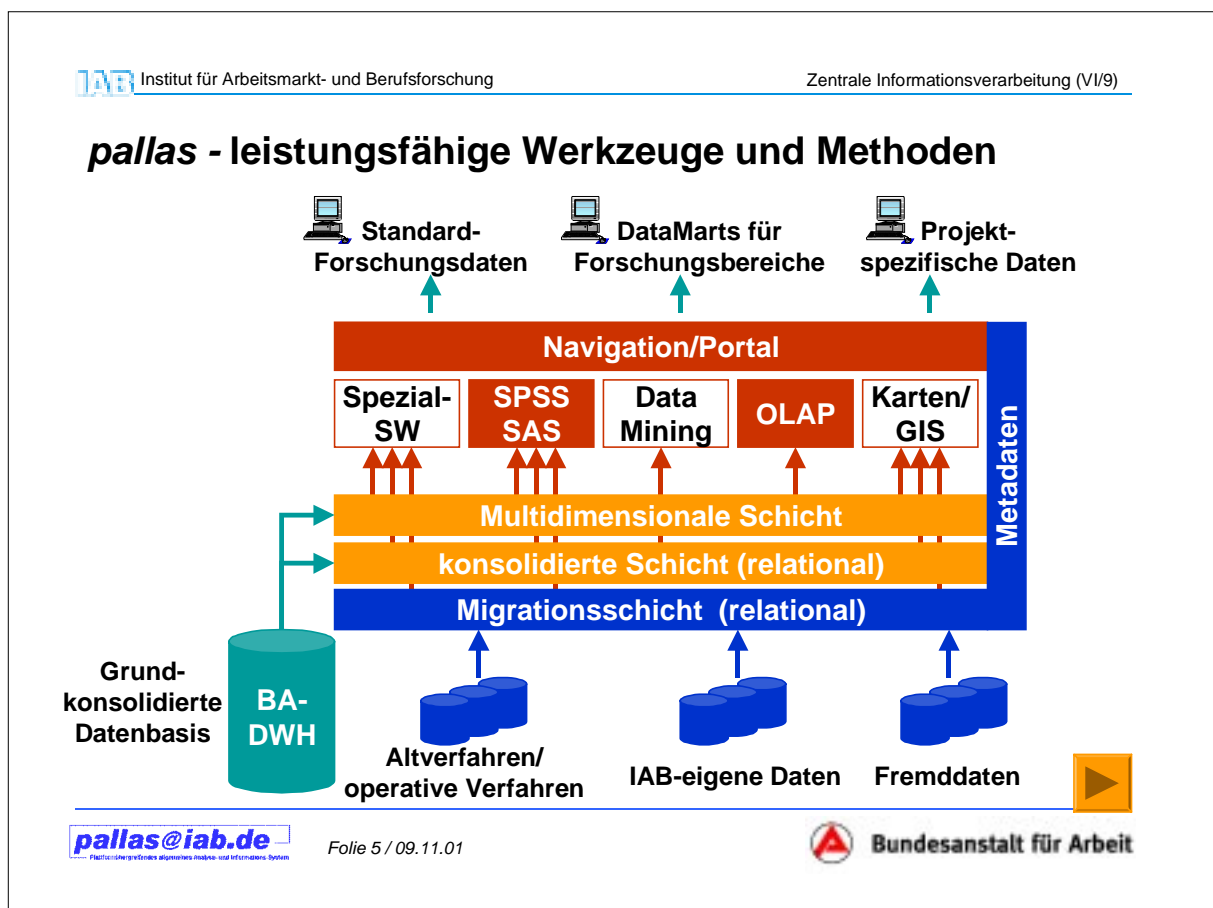
Die derzeit in VI/9 stattfindenden und geplanten Aktivitäten zum Aufbau einer Datenbasis für Zwecke der Evaluationsforschung sind der Kategorie 2 zuzuordnen: Im Rahmen von Machbarkeitsstudien und projektspezifischen Datenbereitstellungsarbeiten werden Datensätze aufbereitet und verknüpft, die bereits jetzt als Auswertungsbasis für die Wirkungsforschung genutzt werden können. Die dabei gewonnenen Erkenntnisse und Ergebnisse sind darüber hinaus wichtige fachliche und datentechnische Vorarbeiten für die Integration dieser Daten in das BA-DataWarehouse.



Größte Priorität beim Aufbau einer Datenbasis für Evaluationszwecke im IAB hat die instrumentenübergreifende Erschließung und Aufbereitung von Individualdatensätzen von Teilnehmern und Vergleichsgruppen zur Ermittlung des Erfolgs von Maßnahmen aktiver Arbeitsmarktpolitik (AAMP). Dass beim Aufbau dieser Datenbasis in der jüngsten Zeit erhebliche Fortschritte gemacht werden konnten, erforderte das Zusammenspiel von drei Parametern, die die notwendige Bedingung für die Bewältigung einer derartigen Herausforderung bilden:

- 1) Verfügbarkeit von Daten in erforderlicher Qualität: Der zentrale Zugang zu den Individualdatensätzen von Maßnahmeteilnehmern für Zu- und Abgänge sowie Bestände ist ab 1.1.2000 möglich. Über die ebenfalls seit diesem Zeitpunkt in den Datensätzen enthaltenen Identifikatoren werden nun auch wichtige Strukturdaten von Maßnahmen mit denen von Maßnahmeteilnehmern verknüpfbar und Informationen über den Verbleib nach der Maßnahme können hinzugefügt werden.

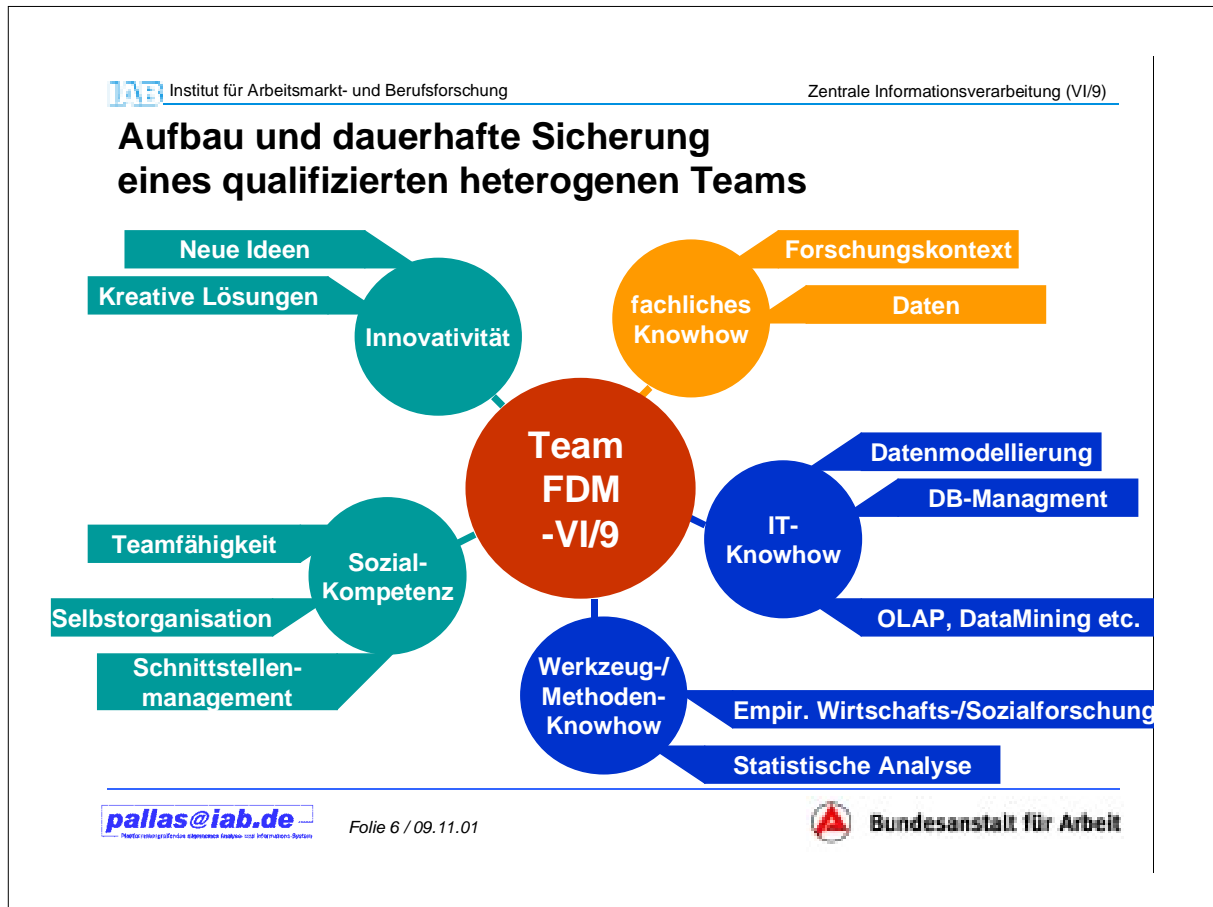
- 2) Verfügbarkeit leistungsfähiger Datenmanagement-Werkzeuge: Das Ziel schneller Umsetzung von Anforderungen der Forschung, insbesondere der Wirkungsforschung, wird durch den Einsatz multifunktionaler integrierter Standard-Softwarewerkzeuge erreicht. Produkte dieser Art garantieren auch bei großen Datenmengen eine hohe Aktions- und Reaktionsgeschwindigkeit, hohen Investitionsschutz bei der Entwicklung von Aufbereitungs- und Auswertungsprozeduren, bieten durch ihre weite Verbreitung gute Rekrutierungsmöglichkeiten für versiertes Personal und ermöglichen ein hohes Maß an Auswertungsflexibilität durch standardisierte Schnittstellen für Daten und Metadatenübergabe an andere Werkzeuge. Die Anwendung dieser Werkzeuge erfordert kein IT-Spezialwissen.
Ein wichtiger methodischer Quantensprung wurde durch die Anwendung des DWH-Konzepts mit seinen Modellierungsansätzen und der mehrschichtigen Architektur erzielt.
- 3) Verfügbarkeit eines qualifizierten interdisziplinären (heterogenen) Datenmanagement-Teams: Interdisziplinär zusammengesetzte Datenmanagement-Teams aus Informatikern und Wirtschafts- und Sozialwissenschaftlern verstehen die Sprache der Fachbereiche besser als „reine“ Teams, die ausschließlich mit IT-Experten besetzt sind. Auf diese Weise ist eine enge Verzahnung der Arbeiten des Teams FDM mit dem Forschungsbereich möglich, Kooperation erfolgt auf gleicher Ebene, Reibungsverluste auf zwischenmenschlicher Ebene durch „Spartendenken“ können minimiert werden. Es ist VI/9 im Zuge der Entwicklungsarbeiten für *pallas* gelungen, ein derartiges Team aufzubauen und längerfristig zu sichern, indem externe Wissenschaftler mit kombiniertem wirtschafts- und sozialwissenschaftlichem und IT-Knowhow in den regulären Mitarbeiterstamm integriert wurden.



Um aufbauend auf der BA-Datenbasis spezifische Forschungsdaten für generelle Fragenstellungen der Arbeitsmarkt- und Berufsforschung, für Forschungs-Schwerpunktthemen wie die Evaluations- oder Regionalforschung oder für konkrete Forschungsprojekte aufzubereiten und bereitzustellen, ist eine moderne Informationstechnologie mit leistungsfähiger Hardware und einem breiten Spektrum an Werkzeugen erforderlich. Im noch im Aufbau befindlichen *plattformübergreifenden allgemeinen Analyse- und Informationssystem – pallas* – setzt das IAB neben klassischer Software aus dem Bereich der statistischen Analyse (neben SAS und SPSS je nach Fragestellung auch single purpose -Produkte) hierzu DataWarehouse-Methoden und –Technologien ein. Diese sind prädestiniert, insbesondere die Verknüpfbarkeit von Daten aus unterschiedlichsten Quellen, performante automatisierte Datenaktualisierungen sowie performante und komfortable multidimensionale (ad-hoc-) Online-Analysen zu realisieren.

Mit Hilfe dieser Technologien werden die aus dem BA-DataWarehouse bzw. aus den Altsystemen der BA abgerufenen Basisdaten in einer konsolidierten relationalen Schicht für Forschungszwecke aufbereitet (z.B. bereinigt, mit weiteren Datensätzen verknüpfbar gemacht, durch Gebietsstands-bereinigungen für lange Zeitreihen vorbereitet). Dabei werden über eine Migrationsschicht weitere Daten integriert, die nicht über das BA-DataWarehouse zur Verfügung stehen. Diese speziell für Forschungszwecke konsolidierte Schicht bildet das Kernstück des IAB-Analyse- und Informationssystems. Aus ihr werden die Daten dann in Form multidimensionaler Würfel für performante Online-Analysen in unterschiedlichsten Verwertungskontexten bereitgestellt (multidimensionale Schicht). Über ein breites Spektrum an Analyse-Werkzeugen können je nach Fragestellung die Daten in den einzelnen Schichten abgefragt und bearbeitet werden.

Da die für Evaluationszwecke erforderlichen Daten zu großen Teilen noch nicht über das BA-DataWarehouse verfügbar sind, gewinnt VI/9 die für die aktuellen Bedarfe der Wirkungsforschung erforderlichen Daten derzeit aus den Altverfahren und führt diese in *pallas* zu einer konsolidierten verknüpfbaren Datenschicht zusammen. Die dabei entwickelten Modellierungs-, Prüf- und Bereinigungsregeln werden später als Anforderung der Forschung an das BA-DataWarehouse übergeben und dort in der konsolidierten Datenbasis umgesetzt, soweit es sich um allgemein relevante Aufbereitungsschritte handelt. Die nur für die Forschung relevanten Aufbereitungen werden auch in Zukunft in *pallas* erfolgen.



Zusammensetzung des interdisziplinären Teams Forschungsdaten-Management (Team FDM) im Informationsverarbeitungsbereich des IAB mit Mitarbeitern sowohl sozial- und wirtschaftswissenschaftlicher als auch informationstechnischer Ausbildung:

Das FDM-Team vereint IT-Kompetenz mit fachlich-methodischer Kompetenz; eine starre Arbeitsteilung zwischen IT und Fachbereich konnte dadurch überwunden werden. Hohe Kompetenz und Motivation zur Lösung komplexer Datenmanagement-Probleme entsteht im Team durch

- Fachverständnis für den projektspezifischen wissenschaftlichen Verwertungskontext ,
- Wissen um den Entstehungskontext der benötigten Daten,
- IT-Kompetenz, insbesondere Kenntnisse aus dem Bereich Datenmodellierung,, Datenmanagement-Systeme, Datenanalysesysteme,
- Kenntnis der Methoden der empirischen Sozial- und Wirtschaftsforschung sowie der Statistik.

Darüber hinaus ist eine hohe kommunikative und soziale Kompetenz des Teams im Rahmen der angestrebten engen Zusammenarbeit sowohl mit IT-Abteilungen der BA als auch mit den IAB-Fachbereichen erforderlich.

Eine flexible, wenn nötig längerfristiger Verstärkung des Teams FDM durch externe Experten ermöglicht

- die Fruchtbarmachung innovativer Ansätze,
 - die Überwindung von Betriebsblindheit
- im jeweils konkreten Forschungsprojekt.

Fazit: Großes Gewicht ist der Zusammensetzung des Datenmanagement-Teams beizumessen, sie stellt einen wesentlichen Erfolgsfaktor für das Gelingen von empirischen Großprojekten dar. Das von derartigen Teams gewonnene **Knowhow** ist über das Projektende hinaus zu **sichern** durch

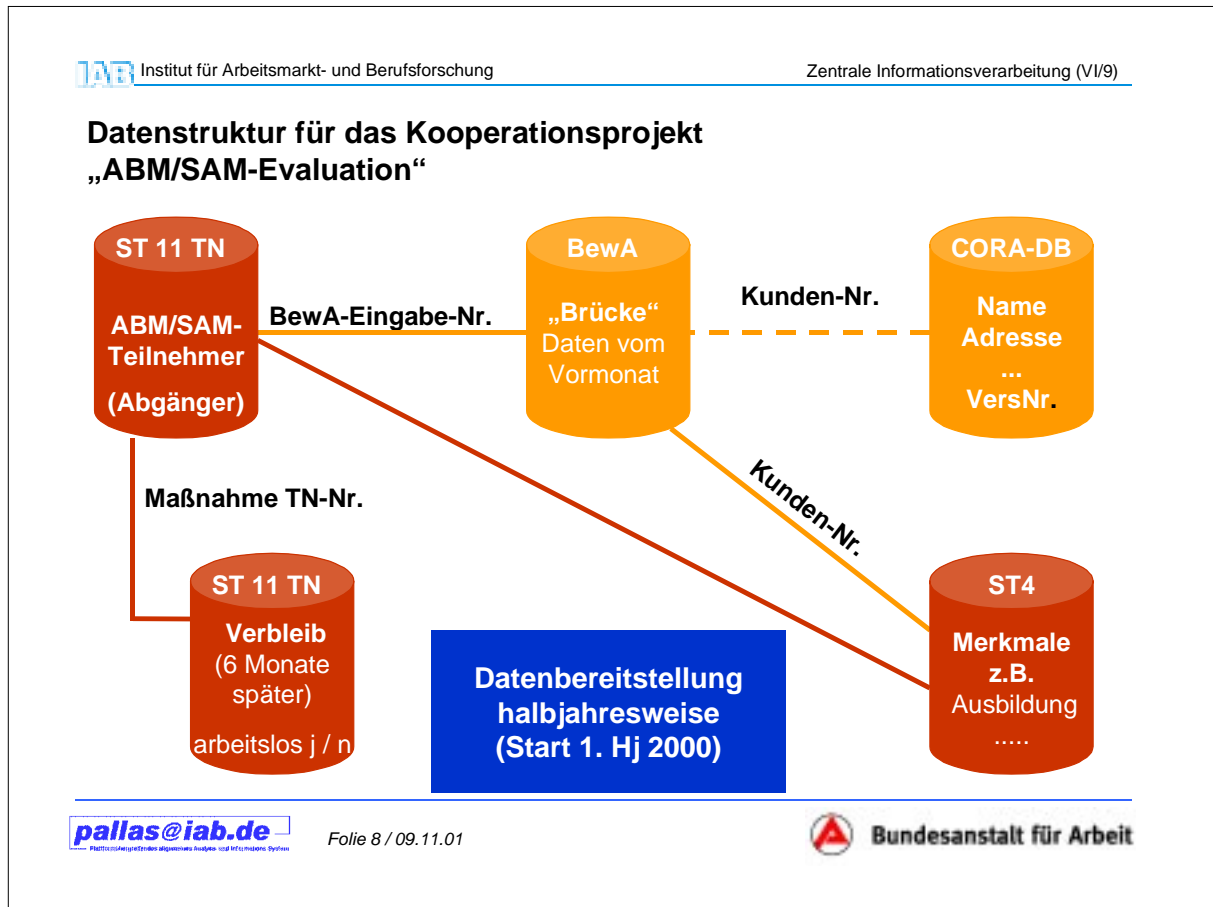
- dauerhafte (projektunabhängige) Sicherung der Personalressourcen eines Kernteams
- Aufbau und Pflege eines auf Dauer gestellten zentralen Dokumentationspools (Metadatenbank).

Anforderungen im Rahmen des ABM/SAM-Kooperationsprojekts

- ▶ **Bereitstellung von Einzeldatensätzen von Teilnehmern an ABM/SAM-traditionell**
 - ▶ angereichert um Informationen zur Person
 - ▶ verknüpft mit Adressdaten für eine Telefonbefragung
 - ▶ mit einer Verknüpfungsmöglichkeit zur Beschäftigtenstatistik
- ▶ **Kontrollgruppenbildung zu den Teilnehmerdaten**
- ▶ **Verbleibsanalysen bis zum aktuellen Rand**

Am Beispiel der Datenerschließungsanforderungen der Wirkungsforschung wird die Funktion des Teams „Forschungsdaten-Managements“ (Team FDM) als „Task Force“ für die Anforderungserfüllung deutlich: Im Rahmen einer Machbarkeitsstudie erschloss das FDM-Team in wenigen Wochen mit Hilfe leistungsfähiger Datenmanagement-Werkzeuge die Datenbasis des IAB-Projekts 10-535 zur Analyse des Eingliederungseffekts von ABM und SAM (Kooperationsprojekt mit Professor Hujer, Universität Frankfurt). Die Konzentration auf die projektspezifischen Bedarfe machte ein unkonventionelles, schnell Alternativen erprobendes Vorgehen möglich, das in kurzer Zeit die zur Analyse der Forschungsfragen erforderlichen Datensätze bereitstellt: Einzeldatensätze von Teilnehmern an ABM/SAM-Maßnahmen werden um Informationen zur Person so weit wie möglich angereichert, Verknüpfungsmöglichkeiten mit Adressdaten für telefonische Befragungen sowie mit Daten der Beschäftigtenstatistik erprobt und hergestellt. Last but not least wird versucht, eine unter gegebenen Umständen optimale Datengrundlage für die methodisch aufwändige Vergleichsgruppenbildung und für Verbleibsanalysen bis zum aktuellen Rand aufzubauen.

Durch eine entsprechende Modellierung wurde sichergestellt, dass die dabei entstandene Datenbasis nicht nur einmalig für dieses Projekt bereitsteht, sondern regelmäßig aktualisiert und für weitere Fragestellungen der Evaluationsforschung verallgemeinert bzw. erweitert werden kann.



Voraussetzung für die Erreichung der Ziele des ABM-/SAM-Kooperationsprojekts mit Prof. Hujer war die Bereitstellung einer Datenbasis für ehemalige ABM-/SAM-Teilnehmer mit der Information des Verbleibs 6 Monate nach Beendigung der Maßnahme sowie Informationen zu persönlichen Merkmalen wie z.B. Ausbildung und –als Basis für die Telefonbefragung – zu Name, Adresse, Telefonnummer. Bis entsprechend verknüpfbare Daten aus dem BA-Data Warehouse zur Verfügung stehen, werden die hierfür erforderlichen Datensätze in der konsolidierten relationalen Datenschicht von *pallas* aus verschiedenen Datenquellen der BA zusammengeführt:

1. Wesentliche Ausgangsdatenquelle ist die Datei „ABM-/SAM-Teilnehmer“ (ST11TN): stehen zur Verfügung?

Diese enthält monatliche Einzeldaten (Zugang, Bestand, Abgang, Verbleib) von Teilnehmern ABM/SAM. Das zu untersuchende ST11TN-Datenmaterial wurde erst Ende des Jahres 2000 von der Statistik freigegeben. Da bereits im April 2001 eine größere Datenmenge für Vorabauswertungszwecke zur Verfügung gestellt werden sollte (etwa 150.000 Teilnehmersätze), konnten bei einer durchschnittlichen Maßnahmedauer von 12 Monaten und der benötigten Verbleibsinformation ein halbes Jahr nach Abgang nur die Abgänge des 1. Halbjahres 2000 einbezogen werden. Unterdessen werden auch die *Zugänge* zu Beginn des Jahres 2000 extrahiert, für die nun ebenfalls Verbleibsinformation vorliegt. *BewA*: Bewerber-Angebot, Daten aus dem Verfahren coARB (dies sind „Vor-Statistik-Daten aus dem Geschäftsfeld der BA).

2. Wichtige persönliche Merkmale der Teilnehmer werden aus der ST4 gewonnen:

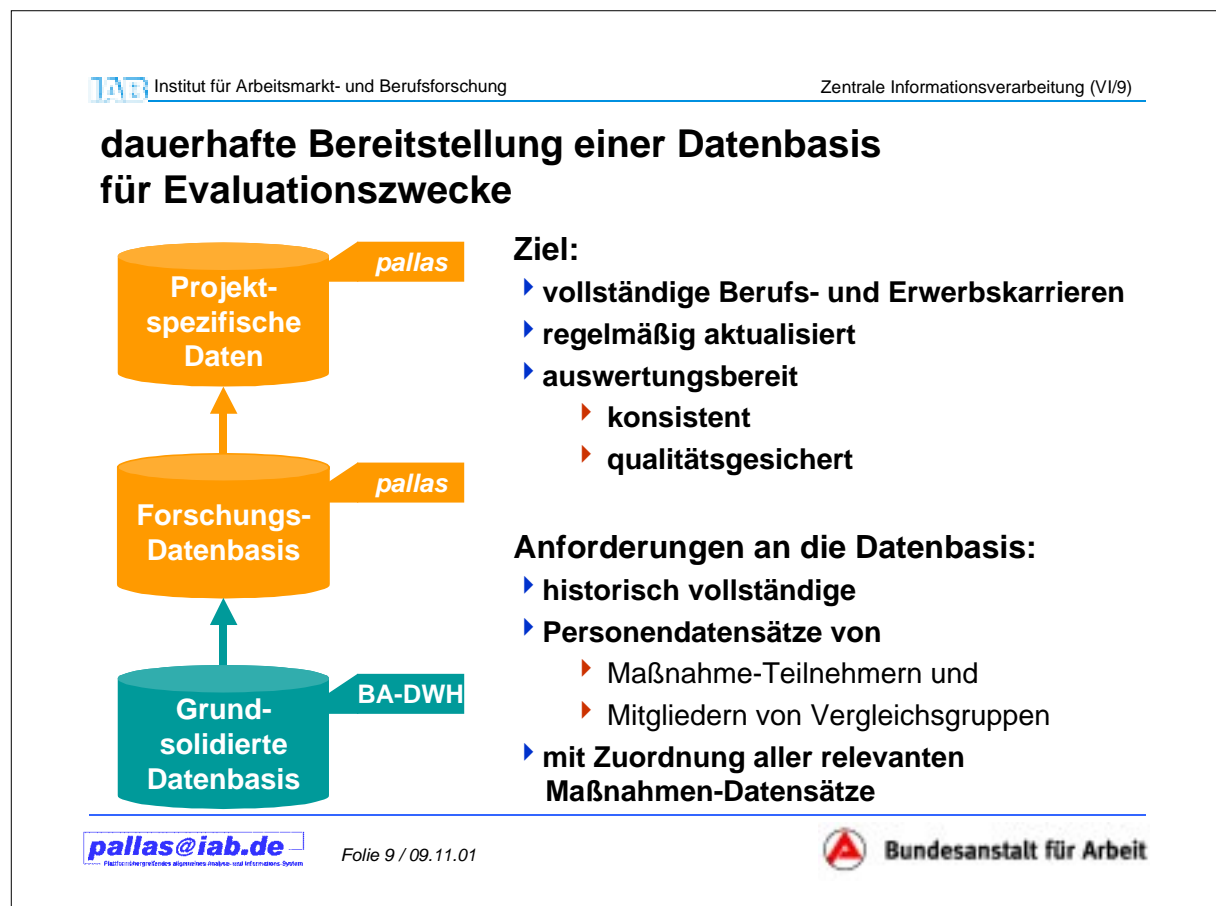
Um wichtige Merkmale für die Evaluation zu erhalten, die nicht in der ST11TN enthalten sind, müssen die ST11TN-Daten durch Daten aus der ST4 ergänzt, die monatliche Einzeldaten (Bestand) von Arbeitssuchenden enthält und aus der *BewA* (Bewerberangebot) gebildet wird. Möglich ist diese Verknüpfung, weil ABM/SAM-traditionell-Teilnehmer arbeitssuchend gemeldet sein müssen (Status „nicht arbeitslos“ in der ST4). Die Verknüpfung erfolgt auf die Vormonatsdatei, weil ein Teilnehmer im Abgangsmonat bei sofortigen Übergang in die Beschäftigung nicht mehr in der ST4 geführt wird.

3. Da eine direkte Verknüpfung zwischen ST11TN und ST4 (noch) nicht möglich ist, wird durch Hinzuziehung von Daten aus der BewA eine Brücke gebildet:

Die ST4-Einzeldaten enthalten ab Januar 2000 die Kunden-Nr als Identifikator, aber nicht die sogenannte BewA-Eingabe-Nr (zusammengesetzt aus Dienststelle, Meldedatum und laufender Nummer). Zwar wurde auch in den ST11TN-Einzeldaten die Kunden-Nr ab Januar 2000 eingeführt, dieses Datum wird jedoch nur bei den Zugängen (und da erst ab April 2000) eingegeben. Das heißt bei Zugängen vor Januar 2000 (aus denen die Datenbasis zunächst gebildet wurde) fehlt diese Kunden-Nr. Da die ST11TN auch die sog. BewA-Eingabe-Nr mitführt, kann über die BewA-Datei (Bewerber-Angebot; Daten aus dem Verfahren coArb), die sowohl die BewA-Eingabe-Nr als auch die Kunden-Nr. enthält, die Verknüpfung zu den ST4-Einzeldaten hergestellt werden.

4. Zuspielden von Adressdaten aus der CORA-DB:

Für die geplanten telefonischen Befragungen von Maßnahmeteilnehmern und Mitgliedern der Vergleichsgruppe sind Adressdaten erforderlich. Diese sind in den bisher betrachteten Datenbeständen nicht enthalten. Um die Teilnehmer-Datensätze mit den erforderlichen Adressdaten anzureichern, wird – über die Brücke BewA-Eingabe-Nr./Kunden-Nr. eine Verknüpfung mit der sog. CORA-Datenbank hergestellt. Die CORA-DB ist die zentralisierte Kundendatenbank, zusammengespielt aus den etwa 180 dezentralen Kundendatenbanken der Arbeitämter.

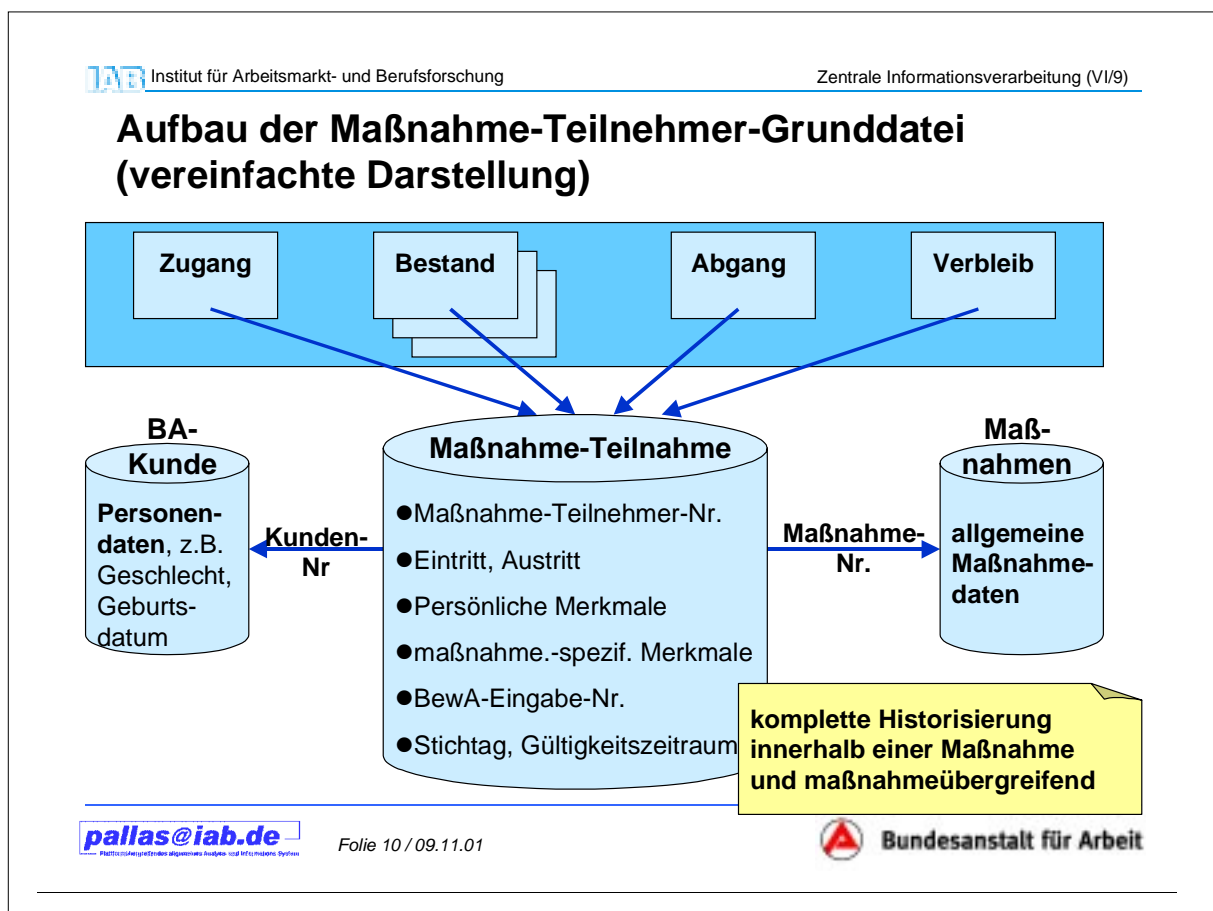


Die Datenbereitstellung im Rahmen des ABM-/SAM-Kooperationsprojekts ist nur der erste Schritt zum Aufbau einer allgemeinen und dauerhaft verfügbaren auswertungsbereiten Datenbasis für Evaluationszwecke. Die nächsten Schritte zum Aufbau der Datenbasis aus dem ABM/SAM-Projekt und ihrer Einbettung in eine instrumentenübergreifende zentral verfügbare „Maßnahme-Teilnehmer-Grunddatei“ sind im IAB bereits in Angriff genommen bzw. geplant:

Ziele für den Aufbau dieser Datenbasis sind:

- möglichst lückenlose Abbildung von Erwerbskarrieren, historisch vollständig
- regelmäßig aktualisierte Daten,
- auswertungsbereite Daten (d.h. gemäß vorgegebener Standards der Fachbereiche konsistent und bereinigt),
- Personendatensätze von Maßnahmeteilnehmern und Mitgliedern von möglichen Kontrollgruppen,
- personenbezogene Zuordnung aller relevanten Maßnahmen (Kontenstruktur der Daten).

Ziel ist es, zukünftig – aufbauend auf den Vorarbeiten des IAB – die Ausgangsdaten für eine derartige Datenbasis über das Datawarehouse der BA bereitzustellen; die für Forschungszwecke aufbereiteten Erwerbskarrieren werden im plattformübergreifenden allgemeinen Analyse und Informationssystem *pallas* des IAB aktualisiert und gepflegt.

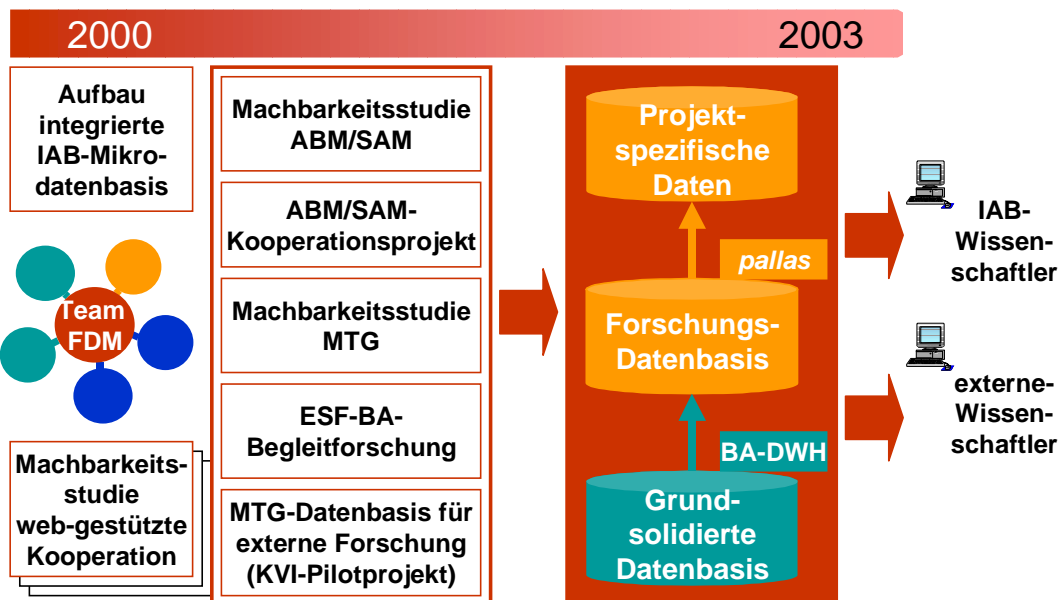


Die Maßnahme-Teilnehmer-Grunddatei (MTG) basiert im Grobmodell auf drei Entitäten: dem BA-Kunden, der Maßnahme und der Maßnahmeteilnahme, die miteinander verknüpft sind. Damit können sowohl Verläufe von Maßnahmen über Eintritts- bzw. Austrittszeitpunkte verfolgt, als auch im herkömmlichen Sinn statistischen Stichtagsdateien rekonstruiert werden.

Man kann sich aber genauso auf den Standpunkt des „BA-Kunden“ stellen und über einen längeren Zeitraum seine Maßnahmeteilnahmen betrachten, aber auch Förderungen zum gleichen Zeitpunkt. Diese Betrachtungsweise ist für Untersuchungen im Rahmen des ESF-BA-Programms unbedingt notwendig, da es sich bei ESF-Förderungen immer um Ergänzungsförderung handelt.

Die monatlichen, statistischen Input-Datensätze mit den Kennungen Zugang, Bestand, Abgang und Verbleib werden so abgespeichert, daß Bestandsredundanzen über das Erzeugen von Gültigkeitszeiträumen im wesentlichen vermieden werden. Gleichzeitig ist aber auch eine Historisierung gewährleistet, weil bei sich ändernden Merkmalen ein neuer Gültigkeitszeitraum aufgemacht wird.

Projekte im IAB zur Erstellung der Forschungsdatenbasis für Evaluationszwecke



Das FDM-Team bedankt sich für Ihre Aufmerksamkeit!

Elisabeth Hummel
 Steffen Kaimer
 Elmar Kellner
 Renate Lindenberg
 Wolfgang Majer
 Wolfgang Mössinger

Martina Oertel
 Armin Rauscher
 Manuela Reith
 Charly Schmidt
 Brigitte Weber
 Manfred Ziegerer

Jürgen Passenberger