Datenarchive

KVI Projekte
Datenservice- und Forschungsdatenzentren

Jürgen Krause
(mit freigegebenen Teilen von 4 der 5 Projekte)

Stand 11. Januar 2003

Universität Koblenz-Landau,
FB Informatik: Institut für Computervisualistik
Informationszentrum Sozialwissenschaften, Bonn

IZA: Internat. Datenservicezentrum (DSZ) Gesamtziele

- Metainformationsportal für Sekundäranalysen zur Bereitstellung von Informationen zur Arbeitsmarktforschung und relevanten Daten
- Die zwanzig wichtigsten Studien sollen identifiziert und ein indirekter Zugang durch eine Schalterstelle ermöglich werden (bei rechtlichen Restriktionen: Daten-Nutzung durch IZA-Gastwiss.modell)
- Deutsche Daten sollen international zugänglich gemacht werden durch Schalterstelle. Neben Umfragedaten gehören hierzu auch prozessgenerierte Daten aus der amtl. Verwaltung und von Unternehmen
- Forschungsdatenzentren (FDZ) und SDZ übergreifend recherchierbare Dokumentation und Standardisierung der Metadaten

  ⇨ Derzeit keine übergreifenden Recherchen möglich

SOEP Soepinfo:
http://panel.gsoep.de/soepinfo/

CESSDA IDC
http://www.nsd.uib.no/Cessda/IDC/

GESIS

Freigegeb./ergänzt 10.1.03 IZA

Jk

IZA: Informationsverdichtung und Integration

Mikro- und Makrodaten werden analysierbar sein: Metadaten unverzichtbares Instrument für umfassendes Datenverständnis.

Grundelemente der Informationsverdichtung DSZ

- Dublin Core
- DSZ-Studien und Variablen Thesaurus
- DSZ-Mikro-und Makrodatenbank

Grundelemente zur Integration DSZ

- Relationale-Datenbank (SQL) und Abbildung durch PHP im Internet

⇨ Vorschlag IZA: XML-Dokumente auf Basis des DC aus SQL-Datenbank generieren

⇨ Problem: technische Standards für Austausch der Informationen zwischen Datenbanken, Abfrage- und Abbildungsort, etc.

IZA: Informationsverdichtung und Integration

Grundelemente der Informationsverdichtung DSZ

- **Dublin Core**

  EU FASTER-Initiative: Metadatenbankformat zur Integration von Mikro- und Makrodaten in einheitlicher Systemarchitektur mit kontrollierten Attributen. Die Minimalanforderung zur Dokumentation von Mikro- und Makrodaten bildet Dublin Core (DC). Das DC Format enthält bspw. Titel, Urheber, Fundstelle, thesaurusbasierte Schlagworte etc

- **DSZ-Studien und Variablen Thesaurus**

  Zunächst - für die zwanzig wichtigsten - soll jede Studie sowie jede einzelne Variable durch einen Thesaurus recherchierbar katalogisiert werden. Der Thesaurus ist bereits in der Entwicklung und als Pilotanwendung verfügbar. Er nutzt den EUROVOC, UNESCO, OECD, JEL, ILO sowie den DDI-kompatiblen CESSDA- bzw. ICPSR-Thesaurus. Die IZA Initiative, den HASSET-Thesaurus aus ESSEX zu nutzen, blieb bislang erfolglos, er soll für das ZA übersetzt werden.

  ⇨ offen, ob DDI anerkannt und umgesetzt von FDZ/DSZ

Faster
http://www.faster-data.org/

DDI
www.icpsr.umich.edu/DDI/

GESIS

Freigegeb./ergänzt 10.1.03 IZA

Jk

IZA: Informationsverdichtung und Integration

Grundelemente der Informationsverdichtung DSZ

- **DSZ-Mikro-und Makrodatenbank**

Integrierte WWW-Oberfläche (Pilot) zur Recherche von Studien-inhalten und Variablen sowie Abbildung und Analyse von Mikro- und Makrodaten. Die Analysemöglichkeiten umfassen einfache deskrip. bis hin zu ökonometrischen Analysen. Zunächst werden die Programmpakete STATA, SPSS und SAS unterstützt.

Grundelemente zur Integration DSZ

Der DC wurde bereits für den Datenbestand des IZA SQL-DB umgesetzt und wird weiter gepflegt. Die Abbildung des DC und der SQL-DB im Internet erfolgt via PHP.

⇨ Planung und Vorschlag Studieninhalte in XML aufzubereiten

DC besteht bereits als Sammlung von XML-Tags, die den Inhalt im Internet recherchierbar macht, DDI hat solche XML-Tags für jeden Aspekt Mikro/Makro incl. Mapping zu DC verfügbar gemacht.

IZA-Metadatabase Pilot

http://metadata.iza.org/

DDI

www.icpsr.umich.edu/DDI/

DDI DC best example

http://govpubs.lib.umn.edu/cbp/ddielements.phtml

GESIS

Freigegeb./ergänzt 10.1.03 IZA

Jk

IZA: Zugangsmodelle

Unter Einhaltung und kooperativer Erweiterung gegebener rechtlicher Nutzungsbestimmungen.

- **Schalterstellenmodell**

Übermittlung von Analyseskripten, Auswertung am DSZ durch deren Mitarbeiter, Ergebnisse zurück an Wissenschaftler senden

- **Gastwissenschaftlermodell**

  In speziell gesicherten Rechnern im DSZ arbeiten nutzungsberechtigte Wissenschaftler als Gäste

Vorraussetzung:
Nutzervereinbarung mit Primärdatendistributoren und FDZ
Standardisierung der Kontrollmechanismen mit z.B. IAB Betriebspanel (mit Nutzerkontrolle, Script- und Analyseprüfung, Zellprüfungen, einschließlich Unterdrückung von Ergebnissen etc.)

⇨ Proprietäre Daten werden nicht weitergegeben und stehen nicht zum Download zur Verfügung

BA: FDZ der Bundesanstalt für Arbeit

Ziel:

- Alle Daten im vollen Merkmalskanon (schwach anonymisiert) als Grundgesamtheit zur Verfügung
- Für alle Wiss. uneingeschränkter Zugang
- Ausbau des Schalterstellenmodells
- **Forschungsprojektdatenbank mit Ergebnissen** der bereitgestellten Daten; Expertenvermittlung
- Keine dezentrale Lösung der Datenaufbereitung und Datenbereitstellung

BA: Arbeitsgrundlagen Schalterstellenmodell

- Codebook incl. Grundauszählungen unter www.betriebspanel.iab.de
- Bisherige Fragebögen, die gedruckt beziehbar
- Testdaten zur externen Programmprüfung für Datenfernverarbeitung (= Paket von Dateien, mit deren Hilfe funktionsfähige Auswertungsprogramme geschrieben werden können; imitieren exakt Strukturen)
- Für Auswertung: SPSS, STATA ....

BA DataWarehouse-Projekt:

Integriertes Modell für die Daten aus den verschiedenen operativen Verfahren; fusionierte integrierte Basis auch für externe Wissenschaftler

Aber: noch keine Schnittstelle für Datenextraktion + stichtagsorientiert (z. B. nicht für Datenaufbereitung Längsschnittstudien)

BA: Stand IT des IAB: pallas.stat

Moderne IT-Struktur zur performanten Erschließung, Aufbereitung, Analyse und Abfrage sehr großer Datenbest. vorhanden (Basis relationale DB)

- Klassische statistische Analyse auf Solaris (Unix): „ZeUS“: SAS, SPSS, STATA, TDA
- Online-Analysesystem auf Windows-Plattform: für online-Abfragen und –analysen aggregierter Daten über verschiedenste Datenquellen

BA: Dokumentation

- Zusätzliche Datenbeschreibungen nötig für externe Wissenschaftler wie: Erhebungskontext, Erhebungs-bogen, Algorithmenbeschreibung, Bereinigungsschritte ....
- Datenbestand derzeit sehr heterogen, gibt noch keine vereinheitlichte DB; nur selten gleiche Identifikatoren (Schlüsselmerkmale), ungleiche Zeitraster + Periodizität

  Z. B.: Man will Erwerbsbiografie von Personen durch Längsschnittdaten abbilden ⇨geht nur über Verknüpfung: Beschäftigungsstatistik, Arbeitslosen-statistik, Maßnahmenstatistik

Forschungsdatenzentrum des Statistischen Bundesamtes

Ziele optimaler Nutzerversorgung:

- **FDZ** als One-Counter-For-The-Customer erste Kontaktstelle für potentielle Nutzer amtl. Mikrodaten
- Bereitstellung Public Use Files + Scientific Use files + Gastaufenthalte (für nicht-anonymisierbare Daten)
- Kontrollierte Datenfernverarbeitung
- Standardisierung der Datendokumentation in der amtlichen Statistik

  ⇨ zugeschnitten auf meistgenutzte Software extern. Wiss.

  ⇨ vollständige Metadaten für Mikrodaten; Metadatenbank
- Veredelung von Mikrodaten: z. B. Analyseroutinen, Modellspezifikationen
- Daten möglichst zeitnah zur Verfügung, neue Produkte

GESIS

Jk

StBA: Netzwerk FDZ, Kooperationen

- Aufbau **Netzwerk aller FDZ als eigenes Angebot**, „das den Zugang zu einem sehr großen Teil dezentral erhobener und gehaltener Einzeldaten ermöglicht“
- Kooperation/Abstimmung mit DA, DSZ, EUROSTAT
- Mitarbeit bei der Entwicklung „virtueller Mikrodaten-banken“ und der hierfür notwendigen Datenstandar-disierung

StBA: Modelle Datenfernverarbeitung

= empirische Forschung mit Orginaldaten, ohne dass externe Wissenschaftler mit den geheimhaltungsbedürftigen Einzeldaten in Kontakt kommen; immer wenn kein Scientific Use File vorliegt

- Sonderaufbereitung durch StBA im Auftrag externer Wissenschaftler (= gängige Praxis)
- In vereinbartem Standard (SAS, SPSS, STATA) geschriebene Auswertungsprogramme: Erstellung durch externe Wissenschaftler, gerechnet bei StBA im abgeschottetem Bereich, Ergebnisse an Wiss., ggf. modifiziert nach Geheimhaltungsvorschriften
- Wie b) aber alle Prozesse automatisiert (ohne Eingriff Personal StBA) wegen Geheimhaltung ⇨ offen, ob hierzu verallgemeinerbare Lösungen möglich (später)

StBA: IT

- Bereitgestellt SAS als Serversoftware sowie SAS, SPSS und STATA (lokal) als in der Wissenschaft intensiver genutzte Software
- Nutzung gemeinsamer Daten- und Metadatenserverstruktur mit dem FDZ der Länder

Metadaten-Informations-System

für die

Forschungsdatenzentren

der

Statistischen Ämter

des Bundes und der Länder

Ausgangssituation

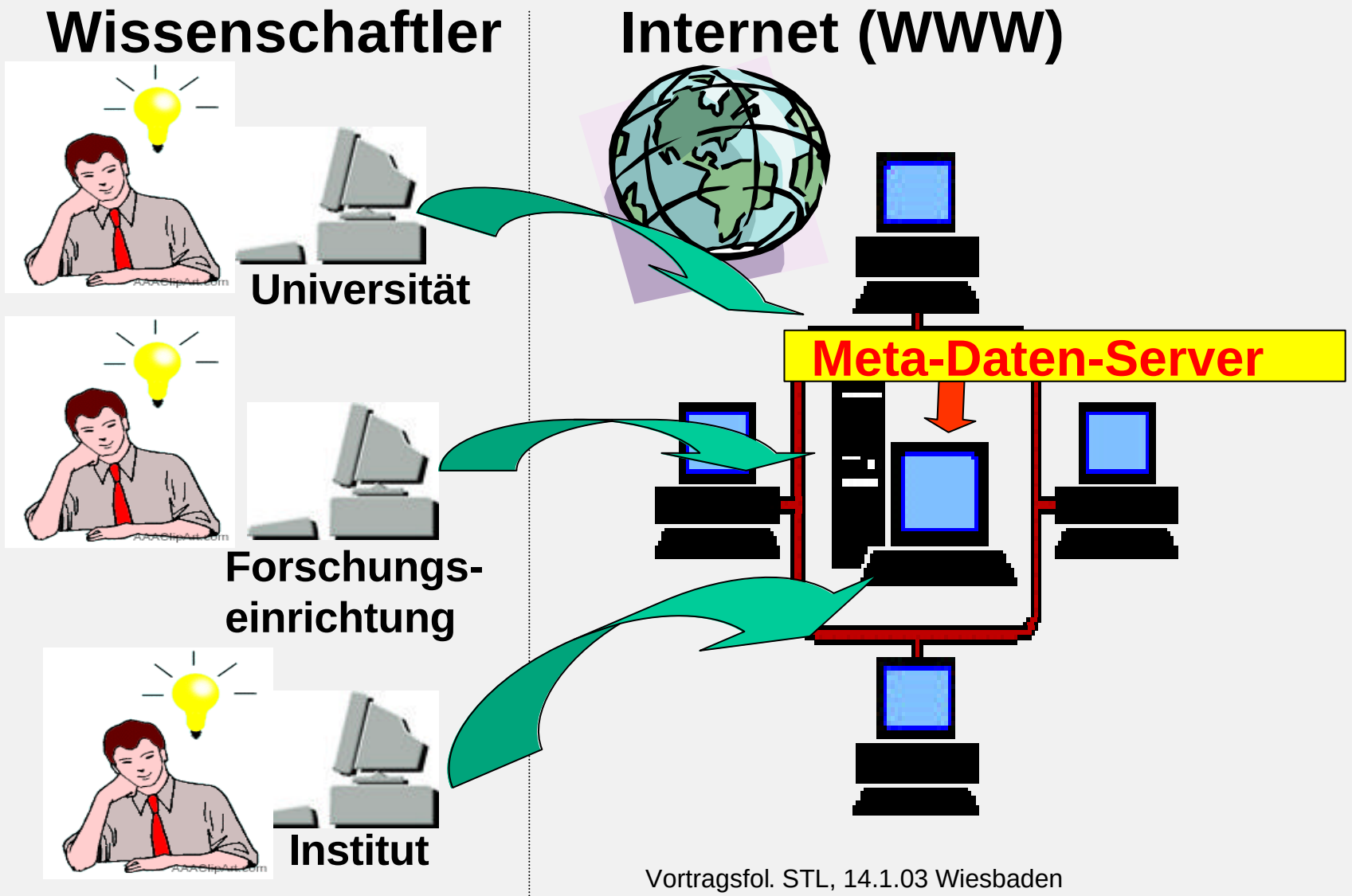
- föderale Organisation der amtlichen Statistik
- rd. 300 verschiedene Einzelstatistiken insges.
- Themenbereiche:
  Bevölkerung, Gesundheitswesen, Bildung, Produzierendes Gewerbe, Handel, Gastgewerbe, Verkehr, Sozialleistungen, Umwelt usw.
- dezentrale und technisch heterogene Metadaten

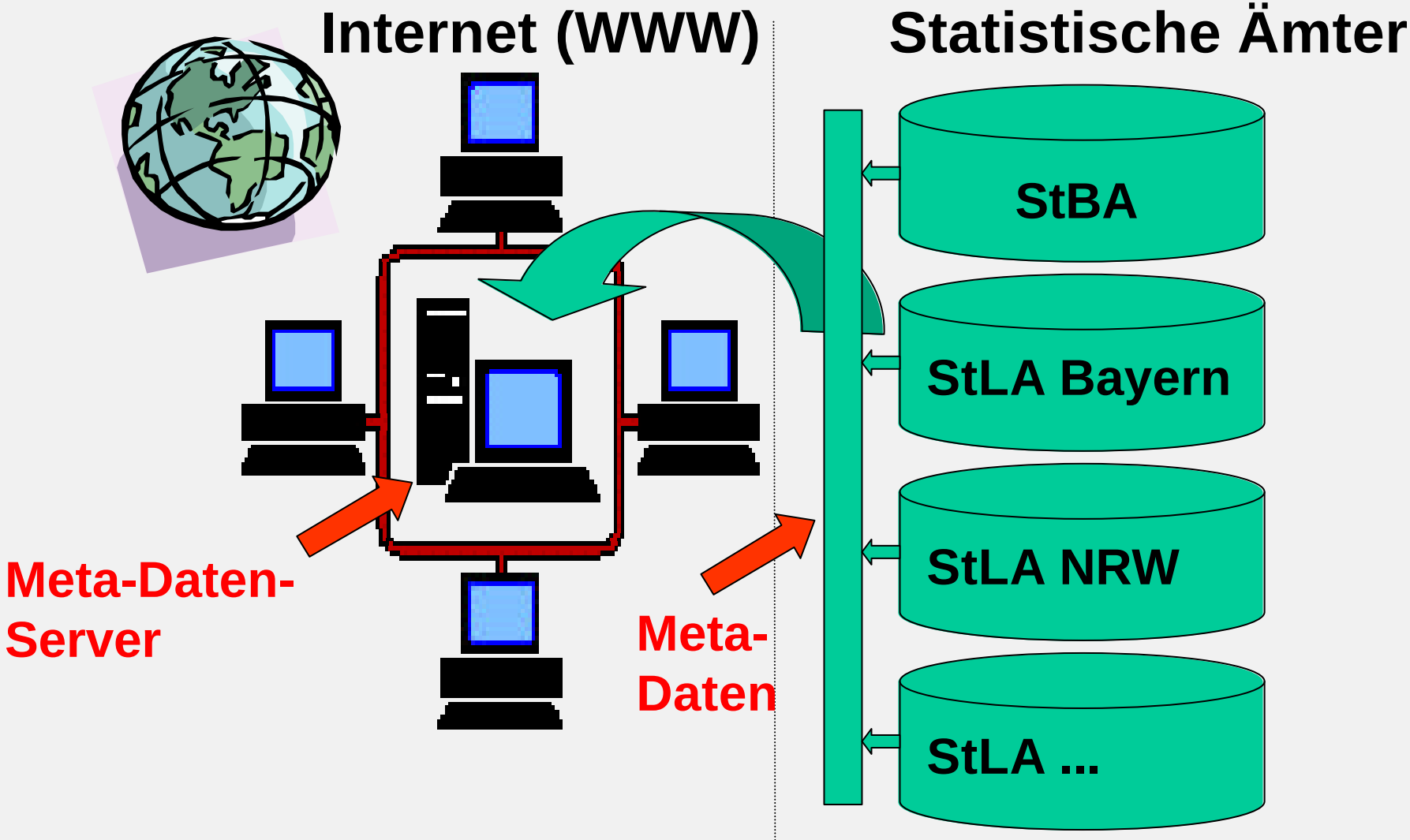
Ziele

- zentrale und nutzerfreundliche Recherche von Metadaten über das WWW

- Kompatibiltät mit Analyse-Software SPSS und SAS sowie Schnittstellenstandard XML

- Metadaten-Import bzw. -Export mit anderen in der amtlichen Statistik eingesetzten Systemen

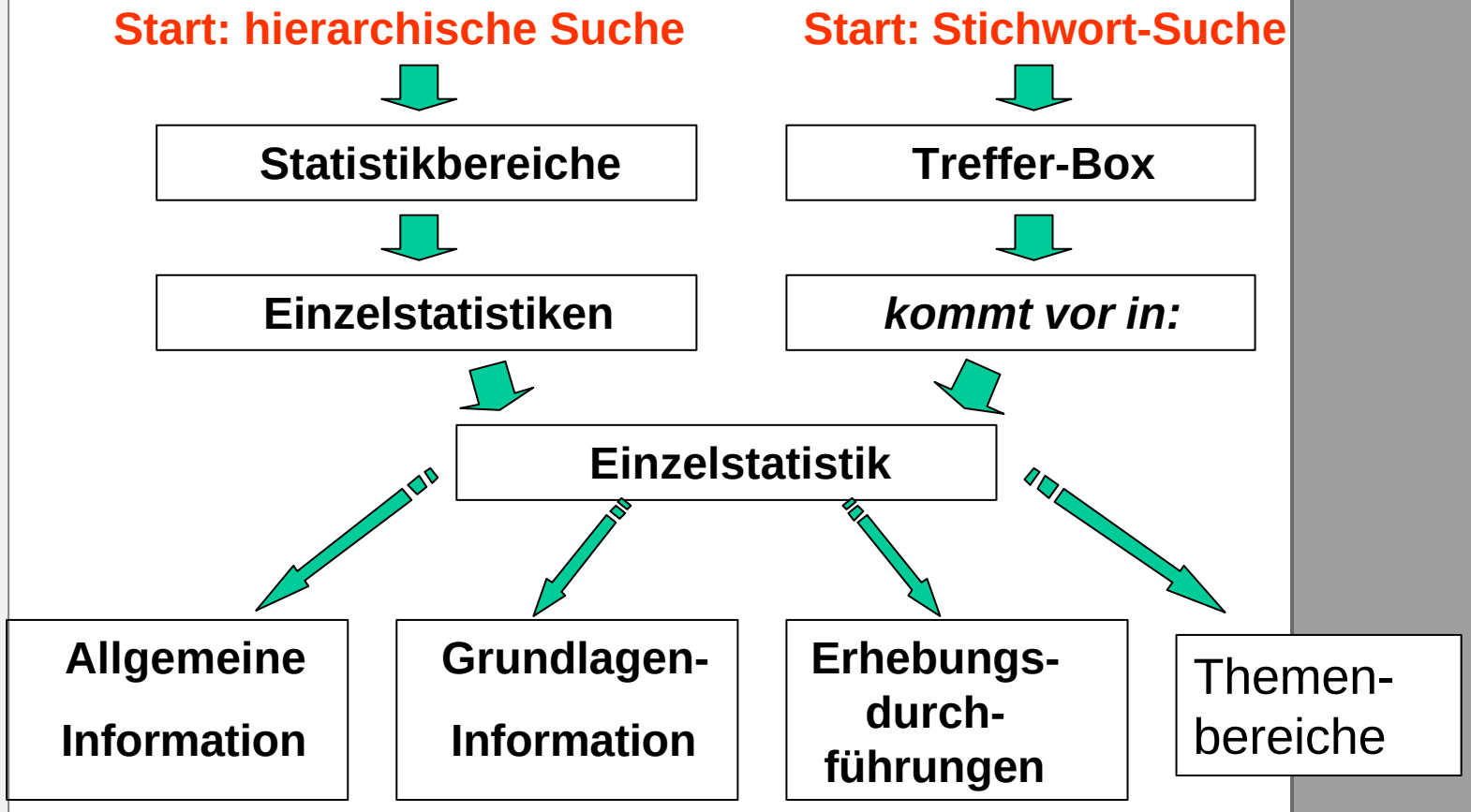

Wissenschaftler
Internet (WWW)
Universität
Meta-Daten-Server
Forschungs-
einrichtung
Institut


Internet (WWW)
Statistische Ämter
StBA
StLA Bayern
StLA NRW
StLA ...
Meta-Daten-Server
Meta-Daten

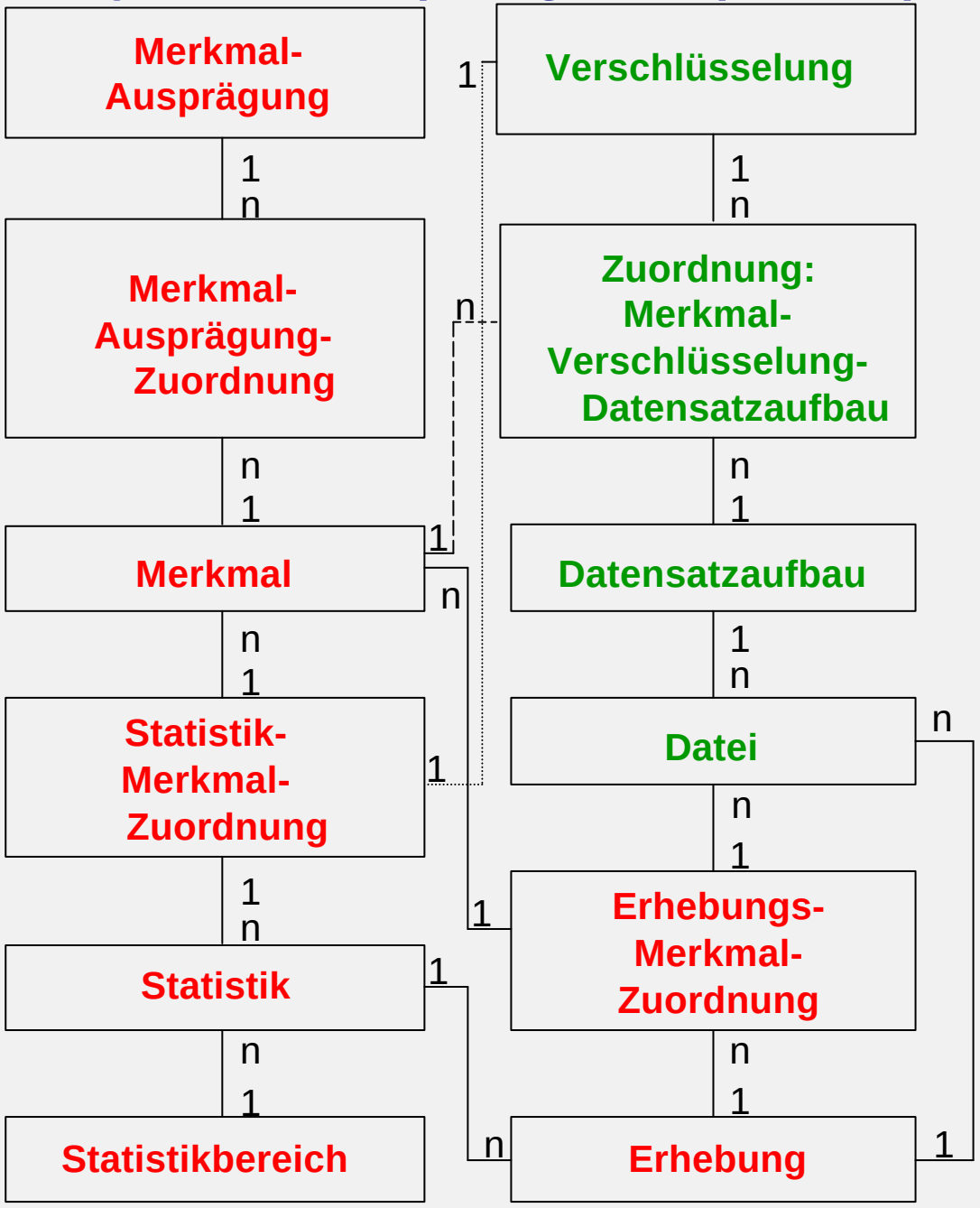
Inhalte - Informationsebenen:

- Grundsätzliche Metadaten zur Informationsebene „Statistik“
- Metadaten zur Informationsebene „Erhebungsdurchführung“
- Metadaten zur Informationsebene „Datei“
- Administrative Informationen

Suchfunktion - Ebenen



- Entity-Relationship-Diagramm (Entwurf) -



Planung für 2003:

- Erstellung eines Feinkonzeptes
- Programmierung der Informations-Module
- Import von Metadaten für priorisierte Statistiken

ZUMA: MISSY

- Informationen von ZUMA noch nicht freigegeben

Offenes Problem: XML-Datenbank als Basis

- DDI/XML legt nahe: auch Abspeicherung XML-DB
- Wahrscheinliche **Vorteile**: Einsparung Transformationsschritt + kürzere Bearbeitungszeiten bei Abruf
- Evtl. **Probleme:**

⇨ verfügbare Produkte auf große Zahl relativ kleiner Dokumente ausgelegt, hier umgekehrt (mehrere Megabyte pro Datensatz) = evtl. ungünstig bei selektivem Zugriff auf Einzelwerte

⇨Offen: Erhaltung der Datenkonsistenz (rel. DB durch Fremdschlüssel, XML durch Schemata)

⇨Prüfen: Suchmöglichkeiten, Zugriff auf Daten unterhalb Dokumentebene (Variablen)

GESIS

Jk

ISSP-Wizard und XML

= Erhebungsdaten der am International Social Survey Program teilnehmenden Länder mit ihren spezifischen Setups auf das Standardsetup des ZA abzubilden.

⇨ Variablen und Werte aufeinander abbilden und auf Variablenebene regelbasierte Plausibilitätsprüfungen und Datenmanipulationen durchführen.

⇨ Relationale DB: effizienter Zugriff auf Einzelwerte + datenbankseitig Konsistenzprüfungen und Datenmanip.

Deshalb: für XML Exportfunktion des ISSP DataWizards nutzen, um die XML-Dokumente zu erzeugen.

Weiterer Vorteil: Der ISSP DataWizard bündelt – im Gegensatz zu datenbankspezifischen SQL-Skripten – das konzeptuelle Wissen über die Datenhomogenisierung in einem externen Modul ⇨über JDBC Standards gleichzeitig auch unabhängig von der konkret eingesetzten DB. Da in Java programmiert, können eine Vielzahl von Technologien /Verfahren eingebunden werden (bis zur Regelverarbeitung), die mit SQL alleine nicht möglich wären.

Zu XML-DB alternative Architektur mit relat. DB

- Speicherung Meta- und Erhebungsdaten in relat. DB + Einbindung in Internet z. B. durch Java Servlets
- Vorteil: Nutzen leistungsfähige Suchfunktionalität auf Daten mit on-the-fly Generierung dynamischer Webinhalte (beschränkt auf ANSI 1992 SQL Standard)
- Externes Repräsentationsformat = XML

  ⇨ Download Daten dann schon in XML; dto. DDI-Dienste

  ⇨ Input Transfer DDI/XML zu rel. DB-Format

  ⇨ Output rel. DB über XSLT konvertierbar auch in PDF, HTML, alternative XML-Schemata (DC, ...)