




Bereinigung der Personendaten an der Universität Bielefeld

Frank Klapper
CIO-IT, Universität Bielefeld
Halle, 13.02.2007



1



Gliederung

- Einige Informationen zum Bielefelder Identity Management Projekt
- Grundsätzliches zur Datenbereinigung
- Statistische Verfahren zur Datenbereinigung
- Ergebnisse der Datenbereinigung
- Initiales Laden des Identity Management Systems



2

Gliederung

- Einige Informationen zum Bielefelder Identity Management Projekt
- Grundsätzliches zur Datenbereinigung
- Statistische Verfahren zur Datenbereinigung
- Ergebnisse der Datenbereinigung
- Initiales Laden des Identity Management Systems



3

Ausgangssituation

- In BI gibt es gut funktionierende Benutzerverwaltungen
 - BenVW im HRZ
 - Ca. 70.000 Personensätze
 - SIS in der Bibliothek
 - Ca. 80.000 Personensätze
 - ...
- Die Datenqualität ist das Hauptproblem
 - Kein (automatischer) Abgleich
 - Manuelle (nicht sauber definierte) Prozesse
 - kein konsequentes Löschen von „alten“ Einträgen



4

Projektidee

- Entwurf und Implementierung von Prozessen, um
 - Personendaten aus verbindlichen Quellen (Mitarbeiterdatei, Studierendendatei, Gästedatei, Selbstbedienung,) **zusammen zu führen.**
 - Alle Nutzer automatisiert mit den Systemzugängen zu versorgen, die sie zur Ausführung ihrer Tätigkeit benötigen.
- Entscheidung für eindeutige Identitäten
 - **Jede real existierende Person soll im Identity Management nur einmal vorhanden sein.**



5

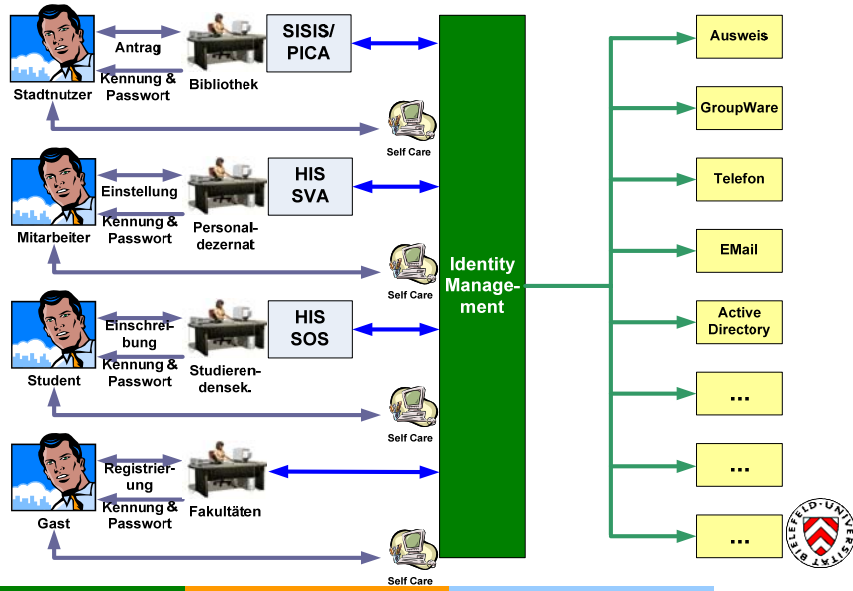
Warum eindeutige Identitäten ?

- Sicherheit
 - Sperrung von Nutzern im Missbrauchsfall
 - Wirtschaftlichkeit
 - Pro User abgerechnete Lizenzen, Kosten für Digitale Zertifikate
 - **Benutzbarkeit**
 - Keine Pflege redundanter Daten
 - Eine Kennung und ein Kennwort für alle Dienste
 - „Dubletten“ im Identity Management führen dagegen zu mehreren Kennungen eines Nutzers für einen Dienst
 - mehrere Emailadressen, mehrere Home-Directories, mehrere elektronische Geldbörsen, ...
- „Dubletten“ sind schwer zu handhaben, überfordern die meisten Nutzer und **verhindern die Systemintegration**



6

Architektur



7

Entscheidung für eine eindeutige Personen-ID

- **Uni-ID**
- Vergabe durch das identity Management (TIM)
- Eigenschaften
 - Lebenslang gültig
 - Unabhängig vom Personenstatus
 - **Ändert sich nie**
- Primäre Einsatzfelder:
 - Personenidentifikation
 - Insbesondere in Identitäten verwaltenden Geschäftsprozessen
 - Datensynchronisation
 - Zwischen den Quellsystemen



8

Gliederung

- Einige Informationen zum Bielefelder Identity Management Projekt
- Grundsätzliches zur Datenbereinigung
- Statistische Verfahren zur Datenbereinigung
- Ergebnisse der Datenbereinigung
- Initiales Laden des Identity Management Systems



9

Entscheidung über den Zeitpunkt der Datenbereinigung

- Bereinigung der Daten
 - vor dem ersten Laden in das Identity Management System
 - direkt in den Quellsystemen
- Später: Housekeeping-Funktionen im Identity Management System vorsehen



10

Quellsysteme

- Quellsysteme in Bielefeld:
 - Studierendenverwaltung (HIS-SOS)
 - Mitarbeiterverwaltung (HIS-SVA)
 - Bibliothekssystem (SISIS)
 - Gästeverzeichnis (BIS)
- Personen sind zum Teil **gleichzeitig** in mehreren Quellsystemen **erfasst**.
- Es gibt keine systemübergreifende eindeutige ID.
- Aber: **Jede real existierende Person soll im Identity Management nur einmal vorhanden sein.**



11

Mapping

- Zuordnung von Datensätzen auf der Basis existierender Attribute
 - Vorname, Nachname, Geburtsdatum, Geschlecht
 - innerhalb jedes Quellsystems und Quellsystem-übergreifend
- Vorab: Daten vergleichbar machen, d.h. Beseitigung von Unstimmigkeiten in den Quellsystemen
 - (1) Erstellung einer einheitlichen Definition der Datenfelder und Bereinigung der davon abweichenden Daten
 - (2) Hinweise auf „schlechte“ Daten durch statistische Verfahren
 - falsche Daten
 - Dopplerund anschließende manuelle Nachbearbeitung



12

Ursachen für „schlechte“ Daten

- Tippfehler
- Buchstaben-/Zahlendreher
- Übertragungsfehler bei der Erfassung
- Inkonsistente Angaben durch die Person selbst
 - vor allem bei Vornamen
- Veralterte Informationen, abweichende Updates
 - besonders problematisch bei Kontaktinformationen
- Fehlende Informationen
- Unterschiedliche Feld-Interpretationen
- ...



13

Bedeutung des Geburtsdatums zur Identifikation von Personen

– Anzahl „Normnamen“ unterschiedliche Geburtsdaten

58.652	1
1.066	2
144	3
35	4
13	5
6	6
2	8
1	9
1	11

➤ Es gibt mehr als tausend Fälle, in denen verschiedene Personen identische Namen (Vorname & Nachname) haben



14

Richtlinie für die Erfassung von Personendaten

- Ein Personenname setzt sich aus den Feldern „Vorname“, „Namenszusatz“ und „Nachname“ zusammen.
- Die Schreibweise ist entsprechend der amtlichen Schreibweise zu wählen.
- Namenszusätze, wie „von“, „de“, „van der“, „de la“ usw. werden in das Feld Namenszusatz eingetragen.
- Für weitere Zusätze, wie Dr., Prof., Freiherr,... muss das Feld „Titel“ benutzt werden.
- Namensergänzungen, wie gen., genannt, jun., sen.,... werden mit einem Leerzeichen getrennt hinter dem Nachnamen aufgenommen.
- Sollte kein Vorname existieren, so ist dieses Feld leer zu lassen.
- Es darf kein Komma in den Namensfeldern vorkommen.

➤ manuelle Nacharbeit in allen Quellsystemen



15

Beispiele für die Erfassung von Personendaten

Klaus Dieter Mayer	Vorname Namenszusatz Nachname	Klaus Dieter Mayer
Pelé	Vorname Namenszusatz Nachname	 Pelé
Pedro de la Rosa sen.	Vorname Namenszusatz Nachname	Pedro de la Rosa sen.
Johann Freiherr von Olpe zur Linde	Vorname Namenszusatz Nachname	Johann Von Olpe zur Linde
Dr. Michaela Zander gen. Forelle	Vorname Namenszusatz Nachname	Michaela Zander gen. Forelle
Zuyj L'Ambert	Vorname Namenszusatz Nachname	Zuyj L' Ambert



16

Gliederung

- Einige Informationen zum Bielefelder Identity Management Projekt
- Grundsätzliches zur Datenbereinigung
- **Statistische Verfahren zur Datenbereinigung**
- Ergebnisse der Datenbereinigung
- Initiales Laden des Identity Management Systems



17

Vorgehen

- Einsatz einer Hilfsdatenbank
 - Arbeitstitel in BI: „Drehscheibe“
- Arbeit mit einem reduzierten Satz an Attributen
 - Vorname, Nachname, Geschlecht, Geburtsdatum, Quell-Id
 - Überführung von Vornamen und Nachnamen in eine normierte Darstellung („Normname“)
- Erstellen von Statistiken, die auf mögliche „schlechte“ Daten hinweisen.
 - mit Hilfe von SQL-Befehlen
 - manuelle Überprüfung und evtl. Nachbearbeitung von „schlechten“ Daten



18

Normnamen

- Normierte Darstellung von Namen
- Notwendig für Vergleichsoperationen
- Regeln:
 - Umwandeln in Kleinbuchstaben
 - Umwandeln von Umlauten, ß, ...
 - Satzzeichen, bekannte Abkürzungen (Dr., Prof.) löschen
 - Leerzeichen löschen
 - ...



19

Statistische Auswertungen: Basis-Statistiken

- Ermittlung der Anzahl der Datensätze pro Quellsystem
 - zum Teil auch Unterscheidung nach Status
- Entfernen verwaister oder veralteter Datensätze
 - Stadtnutzer der UB mit vieljähriger Inaktivität
 - Personen mit Mitarbeiterausweis in der UB, aber ohne SVA-Eintrag
 - ...



20

Statistische Auswertungen: Schreibfehler in Namen

- Verwendung von Normnamen
- Ähnlichkeitsalgorithmus auf Basis der Levenshtein-Differenz
 - Datenbasis: alle Quellsysteme
- Überprüfung aller Datensätze mit einer Ähnlichkeit $> 0,7$
 - sehr optimistische Annahme für die Gleichheit
- Lieferung von Paaren potentiell gleicher Namen an die Quellsysteme zur manuellen Überprüfung
 - Basis der Überprüfung: Aktenlage



21

Statistische Auswertungen: Schreibfehler im Geburtsdatum

- zunächst: Test auf unsinnige Geburtsdaten
- dann: Ähnlichkeitsalgorithmus auf Basis der Levenshtein-Differenz
 - für identische Normnamen
 - Datenbasis: alle Quellsysteme
- Lieferung von potentiellen „schlechten“ Daten an die Quellsysteme zur manuellen Überprüfung
 - Basis der Überprüfung: Aktenlage



22

Statistische Auswertungen: Unterschiedliches Geschlecht

- Unterschiedliches Geschlecht für identische Normnamen
 - Datenbasis: alle Quellsysteme
- Lieferung von potentiellen „schlechten“ Daten an die Quellsysteme zur manuellen Überprüfung
 - Basis der Überprüfung: Aktenlage



23

Statistische Auswertungen: Doppler innerhalb eines Quellsystems

- Identifikation der Datensätze mit gleichem Normnamen und gleichem Geburtsdatum
 - Datenbasis: pro Quellsystem
- manueller Versuch des Mappings innerhalb des Quellsystems
 - es gibt Fälle, in denen ein Mapping aus inhaltlichen Gründen nicht möglich ist

❖ wir haben keinen Fall identifiziert, in dem die Attribute Normnamen, Geschlecht und Geburtsdatum bei **unterschiedlichen** Personen übereinstimmen



24

Gliederung

- Einige Informationen zum Bielefelder Identity Management Projekt
- Grundsätzliches zur Datenbereinigung
- Statistische Verfahren zur Datenbereinigung
- Ergebnisse der Datenbereinigung
- Initiales Laden des Identity Management Systems



25

Anzahl Iterationen

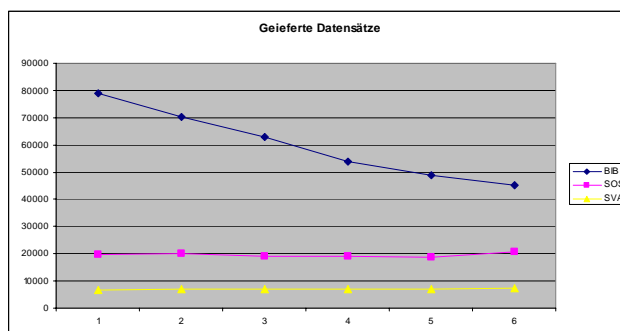
- Wir haben die „Drehscheibe“ sechsmal aufgebaut.
- Anschließend jeweils Datenbereinigung mit spezifischem Fokus.
- ❖ Der Semesterzyklus hat Einfluss auf die Ergebnisse.



26

Letzte statistische Werte – Anzahl Datensätze

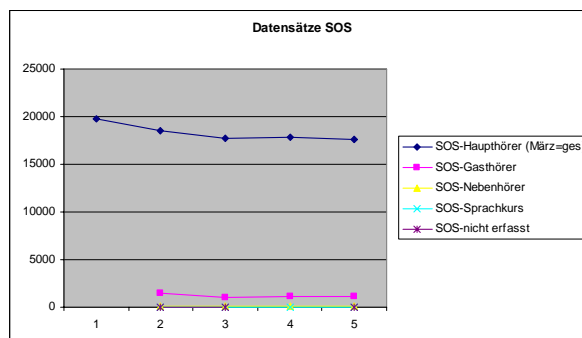
	März	April	Mai	Juni	Juli	November
Bib	78.957	70.292	63.039	53.957	48.950	45.300
SOS	19.764	20.134	18.917	19.008	18.868	20.658
SVA	6.671	7.024	7.085	7.130	7.157	7.270
Summe	105.392	97.450	89.041	80.095	74.975	73.228



27

Letzte statistische Werte – Datensätze SOS

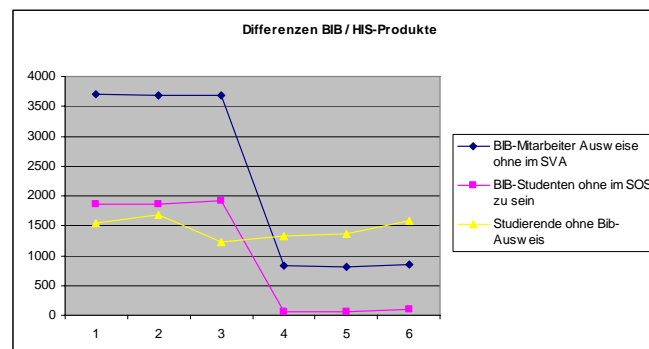
	März	April	Mai	Juni	Juli	November
Haupt Hörer	19.764	18.559	17.758	17.758	17.640	19.288
Gast Hörer		1.467	1.076	1.130	1.140	1.325
Neben Hörer		90	67	73	71	3
Sprachkurs		17	16	17	17	41
Nicht Erfasst		1	0	0	0	1



28

Letzte statistische Werte – Differenzen BIB - HIS

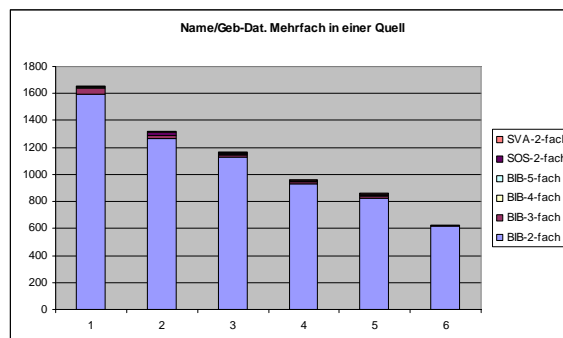
	März	April	Mai	Juni	Juli	November
BIB-MA ohne SVA Eintrag	3.700	3.689	3.690	822	804	848
BIB-Std. ohne SOS Eintrag	1.860	1.857	1.919	55	64	106
Studis ohne BIB-Ausweis	1.536	1.686	1.230	1.334	1.372	1.590



29

Letzte statistische Werte – interne „Doppler“

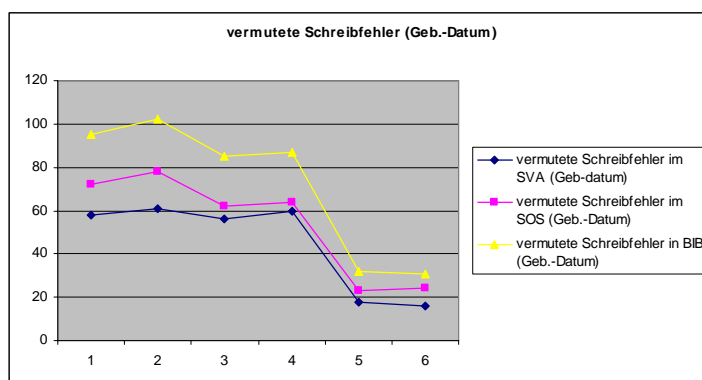
	März	April	Mai	Juni	Juli	November
BIB – 2-fach	1.593	1.264	1.127	928	826	620
BIB – 3-fach	47	23	19	18	16	4
BIB – 4-fach	6	4	4	4	4	0
BIB – 5-fach	2	1	1	0	0	0
SOS – 2-fach	4	17	5	6	6	15
SVA – 2-fach	5	8	8	8	8	1



30

Letzte statistische Werte – vermutete Schreibfehler Geburtsdatum

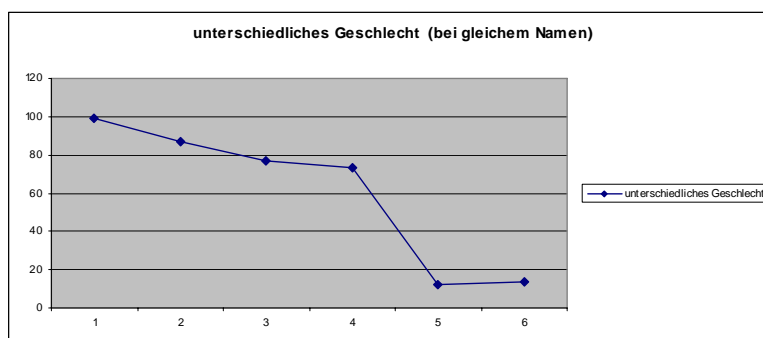
	März	April	Mai	Juni	Juli	November
Bib	95	102	85	87	32	31 (14)
SOS	72	78	62	64	23	24 (9)
SVA	58	61	56	60	18	16 (3)



31

Letzte statistische Werte – unterschiedliches Geschlecht

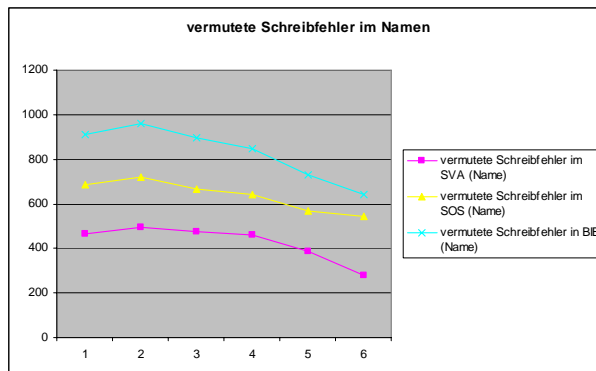
	März	April	Mai	Juni	Juli	November
Untersch. Geschlecht	99	87	77	77	12	14 (4)



32

Letzte statistische Werte – vermutete Schreibfehler im Namen

	März	April	Mai	Juni	Juli	November
Bib	911	959	896	845	731	640
SOS	684	718	664	640	569	543
SVA	467	497	476	461	389	281



33

Gliederung

- Einige Informationen zum Bielefelder Identity Management Projekt
- Grundsätzliches zur Datenbereinigung
- Statistische Verfahren zur Datenbereinigung
- Ergebnisse der Datenbereinigung
- Initiales Laden des Identity Management Systems



34

Generierung der Uni-ID: Verfahren

- Zuweisung in der Drehscheibe
- Bereitstellung über Dateien - mit Tupeln
- Einlesen in die Quellen incl. Füllen der Felder



35

Mergen der Datensätze I

- Während des Initial-Loads
- Berücksichtigt sind:
 - SOS
 - SVA
 - Bibliothek
 - BIS (vergleichbare Datensätze)
- Vordefinierte ‚führende Systeme‘
 - Basierend auf den Aussagen der nächsten Folie



36

Mergen der Datensätze II

- SVA Daten haben immer Gültigkeit
- SOS Daten werden in folgenden Attributen von der BIB überschrieben:
 - PLZ
 - Ort
 - Strasse, Hausnummer
 - Adresszusatz
 - Länderkennung
 - Kontakt e-Mail
- BIS Datenfelder werden (bis auf Titel und die e-Mail Adressen von Nicht-Mitarbeitern) immer überschrieben



37

Identity Management sorgt dafür, dass der richtige Anwender zum richtigen Zeitpunkt die angemessenen Zugriffsrechte erhält.



38