Statistik-Informationssystem zur Analyse von Studienverläufen

Eine Kurzbeschreibung unter Datenschutzaspekten

25. Mai 1993

1. Übersicht

Die derzeitige Situation der Universitäten zwingt dazu, Effizienzuntersuchungen, Qualitätssicherungsmaßnahmen usw. auch auf den Bereich der Lehre anzuwenden. Dies setzt - neben vielem anderen - verläßliche quantitative Informationen über die einzelnen Studiengänge voraus. Die vom Ministerium für Wissenschaft und Forschung im Zusammenhang mit den obligatorischen Lehrberichten abgefragten statistischen Angaben sind Beispiele hierfür.

Für eine wirkliche Qualitätskontrolle sind allerdings sehr viel differenziertere Angaben erforderlich als reine Querschnittsaggregate wie etwa die Anzahl der Absolventen pro Semester o.ä. Man benötigt vielmehr Verlaufsstatistiken über mehrere Jahre hinweg, um etwa feststellen zu können, in welcher Reihenfolge und in welchen zeitlichen Abständen Prüfungen absolviert werden, wie lange Studierende auf Praktikumsplätze warten müssen, welche Prüfungen wie oft und in welchen zeitlichen Abständen wiederholt werden usw. Eine Längsschnittbetrachtung ist auch deswegen unabdingbar, weil nur so festgestellt werden kann, welche Studienplanänderungen (die ja in der Regel mit dem Ziel der Studienzeitverkürzung eingeführt wurden) sich wann und wie ausgewirkt haben.

Derartige Längsschnittanalysen, die ein unabdingbares Instrument zur Steigerung der Effizienz des universitären Lehrbetriebes sind, erfordern vor allem eines: Es muß möglich sein, die Prüfungsergebnisse eines Studierenden *als zum selben Individuum gehörig* zu erkennen. Zur Messung der längerfristigen Auswirkung von Änderungen im Studienplan bzw. der Prüfungsordnung ist es darüber hinaus erforderlich, auch Prüfungsergebnisse solcher Studierenden für statistische Zwecke vorrätig zu halten, die exmatrikuliert wurden, sei es, weil sie ihr Studium abgeschlossen haben, sei es, weil sie die Hochschule gewechselt oder ihr Studium abgebrochen haben.

In den Verwaltungsdatenbeständen für das Personal- und Prüfungswesen erfolgt die Zuordnung der einzelnen Prüfungsereignisse und Personaldaten zur selben Person über die Matrikelnummer. Diese Matrikelnummer ist in den entsprechenden Datenbeständen auch dem Namen, der Anschrift usw. zugeordnet, dient also zur Identifikation einer konkreten Person. Eine Verwendung dieses Zuordnungsschemas zu statistischen Zwecken kommt mithin nicht in Betracht, da dies durch allgemeine Bestimmungen des Datenschutzes ausgeschlossen ist. Eine Speicherung von Prüfungsergebnissen exmatrikulierter Studierender auf dieser Basis verletzt außerdem die spezifischen Bestimmungen des Universitätsgesetzes (§ 125a) und seiner Durchführungsbestimmungen.

Bei dem an der Universität Stuttgart implementierten Prototypen eines Statistik-Informationssystems zur Analyse von Studienverläufen wurde eine Methode gefunden, die einerseits den Anforderungen detaillierter Längsschnittuntersuchungen entspricht, die aber zum anderen auch die Einhaltung aller Datenschutzbestimmungen gewährleistet. Die Grundidee, welche in den folgenden Abschnitten dann näher erläutert wird, ist folgende:

Da für statistische Zwecke sämtliche Angaben, die allein zur Identifikation einer konkreten Person dienen

(Name, Matrikelnummer, Anschrift) irrelevant sind, brauchen diese Daten nicht gespeichert zu werden. Übernommen werden nur die für statistische Auswertungen interessanten Angaben wie Nummer der Prüfung, Prüfungsdatum, Note, Wiederholungsmerkmal, Hochschulsemester, Geschlecht usw. Die Angaben über Prüfungsergebnisse und die sonstigen Merkmale werden also anonymen statistischen Individuen zugeordnet, von denen ein Rückschluß auf konkrete Personen im Prinzip nicht möglich ist.

Das schwierigste Problem besteht nun darin, wie die über die Jahre hin anfallenden einzelnen Prüfungsereignisse, die ja jeweils zu konkreten Personen gehören, dem "richtigen" statistischen Individuum zugeordnet werden können, ohne daß eine rekonstruierbare Beziehung zwischen diesem Individuum und dieser konkreten Person existiert. Die Lösung besteht in einer relativ einfachen Anwendung kryptographischer Methoden, die standardmäßig in Rechnernetzen eingesetzt werden, wenn es darum geht, die Übertragung von Daten gegen unbefugtes "Mithören" zu schützen. Konkret bedeutet dies folgendes: Vor dem Einspeichern in das Statistik-Informationssystem wird die zu einem Prüfungsereignis gehörende Matrikelnummer mit Hilfe einer **Einwegfunktion** in einen sogenannten Ordnungsbegriff umgewandelt. Die Einwegfunktion sorgt dafür, daß verschiedenen Matrikelnummern verschiedene Ordnungsbegriffe zugewiesen werden, doch der Rückschluß von einem Ordnungsbegriff auf die zugrundeliegende Matrikelnummer ist im Einzelfall gänzlich unmöglich und auch bei Vorliegen sehr großer Stichproben nur mit prohibitiv hohem Rechenaufwand zu erzielen. In dem Prototypsystem wird die Einwegfunktion mit Hilfe der Verschlüsselungsvorschrift nach dem sogenannten Data Encryption Standard (DES) realisiert. Dieses Verschlüsselungsverfahren wird zur Datenübertragung bei Banken, im Militärbereich und in anderen Anwendungen mit hohen Sicherheitsanforderungen eingesetzt und gilt nach dem gegenwärtigen Stand der Technik als sehr sicher.

Der für die Anwendung des DES-Verfahrens erforderliche Schlüssel wird nur an einer Stelle verwaltet, und zwar beim Verwaltungs-Informationssystem. Dies ist die Stelle, wo die sensitiven Personaldaten geführt werden, und die von daher ohnehin den höchsten organisatorischen Sicherheitsanforderungen genügen muß. Aus diesem System heraus und in das Statistik-Informationssystem gelangen also nur strikt anonymisierte Daten, die mit dem unumkehrbaren Ordnungsbegriff versehen sind.

Das Statistik-Informationssystem selbst ist auf der Basis eines SQL-Datenbanksystems realisiert, wobei von den flexiblen, wertabhängigen Mechanismen zur Zugriffskontrolle in SQL voller Gebrauch gemacht wird. So ist z.B. durch entsprechende Sichtendefinitionen dafür gesorgt, daß an der Benutzerschnittstelle zur Durchführung der statistischen Auswertungen der Ordnungsbegriff nicht zugänglich ist. Dieser Ordnungsbegriff dient ja nur der systeminternen Verwaltung bei der Einspeicherung neuer Daten; für die Statistik selbst ist er völlig irrelevant und braucht daher an der Schnittstelle auch nicht in Erscheinung zu treten.

Die Benutzerschnittstelle ist außerdem so ausgelegt, daß grundsätzlich keine Einzelinformationen über ein statistisches Individuum ausgegeben werden. Alle Anfrageergebnisse sind Aggregate, die sich auf eine Population von mindestens zehn statistischen Individuen beziehen. Damit ist ausgeschlossen, daß an der Anfrageschnittstelle jemand durch Ausnutzung von (partiellem) Kontextwissen über einzelne Personen eine Zuordnung zu den statistischen Individuen vornehmen kann. Zur weiteren Erhöhung der Sicherheit des Systems ist daran gedacht, die kritische Größe der Ausgabepopulationen nicht mehr konstant auf 10 festzulegen, sondern diese Größe durch eine Zufallszahl mit dem Mittelwert 10 zu ersetzen.

2. Konzeptioneller Aufbau

An der Universität Stuttgart werden zwei unabhängige Systeme geführt, die die nötigen Informationen für Langzeitanalysen der Studienverläufe beinhalten:

- Studenten-Operations-System
- Prüfungs-Operations-System

Beide Systeme sind in ihrer bestehenden Form aus Gründen des Datenschutzes für die genannten Ansprüche nicht geeignet. Bevor eine Konsolidierung der Daten der Systeme stattfinden kann, muß zugesichert sein, daß keine Möglichkeit besteht, auf personenbezogene Daten zuzugreifen. Die Anonymisierung erfolgt in drei Schritten :

1. Ausschreiben der Datenbestände aus den oben genannten Systemen unter Bereinigung personenbezogener Felder wie Name und Adresse. Gleichzeitiges Verschlüsseln der Matrikelnummern unter Verwendung einer Einwegfunktion zur Bildung eines Ordnungsbegriffs.

Dies schließt eine Rückgewinnung der Matrikelnummern aus und garantiert gleichzeitig die Fähigkeit des Systems zur Aktualisierung. Der Schlüssel der Einwegfunktion wird im Rektoramt im Verwaltungsrechenzentrum unter Verschluß gehalten.

2. Auslagern der erstellten Daten auf ein physisch vom Produktionsrechner getrenntes System. Schreiben der Bestände in eine Datenbank mit strikten Zugriffskontrollen.

   Dies hat den Zweck, einen unberechtigten Zugriff **unterhalb** der Anwendungsebene zu verhindern.

3. Erstellen einer Anwendung, die nur vordefinierte Anfragen zuläßt und die aufgrund von unteren und oberen Schranken ein kombiniertes, statistisches *Einkreisen* von Einzelsätzen bei vorhandenem Hintergrundwissen des unberechtigt Anfragenden verhindert.

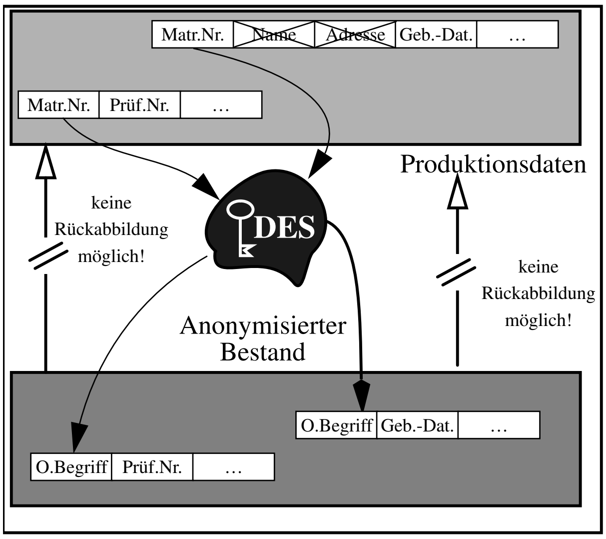
   Diese Vorkehrung ist für die Abwehr sogenannter **Tracker** wichtig. Tracker verfügen über ein gewisses Maß an Hintergrundwissen über die Einzelausprägung der von ihnen gesuchten Personen oder Gegenständen. Dieses Hintergrundwissen kann für den Aufbau von Anfrageserien verwendet werden, um sich so nicht bekannte Einzelinformationen zu verschaffen.

3. Realisierung des Ablaufs

Innerhalb des operativen Ablaufs werden die konkreten Vorkehrungen aus technischer Sicht beschrieben. Die wichtigste Zielsetzung aus diesem Blickwinkel ist, ein Verfahren des Datentransfers und der nachfolgenden Bearbeitung zu entwickeln, das keinerlei unbefugten Zugriff auf die Daten bzw. einen Rückschluß auf die realen Personen zuläßt. Aufbauend auf der strikten Trennung des Produktionssystems und des Informationssystems für die Statistiken gliedert sich der Ablauf wie folgt:

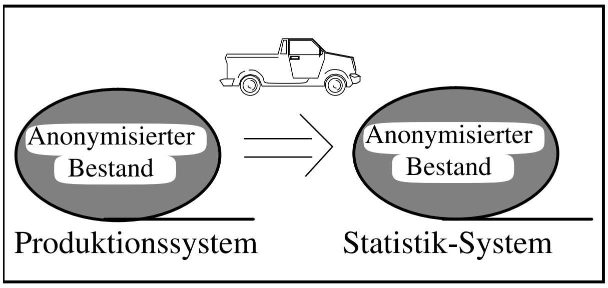
1. Es findet eine Anonymisierung auf der Basis des DES-Algorithmus (data encryption standard) statt. Die statistisch relevanten Daten werden aus den Datenbeständen des Rektoramts selektiert und von nicht benötigten Informationen bereinigt, die Matrikelnummer wird durch die Verschlüsselung ersetzt. Die entstandenen Datensätze werden (derzeit auf der Siemens-Maschine unter BS2000) auf Dateien im ASCII-Format geschrieben.

ABBILDUNG 1. Verschlüsselungsverfahren



2. Die neu erstellten Datenbestände werden noch im Verwaltungssystem auf Bänder ausgelagert, dort aber weiter nicht gespeichert. Diese Bänder werden von einer physikalisch vom Rektoramt getrennten Maschine eingelesen und auf ein UNIX-System transferiert, das eine relationale Datenbank, die für die weitere Bearbeitung zuständig ist, bereit hält.

ABBILDUNG 2. Datentransfer

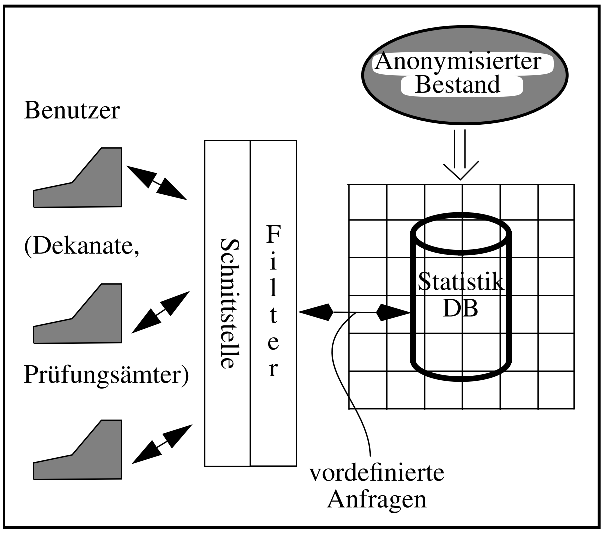


3. Diese Datenbank ist durch eine Reihe von zusätzlich geschaffenen Integritätsregeln so restriktiv aufgebaut, daß der Zugriff auf die Daten nur wenigen Privilegierten gestattet ist; das sind die Bearbeiter, die letztendlich die Statistiken auswerten sollen, momentan ist es nur der Entwickler. Darüber hinaus wird bei Anfragen kein Zugriff auf den Ordnungsbegriff erlaubt.

4. Auf der Ebene der Anwendungsschnittstelle ist dem vorher explizit legitimiertem Endbenutzer nur erlaubt, die vordefinierten Anfragen zu stellen. Dabei werden im wesentlichen Anzeigen in Form von Balkendiagrammen oder ähnlichem zur Verfügung gestellt. Bei numerischen Anfragen gibt es Restriktionen bezüglich der Unter- beziehungsweise Obergrenzen. Diese werden so festgelegt, daß bei

einer Anfrage keine Anzahl kleiner 10 beziehungsweise größer n - 10 herauskommen darf (n steht für die maximale Anzahl der Ausprägungen).

ABBILDUNG 3. Sicherung an den Schnittstellen



Anhand des aufgezeigten Verfahrens wurden die Maßnahmen zur Wahrung der Sicherheitsaspekte aus datenschutzrechtlichen Gesichtspunkten aufgezeigt. Als Anlage werden nun noch die einzelnen, über Bänder transferierten Datenstrukturen aufgelistet. Jede Abbildung zeigt dabei die Struktur der Sätze, wie sie auf den entsprechenden Bändern enthalten sind.

Anlage

ABBILDUNG 1. zeigt Daten der Anmelde- und Notendatei des POS, entspricht den am Prüfungsamt der Universität Stuttgart angemeldeten Prüfungen

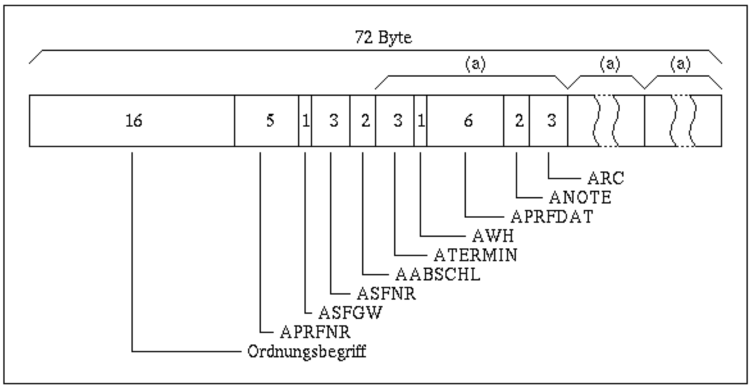


ABBILDUNG 2. zeigt Daten aus der Zeugnisdatei des POS über abgelegte Vor- und Hauptprüfungen

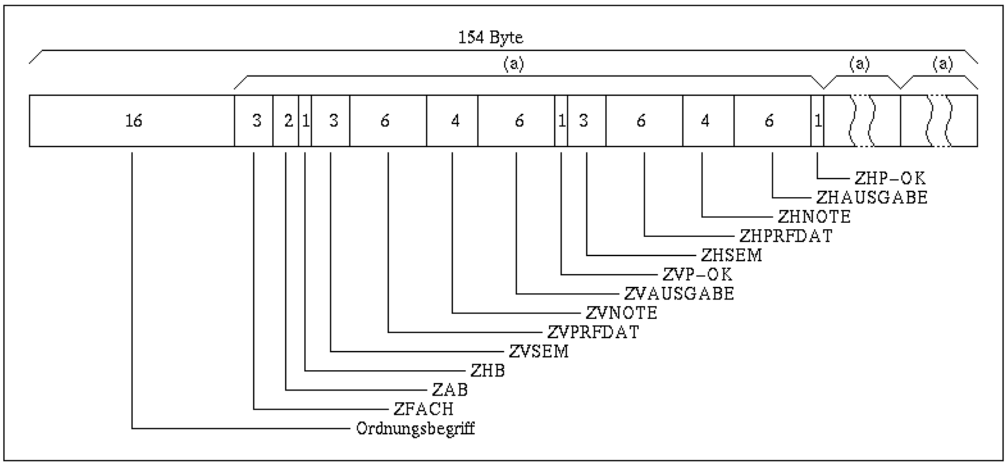


ABBILDUNG 3. zeigt die Zuordnung der Prüfungsnummern zu den Prüfungen

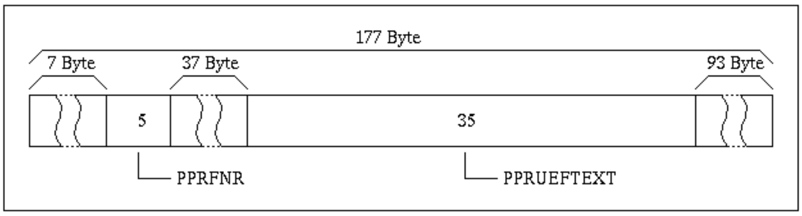


ABBILDUNG 4. enthält Daten der Studierenden, wie Geschlecht, Alter und Staatsangehörigkeit, Immatrikulations- und Exmatrikulationsdatum

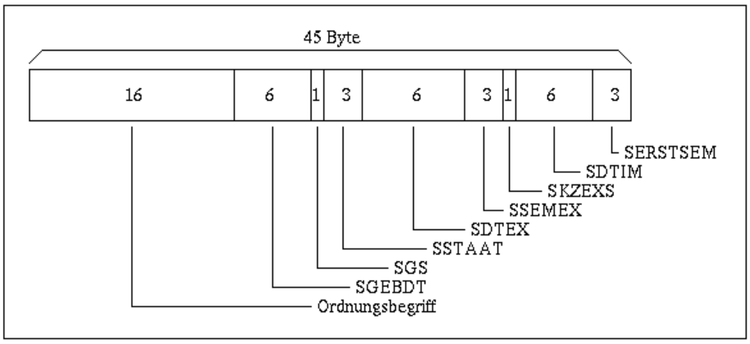


ABBILDUNG 5. zeigt Daten der Vor- und Hauptdiplomsprüfungen

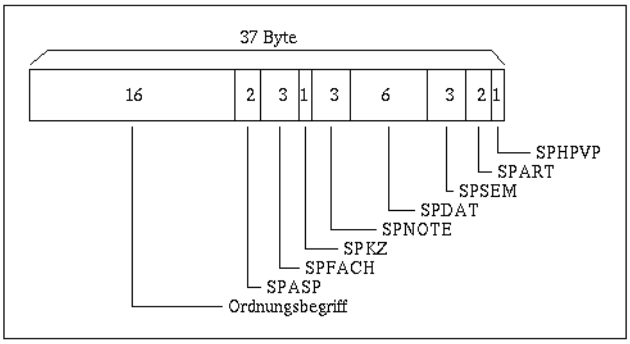
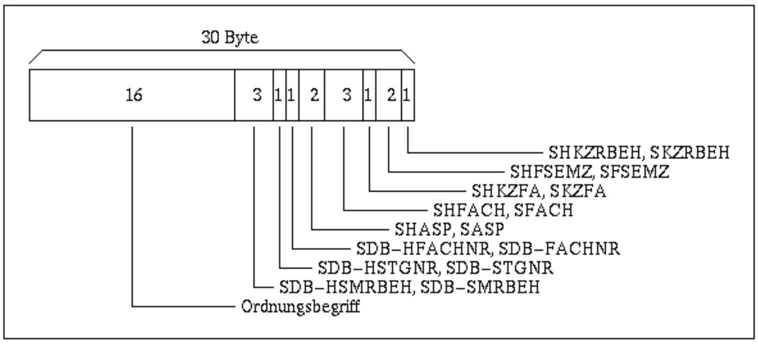


ABBILDUNG 6. zeigt die Studienverläufe der Studierenden auf



Die Kürzel für die Namen der Datenfelder entsprechen den Satzdeklarationen der POS- und SOS-Dateien.