GENESIS:
Sachstandbericht nach Beginn der Arbeiten mit einer Produktionsdatenbank

Die Version 1.0 des Gemeinsamen neuen statistischen InformationsSystems wurde im Juli vergangenen Jahres an alle Statistischen Landesämter ausgeliefert.

Nachdem schon mehrere Artikel über GENESIS erschienen sind, ist es nun soweit, nicht mehr nur über das Konzept zu reden, sondern über die Realisierung. Die Realisierung erstreckt sich nicht nur auf die Programmierarbeiten, sondern auch auf die Bereitstellung der Daten und die ersten Tests unter Produktionsbedingungen.

Im Juli 1997 wurde die GENESIS-Version 1.0 an die Statistischen Ämter ausgeliefert. Im Februar dieses Jahres ist von den Entwicklungsteams die um viele Funktionen ergänzte Version 1.1 freigegeben und ausgeliefert worden. Entwicklungsteams arbeiten in den Ländern Baden-Württemberg, Bayern, Brandenburg und Sachsen. Im Statistischen Bundesamt befindet sich das Entwicklungsbüro, das die Arbeiten nicht nur koordiniert, sondern auch die umfangreichen Zusammenführungen der entwickelten Teile durchführt. Die Länder Niedersachsen und Berlin haben sich nach der Mitarbeit bei der Erstellung des Fachkonzeptes vor Beginn der eigentlichen Programmierarbeiten zurückgezogen.

In dieser ersten Entwicklungsphase wurde der Test des Systems in erster Linie von Mitarbeitern mit informationstechnischer Ausbildung vorgenommen. Dabei ging es im Abnahme- und Gewährleistungstest um die grundsätzliche Funktionalität. Ein System, das so breite Anwendung finden soll wie GENESIS, muß stabil laufen, umfangreiche Plausibilitätsprüfungen enthalten und auf alle Plattformen (Betriebssysteme von MVS, BS2000 und UNIX) portierbar sein.

Für den Test aus fachlicher Sicht wurde im Mai 1997 eine „GENESIS-Nutzergruppe" eingesetzt. Sie hatte und hat den Auftrag, das System aus Anwendersicht zu testen. Ein Zwischenbericht über den Stand der Arbeiten und über die dabei gewonnenen Erfahrungen wurde Anfang 1998 vorgelegt, die Testarbeiten werden mit der neuen GENESIS-Version fortgesetzt. In der Nutzergruppe sind Fachstatistiker aus Hessen, Nordrhein-Westfalen, Brandenburg, Niedersachsen, Rheinland-Pfalz, Baden-Württemberg und Thüringen vertreten. Die Federführung liegt beim Statistischen Bundesamt.

Die Autorin: *Frau Dipl.-Physikerin Madeleine de la Croix ist Referentin im Referat "Anwendungsentwicklung für Landesaufgaben, Dezentrale Datenverarbeitung" des Statistischen Landesamts Baden-Württemberg.*

Die Realisierung aus programmtechnischer Sicht

GENESIS steht gegenwärtig für die Betriebssysteme MVS (IBM), BS2000 (Siemens) und UNIX (SUN) zur Verfügung. Durch den Einsatz der Datenbank ADABAS und der Programmiersprache NATURAL, die Standardwerkzeuge im Verbund sind, wurde damit gewährleistet, daß GENESIS auf all diese Plattformen portierbar und damit in allen Ländern ohne zusätzliche Produkte einsetzbar ist. Das heißt aber auch, daß die Oberfläche zeichenorientiert ist. Rollbalken zur Suche in Listen („Scrollbars"), Schaltflächen für die Maus („Buttons" und „Icons"), automatisch erscheinende Funktionsauswahlen („Pop-up-Menüs") und andere Elemente, die man in vielen graphisch gesteuerten Oberflächen von PC-Produkten kennt, wird man deshalb in der gegenwärtigen Version vergeblich suchen. Doch ist GENESIS vom Konzept her auf eine graphische Oberfläche vorbereitet. Intern ist das System schichtweise modularisiert, so daß eine eindeutige Trennung besteht zwischen Programmen, die auf die Datenbank zugreifen, Programmen, die die Bildschirmoberfläche steuern, und Programmen, die dazwischen liegen und Anwendungsfunktionen durchführen. Für den Einsatz in einer Client-Server-Architektur ist damit eine physische Trennung der Schichten vorgedacht. Die Programme, die die Bildschirmein- und -ausgabe steuern, müßten nur noch an die eingesetzte Netz- und Client-Software angepaßt werden.

Soweit sinnvoll, wurde schon beim Design der Großrechner-Bildschirmmasken eine spätere graphische Oberfläche abgebildet, damit der Nutzer bei einem Wechsel der Plattform alles wiedererkennt. Es gibt eine Cursor-sensitive Menüleiste mit Pull-Down-Menüs. Das Öffnen mehrerer übereinanderliegender Fenster wird intern durch einen Kellerungsmechanismus simuliert. Er ermöglicht den Aufruf verschiedener Anwendungsmasken hintereinander und auch den umgekehrten Weg, die schrittweise Rückkehr in die vorher aufgerufenen Masken. Für Fragen, die der Nutzer beantworten muß, oder für Mitteilungen öffnen sich Dialogboxen.

Unter einer UNIX-Umgebung und dem Einsatz des Produkts NATURAL for Windows wurde im Rahmen des Abnahmetests eine Client-Server-Installation vorgenommen. Die Bildschirm-

masken für den Client (PC) liegen für die Basisversion in „GUIfizierter" Form vor. Diese für die DV-Welt typische Wortschöpfung bedeutet, daß sie nicht alle für eine graphische Oberfläche zu erwartenden möglichen Graphik-Elemente enthalten, daß der Nutzer aber mit der Maus arbeiten kann und die Maske insgesamt so aussieht, wie er es auf einem PC erwartet („GUI" heißt Graphical User Interface).

Mit der Realisierung solch eines verteilten Systems betraten die Statistischen Ämter Neuland und ließen daher die Basisversion extern entwickeln. Mittlerweile haben die Programmierteams in den beteiligten Ländern eigene Komponenten zur Ergänzung des Systems entwickelt. Dabei wurde und wird das Prinzip der Schichtentrennung natürlich beibehalten.

Besonders zu erwähnen ist in diesem Zusammenhang die Beschleunigung der Import-Schnittstelle durch das Entwicklungsteam Bayern. Der „Quick-Import" bildet eine wesentliche Voraussetzung für den Beginn der Arbeiten unter Produktionsbedingungen. Während der realistische Datenbestand aufgebaut wird, wird sich zeigen, ob weitere Tuning-Maßnahmen nötig sein werden.

Um die Entwicklung der Komponenten „Datenexport" und „Regeln" (Darstellung regionaler Hierarchien und arithmetische Operationen auf Tabellenwerte) zusätzlich zu beschleunigen, wurden vom Statistischen Bundesamt und auch vom Bayerischen Landesamt für Statistik und Datenverarbeitung weitere externe Aufträge an die Firma vergeben, die die Basisversion erstellt hatte. Beide Funktionen wurden mit der Version 1.1 zum Teil schon ausgeliefert.

Die verteilte Entwicklung und die gleichzeitige externe Vergabe von Aufträgen zur Erstellung von Systemerweiterungen bedeuten eine besondere Schwierigkeit bei der Fertigstellung solcher Systeme. Andererseits könnte man ein System, das eine so komplexe Leistungsfähigkeit wie GENESIS erreichen soll, aus Kapazitätsgründen kaum in nur einem Land entwickeln.

Das Schichtenkonzept kommt der verteilten Entwicklung entgegen, oft läßt sich aber insbesondere innerhalb der Datenschicht keine klare Trennung der Komponenten einhalten. Plausibel wird das, wenn man zum Beispiel an den Thesaurus denkt. Die Begriffe des Thesaurus werden aus allen GENESIS-Objekten gebildet. Damit sind Eingriffe in sämtliche Datenschichtkomponenten notwendig.

Im DV-Konzept zur Entwicklung wurde dafür einerseits festgelegt, daß jedes Entwicklungsteam Besitzer bestimmter Komponenten ist, andererseits wurden auch „Ausleihmechanismen" beschrieben, um komponentenübergreifendes Arbeiten zu ermöglichen. Das hört sich einfacher an, als es in der Realität dann ist. Programmierer fühlen sich nämlich durch diese Restriktionen in der Arbeit behindert. Sie sind schnell dabei, notwendige Änderungen in Modulen durchzuführen, die einem anderen Entwicklungsteam gehören, weil sie anders die anstehende Aufgabe nicht realisieren können. Oft stellen sie dann hinterher fest, daß der eigentliche Besitzer oder womöglich noch ein drittes Entwicklungsteam ebenfalls Änderungen daran vorgenommen hatten. In der Anfangszeit, als sich noch nicht alle an diese Regeln gewöhnt hatten, wurde oft erst im Entwicklungsbüro festgestellt, daß verschiedene Versionen eines Moduls gleichzeitig vorlagen. Einer mußte dann die Integration aller Änderungen vornehmen. Dabei konnte es im schlimmsten Fall passieren, daß diese sich gegenseitig ausschlossen.

Dazu kommt, daß solche Vereinbarungen bei der externen Entwicklung nicht eingehalten werden konnten. Der externe Partner setzte zudem auf seinem eigenen Versionsstand – der Basisversion – auf, weil er die in der Zwischenzeit im Verbund erstellten Erweiterungen nicht in die Gewährleistung mit einbeziehen wollte und konnte. Das *Schaubild 1* zeigt, wie viele Ergebnisse integriert werden mußten, um die Version 1.1 zusammenzustellen.

Daten für GENESIS

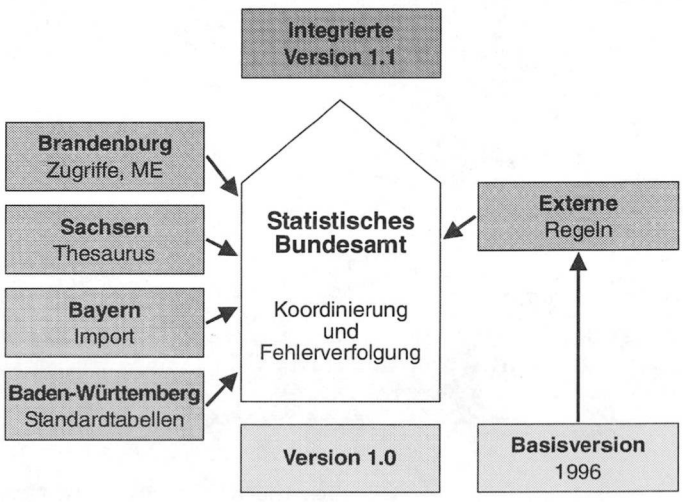
Parallel zu den Entwicklungsteams arbeiten speziell eingesetzte Arbeitsgruppen an der Erstellung verbund-*einheitlicher* Definitionen für die Datenbankinhalte, sowohl im Metadatenbereich als auch für die Datenquader.[1] Die ersten Datenquader, die einheitlich im Verbund geladen werden sollen, wurden mittlerweile von einer Unterarbeitsgruppe der Datenbankreferenten beschrieben. Der erste umfangreiche Arbeitsquader wurde versandt und geladen. Das hört sich recht mager an, beinhaltet aber einige Teilaufgaben, die nicht zu unterschätzen sind.

Zur Datenquaderbeschreibung gehört in GENESIS die komplette Beschreibung seiner Metadaten – das heißt der Erhebung, aus der die in ihm gespeicherten statistischen Werte stammen – und die Beschreibung der Merkmale mit ihren Ausprägungen, die seine Achsen und seine Inhalte bilden. Dafür mußten jeweils verbund-einheitliche Fachschlüssel festgelegt werden.

Für die Spezifizierung eines einzigen Datenquaders erscheint das nun wieder aufwendig. Die Metadaten bilden aber die Grundlage für die Definition aller möglichen weiteren Quader zu dieser Erhebung. Sie können aus dem einmal gespeicherten Metadatenbestand aufgebaut werden.

[1] Die GENESIS-Semantik ist unter anderem von Klaus Engelhardt genau beschrieben und veröffentlicht in: „Bayern in Zahlen", Heft 11/1995, sowie in: „Baden-Württemberg in Wort und Zahl", Heft 6/1996.

Schaubild 1
Die Integrationsarbeiten zu GENESIS Version 1.1



Statistisches Landesamt Baden-Württemberg 144 98

Wie sieht der erste Verbundquader aus?

Der erste Verbundquader enthält tief gegliederte Informationen aus der Bevölkerungsfortschreibung. Den Datenquaderinhalt bildet das „Wert-Merkmal" Bevölkerungsstand zum 31. Dezember.

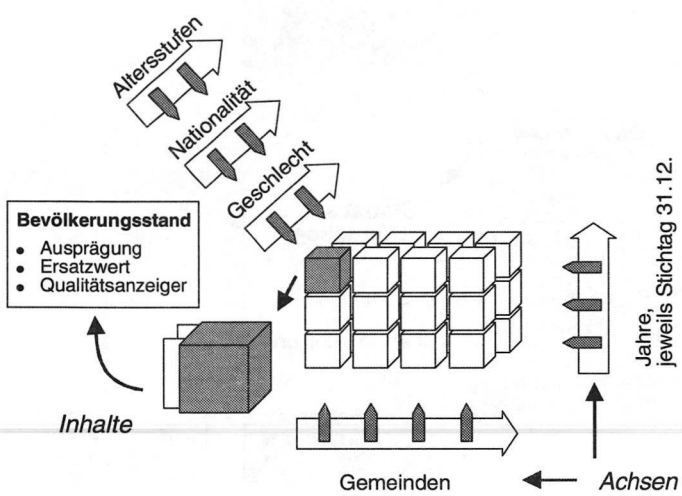
Die Quaderachsen, also die Merkmale, nach denen der Bevölkerungsstand gegliedert ist, sind die Altersgliederung mit 101 Ausprägungen, die Nationalität und das Geschlecht, jeweils in zwei Ausprägungen. Dazu kommen die unverzichtbaren Achsen für die Regionalgliederung und die Zeit *(Schaubild 2)*.

Hier die grundsätzlichen Abrufmöglichkeiten für den ersten Verbundquader:

- Bevölkerungsstand nach Regionalgliederung, Zeit, Nationalität, Geschlecht und Alter
- Bevölkerungsstand nach Regionalgliederung, Zeit, Nationalität und Geschlecht
- Bevölkerungsstand nach Regionalgliederung, Zeit, Nationalität und Alter
- Bevölkerungsstand nach Regionalgliederung, Zeit, Geschlecht und Alter
- Bevölkerungsstand nach Regionalgliederung, Zeit und Nationalität
- Bevölkerungsstand nach Regionalgliederung, Zeit und Geschlecht
- Bevölkerungsstand nach Regionalgliederung, Zeit und Alter
- Bevölkerungsstand nach Regionalgliederung und Zeit

Die Spezifizierung der Datenquader alleine reicht natürlich nicht aus. Die Erstellung der erforderlichen Import-Dateien aus den vorhandenen Datenbeständen stellt für den Verbund und darüber hinaus für jedes einzelne Landesamt eine große Aufgabe dar, die in ihrem Umfang gar nicht überschätzt werden kann.

Schaubild 2
Der erste Verbund-Quader



Statistisches Landesamt Baden-Württemberg 147 98

Der Arbeitskreis Informationstechnik (AKIT) hat schon im Jahr 1996 beschlossen, daß für jede neue Verbundstatistik eine GENESIS-Schnittstelle geschaffen werden muß. Für die bereits bestehenden Verbund-Datenbestände hat das Statistische Bundesamt eine SPLV-Lösung erarbeitet.[2] Eine zusätzlich entwickelte Komponente ermöglicht eine Ausgabe von Daten im GENESIS-Import-Format. Dabei wurden die Möglichkeiten der Tabellengenerierung unter SPLV genutzt.

Möglichkeiten des Werteabrufs

Der Werteabruf in GENESIS bietet bereits mit dem ersten Verbund-Datenquader viele Möglichkeiten, unter anderem den Abruf mit Summierungen über eine oder mehrere Achsen. Aus Fragen der Geheimhaltung und bei sehr umfangreichen Quadern insbesondere auch wegen der Performance wird es aber notwendig sein, entsprechende Summenwerte im System vorzuhalten.

Bei der Tabellengestaltung ist der Nutzer frei,

- welches Merkmal er im Tabellenkopf, in der Vorspalte, als Unter- oder Zwischentitel oder als tabellenübergreifend geltendes Datum anordnen will,
- ob er alle Ausprägungen einer Achse ausgeben will oder nur eine bestimmte Auswahl,
- ob er eine Zeitreihe sehen will, ein Zeitintervall oder nur ein bestimmtes Jahr
- ob er Insgesamt- und Gesamtsummen berechnen will oder
- ob er die Tabelle nach Fachschlüsseln oder Kurztexten angezeigt und sortiert haben will.
- Dazu können verschieden summierte Werte auch nebeneinander in einer Tabelle abgerufen werden, genauso wie es die Möglichkeit gibt, Daten aus verschiedenen Quadern nebeneinander in einer Tabelle darzustellen und
- ab Version 1.1 Regeln in den Abruf einzubeziehen, zum Beispiel zur Zusammenfassung von Regionalmerkmalen.

Die Definition der gewünschten Tabellenstruktur erfordert in der jetzigen Ausbaustufe des Systems noch Kenntnisse des Datenmodells, ist aber leicht erlernbar. Das hat sich bei der ersten Nutzerschulung, die durch das Statistische Landesamt Baden-Württemberg für die Mitglieder der Nutzergruppe durchgeführt wurde, gezeigt. Der Bereich Zugriff- und Zugangsschutz wurde dabei vom Landesamt für Datenverarbeitung und Statistik Brandenburg übernommen.

Dem ungeschulten Nutzer wird es aber schwerfallen, selbst derartige strukturorientierte Recherchen durchzuführen. Deshalb gibt es schon ab Version 1.0 die Möglichkeit, Standardtabellen anzubieten. Das sind Tabellen, die vom Betreiber vorgefertigt werden und nur noch wenige Wahlmöglichkeiten enthalten.

Der Nutzer wählt nur noch eine Tabelle zu einem ihn interessierenden Thema aus. Er muß sich keine Gedanken mehr über die innere Tabellenstruktur machen. Er wird nur noch aufgefordert,

[2] SPLV ist eine vom Statistischen Programmierverbund entwickelte Programmiersprache, die speziell auf Probleme der amtlichen Statistik ausgerichtet ist.

aus einem festgelegten Angebot für die offenen Elemente der Tabelle seine Auswahl zu treffen. Das kann zum Beispiel die Art der Regionalgliederung sein oder ein bestimmtes Jahr. Diese Komponente gehört zu denen, die im Statistischen Landesamt Baden-Württemberg konzipiert und programmiert wurden.

Mit der Version 1.1 wird es im Zusammenhang mit der Realisierung der Recherche-Komponente durch das Statistische Landesamt des Freistaates Sachsen möglich sein, Standardtabellen über Auswahlkriterien im Thesaurus zu finden. Zusätzlich wird die Ansicht von Werten aus einem ausgewählten Datenquader über vom System generierte Tabellen möglich sein. Ein Tabellenassistent, der es auch dem ungeschulten Nutzer ermöglichen soll, Tabellen nach seinem Wunsch zu erstellen, befindet sich noch in der Konzeptionsphase. Die Vorstellungen der Nutzergruppe werden hier einfließen.

Zugangs- und Zugriffsschutz

Voll wirksam ist seit der Version 1.0 schon der Zugangs- und Zugriffsschutz, der die Anforderungen des Datenschutzes und der statistischen Geheimhaltung bei allen Abfragen berücksichtigt. In der GENESIS-eigenen Benutzerverwaltung sind alle Benutzer in Benutzergruppen zusammengefaßt. Jede Benutzergruppe verfügt über ein Zugriffsprofil, in dem der Betreiber sehr differenziert festlegen kann, auf welche Daten die Mitglieder dieser Gruppe Lese- oder Änderungsrechte haben. Das betrifft die Metadaten und auch die statistischen Werte. Dazu kommt die Definition bestimmter Funktionen, die ausgeübt werden dürfen. Ein Beispiel ist die Funktion der Einrichtung neuer Benutzergruppen, die sich der Betreiber sicherlich selber vorbehalten wird. Jeder Benutzer kann sich auch eigene Daten in GENESIS importieren. Sie stehen dann zunächst in seinem „internen Bereich" und sind für keinen anderen Benutzer sichtbar. Erst nach der Freigabe dieser Daten sind diese dann für andere Benutzer zugänglich. Dieser Freigabe muß natürlich eine fachliche Prüfung vorausgehen. Deshalb verfügen nicht alle Benutzer über die Funktion Freigabe in ihrem Zugriffsprofil. Darüber hinaus gibt es noch eine Reihe von Möglichkeiten, einzelnen Benutzern Rechte auf bestimmte Datenbereiche zu verleihen und auch zu verwehren.

In der jetzigen Phase, in der der Datenbestand im Verbund und auch länderspezifisch aufgebaut wird, erscheint es vielleicht lästig, wenn man für die wenigen Nutzer Profile einrichten muß. Aber gerade jetzt kann man Erfahrungen mit der Komplexität dieses Systems sammeln und vielleicht die eine oder andere notwendige Ergänzung oder auch Vereinfachung erkennen.

Die erste Nutzerschulung kann als Beispiel dienen, wie man die Fähigkeiten dieser Komponente nutzen kann. Sie wurde mit der Installation der GENESIS-Produktionsdatenbank durchgeführt. Der Verbundquader ist in Baden-Württemberg bereits auf Gemeindeebene für zehn Jahre geladen (3 Millionen Werte). Es handelt sich dabei um Daten, die nicht für externe Nutzer bestimmt sind.

Für die Schulung wurden noch einige Datenquader importiert, die höher aggregierte Daten enthielten, dazu speziell erstellte Datenquader aus der Veröffentlichung „Lange Reihen". Diese Datenquader wurden für den externen Nutzerkreis der Schulung freigegeben.

Die Schulungsteilnehmer bildeten also eine eigene Gruppe, die in ihrem Profil nur sehr eingeschränkte Rechte erhielt, nämlich Leserechte auf die freigegebenen Datenquader und natürlich auf die zugehörigen Metadaten. Während der Schulung konnte sich jeder Teilnehmer eigene Metadaten bilden. Das Recht, diese Daten für andere Teilnehmer freizugeben, hatten die Schulungsteilnehmer nicht, so daß sie den Produktions-Datenbestand nicht „verunreinigen" konnten. Jeder Teilnehmer konnte aber Tabellen erzeugen und damit Werte aus den freigegebenen Quadern abrufen.

Bei der Schulung hat sich das bestehende Konzept bewährt. Insgesamt sind die ständig vom System durchzuführenden Überprüfungen so umfangreich, daß Überlegungen bestehen, sie in abgegrenzten Bereichen zu lockern, um einen Performance-Gewinn zu erzielen. Das gilt beispielsweise für den Massendatenimport. Der Import wurde gegenüber der Basisversion schon erheblich beschleunigt, angesichts des gewaltigen Umfangs der aus den alten Datenbanken zu überführenden Datenmenge bedeutet jede weitere Beschleunigung aber weitere Zeit- und damit Kostenersparnis.

Weiteres Vorgehen

- Die Testarbeiten durch die Nutzergruppe gehen weiter. Der vorliegende Zwischenbericht wird derzeit ausgewertet.
- Die nächste Integrationsversion 1.2 wird bereits zusammengestellt. Sie enthält Erweiterungen und auch Korrekturen im Bereich der Regeln, der Standardtabellen und des Thesaurus.
- Die darauffolgende Entwicklungsphase wird eine Konsolidierungs- und Tuning-Phase sein.
- Die wichtigsten der anstehenden Aufgaben werden die Beschleunigung des Werteabrufs und die Aktualisierung der online-Dokumentation sein. Weitere Tuning-Maßnahmen werden sich aus dem Bericht der Nutzergruppe ergeben.

Madeleine de la Croix