Erste Erfahrungen mit der Langzeitarchivierung von Datenbanken. Ein Werkstattbericht*

Von Christian Keitel

Die Landesarchivdirektion beschloss Ende 2002, in der baden-württembergischen Archivverwaltung die elektronische Archivierung aufzunehmen. Mit diesem Entschluss verband sich die Hoffnung, gefährdete Unterlagen zu sichern, dabei praktische Erfahrungen zu gewinnen und somit für künftige Aufgaben besser gerüstet zu sein. Der Einstieg sollte zunächst auf niedrigem Niveau erfolgen und eine evolutionäre, schrittweise Lösung der mit der elektronischen Archivierung zusammenhängenden Probleme einleiten. Zuvor waren wir zur Überzeugung gekommen, dass sich eine große Lösung ohne ausreichende praktische Kenntnisse weder konzipieren noch konkret umsetzen lässt.[1] Im Staatsarchiv Ludwigsburg wurde daher im April 2003 eine halbe Stelle zur Archivierung digitaler Unterlagen eingerichtet, die sich stellvertretend für die anderen Staatsarchive um die langfristige Sicherung dieser Dokumente kümmern soll. Im Folgenden sollen die Erfahrungen im Bereich der Eingangsbearbeitung am Beispiel der Volkszählung 1970, die Überlegungen zur Bewertung anhand eines Fachinformationssystems für Lebensmittelüberwachung und Veterinärwesen vorgestellt werden.

Volkszählung 1970

Die Volkszählung 1970 wurde bundesweit durchgeführt. Ihre baden-württembergischen Ausgangsdaten lagen im Statistischen Landesamt (StALA) auf Magnetbändern und in einer selbstdefinierten Bitverschlüsselung vor, sodass die Daten zunächst durch das Landesamt auf neue Datenträger überspielt und in den ASCII-Code migriert werden mussten. Diese Arbeiten wurden im StALA von dem einzigen Mitarbeiter gemacht, der bereits an der Zählung 1970 selbst beteiligt war. Nach seiner in wenigen Jahren anstehenden Pensionierung wäre die erste Migration

---

* Vortrag auf Digitales Verwalten - Digitales Archivieren. 8. Tagung des Arbeitskreises "Archivierung von Unterlagen aus digitalen Systemen" am 27./28.4.2004 in Hamburg.

[1] Christian Keitel, Die Archivierung elektronischer Unterlagen in der baden-württembergischen Archivverwaltung. Eine Konzeption, 12.6.2002, veröffentlicht auf den Internet-Seiten der Staatlichen Archivverwaltung Baden-Württemberg, derzeit unter http://www.lad-bw.de/lad/konzeption.pdf.

also schwierig, wenn nicht unmöglich geworden. Das StALA gab die migrierten Daten im April 2002 ab.

Die Volkszählung 1970 unterscheidet sich vor allem dadurch von älteren Zählungen, dass bei ihr erstmals sämtliche erhobenen Daten maschinell eingelesen und weiterverarbeitet wurden. Die übernommenen Daten stellen daher theoretisch die ganze Zählung dar. Auf der anderen Seite haben die vier zuständigen baden-württembergische Staatsarchive[2] bereits früher 1620 laufende Meter Erhebungsbögen übernommen. Die Informationen lagen folglich sowohl auf Papier als auch in elektronischer Form vor. Bei näherer Betrachtung unterscheiden sich die beiden Überlieferungsformen jedoch in mehrfacher Hinsicht:

- Einige wenige Informationen werden nur in den Erhebungsbögen (Namen), andere nur in den Dateien (z.B. Informationen zum Aufbereitungsprozess, zur Typisierung der Daten) geführt.
- Die digitalen Daten sind fehlerbereinigt und plausibilisiert. Sie stimmen daher in höherem Grad als die Erhebungsbögen selbst mit der abzubildenden Realität überein.
- Nur die digitalen Daten bieten die Möglichkeit einer statistischen Auswertung zu einem vertretbaren Aufwand.
- Der Erhalt des säurehaltigen Papiers der Erhebungsbögen dürfte langfristig kaum zu finanzieren sein.

Die genannten Gründe sprechen eher für eine Archivierung der digitalen Daten. Vor allem jedoch war die Erstellung dieser Daten das Ziel der Volkszählung selbst. Die digitalen Daten können daher als Primärüberlieferung der Volkszählung, die Erhebungsbögen als Hilfsmittel angesehen werden. Soweit die grundsätzlichen Überlegungen. Boten diese Daten aber auch all jene Informationen, die sie versprachen? Waren sie überhaupt les- und verstehbar?

Zeichenformat

Auf den ersten Blick scheinen die Dateien nur Ziffern und einige wenige lateinische Buchstaben zu enthalten. In den Feldern „Hausnummer“ (z.B. 7a) und Priorität werden außerdem auch Buchstaben repräsentiert. Durch eine Abfrage in einem

[2] Generallandesarchiv Karlsruhe, Staatsarchiv Freiburg, Staatsarchiv Ludwigsburg und Staatsarchiv Sigmaringen.

Texteditor lassen sich aber drei Zeichen nachweisen, die nicht vom einfachen 7-Bit-ASCII-Code (128 Zeichen) abgedeckt sind:

Hexadezimal	Dezimal	ISO 8859-1	ANSI	erweiterter ASCII-Satz (DOS)
83	131	-	*ƒ*	â
9F	159	-	Ÿ	*ƒ*
DA	218	Ú	Ú	┌

Diese Zeichen finden keine Entsprechung in den erhaltenen Codelisten. Insgesamt konnten über 66.000 Vorkommen dieser Zeichen nachgewiesen werden. Sollte sich ihr Vorkommen daher über alle Felder erstrecken, wäre der künftige Gebrauch und damit die Archivierung der Datenbank in Frage gestellt. Bis auf eine Ausnahme befanden sich diese Zeichen jedoch in jenen Feldern, mit denen die Pendlerzielgemeinden codiert wurden. [3] Bis auf die Felder der Pendlerzielgemeinden lassen sich also alle anderen der insgesamt 101 Felder im einfachen ASCII-Code darstellen, i.a.W. besteht die grundsätzliche Möglichkeit, dass die in diesen Feldern verwendeten Zeichen auch eine Entsprechung in den Codelisten finden. Für die Darstellung sämtlicher Einträge kann jedoch der für die Archivierung übliche Zeichensatz Latin-1, d.h. ISO 8859-1, nicht verwendet werden, da mit ihm zwei Zeichen nicht repräsentiert werden können.

Dateien und Satzarten

Die Daten sind das Ergebnis zweier Erhebungen und liegen in zwei Satzarten vor. Alle Bewohner Baden-Württembergs mussten die ersten 18 Fragen des Fragebogens beantworten. Neben dieser Totalerhebung mussten 10% der Bevölkerung weitere 18 Fragen beantworten, die zu einer Repräsentativerhebung benötigt wurden. Leider entsprechen sich Satzarten und Erhebungen nicht.[4]

[3] In diese Felder fallen auch einige wenige Vorkommen von „+“ und „]“, die allerdings durch den einfachen ASCII-Code dargestellt werden können.

[4] Datensätze vom Typ 1 beginnen mit „1“ und umfassen das Material der Totalerhebung und zusätzlich einige wenige Angaben aus der Repräsentativerhebung (Geburtsjahr der Kinder 7-12, Details zur Religionszugehörigkeit und Staatsangehörigkeit). Datensätze vom Typ 2 beginnen mit „2“ und enthalten die anderen Daten der Repräsentativerhebung.

	Datensätze	Satzart 1 = Total- und Repräsentativerhebung	Satzart 2 = Repräsentativerhebung
RP 1	3.909.781	3.587.768	322.013
RP 2	2.132.831	1.953.429	179.402
RP 3	2.106.137	1.926.773	179.364
RP 4	1.830.526	1.675.259	155.267
Baden-Württemberg	**9.979.275**	**9.143.229**	**836.046**

Für eine Analyse der Repräsentativerhebung ist es daher stets notwendig, Angaben von beiden Satzarten gemeinsam zu betrachten. Diese Anforderung konnte im vorliegenden Festbreitenformat nicht erfüllt werden. Die Daten werden hier nicht durch Feldtrenner voneinander separiert, sondern durch ihre Position innerhalb der Zeile bestimmt. Bei Satzart 1 stehen die Bytes Nummer 77 bis 80 für die Hausnummer, Nummer 81 für das Geschlecht. Bei der Satzart 2 sind die Bytes anders belegt. Hier repräsentieren die Bytes 80 und 81 das Geburtsjahr des ersten Kindes. Dateien, die unterschiedliche Datensätze im Festbreitenformat besitzen, können nicht mehr ohne weiteres in ein Datenbankprogramm eingelesen werden. Für die Archivierung war es daher nötig, die beiden Satzarten in unterschiedlichen Dateien abzuspeichern - aus den vier Dateien wurden also acht. Nun war es möglich, diese Dateien in ein handelsübliches Datenbankprogramm (MS-ACCESS) einzulesen und dort zu analysieren.

Wie lassen sich einzelne Datensätze eindeutig identifizieren und einem Erhebungsbogen zuweisen? Wie können zwei zusammengehörige Datensätze miteinander verknüpft werden.[5] Interessanterweise konnte das hierzu prädestinierte Feld „Volkszählungskennnummer“ diese Anforderungen nicht erfüllen, da 67 Datensätze vom Typ 1 doppelt vorhanden waren. Die meisten dieser Datensätze bezogen sich zudem auf zwei unterschiedliche Personen. Aus diesem Grunde musste mit dem Feld „Zählerlisten-/Anstaltsnummer“ ein weiteres Merkmal gefunden werden, das dann eine eindeutige Identifikation der einzelnen Datensätze erlaubte. Die acht Tabellen wurden als eigene Dateien im CSV-Format ausgegeben. Nachdem

[5] Selbst bei einer unveränderten Archivierung der Ausgangsdateien hätte die Frage der Verknüpfung gelöst werden müssen, da die Datensätze vom Typ 2 zwar oft unmittelbar den dazugehörigen Datensätzen vom Typ 1 folgten, dies aber keineswegs immer der Fall war.

sich in den Ausgangsdateien (Festbreitenformat) kein Strichpunkt finden ließ, konnte dieser als Feldtrenner verwendet werden. Textbegrenzungszeichen wurden nicht verwendet. Nach dem CSV-Export wurden die Archivdateien mit den Ausgangsdateien verglichen und auf ihre Konsistenz geprüft: Verglichen wurde zunächst die Zahl der Datensätze, dann die Zahl einzelner Zeichen („normaler“ Buchstaben und „Sonderzeichen“). Dabei ergaben sich keine Abweichungen. Die Primärdaten werden in zwei Kopien auf CD-R und DVD+R archiviert. Sie können, wie bereits erwähnt, problemlos in existierende Datenbankprogramme eingelesen werden.

Dokumentation

Insgesamt enthält die Datenbank 101 Felder, von denen 71 codiert, 30 im Klartext vorliegen. Ein Klartextfeld ist z.B. das Feld „Geburtstag“: Eine „27“ steht hier für einen Geburtstag an einem 27sten eines bestimmten Monats in einem bestimmten Jahr. Codiert ist ein Feld dann, wenn die eingetragenen Ziffern erst mittels einer Codeliste verstanden werden können. Im Feld „Geschlecht“ steht beispielsweise eine 1 für „männlich“, eine 2 für „weiblich“. Dokumentiert werden mussten daher vor allem die 71 codierten Felder. Schließlich konnten die Codelisten für 58 Felder eindeutig, die Codes von vier weiteren Feldern mit hoher Wahrscheinlichkeit identifiziert werden. Die Felder mit den aus den Fragebögen übernommenen Angaben sind daher fast vollständig dokumentiert. Undokumentierte Felder beziehen sich v.a. auf geographische Kleinsteinheiten (Nummerierung der Straßenzüge) und nach der Plausibilisierung vorgenommene Typisierungen. Neben der spärlichen Dokumentation, die vom StALA übernommen werden konnten, waren vor allem die 1970 publizierten Codelisten hilfreich. Ein Teil der Informationen konnte auch erschlossen werden.

Verifikation der Daten

Nach der technischen Aufbereitung und Analyse der Daten sowie dem Zusammentragen der Dokumentation konnte die Frage nach der Authentizität und Integrität der Daten gestellt werden. Beinhalten die Daten nur die Inhalte, die sie angeblich repräsentieren? Zunächst wurde überprüft, ob die einzelnen Felder jeweils nur die laut Codeliste oder Klartextdefinition vorgesehenen Einträge enthalten. Das Feld Geburtstag durfte beispielsweise ausschließlich die Ziffern 1 bis 31 führen, im Feld „Geschlecht“ sollte entweder eine 1 oder eine 2 eingetragen sein. Die Daten entsprachend weitgehend, aber nicht vollständig den Erwartungen. Im Feld

Geburtsmonat waren beispielsweise in 15 der 9,9 Millionen Datensätze eine Zahl eingetragen, die höher als 12 war. In einem zweiten Schritt wurden dann die Daten einzelner Felder aufeinander bezogen. Auch hier entsprachen sich die Daten in hohem Maße. 2971 Personen erklärten jedoch, keine Zweitwohnung innezuhaben und spezifizierten dann deren Beschaffenheit. Auch diese scheinbar hohe Zahl bewegt sich jedoch unterhalb des Promillebereichs. Schließlich wurden die digitalen Daten mit den Erhebungsbögen verglichen. In den Datensätzen fanden sich nur einzelne Abweichungen. Ein umfassend abweichender Datensatz konnte nicht festgestellt werden. Die Abweichungen bewegen sich daher auf Feldebene. Von 3260 untersuchten Felder unterschieden sich die Angaben in 61 Feldern. 57 der 61 Abweichungen dienten dabei ganz offenkundig einer absichtlichen Verbesserung der Datenbasis:

- Plausibilisierung (15 Felder): z.B. wurde die Antwort "Gymnasium" auf die Frage 9 (Besuchen Sie gegenwärtig eine Schule? Wenn ja, welche?) bei einer befragten Person des Jahrgangs 1925 in "keine Antwort" geändert.
- Ergänzung der Angabe (7 Felder): z.B. Frage 31 (Praktische Berufsausbildung beendet?): "Keine Antwort" geändert in "nein".
- Reduktion auf eine Angabe (5 Felder): Wenn auf die Frage 8 (Wovon leben Sie überwiegend?) mehrere Antworten gegeben wurden ("Erwerbs-/Berufstätigkeit" und "Unterhalt durch Eltern, Ehemann usw."), wurde nur eine übernommen.
- spezifische Interpretationen (30 Felder): z.B. wurde in den verglichenen Datensätzen "Hausfrau" stets auch als "nicht erwerbstätig" gewertet.

Nur bei vier von 3260 Feldern = 0.123% liegt eine Abweichung vor, die sich vorläufig nicht durch einen gerechtfertigten Eingriff während der Plausibilisierung der Daten erklären lässt. Selbst bei diesen Daten ist allerdings eine Korrektur aufgrund der Erhebungsdaten anderer Zählungen bzw. einer zusätzlichen Recherche nicht unwahrscheinlich. Auf der Basis der erhobenen Stichproben besitzen daher die Datensätze eine höhere Übereinstimmung mit der abzubildenden Realität als die Erhebungsbögen selbst. Damit sind alle eingangs formulierten Anforderungen für eine Archivierung der digitalen Volkszählungsunterlagen erfüllt. Die hier referierten Informationen wurden in der AGÜ (Arbeitsgemeinschaft Überlieferungsbildung) diskutiert. Die Erhebungsbögen sollen nun bis auf wenige Demonstrationsexemplare vernichtet werden. Voraussichtlich werden die Kosten für den Erhalt der digitalen Daten niedriger sein, als die für eine Entsäuerung der Erhebungsbögen zu veranschlagenden Kosten. Die elektronische Archivierung kann daher in einigen Bereichen auch zur Kostenentlastung der Archive beitragen.

Die genannten Arbeiten wurden auf einem Pentium 4 mit 1,7 GHz und 512 MB RAM durchgeführt. Als Software wurde MS-ACCESS, ein Texteditor und ein Dateimanager

verwendet. Ein vergleichbares Soft- und Hardwarepaket kann derzeit für etwa 1000 Euro erworben werden.

BALVI iP und LÜVIS

Neben den seit den siebziger Jahren erstellten Statistiken liegt ein zweiter Schwerpunkt der elektronischen Archivierungsstelle auf den laufenden Fachinformationssystemen. Erste Erfahrungen haben wir hier mit dem System LÜVIS gemacht, das derzeit vom „Entwicklungs- und Betreuungszentrum für Informations- und Kommunikationstechnik des Ministeriums für Ernährung und Ländlichen Raum“ (EBZI) entwickelt wird. LÜVIS steht für „Lebensmittelüberwachungs- und Veterinärdokumentationssystem“ und soll für ganz Baden-Württemberg sämtliche Aufgaben in diesen Bereichen unterstützen.[6] Das Produkt wird von der Firma BALVI programmiert und bundesweit unter der Bezeichnung BALVI iP vertrieben. Gekauft wurde das System bislang von Brandenburg und Niedersachsen, die meisten anderen Bundesländer sind am System interessiert.
Das System basiert auf Oracle 9i, die Daten können über Citrix-MetaFrame via Internet abgerufen und eingegeben werden. In seinem Kern enthält es 270 Tabellen, hinzu kommen dann noch die verknüpften externen Dokumente (v.a. WORD). In anderen Worten ist BALVI ein typisches Fachinformationssystem und damit hinreichend komplex für einen ersten Pilot.

Wo setzt die archivische Bewertung bei einem Fachinformationssystem an?

BALVI besteht aus mehreren Schichten. Auf den zugrundeliegenden Datenhaltungsstrukturen basiert eine Middleware, welche die Daten schließlich an das dem Benutzer sichtbare Frontend weitergibt. Die archivische Beschreibung kann sich nun entweder auf die zugrundeliegenden Datenstrukturen oder auf die den Benutzern gegebenen Sichten beziehen. Das zunächst genannte Vorgehen ermöglicht zwar eine wenig redundante Beschreibung der Strukturen. Zugleich muss sie aber aufwändig ermittelt und später dokumentiert werden, da auf die Hilfsmittel für die Benutzer nicht zurückgegriffen werden kann. Aus diesen Gründen lehnt sich die Beschreibung an die Benutzersichten an. Diese Entscheidung impliziert auch

[6] Unterstützte Fachbereiche: Lebensmittelüberwachung, Weinkontrolle, Tierseuchenüberwachung, Tierseuchenkrisenfall, Tierschutz, Grenz-Kontrolle, Fleischhygiene, Rindfleischetikettierungs-Kontrolle, Legehennenregistrierung, Tierarzneimittel-Kontrolle, Futtermittel-Kontrolle, Handelsklassen-Kontrolle, Pflanzenschutzmittel-Kontrolle, Düngemittel-Kontrolle.

einen erleichterten Datenexport, da die Benutzersichten bereits eine Form der Datenausgabe darstellen.

Wie kann ein Fachinformationssystem überblickt werden?

Wie bereits erwähnt besteht LÜVIS aus über 270 Tabellen. Etliche Informationen werden redundant angeboten (aber nicht gehalten), d.h. ein Großteil der Komplexität ist dem Umstand geschuldet, dass die Dateneingabe für etliche Personengruppen mit jeweils unterschiedlichen Rechten und Aufgaben komfortabel gestaltet werden musste. Aus archivischer Sicht sollen die eingegebenen Daten nach der Übernahme aber nicht mehr verändert werden. Grundsätzlich ist daher eine Reduktion der Komplexität denkbar. Zudem ist es unmöglich, ein derartig komplexes System als Ganzes zu archivieren. Es war daher notwendig, diese Komplexität in mehreren Schritten zu reduzieren.

- Auszeichnung der verzichtbaren Programmteile: LÜVIS besteht im Kern aus einer Oracle-Datenbank, die innerhalb der Vorgangsverwaltung auf extern abgelegte WORD-Dokumente verweist. In diesen kann nach Einspielung verschiedener Datenbankinformationen ein Aktenvermerk angefertigt werden. Da die wesentlichen Informationen in der Datenbank selbst gehalten werden, sollen die WORD-Dokumente nicht archiviert werden. Damit bleiben „nur“ noch die Tabellen übrig.
- Charakterisierung der nicht archivwürdigen Bereiche: Hierzu zählen z.B. Auslagen, Gebühren, Vergütungen, Arbeitszeiterfassung etc.
- Bewertung nach Navigationsobjekten: Die verbleibenden Tabellen und Formulare wurden einzeln bewertet.

Bewertung und Archivierung

Es ergab sich eine dreistufige Gliederung der Daten:

1. Stammdaten
2. Grundinformationen (z.B. Seuchenstatus aller Kühe eines Betriebs)
3. Detailinformationen (z.B. Seuchenstatus einzelner Kühe)

Die Stammdaten umfassen Informationen zu den Betrieben und den einzelnen Kontrolltätigkeiten, d.h. den Betriebsbesuchen. Die wesentlichen Detailinformationen (= Ebene 3) werden noch einmal zusammengefasst in den Grundinformationen wiedergegeben. Zugleich stellen die Detailinformationen die übergroße Mehrzahl der Tabellen. Die Ergebnisse der Bewertung lassen sich wie folgt darstellen:

Die Ebenen 1 und 2 enthalten die zentralen archivwürdigen Informationen. Teilbereiche der Ebene 2 werden nicht archiviert (z.B. Beurteilung von Kampfhunden). Ebene 3 wird nicht archiviert.
Bislang haben die baden-württembergischen Staatsarchive in den einzelnen Bereichen der Lebensmittelüberwachung entweder einige Akten exemplarisch übernommen oder Samples gebildet. Als Alternative bietet es sich nun an, die archivwürdigen Navigationsobjekte alle zwei Jahre mit sämtlichen Datensätzen zu übernehmen. Künftigen Benutzern stünde dann eine Vielzahl neuer Auswertungsmöglichkeiten zur Verfügung. Diese Möglichkeiten legen neben dem Umstand, dass die Lebensmittelversorgung die gesamte Bevölkerung sowohl direkt (durch die alltägliche Nahrungsaufnahme) als auch indirekt (durch die von Nahrungsmitteln ausgelösten Krankheiten und Epidemien) tangiert, eine dauerhafte Archivierung der ausgezeichneten Bereiche nahe.
Die archivwürdigen Informationen lassen sich in 2 Tabellen mit Stammdaten und 16 weiteren, mit den Stammdaten verknüpften Tabellen darstellen. Das EBZI hat in Aussicht gestellt, diese 18 Tabellen im CSV-Format abzugeben.

Resümee

Bei der Bewertung von LÜVIS musste stets die Möglichkeit einer Archivierung und damit die Möglichkeit einer Verknüpfung mitbedacht werden. In anderen Worten waren die archivarischen und die technischen Überlegungen sehr eng miteinander verbunden. Darüber hinaus lassen sich die Erfahrungen des letzten Jahres in zwei Sätzen zusammenfassen: Es ist möglich, mit minimalen Hardware-, Software- und Personalmitteln in die elektronische Archivierung einzusteigen. Die praktischen Erfahrungen verändern die Sicht auf die elektronische Archivierung, die Standards und die angeblichen Unmöglichkeiten erheblich.