Die Archivierung elektronischer Statistiken durch das Landesarchiv Baden-Württemberg

Dr. Christian Keitel

Dr. Christian Keitel leitet den Aufbaustab Digitales Archiv im Landesarchiv Baden-Württemberg. Seit 2002 werden dort auch elektronische Statistiken des Statistischen Landesamts archiviert.

Vor gut 5 Jahren konnte die baden-württembergische Archivverwaltung erstmals ein digitales Archivale übernehmen: Die Volkszählung 1970. Heute liegen auch die Statistiken zu Geburten und Sterbefällen, zur Strafverfolgung und -vollzug und verschiedene andere Zählungen (Volks-, Gebäude- und Wohnungs-, Handwerk- und Arbeitsstättenzählungen) in dem Digitalen Magazin der Archivare. In den Findmitteln des Staatsarchivs Ludwigsburg stehen sie nun neben den vom Statistischen Landesamt übernommenen Papierunterlagen, zu denen auch ältere tabellarische Auswertungen und eine Auswahl der Papierakten zählen. Die sehr gute Zusammenarbeit zwischen dem Statistischen Landesamt und dem seit 2005 unter neuem Namen firmierenden Landesarchiv wird in den nächsten Jahren durch eine bundesweit einheitliche Vereinbarung zwischen den Statistischen Ämtern und den zuständigen Archivverwaltungen auf neue Beine gestellt. Es liegen also genügend Gründe vor, um eine erste Bilanz ziehen zu können.

Anfänge

Aus der Vogelperspektive des Historikers erscheint es nicht als Zufall, dass das erste digitale Archivale aus einem Statistischen Amt stammt. Ähnliches lässt sich am amerikanischen Nationalarchiv beobachten, das seinen Benutzern sogar noch die Lochkarten zum Zensus 1890 anbieten kann. In England setzte die digitale Archivierung ebenso mit Statistiken ein, wie dies in Frankreich, den skandinavischen Ländern oder einzelnen Bundesstaaten der USA der Fall war. Die Ursache hierfür ist in der Form statistischer Informationen zu suchen. Eignet sie sich doch in besonderem Maße für die Datenverarbeitung durch Maschinen. In Stuttgart waren schon 1910 Lochkartenmaschinen zur Informationsverarbeitung eingesetzt worden. Da diese Maschinen sehr früh in die statistischen Ämter gelangten, kamen die maschinenlesbaren Informationen auch mit als erstes in ein Alter, das von Archivaren gemeinhin mit dem Zustand der Archivreife umschrieben wird. Aus diesem Grund stammt nicht nur das erste, sondern auch das älteste digitale Archivale des Landesarchivs aus dem Statistischen Landesamt, die Rede ist von der Volkszählung 1961.

Archivierung

Was verstehen überhaupt die Archivare unter „Archivierung“? Immerhin wird dieses Wort schon seit Jahren viel häufiger durch Software-Firmen und IT-Power-User benutzt. Der Begriff steht dann für ein sicherndes Ablegen über die nächsten 3, 4 oder 5 Jahre. Archivare verstehen unter „Archivierung“ eine Aufbewahrung für unbestimmte Zeit, sie sprechen von der Ewigkeit, zumindest aber von einem Zeitraum, den die ältesten Archivalien in ihren Magazinen bereits durchschritten haben. Beim Landesarchiv Baden-Württemberg wären dies 1 200 Jahre.

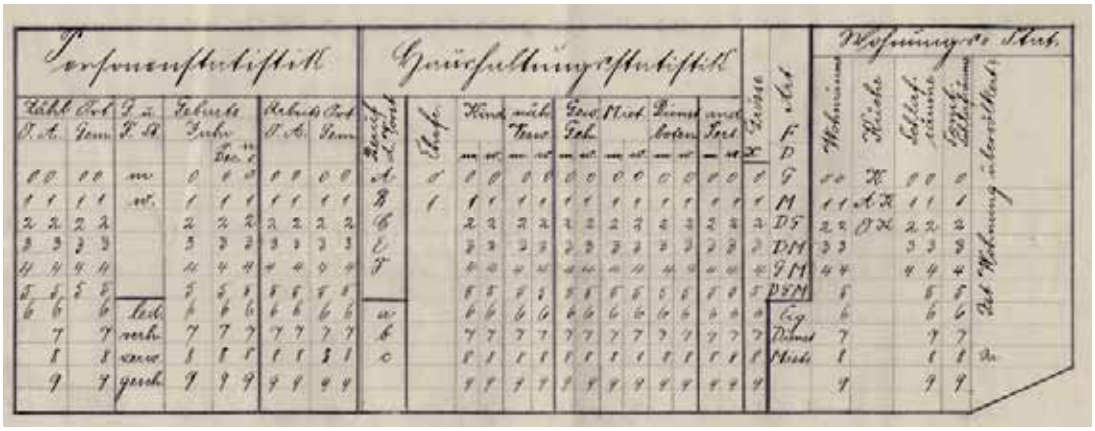

Abbildung 1: Entwurf für die ersten Lochkarten (1910).

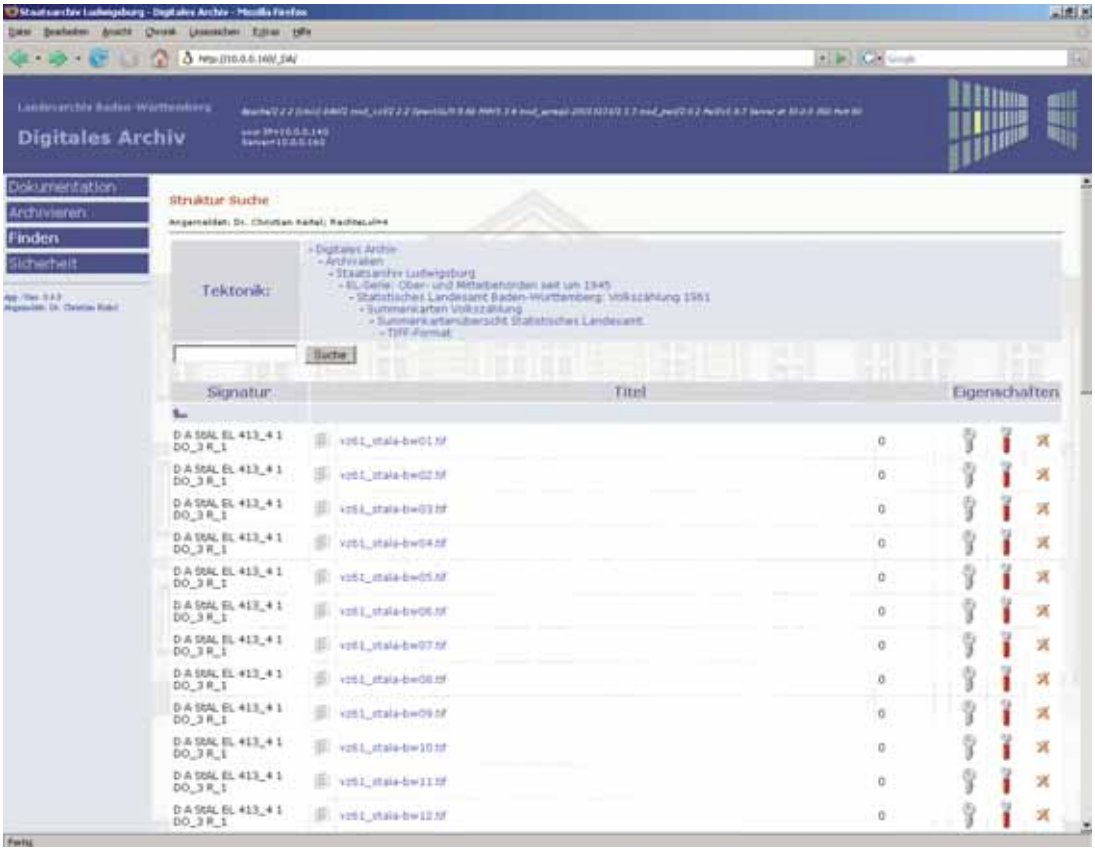

Abbildung 2: Bildschirmansicht aus dem Digitalen Archiv (DIMAG).

Leider währen die Erfahrungen mit der Archivierung digitaler Unterlagen noch nicht ganz so lange. Wie versetzen sich Archivare also in die Lage, nachts noch halbwegs gut schlafen zu können?

Es sind schon die Speichermaterialien, die einer erfolgreichen Umsetzung dieser Aufgabe entgegenstehen. Derzeit existiert kein Datenträger, der digitale Daten über eine sehr lange Zeit sicher speichern kann. Es ist auch unwahrscheinlich, dass in künftigen Jahrhunderten noch ein entsprechendes Laufwerk zum Einlesen dieser Datenträger verfügbar ist. Archive kopieren ihre digitalen Archivalien daher regelmäßig auf aktuelle Trägermaterialien um und verwahren jede Information zum Mindesten in drei identischen Kopien, die wiederum an unterschiedlichen Orten abgelegt werden sollten.

Im Landesarchiv wurden diese Kopien zunächst auf CD gebrannt. Langfristig war dies jedoch eine zunehmend problematische Lösung. Zunächst geben wissenschaftliche Tests die Haltbarkeit von handelsüblichen CDs mit 3 bis 30 Jahren an. Die einzelnen Anbieter bedienen sich unterschiedlicher Hersteller, ohne dass dies zu erkennen wäre. Normmaß wären daher nicht 30 sondern 3 Jahre, und dann wächst der erforderliche Aufwand rasch in sehr bedenkliche Größen. Seit 2006 werden im Landesarchiv Baden-Württemberg daher die digitalen Archivalien in drei Festplattensystemen gespeichert. Das Produktivsystem steht im Staatsarchiv Ludwigsburg, die Backup-Server in Stuttgart und Karlsruhe. Damit ist aber erst die unmittelbare physische Ablage beschrieben. Zugleich muss die Ablage systematisch erfolgen und der Zugang reglementiert werden. Im Landesarchiv Baden-Württemberg konnte 2006 ein Digitales Magazin, kurz DIMAG genannt, entwickelt werden, das diese Anforderungen erfüllt und eine sichere Archivierung digitaler Unterlagen ermöglicht. DIMAG entspricht dem oben genannten Produktivsystem.

Das Digitale Magazin DIMAG

DIMAG ist für alle denkbaren Formen digitaler Archivalien offen und speichert derzeit neben Statistiken auch die Daten aus Datenbanken und Geografischen Informationssystemen, Dokumentmanagementsystemen, digitale Bilder und Einzeldokumente. Zentrale Metadaten werden zur Verwaltung des DIMAG in einer Datenbank abgelegt, zugleich werden aber alle Meta- und Primärdaten durch DIMAG in ein Dateiverzeichnis geschrieben, aus dem sie auch ohne die DIMAG-Software auslesbar sind. Die Daten sind damit prinzipiell transparent zugänglich, sie können auch ohne DIMAG

überleben, aufwendige Exportschnittstellen müssen nicht geschrieben werden. Allerdings gelten diese Freiheiten nur für den Administrator. Der normale Archivar hat keinen direkten Zugriff auf die Dateiverzeichnisse. Er kann nur über DIMAG zu den Daten gelangen. Der einheitliche browserbasierte Zugang bringt verschiedene Vorteile mit sich:

- Der Zugang selbst ist an ein gestuftes System der Rechtevergabe gekoppelt.
- Primärdaten können nicht ohne die Angabe rudimentärer Metadaten abgelegt werden.
- Nach der Einstellungsphase ist es nicht mehr möglich, einzelne Dateien zu ändern oder gar zu löschen.
- Der zentrale Speicher kann in Zukunft von allen Staatsarchiven aus über ein geschütztes Intranet angesprochen werden.

Jede Datei ist durch eine eigene Hashwertdatei beschrieben, auch kleinere Veränderungen lassen sich leicht nachweisen. Ein Protokollierungssystem beschreibt alle wesentlichen Archivprozesse.

Archivierungsstrategie

Mit dem physischen Erhalt der Daten und ihrer Lesbarkeit ist allerdings nur ein Teil der mit der digitalen Archivierung einhergehenden Probleme gelöst. Im Anschluss war die Frage zu klären, ob heutige Datei- und Datenformate von künftigen Computern noch verstanden werden. Unter den möglichen Antworten dominieren zwei Ansätze. Bei der Emulationsstrategie werden die Daten in den Originalformaten zusammen mit der Software und dem Betriebssystem ins Archiv geholt. Werden in Zukunft neue Computer und Betriebssysteme entwickelt, dann muss für diese eine Anpassungssoftware (Emulator) geschrieben werden, damit sie wie ein Rechner aus dem Jahr 2008 reagieren kann. Gelingt dies, dann kann auch noch in ferner Zukunft auf künftigen Computern Windows Vista und Word 2007 aufgerufen werden. Gelingt dies nicht, ist die ganze Information verloren. Das Landesarchiv Baden-Württemberg verfolgt daher die alternative Migrationsstrategie. Dabei werden die Daten in möglichst gut dokumentierten, rechtefreien und weit verbreiteten Formaten übernommen (zum Beispiel CSV, XML, TIFF oder PDF/A), die Software der Behörde jedoch nicht. Die Dateien selbst sollen so gut dokumentiert werden, dass sie auch in ferner Zukunft noch in die dann gängige Software eingespielt werden können. Stirbt eines dieser ins Archiv geholten Formate aus, dann müssen die Dateien vom Archiv in ein neues Format übertragen werden.

Die weitaus meisten Archive setzen auf die Migrationsstrategie. Statistiken werden international wohl ausschließlich mit dieser Strategie archiviert. Sie werden dabei zumeist im CSV-Format, manchmal auch im Festbreitenformat oder in XML abgelegt. Alle drei genannten Formate können wohl noch eine ganze Reihe von Jahren genutzt werden, bevor eine Migration ansteht. Im Anschluss werden dann zwei Textformate vorliegen. Sie ermöglichen die Darstellung derselben Information auf unterschiedliche Weise. Im Laufe der Zeit wird es so eine ganze Reihe verschiedener Ausprägungen eines digitalen Archivales geben, die alle dieselben Inhalte haben. Infolge dieser Eigenschaft müssen die logischen Bestandteile einer Statistik unabhängig von den physischen Erscheinungsformen beschrieben werden. Die logische Beschreibung zielt auf die Inhalte (zum Beispiel Volkszählung 1970), die physische auf die Formate. Während bisher immer ein Objekt in den Lesesaal bestellt werden konnte (zum Beispiel Pergamenturkunde von Graf Ludwig II. von 1451), bedarf es nun einer doppelten Angabe: Welches logische Objekt und welche Erscheinungsform sollen benutzt werden.

Metadaten und Pakete

Mit dem Speicherkonzept und der Migrationsstrategie sind zwei zentrale Anforderungen beschrieben. Was aber nützen physisch lesbare Primärdaten, wenn diese nicht mehr verstanden werden können? Gerade digitale Archivalien benötigen ein besonders umfangreiches Set an Metadaten, damit sie auch in 100 Jahren noch gelesen und benutzt werden können. Bei digitalen Archivalien werden daher nie „nur“ die Primärdaten, sondern immer ganze Informationspakete, bestehend aus Primär- und Metadaten, übernommen, gespeichert und genutzt. Nach Funktion werden Übernahme-, Archivierungs- und Benutzungspakete unterschieden. Das Archivierungspaket umfasst zunächst Kontextinformationen: Um was handelt es sich, verbirgt sich hinter der endlosen Zahlenreihe vielleicht etwas Sinnvolles? Und wenn die Kontextmetadaten dann erklären, dass hier die Volkszählung von 1970, erstellt durch das Statistische Landesamt Baden-Württemberg, vorliegt – wie kann diese Zahlenreihe interpretiert werden? Codelisten sind ebenso in das Archivierungspaket aufzunehmen wie Angaben über die Codierung der einzelnen Ziffern und Buchstaben: Sind die hinter dem Bitstrom zu vermutenden Zeichen nun über ASCII, UNICODE-8

oder UNICODE-16 aufzulösen? Ein kleiner Teil dieser Metadaten wird durch das Archiv selbst erstellt (zum Beispiel die Dokumentation der Archivierung in einem Protokoll). Wichtiger aber ist, dass die für ein Verständnis wesentlichen Metadaten noch in der abgebenden Behörde aufgefunden und an das Archiv übergeben werden können. Ohne diese Informationen sind auch die Primärdaten einer Statistik nicht benutzbar.

Bewertung

Grundsätzlich sieht das Landesarchivgesetz vor, dass alle Unterlagen dem Landesarchiv angeboten werden müssen, wenn sie in der Behörde nicht mehr gebraucht werden. Dennoch wird das Landesarchiv auch bei den digitalen Unterlagen nur einen kleinen Teil dessen übernehmen können, was ihm angeboten wird. Dabei hat sich die Auswahl statistischer Daten stets an den mutmaßlichen Interessen künftiger Benutzer zu orientieren. Inhaltlich bedeutet dies zunächst, dass das Thema auch nach einigen Jahrzehnten noch von einer gewissen Bedeutung sein sollte. Beispielsweise werden die spätmittelalterlichen Bevölkerungslisten noch heute intensiv in den Archiven genutzt. Ähnliches ist von den Volkszählungen 1961, 1970 und 1987 anzunehmen, die in den letzten Jahren ins Landesarchiv übernommen werden konnten. Annehmen können wir, dass auch in Zukunft in Statistiken vor allem ein Abbild von Realitäten gesucht wird. Plausibilisierte Mikrodaten bilden diese Realitäten besser ab als die ihnen vorausgehenden Erhebungsdaten. Auch sollte es möglich sein, die von den Befragten gegebenen Antworten in ihrem ursprünglichen Zusammenhang auszuwerten, also beispielsweise nach dem gleichzeitigen Auftreten von zwei Werten fragen zu können. Aus diesem Grund konzentrieren sich die Aktivitäten des Landesarchivs auf die nicht aggregierten plausibilisierten Mikrodaten. Manche Statistiken sind schon hinreichend über die Veröffentlichungen der Statistischen Ämter beschrieben. Die Entscheidung für eine Archivierung der Mikrodaten muss der Überzeugung folgen, dass diese gegenüber den Veröffentlichungen einen klar erkennbaren Mehrwert enthalten.

Ausblick

Seit dem 1. Januar 2007 ist nun die bereits eingangs erwähnte Vereinbarung zwischen den Statistischen Ämtern und den staatlichen Archivverwaltungen in Kraft. Nach Ablauf einer mindestens 10 Jahre dauernden Verwahrungsfrist sind die Statistiken in einem Archivierungsplan aufzulisten und dem Landesarchiv anzubieten. Das Landesarchiv wird dann die Statistiken bewerten und in Teilen übernehmen. Natürlich ist es von Vorteil, wenn die archivierten Statistiken auch über die Grenzen eines Bundeslandes hinaus miteinander verglichen werden können. Aus diesem Grund erarbeitet derzeit eine Arbeitsgruppe der Archivreferentenkonferenz eine Liste archivwürdiger Statistiken. Neben den aktuellen und künftigen Statistiken beschreibt die Rahmenvereinbarung auch das Vorgehen bei Altstatistiken. Geplant ist eine Bestandsaufnahme der sogenannten Altdatenmaterialien. Es bleibt zu hoffen, dass sich bei möglichst vielen historischen Statistiken nicht nur die Primär- sondern auch die erforderlichen Metadaten erhalten haben. Der Archivar kennt gegen den Zahn der Zeit nur eine Medizin: Die Bestandsaufnahme der Altstatistiken, ihre Bewertung und schließlich die Übernahme der archivwürdigen Daten sollte so rasch wie möglich vorgenommen werden. ■

kurz notiert ...

Fußball-EM 2008: Über 800 000 Einwohner in Baden-Württemberg kommen aus einem der 15 ausländischen Teilnehmerländer

In Baden-Württemberg leben mehr als 801 000 Personen, die die Staatsangehörigkeit eines der insgesamt 15 ausländischen Teilnehmerländer bei der Fußball-Europameisterschaft 2008 besitzen. Die ausländische Fan-Gemeinde bei der diesjährigen Fußball-Europameisterschaft ist deutlich größer als bei der Fußball-Weltmeisterschaft vor 2 Jahren. Seinerzeit hatten lediglich gut 500 000 der hier lebenden ausländischen Einwohner den Pass eines der 31 ausländischen Teilnehmerländer der Weltmeisterschaft.

Mit über 292 000 Personen besitzt die Türkei bei der Fußball-Europameisterschaft unter den teilnehmenden ausländischen Nationen die mit Abstand größte Fan-Gemeinde in Baden-Württemberg, gefolgt von Italien (164 000 Einwohner), Kroatien (75 000 Einwohner) und dem amtierenden Europameister Griechenland (70 000 Einwohner). Allein auf diese vier genannten Nationen entfallen zusammen 75 % der hier lebenden Ausländerinnen und Ausländer aller Teilnehmerländer. Zahlenmäßig vergleichsweise schwach vertreten sind die Teilnehmerländer Schweden, Tschechien und die Niederlande, von denen jeweils weniger als 10 000 Personen in Baden-Württemberg wohnen. ■