DIMAG und IngestList
Übernahme, Archivierung und Nutzung von digitalen Unterlagen im Landesarchiv Baden-Württemberg

Von Christian Keitel und Rolf Lang

Die Software-Programme DIMAG und IngestList wurden im Rahmen des Projekts *Konzeption für ein digitales Landesarchiv* entwickelt,[1] in dem die Grundlagen für die digitale Archivierung im Landesarchiv Baden-Württemberg gelegt werden sollten. Das zentrale Projektziel war, in den Feldern der Übernahme und Archivierung digitaler Unterlagen die Handlungsfähigkeit des Landesarchivs herzustellen. Außerdem sollte es von Anbeginn an möglich sein, die übernommenen Unterlagen auch nutzen zu können. Weitergehende Nutzungsmöglichkeiten, wie sie beispielsweise bereits das National Digital Archive of Datasets im Internet anbietet,[2] sollten erst zu einem späteren Zeitpunkt realisiert werden.

Für die Übernahme digitaler Unterlagen setzt das Landesarchiv Baden-Württemberg das Programm IngestList ein. Die Archivierung erfolgt im Digitalen Magazin DIMAG. Die Funktionsweise der beiden Programme bildet die Prämissen ab, mit denen das Projekt konzipiert wurde. In den Prämissen spiegeln sich wiederum die Rahmenbedingungen des Landesarchivs einerseits und die spezifischen Anforderungen der digitalen Unterlagen andererseits wider. In DIMAG haben sich die parallel entwickelte Metadatenkonzeption und das darin begründete Repräsentationenmodell niedergeschlagen. IngestList kann als ein Antwortversuch auf die Frage verstanden werden, wie digitale Unterlagen glaubwürdig übernommen und erhalten werden können.

Organisation der digitalen Archivierung

Digitale Archivierung wurde im Landesarchiv Baden-Württemberg nie als die Angelegenheit einer von den restlichen Abteilungen und Referaten isolierten Stelle begriffen. Vielmehr sollte die Aufgabe unter möglichst weitgehender Beteiligung der sechs Staatsarchive entwickelt werden.[3] Sie sind nicht nur für die Bewertung der digitalen Unterlagen zustän-

[1] Das Projekt ging zunächst von September 2005 bis Ende 2008 und wurde dann bis Ende 2009 verlängert. Dem Projektteam gehörten Dr. Kai Naumann, Rolf Lang und Dr. Christian Keitel (Leitung) an.

[2] URL: http://www.ndad.nationalarchives.gov.uk/ (11. Dezember 2009).

[3] Zur Aufgabenverteilung vgl. Christian *Keitel*: Die Archivierung elektronischer Unterlagen in der baden-württembergischen Archivverwaltung. Eine Konzeption. 2002. URL: http://www.landesarchiv-bw.de/sixcms/media.php/25/keitel_elektronische_konz.pdf (1. Dezember 2009).

dig. Diese werden auch in den bereits bestehenden Beständeübersichten und Findmitteln der Häuser nachgewiesen. Die theoretisch denkbaren anderen Lösungen, also eine zentrale Beständeübersicht für die digitalen Unterlagen oder eine solche für jedes einzelne Haus, bildeten nie eine ernsthafte Alternative. Dieser Ansatz legt es nahe, dass die von den Häusern bewerteten und nachgewiesenen Archivalien auch in den Lesesälen der Häuser einsehbar sein sollten. Die Nutzerinnen und Nutzer sollen in anderen Worten die Möglichkeit haben, sämtliche von einer Stelle übernommenen Unterlagen gemeinsam auszuwerten, sie sollen also abwechselnd in die Papierakte und auf den Computermonitor mit dem dort aufscheinenden Fachverfahren sehen können. Möglich werden soll dieses in den Lesesälen der Staatsarchive. Wäre es dann nicht am einfachsten, wenn die digitalen Unterlagen ebenso wie ihre konventionellen Geschwister in den Staatsarchiven selbst archiviert werden würden? Diese Vorstellung erschien 2005 dann doch als zu utopisch. Zwischen 2002 und 2012 muss das Landesarchiv Baden-Württemberg 20 Prozent seines Personals abbauen und denselben Anteil auch bei den Sachmitteln einsparen. In einer solchen Situation ist es nicht vorstellbar, dass zu den von immer weniger Händen zu erledigenden bestehenden Fachaufgaben weitere komplexe und schulungsintensive Aufgaben hinzukommen. Stattdessen sollten Aufbereitung, Archivierung und digitale Bestandserhaltung zentral erledigt werden. Auch in dieser Vorgabe schlägt sich die Struktur des Landesarchivs nieder, hat dieses doch seit Langem zentrale Abteilungen gehabt, die gerade solche Aufgaben übernommen haben, die eben nicht an das räumliche Umfeld der einzelnen Häuser gebunden waren und rationell an einer Stelle für alle gemeinsam erledigt werden konnten.

Ein derartiges Zusammenspiel von zentralen und dezentralen Einheiten ließ sich in vergangenen Zeiten gut über Briefe und wechselseitige Besuche organisieren. In der digital gewordenen Welt mussten die Mittel angepasst werden. Ein Ansatzpunkt war die Verständigung auf ein archivisches Intranet und die Übermittlung geschützter Informationen über speziell gesicherte Internetprotokolle.[4] Aber auch die Archivierungssoftware sollte in die Lage versetzt werden, einen Aufruf von ganz unterschiedlichen Punkten aus zu ermöglichen. Das Archivierungssystem wurde daher von vornherein als ein System konzipiert, das sich im Browser aufrufen lässt.

DIMAG existiert heute als Intranetserver am Standort Ludwigsburg. Die Kommunikation der Anwender mit dem Server erfolgt über eine Webschnittstelle, sie basiert auf dem 256 Bit stark verschlüsselten https-Protokoll. Damit ist die Öffnung ins Internet technisch möglich und sicher. Voraussetzung für die Nutzung sind zum einen die Freischaltung des Ports 443 für wohl definierte IPs (Nutzer) und zum anderen personenbezogene Zugangsdaten bei DIMAG.

Ähnlich verhält es sich mit dem ftp-Zugang. Die Kommunikation darf nicht transparent erfolgen. Sowohl die Log-in-Zugangsdaten als auch die transferierten Dateien müssen hochgradig verschlüsselt werden. Dies wird erreicht mit dem sftp (secure shell ftp)[5]-

[4] Christian *Keitel:* Zugänglichkeit contra Sicherheit? Digitale Archivalien zwischen Offline-Speicherung und Online-Benutzung. 2002. URL: http://www.landesarchiv-bw.de/sixcms/media.php/25/zugaenglichkeit%20contra%20sicherheit.pdf (1. Dezember 2009). Auch Christian *Keitel*, wie Anm. 3.

[5] URL: http://de.wikipedia.org/wiki/SSH_File_Transfer_Protocol (21. Dezember 2009).

Protokoll. Dieses hat auch Vorteile bezüglich der Portfreischaltung, da nur ein Port (220) verwendet wird. Andere ftp-Verfahren verwenden für den Datentransfer die zur Laufzeit ausgehandelten Port-Nummern.

Jede weitere Kommunikation mit dem Server – ausgenommen direkt an der Konsole in Ludwigsburg – geschieht verschlüsselt. Dies gilt für den Servicezugang wie auch für das Back-up-Verfahren über rsync oder Tivoli.[6]

Aufgaben des Archivierungssystems

Welche Funktionen sollte das Archivierungssystem nun abdecken? Zur Beantwortung dieser Frage wird in aller Regel ein Blick auf das Funktionsmodell des OAIS-Standards geworfen.[7] Das Modell unterscheidet dabei zwischen fünf zentralen Einheiten: Übernahme und Aufbereitung (Ingest), Nachweis (data management), Archivierung (Archival Storage), Bestandserhaltung (Preservation Planning) und Nutzung (Access). 2005 besaß das Landesarchiv mit MIDOSA 21 ein System, dessen Bestandteile bereits die Bereiche Data Management (scopeArchiv) und Access (OLF21) abdeckten. Wäre es nicht am einfachsten, in diesen Programmen auch die Archivierung vorzunehmen? Schon bei der Planung des Projekts war klar, dass die bestehenden Produkte die spezifischen Anforderungen des Bereichs nicht erfüllen können. Es sollte also eine eigene Softwarelösung für die Archivierung beschafft oder entwickelt werden.

Im Archivierungssystem sollten sämtliche von den baden-württembergischen Staatsarchiven übernommenen digitalen Unterlagen archiviert werden. In manchen Papierarchiven ist nun das Phänomen bekannt, dass zwischen Übernahme und Erschließung ein längerer Zeitraum liegen kann. Im ungünstigsten Fall werden die übernommenen Akten, Karten und Bände vergessen und nach vielen Jahren als kaum mehr einzuordnender Stapel wiederentdeckt. Auch bei digitalen Unterlagen gibt es die Phase der Aufbereitung, in der die übernommenen Unterlagen erst noch für die Archivierung selbst vorbereitet werden müssen. Hier würde ein großer zeitlicher Abstand zwischen der Übernahme einerseits und der Aufbereitung und Erschließung andererseits die Archivierung in den meisten Fällen verhindern, da viele wesentliche Metadaten nur unmittelbar nach der Übernahme zusammengetragen und überprüft werden können. Entsprechende Datenverluste werden dann wahrscheinlich, wenn digitale Archivalien längere Zeit bis zur Aufbereitung in Form von Wechseldatenträgern oder auf anderen Medien außerhalb des Archivierungssystems zwischengelagert werden. Das Archivierungssystem sollte die digitalen Unterlagen daher bereits unmittelbar nach ihrer Übernahme zusammen mit einigen rudimentären Metadaten aufnehmen können. Es sollte also neben der Archivierung (Archival Storage) auch die Aufbereitung und damit den zweiten Teil des Ingests umfassen.

---

[6] URL: http://de.wikipedia.org/wiki/Rsync (21. Dezember 2009); URL: http://www-142.ibm.com/software/products/de/de/tivostormana (21. Dezember 2009).

[7] Open Archival Information System, kurz OAIS (ISO 14721), URL: http://public.ccsds.org/publications/archive/650x0b1.pdf, S. 1–3 (7. Dezember 2009).

Mit den genannten Prämissen ging das Projektteam 2006 daran, die auf dem Markt befindlichen Systeme FEDORA, DAITSS und DSpace zu testen.[8] Diese unter anderem von Bibliotheken eingesetzten Systeme erfüllten die spezifischen archivischen Anforderungen nicht. Andere Systeme wie DIAS kamen wegen der sogenannten Black-Box-Problematik nicht infrage – Speicherung und Speicherorte sind für das Archiv nicht nachvollziehbar. Das Landesarchiv Baden-Württemberg entschloss sich daher, selbst ein Archivierungssystem zu entwickeln. Das Digitale Magazin DIMAG sollte mit den bereits bestehenden Komponenten von MIDOSA 21 (scopeArchiv und OLF21) verbunden werden. Für die Entwicklungsphase erschien es allerdings sowohl wegen der Flexibilität als auch wegen des Aufwands sinnvoller, das Archivierungssystem zunächst als Stand-Alone-System aufzubauen.

Das Digitale Magazin DIMAG

Ein Archivierungssystem muss die digitalen Archivalien sicher verwahren, und da zum Verwahren auch das Speichern der Daten gehört, wurde gleich zu Beginn des Projekts über die möglichen Speicher diskutiert. In der Diskussion standen Magnetbandkassetten und Festplattensysteme; die Entscheidung fiel zugunsten der Letzteren aus. Magnetbandkassetten boten 2005 zwar einen günstigeren Speicherplatz als Festplatten an. Das preisliche Verhältnis erschien aber mittelfristig nicht genügend stabil, um als wesentliches Argument gewichtet zu werden. Festplatten sind flexibler, sie erleichtern es, die gerade während einer Einführungsphase zu erwartenden häufigen Änderungen vorzunehmen. Sie sind außerdem schneller im Zugriff und ermöglichen nach einigen Jahren auch einen schnelleren Umstieg auf künftige Speichermedien (Datenträgermigration).

Es liegt auf der Hand, dass Archivierungssysteme von den zu verwendenden Speichermedien abhängig sind. Beide zählen mindestens teilweise zum OAIS-Bereich Archival Storage. Auf der anderen Seite adressieren Archivierungssysteme einerseits und Datenträger andererseits unterschiedliche Aufgaben. Auch ist es mittel- und langfristig erstrebenswert, das System zur Verwaltung der Archivalien und die zu verwendenden Datenträger vollständig zu entkoppeln. Die Einführung neuer Speichermedien würde dann keinen Austausch des Archivierungssystems erfordern.[9] Zukünftig wird eine klare Unterscheidung zwischen den beiden Verantwortungsbereichen vorgenommen werden. Auf der einen Seite steht das Archivierungssystem DIMAG, welches die einheitliche Sicherung und Beschreibung von digitalen Objekten vornimmt, und auf der anderen der Betrieb beziehungsweise

[8] FEDORDA, URL: http://fedoraproject.org/ (21. Dezember 2009); DAITSS, URL: http://www.fcla.edu/digitalArchive/ (21. Dezember 2009); Dspace, URL: http://www.dspace.org (21. Dezember 2009).

[9] Vgl. zum Beispiel Jim *Linden*, Sean *Martin*, Richard *Masters* and Rorderic *Parker*: The lange-scale archival storage of digital objects. DPC Technology Watch Series Report 04-03. Februar 2005. URL: http://www.dpconline.org/technology-watch-reports/download-document/89-the-large-scale-archival-storage-of-digital-objects.html (21. Dezember 2009).

die Serveradministration. Letztere wird voraussichtlich in ein landeseigenes Rechenzentrum ausgelagert werden.

Bei der Konzeption von DIMAG war wichtig, das System redundant in seinen Einzelkomponenten aufzubauen. Ein möglicher Ausfall einer Einzelkomponente hat damit keine schwerwiegenden Konsequenzen. Diese Redundanz ist sowohl in der Hardware-Architektur als auch im Software-Design realisiert. Im Folgenden einige Beispiele hierfür:

Das verwendete Dateisystem ist ein Journaling-Dateisystem, entwickelt von Hans Reiser, welches zusätzliche Ausfallsicherheit bietet. So werden alle Änderungen vor dem eigentlichen Schreiben in einem dafür reservierten Speicherbereich, dem Journal, aufgezeichnet. Sollte nun es zu einem Ausfall kommen, so ist es jederzeit möglich, einen konsistenten Zustand der Daten zu rekonstruieren. Erfolgte der Ausfall während des Beschreibens der Journaldatei, wird dies beim Neustart ignoriert; erfolgte der Abbruch während der Übertragung vom Journal ins Dateisystem, so wird dieses erkannt und der ursprüngliche Zustand kann vollständig wiederhergestellt werden.

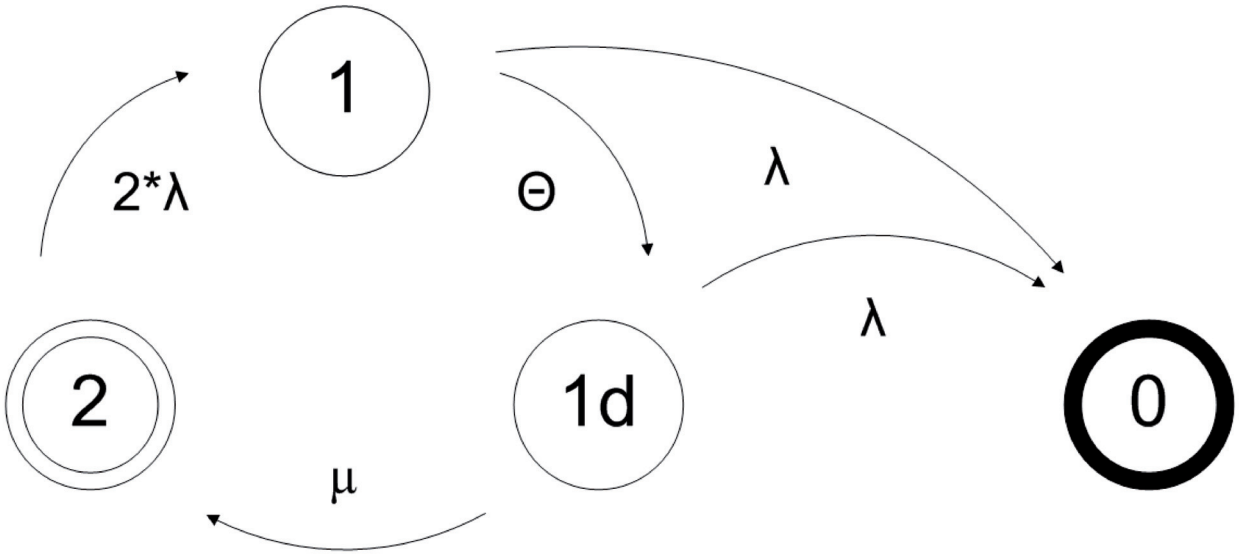
Bereits das Festplattensystem ist redundant als RAID 5 (redundant array of independant disks) aufgebaut. Dies ist eine sehr beliebte Variante der verschiedenen RAID-Verfahren und beruht auf Striping & Distributed Parity. Dabei werden die Nutzdaten über mehrere Festplatten von dem Raid-Controller aufgeteilt. Zusätzlich werden Parity-Blöcke gebildet, welche so geschickt auf den Festplatten verteilt werden, dass durch den Ausfall einer Festplatte das System mithilfe der Parity-Blöcke wieder hergestellt werden kann. Die nutzbare Gesamtkapazität errechnet sich aus der Formel (Anzahl der Festplatten – 1) * (Kapazität der kleinsten Festplatte). Im Fall von DIMAG sind dies derzeit fünf Terabyte.

Eine weitere wesentliche Grundentscheidung war der Aufbau von DIMAG als Datenbank- und Dateisystem. Dies wird in vielen Systemen so praktiziert, die Unterschiede sind nur: werden Metadaten, Beziehungen und Primärdaten redundant geführt oder nicht. So werden beispielsweise in DSpace die Metadaten nur in der Datenbank geführt, von der ein eindeutiger Schlüssel ins Dateisystem zu den Primärdaten führt. Ein Ausfall der Datenbank hätte daher einen irreparablen Datenverlust zur Folge.

In DIMAG werden alle Meta-, Beziehungs- und Primärdaten im Dateisystem geführt. In der Datenbank sind Metadaten und Beziehungen nochmals redundant vorhanden. Primärdaten werden nicht mehr redundant in der Datenbank vorgehalten, obwohl dies bis zu einer bestimmten Größe denkbar wäre.

Kritische Stimmen können nun fragen: Reicht denn eine Sicherung aus? Hierzu möchten wir das von Markov entwickelte Modell von zwei gleichartigen Speichersystemen heranziehen. Markov beschreibt dieses Modell mittels vier Zuständen.

Im Ausgangszustand (2) sind beide Speicher in Ordnung. Im Zustand (1) ist ein Speichersystem ausgefallen. Dies wird mit der MTTF (mean time to failure)-Angabe quantifiziert. Ein weiterer Folgezustand (1d) ist erreicht, wenn mit der MTTFD (mean time to failure detection) das Problem erkannt ist. Erst dann kann gehandelt werden und das System mit der MTTR (mean time to repair) in den gewünschten Ausgangszustand zurückgeführt werden. Interessant ist nun, wenn im Zustand 1 oder 1d das noch intakte System auch ausfällt. Dann bedeutet dies Totalverlust trotz Sicherung. Dieses Modell erlaubt es nun, den



$$L = \frac{(\theta+\lambda)(\mu+\lambda)+2\lambda(\mu+\lambda)+2\lambda\theta}{2\lambda^2(\theta+\mu+\lambda)}$$

¶

L = Time to lose all data

$\mu = 1/$ mean time to repair (MTTR)

$\lambda = 1/$ mean time to failure (MTTF)

$\theta = 1/$ mean time to failure detection (MTTFD)

Totalverlust zu berechnen und kommt beispielsweise mit den Angaben (MTTF = 3 Jahre, MTTR = 50 h, MTTRD = 14 Tage) auf 106 Jahre.

Einige prinzipielle Anmerkungen hierzu: Es ist natürlich nur ein einfaches Modell, welches beispielsweise die sogenannte Badewannenkurve der Hardware-Ausfälle nicht berücksichtigt. Diese sieht drei Bereiche vor: häufige Frühausfälle zum Beginn, seltene statistische Ausfälle während einer langen Laufzeit und am Lebensende eine stetige Zunahme durch Verschleiß. Auch von hp gibt es hierzu kritische Anmerkungen.[10]

Markov selbst beschreibt darüber hinaus weitere und komplexere Modelle mit drei Speichersystemen, welche zu deutlich längeren Zeiten bis zum Totalverlust führen.[11]

Das Landesarchiv Baden-Württemberg hat sich daher für doppelte Datensicherung mit zwei unterschiedlichen Verfahren entschieden. Jede Datei sollte demnach auf dem Produktionssystem im Staatsarchiv Ludwigsburg und außerdem in zwei Kopien vorliegen. Das Besondere daran ist, dass vor der eigentlichen Sicherung die Integrität der Daten des Produktionssystems anhand der md5-Dateien überprüft wird. Md5 ist eine weitverbreitete kryptografische Hashfunktion, die aus einer beliebigen Datei einen 128-Bit-Hashwert erzeugt. Dieser ist nicht eindeutig, da es nur $3{,}4 * 10^{38}$ verschiedene md5-Werte gibt (Zahlenvergleich: Masse der Erde $5{,}9 * 10^{27}$ Gramm). Er stellt aber sicher, dass minimalste Änderungen innerhalb einer Datei einen vollkommen anderen md5-Wert erzeugen. Erst wenn aus der Sicht von DIMAG alle Dateien den gleichen md5-Wert haben, wie er beim Ingest abgelegt wurde, startet die Datensicherung.

Damit wird verhindert, dass zerstörte Daten auch auf das Sicherungssystem gelangen und dort im schlechtesten Fall unverfälschte Dateien überschreiben.

Metadatenkonzept

DIMAG hat unter anderem die Aufgabe, sämtliche Ansichts- und Bearbeitungsvorgänge über eine Rechteverwaltung zu kanalisieren. Mindestens die Archivare sollten unter Einhaltung bestimmter Berechtigungen die Möglichkeit haben, digitale Archivalien innerhalb von DIMAG aufzurufen. DIMAG muss daher bei seiner Rechteverwaltung sowohl zwischen den Nutzern als auch zwischen den Objekten und, weitergehend, zwischen Metadaten und Primärdaten unterscheiden können. Und wenn ein Nutzer dann mit seinen Berechtigungen versehen auf das System sieht, sollte er auch die Möglichkeit haben, anhand von Metadaten das ihn Interessierende zu finden. Das System musste daher mehr tun als nur eine große Menge nicht unterscheidbarer Objekte anhand von IDs zu verwalten. Stattdessen zwangen Rechteverwaltung und Recherchefunktionalität, im Archivierungssystem zwischen den verschiedenen Einzelteilen anhand der Metadaten zu unterscheiden. Metadaten waren aber noch aus einem weiteren Grund für DIMAG „lebensnotwendig“: Schließlich sollte die ganze Aufbereitungsphase im DIMAG erfolgen. DIMAG musste da-

[10] ‘Reliability of Markov Models are Becoming Unreliable’, URL: http://www.usenix.org/events/fast08/wips_posters/slides/greenan.pdf (21. Dezember 2009).

[11] URL: http://www.ics.forth.gr/isl/publications/paperlink/Reliability%20modelling.pdf (21. Dezember 2009).

her alle Metadaten einfordern, die für ein Archivierungspaket (OAIS: Archival Information Package) benötigt wurden. DIMAG konnte daher nur dann aufgebaut werden, wenn sich das Projektteam über die notwendigen Metadaten und ihren Zusammenhang Klarheit verschaffen konnte.

Das Metadatenkonzept für digitale Unterlagen musste drei Gegebenheiten Rechnung tragen. Berücksichtigt werden mussten die Eigenschaften der zur Archivierung anstehenden Unterlagen selbst, die bereits bestehenden und in den Findmitteln geronnenen Metadatenstrukturen des Landesarchivs und die zu berücksichtigenden Arbeitsabläufe.

Welche digitalen Unterlagen standen im Landesarchiv Baden-Württemberg zur Archivierung an? Schon in der Projektplanung 2005 zeichnete sich ab, dass in den nächsten Jahren so unterschiedliche digitale Objekte wie digitale Fotos, Daten aus Fachverfahren und Geoinformationssystemen oder einfache Textdokumente archiviert werden sollten. Das Archivierungssystem sollte daher von vornherein für sämtliche denkbaren digitalen Unterlagen offen sein. Entsprechend offen musste denn auch das Metadatenkonzept entwickelt werden. Die DOMEA-Abstufung zwischen Akte, Vorgang und Dokument sollte zwar abbildbar sein, aber keine feste Vorgabe darstellen.

Was sind nun die Gemeinsamkeiten von allen digitalen Unterlagen? An erster Stelle kann hier die Trennung zwischen Information und Ausprägung genannt werden. Erhalten werden sollen in erster Linie Informationen. Diese können nur in einer materialisierten Form überleben – das heißt auf einem bestimmten Datenträger und in einem spezifischen Dateiformat –, aber gerade diese Erscheinungsformen haben nur eine sehr kurze Lebensdauer. Die zu erhaltenden logischen Informationen müssen daher immer wieder in neue physische Formen überführt werden (Migrationsstrategie). In Anlehnung an PREMIS wurde die Erscheinungsform einer Informationseinheit Repräsentation genannt. Diese Repräsentation kann beliebig viele Dateien in unterschiedlichen Dateiformaten enthalten. Repräsentation und Informationseinheit werden in DIMAG durch eigene Datensätze oder XML-Dateien beschrieben. Zusammen ergeben sie das digitale Archivale. In einer grafischen Darstellung hängen von der Informationseinheit (Digitales Objekt) auf einer sich unmittelbar anschließenden, tiefer liegenden Ebene beliebig viele Repräsentationen ab. Das Digitale Objekt selbst ist wiederum in die Tektonik der Staatsarchive eingebaut. Damit sind die Digitalen Objekte auf derselben Ebene wie die konventionellen Archivalien angesiedelt, sie werden entsprechend der Vorgabe zusammen und in einem System nachgewiesen.

In den Repräsentationen liegen die übernommenen Primärdaten. Jede Primärdatendatei wird durch eine Metadatendatei im XML-Format beschrieben. Metadatendateien beschreiben auch die Repräsentationen, die Digitalen Objekte und alle höher liegenden Einträge in Klassifikation und Tektonik. Zusammen mit den Primärdaten ergeben sie das Archivierungspaket. Aus den Archivierungspaketen kann DIMAG verschiedene Nutzungspakete generieren. Derzeit ermöglicht das System die Auswahl zwischen den Optionen *Metadatenliste* (in HTML oder als URL), *Metadaten* (im METS-Format) und *Primär- und Metadaten* – im METS-Format mit Base64-Codierung der Primärdaten, als Verzeichnisbaum für ein lokales Web oder bei einer einzelnen Primärdatendatei mit einem reduzierten Metadatensatz im txt-Format. Die Nutzung erfolgte bisher auf den Rechnern im Lesesaal des Staatsarchivs Ludwigsburg oder durch Abgabe eines Wechseldatenträgers.

Authentizität

Die leichte Veränderbarkeit digitaler Unterlagen generell und die Notwendigkeit, die zu archivierenden Informationen immer wieder in neue Repräsentationen zu überführen, stellen die Archivarinnen und Archivare vor besondere Herausforderungen. Wie kann künftigen Nutzern glaubwürdig versichert werden, dass die ihnen vorgelegten Unterlagen die vom Archiv vor langer Zeit übernommenen Informationen enthalten, dass diese Informationen sich weder vermehrt noch verringert oder verändert haben? Zunächst muss das Archiv seine Handlungen und Geschäftsregeln dokumentieren. Das Landesarchiv Baden-Württemberg erledigt dies durch den DIMAG-Bereich *Dokumentation*, der genau diese allgemeinen Angaben zur Archivierung enthält. Sie werden in DIMAG wie gewöhnliche digitale Archivalien verwahrt. Zum Zweiten besitzt jedes in DIMAG gespeicherte digitale Archivale ein Protokoll, in dem alle wesentlichen Vorkommnisse im Lebenslauf dieses Objekts festgehalten werden. Drittens müssen digitale Archivalien nach jeder Veränderung validiert werden; es muss geprüft werden, ob das Ergebnis auch den Absichten entsprechend ausgefallen ist. Konzeptionell konnte die Aufgabe wie folgt beschrieben werden: Der Lebenslauf digitaler Archivalien ist eine Abfolge von Phasen, in denen das Objekt unverändert bleibt, und Phasen, in denen des verändert wird. Im ersten Fall können Hashwerte die Unversehrtheit des Archivales belegen. Im zweiten Fall ist es möglich, von einem Transfer zu sprechen. Ein Transfer findet statt bei der Übernahme von der abgebenden Stelle ins Archiv, er findet bei einer Datenträgermigration ebenso wie bei einer Dateiformatmigration statt. Immer ist es unklar, ob das Ergebnis den Vorgaben entspricht. Seine Validierung kann durch den Vergleich mit einer externen Vorgabe – zum Beispiel ein Format-Standard – oder durch einen Vergleich mit der vor dem Transfer bestehenden Ausprägung der Information erfolgen. Hierfür werden vor und nach dem Transfer dieselben Metadaten abgefragt und diese beiden Ergebnisse dann miteinander verglichen. Diese Aufgabe wird von dem Programm IngestList übernommen. Außerdem ermöglicht IngestList den automatisierten Datenimport nach DIMAG.

Einfacher Ingest nach DIMAG

Um einzelne digitale Unterlagen nach DIMAG einzustellen, gibt es ein zentrales Formular. Damit werden sowohl die Metadaten als auch die Primärdaten erfasst. Eine Ablage der Primärdaten ohne Metadaten wird vom System verhindert. Viele technische Metadaten werden selbsttätig erfasst. Wer hat beispielsweise wann was eingestellt. Erfasst werden aber auch weitere für die Bestandserhaltung wichtige Werte wie der md5-Hashwert der Primärdaten, die Bestimmung des Datenformats inklusive einer Referenz zum Formatregister Pronom sowie die Erhebung des Zeichenformats. Alle diese Metadaten stehen in einem hierarchischen Zusammenhang und werden von einer höheren Instanz auf die darunterliegende vererbt.

Natürlich ist es gut, wenn einzelne Primärdateien archiviert werden können. Für größere Datenmengen ist dies aber unzureichend. Hier benötigt man ein Verfahren, bei dem mit einem Rutsch eine Vielzahl von Primärdaten archiviert werden können.

Hierzu gibt es nun bei DIMAG einen sogenannten workspace. Dorthin können die Daten über sshftp transferiert werden. Die massenhafte Übernahme erfolgt anschließend, indem die workspace-Daten von der Tektonik ausgehend verortet werden. Die Erstellung und der Transfer der Daten erfolgt mit IngestList.

Ingest mit IngestList

IngestList ist eine javabasierte Eigenentwicklung des Landesarchivs zur Übernahme von Daten aus Fachverfahren (Datenbanken) und Dateien.[12] Das Programm wird sowohl von den abgebenden Stellen als auch vom Landesarchiv Baden-Württemberg eingesetzt. Es verwendet intern openSource-Klassen und Methoden von DROID,[13] bereitgestellt von *The National Archives* und jHove von *JSTOR* und der *Harvard University Library*.

Unter sourceForge sind zwei Flash-Demos eingestellt, welche schrittweise die verschiedenen Möglichkeiten der IngestList Nutzung aufzeigen.[14] Das Erste zeigt die grafische Oberfläche von IngestList.[15] Gezeigt wird hier exemplarisch, wie IngestList genutzt werden kann. Zuerst wird die aktuelle Signaturdatei von Pronom über den Webservice der National Archives abgefragt. Dann erfolgt eine verbale Beschreibung zur abgebenden Stelle: Wer hat wann was abzugeben. Zum Schluss werden die technisch erfassbaren Metadaten einer Dateisammlung automatisiert erhoben und bei Bedarf manuell ergänzt.

IngestList ist auch in der Lage, eine Verbindung zu Datenbanken (MySQL, MsAccess, Oracle) aufzubauen. Grundlage ist hier eine JDBC-Verbindung, welche künftig auch die Anbindung an weitere Datenbank Systeme ermöglicht. Alle Tabellen oder Views werden in IngestList zunächst aufgelistet und dann je nach Bedarf selektiert, gezählt und exportiert. Der Export einer Tabelle oder eines Views umfasst neben den Daten in der CSV-Darstellung die Datenbankbeschreibung der Tabelle, die Anzahl der in der Datenbank gezählten Zeilen und Spalten sowie das verwendete SQL mit möglichen Spalten- / Zeileneinschränkungen. Damit kann zu einem späteren Zeitpunkt ein Vergleich vorgenommen werden, ob die Datenmenge in der CSV-Tabelle noch den ursprünglich in der Datenbank erfassten Werten entspricht.

---

[12] URL: http://sourceforge.net/projects/ingestlist (21. Dezember 2009).
[13] URL: http://sourceforge.net/projects/droid/ (21. Dezember 2009).
[14] URL: http://hul.harvard.edu/jhove/index.html (21. Dezember 2009).
[15] URL: http://mesh.dl.sourceforge.net/project/ingestlist/IngestList/2009-10-21/flash/IngestListGUI.swf (21. Dezember 2009).

Die zweite Flash-Demo zeigt die Kommandozeilenversion von IngestList.[16] Diese benötigt keine menschlichen Interaktionen und ist gedacht für die Einbindung in bestehende Programme.

Ein Export der Daten von IngestList in eine Datenbank ist nicht vorgesehen. Dies ist auch nicht nötig, da Datenbanken selbst in der Lage sind, einfache Formate wie CSV zu importieren. Es bietet zudem den Vorteil, dass IngestList kein Schreibrecht für die Datenbank mit den zu übernehmenden Daten benötigt. Dadurch wird der Einsatz von IngestList in der abgebenden Stelle erleichtert, lässt doch niemand gerne ein ihm kaum bekanntes Programm mit Schreibberechtigung auf sein Fachverfahren zugreifen. Die fehlende Hemmschwelle erhöht so die Chancen des Landesarchivs, mit IngestList weitere Fachverfahren übernehmen zu können.

[16] URL: http://ovh.dl.sourceforge.net/project/ingestlist/IngestList/2009-10-21/flash/try.swf (21. Dezember 2009).