Auszug aus:
Katharina Ernst (Hrsg.): Erfahrungen mit der Übernahme digitaler Daten. Bewertung, Übernahme, Aufbereitung, Speicherung, Datenmanagement. Veröffentlichungen des Archivs der Stadt Stuttgart Bd. 99. Stuttgart 2007 (ISBN 978-3-89850-980-0).

– Speicherung –

Konzeption und Aufbau eines digitalen Archivs: Von der Skizze zum Prototypen

Christian Keitel, Rolf Lang, Kai Naumann

Gegen Ende des Jahres 2005 begann im Landesarchiv Baden-Württemberg das Projekt „Langzeitarchivierung digitaler Unterlagen“. Bis Dezember 2008 sollen die Möglichkeiten für eine dauerhafte Archivierung digitaler Unterlagen erkundet und verschiedene Umsetzungsmöglichkeiten entwickelt werden. Die Projektgruppe besteht aus zwei Archivaren und einem Informatiker. Die Projektkonzeption kennt vier zentrale Aufgabenbereiche: Bewertung und Übernahme, Aufbereitung, Archivierung und Benutzung. Mit der Metadatenkonzeption und dem Digitalen Magazin, kurz dimag, werden hier die Ergebnisse im Bereich der Archivierung vorgestellt[1].

1.

Wie wünschen wir uns die Entwicklung eines digitalen Archivs? Zunächst setzt sich der Archivar an seinen Schreibtisch und formuliert eine Prämisse nach der anderen, wägt die Einzelteile sorgsam gegeneinander ab und formuliert daraus ein umfassendes organisatorisch-technisch-finanzielles Gesamtkonzept. In einem zweiten Schritt setzt der Kollege eine Prämisse nach der anderen in entsprechenden Metadaten um. Nachdem auch das Metadatenkonzept steht, baut er sein digitales Archiv Bauteil um Bauteil zusammen.
In dem Projekt führten zwar manche Grundsätze zu einem Metadatum und dann zu einem entsprechenden dimag-Bauteil. Manchmal schärfte aber auch das Metadatum eines überseeischen Standards erst unseren Sinn für eine bestimmte Notwendigkeit, bevor diese im dimag umgesetzt wurde. Manchmal bildete sich auch erst während der Implementierung des einen Grundsatzes ein zweiter, weiterer heraus. Zahllose andere Varianten sind nicht nur denkbar, sie haben sich auch ereignet. Theorie und Praxis beeinflussen sich wechselseitig und können daher sinnvoll nur gemeinsam entwickelt werden. Nicht wenige Fragen haben sich erst im Laufe der Zeit ergeben; vieles ist abhängig von den dann schon getroffenen Entscheidungen und gemachten Erfahrungen. Es ist schlicht unmöglich, zunächst vollständige Baupläne anzulegen, bevor man sich an die Umsetzung dieser Pläne macht. Die erste Version des dimags steht und läuft im Alltagseinsatz. Allerdings ist der Prozess noch lange nicht abgeschlossen. Eher kann die Entwicklung eines digitalen Archivs als ein mehrstufiges Prototyping verstanden werden.

2.

Zentral für alle Projekte zum Aufbau eines digitalen Archivs ist das Thema „Sicherheit“. Auch nach der fünften Version des dimag soll noch genügend Grund zur Annahme bestehen, dass die Daten bereits in ihrer innerarchivischen Frühgeschichte sicher und vollständig archiviert – und das heißt, im dimag gespeichert – wurden. Einige Bedingungen müssen daher von Anfang an erfüllt werden, und sie können nur im dimag und nicht außerhalb des Systems erfüllt werden. Schon aus diesem Grund ist es wichtig, auch solche Zugänge rasch in das dimag zu bringen, die noch aufbereitet werden müssen. In der Papierarchivierung ist die Alternative hierzu gut bekannt. Da finden sich dann in manchen Ecken und Winkeln des Archivs längst vergessene Stapel mit Zugängen aus vergangenen Jahrzehnten, die immer noch nicht erschlossen werden konnten. Vergleichbare Disketten-, CD- oder DVD-Gräber sollten nach Möglichkeit erst gar nicht entstehen.
Der Zugang zu einem digitalen Archiv sollte limitiert sein. Dimag verfügt daher über eine eigene Rechteverwaltung. In regelmäßigen Abständen überprüft das System die Integrität der Daten mittels eines Hashverfahrens. Das ganze System wurde auf der Basis von Linux, PHP, MySQL und Apache programmiert, weshalb die Abhängigkeit von proprietären Systemen denkbar gering ist.
Zu Beginn des Projekts wurde die Entscheidung getroffen, zumindest während der Projektlaufzeit die Daten auf RAID-Systemen, also auf Festplatten, abzulegen. Für diese Architektur sprachen die Überschaubarkeit, der moderate Preis und die leichte Migrierbarkeit auf andere Datenträger. Nicht zuletzt sind RAID-Systeme sehr flexibel, was den Anforderungen unseres Projekts sehr entgegen kommt. Ob beispielsweise einzelne Metadatenfelder wirklich benötigt werden, erweist sich z. B. erst nach den in

[1] Zu den bisherigen Ergebnissen im Bereich der Bewertung und Übernahme vergleiche den Artikel „Handlungsfähige Archive: Erfahrungen mit der Bewertung und Übernahme digitaler Unterlagen“ im vorliegenden Band.

der Praxis gemachten Erfahrungen. Offen ist zum jetzigen Zeitpunkt, ob die digitalen Archivalien nach Projektabschluss weiterhin auf Festplatten oder auch oder ausschließlich auf Band gesichert werden sollen. Vielleicht werden auch manche Daten, insbesondere Bilder, auf Mikrofilmen ausbelichtet, um sie bei einer künftigen Benutzung wieder zu redigitalisieren. Bei diesem Konversionsverfahren kommen dann die Ergebnisse des ARCHE-Projekts zum Einsatz[2].
Das Produktivsystem des dimag steht im Staatsarchiv Ludwigsburg, eine Kopie geht nach Stuttgart in die ehemalige Landesarchivdirektion und eine weitere Kopie an eine Maschine im Hauptstaatsarchiv Stuttgart. Die Anforderungen an die Umsetzung waren: Der Transport sollte sicher sein und automatisiert ablaufen, eine vorübergehende Zwischenspeicherung wurde ausgeschlossen. Auf den beiden Backup-Servern laufen mit Suse Linux 10.0 und Windows Server 2003 zwei unterschiedliche Betriebssysteme, die das dimag zu unterschiedlichen Zeiten nach unterschiedlichen Verfahren ansteuert. Derzeit werden in dem einem Verfahren mit jedem Backup alle Daten übertragen (als tar-Paket), im anderen wird das Backup inkrementell ergänzt (rsync). Die Datenübertragung erfolgt selbstverständlich verschlüsselt.
Was passiert nun, wenn im Produktivsystem, also in Ludwigsburg, eine Datei beschädigt wird und diese dann auf die Backup-Server übertragen wird? Nach einer solchen Aktion existieren nur noch drei beschädigte Dateien. Die nicht beschädigten Versionen der Datei haben zwar zunächst noch auf den beiden Stuttgarter Servern überlebt, wurden dann aber im nächsten Backup überschrieben. In dem beschriebenen Fall ist die vielleicht größte Gefahr bei einer Speicherung digitaler Daten auf Festplatte beschrieben. Vor jedem Backup überprüft dimag daher zunächst die Integrität der Ludwigsburger Daten mittels Hashwerten. Zu jeder „normalen“ Datei (Primär- und Metadaten) wurde daher eine MD5-Datei angelegt, die den Hashwert dieser „normalen“ Datei enthält. Erst wenn das Licht auf grün gestellt wurde, also von allen „normalen“ Dateien neue Hashwerte gebildet und ohne Fehlermeldung mit den Ausgangs-Hashwerten verglichen wurden, startet der Backup-Prozess. Über Start, Abschluss der Integritätsprüfung und Ende des Backups unterrichtet je eine Mail die zuständigen Mitarbeiter. Als weitere Sicherheitsstufe finden die beiden Backupverfahren zu unterschiedlichen Zeiten statt.
Ein digitales Archiv hat unterschiedliche Aufgaben zu erfüllen. Zunächst sollten die Daten langfristig sicher abgelegt werden. Hierfür eignet sich ein Dateisystem eher als eine Datenbank. Auch wenn die dimag-Software komplett ausfallen sollte – die Daten können dennoch dem Dateisystem entnommen werden. Auf der anderen Seite gibt es legitime Bedürfnisse nach Recherche, die wiederum besser von einer Datenbank als von einem Dateisystem befriedigt werden. Im dimag finden sich beide Welten wieder. Zwar sind alle Meta- und Primärdaten im Dateisystem, ein Teil der Metadaten ist aber darüber hinaus auch in der Datenbank. Die Einheitlichkeit der Daten wird durch ein zentrales Eingabeformular gewahrt. In dieses Formular müssen alle Daten eingegeben werden, einen anderen Weg in das dimag kennt das Frontend des Systems nicht. Nach dem Absenden des Formulars werden die Daten gleichzeitig in das Dateisystem und die Datenbank eingetragen.

3.

Wie bereits erwähnt sollte ein derartiges System sowohl flexibel sein als auch seine Daten integer verwahren. Im dimag werden daher nicht alle Informationen eines Archivobjekts in eine Datei gepackt, sondern die Informationen in Dateien (für das Dateisystem) bzw. Datensätze (in der Datenbank) aufgespalten. Diese Aufteilung erspart zahlreiche Redundanzen. Vom Prinzip her ist das System so konstruiert, dass sowohl im Dateisystem als auch in der Datenbank jede Information nur einmal vorkommt. Der Archivar erkennt hier sofort den Einfluss von ISAD(G). Anders formuliert bedeutet dies, dass Angaben zum Bestand nicht in jedem Archivobjekt dieses Bestands wiederholt werden müssen. Ähnliches gilt auch für Codelisten und weitergehende Dokumentation. Auf jeden Fall muss bei Änderungen immer nur eine Datei respektive ein Datensatz bearbeitet werden.
Die digitalen Objekte werden in die Tektonik der einzelnen Staatsarchive eingefügt. Diese archivpolitische Grundsatzentscheidung bringt mehrere Vorteile mit sich.

- Zunächst muss kein eigenes, zweites Nachweissystem neben den bestehenden Beständegliederungen aufgebaut werden.
- Digitales Archivgut wird zusammen mit konventionellen Archivalien nachgewiesen. Damit werden auch die Übergänge zwischen beiden Welten, die allseits gefürchteten Hybridakten, beherrschbar.

[2] http://www.landesarchiv-bw.de > Aktuelles > Projekte.

- Drittens stellt die Beständegliederung ihrerseits wesentliche Informationen zur Einordnung des digitalen Objekts. Diese Kontextinformationen sind daher auch wesentliche Bestandteile der Archivierungs- und später auch der Benutzungsobjekte. Wie alle Metadaten können auch sie sich grundsätzlich ändern. Als Beispiel sei die Zuordnung zu einem anderen Bestand erwähnt. Die dann erforderliche Flexibilität wurde bereits erwähnt.

Die wichtigste Funktion von dimag ist es, das Landesarchiv auch bei der langfristigen Aufbewahrung digitaler Unterlagen handlungsfähig zu machen. Es genügt eben nicht, nur eine gesetzliche Zuständigkeit zu postulieren, die Archive müssen tatsächlich in der Lage sein, diese Objekte speichern zu können. Und seitdem wir diesen Nachweis erbringen können, bieten Behörden auch Dinge an, deren Existenz im Landesarchiv bis vor kurzem noch unbekannt war.

Sowohl unter dem Aspekt der eigenen Handlungsfähigkeit als auch hinsichtlich der Signalwirkung gegenüber den Behörden war es von zentraler Bedeutung, dass dimag alle vorstellbaren Unterlagentypen aufnehmen kann. Technik und Metadaten sind so ausgelegt, dass in dem System neben Fachverfahren auch einfache Bilder, neben elektronischen Akten auch die internen Mitteilungsblätter des Landeskriminalamts gespeichert werden können.

Digitale Unterlagen werden ebenso benannt und eingeordnet wie ihre konventionellen Brüder und Schwestern, sie stehen ebenfalls auf der Verzeichnungsstufe der „Archivalieneinheit“. Ihre Signatur ist zweiteilig, nach der Bestandssignatur wird die Bestellnummer eingeleitet durch ein DO, das für digitales Objekt stehen mag. Aufgrund dieser Kennung sind sämtliche digitalen Objekte schnell identifizierbar, es ist also nicht notwendig, digitale Bestände anzulegen. Wenn das Landesarchiv künftig die internen Mitteilungsblätter des Landeskriminalamts also in digitaler Form übernimmt, können sie demselben Bestand und derselben Gliederungsgruppe zugewiesen werden, die auch die papiernen Mitteilungsblätter schon enthalten.

4.

Zugänge werden grundsätzlich rasch und unverändert im dimag abgelegt. Auf diese Weise enthält das Archiv natürlich auch die Formate von MS-Word und MS-Excel, die nicht unbedingt als sehr archivierungsfreundlich gelten. Im Zuge der Aufbereitung werden sie dann in Formate überführt, die sich eher für die Archivierung eignen. Mit dem Verfahren verbinden sich mehrere Vorteile. Bei Fragen der Glaubwürdigkeit ist ein Hinweis auf das übergebene „Original“ möglich und für den Fall, dass sich die Archivversion als fehlerhaft herausstellen sollte, kann sie für eine gewisse Zeit erneut aus dem Ausgangsmaterial erstellt werden.

Im Ergebnis liegt daher ein Objekt in verschiedenen Erscheinungsformen oder Repräsentationen (s. u.) vor. Die Unterschiede betreffen sowohl den Umfang der Metadaten als auch die interne Struktur der Objekte. Natürlich werden die übergebenen Metadaten oft noch mit weiteren Informationen angereichert oder die unstrukturiert übernommenen Zusatzinformationen in die einzelnen Felder des Erfassungsformulars überführt. Im Umkehrschluss bedeutet dies, dass dimag auch solche Objekte akzeptiert, bei denen die Metadaten noch nicht in einer perfekten Form vorliegen.

Das älteste digitale Archivale im Landesarchiv ist die Volkszählung 1961[3]. Wir übernahmen vom Statistischen Landesamt eine Datei im Festbreitenformat, die 108 verschiedene Arten von Summenkarten, also aggregierte Auswertungen der Zählung enthält. Die Datei wurde zunächst als ein digitales Objekt abgelegt. Nach der Aufbereitung wurde das Objekt in 108 unterschiedliche Objekte aufgespalten, die zusammengenommen dieselbe Information enthalten wie das Ausgangsobjekt. Zu dem Zugangsobjekt wurden also entsprechend viele Unterobjekte gebildet. In anderen Worten: ein digitales Objekt kann weitere digitale Objekte enthalten.

[3] Kai Naumann, Älteste digitale Archivquelle der Bundesrepublik gesichert: Daten der Volkszählung von 1961 für das Land Baden-Württemberg übernommen und aufbereitet, in: *Der Archivar* 60 (2007), S. 53 f.

– Speicherung –

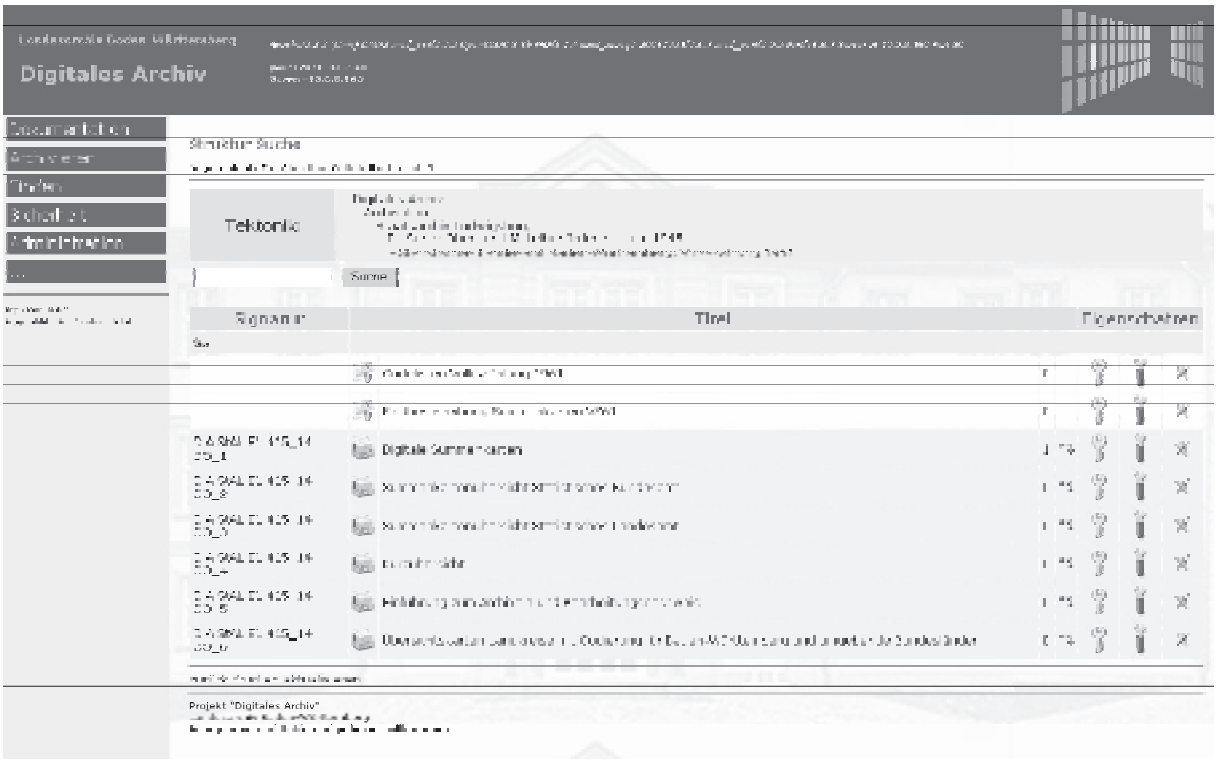

Abb. 1 Volkszählung 1961 im dimag

5.

Überhaupt birgt der Aufbereitungsprozess einige Tücken, die hier unter den Stichworten „Flexibilität“ und „Stabilität“ nur kurz gestreift werden können. Zunächst wünscht sich der Archivar natürlich ein möglichst flexibles System. Im Zuge der Aufbereitung muss es möglich sein, die Metadaten zu korrigieren, an der besten Formulierung zu arbeiten oder einen fehlerhaften Upload der Primärdaten zu eliminieren. Es erscheint nicht sinnvoll, diese Prozesse umfassend für die Nachwelt zu überliefern. Anders sieht es nach Abschluss der Aufbereitung aus. Nun sollen die Unterlagen sicher abgelegt sein. Gleichwohl ergeben sich auch dann noch Momente, in denen etwas nachgetragen oder korrigiert werden sollte. Diese Möglichkeit kollidiert mit der Forderung, unberechtigte Änderungen rückgängig machen zu können. Den soeben beschriebenen Anforderungen sollen zwei Mittel abhelfen. Zunächst haben alle Unterlagen den Status „In Bearbeitung“, d.h. der Archivar hat die von ihm benötigte Flexibilität. Im Anschluss wird der Status auf „abgeschlossen“ gestellt. Nach dem Umstellen des Status’ kann dieser nicht mehr auf „In Bearbeitung“ gestellt werden. Ab jetzt führt jede Änderung zu einer neuen Version. Ältere Versionen werden beibehalten. Die Flexibilität bleibt also grundsätzlich bestehen, unerlaubte Änderungen führen aber nicht zum Ausfall der korrekten Versionen.
Um künftig glaubwürdig auf das ius archivi verweisen zu können, ist es erforderlich, bereits von Anbeginn an die eigenen innerarchivischen Handlungen nachvollziehbar und transparent zu halten. Dieses Ziel soll durch drei Vorgaben erreicht werden. Zunächst soll das Digitale Magazin nicht nur die digitalen Archivalien, sondern auch die für das Archiv konstitutiven Dokumente enthalten. In einem eigenen Bereich werden daher Schemata oder Festlegungen auf Zeichen- und Dateiformate genauso wie die digitalen Archivalien abgelegt und archiviert. Die digitalen Objekte können, wie bereits gesagt, nicht ohne Metadaten abgelegt werden. Vielleicht kann auf diese Weise die Überlieferung von nicht mehr auflösbaren Zahlenkolonnen oder Textwüsten vermieden werden.
Dokumentiert werden soll schließlich auch die Archivierung selbst. Die dabei anfallenden Metadaten unterscheiden sich von den anderen Metadaten. Sie fallen in Datenbanksystemen normalerweise in mehreren kryptischen, kaum lesbaren Logdateien an. Sowohl die Aufsplitterung als auch die schlechte Lesbarkeit sollten vermieden werden. Aus diesem Grund besitzt jedes digitale Objekt genau eine Protokolldatei, die nur fünf Felder und das Format XML umfasst, folglich bewusst sehr einfach gehalten wurde. In diese Datei werden nicht alle, sondern nur die wesentlichen Prozesse geschrieben. Beispielsweise werden die Backups nicht vermerkt.

Die meisten der für ein Archivale wesentlichen Prozesse werden automatisch in diese Protokolldatei geschrieben. In manchen Fällen ist es möglich, diesen Einträgen eine händische Ergänzung hinzuzufügen. Ausschließlich händische Einträge sind ebenfalls möglich (z. B. eine Notiz). Es ist in keinem Fall möglich, einmal angefertigte Einträge zu verändern. Dies gilt auch für die Fälle, in denen ein automatischer Eintrag händisch ergänzt wurde. Zu jedem Protokoll wird ein Hashwert ermittelt, der unmittelbar nach der Erweiterung des Protokolls neu berechnet wird. Protokolldateien fallen daher ebenso wie alle anderen Dateien unter die regelmäßigen Integritätsprüfungen. Alle Prozesse oberhalb der digitalen Objekte werden in einem Archivprotokoll festgehalten.

6.

Der Aufbau des Archivs kann anhand von zwei Bildern erklärt werden. Zunächst besteht das Dateisystem natürlich nur aus einer mehr oder weniger großen Menge an Dateien. Diese Dateien kann man bildlich als Schachteln oder vornehmer als Container verstehen. Die Schachteln verpacken die Primärdaten. Da sie ausnahmslos in XML angelegt sind, müssen sie wenigstens kurz- und mittelfristig nicht migriert werden. In der letzten Schachtel stecken schließlich die Primärdaten und weitere Metadaten, die dann in jedem erdenklichen Format vorliegen können und vermutlich wesentlich schneller als die Schachtel-XMLs migriert werden müssen. Diese logische Sicht kann auch durch eine eher physische ergänzt werden. Sie macht den Blick auf das Skelett, also die tragenden Einheiten frei.

Die Skelettsicht bildet die Struktur des Dateisystems im dimag ab. Über allem schwebt die hier nicht wiedergegebene Beständegliederung der Archive. Vom Archiv kommen wir zu den Beständen, schließlich zu einzelnen Gliederungspunkten, unterhalb denen dann die einzelnen Archivalieneinheiten angesiedelt sind. Auf dieser Ebene finden sich sowohl die digitalen Objekte als auch die Papierakten und die Hybridobjekte. Die digitalen Objekte zerfallen in eine oder mehrere Repräsentationen. Jede dieser von PREMIS[4] inspirierten Erscheinungsformen enthält dieselbe Information, die Repräsentation 1 ist grundsätzlich der Zugang. Eine Repräsentation kann schließlich eine beliebige Zahl an Dateien enthalten, die je nach Status in unterschiedlichen Versionen abgespeichert werden. In Abb. 2 werden die einzelnen Kästchen innerhalb eines digitalen Objekts durch jeweils eine eigenständige XML-Datei dargestellt.

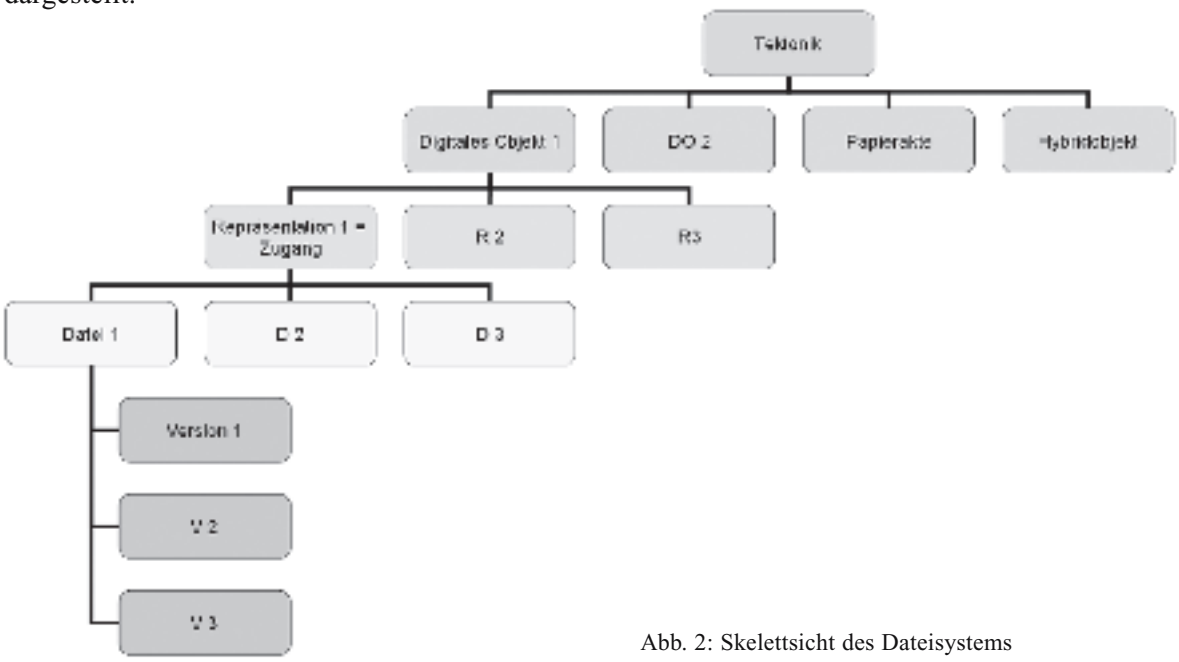

Abb. 2: Skelettsicht des Dateisystems

[4] http://www.oclc.org/research/projects/pmwg/.

Abschließend soll hier noch ein Beispiel für die Schachtelsicht gegeben werden, in die zusätzlich auch die Protokoll- und MD5-Dateien aufgenommen wurden, die also in anderen Worten alle Dateien des digitalen Objekts geordnet wiedergibt.

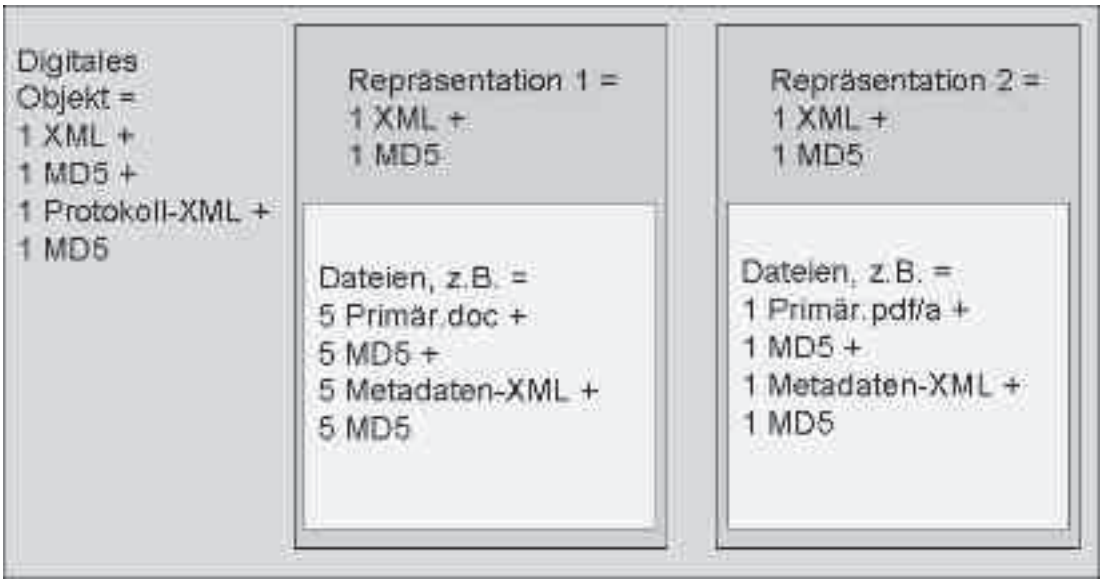

Abb. 3 Erweiterte Schachtelsicht

Die Abbildung zeigt ein digitales Objekt mit zwei Repräsentationen. Das Objekt wird durch eine XML-Datei beschrieben, die auf es bezogenen Prozesse durch eine Protokolldatei. Zu beiden Dateien existiert – ebenso wie zu allen anderen Nicht-MD5-Dateien im Archiv – je eine MD5-Datei. Die Repräsentationen 1 und 2 sind ihrerseits durch eine XML-Datei beschrieben und durch eine MD5-Datei in ihrer Integrität gesichert. Die erste Repräsentation enthält den Zugang, d. h. fünf Primärdaten-Dateien, die jeweils durch eine Metadaten-XML beschrieben werden und 10 weitere MD5-Dateien. Die zweite Repräsentation enthält das aufbereitete Material in einer PDF/A-Datei, wir haben daher nur eine beschreibende Metadaten-XML und nur zwei MD5-Dateien.
Jede Schachtel enthält Informationen, die auf alle in diesem Rechteck enthaltenen Informationseinheiten zutreffen. Die Angaben für das digitale Objekt werden also beispielsweise nicht mehr in den Metadaten zu jeder Repräsentation wiederholt.

7.

Die vorgestellten Konzepte und Umsetzungen sollten zum Abschluss mit einem großen Ausrufezeichen versehen werden: Entstehungszeit ist das Jahr 2006. Vielleicht sind sie auch noch in fünf oder zehn Jahren gültig. Denkbar ist aber auch eine stark abweichende Ausgestaltung digitaler Archivierung. Derzeit scheinen die aufgebauten Strukturen tragfähig zu sein. Dimag läuft bereits seit Mitte 2006 im Echtbetrieb. Gegen Ende des Jahres archivierte es 13.900 digitale Objekte oder 24,6 Mio. Datensätze[5] auf 11,5 GB.

[5] Es wurde pro digitalem Objekt jeweils nur eine Repräsentation gezählt.