Langzeitarchivierung im Digitalen Monumentalbau-Archiv MonArch

Alfons Ruch

Lehrstuhl für Informationsmanagement, Universität Passau

Alfons.Ruch@uni-passau.de

Technischer Bericht, Nummer MIP-0912
Fakultät für Informatik und Mathematik
Universität Passau, Deutschland
August 2009

Inhaltsverzeichnis

1 Übersicht

In diesem Dokument wird ein erster Ansatz zur Langzeitarchivierung im Kontext des Forschungsprojekts MonArch[1] [Mon] definiert.
Es werden die empfohlenen Dateiformate für das DMA, die Sicherung der Informationen im Archiv-Verbund durch Replikation und ein Weiterentwicklungskonzept für das DMA gemäß Referenzmodell „Open Archival Information System“ (OAIS ISO-Standard 14721:2003) beschrieben. Das DMA soll im weiteren Verlauf der Projektarbeiten so erweitert werden, dass die Nutzerinnen und Nutzer aktiv bei der Langzeitarchivierung der Daten unterstützt werden. Dies umfasst unter anderem die Überprüfung der im Archiv verwalteten Dateiformate und deren Dokumentation.

2 Backupkonzept

Im folgenden wird das Backupkonzept für das Digitale Monumentalbau-Archiv (DMA) vorgestellt, für eine ausführliche Beschreibung wird auf [RSF09c] verwiesen.
Bei der Sicherung werden zwei Kategorien des Datenbestandes unterschieden: der physische Datenbestand und der logische Datenbestand. Unter dem physischen Datenbestand versteht man alle auf einem Datenträger gespeicherten Nutzdaten. Hierbei sind die Dokumente, Navigationskarten, Kartierungen und die in der Datenbank gespeicherten Metadaten zu nennen. Unter den logischen Datenbestand fallen alle Daten, die notwendig sind, um die Nutzdaten des Archivs technisch zu interpretieren. Hierunter fallen die Programmversion des DMA, das Schema der Datenbank und der Quellcode der Archivsoftware.
Die Sicherungsstrategie des DMA sichert sowohl den physischen als auch den logischen Datenbestand. Die Sicherung des physischen Datenbestandes stellt die Wiederherstellung der in das Archiv ein gepflegten Nutzdaten sicher. Die Sicherung des logischen Datenbestandes stellt die Wiederherstellung der DMA-Installation sicher. Für die Sicherung des physischen Datenbestandes setzt das Archivsystem eine differenzielle Sicherung ein. Um eine Sicherung des physischen Datenbestandes zu interpretieren, ist die korrespondierende Sicherung des logischen Datenbestandes notwendig.
Hierbei ist zu beachten, dass diese Sicherungsstrategie eine reine Sicherung des operativen Datenbestandes darstellt. Dies löst aber nicht die Herausforderungen wie Formatalterung oder Hardwarealterung. In den folgenden Abschnitten werden die notwendigen Schritte für die Langzeitspeicherung und Langzeitarchivierung des DMA Datenbestandes dargestellt.

[1]Dieses Projekt ist gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter dem Aktenzeichen FR 1012/8-1.

3 Dateiformate der gespeicherten Dateien

In diesem Abschnitt werden die empfohlenen Dateiformate für das Digitale Monumentalbau-Archiv beschrieben. Dokumente werden in dem jeweiligen aktuellen Standardformat archiviert, um den zukünftigen Zugriff auf diese Informationen zu sichern. Das Archivsystem ist technisch nicht an diese Formate gebunden, sondern kann zu jeder Zeit um andere Formate ergänzt und erweitert werden.

3.1 Dateiformate für Textdokumente

Bei Textdokumenten wird als Standardformat [PDF/A] (ISO 19005-1:2005) empfohlen. PDF/A hat sich in der Industrie als Standard für die Langzeitarchivierung von Textdokumenten durchgesetzt. In Abschnitt 4.1 wird das Emulations-/Migrationsvorgehen für Textdokumente vorgestellt, den Inhalt der Textdokumente und die Informationen über deren Struktur zu sichern.

3.2 Dateiformate für Bilder

Das Digitale Monumentalbau-Archiv folgt den Empfehlungen der Deutschen Forschungsgemeinschaft (DFG) aus [For09b] und schreibt für Bilder das Format TIFF [TIFF] verpflichtend vor. Der Ablauf der Speicherung ist in Abbildung 1

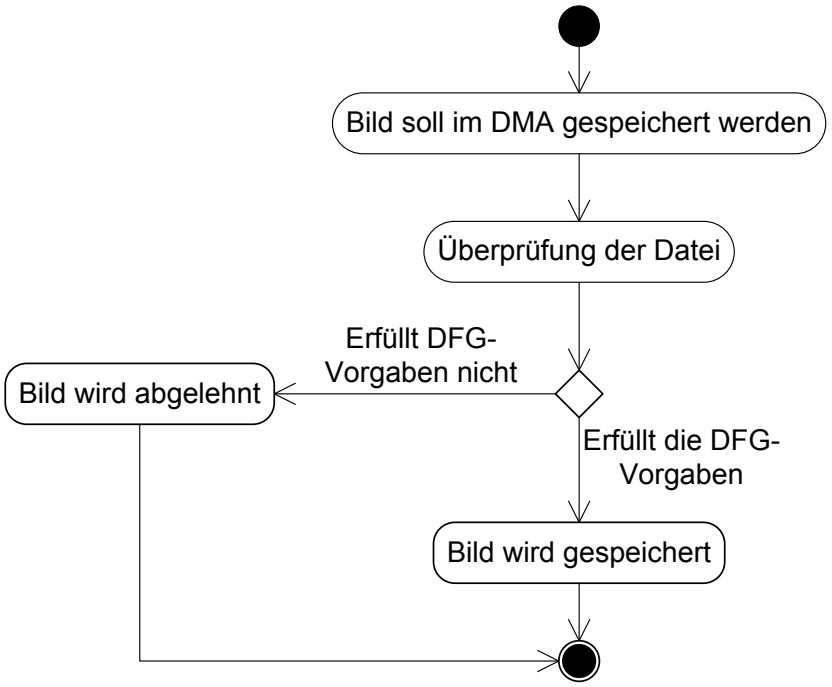

Abbildung 1: Speicherung von Bildern

dargestellt. Das seit den 1980er Jahren bekannte TIFF-Format hat sich als einer der wichtigsten Standards etabliert.
Die DFG empfiehlt für die Digitalisierung folgende Eigenschaften [For09b]:

- Format:
  - Graustufen- oder Farbbilder: TIFF (uncompressed)
  - bitonale Bilder: TIFF Group 4 Kompression
- Auflösung:
  - Graustufen- oder Farbbilder: 300 dpi
  - Handschriften oder Kartenwerken: 400 dpi
  - bitonale Bilder: 600 dpi
- Farbtiefe: mindestens 24 bit

Aus den eingefügten Bildern können verschiedene Derivate (JPEG, GIF usw.) erstellt werden, abhängig von der gewählten Präsentation und geplanten Verwendung (siehe Abbildung 2). Die notwendigen Schritte im Erstellungsprozess vom *Master* zum *Derivat* werden mit einer XML-Datei definiert und dokumentiert. Falls die Bilder in einer verlustfreien Kompression und im TIFF Format vorliegen, kann man nach heutigem Stand [For09b] mit hoher Sicherheit die Bilder zu einem späteren Zeitpunkt wiederverwenden. Insbesondere erlaubt eine verlustfreie Kompression eine Migration in ein anderes Bildformat ohne Qualitätsverlust, falls dies einmal notwendig werden sollte. Um dieses Vorgehen zusätzlich abzusichern, wird im DMA die Original-Definition der Dateiformate archiviert, wie sie von ihren Herstellern dokumentiert wurde.
Im DMA wird in Anlehnung an [For09b] als Arbeitsformat der Bilder das Format JPEG (ISO/IEC 10918-1) [JPEG] empfohlen. JPEG wird von der *Joint Photographic Experts Group* entwickelt und gepflegt. Die JPEG Version eines Bildes wird aus dem TIFF Master erzeugt (siehe Abbildung 2).
Als weiteres mögliches Primärdatenformat für die Archivierung stehen die Rohdaten (RAW) der Bilder zur Diskussion. Aktuell gibt es hierfür keinen einheitlichen Standard. So hat das von Adobe definierte Format *Digital Negative (DNG)* [DNG] das höchste Potential, in Zukunft ein de-facto Standard zu werden, und sollte deswegen weiter beobachtet werden. Die Erfahrung aus der Praxis zeigt aber, dass in den meisten Fällen keine Rohdaten zur Verfügung stehen, was derzeit gegen eine Empfehlung von Rohdaten als Hauptstandard für Bilder im Archiv spricht.

3.3 Dateiformate für Kartierungen

Für die Verwaltung von Kartierungen wird das Format DWG/DXF [DXF] verwendet. DWG und DXF wurden von AutoDesk für AutoCAD entwickelt und haben sich als de-facto Standard für Vektorgraphiken in der Industrie durchgesetzt. Ein Vorteil bei der Verwendung von DXF im Archiv ist die Speicherung der

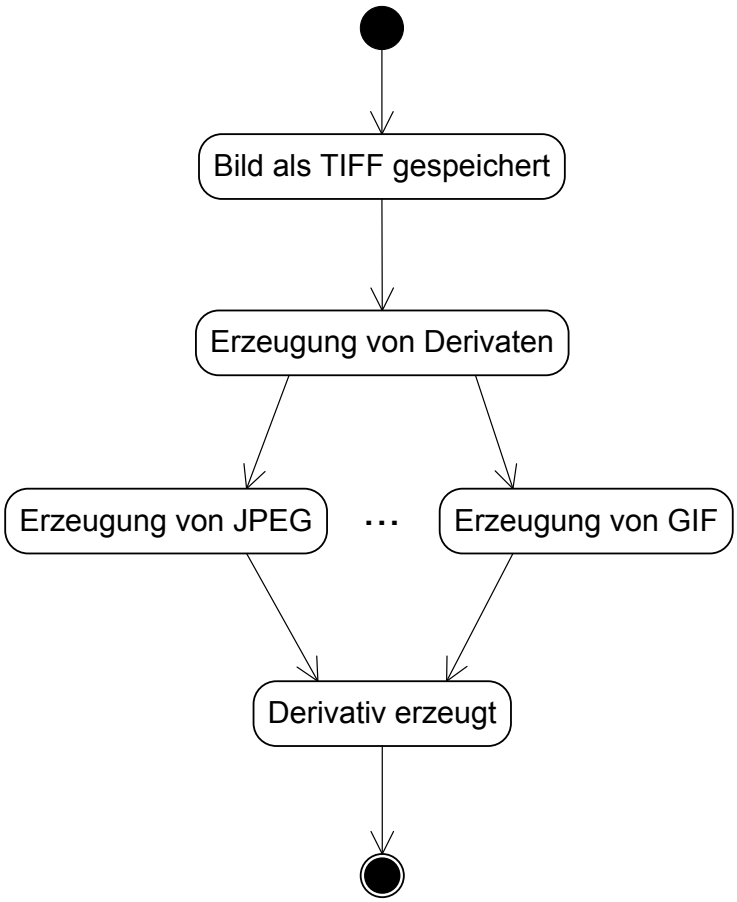

Abbildung 2: Erzeugung von Derivativen

Karten-Informationen im ASCII Format. Dies erlaubt den Zugriff auf die Informationen ohne ein CAD-Programm. Sollte daher zu einem späteren Zeitpunkt kein passendes CAD-Programm zur Verfügung stehen, kann man trotzdem auf alle Informationen der DXF-Kartierung zugreifen.

4 Kombinierter Emulations-/Migrationsansatz

In diesem Abschnitt wird der kombinierte Emulations-/Migrationsansatz für die Speicherung der Digitalen Objekte und der Formate beschrieben.

4.1 Speicherung von Textdokumenten

Bei der Speicherung von Textdokumenten werden die in der Datei enthaltenen Zeichen redundant als XML-Datei in einer UTF-8 Codierung gespeichert. Für diese strukturierten Metadaten wird die Liste von Bezeichnungen des DFG-Viewers[2] als Vorlage genommen. Dieses Vorgehen sichert den Inhalt der Datei in einem allgemeingültigen Format und erlaubt, sollte das ursprüngliche Erstellungsprogramm nicht mehr existieren, einen Zugriff auf die Informationen mit einfachsten Mitteln.

4.2 Archivierung der Formatdefinitionen

Die Definition der Dateiformate wird selbst im Archivsystem gespeichert. Dabei wird auf die Formatbeschreibung der Hersteller zurückgegriffen. Bei der Formatbeschreibung ist aber nicht nur die Information über das Erscheinungsbild interessant, sondern auch die Information, wie man auf die Inhalte eines bestimmten Formats zugreifen kann. Daher sollen die notwendigen Informationen über Syntax und Aufbau der Dateien im DMA ebenfalls gesichert werden. Dies erlaubt bei einem etwaigen Reverse Engineering die Entwicklung eines Parsers, falls dies zu einem späteren Zeitpunkt notwendig sein sollte. Dieses Vorgehen unterstützt die nachträgliche Entwicklung einer Anwendung, die das Verhalten des ursprünglichen Programms emuliert und die Daten interpretieren kann.

5 Dokumentation des Datenmodells

Das im Digitalen Monumentalbau-Archiv verwendete Datenmodell wird von der Entwicklung bis hin zum Betrieb des DMA dokumentiert. Für die Dokumentation werden standardisierte Modelle verwendet.

5.1 Dokumentation des Datenbankmodells

Das Datenmodell der zugrunde liegenden Datenbank wird mit Hilfe der zu seiner Definition verwendeten Datenbankbefehle in der ISO-zertizierten Sprache SQL [Tür03, SQL03] dokumentiert. Die SQL-Befehle werden als SQL-Skript in einer Textdatei gespeichert, die im Archivsystem verwaltet wird und somit einen Teil der Selbst-Dokumentation des Digitalen Monumentalbau-Archiv bildet.

5.2 Beschreibung der Metadaten durch Ontologien

Die im Archiv vorkommenden Metadaten werden mit den Ontologie Beschreibungssprachen: RDF (Resource Description Framework) [RDF] und OWL (Web Ontology Language) [OWL] definiert. Diese W3C Standards stellen mit ihrem selbstbeschreibenden und menschenlesbaren Charakter sicher, dass die semantischen Informationen des DMA wohl definiert und dokumentiert sind. Eine ausführlich Beschreibung des Datenmodells ist in [RSF09b] zu finden.

[2] http://www.dfg-viewer.de

5.3 Exportfunktion der Metadaten

Die im Digitalen Monumentalbau-Archiv vorhandenen Metadaten werden in verschiedenen Formaten exportiert. Im aktuellen Entwicklungsstand werden die Metadaten, wie in Abschnitt 4 beschrieben, im OWL Format über einen Webservice zur Verfügung gestellt. Es ist geplant, den Webservice in dem Sinne weiter auszubauen, dass die Metadaten im METS-Format [Fed07] und im MuseumDAT Format [museumDAT] exportiert werden. Der METS Export erfüllt insbesondere eine Vorgabe der DFG [For09b]. Das MuseumDAT Format ist zu dem CIDOC RM [CDG+06] konform und stellt das Austauschformat mit anderen Bildarchiven, insbesondere dem Marburger Index [Bra07] dar.

6 Replikation der Informationen

MonArch sieht eine Peer-to-Peer-Vernetzung (P2P) von Archiv- und Rechercheservern vor. Die in dem Digitalen Monumentalbau-Archiv verwalteten Informationen werden in dem Peer-to-Peer Netzwerk repliziert. Replizierbar sind u.a. die Gebäudestruktur, die Themen (beschrieben durch Ontologien) und die eigentlichen Dokumente.
In günstigen Fällen erlaubt die Replikation die vollständige Wiederherstellung der Informationen eines bestimmten Archivs in einem Archiv-Verbund durch die Kombination der Informationen der beteiligten Peers. Replikationsstrategien für eine Peer-to-Peer Umgebung wurden zum Beispiel in Systemen wie LOCKSS [MRG+05] erprobt. LOCKSS verwendet für die Kommunikation und die Abstimmung der Informationen im P2P-Netz das so genannte *Option Poll Protocol*. Dieses Protokoll erlaubt die Wiederherstellung einer lokal bei einem Clienten verloren gegangenen Information durch ein Abstimmungsverfahren mit den anderen Clienten des Netzwerks. Das Abstimmungsverfahren wird in Abbildung 3 gezeigt.
Die Deutsche Forschungsgemeinschaft empfiehlt, LOCKSS zur Unterstützung der Langzeitarchivierung einzusetzen [For09a, SG08]. Das MonArch-Archiv folgt mit seiner Peer-to-Peer Struktur und der darauf beruhenden Replikationsstrategie dieser Empfehlung. Eine Untersuchung, welches Replikationsprotokoll sich

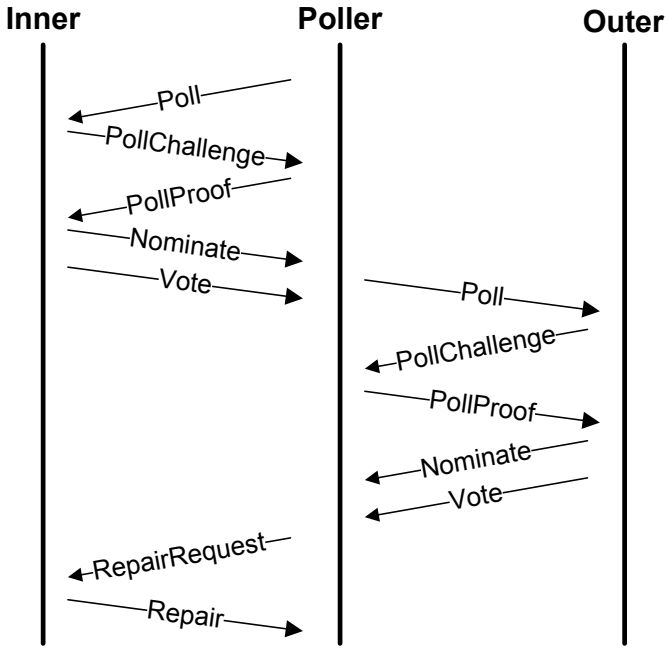

Abbildung 3: P2P-Protokoll eines LOCKSS Netzwerks

für den MonArch-Verbund konkret anbietet, ist Gegenstand künftiger Arbeiten.

7 Überprüfung der Workflows

Gemäß dem Referenzmodell des Open Archival Information System (OAIS: ISO 14721:2003) [OAIS, Lav04] werden im Digitalen Monumentalbau-Archiv die Informations-Lebenszyklen und die zugehörigen Workflows überprüft.

7.1 Referenzmodell des OAIS

Im folgenden Abschnitt wird das Referenzmodell des Open Archival Information System (OAIS) vorgestellt.

7.1.1 Information Modell

Im OAIS ist das logische Modell für die archivierten Informationen aus verschiedenen Konzepten aufgebaut. Das fundamentale Konzept besteht darin, dass *Information Objects* aus der Kombination von *Data Objects* und *Representation*

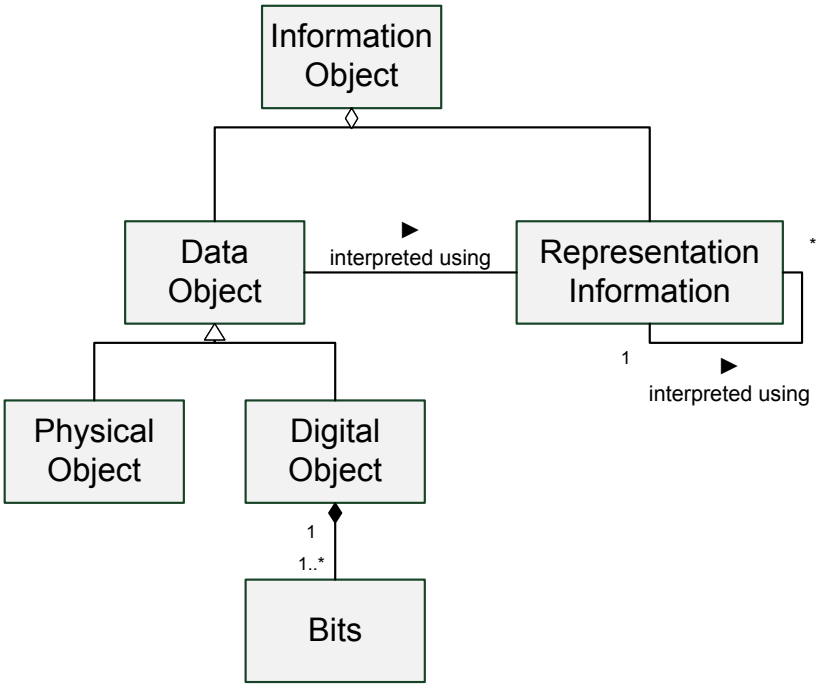

Abbildung 4: OAIS Information Object

Information bestehen. Dies wird in dem UML Diagramm in Abbildung 4 verdeutlicht [OAIS Standard]. Ein *Data Object* kann ein physikalisches oder digitales Objekt sein, wobei das *Digital Object* eindeutig aus den zugeordneten *Bit*s besteht. Die *Representation Information* erlaubt die Interpretation der Daten zu aussagekräftigen Informationen. Hierbei kann *Representation Information* ihrerseits unter dem Einsatz anderer *Representation Information* interpretiert werden. Zum Beispiel kann eine *Representation Information* eine Bit-Sequenz als Zahlenwert mit seinem Datentyp, wie Integer oder Float, beschreiben, während eine weitere *Representation Information* diesem Datentyp die Semantik einer Tabelle von Messdaten zukommen lässt.
Die nächste konzeptionelle Struktur ist das *Information Package*, dargestellt in Abbildung 5 [OAIS Standard]. Das *Information Package* besteht aus zwei Typen von *Information Objects*, der *Content Information* und der *Preservation Description Information (PDI)*. Die *Content Information* ist die Menge an Informationen, welche vom OAIS bewahrt werden soll. Hierbei kann es sich, wie in Abschnitt 7.1.1 beschrieben, um physische und digitale *Data Objects* handeln. Die *Preservation Description Information* enthält Informationen, welche für das Verständnis der *Content Information* benötigt wird.
Das *Information Package* kann mit zwei Typen von *Information Objects* assoziiert werden, der *Packaging Information* und der *Package Description*. Die *Packaging Information* enthält die Information darüber, wie die Komponenten des Pakets auf identifizierbare Einheiten auf spezifischen Medien abgebildet werden können. Die *Package Description* beschreibt das *Content Information*

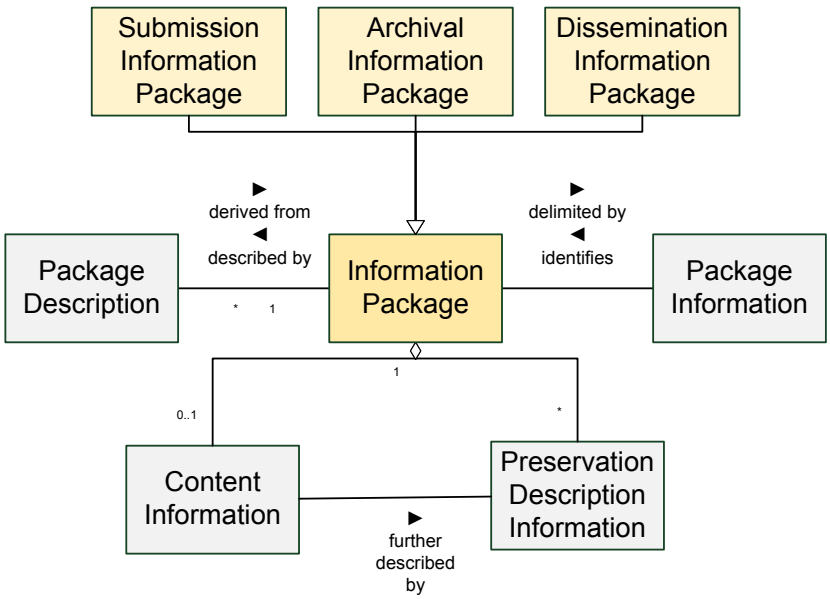

Abbildung 5: OAIS Information Package

Object um einen effizienten Zugriff zu ermöglichen.
Das *Information Package* lässt sich in *Submission Information Package*, *Archival Information Package* und *Dissemination Information Package* unterteilen, entsprechend der Funktion des Archivierungsprozesses in dem das *Information Package* eingesetzt wird. Die hierbei definierten Interaktionen werden im nächsten Abschnitt vorgestellt.

7.1.2 Interaktionen

Die Interaktionen der OAIS-Umgebung mit ihren Informations-Lieferanten (*Producer*) und -Nutzern (*Consumer*) ist in Abbildung 6 dargestellt [OAIS Standard]. Der *Producer* übergibt SIPs *(Submission Information Package)* an das OAIS, das OAIS verwaltet die Informationen als AIPs (*Archival Information Package*) und übermittelt sie als DIPs (*Dissemination Information Package*) an die *Consumer*.

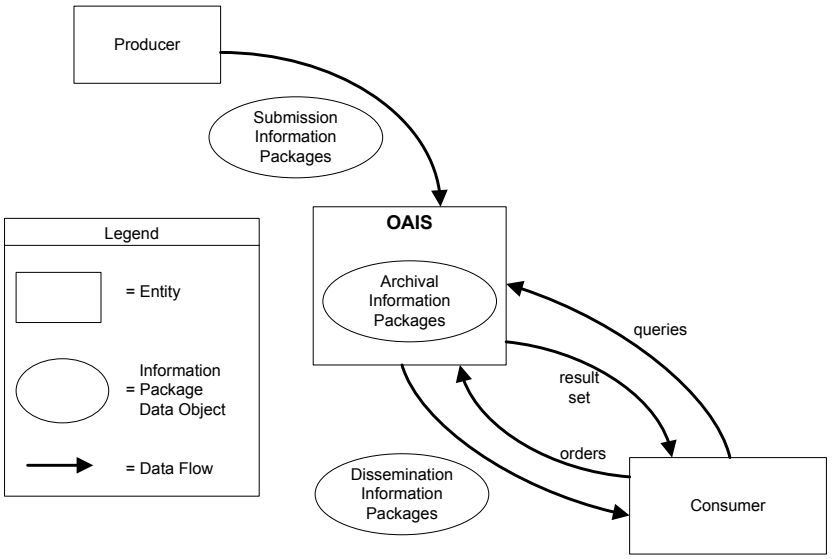

Abbildung 6: OAIS Interactions

7.1.3 Funktionale Modell

Das funktionale Modell des OAIS ist in Abbildung 7 gemäß [OAIS Standard] zu sehen, wo auch die wesentlichen Informationsflüsse angegeben sind.
Die einzelnen Einheiten haben folgende Funktionalitäten (diese Ausführungen basieren auf [SK08]):

Ingest umfasst Funktionen und Dienste zur Übernahme der SIPs von den Produzenten (oder internen Einheiten unter Kontrolle der *Administration*)

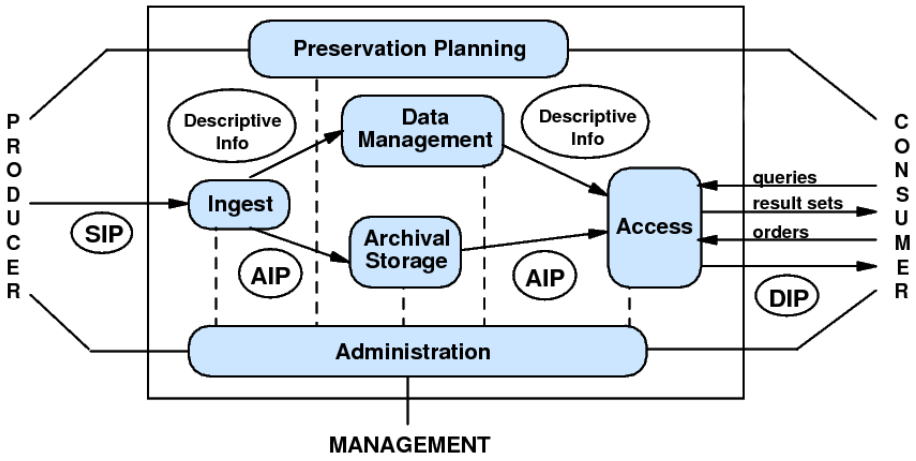

Abbildung 7: OAIS Functional Entities

und zur Vorbereitung der Inhalte für ihre Speicherung und Verwaltung im Archiv. Die Funktionen von *Ingest* sind: SIPs übernehmen; Qualität der SIPs überprüfen; den internen Formaten und Normen entsprechende AIPs generieren; Auszeichnungsmetadaten (*Descriptive Information*) aus den AIPs extrahieren und in die Archivdatenbank (siehe *Data Management*) einfügen; Updates von *Archival Storage* und *Data Management* koordinieren.

Archival Storage bündelt Funktionen und Dienste zur eigentlichen Speicherung und Bereitstellung der AIPs. Die Funktionen von *Archival Storage* sind: AIPs von *Ingest* entgegennehmen und in den Speicher einfügen; Speicher verwalten; Speichermedien aktualisieren; Fehlersuche und -korrektur durchführen; Wiederherstellung von Daten ermöglichen; AIPs für die *Access* Komponente bereitstellen.

Data Management bietet Funktionen und Dienste zur internen Verwaltung und Nutzung von *Descriptive Information* über die im Archiv gespeicherten digitalen Informationsobjekte. Metadaten-Anfragen an die interne Datenbank werden entgegengenommen, bearbeitet und in geeigneter Form weitergeleitet. Daneben werden Verwaltungsinformationen für die Administration gehalten und Aufgaben der Datenbankadministration übernommen.

Die Administration ist für die Steuerung und Überwachung aller Abläufe in der OAIS-Umgebung verantwortlich. Es werden die Beziehungen und der Datenaustausch mit der Außenwelt von der *Administration* koordiniert. Außerdem ist die *Administration* für die Konfiguration von Hard- und Software sowie die Vergabe von Zugriffsrechten zuständig. Zu den Funktionen der Administration gehören: Aushandeln von Submission Agreements mit

Informations-Lieferanten (producer); Überprüfung, ob die gelieferten Informationen den Standards des Archivs entsprechen; Durchführung der erforderlichen Datenmigrationen und -aktualisierungen; Definition von Standards, Richtlinien und Kundenservices.

Preservation Planning ist für die Beobachtung der Technologieentwicklung und die Definition von Strategien zur Reaktion auf Technologieänderungen zuständig. Es werden Empfehlungen zu den internen Standards und Richtlinien entwickelt, Maßnahmen zur Erhaltung der Brauchbarkeit der gespeicherten Information definiert und ggf. notwendige Datenmigrationen und Kopiervorgänge geplant.

Access stellt eine Schnittstelle für den Zugriff von Informationsnutzern auf das Archiv bereit. Es werden Anfragen und Bestellungen (dissemination requests) entgegen genommen und prozessiert. Für die gefundenen Informationen werden DIPs generiert, Antwortmengen erzeugt und ggf. Reportinformation bereit gestellt.

Common Services ist eine zusätzliche Einheit, die in Abbildung 7 nicht dargestellt ist, da sie über das gesamte Archiv verteilt ist. Hierunter fallen Betriebssystemdienste, Netzwerkdienste und Sicherheitsdienste (wie Räume mit Zugangskontrolle usw.).

Für eine vollständige und detaillierte Beschreibung wird auf die Definition des

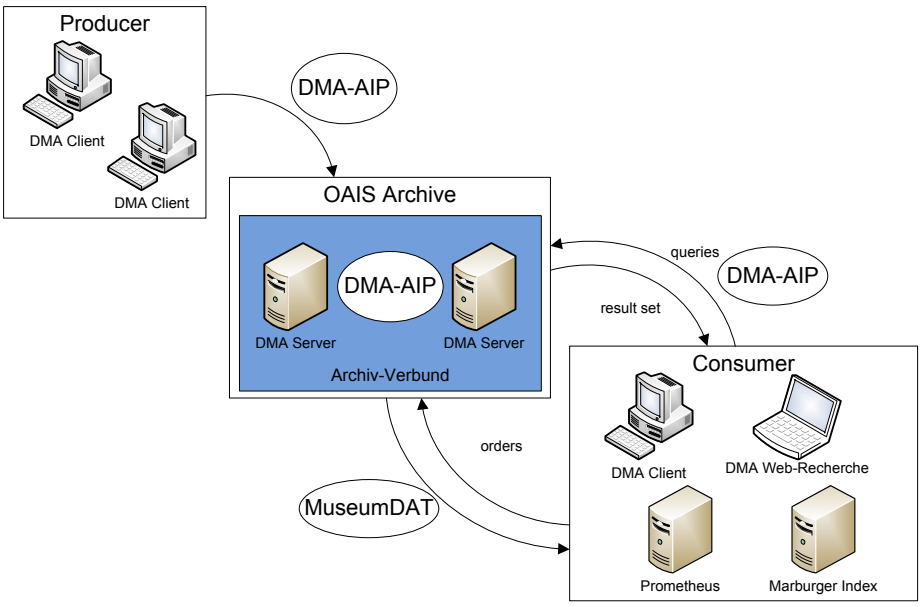

Abbildung 8: DMA Architektur OAIS

OAIS [OAIS Standard] verwiesen. Im Folgenden wird die Interpretation des OAIS für das Digitale Monumentalbau-Archiv beschrieben.
Die Komponenten der DMA Architektur nehmen die folgenden Rollen des OAIS Modells ein (siehe Abbildung 8): Die *DMA-Clients* haben eine Doppel-Rolle als *Producer* und *Consumer* inne, der *DMA-Server* bzw. der von diesen gebildete *Archiv-Verbund* die Rolle des *OAIS Archives* und die *DMA Web-Recherche* die Rolle eines *Consumers*. Zusätzlich nehmen die externen Archive wie *Prometheus* [Nem01] und *Marburger Index* [Bra07] die Rolle weiterer *Consumer* ein. Die ausgetauschten *Information Packages* werden in den nächsten Abschnitten beschrieben.

7.2 Einfügen von Informationen (Ingest)

Das DMA ermöglicht die Auszeichnung von digitalen Objekten mit Metadaten, die in Taxonomien (Ontologien) organisiert sind. Diese Informationen definieren zusammen die DMA-AIPs, die aus dem digitalen Objekt und den Metadaten bestehen (siehe Abbildung 9). Die Metadaten werden in einfache und strukturierte Metadaten unterteilen. Zu den einfachen Metadaten zählen Angaben wie Autor, Kommentare, Einfügedatum usw., zu den strukturierten Metadaten gehört die Auszeichnung mit der Partonomie oder der Taxonomie. Der DMA-Client berei-

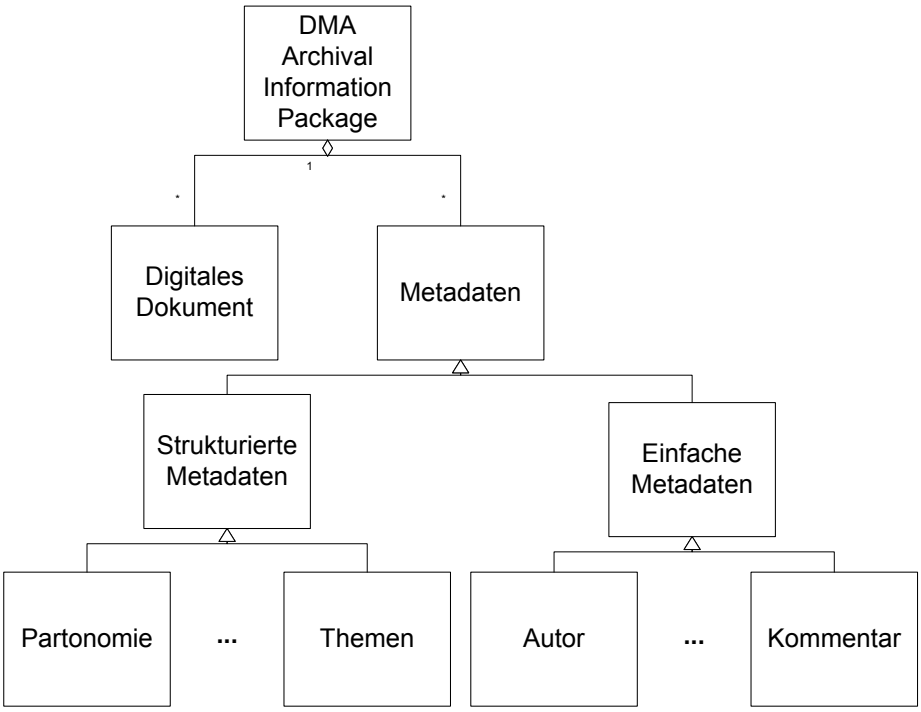

Abbildung 9: DMA Archival Information Package

tet seine *Submission Information Packages* für das *Ingest* so vor, dass die SIPs dem Aufbau des *DMA Archival Information Package* entsprechen. Beim *Ingest* werden die Informationen auf ihre Korrektheit und Vollständigkeit geprüft. Im Detail werden Dokumente auf das richtige Dateiformat und die angegebenen Metadaten auf Konsistenz untersucht. So wird beispielsweise verhindert, dass ein Bild als BMP-Datei gespeichert wird, obwohl eine TIFF Datei als Standard definiert ist. Die Auszeichnung wird durch logische Inferenz auf den Ontologien auf ihre Konsistenz geprüft. Wenn notwendig, wird der Benutzer vor Inkonsistenzen gewarnt.
Bei der Erzeugung der DMA-AIPs werden zusätzliche *Descriptive Information* extrahiert. Zum Beispiel werden Thumbnails aus den Bilddateien erzeugt, die Thumbnails werden dem Benutzer als Vorschau der eigentlichen Datei angezeigt. Eine Besonderheit besteht in der Kombination aus dem DMA und dem Kartierungsformat des MMSArchive, da hier die Glossardefinitionen und Themenauszeichnungen aus den Kartierungen extrahiert werden. Diese Auszeichnungen werden dann in den DMA Glossar- und Themensatz integriert und somit im Archiv-Verbund standardisiert.

7.3 Speicherung im Archiv (Archival Storage)

Zur Verarbeitung der DMA-AIPs ist der DMA-Server in einer Schichtenarchitektur aufgebaut (siehe Abbildung 10).
So werden nach Annahme der DMA-AIPs über die *Transaktionsmanagement-Schicht*, die erhaltenen DMA-AIPs in der *Analyse-Schicht* auf potentielle Übertragungsfehler überprüft: Korrektheit der Datentypen, Konsistenz mit dem Spei-

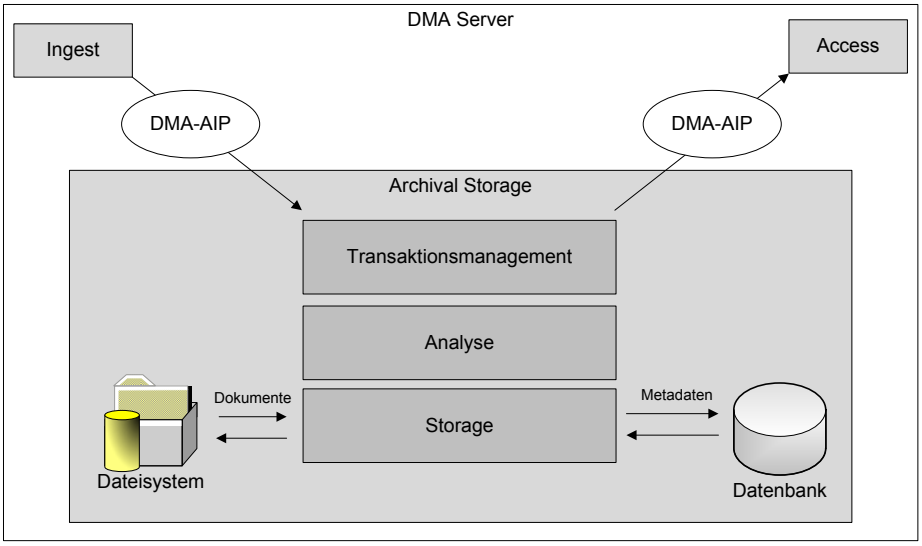

Abbildung 10: DMA Schichten Architektur

chermodell. Wenn möglich, werden die Fehler korrigiert, anderenfalls werden die DMA-AIPs erneut von *Ingest* angefordert. Über die *Storage-Schicht* werden die DMA-AIPs in die Datenspeicher abgelegt. Die Metadaten werden in der Datenbank, die Dokumente auf dem Dateisystem gespeichert, wobei dieser Teil des Dateisystems unter der Verwaltung des DMA-Servers steht. Zukünftig sollen zusätzlich die Inhalte von Textdokumenten ausgelesen und als XML-Datei im Dateisystem abgelegt werden, wie in Abschnitt 4.1 beschrieben.
Die Übertragungsform der AIPs im DMA ist durchgängig mit RDF [RDF] umgesetzt und wird in diesem Format der *Access*-Funktion zur Verfügung gestellt.
Die im Digitalen Monumentalbau-Archiv gespeicherten Daten werden auf verschiedene Art und Weise gesichert. So wird zum einen der lokale Datenbestand mit regelmäßigen Backups abgesichert (siehe Abschnitt 2). Zum anderen wird der lokale Datenbestand über Replikation im Archiv-Verbund verteilt und redundant verwaltet. Die Replikation wird in Abschnitt 6 dargestellt. Die Wiederherstellung der lokalen Daten kann durch die regelmäßigen Backups (Abschnitt 2) oder durch die replizierten Daten (Abschnitt 6) erfolgen. Eine weitere Empfehlung ist die Speicherung und Archivierung von Informationen auf Mikrofilm, die in Abschnitt 7.7 näher beschrieben wird.
Der Archivalienbestand auf externen Medien, wie DVDs, wird vom DMA beobachtet. Das Archivsystem kennt sowohl den Standort als auch den aktuellen Stand des Untersuchungszyklus der externen Medien. Der Untersuchungszyklus wird von der *Administration* für die externen Medien bestimmt. So kann zum Beispiel festgelegt sein, dass DVDs alle sechs Monate daraufhin untersucht wer-

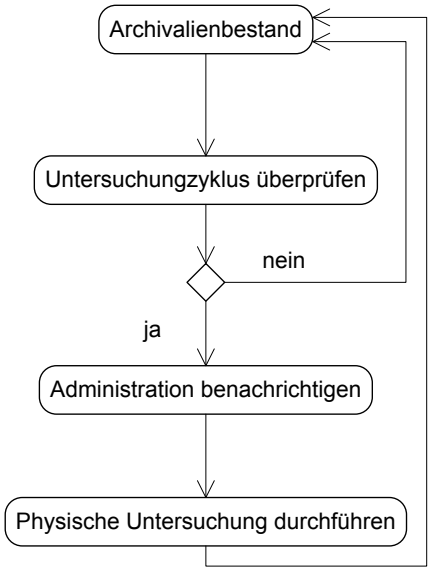

Abbildung 11: DMA Workflow Archivalien Untersuchung

den müssen, ob ihre Oberfläche Schäden genommen hat und ob die Inhalte lesbar sind. Wenn die Untersuchung eines Mediums ansteht, wird die *Administration* darüber benachrichtigt. Der Ablauf einer solchen Untersuchung wird in Abbildung 11 verdeutlicht.

7.4 Verwalten von Informationen (Data Management)

Das Archivsystem überwacht die gespeicherte *Descriptive Information* zu den *DMA Archival Information Packages*, insbesondere der Metadaten der Dateien und der definierten Dateiformate.
Das DMA bietet eine mehrdimensionale Anfragemöglichkeit, um die *Descriptive Information* über verwaltete AIPs zu erhalten. Durch die Kombination von strukturierten Metadaten, wie Partonomie und Themen, und nicht-strukturierten Metadaten, wie Kommentare und Zeitangaben, kann die gewünschte *Descriptive Information* bei einer Suchanfrage gezielt eingeschränkt werden. Ein Beispiel für eine solche Suchanfrage wird in Abbildung 12 gezeigt. Hier wird die *Descriptive Information* durch eine Anfrage über die Partonomie und spezielle Themen für Dokument- und Schadenstypen gefiltert und mit einem nicht-strukturierten Metadatum, einer Zeitangabe, weiter eingeschränkt.
Das DMA soll künftig dahingehend erweitert werden, dass es auch die Definitionen der Dateiformate archiviert. Werden zum Beispiel TIFF oder JPEG Dateien im Archiv verwaltet, muss auch die Definition dieser Dateiformate im Archiv vorhanden sein. Zusätzlich überprüft das DMA in regelmäßigen Abständen, ob die verwalteten Dateiformate dem aktuell definierten Standard des Archivs entsprechen. Diese Überprüfung wird insbesondere auch bei Änderungen und Ergänzungen der Definitionen ausgeführt. Die im Archiv verwendeten Standards werden einzig von der *Administration* bestimmt. Ein Nutzer kann zwar

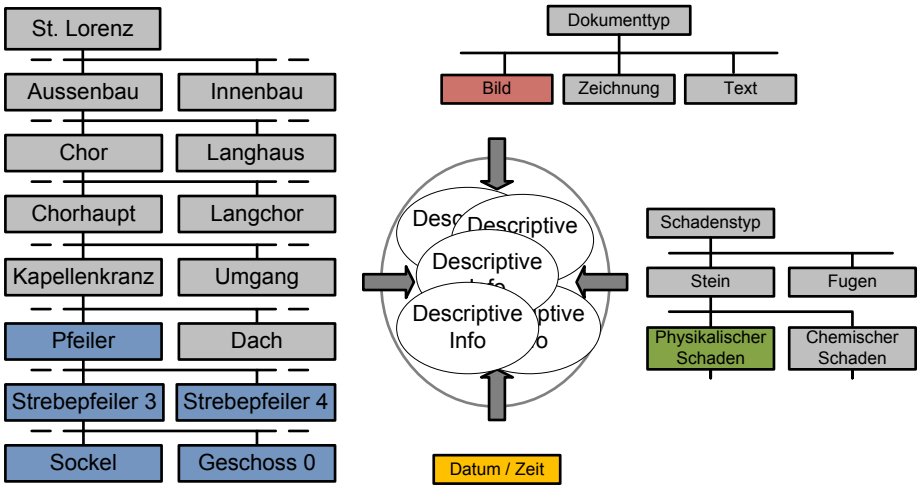

Abbildung 12: DMA Mehrdimensionale Filter

sehen, welche Standards für das Archiv definiert sind, diese aber nicht ändern. Die Dokumentation der definierten Standards wird in der Datenbank des DMA gespeichert.
Sollte der Fall eintreten, dass ein Dateiformat nicht mehr dem von der *Administration* aktuell festgelegten Stand entspricht, kann eine entsprechende Übersicht über die betroffenen Dateien generiert werden. Die *Administration* hat dann die Aufgabe, geeignete Maßnahmen, beispielsweise eine Migration, vorzunehmen. Der konkrete Ablauf der Migration kann erst zum jeweiligen Zeitpunkt bestimmt und definiert werden [Coy06].

7.5 Administration (Administration)

Es sollte mindestens einen dedizierten Administrator geben, der für die Überprüfung und das Monitoring des Archivs und des Datenbestands zuständig ist. Die Aufgaben für den operativen Betrieb des DMA sind in [RSF09b] beschrieben.
Weitere Aufgaben ergeben sich im Rahmen der Langzeitarchivierung. So muss der Administrator die Definitionen der Dateiformate verwalten und notwendige Migrationspläne erstellen, der Prozess wird in Abbildung 13 dargestellt.

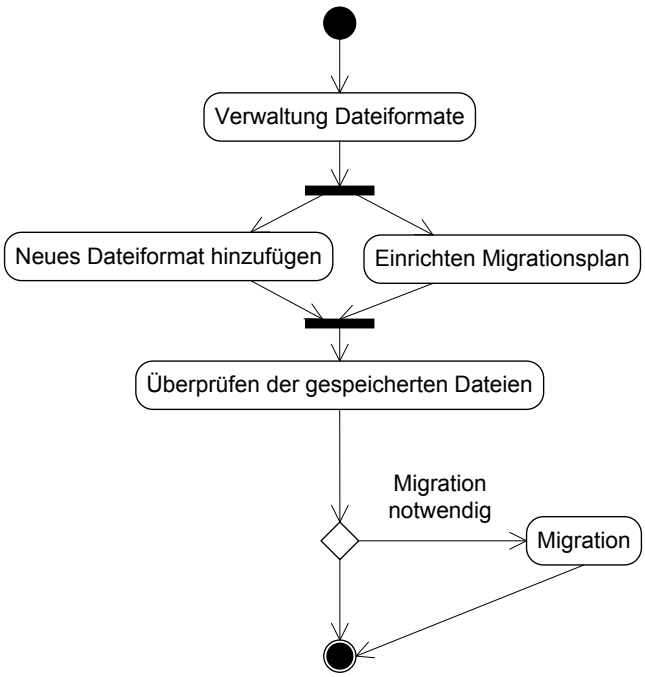

Abbildung 13: DMA Prozess Administration

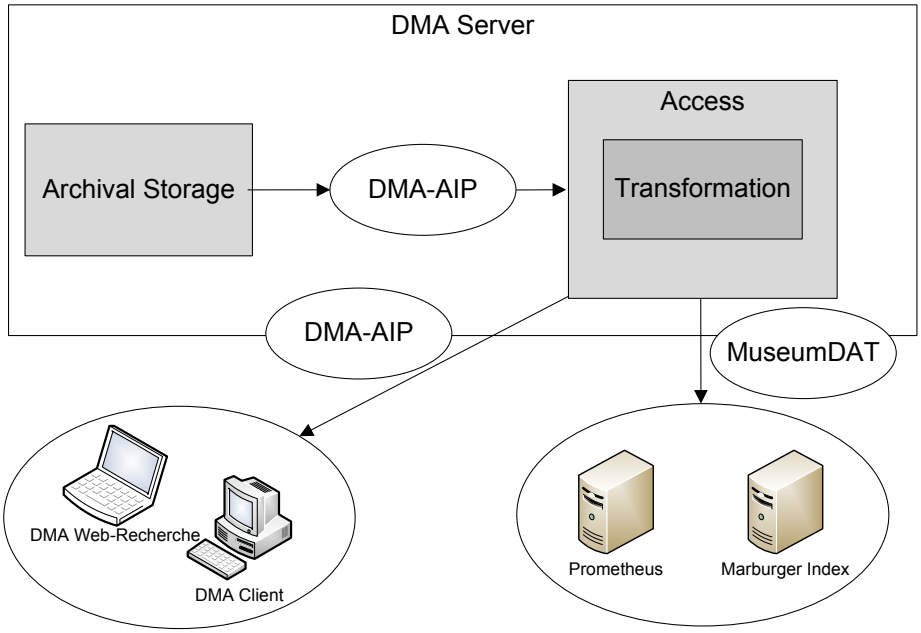

Abbildung 14: DMA Access Komponente

7.6 Zugriff auf Informationen (Access)

Für den Zugriff auf die im DMA gespeicherten Informationen werden verschiedene Schnittstellen bereitgestellt. Die *Access* erhält die DMA-AIPs in der internen RDF Darstellung, und kann diese entsprechend für den *Consumer* aufbereiten (siehe Abbildung 14). Ist der *Consumer* ein DMA-Client, können die angefragten AIPs direkt in der RDF Darstellung weitergeben werden. Somit sind hier die DMA-AIPs und DIPs identisch. Der DMA-Client präsentiert dann die Informationen in den DIPs dem anfragenden Anwender.
Anders steht es um *Consumer*, z.B. Marburger Index und Prometheus. Diese benötigen ein angepasstes Format auf Basis des CIDOC RM [CDG+06]. Hierfür wird zum nächsten Projektabschnitt ein WebService entwickelt, der die DIPs im Format des Metadatenstandards MuseumDAT [museumDAT] erstellt und ausliefert.
Eine andere Art der Schnittstelle bilden eine Reihe von Webservices, die einen Export in verschiedene Metadatenstandards anbieten. So wurde ein Webservice für den Export in die Ontologiebeschreibungssprachen RDF [RDF] und OWL [OWL] umgesetzt, um einen allgemeingültigen Metadatenstandard zu unterstützen[3]. Dies erlaubt einen umgehenden Datenaustausch ohne einen langwierigen Entwicklungsprozess für einen spezifischen Metadatenstandard zu durchlaufen.

[3]www.monarch.uni-passau.de im Bereich „Internet-Recherche“

Der Zugriff auf die Daten des DMA wird nicht nur Spezialisten, sondern auch Laien ermöglicht. Hierfür wird eine Web-Recherche implementiert, die einen Zugriff für jeden ermöglicht, der einen Internetanschluss besitzt. Diese Web-Recherche stellt eine spezielle Art des DMA-Client dar.

7.7 Sicherung der Informationen auf Mikrofilm

Abbildung 15 zeigt den für die Sicherung von Informationen auf Mikrofilm geplanten Workflow.
Zu Beginn des Arbeitsablaufs steht die *Digitalisierung* der Archivalien, hierbei sind insbesondere die Vorgaben der DFG [For09b] und aus dem Abschnitt 3.2 zu beachten. Das so entstandene *Masterdigitalisat* wird unter Angabe von weiteren Metadaten in den *DMA-Client eingefügt*. Bei den *Metadaten* müssen mindestens Angaben zu Titel, Erscheinungsort, Erscheinungsjahr und Beschreibung gemacht werden, konform zu den Vorgaben der DFG zu Metadatenangaben im METS Standard [For09b, Fed07]. Diese Informationen werden als DMA-AIP (siehe Abbildung 8) über *Ingest* in den *Archival Storage* weitergeleitet. *Ingest*

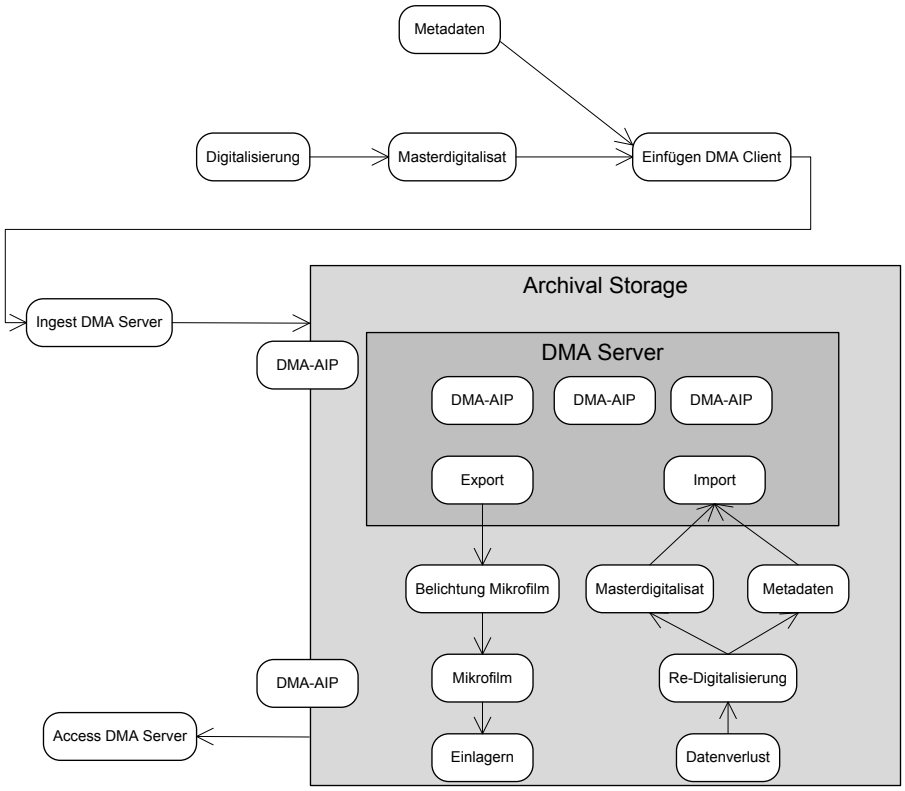

Abbildung 15: DMA Workflow Sicherung auf Mikrofilm

überprüft zusätzlich die Einhaltung der Vorgaben an das Masterdigitalisat und die Metadaten (Abschnitt 7.2). Im *Archival Storage* gibt es eine *Export*- und *Import*-Schnittstelle, welche die DMA-AIPs für eine Belichtung auf Mikrofilm vorbereiten bzw. wieder DMA-AIPs erstellen. Der *Export* übernimmt folgende Aufgaben: Konsolidierung der Metadaten und Organisation der Informationen für die *Belichtung auf Mikrofilm*. Bei der Konsolidierung der Metadaten wird insbesondere die Eindeutigkeit der Identifier und die Vollständigkeit der Metadaten überprüft. Bei der Organisation der Informationen wird die Anordnung und Aufteilung der Dokumente und Bilder mit den Metadaten auf dem Mikrofilm erstellt. Die Metadaten werden in Form von XML-Code mit den Dublin Core und METS Informationen exportiert. Somit wird bei der Belichtung eine fortlaufende Kombination aus Bildern und Dokumenten zusammen mit den passenden Metadaten auf Mikrofilm gespeichert. Die hierbei entstehenden Mikrofilme können dann in einem geeigneten physischem Archiv eingelagert werden. Bei einem Datenverlust werden die Informationen auf den Mikrofilmen *re-digitalisiert*. Hierbei werden die *Masterdigitalisate* und die zugehörigen *Metadaten* erfasst und an den *Import* weitergegeben. Der *Import* erzeugt daraus wieder die entsprechenden DMA-AIPs. Danach stehen die DMA-AIPs im *Archival Storage* zum Beispiel für den *Access* wieder zur Verfügung.

8 Zusammenfassung

Es wurde ein erster Ansatz zur Langzeitarchivierung im Digitalen Monumentalbau-Archiv beschrieben.
Dazu wurden die empfohlenen Dateiformate, wie TIFF für Bilder, PDF/A für Textdokumente oder DWG Dateien für Kartierungen vorgestellt. Des weiteren wurde vorgestellt in welchen Teilen das DMA das OAIS unterstützt und wie sinnvolle Erweiterungen aussehen können.
Es sollte beachtet werden, dass die Langzeitarchivierung in einem Archivsystem nicht allein von dessen technischen Möglichkeiten abhängt, sondern auch davon, wie gut die verwaltenden Organe, z.B. die Administratoren und Informationsverwalter, diese Möglichkeiten einsetzen.

Literatur

[Bra07] Christian Bracht. Bildarchiv Foto Marburg, Deutsches Dokumentationszentrum für Kunstgeschichte. *Rundbrief Fotografie*, 14:15–19, 2007.

[CDG+06] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. Definition of the CIDOC Conceptual Reference Model, October 2006.

[Coy06] Prof. Dr. Wolfgang Coy. nestor - materialien 5 - perspektiven der langzeitarchivierung multimedialer objekte, 2006.

[DNG] Adobe. Digital Negative (DNG) Specification, June 2009.

[DXF] Autodesk. AutoCAD 2009 DXF Specifications, DXF Reference, `http://images.autodesk.com/adsk/files/acad_dxf0.pdf`.

[Fed07] Digital Library Federation. Metadata encoding and transmission standard (METS), 2007.

[For09a] Deutsche Forschungsgemeinschaft. Aktionslinie 11: Langfristarchivierung digitaler Publikationen. `http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/digitale_information/foerderbereiche/elektronische_publikationen.html`, 2009.

[For09b] Deutsche Forschungsgemeinschaft. *DFG-Praxisregeln „Digitalisierung“*. Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS):, April 2009.

[JPEG] Independent JPEG Group (IJG) and Tom Lane. JPEG, ISO/IEC 10918-1.

[Lav04] Brian F. Lavoie. The open archival information system reference model: Introductoryguide. Technical report, Office of Research, OCLC Online Computer Library Center, Inc., 2004.

[Mon] MonArch-Projekt: www.monarch-project.eu und www.monarch.uni-passau.de.

[MRG+05] Petros Maniatis, Mema Roussopoulos, T. J. Giuli, David S. H. Rosenthal, and Mary Baker. The LOCKSS peer-to-peer digital preservation system. *ACM Trans. Comput. Syst.*, 23(1):2–50, 2005.

[museumDAT] Regine Stein, Axel Ermert, Jürgen Gottschewski, Monika Hagedorn-Saupe, Regine Heuchert, Hans-Jürgen Hansen, Angela Kailus, Carlos Saro, Regine Scheffeland, Gisela Schulte-Dornberg, Jörn Sieglerschmidt, and Axel Vitzthum. MuseumDAT - XML Schema zur Bereitstellung von Kerndaten in museumsübergreifenden Beständen.

[Nem01] Nemitz, Jürgen and Thaller, Manfred. Gleiche unter Gleichen: prometheus. Informationssysteme ohne Zwang zur Vereinheitlichung. *EDV Tage Theuern*, 2001.

[OAIS] International Organization for Standardization. OAIS: ISO 14721:2003, 2003.

[OAIS Standard] Consultative Commitee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*, 2002.

[OWL] W3C. OWL (Web Ontology Language), w3c recommendation 10 february 2004 , `http://www.w3.org/TR/owl-features/`.

[PDF/A] International Organization for Standardisation. PDF/A, ISO 19005-1:2005, 2005.

[RDF] W3C. RDF vocabulary description language 1.0: RDF Schema , W3C recommendation 10 february 2004, `http://www.w3.org/TR/rdf-schema/`.

[RSF09b] Alfons Ruch, Alexander Stenzer, and Burkhard Freitag. Betriebskonzept des Digitalen Monumentalbau-Archivs MonArch. Technical Report MIP-0911, Universität Passau, 2009.

[RSF09c] Alfons Ruch, Alexander Stenzer, and Burkhard Freitag. Backupkonzept des Digitalen Monumentalbau-Archivs MonArch. Technical Report MIP-0913, Universität Passau, 2009.

[Ru08] Alfons Ruch, editor. *Digital Preservation, Seminarband*. Universität Passau, Lehrstuhl für Informationsmanagement, 2008.

[SG08] Michael S. Seadle and Elke Greifeneder. In archiving we trust: Results from a workshop at Humboldt University in Berlin. *First Monday*, 13(1), 2008.

[SK08] Sergiy Kolesnikov. Open Archival Information System (OAIS). In Ruch [Ru08].

[SQL03] International Standardization Organization. SQL:2003 ISO/IEC 9075:2003.

[TIFF] Adobe Developers Association. TIFF, Revision 6.0, June 3 1992.

[Tür03] Can Türker. *SQL:1999 & SQL:2003 - Objektrelationales SQL, SQLJ & SQL/XML.* dpunkt.verlag, 2003.