H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.0

Kapitel 12
Technischer Workflow

nestor

vwh

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Korrektur), Martina Kerzel (Bildbearbeitung) und Jörn Tietgen (Layout und Formatierung des Gesamttextes) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte dieses Buchs stehen auch als Onlineversion
(http://nestor.sub.uni-goettingen.de/handbuch/)
sowie über den Göttinger Universitätskatalog (http://www.sub.uni-goettingen.de) zur Verfügung.
Die digitale Version 2.0 steht unter folgender Creative-Commons-Lizenz:
„Attribution-Noncommercial-Share Alike 3.0 Unported“
http://creativecommons.org/licenses/by-nc-sa/3.0/

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.
© Verlag Werner Hülsbusch, Boizenburg, 2009
www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: Kunsthaus Schwanheide

Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-940317-48-3

URL für Kapitel 12 „Technischer Workflow“ (Version 2.0): urn:nbn:de:0008-20090811614
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20090811614

Gewidmet der Erinnerung an Hans Liegmann (†), der als Mitinitiator und früherer Herausgeber des Handbuchs ganz wesentlich an dessen Entstehung beteiligt war.

12 Technischer Workflow

12.1 Einführende Bemerkungen und Begriffsklärungen

Reinhard Altenhöner

Immer dann, wenn Termini und Methoden in den Kontext eines spezifischen Themas oder spezialisierter Abläufe eingeführt werden, entsteht Bedarf für einen der eigentlichen Beschäftigung mit dem Gegenstand vorhergehenden Definitions- und Klärungsschritt. So auch in diesem Fall, wenn generelle Methoden zur Beschreibung und zur Modellierung von Abläufen auf die Langzeitarchivierung übertragen werden. Konkret: Die Langzeitarchivierung als relativ neues Arbeitsgebiet, in dem bislang der Schwerpunkt stark auf forschungsnahen oder gar experimentellen Ansätzen lag, wird beim Übergang zu produktiven Systemen und operativen Ablaufproblemen mit neuen Aufgabenstellungen konfrontiert. Und bislang steht für diesen Übergang keine spezifische Methodologie zur Verfügung, die im Sinne eines Vorgehensmodells konkrete Schritte für die Workflowentwicklung benennt.

Beim Übergang in den Langzeitarchivierungsbetrieb geht es um umfassende Arbeitsabläufe, um die massenhafte Prozessierung von (automatisierten) Arbeitsschritten und es wird sinnvollerweise auf das Erfahrungswissen und die

Methodik aus anderen Arbeitsbereichen und Geschäftsfeldern zurückgegriffen, um spezifische Antworten für eine Umsetzung im Umfeld der Langzeitarchivierung zu entwickeln. Günstig ist in diesem Zusammenhang, dass der Bewusstheitsgrad, mit dem Arbeitsprozesse im kommerziellen Kontext – oft über aufwändige Beratungsdienste durch einschlägige Anbieter - organisatorisch und technisch modelliert bzw. erneuert werden, hoch ist. Das gilt sicher generell für das Thema (technische) Prozessorganisation, um so mehr aber für das Arbeitsfeld der Langzeitarchivierung, das insbesondere in Bibliotheken, Archiven und Museen zunehmend wichtiger wird, das aber bislang bis auf wenige Ausnahmen noch nicht in größerem Umfang etabliert und in die allgemeinen Arbeitsabläufe generell integriert ist. Es folgen daher hier zunächst einige einführende Begriffsklärungen, die dann im nächsten Schritt für die konkrete Thematik Langzeitarchivierung methodisch-konzeptionell aufgegriffen werden, um schließlich in einem weiteren Schritt den bislang erreichten Praxisstand an einigen Beispielen etwas eingehender zu betrachten. Ergänzend noch der Hinweis, dass in diesem Handbuch zwischen dem organisatorischen[1] und dem technischen Workflow differenziert wird.

Der Begriff des Workflow wird im Deutschen im Allgemeinen mit dem Begriff des Geschäftsprozesses gleichgesetzt. Aus der abstrahierenden Beschreibung von Einzelfällen entsteht die Basis dafür, Abläufe systematisch als Arbeits- oder Geschäftsprozess zu beschreiben, um zum Beispiel daraus Schulungsmaterial für MitarbeiterInnen zu generieren, aber auch um Schwachstellen zu identifizieren oder neue Fallgruppen zu integrieren. Für die Etablierung neuer Geschäftsprozesse, für die bislang keine Vorlagen oder Matrizen existieren, wird auf die Ergebnisse aus einem Anforderungserhebungsprozess zurückgegriffen; das Requirements Engineering bildet einen eigenen methodischen Ansatz zur systematischen Aufarbeitung der zu lösenden Ausgangssituation. Mit der unterhalb der Ebene des Geschäftsprozesses liegenden Ebene der Arbeitsschritte – der Arbeits/Geschäftsprozess (work process) ist als eine geordnete Folge von Arbeitsschritten definiert - wird ein relativ hoher Detaillierungsgrad angestrebt, der es erlaubt, auf feingranularer Stufe Abläufe differenziert zu verstehen.

Hat man regelbasiert Abläufe beschrieben, tut sich erst die Möglichkeit auf, Geschäftsprozesse zu planen, bewusst systematischer einzugreifen, Teile oder ganze Abläufe neu zunächst abstrakt zu modellieren und dann zum Beispiel in Form von Arbeitsanweisungen praktisch umzusetzen. Auf diese Weise werden Abläufe steuerbar, sie können „gemanaged“ werden. In diesen Prozessen wer-

1 Vgl. hierzu auch den von den Herausgebern dieses Handbuchs vorgesehenen Artikel zu organisatorischen Aspekten des Workflow.

den dann Dokumente, Informationen oder auch Aufgaben und Objekte von einem Teilnehmer zum anderen gereicht, die dann nach prozessorientierten Regeln bearbeitet werden. In klassischer Definition wird der Workflow übrigens häufig mit der teilweisen oder vollständigen Automatisierung eines Geschäftsprozesses gleichgesetzt.[2] Dahinter steht die Ansicht, den Reorganisationsbedarf in Institutionen mit der Einführung von IT-gestützten Verfahren bedienen zu können mit der manchmal fatalen Folge, dass anstelle einer eingehenden Analyse der Ausgangssituation die gegebene Organisation an ein gegebenes IT-Verfahren angeglichen wird.

Enger auf den Bereich der öffentlichen Verwaltung bezogen und so auch in Bibliotheken gebraucht ist der Begriff des „Geschäftsgangs", in diesen Einrichtungen häufig festgemacht am Bearbeitungsobjekt, in der Regel Büchern oder auch Akten und dem Weg dieser Objekte durch die einzelnen Phasen seiner Bearbeitung. Gemeint ist hier letztlich – trotz der verwaltungstypischen Fokussierung auf die bearbeiteten Objekte – der Arbeitsablauf/Geschäftsprozess als die Gesamtheit aller Tätigkeiten zur Erzeugung eines Produktes bzw. zur Erstellung einer Dienstleistung.[3]

Ein „Workflow-System" bezeichnet dagegen explizit die IT-gestützte integrierte Vorgangsbearbeitung, in der Datenbank, Dokumentenmanagement und Prozessorganisation in einem Gesamtkonzept abgebildet werden.[4] Abläufe werden also technisch unterstützt, wenn nicht sogar überhaupt mit Hilfe technischer Werkzeuge und Methoden betrieben.

Aber auch die Modellierung / Aufnahme von Geschäftsprozessen selbst kann toolunterstützt erfolgen; solche Geschäftsprozeßmanagement-Tools dienen der Modellierung, Analyse, Simulation und Optimierung von Prozessen. Die entsprechenden Applikationen unterstützen in der Regel eine oder mehrere Methodiken, ihr Funktionsspektrum reicht von der Ist-Aufnahme bis zur Weitergabe der Daten an ein Workflow-Management-System. Im Mittelpunkt stehen dabei die Organisation, Aufgaben bzw. Ablauf der Aufgaben und die zugrundeliegenden Datenmodelle. Mit der Schnittstelle solcher Tools zum Beispiel zu Workflow-Management-Systemen beschäftigt sich die Workflow-Management-Coalition[5], die sich insbesondere die Austauschbarkeit der Daten und damit die Interoperabilität zwischen unterschiedlichen, zum Teil spezialisierten

2 Martin (1999), S. 2.

3 Verwaltungslexikon (2008), Eintrag Workflow. Damit der englischen Ausgangsbedeutung des Begriffs folgend.

4 Verwaltungslexikon (2008), aaO.

5 http://www.wfmc.org/
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

Tools durch entsprechende Standardisierungsanstrengungen auf die Fahnen geschrieben hat.

Der Begriff des „technischen Workflows“ schließlich wird im Allgemeinen primär für die Abläufe verwandt, die einen hohen Automatisierungsgrad bereits haben oder wenigstens das Potential dazu. Entsprechend bezeichnet man mit dem Begriff des „Technischen Workflow-Management“ die Systeme, die durch eine geringe Involviertheit von Menschen und eine hohe Wiederholbarkeit bei geringen Fehlerquoten gekennzeichnet sind.

Damit ist klar, dass der Begriff des technischen Workflow im Kontext der Langzeitarchivierung geradezu programmatischen Charakter hat, da angesichts der großen Objektmengen und ihrer prinzipiell gegebenen Eigenschaften als digitale Publikation ein hoher Automatisierungsgrad besonders bedeutsam ist. Und gleichzeitig liegt es nahe, sich bewusst auf Methoden und Werkzeuge aus dem Bereich des (technischen) Workflowmanagement zu beziehen.

12.2 Workflow in der Langzeitarchivierung: Methode und Herangehensweise

Reinhard Altenhöner

Die allmähliche Einführung der Langzeitarchivierung in das reguläre Auftragsportfolio von Bibliotheken und anderen Kulturerbeeinrichtungen mit immer höheren Bindungsquoten von Personal und anderen Ressourcen erzeugt(e) zunächst neue, häufig isolierte und händisch durchgeführte Abläufe, verändert aber auch in einer ganzheitlichen Betrachtung Arbeitsabläufe und die sie modellierenden Geschäftsprozesse. So ist schon für sich die Einspielung von Daten in ein Langzeitarchiv ein komplexer Vorgang, in dem eine ganze Reihe von aufeinander bezogenen bzw. voneinander abhängenden Aktivitäten ablaufen. Vor allem aber die zunehmende Relevanz der technischen und operativen Bewältigung der Aufgabe verlangt nach einer systematischen Modellierung der Geschäftsprozesse, also dem Einstieg in ein systematisches (technisches) Workflowmanagement. Es gilt allerdings festzustellen, dass selbst in Einrichtungen, die bereits seit einigen Jahren Erfahrungen mit dem Betrieb von Langzeitarchiven und ihrer Integration in die jeweilige Systemlandschaft gesammelt haben, häufig noch isolierte Bearbeitungsketten ablaufen, die zudem keinesfalls wirklichen Vollständigkeitsgrad haben, also alle Anforderungs- /arbeitsfelder abdecken und außerdem vielfach noch manuelle Eingriffe erfordern, insbesondere auf dem Gebiet des Fehlermanagements.

Diese Feststellung bedeutet aber auch, dass der Erfahrungshorizont zum technischen Workflow und insbesondere zum Management insgesamt noch gering ist, also hier noch konkrete Erfahrungen vor allem im Umgang mit großen Mengen und insbesondere auch im automatisierten Qualitätsmanagement gewonnen werden müssen. Insofern hat die Beschäftigung mit dem technischen Workflow derzeit noch viele theoretische Elemente und ‚propädeutischen' Charakter.

In einer Situation, in der verschiedene (bereits existente und neu entwickelte) Arbeitsprozesse ineinander greifen und auch verschiedene Organisationseinheiten an ein und demselben Vorgang beteiligt sind, ist die Modellbildung im Sinne der Geschäftsprozessmodellierung ein Beitrag zu einer umfassenden Optimierung. Damit befinden sich Bibliotheken, Archive und Museen in einer Situation, die man mit den Anstrengungen der Privatwirtschaft Anfang der 1990er Jahre vergleichen kann, als dort die Modellierung von Geschäftsprozessen unter verschärften Wettbewerbs- und Kostendruckbedingungen systematischer als zuvor angegangen wurde. Auch wenn im öffentlich finanzierten Um-

feld in besonderem Maße historisch geprägte Organisationsformen gegeben sind, die eine vorgangsbezogene Sicht erschweren, führt an der grundsätzlichen Anforderung der Neu-Modellierung aus systematischer Sicht kein Weg vorbei. Diese wird im Umfeld des technischen Workflow immer stark auch von der informationstechnischen Entwicklungsseite getrieben sein, denn Ziel der Geschäftsprozessmodellierung ist letztlich ihre technische Abbildung.

Übergeordnete Ziele dieses Herangehens, also der systematischen Modellierung und eines methodenbewussten Workflowmanagements sind[6]:

- Verbesserung der Prozessqualität
- Vereinheitlichung der Prozesse
- Schnellere und zuverlässigere Bearbeitung von Aufträgen (extern und intern)
- Reduzierung der Durchlaufzeiten
- Kostenreduktion
- Verbesserte Verfügbarkeit von Information / Dokumentation
- Erhöhte Prozessflexibilität
- Erhöhung der Transparenz der Prozesse (Statusermittlung, Dokumentation von Entscheidungen), Qualitätssicherung
- Automatische Eingriffsmöglichkeiten: Dokumentation, Eskalation bei Zeitüberschreitungen, Verteilung von Aufgaben und Verantwortlichkeiten
- Vermeidung von Redundanz, mangelnder Aktualität und Inkonsistenz durch Mehrfachschritte

Natürlich lassen sich kleine isolierte Prozesse oder Prozesselemente durch individuelle Programmierung jeweils neu umsetzen. Dies geschah in der Vergangenheit vielfach für einzelne Objektklassen oder auch einzelne Datenübergabe- oder –tauschprozesse. Aber schon beim Zusammenführen bzw. Hintereinandersetzen der einzelnen Teilschritte bedarf es einer Gesamtlogik für das Management des Ablaufs dieser Schritte. Fehlt diese Logik, entstehen letztlich viele immer wieder manuelle neu anzustoßende Teilkonstrukte mit dazu häufig proprietären „Konstruktions“elementen. Schon insofern ist die systematische Analyse verschiedener wiederkehrender Arbeitsabläufe ein sinnvoller Ansatz, um so zur Modellierung auch komplexer Vorgänge im Bereich der Langzeitarchivierung zu kommen.

6 Die folgende summarische Zusammenstellung betrifft sowohl organisatorische wie technische Aspekte des Workflowmanagements. Eine Trennung ist theoretisch zwar möglich, praktisch aber nicht sinnvoll.

Ziel dieses systematischen Ansatzes ist es, Services zu definieren, die auch in anderen Kontexten (wieder) verwendbar sind. Sie bilden Arbeitsschritte granular ab, die so in verschiedenen Umfeldern vorkommen (können), beispielsweise das Aufmachen eines Bearbeitungsfalls für ein Objekt und die IT-gestützte Verwaltung verschiedener Be-/Verarbeitungsschritte dieses Objekts. In dieser Perspektive entsteht der Geschäftsprozess für eine Klasse von Objekten aus der Zusammenfügung verschiedener Basisservices, die miteinander interoperabel sind. Dass diese Herangehensweise sehr stark mit dem Modell der Serviceorientierten Architektur (SOA) bei der Entwicklung IT-basierter Lösungen korrespondiert, ist dabei kein Zufall. Voraussetzung dafür ist aber die systematische Modellierung der Arbeits- oder Geschäftsprozesse, die vorgeben, welche Services wann und wie gebraucht werden. Die Prozessmodellierung bildet also die Basis für die Implementierung, die Prozesse selbst dienen der Orchestrierung, dem Zusammenspiel und der Aufeinanderabstimmung der Services. In einem optimalen (Infrastruktur)-Umfeld können so die Arbeitsschritte als kleinere Einheit eines Geschäftsprozesses verschiedene Services lose zusammenbringen.

Der Ansatz, Services nachnutzbar zu gestalten, bezieht sich in der Regel auf eine Organisation. Zwar wird immer wieder versucht, Geschäftsprozesse aus einem institutionellen Umfeld auf ein anderes zu übertragen, allerdings erweist sich dies in der Praxis als außerordentlich schwierig: Zu stark sind die Abweichungen der einzelnen Arbeitsschritte voneinander und zu unterschiedlich die jeweiligen Prioritäten und Schwerpunktsetzungen in den einzelnen Institutionen. Hinzu kommt außerdem noch, dass der Prozess der Modellierung und Ausgestaltung von Geschäftsprozessen selbst erhebliche Erkenntnisgewinne in der jeweiligen Organisation mit sich bringt, die für eine erfolgreiche Einführung neuer oder veränderter Geschäftsprozesse unverzichtbar sind.

Die Informatik hat für die Modellierung und Notation von Geschäftsprozessen verschiedene methodische Herangehensweisen entwickelt, zum Beispiel die Ereignisgesteuerten Prozessketen (EPK), eine von Scheer und Mitarbeitern entwickelte Sprache zur Modellierung von Geschäftsprozessen[7] und vor allem die Unified Modeling Language (UML) der Object Management Group (OMG), die in der Praxis heute dominierende (technische) „Sprache“ für die Modellierung von Daten, Verhalten, Interaktion und Aktivitäten.[8]

Legt man also UML als Syntax fest, sind noch methodische Festlegungen für die Herangehensweise zu treffen und es liegt nahe, sich für die vorbereitende Modellierung von technischen Abläufen in der Langzeitarchivierung am OAIS-

7 Keller (1992)

8 OMG Infrastructure (2007) und OMG Superstructure (2007)

Modell zu orientieren, das die prinzipiellen Aufgaben im Umfeld der Langzeitarchivierung in funktionaler Perspektive beschreibt und an anderer Stelle dieser Enzyklopädie ausführlich beschrieben wird.[9] Für den Bereich des Ingest einzubeziehen ist der Producer-Archive Interface Methodology Abstract Standard" (CCSDS 651.0-B-1), der insbesondere Validierungsmechanismen betrachtet.[10]

Einzelne Funktionen lassen sich so vor der Folie bisher bereits gemachter Erfahrungen allgemein beschreiben. Beispiele für diese übergreifenden Basisprozesse sind (ich nenne nur Beispiele für unmittelbar aus dem Kontext der Langzeitarchivierung heraus relevante Prozesse):

- Plattform- und Systemübergreifendes Taskmanagement
- Daten- und Objekttransfer-Mimik (z.B. OAI, ORE)
- Extraktion und Generierung von Metadaten (METS, LMER)
- Validierung von Dokumentformaten (z.B. JHOVE)
- Persistente Adressierung und Zugriffsmanagement auf Objektebene
- Speicherprozesse
- ID-Management
- Inhaltsauswahl / Basisrecherche
- Migrationsprozesse / Formatkonvertierungen
- On-the-fly-Generierung einer Bereitstellungsumgebung

9 Vgl. hierzu den entsprechenden Artikel von Nils Brübach / Manuela Queitsch / Hans Liegmann (†) in dieser Enzyklopädie als Kapitel 4: „Das Referenzmodell OAIS - Open Archival Information System"

10 Vgl. hierzu http://public.ccsds.org/publications/archive/651x0b1.pdf

12.3 Technisches Workflowmanagement in der Praxis: Erfahrungen und Ergebnisse

Reinhard Altenhöner

Massenprozesse in der Langzeitarchivierung sind noch wenig etabliert; daher ist der Umfang praktischer Erfahrungen noch begrenzt. Wichtige Erkenntnisse konnte sowohl in der technischen Workflowentwicklung als auch in der praktischen Umsetzung die niederländische Nationalbibliothek sammeln. Auch in der Deutschen Nationalbibliothek liegen erste Erfahrungen vor[11]: Nach einer Gesetzesnovelle Mitte des Jahres 2006 hat sie die Zuständigkeit für die Erhaltung der Langzeitverfügbarkeit deutscher Online – oder Netzpublikationen erhalten und steht nun vor sehr konkreten Herausforderungen, die derzeit zu einer umfassenden Reorganisation des technischen Workflow führen.[12] Mit dem Inkrafttreten des neuen Gesetzes und der damit verbundenen deutlich erweiterten Verpflichtung, die Aufgabe der Langzeitarchivierung zu erfüllen, stellt sich hier die Frage in einer neuen Dimension: Wie wird die Bibliothek die neuen Abläufe organisieren, welche technischen Methoden und Anwendungen werden im Massenverfahren eingesetzt? Da gleichzeitig die alten Arbeitsabläufe und –verfahren weiterlaufen, stellt sich die Frage der Integration in ganz anderer Weise. Zwar ist die Bibliothek in der glücklichen Situation, für die neuen Aufgaben zusätzliche Ressourcen erhalten zu haben, doch würden diese nicht eine nahtlose Imitation des organisatorisch-operativen Workflows auf Basis der existierenden Systeme abdecken – das ergibt sich schon aus den Mengen, um die es geht.

Die Königliche Bibliothek der Niederlande (KB) betreibt seit dem Jahr 2003 das OAIS-kompatible Archivierungssystem DIAS der Firma IBM operativ und hat im Laufe der gewonnenen Erfahrungen insbesondere organisatorisch eine ganze Reihe von Anpassungen vorgenommen.[13] Technisch gesehen wurde

---

11 Vgl. hierzu den einführenden Artikel von Maren Brodersen / Sabine Schrimpf im 18. Kapitel „Praxisbeispiele“ dieser Enzyklopädie unter dem Titel „Langzeitarchivierung von elektronischen Publikationen durch die Deutsche Nationalbibliothek“:

12 Es sei angemerkt, dass es eine ganze Reihe von weiteren Publikationen zum Thema gibt. So stellte etwa Clifton (2005) Worksflows der australischen Nationalbibliothek vor; diese beziehen sich allerdings auf die manuelle Behandlung von Objekten mittels einzelner Tools. Seit 2007 läuft in der australischen Nationalbibliothek ein Projekt zur Etablierung und IT-basierten Unterstützung der Datenmigration von physischen Datenträgern; noch ist es zu früh, um die Übertragbarkeit bzw. Nachnutzbarkeit des Ansatzes beurteilen zu können, vgl. http://prometheus-digi.sourceforge.net/

13 KB (2008)

eine auch in der KB weitgehend isolierte eigene Entwicklung aufgesetzt, die nur in geringem Maße an die sonstigen Abläufe der Bibliothek angebunden ist. Schwerpunkt liegt auf dem Ingest-Prozess, also dem Einspielen des in der Regel von Verlagen bereitgestellten publizierten Materials in das Archiv. Dieses erfolgt weitgehend automatisiert und es ist der Niederländischen Nationalbibliothek sehr schnell gelungen, die Fehlerquoten auf niedrige Promillebereiche zu drücken. Inzwischen sind mehr als zehn Millionen Objekte eingespielt, darunter auch (allerdings wenige) komplexe Objekte wie historische CD-ROMs. Für alle Objekte – es handelt sich in der weit überwiegenden Zahl um PDF-Dateien – gilt, dass in der eigentlichen Langzeitarchivumgebung nur rudimentäre Metadateninformationen gespeichert werden; die bibliographischen Informationen werden über das Recherchesystem der KB zur Verfügung gestellt.

Insgesamt ist es der KB gelungen, den technischen Workflow relativ unkompliziert und damit effizient und für hohe Durchsatzmengen geeignet zu gestalten. Dies war auch deswegen möglich, weil die Zahl der Lieferanten in das System in den Niederlanden zumindest in der Startsituation klein war, da wenige große Verlage bereits einen überwiegenden Anteil am Publikationsvolumen der Niederlande haben.

In Deutschland stellt sich die Situation anders dar: Hier bestimmen in einer zum Teil noch sehr traditionell geprägten Veröffentlichungslandschaft viele Verleger das Bild. Ausgangspunkt für die Deutsche Nationalbibliothek bei der Neukonzipierung ihrer technischen Abläufe war eine Situation, in der für die Verarbeitung von Online-Dokumenten bereits eine Vielzahl von mehr oder weniger halbautomatischen Verfahren für Netzpublikationen, Online-Dissertationen und weitere Materialien existierte. Diese historisch gewachsenen Strukturen standen nebeneinander, d.h. – nicht untypisch für Gedächtnisorganisationen im öffentlichen Kontext – die einzelne Objektklasse war der definitorische Ausgangspunkt für einen hochspezialisierten Workflow. Ziel war und ist daher die Schaffung eines automatischen, einheitlichen Verfahrens mit der Übergabe der Archivobjekte an das im Rahmen des Projekts kopal entstandene Archivsystem und die dort entstandenen Verfahren.[14] Davon betroffen sind sowohl der Ingest wie aber auch der Zugriff auf die Objekte: Aus der Langzeitarchivlösung kopal werden Objekte an den Arbeitsplatzrechner übergeben oder über das in der Konzeptionsphase befindliche Bereitstellungssystem zur Verfügung gestellt. Dabei sind zahlreiche Arbeitsbereiche in der DNB involviert: neben dem bibliographischen System sind dies die Fachbereiche, externe Ablieferer,

14 kopal (2008)

aber auch die für die digitalen Dienste der DNB Verantwortlichen. Insofern ist hier vieles noch offen und ein Werkstattbericht mag dies illustrieren:[15]

Für den Transfer und das Angebot von Objekten auf elektronischen Materialien auf physischen Datenträgern (d.h. CD- bzw. DVD-Veröffentlichungen) existiert ein älterer, segmentierter Workflow, der nun aufgrund der Anforderungen seitens Archivsystem und künftiger Bereitstellung anzupassen ist. Nach Erstellung der Images der Daten und einer Analyse des vorhandenen Materials wurde daher ein Änderungs- und Ergänzungsvorschlag für den integrierten Workflow dieser Materialgruppe erarbeitet.

Ebenso wird der Workflow für genuin online vorliegende Netzpublikationen unter Einbeziehung der Anforderungen der Langzeitarchivierung neu gestaltet und auf die Schnittstellen des Archivsystems angepasst. Dabei ergeben sich eine ganze Reihe von Problemen: So entsprechen fortlaufende Publikationen (vor allem elektronische Zeitschriften-Artikel) und die künftigen zu archivierenden Objekte häufig nicht der aktuellen Abbildung im Online-Katalog. Bibliografische Metadaten von Archivobjekten müssen aber künftig im bibliografischen System abgebildet werden, um einen einheitlichen Zugang zu gewährleisten. Dazu müssen eine Festlegung von Erschließungsvarianten und ein Mapping von Archivobjekten auf Katalogobjekte erfolgen, letztlich also eine klare Definition der Granularität von Objekten und ihrer Abbildung gefunden werden.

Das URN-Management in der DNB wurde bereits erweitert und vor allem technisch so weiterentwickelt, dass eine Einbindung in andere Arbeitszusammenhänge/Module erfolgen kann. Da jedes Objekt zum Einspielen in das Archiv einen Persistent Identifier benötigt, erfolgt für bereits gesammelte Objekte ohne URN eine retrospektive Vergabe der URN. Alle neuen Objekte müssen entweder mit URN geliefert werden bzw. bei Eingang/Bearbeitung einen URN erhalten, was dem künftigen Verfahren entspricht.

Wesentliche Voraussetzungen für die Einbindung des Archivs in die Geschäftsumgebung der Institution liegen mittlerweile vor oder werden gerade geschaffen. Insbesondere die Kernelemente des Produktionssystems laufen, das produktive Einspielen von Material wurde und wird erprobt, nötige Weiterentwicklungen (z.B. noch fehlende Module zur Auswertung von Dateiformaten) wurden und werden ermittelt und Änderungen / Anpassungen in diversen Workflows der traditionellen Bearbeitung wurden bereits angestoßen. Weitere Aufgaben betreffen in hohem Maße die Übergabe des kopal-Systems, die Eta-

15 Wollschläger (2007), S. 18ff.

blierung einer ständigen Arbeitseinheit sowie die retrospektive Aufarbeitung des früher bereits in die Bibliothek gelangten Materials.

Hinter diesen Bemühungen steht der Anspruch, die neuen, mit der Gesetzesnovelle übernommenen Aufgaben, die weit über das Arbeitsfeld der Langzeitarchivierung hinausgehen, in einem ganzheitlichen technischen Workflow abzubilden. In dessen Mittelpunkt stehen aktuell die Übernahme von elektronischen Objekten mit möglichst breiter Nachnutzung vorhandener Metainformationen und die Integration der Abläufe in die Arbeitsumgebung der DNB.

Die praktischen Erfahrungen an der DNB insbesondere für diesen Bereich belegen den besonderen Bedarf für eine bewusste Modellierung der Geschäftsprozesse, die in der Vergangenheit häufig nur unvollkommen gelungen ist. Im Ergebnis standen isolierte, von nur wenigen Personen bediente und bedienbare Abläufe mit einem hohen manuellen Eingriffs- und Fehlerbehandlungsbedarf. Ohne dass heute bereits ein komplettes Profil der zukünftigen technischen Workflow-Umgebung zitierfähig vorliegt, kann doch gesagt werden, dass ein methodisch bewusstes, in enger Kooperation von Bedarfsträger und Informationstechnik ablaufendes Vorgehen zu deutlich klareren Vorstellungen darüber führt, wie die wesentlichen Arbeitsschritte exakt aussehen und wie sie adäquat so abgebildet werden, dass die entstehenden Services auch langfristig und damit über ihren aktuellen Entstehungshintergrund hinaus genutzt werden.

Dass dabei für eine technische Arbeitsumgebung besondere Anforderungen an die Flexibilität und die Orientierung an offenen Standards gelten, liegt auf der Hand und hat wesentlich die Entwicklungsleitlinien für kopal mitbestimmt.[16]

Quellenangaben

Clifton, Gerard: *Safe Havens In A Choppy Sea: Digital Object Management Workflows At The National Library of Australia* (2005), Beitrag zur iPRES - International Conference on Preservation of Digital Objects, Göttingen (September 15, 2005). In: http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Safe%20Havens%20In%20A%20Choppy%20Sea%20Digital%20Object%20Management%20Workflows%20At%20The%20National%20Library%20of%20Australia%20-%20Gerard%20Clifton.pdf

16 kopal (2008a)

Keller, Gerhard / Nüttgens, Markus / Scheer, August-Wilhelm (1992): *Semantische Prozessmodellierung auf der Grundlage „Ereignisgesteuerter Prozessketten (EPK)*. In: A.-W. Scheer (Hrsg.): Veröffentlichungen des Instituts für Wirtschaftsinformatik, Heft 89, Saarbrücken. Online in: http://www.iwi.uni-sb.de/Download/iwihefte/heft89.pdf

Königliche Bibliothek der Niederlande (KB): *How the e-Depot works* (2008) In: http://www.kb.nl/dnp/e-depot/dm/werking-en.html

Königliche Bibliothek der Niederlande (KB): *The e-Depot system (DIAS)* (2008) In: http://www.kb.nl/dnp/e-depot/dias-en.html

Kopal (2008): Projekthompage. In: http://kopal.langzeitarchivierung.de/

Kopal (2008a): *kopal: Ein Service für die Langzeitarchivierung digitaler Informationen*. In: http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf

Martin, Norbert (1999): *Und wie kommt die Dissertation auf den Server? Gedanken zum Workflow. Vortrag auf der IuK-Tagung „ Dynamic Documents „ vom 22.-24.3.1999 in Jena*. In: http://edoc.hu-berlin.de/epdiss/jena3/workflow.pdf

OMG Infrastructure (2007). *UML Infrastructure Specification, v2.1.2*. OMG document formal/07-11-04. In: http://www.omg.org/docs/formal/07-11-04.pdf

OMG Superstructure (2007). *UML Superstructure Specification, v2.1.2*. OMG document formal/07-11-02. In: http://www.omg.org/docs/formal/07-11-02.pdf

Stapel, Johan: *The KB e-Depot. Workflow Management in an Operational Archiving Environment* (2005). Beitrag zur iPRES - International Conference on Preservation of Digital Objects, Göttingen (September 15, 2005). In: http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Workflow%20Management%20In%20An%20Operational%20Archiving%20Environment%20-%20Johan%20Stapel.pdf

Verwaltungslexikon (2008) *Management und Reform der öffentlichen Verwaltung* (2008) In: http://www.olev.de/w.htm

Wollschläger, Thomas (2007): *„kopal goes live“*. In: Dialog mit Bibliotheken 19 (2007), H.2, S. 17 – 22

Workflow Management Coalition (2008) – Website. In: http://www.wfmc.org/

12.4 Systematische Planung von Digitaler Langzeitarchivierung

Carmen Heister, Hannes Kulovits, Christoph Becker und Andreas Rauber

Durch ständige technologische Veränderungen weisen digitale Objekte eine geringe Lebensdauer auf. Digitale Langzeitarchivierung ist somit zu einer dringlichen Aufgabe geworden. Zur langfristigen Bewahrung digitaler Objekte müssen diese mit Tools zur Langzeitarchivierung bearbeitet werden. Die Wahl eines spezifischen Tools für die Format-Migrationen oder Emulationen und die Einstellung spezifischer Parameter ist jedoch eine sehr komplexe Entscheidung. Die Evaluierung, ob und zu welchem Grad potentielle Alternativen spezifische Anforderungen erfüllen und die Erstellung eines soliden Plans zur Erhaltung einer bestimmten Gruppe von Objekten lässt sich als „Planung von Langzeitarchivierung" zusammenfassen. Derzeit wird die Langzeitarchivierungsplanung manuell, meist ad-hoc, mit wenig oder keiner Softwareunterstützung durchgeführt. Dieses Kapitel stellt einen Workflow vor, der hilft, diesen Planungsprozess zu systematisieren.

Einführung

Es gibt eine Reihe von Strategien und Tools, welche die digitale Langzeitarchivierung unterstützen, jedoch fehlt oftmals eine Entscheidungshilfe für die Auswahl der optimalen Lösung. Für die Wahl einer geeigneten Archivierungsstrategie und eines konkreten Tools müssen komplexe Anforderungen bedacht werden. Sorgsame Dokumentation und gut definierte Vorgehensweisen sind nötig um sicherzustellen, dass das Endergebnis zur Planung von Erhaltungsmaßnahmen den Anforderungen der jeweiligen Einrichtung, insbesondere den Nutzern der Objekte („Designated Community") entspricht. Dies ist auch eine der Kernaufgabe von TRAC[17] und nestor[18].

Eine sorgfältige Planung der digitalen Langzeitarchivierung unterstützt den Entscheidungsprozess zur Auswahl der optimalen Lösung, indem im Planungsprozess verfügbare Lösungsmöglichkeiten gegen klar definierte und messbare Kriterien evaluiert werden. Sie stellt eine Kerneinheit des Open Archival Information System (OAIS) Referenzmodells dar[19], insbesondere im Funktionsmodell Preservation Planning – siehe Kapitel 4. Die Planung besteht aus einem konsistenten Workflow, der idealerweise zu einem konkreten Langzeitarchivie-

17 OCLC (2007)
18 nestor (2006)
19 CCDS (2007)

rungsplan *(„preservation plan“)* führt. Für die Planung der digitalen Langzeitarchivierung muss der Planungsbeauftragte über mögliche Lösungswege, die auf die betreffenden Objekte anwendbar sind, informiert sein. Es wird ein vorzugsweise automatisierter Vergleich von Dokumenten und Objekten vor und nach der Verwendung einer Archivierungsstrategie (z.B. einer Migration oder Emulation) benötigt, um die Qualität der verwendeten Erhaltungsmaßnahme *(„preservation action“)* zu evaluieren. Der Prozess der zur Auswahl der Erhaltungsmaßnahme geführt hat, sollte darüber hinaus wiederholbar und auch gut dokumentiert sein, um die Nachvollziehbarkeit sowohl der zu Grunde liegenden Entscheidungen als auch der Gründe für die Wahl der Erhaltungsmaßnahme zu gewährleisten. Der im Folgenden vorgestellte Workflow stellt eine Konkretisierung der funk-

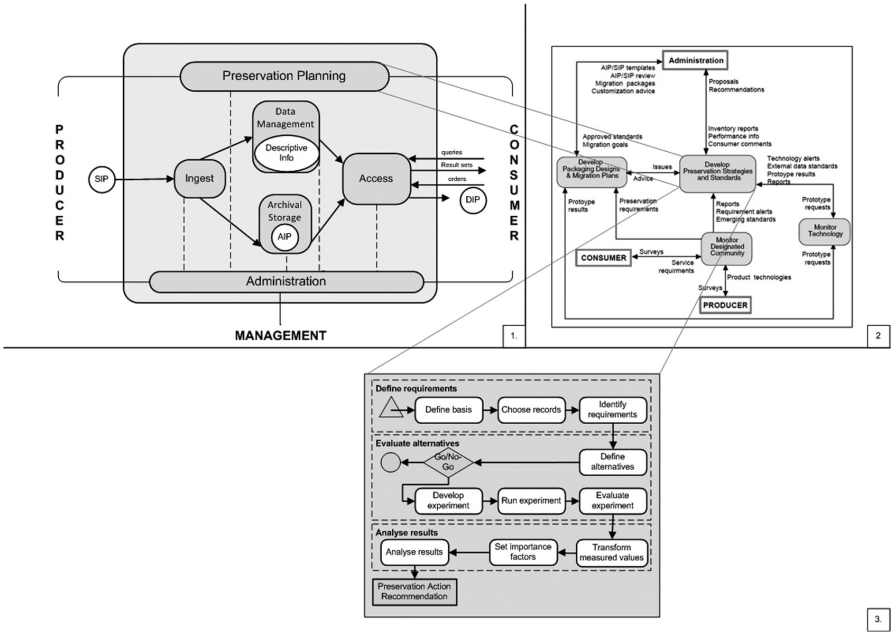

Abbildung 1: OAIS-Modell < Funktionale Entität „Preservation Planning“ < Planungs-Workflow

tionalen Komponente *„Develop Preservation Strategies and Standards“* aus dem als ISO 14721 verabschiedeten OAIS Modell „Preservation Planning“ dar (Abbildung 1). Der Workflow wurde ursprünglich im Rahmen des Preservation

Clusters des EU NoE DELOS[20] (Network of Excellence on Digital Libraries)[21] konzipiert und nachfolgend im Rahmen des EU Projektes Planets[22] (Preservation and Long-Term Access via Networked Services) verfeinert.[23] Der Workflow basiert auf der Nutzwert-Analyse, einem Verfahren ähnlich der Kosten-Nutzen-Rechnung, kombiniert mit experimenteller Evaluierung.[24]

Der PLANETS Workflow zur Langzeitarchivierung

Anforderungserhebung („Define requirements")

Die Phase 1 des Planungsverfahrens ist die Anforderungserhebung. Dazu gehören das Sammeln von Anforderungen von einer möglichst breiten Nutzergruppe (Abbildung 2), sowie der Faktoren der institutionellen Umgebung, welche die Langzeitarchivierung beeinflussen.

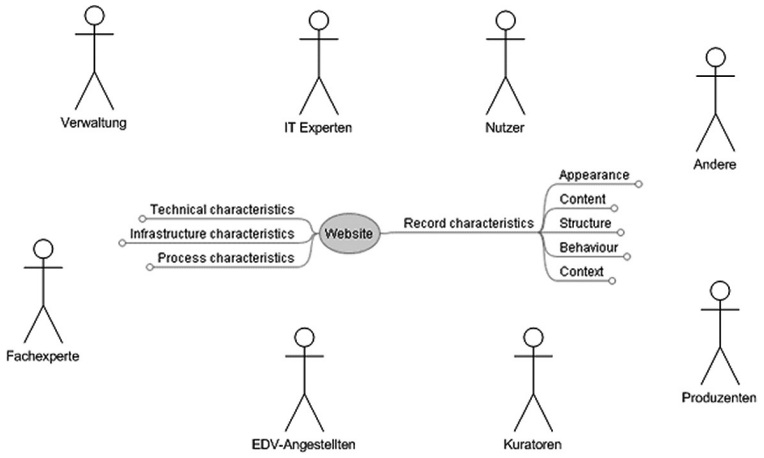

Abbildung 2: Experten, die die Anforderungen auswählen

20 http://www.delos.info/
21 Strodl (2006)
22 http://www.planets-project.eu
23 Farquhar (2007)
24 Rauch (2004)

Evaluierung der Alternativen („Evaluate alternatives")

Die zweite Phase besteht in der Auswahl der in Frage kommenden Strategien, ihrer experimentellen Anwendung auf ausgewählte Beispielobjekte und der Evaluierung der Alternativen bezüglich der definierten Anforderungen.

Analyse der Ergebnisse („Analyse results")

In der dritten Phase werden die Alternativen in ihren Stärken und Schwächen verglichen und analysiert. Auf dieser Basis sind dann fundierte und gut dokumentierte Entscheidungen zur Auswahl der optimalen Strategie möglich.

Erstellen eines Plans zur Langzeitarchivierung

(„Build preservation plan")

Der Plan zur Langzeitarchivierung wird in der vierten Phase in der funktionalen Entität „*Develop Packaging Designs & Migration Plans*" im OAIS-Model nach Genehmigung der empfohlenen Strategie in „*Administration*" erstellt. Er legt fest, welche Archivierungsmaßnahmen wie und von wem durchgeführt werden sollen. Änderungen an den Objekten, eine veränderte Umgebung oder neue Technologien machen es unter Umständen notwendig den Plan anzupassen. Eine Überwachung dieser Parameter und daraus resultierende Veränderungen am Plan bewirken einen ständigen Kreislauf im Planungsprozess.

Detaillierte Beschreibung des Workflows

Im folgenden Abschnitt wird auf die drei Kernphasen des Workflows genauer eingegangen, da sich dieses Kapitel auf die Planungsphasen konzentriert.

Festlegen der Grundlagen („Define basis")

Im ersten Schritt der Phase 1 wird der Kontext des Planungsvorhabens dokumentiert. Dies beinhaltet den Namen des Planes sowie den Namen der Planungsverantwortlichen. Es wird der organisatorische Rahmen dokumentiert, welche Planungsziele die jeweilige Institution hat, was der Planungsgrund ist, welche Zielgruppe angesprochen wird, welche institutionellen Richtlinien zur Langzeitarchivierung existieren (vgl. Kap. 4.2) und welche rechtlichen Bedingungen, personellen sowie finanziellen Ressourcen und organisatorischen Einschränkungen für die Planung wichtig sind.

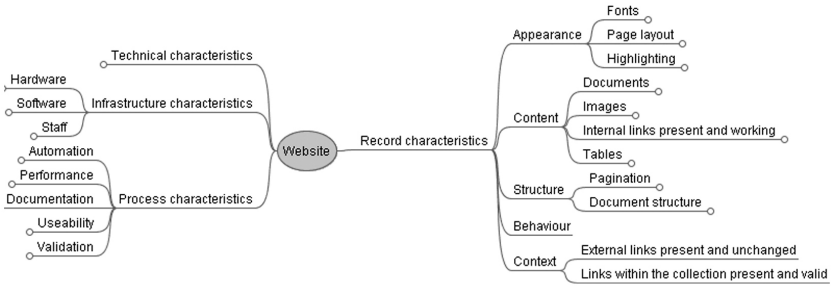

Abbildung 3: Anforderungsform als Mindmap

Auswahl der Datensätze („Choose records")

Im zweiten Schritt werden repräsentative Beispielobjekte ausgewählt, welche die essenziellen Charakteristiken der gesamten Kollektion abdecken. In einem Planungsszenario für die Langzeiterhaltung von digitalen Dissertationen wären das beispielsweise: Eine Dissertation, die sehr viele Abbildungen enthält, eine sehr große bzw. sehr kleine Datei, eine Dissertation, die mathematische Gleichungen/Abbildungen enthält, und eine Datei, die interaktive Elemente beinhaltet. Diese Beispielobjekte werden im späteren Verlauf zur Evaluierung potenzieller Lösungen herangezogen. Üblicherweise werden drei bis fünf Objekte ausgewählt.

Identifizieren der Anforderungen („Identify requirements")

Das Ziel dieses entscheidenden Schrittes ist die Dokumentation der Anforderungen für eine Archivierungsstrategie in klarer und eindeutiger Weise. Diese explizite Definition z.B. der bewahrten Eigenschaften ist eine Kernfordernis sowohl des nestor Kriterienkataloges (Punkt 9.3) [25] als auch des „TRAC"-Kataloges (Punkt B 2.1.)[26]. Allgemeine Ziele und detaillierte Anforderungen werden in einer Baumstruktur, dem sogenannten Kriterienbaum („*Requirements tree*", „*Objective tree*"), gesammelt und strukturiert (Abbildung 3). Die Inhalte des Kriterienbaumes bilden die Basis der quantitativen Evaluierung und ermöglichen dadurch eine objektive Entscheidungsfindung. Während sich die Kriterien im Allgemeinen je nach Kontext und Szenario relativ stark unterscheiden, können einige allgemein gültige Prinzipien festgehalten werden - so hat es sich als

25 nestor (2006)
26 OCLC (2007)

zweckmäßig erwiesen, die Bäume auf der obersten Ebene in vier Hauptkategorien zu unterteilen:

- *Objekteigenschaften („Object characteristics")* beschreiben Anforderungen, inwieweit visuelle und inhaltliche Erlebnisse des Benutzers bei der Betrachtung eines digitalen Objektes erhalten bleiben. Zur Beschreibung der wesentlichen Eigenschaften werden primär fünf Aspekte eines digitalen Objektes herangezogen: Inhalt, Aussehen, Struktur, Verhalten und inhaltlicher Kontext (Metadaten). Bei der nachfolgenden experimentellen Analyse wird gemessen, wie gut diese Eigenschaften bei Anwendung der Erhaltungsmaßnahme erhalten bleiben.
- *Datensatzeigenschaften („Record characteristics")* beschreiben den technischen Kontext der Objekte beziehungsweise die verteilten Strukturen. So können z.B. in Powerpoint-Präsentationen Foliensätze, Videos oder Bilder über eine Verlinkung eingebettet sein oder Webseiten aus zahlreichen Komponenten wie z.B. „Styles-sheets", „Bildern", etc. aufgebaut sein. Diese Kriterien werden unter Umständen vom Anwender nicht unmittelbar wahrgenommen, wenn er mit dem digitalen Objekt zu tun hat. Trotzdem sind sie notwendig, um das Objekt in den Originalzustand zu überführen und damit seine ursprüngliche Wirkung und integrierte Funktionalität wieder herzustellen.
- *Prozesseigenschaften („Process characteristics")* beziehen sich auf den Prozess beziehungsweise das Tool selbst. Sie beinhalten unter anderem die Skalierbarkeit eines Verfahrens auf große Datenmengen oder die Komplexität eines Verfahrens, aber auch Aspekte der Automatisierbarkeit, inwieweit manuelle Eingriffe notwendig sind, etc.
- *Kosten („Costs")* spielen normalerweise eine wichtige Rolle beim Treffen der Entscheidungen. Sie können im Prinzip bei den jeweiligen Bereichen im Baum aufgeführt werden. Aus Gründen der besseren Gliederung werden sie aber meist in einem eigenen Zweig gebündelt. Sie können in technische Kosten und Personalkosten unterteilt werden sowie in Initialkosten und laufende Ausgaben.

Ein Kriterienbaum unterstützt die Identifikation von Kriterien und wird jeweils an einzelnen Stellen erweitert, an anderen reduziert, falls das eher den Vorstellungen einer Organisation entspricht. Zur vollständigen Identifikation der Kriterien ist meist eine ausführliche Recherche relevanter Literatur für das jeweilige Anwendungsgebiet sowie eine detaillierte Analyse der zu bewahrenden Objekte und Anforderungen erforderlich.

Die Kriterienbäume werden üblicherweise in Workshops erstellt, bei denen Anwender, Techniker und Archivierungsexperten zusammenarbeiten, um die relevanten Anforderungen zu ermitteln und zu strukturieren. Ein zentrales Element der Anforderungsanalyse in diesem Zusammenhang ist stets die quantitative Natur der Nutzwertanalyse. Jede Anforderung sollte soweit als möglich objektiv messbar gemacht werden. Daher wird jedem Kriterium in der untersten Ebene eine Skala zugewiesen, auf der die Erfüllung dieses Kriteriums gemessen wird. Soweit wie möglich sollten diese Kriterien objektiv und automatisch messbar sein, z.B. in Euro pro Jahr oder als prozentuelle Abweichung von der ursprünglichen Auflösung eines Bildes in Bildpunkten. In manchen Fällen müssen jedoch (semi-) subjektive Ordinalskalen zum Zuge kommen. Ein Beispiel dafür ist die Akzeptanz oder der Grad der Offenheit und Standardisierung eines Dateiformates.

Der erstellte Baum ist unabhängig von den betrachteten Alternativen; er dokumentiert die individuellen Anforderungen einer Institution oder Person in Bezug auf die langfristige Archivierung einer bestimmten Kollektion digitaler Objekte. Typischerweise enthalten entsprechende Bäume zwischen 20 und 150 Kriterien auf 3 bis 5 Ebenen. Die Anzahl der Kriterien hängt vor allem von der Art der Objekte ab – je komplexer die Inhalte sind, die in den Objekten abgebildet sind, desto aufwändiger ist die Erstellung des Baumes. Diese Komplexität spiegelt sich dann auch in der Anzahl der Kriterien wider.

Alternativen definieren („Define alternatives")

Dieser Schritt betrachtet in Frage kommende Alternativen, wie beispielsweise Migration (vgl. Kapitel 12.2) oder Emulation (vgl. Kapitel 12.3). In diesem Schritt werden die verfügbaren Tools für die in Frage kommenden Strategien ausgewählt. Die Alternativen werden in diesem Schritt ausführlich beschrieben: Name der Alternative, Beschreibung der Alternative, Gründe, warum sie gewählt wurde, Konfigurierungsumgebung und Ressourcen, die für die Ausführung und Evaluierung nötig sind. Wichtig sind insbesondere die Versionsnummer eines Tools, die Parameter-Einstellungen, das installierte Betriebssystem, die Schriftarten, Programmbibliotheken etc.

Fortfahren / Abbruch („Go/No-Go")

Unter Berücksichtigung der definierten Anforderungen, der Alternativen und einer Einschätzung der benötigten Ressourcen wird in diesem Schritt entschieden, ob der Prozess der Evaluierung fortgesetzt, abgebrochen oder verschoben werden soll. Außerdem wird entschieden, welche der aufgelisteten Alternativen

evaluiert werden sollen. Pro Alternative wird dokumentiert, weshalb sie in die engere Wahl gekommen ist oder verworfen wird. Beispielsweise kann es sein, dass für eine Alternative Hardware benötigt wird, die in der Anschaffung für die jeweilige Institution von vornherein viel zu teuer ist: Aus Kostengründen kann diese Alternative nicht evaluiert werden. Dieser Grund für die Entscheidung wird dann dokumentiert. Eine weitere Möglichkeit kann sein, dass eine neue Version eines Tools in naher Zukunft verfügbar sein wird. Diese Alternative kann dann in die Liste aufgenommen, die Evaluierung jedoch auf einen späteren Zeitpunkt verschoben werden („*Deferred-go*").

Experiment entwickeln („Develop experiment")

Um reproduzierbare Ergebnisse zu gewährleisten, wird in diesem Schritt ein Entwicklungsplan für jede Alternative spezifiziert, die das Experiment-Umfeld und die Art und Weise der Evaluierung mit einschließt. Dies umfasst die Rechnerumgebung, auf der die Experimente durchgeführt werden, die Konfiguration und das Aufsetzen der Messinstrumente (Zeitmessung etc.). Im Idealfall ist eine standardisierte Test-Umgebung vorhanden.

Experiment durchführen („Run experiment")

Die betrachteten Alternativen werden nun in einem kontrollierten Experiment auf die gewählten Beispielobjekte angewandt. Das heißt, die Objekte werden mit den ausgewählten Tools migriert oder in den jeweiligen Emulatoren geöffnet. Dabei anfallende Fehlermeldungen bzw. Zeitmessungen sowie Ausgaben in Protokolldateien werden erfasst. Auch dieser Schritt kann durch die Verwendung von in zentralen Registries erfassten Tools, die über Webservices standardisiert aufgerufen werden können, drastisch vereinfacht werden.

Experimente evaluieren („Evaluate experiments")

Um festzustellen, zu welchem Grad die Anforderungen im Kriterienbaum von den einzelnen Alternativen erfüllt werden, werden die Ergebnisse der Experimente evaluiert. Hierfür wird jedes einzelne Blatt im Kriterienbaum für jedes Objekt evaluiert. Die Evaluierung kann zum Teil automatisiert durch Analysetools unterstützt werden, welche die signifikanten Eigenschaften der Objekte vor und nach der Anwendung der Tools vergleichen und die Ergebnisse dokumentieren.

Umwandeln/Gleichsetzung der gemessenen Werte („Transform measured values")

Nach der Evaluierung der Kriterien im Kriterienbaum sind diese in unterschiedlichen Skalen (z.B. EURO, Sekunden, Farbe: ja/nein) definiert. Damit die Kriterien vergleichbar und aggregierbar werden, wird pro Skala eine Transformationstabelle spezifiziert, welche die Werte der Messskala auf eine einheitliche Zielskala, den sogenannten Nutzwert abbildet. Die Zielskala ist üblicherweise eine Zahl zwischen 0 und 5, wobei 5 der beste Wert ist, während 0 ein nicht akzeptables Ergebnis darstellt.[27]

Das Kriterium „Proprietäres Dateiformat" mit einer Boolean Skala „Yes/No" kann je nach Szenario unterschiedlich transformiert werden. Bei einer Transformation von „No" auf den Wert „eins" und „Yes" auf den Wert „fünf", wäre ein proprietäres Dateiformat zwar akzeptabel aber niedrig bewertet. Jedoch bei einer Transformation von „No" auf den Wert „null" (und „Yes" auf den Wert „fünf") wäre ein proprietäres Dateiformat ein Ausschlusskriterium für die gesamte Alternative.

Alternative	Total Score Weighted Sum	Total Score Weighted Multiplication
PDF/A (Adobe Acrobat 7 prof.)	4.52	**4.31**
PDF (unchanged)	**4.53**	0.00
TIFF (Document Converter 4.1)	4.26	3.93
EPS (Adobe Acrobat 7 prof.)	4.22	3.99
JPEG 2000 (Adobe Acrobat 7 prof.)	4.17	3.77
RTF (Adobe Acrobat 7 prof.)	3.43	0.00
RTF (ConvertDoc 4.1)	3.38	0.00
TXT (Adobe Acrobat 7 prof.)	3.28	0.00

Abbildung 4: Evaluierungsergebnisse elektronischer Dokumente

Wertigkeiten festlegen („Set importance factors")

Die Kriterien, die im Kriterienbaum festgelegt worden sind, haben nicht alle die gleiche Wertigkeit für den Planenden. In diesem Schritt wird daher eine relative Gewichtung der Kriterien auf allen Ebenen durchgeführt, um der unterschiedlichen Bedeutung der einzelnen Ziele Rechnung zu tragen. Sind beispielsweise für eine Institution die Kosten sehr wichtig, werden sie in der Gewichtung höher gestuft als beispielsweise bestimmte Objekteigenschaften. Eine Institution, die beispielsweise eine sehr große Anzahl an Objekten migrieren muss, wird auf der höchsten Ebene des Kriterienbaums die Prozesseigenschaften etwas höher

27 Becker (2007)

gewichten als die übrigen. Folgende Gewichtung wäre denkbar: Objekteigenschaften (20%), Datensatzeigenschaften (20%), Prozesseigenschaften (40%) und Kosten (20%). Damit haben gute bzw. schlechte Prozesseigenschaften einen größeren Einfluss auf das Endergebnis.

Evaluierungsergebnisse analysieren („Analyse evaluation results")

Im abschließenden Schritt werden die Ergebnisse aller Alternativen berechnet und aggregiert, um eine Kennzahl zu schaffen, die zum Vergleich der Alternativen herangezogen werden kann. Dabei können verschiedene Aggregationsmechanismen verwendet werden. Die wichtigsten Aggregationsmechanismen sind die Aufsummierung und die Multiplikation. Bei der Aufsummierung werden die transformierten Ergebniswerte jeder Alternative mit dem relativen Gewicht des entsprechenden Kriteriums multipliziert und über die Hierarchie des Baumes hinweg aufsummiert. Dadurch ergibt sich auf jeder Ebene eine Kennzahl zwischen null und fünf, die dem Erfüllungsgrad der entsprechenden Anforderung durch die betrachtete Alternative entspricht. Bei der Multiplikation dagegen werden die transformierten Werte mit dem relativen Gewicht potenziert und über die Hierarchie des Baumes hinweg multipliziert. Wiederum ergibt sich auf jeder Ebene eine Kennzahl zwischen null und fünf. Der wesentliche Unterschied zur Aufsummierung besteht darin, dass ein einzelnes nicht-akzeptiertes Kriterium zu einem Totalausfall der Alternative führt, da durch die Multiplikation der Wert „null" bis in den Wurzelknoten durchschlägt. Das Ergebnis sind aggregierte Ergebniswerte für jeden Teilbaum des Kriterienbaumes und für jede Alternative. Eine erste Reihung der Alternativen kann auf den aufsummierten und multiplizierten Kennzahlen geschehen. Abbildung 4 zeigt die Bewertung von verschiedenen Alternativen mit Hilfe der zwei Aggregationsmethoden „Gewichtete Summe" und „Gewichtete Multiplikation". Der Hauptunterschied dieser zwei Aggregationsmethoden liegt in der Einflussnahme von nicht erfüllten Kriterien auf das Bewertungsergebnis der Alternative. Bei der Multiplikation scheiden Alternativen aus, d.h. sie werden mit 0 bewertet, falls ein oder mehrere Mindestkriterien nicht erfüllt werden. Die Alternativen RTF und TXT scheiden beispielsweise aus, weil sie große Nachteile in der Erhaltung der Struktur des Dokuments aufweisen. Die Alternative PDF („unchanged") scheidet bei der Aggregationsmethode Multiplikation aus, da das essentielle Kriterium der Verhinderung von eingebetteten Skripten nicht erfüllt wird. Bei Aufsummierung wird die Alternative PDF („unchanged") mit 4.53 knapp am höchsten bewertet, da nicht erfüllte Mindestkriterien kein Ausscheiden der Alternative verursachen, sondern normal in die Berechnung ein-

fließen. Unter Berücksichtigung der Ergebnisse der beiden Aggregationsmethoden kann eine genaue Analyse der Stärken und Schwächen jeder Alternative durchgeführt werden.

Das Ergebnis dieses Planungsprozesses ist eine konzise, objektive und dokumentierte Reihung in Frage kommender Alternativen für ein betrachtetes Archivierungsproblem unter Berücksichtigung der spezifischen situationsbedingten Anforderungen. Welche Lösung tatsächlich umgesetzt wird, hängt von den begleitenden Umständen ab. Aus der Nutzwertanalyse lässt sich jedoch eine klare Empfehlung ableiten, die mit direkt sichtbaren Argumenten hinterlegt und sorgfältig abgewogen ist und sich daher sehr gut als Entscheidungsgrundlage eignet. Durch die Darstellung sowohl allgemeiner als auch detaillierter Ergebniszahlen aus standardisierten und reproduzierbaren Testbedingungen wird eine solide Basis geschaffen, auf der wohlüberlegte und dokumentierte Entscheidungen getroffen werden können.

Das Planungstool Plato

Das EU-Projekt PLANETS entwickelt eine verteilte, serviceorientierte Architektur mit anwendbaren Services und Tools für die digitale Langzeitarchivierung[28]. Plato (PLANETS Preservation Planning Tool) (vgl. Kapitel 13.2) ist ein in PLANETS entwickeltes Planungstool, das den oben beschriebenen, in drei Phasen unterteilten Workflow implementiert und zusätzlich externe Services integriert, um den Prozess zu automatisieren.[29]

Eines dieser Services ist DROID (Digital Record Object Identification) von den National Archives UK. Damit kann automatisch die Bezeichnung des Dateiformats, die Version, der MIME-Type (Multipurpose Internet Mail Extensions) und der PUID (PRONOM Persistent Unique Identifier) ermittelt werden. Ein weiteres integriertes Service ist die Beschreibung des digitalen Objektes im XCDL-Format. Dieses Service wurde von der Universität Köln entwickelt und wandelt die ausgewählten Objekte in ein XCDL-Format um, welches für die spätere Evaluierung notwendig ist [5]. Zudem integriert Plato mehrere Registries, aus denen zu den Beispielobjekten passende Erhaltungsmaßnahmen ausgewählt und automatisch auf die Beispielobjekte angewendet werden können. Bestimmte Objekteigenschaften können automatisch gemessen und evaluiert werden.

Durch die Zuhilfenahme von frei verfügbaren Frameworks wie z.B. Java Server Faces und AJAX wurde Plato als eine J2EE-Web-Applikation entwickelt,

28 Becker (2008b)

29 Becker (2008a) Strodl, (2007)

die frei verfügbar für Planungsvorhaben zur digitalen Langzeitarchivierung genutzt werden kann. [30]

Zusammenfassung

In diesem Kapitel wurde der Planets Workflow zur Planung digitaler Langzeitarchivierungsvorhaben vorgestellt. Dieser Workflow ist die konkrete Ausarbeitung der Kerneinheit „Preservation Planning" des mit dem ISO Standard 14721 verabschiedeten OAIS-Modells. Der Workflow erfüllt nach derzeitigem Wissenstand in den entsprechenden Bereichen die Anforderungen von Initiativen zur Zertifizierung und Validierung von vertrauenswürdigen Archiven, insbesondere nach TRAC[31] und dem nestor - Kriterienkatalog für vertrauenswürdige digitale Langzeitarchive[32].

Literaturverzeichnis

Becker, Christoph, Rauber, Andreas (2007): *Langfristige Archivierung digitaler Fotografien*. Wien.

Becker, Christoph et al. (2008a): *Distributed Preservation Services: Integrating Planning and Actions*. In: Christensen-Dalsgaard, Birte et al.: Research and Advanced Technology for Digital Libraries. Proceedings of the 12th European Conference on Digital Libraries (ECDL 2008). Berlin, Heidelberg: Springer-Verlag. S. 25-36.

Becker, Christoph et al. (2008b): *Plato: a service-oriented decision support system for preservation planning*. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. 2008. S. 367-370.

Becker, Christoph et al. (2008c): *A Generic XML Language for Characterising Objects to Support Digital Preservation*. In: Proceedings of the 2008 ACM symposium on Applied computing. 2008. S. 402-406.

CCDS Consultative Committee for Space Data Systems (Hrsg.) (2002): *Reference model for an open archival information system (OAIS)* / Consultative Committee for Space Data Systems. public.ccsds.org/publications/archive/650x0b1.pdf

Farquhar, Adam., Hockx-Yu, Helen (2007) *Planets: Integrated services for digital preservation*. In: International Journal of Digital Curation, 2. (2007). S. 88-99.

30 http://www.ifs.tuwien.ac.at/dp/plato

31 OCLC (2007)

32 nestor (2006)

National Library of Australia, Unesco. Information Society Division (Hrsg.) (2005): *Guidelines for the preservation of digital heritage. Prepared by the National Library of Australia*. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (Hrsg.) (2006): *Kriterienkatalog vertrauenswürdige digitale Langzeiarchive*. Version 2. (nestor-Materialien 8). Frankfurt am Main: nestor. www.langzeitarchivierung.de/downloads/mat/**nestor**_mat_08.pdf

OCLC Online Computer Library Center, CRL The Center for Research Libraries (Hrsg.) (2007): *Trustworthy Repositories Audit & Certification* (TRAC): Criteria and Checklist. Chicago, Dublin: Center for Research Libraries, OCLC Online Computer Library Center. http://www.crl.edu/PDF/trac.pdf

Rauch, Carl, Rauber, Andreas (2004): *Preserving digital media: Towards a preservation solution evaluation metric*. In: Chen, Zhaoneng et al.: Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004). Berlin: Springer. S. 203-212.

Strodl, Stephan et al. (2006): *The DELOS Testbed for Choosing a Digital Preservation Strategy*. In: Lecture Notes in Computer Science: Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL 2006). Berlin, Heidelberg: Springer. S. 323-332.

Strodl, Stephan et al. (2007): *How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure*. In: Proceedings of the ACM IEEE Joint Conference on Digital Libraries. 2007. S. 29 - 38.