Reinhard Altenhöner, Alfred Kranstedt

SHAMAN

Sustaining Heritage Access through Multivalent Archiving

SHAMAN als Verbundprojekt

Im März dieses Jahres hat das Projekt SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) seine Arbeit aufgenommen. Dieses thematisch im Bereich der Langzeitarchivierung elektronischer Dokumente angesiedelte Verbundprojekt wird von der EU-Kommission mit 8,4 Mio. Euro gefördert und ist auf vier Jahre angelegt. SHAMAN wird von einem Konsortium aus 18 Institutionen und Unternehmen aus acht europäischen Ländern getragen[1]. Innerhalb dieses Konsortiums übernimmt die Deutsche Nationalbibliothek (DNB), organisatorisch vertreten durch die Abteilung Informationstechnik, die Verantwortung für die Leitung und Durchführung des Arbeitspakets »Dokumenterzeugung, Archivierung, Zugriff und Nachnutzung im Kontext von Gedächtnisorganisationen für wissenschaftliche und behördliche Sammlungen«.

Die Zielsetzung von SHAMAN besteht in der Schaffung von übergreifenden Rahmenbedingungen für die Entwicklung digitaler Archivierungssysteme der nächsten Generation. Auf der Basis einer Analyse bestehender Systeme und institutioneller Ansätze, Technologien und Archivierungsprozesse wird die Entwicklung eines umfassenden, letztlich internationalen Ansatzes für eine vernetzte Archivierungsinfrastruktur angestrebt. Ausgehend vom Reference Model for an Open Archival Information System (OAIS)[2] soll ein offenes und erweiterbares »Digital Preservation Framework« entstehen, das alle für die Langzeitarchivierung notwendigen Komponenten, Dienste, Schnittstellen und Spezifikationen im Rahmen einheitlicher und umfassender Standards nachnutzbar definiert. Einheitliche Schnittstellen und anerkannte sowie langfristig stabile Standards bilden die Grundlage für eine schrittweise Vernetzung dezentral vorhandener oder entstehender Archivsysteme. Mit der Vernetzung dieser Systeme unter Nutzung so genannter GRID-Technologien wird der Aufbau einer verteilten Archivierungsinfrastruktur betrieben, die eine kooperative, arbeitsteilige und effiziente Bewältigung der ressourcenintensiven und komplexen Aufgaben der Langzeitarchivierung ermöglichen wird[3]. Dies umfasst nicht nur den reinen Datenaustausch zwischen Institutionen und deren jeweiligen Archivsystemen, sondern auch die kooperative Realisierung und Nutzung standardisierter Dienstleistungen auf der Basis dieser Daten sowie ein verbindliches Zugriffs- und Rechtemanagement auf der Grundlage der jeweils gültigen gesetzlichen Bestimmungen.

Nutzung von GRID-Technologien

Die in SHAMAN entwickelten Konzepte, Technologien und Dienste werden prototypisch implementiert und in Testumgebungen und Praxisszenarien evaluiert. SHAMAN wird also nicht nur einen theoretischen Rahmen, sondern auch konkrete Infrastrukturkomponenten, exemplarische Dienste und prototypische Software liefern. Die DNB übernimmt dabei die Verantwortung für den Aufbau und Test eines SHAMAN-Prototypen speziell für Gedächtnisorganisationen. Sie setzt so ihren Anspruch um, zumindest auf der nationalen Ebene eine der führenden Einrichtungen auf dem Gebiet der Langzeitarchivierung und -verfügbarkeit unter den Gedächtnisorganisationen zu sein. Dazu gehört auch, sich bereits heute mit den technischen Anforderungen an zukünftige Systeme und eine kooperativ verteilte Infrastruktur zu beschäftigen.

SHAMAN-Prototyp

Internationaler Ansatz für vernetzte Architekturstruktur

Die Arbeit im SHAMAN-Projekt profitiert von den vielfältigen Vorarbeiten und Erfahrungen der beteiligten Partner aus den Arbeitsfeldern digitale Bibliotheken, nachhaltige Archivtechnologien und GRID-Infrastrukturen. Die DNB wird Ergebnisse aus dem Langzeitarchivierungsprojekt kopal[4] sowie dem Kompetenznetzwerk nestor[5] einbringen. Mit dem gemeinsam mit der Staats- und Universitätsbibliothek Göttingen (SUB), der Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG) und IBM entwickelten kopal-System steht erstmalig in Deutschland, aber auch in der internationalen Perspektive ein bewusst kooperativ

kopal-System als Ausgangspunkt

betriebenes Archivierungssystem für elektronische Dokumente zur Verfügung, dessen Komponenten als Ausgangspunkt für die weiteren Entwicklungsschritte in SHAMAN herangezogen werden können. Die gemeinsame Nutzung eines (mandantenfähigen) Archiv-Backends, lokalisiert bei einem Vertragspartner, ist ein Beispiel für den Erfolg eines kooperativen Ansatzes in der Langzeitarchivierung. Der SHAMAN-Ansatz einer GRID-basierten Infrastruktur für die Langzeitarchivierung kann als konsequente Weiterentwicklung dieses Gedankens angesehen werden.

Die im SHAMAN-Projekt geplante Vernetzung von Archivierungssystemen hat nicht vorrangig zum Ziel, allen Teilnehmern den Zugriff auf alle Daten zu ermöglichen, sondern vielmehr die effiziente Bewältigung gemeinsamer Aufgaben unter Nutzung verteilter Ressourcen. So werden im Rahmen der Pflege des archivierten Datenmaterials zukünftig vielfältige und zum Teil sehr umfangreiche Arbeiten anfallen, die alle archivierenden Institutionen gleichermaßen betreffen, z. B. die Überwachung obsolet werdender Dateiformate oder die Migration großer Datenbestände. Eine kooperative und arbeitsteilige Bearbeitung auf der Basis verbindlicher Vereinbarungen und Standards verspricht hierbei große Synergieeffekte und eine zusätzliche Qualitätssicherung. Mit dem kopal-Projekt wurden bereits wichtige Voraussetzungen geschaffen, die aber auch für andere Nutzungsszenarien zugänglich gemacht und gleichzeitig erweitert werden müssen.

Nutzung verteilter Ressourcen

Der Aufbau einer gemeinsamen Archivierungsinfrastruktur über Instituts- und Ländergrenzen hinweg umfasst erheblich mehr Aspekte als die reine Steuerung von Datenströmen. So sind Standards für Aufbau und Struktur der digitalen Archivpakete einschließlich einheitlicher Metadatenschemata zu definieren. Eine gemeinsame Schnittmenge der notwendigen Arbeitsabläufe (Workflows) für die Erstellung von Archivpaketen (Ingest) einerseits und der Zugriff (Access) auf diese andererseits ist zu identifizieren und zu formalisieren. Im Hinblick auf eine nachhaltige Bewahrung und Verfügbarkeit digitaler Objekte ist zu klären, welche Informationen über diese Objekte und ihren Kontext erfasst werden müssen und welche Strategien zukünftig eine verlässliche Interpretation und Darstellung der Objekte auf der Basis dieser Informationen sicherstellen.

Klärung zukünftiger Strategien

Um das ehrgeizige Ziel einer vernetzten Archivinfrastruktur auf der Basis eines umfassenden Digital Preservation Environment Framework zu erreichen, hat SHAMAN vielfältige Einzelaufgaben zu bearbeiten. Dazu gehören u. a.:

- Die Identifikation repräsentativer Information über die zu archivierenden Objekte. Dies umfasst nicht nur Informationen zu Inhalt, Struktur und Format digitaler Objekte und ihre Speicherung in einheitlichen Metadaten. Darüber hinaus soll geklärt werden, welche Kontextinformationen über das Objekt und insbesondere die Archivierungsumgebung selbst notwendig sind, um eine Rekonstruktion und Anwendung durch zukünftige Nutzer sicherstellen zu können.
- Die Identifikation und Formalisierung der für die Archivierung notwendigen Prozesse. Hinzu kommen erforderliche Prozesse und geeignete Verfahren zur Verwaltung von Archivsystemen.
- Der Aufbau einer GRID-Infrastruktur. Aufbauend auf einem geeigneten GRID-Framework müssen technische Verfahren für die Abbildung von Standardprozessen und Workflows in diesem Framework entwickelt werden.
- Die Entwicklung von Strategien der Langzeiterhaltung digitaler Objekte, die eine zukünftige Interpretierbarkeit und Darstellung der Objekte sicherstellen. Mögliche Strategien umfassen Datenmigrationen, Systememulationen aber auch Ansätze auf der Basis universaler virtueller Maschinen.

Was versprechen sich die SHAMAN-Projektpartner vom Einsatz der GRID-Technologie?

GRID-Technologien finden seit einigen Jahren gerade für die Bewältigung komplexer, ressourcenintensiver Aufgaben zunehmend Beachtung. Unter einem Computational GRID (in Analogie zum Power Grid = Stromnetz) versteht man eine Software-Infrastruktur, welche die gemeinsame Nutzung verteilter Ressourcen (Software, Daten, Speicher, Rechenleistung etc.) ermöglicht. Im Sinne einer effizienten Ressourcenauslastung und eines Interessenausgleichs zwischen den beteiligten Part-

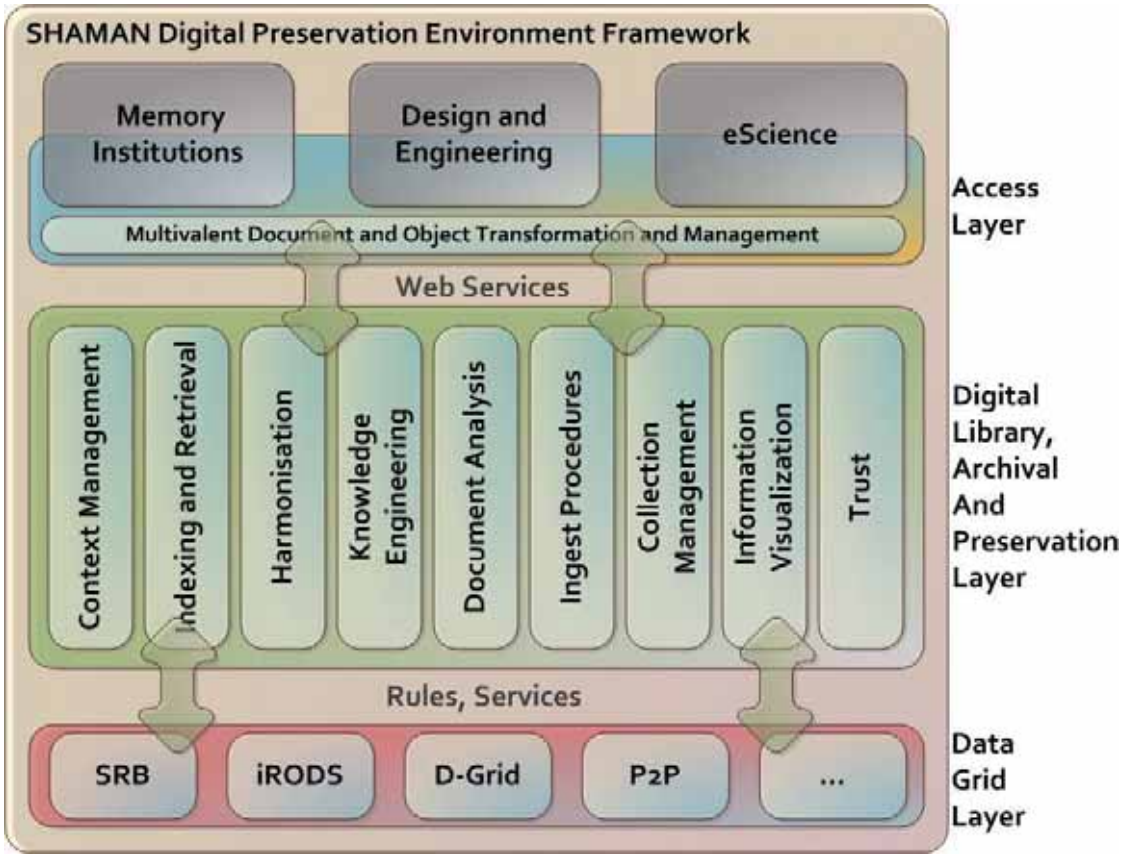
GRID-Technologien

nern erfolgt diese gemeinsame Nutzung im Rahmen kontrollierter Abläufe. Anbieter und Nutzer der Ressourcen einigen sich auf verbindliche Richtlinien, welche Ressourcen geteilt werden, wem die Nutzung erlaubt ist und unter welchen Bedingungen diese erfolgt. Diese Richtlinien werden als Teil der GRID Infrastruktur implementiert und verwaltet. Man kann sich die GRID Software als eine Schicht zwischen den angebotenen Ressourcen und den Nutzern (Personen/Institutionen wie auch Systeme) vorstellen, welche verteilt vorliegende Ressourcen koordiniert, um wiederum komplexe Dienstleistungen zu ermöglichen und dabei offene, standardisierte Protokolle und Schnittstellen verwendet. Auf nationaler wie internationaler Ebene gibt es seit einigen Jahren vielfältige Bestrebungen GRID Infrastrukturen aufzubauen. Genannt sei an dieser Stelle nur die D-GRID Initiative der Bundesregierung[6], welche seit 2005 in mehreren Großprojekten die Entwicklung von GRID-Basistechnologie sowie ihre Anwendung in verschiedenen wissenschaftlichen Kontexten fördert.

SHAMAN wird diese GRID-Technologie auch für die Vernetzung von Langzeitarchivierungssystemen nutzbar machen. Hierbei können die Projektpartner auf einschlägige Vorarbeiten zurückgreifen[7].

Im Rahmen von SHAMAN werden verschiedene GRID-Software-Frameworks auf ihre Eignung für die Langzeitarchivierung evaluiert. Hierbei stehen vor allem Verfahren im Fokus, die eine komfortable und flexible Implementierung von Archivierungs-Workflows und Diensten im Rahmen des Datenzugriffs und der Datenerhaltung erlauben. Ausgangspunkt wird zunächst die unter Federführung des San Diego Supercomputer Center SDSC entwickelte Software iRODS[8] sein, die wiederum auf der Daten-GRID-Software Storage Ressource Broker (SRB) aufsetzt. Sie ermöglicht die regelbasierte Implementierung von Standardoperationen und Dienstleistungen an jedem einzelnen Knoten des GRID-Netzwerkes. Dienste werden als (Micro-) Services realisiert, ihre Anwendbarkeit über Regeln konfiguriert, die von einer so genannten Rule-Engine konsistent auswertet werden. Mit diesem Verfahren lässt sich eine wachsende Sammlung von

iRODS-Software



Architekturschema des SHAMAN - Digital Preservation Environment Framework

Diensten realisieren, die einzelne Aufgaben der Archivierung kapseln und über so genannte Web Services verschiedenen Institutionen zur Verfügung stehen.

Um einen möglichst fundierten Überblick über die relevanten Dienste zu bekommen, wird SHAMAN in seinen ersten Arbeitspaketen die notwendigen Workflows für den Zugriff auf Archivsysteme (Ingest und Retrieval) wie auch geeignete Routinen für den Erhalt des Datenmaterials (Preservation) weiter vereinheitlichen und formalisieren. Hierfür werden typische Nutzungsszenarien ermittelt und ein Katalog von Anforderungen an Langzeitarchivierungssysteme definiert. Diese Arbeiten fließen in eine SHAMAN-Referenzarchitektur auf der Basis des SHAMAN Digital Preservation Framework ein, s. Abb. In dieser Referenzarchitektur definieren die auf Aufgaben der Langzeitarchivierung spezialisierten Dienste einen Service-Layer, der über dem Daten-Layer angesiedelt ist.

Anforderungskriterien definieren

Auf einer Daten-Grid-Schicht werden verschiedene verteilt lokalisierte Dienste realisiert, die von unterschiedlichen Institutionen gemeinsam über Web Services genutzt werden.

Im Hinblick auf eine langfristige Interpretierbarkeit des gespeicherten Datenmaterials sollen im Rahmen von SHAMAN zwei grundsätzliche Fragen geklärt werden:

Grundsatzfragen

- Welche Informationen sind unerlässlich für die spätere Rekonstruktion der archivierten Objekte und müssen deshalb dem Objekt bei der Archivierung mitgegeben werden?
- Welche Strategien und Verfahren der Datenpflege und des Datenzugriffs stellen zukünftig auch die Interpretierbarkeit und Darstellbarkeit möglicherweise obsolet gewordener Datenformate sicher - effizient, komfortabel und möglichst verlustfrei?

Während Strategien und Verfahren im Laufe der Jahre zumindest bedingt dem sich weiterentwickelnden Stand der Technik angepasst werden können, sollte die erste Frage schon im Vorfeld der Archivierung geklärt werden, um mögliche Informationsverluste bzw. aufwändige Nacharchivierungen zu vermeiden.

Im Gegensatz zu früheren Projekten will SHAMAN bei dieser Frage verstärkt den Kontext von digitalen Archivalien in den Blick nehmen. Damit sind nicht nur Informationen über Dateiformate und technische Mindestanforderungen für das Darstellen eines Objektes gemeint, sondern auch Informationen über das Archivsystem selbst und die Prozesse, die das Objekt bei der Archivierung durchlaufen hat. Der Grundgedanke hinter dieser Ausweitung des Kontextbegriffes ist, das Archivierungsprozesse die Archivalien, insbesondere die in ihnen enthaltenen und erst während der Archivierung generierten Meta-Informationen prägen, aber selber einem historischen Wandel unterworfen sind. Im kopal-System werden beispielsweise technische Metadaten automatisiert unter Nutzung der Software JHOVE[9] generiert. Was dabei erfasst wird, hängt davon ab, nach welchen Informationen die Software in den Originaldateien sucht, und nach welchen Kriterien sie die Informationen sortiert. Diese Kriterien könnten sich aber zukünftig ändern.

Digitale Archivalien im Fokus

In der Konsequenz würde dieser Gedanke bedeuten, dass sich das Archivsystem selbst archivieren müsste, praktisch also seine Kontextabhängigkeit bestimmen und für zukünftige Anwender deklarieren müsste. Da dies nicht umsetzbar ist, setzt sich SHAMAN das Ziel, im Rahmen des anvisierten Framework of Preservation fundierte Vorschläge für die automatisierte Erfassung und Speicherung eines (gegenüber bisherigen Ansätzen erweiterten und daher nun ausreichenden) Minimalsatzes von Kontextinformationen vorzulegen. Auch diese zusätzlichen Kontextinformationen sollen formalisiert, damit maschinell auswertbar und über geeignete Metadatenschemata mit den Archivobjekten verknüpft werden.

Auf dem Gebiet der Metadatenschemata kann die DNB umfangreiche Vorarbeiten u. a. aus dem Projekt kopal einbringen. Mit dem Universal Object Format (UOF)[10] basierend auf dem Metadatenformat METS[11] liegt ein erprobtes Format vor, das flexibel erweiterbar ist und sich somit auch für die Einbindung weiterer Kontextinformationen eignet.

Universal Objekt Format

Im Hinblick auf die langfristige Interpretierbarkeit digitaler Archivalien hat sich das Projekt SHAMAN darüber hinaus vorgenommen, verschiedene Strategien und Lösungsansätze wie die Migration obsoleter Datenformate einerseits oder die Emulation veralteter Software- und Hardwareumgebungen andererseits an konkreten Fallszenarien zu evaluieren. Dabei soll auch die Transformation und

Darstellung elektronischer Dokumente auf der Basis universaler Objektmodelle getestet werden. Dieser Ansatz, zu dem unter dem Namen »Multivalent« erste Implementationen existieren[12], versucht die Vorteile beider Strategien in einem einheitlichen Verfahren zusammenzuführen.

Multivalent

Multivalent begreift sich als eine Weiterentwicklung der ursprünglich von IBM maßgeblich mitentwickelten Idee einer Universalen Virtuellen Maschine UVM[13]. Hierunter ist eine Abstraktionsschicht zu verstehen, die beliebige Hardware bzw. Betriebssysteme auf ein einheitliches (virtuelles) Computermodell abbildet. Software, die für eine UVM geschrieben wird, ist somit auf jedem System lauffähig, vorausgesetzt es existiert eine UVM-Implementierung für dieses System. Multivalent definiert zusätzlich eine weitere Abstraktionsschicht mit einem universalen Objektmodell und stellt Werkzeuge für den Zugriff und die Darstellung von Instanzen dieses Modells zur Verfügung[14]. So genannte Media-Adapter implementieren die Abbildungen von den bestehenden Dateiformaten auf das universelle Objektformat.

Welche Perspektiven ergeben sich mittelfristig aus dem SHAMAN-Projekt für die DNB?

Perspektiven für DNB

Zunächst bietet sich für die DNB die Chance, die Erfahrungen wie auch die bisher entwickelten Verfahren und Systembausteine für die Langzeitarchivierung in einen größeren internationalen Zusammenhang einzubringen und damit eine breite Nachnutzung zu ermöglichen. Die notwendige Weiterentwicklung dieser Verfahren und Systeme in enger internationaler Abstimmung umfasst eine zusätzliche Qualitätssicherung und verhindert divergierende Entwicklungsstränge an anderen Orten. Gleichzeitig wird damit sichergestellt, dass die Langzeitarchivierungsinfrastruktur der DNB auch zukünftig analog zu internationalen Entwicklungen ausgebaut wird.

Der prototypische Aufbau von GRID-Infrastrukturen im EU-geförderten Projekt SHAMAN erlaubt das Sammeln wertvoller Erfahrungen mit dieser Schlüsseltechnologie und die eingehende Erprobung einzelner Systembausteine. In ihrem konzeptuellen Ansatz erscheint diese Technologie geeignet für eine schrittweise und kontrollierte Vernetzung der Langzeitarchivierungssysteme der DNB mit vergleichbaren Systemen anderer Institutionen – auch und gerade im Kontext von Anforderungen aus Industrie und Privatwirtschaft wie Versicherungen und Banken.

Die Einbindung in eine gemeinsame Infrastruktur erlaubt die kooperative Nutzung von Ressourcen bzw. die Auslagerung aufwändiger Standardprozesse und bietet somit ein hohes Potenzial an Synergien. Konkrete und in absehbarer Zeit umgesetzte Beispiele dafür sind der gemeinsame Aufbau einer Format-Registry, die Beobachtung der Formatentwicklung oder die Bewältigung umfangreicher Formatmigrationen.

Damit werden die Voraussetzungen geschaffen, die im Rahmen des gesetzlichen Auftrages der DNB notwendige dauerhafte Archivierung elektronischer Publikationen aller Art mit ihrem immensen Datenvolumen und heterogener Struktur mit

einem vertretbaren Aufwand zu ermöglichen.
Ganz konkret können aber auch schon während der Projektlaufzeit Softwarekomponenten der Projektpartner in einer »Laborumgebung« auf ihre Verein-barkeit mit dem bestehenden System und ihre Eignung für die Aufgaben der DNB evaluiert und gegebenenfalls in die bestehenden Systeme eingepflegt werden.

Anmerkungen

1 Weitere Informationen über das Konsortium von SHAMAN sowie seine Aufgaben finden sich auf der offiziellen Webseite des Projektes unter: <http://www.shaman-ip.eu/>

2 OAIS – Reference Model for an Open Archival Information System. Ein ursprünglich von der NASA initiiertes Referenzmodell für die Langzeitarchivierung elektronischer Dokumente. OAIS hat den Status eines ISO-Standards erreicht (ISO-Standard 14721:2003). Öffentlich verfügbar unter: <http://public.ccsds.org/publications/archive/650x0b1.pdf>

3 Vgl. hierzu auch: Altenhöner, Reinhard; Kett, Jürgen: Grid. In: Dialog mit Bibliotheken, 18 (2006) 1, S. 35 – 41.

4 Über kopal ist an dieser Stelle bereits ausführlich berichtet worden: Wollschläger, Thomas: Kopal goes lieve. In: Dialog mit Bibliotheken, 19 (2007) 2, S. 17 – 22.
Altenhöner, Reinhard: Das kopal-Projekt des Bundesministeriums für Bildung und Forschung. In: Dialog mit Bibliotheken, 17 (2005) 1, S. 21 – 34.

5 NESTOR – Kompetenznetzwerk Langzeitarchivierung, siehe: <http://www.langzeitarchivierung.de>

6 D-GRID Initiative, gefördert durch die Bundesregierung seit 2005: <http://www.d-grid.de/>

7 Exemplarisch genannt seien Erfahrungen mit den Softwareumgebungen iRODS und SRB (Universität Liverpool, Fernuniversität Hagen), s. folgende Anm. sowie die Beteiligung der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) am Projekt Text-GRID.

8 iRODS – Daten-GRID Software auf der Basis des »Storage Ressource Broker« (SRB) entwickelt vom San Diego Supercomputer Center: <https://www.irods.org>

9 JHOVE – JSTOR / Harvard Object Validation Environment. Software für die automatische Identifikation, Validierung und Charakterisierung von Formaten digitaler Objekte. Siehe: <http://hul.harvard.edu/jhove/>

10 UOF – Universal Object Format: <http://kopal.langzeitarchivierung.de/downloads/kopal_Universelles_Objektformat.pdf>

11 METS – Metadata Encoding & Transmission Standard: <http://www.loc.gov/standards/mets/>

12 <http://multivalent.sourceforge.net/>

13 UVM – Universal Virtual Machine, oder auch Universal Virtual Computer UVC, ein speziell auch für die Langzeitarchivierung entwickeltes Konzept einer universalen Computerplattform, siehe: <http://www-05.ibm.com/nl/dias/resource/uvc.pdf>

14 Da zum jetzigen Zeitpunkt kaum ausreichende Implementierungen der UVM existieren, haben sich die Anbieter von Multivalent für die Java Virtual Machine, die sie als eine Approximation der UVM ansehen, als Plattform für ihre Software entschieden.