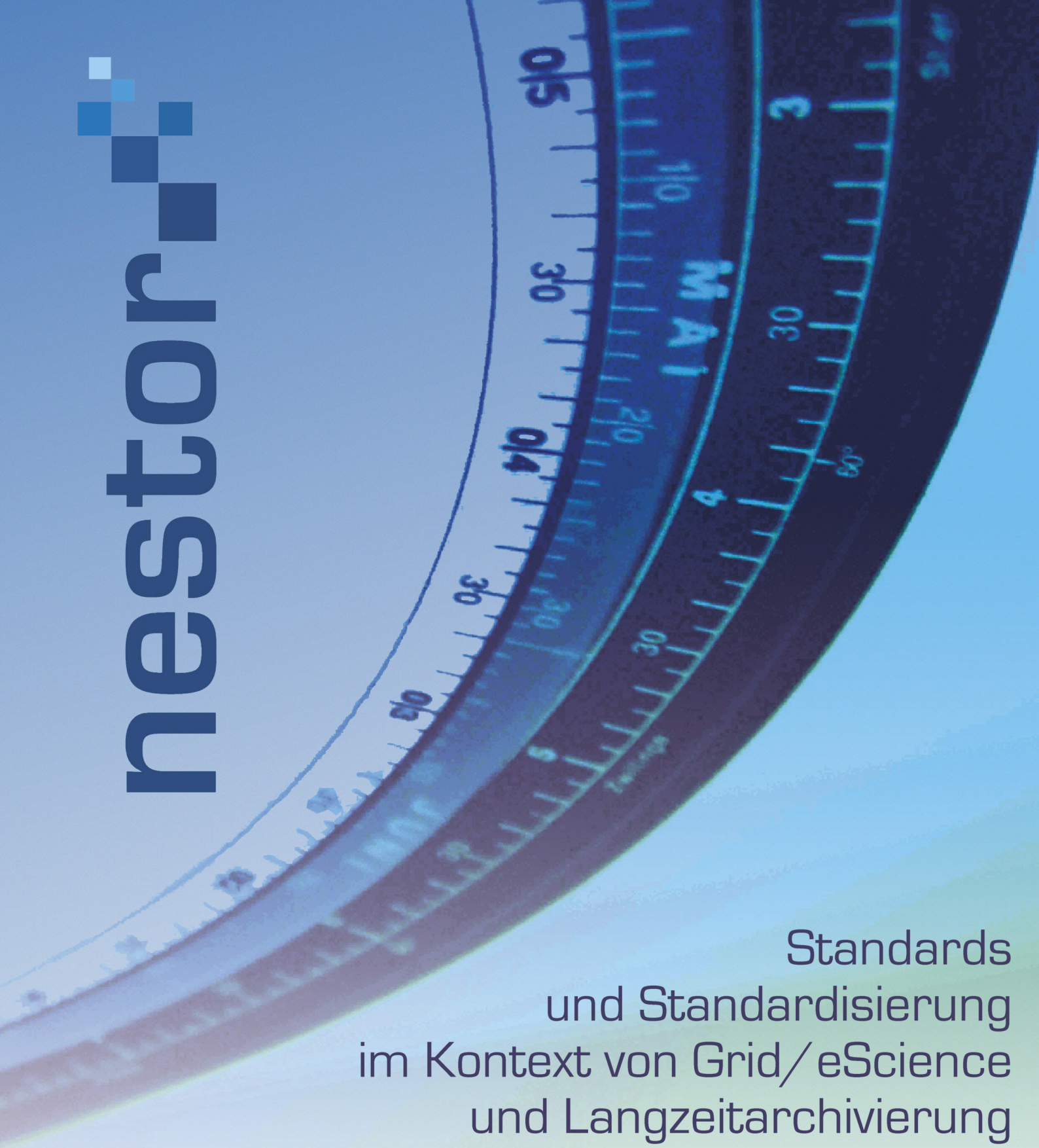

nestor

Standards und Standardisierung im Kontext von Grid/eScience und Langzeitarchivierung

Uwe M. Borghoff und Peter Rödig

Universität der Bundeswehr München
Fakultät für Informatik - Institut für Softwaretechnologie

nestor-materialien 11

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Standards und Standardisierung in Kontext von Grid/eScience und Langzeitarchivierung

Uwe M. Borghoff
Peter Rödig

Universität der Bundeswehr München

nestor-materialien 11

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Herausgegeben von

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

nestor - Network of Expertise in Long-Term Storage of Digital Resources

http://www.langzeitarchivierung.de

Projektpartner:

Bayerische Staatsbibliothek, München
Bundesarchiv
Deutsche Nationalbibliothek (Projektleitung)
FernUniversität in Hagen
Humboldt-Universität zu Berlin - Computer- und Medienservice / Universitätsbibliothek
Institut für Museumsforschung, Berlin
Niedersächsische Staats- und Universitätsbibliothek, Göttingen

Betreuer dieser Veröffentlichung:

Bayerische Staatsbibliothek München
Dr. Astrid Schoger
Tobias Beinert
Dr. Thomas Wolf-Klostermann

URN: urn:nbn:de:0008-2009012103
http://nbn-resolving.de/urn:nbn:de:0008-2009012103

Grid-Technologie und Langzeitarchivierung in nestor

Die modernen Informationstechnologien haben in allen Lebensbereichen starke Veränderungen bewirkt. Besonders stark beeinflusst sind die Wissenschaften, die auch eine treibende Kraft dieser Entwicklungen sind und immer größere Anforderungen an Rechner, Speicher und IT-Werkzeuge stellen. In neuen Experimenten der Teilchenphysik werden kaum bewältigbare Datenmengen für Tausende von Wissenschaftlern produziert, Klimaforscher berechnen immer detailliertere Modelle des Systems Erde, und die Geisteswissenschaften beginnen riesige digitale Sammlungen von Kulturgütern mit Rechnern zu analysieren.
Die Grid-Technologie zur Aufteilung der Aufgaben auf viele verteilte IT-Ressourcen ist ein Mittel, um den Herausforderungen dieser neuen, als e-Science bezeichneten wissenschaftlichen Arbeitsweise gerecht zu werden.

nestor und Wissenschaftler weltweit haben immer wieder darauf hingewiesen, dass mit der Zunahme der Bedeutung digitaler Daten auch die Notwendigkeit wächst, ihre langfristige Nutzbarkeit zu sichern. Bei der Grid-Technologie ergibt sich die chancenreiche Situation, dass nicht nur wertvolle und zu erhaltende Daten produziert werden, sondern auch Mittel bereit gestellt werden, die für die Herausforderung der Langzeitarchivierung großer und komplexer Datenmengen nutzbar sein können. Die klassischen Gedächtnisorganisationen - wie Bibliotheken, Archive und Museen - und die neuen Gedächtnisorganisationen - wie Daten- und Rechenzentren - können wechselseitig voneinander profitieren.

Um dieses Potenzial auszuloten, hat nestor in seiner zweiten Projektphase eine Arbeitsgruppe mit Fachleuten aus klassischen Gedächtnisinstitutionen und aus e-Science- und grid-engagierten Institutionen initiiert und drei Expertisen in Auftrag gegeben. Diese Expertisen untersuchen den Ist-Stand und die Anforderungen und Ziele für das Zusammenspiel von e-Science-/Grid-Technologie und Langzeitarchivierung unter drei Gesichtspunkten:
Welche Anforderungen gibt es für die Archivierung von Forschungsdaten?
Was sind die möglichen Synergien, die angestrebt werden sollten?
Und auf welche Standards können weitere Arbeiten in diesen Bereich aufgebaut werden und welche sind gegebenenfalls noch zu entwickeln?

Neben der Untersuchung des Standes der Technik, sind einige Projekte der deutschen Grid-Initiative D-Grid befragt worden. nestor wird in seiner Grid-/eScience-Arbeitsgruppe die Ergebnisse der Expertisen aufnehmen und versuchen, eine Landkarte für die weiteren Entwicklungsperspektiven zu zeichnen.

e-Science-/Grid-Technologie und Langzeitarchivierung sind relativ neue Forschungsbereiche, die sich sehr schnell entwickeln. Einzelne Fragen, die von nestor Mitte 2006 formuliert wurden, als die ersten Projekte der deutschen Grid-Initiative D-Grid gerade gestartet waren, stellen sich heute, wo bald schon die dritte Generation von D-Grid-Projekten beginnt, unter den veränderten Bedingungen möglicherweise anders dar.
Die Expertisen müssen daher auch vor ihrem Entstehungshintergrund betrachtet werden. Derzeit liefern sie eine Beschreibung sinnvoller und notwendiger Entwicklungen. Wenn sie in naher Zukunft „veralten“, weil sie zur erfolgreichen Zusammenarbeit von e-Science/Grid und Langzeitarchivierung beigetragen haben, dann haben sie ihren Sinn erfüllt.

Standards und Standardisierung im Kontext von Grid/eScience und Langzeitarchivierung

Universität der Bundeswehr München
Fakultät für Informatik
Institut für Softwaretechnologie

Uwe M. Borghoff
Peter Rödig

Betreuung: Bayerische Staatsbibliothek München
Dr. Astrid Schoger
Tobias Beinert
Dr. Thomas Wolf-Klostermann

Inhaltsverzeichnis

1 Zusammenfassung

Standards sind unverzichtbare Bausteine für das Funktionieren komplexer technischer und organisatorischer Systeme. Neben der Sicherstellung der grundsätzlichen Funktionsfähigkeit liefern Standards einen Beitrag sowohl zur Vertrauenswürdigkeit als auch zur Wirtschaftlichkeit – Eigenschaften, die von einem System zum Langzeiterhalt digitaler Information von Nutzern und Betreibern erwartet werden. Die Standardisierung im Bereich der Langzeitarchivierung unter Berücksichtigung internationaler Aktivitäten voranzutreiben ist eine Aufgabe der nationalen Initiative nestor. Des Weiteren zählen zum Betrachtungsumfang von nestor auch digitale Informationen aus dem Bereich der Naturwissenschaften und Technik. Die dortigen Inhalte, Methoden, Techniken und Rahmenbedingungen unterscheiden sich in weiten Bereichen deutlich von denen anderer Disziplinen. Diese Unterschiede haben einerseits Auswirkungen auf die jeweils verwendeten Standards für die Handhabung von Information und andererseits sind diese Unterschiede bei der Entwicklung neuer Standards für die Langzeitarchivierung zu berücksichtigen. Natürlich sollen auch die Gemeinsamkeiten erkennbar werden. Beispielsweise ist die eindeutige Benennung von digitalen Objekten oder das Auffinden von Information in verteilten Umgebungen, die von unterschiedlichen autonomen organisatorischen Einheiten betrieben werden, zumindest konzeptionell, ein gemeinsames Anliegen.

In dieser Expertise war zu untersuchen, wie sich die Anwendungen von Grid-Technologien auf die Standards traditioneller Gedächtnisorganisationen auswirken. Die Studie zeigt auf, wo sich Probleme mit der Verwendung bisheriger Standards ergeben können und wo Handlungsbedarf erkennbar ist. Hochgradige Verteilung und Virtualisierung von Rechen- und Speicherkapazität und die damit verbundenen Anwendungen sind eine neue Herausforderung für traditionelle Gedächtnisorganisationen. Inhalte, Methoden, Techniken und Rahmenbedingungen der Informationsverarbeitung im Bereich eScience unterscheiden sich ebenfalls deutlich von traditionellen Archivierungsszenarien. Zur Bewertung von Standards stützt sich diese Expertise auf die Vorarbeiten des GeoForschungsZentrums Potsdam, die in der nestor-Expertise „Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Daten" dokumentiert sind [Klum2007]. Besondere Anforderungen ergeben sich aus der Größe bzw. aus der hohen Anzahl der digitalen Objekte und ihrer ggf. verteilten Speicherung und Administration. Die Entstehung und Prozessierung der Objekte ist ebenfalls sehr spezifisch, z.B. wegen der Erfassung durch Messgeräte und wegen der von verschiedenen Forschergruppen durchgeführten mehrfachen Prozessierung mittels komplexer Verfahren. Des Weiteren wird teilweise auf die explizite Speicherung von Zwischenergebnissen aus Aufwandsgründen verzichtet, wenn diese Ergebnisse durch eine erneute Berechnung wieder herstellbar sind. Können die Ergebnisse für die Langzeitarchivierung nicht explizit gemacht werden, müssten die Berechnungsvorschriften (z.B. in Form von Software) langfristig aufbewahrt werden. Eine weitere wichtige Besonderheit sind der hohe Interpretationsbedarf der Inhalte digitaler Objekte und ihre semantische Komplexität.

Diese Sachverhalte sind zu analysieren und den Standards der Gedächtnisorganisationen gegenüberzustellen. Die Studie betrachtet schwerpunktmäßig Standards und Standardisierungsorganisationen wie METS (Metadata Encoding and Transmission Standard) von LOC (Library of Congress), PREMIS (Data Dictionary for Preservation Metadata) von OCLC (Online Computer Library Center), UOF (Universelles Objektformat) und NBN (National Bibliographic Number) von DNB (Deutsche Nationalbibliothek), Handle System von

CNRI (Corporation for National Research Initiatives), DOI von IDF (International DOI Foundation), PRONOM von TNA (The National Archive), GDFR (Global Digital Format Registry) von HUL (Harvard University Library).

Konsequenzen für die Standardisierung ergeben sich auch aus den Szenarien einer Einbindung von Gedächtnisorganisationen in Grid-basierte eScience-Infrastrukturen, wie sie in der Studie „Synergiepotentiale zwischen Grid- und eScience-Technologien für die digitale Langzeitarchivierung“ der FernUniversität Hagen dargestellt sind [Schi2008]. Hierzu wären die entsprechenden Schnittstellen in Form von Middleware zu realisieren. Wie im Bereich der Langzeitarchivierung hat die Standardisierung noch nicht das gewünschte Maß erreicht. So hat sich insbesondere das OGF (Open Grid Forum) vorgenommen, die Standardisierung im Bereich Grid voranzubringen. Angestrebt wird hierbei die Abstützung auf Standards des Internets und der so genannten serviceorientierten Architekturen. Bedeutende Standardisierungsorganisationen auf diesen Gebieten sind IETF (The Internet Engineering Task Force), OASIS (Organization for the Advancement of Structured Information Standards) und W3C (World Wide Web Consortium). Schließlich werden exemplarisch noch Standards berücksichtigt, die im Bereich D-Grid bereits Anwendung finden bzw. auf die von der nestor-AG „Grid/eScience und LZA“ hingewiesen wurde.

Darüber hinaus sind Potenziale von Grid-Technologien für die Archivierung im traditionellen Umfeld erkennbar, die mit der Einführung entsprechender Middleware genutzt werden könnten. Hier wären dieselben technischen Standards relevant wie bei einer Integration von Grid-basierten eScience-Infrastrukturen.

Dazu ist die Studie wie folgt aufgebaut: Zuerst wird das Konzept der Vertrauenswürdigkeit digitaler Langzeitarchive vorgestellt. Die Kriterien der Vertrauenswürdigkeit erlauben dem weniger technisch orientierten Leser einen leichteren Einstieg sowie einen einfacheren Vergleich mit Konzepten von (High-Level)-Grid-Architekturen und von Standards. Vertrauenswürdigkeit, mit den Prinzipien der Dokumentation, Transparenz und Angemessenheit, ist ein Maßstab für die Qualität eines Archivs, an dem Standards auszurichten sind. Ausgehend vom Prinzip der Vertrauenswürdigkeit wird die Rolle von Standards diskutiert. Entsprechend der Sicht einer Gedächtnisorganisation liegt der Schwerpunkt der Expertise auf dem langfristigen Erhalt von Information. Zur Erläuterung dieser Kernproblematik wird speziell auf das Informationsmodell des ISO-Referenzmodells OAIS (Open Archival Information System) Bezug genommen. Unabhängig von einer konkreten Implementierung muss das Konzept des Informationspaketes realisiert werden; d.h. die Daten sind dauerhaft mit genügend Information zu verknüpfen, um sie langfristig verfügbar und interpretierbar zu machen. Darüber hinaus dient OAIS als Ordnungsschema für die Auswertung der oben genannten nestor-Studien und der bereits genannten Standards. Anforderungen und Standards werden soweit sinnvoll den Modellelementen von OAIS zugeordnet. Anhand dieser Strukturierungen werden schließlich Lücken und Problembereiche der Standardisierung aufgezeigt und Handlungsempfehlungen abgeleitet. Die Aussagen hierzu werden möglichst konkret gefasst. Doch im Rahmen dieser Studie ist eine gewisse Vereinfachung und Generalisierung unvermeidbar. Daher sind die Problembereiche und Handlungsempfehlungen ggf. im jeweiligen, möglicherweise sehr spezifisch technischen und organisatorischen Anwendungskontext einzuordnen und zu vertiefen.

Da die formale Standardisierung in Bezug auf die digitale Langzeitarchivierung noch weitgehend am Anfang steht, wird der Begriff Standards hier im weiteren Sinne verwendet. Die Betrachtung von Standards aus dem Bereich der traditionellen Gedächtnisorganisatio-

nen impliziert nicht, dass die Archivierung auch in diesem Bereich organisatorisch angesiedelt sein muss. Vielmehr geht es entsprechend dem Auftrag von nestor darum, den Erfahrungsaustausch zwischen verschiedenen Organisationen zu fördern, um die Aufgabe der Langzeitarchivierung in einer kooperativen und damit effizienteren und nachhaltigen Art und Weise zu lösen.

2 Vertrauenswürdigkeit als Qualitätsmaßstab für digitale Langzeitarchive

Digitale Langzeitarchive entziehen sich momentan noch einer direkten Bewertung ihrer Qualität. Ursachen hierfür sind u.a. die inhärente Komplexität digitaler Objekte, die zahlreichen offenen Fragen sowie eine Leistungserbringung, die in der Zukunft liegt und im Extremfall keinen Endtermin kennt. Um diesen Herausforderungen zu begegnen, wurde das Konzept der Vertrauenswürdigkeit von digitalen Langzeitarchiven entwickelt. Dabei wird Vertrauenswürdigkeit als die Eigenschaft eines Systems angesehen, gemäß seinen Zielen und Spezifikationen zu operieren – d.h. es tut genau das, was es zu tun vorgibt.

Um die Vertrauenswürdigkeit einer Bewertung zugänglich zu machen, hat die nestor AG „Vertrauenswürdige Archive – Zertifizierung“ einen Kriterienkatalog entwickelt, der einen gesamtheitlichen Blick auf digitale Langzeitarchive wirft; die formulierten Kriterien umfassen Aspekte der Organisation, der zu bewahrenden Information sowie der technischen Infrastruktur und Sicherheit [nestor2008]. Bewertbarkeit setzt ein gemeinsames Verständnis über Grundkonzepte der Langzeitarchivierung voraus. Aus diesem Grund orientiert sich der Kriterienkatalog am OAIS-Referenzmodell, soweit von der AG als sinnvoll erachtet. Die Kriterien sind abstrakt gefasst, um einen längeren Gültigkeitszeitraum sicherzustellen und um die Anwendbarkeit nicht auf spezifische Archivierungsszenarien einzuengen. Für die Anwendung der Kriterien sind einige Grundprinzipien aufgestellt. Dazu zählen die *Dokumentation* von Zielen, Konzepten, Spezifikationen und Implementierungen, um die Schlüssigkeit eines Langzeitarchivs bewerten zu können. Weitere Prinzipien sind *Transparenz*, hier die direkte oder indirekte Zugänglichkeit der Dokumentation, sowie *Angemessenheit*, d.h. dass statt absoluter Maßstäbe die jeweiligen Ziele und Aufgaben eines Archivs als Maß für eine Bewertung heranzuziehen sind.

Durch die Abstraktion schließt der Kriterienkatalog keine Organisationsformen und keine Techniken von vornherein aus. Maßgebend ist die Erfüllbarkeit der Kriterien. In Bezug auf digitale Information adressiert der Kriterienkatalog folgende Eigenschaften: Integrität, Authentizität, Verfügbarkeit, Vertraulichkeit und speziell die langfristige Auffindbarkeit, Identifizierbarkeit und Interpretierbarkeit.

In diesem Zusammenhang spielen Standards eine nützliche Rolle, da sie der Vertrauenswürdigkeit zu Grunde liegende Kriterien zum Thema haben bzw. haben sollten. Die Abstützung der Langzeitarchivierung auf Standards und der Nachweis der konformen Nutzung würde eine Bewertung der Vertrauenswürdigkeit deutlich erleichtern. Offene Standards unterstützen auch das Prinzip der *Dokumentation* und *Transparenz*. Das folgende Kapitel beschreibt die Rolle von Standards etwas näher.

3 Die Rolle von Standards

Bedeutung von Standards

Standards sind unabdingbare Voraussetzung für kompatible und interoperative Systeme aller Art. Standards fördern die Wiederverwendbarkeit und Austauschbarkeit von Komponenten, erleichtern die Bewertbarkeit und Vergleichbarkeit von Produkten und von Dienstleistungen und bieten verlässliche Vorgaben für System- und Produktentwickler. Öffentlich verfügbare und realistisch umsetzbare Vorgaben sind Basis für konkurrierende Implementierungen und somit für einen funktionierenden Markt.

Darüber hinaus waren Standards häufig die Grundlage für vollkommen neue ökonomische und wissenschaftliche Modelle. Ein beeindruckendes Beispiel ist das Protokoll TCP/IP, welches heute praktisch von jedem Rechner und vielen Peripheriegeräten verstanden wird und welches die Basis für vielfältige Dienste und Produkte der verteilten Datenverarbeitung bildet und schließlich mit HTTP und HTML bzw. XML einen neuen Höhepunkt mit großer praktischer Bedeutung gefunden hat.

Gelingt es, die Standardisierung bezüglich der Flexibilisierung der Konfigurierbarkeit von IT-Ressourcen weiterzuentwickeln, lassen sich neue Modelle der Zusammenarbeit wirtschaftlich realisieren. Serviceorientierung und Grid haben die Flexibilisierung zum Konzept.

Die Bedeutung von Standards für das Funktionieren und für die Akzeptanz von Grid-Infrastrukturen wird an mehreren Stellen behandelt.

Ian Foster hebt in seiner Kolumne „What is the Grid? A Three Point Checklist“ in Grid Today [Fost2002] die kritische Rolle von Standards hervor. Neben der Offenheit und Allgemeingültigkeit von Schnittstellen, Protokollen und Policies ist ihre Standardisierung elementar für die Vision Grid, denn Standards erlauben die gemeinsame Nutzung von Ressourcen in einer dynamischen Art und Weise für jeden, der daran interessiert ist. Außerdem ermöglichen Standards die Entwicklung von allgemein nutzbaren Diensten und Werkzeugen.

Des Weiteren wird die Standardisierung in zahlreichen Ausführungen zum Thema Grid aufgegriffen (z.B. in [Ried2005], [OGSA-SCRN 2005], [GTK2005], [Gent2007], [Schu2008]). Die Initiative D-Grid hat eine Gruppe eingerichtet, die sich explizit der Nachhaltigkeit widmet.

Einteilung von Standards

Eine einheitliche Begrifflichkeit bezüglich Standards ist nicht erkennbar. Zahlreiche nationale und internationale Standardisierungsorganisationen bzw. Standardentwicklungsorganisationen, abweichende Rechtssysteme und Kulturen sowie äußerst unterschiedliche Gegenstände der Standardisierung tragen zu einem uneinheitlichen Bild bei.

Als Startpunkt für eine Einteilung werden Ausführungen aus dem Projekt COPRAS (Cooperation Platform for Research and Standards) herangezogen. Standards werden dort einerseits in formale Standards, informale Standards, private Spezifikationen unterteilt, und andererseits in normative Standards und informative Standards. Normative Standards schreiben vor, in welcher Form ein Standardisierungsgegenstand mit etwas übereinstim-

men soll. Informative Standards hingegen liefern nützliche Informationen und Anleitungen. Formale Standards fallen nur unter dem Typ der normativen Standards und werden von einer nationalen (z.B. DIN), regionalen (z.B. CEN[1]) oder internationalen (z.B. ISO[2]) Standardisierungsorganisation (Standards Body) entwickelt und durchlaufen den jeweiligen formalen Anerkennungsprozess. Diese Art wird auch als de jure bezeichnet. Informale Standards können vom Typ normativ oder informativ sein. Sind sie normativ, wird die Bezeichnung technische Spezifikation verwendet und sie werden von einer formalen Standardisierungsorganisation entwickelt oder von einer Standardentwicklungsorganisation (Standard Developer Organization – SDO) wie W3C, IEEE, IETF. Sie beruhen auf Konsens der Mitglieder oder der am Standardisierungsprozess Beteiligten und unterliegen den jeweiligen Anerkennungsprozessen. Ist der Inhalt nur informativ, genügt Konsens, und als Bezeichnung dient Empfehlung (Recommendation) oder Bericht (Report). Private Spezifikationen sind durch eine geschlossene Mitgliedschaft (einzelne Firma, Industriekonsortium) gekennzeichnet und können sowohl normativ sein (als Spezifikation bezeichnet) als auch informativ (als Bericht, Empfehlung, Verhaltensnorm o.ä. bezeichnet).

Auffallend an obiger Einteilung ist die Einschränkung des Begriffs Standard auf formale Standards, die von den meisten SDOs nicht befolgt wird. In Deutschland ist für formale Standards der Begriff Norm als Ergebnis konsensbasierter Normungsarbeit üblich. Neben den „echten" Standards, sehen auch die formalen Standardisierungsorganisationen zusätzlich abgeschwächte Formen vor, um insbesondere Gebieten mit kurzen Innovationszyklen gerecht zu werden. Eine Beschleunigung des Verfahrens kann u.a. durch den (vorläufigen) Verzicht auf einen breiten Konsens erreicht werden. DIN hat hierzu beispielsweise die PAS (Publicly Available Specification) eingeführt, deren Erarbeitungsprozess *Standardisierung* genannt wird.

Die in dieser Expertise betrachteten Standards werden nicht auf eine bestimmte Klasse eingeschränkt; der erkennbare Charakter als Konvention sowie die Schriftlichkeit und Öffentlichkeit war Mindestvoraussetzung für die Einstufung als Standard.

Transparenz und *Dokumentation* sind Prinzipien für vertrauenswürdige digitale Langzeitarchive. Standards liefern hierzu einen Beitrag, insbesondere, wenn sie als *offen* eingestuft werden können. Standards aus der IT-Welt schmücken sich oft mit diesem Attribut. Eine genaue und einheitliche Definition hierfür gibt es jedoch nicht. Verschiedene SDOs geben mehr oder weniger umfangreiche Kriterien an. IETF versteht darunter Standards, die eine gewisse Stabilität erreicht haben, offen zugänglich sind (ggf. gegen Bezahlung) und von einer allgemein anerkannten Standardisierungsgruppe erstellt wurden. ITU (International Telecommunication Union)[3] nimmt u.a. die langfristige Pflege und Unterstützung als Kriterium auf. Teilweise wird der Begriff *Offen* im Namen einer Norm geführt und dort individuell abgegrenzt, z.B. in OAIS. CCSDS weist dort darauf hin, dass nicht ein frei zugängliches Archiv gemeint sei, sondern dass der aktuelle Standard wie auch künftige damit verbundene Standards in einem offenen Forum entwickelt werden würden.

Probleme und Grenzen der Standardisierung

Standardisierungsprozesse sind nicht frei von Problemen. Zeitliche Verfügbarkeit von Standards und die inhaltliche Abgrenzung sind nicht immer ideal. Das Konsensprinzip

[1] http://www.cen.eu
[2] http://www.iso.ch
[3] http://www.itu.int

erfordert ggf. sehr aufwändige Abstimmungsprozesse. Zwar gibt es verkürzte Standardisierungsprozesse, gleichzeitig wird aber wieder beklagt, dass für eine gründliche Begutachtung nicht ausreichend Zeit sei. Wirtschaftliche oder sonstige Interessen führen teilweise zu konkurrierenden Inhalten oder unnötigem Umfang von Ansätzen. Aus unternehmerischer Sicht ist es verständlich, dass bereits getätigte Investitionen in einen Standard einfließen sollen. Die Abgrenzung von Inhalten und die zeitliche Synchronisation können auch durch die Vielzahl der Standardisierungsorganisationen negativ beeinflusst werden. Auf jeden Fall ist das Prozedere der Standardisierung und der Aufbau der Standards sehr unterschiedlich. Aus sachlichen Gründen mag eine Aufgabenteilung wegen der Komplexität und Breite der Inhalte gerechtfertigt sein, vorteilhaft wäre jedoch, wenn allgemein respektierte Referenzmodelle zu Grunde lägen, an denen sich die Standardisierung ausrichtet. Positiv zu vermerken ist, dass die SDOs gegenseitige Beziehungen (Liasons) in einer mehr oder weniger institutionalisierten Form unterhalten, und sich zumindest teilweise um eine Aufgabenteilung bemühen.

Die Offenheit von Standards ist nicht nur eine rein definitorische Angelegenheit, sondern kann weitgehende rechtliche und wirtschaftliche Konsequenzen haben. Versteckte Patentente oder sonstige Hindernisse, die z.B. Mitbewerber bei einer Implementierung behindern, können sich nachteilig auf die Zuverlässigkeit und Wirtschaftlichkeit der Langzeitarchivierung auswirken. Vorteilhaft ist, dass sich die Standardisierungsorganisationen um mehr Transparenz und auch Einheitlichkeit bei der Behandlung und Darstellung von Rechten (Intellectual Property Rights – IPR) bemühen.

Problematisch sind die hohe Anzahl der Standards, der teilweise extreme Umfang sowie Inhalte, die tiefgehende Vorkenntnisse erfordern, z.B. auf dem Gebiet der Logik oder Mathematik. Um einen besseren Überblick zu gewinnen und um Standards nach Fachdomänen zu strukturieren bietet sich der Rückgriff auf Architekturmodelle an. In dieser Expertise dient das OAIS-RM als Orientierungshilfe.

4 Standard Developer Organizations (SDO) und Standards im Überblick

4.1 OGF

Im Rahmen der Standardisierung von Grid-Technologien nimmt das Open Grid Forum (OGF)[4] eine zentrale Rolle ein. Grid Systeme und Anwendungen zielen darauf ab, Ressourcen und Dienste innerhalb einer verteilten, heterogenen, dynamischen *virtuellen Organisation* zu integrieren, zu virtualisieren und zu verwalten. Standardisierung wird als Schlüssel gesehen, um diese Vision zu realisieren. Zur Strukturierung der Standardisierung hat OGF die Kernfähigkeiten und das Verhalten eines Grids in einer serviceorientierten Architektur zusammengefasst (Open Grid Services Architecture – OGSA) [OGF-OGSA 2008]. Die Architektur soll eine Reihe von Anforderungen erfüllen: Interoperabilität und Unterstützung dynamischer und heterogener Umgebungen, die gemeinsame Nutzung von Ressourcen über Organisationsgrenzen hinweg, Optimierung der Nutzung von Ressourcen, die Zusicherung bestimmter Qualitäten von Diensten, die Ausführung von Aufträgen, Datendienste, Sicherheit, Reduktion der Kosten für die Administration, Skalierbarkeit, Verfügbarkeit sowie einfache Bedienung und Erweiterbarkeit. Diese Anforderungen würde man auch an die Infrastruktur eines vertrauenswürdigen digitalen Langzeitarchivs stellen.

[4] http://www.ogf.org

Um nun die Anforderungen zu erfüllen umfasst die Architektur eine Menge von Fähigkeiten: 1) Dienste für das Management der Ausführung (Execution Management) z.B. das Ausführen von Aufträgen, 2) Datendienste (Data Services) zum Verwalten, Abfragen und Ändern von Datenressourcen, 3) Dienste für die Verwaltung von Ressourcen (Resource Management Services) wie physische Ressourcen (Hardware) aber auch höherwertigere Dienste, 4) Dienste für die Sicherheit (Security Services) wie Authentifizierung und Autorisierung, 5) Dienste zum Selbstmanagement (Self-Management Services) wie Selbstheilung- und -optimierung, 6) Informationsdienste (Information Services) um Informationen über Dienste, Anwendungen und Ressourcen zu beschreiben und anzufragen. Um die bisher genannten Dienste zu unterstützen, beinhaltet OGSA so genannte Infrastrukturdienste (Infrastructure Services) wie die Benamung (Naming) von OGSA-Entitäten, die Darstellung von Zuständen (Representation of States), die Benachrichtigung (Notification), die Sicherheit (Security), Transaktionen (Transaction), Orchestrierung (Orchestration).

Wegen der anspruchsvollen Ziele und des großen Leistungsumfangs, wird durch so genannte Profile eine standardisierte Möglichkeit eröffnet, konsistente Teilkonfigurationen zu definieren. Somit kann der Reifegrad von Standards und der jeweilige Bedarf an Diensten bei der Zusammenstellung von Standards für ein interoperables System berücksichtigt werden. OGSA-Profile haben wie OGSA-Spezifikationen normativen Charakter.

Da realistischerweise nicht alle Komponenten gleichzeitig in Spezifikationen umgesetzt werden können, nimmt das OGF in einem im Strategiepapier [OGF2007a] eine Priorisierung von Aktivitäten bis 2010 vor, wobei sehr spezielle Architekturen ausgeschlossen sind wie z.B. SETI@HOME. Auf der Prioritätenliste stehen:

- Sicherheit im Grid (Grid Security), um Daten sicher zu transferieren, Nutzer zu authentifizieren und den Zugriff zu Ressourcen zu autorisieren

- Bereitstellung von Anwendungen (Application Provisioning), um Software einschließlich Zustand im Lebenszyklus aufzufinden, bereitzustellen und zu verwalten

- Einlieferung von Rechenaufträgen (Job Submission), um Jobs einzuliefern, den Status laufender Jobs abzufragen, laufende Jobs abzubrechen, welche in einer verteilten Umgebung ausgeführt werden

- Dateitransfer (File Movement), um Daten zu transferieren und dabei zu verwalten, einschließlich der Möglichkeit des Abbruchs, Zurückstellens und Wiederaufnehmens von Operationen

- Bereitstellung von Daten (Data Provisioning) zur Handhabung von Dateien, Datenbanken, Caching, Transfer, Metadaten und Föderation sowohl auf der Datenebene als auch auf der Speicherebene

- Programmierschnittstellen für Grid-Anwendungen (Grid Application Programming Interfaces), um Programmierschnittstellen und Abstraktionen zur Verfügung zu stellen, die Stabilität gegen Änderungen von Middleware-Technologien und zugrunde liegender Protokolle bieten.

Eine Reihe von Spezifikationen ist bereits erstellt und öffentlich verfügbar. Weitere Typen von Dokumenten, neben der oben näher erläuterten OGSA, geben nähere Information zu

den einzelnen Diensten der Architektur, zu Anwendungsfällen, zur Standardisierungsmethodik und zum Stand der Standardisierungsaktivitäten.

4.2 OASIS

OASIS (Organization for the Advancement of Structured Information Standards)[5] ist ein Non-profit-Konsortium, das sich zum Ziel gesetzt hat, die Entwicklung von offenen Standards für die Interoperablität von verteilten Informationsinfrastrukturen und Anwendungen im Internet oder innerhalb von Organisationen zu betreiben oder zu unterstützen. Schwerpunkte bilden Standards für Web-Services, Sicherheit, e-Business, Logistik und den öffentlichen Bereich.

Wegen der Relevanz für die Umsetzung von Grid-Architekturen, insbesondere von komplexeren Kommunikationssituationen unter Beachtung von Sicherheitsbedingungen, nehmen OGSA-Spezifikationen Bezug auf OASIS-Standards. OASIS ist auch die Heimat des relativ populären Standards Open Document Format for Office Applications (OpenDocument), der auch als ISO-Standard verabschiedet wurde [ISO/IEC 26300:2006].

4.3 W3C

Durch die Entwicklung von Standards und Empfehlungen sowie weitere Aktivitäten will das W3C (World Wide Web Consortium)[6] den sozialen Wert des Webs steigern und sicherstellen, indem die Zugänglichkeit für jeden und überall ermöglicht wird nach dem Motto „Web for Everyone“.

Durch etliche Standards ist diese Mission bereits Realität geworden. Einen wesentlichen Beitrag hat sicherlich HTML (Hypertext Markup Language) geliefert. Durch XML und XML Schemata sowie durch Transformations- und Abfragesprachen wie XSLT (XSL Transformations, XSL steht für Extensible Stylesheet Language) und XQuery wird eine strukturiertere und somit ausdrucksstärkere und verlässlichere Beschreibbarkeit und Verarbeitbarkeit von Daten geboten. Darüber bietet W3C offene Formate für die portable Darstellung von Bildern (PNG – Portable Network Graphics), Vektorgrafiken (SVG – Scalable Vector Graphics) und bestimmten Multimediaobjekten (SMIL – Synchronized Multimedia Integration Language).

Weitere Standards dienen der Realisierung von Web Services (WS) wie SOAP, WSDL (Web Services Description Language), WS Addressing und WS Policy.

Das *semantische Web* soll als universelles Medium für den Austausch von Daten im wissenschaftlichen, kommerziellen und kulturellen Bereich dienen. Maschinelles Verarbeiten und „Verstehen“ von Daten ist eine Voraussetzung. Wichtige Komponenten hierzu sind gegenwärtig RDF (Resource Description Framework), RDF Schema und OWL (Web Ontology Language). Die Arbeiten im W3C zu Metadaten sind in den Aktivitäten zum semantischen Web aufgegangen.

Spezifikationen und oder Sätze von Richtlinien (guidelines) werden als *Recommendation* bezeichnet und ähneln Standards anderer Organisationen.

[5] http://www.oasis-open.org
[6] http://www.w3c.org

4.4 IETF

Die IETF (Internet Engineering Task Force)[7] beschäftigt sich mit der Weiterentwicklung der Internet-Architektur als auch mit dem konkreten technischen Betrieb des Internets. mit Protokollen für den Transport und für die Vermittlung, mit Sicherheitsmechanismen und mit der Identifikation und Verwaltung von Ressourcen. IETF hat somit die Federführung für elementare Standards in verteilten Umgebungen wie TCP/IP (Transmission Control Protocol/Internet Protocol), FTP (File Transfer Protocol), LDAP (Lightweight Directory Access Protocol). Besonders populär ist das HTTP (Hyper Text Transfer Protocol).

Die Standards der IETF werden als *Internet Standard* oder auch *Full Standard* bezeichnet. Weitere Dokumente, die wie die Standards unter den Oberbegriff Request for Comments (RFC) fallen, liefern allgemeine Einführungen, Übersichten sowie Hilfestellung für die Implementierung.

4.5 Weitere Organisationen

Für das Funktionieren einer vernetzten Infrastruktur spielen weitere Organisationen eine mehr oder weniger große Rolle. Standards mit relativ engem Bezug zu Grid werden nun kurz vorgestellt.

IEEE (ursprünglich für Institute of Electrotechnical and Electronics Engineers)[8] ist für zahlreiche technisch orientierte Standards verantwortlich. Spezifikationen von IEEE sorgen für den ungehinderten physischen Transport der Daten als Basis jeder Art von Vernetzung. Darüber hinaus ist IEEE zusammen mit der Open Group verantwortlich für den Standard POSIX (Portable Operating System Interface), welcher sowohl für Teile des O-AIS-RM als auch der OGSA als Anhalt dient. Viele Anwendungen der eScience sind auf eine einheitliche und definierte Darstellung von Gleitkommazahlen und zugehörige Operationen angewiesen. Standards der IEEE versuchen dieses schwierige Problem in den Griff zu bekommen.

SNIA (Storage Networking Industry Association)[9] ist eine Nonprofit-Organisation, die sich primär mit der Standardisierung der Datenspeicherung und der Datenverwaltung beschäftigt. Die Aktivitäten zielen auf eine Verbesserung der Virtualisierung, Zuverlässigkeit, Sicherheit sowie Kommunikation zwischen Anwendungs- bzw. Verwaltungssoftware und Speichersystemen in verteilten und heterogenen Umgebungen.

WS-I (Web Services Interoperability Organization)[10] ist eine offene Industrie-Organisation, die sich die Sicherstellung der Interoperabilität von Web Services zum Ziel gesetzt hat. Die Vielzahl von Standards wird als eine Herausforderung für das Zusammenspiel von Web-Diensten von verschiedenen Herstellern und in unterschiedlichen Systemumgebungen gesehen. So genannte *Profile*, die auf Best Practices für ausgewählte Standards basieren, dienen als Anleitung für die Entwicklung (Implementierung) und den Betrieb von interoperablen Web Services. Anwendungsbeispiele, die auf unterschiedlichen Umgebungen beruhen, sowie Testwerkzeuge sollen die Feststellung der Übereinstimung mit Richtlinien der WS-I erleichtern.

[7] http://www.ietf.org
[8] http://www.ieee.org
[9] http://www.snia.org
[10] http://www.ws-i.org

5 OAIS-Referenzmodell als konzeptuelle Basis und Ordnungsschema

Das OAIS-Referenzmodell hat in der Langzeitarchivierungsszene Beachtung gefunden. Zahlreiche Veröffentlichungen behandeln dieses Modell und die beiden vorgelagerten nestor-Expertisen des GFZ und der FUH stellen bereits wesentliche Teile vor [Klum2007], [Schi2008]. Die folgenden Ausführungen konzentrieren sich daher auf Aspekte, die im Kontext dieser Expertise vertieft werden bzw. die den Gesamtüberblick erleichtern.

5.1 OAIS im Überblick

Das OAIS-RM beschreibt die Architektur eines Archivs, bestehend aus einer Organisation von Personen und Systemen, die die Verantwortung für den Erhalt und für die Bereitstellung von Information für eine bestimmte Zielgruppe übernommen hat. Die Beschreibung der Modellelemente liegt auf der konzeptionellen Ebene. Die Autoren weisen ausdrücklich darauf hin, dass das Modell weder einen Entwurf einer Implementierung noch eine Implementierung selbst spezifiziere. Implementierungsbezogene Funktionen eines Langzeitarchivs sind jedoch im Modell enthalten, wie das Verwalten von Speicherhierarchien oder die Bereitstellung allgemeiner Dienste (Common Services), wie sie üblicherweise von einem Betriebssystem erbracht werden. Solche Dienste könnte künftig eine Grid-Infrastruktur realisieren. Das OAIS-RM berücksichtig außerdem die bereits hochgradige Verteilung digitaler Information sowie die kooperative Organisation von Archiven. Virtuelle Organisationen (VO) auf Seite des Informationsproduzenten /-konsumenten bzw. des Archivs kommen jedoch in der Begrifflichkeit noch nicht vor.

Neben einem grundsätzlichen Modell für ein Archiv, beschreibt das OAIS-RM ausführlich ein Funktionsmodell und ein Informationsmodell.

5.2 Funktionsmodell

Das Funktionsmodell umfasst auf der obersten Ebene sechs funktionale Entitäten, nämlich Ingest, Archival Storage und Access sowie zur Unterstützung Datenmanagement, Preservation Planning und Administration. Vereinfacht nimmt Ingest die SIPs (Submission Information Package) (Erläuterung zu den Informationspaketen folgt im Informationsmodell) entgegen und generiert AIPs (Archival Information Package), welche dem Archival Storage zur permanenten Speicherung übergeben werden. Die Funktion Access nimmt die Anfragen der Konsumenten entgegen und generiert aus den AIPs, das sie von der Funktion Archival Storage auf Anfrage erhält, die DIPs (Dissemination Information Package) und liefert sie aus. Eine wesentliche Aufgabe des Datenmanagements ist die Verwaltung deskriptiver (inhaltsbeschreibender) Information. Die Funktion Administration dient der Verwaltung des gesamten Archivsystems und Preservation Planning behandelt spezifische Planungsaufgaben der Langzeitarchivierung u.a. das Entwickeln von Erhaltungsstrategien und Standards. Neben diesen sechs Entitäten thematisiert das OAIS-RM Querschnittsfunktionen, die im Folgenden etwas näher aufgeschlüsselt werden, da sie eine Brücke zu den Diensten bilden, die auch zum Umfang von Grid-Middleware gezählt werden können.

5.3 Common Services

Obwohl sich das OAIS-RM auf einer eher abstrakten Ebene bewegt, beschreiben die *Querschnittsdienste* (Common Services) solche unterstützende Dienste, die „moderne“ verteilte

Anwendungen erwarten. Die Ausführungen stützen sich dabei teilweise auf Vorarbeiten wie im POSIX OSE Referenz Modell beschrieben. Das OAIS-RM nennt drei Gruppen von Diensten: 1) Dienste eines Betriebssystems (Operating System Services): Diese umfassen Kernoperationen wie das Erzeugen und Verwalten von Prozessen, die Ausführung von Programmen oder das Verwalten von Dateien und Verzeichnissen. Zu dieser Gruppe zählen auch Kommandos und Dienstprogramme u.a. zum Anzeigen, Vergleichen und Editieren von Dateien, die Ausführung von Kommandoskripten und der Zugriff auf Umgebungsinformation. Weitere Dienste betreffen das Systemmanagement einschließlich Konfiguration und Leistungsmanagement von Geräten, Dateisystemen, Abrechnung, Warteschlangen und Backup. Dazu zählen auch Dienste, die sich mit der Systemsicherheit beschäftigen. Schließlich gehören zu dieser Gruppe noch die Echtzeit-Erweiterungen. 2) Netzwerkdienste (Network Services): Diese Gruppe von Diensten unterstützt den Datenzugriff und die Interoperabilität von Anwendungen in heterogenen Netzwerkumgebungen. Dazu zählen die Datenkommunikation für eine verlässliche und transparente Punkt-zu-Punkt-Übertragung von Daten in Netzwerken sowie der transparente Zugriff auf Dateien in heterogenen Netzwerken. Des Weiteren soll die Einbindung von PCs und Mikrocomputern unterstützt werden, deren Betriebssysteme nicht offen dokumentiert bzw. formal spezifiziert sind. Entfernte Prozeduraufrufe (Remote Procedure Call – RPC) und Dienste für die Netwerksicherheit, also für die Absicherung der Kommunikation zwischen Sendern und Empfängern im Netzwerk, zählen ebenfalls zu dieser Gruppe. 3) Dienste zur Sicherheit (Security Services): Diese Dienste schützen sensitive Daten und ihre Handhabung innerhalb des Informationssystems. Dazu gehören Dienste zur Identifikation bzw. Authentifizierung, zur Zugriffskontrolle, zur Sicherung der Integrität von Daten, zur Vertraulichkeit von Daten und zur Nichtabstreitbarkeit durch beteiligte Instanzen. Durch den letzten Dienst wird einerseits sichergestellt, dass der Empfänger einen Nachweis der Quelle erhält, so dass der Sender das Verschicken nicht leugnen kann. Andererseits erhält der Sender einen Nachweis über das Versenden der Daten, so dass der Empfänger den Erhalt nicht abstreiten kann.

5.4 Informationsmodell und logisches Datenmodell

Eine weitere Säule des OAIS-Referenzmodells ist das Informationsmodell. Es verdeutlicht, dass der Informationserhalt die Kernaufgabe eines OAIS-konformen Archivsystems ist. Hierfür ist es unabdingbar, dass Datenobjekte, die aus einer Sequenz aus Bits bestehen, mit ausreichend Information versehen werden, um durch Interpretation Inhaltsinformation (Informationsobjekte) zu generieren. Diese zusätzliche Information wird als Repräsentationsinformation (Representation Information) bezeichnet und benötigt ggf. zur Gewinnung von Information aus ihrer zu Grunde liegenden Repräsentation, z.B. in Form digitaler Daten, selbst wieder Information zur Interpretation. Über die Inhalte der Datenobjekte und über die Darstellung der Repräsentationsinformation macht das OAIS-RM keine Annahmen. Die Wahl objektorientierter Beschreibungsmittel im OAIS-RM impliziert nicht, dass auch die Inhaltsobjekte oder die Repräsentationsinformation in dieser Form repräsentiert sein müssten. Physische Objekte sind als Datenobjekte ebenfalls zugelassen. Somit berücksichtigt das OAIS-RM auch hybride Sammlungen, wie sie sowohl in klassischen Archiven und Bibliotheken als auch in naturwissenschaftlichen Sammlungen vorkommen. Die Repräsentationsinformation kann in sehr unterschiedlicher Form vorliegen, z.B. als formale Spezifikation eines Formats ggf. auf Papier oder als Emulationssoftware, die in einer aktuellen Umgebung läuft. Häufig werden Repräsentationsinformationen mittels Metadaten darge-

stellt, in der Hoffnung, geeignete Konzepte zu bezeichnen, mit deren Hilfe die gespeicherte Bitsequenz interpretiert werden kann.

Um Inhaltsinformationen langfristig verstehen zu können, führt das OAIS-RM zusätzlich vier weitere beschreibende Informationstypen ein, die unter dem Begriff Preservation Description Information (PDI) zusammengefasst sind. Referenzinformationen (Reference Information) sorgen für die – auch externe – Identifikation von Inhaltsinformationen. Kontextinformationen (Context Information) beschreiben den Bezug der Inhaltsinformation zur Außenwelt, u.a. warum die Information erzeugt wurde und wie sie mit anderen externen Inhaltsinformationsobjekten in Beziehung steht. Information zur Provenienz (Provenance Information) als spezielle Form der Kontextinformation dokumentiert die Historie eines Informationsobjekts wie die Quelle sowie Änderungen und Obhut seit der Entstehung. Die so genannte Fixity Information liefert die Prüfungen für die Integrität der Daten oder Schlüssel zur Validierung bzw. Verifikation, um sicherzustellen, dass die Inhaltsinformationen nicht in einer undokumentierten Art und Weise geändert wurden.

Die soeben beschriebenen Informationen (PDI) zusammen mit den Inhaltsinformationen bilden ein Informationspaket (Information Package). Für die Identifizierbarkeit dieser Entität und die Zusammengehörigkeit der Komponenten sorgt die so genannte Packaging Information. Die Pakete selbst werden ergänzend durch die Package Description beschrieben.

Das OAIS-RM unterscheidet drei Typen von Informationspaketen: Das Submission Information Package (SIP) wird vom Produzenten dem Archiv übersandt. Durch das Archiv wird ein oder mehrere SIPs in ein Archival Information Package (AIP) transformiert, insbesondere ist hierbei für vollständige PDI zu sorgen. Das AIP bzw. Teile oder Sammlungen davon, die dem Konsumenten bereitgestellt werden, bezeichnet das OAIS-RM als Dissemination Information Package (DIP).

Ein AIP unterscheidet sich von den anderen Pakettypen durch die geforderte Vollständigkeit der PDI. Virtuelle AIPs sind durchaus zugelassen, sofern die OAIS-Funktion Archival Storage für die erforderliche Abgrenzung und Identifikation sorgt. Darüber hinaus ist jedes AIP mit einer strukturierten Form deskriptiver Information, nämlich Package Description, zu verbinden, um für den Konsumenten relevante Information lokalisieren, analysieren und anfordern zu können. Die Package Description Information ist auch Basis für Zugriffshilfen (Access Aids), die ein AIP vor einer Auslieferung ggf. einer Transformation unterziehen, wie Teilmengenbildung, Subsampling oder Formatkonvertierung. Solche Transformationen können auch die PDI betreffen.

Des Weiteren führt das OAIS-RM für das AIP zwei Spezialisierungen ein, nämlich Archival Information Unit (AIU) und Archival Information Collection (AIC). Damit sollen unterschiedlich komplexe Archivierungssituationen bezüglich des Zugriffs einerseits und der Langzeiterhaltung andererseits gezielter behandelt werden können. Ein AIU umfasst genau ein Objekt mit Inhaltsinformationen, dem genau ein Satz an PDI zugeordnet ist. AIUs bilden also die atomaren Archivierungseinheiten in einem OAIS während ein AIC weitere AICs bzw. AIUs enthalten, die jeweils mit einem eigenen Satz an PDI versehen sind. Diese Komponenten bilden bezüglich eines AIC die Inhaltsinformation (Content Information), die wiederum mit PDI beschrieben wird. Die zugehörige Paketbeschreibung, Collection Description genannt, enthält Informationen über die Sammlung als Ganzes, und optional jeweils separate Beschreibungen der Komponenten. Mittels der Paketbeschreibung der Sammlung können komplexe Zugriffssituationen dokumentiert werden, insbesondere wenn

ein geeigneter Zugriff (Navigation) durch Einzelbeschreibungen der Pakete bzw. durch zugrunde liegende oder verfügbare Containerformate nicht darstellbar ist. Des Weiteren können mittels der Paketbeschreibungen für Sammlungen so genannte Access Collections definiert werden. Sie dienen dazu, bestimmte neue Sichten auf eine Sammlung, z.B. auf Basis eines Datamining-Prozesses, darzustellen, die für die Informationserschließung – temporär oder auch dauerhaft – von Nutzen sind. Access Collections dienen auch zur Beschreibung von Sammlungen, die erst künftig angelegt bzw. vervollständigt werden.

5.5 OAIS im Kontext von Grid/eScience

Das OAIS-RM ist eine geeignete Basis auch für Anwendungen im Bereich eScience. Das Informationsmodell macht keine Annahmen über Art und Umfang der Daten, über die Komplexität von Sammlungen und die durch sie repräsentierte Information. Das Informationsmodell legt den Grundstein für modular aufgebaute Informationsobjekte sowie für modulare Metadatenschemata. Das Funktionsmodell lässt außerdem genügend Spielraum für eine Grid-Implementierung. Die Bezugnahme auf nicht ganz aktuelle bzw. nicht ganz abstrakte Architekturen der Verteilung und Kooperation stellt keine besondere Einschränkung dar.

6 Wechselwirkung zwischen Langzeitarchivierung und eScience / Grid

6.1 Expertise: Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Daten

Diese nestor-Expertise untersucht aus technologischer wie organisatorisch-strategischer Perspektive, ob existierende eScience-Infrastrukturen in rohdatenproduzierenden Communities den Anforderungen zur Langzeitarchivierung gerecht werden, und ob die Erfahrungen der Communities im Bereich der Grid-Technologien auf Organisationen und Systeme zur digitalen Langzeitarchivierung übertragen werden können.

Im Folgenden sind die Inhalte der Expertise wiedergegeben, um unter Standardisierungsaspekten die Bezugnahme übersichtlicher zu gestalten.

Generell ist eScience charakterisiert durch eine hohe semantische Komplexität, mit der Daten, Dokumente und interaktive Werkzeuge zur Bearbeitung verknüpft werden. Die Datenmengen bleiben jedoch vergleichsweise gering. Semantische Komplexität tritt aber auch bei D-Grid Projekten auf. Zur Beschreibung von Objektbeziehungen werden Standards des Semantic Webs verwendet: RDF, OWL und darauf aufbauend SKOS (Simple Knowledge Organisation System). Dabei sind auch Objekte nicht-digitaler Art zu berücksichtigen.

Die Entwicklung von Grid wird angetrieben durch den hohen Bedarf an Rechen-, Speicher und Bandbreitenressourcen vorwiegend aus Forschungsprojekten der Naturwissenschaften, aber auch in anderen Disziplinen zeichnet sich hoher Ressourcenbedarf ab wie in der linguistischen Analyse von Texten. Die Datenmengen der Projekte gehen deutlich über das dort bisher übliche Maß hinaus.

Nun zu den Einzelheiten der anonymisierten Ergebnisse der in der Studie durchgeführten Befragung.

Herausforderung an die Archivtechnologie

Komplexität und Größe (im Petabyte-Bereich) der digitalen Objekte sowie hoher Ressourcenbedarf für Medien- und Formatmigration werden als Herausforderung gesehen. Daten-Grids sind gekennzeichnet durch Beschränkung auf Community-Grids, durch fehlende Standards für interoperable Kataloge zum Nachweis von Datenbeständen und Diensten und durch fehlende Autorisierungssysteme. Im Bereich eScience wird das Fehlen einheitlicher Schnittstellen beklagt.

Die Dauer der geforderten Aufbewahrung ist uneinheitlich und teilweise durch die intellektuelle Schöpfungshöhe und gesetzliche Vorgaben bestimmt. Die Auswahl der zu archivierenden Daten richtet sich nach einem angenommenen Wert, der u.a. bestimmt wird durch die intellektuelle Schöpfungshöhe oder die Wiederholbarkeit der Datengewinnung.

Bezüglich der Datei- und Medientypen stellt die Expertise fest, dass die Nachhaltigkeit „technischer“ Formate nicht bedacht werde, Format- und Informationsmodell der Daten oftmals nicht dokumentiert seien und Fragen nach der Tauglichkeit für die LZA spätestens bei der Überführung in ein Archiv zu klären wären. Im Bereich eScience wurde geäußert, dass die im Grid speicherbaren Daten wegen ihrer Heterogenität nicht den Anforderungen entsprächen. Ein Übergang von der Daten- zur Objektorientierung wird vorgeschlagen, um der Architektur des OAIS-RM besser gerecht zu werden.

Beklagt wurden auch Mängel bezüglich der Vertrauenswürdigkeit von Datengrids für die Langzeitarchivierung, der Granularität einer Rechteverwaltung, der Stabilität und der Nutzerfreundlichkeit von Werkzeugen. Auf die wichtige Rolle der Visualisierung und Prozessierung im Umfeld von Grid-Technologien wird hingewiesen. Des Weiteren wird die Frage nach der Langzeitarchivierung von Anwendungen und Quellkode aufgeworfen. Außerdem wurden Wünsche geäußert zu einer standardisierten Vorschau, zur Interpolation von Raum, Zeit oder anderen Dimensionen sowie zur Auswahl und Referenzierbarkeit von Teilmengen.

Weitere Themen sind der Zugriff auf verteilte Quellen, einheitliche Schnittstellen zu Archiven und deren Interoperabilität, sowie Formate. Neben disziplinspezifischen Formaten überwiegen Text- und Bildformate, Community-Grids verfügen über große Datenbestände im Binärformat. Eine reine Bitstream-Preservation wird als nicht ausreichend eingestuft mit der Folge, dass Anwendungen aufzubewahren sind und die Notwendigkeit einer Hardware-Emulation besteht. Schließlich wurde noch angemerkt, dass Standardisierung die wissenschaftliche Zusammenarbeit nicht behindern solle.

Herausforderung Metadaten

Metadaten werden häufig als lästig und nicht adäquat empfunden. Geeignete Werkzeuge für das Erstellen und Editieren fehlen und Standards sind entweder zu einfach oder noch häufiger zu komplex. Metadaten dienen meist dem Auffinden von Information und konzentrieren sich auf Kataloge, Inhalte und Zugriffsrechte. Nicht angesprochen werden hingegen Herkunft der Daten, Lizenzierung, Formate und semantische Verweise. Offensichtlich besteht auch das Missverständnis, dass eine Einigung auf einen einzigen Standard und eine Lesbarkeit durch Menschen nötig sei.

Angaben zu Dateitypen sind häufig nur implizit; man stützt sich auf MIME-Typen ab und hofft auf die Interpretation durch die Anwendung. Die Beschreibung der Herkunft von

Daten wird trotz ihrer Notwendigkeit, z.B. zur Angabe instrumenteller Parameter, eher vernachlässigt, d.h. die Gewinnung und Bearbeitung der Daten wird eher selten archiviert oder in Metadaten gefasst. Auf PREMIS als Möglichkeit wird explizit verwiesen, aber auch auf die dort fehlende Fähigkeit, Prozessketten vollständig abzubilden. Ein ausgeprägteres Bewusstsein besteht bei den eScience-Projekten, was sich in Prozessontologien und Versionierungen von Daten und Metadaten widerspiegelt. Der Autor der Expertise weist darauf hin, dass Grid vorteilhaft für die strukturierte Archivierung von Metadaten zur Herkunft und Verarbeitung sei, weil beides kodiert wird. Bemängelt wird jedoch die Ausrichtung der Workflows am e-Business. Die Befragung lieferte keine Nennung von Best-Practice-Beispielen, jedoch erfolgte ein Hinweis auf DFDL (Data Format Description Language) des OGF (Open Grid Forum).

Herausforderung Semantic Web

Generell zeichnen sich Grid-Projekte durch eine geringere semantische Komplexität aus und stehen im Gegensatz zu eScience-Projekten mit stets hoher Komplexität. Einige Grid-Projekte behandeln aber gleichzeitig hohe Datenvolumina und hochgradig vernetzte Objekte. Bei Grid-Projekten und weniger ontologieorientierten Vorhaben steht eher die Datenintegration im Vordergrund. Konkret wird die Beziehung zwischen Datenobjekten implizit über Identifier oder explizit mit RDF und OWL beschrieben. Die semantische Integration bei Grid wird von eScience als mangelhaft eingeschätzt. Bei der Zusammenführung von Grid- und Sementic-Web-Technologien wird Forschungsbedarf erkannt. Eine weitergehende Nutzung des Semantic Web könnte die Beschreibung von implizitem Wissen, wie z.B. Prozesswissen, oder die Verknüpfung mit nicht digitalen Objekten sein, wobei der Hinweis erfolgte, dass implizites Wissen (Prozesswissen) als Wettbewerbsvorteil zu sehen sei.

Herausforderung Zugang zu Daten und Rechteverwaltung

Die Zugänglichkeit zu Daten wird grundsätzlich bejaht und Einschränkungen des Zugangs sind im Wesentlichen rechtlich begründet. Einige Besonderheiten wurden genannt wie der Wandel der Rechte in Abhängigkeit der Arbeitsphasen, die Publikation von Daten ohne Fachliteratur, die Vererbung von Rechten über lange Zeiträume, die Archivierung der Zugriffregelungen und die Nachsignierung. Beim Management von Rechten in verteilten Umgebungen werden Defizite genannt wie mangelnde Granularität, Verwaltbarkeit und Zuverlässigkeit. DRM (Digital Rights Management) hingegen ist eher von untergeordneter Bedeutung. Standardisierte Lizenzen sind wegen ihrer Maschinenlesbarkeit zu bevorzugen.

Herausforderung Organisation und Nachhaltigkeit

Die Befragung ergab, dass nur eine Minderzahl der Projekte eine Policy zur Langzeitarchivierung hat, eine vergleichbare Anzahl ist dabei, eine zu entwickeln. Wenn keine gesetzlichen Regelungen einzuhalten sind, orientieren sich die Projekte an DFG-Empfehlungen oder entsprechende Regelungen anderer Wissenschaftsorganisationen. Die Diskrepanz zwischen der Kurzfristigkeit von Projekten und der Erfordernis einer nachhaltigen Infrastruktur wirft neue Fragen auf. Außerdem werden Unsicherheiten über die Vertrauenswürdigkeit von Archiven sichtbar insbesondere, wenn rechtlich begründete Pflichten bestehen. Vom Autor als bemerkenswert eingestuft wird die Aussage, dass die Abrechnung von Leistungen im Grid innerhalb von virtuellen Organisationen und mit Dritten noch nicht als offene Frage wahrgenommen wird. Ein weiteres Problem wird in der Nutzung der pilothaft

aufgebauten Systeme als Berechtigung für ihre Existenz gesehen. Die Möglichkeit für Wissenschaftler, ihre Ergebnisse langfristig zu archivieren, könnte ein Anreiz sein.

Zusammenfassung in Hinblick auf eine Standardisierung

Die vorgestellte Studie zeigt eine Reihe von Anforderungen, die sich sowohl auf der technischen und konzeptionellen als auch auf der organisatorischen Ebene bewegen. Obwohl sich in einzelnen Communities bereits Best Practices herausbilden, ist eine umfassende, auch über Community-Grenzen hinausgehende Standardisierung bezüglich Langzeitarchivierung nicht erkennbar. Des Weiteren sind keine systematischen Aktivitäten genannt, die der Vorbereitung einer Standardisierung dienen, wie z.B. eine systematische und vertiefte Aufnahme des IST-Zustandes und eine hinreichend formale Darstellung der Anforderungen, die u.a. Aussagen zur grundsätzlichen Eignung von Produkten und Prozessen für eine Standardisierung erlauben sollten. Gelegentlich wird auf Standards hingewiesen, die außerhalb von eScience für die Langzeitarchivierung verwendet werden.

6.2 Expertise: Synergiepotenziale zwischen Grid- und eScience-Technologien für die digitale Langzeitarchivierung

Die Studie zeigt der FernUniversität Hagen mögliche Potenziale zwischen Grid- und eScience-Technologien für die digitale Langzeitarchivierung sowie Konzepte zur Integration auf [Schi2008]. Die Studie untersucht also einerseits, wie Archive in eine Grid-Umgebung eingebunden werden können, so dass ein Mehrwert für die Nutzer des Grids entsteht, und andererseits, wie für Zwecke der Langzeitarchivierung Grid-Technologien genutzt werden können.

Zur Einführung charakterisiert die Studie die Aufgaben und Probleme der digitalen Langzeitarchivierung und gibt eine knappe Einführung in das Informations- und Funktionsmodell des OAIS-Referenzmodells. Es folgt eine Vorstellung von Systemen die sich an OAIS orientieren, nämlich BABS (Bibliothekarisches Archivierungs- und Bereitstellungssystem)[11] der Bayerischen Staatsbibliothek München und kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen)[12] der Deutschen Nationalbibliothek, der Niedersächsischen Staats- und Universitätsbibliothek, der GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung) und IBM. Des Weiteren werden Eigenschaften institutioneller Repositorien skizziert und auf die Open Archive Initiative (OAI) und die Möglichkeit des Metadatenexports mittels OAI-PMH (Preserevation Metadata Harvesting) hingewiesen.

Im Weiteren erläutern die Autoren die Begriffe Middleware, Grid und eScience. Middleware ist eine Softwareschicht, die die Heterogenitäten und Komplexität von vernetzten Rechner- und Speichersystemen vor dem Nutzer verbirgt und eine einheitliche Sicht auf die Ressourcen bereitstellt. Entsprechend den zugrunde liegenden Paradigmen findet eine Unterscheidung zwischen Peer-to-Peer (P2P) und Grid statt. Probleme von P2P Lösungen für die LZA sehen die Autoren in der eingeschränkten Leistungsfähigkeit der beteiligten Rechner und der Netzwerkanbindung, in der Verfügbarkeit der Rechner und der daraus entstehenden Notwendigkeit der aufwändigen Verteilung von Rechenaufgaben und der Anlage zahlreicher Kopien (Replikate). Es werden zwei Lösungen zur Archivierung vorge-

[11] http://www.babs-muenchen.de
[12] http://kopal.langzeitarchivierung.de

stellt, die sich trotz dieser Nachteile des P2P-Ansatzes bedienen, nämlich LOCKSS[13] und OceanStore[14].

Für Grid-Systeme werden verschiedene Definitionen vorgestellt und deren Architektur anhand eines Schichtenmodells erläutert. Charakteristisch für Grid sind die Abstraktion von technischen Details und die gemeinsame Nutzung der eingebundenen Ressourcen durch viele Nutzer. Entsprechend der hauptsächlich bereitgestellten Ressourcen, findet sich in der Literatur eine Einteilung in Rechengrids (Compute Grid) und Datengrids (Data Grid). Dauerhaftigkeit, Nachvollziehbarkeit und Kostenvorteile der angebotenen Dienste gehören zu den geforderten Eigenschaften. Interoperabilität, Portierbarkeit und Wiederverwendbarkeit sind durch entsprechende Standards und Schnittstellenspezifikation erreichbar. In diesem Zusammenhang wird auf die Bedeutung der Open Grid Services Architecture (OGSA) hingewiesen. Wegen ihrer *Zustandslosigkeit* sind die üblichen Web-Dienste nicht geeignet, um den Datenverkehr zwischen den Diensten sinnvoll zu bewältigen. Die notwendige Erweiterung legt WSRF (Web Services Resource Framework) in einer standardisierten Weise fest.

Folgende drei Rechengrids stellt die Studie vor: 1) Globus mit den Komponenten gsiFTP (vormals GridFTP), OGSA-DAI (Database Access and Integration Service) und DRS (Data Replication Service), bestehend aus RFT (Reliable File Transfer Service) und RLS (Replication Location Service), 2) UNICORE als betriebsfertige Middleware, die sich ebenso wie Globus auf OGSA und WSRF abstützt, 3) gLite mit seiner Möglichkeit, Komponenten anderer Grid-Middleware einzubinden.

Datengrids haben den Zweck, Datensammlungen aus verschiedenen Anwendungsdomänen für eine gemeinsame Nutzung innerhalb von Communities zusammenzuführen, ohne sich bei der jeweils anderen Domäne als Benutzer explizit anmelden zu müssen. Trotz ihrer großen Bedeutung sowohl für Rechengrids als auch für eScience sind sie noch unterrepräsentiert. Die Expertise beschreibt folgende drei Produkte: 1) dCache: Neben der Herstellung von Transparenz unterstützt diese Middleware die Verwaltung von Replikaten sowie den Zugang zu unterschiedlichen hierarchischen Speicherverwaltungssystemen (Hierarchical Storage Management – HSM) wie TSM (Tivoli Storage Manager), Open Storage Manager (OSM) und High Performance Storage System (HPSS). Kerberos und die Verwendung von SSL-Protokollen (Secure Sockets Layer) erhöhen die Sicherheit. Die Abstützung auf verschiedene FTP-Derivate (GsiFTP) und weiterer Standards wie LDAP erhöht die Flexibilität. 2) SRB (Storage Resource Broker): Charakteristisch für dieses System ist die Bereitstellung eines Metadatenkataloges (Metadata Catalog – MCAT), auf Basis unterschiedlicher Datenbankmanagementsysteme, und die Organisation der Daten in hierarchisch aufgebauten Sammlungen. Die Authentifizierung geschieht Client-basiert mittels GSI (Grid Security Infrastructure) von Globus oder über das DBMS, das den MCAT realisiert. Der Netzwerkverkehr kann über die Verschlüsselung Encrypt1 gesichert werden. 3) Nirvana ist die kommerzielle Variante von SRB, wobei sich die Varianten zunehmend entfernen. 4) iRODS (i Rule Oriented Data System) ist ebenfalls eine Variante von SRB und erlaubt mittels regelorientierter Programmierung (Rule Oriented Programming – ROP) eine feingranulare Anpassung an individuelle Bedürfnisse. Das Konzept der Anpassbarkeit an neue Systemumgebungen und neue Anforderungen wird als eine vorteilhafte Eigenschaft von Grid-Middleware eingestuft. iRODS befindet sich noch in einem frühen Entwicklungsstadium.

[13] http://www.lockss.org
[14] http://oceanstore.cs.berkeley.edu

eScience bedeutet eine orts- und/oder zeitunabhängige Zusammenarbeit in Kerngebieten der Wissenschaften und eine Verfügbarkeit der dafür benötigten Werkzeuge und Infrastrukturen, welche den Austausch von Nachrichten und Daten und eine gemeinsame Nutzung von Ressourcen wie Rechnersysteme, Netzwerke, Instrumente erlauben. Die offensichtliche Korrelation zwischen diesen Anforderungen und den Eigenschaften von Grid-Systemen legt eine Nutzung dieser Systeme nahe. In Deutschland schafft die Initiative D-Grid Grundlagen für den integrativen Einsatz dieser Technologien. Auf das bisherige Fehlen einer einheitlichen Plattform für die Archivierung von Projektergebnissen und wichtigen Datenbeständen weisen die Autoren hin. Unter dem Gesichtspunkt, Synergiepotenziale bezüglich einer einheitlichen und allgemein verwendbaren Kommunikationsstruktur zu erkennen, werden wegen ihrer kollaborativen und wissenserzeugenden Komponenten folgende vier Projekte vorgestellt: eSciDoc, ONTOVERSE, WIKINGER, Im Wissensnetz.

Die Befragung der drei Community-Grid-Projekte AstroGrid-D, C3-Grid und TextGrid diente einer Vertiefung der Analyse zu Synergiepotenzialen. Neben dem status quo umfasst die Erhebung Anforderungen an eine künftige Langzeitarchivierung. Die Befragung lieferte auch einige explizite Nennungen verwendeter Standards.

Auf Basis der vorhergehenden Analysen zeigt die Expertise konkrete Synergiepotenziale auf und skizziert eine Infrastruktur zur Langzeitarchivierung. Diese berücksichtigt, dass unterschiedliche, teilweise bereits bestehende, im Allgemeinen heterogene Archive einzubinden sind. Die Serviceorientierung dient hierfür als Architekturkonzept. Als konkretes Protokoll zum Anbieten und Nutzen von Diensten werden Web Services (WS) vorgeschlagen, die auch in Grid-Architekturen als Standard akzeptiert sind und in den analysierten Archiven bzw. in den übliche IR-Produkten enthalten sind.

Diese Architektur bildet die Grundlage für Handlungsempfehlungen zur Umsetzung der identifizierten Synergiepotenziale: 1) Grid nutzt LZA: nahtlose Integration bestehender und neuer Archive, kombinierte Einreichung von Publikationen und zugehörigen Daten, bessere Auslastung von Grid-Ressourcen. 2) LZA nutzt Grid: Ausführung von Formatvalidierungen im Grid, Test der Archivfähigkeit bzw. Migration vor Ingest, Vorbereitung von SIPs, zusätzliche Metadaten durch Indexierung, ontologiebasierte Analyse, metadatenbasierte Suche im Grid, Formatmigration im Grid, Migration eingebetteter Formate, regelmäßige Validierung und Migration von AIP und schließlich Unterstützung der Archive durch Grid-Speicherressourcen. Die Autoren weisen auf zu beachtende Punkte hin, wie 1) die Orthogonalität der Dienste bei gleichzeitiger Vermeidung von überhöhtem Kommunikationsaufwand durch eine ungeeignete Modularisierung, 2) die Konfigurierbarkeit um ggf. auch eine interdisziplinäre Wissensvernetzung zu realisieren, 3) Spezifikation von Funktionen und Parameter für WS und Herleitung von Standards, 4) Entwicklung von Methoden zum Scheduling für eine schnelle Reaktion der Webservices und eine optimale Auslastung der Ressourcen.

Die Handlungsempfehlungen beinhalten auch den Aufbau eines Testbeds, um die Performanz und Akzeptanz einer serviceorientierten LZA-Infrastruktur zu evaluieren und zu optimieren sowie einen Piloten und nachfolgend einen produktiven Betrieb vorzubereiten. Als Systeme für ein Testbed kommen wegen ihrer Zugänglichkeit über WS in Frage: kopal mit der Schnittstelle koLibRI oder Datenarchive des IVOA (International Virtual Observatory Alliance) oder WDCC (World Data Center for Climate), aber auch institutionelle Re-

positorien (IR) wie dSpace[15] oder Fedora[16]. Letztere könnten auch dazu genutzt werden, ein vorhandenes Datengrid, wie SRB oder dCache, anzubinden.

Zusammenfassung in Hinblick auf eine Standardisierung

Die vorgestellte Studie stellt eine Reihe von Grid-Technologien und Produkten vor. Um die genannten Potenziale wirtschaftlich nutzen zu können, steht eine Standardisierung außer Zweifel (vgl. auch Aussagen und Standardisierungsaktivitäten des OGF). Die Studie umfasst auch eine Befragung von drei Community-Grid-Projekten, welche Aussagen zur derzeitigen Archivierung und zu Anforderungen an eine künftige Langzeitarchivierung macht und erste Hinweise für eine Standardisierung liefert. Einige Standards konzeptioneller (inhaltsbeschreibend wie TEI[17] oder ISO 19115) und technischer Art (OGSA-DAI) werden explizit genannt. Hinter TIBORDER verbirgt sich DOI, also die Identifizierung und Katalogisierung digitaler Objekte.

6.3 Weitere Quellen

Weitere Hinweise gibt die nestor-Expertise *Langzeitarchivierung von Rohdaten* [Seve2006]. Einige Ergebnisse seien hier nochmals hervorgehoben. Bezüglich der Daten sagt die Studie aus, dass die Dateiformate weitestgehend projektspezifisch und nur teilweise innerhalb der Disziplinen vereinheitlicht seien. Bemühungen um eine Reduzierung der Formate werden jedoch genannt. Außerdem überwiegt das Binärformat für Rohdaten. Des Weiteren stellen die Autoren fest, dass die Mehrzahl der Daten nur mittels spezieller Software intellektuell erschließbar seien. Aus diesen beiden Aussagen ergeben sich erhebliche Konsequenzen für die Bereitstellung von Repräsentationsinformationen in einem Archiv. Darüber hinaus wird eine umfassende Standardisierung für Daten und Metadaten für nicht durchführbar erachtet. Eine weitere Aussage betrifft die Sorge um die Sicherheit der Daten. Hierbei wird erkennbar, dass die Vertrauenswürdigkeit von Archiven für den Aufbau einer LZA-Infrastruktur ein Thema ist.

Weitere Aussagen zur Archivierung finden sich in den Dokumenten des OGF. Ein Memo sammelt Anwendungsfälle der eScience-Community bezüglich Grid-Netzwerkdiensten [OGF-HPRG 2007]. Aussagen zu Rechenleistungen, Speicherkapazitäten, Bandbreiten für die Übertragung sowie Zuverlässigkeit und Sicherheit überwiegen. Ein Beitrag im Zusammenhang mit Networked Supercomputing thematisiert auch Speichernetzwerke. Sofern sie auf Dateiebene (im Gegensatz zur Blockebene) arbeiten, bieten sie an zentraler Stelle eine Möglichkeit, Dateien für einen *Archivierungsdienst* zu verwalten.

Ein weiteres Memo der OGF beschäftigt sich intensiv mit den Anforderungen an ein digitales Langzeitarchiv [OGF-LTDAR 2005]. Ausgehend von zwei Kernforderungen, nämlich der Minimierung der Kosten und der Risiken des Betriebs eines Archivs, leitet der Autor 16 Einzelforderungen ab, um eine OAIS-konforme Architektur zu implementieren. Danach folgt eine Abbildung der Forderungen auf die einzelnen Standardisierungsaktivitäten des OGF. Integrationsbedarf sieht der Autor u.a. bei den Architekturen zum Datenmanagement aus der Welt des Grids und der Bibliotheken. Genannt werden in diesem Zusammenhang METS, AIPs, OAI-PMH sowie Metadaten zur Authentizität und zur Integrität.

[15] http://www.dspace.org
[16] http://www.fedora.info
[17] Text Encoding Initiative siehe http://www.tei-c.org

7 Standards im Einzelnen

Die Betrachtung der Standards soll ein Bild der aktuellen Situation der Standardisierung liefern und Anhaltspunkte geben für weitere Aktivitäten im Zusammenhang mit Standards.

Die untersuchten Standards beziehen sich schwerpunktmäßig auf das Informationsmodell des OAIS-Referenzmodells. Die Umsetzung des Informationsmodells ist wesentlicher Bestandteil eines OAIS-konformen Langzeitarchivs. Vertieft betrachtet werden also Standards, die sich mit der Interpretierbarkeit digitaler Daten und der Verwaltbarkeit digitaler Daten, die Informationspakete repräsentieren, beschäftigen. Bekannte Standards zur Langzeitarchivierung wurden weitgehend von traditionellen Gedächtnisorganisationen entwickelt, wobei der Fokus (bisher) naturgemäß nicht auf wissenschaftlichen Daten und der zugrunde liegende Infrastruktur wie Grid oder gar Messgeräten liegt. Daher wird, jeweils nach Vorstellung dieser Standards, eine Diskussion geführt, wie es mit einer Eignung im Kontext eScience und Grid steht, wobei auf Aussagen der Studie des GFZ [Klum2007] Bezug genommen wird. Da auch im Bereich Wissenschaft (eScience) die Notwendigkeit erkannt ist, Daten mit weiterer Information zur „Selbstdokumentation“ versehen und in eine geordnete Form zu bringen, werden exemplarisch einige Standards aus diesem Bereich besprochen wie spezifische Datenformatbeschreibungssprachen. Solche Standards müssten von jenen Gedächtnisorganisationen berücksichtigt werden, die Verantwortung für wissenschaftliche Daten übernehmen.

Da beide Bereiche formale (implementierbare) Beschreibungsmittel benötigen, werden einige Basisstandards wie XSchema vorgestellt. Letztlich sollen Informationen auch ausgetauscht und geteilt werden können. Daher finden sich auch Standards, die eine gemeinsame und sichere Nutzung (z.B. OGSA-DAI) bzw. einen Austausch (z.B. SOAP) von Daten erlauben. Auch zu diesen Standards finden sich Anmerkungen in Zusammenhang mit Grid/eScience und Langzeitarchivierung. Etliche Standards bieten durchaus die Chance, die in der Studie des GFZ [Klum2007] angesprochenen Probleme zu lösen, wie die Verwaltung von Rechten in verteilten Umgebungen.

Als Raster für die Einteilung des Kapitels dienen Elemente des OAIS-Referenzmodells. Im Anhang finden sich eine Zusammenstellung der Standards und ihre Zuordnung zu den OAIS-Elementen.

7.1 Erhaltungsinformation (Representation Information und PDI)

Die Interpretierbarkeit von Bitsequenzen ist ein Kernproblem der Langzeiterhaltung digitaler Information. Erst durch Interpretationsprozesse entsteht aus den Daten Information. Um den Interpretationsprozess zu ermöglichen, sind Repräsentationsinformationen erforderlich sowie weitere Informationen, die die Qualität dieses Prozesses sicherstellen bzw. bewertbar machen. Diese Informationen lassen sich unter dem Begriff Erhaltungsinformation zusammenfassen. Sie werden häufig in Form von Daten dargestellt, die unter dem Begriff Metadaten eingeordnet werden. Die Begrifflichkeit für diese spezifischen Metadaten ist jedoch nach wie vor uneinheitlich und nicht deutlich abgegrenzt (technische Metadaten, Verwaltungsmetadaten, Erhaltungsmetadaten u.ä.). Nachfolgend wird PREMIS genauer betrachtet, da dort das Thema Erhaltungsinformation in einer Ausführlichkeit behandelt wird, die auch bei einer Bewertung oder Entwicklung anderer Ansätze hilfreich ist.

7.1.1 PREMIS

Eine durch die OCLC und RLG eingerichtete Arbeitsgruppe namens PREMIS[18] (Preservation Metadata Implementation Strategies) veröffentlichte im Mai 2005 ein Data Dictionary, welches einen implementierbaren, breit nutzbaren Kern von Erhaltungsmetadaten definiert [PREMIS2005]. Neben der reinen Definition von Datenelementen findet sich eine ausführliche Diskussion zu speziellen Problempunkten der Langzeitarchivierung. Die Autoren wiesen darauf hin, dass es sich nicht um ein fixiertes und endgültiges Data Dictionary handle, sondern um einen konsolidierten Ausgangspunkt für Erweiterungen und Verbesserungen auf der Basis von Erfahrungen und Rückmeldungen. Seit März 2008 liegt nun nach intensiven Revisionsaktivitäten, unterstützt durch die Library of Congress, die Version 2.0 vor [PREMIS2008]. Die konzeptionellen Aspekte der Erhaltung in PREMIS können als eine Konkretisierung der Konzepte aus OAIS, insbesondere der Repräsentationsinformation und PDI, eingeordnet werden. Darüber hinaus thematisiert PREMIS die Implementierung des Data Dictionary; eine Architektur für die Implementierung wird jedoch nicht vorgeschrieben. Für die Version 2.0 liegt der Draft eines XML-Schemas vor.

PREMIS definiert Erhaltungsmetadaten (preservation metadata) als jene Information, die ein Repository nutzt, um den Erhaltungsprozess zu unterstützen. Speziell werden Metadaten genannt, die die Funktionsfähigkeit (viability), Darstellbarkeit (renderability), Verständlichkeit (understandability) und Identität (identity) in einem Erhaltungskontext bewahren. Besondere Aufmerksamkeit erfuhr die Dokumentation der Provenienz und der Beziehungen unterschiedlicher Objekte, insbesondere jener im Repository.

Das Grundgerüst für das Datenmodell bilden fünf Typen von Entitäten (Entities): die intellektuelle Entität, das digitale Objekt oder kurz Objekt (Digital Object bzw. Object), das Ereignis (Event), die Rechte (Rights bzw. Rights Statements) und der Handelnde (Agent). 1) Die intellektuelle Entität ist als zusammenhängende Menge von Inhalt (Content) definiert, der vernünftigerweise als Einheit beschrieben wird. Ein Objekt wird mit einer Einheit von Information in digitaler Form gleichgesetzt. 2) Das Ereignis ist eine Aktion, die mindestens ein dem Repository bekanntes Objekt oder Handelnden einschließt. 3) Ein Handelnder ist eine Person, Organisation oder Software-Programm/-System, verbunden mit Erhaltungsereignissen im Leben eines Objekts. 4) Rechte sind Erklärungen (Assertion) zu einem oder mehr Rechten bezüglich eines Objekts und/oder eines Handelnden.

Eine Beziehung (Relationship) ist eine Verbindung (Association) zwischen Instanzen von Entitäten, explizit zwischen zwei oder mehr Objekten oder Entitäten unterschiedlichen Typs. (Anmerkung: Dass aufgrund der Definition von intellektuellen Entitäten diese Typen in einer Ganzes-Teile-Beziehung stehen dürfen, wird hier nicht mehr aufgegriffen). Die bereits erwähnten semantischen Einheiten beschreiben Eigenschaften einer Entität und besitzen Werte (Value). Semantische Einheiten werden außerhalb konkreter Entitätstypen definiert. Die Zuordenbarkeit (Kardinalität) zu Entitätstypen wird jeweils in den semantischen Einheiten spezifiziert. Teilweise sind semantische Einheiten in Behältern (Container), die selbst keine Werte besitzen, zusammengefasst, wobei die Teile in dieser Rolle dann als semantische Komponenten (Semantic Component) bezeichnet werden.

Die Entität Objekt hat folgende drei Untertypen: Datei (File), Bitstrom (Bitstream) und Repräsentation (Representation). 1) Eine Datei ist eine geordnete Sequenz von Bytes, die dem Betriebssystem bekannt ist, und besitzt ein Dateiformat, Zugriffserlaubnisse und Sys-

[18] http://www.loc.gov/standards/premis

temstatistiken. 2) Ein Bitstrom bezeichnet zusammenhängende oder nicht zusammenhängende Daten innerhalb einer Datei, die für Erhaltungszwecke gemeinsame Eigenschaften aufweisen. Ein Bitstrom kann nicht ohne weitere Maßnahmen in eine eigenständige Datei überführt werden. Sollte dies ohne weiteres Hinzufügen von Information, ggf. mittels Transformation, möglich sein, so definiert PREMIS hierfür den Begriff Dateistrom (Filestream). 3) Eine Repräsentation ist diejenige Menge von Dateien, einschließlich struktureller Metadaten, die eine vollständige und vernünftige Darstellung einer intellektuellen Entität erfordert.

Die PREMIS Arbeitsgruppe gelangte zu der Erkenntnis, dass sich die meisten Beziehungen zwischen Objekten in folgende drei Typen einteilen ließe: Strukturelle Beziehung (Structural Relationship), Ableitungsbeziehung (Derivation Relationschip) und Abhängigkeitsbeziehung (Dependency Relationship). 1) Strukturelle Beziehungen beschreiben Beziehungen zwischen Teilen von Objekten. 2) Ableitungsbeziehungen resultieren aus einer Replikation oder Transformation eines Objekts. Der intellektuelle Inhalt wird durch die Ableitung nicht verändert. 3) Eine Abhängigkeitsbeziehung besteht, wenn ein Objekt ein anderes benötigt, um seine Funktion, seine Bereitstellung oder den Zusammenhang des Inhalts zu unterstützen. Im Data Dictionary sind die Abhängigkeitsbeziehungen ein Teil der Umgebungsinformation (Environment Information), um eine zusammengefasste Darstellung aller Abhängigkeiten, einschl. Hardware, zu erlauben. Die Beziehungstypen *strukturell* und *Ableitung* hingegen sind in einer eigenen semantischen Komponente in einem eigenen Container vorgesehen, wobei sie dort nur als Wert für eine High-Level-Kategorisierung einer Beziehung vorgeschlagen werden, die durch genau einen Subtyp näher spezifiziert werden muss. Für den Subtyp werden als Werte vorgeschlagen: has sibling, is part of, has part, is source of, has source, has root, includes, is included in. Die semantische Komponente kann allen Objekttypen – auch mehrfach – zugeordnet werden, wobei die Verwendung der vorgeschlagenen Subtypen bestimmten Einschränkungen unterliegt (z.B. has root nur für Repräsentationen).

Im Folgenden werden wichtige Abgrenzungen und Diskussionspunkte aus Sicht PREMIS besprochen. So schließt die Arbeitsgruppe deskriptive Metadaten, die sie den intellektuellen Entitäten zuordnet, von der Betrachtung aus, da sich bereits andere Standards und Initiativen dieser häufig fachspezifischen Daten widmen. Aus ähnlichen Gründen wird auch auf eine Detaillierung der Handelnden verzichtet. Eine Beschränkung auf erhaltungsspezifische Gesichtspunkte erfolgt ebenfalls bei den Rechten. Hier steht also die Frage im Vordergrund, welche Änderungen ein Repository an den Objekten vornehmen darf. Obwohl PREMIS technische, formatspezifische Metadaten für die meisten Erhaltungsstrategien als wichtig erachtet, werden aus Ressourcengründen nur Metadaten definiert, die der Erwartung nach für Objekte aller Formate zutreffen. Außerdem wird die Definition von Metadaten zur detaillierten Beschreibung von Datenträgern und Hardware entsprechenden Spezialisten überlassen. Bezüglich Geschäftsregeln ist nur das Erhaltungsniveau (preservation level) eines Objekts vorgesehen.

Besondere Aufmerksamkeit widmet PREMIS der Information über Formate. Der Begriff Format ist dort definiert als die Organisation digitaler Information entsprechend einer festgesetzten formalen oder informalen Spezifikation. Die semantische Einheit namens Format kann den Entitäten Datei (hier einschließlich Dateistrom) oder Bitstrom zugeordnet werden. Da sich Formate auch auf Bitströme beziehen, wird der Name Dateiformat vermieden. Die Nutzung von MIME-Typen oder Dateierweiterungen wird ohne Versionsangaben als unzureichend eingestuft. Da künftige, zentral gepflegte Formatverzeichnisse als beste Lö-

sung eingeschätzt werden, sind im Data Dictionary Verweismöglichkeiten vorgesehen. Bei spezifisch angepassten Formaten, hier als *Profile* bezeichnet, wird die Verwendung der spezifischsten Bezeichnung vorgeschlagen. Was die spezifischste Bezeichnung ist, ist nach Erkenntnis der PREMIS Arbeitsgruppe eine Auffassungsfrage[19]. Die explizite Beschreibbarkeit einer spezifischen Profilierung ist jedoch nicht vorgesehen.

Ein weiterer wesentlicher Bestandteil der Erhaltungsmetadaten kann aus Sicht von PREMIS die technische Umgebung sein. Doch durch die wiederholte Zerlegbarkeit der Umgebung in Komponenten kann die Beschreibung sehr komplex werden. Obwohl im Datenmodell erlaubt, wird nicht unbedingt die Beschreibung aller möglichen Umgebungen (Hard- und Softwarekombinationen) gefordert, sondern die Angabe bestimmter Kategorien von Umgebungen, wobei zumindest eine Minimalumgebung enthalten sein sollte. Sind jedoch bestimmte signifikante Eigenschaften eins Objekts zu erhalten, ist eine entsprechend höherwertigere Kategorie anzugeben. Auf die Zuordenbarkeit von Umgebungsinformation zu Bitströmen wird hingewiesen, da nicht in allen Fällen das Dateiformat die entsprechende Kenntnis zur Interpretation verfügt. Eine Zuordenbarkeit zu Repräsentationen ist ebenfalls zulässig, um zwischen der Interpretation aggregierter Objekte und einzelner Bestandteile, hier Dateien, unterscheiden zu können. Obwohl Metadaten zur Umgebung als kritisch angesehen werden, ist die Angabe dieser semantischen Einheit optional, weil nicht abschließend gesagt werden kann, ob sie für alle, auch künftigen, Erhaltungsstrategien[20] notwendig ist. Für das Etablieren praktischer Methoden für das Sammeln, Speichern und Pflegen solcher Metadaten wird noch hoher Arbeitsaufwand erwartet.

Für verschlüsselte und komprimierte Objekte wird ein Zwiebelmodell mit Kompositionsebenen (composition level) eingeführt, um die Reihenfolge der Interpretationsschritte anzugeben und um Erhaltungsmetadaten genau den entsprechenden Ebenen zuzuordnen. Verpackungsbeziehungen, wie das Enthaltensein von Bitströmen und Dateiströmen in Dateien oder wie bei gepackten Dateien (ZIP, tar), sind trotz Ähnlichkeit unterschiedlich zu diesem Zwiebelmodell, weil die enthaltenen und die umfassenden Komponenten jeweils eigenständige Objekte sind.

Die Unveränderlichkeit (Fixity), Integrität (Integrity) und Authentizität (Authenticity) von Objekten nehmen einen hohen Stellenwert ein. Unveränderlichkeit bedeutet den Ausschluss unautorisierter oder undokumentierter Änderung von Objekten. Als Schlüsselindikatoren für Integrität werden die Identität und Validität (Übereinstimmung mit einer Spezifikation) einer Datei angeführt. Bezüglich der Entität Repräsentation werden die Vollständigkeit und richtige Bezeichnung der Dateien auf Basis zugehöriger struktureller Metadaten angeführt. Authentizität beschreibt den Sachverhalt, dass ein Objekt auch das ist, was es vorgibt zu sein. Dies bezieht sich sowohl auf die Quelle als auch auf die Integrität eines digitalen Objekts.

Bezüglich digitaler Signaturen orientiert sich PREMIS bei der Benennung und Strukturierung semantischer Einheiten an XMLDsig (XML Signature Syntax and Processing). Eine vollständige Übernahme scheidet wegen einer zu spezifischen Kodierung und Validierung aus.

[19] Hier könnte man auf Basis der Erkenntnisse aus der objektorientierten Modellierung bezüglich der Spezialisierung deutlich mehr Klarheit schaffen.
[20] Eine kompakte und etwas tiefer gehende Darstellung von Erhaltungsstrategien findet sich z.B. in [Borg2006].

Des Weiteren widerspricht die dortige Art der Signierung, welche eine Zusammenfassung von Datenobjekten (Dateien bzw. Bitströme) erlaubt, der Philosophie von PREMIS[21].

Im Gegensatz zur ersten Version enthält PREMIS nun Mechanismen zur formalisierten (kontrollierten) Erweiterung. Hierzu werden vordefinierte semantische Einheiten bereitgestellt, die explizit als Erweiterungen gekennzeichnet sind. Die Erweiterungen beziehen sich im Wesentlichen auf semantische Einheiten für die Entität Objekt; im Einzelnen sind dies: Signifikante Eigenschaften, Objektcharakteristiken, erzeugende Anwendung, Umgebung und Information zur Signatur. Für die Entität Ereignis können die Ergebnisdetails erweitert werden. Schließlich ist eine Erweiterung der Entität Rechte möglich.

PREMIS im Kontext von Grid/eScience

Eine besondere Stärke von PREMIS liegt darin, erhaltungsspezifische Aspekte zu identifizieren und zu strukturieren, was u.a. dazu beiträgt, Metadatenschemata modularer zu gestalten. Innerhalb von PREMIS wird die Modularität erreicht durch die Einführung von Entitätstypen bzw. -untertypen und zugehörige typisierte Beziehungen zwischen den Instanzen sowie durch semantische Einheiten, die entsprechend ihrer Logik zusammengefasst sind. Die Elemente von PREMIS erlauben auch die Beschreibung komplexer Szenarien der Langzeitarchivierung. Die Verwendung von Ereignissen und Ableitungsbeziehungen gestattet z.B. die Beschreibung komplexer, mehrstufiger und ggf. verzweigter Transformationen. Die neu eingeführten Erweiterungsmöglichkeiten unterstützen nun auch die systematische Einbindung sehr spezieller Repräsentationsinformation und Preservation Description Information (PDI) entsprechend dem OAIS-RM in einen einheitlichen Rahmen. Natürlich obliegt es den einzelnen Communities die entsprechenden Erweiterungsmodule zu entwickeln oder zu benennen. Zur Beschreibung der Umgebung könnten z.B. Schemaelemente übernommen werden, wie sie in der JSDL (Job Submission Description Language) des OGF zur Anforderungsbeschreibung von Rechenaufträgen an Computer-Ressourcen spezifiziert sind. Vorteilhaft ist, dass die dort beschreibbaren Anforderungen sich nicht nur auf Grid-Umgebungen beschränken.

Mit PREMIS steht insgesamt ein sehr flexibles Grundgerüst zur Verfügung, das jedoch entsprechend den Anforderungen – über ein kontrolliertes Vokabular hinausgehend – als Gesamtes profiliert werden sollte.

7.1.2 NLNZ / LMER

Einen weiteren Ansatz, der sich auf Erhaltungsinformation konzentriert, wurde nach etlichen Vorarbeiten 2003 von der Nationalbibliothek von Neuseeland (NLNZ) veröffentlicht [NLNZ2003]. Implementierbarkeit und Berücksichtigung von Standards und internationaler Aktivitäten waren ein Anliegen der Autoren. Der Bezug zu OAIS ist deutlich erkennbar und übersichtlich dokumentiert und auf der Detailebene werden z.B. Metadatenelemente aus NISO Z39.87 übernommen. Bestimmte Aspekte der Langzeiterhaltung werden von der minimalen Metadatenmenge ausgeschlossen und dem Datenmanagement zugeschlagen. Das Management von Rechten wurde, von ganz spezifischen Aspekten abgesehen, von vornherein ausgeschlossen.[22]

[21] Diese Einschränkung kann auch nicht durch den PREMIS-Erweiterungsmechanismus für die Entität Rechte umgangen werden, da die Erweiterung nicht einer PREMIS-Repräsentation zugeordnet werden darf.
[22] Für technische Aspekte, die einen Zugriff verhindern und einen rechtlichen Grund haben können, ist ein Attribut vorgesehen.

Den Kern des Metadatenschemas bildet der Erhaltungs-Master (Preservation Master), der die bestmögliche Erstellung eines erhaltungsfähigen Objekts repräsentiert basierend auf einem Original, welches dem Archiv übergeben bzw. welches vom Archiv akquiriert wurde. Das NLNZ-Modell nimmt an, dass der Preservation Master im Laufe der Zeit zu modifizieren oder gar zu ersetzen sei. Dabei existiert jeweils nur ein Master; er stellt den Gegenstand maximaler Erhaltungsanstrengung dar im Gegensatz zu eventuell alternativen Manifestationen z.B. für die Auslieferung. Demzufolge bildet der Master das Bezugsobjekt für die Erhaltungsmetadaten. Neben diesem logischen Objekt (Digital Object), das auch die Referenz für deskriptive Metadaten bildet, werden noch drei weitere Entitäten in das Modell aufgenommen, nämlich Datei (File), Metadatenänderung (Metadata Modification) und Prozess (Process). Die Entität Datei beschreibt die technischen Aspekte von Dateien – als unterste (feinste) Ebene eines digitalen Objekts – wie Dateiformat und Dateiformatversion. Die Typen Bild (Image), Audio, Video und Text enthalten Attribute für Detailinformationen während für Datensets (Datasets) und Systemdateien (System File) keine Attribute spezifiziert sind. Prozesse beschreiben ändernde und nicht-ändernde Operationen auf dem Master und umfassen u.a. benutzte Soft- und Hardware sowie die Einzelschritte eines Prozesses. Die NLNZ beschränkt sich hierbei auf Änderungen, die innerhalb des Archivs stattfinden. Die Entität Metadatenänderung dient der Aufzeichnung von Änderungen bzw. Verbesserungen an den Metadaten.

Für das digitale Objekt ist eine Einteilung in einfache Objekte (simple object), komplexe Objekte (complex object) und Objektgruppe (object group) vorgesehen. Das einfache Objekt wird genau durch eine Datei repräsentiert während das komplexe Objekt aus mehreren Dateien besteht, die durch Abhängigkeiten gekennzeichnet sind, die für das „Funktionieren“ des logischen Objekts nötig sind. Eine Detaillierung dieser logischen Abhängigkeiten nimmt das Modell nicht vor. Die Objektgruppe beschreibt hingegen nur eine Menge von Objekten und dient wohl eher der physischen Partitionierung von Objekten bzw. einer Normalisierung[23] bei der Zuordnung von Metadaten.

LMER

LMER (Long-term Preservation Metadata for Electronic Resources) ist eine deutsche Entwicklung, die sich am neuseeländischen Ansatz orientiert [DDB2005]. Eine Reihe von Verbesserungen kennzeichnet LMER; im Wesentlichen: 1) Das Schema beschränkt sich einerseits auf einen Kern von Elementen, welche für jeden Dokumententyp gültig sind, und führt anderseits neue Elemente ein, um Formate präziser beschreiben zu können. 2) LMER erlaubt die Angabe von Referenzen, die der Kopplung zu externen Formatverzeichnissen dienen. 3) LMER besitzt einen modularen Aufbau, der für die Repräsentation von Informationspaketen geeignet ist.

LMER wird im Universellen Objektformat im System kopal genutzt, um METS mit Erhaltungsmetadaten zu versorgen, und ist im Rahmen der Registrierung von METS-Profilen bei der Library of Congress als *Erweiterungsschema* verzeichnet.

LMER im Kontext von Grid/eScience

Der Vorteil von LMER ist in der Konzentration auf Kernaspekte der Langzeitarchivierung und in der hohen Konkretisierung bestimmter Teile zu sehen. Im Bereich Grid/eScience könnten sich jedoch Nachteile aus der mangelnden (Vor-)Strukturierung von Konzepten

[23] hier im Sinne einer Redundanzvermeidung

zur Modellierung komplex aufgebauter Objekte und ihrer Abhängigkeiten ergeben. Als negativ für die Beherrschung der ausgeprägten Heterogenitäten im eScience-Bereich könnten sich die fehlenden expliziten Profilierungs- und Erweiterungsmechanismen erweisen. Die Behandlung solcher Fälle war aber auch kein Entwurfsziel von LMER. Eine alleinige Verwendung zur Beschreibung von Erhaltungsinformation wird daher im Allgemeinen nicht ausreichend sein. Das zugehörige XML-Schema kann jedoch in andere Schemata eingebunden werden, und umgekehrt kann LMER beliebige XML-Metadaten für einzelne Dateien aufnehmen.

7.1.3 Formatverzeichnisse

Formate stellen in mehr oder großem Umgang die Repräsentationsinformation in einem Informationspaket dar. Um eine ständige Neudefinition von Formatbeschreibungen in Metadatenschemata und vielfache Wiederholungen der zugehörigen Daten zu vermeiden, bietet sich der Aufbau von allgemein zugänglichen Verzeichnissen an. Wegen ihrer Bedeutung für die Vollständigkeit von Informationspaketen ist der Zugriff auf solch einen Dienst langfristig und in der geforderten Qualität sicherzustellen. So müssen z.B. Formatspezifikationen lesbar und verständlich gehalten werden. Für die Verwaltung, ob zentral oder dezentral, ist ein definiertes Verfahren vorzugeben und aus Gründen der Vertrauenswürdigkeit zu veröffentlichen. Der Aufbau von Formatverzeichnissen bietet auch einen geeigneten Rahmen, den Begriff Format deutlicher zu fassen. Vorteilhaft für ein Formatverzeichnis wäre es, Informationen über konkrete Profilierungen zu den jeweiligen Formaten anzubieten.

PRONOM

Der Entwicklung von PRONOM[24] liegt die Erkenntnis des Nationalarchivs von Großbritannien (The National Archive – TNA) zu Grunde, dass für die Wiederherstellung von Information der verlässliche Zugriff auf technische Information über die im Archiv vorhandenen elektronischen Aufzeichnungen erforderlich ist. Das ursprünglich als interne Hilfe geplante technische Verzeichnis steht nun im Web für die Langzeitarchivierungs-Community zur Verfügung und ist zwischenzeitlich mit einer stattlichen Anzahl von Daten gefüllt. Volle Kompatibilität zu anderen technischen Verzeichnissen u.a. zu GDFR ist vorgesehen.

Kernentitäten des Informationsmodells PRONOM 4 [TNA2005] sind die technische Komponente (Technical Component), Akteuer (Actor), Dokumentation (Documentation), Rechte (Intellectual Property Rights – IPR) und Identifikator (Identifier) und entsprechende Beziehungen. Die Entität Komponente ist in weitere unterteilt: das Dateiformat (File Format), die Software-Komponente, die Hardware-Komponente, das Speichermedium, die Zeichenkodierung (Character encoding), der Typ der Kompression (Compression Type), die interne Signatur (Internal Signature), Bytefolge (Byte Sequence) (hier die Repräsentation der internen Signatur), die externe Signatur, der Name (zur Benennung technischer Komponenten einschließlich Version). Des Weiteren können den technischen Komponenten Klassifizierungsschemata zugeordnet werden. Außerdem ist eine Einteilbarkeit in Familien vorgesehen. Schließlich ist noch die Identifizierung und Beschreibung von Referenzdateien Bestandteil des Modells. Mit den Beziehungen kann nun modelliert werden, welche Leistungen eine bestimmte Software-Komponente bezüglich eines Dateiformats

[24] http://www.nationalarchives.gov.uk/pronom

erbringen kann, wie z.B. das Anzeigen oder das Extrahieren von Metadaten, oder welche Anforderungen von technischen Komponenten an Hardware bzw. Software bestehen.

Zur dauerhaften Identifizierung von Entitäten wurde eine Schema namens PUID (PRONOM Persistent Unique Identifier) entwickelt. Auf die Anreicherung oder Anreicherbarkeit mit zusätzlicher Information sowie auf eine Dereferenzierbarkeit zur Lokalisierung wird verzichtet. Bisher werden Identifikatoren für Dateiformate vergeben (die „einfache" Syntax sieht hier eine Differenzierung nach Entitätstypen vor), wobei jeweils das spezifischste Format zu Grunde liegt und Versionen eine eigene Identität rechtfertigen. Ausdifferenzierungen (Varianten) innerhalb einer Dateiformatversion (z.B. die Nutzung verschiedener Kompressionsverfahren innerhalb des Formats TIFF), unterschiedliche Kodierungsschemata für Zeichen (Character encoding) und die Anordnung der Bytes (Byte Order) bilden in PRONOM keine eigene Version. Bei Containerformaten mit variablen Inhalten kann man sich auf die IDs der eingeschlossenen Formate abstützen. Eine Klassifizierung der Formate und die explizite Definition von Beziehungen ist nicht Bestandteil des Identifikationsschemas. PUIDs können als info-URI-Schema [RFC4452] ausgedrückt werden; die Einbindung in einen Resolver-Dienst wird noch nicht angeboten.

GDFR

Die Harvard University Library (HUL) entwickelt in Zusammenarbeit mit dem Online Computer Library Center (OCLC) ein weiteres Verzeichnis, um Repräsentationsinformation im Sinne von OAIS zur Interpretation digitaler Objekte nachhaltig bereitzustellen. Das geplante GDFR (Global Digital Format Registry)[25] zeichnet sich durch einen verteilten Ansatz aus, der die dezentrale, aber koordinierte Pflege und Nutzung erlaubt sowie als Gesamtsystem Robustheit gegenüber lokalem, durch Kurzfristigkeit geprägtem Handeln bietet. Neben der Entwicklung eines Datenmodells zur Beschreibung digitaler Formate ist also der Entwurf entsprechender Regeln, Dienste und Protokolle vorzunehmen, um die Qualität der Inhalte zu sichern. Um beispielsweise die Beschreibung von Formaten mit lokaler Bedeutung zu ermöglichen bzw. um eine schnelle Verbreitung von Beschreibungen aus offensichtlich verlässlichen Quellen nicht zu behindern, sind Beiträge ohne Review-Prozess zugelassen. Diese Regel orientiert sich am differenzierten Vorgehen bei der Registrierung von MIME-Typen durch IETF einerseits und durch Hersteller/Personen anderseits. Vorhandene Standards finden bei der Entwicklung Berücksichtigung wie z.B. OASIS/ebXML Registry Information Model und ANSI X3.285[26] beim Dienstemodell sowie ISO/IEC 11179 Metadata Registries [ISO/IEC 11179] und wiederum OASIS/ebXML beim Datenmodell. Angesichts der zunehmenden Komplexität von Formaten kann sich die Schaffung formaler Grundlagen, wie in GDFR in Angriff genommen, als großer Vorteil erweisen, insbesondere dann, wenn Erhaltungsmaßnahmen automatisiert durchgeführt werden sollen.

Der Formatbegriff beruht auf einem formalen Modell, welches aus vier Entitäten besteht, nämlich dem Informationsmodell (IM), dem semantischen Modell (SM), dem syntaktischen Modell (CM) und dem serialisierten Bytestrom (SB). Das IM umfasst eine Klasse austauschbaren Wissens und das SM eine Menge semantischer Informationsstrukturen, die die Bedeutung des IM realisieren können [HUL2007]. Das CM besteht aus einer Menge von syntaktischen Dateneinheiten, die das SM ausdrücken können. Der SB manifestiert schließlich in Form einer Bytesequenz das CM. Zwischen den Entitäten, die auf unterschiedliche Abstraktionsebenen angesiedelt sind, können nun Abbildungen definiert wer-

[25] http://hul.harvard.edu/gdfr
[26] X3 ist jetzt incits (InterNational Committee for Information Technology Standards)

den. Diese drei Abbildungen zwischen den vier Abstraktionsebenen sind geeignet, ein Format zu charakterisieren. Auf dieser formalen Grundlage können schließlich Beziehungen zwischen Formaten klarer definiert werden, im Fall von GDFR nämlich Erweiterung (Extension), Einschränkung (Restriction), Abänderungen (Modification), Enthaltensein (Containment), Gleichheit (Equivalence), Version (Version) und schließlich noch Affinität (Affinity), welche eine bestimmte technische Ähnlichkeit ausdrückt, die hier, wie die Versionsbeziehung auch, als nicht strikt formal fassbar eingestuft wird.

Zum Umfang von GDFR gehört auch die eindeutige und persistente Identifikation bestimmter Ressourcen wie registrierte digitale Formate, Klassifikationen der Formate oder Knoten in einem GDFR-Netzwerk [HUL2006]. Auf die Anreicherung oder Anreicherbarkeit mit zusätzlicher Information sowie auf eine Dereferenzierbarkeit zur Lokalisierung wird verzichtet. Die Identifikatoren können als info-URI-Schema [RFC4452] ausgedrückt werden.

7.1.4 Format- und Schemabeschreibungsmittel

Formate und Schemata sind unbestritten Grundlagen, um digital repräsentierte Inhalte effizient zu interpretieren. Anwendungsfälle, die sich durch einen hohen Grad an Allgemeingültigkeit auszeichnen, sind befriedigend durch standardisierte Formate abgedeckt. Steigen jedoch die Ansprüche an die Darstellung der Inhalte und die beschreibenden Information, entsteht Bedarf, individuelle Formate zu definieren. Hohe Spezialisierung, mangelnde Zeit oder (noch) unbekannte Adressaten können einen Standardisierungsvorgang als ungeeignet erscheinen lassen. Einheitliche und standardisierte Beschreibungsmittel können jedoch den Austausch in einer verteilten, heterogenen und dynamischen Umgebung erleichtern. Auch aus Gesichtspunkten der Langzeitarchivierung wäre es wünschenswert, über solche Mittel zu verfügen, um Formatbeschreibungen und somit Repräsentationsinformation zu vereinheitlichen, was ein technisches Monitoring und eine Automatisierung von Erhaltungsmaßnahmen erleichtern würde.

Datendefinitionssprachen

So genannte Datendefinitionssprachen (Data Definition Language – DDL) werden häufig im Kontext von Datenbanksystemen genutzt, um Datenstrukturen und weitere Elemente, die innerhalb eines Systems erlaubt sind, zu definieren. Je nach Datenmodell stehen bestimmte logische Konstrukte zur Verfügung, die definiert werden können, um die semantischen Ebene zu repräsentieren. Mittels so genannter Datenmanipulationssprachen (Data Manipulation Language – DML) können Datenobjekte eingefügt, geändert oder gelöscht werden. Die Beschränkung auf bestimmte logische Elemente zu Lasten der Ausdrucksfähigkeit erlaubt einen überschaubareren Sprachumfang mit entsprechender Vereinfachung der Implementierbarkeit, der Optimierbarkeit der Ausführung und letztlich der Handhabung für den Endnutzer und Anwendungsentwickler.

SQL DDL

Im Bereich der relationalen Datenbanken hat sich SQL als Datendefinitions- und Datenmanipulationssprache durchgesetzt. In relationalen Datenbanken bilden Relationen (im mathematischen Sinne), auch Tabellen genannt, die Kernelemente. Ein Tupel (Zeile) repräsentiert i.d.R. jeweils die Instanz einer Entität oder eine Beziehung zwischen Instanzen) und die Spalten die Attribute mit den entsprechenden Werten der Instanzen. Weitere wichtige Elemente sind Sichten (View), die mittels eines SQL DML-Befehls eine virtuelle Re-

lation bilden, die bei Bedarf auch materialisiert, also physisch persistent gespeichert werden kann. Mit so genannten Beschränkungen (Constraint) können sowohl einfache strukturelle Abhängigkeiten zwischen Tupeln (Zeilen), wie die Einmaligkeit eines Wertes innerhalb einer Spalte, als auch zwischen Tupeln unterschiedlicher Relationen, wie Referenzen basierend auf der Wertegleichheit bestimmter Felder, definiert werden. Die Definition von Attributen als Schlüssel kann als eine spezielle Beschränkung gesehen werden. Außerdem können komplexere Beschränkungen beschrieben werden, die den Sprachumfang von SQL, ggf. unter Einbeziehung einer prozeduralen Spracherweiterung, ausschöpfen. Um die Ausdrucksstärke von SQL zu erhöhen, führte der Standard im Laufe der Zeit eine Reihe von Erweiterungen ein. Selbstdefinierbare und strukturierte Datentypen führen beispielsweise zu einer kompakteren Datenstruktur und das Konzept der Vererbung (strukturelle Objektorientierung) vermindert die Kluft zur objektorientierten Programmierung.

Obwohl zu den Zielen von Datenbankmanagementsystemen die Abstraktion von der physischen Ebene gehört, finden sich in SQL DDL auch Elemente, die dieser Ebene zuzuordnen sind, wie die Bestimmung des initialen Speicherplatzbedarfes für eine Relation oder Anweisungen zu Caching-Strategien.

Üblicherweise werden alle Definitionen in einem besonderen Bereich der Datenbank (Data Dictionary, auch Schemakatalog genannt) abgelegt, der mit den entsprechenden Privilegien abgefragt werden kann. Eine Inspektion eines Datenbankschemas ist somit möglich. Ein vollständiges semantisches Modell wird sich daraus jedoch ohne weitere Beschreibungsmittel in der Regel nicht herleiten lassen. Zur Vereinheitlichung der bisher sehr produktspezifischen DDs in SQL-Datenbanken dient der ISO-Standard SQL Schemata [ISO/IEC 9075-11] als Teil der SQL Sprachdefinition.

XML Schema

XML erfreut sich bei der Modellierung und beim Austausch von Daten größter Beliebtheit. Mittels Document Type Definitions (DTD), die Teil der W3C Empfehlung für XML sind, können Regeln vorgegeben werden, denen ein konkretes XML-Dokument zu gehorchen hat. Als Schemabeschreibungssprache kann DTD im Wesentlichen Elemente und Attribute einschließlich ihrer Anordnung deklarieren und einige grundlegende Anforderungen an Inhalte definieren. Für viele Anwendungen ist der Grad an Formalisierung und Modularisierung von XML-Dokumenten mittels DTD nicht ausreichend, insbesondere wenn die Dokumente maschinell weiterverarbeitet werden sollen und hierfür an die Typen von Programmiersprachen zu binden sind. Mit W3C XML Schema steht eine deutlich mächtigere Schemabeschreibungssprache zur Verfügung. Ein umfangreiches System zur Definition von Datentypen basierend auf vorgegebenen Grundtypen erleichtert die Sicherung der Datenintegrität und Zuverlässigkeit zugehöriger Operationen. Zusätzliche Regeln zur Einschränkung wie Wertebereiche, Schlüsseleigenschaften oder reguläre Ausdrücke erlauben eine weitere Präzisierung. Die Einbindbarkeit und gleichzeitige Anpassbarkeit fremder Schemata unterstützt eine Modularisierung und Wiederverwendung. Im Gegensatz zu DTD kann XML Schema mit Namensräumen umgehen. Da Schemadefinitionen selbst XML-Dokumente sind, können sie mit allen XML-fähigen Werkzeugen bearbeitet werden, was einen Beitrag zur Vereinfachung und Vereinheitlichung der Werkzeuge liefert.

Schematron

Neben DTD und XML Schema existieren noch weitere Schemabeschreibungssprachen wie RELAX NG [ISO/IEC 19757-2:2003]. An dieser Stelle soll Schematron [ISO/IEC 19757-

3:2006] etwas näher erläutert werden, da es sich durch seinen regelbasierten Ansatz unterscheidet. Dieser deklariert kein Schema, sondern definiert mittels Pfadausdrücken von W3C XPath und Ausdrücken der Transformationssprache W3C XSLT Muster, die von einem XML-Dokument (Instanz) einzuhalten sind. Es wird also getestet, ob bestimmte Eigenschaften erfüllt werden oder nicht. Anforderungen, wie sie eine Schemadefinitionssprache vorschreiben, z.B. welche Attribute ein Element enthält oder welche Unterelemente ein Element umfasst, können durch die Formulierung von Regeln (Tests) in Schematron, die auf ein XML-Dokument angewendet werden, geprüft werden. Der Vorteil von Schematron gegenüber XML Schema liegt in der Prüfbarkeit wechselseitiger Abhängigkeiten (co-occurences) von Elementen und Attributen ggf. über verschiedene Dokumente hinweg. Ist z.B. ein Objekt durch ein entsprechendes Attribut als Buch klassifiziert, kann überprüft werden, ob ein Attribut existiert, das die ISBN beinhaltet. Entsprechend den Möglichkeiten von XPath sind sehr komplexe Regeln deklarierbar, womit eine flexiblere (dynamischere) Gestaltung von Schemata gegeben ist als bei rein grammatikbasierten Ansätzen. Die Anpassung von Schemata an individuelle Bedürfnisse, die Aktualisierung von Schemata oder die Prüfung von Dokumenten ohne Schemata oder mit schlecht strukturierten Schemata wird deutlich erleichtert. Gewollte oder ungewollte Redundanzen sind leichter in den Griff zu bekommen, sofern sich die Abhängigkeiten mit XPath und XSLT beschreiben lassen. Die Architektur von Schematron erlaubt sowohl den alleinigen Einsatz als auch die Kombination mit anderen Sprachen wie Relax NG oder dem aktuell dominierenden XML Schema.

Angesichts des hohen Nutzungsgrades von XML sowohl für Daten als auch für Metadaten, ist dem Entwurf und der Kontrolle von XML-Dokumenten höchste Aufmerksamkeit zu widmen. Sprachen zur Deklaration und zur Validierung von Schemata spielen hierfür eine elementare Rolle zur Umsetzung von Qualitätsanforderungen. Fehlende Entwurfs- und Anwendungsregeln können jedoch wegen der Flexibilität und Mächtigkeit der Sprachen zu einer Gefährdung von Qualitätszielen führen.

ASN.1

Der Standard ASN.1 (Abstract Syntax Notation One) hat seine Wurzeln in der Telekommunikation, mit dem Bedürfnis, Nachrichten plattformunabhängig und höchst zuverlässig austauschen zu können. Hierfür erlaubt ASN.1 die programmiersprachenunabhängige Spezifikation von Datenstrukturen. Aufbauend auf primitiven Typen können ähnlich wie in imperativen Programmiersprachen komplexe Typen definiert werden. Die Zusammenfassbarkeit von Typen in Modulen unterstützt die Einheitlichkeit in größeren Anwendungen. Ein wesentliches Element ist die standardisierte Abbildung der abstrakt notierten Datenstrukturen auf eine Bitsequenz. Um verschiedene Randbedingungen zu berücksichtigen, beinhaltet der Standard unterschiedliche Verfahren der Kodierung, um beispielsweise eine möglichst kompakte Form zu erreichen oder um durch Eindeutigkeit die digitale Signierbarkeit in verteilten und heterogenen Umgebungen sicherzustellen. Wegen dieser Eigenschaften ist ASN.1 die Basis etlicher Internetstandards wie X.509, LDAP oder SMTP (Simple Mail Transfer Protocol).

Die Anbindung an die XML-Welt ist ebenfalls Gegenstand der Standardisierung geworden. Es bietet sich an, ASN.1 Datenstrukturen nach XML zu übersetzen, um die Vielzahl der hierfür vorhandenen Werkzeuge zu nutzen. Ein weiterer Vorteil ist die Steigerung der Speicher- und Verarbeitungseffizienz durch eine standardisierte binäre statt textuelle Repräsentation von XML-Infosets, was sich insbesondere bei großen Objekten auszahlt [I-

SO/IEC 24824-1:2007]. Dieser Mechanismus wird u.a. zur interaktiven Verarbeitung von 3D-Objekten genutzt, wie im Nachfolgestandard zu VRML spezifiziert [ISO/IEC 19776-3:2007].

In den oben geschilderten Anwendungen dient ASN.1 der Realisierung von *Common Services*. Die Spezifikationsmöglichkeiten erlauben auch die Modellierung und Implementierung von anwendungsbezogenen Datenstrukturen in stark verteilten Umgebungen mit vielen Systemen zur Erfassung, Speicherung und Auswertung von Daten wie in der Bioinformatik. NCBI[27] nutzt beispielsweise ASN.1 zur Speicherung und zum Retrieval von Daten der Biotechnologie. Je nach Endanwendung werden aus diesem „Kernformat" für den Nutzer geeignete Formate erzeugt.

Data Format Definition Language (DFDL)

Einen weitergehenden Ansatz verfolgt die DFDL-Arbeitsgruppe innerhalb des Open Grid Forums. Sie hat zum Ziel, eine universelle Beschreibungsmöglichkeit für „alle möglichen" Datenformate zu entwickeln. Dies wird als Voraussetzung gesehen, um individuelle Formate zu definieren und in einer typischen Grid-Umgebung ohne vorherige Vereinbarung auszutauschen. Es können also Formate eingeführt werden, die der aktuellen Problemstellung angepasst sind, da sie beispielsweise einen geringen Speicherbedarf erfordern, eine optimierte Bearbeitung erlauben oder sich aufgrund einer vorhergehenden Anwendungsentwicklung als nützlich erwiesen haben. Binärformate sollen ebenso beschreibbar sein wie textbasierte Formate, unabhängig davon ob sie für kommerzielle oder wissenschaftliche Anwendungen gedacht sind. DFDL ist jedoch nicht als Transformationssprache für Formate gedacht.

Im Folgenden werden einige Kerngedanken von DFDL skizziert [OGF-DFDL 2008]. Um die Selbsterklärung eines Formats (zumindest bezüglich der syntaktischen Ebene) zu erreichen, wird es in Form eines eingeschränkten XML-Schemas spezifiziert, welches sowohl dem Parsen als auch der Formatierung dient. Dabei wird eine Trennung zwischen logischem Modell und physischer Repräsentation vorgenommen (vgl. hierzu auch das Formatmodell in GDFR). Zur Modellierung der logischen Ebene dienen ausgewählte Konstrukte aus XML Schema und die Beschreibung der physischen Ebene erfolgt mittels so genannter Format-Annotationen, die einen Großteil von Formatbeschreibungen abdecken kann. Zahlen als logische bzw. abstrakte Konstrukte können beispielsweise auf vielfältige Weise sinnvoll repräsentiert werden. Im Laufe der Zeit sind zahlreiche binäre standardisierte und nicht-standardisierte (maschinenspezifische) Formate entstanden; ebenso sind etliche textuelle Darstellungen in Gebrauch z.B. für die Bezeichnung von Exponenten, zur Trennung von Zahlenbestandteilen oder für die Anzahl führender Nullen. Für eine allgemeine Formatbeschreibungssprache wäre hier die Einschränkung, die durch XML Schema für die Repräsentation von Zahlen vorgegeben sind, nicht geeignet. Ähnliches gilt für die Darstellung von Text (string), die sich einer Reihe von Alphabeten bedienen kann. Dies sind natürlich nicht die einzigen Typen von Repräsentationsalternativen. Wird z.B. für die Kodierung jeweils mehr als Byte benötigt, muss für die richtige Interpretation deren Reihenfolge (Byte order) bekannt sein. Um Formate zu beschreiben, sind weitere Beschreibungsmittel nötig. So werden aus einfachen Datentypen komplexere gebildet (das Format kann als Datentyp betrachtet werden), wobei sich die Bildungsregeln abstrakt mit ausgesuchten XML-Schema-Konstrukten beschreiben lassen. Für die physische Repräsentation

[27] National Center for Biotechnology Information – National Library of Medicine, http://www.ncbi.nlmnih.gov

steht wieder eine Vielzahl von Optionen offen. So können beispielsweise die Elemente einer Sequenz durch ein bestimmtes Zeichen getrennt und die Sequenz als Ganzes durch ein weiteres spezielles Zeichen abgeschlossen sein. Eine solche Repräsentation ergibt bei stark schwankender Länge der Elemente eine sehr kompakte Repräsentation, die aber bei großen Vektoren und bei einem gewünschten direkten Zugriff auf ein bestimmtes Element einen unakzeptablen Aufwand erfordern kann. Jedoch erleichtert die Trennung von Logik und physischer Repräsentation die Definition alternativer Formate. Im obigen Beispiel der Sequenz wäre u.a. zu definieren, welche Zeichen zur Trennung der einzelnen Elemente dienen und mit welchem Zeichensatz diese ggf. zu kodieren wären, sofern die standardmäßige Darstellung in UTF-8 (8-bit Unicode Transformation Format) nicht in die Anwendungsumgebung passt. Solche Angaben werden in den entsprechenden Attributen der Format-Annotation für Konstrukte aus XML Schema untergebracht. Besondere Aufmerksamkeit ist dabei der Regelung der Geltungsbereiche (Scope) solcher Annotationen zu widmen; so kann z.B. in einer weiteren Sequenz ein anderes Trennzeichen geeignet oder aufgrund einer Strukturierung, die mit Positionen in der Bitsequenz arbeit, sogar überflüssig sein.

Format-Spezifikationen können eine äußerst komplexe Form annehmen. Etliche Punkte für die Standardisierung einer Datenformatbeschreibungssprache sind deshalb noch in der Diskussion bzw. vorerst zurückgestellt. Eine detaillierte und erfreulich offene Diskussion findet sich auf den Internet-Seiten der DFDL-WG.

Weitere Anstrengungen zur Standardisierung auf diesem Gebiet zählt Alan Powell vom OGF in einem Fortschrittsbericht[28] auf, nämlich:

Earth Science Markup Language (ESML)[29]

Auch hier liegt die Auffassung zu Grunde, dass sich innerhalb der Community kein standardisiertes Format erzwingen lässt. Auf der Basis des ESML-Schemas können Nutzer die Struktur ihrer Dateien beschreiben. Der Schwerpunkt liegt auf Binär- und ASCII-Darstellungen. Darüber hinaus gibt es Metadatenschemata zu den Formaten HDF-EOS und GRIB [Rama2003]. Mit Hilfe der ESML-Bibliothek und der Beschreibungsdatei können nun Anwendungen Dekodierung vornehmen. Zusätzlich zu den syntaktischen Informationen können mit Hilfe von Marken die Verbindungen zu Ontologien aus Fachgebieten hergestellt werden, um die inhaltliche Interpretation maschinell zu unterstützen.

Extensible Scientific Interchange Language (XSIL)[30]

Bei diesem Ansatz können Container definiert werden, die Daten in Form von Feldern (nahe dem Konzept der Arrays in C und Fortran) und Tabellen sowie jeweils deren Beschreibungen enthalten können. Alternativ sind Verweise zugelassen, wobei auch auf Dateien mit binären Inhalten gezeigt werden kann. Die Container können zum Aufbau hierarchischer Strukturen selbst wieder Container enthalten.

[28] Data Format Description Language (DFDL) – Progress Update, Abschnitt Extra Information, https://forge.ggf.org/sf/docman/do/downloadDocument/projects.dfdl-wg/docman.root.current_0/doc15029/1
[29] http://esml.itsc.uah.edu/index.jsp
[30] http://www.cacr.caltech.edu/SDA/xsil

Binary Format Description (BFD) language[31]

Hierbei handelt es sich um eine Erweiterung von XSIL. So können z.B. durch die Einführung der Kontrollstruktur *if* auf Basis entsprechender Parameter geeignete Formatvarianten ausgewählt werden.

Binary XML Description Language (BinX)[32]

Die Entwickler sehen den Schwerpunkt ihrer Arbeit auf der Beschreibung binärer Formate insbesondere für numerische Anwendungen, doch eine Ausweitung auf Textformate ist vorgesehen. Als komplexe Datentypen können Datensätze (Records), Felder (Arrays) und Vereinigungen (Union)[33], wiederum in Anlehnung an die Datentypen imperativer Programmiersprachen, definiert werden.

Insgesamt sind diese Ansätze deutlich weniger generell als bei DFDL. Der Umfang elementarer Typen und der Bildungsregeln für komplexe Typen und Formate ist eingeschränkt und orientiert sich vorwiegend an naturwissenschaftliche Anwendungen. Obwohl die zu definierenden Konzepte sich sehr ähneln, unterscheiden sich die Konventionen zur Beschreibung deutlich.

Unter Langfristgesichtpunkten ist nach heutigem Kenntnisstand eine Anhäufung derartiger Beschreibungen als ungünstig zu betrachten, so dass eine Standardisierung Vorteile bietet. Andererseits ist ein relativ genereller Ansatz wie in DFDL relativ komplex, obwohl er gar nicht alle Schemabeschreibungen ablösen möchte.

7.2 Beschreibung von Informationspaketen (Packaging Information)

Die Bildung von archiv- und austauschfähigen Informationspaketen im Sinne von OAIS ist eine weitere Kernaufgabe der Langzeitarchivierung. Die logische Zusammengehörigkeit verteilt gespeicherter Daten und die Sicherstellung ihrer Interpretierbarkeit der damit verbundenen Ordnungsstrukturen und Inhalte sind Grundvoraussetzungen für die Nachnutzbarkeit von Inhaltsdaten. Dem Standard METS wird im Folgenden breiter Raum eingeräumt, weil er national und international eine große Bedeutung für digitale Bibliotheken und für die Langzeitarchivierung gewonnen hat.

7.2.1 METS

Diese Spezifikation dient sowohl der Verwaltung digitaler Objekte innerhalb eines Archivs bzw. einer digitalen Bibliothek als auch dem expliziten Austausch von digitalen Objekten. Sicherung der Integrität verteilt gespeicherter Inhalte sowie die gezielte und dauerhafte Zuordnung von Metadaten werden unterstützt. METS definiert als Speicher- und Transfersyntax von wenigen Ausnahmen abgesehen selbst keine Metadaten; die Spezifikation unterstützt jedoch die Einteilung von Metadaten in verschiedene Kategorien. Somit kann METS den Containerformaten zugeordnet werden. Moore bezeichnet METS in [Moor2003a] sogar als eine Variante einer digitalen Ontologie, die auf Objektebene (i.W. Dateiebene) ansetzt. METS ist somit ein Mittel um ein Informationspaket im Sinne von OAIS aufzubauen. Schwerpunktmäßig sind mit METS die strukturellen Aspekte darstell-

[31] http://collaboratory.emsl.pnl.gov/sam/bfd
[32] http://www.edikt.org/binx
[33] im Sinne varianter Datensätze, die mit einem typisierten Diskriminator unterschieden werden können

bar, also die Beziehungen der Inhaltsobjekte untereinander als auch der Beziehungen zu Metadaten.

Die Entwicklung von METS wird von DLF (Digital Library Foundation) gesponsert und vom METS Editorial Board vorgenommen und koordiniert. Die LOC (Library of Congress) hat die Rolle der Maintenance Agency übernommen. Dort finden sich auch ausführliche Informationen einschließlich registrierter Profile. Darüber hinaus ist METS bei NISO formal registriert.

Die Verknüpfung von Inhaltskomponenten untereinander und mit Metadaten ist eine Stärke von METS. Dies ist vorteilhaft, weil in der Praxis unterschiedliche Formen der Repräsentation von Inhalten spezifische Dateiformate erfordern oder Inhalte aus verschiedenen Gründen auf mehrere Dateien aufgeteilt werden. Dateien, die Inhalte repräsentieren, können zu Gruppen zusammengefasst werden, die z.B. die unterschiedliche Nutzung widerspiegeln wie Vorschaubild oder Masterbild. Die Inhalte können extern in Dateien gehalten werden. Folgende Mechanismen der Referenzierung sind hierfür benennbar: ARK, URN, URL, PURL, HANDLE, DOI und OTHER, wobei der letzte Typ in einem eigenen Metadatenfeld beschrieben werden sollte. Außerdem können Inhaltsdaten inline genommen werden, wobei jedoch nicht in XML kodierte Inhalte eine Base64-Kodierung[34] erfordern.

Da den Dateibezeichnern IDs zugeordnet werden, können unter Vermeidung von Redundanzen logische und physische Strukturen bezüglich der Inhalte beschrieben werden. Die Elemente, die die Struktur beschreiben, nutzen hierfür die entsprechenden IDs. Sollen z.B. Dateien stets in einer bestimmten Reihenfolge oder Hierarchie erscheinen, wird dies durch die entsprechende Notation der Strukturelemente sichergestellt. Man erspart sich hiermit z.B. die Mühen, Unsicherheiten und Beschränkungen die aufträten, wenn man derartige Sachverhalte auf Dateisystemebene modellieren würde[35]. Ein weiterer Vorteil ist, dass beliebig viele Strukturen angelegt werden können und somit unterschiedlichste Sichten modellierbar sind. Durch eine explizite Verlinkung der Strukturelemente können auch Verbindungen nachgebildet werden, wie sie für Web-Seiten und Hypermedia-Objekte typisch sind. Darüber hinaus können auch Teile von Dateien direkt und entsprechend dem Dateiformat in einer typisierten Weise angesprochen werden. Im primitivsten Fall kann ein Offset und eine Länge in Bytes angegeben werden. Bei Markup-Sprachen ist die Nutzung von Element-IDs möglich und bei entsprechenden Formaten können auch Raum- und Zeitwerte herangezogen werden.

Eine hohe Flexibilität und Zielgenauigkeit bei der Vergabe von Metadaten wird erreicht, weil die Strukturelemente auf Metadaten verweisen können, da diese ebenfalls mit IDs zu versehen sind. Darüber hinaus können Metadaten in Kategorien eingeteilt werden. Vorgesehen sind deskriptive (inhaltsbeschreibende) Metadaten und administrative Metadaten, die wiederum unterteilt sind in Metadaten zur Beschreibung der technischen Aspekte, der Rechte, der ggf. analogen Quelle und der Provenienz (Lifecycle) eines Objektes und seiner Bestandteile. Ist z.B. derselbe Inhalt technisch unterschiedlich repräsentiert, können spezifische technische Metadaten zugeordnet werden, ohne den Inhalt mehrmals mit deskriptiven Metadaten beschreiben zu müssen.

[34] Der Standard XML Schema würde auch noch HexBinary zulassen. Ein Oktet binärer Daten wird dabei auf zwei Hexadezimalzahlen abgebildet.
[35] Gedacht ist hier z.B. an die Beschreibung beliebig tiefer Hierarchien oder das Erzwingen einer bestimmten Sortierreihenfolge.

Metadaten können extern, also außerhalb eines METS-Dokuments gehalten werden, indem auf sie verwiesen wird. Folgende Typen von Mechanismen sind in METS benennbar: ARK, URN, URL, PURL, HANDLE, DOI und OTHER, wobei der letzte Typ in einem eigenen Metadatenfeld beschrieben werden sollte. Wie die Inhaltsdaten können auch die Metadaten inline genommen werden. Dabei können die Metadaten in XML-kodierter Form vorliegen oder in binärer bzw. textueller Form. Bei nicht vorhandener XML-Kodierung ist eine Base64 Darstellung nötig. Diese Fallunterscheidung wird in METS explizit kenntlich gemacht. Für eine Reihe von Metadatenschemata, die vom Editorial Board als bedeutend eingeschätzt werden, steht in METS ein Vokabular bereit, nämlich für die deskriptiven Metadaten DDI, DC, EAD, FCDC, LOM, MARC, MODS, TEI Header sowie VRA und für die administrativen Metadaten LCAV, NISO Technical Metadata for still images und PREMIS.

Des Weiteren können für die Inhaltsdateien die Transformationen *Entpacken* und *Entschlüsseln* und deren Reihenfolge angegeben werden, um zu spezifizieren, wie Inhalte in eine darstellbare Form zu bringen sind. Außerdem kann jedem Inhaltsobjekt über Verweise beliebiges Verhalten, also nicht nur Methoden zum Entpacken und Entschlüsseln, zugeordnet werden. Dazu ist auf alle Fälle eine Schnittstellendefinition erforderlich. Optional kann auch auf Objekte verwiesen werden, die das in der oder den Schnittstellen abstrakt definierte Verhalten direkt (z.B. von ausführbaren Kode) oder indirekt (z.B. als Verweis auf einen Web-Service) realisieren. Die Benennbarkeit der Mechanismen der Referenzierung entsprechen denen der externen Metadaten. Auf diese Art können auch Verweise auf Verfahren für die o.g. zwei Transformationen explizit angegeben werden. Mit der Zuordenbarkeit von Verhalten kann also auch diese Form der Repräsentationsinformation in einem einheitlichen Containerformat behandelt werden.

Über so genannte METS Pointer sind METS-Dokumente in einer standardisierten und einheitlichen Weise verknüpfbar. Somit können z.B. übergeordnete Paketstrukturen, wie sie z.B. in OAIS als AIU, AIC und Access Collections vorgesehen sind, einheitlich repräsentiert werden.

Profilierung von METS

Die Flexibilität und Mächtigkeit von METS kann zu unnötiger Komplexität führen, die z.B. mit entsprechendem Aufwand für das Anlegen von effizienten Speicherstrukturen oder die Entwicklung von Tools für die Erstellung, Abfrage und Bearbeitung von METS-Objekten verbunden ist. Nicht gewollte Freiheitsgrade sind eine Gefahr für die Konsistenz von METS-Objekten. Daher ist die Möglichkeit vorgesehen, Profile zu erstellen, welche eine Anpassung an spezifische Belange erlauben. Einschränkungen können u.a. betreffen: die zulässigen Metadatenschemata, die erlaubten Dateiformate für Inhalte, Metadaten und Ausführungsmechanismen, den strukturellen Aufbau von digitalen Objekten, die Zulässigkeit externer/interner Objekte, die Verwendung kontrollierter Vokabularien, die Einbeziehung von Schemaerweiterungen, wobei diese Erweiterungen zu veröffentlichen sind. Für ausgewählte Schemata stehen durch das METS Editorial Board bestätigte Erweiterungen zur Verfügung, nämlich für deskriptive Metadaten DC, MARCXML, VRA Core und für die administrativen Metadaten Schema for Technical Metadata for Text, Schema for Rights Declaration, MIX und die vier PREMIS Entitäten (Objekte, Rechte, Ereignisse, Handelnde) einschließlich eines Containers für diese Entitäten. Um die Profile strukturierter beschreiben zu können, ist ein XML-Schema vorgegeben. Ein Vorgehen zur strengen Formalisierung der Profilierung ist bisher aber nicht angegeben.

Eine *Schemaerweiterung* zur Rechteverwaltung mit METS

Die Stanford University Libraries[36] haben eine *Schemaerweiterung* publiziert, die die Angabe von Metadaten für die Verwaltung geistiger Rechte an Objekten (auch assets genannt) bzw. Teilen davon erlauben. Da das Rechtemodell schlank sein sollte, ist die Menge der Metadaten minimal gehalten. Die Metadaten sind in drei Hauptgruppen eingeteilt. In der *Rechteerklärung* wird das geistige Eigentum, das mit den Objekten oder Teilen verbunden ist, beschrieben. Hierzu stehen u.a. vordefinierte Kategorien wie *copyrighted* oder *public domain* zur Verfügung. Des Weiteren können Personen oder Institutionen als Besitzer der geistigen Rechte erfasst werden. Der so genannte *Kontext* beschreibt schließlich, wer welche Rechte besitzt und wer welchen Beschränkungen unterliegt. Hierzu stehen vordefinierte Nutzer bzw. Nutzergruppen bereit, wie *allgemeine Öffentlichkeit* oder *akademischer Bereich*. Erlaubte Nutzungen sind ebenfalls vordefiniert, wie das Anzeigen oder *Kopieren*, die durch *Einschränkungen* wie *Format* oder *Qualität* ergänzt werden können.

METS in Anwendungen

Mets ist bereits in zahlreichen kommerziellen und nicht-kommerziellen Produkten im Einsatz. Zwei Anwendungen werden im Folgenden kurz vorgestellt.

UOF

Das Paketformat UOF (Universal Object Format bzw. Universelles Objektformat) wurde im Rahmen des nationalen Projekts kopal entwickelt für Austausch und Archivierung (zu kopal siehe auch [Schi2008]). Anforderungen waren offene Schnittstellen für Einlieferungs- und Auslieferungspakete im Sinne von OAIS sowie eine Abstützung auf Standards als auch Flexibilität, um alle Arten von digitalen Objekten einschließlich Metadaten erfassen zu können. METS wurde als geeignete Basis erachtet, weil es spezifische technische Metadaten aufnehmen und Objekte aller Dateiformate in beliebiger Anzahl und Struktur behandeln kann. Zur Beschreibung von Metadaten innerhalb von METS dient das Schema LMER (Long-term Preservation Metadata for Electronic Resources), welches auf einer Entwicklung der Nationalbibliothek von New Zealand beruht. Es wurde der Struktur von METS angepasst, insbesondere sind Elemente entfernt, die bereits vorhanden sind. PREMIS als Alternative zu LMER schied aus, da zur Entwicklungszeit noch keine Ergebnisse veröffentlicht worden waren. Ein UOF-Paket besteht aus einer einzigen gepackten Datei, auf dessen Wurzelebene sich genau eine METS-Datei befinden muss. Konzeptionell können beliebig viele Dateien und Dateiordner in ein UOF-Paket aufgenommen werden [Stei2006a].

Für das UOF wird bezüglich METS eine Reihe von Einschränkungen festgelegt, welche in Form eines Profils bei der Library of Congress registriert sind (siehe auch [Stei2006b]). So müssen sich alle Metadaten innerhalb der METS-Datei befinden. Referenzierte Inhaltsdateien müssen mit URL adressiert sein und lokal vorliegen. Wird auf externe Archivobjekte Bezug genommen, so soll dies einen technischen Zusammenhang widerspiegeln, wie z.B. eine zu Grunde liegende Schemadefinition. Inhaltliche Zusammenhänge hingegen sollten mittels der deskriptiven Metadaten beschrieben werden. Weitere Regelungen betreffen die Vergabe externer und interner Identifikatoren für die Objekte.

[36] http://cosimo.stanford.edu/sdr/metsrights.xsd

Weitere Einschränkungen können sich durch die jeweils verwendete technische Umgebung (Archivsystem) ergeben, wie z.B. durch DIAS im Projekt kopal (siehe auch [Stei2006b]). So ist z.B. die Art und Anzahl von Metadatenschemata zwar nicht eingeschränkt, aber in DIAS wird nur eine Teilmenge bestimmter Schemata für das so genannte Datenmanagement genutzt. Einschränkungen können sich auch durch die gewählte Implementierung ergeben. Als Packformat sind bestimmte Versionen von ZIP (unkomprimiert oder mit Standardkompression) und TAR erlaubt. Die Anzahl erlaubter Dateien ist generell auf 5000 beschränkt und bei Verwendung von ZIP darf die Größe einer Datei 2 GB nicht überschreiten [IBM2006a, IBM2007b].

BABS

BABS (Bibliothekarisches Archivierungs- und Bereitstellungssystem) dient an der Bayerischen Staatsbibliothek München zur Langzeitarchivierung einer großen Anzahl digitaler Objekte unterschiedlichster Art. Dazu gehören z.B. Retrodigitalisate, der Output aus dem Kooperationsprojekt mit Google, sowie E-Books und E-Zeitschriften und künftig auch verstärkt Web-Seiten und Online-Datenbanken. Die nestor-Expertise der FUH [Schi2008] hat die technische Grundarchitektur bereits vorgestellt; hier soll der Aufbau und die Verwaltung von Informationspaketen mit Hilfe von METS thematisiert werden. Die Architektur von BABS stützt sich neben dem Speichersystem auf zwei Softwarekomponenten: 1) ZEND (Zentrale Nachweis- und Erfassungsdatenbank), die alle bibliografischen, technischen, administrativen und strukturellen Metadaten der Retrodigitalisate verwaltet, und den gesamten Lebenslauf der digitalen Objekte von Produktion über Bereitstellung zur Archivierung unterstützt und dokumentiert. 2) DigiTool, ein kommerzielles Produkt der Firma Ex Libris zur Verwaltung heterogener digitaler Objekte und der zugehörigen bibliografischen, technischen, administrativen und strukturellen Metadaten.

Beide Systeme nutzen zur Darstellung komplexer Objekte METS. Zur Speicherung struktureller Beschreibungen stützt sich ZEND zwar auf TEI-XML mit ebind-Elementen ab, doch zur Unterstützung des DFG-Viewers können die entsprechenden METS-Dateien zum Anforderungszeitpunkt generiert werden. Ziel dieses Viewers ist die vereinheitlichte Anzeige von Digitalisaten aus dezentralen Bibliotheksrepositorien. Dazu dient ein METS-Profil, das den Umfang der Metadaten und den strukturellen Aufbau der Objekte festlegt. Beim Entwurf war auf die unterschiedlichen Erschließungsmöglichkeiten seitens der Produzenten zu achten, weshalb das Profil eine bestimmte Bandbreite zulässt. Ebenso sollten Erweiterungsmöglichkeiten bezüglich der Navigation, die über die Fähigkeiten eines Pageturners hinausgehen, berücksichtigt werden. Für die Bereitstellung bibliografischer Daten dient die typisierte Einbindbarkeit von Metadaten durch METS.

Auch die Firma Ex Libris hat für DigiTool ein METS-Profil entwickelt und bei der Library of Congress registrieren lassen. Es unterstützt die strukturelle Beschreibung von intellektuellen Entitäten in logischer und physischer Sicht, die aus Objekten mit unterschiedlichen Inhaltsrepräsentationen (Formaten) zusammengesetzt sind (Ex Libris – DigiTool Multipage Entity). Das Produkt DigiTool umfasst sowohl einen METS-Editor als auch einen METS-Viewer.

METS im Kontext von Grid/eScience

Die zuverlässige und leicht handhabbare Organisation von Inhaltsdaten und Metadaten ist sicherlich auch ein Anliegen im Bereich Grid/eScience. Nach Moore und Merzky [Moor2003b] gehören Datencontainer zu den Kernfähigkeiten eines persistenten Archivs

innerhalb eines virtuellen Datengrids. Hierfür werden technische Gründe angeführt, wie die Vereinfachung des Datentransfers und des Speichermanagements. Neben der Zusammenfassung mehrerer digitaler Entitäten in einer einzigen Datei werden an dieser Stelle keine weiteren Forderungen spezifiziert. Die Studie des GFZ [Klum2007] berichtet, dass in den Projekten teilweise sehr viele digitale Einzelobjekte anfielen sowie dass die Bedeutung von Metadaten erkannt sei. Auf die Heterogenität der Daten und Dokumente, die sich in der Verwendung unterschiedlicher Dateiformate niederschlägt, wird hingewiesen. Sie ist ein Hindernis bei der Gewinnung von Information. Des Weiteren spielen Binärdaten eine deutlich wichtigere Rolle als bei den herkömmlichen Archivierungsszenarien. Ebenso wird der Bedarf an einer höheren Granularität für die Rechteverwaltung sichtbar, was eine gezieltere Zuordnung von Metadaten erfordert. Die Kodierung von Metadaten in Verzeichnis- oder Dateinamen zählt Ian Foster zum Messy-Data-Problem[37]. Der Rückgriff auf solche Notlösungen ist ein deutlicher Hinweis, dass ausreichende und einfach handhabbare Mittel nicht zur Verfügung stehen oder bekannt sind.

Bestimmte Eigenschaften von METS können hierfür Hilfe leisten. So bietet die Flexibilität eine gute Anpassbarkeit an spezifische Anforderungen von Communities. Vorteilhaft ist die Definierbarkeit jeweils unabhängiger physischer und logischer Strukturen für die Inhaltsobjekte in einer nachvollziehbaren Art und Weise. Ein großes Plus ist auch die gezielte und modulare Zuordenbarkeit von Metadatenobjekten. Durch diese Modularität und Nutzung von Referenzen können auf Ebene von METS-Objekten Redundanzen bei der Vergabe von Metadaten vermieden oder reduziert werden, was sich insbesondere bei einer sehr großen Anzahl von Inhaltsobjekten als nützlich erweisen kann. Positiv ist auch zu bewerten, dass der Zugriff auf Teilbereiche von Inhaltsobjekten beschreibbar ist. Dadurch lassen sich, wenn auch nur statisch, Ausschnitte definieren, mit denen insbesondere für große Einzelobjekte eine Vorschau realisiert werden kann. Aufgrund der Modularität bei der Metadatenvergabe, wäre hierfür auch eine spezifische Rechtevergabe möglich.

Durch die Spezifizierbarkeit internetfähiger Referenzen auf wesentliche Komponenten eines METS-Objektes wird der kontrollierte Aufbau virtueller Pakete begünstigt. Zu diesen Komponenten zählen die Inhaltsobjekte, Metadatenobjekte, Schnittstellenobjekte, Objekte mit Ausführungsmechanismen und METS-Objekte. Die letzte Möglichkeit vereinheitlicht die inkrementelle Archivierung komplexer Sammlungen über einen ggf. langen Zeitraum. So können durch das Anlegen eines übergeordneten Objektes Änderungen an einem Punkt konzentriert und Manipulationen an bereits eingelieferten Objekten vermieden werden. Für solch ein übergeordnetes Objekt bietet sich die Entwicklung einer spezifischen Schemaerweiterung für die administrativen Metadaten an. Die explizite Referenzierbarkeit von Verhalten ist ebenfalls positiv einzuschätzen, insbesondere wenn für die Interpretation der Inhalte spezifischere Verfahren nötig sind, die nicht zum „Allgemeingut“ gezählt werden können und nicht unbedingt Kandidaten für ein allgemeines Formatverzeichnis (Registry) sind. Eine Vereinheitlichung würde die Einführung von Workarounds verhindern, wie das „Hineinkodieren“ von Versionen oder Varianten in die Begriffe eines kontrollierten Vokabulars. Eine Standardisierung würde langfristig ein technisches Monitoring im Sinne von OAIS deutlich erleichtern. Als weiterer Vorteil ist die Abstützung auf (wenige) Standards des W3C zu sehen. Transformations- und Abfragesprachen wie XSLT oder XML Query unterstützen die Verwaltung von XML Schemata wie die Transformation zwischen den einzelnen Pakettypen des OAIS.

[37] Im Rahmen des Vortrags „Swift: Fast, Reliable, Loosely Coupled Parallel Computation“ von Ian Foster am Leibniz Rechenzentrum in Garching am 30.04.2007

Neben den Vorteilen sind bei METS auch Nachteile zu erkennen. Unter Gesichtspunkten der Langzeiterhaltung wäre eine einheitliche Möglichkeit für eine präzisere Typisierung der referenzierten bzw. beinhalteten Metadatenschemata nützlich, insbesondere unter der Annahme, dass wissenschaftliche Daten empfindlicher für Fehlinterpretationen sind und dass über kurz oder lang zahlreiche Schemaausprägungen für Metadaten entstehen. Verteilte Projektbearbeitung erhöht die Wahrscheinlichkeit der Verwendung unterschiedlicher Schemata. Ebenfalls als problematisch könnten sich die Profilierungsmöglichkeiten von METS erweisen, wenn sie in unkoordinierter Weise genutzt werden, so dass eine unnötig hohe Anzahl an Varianten entsteht, z.B. durch sich überschneidende Schemaerweiterungen. Hier ist sicherlich Potenzial für eine künftige Vereinheitlichung bzw. Generalisierung vorhanden, auch für wissenschaftliche Prozesse und Daten. Des Weiteren wäre angesichts der Heterogenität und hohen Anzahl von digitalen Objekten eine stärkere Formalisierung der Profilierung wünschenswert. Eine Nutzung der Modularisierungsmöglichkeiten des W3C-Standards XML Schema bietet sich hierfür an. Für ein universelles Containerformat wäre auch die standardisierte Einbeziehung von Binärdaten ohne Umkodierung in eine speicherplatzfressende Darstellung wünschenswert, sofern ein höherer Grad an Kapselung, wie vielfach für eine Langzeitarchivierung praktiziert oder als zuverlässiger erachtet, erreicht werden soll. Das UOF z.B. kann im Gegensatz dazu Daten in beliebiger Form aufnehmen, weil METS hier nur eine Komponente des Pakets bildet. Falls wissenschaftliche Daten auch nach der Arbeitsphase eine gewisse Lebendigkeit haben, wäre auch eine bessere Unterstützung der Versionierung wünschenswert, z.B. durch die Darstellbarkeit expliziter Beziehungen auf Ebene der möglichen Komponenten von METS[38]. Dieser Wunsch gilt auch für die Adressierung von Profilversionen[39].

Darüber hinaus ist festzustellen, dass METS in Konkurrenz zu anderen Metadatenschemata steht, die ebenfalls die Beschreibung struktureller Aspekte abdecken. Hierzu zählen z.B. PREMIS und LMER. Die Auflösbarkeit eventuell genutzter Referenzen und die Definition nicht-struktureller Metadaten liegen per Definition außerhalb der Standardisierung von METS.

7.2.2 XFDU

XFDU (XML Formatted Data Unit) ist ein Containerformat, das vom CCSDS spezifiziert wird. Im Anhang zum White Book aus dem Jahr 2004 [CCSDS2004] wird begründet, warum METS als Basis für die Entwicklung eines derartigen Formats vorgeschlagen wurde. Die hohe Flexibilität der Verknüpfung von Inhaltsdaten und Metadaten wird ebenso wie der Bezug zum OAIS-Referenzmodell als Vorteil gesehen. Als nachteilig werden der mehr konzeptionelle Charakter von METS und die Art der Abbildung der OAIS Referenzinformation eingeordnet. Tatsächlich wird der Formalisierung eine stärkere Aufmerksamkeit gewidmet. Außerdem besitzt XFDU einen stärkeren Bezug zu OAIS, was die Kategorien (z.B. Repräsentationsinformation, PDI) und Klassen (z.B. Fixity, Provenance) des dortigen Datenmodells anbelangt.

Seit September 2008 liegt die Empfehlung als Blue Book vor [CCSDS2008]. Somit zeichnet sich bezüglich der Standardisierung eine Konsolidierung ab.

[38] Ansätze finden sich im folgenden Profil: ECHO Dep Generic METS Profile for Preservation and Digital Repository Interoperability.

[39] vgl. hierzu auch Ausführungen zu Profilen von Brian Tingle am METS Opening Day an der SUB Göttingen am 07.05.2007, siehe
http://www.loc.gov/standards/mets/presentations/Brian_Tingle_profiles_germany.ppt

XFDU im Kontext von Grid/eScience

Im Wesentlichen sind die Ausführungen zu METS übertragbar. Entsprechend dem Hintergrund von CCSDS wird die (noch) unzureichende Behandlung von Binärdaten in XML betont. Ein besonderer Schwerpunkt wird in der Beschreibung und Anwendung von ausführbaren *Verhalten* gesehen. Voraussetzung für die Anwendbarkeit ist, dass ausführbarer Kode (Software) tatsächlich zur Verfügung steht, entweder im Paket eingeschlossen (inline) oder über eine Referenz. Somit könnten Inhaltsdaten nach Bedarf generiert werden. Die Umsetzung der konkreten Ausführbarkeit ist für künftige Versionen von XFDU vorgesehen. Bis dahin dient ein *abstraktes* Element als „Platzhalter". In [CCSDS2004] wird zur Umsetzung die Möglichkeit empfohlen, Software-Einheiten mittels des Distributionswerkzeuges GPT (Grid Packaging Toolkit) darzustellen. Die Planung für künftige Versionen umfasst auch eine präzisere Beschreibung der Objektbeziehungen auf Basis von RDF und OWL.

7.2.3 Weitere Möglichkeiten der Paketbeschreibung

Der Bedarf Daten zusammenfassen und als Einheit zu behandeln ist nicht neu. Die Vereinfachung der Datensicherung und der Archivierung sowie des Transports von Daten waren wesentliche Motivation. Die Verteilung von Software trat als weitere Aufgabe hinzu. Zunächst wird ein „Klassiker" skizziert und diskutiert.

ZIP

Ein weit verbreitetes Containerformat mit der Option der Komprimierung und Verschlüsselung ist ZIP. Die Kontrolle über die Spezifikation wird von der Firma PKWARE ausgeübt. Grundsätzlich können beliebige Dateien sowie Dateiverzeichnisse zahlreicher Betriebssysteme in eine Datei gepackt werden. Außerdem können mehrere Datenträger überspannt (Multivolume) oder Segmente bestimmter Größe vorgegeben werden[40]. Das Format erlaubt den wahlfreien Zugriff auf einzelne Dateien, da ein zentrales Verzeichnis angelegt wird. Neuere Versionen erweitern die bisherigen und harten, konzeptionell bedingten Größenbeschränkungen erheblich. Ebenso erhöhte sich die Interoperabiltät durch die Einführung von UTF-8 für Dateinamen.

In eine ähnliche Kategorie fallen die aus der Unix-Welt stammenden Formate tar (Tape Archive) und cpio (copy in, copy out) sowie das weniger populäre PAX (Portable Archive Exchange) als Vereinigung von tar und cpio. PAX ist Bestandteil von POSIX und somit auch formal standardisiert. Verschlüsselung und Kompression sind bei diesen drei Formaten nicht Gegenstand der Spezifikation.

ZIP im Kontext von Grid/eScience

Nachdem die Größenbeschränkungen gefallen sind und die Interoperabilität erhöht wurde, sind die Voraussetzungen für eine Eignung als Paketformat im Grid/eScience-Feld gegeben. Die Möglichkeit des direkten Zugriffs auf Einzelobjekte kann sich bei großen Archivdateien als vorteilhaft erweisen. Die Nutzung der Verschlüsselungs- und Kompressionstechniken ist im Einzelfall genau abzuwägen. Anwendungsspezifische bzw. objektspezifische oder dateiübergreifende Kompressions- bzw. Reduktionsverfahren bringen gegebe-

[40] Erlaubt sind (2 hoch 32)-1 Segmente mit maximal (2 hoch 32) Byte also ca. 4,29 GB (SI).

nenfalls einen deutlich höheren Effekt. In diesem Fall wäre aber die Angabe weiterer Metadaten erforderlich, was mit ZIP nicht in einer standardisierten Weise möglich ist.

In Hinblick auf die Langzeitarchivierung und die Anwendung in heterogenen Organisationsformen sind die fehlende Zuordenbarkeit und die ungenügende Beschreibbarkeit logischer Strukturen als Nachteil dieses Formats zu betrachten. Hier wird klar erkennbar, dass beim Entwurf von ZIP Aspekte des Backups und Austauschs von Dateien und Dateiverzeichnissen im Fokus standen und nicht die Modellierung von Information und deren Verwaltung. Web-spezifische Anforderungen, wie Streamingfähigkeit oder die Auszeichnung mit spezifischen Metadaten, wie MIME-Typen, sind bisher nicht berücksichtigt. Die Möglichkeiten zur Selbstdokumentation und Abgeschlossenheit, wie sie häufig für eine Langzeitarchivierung gefordert werden, sind in einer standardisierten Weise nur eingeschränkt gegeben.

Eine Indiz für eine Bedeutung auch in Grid-Umgebungen ist die geplante Erstellung bzw. die Auflösung einer Archivdatei unter Nutzung des „Algorithmus“ Info-ZIP in der Komponente Activities der Middleware OGSA-DAI [OGSA-DAI2007].

MPEG-21 DID

MPEG-21 besteht aus einer Serie von Standards, die ein abgestimmtes und modulares Rahmenwerk[41] für die Bereitstellung und Nutzung multimedialer Inhalte in einer vernetzten Umgebung mit unterschiedlichen Geräten und für unterschiedliche Communities liefert. Teil 2 der Serie definiert ein konzeptuelles Modell für digitale Objekte, als Digital Item bezeichnet, sowie deren Repräsentation mit Hilfe von XML durch die Digital Item Declaration Language (DIDL) [ISO/IEC 21000-2:2005]. Die Flexibilität des Modells entsteht durch den beliebig tief schachtelbaren (rekursiven) Aufbau für ausgewählte Elemente. Dabei kann den umschließenden Elementen jeweils ein Deskriptor zur Beschreibung der umschlossenen Elemente zugeordnet werden. Die eigentlichen Inhalte sind zusätzlich nochmals in Fragmente unterteilbar. Die Referenzierungsmöglichkeit von Inhalten erlaubt deren Speicherung auch außerhalb des DIDL-Dokuments. Durch die Angabe von Regeln für umschließende Elemente kann bestimmt werden, ob die betreffenden Elemente Teil des Objektes sein sollen oder nicht. Dieser Mechanismus gestattet die dynamische Konfiguration von digitalen Objekten.

MPEG-21 DID im Kontext von Grid/eScience

Die Fähigkeiten von MPEG-21 DID fanden auch in der LZA-Community Beachtung. Eine Reihe von Vorteilen ist in [Beka2005] aufgezählt. Hierzu gehören eine wahrscheinlich starke Unterstützung durch die Industrie, was sich positiv auf die Verfügbarkeit künftiger Migrationswerkzeuge auswirken sollte, sowie das abstrakte Datenmodell, das eine Abbildung auf künftige Technologien der Serialisierung erleichtert. Eine positive Einschätzung erfährt auch die leichte Umsetzbarkeit des OAIS-Informationsmodells bzw. der OAIS-Informationspakete.

Falls Forderungen nach einer feingranularen Verwaltung von heterogenen Informationen bestehen, stellt MPEG-21 DID eine untersuchenswerte Alternative dar. Außerdem erleichtert der hohe Abstraktionsgrad die Nutzung hocheffizienter (künftiger) Serialisierungstechniken, falls sehr große Objekte und spezielle Zugriffsmechanismen erforderlich sind.

[41] vgl. z.B. Nutzung der MPEG-21-Rechtesprache in DOI

Grundsätzlich bleibt auch hier das Problem, Profile zu entwickeln, die einen geregelten Austausch bzw. eine Integration über Organisationsgrenzen hinweg sicherstellen.

7.3 Pakettransformation (Ingest, Access) und Paketrepräsentation

Pakettransformationen sind Kernprozesse im OAIS-RM. Für die eigentliche Archivierung werden die Einlieferungsinformationspakete (SIP) beim Ingest in Archivinformationspakete (AIP) transformiert und beim Access werden daraus ggf. nach vielen technologischen Innovationszyklen Auslieferungsinformationspakete (DIP) generiert. Logische Vollständigkeit und informationserhaltende Transformationen einschließlich der Erstellung von SIPs sind insbesondere bei komplexen Paketen eine große Herausforderung. Komplexität kann sich ergeben durch uneinheitliche Repräsentationen der zugrunde liegenden Daten, durch eine schwierige Interpretierbarkeit der Daten, durch eine vielstufige und langfristige Entstehungsgeschichte als auch durch die physische Größe oder hohe Anzahl von Einzelobjekten.

Daten in verteilten Umgebungen stellen nach wie vor ein großes Problem für die Gewinnung von Information dar. Etliche Standards haben bereits dazu beigetragen, die Transparenz zu verbessern. Insbesondere können technische Heterogenitäten gut vor Anwendungen verborgen werden. Mit der Einführung verteilter Dateisysteme und verteilter Datenbankenmanagementsysteme, die weitgehend von einem darunter liegenden Betriebssystem abstrahieren, konnte ein hohes Maß an Verteilungstransparenz hergestellt werden, insbesondere, wenn man nicht gezwungen war, eine Produktlinie zu verlassen. Fehlende bzw. nicht rechtzeitig verfügbare Standards oder Herstellerinteressen einerseits und die Notwendigkeit zu kooperieren andererseits führten zur Entwicklung von Middleware-Produkten. Diese können teilweise auf Standards zurückgreifen, die versuchen, die entstandenen Heterogenitäten nachträglich zu überwinden oder selbst Gegenstand einer Standardisierung sind wie z.B. CORBA (Common Object Request Broker Architecture). Die Konvertierung von Datentypen verschiedener Datenbanksysteme und Programmiersprachen ist eine wichtige Voraussetzung, Heterogenitäten zu verbergen. Ein weiteres Ziel der Vereinheitlichung sind komplexe Datentypen, wie sie sich durch unterschiedliche Paradigmen ergeben (z.B. objektorientierte vs. relationale vs. grammatikbasierte Ansätze). Entsprechend den unterschiedlichen Ansätzen und – zusätzlich – ihren unterschiedlichen Ausprägungen erweisen sich auch Datendefinitions- und Datenmanipulationssprachen als sehr heterogen.

Zur Sicherung der Integrität der Daten sind neben korrekten Transformationen weitere Maßnahmen erforderlich. Transaktionen zuverlässig durchzuführen ist hierfür ein wichtiger Dienst. Zu den kritischen Eigenschaften zählt, dass die Durchführung der Transaktionen komplett ist, d.h. alle Einzelschritte, die für eine logische Operation nötig sind werden tatsächlich als Gesamtheit durchgeführt oder im Fehlerfall unterlassen (Atomarität). Des Weiteren hat eine Transaktion das System in einen konsistenten Zustand zu verlassen. So dürfen Eigenschaften, wie die Eindeutigkeit von Schlüsseln, nicht verletzt werden. Die Regeln hierfür müssen dem Transaktionsprozess natürlich bekannt sein (Konsistenz). Transaktionen dürfen sich nicht gegenseitig in einer unbeabsichtigten Weise beeinflussen (Isoliertheit). Schließlich wird erwartet, dass eine erfolgreich abgeschlossene Transaktion im System auch festgeschrieben wird (Dauerhaftigkeit). Die Umsetzung dieser Forderungen (als ACID-Eigenschaft bekannt) ist in verteilten und heterogenen Umgebungen eine große Herausforderung. Auch zum Thema Transaktionen gibt es Aktivitäten zur Standardisierung z.B. in POSIX.

Zur Sicherung der Integrität (und Vertraulichkeit) gehört auch der Schutz vor unberechtigten Operationen. Die Deklaration und Überwachung von Rechten ist durch sehr unterschiedliche Konzepte und Realisierungen gekennzeichnet. Dateisysteme erlauben nur eine sehr grobe Rechteverwaltung im Vergleich zu Datenbankmanagementsystemen. Dort können Rechte sehr feingranular verwaltet werden. Da Rechte auch in Abhängigkeit von konkreten Werten ausgewählter Attribute oder Berechnungen ermittelt werden können, lässt sich leichter ein flexibles Rechtemanagement realisieren. Nachteilig ist dabei die Bindung an ein bestimmtes Modell der Datenhaltung.

Zur Integrität der Daten tragen natürlich auch „physische“ Maßnahmen bei, wie das Anlegen von Replikaten, die darüber hinaus für eine bessere Verfügbarkeit (Fail over, Zugriffsgeschwindigkeit) sorgen können. Natürlich müssen Replikate in ein Transaktionskonzept einbezogen werden (siehe auch die Ausführungen in [Schi2008]).

Verfügbarkeit und Qualität von Daten und ihren Informationsgehalt zu sichern und gegenüber anderen Architekturen auch zu verbessern ist ein zentrales Anliegen der Entwicklung und Standardisierung im Grid-Bereich. Hierzu wird auch Bezug auf Standards konventioneller, auch verteilter Architekturen genommen.

7.3.1 OGSA Data Architecture

Wie in den Expertisen des GFZ und der FUH angedeutet, sind übergreifende Datengrids trotz ihrer Bedeutung noch nicht sehr verbreitet [Klum2008][Schi2008]. Eine Ursache liegt sicherlich in der Komplexität der Problematik, die nicht nur mit schwierigen Fragestellungen der integritätserhaltenden Transformation und Integration von Inhalten zu tun hat, sondern auch mit Themen der Effizienz und Sicherheit sowie mit bereits etablierten Systemen und mit nicht unbedingt optimal abgestimmten Standards. Mit dieser Problematik beschäftigt sich innerhalb des OGF die OGSA Data Working Group. Sie hat ein Rahmenwerk zur Datenarchitektur entwickelt, das auf einem abstrakten Niveau Schnittstellen (interfaces), Verhalten (behavior) und Bindungen (bindings) für die Behandlung von Daten innerhalb der Architektur OGSA beschreibt [OGSA Data WG 2007]. Die Architektur umfasst die Aspekte Speicherung, Transport, Zugriff, Replikation, Caching und Föderation bezüglich Dateien und Datenbanken. Aussagen zu Repräsentationsinformation, PDI und Beschreibungsinformation (vgl. OAIS) finden sich unter dem Begriff Datenbeschreibung (data description) mit den Punkten Formatbeschreibung, Ressourcenbeschreibung (resource description), Beschreibungen Dritter (third-party description) und Provenance. Beschreibungen Dritter beziehen sich auf inhaltliche Beschreibungen durch Konsumenten, die die Daten nicht besitzen oder verwalten, oder auf Beschreibungen, die sich im Laufe der Zeit ändern. Im Sinne des OAIS-Informationsmodells könnte diese Art der Beschreibung unter Collection Information fallen (vgl. OAIS-Informationsmodell). Die Benamung (naming) von Daten-Entitäten wird in einem Rahmen betrachtet, der alle Typen von Entitäten in einem Grid betrachtet wie z.B. Verzeichnisse, Namensräume, Transaktionen oder Vokabularien, wobei Anforderungen an eine Benamung formuliert sind. Des Weiteren umfasst das Rahmenwerk auch Aussagen zur Standardisierung. Insbesondere im Bereich der Föderation wird noch großer Handlungsbedarf gesehen, so dass die Grundlagen für eine Standardisierung noch fehlen. In diesem Zusammenhang wird auf die Arbeiten zu OGSA DAI (Web Services Data Access and Integration) verwiesen. Ziel der Architekturbeschreibung ist es auch, aufzuzeigen, wie Standards (Spezifikationen) zusammenpassen, um eine zusammenhängende Grid-Architektur zu bilden. Im Anhang des Rahmenwerks findet sich eine Zu-

sammenstellung relevanter Standards, die nach den Komponenten der Architektur gegliedert ist.

7.3.2 OGSA-DAI

Der Zugriff auf verteilte Datenbestände und deren Integration spielt eine fundamentale Bedeutung für den Informationsaustausch. Datenbanksysteme spielen in vielen Anwendungen eine wichtige Rolle, nicht nur für die Inhaltsdaten, sondern auch für Metadaten aller Art als auch für die Verwaltung von Systemen wie verteilte Dateisysteme oder das Datenmanagement nach OAIS. Sie sind daher in eine Architektur mit einzubinden. Die Arbeitsgruppe OGSA DAI hat sich dieser Aufgabe angenommen und eine Reihe generischer Schnittstellen für den Datenzugriff spezifiziert, die als Web Service angeboten werden sollen [OGSA-DAIS 2006]. Unter Zugriff wird nicht nur das Lesen, sondern auch das Ändern und Einfügen von Daten verstanden. In diesem Zusammenhang ist natürlich in den Diensten zu regeln, welche Arten von Operationen ausgeführt werden dürfen. Die Spezifikation trennt in ein abstraktes Modell der Dienste und in eine Repräsentation für die Ausführung in WSDL. Unter Einschränkung von Funktionalität ist auch der Fall eines Zugriffs ohne WSRF vorgesehen. Um unnötigen Datenverkehr zu vermeiden dient das Konzept des indirekten Zugriffs. Damit können Ergebnisse einer Anfrage am Ort der Quelle verbleiben und mit einem weiteren Dienst angesprochen werden. Zum Umfang des Standards gehört jedoch nicht, die unterschiedlichen Datenmodelle und DMLs zu vereinheitlichen und die Heterogenität der darunter liegenden Datenhaltungssysteme zu verbergen. Dies ist einem Standard vorbehalten, der sich mit einer höheren Ebene (Information) befasst. Die spezifizierten Kern-Dienste können entsprechend der genutzten Datenhaltungssysteme spezialisiert werden. Für relationale und XML-basierte Datenressourcen wurden im Rahmen von OGSA zwei weitere Spezifikationen veröffentlicht [OGSA-DAIR 2006], [OGSA-DAIX 2006].

Eng mit der Entwicklung mit diesen Spezifikationen stehen das Projekt und die Middlewaresoftware OGSA-DAI[42]. Die Software, die gegenwärtig auf den Spezifikationen aus dem Jahr 2003 beruht, erlaubt eine Einbindung unterschiedlichster Datenressourcen. Hierzu bietet das Produkt zahlreiche vorgefertigte Transformationen zwischen den verschieden Datenmodellen an. Die Definierbarkeit von Workflows ermöglicht die Anordnung von Aktivitäten in einem einzigen Auftrag. Damit können datenzentrierte Workflows realisiert werden wie z.B. Datenzugriffe mit Übergabe der Ergebnisse an eine Transformation sowie das anschließende Verteilen der Ergebnisse. Diese Zusammenfassbarkeit trägt auch dazu bei, Overhead für die Kommunikation zu vermeiden. Das Zusatzwerkzeug OGSA-DQP (distributed query processing) unterstützt die Virtualisierung von Datenbanken, indem ein komplexer Auftrag zusammengefasst an einen OGSA-DAI-Server gerichtet werden kann und dann durch OGSA-DQP an die entsprechenden physischen Datenquellen weitergeleitet wird. WSDL und WSRF für die Kommunikation und JDBC (Java Database Connectivity) für den Zugriff auf relationale Datenbanken sind Grundlagen.

Unter Langzeitgesichtspunkten wirft sich die Frage auf, welchen Grad an Heterogenität und Virtualisierung man sich insbesondere für AIPs leisten kann. Datenmodelle, Standards und Produkte sind nach wie vor einem Wandel unterworfen, den eine Middleware zu berücksichtigen hätte. Ein Blick in die Dokumentation von OGSA-DAI (Abschnitt Dependencies) oder von ähnlichen Produkten gibt einen Einblick in die Problematik der Siche-

[42] http://www.ogsadai.uk

rung der Kompatibilität beteiligter Komponenten bereits innerhalb eines relativ schmalen Zeitfensters. Ein großer Vorteil solcher Produkte ist bezüglich der Langzeitarchivierung eher in der Konsolidierung und Normalisierung von Daten wie sie im Rahmen der Erstellung von SIP oder spätestens AIP geschehen könnte. Der Grad an Virtualisierung sollte auf ein vertrauenswürdiges Maß beschränkt werden. Bei der Normalisierung ist zu beachten, dass ggf. Informationen verloren gehen können, entweder aus theoretischen Überlegungen oder aufgrund unvollständiger Produkte. Im Falle von OGSA-DAI könnte man, soweit theoretisch möglich, die fehlenden Funktionen selbst ergänzen.

7.3.3 EXI

XML hat sich ohne Zweifel zu einer lingua franca im WWW entwickelt. Dies gilt für Metadaten unterschiedlichster Zwecke, für Nachrichtenbeschreibungen als auch für Inhalte insbesondere semistrukurierter Art. Der Overhead von XML bedingt jedoch Einschränkungen wie erhöhten Bedarf an Speicherplatz und Bandbreiten, die insbesondere bei großen Datenmengen oder vielen Einzelobjekten ins Gewicht fallen.

Die W3C-Arbeitsgruppe Efficient XML Interchange Format hat deshalb Anforderungen an eine kompaktere Darstellung von XML unter Betrachtung konkreter Anwendungsfälle formuliert und schließlich einen Draft für ein entsprechendes Format namens EXI erarbeitet [W3C EXI 2008]. Neben zahlreichen geforderten Eigenschaften, wie z.B. eine möglichst nahtlose Einbettung in XML-Technologien, wird durch verschiedene Mechanismen eine deutlich höhere Kompaktheit erreicht. So stehen vorab Optionen zur Verfügung, bestimmte Bausteine von XML (XML Items) zu ignorieren, weil sie für die jeweilige Anwendung nicht mehr gebraucht werden oder weil sie per Konvention ausgeschlossen sind (z.B. in SOAP). Beispiele für solche Bausteine sind Kommentare, Präfixe für Namensräume oder Verarbeitungsanweisungen (Processing Instructions). Außerdem kann für den notwendigen Header im EXI-Format zur Identifizierung und globalen Beschreibung des Formats statt der XML-Darstellung alternativ eine sehr kompakte Form gewählt werden. Kernkonzept des Formats ist die Darstellung von XML Items, wie ein Attribut eines Elements, als eine Sequenz so genannter Ereignisse (Events), denen gegebenenfalls Werte zugeordnet sind. Das Konzept der Ereignisse ist ähnlich wie in SAX (Simple API for XML). Aufgabe ist es nun, entsprechend der Häufigkeitswahrscheinlichkeit der Ereignisse unterschiedlich lange Codes zu verwenden (Entropiekodierung) und sich wiederholende Elemente mittels möglichst kompakter Referenzen zu ersetzen. Die zugrunde liegenden Grammatiken nutzen hierfür das Wissen über die Struktur eines XML-Dokuments. Die dynamische Erweiterung der Ausgangsgrammatik, die zunächst eine generische Dokumentstruktur annimmt, führt zu einer Erhöhung der Effizienz. Eine Effizienzsteigerung kann auch erreicht werden, wenn Informationen durch Schemadefinitionen verfügbar sind, die in der Form von XML Schema und Datenstrukturen, XML DTD und RELAX NG schemas [ISO/IEC 19757-2:2003] oder einer regulären Sprache vorliegen. Hierbei ist auch vorgesehen, Dokumente zu verarbeiten, die die Schemadefinition verletzen bzw. undefinierte Erweiterungen enthalten. In diesem Fall wird dann auf die Ausgangsgrammatik zurückgeschaltet. Zur Vorbereitung und Verbesserung der optionalen Standardkompression (Standard Deflate Compressed Data Format nach IETF RFC 1951) werden, unabhängig von der Nutzung einer Schemadefinition, getrennte Kanäle angelegt. Im Wesentlichen sind es pro Block aus der EXI-Sequenz ein Kanal mit Strukturinformationen und mehrere Kanäle mit Inhaltsinformationen. Die Größe des Blocks kann variiert werden. Enthalten die Kanäle zu wenig Werte, werden sie zur Vermeidung von unangemessenem Overhead zusammengefasst. Der Draft sieht außerdem vor, dass selbstdefinierte Datentypen – neben

den vordefinierten Datentypen des EXI-Formats – und entsprechende Codecs (Encoder/Decoder), z.B. für eine spezifische Komprimierung, nach Bekanntgabe im Header als so genannte pluggable CODECs, verwendet werden können.

EXI im Kontext von Grid/eScience

Zur grammatikbasierten Kompression existiert eine Reihe von Lösungen, so dass aus Gesichtspunkten der Langzeitarchivierung eine Standardisierung zu begrüßen ist. In datenintensiven Anwendungen werden der erhöhte Speicherplatz- und der Bandbreitenbedarf als Argument gegen eine Nutzung von XML in datenintensiven Anwendungen angeführt (vgl. z.B. Severins). Hier bietet die vorgeschlagene Lösung von EXI eine betrachtenswerte Alternative zu anderen Formaten, insbesondere, wenn die Daten einer stark unterschiedlichen Strukturierung unterliegen und entsprechend ausgezeichnet werden sollen. Vorteilhaft dieser Standardlösung ist auch, dass man auf Grund der kombinierbaren Grammatiken mit einem einzigen Parser auskommt, der zusätzlich selbst kompakt gehalten werden kann und somit auch in einfacheren Umgebungen (z.B. Messgeräten) lauffähig ist. Die direkte Unterstützung von Gleitkommazahlen nach IEEE 754 ist jedoch noch in der Diskussion, da diese Zahlenrepräsentation nicht in allen Umgebungen wie z.B. mobile Geräte üblich ist. Eine Unterstützung dieses Typs durch das EXI-Format würde die jeweils erforderliche Umkodierung ersparen. Weitere Überlegungen beziehen sich auf die Indexierung von Elementen der EXI-Sequenz. Diese würde einen wahlfreien Zugriff erlauben sowie einen sequenziellen Zugriff beschleunigen sofern bestimmte Abschnitte übersprungen werden können. Sehr große Objekte könnten durch die Indexierung bei Bedarf direkt nach logischen (anwendungsbezogenen) Gesichtspunkten für eine Prozessierung zerlegt werden. Die Möglichkeit, selbstdefinierte Datentypen mit spezifischen Codecs im EXI-Format einzuführen, erlaubt auf dieser Ebene der Paketbildung eine (nachträgliche) Anwendung von Kompressions- bzw. Reduktionsverfahren, die für wissenschaftliche oder technische Zwecke besonders vorteilhaft sind bzw. speziell dafür entwickelt wurden. Nachteilig für eine verteilte Umgebung ist jedoch, dass diese Erweiterungsmöglichkeit zusätzliche Absprachen zwischen Sender und Empfänger erfordert, die im EXI-Format nicht notierbar sind, da dort nur der nutzerdefinierte Typ und ein Name für den Codec vorgesehen sind. Zur effizienten Verarbeitung von XML-Dokumenten siehe auch die Ausführungen unter ASN.1.

7.4 Identifikation (Reference Information)

Im OAIS-RM identifiziert (und beschreibt ggf.) die Referenzinformation (Reference Information) einen oder mehrere Mechanismen, wie für eine Inhaltsinformation (Content Information) zugeordnete Identifikatoren geliefert werden. Die Referenzinformation liefert auch jene Identifikatoren, die außen stehenden Systemen den eindeutigen Verweis auf eine Inhaltsinformation erlauben. Entsprechend dem OAIS-RM umfasst Inhaltsinformation neben den Daten auch die zugehörige Repräsentationsinformation. Somit ist in diesem Referenzmodell geklärt, welcher Gegenstand (Ressource) identifiziert werden soll. Für die Identifikation von Objekten sind zahlreiche Mechanismen verfügbar oder in Entwicklung. Über den Leistungsumfang von Identifikatoren bestehen unterschiedliche Auffassungen z.B. über den Umfang an Information, den ein Identifikator mit sich tragen soll. Einhelligkeit herrscht darüber, dass für Objekte im Internet die Identifikation über Lokalisatoren unter Aspekten der Nachhaltigkeit keine geeignete Lösung sein kann.

Der folgende Abschnitt stellt wichtige Grundmechanismen für digitale Objekte im Internet vor und zeigt konkrete Nutzungen.

Zahlreiche Basismechanismen sind in den Dokumenten der IETF beschrieben. Sie berücksichtigen Gesichtspunkte digitaler Repräsentationen von Informationen im Internet. Ein URI (Uniform Resource Identifier) [RFC3986] ist eine kompakte Sequenz von Zeichen, die eine abstrakte oder physikalische Ressource identifiziert. Um eine maschinelle und weltweite domänenunabhängige Verarbeitung mit vernünftigem Aufwand sicherzustellen, ist eine möglichst einheitliche (generische) Syntax spezifiziert, die individuellen Bedürfnissen anpassbar ist. Die Spezifikation lässt die Art der Ressourcen offen und impliziert nicht notwendigerweise Zugreifbarkeit auf die identifizierte Ressource. Ein URI besitzt die Komponenten *Schema, Authority, Pfad, Abfrage* und *Fragment*. Die Angabe von Schemata erlaubt die Anpassung der Syntax und die nähere Beschreibung der Semantik der einzelnen Bestandteile. Ausprägungen dieser generischen Definition können sowohl URNs (Uniform Resource Name) als auch URLs (Uniform Resource Locator) sein. URNs [RFC2141] sollen der dauerhaften (persistenten) und ortsunabhängigen Identifikation von Ressourcen dienen. URLs liefern nach [RFC3986] als Unterklasse von URIs Mittel zur Lokalisierung von Ressourcen, indem sie den *primären* Zugriffsmechanismus beschreiben. URIs können beide Klassen von Identifikatoren enthalten. Handle-Systeme bieten einen weiteren Ansatz, einen globalen Namensraum für digitale Objekte zu verwalten. Diese Lösung verzichtet auf den Rückgriff auf URI-Schemata.

7.4.1 NBN

Die NBN (National Bibliographic Booknumber) ist eine Entwicklung der bibliografischen Community, um persistente und eindeutige Identifikatoren von elektronischen als auch gedruckten oder in sonstiger physikalischen Form vorliegenden Materialien zu unterstützen. RFC 3188 [RFC3188] diskutiert als informelles Dokument, wie NBNs innerhalb von URNs genutzt werden können. Die NBN ist ein generischer Name für eine Gruppe von Identifikationssystemen, die ausschließlich von Nationalbibliotheken benutzt werden, um Publikationen zu identifizieren, denen insbesondere ein geeigneter Identifikator fehlt, oder deskriptive Metadaten, die eine Ressource beschreiben. Im Zusammenhang mit URNs stellen NBNs einen Namensraum dar, der in einen Unterraum, z.B. für vertrauenswürdige Partner, unterteilt sein kann. Bei der Auflösung von URNs für nicht-elektronische Dokumente wird erwartet, dass Daten aus den nationalen bibliografischen Katalogen einschließlich des Standortes geliefert werden. Die zu identifizierenden Objekte sollen zwar *endlich* sein und eine *handhabbare* Größe aufweisen, aber eine Auflösung von Hierarchien ist durchaus vorgesehen.

Die Deutsche Nationalbibliothek (DNB)[43] definiert Regeln zum Aufbau und zur Verwaltung von URNs und ist für die Vergabe von Unterräumen zuständig. Bei der DNB ist der URN Bestandteil der Titelaufnahme. Darüber hinaus betreibt die DNB eine technische Infrastruktur u.a. zur Vergabe, Pflege und Auflösung von URNs.

7.4.2 Handle Systeme

Handle Systeme[44], wie z.B. in [RFC3650] erläutert, bieten einen weiteren Ansatz, Objekte in verteilten Umgebungen weltweit eindeutig zu identifizieren. Handles sind Namen für

[43] http://www.persistent-identifier.de
[44] siehe auch http://www.handle.net

elektronische Ressourcen und werden in verteilten Computersystemen gespeichert. Ein entsprechendes Protokoll löst Handles in jene Information auf, die nötig ist, um eine Ressource zu lokalisieren und um auf sie zugreifen bzw. sie anderweitig nutzen zu können. Da Handles auch auf Attribute oder sonstige Dienste verweisen dürfen, sind dauerhafte und sichere Einstiegspunkte für digitale Ressourcen realisierbar. Die hierfür erforderlichen Attribute wie der Speicherort oder sonstige Zustandsinformationen können ohne Änderung der Handles angepasst werden. Syntaktisch bestehen Handles aus zwei Teilen, nämlich aus Handle Naming Authority (Präfix) und Handle Local Name (Suffix), wobei das Präfix global und das Suffix lokal eindeutig sein muss. Naming Authorities als organisatorische Einheiten können hierarchisch gegliedert, mit unterschiedlichen Handle-Systemen realisiert und unabhängig administriert werden. Entsprechend der Zweiteilung gibt es für die obere Ebene einen globalen Verzeichnisdienst, der ein globales Verzeichnis verwaltet, das jene Informationen über lokale Handle-Dienste enthält, die nötig sind, um das Präfix aufzulösen.

CNRI (Corporation for National Research Initiatives) betreibt im Rahmen von *The Handle System* ein globales Verzeichnis (Global Handle Registry – GHR). Im Bereich institutioneller Repositories erlaubt DSpace das Handle-System zur dauerhaften Identifikation digitaler Dokumente zu nutzen. Auch im Rahmen von Grid-Anwendungen wird die Nutzung von Handles untersucht, um das Veröffentlichen und Finden von Metadaten über Grid-Ressourcen zu erleichtern (vgl. Handle System – Globus Toolkit Integration Project)[45].

7.4.3 DOI

DOI ist eine weitere Lösung, die von der IDF (International DOI Foundation)[46] entwickelt und betreut wird. Durch die Abstützung auf das CNRI Handle-System kann eine Entität in mehrere Entitäten aufgelöst werden (Multiple Resolution). Dieser Mechanismus erlaubt neben der Angabe der Örtlichkeit (z.B. in Form eines URL), die Versorgung mit weiteren Informationen über eine Ressource unterschiedlichster Art, wie die Angabe von deskriptiven Metadaten, von Verweisen auf alternative Darstellungen oder von inhaltlich zusammengehörigen Dokumenten. Zur Systematisierung und Standardisierung dieses Prozesses hat IDF ein Rahmenwerk für Metadaten entworfen, welches ein umfangreiches Datenmodell beinhaltet, das sich teilweise auf bestehende Standards abstützt wie die Rechtebeschreibung in MPEG-21 [ISO/IEC 21000-6:2004]. Zur Verbesserung der Interoperabilität und Administrierbarkeit ist dieses Modell abgestuft. Ein Kernschema legt die elementaren Eigenschaften einer Ressource fest, um Interoperabilität und ein Mindestmaß an deskriptiven Metadaten zu garantieren. Das indecs Data Dictionary (iDD) ist eine weitere Komponente des Datenmodells. Hier sollen die Definitionen und Beschreibungen aller Metadatenelemente unabhängig von spezifischen Metadatenschemata zur Sicherung der *semantischen Integrität* eines DOI-Systems verzeichnet sein. Schließlich ist zum Austausch von Metadaten vor allem zwischen Registrierungsagenturen (Registration Agency – RA)[47] ein Austauschformat spezifiziert.

Neben diesen technischen Spezifikationen besitzt das DOI System auch einen organisatorischen Unterbau mit der entsprechenden Policy zur Sicherstellung der nötigen Verwaltungsprozesse, wie die Registrierung, und des Betriebs der technischen Infrastruktur z.B. für die Auflösung. Die Erhebung von Beiträgen soll die Nachhaltigkeit der Lösung sichern.

[45] die News hierzu enden jedoch im Jahr 2005
[46] http://www.idf.org
[47] in Deutschland TIB (Technische Informationsbibliothek)

IDF erwartet die Verabschiedung als ISO-Standard für Ende 2008 bzw. Anfang 2009[48]. Seit April 2008 ist die Abstimmungs- und Kommentierungsphase des Committee Drafts (CD) abgeschlossen. Eine Anwendung von DOIs für wissenschaftliche Daten unter dem Gesichtspunkt der Zitierbarkeit findet sich in [Bras2007].

7.4.4 ARK

ARK (Archival Resource Key) ist ein weiterer Ansatz, auf Informationsressourcen dauerhaft zugreifen zu können [Kunz2008]. Der Ansatz beruht auf dem Grundprinzip, dass die Dauerhaftigkeit rein die Angelegenheit eines Dienstes sei und nicht die Eigenschaft eines Objekts und auch nicht durch die Syntax eines Namensschemas zu bewerkstelligen sei. Was ein Identifikator bestenfalls leisten kann ist, dass er den Nutzer zu einem Dienst führt, der eine zuverlässige Referenz bietet. Unter dieser Annahme sehen die Entwickler von ARK u.a. Indirektionsverfahren, wie sie z.B. für PURLs, Handles oder URNs nötig sind, eher kritisch, da sie eine langfristige Pflege erschweren.

ARKs sollen eine verschlankte Lösung bieten, die sich auf die Kerninfrastruktur des Internets beschränkt. Dabei wird der Rückgriff auf URLs als akzeptabel betrachtet. In Ergänzung zum direkten Objektzugriff spezifiziert ARK für die langfristige Vertrauenswürdigkeit zwei Zusatzleistungen, die mit dem URL angefordert werden können: erstens Metadaten über ein Objekt in einer kompakten Form und zweitens eine abgestufte Aussage des Providers (im technischen Sinne) zum Grad der Dauerhaftigkeit. Für diese beiden Zusatzinformationen sind somit keine eigenen Identifikatoren erforderlich.

Das URL-Namensschema besteht aus zwei, durch die Marke ark getrennten Teilen. Der erste deckt den technologiebezogenen Anteil mit Protokoll und Hostnamen ab und gilt als ersetzbar und optional. Der zweite Teil enthält den dauerhaften und global eindeutigen Namen der Organisation, die für diesen Teil zuständig ist, sowie den Namen der Informationsressource. Der optionale Qualifier erlaubt weitere Angaben zum Objekt wie Teilehierarchien oder Varianten. Eine teilweise Formalisierung von weit verbreiteten Konventionen für URLs erlaubt die Interpretation eines Qualifiers, wodurch auf die Inspektion zugehöriger Metadaten verzichtet werden kann. Der Qualifier muss nicht dauerhaft sein und kann somit z.B. technologischen Weiterentwicklungen angepasst werden.

Identifikation im Kontext von Grid/eScience

Die Identifikation von Objekten stellt weiterhin – selbst für konventionelle digitale Dokumente – eine große Herausforderung dar. Systeme zur Identifikation von digitalen Objekten sind schwierig einzuordnen und zu bewerten, wie beispielsweise die kontroversen Diskussionen innerhalb der IETF zeigen.

Das OAIS-RM trägt zur Klärung bei, was überhaupt zu identifizieren ist. Selbstverständlich bleibt die Problematik, Datenobjekte und Repräsentationsinformation geeignet zu bestimmen und einzugrenzen. Besonders in eScience-Anwendungen können sich komplexe Situationen für die Identifikation von Inhaltsinformation ergeben, wenn die Repräsentationsinformation aus unterschiedlichen und ggf. sich weiterentwickelnden Komponenten besteht und wenn sich Inhaltsdaten auf verteilen und heterogenen Systemen befinden. Hier kommen ggf. noch die Namensauflösungen virtueller Datei- oder Datenbanksysteme ins Spiel. Für die einzelnen Komponenten müsste ein Lifecycle-Modell vorhanden sein, um

[48] siehe DOI News 2008 http://www.doi.org/news/DOINewsJun08.html

die Integrität der Informationspakete auf Dauer sicherzustellen. Erschwerend können Informationsobjekte auch physische Datenobjekte wie Erkundungsproben umfassen. In den geschilderten Situationen könnte sich die Anreicherung eines reinen Identifikationssystems mit Zusatzinformationen als vorteilhaft erweisen. Standardisierte Aussagen zum Status von Komponenten und Hinweise auf alternative Möglichkeiten für die Vervollständigung eines Informationspaketes als Rückfalllösung würden einen Beitrag zur Vertrauenswürdigkeit digitaler Langzeitarchive liefern. Mit dieser Problematik setzt sich auch die nestor-AG „Langzeitarchivierungsstandards“ auseinander.

7.5 Zugangshilfen (Finding Aids, Descriptive Information)

Die Auffindbarkeit von Information ist ein weiterer Baustein eines vollständigen und vertrauenswürdigen Langzeitarchivs und hängt eng mit der Identifikation zusammen. Die Auffindbarkeit erfordert eine inhaltliche, formale und strukturelle Beschreibung der Information. Die eigentliche Erschließung unterschiedlichster fachlicher Inhalte soll hier nicht Thema sein, sondern die Beschreibung von strukturellen Beziehungen von Objekten, um zu erkennen, wie ein Aufbau von Informationspakten bzw. Informationssammlungen realisiert werden und wo Probleme für die Langzeitarchivierung liegen können.

Kataloge stellen nach wie vor wichtige Instrumente der Informationserschließung dar. Zahlreiche Standards und Regelungen führen zu einer Erschließung, die durch ein hohes Maß an Vollständigkeit, Einheitlichkeit und Interoperabilität gekennzeichnet ist. Der Standard MARC für die Repräsentation und für den Asstausch bibliografischer (und damit verbundener) Information ist die Basis für unterschiedliche Weiterntwicklungen. wie für den Standard MARC 21, der bis dahin entstandene Varianten harmonisiert und eine Internationalisierung vereinfacht. MARC XML (MARC 21 XML Schema) ist die Abbildung von MARC 21 Daten auf ein XML-Schema. MODS ist eine Entwicklung, die den Web-Anwendungen besser entgegen kommt und insbesondere Vorteile bietet, wenn auf bereits bestehende Daten in MARC 21 zurückgegriffen werden kann, da MODS eine Teilmenge von MARC 21 enthält. MODS besitzt mehr Elemente als Dublin Core, gibt aber unter dem Stichwort MODS lite Hinweise wie eine Abbildung aussehen kann. MODS ist für METS ein „bekannter“ Standard und ist als SRU Record Schema gelistet. SRU (Search/Retrieval via URL) ist ein einfaches Protokoll zum Absetzen von Anfragen an datensatzorientierte Verzeichnisse, wie bibliothekarische Kataloge oder Verzeichnisse über Web-Seiten. Kern ist hierfür die Abfragesprache CQL (Contextual Query Language), welche versucht die Lücke zwischen den komplexen Abfragesprachen, wie SQL oder XQuery, und den primitiven Sprachen, wie bei Google verwendbar, zu schließen.

Deskriptive Metadaten im Kontext der Langzeitarchivierung

Praktisch alle Standards für deskriptive Metadaten enthalten Mittel zur strukturellen Beschreibung von Objektbeziehungen digitaler und nicht-digitaler Art. Damit eröffnet sich auch eine Vielzahl von Möglichkeiten, Informationspakete und -sammlungen im Sinne von OAIS zu beschreiben. Dieser Umstand birgt die Gefahr nicht (maschinell) erkennbarer Redundanzen und künftiger Interpretationsprobleme infolge impliziter Annahmen, was zu Informationsverlust führen kann, insbesondere wenn Sammlungen sukzessive über einen längeren Zeitraum aufgebaut werden oder wenn in komplexen Objekten oder Sammlungen Migrationen nötig werden. Eine strikte „Arbeitsteilung“ und eine möglichst explizite Beschreibung von Objektbeziehungen wären hilfreich. Unvermeidbare Redundanzen ließen sich mit einem regelbasierten Ansatz unter Kontrolle bringen. Ein praktisches Hilfsmittel

könnte die Erstellung von standardübergreifenden Profilen sein, ähnlich denen in serviceorientierten Architekturen, um die (historisch gewachsene) Vielfältigkeit zu kanalisieren. Eine weitere Problemstellung betrifft die Navigierbarkeit in komplex aufgebauten Objekten und die Auswahl der dafür nötigen Sprachen. Hier stellt sich die Frage nach den Grenzen einer einfachen Sprache wie CQL.

7.6 Allgemeine Dienste (Common Services)

Der folgende Abschnitt stellt Standards vor, die für die Realisierung des Informationsaustausches zwischen Archiven und Produzenten (Ingest) bzw. Konsumenten (Access) sowie zwischen kooperierenden Archiven oder sonstigen Dienstleistern relevant sind oder relevant werden könnten.

7.6.1 WS-Basisdienste

Standardisierte Dienste und Protokolle oberhalb der *Transport- und Vermittlungsschicht* sind Voraussetzung, um den Aufwand für die Herstellung von Interoperabilität unterschiedlicher und verteilter Anwendungen zu beschränken. Einfachheit sowie Sprach- und Herstellerunabhängigkeit sind neben dem Leistungsumfang entscheidend für einen Erfolg.

Große Erwartungen sind mit der serviceorientierten Architektur verbunden, deren Grundbausteine im Folgenden skizziert werden. Möglichst hohe Abstraktion von konkreten Implementierungen, Erweiterbarkeit sowie strukturierte Beschreibung der jeweiligen Leistung ergänzbar um Annotationen auf Basis von XML charakterisieren die Standards. SOAP, WSDL und UDDI werden häufig im Zusammenhang genannt, obwohl sie nicht von der gleichen SDO stammen. An dieser Stelle soll erwähnt werden, dass sich eine SOA auch mit anderen Standards bzw. ohne diese Standards realisieren lässt. Ein alternativer Ansatz läuft unter dem Begriff REST (Representational State Transfer), der zwar flexibel ist und sich auf wenige Basistechnologien stützt, aber bezüglich Effizienz deutliche Grenzen aufweist.

SOAP

Im Zusammenhang mit Web Services hat der W3C-Standard SOAP (vormals als Akronym für Simple Object Access Protocol) eine dominierende Rolle für die Nachrichtenübermittlung übernommen. Grundsätzlich kann SOAP oberhalb (oder auch innerhalb) zahlreicher Transportprotokolle verwendet werden, doch am häufigsten ist die Bindung an HTTP bzw. HTTPS, welche in der Spezifikation auch näher ausgeführt wird. Der Nachrichtenaustausch kann synchron oder asynchron erfolgen. Der Standard erlaubt auch selbst definierbare Austauschmuster (Message Exchange Pattern – MEP). Auf der Basis eines abstrakten Frage-Antwort-Musters definiert der Standard eine einheitliche und programmiersprachenunabhängige Darstellung von RPCs, einem in der Programmierwelt häufig genutzten Kommunikationsmuster für synchronen Nachrichtenaustausch. Zur SOAP-Spezifikation gehört auch ein Fehlerbehandlungsmechanismus. Die Darstellung der gesamten Nachricht (Envelope, Header, Body) erfolgt in XML, was ggf. zu einem unakzeptablen Overhead führen kann. Unterschiedliche Standardisierungsansätze versprechen Abhilfe (vgl. z.B. Ausführungen zu ASN.1).

WSDL

Ein weiterer Baustein in einer serviceorientierten Architektur ist die Beschreibung von Diensten, um die verfügbaren Leistungen für potenzielle Konsumenten verständlich darzustellen. Mit dem W3C-Standard WSDL (Web Services Description Language) können als XML-Dokument alle öffentlichen Schnittstellen eines Dienstes in abstrakter Form beschrieben werden einschließlich von Angaben, nach welchen Kommunikationsmustern (MEP) einzelne Operationsaufrufe abgearbeitet werden. MEPs können Entwickler selbst definieren, doch gebräuchliche Muster sind bereits vordefiniert. Diese abstrakten Beschreibungen müssen zur Realisierung an ein konkretes Nachrichtenformat und Transportprotokoll *gebunden* werden. Standardmäßig ist die Bindung an SOAP bzw. SOAP und HTTP(S) wie alternativ die direkte Bindung an HTTP(S) vorgesehen. Prinzipiell ist auch die Bindung an andere Protokolle realisierbar. Ein weiteres Element im WSDL-Dokument bestimmt durch die Angabe von Netzwerkadressen, wo der Dienst ausgeführt werden kann. Ein Dokumentationsteil bietet Raum für eine ausführliche Beschreibung des Dienstes, und ein Mechanismus zur Fehlerbehandlung gehört ebenfalls zum Umfang dieses W3C-Standards.

UDDI

Als nützlich für eine serviceorientierte Architektur wurde auch eine standardisierte Plattform für das Veröffentlichen und Auffinden von statischen Diensten und Diensteanbietern erachtet. UDDI (Universal Description, Discovery and Integration) bietet eine durch OASIS standardisierte Lösung an. WS-I Basic Profile macht diesen Standard zur Vorgabe, falls ein Verzeichnisdienst genutzt werden soll. Die Architektur von UDDI ist für verteilte und offene Umgebungen ausgelegt, so dass ein Netzwerk von öffentlichen Verzeichnissen aufgebaut und verwaltet werden kann. UDDI ist jedoch auch für eine organisationsinterne (private), ggf. verteilte Nutzung einsetzbar. Kernaufgabe ist die Registrierung von Diensten; diese müssen aber nicht unbedingt mit WSDL formuliert sein. Hierzu gehören Angaben zu Methoden und Parametern wie auch zur technischen Realisierung eines Dienstes, wie Protokolle oder Formate für den Nachrichtenaustausch. UDDI sieht auch die strukturiere Beschreibung von Diensteanbietern vor. Darüber hinaus definiert der Standard Regeln zur Identifikation von Einträgen und bietet formalisierte Hilfsmittel zur Kategorisierung von Einträgen, um insbesondere den Einstieg in eine Suche zu erleichtern. Zur Veröffentlichung, zur Suche und zur Verwaltung von Einträgen und der Kooperation von Verzeichnissen im Sinne der Serviceorientierung sind Programmierschnittstellen spezifiziert.

WS-Basisstandards im Kontext von Grid/eScience

Den drei genannten Standards gingen eine rege Diskussion und etliche unterschiedliche Implementierungen voraus, die neben Einzelfragen des Leistungsumfangs auch grundsätzliche Fragen einer verteilten Architektur betreffen. Zu den grundsätzlichen Themen gehört die Notwendigkeit der Beschreibung von Diensten in Form einer Datei wie in WSDL oder die Einführung eines „schwergewichtigen“ Protokolls wie SOAP (im Standard als leichtgewichtig bezeichnet). Spezielle Anforderungen an diese Standards, die ursprünglich durch Business-to-Business-Anwendungen motiviert waren, kommen aus dem Bereich Grid, insbesondere wegen der Forderung nach hohem Durchsatz, wegen der hochgradigen Verteilung und der dynamischen Belegung und Freigabe von Ressourcen. So können diese Standards nur für Teilprobleme oder nach Erweiterung bzw. Anpassung übernommen werden.

Mit Grimoires[49] beispielsweise steht ein UDDI-konformes Verzeichnis zur Verfügung, das u.a. die im Grid häufige Authentifizierung über Zertifikate unterstützt als auch die Registrierung von Workflows als spezielle Dienste. Spezifische Anforderungen haben auch zu eigenständigen Standards geführt. Die *Zustandslosigkeit* von Web Services (SOAP) und die Herstellung von *Zustandsbehaftung*, wie in üblichen Web-Anwendungen mit Hilfe unterschiedlicher Mittel (Cookies, Session-IDs) realisiert, ist für Grid-Anwendungen nicht geeignet. Dieser Mangel führte zur Entwicklung des Standards WSRF, der im Folgenden noch näher beschrieben wird.

Trotz der Zuordnung der besprochenen Standards zum OAIS-Modellelement Common Services, besteht direkter Bezug zu den konzeptionellen Bausteinen wie Representation Information, Preservation Description Information und Descriptive Information. Dies liegt in der Philosophie von SOA, einzelne Dienste zu höherwertigen Diensten zusammenzusetzen. Dies erfordert Mechanismen für die Darstellung und den Austausch von „Information". Beschreibungen in WSDL und UDDI geben (öffentlich) Auskunft über die Bedeutung von Diensten und Methoden und somit ggf. auch wie Daten verarbeitet werden können oder tatsächlich wurden. Die Ds (für Description) in den Standardnamen weisen also darauf hin, dass direkte Beiträge zu Erhaltungsmetadaten möglich sind. Die Verarbeitbarkeit und das Verarbeiten von Datenobjekten durch unterschiedliche Services kann Information zur Bedeutung dieser Objekte und zur Verarbeitungshistorie (Provenance) geben. Umfangreiche Beschreibungsmöglichkeiten bestehen in UDDI, womit auch deskriptive Metadaten untergebracht werden können. In konventionellen Umgebungen bestehen ebenfalls Dokumentationsmöglichkeiten, um die Bedeutung von Operationen und damit auch der zu verarbeitenden Objekte zu beschreiben, teilweise mit adäquateren Mitteln als mit (bisherigen) Sprachen der Serviceorientierung. Der größte Nachteil ist jedoch die Uneinheitlichkeit, z.B. durch die Bindung an bestimmte Programmier- oder Spezifikationssprachensprachen, was eine Übernahme in Schemata für Erhaltungsmetadaten deutlich erschwert.

WSRF

Der OASIS-Standard WSRF (Web Services Resource Framework) spezifiziert einen mit Web Services kompatiblen Mechanismus, um den Informationsaustausch zwischen einzelnen Aufrufen (*Zustandsbehaftung*) einheitlicher zu organisieren. Hierzu spezifiziert WSRF ein XML-Dokument (Resource Properties Document), das eine durch WSDL referenzierte Ressource nach einem einheitlichen Grundschema beschreibt. Dieses XML-Dokument, welches zusammen mit dem WSDL-Dokument als WS-Resource bezeichnet wird, dient als Basis, um die Kommunikation mit einer Ressource durch eine definierte Menge von Nachrichten einschließlich Fehlermeldungen zu vereinheitlichen. Die Interfaces der Web Services werden also in einer standardisierten Weise erweitert. Die Nachrichten umfassen die Abfrage und Aktualisierung von Eigenschaften einer Ressource sowie den Austausch eines gesamten beschreibenden Dokuments. Der standardisierte Zugriff auf aktuelle Zustandsinformation von Ressourcen und die einheitliche Verwaltung von Identitäten erweisen sich als besonders vorteilhaft, wenn eine Anwendung viele einzelne Ressourcen in verteilten Umgebungen, wie z.B. im Grid, verwalten soll. Des Weiteren kann eine Anwendung in standardisierter Weise bestimmen, ob sie über Änderungen von Eigenschaften einer WS-Resource entsprechend dem Subscriber-Modell informiert werden möchte. Der Standard gibt außerdem einer Anwendung die Möglichkeit, den Lebenszyklus einer WS-Resource zu steuern. Wegen der Vielfältigkeit möglicher Szenarien, eine WS-Resource ins Leben zu

[49] http://twiki.grimoires.org/pub/Grimoires/doc/intro.html

rufen, beschränkt sich die Spezifikation jedoch auf die Beendigung durch sofortige Zerstörung oder Angabe bzw. Aktualisierung von Zeitpunkten.

Weitere SOA-Standards

Das Funktionieren serviceorientierter Architekturen erfordert weitere Standards, um das Zusammenspiel der Dienste zu garantieren. Das Ausführen von Geschäftsprozessen, das zuverlässige Ausführen von Transaktionen, das Verwalten von Ressourcen und die Sicherung von Qualität sind weitere Themen der Standardisierung. Normierte *Profile*, wie z.B. von WS-I entwickelt, erleichtern die Einrichtung kompatibler Konfigurationen. Für die Vertrauenswürdigkeit einer verteilten Architektur spielt das Thema Sicherheit eine entscheidende Rolle, welches im folgenden Abschnitt kurz erläutert wird.

7.6.2 Sicherheit

Sicherheit in verteilten und dynamischen Systemen stellt eine extreme Herausforderung für den Entwurf und Umsetzung von Sicherheitsarchitekturen dar. Laufzeiteffizienz, einfache Verwaltbarkeit und Nutzerkomfort sind wichtige Randbedingungen. Zur Aufwandsbegrenzung gehört auch eine einfache Implementierbarkeit. Im Allgemeinen umfasst Sicherheit die Vertraulichkeit, Authentizität, Integrität und Verfügbarkeit der Daten. Außerdem soll der Sender bzw. der Empfänger von Nachrichten das Versenden bzw. das Empfangen nicht abstreiten können (Nicht-Anfechtbarkeit). Anonymität ist als neues Sicherheitsziel hinzugetreten.

Basistechniken

Um diese Sicherheitsziele zu erfüllen, haben sich bereits Standards für verteilte Systeme etabliert, die jeweils im Stande sind, einzelne Aspekte abzudecken. Kryptografische Verfahren, insbesondere Verschlüsselungsverfahren, sind wichtige Grundmechanismen, um Information vor ungewollten Einblicken zu schützen. Symmetrische und asymmetrische Verfahren bilden die zwei Hauptkategorien. Vertreter für die erste Kategorie sind DES (Data Encryption Standard) und der Nachfolger AES (Advanced Encryption Standard), und für die zweite Kategorie Elgamal und, deutlich dominierend, RSA. Symmetrische Verfahren nutzen für die Ver- und Entschlüsselung denselben Schlüssel während asymmetrische Verfahren je einen *öffentlichen* und einen *privaten* Schlüssel pro Kommunikationspartner vorsehen. Ein weiterer Baustein zur Erzielung von Sicherheit ist die Authentifizierung als Überprüfung einer behaupteten Identität eines Subjekts (Personen, Software). Drei Kategorien lassen sich unterscheiden. Eine einfach zu verwaltende Kategorie beruht auf Verfahren, denen ein gemeinsames Geheimnis zu Grunde liegt, wie z.B. Passwörter oder PINs. Die nächste Kategorie bedient sich *öffentlicher Schlüssel.* Mit Hilfe des öffentlichen Schlüssels eines Subjekts, der seine Nachricht oder Teile davon mit seinem privaten Schlüssel verschlüsselt und versendet, kann die Identität des Senders überprüft werden. Für große Nutzerzahlen ist dieses Vorgehen schwer zu verwalten, so dass als weitere Kategorie *zertifikatbasierte* Verfahren entwickelt wurden, bei denen für die Nutzer ein vertrauenswürdiger Dritter (Certification Authority – CA) ein Zertifikat ausstellt. Dieses Zertifikat wird mit dem privaten Schlüssel der CA signiert und die öffentlichen Schlüssel der CAs werden als bekannt vorausgesetzt. Voraussetzung für dieses Verfahren ist eine Public-Key-Infrastruktur (PKI). Der wichtigste Standard hierfür läuft unter dem Kürzel X.509. Eine Alternative zur PKI ist in Kerberos [RFC4120] realisiert, bei dem sich ein Nutzer mittels eines Passwort-Verfahrens bei einer Stelle (Key Distribution Center – KDC) anmeldet, die dann einen Nutzer mit einem temporären Schlüssel (Session Key) versorgt, der mit dem

öffentlichen Schlüssel dieser Stelle verschlüsselt ist. Kerberos ist als Sicherheitsmechanismus z.B. im Network File System (NFS) vorgesehen [RFC3010]. Ein weiterer Baustein in einer Sicherheitsarchitektur muss die Verletzung der Integrität einer Nachricht (Message Integrity) feststellen können. Kryptografische Hash-Funktionen wie MD5[50], SHA-1[51] oder SHA-256[52] sind Mittel, um bösartige Manipulationen zu erkennen. Da der Hash-Wert von jedermann erzeugt werden kann, der die Nachricht kennt, sind weitere Vorkehrungen nötig, wie die Verschlüsselung der Nachricht und des Hash-Wertes. Eine weitere Möglichkeit, als Message Authentication Code (MAC) bezeichnet, ist die Nutzung von Hash-Funktionen mit einem gemeinsamen geheimen Schlüssel [RFC2104].

SAML

Serviceorientierte Architekturen, benötigen Ansätze für die Handhabung von Identitäten, die der losen Kopplung von Anwendungen und deren Verbindung mit unterschiedlichen Nachrichtenprotokollen gerecht werden. Um Dienste in Bereichen mit unterschiedlichen Anwendungen, Verwaltungen und Sicherheitsregeln einfach nutzen und verwalten zu können, ist der standardisierte Austausch von Informationen zur Identität erforderlich. Der OASIS-Standard SAML (Security Assertion Markup Language) [OASIS-SAML2005] bietet hierfür Unterstützung. Single Sign-On bringt dem Nutzer (genereller: *Subjekt*) erhebliche Vereinfachung, in dem er sich nur einmal bei einem so genannten Identity Provider einloggt, um dann von unterschiedlichen Service Providern erkannt zu werden. Falls die konkrete Identität eines Nutzers nicht von Bedeutung ist oder die Privatheit gewahrt werden soll, kann eine *attributbasierte* Autorisierung erfolgen, d.h. bestimmte Attribute, die einem Subjekt zugeordnet sind, wie die Zugehörigkeit zu einem Institut und eine bestimmte Position, reichen zur Erkennung aus. Neben diesen Informationen über die Identität und über die Eigenschaften eines Subjekts, können Aussagen zur Autorisierung hinzugefügt werden. Diese drei Typen von Angaben bezeichnet der Standard als *Zusicherungen* (Assertion) und werden von einer SAML Authority gemacht. Auf Grund dieser Ausdrucksstärke sind Zusicherungen auch als Sicherheits-Token wie z.B. X.509-Zertifikate geeignet, um z.B. die Sicherheit von SOAP-Nachrichten zu erhöhen. Der Standard spezifiziert außerdem Protokolle zum Austausch von Zusicherungen zwischen den einzelnen definierten Rollen der SAML-Architektur sowie die Abbildung (*Bindung*) dieser Protokolle auf standardisierte Protokolle zum Nachrichtenaustausch und zur Kommunikation (wie SOAP und HTTP). Schließlich beschreibt der Standard noch *Profile*, um den Umfang einzuschränken bzw. zu erweitern. So können Attribute in den Zusicherungen so profiliert werden, dass sie den Attributen in LDAP entsprechen.

XACML

Sicherheitsregeln in einer verteilten und heterogenen Umgebung zu formulieren, durchzusetzen und zu pflegen, erweist sich als aufwändig und risikobehaftet. Abhilfe kann der von OASIS herausgegebene Standard XACML (eXtensible Access Control Markup Language) [OASIS-XACML2005] bieten, indem er eine ausdrucksstarke und flexible Sprache zur Formulierung von Sicherheits-Policies definiert. Mit der angebotenen Sprache ist es möglich, eine Vielzahl etablierter Methoden der Zugangskontrolle abstrakt zu beschreiben. So können individuelle Regeln und Policies mittels vor- oder selbstdefinierter Kombinationsregeln in einem Satz zusammengefasst werden. Beispielsweise kann damit festgelegt wer-

[50] http://www.ietf.org/rfc/rfc1321
[51] http://www.ietf.org/html/rfc3174
[52] http://csrc.nist.gov/publications/fips/fips180-3/fips180-3_final.pdf

den, dass eine Regel, die zur Ablehnung eines Zugriffs geführt hat, nicht durch eine weitere Regel wieder außer Kraft gesetzt werden kann. Eine weitere Fähigkeit der Sprache besteht darin, verschiedenen Subjekten, die an einer Aktion beteiligt sein sollen, mit unterschiedlichen Rechten zu versehen. Somit kann man z.B. definieren, dass mindestens zwei Personen an einer kritischen Transaktion beteiligt sein müssen. Hohe Flexibilität wird durch die Zuordenbarkeit von Attributen zu den im Standard spezifizierten Entitäten wie *Subjekt*, *Ressource*, *Aktion* und *Umgebung* erreicht. Eine wichtige Anwendung hierzu ist die rollenbasierte Zugriffskontrolle. Liegt die Ressource als XML-Dokument vor, kann mittels XPath direkt auf die dortigen Attribute zugreifen, was die Realisierung eines *inhaltsbasierten* Zugriffs erleichtert. Außerdem definiert der Standard auch den Zugriff auf in LDAP abgelegte Attribute. Um eine Entscheidung zur Autorisierung treffen zu können, sind Regeln und Policies trotz heterogener Umgebungen zuverlässig auszuwerten; d.h. für die Attribute der einzelnen Entitäten waren im Standard plattformunabhängige Typen und Operationen zu definieren. Neben einen einfachen Wertevergleich kennt der Standard arithmetische, mengenorientierte und boolesche Funktionen, wobei Anleihen aus den Standards XQuery, XPath und MathML genommen werden. Des Weiteren bietet XACML Beschreibungshilfen für die Verwaltung von Policies in verteilten Umgebungen. Weitere Co-Standards beschreiben die Anwendung des Standards für verschiedene Randbedingungen und Szenarien. Ein Profil beschreibt erstens, wie SAML zum Transport von Informationen, die in XACML formuliert sind, genutzt werden kann, und zweitens, wie die dortigen Mechanismen zur Erstellung, Abfrage und Validierung von *Zusicherungen* einsetzbar sind [OASIS-SAMLprofile2005]. Damit ist eine Möglichkeit beschrieben, die im Entwurf von XACML nicht vorgesehenen Eigenschaften zu ergänzen, und mittels der in SAML definierten Protokollbindungen zu operationalisieren.

Der OGC-Standard GeoXACML ist eine Erweiterung von XACML, der hilft, den Zugriff auf geografische Daten zu regeln, sofern rechtliche, kommerzielle oder sonstige Anforderungen wie die Reife von Forschungsarbeiten bestehen [OGC2008]. Hierzu definiert der Standard geometrische Datentypen und topologische Funktionen, um raumbezogene Zugriffsregeln formulieren zu können. Diese Definitionen können wiederum auf ein standardisiertes Architekturmodell der OGC zurückgreifen. Diese Erweiterung ist ein Beispiel, wie sich ein feingranularer Zugriff in einer verteilten Umgebung realisieren lässt.

Sicherheitsstandards im Kontext Grid/eScience und Langzeitarchivierung

Der Druck, Sicherheitsstandards zu entwickeln, ist hoch, wie u.a. die Prioritätenliste des OGF zeigt. Erfreulicherweise kann eine Konvergenz bei der Entwicklung von Standards und deren Integration auch in Grid-Produkten festgestellt werden. In die Entwicklung des Standards SAML sind Beiträge der Liberty Alliance[53] eingeflossen. Die Spezifikation ID-FF (Liberty Identity Federated Framework) unterstützt SAML 2.0. Zuarbeit leistete auch die Middleware-Initiative Internet2[54]. Das Open-Source-Produkt Shibboleth ist u.a. kompatibel zu SAML 2.0 und die GridShib SAML Tools widmen sich der Integration von Globus Toolkit und Shibboleth. Zur Verwendung von SAML in Grids siehe [Ried2008].

Die bisherigen Standards sind in der Lage eine Reihe von Forderungen, die aus den Bereichen eScience und Grid stammen, sowohl aus Sicht der Nutzer als auch der Betreiber abzudecken. Single Sign-On, als eine Hauptanwendung von SAML, kann nun in einer einheitlicheren Form realisiert werden, und XACML erleichtert eine feingranulare Rechtever-

[53] http://www.projectliberty.org
[54] http://www.internet2.edu

gabe. Die Berücksichtigung bestehender Sicherheitsstandards durch Erweiterbarkeit bzw. durch die explizite Spezifikation von Schnittstellen und Profilen erleichtert die Einführung in bereits vorhandene Infrastrukturen. Die Entwicklung der Standards befindet sich jedoch noch in einer eher lebhaften Phase. Grundsätzliche Fragestellungen ergeben sich in hochgradig dynamischen Organisationsformen, in den sich Mitglieder mehr oder weniger beliebig anschließen und sich wieder verabschieden können. Eine Kernfrage ist dabei festzustellen, inwieweit sich die Beteiligten vertrauen können. Für die Formulierung, Verhandlung und Verwaltung von Vertrauen mit Hilfe so genannter Trust Management Systeme (TMS) zeichnet sich noch deutlicher Forschungsaufwand ab [Chak2007], so dass hier noch keine Standardisierung in Sicht ist.

Sicherheit ist ein Kriterium für vertrauenswürdige Langzeitarchive; eine weitere Forderung ist die angemessene Zugänglichkeit zur Information, die aus technischen oder auch wirtschaftlichen Gründen durch Sicherheitsmassnahmen nicht konterkariert werden darf. Die Archivierung digital signierter Daten ist ein wichtiges Problemfeld, das aber bereits Gegenstand der Standardisierung ist (siehe u.a. [RFC4998]).

Neben technischen Komponenten gehören zu einer Sicherheitsarchitektur auch organisatorische und soziale Aspekte. Eine Facette davon spricht die nestor-Studie des GFZ an, nämlich das Misstrauen, das einer zentralen Administration entgegengebracht wird [Klum2007]. Sicherheit gesamtheitlich zu betrachten, spiegelt sich auch in der internationalen Standardisierung wider. Der ISO-Standard 27001 [ISO/IEC 27001:2005] spezifiziert Anforderungen für die Herstellung, Einführung, Betrieb, Überwachung, Wartung und Verbesserung eines dokumentierten Informationssicherheits-Managementsystems unter Berücksichtigung von Risiken innerhalb einer gesamten Organisation, und der ISO-Standard 27002 [ISO/IEC 27002:2005] stellt Richtlinien und allgemeine Prinzipien auf für die Initiierung, Einrichtung, Wartung und Verbesserung eines Informationssicherheits-Managements in einer Organisation.

Zu einer Sicherheitsarchitektur gehört auch eine Komponente zur Beobachtung (Monitoring) des Systems. Die Aufzeichnung und Analyse des Gebrauchs aller gefährdeten Ressourcen erlaubt das Erkennen missbräuchlicher Nutzung. Darüber hinaus dient das Monitoring der Leistungsanalyse, um eine statische oder dynamische Optimierung bzw. eine Selbstheilung zu ermöglichen. Eine weitere Anwendung ist die Erfüllung von gesetzlichen Auflagen zur Dokumentation und natürlich die automatische Generierung von Provenance Information. Schließlich können die aufgezeichneten Daten der Abrechnung von Leistungen dienen.

7.6.3 Abrechnung (Accounting)

Obwohl die Abrechnung von Leistungen im Grid laut nestor-Studie der GFZ [Klum2007] noch nicht als offene Frage wahrgenommen wird, sollen dieses Thema und Standardisierungsaktivitäten hier kurz behandelt werden. Die Verteiltheit und die Vielzahl unterschiedlicher Leistungen im Grid und die Quantifizierung ihrer Güte wie auch die Organisationsformen der Nutzer und Anbieter sind eine Herausforderung für die Standardisierung der Abrechnung von Leistungen (Accounting). Nutzer und Ressourcen sind zu identifizieren und einheitlich zu beschreiben, und Informationen über erbrachte bzw. mangelhaft oder nicht erbrachte Leistung Leistungen sind zu definieren, so dass sie als Abrechnungsgrundlage dienen können. Ein Mechanismus muss dafür sorgen, dass diese Einzelinformationen zusammengeführt und ausgewertet werden können. Bestehende Standards sind bereits in

der Lage, Teilaufgaben zu unterstützen. SAML z.B. erlaubt die Verwaltung von Identitäten in verteilten Umgebungen, Services können sich mittels WSRF abrechnungsrelevante Daten über Einzelaufrufe hinweg merken und der Standard WS-Management der DMTF (Distributed Management Task Force) [DMTF2008] vereinfacht die Einbettung in ein ressourcen- und aufgabenübergreifendes Managementsystem. Eine Bestandsaufnahme zu Accounting-Systemen für das Grid im Zusammenhang mit Standardisierungsaktivitäten findet sich in [Göhn2006]. Im Rahmen der OGSA läuft diese Thematik unter dem Begriff Resource Usage Service (RUS). Eine Roadmap der zuständigen Arbeitsgruppe (RUS-WG) zeigt deutlich die Abstützung auf Standards aus dem Bereich WS. Dies kann als weiteres Indiz für eine Konvergenz im Bereich der Standardisierung und die integrierende Wirkung serviceorientierter Techniken gewertet werden.

Die Abrechnung von bestimmten Leistungen für die Langzeitarchivierung wirft jedoch zusätzliche Fragen für die Abrechnung von Leistungen auf. Eine Besonderheit liegt in der „dauerhaften“ Allokierung von Ressourcen, insbesondere von Speicherkapazitäten, und in der schweren Quantifizierbarkeit der Qualität von Diensten, die den Erhalt von Information jenseits der Datenebene (Bitstream Preservation) sichern sollen.

8 Resümee, Hinweise und Handlungsempfehlungen

Im vorherigen Kapitel wurde eine Reihe von Standards entlang der Modellelemente des OAIS-RM erläutert und im Kontext von Langzeitarchivierung und Grid/eScience im Detail diskutiert. Die Darstellung gibt einen etwas tieferen Einblick in die Vielzahl von Einzelfragen und in den Stand der Standardisierung.

Im Folgenden wird die Situation bezüglich der Standardisierung zusammengefasst. Dann folgen Hinweise für die Fälle, in denen Gedächtnisorganisationen mit Anforderungen aus dem Bereich eScience/Grid konfrontiert werden bzw. Grid-Technologien für die Langzeitarchivierung nutzen wollen. Handlungsempfehlungen zur Standardisierung schließen das Kapitel ab.

8.1 Stand der Standardisierung

Die Lage der Standardisierung ist sowohl für die Grid-Technologien als auch in der Langzeitarchivierung nicht zufrieden stellend. Ursachen sind in der technischen Komplexität und in der Neuartigkeit der Technologien bzw. der Aufgabenstellung zu sehen. Offene Fragestellungen widersprechen schließlich dem Prinzip von Standards, einen anerkannten bzw. konsolidierten Stand der Technik wiederzugeben.

Bei einer Bewertung des Standes der Standardisierung im Grid-Bereich sollte im Auge behalten werden, dass es sich um ein neues und sehr umfassendes Paradigma handelt und teilweise sehr komplexe Fragestellungen vorab zu lösen sind wie z.B. das Scheduling zur optimalen Nutzung technischer Ressourcen. Ein wissenschaftlicher Workflow hat deutlich andere Anforderungen als einer im e-Business, so dass nicht ohne weiteres auf die dort existierenden Standards zurückgegriffen werden kann (eine Darstellung der Unterschiede findet sich z.B. in [Rado2007]). Anders als bei vielen anderen technischen Paradigmen der IT, wie z.B. bei der objektorientierten Programmierung, steht das Grid-Paradigma von Anfang an in einer starken Wechselwirkung mit Formen der Projektabwicklung und Organisation.

Positiv zu bewerten ist, dass mit OGSA (Open Grid Services Architecture) des OGF (Open Grid Forum) ein umfassendes Rahmenwerk zur Verfügung steht, das den Standardisierungsprozess in eine effektive Form bringen kann. Zwar sind bisher nur relativ wenige Spezifikationen und *erläuternde* Dokumente verfügbar, und die Autoren erwarten noch einen Zeitbedarf von etlichen Jahren für eine vollständige Abdeckung von OGSA durch *normative* Dokumente, dennoch finden OGSA und zugehörige Spezifikationen Beachtung und Unterstützung durch wichtige Grid-Projekte [OGF-OGSA 2008]. Die Ausrichtung der Grid-Architektur an Web Services für den Aufruf von Methoden und den Austausch von Daten in lose gekoppelten Systemen kann als weiterer Vorteil gesehen werden. Für die Beschreibung von Diensten, den Austausch von Nachrichten und die Strukturierung von Daten kann, wenn auch mit Einschränkungen, auf bestehende Basisstandards zugegriffen werden. Teilweise sind diese Standards schon in Systemen zur Langzeitarchivierung im Einsatz bzw. in den zu Grunde liegenden Produkten enthalten. Dabei darf jedoch nicht übersehen werden, dass sich neue Abhängigkeiten ergeben und dass auch bei den Web Services sich die Standardisierung noch in einer sehr lebhaften Phase befindet und eher von Anwendungen des e-Business getrieben wird.

Bei der Entwicklung von Standards im Grid-Bereich spielen Aspekte der Langzeitarchivierung bisher eine untergeordnete Rolle im Vergleich zu Themen wie Performanz, Zuverlässigkeit, Skalierbarkeit und Kollaboration. Hier bieten die Erfahrungen und Standardisierungsaktivitäten (traditioneller) Gedächtnisorganisationen eine geeignete Ergänzung. Wie die Ausführungen im vorherigen Kapitel gezeigt haben, beschäftigen sich diese Organisationen weltweit intensiv mit der Erhaltung digitaler Information und mit der hierfür erforderlichen Standardisierung. Aus der obigen Diskussion ist erkennbar, dass die bisherigen und in Entwicklung befindlichen Standards durchaus Potenzial bieten, um Anforderungen aus dem Bereich Grid/eScience zu bedienen. Doch ähnlich wie beim Grid-Computing fehlt den Lösungen zur Langzeitarchivierung die nötige Reife, um alle Funktionen eines Archivsystems durch normative Standards abzudecken. Auf den teilweise vorläufigen bzw. experimentellen Charakter von Gegenständen der Standardisierung wird in einigen Standards sogar ausdrücklich hingewiesen.

Eine weitere Ursache für die unbefriedigende Lage ist in den bestehenden Organisationsformen zu erkennen, die eine optimale Standardisierungsarbeit der Gedächtnisorganisationen verhindern. Der Nutzen bzw. die Notwendigkeit von Standards wird zwar erkannt, aber die erforderlichen Ressourcen und Strukturen stehen für diese anspruchsvolle Aufgabe, zumindest in Deutschland[55], nicht zur Verfügung. Somit können die notwendigen Prozesse zur Entwicklung und zur nationalen und internationalen Abstimmung nicht im ausreichenden Maße stattfinden. Eine umspannende und bereits detaillierte Architektur, vergleichbar mit OGSA, und ein zugehöriger methodischer Rahmen sind für die Langzeitarchivierung noch nicht erarbeitet, obwohl das OAIS-RM (Open Archival Information System – Referenzmodell) eine geeignete Ausgangsbasis bietet. Fälle redundanter, verspäteter, inhaltlich nicht adäquater Standards könnten vermieden werden[56]. Und nach wie vor fehlt als Arbeitsgrundlage eine klare, konsistente und akzeptierte Taxonomie für die digitale Langzeitarchivierung (siehe u.a. [Rieg2008]). Der Gebrauch von Kernbegriffen wie digitales Objekt, digitale Entität, Manifestation, Repräsentation, Inhalt, Content, Version, Migration ist uneinheitlich oder bleibt vage.

[55] Ein Hinweis auf die angemessene Ausstattung beim Vorhaben GDFR sei an dieser Stelle erlaubt.
[56] siehe hierzu z.B. Olaf Brandt, Markus Enders, SUB Göttingen, PREMIS and METS, METS Opening Day, 07.05.2007

8.2 LZA-Unterstützung für die eScience-Community

Durch die Anforderungen aus dem eScience-Bereich und aus der Nutzung von Grid-Technologien entstehen für Gedächtnisorganisationen neue Fragestellungen. Praktiken und Standards für den Umgang mit digitalen Objekten traditioneller Gedächtnisorganisationen wie Bibliotheken oder Staatsarchive sind nur bedingt anwendbar. Wesentliche Gründe hierfür sind nachfolgend kompakt zusammengefasst, wobei zu berücksichtigen ist, dass diese Aussagen stark von der jeweiligen Community abhängen und dass durch fachliche Anforderungen und rechtliche oder wirtschaftliche Rahmenbedingungen sehr spezifische Gründe hinzutreten können:

- Die Größe der digitalen Objekte. Prinzipien der Atomarität, die die Ein- und Auslieferung von Objekten bei Änderungen, Erweiterungen oder Anfragen als Ganzes fordern, können sich als nachteilig erweisen.

Dieses Problem kann teilweise mit technischen Mitteln durch die Bereitstellung von Ressourcen überwunden werden und ist daher eher wirtschaftlicher Art. Möglichkeiten der Effizienzsteigerung auf technischer Ebene sind nach wie vor Gegenstand der Forschung und der Standardisierung.

Wird aus Effizienzgründen die Änderbarkeit von Teilen von Informationspaketen zugelassen, sind Fragen der Identität und Authentizität sowie der Aufzeichnung der Provenance zu klären.

Die folgenden Probleme liegen auf der Ebene Information – und Wissen, da ein Bezug zur *Knowledge Base* der Designated Community herzustellen ist. Die Communities der eScience sind in sich und in ihrer Gesamtheit durch die hohe Spezialisierung sehr heterogen, was eine Standardisierung bzw. die Erkennung von Grenzen einer Standardisierung deutlich erschwert.

- Abgrenzbarkeit der Objekte. Es fehlt der diskrete Charakter üblicher Dokumente (zeitlicher und räumlicher Abschluss, diskrete Dimensionen der Inhalte, einfache Definierbarkeit statischer physischer und logischer Teilobjekte wie Seiten oder Absätze). Daraus ergeben sich besondere Anforderungen an die Identifikation, Navigation, Transformation und Verbindung mit anderen Objekten.

- Hoher oder sehr spezifischer Interpretationsbedarf der Daten. Hieraus folgen besondere Anforderungen an die Formulierung und Darstellung von Repräsentationsinformation.

- Komplexe Prozessierung digitaler Objekte – auch schon vor der Einlieferung in das Archiv. Daraus ergeben sich erhöhte Ansprüche an die Beschreibung von Provenance Information.

Andererseits besteht mit den bisherigen Standards ein ausbaufähiges Grundgerüst zur Verfügung. Um die Standardisierung bezüglich der Langzeitarchivierung genauer zu bewerten und zu vervollständigen, wären grundsätzliche Aufgaben eines Archivs in einer eScience Umgebung zu klären. Die Eigenschaften der Informationsobjekte und die damit verbundenen Erhaltungsmaßnahmen stehen in einer engen Wechselwirkung mit den dortigen Prozessen und Aufgabenstellungen. Dies betrifft insbesondere die Rollen für Verantwortlichkeiten und Durchführung bei der Erstellung von Submission Information Packages (SIP) und bei Ingest-Prozessen, also beim Generieren von Archival Information Packages (AIP). Diese Prozesse sind äußerst kritisch für die Qualität eines digitalen Langzeitarchivs. Ent-

scheidend ist hier, das richtige Maß an Repräsentationsinformation zu finden und diese selbst in eine für die Langzeitarchivierung geeignete Form zu bringen. Die oben aufgezählten Unterscheidungsmerkmale erfordern nach heutigem Stand der Technik eine enge Abstimmung zwischen Produzenten von Information und Archivorganisationen. Beziehungen und Interaktionen zwischen Produzenten und Archiven identifiziert, definiert und strukturiert als Ergänzung zum OAIS-RM der ISO-Standard Producer-archive interface – methodology abstract standard [ISO/IEC 20652:2006]. Dieser sehr generische Standard wäre für Anwendungsfälle im Bereich eScience auszugestalten.

Zu einem vertrauenswürdigen Archiv gehört auch, dass die Informationen dem Nutzer bereitgestellt werden können, also dass angemessene Dissemination Information Packages (DIP) lieferbar sind. Damit sind etliche Fragen verbunden, die Auswirkungen auf die zu unterstützenden oder zu entwickelnden Standards haben. So ist z.B. zu klären, ob ein Objekt physisch als Ganzes oder in vordefinierten oder vom Nutzer definierbaren Teilen ausgeliefert werden soll. Standardisierte Protokolle, die entsprechende Bandbreiten erlauben, wie z.B. gsiFTP, oder Abfragesprachen, entweder generischer Art wie XQuery oder SQL bzw. spezialisierter Art wie ADQL (Astronomical Data Query Language), bieten sich als Möglichkeit an. Des Weiteren stellt sich die Frage, ob *ausführbare* Repräsentationsinformation (z.B. spezialisierte Visualisierungssoftware) im Archiv oder außerhalb des Archivs bereitgestellt (und gepflegt) bzw. verarbeitet werden soll, um vollständige Informationsobjekte zu erzeugen. Bei anspruchsvollen Anwendungen wie z.B. bei der interaktiven Visualisierung sind auch komplexere technische Formen der Arbeitsteilung möglich. So können Befehle für mit OpenGL generierte drei-dimensionale Graphiken serverseitig im Archiv abgearbeitet werden, um die Übertragung zu einem entfernten Rechner des Nutzers auf zwei-dimensionale Grafiken zu beschränken, falls die dortige Grafikhardware und die Bandbreiten nicht ausreichend sind.

Durch die Ausrichtung der Grid-Architekturen am Paradigma der Serviceorientierung und durch die Unterstützung der Grid-Middleware gängiger Basisstandards ist die technische Unterstützung als deutlich einfacher einzustufen als eine „semantische“ (also jenseits der so genannten Bitstream-Erhaltung). Standards wie WSDL, SOAP, JDBC werden von zahlreichen konventionellen Produkten unterstützt, wobei sich aber die Fähigkeit, auch viele bzw. große Datenobjekte effizient (oder überhaupt) zu verarbeiten je nach Produkt deutlich unterscheiden kann.

Die geschilderten Entscheidungssituationen unterstreichen die Notwendigkeit für Archive, die Bedürfnisse und Fähigkeiten ihrer Kundschaft (Designated Community) sowie deren Standards bzw. gängige Techniken zu kennen und zu beobachten.

8.3 Nutzung von Grid-Technologien für die Langzeitarchivierung

Die nestor-Studie der Fernuniversität Hagen zeigt Nutzungsmöglichkeiten von Grid für die Langzeitarchivierung („LZA nutzt Grid“) [Schi2008]. Dabei kann zwischen der Erledigung rechenintensiver Aufgaben wie Formatvalidierung oder Formatmigration einerseits und der Unterstützung durch Speicherressourcen im Grid andererseits unterschieden werden, sofern nicht beide Arten von Leistungen benötigt werden. So kann z.B. auch die Validierung „kleiner“ Objekte äußerst rechenintensiv ausfallen. Die Nutzung von Grid-Technologien bietet natürlich auch Potenzial für LZA-Organisationen, deren Konsumenten bzw. Archivobjekte nicht aus dem eScience-Bereich stammen.

Nutzung von Grid-Speicherressourcen und Grid-Datenintegrationsfunktionen

Während bei der Standardisierung im Bereich des Datenmanagements eine gewisse Reife erkennbar ist, bringt das Fehlen einer zentralen Verwaltung der Infrastruktur und seiner Nutzung noch Probleme mit sich. Fragen zur Sicherheit, Abrechnung und der langfristigen Stabilität sind noch offen, was sich zwangsläufig im Fortschritt der Standardisierung niederschlägt. Aus Sicht der Vertrauenswürdigkeit spielt Sicherheit die größte Rolle, vorausgesetzt, dass die Instabilitäten nicht zu Verlusten an Integrität oder zu einem unzumutbaren Antwortverhalten führen. Die Wahrung des Urheberrechts und des Datenschutzrechts (oder ähnlicher Gesetze und Vorschriften aus dem Bereich der Archivierung) lässt sich bisher in einer Grid-Umgebung nicht sicherstellen. Komplizierter wird die Situation dadurch, dass nicht nur Inhaltsdaten sondern auch Softwareprodukte ggf. mit sehr unterschiedlichen Lizenzmodellen durch das Netz wandern und zwischengespeichert werden. Bei einem offenen (broad) Grid ist außerdem zu berücksichtigen, dass unterschiedliche rechtliche Geltungsbereiche durchschritten werden (vgl. hierzu auch die Ausführungen in [Baun2007]).

Um Grid-Technologien für die Speicherung dennoch vorzeitig zu nutzen, wäre eine Reduzierung der Universalität des Grid-Ansatzes im Sinne eines *narrow grid*. Der Nutzerkreis könnte bei diesem Ansatz auf ein Rechen- oder Datenzentrum beschränkt werden, oder etwas weiter gefasst, auf mehrere Institutionen, die untereinander in einer (vertraglich) geregelten Beziehung stehen aber durchaus geografisch weit verteilt sein könnten. Die mangelnde Unterstützung durch Standards und entsprechende Implementierungen könnte in dieser Situation durch individuelle Absprachen einschließlich Verzicht auf bestimmte Funktionalitäten ausgeglichen werden. Bestehende Speicherinfrastrukturen könnten mit Grid-Technolgien weiter (mit-)genutzt werden. Die Middlewareprodukte unterstützen eine Vielzahl von Dateisystemen und teilweise außer Dateien auch Datenelemente wie Binary Large Objects (BLOBs wie im SQL-Standard definiert) oder Realtime-Datenströme. Des Weiteren unterstützen die meisten Middlewareprodukte gängige hierarchische Speichermanager (HSM) oder zumindest Bandbibliotheken (Tape Library), indem sie entsprechende Schnittstellen anbieten. Hierbei ist jedoch zu erkennen, dass die Arbeitsteilung zwischen Grid-Midlleware und der Verwaltung der physischen Speicher sehr unterschiedlich gestaltbar ist. Unter Gesichtspunkten der langfristigen Datenhaltung ist darauf zu achten, dass Informationen zum Zustand eines Mediums (wie *Health and Fault*) geeigneten Beobachtungsprozessen (Monitoring) zur Verfügung gestellt werden.

Insgesamt ist festzustellen, dass sich die Produkte zur Speicherung konzeptionell als auch in ihrer Umsetzung (noch) deutlich unterscheiden. Die angebotenen Schnittstellen sowohl in Richtung Anwendung (z.B. die Strukturierung von Sammlungen, die Anreicherbarkeit mit Metadaten, die Definierbarkeit komplexer Transaktionen) als auch in Richtung physische Speicherung (z.B. unterstützte Transportprotokolle, Speichermedien, Speicherverwaltungssysteme) sind auch bei konzeptioneller Ähnlichkeit sehr unterschiedlich. Die Einbindung in eine standardisierte Grid-Umgebung mit z.B. einheitlicher Verwaltung oder einheitlichen Workflows ist noch nicht sehr ausgeprägt. Erfreulicherweise gibt es Zusagen, Konformität mit OGSA zu erreichen, wie beispielsweise für den Storage Resource Broker (SRB). Eine vereinfachte Verwaltung der Speicherressourcen wäre von Vorteil, da z.B. die Rechteverwaltung, das Anlegen und Synchronisieren von Replikaten oder die Zuteilung von Ressourcen nach einem einheitlichen Konzept und auf Basis standardisierter Schnittstellen und Sprachen möglich sind.

Neben der Nutzung von Grid-Technologien zur Herstellung virtueller Dateisysteme, die weitgehend inhaltsneutral sind, lassen sich auch Heterogenitäten in den gegebenenfalls

verteilten Datenquellen verbergen. Diesbezügliche Standardisierungsarbeiten des OGF laufen unter dem Thema *Data Access and Integration Services*. Seit Mai 2008 steht das Produkt OGSA-DAI WS DAIX[57] als Implementierung der entsprechenden Spezifikation des OGF zur Verfügung. Inwieweit jedoch dieses Vorgehen im Sinne einer Langzeitarchivierung für die Repräsentation von Informationspaketen sinnvoll ist, muss im Einzelfall geklärt werden. Greifen z.B. nach Abschluss eines Projektes keine spezialisierten Anwendungen mehr auf Teilobjekte zu, bietet sich als Alternative eine explizite *Normalisierung* (Vereinheitlichung) an. Komplizierte Transformationen könnten somit zeitnah auf Informationserhaltung geprüft werden und eine dauerhafte Beobachtung der Middleware (Verhalten bei neuen Versionen oder nach einer Portierung) entfiele.

Da Middleware zur Datenverwaltung als eigenständige Produkte verfügbar sind, ist es nicht notwendig, eine komplette Grid-Infrastruktur zu installieren. Ein weiterer Vorteil ist, dass sich die Produkte auf gängige Standards stützen, die in vielen Umgebungen zur Langzeitarchivierung bereits vorhanden sind, wie Java-Laufzeitumgebungen oder relationale Datenbanken. Doch bei der Verwendbarkeit unterschiedlicher relationaler Datenbankprodukte machen die Anbieter bzw. Entwickler von Middleware deutliche Einschränkungen, was die Grenzen einer Standardisierung (hier SQL), zumindest bei komplexen Produkten wie Datenbankmanagementsysteme, vor Augen führt.

Nutzung von Rechenressourcen

Ähnlich wie bei der Nutzung von Speicherressourcen macht sich die noch nicht weit fortgeschrittene Standardisierung beim Gebrauch von Rechenkapazitäten im Grid bemerkbar. Die Formulierbarkeit und Durchführung von Workflows, z.B. um komplette Ingest-Prozesse mit Formattransformation und Qualitätssicherung als Transaktion abzubilden, ist stark von der jeweiligen Umgebung abhängig. Die vorhandenen Anwendungsprogramme sind weitgehend auf spezifische wissenschaftliche Bedürfnisse zugeschnitten und zumindest nicht typisch für Archivsysteme, insbesondere wenn „unterstellt“ wird, dass Archive keine primären Informationsobjekte erzeugen und somit auch keine Instrumente und Software zur Erfassung betreiben. Trotz dieser Ausrichtung auf häufig technisch-naturwissenschaftliche Fragestellungen, sind im Grid auch Anwendungen zu finden, die für Zwecke der Langzeitarchivierung nützlich sein könnten, insbesondere für den Aufbau deskriptiver Information. Rechenintensive Methoden der Mustererkennung wie die Clusteranalyse können z.B. helfen, die Inhalte herkömmlicher Archivgüter wie digitalisierte Handschriften oder Landkarten zu erschließen. Ob solche Methoden nur intern genutzt oder auch den Endnutzern direkt zur Verfügung gestellt werden, ist eine Frage des „Geschäftsmodells“ einer Organisation, welches auch den Umfang relevanter Standards bestimmt (vgl. Rolle von Gedächtnisorganisationen in eScience-Umgebungen).

Wie bei der Speicherung von Daten im Grid stellen sich auch hier Fragen der Sicherheit, Zuverlässigkeit und Abrechnung. Um Rechenkapazitäten im Sinne eines Grids vorzeitig zu nutzen, bietet sich die Beschränkung auf eine (größere) Organisation oder einen überschaubaren Verbund an. Zur Nutzung müssten dann neben der Bereitstellung der physischen Ressourcen weitere Schritte durchgeführt werden. So wären die entsprechenden Programme zu entwickeln und in Diensteverzeichnissen wie UDDI bekannt zu machen. Bestimmte Services könnten auch den Produzenten von Information angeboten werden, um beispielsweise geeignete SIPs zu erstellen oder deren Validität zu prüfen. Des Weiteren wäre eine Einbettung der Grid-Dienste in die Workflows von Gedächtnisorganisationen

[57] http://www.ogsadai.org

erforderlich. Falls AIPs im Grid manipuliert werden, wären zur Vervollständigung von AIPs Prozeduren zu entwickeln, die relevante Verarbeitungsschritte im Grid als Provenance Information dokumentieren. D.h. diese Informationen müssten auf die Modelle bestehender Metadatenstandards abgebildet oder Schemata müssten entsprechend ergänzt werden. Statusinformationen, die Auskunft über den Erfolgsgrad von Operationen geben, liefern einen zusätzlichen Beitrag zur Vertrauenswürdigkeit.

Ggf. sind die nötigen Speicherressourcen eines Grid einzubinden (siehe oben), wobei zu berücksichtigen ist, dass der Aufwand für die Integrierbarkeit (noch) von den jeweils gewählten Rechengrid-Produkten abhängt.

8.4 Handlungsempfehlungen zur Standardisierungsarbeit

Standardisierungsaktivitäten sollten als integraler Bestandteil einer wie auch immer strukturierten Gedächtnisorganisation betrachtet werden. Schließlich umfasst im OAIS-RM die Funktion *Preservation Planning* auch die Entwicklung von Standards.

Methodik

Zur Beherrschung der Komplexität sollten verstärkt Architekturmodelle entwickelt werden, welche ein System auf einer jeweils geeigneten Abstraktionsebene in überschaubare Komponenten zerlegen und jeweils einen geeigneten Bezugsrahmen (Referenz) bieten. Das OAIS-Referenzmodell, das Modell zur Serviceorientierung von OASIS (Organization for the Advancement of Structured Information Standards) oder OGSA (Open Grid Services Architecture) liefern konkrete Ausgangspunkte für eine weitere Architekturbildung und somit für eine Strukturierung künftiger Standardisierungsaktivitäten bzw. für eine Zuordnung und Bewertung von Standards für die Langzeitarchivierung. Das OAIS-RM Modell beinhaltet bereits explizit Hinweise für eine weitere Standardisierung. Eine strukturierte und einheitlich notierte Darstellung von Standards, insbesondere der vielfältigen Abhängigkeiten, würde die *Erhaltungsplanung* erleichtern. Das Vorgehen der OGF bei der Verfolgung eigener und externer Standards, wie in der neuesten Roadmap für OGSA-Standards dargestellt, kann als Anregung dienen [OGF-OGSA 2008].

Ein modellhafter Ansatz unterstützt die Entwicklung adäquater und modularer Standards und somit auch eine einfachere Bewertung, Anwendung und Wartung. Positive Ansätze in diese Richtung sind bereits zu erkennen, z.B. bei der Definition neuer Formate wie MPEG-21, wo abstrakte Modelle gebildet werden, die die Aspekte der Repräsentation von Information von der Repräsentation von Daten trennen. Solche Abstraktionen würden auch helfen, das Thema Informationserhalt, z.B. bei unvermeidbaren Transformationen (Migrationen), besser in den Griff zu bekommen. Anhand von Modellen strukturierte Standards tragen dazu bei, ein *technisches Monitoring* im Sinne von OAIS zu erleichtern.

Konkrete Gegenstände einer gemeinsamen Standardisierung

Mittelknappheit bei den Gedächtnisorganisationen wird eine Etablierung optimaler Organisationsformen auch künftig verhindern, aber eine Verstetigung der Kooperation von Gedächtnisorganisationen unterschiedlichster Domänen könnte Synergiepotenziale auch für die Standardisierung nutzbar machen, indem gemeinsame Problemstellungen erkannt und bearbeitet werden. Die Aktivitäten der nestor-AG „Grid/eScience und LZA“ zeigen, dass trotz spezifischer Belange Gemeinsamkeiten sowohl auf konzeptioneller Ebene als auch

bei sehr konkreten Fraugestellungen existieren. Als Beispiele für mögliche kooperative Standardisierungsaktivitäten seien hier genannt:

- Erarbeitung eines Konzepts zur „Identität“ und Methoden für eine persistente Identifikation und Lokalisierbarkeit von Information. Ein solches Konzept dient auch als Basis für den Aufbau virtueller Sammlungen. Ontologiebasierte Objektbeziehungen könnten sich auf definierte und identifizierbare Ressourcen beziehen.

- Definition generischer Formate für (virtuelle) Container zur nachhaltigen Sicherung der Integrität und vereinfachten Verwaltung von Informationspaketen.

- Entwicklung von konzeptuellen Modellen für komplexe Objekte einschließlich Standardoperationen unter Berücksichtigung eines Versions- und Variantenkonzepts sowie einer Referenzierbarkeit von Teilobjekten.

- Aufbereitung von Repräsentationsinformation und ihre zuverlässige Bereithaltung in Repositorien. Aufwändige Emulationen, als eine Ausprägung von Repräsentationsinformation, könnten als Teil eines verteilten Verzeichnisses in einer Grid-Umgebung laufen.

Eine einheitliche und systematische Definition und Darstellung von Kernkonzepten, die über den Detaillierungsgrad des Informationsmodells im OAIS-RM hinausgehen, wäre eine wichtige Grundlage. Der uneinheitliche Gebrauch von Begriffen und die uneinheitliche Notation in den verschiedenen Standards stören das Verstehen erheblich. Hier böte sich ein Feld, formale Wissensrepräsentationssprachen (des Webs oder andere) anzuwenden und zu erproben. Außerdem stehen bereits zahlreiche standardisierte (auch weniger formale) Beschreibungsmittel zur Verfügung. In GLUE [OGSA-GLUE 2008] wird demonstriert, wie man mittels der standardisierten Modellierungssprache UML (Unified Modeling Language)[58] Systemteile und ihre Zusammenhänge mit hoher Präzision, aber von konkreten Datenmodellen abstrahierend, kompakt beschreiben kann.

Zusammenarbeit mit Standardisierungsorganisationen und SDOs

Vorteilhaft wäre auch eine intensivere und breitere Institutionalisierung der Verbindung von Gedächtnisorganisationen jeder Art mit Standardisierungsorganisationen bzw. SDOs, da viele Standards auf die Bewältigung der Langzeitarchivierung Einfluss haben, sowohl was die Hilfsmittel als auch was die zu bewahrende Information selbst betrifft. Die Langzeitarchivierung wird jedoch nur in den seltensten Fällen als Entwurfsziel von Standards genannt. Eine Verschlankung oder Ergänzung von Standards kann Erleichterung schaffen. Aktivitäten wie bei PDF/A [ISO 19005-1:2005] sind auch auf anderen Gebieten denkbar so z.B. bei Werkzeugen bzw. Regeln für Transformationen, die der dauerhaften Generierbarkeit virtueller Objekte dienen, die für sich einen hohen Wert für den Informationserhalt haben und aus bestimmten Gründen nicht materialisiert, also nicht explizit gespeichert werden können. Die Anpassung von Standards kann sich zwar als sehr problematisch erweisen, aber entsprechende Modelle zum Informationserhalt bieten hierfür Unterstützung. In Abfragesprachen für Inhaltsdaten könnte beispielsweise der Verzicht auf Änderungsfunktionen und Optimierungsdirektiven akzeptabel sein.

[58] http://www.uml.org

Sammlung und Aufbereitung von Anwendungsfällen

Der Rückgriff auf eine ausreichende Menge von dokumentierten, modellierten und möglichst detaillierten Anwendungsfällen (Use Cases) oder Problemfällen wäre ein weiterer Beitrag zur Verbesserung der Standardisierung. Im Rahmen der Entwicklung von OGSA ist nun eine Reihe von Nutzungsszenarien dokumentiert [OGF-HPRG 2007]. Leider ist dort das Thema Archivierung nur am Rande erwähnt. Eine detailliertere und modellhafte Sammlung zu einer spezifischeren Aufgabe der Langzeitarchivierung wurde im Rahmen der Entwicklung des verteilten Verzeichnisses GDFR (Global Digital Format Registry) erstellt. Sie zeigt eine hohe methodische Qualität und könnte als Orientierungshilfe bei anderen Entwicklungen dienen, wobei jedoch zu berücksichtigen ist, auf welchem Abstraktionsniveau man sich befindet.

Begleitung durch Testbeds

Schließlich können bei der Erarbeitung (oder Bewertung) bestimmter Standards begleitende Versuchsumgebungen (Testbeds) eine entscheidende Rolle für den Erfolg spielen, insbesondere für Aspekte, die sich einer theoretisch herleitbaren Aussage entziehen z.B. zum Laufzeitverhalten von Implementierungen, zum Aufwand für eine Realisierung oder allgemein zur Akzeptanz bei Nutzern oder Entwicklern.

Danksagung

Wir bedanken uns für die Beauftragung der Expertise bei der nestor-Initiative sowie für die Betreuung bei der Bayerischen Staatsbibliothek München insbesondere bei Frau Dr. Schoger, Herrn Beinert und Herrn Dr. Wolf-Klostermann. Die zahlreichen Diskussionsbeiträge der nestor-AG „Grid/eScience und LZA“ waren äußerst hilfreich. Herrn Altenhöner und Herrn Steinke von der Deutschen Nationalbibliothek danken wir für die gründliche Durchsicht des Berichtsentwurfs und ihre Anregungen. Wir bedanken uns ebenfalls bei Herrn Lobontu vom Forschungszentrum Karlsruhe für die Unterstützung bei der Recherche zu Grid-Produkten und zu Standardisierungsaktivitäten in Deutschland.

9 Abkürzungen

ACID	Atomicity, Consistency, Isolation, Durability	IT-Grundbegriff
ADQL	Astronomical Data Query Language	Standard
AES	Advanced Encryption Standard	Standard / Sicherheit
AIC	Archival Information Collection	Standard / OAIS
AIP	Archival Information Package	Standard / OAIS
AIU	Archival Information Unit	Standard / OAIS
ANSI	American National Standards Institute	SO
API	Application Programming Interface	IT-Grundbegriff
ARK	Archival Resource Key	Standard
ASN.1	Abstract Syntax Notation Number One	Standard
BABS	Bibliothekarisches Archivierungs- und Bereitstellungssystem	SW-Produkt
BES	Basic Execution Services	Standard
BFD	Binary Format Description	Standard
BinX	Binary XML Description Language	Standard
BLOB	Binary Large Object	IT-Grundbegriff
BPEL	Business Process and Execution Language	Standard
C3	Collaborative Climate Community Data and Processing	Projekt
CA	Certification Authority	IT-Grundbegriff / Sicherheit
CALTECH	California Institute of Technology	SDO
CCSDS	Consultative Committee for Space Data Systems	SDO
CD	Committee Draft	ISO-Standardisierung
CDL	California Digital Library	SDO
CEN	European Committee for Standardization	SO
CIM	Common Information Model	Standard
CLIR	Council on Library and Information Resources	Fachorganisation
CNRI	Corporation for National Research Initiatives	SDO
CORBA	Common Object Request Broker Architecture	Standard
cpio	copy in, copy out	SW-Produkt
CQL	Contextual Query Language	Standard
CVS	Collaborative Visualization Service	Grid
DAI	Data Access and Integration	SW-Produkt / Grid
DAIR	Data Access and Integration – The Relational Realization	SW-Produkt / Grid
DAIX	Data Access and Integration – The XML Realization	SW-Produkt / Grid
DCMES	Dublin Core Metadata Element Set	Standard
DCMI	Dublin Core Metadata Initiative	SDO
DD	Data Dictionary	IT-Grundbegriff
DDI	Data Documentation Initiative	SDO
DDL	Data Definition Language	IT-Grundbegriff
DES	Data Encryption Standard	Standard / Sicherheit
DFDL	Data Format Description Language	Standard
DFG	Deutsche Forschungsgemeinschaft	Organisation
DGIWG	Digital Geospatial Information Working Group	SDO
DIAS	Digital Information Archiving System	SW-Produkt
DIDL	Digital Item Declaration Language	Standard

DIN	Deutsches Institut für Normung	SO
DIP	Dissemination Information Package	Standard / OAIS
DML	Data Manipulation Language	IT-Grundbegriff
DMTF	Distributed Management Task Force	SDO
DNB	Deutsche Nationalbibliothek	SDO
DQP	Distributed Query Processing	SW-Produkt / Grid
DRM	Digital Rights Management	IT-Grundbegriff
DRS	Data Replication Service	SW-Produkt / Grid
DTD	Document Type Definition	IT-Grundbegriff
EAD	Encoded Archival Description	Standard
Ebind	Electronic Binding DTD	Standard
ebXML	Electronic Business using eXtensible Markup Language	Standard
ECDL	European Conference on Digital Libraries	Konferenz
ESML	Earth Science Markup Language	Standard
ETSI	European Telecommunications Standards Institute	SDO
EXI	Efficient XML Interchange Format	Standard
FCDC	Federal Geographic Data Committee	SDO
FDL	Format Description Language	IT-Grundbegriff
Fedora	Flexible Extensible Digital Object and Repository Architecture	SW-Produkt
FTP	File Transfer Protocol	Standard
FZJ	Forschungszentrum Jülich	Organisation
GDFR	Global Digital Format Registry	Standard
GGF	Grid File System	Grid
GHR	Global Handle Registry	SW-Produkt /Grid
GPT	Grid Packaging Toolkit	SW-Produkt / Grid
GRIB	GRIddedBinary	Standard
GSI	Grid Service Infrastructure	Grid
GT	Global Toolkit	SW-Produkt / Grid
GWDG	Gesellschaft für wissenschaftliche Datenverarbeitung	Organisation
HDF-EOS	Hierarchical Data Format – Earth Observing System	Standard
HPSS	High Performance Storage System	SW-Produkt
HSM	Hierarchical Storage Management	IT-Grundbegriff
HTTP	Hyper Text Transfer Protocol	Standard
HTTPS	Hyper Text Transfer Protocol Secure	Standard
HUL	Harvard University Library	SDO
IDF	International DOI Foundation	SDO
IEEE	Ursprünglich für: Institute of Electrical and Electronics Engineers	SDO
IETF	Internet Engineering Task Force	SDO
incits	InterNational Committee for Information Technology Standards	SDO
IP	Information Package	Standard / OAIS
IPR	Intellectual Property Rights	Grundbegriff
IR	Institutional Repository	IT-Grundbegriff
IRI	Internationalized Resource Identifier	Standard
ISBN	Internationale Standardbuchnummer	Standard
ITSC	Information Technology and Systems Center – The University of Alabama	SDO
ITU	International Telecommunication Union	SDO

IVOA	International Virtual Observatory Alliance	SDO
JDBC	Java Database Connectivity	Standard
JSDL	Job Submission Description Language	Standard
JSR	Java Specification Request	Standard
JTC	Joint Technical Committee	ISO-Standardisierung
KDC	Key Distribution Center	IT-Grundbegriff / Sicherheit
koLibRI	kopal Library for Ingest and Retrieval	SW-Produkt
kopal	Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen	Projekt
LCAV	Library of Congress Audiovisual	Standard
LDAP	Lightweight Directory Access Protocol	Standard
LMER	Langzeitarchivierungsmetadaten für elektronische Ressourcen	Standard
LOC	Library of Congress	SDO
LOCKSS	Lots of Copies Keep Stuff Save	SW-Produkt
LOM	Learning Object Metadata	Standard
LZA	Langzeitarchivierung	Grundbegriff
MAC	Message Authentication Code	Standard / Sicherheit
MACE	Middleware Architecture Committee for Education	SDO
MARC	Maschine Readable Cataloging	Standard
MathML	Mathematical Markup Language	Standard
MCAT	Metadata Catalog	SW-Produkt / Grid
MD	Message Digest	Standard
MEP	Message Exchange Pattern	Standard
METS	Metadata Encoding and Transmission Standard	Standard
MIME	Multipurpose Internet Mail Extensions	Standard
MIX	Niso Metadata for Images in XML Schema	Standard
MODS	Meta Object Description Language	Standard
MPEG	Moving Picture Experts Group	SDO
NBN	National Bibliographic Number	Standard
NCBI	National Center for Biotechnology Information	SDO
NFS	Network File System	Standard
NISO	National Information Standards Organization	SDO
OAI	Open Archives Initiative	SDO
OAIS	Open Archival Information System	Standard
OASIS	Organization for the Advancement of Structured Information Standards	SDO
ODBC	Open Database Connectivity	Standard
OGC	Open Geospatial Consortium	SDO
OGF	Open Grid Forum	Grid / SDO
OGSA	Open Grid Service Architecture	Grid
OGSI	Open Grid Service Infrastructure	Grid
OpenGL	Open Graphics Library	Standard
OSE	Open System Environment	Standard
OSM	Open Storage Manager	SW-Produkt
OWL	Web Ontology Language	Standard
P2P	Peer-to-Peer	IT-Grundbegriff
PAS	Publicly Available Specification	Standardisierung
PAX	Portable Archive Exchange	Standard
PDI	Preservation Description Information	Standard / OAIS
PKI	Public-Key-Infrastruktur	Standard / Sicherheit

PMH	Preserevation Metadata Harvesting	Standard
PNG	Portable Network Graphics	Standard
POSIX	Portable Operating System Interface	Standard
PREMIS	Preservation Metadata Implementation Strategies	Standard
PUID	PRONOM Persistent Unique Identifier	Standard
RDBMS	Relationales Datenbankmanagementsystem	IT-Grundbegriff
RDF	Resource Description Framework	Standard
REST	Representational State Transfer	Standard
RFC	Request for Comments	IETF-Standardisierung
RLS	Replication Location Service	SW-Produkt / Grid
RM	Referenzmodell	Grundbegriff
ROP	Rule Oriented Programming	SW-Produkt / Grid
RPC	Remote Procedure Call	IT-Grundbegriff
RSA	Rivest, Shamir, Adleman	Standard
RUS	Resource Usage Service	Grid
SAGA	Simple API for Grid Applications	Grid
SAML	Security Assertion Markup Language	Standard
SAX	Simple API for XML	Standard
SC	Sub Committee	ISO-Standardisierung
SDO	Standards Developing Organization	Grundbegriff
SHA	Secure Hash Algorithm	Standard
SI	Système international d'unités	Standard
SIP	Submission Information Package	Standard / OAIS
SKOS	Simple Knowledge Organisation System	Standard
SMI	Storage Management Initiative	SDO
SMTP	Simple Mail Transfer Protocol	Standard
SNIA	Storage Networking Industry Association	SDO
SO	Standards Organization	Grundbegriff
SOAP	Ursprünglich für: Simple Object Access Protocol	Standard
SRB	Storage Resource Broker	SW-Produkt / Grid
SRU	Search/Retrieval via URL	Standard
SSL	Secure Sockets Layer	Standard
SVG	Scalable Vector Graphics	Standard
TAR	Tape Archive	SW-Produkt
TCP/IP	Transmission Control Protocol/Internet Protocol	Standard
TEI	Text Encoding Initiative	Standard
TIFF	Tagged Image File Format	Standard
TMS	Trust Management System	Grid
TSM	Tivoli Storage Manager	SW-Produkt
UDDI	Universal Description, Discovery and Integration	Standard
UML	Unified Modeling Language	Standard
UNICORE	Uniform Interface to Computing Resources	SW-Produkt / Grid
UOF	Universal Object Format	Standard
URI	Uniform Resource Identificator	Standard
URL	Uniform Resource Locator	Standard
URN	Uniform Resource Name	Standard
UTF	Unicode Transformation Format	Standard
VO	Virtuelle Organisation	Grundbegriff
VRA	Visual Resources Association	SDO
VRML	Virtual Reality Modeling Language	Standard
W3C	World Wide Web Consortium	SDO

WD	Working Draft	ISO Standardisierung
WDCC	World Data Center for Climate	SDO
WS	Web Services	Standard
WSDL	Web Services Description Language	Standard
WS-I	Web Services Interoperability Organization	SDO
XACML	eXtensible Access Control Markup Language	Standard
XFDU	XML Formatted Data Unit	Standard
XML	Extensible Markup Language	Standard
XMLDsig	XML Signature Syntax and Processing	Standard
XPath	XML Path Language	Standard
XQuery	XML Query Language	Standard
XSIL	Extensible Scientific Interchange Language	Standard
XSL	Extensible Stylesheet Language	Standard
XSLT	XSL Transformation	Standard
ZEND	Zentrale Nachweis- und Erfassungsdatenbank	SW-Produkt

10 Literatur

Baun2007	Christian Baun, Arial Garcia, Wolfgang Gentzsch, Kennzeichen D, in: iX, 12.2007
Beka2005	Jeroen Bekaert, Xiaoming, Herbert van de Sompel, Representing Digital Assets for Long-Term Preservation using MPEG-21 DID, PV 2005, 2005 http://arxiv.org/ftp/cs/papers/0509/0509084.pdf
Borg2006	Uwe M. Borghoff, Peter Rödig, Jan Scheffczyk, Lothar Schmitz, Long-Term Preservation of Digital Documents – Principles and Practices, Springer Verlag, 2006
Bras2007	Jan Brase, Jens Klump, Zitierfähige Datensätze – Primärdaten-Management durch DOIs, in: WissKom 2007: Wissenschafts-kommunikation der Zukunft, 11.2007 http://edoc.gfz-potsdam.de/gfz/display.epl?mode=doc&id=10493
CCSDS2004	XML Formatted Data Unit (XFDU) – Structure and Construction Rules, Draft Recommended Standard, White Book, 09.2004 http://sindbad.gsfc.nasa.gov/xfdu/pdfdocs/iprwbv2a.pdf
CCSDS2008	XML Formatted Data Unit (XFDU) – Structure and Construction Rules, Recommended Standard, Blue Book, 09.2008 http://public.ccsds.org/publications/archive/661x0b1.pdf
Chak2007	Anirban Chakrabarti, Grid Computing Security, Springer Verlag, 2007
DDB2005	Die Deutsche Bibliothek, LMER – Langzeitarchivierungsmetada-ten für elektronische Ressourcen, 07.04.2005 http://www.d-nb.de/standards/pdf/lmer12.pdf
DLM2008	DLM-Forum, MoReq2 Specification – Model Requirements Specification for the Management of Electronic Records, 13.02.2008 http://ec.europa.eu/transparency/archival_policy/moreq/doc/moreq2_spec.pdf
DMTF2008	DMTF, Web Services for Management (WS-Management) Speci-fication, 12.02.2008 http://www.dmtf.org/standards/published_documents/DSP0226_1.0.0.pdf
Fost2002	Ian Foster, What is the Grid? A Threepoint checklist, GRIDtoday, Vol. 1, No. 6, 22.07.2002 http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf
Gent2007	Wolfgang Gentzsch, Grid Infrastructures and Standards – Exam-ple: D-Grid, 4th e-Infrastructure Concertation Meeting, ETSI, 05.-06.12.2007 http://portal.etsi.org/Docbox/Workshop/2007/200712_ECCONCERTATION/GENTZSCH%20KEYNOTE%20e-Inf%20Concertation.pdf
Göhn2006	Mathias Göhner, Status Quo: Accounting im Bereich des Grid Computing, Universität der Bundeswehr München, Bericht Nr. 2006-3, 10.2006
GTK2005	Globus Toolkit Version 4 Grid Security Infrastructure: A Stan-dards Perspective, The Globus Security Team, Version 4, 12.09.2005 http://www.globus.org/toolkit/docs/4.0/security/GT4-GSI-

	Overview.pdf
HUL2006	HUL, Global Digital Format Registry (GDFR) – Identifiers, Version 1.0.1 (Draft), 18.09.2006 www.gdfr.info/docs/GDFR-Identifiers-1_0_1.pdf
HUL2007	HUL, Global Digital Format Registry (GDFR) – Format Model and Relationships, Version 1.0.7 (Draft), 11.05.2007 http://hul.harvard.edu/gdfr/documents/GDFR-Format-Model-and-Relationship-1_0_7.rtf
IBM2006a	H. Verhoeven, IBM, Digital Information Archiving System – SIP Interface Specification Vers. 2.5, 12.01.2006 http://kopal.langzeitarchivierung.de/downloads/kopal_DIAS_SIP_Interface_Specification.pdf
IBM2006b	H. Verhoeven, IBM, Digital Information Archiving System – DIP Interface Specification Vers. 2.6, 10.02.2006 http://kopal.langzeitarchivierung.de/downloads/kopal_DIAS_DIP_Interface_Specification.pdf
IEEE1995	IEEE Std 1003.0-1995 IEEE Guide to the POSIX Open System Environment (OSE)-Description Hinweis: zwischenzeitlich zurückgezogen
IEEE754	ANSI/IEEE Std 754-2008 – IEEE Standard for Binary Floating-Point Arithmetic http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4610935
ISO 19005-1:2005	ISO 19005-1:2005 Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1), 15.07.2008
ISO/CD 26324	ISO/CD 26324 Information and documentation – Digital object identifier system, 26.04.2008
ISO/IEC 19757-2:2003	ISO/IEC 19757-2:2003 Information technology – Document Schema Definition Language (DSDL) – Part 2: Regular-grammar based validation – RELAX NG, 28.11.2003
ISO/IEC 19757-3:2006	Information technology – Document Schema Definition Language (DSDL) – Part 3: Rule-based validation – Schematron, 24.05.2006
ISO/IEC 19776-3:2007	ISO/IEC 19776-3:2007 Information technology – Computer graphics, image processing and environmental data representation – Extensible 3D (X3D) encodings – Part 3: Compressed binary encoding, 20.09.2007
ISO/IEC 20652:2006	ISO/IEC 20652:2006 Space data and information transfer systems – Producer-archive interface – Methodology abstract standard, 30.01.2006
ISO/IEC 21000-2:2005	ISO/IEC 21000-2:2005 Information technology – Multimedia framework (MPEG-21) – Part 2: Digital Item Declaration, 06.10.2005
ISO/IEC 21000-6:2004	ISO/IEC 21000-6:2004 Information technology – Multimedia framework (MPEG-21) – Part 6: Rights Data Dictionary, 17.05.2004
ISO/IEC 24824-1:2007	ISO/IEC 24824-1:2007 Information technology – Generic Applications of ASN.1 – Fast Infoset, 04.05.2007
ISO/IEC 26300:2006	ISO/IEC 26300:2006 Information technology – Open Document Format for Office Applications (OpenDocument) v1.0, 30.11.2006
ISO/IEC 27001:2005	ISO/IEC 27001:2005 Information technology – Security techniques – Information security management systems – Requirements, 14.10.2005
ISO/IEC 27002:2005	ISO/IEC 27002:2005 Information technology – Security tech-

	niques – Code of practice for information security management, 22.04.2008
ISO/IEC 9075-11	ISO/IEC 9075-11 Information technology – Database languages – SQL Part 11: Information and Definition Schemas (SQL Schemata), 11.07.2008
ISO/IEC 11179	ISO/IEC 11179 Information technology – Metadata Registries (MDR) Standard mit 6 Teilen
Klum2007	Jens Klump, nestor – materialien 9 – Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten, GeoForschungsZentrum Potsdam (GFZ), 26.10.2007 http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_09.pdf
Kunz2008	J. Kunze, R. Rodgers, The ARK Identifier Scheme, Network Preservation Group, California Digital Library, US National Library of Medicine, 22.05.2008 http://www.cdlib.org/inside/diglib/ark/arkspec.html
Moor2003a	Reagan W. Moore, Preservation of Data, SDSC Technical Report, San Diego, 06.2003 http://www.npaci.edu/DICE/Pubs/data-preservation.doc
Moor2003b	Reagan W. Moore, Andre Merzky, Global Grid Forum, Persistent Archive Concepts, 25.12.2003 http://www.npaci.edu/DICE/Pubs/Data-PAWG-PA.doc
nestor2008	nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung, nestor – materialien 8 – Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, Version II, 09.2008 http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_08.pdf
NLNZ2003	National Library of New Zealand, Metadata Standards Framework – Preservation Metadata (Revised), 06.2003 http://www.natlib.govt.nz/downloads/metaschema-revised.pdf
OASIS-SAML2005	Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0, 15.03.2005
OASIS-SAMLprofile2005	OASIS, SAML 2.0 profile of XACML v2.0, 01.02.2005
OASIS-XACML2005	OASIS eXtensible Access Control Markup Language (XACML) Version 2.0, 01.02.2005
OGC2008	Open Geospatial Consortium, Geospatial eXtensible Access Control Markup Language (GeoXACML), 20.02.2008
OGF-OGSA 2008	OGF, Defining the Grid: A Roadmap for OGSA Standards, Version 1.1, 12.02.2008
OGF2007a	OGF, Technical Strategy for the Open Grid Forum 2007-2010, 02.01.2007
OGF-DFDL 2008	OGF FDL-WG, Data Format Description Language (DFDL) v1.0 – Core Specification (Internal Committee Working Document), 07.02.2008
OGF-HPRG 2007	OGF, Grid Network Services Use Cases from the e-Science Community, 12.12.2007
OGF-LTDAR 2005	OGF, Long-Term Digital Archive Requirements, 03.10.2005
OGSA-DAI 2006	OGSA DAIS WG, Web Services Data Access and Integration – The Core (WS-DAI) Specification, Version 1.0, 20.07.2006
OGSA-DAI 2007	Kostas Karasavas, OGSA-DAI – Redesigned and New Activities, 16.05.2007
OGSA-DAIR 2006	OGSA DAIS WG, Web Services Data Access and Integration –

	The Relational Realization (WS-DAIR) Specification, Version 1.0, 05.09.2006
OGSA-DAIX 2006	OGSA DAIS WG, Web Services Data Access and Integration – The XML Realization (WS-DAIX) Specification, Version 1.0, 05.09.2006
OGSA-GLUE 2008	OGSA, GLUE Specification v. 2.0, 20.05.2008
OGSA-SCRN 2005	SDOs Collaboration on networked Resourced Management (SCRM) – WG, Jay Unger, Mark Carlson, Hiro Kishimoto, 27.06.2005
Prod2007	Radu Prodan, Thomas Fahringer, Grid Computing – Experiment, Tool Integration, and Scientific Workflows, LNCS 4340, Springer, 2007
Rama2003	Rahul Ramachandran, Andrew McDowell, Xiang Li, Sunil Movva, Matt He, Earth Science Markup Language – Schema Documentation for v3.0, 03.11.2003 http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01430074
RFC2104	IETF, Request for Comments 2104, HMAC: Keyed-Hashing for Message Authentication, 02.1997
RFC2141	IETF, Request for Comments 2141, URN Syntax, 05.1997
RFC3010	IETF, Request for Comments 3010, Network File System (NFS) version 4 Protocol, 04.2003
RFC3188	IETF, Request for Comments 3188, Using Bibliographic National Numbers as Uniform Resource Names, 10.2001
RFC3650	IETF, Request for Comments 3650, Handle System Overview, 11.2003
RFC3986	IETF, Request for Comments 3986, Uniform Resource Identifier (URI): Generic Syntax, 01.2005
RFC4120	IETF, Request for Comments 4120, The Kerberos Authentication Service (V5), 07.2005
RFC4452	IETF, Request for Comments 4452, The “info” URI Scheme for Information Assets with Identifiers in Public Namespaces, 04.2006
RFC4998	IETF, Request for Comments 4998, Evidence Record Syntax, 08.2007
Ried2005	Morris Riedel, Daniel Mallmann, Standardization Processes of the UNICORE Grid System, Proceedings of 1st Austrian Grid Symposium 2005, Forschungszentrum Jülich, 2006 http://www.unicore.eu/documentation/files/riedel-2006-SPU.pdf
Ried2008	Morris Riedel, SAML – Create and Exchange Security Information in Grids, FZJ, JSC, 3. D-Grid Security Workshop, 01.-02.04.2008 www.medigrid.de/u_veranst/080401_security_ws/SAML_Riedel_080402.pdf
Rieg2008	Oya Y. Rieger, Preservation in the Age of Large-Scale Digitization – A White Paper, CLIR Report, 02.2008 http://www.clir.org/pubs/reports/pub141/pub141.pdf
Sant2008	Mathias Santos de Brito, Liria Matsumoto Sato, Extending OGSA-DAI possibilities with JDBC Driver, 11th IEEE International Conference on Computational Science and Engineering, 07.2008 http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04578228
Schi2008	Udo Hönig, Wolfram Schiffmann, Expertise Synergiepotentiale zwischen Grid- und E-Science-Technologien für die digitale

	Langzeitarchivierung, 16.04.2008
Scho2008	Tobias Scholl, Benjamin Gufler, Jessica Müller, Angelika Reiser, Alfons Kemper, P2P-Datenmanagement für e-Science-Grids, in: Datenbankspektrum 09.2008
Schu2008	Bernd Schuller, UNICORE Version 6 – Architecture, UNICORE Migration Workshop, 29.10.2008 http://www.unicore.eu/documentation/tutorials/unicore6/files/01_Schuller.pdf
Seve2006	Thomas Severins, Eberhard R. Hilf, nestor – materialien 6 – Langzeitarchivierung von Rohdaten, Carl von Ossiezky Universität Oldenburg, 2006 http://edoc.hu-berlin.de/series/nestor-materialien/6/PDF/6.pdf
Stei2006a	Tobias Steinke, The Universal Object Format – An Archiving and Exchange Format for Digital Objects, ECDL 2006, LNCS 4172, 2006 http://www.springerlink.com/content/u7l3326h17264051/fulltext.pdf
Stei2006b	Tobias Steinke, Universelles Objektformat – Ein Archiv- und Austauschformat für digitale Objekte, Frankfurt am Main, 2006 http://kopal.langzeitarchivierung.de/downloads/kopal_Universelles_Objektformat.pdf
TNA2005	The National Archives, PRONOM 4 Information Model, 04.01.2005 http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_info_model.pdf
TNA2006	The National Archives, Digital Preservation Technical Paper 2 – The PRONOM PUID Scheme: A scheme of persistent unique identifiers for representation information, 27.07.2006 http://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf
W3C EXI 2008	W3C, Efficient XML Interchange (EXI) Format 1.0, W3C Working Draft 26.04.2008

ISO	http://www.iso.org
OASIS	http://www.oasis-open.org
OGF / OGSA	http://www.ogf.org
RFC	http://www.rfc-editor.org
W3C	http://www.w3c.org

11 Anlagen

11.1 SDOs auf dem Gebiet LZA und ihre Standards

SDO	Standard	SO / Gültigkeitsbereich	Zuordnung
CCSDS	OAIS	ISO	Referenzmodell (i.W. technisch-konzeptionell)
CCSDS	XFDU	ISO	IP-Struktur
CDL	ARK	IETF	Reference Information
CNRI	Handle System	de facto, RFC Handle System Overview RFC 3650, Handle System Namespace and Service Definition RFC 3651, Handle System Protocol (vers 2.1) Specification RFC 3652	Reference Information
DCMI	DC	de facto / Ausschnitt als: ISO 15836:2003 Information and documentation – The Dublin Core metadata element set, 01.07.2008 (in der Revision) ANSI/NISO Standard Z39.85-2007, 05.2007 RFC 5013, 08.2007 (informational)	Descriptive Information, auch PDI z.B. Context Information
DNB	UOF	Community	IP-Struktur, AIC, AIU
DNB	NBN	Community / RFC 3188 (informational)	Reference Information
DNB	LMER	Community	Representation Information, PDI
HUL, OCLC	GDFR	Community	Representation Information
IDF	DOI System	ISO/CD 26324: Information and documentation – Digital object identifier system nutzt *Handle* und Teile von MPEG-21	Reference Information, zusätzlich auch PDI
LOC	METS	Community	IP-Struktur, AIC, AIU
LOC	MARC / MARCXML	Vgl. ISO 2709:2008 Information and documentation – Format for information exchange, 30.06.2008 ISO 25577:2008 Information and documentation – MarcXchange, 25.11.2008	Descriptive Information, auch PDI z.B. Context Information
LOC	MODS	Community	Descriptive Information, auch PDI z.B. Context Information
LOC	SRU, CQL	Community	Finding Aids, Access
OCLC	PREMIS	Community	Representation Information, PDI
TNA	PRONOM	Community	Representation Information

11.2 SDOs auf dem Gebiet Grid und IT und ihre Standards

SDO	Standard	Zuordnung
CCSDS	XML Formatted Data Unit (XFDU)	IP-Struktur, AIC, AIU
DMTF	Common Information Model (CIM)	Common Services
DMTF	Web Based Enterprise Management (WBEM)	Common Services
DMTF	Web Services for Management (WS-Management)	Common Services
IEEE	Portable Operating System Interface (POSIX)	Common Services
IEEE	Binary Floating-Point Arithmetic	Representation Information
IETF	IRI, URI, URN, URL	Reference Information
IETF	Hyper Text Transfer Protocol (HTTP)	Common Services
IETF	Hyper Text Transfer Protocol Secure (HTTPS)	Common Services
IETF	8-bit Unicode Transformation Format (UTF-8)	Common Services
IETF	Network File System (NFS)	Common Services
IETF	Sicherheitsbasisstandards	Common Services
ISO JTC 1 / SC 29 MPEG	MPEG-21	IP-Struktur, auch Representation Information und PDI
ISO JTC 1 / SC 34 OASIS	RELAX NG	Common Services
ISO JTC 1 / SC 34	Schematron	Common Services
ISO JTC 1 / SC 32	SQL Data Definition Language	Common Services
ISO JTC 1 / SC 32	SQL Schemata	Common Services
ITU	Abstract Syntax Notation One (ASN.1)	Representation Information
ITU	X.509 Public-key and attribute certificate frameworks	Common Services
Microsoft	Open Database Connectivity (ODBC)	Common Services
OASIS	eXtensible Access Control Markup Language (XACML)	Common Services
OASIS	Security Assertion Markup Language (SAML)	Common Services
OASIS	Web Services Resource Framework (WSRF)	Common Services
OASIS	Universal Description, Discovery and Integration (UDDI)	Access, Finding Aids
OGF	OGSA-xx, u.a.: Basic Execution Services (BES), ByteIO, Data Format Description Language (DFDL), Job Submission Control Language (JSCL), Web Services Data Access and Integration – The Relational Realization (WS-DAIR) Specification	i.W. Common Services (auch technisch-konzeptionelle Referenzmodelle)
PKWARE	ZIP	IP-Struktur
Silicon Graphics	Open Graphics Library (OpenGL)	Common Services

SNIA	Storage Management Technical Specification,	Archival Storage
SNIA	ANSI INCITS 388-2004	Archival Storage
Sun Micro-systems	Java Specification Request (JSR) 168 Portlet Specification	Common Services
Sun Micro-systems	Java Database Connectivity (JDBC)	Common Services
W3C	XML Signature Syntax and Processing	Common Services
W3C	Web Services Description Language (WSDL)	Common Services
W3C	Web Services Addressing (WS Addressing)	Reference Information
W3C	XML Schema	Common Services
W3C	SOAP	Common Services
W3C	XML Path Language (XPath)	Common Services
W3C	XML Query Language (XQuery)	Common Services
W3C	XSL Transformation (XSLT)	Common Services
W3C	Efficient XML Interchange Format (EXI)	Common Services
W3C	XML Signature Syntax and Processing	Common Services

Anmerkung: Sprachen werden vereinfacht den Common Services zugeordnet.

11.3 SDOs mit Domänenbezug (eScience-Bereich, wissenschaftlich) und ihre Standards

SDO	Standard	Zuordnung
CALTECH	Extensible Scientific Interchange Language (XSIL)	Wissenschaftliche Daten / OAIS-IP
ITSC	Earth Science Markup Language (ESML)	Earth Science / OAIS-IP
IVOA	Astronomical Data Query Language (ADQL)	Astronomie / OAIS (Implementierung): Access
TC 211 Geographic information / Geomatics	ISO 19115 Geographic information – Metadata	Geoinformation / OAIS: Representation Information, PDI
TC 211 Geographic information / Geomatics	ISO 19139 Geographic information – Metadata – XML Schema implementation.	Geoinformation / OAIS (Implementierung): Representation Information, PDI
U.S. Dept of Energy – Collaboratory	Binary Format Description (BFD) language	Erweiterung von XSIL

11.4 Nutzung von Standards durch Entwickler / Projekte / Anbieter

Die Tabelle gibt einen groben Überblick über die konkrete Verwendung von Standards durch Entwickler, Projekte bzw. Anbieter. Natürlich bedeutet die Anwendung von Standards noch keine (technische) Interoperabilität. Unvollständige, fehlerhafte oder bezüglich Speicher bzw. Laufzeit ineffiziente Implementierungen sowie unterschiedliche Versionen bzw. Varianten (Profile, Level u.ä.) können in der Praxis schon auf dieser eher technischen Ebene zu erheblichen Einschränkungen führen. Im Fall kompatibler Basistechnologien können unterschiedliche konzeptionelle Modelle die „direkte" Interoperablität verhindern[59]. Die Empfehlungen in dieser Expertise geben Hinweise zu Ansatzpunkten für eine Standardisierung auf konzeptioneller Ebene bezüglich Langzeitarchivierung (vgl.8.4).

Wegen der Bedeutung von relationalen Datenbankmanagementsystemen (RDBMS) für die dauerhafte Vollständigkeit von *Informationspaketen* sowohl in „traditionellen" Architekturen der LZA als auch in neueren Architekturen in eScience-Umgebungen (vgl. hierzu z.B. [Scho2008] zur Rolle von SQL in P2P-Systemen) sind exemplarisch zwei High-End-Produkte aufgenommen. Eine Diskussion zur Bedeutung einer Einbindung von RDBMS in die Grid-Middleware OGSA-DAI via JDBC findet sich in [Sant2008]. Oracle und IBM sind Mitglieder im OGF, IBM im UNICORE Forum und im Globus Consortium[60].

Entwickler / Projekte	Produkt	Standard
Globus Alliance	GT	CNRI Handle
Globus Alliance	GT	SOAP
Globus Alliance	GT	WSDL
Globus Alliance	GT	WSRF
Globus Alliance	GT	JDBC
Globus Alliance	GT	X.509
Globus Alliance	GT	SAML
Gridsphere Project	GridSphere Portal Framework	JSR 168
IBM	DB2 + WebSphere	SQL mit XML Unterstützung[61]
IBM	DB2 + WebSphere	X.509
IBM	DB2 + WebSphere	SAML
IBM	DB2 + WebSphere	JDBC
IBM	DB2 + WebSphere	UDDI
IBM	DB2 + WebSphere	LDAP
IBM	DB2 + WebSphere	SOAP
IBM	DB2 + WebSphere	WSDL
Internet2 Initiative[62]	Shibboleth	SAML
Oracle	RDBMS + AS	SQL mit XML Unterstützung
Oracle	RDBMS + AS	X.509

[59] Vgl. hierzu u.a. die Ausführungen unter http://www.unicore.eu/documentation zur Interoperabilität von UNICORE 6 und Globus Toolkit.
[60] http://www.globusconsortium.org
[61] Vgl. hierzu auch ISO/IEC 9075-14:2003 Information technology – SQL – Part 14: XML-Related Specifications (SQL/XML).
[62] Vgl. auch Vascoda, dortige Definition für Handle: Fingierter Name, der anstelle eines realen Personennamen angegeben werden kann, um aus Datenschutzgründen eine gewisse Anonymität zu erreichen. In Shibboleth kann ein persistentes Pseudonym verwendet werden, das nur im Missbrauchsfall aufgelöst wird.

Oracle	RDBMS + AS	SAML
Oracle	RDBMS + AS	UDDI
Oracle	RDBMS + AS	LDAP
Oracle	RDBMS + AS	SOAP
Oracle	RDBMS + AS	WSDL
Oracle	RDBMS + AS	JDBC
UNICORE	UNICORE	X.509
UNICORE	UNICORE	XACML
UNICORE	UNICORE	SAML
UNICORE	UNICORE	OGSA JSDL
UNICORE	UNICORE	OGSA BES
UNICORE	UNICORE	OGSA ByteIO
UNICORE	UNICORE	SOAP
UNICORE	UNICORE	JAR / Signed Java Objects

AS Application Server (Teilweise Middleware-Produkt *Fusion* erforderlich)
GT: Globus Toolkit

Ein Ausblick auf die Umsetzung weiterer Standards in UNICORE findet sich in [Schu2008]. Mit Hilfe von Plug-In-Punkten können weitere Standards zur Anwendung kommen wie z.B. GridFTP.