H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 9

Access

nestor

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

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

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

Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.

Die digitale Version 2.3 steht unter folgender Creative-Commons-Lizenz:
„Namensnennung-Keine kommerzielle Nutzung-Weitergabe unter gleichen Bedingungen 3.0 Deutschland“
http://creativecommons.org/licenses/by-nc-sa/3.0/de/

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

URL für Kapitel 9 „Access“ (Version 2.3): urn:nbn:de:0008-2010062482
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010062482

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

9 Access

9.1 Einführung

Karsten Huth

Der Titel dieses Kapitels ist ein Begriff aus dem grundlegenden ISO Standard OAIS. Access steht dort für ein abstraktes Funktionsmodul (bestehend aus einer Menge von Einzelfunktionalitäten), welches im Wesentlichen den Zugriff auf die im Archiv vorgehaltenen Informationen regelt. Das Modul Access ist die Schnittstelle zwischen den OAIS-Modulen „Data Management", „Administration" und „Archival Storage".[1] Zudem ist das Access-Modul die Visitenkarte eines OAIS für die Außenwelt. Nutzer eines Langzeitarchivs treten ausschließlich über dieses Modul mit dem Archiv in Kontakt und erhalten gegebenenfalls Zugriff auf die Archivinformationen. In der digital vernetzten Welt kann man davon ausgehen, dass der Nutzer von zu Hause aus über ein Netzwerk in den

1 Consultative Committee for Space Data Systems (Hrsg.) (2002): *Reference Model for an Open Archive Information System: Blue Book*. Washington, DC. Page 4-14ff
Alle hier aufgeführten URLs wurden im Mai 2010 auf Erreichbarkeit geprüft.

Beständen eines Archivs recherchiert. Entsprechende technische Funktionalitäten wie Datenbankanfragen an Online-Kataloge oder elektronische Findmittel werden bei vielen Langzeitarchiven zum Service gehören. Die Möglichkeit von Fernanfragen an Datenbanken ist jedoch keine besondere Eigenart eines Langzeitarchivs. Wesentlich sind folgende Fragen:

- Wie können die Informationsobjekte (z. T. auch als konzeptuelle Objekte bezeichnet) dauerhaft korrekt adressiert und nachgewiesen werden, wenn die logischen Objekte (z.B. Dateien, Datenobjekte) im Zuge von Migrationen technisch verändert werden und im Archiv in verschiedenen technischen Repräsentationen vorliegen?[2]
- Wie kann der Nutzer erkennen, dass die an ihn gelieferte Archivinformation auch integer und authentisch ist?[3]
- Wie kann das Archiv bei fortwährendem technologischem Wandel gewährleisten, dass die Nutzer die erhaltenen Informationen mit ihren verfügbaren technischen und intellektuellen Mitteln auch interpretieren können?

Erst wenn sich ein Archiv in Bezug auf den Zugriff mit den oben genannten Fragen befasst, handelt es strategisch im Sinne der Langzeitarchivierung. Die entsprechenden Maßnahmen bestehen natürlich zum Teil aus der Einführung und Implementierung von geeigneten technischen Infrastrukturen und Lösungen. Da die technischen Lösungen aber mit der Zeit auch veralten und ersetzt werden müssen, sind die organisatorisch-strategischen Maßnahmen eines Archivs von entscheidender Bedeutung. Unter diesem Gesichtspunkt sind Standardisierungen von globalen dauerhaften Identifikatoren, Zugriffsschnittstellen, Qualitätsmanagement und Zusammenschlüsse von Archiven unter gemeinsamen Zugriffsportalen eine wichtige Aufgabe für die nationale und internationale Gemeinde der Gedächtnisorganisationen.

---

2 vgl. Funk, Stefan: *Kap 7.2 Digitale Objekte und Formate*

3 nestor - Materialien 8: nestor - Kompetenznetzwerk Langzeitarchivierung / Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, Version 1 (Entwurf zur Öffentlichen Kommentierung), Juni 2006, Frankfurt am Main : nestor c/o Die Deutsche Bibliothek, http://www.langzeitarchivierung.de/publikationen/expertisen/expertisen.htm#nestor-materialien8; Punkt 6.3 S. 16. Als Version 2 unter http://nbn-resolving.de/urn:nbn:de:0008-2008021802
abrufbar Es ist davon auszugehen, dass in den nächsten Jahren eine Klärung erfolgt, welche weiteren Persistenten Identifikatoren – außer den in diesem Kapitel vertieft Behandelten URN und DOI – in der Anwendung Bedeutung erhalten werden.

9.2 Workflows für den Objektzugriff

Dirk von Suchodoletz

Es genügt nicht, lediglich ein digitales Objekt bit-getreu zu bewahren, sondern es sind Vorkehrungen zu treffen, um dieses Objekt zu einem späteren Zeitpunkt wieder darstellen zu können. Hierzu dienen bestimmte Workflows, die ein Langzeitarchiv implementieren sollte. Deshalb beschäftigt sich dieser Abschnitt mit der theoretischen Fundierung und Formalisierung technischer Abläufe, wie sie beispielsweise mit dem DIAS-Projekt zu Beginn des Jahrtausends an der Königlichen Bibliothek der Niederlande eingeführt wurden[4].

Der zentrale Ausgangspunkt der Überlegungen liegt darin begründet, dass digitale Objekte nicht allein aus sich heraus genutzt oder betrachtet werden können. Stattdessen benötigen sie einen geeigneten Kontext, damit auf sie zugegriffen werden kann. Dieser Kontext, im Folgenden Erstellungs- oder Nutzungsumgebung genannt, muss geeignete Hardware- und Softwarekomponenten so zusammenfügen, dass je nach Typ des digitalen Objekts dessen Erstellungsumgebung oder ein geeignetes Äquivalent erzeugt wird. Für diese Schritte der Wiederherstellung sind seitens des Archivbetreibers geeignete Workflows (Arbeitsabläufe) vorzusehen. Um diese beschreiben zu können, sind sogenannte „View-Paths“ ein zentrales Konzept. Diese Darstellungspfade liefern die grundlegenden Anweisungen zur Konstruktion geeigneter technischer Workflows für das Sichtbarmachen oder Ablaufenlassen verschiedener digitaler Objekte.

Den Betreibern eines digitalen Langzeitarchivs wachsen aus diesen Überlegungen verschiedene Aufgaben zu. Hierzu zählen die Bestimmung des Typs eines Objekts bei der Einstellung ins Archiv (dem sog. Ingest in der OAIS-Terminologie) und die Beschaffung und Ablage der notwendigen Metadaten, auf die an anderer Stelle in diesem Handbuch ausführlich eingegangen wird.

Für den späteren Objektzugriff spielt die Überprüfung, inwieweit im Langzeitarchivierungssystem eine für diesen Objekttyp passende Nutzungsumgebung vorhanden ist, eine besondere Rolle. Deren technischer Workflow wird nachfolgend näher ausgeführt. Dabei können View-Path und Nutzungsumgebung je nach Art der betreibenden Organisation und ihrer spezifischen Anforderungen, die typischerweise durch „Significant Properties“[5] beschrieben

4 Vgl. van Diessen; Steenbakkers 2002 und van Diessen 2002, S. 16f

5 Vgl. http://www.significantproperties.org.uk/index.html sowie http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.aspx

werden, unterschiedlich aussehen. Dies resultiert je nach Benutzergruppe oder Einrichtung in verschiedene Kriterien, nach denen Darstellungspfade bestimmt werden. Es lassen sich drei wesentliche Phasen voneinander unterscheiden (Abbildung 1):

- Notwendige Arbeitsschritte und Handlungen bei der Objektaufnahme in das OAIS-konforme Archiv.
- Workflows, die im Laufe des Archivbetriebs umzusetzen sind.
- Abläufe für den Objektzugriff nach der Objektausgabe an den Endbenutzer.

Der View-Path zum Zugriff auf die unterschiedlichen, im Langzeitarchiv abgelegten Objekttypen ist im Moment der Archivausgabe festzulegen. An dieser Stelle müssen diverse Workflows implementiert sein, die es erst erlauben, dass ein späterer Archivnutzer tatsächlich auf das gewünschte Objekt zugreifen kann. Hierbei spielt die auf das Objekt angewendete Langzeitstrategie, ob Migration oder Emulation, keine Rolle (Abbildung 2). In jedem Fall muss das Archivmanagement dafür sorgen, dass eine passende Nutzungsumgebung bereitgestellt wird.

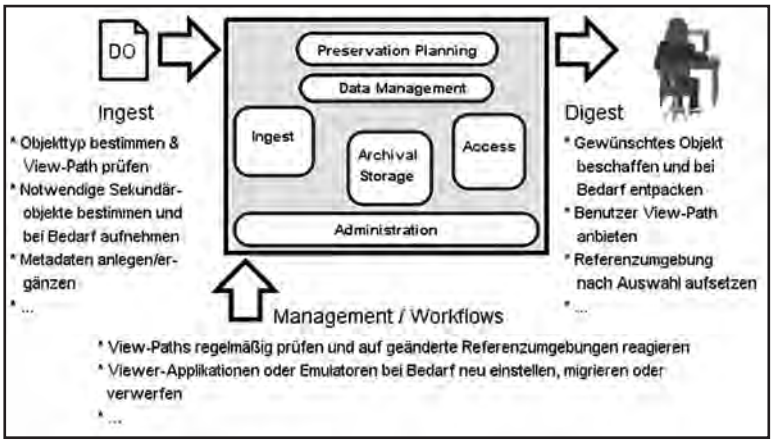

Abbildung 1: Bei der Unterhaltung eines digitalen Langzeitarchivs sind eine Reihe verschiedener technischer Workflows festzulegen und umzusetzen.

Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

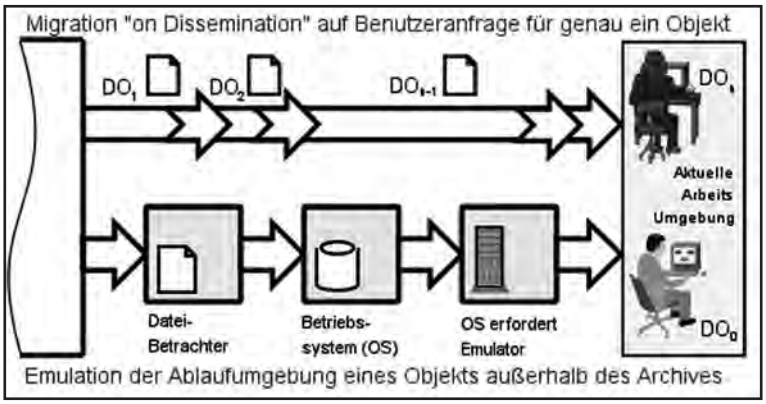

Abbildung 2: Unabhängig von der eingesetzten Archivierungsstrategie eines Objekts muss für einen geeigneten Zugriff gesorgt werden.

Formalisierter Zugriff auf archivierte Objekte

Die Wiederherstellung von Nutzungsumgebungen oder entsprechender Äquivalente lässt sich durch View-Paths formalisieren. Diese Wege starten vom darzustellenden oder auszuführenden digitalen Objekt. Sie reichen je nach angewendeter Archivierungsstrategie über verschiedene Zwischenschritte bis in die tatsächliche Arbeitsumgebung des Archivnutzers. Da digitale Archivalien, im Zusammenhang mit View-Path auch Primärobjekte genannt, nicht aus sich allein heraus genutzt werden können, werden weitere digitale Objekte benötigt. Hierzu zählen Viewer, Hilfsapplikationen, Betriebssysteme oder Emulatoren. Sie sind als Hilfsmittel, im Folgenden als Sekundärobjekte bezeichnet, nicht von primärem Archivierungsinteresse, jedoch zwingend ebenfalls zu berücksichtigen.

Das Konzept der Darstellungspfade wurde ursprünglich im Zuge des eingangs genannten DIAS-Projektes an der Königlichen Bibliothek der Niederlande entwickelt[6]. Die Abbildung 3 zeigt einen typischen View-Path-Verlauf ausgehend vom digitalen Objekt. Im Beispiel wurde es mittels einer bestimmten Software erzeugt, die ihrerseits wiederum auf einem Betriebssystem ausgeführt werden kann, das seinerseits wegen des Nicht-Vorhandenseins des originalen Rech-

6 Vgl. http://www.kb.nl/hrd/dd/dd_onderzoek/preservation_subsystem-en.html sowie van Diessen; Steenbakkers (2002).

ners einen Hardwareemulator erfordert. Dieser läuft als Applikation in der Arbeitsumgebung des Archivnutzers.

Abbildung 3: Das Beispiel zeigt einen generellen View-Path zum Rendering/Ablaufenlassen eines digitalen Objekts eines bestimmten Typs unter Einsatz der Emulationsstrategie.

Weitere View-Path-Beispiele

Ein View-Path eines migrierten Objekts wäre entsprechend kürzer: Da eine in der aktuellen Arbeitsumgebung ablaufende Applikation genutzt werden kann, muss lediglich diese passend zum jeweiligen Objekttyp bereitgestellt werden. Umgekehrt kann sich ein View-Path bei geschachtelter Emulation verlängern: Steht der im obigen Beispiel genannte Hardwareemulator nur für ein älteres Betriebssystem statt der aktuellen Umgebung zur Verfügung, würde ein Zwischenschritt aus diesem obsoleten Betriebssystem mit passendem Emulator für die aktuelle Umgebung angehängt werden (Abbildung 4 unten).

Während der Ausgangspunkt des View-Paths durch das Primärobjekt fixiert ist, wird sich, erzwungen durch den technologischen Fortschritt und die sukzessive Obsoleszenz vorhandener Rechnerplattformen, der Endpunkt des View-Path im Zeitablauf verschieben. Weiterhin hängt die Länge eines Darstellungspfades vom Typ des Objekts ab: Ist eine Applikation wie z.B. eine alte Datenbank von primärem Interesse, so entfällt beispielsweise der Zwischenschritt der Erstellungs- oder Anzeigeapplikation, da sie direkt auf einem Betriebssystem ausgeführt werden kann.

View-Paths müssen nicht automatisch eindeutig bestimmt sein. Aus Sicht des Archivmanagements bieten sich generell folgende Szenarien für Darstellungspfade an:

- Es existiert zum gegebenen Zeitpunkt ein Weg vom Primärobjekt zu seiner Darstellung oder Ausführung.

- Es existieren mehrere verschiedene View-Paths. Diese sind mit geeigneten Metriken zu versehen. Diese erlauben die Bewertung zur Verfügung stehender Alternativen und werden weiter hinten besprochen.
- Es kann Archivobjekte geben, zu denen zu bestimmten Zeitpunkten kei-

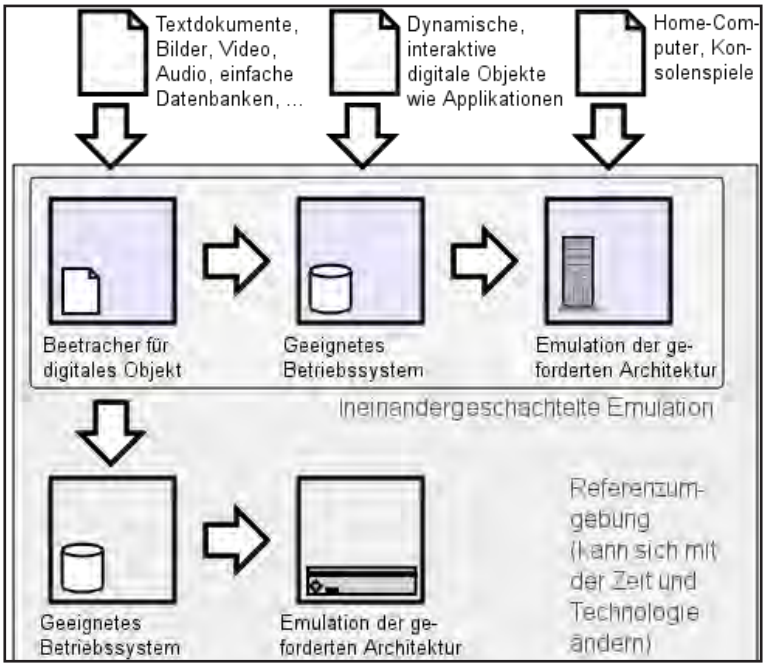

Abbildung 4: Die Länge eines View-Path hängt sowohl vom Typ des Objekts als auch von der eingesetzten Archivierungsstrategie ab.

ne View-Paths konstruierbar sind.

Referenzumgebung - Endpunkt von View-Paths

Referenzumgebungen haben jeweils festgelegte Eigenschaften, die sich aus der Definition ihrer Hard- und Softwareumgebung bestimmen. Dieses sind die aktuellen Arbeitsumgebungen, in denen sich Archivbenutzer bewegen. Sie ändern sich zwangsläufig im Laufe der Zeit und sind unter anderem durch die Beschaffungspolitik der jeweiligen Gedächtnisorganisation determiniert. View

Paths sollten sich auf wohldefinierte Referenzumgebungen[7] beziehen.[8] Geeignete Referenzumgebungen versuchen deshalb, in möglichst kompakter und gut bedienbarer Form ein ganzes Spektrum von Nutzungsumgebungen zur Verfügung zu stellen. Dabei sollte die Basisplattform möglichst der jeweils aktuellen Hardware mit jeweils üblichen Betriebssystemen entsprechen. Das verhindert einerseits das Entstehen eines Hardwaremuseums mit hohen Betriebskosten und andererseits findet sich der Benutzer zumindest für das Basissystem in gewohnter Umgebung wieder.

Eine Referenzumgebung sollte in der Lage sein, neben der jeweiligen Nutzungsumgebung zusätzlich die notwendigen Hinweise zum Einrichten und zur Bedienung bereitzustellen. Dies schließt den geeigneten Zugriff auf die Objektmetadaten mit ein. Weitere Kriterien liegen in der Güte der Darstellung der Nutzungsumgebung, was durch den jeweils eingesetzten Emulator und seine Benutzer-Interfaces in der jeweiligen Referenzumgebung mitbestimmt wird.

Andererseits können Referenzumgebungen als Endpunkte eines View-Paths

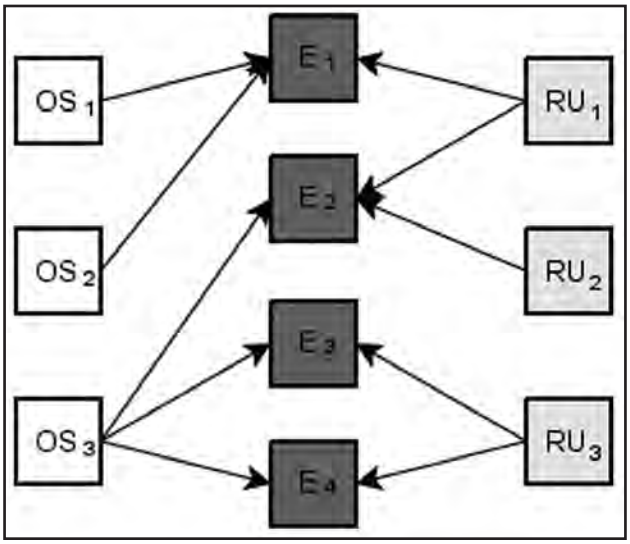

Abbildung 5: Die Auswahl der Emulatoren (E) wird in vielen Fällen nicht nur durch das Betriebssystem (OS) im View-Path, sondern auch von der anderen Seite durch die verfügbaren Referenzumgebungen (RU) beeinflusst.

7 Gedächtnisorganisationen haben typischerweise bestimmte Arbeitsumgebungen an ihren Arbeitsplätzen für ihre eigenen Archivare oder externe Nutzer (Abbildung 1). Diese Umgebungen können sie in einem gewissen Umfang selbst mitbestimmen. Vgl. a. van Diessen 2002.

8 van Diessen (2002b)

diesen umgekehrt mitbestimmen (Abbildung 5). Wenn technische, finanzielle oder personelle Restriktionen bestimmte Referenzumgebungen, die potenziell nachgefragt werden, nicht erlauben, kann dies die Wahl eines Emulators wesentlich beeinflussen. Die Rekursion kann sich unter Umständen noch weiter auf frühere Entscheidungsknoten auswirken.

Wegen ihrer spezielleren Anforderungen, die durch die eingesetzten Emulatoren und Viewer beeinflusst werden, ist es für die Betreiber von Langzeitarchiven vielfach sinnvoll, eine für die Objekte ihres Archivs relevante Referenzplattform selbst zu definieren und bereitzustellen. Diese unterscheidet sich je nach Anwendung und Gedächtnisorganisation: Bibliotheken und Archive brauchen in erster Linie Viewer für ihre migrierten statischen Objekte, die sie ihren Nutzern innerhalb ihrer Recherchesysteme anbieten wollen. Darüber hinaus können Emulatoren für nicht-migrierbare Archivinhalte und als Kontrollwerkzeug für mehrfach migrierte Objekte benötigt werden.[9]

Technische Museen oder Ausstellungen leben eher von interaktiven Objekten. Die Referenz-Workstation ist entsprechend den zu zeigenden Exponaten zu bestücken. Ähnliches gilt für multimediale Kunstobjekte. Hier könnten die Significant Properties jedoch sehr spezielle Anforderungen an eine möglichst authentische Präsentation stellen.

Für Firmen oder Institutionen werden in den meisten Fällen lediglich View-Paths erforderlich sein, die sich bereits mittels X86-Virtualisierern[10] komplett erstellen lassen. Die erwarteten Objekte sind eher statischer Natur und wurden typischerweise auf PCs verschiedener Generationen erstellt. Generell muss es sich bei den eingesetzten Referenz-Workstations nicht um die jeweils allerneueste Hardwaregeneration handeln. Stattdessen sollte jene Technologie angestrebt werden, die einen optimalen Austausch erlaubt und den Anforderungen der jeweiligen Nutzer gerecht wird.

Je nach Art des Archivs kann ein Datenaustausch mit der Außenwelt erforderlich werden: Nicht alle Objekte müssen bereits von Anfang an im Archiv abgelegt sein. Es können später immer wieder Objekte, beispielsweise aus Nachlässen von Wissenschaftlern, Künstlern oder Politikern auftauchen. In solchen

9 Die Darstellung eines Objekts via Emulation wird typischerweise deutlich aufwändiger sein. Gerade für häufig nachgefragte, statische Objekte bietet sich deshalb die Migration an. Bei Zweifeln an der Authentizität kann mittels Emulation das Ergebnis des n-ten Migrationsschritts mit dem unveränderten Originalobjekt verglichen werden.

10 Softwareprodukte wie VMware oder VirtualBox erlauben das Nachbilden eines X86er PCs auf einer X86er 32 oder 64-bit Maschine. Solange die Treiberunterstützung besteht, können ältere Betriebssysteme, wie Windows95 oder 2000 bequem innerhalb einer Applikation (dem Virtualisierer) auf dem Standard-Desktop des Benutzers ausgeführt werden.

Fällen wird es zu Zwecken der Datenarchäologie[11] von Interesse sein, externe Objekte in bestehende Workflows einspeisen zu können. Umgekehrt sollen vielleicht Objekte für externe Nutzung speicher- oder ausdruckbar sein.

Für alle Schritte muss ein ausreichendes Bedienungswissen vorhanden sein. Hierzu werden sich neue Berufsfelder, wie das eines digitalen Archivars herausbilden müssen, um ausgebildetes Fachpersonal auch für recht alte Nutzungs-

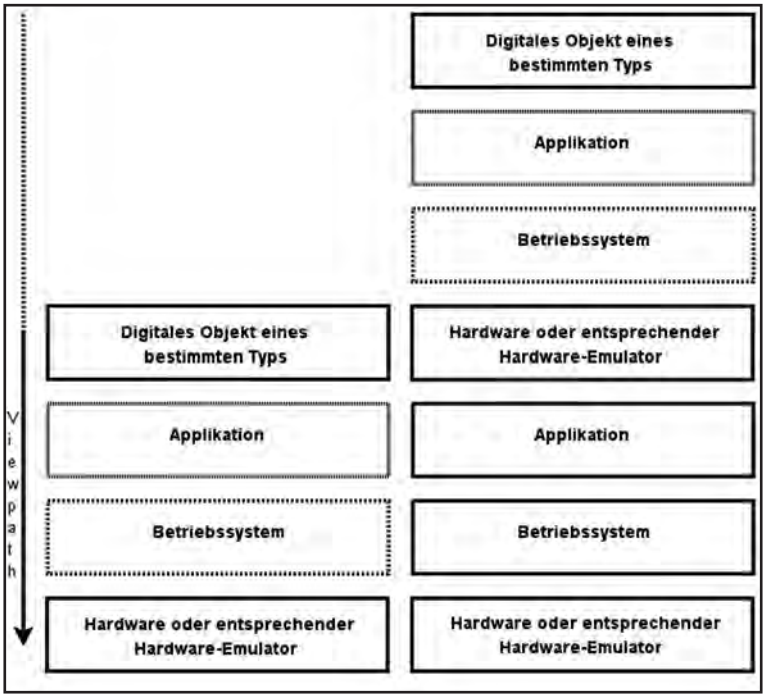

Abbildung 6: Ein View-Path in einer Darstellung als Schichtenmodell. Je nach Typ des Objekts muss nicht jede Schicht durch eine Softwarekomponente repräsentiert sein.

umgebungen vorhalten zu können. Selbst wenn diese nicht mehr direkt auf der Originalhardware ablaufen, so müssen sie innerhalb der Emulatoren bedient werden können.

View-Path-Varianten und -Auswahl

Einen View-Path kann man als Entscheidungsbaum interpretieren, an dessen Wurzel das interessierende Primärobjekt steht. Ein Blatt ohne weitere Verzwei-

11 Im Sinne des Umgangs (Bestimmung) mit sehr alten, lange schon nicht mehr zeit-typischen digitalen Objekten.

gungen stellt das Ende des Pfades in Form einer gültigen Referenzumgebung dar. Zur Veranschaulichung des Aufbaus der geforderten Nutzungsumgebung sollte man sich am besten ein Schichtenmodell vorstellen, wie es in Abbildung 6 präsentiert wird.
Eine ganze Reihe digitaler Objekte können durch mehr als eine Applikation dargestellt oder ablauffähig gemacht werden. Dabei können die Ergebnisse in Authentizität, Komplexität oder Qualität differieren. Aus diesen Überlegungen folgen Pfadverzweigungen und auf der Schicht der Applikation eine Auswahlmöglichkeit. Ähnliches wiederholt sich für die Anforderung der Applikation nach einem Betriebssystem. Auf dieser Ebene kann erneut eine Verzweigung auftreten. Die Rekursion setzt sich mit dem Betriebssystem und einer Menge an geeigneten Hardwareemulatoren fort.

Da der technologische Fortschritt wesentlichen Einfluss auf die Referenzumgebung hat und nur bedingt durch den Archivbetreiber beeinflusst werden kann, bestimmen sich View-Path und Referenzumgebung gegenseitig. Auf den Zwischenschichten stehen Betriebssysteme und Hardwareemulatoren über Gerätetreiber in Abhängigkeit zueinander.

Die Modellierung des View-Paths in Schichten erfolgt nicht eng fixiert: So reduziert sich beispielsweise bei einem digitalen Primärobjekt in Form eines Programms die Zahl der Schichten. Ähnliches gilt für einfache Plattformen wie bei Home-Computern, wo keine Trennung zwischen Betriebssystem und Applikation vorliegt (Abbildung 4 Mitte). Darüber hinaus können Schichten wiederum gestapelt sein, wenn es für einen bestimmten Emulator erforderlich wird, seinerseits eine geeignete Nutzungsumgebung herzustellen, was im rechten Teil von Abbildung 6 gezeigt wird.

Metriken als Entscheidungskriterien

Eine sinnvolle Erweiterung des etwas starren Ansatzes im ursprünglichen DIAS-Preservation-Modell (van Diessen; Steenbakkers 2002) könnte in der Gewichtung der einzelnen View-Path-Varianten liegen. Dies ließe sich durch eine beschreibende Metrik formalisieren. Gerade wenn an einem Knoten mehr als eine Option zur Auswahl steht (Abbildung 7), erscheint es sinnvoll:

- Präferenzen der Archivnutzer beispielsweise in Form der Auswahl der Applikation, des Betriebssystems oder der Referenzplattform zuzulassen.
- Gewichtungen vorzunehmen, ob beispielsweise besonderer Wert auf die Authentizität der Darstellung (van Diessen; van der Werf-Davelaar 2002) oder eine besonders einfache Nutzung gelegt wird.
- Vergleiche zwischen verschiedenen Wegen zuzulassen, um die Sicherheit

und Qualität der Darstellung der Primärobjekte besser abzusichern.
- Den Aufwand abzuschätzen, der mit den verschiedenen Darstellungspfaden verbunden ist, um bei Bedarf eine zusätzliche ökonomische Bewertung zu erlauben.

Ein Ergebnis könnten mehrdimensionale Metriken sein, die mit den Objektmetadaten gespeichert und durch das Archivmanagement regelmäßig aktualisiert werden. Hierzu sollte eine Rückkopplung mit den Archivbenutzern erfolgen. So könnten in die Aktualisierungen Bewertungen seitens der Nutzer einfließen, die auf dem praktischen Einsatz bestimmter View-Paths sowie ihrer Handhabbarkeit und Vollständigkeit beruhen.

Aggregation von View-Paths

Wie erläutert, kann für bestimmte Objekttypen mehr als ein Darstellungspfad existieren. Dieses kann die Wahrscheinlichkeit des langfristig erfolgreichen Zugriffs verbessern – jedoch zum Preis potenziell höherer Kosten. Ausgehend vom Objekttyp und der eventuell notwendigen Applikation erhält man weitere View-Paths für andere Objekte, die automatisch anfallen: Ein einfaches Beispiel demonstrieren die sogenannten Office-Pakete, Zusammenstellungen verschiedener Applikationen. Sie können mit einer Vielfalt von Formaten umgehen – nicht nur mit denen der enthaltenen Teilkomponenten, sondern über Importfilter hinaus mit einer Reihe weiterer Dateiformate.[12]

Diese Betrachtungen können dazu dienen, eine potenziell überflüssige Referenzplattform zu identifizieren, die nur für ein bestimmtes Objekt vorgehalten wird, das aber auf alternativen Wegen ebenfalls darstellbar ist. So muss beispielsweise zur Betrachtung eines PDFs nicht eine besonders seltene Plattform genutzt werden, wenn ein gleichwertiger Viewer auf einer mehrfach genutzten anderen Plattform ebenfalls ablauffähig ist (Abbildung 7).

Schwieriger wird jedoch eine Zusammenlegung, wenn von den alternativen View-Paths nicht bekannt ist, ob sie ebenfalls zu 100% das gefragte Objekt rekonstruieren. Ein typisches Beispiel ist der Import von Word-Dokumenten in einer anderen als der Erstellungsapplikation. An diesem Punkt könnten ebenfalls die im vorangegangenen Abschnitt vorgenommenen Überlegungen zu Benutzerrückmeldungen in Betracht gezogen werden.

---

12 Ein Paket wie OpenOffice kann bezogen auf seine Importfilter für die verschiedenen Teilanwendungen wie Textverarbeitung, Tabellenkalkulation oder Präsentation inklusive der Unterstützung verschiedener älterer Versionen schnell über 100 verschiedene Formate statischer Objekte lesen.

Mit jedem View-Path sind bestimmte Kosten verbunden, die vom Archivmanagement beachtet werden müssen. Diese lassen sich für jede Schicht eines

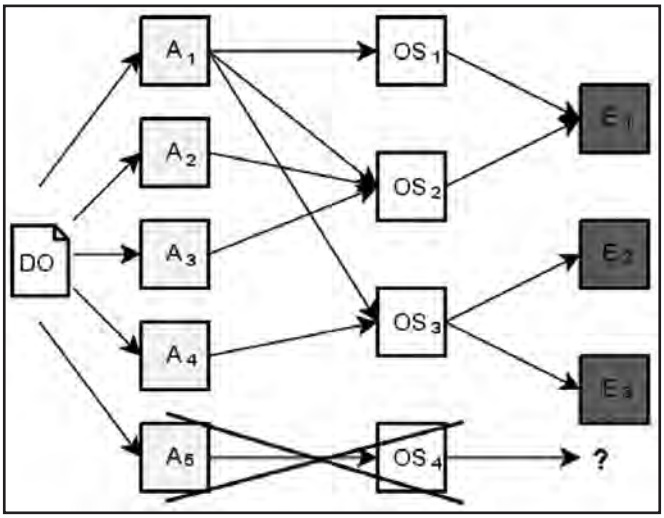

Abbildung 7: Redundante View-Paths zur Darstellung oder zum Abspielen eines digitalen Objekts (DO). Dieses kann im Beispiel von fünf verschiedenen Applikationen (A) geöffnet werden, die auf vier verschiedenen Betriebssystemen (OS) laufen können, wobei für ein OS kein geeigneter Emulator (E) bereitsteht.

View-Paths abschätzen. Übersteigen sie eine gewisse Schwelle, könnten ökonomische Erwägungen an die Aufnahme von Primärobjekten in das Archiv und das Archivmanagement geknüpft werden. In solchen Fällen könnten bestimmte Dateiformate abgelehnt werden, weil deren spätere Rekonstruktion zu hohe Aufwendungen erwarten lässt.

So sollte je nach Situation oder Archiv hinterfragt werden, einen Darstellungspfad aufrecht zu erhalten, wenn sinnvolle Alternativen existieren. Liegt beispielsweise ein Primärobjekt nach dem PDF 1.0 Standard vor, welches mit einem Werkzeug in einer Windows 3.11 Nutzungsumgebung erzeugt wurde, muss deshalb diese Umgebung nicht zwingend erhalten werden. Unter bestimmten Bedingungen kann auf einen View-Path verzichtet werden:

- Es existiert eine ausreichende Anzahl weiterer, gleichwertiger und dabei einfacherer Darstellungspfade.
- Wegen der guten und vollständigen Beschreibung des Formats ist es deutlich einfacher, einen Viewer für die jeweils aktuellen Arbeitsumge-

bungen zu migrieren, als alte Nutzungsumgebungen durch Emulatoren zu sichern.

- Dieser Objekttyp ist der einzige, der eine Windows 3.11 Umgebung potenziell verlangt.

- Es gibt kein spezielles Interesse, diese Nutzungsumgebung aufrecht zu erhalten, da sie nicht im Fokus der Institution liegt.

Ein solches Vorgehen ließe sich auf andere View-Paths ausdehnen, die für eine Reihe von Dateiformaten und Applikationen Apple- oder alternativ Microsoft-Betriebssysteme voraussetzen. Wenn beispielsweise kein gesondertes Bedürfnis speziell nach dedizierten Apple-Nutzungsumgebungen besteht, weil beispielsweise die Art der Benutzerinteraktion, das Aussehen der grafischen Oberfläche und der Applikation in dieser von eigenständigem Interesse sind, könnte ein solcher Zweig im Archiv geschlossen werden. Besonders gut lässt sich dies am OpenOffice veranschaulichen, welches für etliche kommerzielle und freie UNIX-Varianten, wie BSD, Solaris oder Linux, Mac-OS X und selbstredend für die Windows-Betriebssysteme angeboten wird.

Ähnlich liegt der Fall in der trivialen Vervielfachung der View-Paths durch unterschiedliche Anpassungen von Programmen und Betriebssystemen an nationale und regionale Besonderheiten. Hierzu zählen natürliche Sprachen, spezielle Zeichen oder auch Währungen und Einheiten. Während Open-Source-Betriebssysteme und Applikationen schon lange mehrere Sprachpakete in einer Installation erlauben, setzte dies sich bei kommerziellen Systemen erst recht spät durch. Das Bedürfnis von einem Betriebssystem wie Windows 2000 alle verschiedenen Landesversionen aufzubewahren, lässt sich vermutlich besser in Kooperationen ähnlicher Gedächtnisinstitutionen über Landesgrenzen hinweg erreichen. Dies würde zudem Kosten und Aufwand für redundante Lizensierungen reduzieren helfen.

Einfacher und damit oft günstiger zu pflegenden View-Paths kann der Vorrang vor anderen eingeräumt werden. Jedoch sind auch hier Voraussagen schwierig und es kann die Gefahr bestehen, dass sich mit dem Wandel der Referenzumgebungen die Kostenstrukturen erneut ändern. Andererseits lassen sich Vorkehrungen treffen, dass seltenere View-Paths zumindest an spezialisierten Institutionen mit besonderem Sammelauftrag und entsprechender Finanzierung weiter betreut werden.

Die verschiedenen Strategien der Langzeitarchivierung der Emulatoren – Migration oder Schachtelung generieren unterschiedliche Aufwendungen im Archivbetrieb:

- Die geschachtelte Emulation sorgt für eher längere Darstellungspfade bei geringem Migrationsbedarf. Der Aufwand entsteht beim zunehmend komplexer werdenden Zugriff auf das Primärobjekt.

- Die Migration von Emulatoren, Universal Virtual Machines und modulare Ansätze[13] generieren eher kurze View-Paths bei einfacherem Zugriff auf das Primärobjekt. Jedoch liegt der Aufwand im regelmäßigen Update aller benötigter Emulatoren oder ihrer Teilkomponenten.

Der erste Ansatz ist aus diesem Grund eher für Objekte mit seltenem Zugriff oder Institutionen mit kleinen, speziell ausgebildeten Nutzerkreisen wie Archive, geeignet. Die zweite Strategie passt sicherlich besser auf viel genutzte

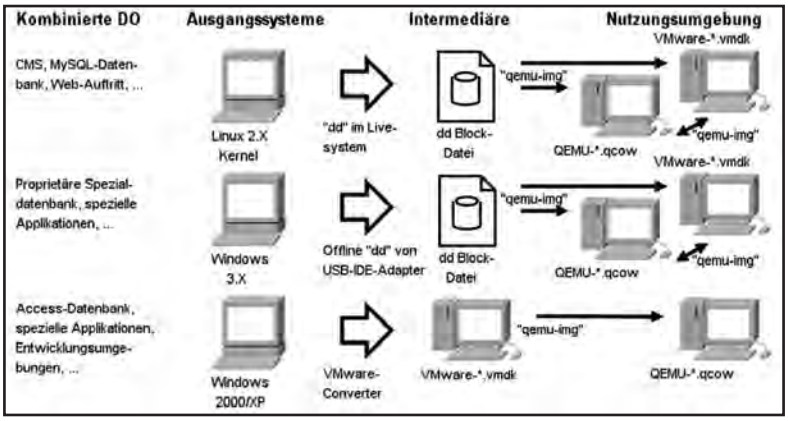

Abbildung 8: In Abhängigkeit vom zu archivierenden Gesamtsystem kombinierter digitaler Objekte existieren verschiedene Wege zur Archivierung.

Objekttypen mit größeren Anwenderkreisen. Eine Reihe von Kostenfaktoren kann durch gemeinschaftliche Anstrengungen und verteilte Aufgaben reduziert werden.

Kombinierte Archivobjekte

Primärobjekte mit bereits im Paket enthaltenen notwendigen Sekundärobjekten wie bestimmte Metadaten zur Beschreibung, Viewer und Hilfsprogramme, zusammengeführt in einem Archival Information Package (AIP), können bereits

13 Vgl. van der Hoeven; van Diessel; van der Meer (2005).

aufbereitet vorgehalten werden.

Solche kombinierten Objekte könnten aus der Überführung eines physischen Systems in ein virtuelles entstehen. Dieses Vorgehen lässt sich beispielsweise auf sehr spezielle dynamische Objekte, wie Datenbanken von Content Management Systemen (CMS) für WWW-Auftritte, Firmendaten aus Produktionsanlagen oder Entwicklungsabteilungen vorstellen (Abbildung 8): Statt Daten, Applikation und notwendiges Betriebssystem im Zuge der Archivierung voneinander zu lösen, könnte eine komplette Maschine in eine emulierte Umgebung überführt werden. Damit bleibt die Serverfunktionalität mit ihren gesamten Einstellungen als eine Einheit erhalten. Das Verfahren kann auf diese Weise sicherstellen, dass ein AIP über Archive hinweg ohne einschränkende Nebenbedingungen ausgetauscht werden kann.

Das Zusammenfassen bestimmter View-Paths in einer gemeinsamen Umgebung könnte einerseits für bestimmte Forschungsthemen und andererseits für eine Vereinfachung der View-Path-Erstellung für den Benutzerbetrieb sinnvoll sein.

Eine etwas anders gelagerte Zusammenfassung schlagen Reichherzer[14] und Brown (2006) vor, um auf Datensätze der öffentlichen Administration zu späteren Zeitpunkten zugreifen zu können. Solche typischerweise in einem komprimierten Archiv zusammengefassten Daten von Erhebungen, Statistiken oder Auswertungen umfassen eine Reihe typischer Objektformate einer gewissen Epoche. Diese könnten in einem gemeinsamen Container untergebracht sein und Hilfsprogramme, wie den Dokumentenausdruck nach Postscript oder PDF enthalten. Passende virtuelle Drucker, die durch Postscript- oder PDF-Generatoren repräsentiert werden, können fehlende Exportfilter alter Applikationen ersetzen.

View-Paths als OAIS-Workflows

Im OAIS-Referenzmodell übernimmt das Management des digitalen Langzeitarchivs eine Reihe von Aufgaben, die sich mit dem Lifecycle-Management und damit verbundenen Workflows von Primärobjekten befassen. Hierfür wurden besonders die Arbeitsprozesse für die Objektausgabe betrachtet. Für das langfristige Management wird insbesondere die Aufgabe des Preservation Planning interessant.[15]

Ein zentrales Moment ist die regelmäßige Kontrolle der View-Paths bei

14 Die Idee hierzu findet sich bereits bei Rothenberg (1998) sowie Reichherzer und Brown (2006).

15 van Diessen (2002a)

einem Wechsel der Referenzumgebung als Bezugsgröße. Jeder Plattformwechsel stellt neue Anforderungen für die Wiederherstellung von Nutzungsumgebungen. Bei dieser Überprüfung handelt es sich um einen iterativen Prozess, der über alle registrierten Objekttypen des Archivs abläuft. Hierfür ist jeweils eine passende Strategie für den Übergang von einer Referenzumgebung auf eine neue zu suchen. Generell gilt: Neue Objekttypen und Dateiformate erfordern neue Darstellungspfade.

Es ergeben sich verschiedene Anforderungen an den Archivbetrieb:

- Erstellung eines Hintergrundarchivs - In diesem werden die einzelnen Elemente des View-Path dauerhaft abgelegt. Sie werden dann genauso behandelt wie Primärobjekte. An dieser Stelle kann überlegt werden, ob bestimmte Einzelobjekte, wie Emulatoren, spezifische Hilfsprogramme und Beschreibungen in einem AIP gebündelt oder einzeln abgelegt werden.
- Betrieb eines Online-Archivs für den Direktzugriff - Für häufig nachgefragte Sekundärobjekte kann es sinnvoll sein, diese zusätzlich zum Langzeitarchiv in einem speziellen Archiv, wie einem aktuellen Dateisystem einer Referenzumgebung, vorzuhalten. Das kann einerseits das Langzeitarchiv entlasten und andererseits zu einem beschleunigten Ablauf der View-Path-Erstellung führen.
- Anlage eines View-Path-Caches - Für häufiger nachgefragte und aufwändiger zu generierende Darstellungspfade kann die Vorhaltung vorbereiteter Nutzungsumgebungen den Aufwand für Nutzer und Archivbetreiber reduzieren. Diese Caches könnten als Teil des Online-Archivs oder direkt auf der Referenzplattform abgelegt sein.

Die aufgezeigten Überlegungen haben klare ökonomische Implikationen für die mit digitalen Objekten befassten Gedächtnisorganisationen. Sie werden sich im Zuge der technischen Workflows, die sich mit der Wiederherstellung von Nutzungsumgebungen befassen, einen Teil der notwendigen Entwicklungen der Emulatoren und Viewer selbst leisten oder diese Leistung am Markt einkaufen müssen. Entstehen bei den angestrebten Entwicklungen offene Standards und Werkzeuge, wie PRONOM[16] oder Dioscuri,[17] können sich einerseits die Belastungen der Einzelinstitution in Grenzen halten und andererseits verbindliche Verfahren entwickeln, die von einer breiten Anwendergemeinschaft unterstützt werden.

16 Vgl. http://dioscuri.sourceforge.net/

17 Vgl. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Literatur

The Preservation Manager for the e-Depot: http://www.kb.nl/hrd/dd/dd_onderzoek/preservation_subsystem-en.html

van Diessen, Raymond; Steenbakkers, Johan F. (2002): *The Long-Term Preservation Study of the DNEP project - an overview of the results,* The Hague, The Netherlands; http://www.kb.nl/hrd/dd/dd_onderzoek/reports/1-overview.pdf

van Diessen, Raymond (2002): *Preservation Requirements in a Deposit System*, The Hague, The Netherlands: http://www.kb.nl/hrd/dd/dd_onderzoek/reports/3-preservation.pdf

van der Hoeven, Jeffrey; van Diessen, Raymond; van der Meer; K. (2005):*Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects*, Journal of Information Science, Vol. 31, No. 3, 196-208; DOI: 10.1177/0165551505052347

van Diessen, Raymond; van der Werf-Davelaar, Titia (2002): *Authenticity in a Digital Environment,* The Hague, The Netherlands, http://www.kb.nl/hrd/dd/dd_onderzoek/reports/2-authenticity.pdf

Rothenberg, Jeff (1998): Avoiding Technological Quicksand: *Finding a Viable Technical Foundation for Digital Preservation in „The State of Digital Preservation: An International Perspective"*, Washington, http://www.clir.org/pubs/reports/rothenberg/contents.html

Reichherzer, Thomas; Geoffrey, Geoffrey (2006): *Quantifying software requirements for supporting archived office documents using emulation*, International Conference on Digital Libraries 06: Proceedings of the 6th ACM/IEEE-CS joint Conference on Digital Libraries; S. 86-94

von Suchodoletz, Dirk (2009): *Funktionale Langzeitarchivierung digitaler Objekte - Erfolgsbedingungen des Einsatzes von Emulationsstrategien*, urn:nbn:de:0008-2008070219, ISBN 978-3-86727-979-6

9.3 Retrieval

Matthias Neubauer

Genauso wichtig wie die sichere Archivierung der digitalen Objekte ist auch die Möglichkeit, diese Objekte wieder aus dem Archiv herauszuholen und zu nutzen. Dabei muss gewährleistet sein, dass die Objekte den Zustand und den Informationsgehalt zum Zeitpunkt des Einspielens in das Archivsystem widerspiegeln. Im Idealfall sollte das Objekt noch exakt so abrufbar sein, wie es einmal in das Archiv eingespielt wurde. Je nach Verwendungszweck kann es jedoch auch sinnvoll sein, eher eine migrierte Form eines Objektes abzurufen. Einige wichtige Punkte, die es beim Zugriff von archivierten Objekten zu beachten gilt, sollen im Folgenden kurz erläutert werden.

Objektidentifikation

Zunächst ist eine eindeutige Identifikation des abzurufenden Objektes wichtig. Zu dieser Thematik existieren vielerlei Lösungen und Philosophien. Einige werden in den folgenden Kapiteln zum Thema „Persistent Identifier" vorgestellt. Grundsätzlich muss es anhand der verwendeten Identifizierungen möglich sein, jedwede Form und Version eines digitalen Objektes aus dem Langzeitarchiv abzurufen. Dies kann gegebenenfalls auch durch eine Kombination von externen und internen Identifikatoren realisiert werden.

Datenkonsistenz

Die Unversehrtheit der Daten hat höchste Priorität. Innerhalb des Archivs sollte durch geeignete Routinen zwar sichergestellt sein, dass der originale digitale Datenstrom erhalten bleibt. Jedoch können auch - und vor allem - bei der Übertragung der Daten aus dem Archiv heraus Inkonsistenzen durch Übertragungsfehler oder andere Störeinflüsse entstehen. Idealerweise sollte daher bei jedem Zugriff auf ein Archivobjekt über Checksummenvergleiche die Unversehrtheit der Daten sichergestellt werden. Je nach Art und Status der Daten kann diese Überprüfung auch nur stichprobenartig erfolgen.

Versionsmanagement

Je nach Verwendungszweck der Daten kann es entweder sinnvoll sein, das ursprüngliche Originalobjekt aus dem Archiv herauszuholen oder aber auch eine migrierte Form zu nutzen. Die höchste Authentizität wird man sicherlich mit

dem ursprünglichen Objekt erreichen, jedoch kann es sich auf zukünftigen Systemen sehr schwierig gestalten, die erhaltenen Daten aufzubereiten und zu nutzen (mehr darüber im Kapitel über Emulation und Migration). Ein gutes Langzeitarchivierungssystem sollte nach Möglichkeit sowohl Originalversion und letzte Migrationsform, als auch alle dazwischen liegenden Objektversionen zugreifbar halten, um eine vollkommene Transparenz und Rekonstruierbarkeit zu gewährleisten.

Interpretation und Aufbereitung der Daten

Sofern das digitale Objekt zum Zweck einer Präsentation oder Weiternutzung abgerufen wurde, muss es durch geeignete Methoden aufbereitet und verfügbar gemacht werden. Schon beim Einspielen der Daten in das Archivsystem ist daher darauf zu achten, dass man die Struktur des Objektes in den beiliegenden Metadaten dokumentiert. Zudem kann es notwendig sein, die innerhalb eines Archivsystems verwendeten Schlüsselnummern zur eindeutigen Identifikation von Dateiformaten zu entschlüsseln und auf ein anderes System einzustellen.

Caching

Unter dem Begriff „Caching" versteht man die Pufferung oft genutzter Daten in einem schnell verfügbaren und hochperformanten Zwischenspeicher. Im Falle des Retrieval aus einem Langzeitarchivierungssystem ist dies dann sinnvoll, wenn die Archivobjekte auch als Basis für Präsentationssysteme und den täglichen Zugriff dienen sollen. Um das Archivsystem nicht mit unnötigen Anfragen nach häufig genutzten Objekten zu belasten, wird ein lokaler Zwischenspeicher angelegt, der stark frequentierte Objekte vorhält und gegebenenfalls mit einer neuen Version innerhalb des Archivsystems synchronisiert beziehungsweise aktualisiert. Bei einem Zugriff auf das Objekt wird also nicht direkt das Archivsystem angesprochen, sondern zuerst geprüft, ob das Objekt bereits in der gewünschten Version lokal vorliegt. Eine kurze Kommunikation mit dem Archivsystem findet lediglich statt, um den Status und die Konsistenz des lokal vorliegenden Objektes zu validieren.

Sichere Übertragungswege

Um die Datensicherheit und den Datenschutz zu gewährleisten, sind sichere Übertragungswege zwischen dem Langzeitarchivierungssystem und dem zugreifenden System unerlässlich. Zwar kann eine etwaige Manipulation der Daten und Objekte durch die bereits angesprochene Checksummenüberprüfung erkannt werden, jedoch schützt dies nicht vor dem unerlaubten Zugriff Dritter

auf die Objekte des Archivsystems. Dies kann sowohl über sogenanntes Abhören der Datenleitung geschehen, als auch dadurch, dass unbefugte Dritte an Zugangsdaten und Netzwerkadressen des Archivsystems gelangen. Hier ist es daher sinnvoll, mit eindeutigen Befugnissen, sicheren Übertragungsprotokollen (wie HTTPS oder SFTP) und idealerweise Signaturschlüsseln und restriktiven IP-Freigaben zu arbeiten.

Datenübernahme in ein neues Archivsystem

Ein digitales Langzeitarchivsystem sollte die Möglichkeit bieten, alle Objekte zum Zwecke einer Migration auf ein neues oder anderes Archivsystem als Gesamtpaket oder als einzelne Objekte abzurufen. Verbunden mit dem einzelnen Objekt oder dem Gesamtpaket sollten auch alle gesammelten Metadaten sein. Sie sollten nach Möglichkeit komplett in das neue Archivsystem übernommen werden.

Diese Punkte sollten bei der Planung und Umsetzung von Zugriffsstrategien auf ein Archivsystem beachtet und mit einbezogen werden. Für individuelle Lösungen werden sicherlich auch noch weitere Faktoren eine Rolle spielen. Die jeweiligen Implementierungen sind natürlich auch stark von dem verwendeten Archivsystem abhängig.

9.4 Persistent Identifier (PI) – ein Überblick

Kathrin Schroeder

Warum Persistent Identifier?

Wer eine Printpublikation bestellt, kennt i.d.R. die ISBN – eine weltweit als eindeutig angesehene Nummer. Damit kann die Bestellung sicher ausgeführt werden. Eine ähnliche Nummerierung bieten Persistent Identifier für elektronische Publikationen, die im Internet veröffentlicht werden. Damit können sehr unterschiedliche digitale Objekte wie z.B. PDF-Dokumente, Bilder, Tonaufnahmen oder Animationen dauerhaft identifiziert und aufgefunden werden.

Als „ISBN für digitale Objekte" sind die gängigen Internetadressen, die Uniform Resource Locators (URL) nicht geeignet, da diese sich zu häufig ändern.[18] Stabile, weltweit eindeutige Identifier sind für ein digitales Langzeitarchiv unumgänglich, wie dies z.B. auch aus dem OAIS-Referenzmodell hervorgeht. Ein von außen sichtbarer stabiler Identifier ist für die zuverlässige Referenzierung sowie für die sichere Verknüpfung von Metadaten mit dem Objekt wichtig.

Kriterien

Kriterien an PI-Systeme können sehr unterschiedlich sein. Exemplarisch sind Kriterien, die in Der Deutschen Nationalbibliothek für die Entscheidung für ein PI-System zugrunde gelegt wurden, aufgeführt.

Standardisierung

- Verankerung in internationalen Standards

Funktionale Anforderungen

- Standortunabhängigkeit des Identifiers
- Persistenz
- weltweite Eindeutigkeit
- Der Identifier ist adressierbar und anklickbar (Resolving).

18 Weiterführende Informationen zu „Adressierung im Internet und Leistungsgrenzen standortgebundener Verweise" vgl. http://www.persistent-identifier.de/?link=202

- Es kann von einem PI gleichzeitig auf mehrere Kopien des Dokumentes (1:n-Beziehung) verwiesen werden.

Flexibilität, Skalierbarkeit

- Das PI-System ist skalierbar und
- flexibel in der PI-Anwendung selbst, d.h. es können neue Funktionalitäten hinzukommen, ohne die Konformität zum Standard zu gefährden.

Technologieunabhängigkeit und Kompatibilität

- Das PI-System ist generisch sowie protokoll- und technologieunabhängig als auch
- kompatibel mit existierenden Anwendungen und Diensten wie z.B. OpenURL, SFX, Z39.50, SRU/SRW.

Anwendung, Referenzen

- Wie verbreitet und international akzeptiert ist das PI-System?

Businessmodell und nachhaltiger Bestand

- Folgekosten (Businessmodell), Nachhaltigkeit des technischen Systems

PI-Beispiele

Nachfolgend werden die gegenwärtig als Persistent Identifier bekannten und publizierten Systeme, Spezifikationen und Standards tabellarisch vorgestellt. Zu Beginn wird das einzelne PI-System optisch hervorgehoben („Kürzel – vollständiger Name"). Die PI-Systeme sind alphabetisch geordnet.

Jede Tabelle beinhaltet die nachfolgenden Elemente:

Kurzbezeichnung	*allgemein verwendete oder bekannte Abkürzung des PI-Systems*
Erläuterung	*kurze, allgemeine inhaltliche Erläuterungen über das Ziel sowie die Funktionalitäten des PI-Systems*
Syntax	*Darstellung der allgemeinen Syntax des PIs Zusätzlich wird der jeweilige PI als URN dargestellt.*
Beispiel	*ein oder mehrere Beispiele für einen PI*
Identifizierung / Registry	*kurze Angaben, was mit dem PI identifiziert wird und ob ein Registry gepflegt wird*

Resolving	*Wird ein Resolving unterstützt, d.h. kann der Identifier in einer klickbaren Form dem Nutzer angeboten werden*
Anwender	*Anwendergruppen, Institutionen, Organisationen, die das PI-System unterstützen, z.T. erfolgt dies in Auswahl*
Tool-Adaption	*Vorhandene Tools, Adaption in Digital Library Tools oder anderen Content Provider Systemen*
Referenz	*Internetquellen, Die Angabe erfolgt in Form von URLs*

ARK - Archival Resource Key

Kurzbezeichnung	*ARK*
Erläuterung	*ARK (Archival Resource Key) ist ein Identifizierungsschema für den dauerhaften Zugriff auf digitale Objekte. Der Identifier kann unterschiedlich verwendet werden: Als Link* *•von einem Objekt zur zuständigen Institution,* *•von einem Objekt zu Metadaten und* *•zu einem Objekt oder dessen adäquater Kopie.*
Syntax	*[http://NMAH/]ark:/NAAN/Name[Qualifier]* *NMAH: Name Mapping Authority Hostport* *ark: ARK-Label* *NAAN: Name Assigning Authority Number* *Name: NAA-assigned* *Qualifier: NMA-supported*
Beispiel	*http://foobar.zaf.org/ark:/12025/654xz321/s3/f8.05v.tiff* *Als URN:* *urn:ark:/12025/654xz321/s3/f8.05v.tiff*
Identifizierung / Registry	*- ARK-Vergabe für alle Objekte* *- zentrales Registry für Namensräume*

Resolving	*Ja, ein zentrales Register der ARK-Resolving-Dienste soll in einer „globalen Datenbank“ erfolgen, die gegenwärtig nicht von einer internationalen Agentur wie z.B. der IANA betreut wird.*
Anwender	*15 angemeldete Institutionen: (Eigenauskunft) Darunter: California Digital Library, LoC, National Library of Medicine, WIPO, University Libraries Internet Archive, DCC, National Library of France*
Tool-Adaption	*Entwicklung der California Digital Library: Noid (Nice Opaque Identifier) Minting and Binding Tool*
Referenz	*http://www.cdlib.org/inside/diglib/ark/*
Bemerkungen	*Allerdings muss bei Kopien der spezif. Resolving-Service angegeben werden.*

DOI – Digital Object Identifier

Kurzbezeichnung	*DOI*
Erläuterung	*Anwendungen von Digital Object Identifiers (DOI) werden seit 1998 durch die International DOI Foundation (IDF) koordiniert. Dem DOI liegt ein System zur Identifizierung und dem Austausch von jeder Entität geistigen Eigentums zugrunde. Gleichzeitig werden mit dem DOI technische und organisatorische Rahmenbedingungen bereitgestellt, die eine Verwaltung digitaler Objekte sowie die Verknüpfung der Produzenten oder Informationsdienstleistern mit den Kunden erlauben. Dadurch wird die Möglichkeit geschaffen, Dienste für elektronische Ressourcen, die eingeschränkt zugänglich sind, auf Basis von DOIs zu entwickeln und zu automatisieren.* *Das DOI-System besteht aus den folgenden drei Komponenten:* *• Metadaten,* *• dem DOI als Persistent Identifier und* *• der technischen Implementation des Handle-Systems.* *Institutionen, die einen Dienst mit einem individuellen Profil aufbauen wollen, können dies in Form von Registration Agencies umsetzen. Das bekannteste Beispiel ist CrossRef, in dem die Metadaten und Speicherorte von Referenzen verwaltet und durch externe Institutionen weiterverarbeitet werden können.* *Die DOI-Foundation ist eine Non-Profit-Organisation, deren Kosten durch Mitgliedsbeiträge, den Verkauf von DOI-Präfixen und den vergebenen DOI-Nummern kompensiert werden.* *Die Struktur von DOIs wurde seit 2001 in Form eines ANSI/NISO-Standards (Z39.84) standardisiert, welche die Komponenten der Handles widerspiegelt:*

Syntax	*Präfix / Suffix*
Beispiel	*10.1045/march99-bunker* *Der Zahlencode “10” bezeichnet die Strings als DOIs, die unmittelbar an den Punkt grenzende Zahlenfolge “1045” steht für die vergebende Institution z.B. eine Registration Agency. Der alphanumerische String im Anschluss an den Schrägstrich identifiziert das Objekt z.B. einen Zeitschriftenartikel.* *Als URN:* *urn:doi:10.1045/march99-bunker*
Identifizierung / Registry	*- DOI-Vergabe für alle Objekte* *- zentrale Registrierung von Diensten,* *- Nutzer müssen sich bei den Serviceagenturen registrieren*
Resolving	*- Ja, Handle-System als technische Basis* *- Zentraler Resolving-Service* *- verschiedene, nicht kommunizierte dezentrale Dienste*
Anwender	*- 7 Registration Agencies (RA)* *Copyright Agency, CrossRef, mEDRA, Nielson BookData, OPOCE, Bowker, TIB Hannover* *- CrossRef-Beteiligte: 338* *CrossRef-Nutzer* *- Bibliotheken (970, auch LoC)* *- Verlage (1528)*
Tool-Adaption	*Tools, welche die Nutzung von DOIs vereinfachen und die Funktionalität erweitern: http://www.doi.org/tools.html* *Digital Library Tools von ExLibris*
Referenz	*http://www.doi.org*
Bemerkungen	*- DOIs sind URN-konform.* *- kostenpflichtiger Service* *- gestaffelte Servicegebühren*

ERRoL - Extensible Repository Resource Locator

Kurzbezeichnung	*ERRoL*
Erläuterung	*Ein ERRoL ist eine URL, die sich nicht ändert und kann Metadaten, Content oder andere Ressourcen eines OAI-Repositories identifizieren.*
Syntax	*„http://errol.oclc.org/“ + oai-identifier*
Beispiel	*http://errol.oclc.org/oai:xmlregistry.oclc.org:demo/ISBN/0521555132.ListERRoLs* *http://errol.oclc.org/oai:xmlregistry.oclc.org:demo/ISBN/0521555132.html* *http://errol.oclc.org/ep.eur.nl/hdl:1765/9*
Identifizierung / Registry	*OAI Registry at UIUC (Grainger Engineering Library Information Center at University of Illinois at Urbana-Champaign) http://gita.grainger.uiuc.edu/registry/ListRepoIds.asp?self=1*
Resolving	*http-Redirect*
Anwender	*Nicht zu ermitteln*
Tool-Adaption	*DSpace*
Referenz	*http://errol.oclc.org/* *http://www.oclc.org/research/projects/oairesolver/*
Bemerkungen	*Erscheint experimentell. Kein echter Persistent Identifier, da URLs aktualisiert werden müssen .*

GRI – Grid Resource Identifier

Kurzbezeichnung	*GRI*
Erläuterung	*Die Spezifikationen definieren GRI für eindeutige, dauerhafte Identifier für verteilte Ressourcen sowie deren Metadaten.*
Syntax	*s. URN-Syntax*
Beispiel	*urn:dais:dataset:b4136aa4-2d11-42bd-aa61-8e8aa5223211* *urn:instruments:telescope:nasa:hubble* *urn:physics:colliders:cern* *urn:lsid:pdb.org:1AFT:1*
Identifizierung / Registry	*s. URN*

Resolving	*Im Rahmen von applikationsabhängigen Diensten wie z.B. Web-Services.*
Anwender	*School of Computing Science, University of Newcastle upon Tyne, Arjuna Technologies http://www.neresc.ac.uk/projects/gaf/*
Tool-Adaption	*http://www.neresc.ac.uk/projects/CoreGRID/*
Referenz	*http://www.neresc.ac.uk/ws-gaf/grid-resource/*
Bemerkungen	*GRI sind URN-konform.*

GRid - Global Release Identifier

Kurzbezeichnung	*GRid*
Erläuterung	*GRid ist ein System, um Releases of Tonaufnahmen für die elektronische Distribution eindeutig zu identifizieren. Das System kann Identifizierungssysteme in der Musikindustrie integrieren. Dazu gehören ein Minimalset an Metadaten, um Rechte (DRM) eindeutig zuordnen zu können.*
Syntax	*A Release Identifier consists of 18 characters, and is alphanumeric, using the Arabic numerals 0 to 9 and letters of the Roman alphabet (with the exception of I and O). It is divided into its five elements in the following order:* • *Identifier Scheme* • *Issuer Code* • *IP Bundle Number* • *Check Digit*

Beispiel	*A1-2425G-ABC1234002-M* *A1 - Identifier Scheme (i.e. Release Identifier for the recording industry)* *2425G - Issuer Code – (for example ABC Records)* *ABC1234002 - IP Bundle Number (for example an electronic release composed of a sound and music video recording, screensaver, biography and another associated video asset)* *M - Check Digit*
Identifizierung / Registry	*RITCO, an associated company of IFPI Secretariat, has been appointed as the Registration Agency.*
Resolving	*Resource Discovery Service*
Anwender	*Unklar*
Tool-Adaption	*unklar*
Referenz	*ISO 7064: 1983, Data Processing – Check Character Systems* *ISO 646: 1991, Information Technology – ISO 7-bit Coded Character Set for Information Exchange.*
Bemerkungen	*Kostenpflichtige Registrierung (150 GBP) für einen Issuer Code für 1 Jahr.*

GUUID / UUID

Kurzbezeichnung	*GUUID / UUID*
Erläuterung	*GUIDs (Globally Unique IDentifier) sind unter der Bezeichnung „UUID“ als URN-Namespace bereits bei der IANA registriert. Aufgrund des Bekanntheitsgrades werden diese erwähnt.* *Ein UUID (Universal Unique Identifier) ist eine 128-bit Nummer zur eindeutigen Identifizierung von Objekten oder anderen Entities im Internet.* *UUIDs wurden ursprünglich in dem Apollo Computer-Netzwerk, später im Rahmen der Open Software Foundation's (OSF), Distributed Computing Environment (DCE) und anschließend innerhalb der Microsoft Windows Platforms verwendet.*
Syntax	*s. URN-Syntax*
Beispiel	*urn:aps:node:0fe46720-7d30-11da-a72b-0800200c9a66*
Identifizierung / Registry	*URN-Namespace-Registry*
Resolving	*Kein*
Anwender	*Softwareprojekte*
Tool-Adaption	*UUID-Generatoren: http://kruithof.xs4all.nl/uuid/uuidgen* *http://www.uuidgenerator.com/* *http://sporkmonger.com/*
Referenz	*http://www.ietf.org/rfc/rfc4122.txt*
Bemerkungen	*In der Spezifikation wird ein Algorithmus zur Generierung von UUIDs beschrieben. Wichtig ist der Ansatz, dass weltweit eindeutige Identifiers ohne (zentrale) Registrierung generiert und in unterschiedlichen Applikationen sowie verschiedene Objekttypen verwendet werden können. Wobei deutlich gesagt wird, dass UUIDs *nicht* auflösbar sind.*

Handle

Kurzbezeichnung	*Handle*
Erläuterung	*Das Handle-System ist die technische Grundlage für DOI-Anwendungen. Es ist eine technische Entwicklung der Corporation for National Research Initiatives. Mit dem Handle-System werden Funktionen, welche die Vergabe, Administration und Auflösung von PIs in Form von Handles erlauben, bereitgestellt. Die technische Basis bildet ein Protokoll-Set mit Referenz-Implementationen wie z.B. DOI, LoC.*
Syntax	*Handle ::= Handle Naming Authority "/" Handle Local Name* *Das Präfix ist ein numerischer Code, der die Institution bezeichnet. Das Suffix kann sich aus einer beliebigen Zeichenkette zusammensetzen.*
Beispiel	*Als URN: urn:handle:10.1045/january99-bearman*
Identifizierung / Registry	*Zentrales Handle-Registry für die Präfixe.*
Resolving	*Handle-Service*
Anwender	*DOI-Anwender, LoC, DSpace-Anwender*
Tool-Adaption	*DSpace*
Referenz	*http://www.handle.net*
Bemerkungen	*Handles sind URN-konform.*

InfoURI

Kurzbezeichnung	*InfoURI*
Erläuterung	*InfoURI ist ein Identifier für Ressourcen, die über kein Äquivalent innerhalb des URI-Raumes verfügen wie z.B. LCCN. Sie sind nur für die Identifizierung gedacht, nicht für die Auflösung. Es ist ein NISO-Standard.*
Syntax	*„info:“ namespace „/“ identifier [„#“ fragment]* *info-scheme = “info”* *info-identifier = namespace “/” identifier* *namespace = scheme* *identifier = path-segments*
Beispiel	*info:lccn/n78089035* *Als URN:* *urn:info:lccn/n78089035*
Identifizierung / Registry	*Zentrales Registry für Namespaces*
Resolving	*nein*
Anwender	*18 Anwender:* *LoC, OCLC, DOI etc.*
Tool-Adaption	*Entwicklung für die Adaption von OpenURL-Services*
Referenz	*http://info-uri.info/*
Bemerkungen	*Zusammenarbeit mit OpenURL.*

NLA - Australische Nationalbibliothek

Kurzbezeichnung	*Keine vorhanden, aber die Identifier beginnen mit NLA*
Erläuterung	
Syntax	*Abhängig von den einzelnen Typen elektronischen Materiales werden die Identifier nach verschiedenen Algorithmen gebildet.* *Beispiel* *Collection Identifier* *nla.pic, nla.ms, nla.map, nla.gen, nla.mus, nla.aus, nla.arc*
Beispiel	*Manuscript Material* *collection id-collection no.-series no.-item no.-sequence no.- role code-generation code* *nla.ms-ms8822-001-0001-001-m*
Identifizierung / Registry	*Objekte, die archiviert werden. Es existiert ein lokales Registry.*
Resolving	*Ja, für die lokalen Identifier*
Anwender	*ANL, Zweigstellen, Kooperationspartner*
Tool-Adaption	
Referenz	*http://www.nla.gov.au/initiatives/persistence.html*
Bemerkungen	*Dies ist eine Eigenentwicklung. Es werden keine internationalen Standards berücksichtigt.*

LSID - Life Science Identifier

Kurzbezeichnung	*LSID*
Erläuterung	*Die OMG (Object Management Group) spezifiziert LSID als Standard für ein Benennungsschema für biologische Entitäten innerhalb der "Life Science Domains" und die Notwendigkeit eines Resolving-Dienstes, der spezifiziert, wie auf die Entitäten zugegriffen werden kann.*

Syntax	*The LSID declaration consists of the following parts, separated by double colons:* • *"URN"* • *"LSID"* • *authority identification* • *namespace identification* • *object identification* • *optionally: revision identification.* *If revision field is omitted then the trailing colon is also omitted.*
Beispiel	*URN:LSID:ebi.ac.uk:SWISS-PROT. accession:P34355:3 URN:LSID:rcsb. org:PDB:1D4X:22 URN:LSID:ncbi.nlm.nih. gov:GenBank.accession:NT_001063:2*
Identifizierung / Registry	*s. URN*
Resolving	*DDDS/DNS, Web-Service*
Anwender	*undurchsichtig*
Tool-Adaption	
Referenz	*http://xml.coverpages.org/lsid.html* • *"OMG Life Sciences Identifiers Specification." - Main reference page.* • *Interoperable Informatics Infrastructure Consortium (I3C)* • *Life Sciences Identifiers. An OMG Final Adopted Specification which has been approved by the OMG board and technical plenaries. Document Reference: dtc/04-05-01. 40 pages.* • *LSID Resolution Protocol Project. Info from IBM.* • *"Identity and Interoperability in Bioinformatics." By Tim Clark (I3C Editorial Board Member). In Briefings in Bioinformatics (March 2003).* *"Build an LSID authority on Linux." By Stefan Atev (IBM)*
Bemerkungen	

POI - PURL-Based Object Identifier

Kurzbezeichnung	*POI*
Erläuterung	*POI ist eine einfache Spezifikation als Resource-Identifier auf Grundlage des PURL-Systems und ist als „oai-identifier“ für das OAI-PMH entwickelt worden. POIs dienen als Identifier für Ressourcen, die in den Metadaten von OAI-konformen Repositories beschrieben sind. POIs können auch explizit für Ressourcen verwendet werden.*
Syntax	*“http://purl.org/poi/”namespace-identifier “/” local-identifier* *namespace-identifier = domainname-word “.” domainname* *domainname = domainname-word [“.”domainname]* *domainname-word = alpha *(alphanum | “-”)* *local-identifier = 1*uric*
Beispiel	*http://www.ukoln.ac.uk/distributed-systems/poi/*
Identifizierung / Registry	*kein*
Resolving	*Ja, wenn dieser über das OAI-Repository bereitgestellt wird, wobei der PURL-Resolver empfohlen wird.*
Anwender	*unklar*
Tool-Adaption	*POI-Lookup-Tools* *http://www.rdn.ac.uk/poi/*
Referenz	*POI Resolver Guidelines* *http://www.ukoln.ac.uk/distributed-systems/poi/resolver-guidelines/* *“The PURL-based Object Identifier (POI).”* *By Andy Powell (UKOLN, University of Bath), Jeff Young (OCLC), and Thom Hickey (OCLC). 2003/05/03. http://www.ukoln.ac.uk/distributed-systems/poi/*
Bemerkungen	

PURL – Persistent URL

Kurzbezeichnung	*PURL*
Erläuterung	*PURL (Persistent URL) wurde vom „Online Computer Library Center“ (OCLC) 1995 im Rahmen des „Internet Cataloging Projects“, das durch das U.S. Department of Education finanziert wurde, eingeführt, um die Adressdarstellung für die Katalogisierung von Internetressourcen zu verbessern. PURLs sind keine Persistent-Identifier, können jedoch in bestehende Standards wie URN überführt werden. Technisch betrachtet wird bei PURL der existierende Internet-Standard „HTTP-redirect“ angewendet, um PURLs in die URLs aufzulösen.*
Syntax	*http://purl.oclc.org/docs/help.html* *- protocol* *- resolver address* *- name*
Beispiel	*http://purl.oclc.org/keith/home* *Als URN:* *urn:/org/oclc/purl/keith/home*
Identifizierung / Registry	*Kein Registry*
Resolving	*ja, jedoch wird nur ein lolaker Resolver installiert.*
Anwender	*Keine Auskunft möglich (lt. Stuart Weibel)* *- OCLC* *- United States Government Printing Office (GPO)* *- LoC*
Tool-Adaption	*PURL-Software*
Referenz	*http://purl.org*

Bemerkungen	*- kein zentrales Registry* *- Die genaue Anzahl von vergebenen PURLs ist unbekannt. ??* *- Ein Test der DOI-Foundation ergab, dass nur 57% der getesteten PURLs auflösbar waren.* *- Experimentell von OCLC eingeführt.* *- Es ist keine Weiterentwicklung vorgesehen.*

URN – Uniform Resource Name

Kurzbezeichnung	*URN*
Erläuterung	*Der Uniform Resource Name (URN) existiert seit 1992 und ist ein Standard zur Adressierung von Objekten, für die eine institutionelle Verpflichtung zur persistenten, standortunabhängigen Identifizierung der Ressourcen besteht. URNs wurden mit dem Ziel konzipiert, die Kosten für die Bereitstellung von Gateways sowie die Nutzung von URNs so gering wie möglich zu halten - vergleichbar mit existierenden Namensräumen wie z.B. URLs. Aus diesem Grund wurde in Standards festgelegt, wie bereits existierende oder angewendete Namensräume bzw. Nummernsysteme einfach in das URN-Schema sowie die gängigen Protokolle wie z.B. HTTP (Hypertext Transfer Protocol) oder Schemas wie z.B. URLs integriert werden können.* *Der URN als Standard wird von der Internet Engineering Task Force (IETF) kontrolliert, die organisatorisch in die Internet Assigned Numbering Authority (IANA) eingegliedert ist. Sie ist für die Erarbeitung und Veröffentlichung der entsprechenden Standards in Form von "Request for Comments" (RFCs) zuständig. Diese umfassen die folgenden Bereiche:* • *URN-Syntax (RFC 2141),* • *funktionale Anforderungen an URNs (RFC 1737),* • *Registrierung von URN-Namensräumen (z.B. RFCs 3406, 2288, 3187, NBN: 3188),* • *URN-Auflösungsverfahren (RFCs 3401, 3402, 3403, 3404).*

Syntax	*URN:NID:NISS* *URNs bestehen aus mehreren hierarchisch aufgebauten Teilbereichen. Dazu zählen der Namensraum (Namespace, NID), der sich aus mehreren untergeordneten Unternamensräumen (Subnamespaces, SNID) zusammensetzen kann, sowie der Namensraumbezeichner (Namespace Specific String, NISS).*
Beispiel	*urn:nbn:de:bsz:93-opus-59* *Als URL / URI:* *http://nbn-resolving.de/* *urn:nbn:de:bsz:93-opus-59* *Als OpenURL:* *http.//[openURL-service]?identifier=* *urn:nbn:de:bsz:93-opus-59* *Als InfoURI:* *info:urn/urn:nbn:de:bsz:93-opus-59* *Als ARK:* *http://[NMAH]ark:/NAAM/* *urn:nbn:de:bsz:93-opus-59* *Als DOI:* *10.1111/urn:nbn:de:bsz:93-opus-59*
Identifizierung / Registry	*Überblick über den Status registrierter URN-Namensräume (unvollständig)* *http://www.iana.org/assignments/* *urn-namespaces/*
Resolving	*Es gibt mehrere Möglichkeiten:* *- http-Redirect (Umleitung der URN zur URL)* *- DNS (Domain Name System)*

Anwender	*CLEI Code* *IETF* *IPTC* *ISAN* *ISBN* *ISSN* *NewsML* *OASIS* *OMA* *Resources* *XML.org* *Web3D* *MACE* *MPEG* *Universal Content Identifier* *TV-Anytime Forum* *Federated Content* *Government (NZ)* *Empfehlung: OAI 2.0: oai-identifier als URNs verwenden* *NBN:* *Finnland,* *Niederlande,* *Norwegen,* *Österreich,* *Portugal,* *Slovenien,* *Schweden,* *Schweiz,* *Tschechien,* *Ungarn,* *UK*
Tool-Adaption	*OPUS, DigiTool (ExLibris), Miless*

Referenzen	*Internetstandards:* *http://www.ietf.org/rfc/rfc1737.txt* *http://www.ietf.org/rfc/rfc2141.txt* *http://www.ietf.org/rfc/rfc3406.txt* *http://www.ietf.org/rfc/rfc288.txt* *http://www.ietf.org/rfc/rfc3187.txt* *http://www.ietf.org/rfc/rfc3188.txt* *http://www.ietf.org/rfc/rfc3401.txt* *http://www.ietf.org/rfc/rfc3402.txt* *http://www.ietf.org/rfc/rfc3403.txt* *http://www.ietf.org/rfc/rfc3404.txt* *URN-Prüfziffer Der Deutschen Bibliothek:* *http://www.pruefziffernberechnung.de/U/URN.shtml*
Bemerkungen	*Innerhalb der URNs sind sowohl die Integration bereits bestehender Nummernsysteme (z.B. ISBN) als auch institutionsgebundene Nummernsysteme auf regionaler oder internationaler Ebene als Namensräume möglich. Dazu zählt auch die „National Bibliography Number“ (NBN, RFC 3188), ein international verwalteter Namensraum der Nationalbibliotheken, an dem Die Deutsche Bibliothek beteiligt ist.*

XRI - Extensible Resource Identifier

Kurzbezeichnung	*XRI*
Erläuterung	*XRI wurde vom TC OASIS entwickelt. XRI erweitert die generische URI-Syntax, um "extensible, location-, application-, and transport-independent identification that provides addressability not just of resources, but also of their attributes and versions." zu gewährleisten. Segmente oder Ressourcen können persistent identifiziert und/oder zu adressiert werden. Die Persistenz des Identifiers wird mit den Zielen der URNs gleichgestellt.*
Syntax	*xri: authority / path ? query # fragment*
Beispiel	*xri://@example.org*agency*department/ docs/govdoc.pdf* *XRI mit URN:* *xri://@example.bookstore/!(ur n:ISBN:0-395-36341-1)*
Identifizierung / Registry	*nein*
Resolving	*OpenXRI.org server*
Anwender	*12 Förderer* *http://www.openxri.org/*
Tool-Adaption	

Referenz	*http://www.openxri.org/* • *"OASIS Releases Extensible Resource Identifier (XRI) Specification for Review." News story 2005-04-07.* • *XRI Generic Syntax and Resolution Specification 1.0 Approved Committee Draft. PDF source posted by Drummond Reed (Cordance), Tuesday, 20 January 2004, 03:00pm.* • *XRI Requirements and Glossary Version 1.0. 12-June-2003. 28 pages. [source .DOC, cache]* • *OASIS Extensible Resource Identifier TC web site* • *XRI TC Charter* • *"OASIS TC Promotes Extensible Resource Identifier (XRI) Specification." News story 2004-01-19. See also "OASIS Members Form XRI Data Interchange (XDI) Technical Committee."*
Bemerkungen	

Referenzen

Beschreibung	Referenz
Überblicksdarstellung von PI-Systemen des EPICUR-Projektes	http://www.persistent-identifier.de/?link=204
PADI – Preserving Access to Digital Information	http://www.nla.gov.au/padi/topics/36.html
nestor-Informationsdatenbank, Themenschwerpunkt: Persistente Identifikatoren	http://nestor.sub.uni-goettingen.de/nestor_on/browse.php?show=21
ERPANET Workshop „Persistent Identifier“, 2004	http://www.erpanet.org/events/2004/cork/index.php

9.4.1 Der Uniform Resource Name (URN)

Christa Schöning-Walter

Damit digitale Objekte auf Dauer zitierfähig sind, müssen stabile Referenzen vorhanden sein, die auch über technische und organisatorische Veränderungen hinweg eine verlässliche Adressierung ermöglichen. Grundlegend für den dauerhaften Zugang ist die Langzeitverfügbarkeit der digitalen Objekte an sich. Die Speicherung in vertrauenswürdigen Archiven ist dafür eine unabdingbare Voraussetzung. Persistent Identifier *haben in diesem Kontext die zentrale Funktion, die Objekte digitaler Sammlungen langfristig und weltweit eindeutig zu identifizieren.*

Sammlung und Langzeitarchivierung von Netzpublikationen bei der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek (DNB) hat den Auftrag, das kulturelle und wissenschaftliche Erbe Deutschlands in seiner seit 1913 veröffentlichten Form zu sammeln, dauerhaft zu bewahren und für die Nutzung zugänglich zu machen. Seit dem Inkrafttreten des *Gesetzes über die Deutsche Nationalbibliothek* vom 22. Juni 2006 gehören auch Netzpublikationen zum Sammelauftrag.[19] Als Netzpublikationen gelten jegliche Darstellungen in Schrift, Bild oder Ton, die in öffentlichen Netzen bereitgestellt werden. Elektronische Zeitschriften, E-Books, Hochschulprüfungsarbeiten, Forschungsberichte, Kongressschriften und Lehrmaterialien gehören ebenso dazu wie Digitalisate alter Drucke, Musikdateien oder Webseiten. Am 17. Oktober 2008 ist zudem die Pflichtablieferungsverordnung neu gefasst worden. Sie konkretisiert den gesetzlichen Sammelauftrag der DNB.

Um die Benutzbarkeit ihrer digitalen Sammlungen auch in Zukunft gewährleisten zu können, engagiert sich die DNB auf dem Gebiet der Langzeitarchivierung. Grundlagen für die Langzeiterhaltung und Langzeitbenutzbarkeit digitaler Objekte sind in dem vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekt „kopal – Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen“ entwickelt worden (*http://kopal.langzeitarchivierung.de*). Fortlaufende Anpassungen der Datenbestände an den Stand der Technik (Migrationen, Emulationen) sollen dafür sorgen, dass digitale Samm-

19 http://www.d-nb.de/netzpub/index.htm

lungen trotz Weiterentwicklungen bei den Hard- und Softwaresystemen verfügbar bleiben.

Im Zuge der Langzeitarchivierung muss neben dem Erhalt der digitalen Daten an sich und ihrer Interpretierbarkeit auch die Identifizierbarkeit der Objekte sichergestellt werden. Eindeutige Bezeichner, die über den gesamten Lebenszyklus hinweg mit den Netzpublikationen und ihren Metadaten verbunden bleiben, ermöglichen es, die Objekte in digitalen Sammlungen persistent (dauerhaft) zu identifizieren – d.h. sie dem Nutzer auch über Systemgrenzen und Systemwechsel hinweg verlässlich zur Verfügung zu stellen.

Die DNB verwendet als Schema für die Identifizierung digitaler Ressourcen den *Uniform Resource Name* (URN). Das Konzept zur Langzeitarchivierung beinhaltet, dass alle Netzpublikationen, die bei der DNB gesammelt, erschlossen und archiviert werden, zwingend einen Persistent Identifier benötigen.[20] Die Zuordnung erfolgt spätestens nach dem Erhalt einer Netzpublikation durch die DNB. Der Service zur Registrierung und Auflösung von URNs ist ein kooperativer Dienst, der von den verlegenden Institutionen mit genutzt werden kann. Idealerweise erfolgt die URN-Vergabe schon im Zuge der Veröffentlichung einer Netzpublikation. Dann ist es möglich, alle Speicherorte von vornherein mit in das System aufzunehmen.

Das URN-Schema

Die Wurzeln des URN reichen zurück bis in die frühen 1990er Jahre. Das Funktionsschema gehört zu den Basiskonzepten, die im Zusammenhang mit dem Entwurf einer Architektur für das *World Wide Web* (WWW) entstanden sind. Der URN ist ein *Uniform Resource Identifier* (URI). URIs werden im globalen Informations- und Kommunikationsnetz für die Identifizierung jeglicher zu adressierender physikalischer oder abstrakter Ressourcen benutzt (z.B. für den Zugriff auf Objekte, den Aufruf von Webservices, die Zustellung von Nachrichten etc.).[21] Das *World Wide Web Consortium* (W3C) betont die besondere Bedeutung dieser Technologie:

> *The Web is an information space. [...] URIs are the points in that space. Unlike web data formats [...] and web protocols [...] there is only one Web naming/addressing technology: URIs.*[22]

20 Schöning-Walter (2002)

21 Tim Berners-Lee (2005): Uniform Resource Identifier (URI) – Generic Syntax. http://www.ietf.org/rfc/rfc3986.txt

22 W3C: Web Naming and Addressing Overview. http://www.w3.org/Addressing/.

Das URN-Schema hat den Status eines durch die *Internet Assigned Numbers Authority* (IANA) legitimierten de facto-Internetstandards.[23] Beschrieben ist das Schema in verschiedenen so genannten *Requests for Comments* (RFCs), wie sie üblicherweise von den Arbeitsgruppen der *Internet Engineering Task Force* (IETF) veröffentlicht werden.

URNs sind als weltweit gültige, eindeutige Namen für Informationsressourcen im WWW konzipiert worden. Die Entwicklung zielte darauf, Unabhängigkeit vom Ort der Speicherung und vom Zugriffsprotokoll zu erreichen (*RFC 2141*, URN Syntax).[24]

> *Uniform Resource Names (URNs) are intended to serve as persistent, location-independent resource identifiers and are designed to make it easy to map other namespaces (that share the properties of URNs) into URN-space. Therefore, the URN syntax provides a means to encode character data in a form that can be sent in existing protocols, transcribed on most keyboards, etc.*

Das Schema fächert sich in sogenannte Namensräume auf, die ebenfalls bei IANA registriert werden müssen.[25] Mit der Verzeichnung eines Namensraums werden sowohl der Geltungsbereich (welche Art von digitalen Objekten soll identifiziert werden) als auch spezifische Regeln festgelegt (*RFC 3406*, URN Namespace Definition Mechanisms).[26] Zu den bisher bei IANA angemeldeten URN-Namensräumen gehören u.a. (Stand: 15. Februar 2009):

- urn:issn – International Serials Number (*RFC 3044*),
- urn:isbn – International Standards Books Number (*RFC 3187*),
- urn:nbn – National Bibliography Number (*RFC 3188*),
- urn:uuid – Universally Unique Identifiers (*RFC 4122;* für verteilte Softwaresysteme),
- urn:isan – International Standard Audiovisual Number (*RFC 4246*).

RFC 1737 (Functional Requirements for URNs) beschreibt die Anforderungen, die grundsätzlich jeder URN-Namensraum erfüllen muss.[27] Das sind:

---

23 http://www.ietf.org/rfc.html

24 Ryan Moats (1997): URN Syntax. http://www.ietf.org/rfc/rfc2141.txt

25 Die bei IANA registrierten URN-Namensräume sind verzeichnet unter http//www.iana.org/assignments/urn-namespaces.

26 Leslie L. Daigle et. al. (2002): URN Namespace Definition Mechanisms. http://www.ietf.org/rfc/rfc3406.txt.

27 Larry Mainter, Karen Sollins (1994): Functional Requirements for Uniform Resource Names. http://www.ietf.org/rfc/rfc1737.txt.

- *Global Scope*: Gültigkeit der Namen weltweit,
- *Global Uniqueness*: Eindeutigkeit der Namen weltweit,
- *Persistence*: Gültigkeit der Namen auf Dauer,
- *Scalability*: das Namensschema muss beliebig viele Objekte bezeichnen können,
- *Legacy Support*: schon vorhandene Bezeichnungs- und Nummerierungssysteme, Zugriffsschemata oder Protokolle müssen eingebettet werden können, sofern sie regelkonform sind,
- *Extensibility*: das Namensschema muss bei Bedarf weiterentwickelt werden können,
- *Indepencence*: die beteiligten Institutionen selbst legen die Regeln für das Namensschema fest,
- *Resolution*: die Auflösung von URNs in Zugriffsadressen erfolgt über einen Resolvingdienst.

URNs verweisen nicht selbst auf die Informationsressourcen. Ein zwischengeschalteter Resolvingmechanismus führt die Auflösung durch (*RFC 2276*, Architectural Principles of URN Resolution).[28] In der Regel werden URNs über ein Register in URLs umgewandelt. Dieses Prinzip ermöglicht es, den Aufwand für die Pflege von Zugriffsadressen relativ gering zu halten.

Die National Bibliography Number

Die *National Bibliography Number* (NBN) ist ein registrierter URN-Namensraum mit maßgeblicher Bedeutung für den Bibliotheks- und Archivbereich.[29] Das Konzept beruht auf einer Initiative der *Conference of European National Librarians* (CENL) und wurde im Jahr 2001 unter Federführung der Finnischen Nationalbibliothek entwickelt, um digitale Publikationen in den Nationalbibliografien verzeichnen zu können (*RFC 3188*, Using National Bibliography Numbers as URNs).[30]

Die NBN ist international gültig. Auf nationaler Ebene sind üblicherweise die Nationalbibliotheken für den Namensraum verantwortlich. In Deutsch-

28 Karen Sollins (1998): Architectural Principles of Uniform Resource Name Resolution. http://www.ietf.org/rfc/rfc2276.txt.

29 Hans-Werner Hilse, Jochen Kothe (2006): Implementing Persistent Identifiers. Overview of concepts, guidelines and recommendations. London: Consortium of European Research Libraries. Amsterdam: European Commission on Preservation and Access. urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

30 Juha Hakala (1998): Using National Bibliography Numbers as Uniform Resource Names. http://www.ietf.org/rfc/rfc3188.txt.

land hat die DNB die Koordinierungsfunktion übernommen. In den Jahren 2002 bis 2005 wurde im Rahmen des BMBF-Projekts „EPICUR – Enhancements of Persistent Identifier Services" eine Infrastruktur aufgebaut (*http://www.persistent-identifier.de*).[31] Die DNB betreibt seither einen URN-Resolver für Deutschland, Österreich und die Schweiz. Beim Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) existiert dazu ein Spiegelsystem, das bei Bedarf den Resolvingservice übernehmen kann.

Die Entwicklungen in EPICUR waren vorrangig auf die Verzeichnung von Hochschulschriften ausgerichtet. In Zusammenarbeit mit Hochschulbibliotheken, Verbundzentralen und Forschungseinrichtungen wurden URN-Registrierungsverfahren entwickelt, die mittlerweile stark vereinheitlicht und in der Praxis recht gut etabliert sind. Die Regeln zur Vergabe von URNs für Hochschulschriften sind in der *URN-Strategie der Deutschen Nationalbibliothek* beschrieben.[32] Auch die Schweizerische Nationalbibliothek (NB) nimmt seit mehreren Jahren aktiv am URN-Verfahren teil. Seit August 2008 gibt es ein Handbuch zur Anwendung der NBN in der Schweiz.[33]

Die DNB hat die Verwendung des NBN-Schemas mittlerweile auf alle Netzpublikationen ausgedehnt, die im Rahmen ihres erweiterten Sammelauftrags erschlossen werden. Dies hat zur Folge, dass auch fortlaufende Sammelwerke (z.B. Zeitschriften oder Schriftenreihen), granulare Erscheinungsformen (z.B. die Beiträge in einer Zeitschrift oder die Einzelseiten eines Digitalisats) sowie dynamische Publikationsformen (z.B. veränderliche Webseiten) schrittweise mit in die URN-Strategie eingebunden werden müssen. Jedes digitale Objekt, das einzeln identifizierbar und adressierbar sein soll, muss als inhaltlich stabile, eigenständige Einheit in das Langzeitarchiv aufgenommen werden und benötigt für den Zugang einen eigenen Persistent Identifier.

Die Erschließung elektronischer Zeitschriften bei der DNB wird gegenwärtig dahingehend neu ausgerichtet, in der Zukunft auch Fachartikel in der Nationalbibliografie zu verzeichnen und mit einem URN zu kennzeichnen. Im Förderprogramm *Kulturelle Überlieferung* der Deutschen Forschungsgemeinschaft (DFG) wird sogar eine persistente Identifizierung bis auf die Ebene der Einzelseiten digitalisierter Drucke gefordert.[34] Vor diesem Hintergrund erprobt die

---

31 Kathrin Schroeder (2005): EPICUR. In: Dialog mit Bibliotheken. 2005/1. S. 58-61.

32 EPICUR: Uniform Resource Names (URN) – Strategie der Deutschen Nationalbibliothek (2006). http://www.persistent-identifier.de/?link=3352. urn:nbn:de:1111-200606299.

33 e-Helvetica URN-Handbuch. August 2008. Version 1.0. http://www.nb.admin.ch/slb/slb_professionnel/01693/01695/01706/index.html?lang=de.

34 Deutsche Forschungsgemeinschaft: Praxisregeln im Förderprogramm *Kulturelle Überlieferung*. Kap. 5.7. Zitieren, persistente Adressierung. http://www.dfg.de/forschungsfoerderung/

ULB Sachsen-Anhalt in einem Modellprojekt die Vergabe von NBNs für die digitalisierten Drucke aus der Sammlung Ponickau (es handelt sich um ca. 10.000 Drucke mit insgesamt ca. 600.000 Seiten).[35]

Die Struktur der NBN

Das URN-Schema ist streng hierarchisch aufgebaut und gliedert sich in ein Präfix und ein Suffix. Wie andere URI-Schemata auch (z.B. *http:*, *ftp:*, *mailto:*) wird der URN durch seine Bezeichnung gekennzeichnet, gefolgt von einem Doppelpunkt. Der grundsätzliche Aufbau eines URN lautet:

urn:[NID]:[SNID]-[NISS]

Präfix:

- NID *Namespace Identifier (hier: nbn)*
- SNID *Subnamespace Identifier*

Suffix:

- NISS *Namespace Specific String*

Durch Gliederung in Unternamensräume (Subnamespaces) kann die auf internationaler Ebene eingeleitete hierarchische Strukturierung auf nationaler Ebene weiter fortgesetzt werden. Ein zentrales Strukturelement ist das Länderkennzeichen. Ein URN, der mit urn:nbn:de beginnt, drückt aus, dass es sich um eine NBN für eine in Deutschland veröffentlichte Publikation handelt (urn:nbn:ch gilt für die Schweiz, urn:nbn:at für Österreich), die über den URN-Resolver bei der DNB aufgelöst werden kann.

Die Option zur Gliederung der NBN in Unternamensräume wird genutzt, um interessierten Institutionen die Möglichkeit einzuräumen, selbst die Persistent Identifier für ihre Netzpublikationen zu vergeben. Für Deutschland, Österreich und die Schweiz erfolgt die Registrierung von Unternamensräumen bei der DNB. Die Bezeichnung muss eindeutig sein. Als Kennzeichen für Unternamensräume (Subnamespace Identifier) können verwendet werden:

---

formulare/download/12_151.pdf.

35 Dorothea Sommer, Christa Schöning-Walter, Kay Heiligenhaus (2008): URN Granular: Persistente Identifizierung und Adressierung von Einzelseiten digitalisierter Drucke. Ein Projekt der Deutschen Nationalbibliothek und der Universitäts- und Landesbibliothek Sachsen-Anhalt. In: ABI-Technik. Heft 2/2008. S. 106-114.

- das Bibliothekssigel oder die ISIL (*International Standard Identifier for Libraries and Related Organizations*),[36] ggf. kombiniert mit dem Kürzel des Bibliotheksverbundes,
- eine 4-stellige (fortlaufende) Nummer oder
- eine alphanumerische Zeichenkette.

Die Möglichkeiten der Differenzierung eines Namensraums sind vielfältig. Das Präfix in seiner Gesamtheit hat letztlich die Funktion, den Anwendungsbereich zu spezifizieren (Wofür wird der Name benutzt? Wer ist verantwortlich?) und ist definiert als derjenige Teil des URN, der im Resolver als fester Bestandteil des Namens verzeichnet ist. Beispiele zulässiger Namensräume sind:

- urn:nbn:de:101[37]
- urn:nbn:de:0008[38]
- urn:nbn:de:0100[39]
- urn:nbn:de:tuda[40]
- urn:nbn:de:gbv:3[41]
- urn:nbn:de:gbv:3:1[42]

Demgegenüber kennzeichnet das Suffix das digitale Objekt an sich (NISS, Namespace Specific String). Die Regeln für die Bildung des Suffix werden von der Institution festgelegt, die den Unternamensraum besitzt. Bereits existierende Nummerierungssysteme können eingebettet werden, wenn sie mit den Konventionen des URN-Schemas übereinstimmen.

Erlaubte Zeichen für die Bildung einer NBN sind alle alphanumerischen Zeichen und zusätzlich einige Sonderzeichen. Nach den in EPICUR festgelegten Regeln ist die letzte Ziffer immer eine automatisch berechnete Prüfziffer.[43] Beispiele zulässiger Namen sind:

36 http://sigel.staatsbibliothek-berlin.de/isil.html
37 Namensraum der DNB
38 Namensraum des nestor-Projekts
39 Namensraum des Suhrkamp-Verlags
40 Namensraum der TU Darmstadt
41 Namensraum der ULB Sachsen-Anhalt
42 Namensraum der ULB Sachsen-Anhalt für die Digitalisate der Sammlung Ponickau
43 EPICUR: Beschreibung des Algorithmus zur Berechnung der Prüfziffer. http://www.persistent-identifier.de/?link=316

- urn:nbn:ch:bel-110142[44]
- urn:nbn:de:bsz:16-opus-88271[45]
- urn:nbn:de:gbv:7-isbn-90-6984-508-3-8[46]
- urn:nbn:de:tib-10.1594/WDCC/EH4106_6H_1CO2IS92A_U306[47]

Registrierung und Bekanntmachung der NBN

Die NBN wird aktiv, sobald der URN und mindestens ein gültiger URL im Resolver verzeichnet sind. Idealerweise ist der URN einer Netzpublikation bereits auflösbar, sobald er erstmals bekannt gemacht wird (z.B. durch Verzeichnung auf einer Webseite oder in einem Online-Katalog).
Für die Übermittlung von URNs und URLs an das Resolvingsystem stehen verschiedene Transferschnittstellen und ein standardisiertes Datenaustauschformat zur Verfügung. Mögliche Registrierungsverfahren sind:

- OAI-Harvesting: Das Verfahren eignet sich besonders für Massendaten. Grundlage bildet das im Projekt EPICUR entwickelte Datenaustauschformat xepicur.[48] Neue Daten werden täglich eingesammelt, teilweise sogar mehrmals täglich.
- Mailverfahren: Die automatische Registrierung kann auch in Form einer elektronischen Nachricht erfolgen, indem eine xepicur-Datei als Dateianhang an den URN-Dienst gesendet wird.
- EPICUR-Frontend: Das Webinterface ermöglicht die manuelle Erfassung einzelner URNs.[49]

Mit Ablieferung einer Netzpublikation bei der DNB wird der URN in die Nationalbibliografie und den Online-Katalog übernommen und anschließend über die Datendienste weiter verbreitet. Die verlegende Stelle ihrerseits kann den URN durch Einbettung in die Publikation oder Verzeichnung auf einer vorgeschalteten Webseite bekannt machen – und damit die Nutzung des Persistent Identifier unterstützen.

---

44 Heft einer Verlagszeitschrift, URN vergeben von der Schweizerischen Nationalbibliothek
45 Dissertation, URN vergeben von der Universität Heidelberg
46 Forschungsbericht, URN vergeben von der SUB Göttingen
47 Forschungsdatensatz des World Data Center for Climate (WDCC), URN vergeben von der DOI-Registrierungsagentur bei der TIB Hannover
48 xepicur - XML-Datentransferformat zur Verwaltung von Persistent Identifiers. http://www.persistent-identifier.de/?link=210.
49 Persistent Identifier – Management der Deutschen Nationalbibliothek. https://ssl.nbn-resolving.de/frontend/

Für sammelpflichtige Netzpublikationen erfüllt die DNB die Aufgabe der Langzeitarchivierung und ergänzt in der URN-Datenbank die Archivadresse, sobald die Publikation in ihren Geschäftsgang gelangt. Noch nicht registrierte URNs werden bei der Erschließung nachträglich verzeichnet. Besitzt eine Netzpublikation keinen URN, dann übernimmt die DNB die Zuordnung dieses eindeutigen Namens, um die Identifizierbarkeit aller digitalen Objekte im Langzeitarchiv gewährleisten zu können.

Auflösung der NBN

Der URN-Resolver sorgt für die Auflösung der Namen in Zugriffsadressen. Ein URN muss dafür mit der Basisadresse des Resolvers (*http://nbn-resolving.de*) verknüpft werden:

- http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8 oder
- http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

Diese Funktionalität ist in der Regel direkt eingebettet in die Systeme, die den URN entweder anzeigen oder ihn für den Zugriff auf ein digitales Objekt benutzen (z.B. Dokumentenserver, Online-Kataloge etc.). Für einen individuellen Zugriff kann die EPICUR-Webseite benutzt werden.[50] Die handelsüblichen Browser unterstützen die Auflösung des URN bisher nicht.

Abhängig von den benutzten Parametern wird entweder der direkte Zugang zum digitalen Objekt hergestellt oder der Resolver gibt die Liste aller registrierten URLs zurück. Bei Vorhandensein mehrerer URLs existiert ein Standardverhalten des Resolvers. Vorrangig wird der URL mit der höchsten Priorität aufgelöst. Falls dieser URL nicht erreichbar ist, wird der URL mit der nächsten Priorität benutzt. Durch Pflege des URN (Aktualisierung der Zugriffsadressen bei Veränderungen) lässt sich erreichen, dass die Verknüpfung mit dem Original erhalten bleibt, bis die Netzpublikation vor Ort (auf dem Hochschulschriftenserver, auf dem Verlagsserver, im Institutional Repository etc.) nicht mehr verfügbar ist. Die Archivversion bei der DNB erlangt erst dann Bedeutung für den Zugang, wenn andere Zugriffsmöglichkeiten nicht mehr funktionieren.

50 http://www.persistent-identifier.de/?link=610

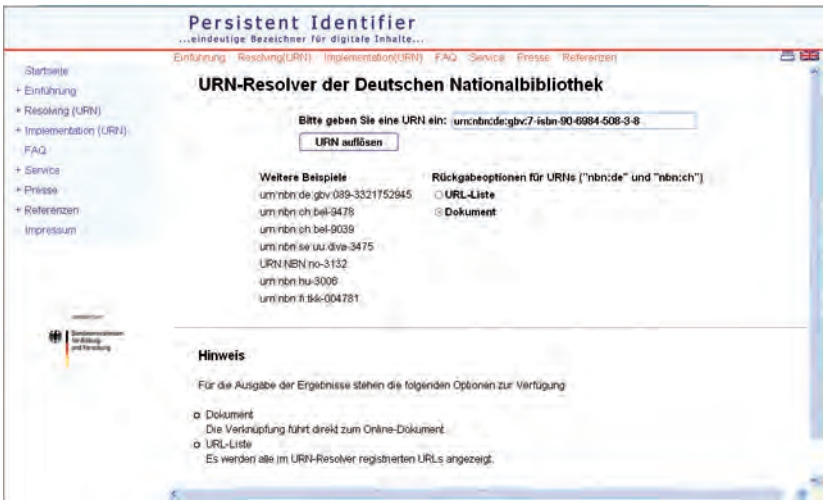

Abb.: URN-Auflösung über die EPICUR-Webseite

Zusammenfassung

Durch Nutzung von Persistent Identifiern lassen sich die Nachteile einer standortbezogenen Identifizierung und Adressierung digitaler Objekte weitgehend überwinden. Während die genaue Speicheradresse (URL) meistens nicht auf Dauer benutzbar ist, behalten URN-basierte Referenzen in Online-Katalogen, Bibliografien, Portalen oder Publikationen auch dann ihre Gültigkeit, wenn sich der Speicherort einer Netzpublikation verändert. Das hat auch Vorteile für die Zitierfähigkeit digitaler Quellen in der Praxis des wissenschaftlichen Arbeitens.
Persistenz ist keine Eigenschaft der URNs an sich. Sie erfordert abgestimmte Regeln sowie eine Pflege der Daten im Resolvingsystem. Die Infrastruktur muss in der Lage sein, die URNs solange nachzuweisen und aufzulösen, wie die Netzpublikationen selbst oder Referenzen darauf irgendwo existieren. Sind diese Voraussetzungen erfüllt, dann kann im Zusammenspiel mit der Langzeitarchivierung auch für die Objekte digitaler Sammlungen über lange Zeiträume hinweg der Zugang gewährleistet werden.

Quellenangaben

Berners-Lee, Tim (2005): *Uniform Resource Identifier (URI) – Generic Syntax*. http://www.ietf.org/rfc/rfc3986.txt

Daigle, Leslie L et. al. (2002): *URN Namespace Definition Mechanisms*. http://www.ietf.org/rfc/rfc3406.txt

e-Helvetica *URN-Handbuch*. (2008). Version 1.0. http://www.nb.admin.ch/slb/slb_professionnel/01693/01695/01706/index.html?lang=de

EPICUR: *Uniform Resource Names (URN) – Strategie der Deutschen Nationalbibliothek* (2006). http://www.persistent-identifier.de/?link=3352. urn:nbn:de:1111-200606299.

Hakala, Juha (1998): *Using National Bibliography Numbers as Uniform Resource Names*. http://www.ietf.org/rfc/rfc3188.txt.

Hilse, Hans-Werner; Kothe, Jochen (2006): *Implementing Persistent Identifiers. Overview of concepts, guidelines and recommendations. London: Consortium of European Research Libraries*. Amsterdam: European Commission on Preservation and Access. urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

Mainter, Larry ; Sollins, Karen (1994): *Functional Requirements for Uniform Resource Names*. http://www.ietf.org/rfc/rfc1737.txt

Moats, Ryan (1997): URN Syntax. http://www.ietf.org/rfc/rfc2141.txt

Schöning-Walter; Christa (2008): *Persistent Identifier für Netzpublikationen*. In: Dialog mit Bibliotheken. 2008/1. S. 32-38.

Schroeder, Kathrin (2005): *EPICUR*. In: Dialog mit Bibliotheken. 2005/1. S. 58-61.

Sollins, Karen (1998): *Architectural Principles of Uniform Resource Name Resolution*. http://www.ietf.org/rfc/rfc2276.txt

Sommer, Dorothea; Schöning-Walter, Christa; Heiligenhaus, Kay (2008): *URN Granular*: Persistente Identifizierung und Adressierung von Einzelseiten digitalisierter Drucke. Ein Projekt der Deutschen Nationalbibliothek und der Universitäts- und Landesbibliothek Sachsen-Anhalt. In: ABI-Technik. Heft 2/2008. S. 106-114.

9.4.2 Der Digital Objekt Identifier (DOI)

Jan Brase

Der Digital Object Identifier (DOI)

Der Digital Object Identifier (DOI) wurde 1997 eingeführt, um Einheiten geistigen Eigentums in einer interoperativen digitalen Umgebung eindeutig zu identifizieren, zu beschreiben und zu verwalten. Verwaltet wird das DOI-System durch die 1998 gegründete International DOI Foundation (IDF).[51]

Der DOI-Name ist ein dauerhafter persistenter Identifier, der zur Zitierung und Verlinkung von elektronischen Ressourcen (Texte, aber Forschungsdaten oder andere Inhalte) verwendet wird. Über den DOI-Namen sind einer Ressource aktuelle und strukturierte Metadaten zugeordnet.

Ein DOI-Name unterscheidet sich von anderen, gewöhnlich im Internet verwendeten Verweissystemen wie der URL, weil er dauerhaft mit der Ressource als Entität verknüpft ist und nicht lediglich mit dem Ort, an dem die Ressource platziert ist.

Der DOI-Name identifiziert eine Entität direkt und unmittelbar, also nicht eine Eigenschaft des Objekts (eine Adresse ist lediglich eine Eigenschaft des Objekts, die verändert werden und dann ggf. nicht mehr zur Identifikation des Objekts herangezogen werden kann).

Das IDF-System besteht aus der „International DOI Foundation“ selbst, der eine Reihe von Registrierungsagenturen („Registration Agencies“; RA) zugeordnet sind. Für die Aufgaben einer RA können sich beliebige kommerzielle oder nicht kommerzielle Organisationen bewerben, die ein definiertes Interesse einer Gemeinschaft vorweisen können, digitale Objekte zu referenzieren.

Technik

Das DOI-System baut technisch auf dem Handle-System auf. Das Handle System wurde seit 1994 von der US-amerikanischen Corporation for National Research Initiatives (CNRI)[52] als verteiltes System für den Informationsaustausch entwickelt. Handles setzen direkt auf das IP-Protokoll auf und sind eingebettet in ein vollständiges technisches Verwaltungsprotokoll mit festgelegter Prüfung der Authentizität der Benutzer und ihrer Autorisierung. Durch das Handle-Sy-

51 http://www.doi.org/

52 http://www.cnri.reston.va.us/ bzw. http://www.handle.net

stem wird ein Protokoll zur Datenpflege und zur Abfrage der mit dem Handle verknüpften Informationen definiert. Diese Informationen können beliebige Metadaten sein, der Regelfall ist aber, dass die URL des Objektes abgefragt wird, zu dem das Handle registriert wurde. Weiterhin stellt CNRI auch kostenlos Software zur Verfügung, die dieses definierte Protokoll auf einem Server implementiert (und der damit zum sog. Handle-Server wird).

Ein DOI-Name besteht genau wie ein Handle immer aus einem Präfix und einem Suffix, wobei beide durch einen Schrägstrich getrennt sind und das Präfix eines DOI-Namens immer mit „10." Beginnt. Beispiele für DOI-Namen sind:

doi:10.1038/35057062
doi:10.1594/WDCC/CCSRNIES_SRES_B2

Die Auflösung eines DOI-Namens erfolgt nun über einen der oben erwähnten Handle-Server. Dabei sind in jedem Handle-Server weltweit sämtliche DOI-Namen auflösbar. Dieser große Vorteil gegenüber anderen PI-Systemen ergibt sich einerseits durch die eindeutige Zuordnung eines DOI-Präfix an den Handle-Server, mit dem dieser DOI-Name registriert wird und andererseits durch die Existenz eines zentralen Servers bei der CNRI, der zu jedem DOI-Präfix die IP des passenden Handle-Servers registriert hat. Erhält nun ein Handle-Server irgendwo im Netz den Auftrag einen DOI-Namen aufzulösen, fragt er den zentralen Server bei der CNRI nach der IP-Adresse des Handle-Servers, der den DOI-Namen registriert hat und erhält von diesem die geforderte URL.

DOI-Modell

Die Vergabe von DOI-Namen erfolgt wie oben erwähnt nur durch die DOI-Registrierungsagenturen, die eine Lizenz von der IDF erwerben. Dadurch wird sichergestellt, dass jeder registrierte DOI-Name sich an die von der IDF vorgegebenen Standards hält. Diese Standards sind als Committee Draft der ISO Working Group TC46 SC9 WG7 (Project 26324 Digital Object Identifier system) veröffentlicht und sollen ein anerkannter ISO Standard werden. Zum Stand 02/09 gibt es 6 DOI-Registrierungsagenturen, die teilweise kommerzielle, teilweise nicht-kommerzielle Ziele verfolgen. Bei den Agenturen handelt es sich um

- CrossRef[53], mEDRA[54] und R.R. Bowker[55] als Vertreter des

53 http://www.crossref.org/
54 http://www.medra.org/
55 http://www.bowker.com/

Verlagswesens,

• Wanfang Data Co., Ltd[56] als Agentur für den Chinesischen Markt,

• OPOCE (Office des publications EU)[57], dem Verlag der EU, der alle offiziellen Dokumente der EU registriert

• TIB/DataCite[58] als nicht-kommerzielle Agentur für Forschungsdaten und wissenschaftliche Information

Dieses Lizenz-Modell wird häufig gleichgesetzt mit einer kommerziellen Ausrichtung des DOI-Systems, doch steht es jeder Registrierungsagentur frei, in welcher Höhe sie Geld für die Vergabe von DOI-Namen verlangt. Auch muss berücksichtigt werden, dass – anders als bei allen anderen PI-Systemen – nach der Vergabe von DOI-Namen durch die Verwendung des Handle-Systems für das Resolving- bzw. für die Registrierungs-Infrastruktur keine weiteren Kosten entstehen.

Die TIB als DOI Registrierungsagentur für Forschungsdaten

Der Zugang zu wissenschaftlichen Forschungsdaten ist eine grundlegende Voraussetzung für die Forschungsarbeit vor allem in den Naturwissenschaften. Deshalb ist es notwendig, bestehende und zum Teil auch neu aufkommende Einschränkungen bei der Datenverfügbarkeit zu vermindern.
Traditionell sind Forschungsdaten eingebettet in einen singulären Forschungsprozess, ausgeführt von einer definierten Gruppe von Forschern, geprägt von einer linearen Wertschöpfungskette:

Experiment ⇒ Forschungsdaten ⇒ Sekundärdaten ⇒ Publikation
Akkumulation Datenanalyse Peer-Review

56 http://www.wanfangdata.com/
57 http://www.publications.eu.int/
58 http://www.datacite.org

Durch die Möglichkeiten der neuen Technologien und des Internets können einzelne Bestandteile des Forschungszyklus in separate Aktivitäten aufgeteilt werden (Daten-Sammlung, Daten-Auswertung, Daten-Speicherung, usw.) die von verschiedenen Einrichtungen oder Forschungsgruppen durchgeführt werden können. Die Einführung eines begleitenden Archivs und die Referenzierung einzelner Wissenschaftlicher Inhalte durch persistente Identifier wie einen DOI-Namen schafft die Möglichkeit anstelle eines linearen Forschungsansatzes, den Wissenschaftlerarbeitsplatz einzubinden in einen idealen Zyklus der Information und des Wissens (siehe Abbildung 1), in dem durch Zentrale Datenarchive als Datenmanager Mehrwerte geschaffen werden können und so für alle Datennutzer, aber auch für die Datenautoren selber ein neuer Zugang zu Wissen gestaltet wird.

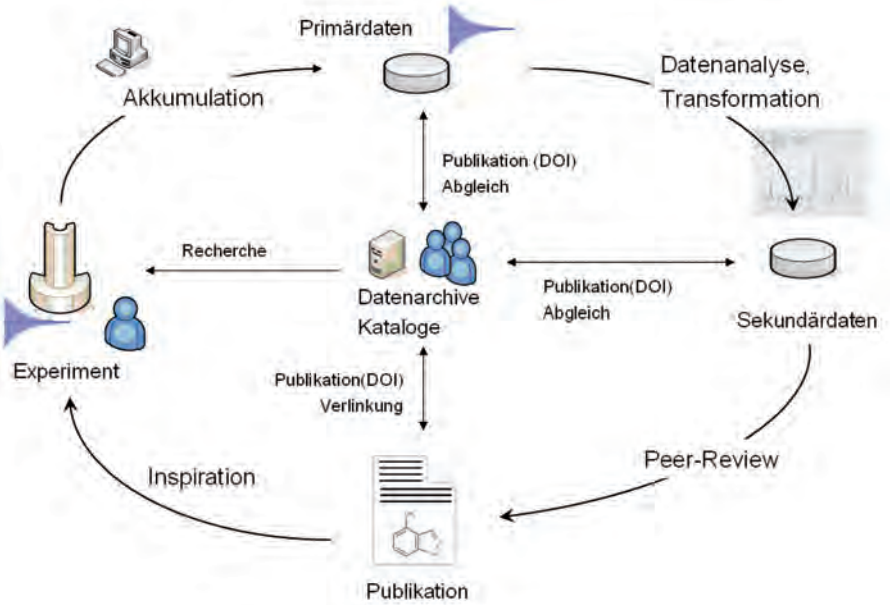

Abbildung 1: Ein idealer Zyklus der Information und des Wissens

Der DFG-Ausschuss „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ hat 2004 ein Projekt[59] gestartet, um den Zugang zu wissenschaftlichen Forschungsdaten zu verbessern. Aus diesem Projekt heraus ist die TIB seit Mai 2005 weltweit erste DOI-Registrierungsagentur für wissenschaftliche Daten. Beispielhaft im Bereich der Geowissenschaften werden Forschungsdatensätze

59 http://www.std-doi.de

registriert. Die Datensätze selber verbleiben bei den lokalen Datenzentren und die TIB vergibt für jeden Datensatz einen DOI-Namen.

Der Datensatz wird somit eine eigene zitierfähige Einheit. Mittlerweile wurden über dieses System über 600.000 Datensätze mit einer DOI versehen und zitierfähig gemacht. Die Metadatenbeschreibungen der Datensätze werden zentral an der TIB gespeichert. Diese Beschreibungen enthalten alle Angaben, die nach ISO 690-2 (ISO 1997) zur Zitierung elektronischer Medien verlangt werden.

Abbildung.2: Anzeige eines Forschungsdatensatzes im Online-Katalog der TIB Hannover

Zusätzlich werden Sammlungen oder Auswertungen von Forschungsdatensätzen auch in den Katalog der TIB aufgenommen. Die Anzeige eines Forschungsdatensatzes im Katalog der TIB sehen sie in Abbildung 2.

Die DOI Registrierung erfolgt bei der TIB immer in Kooperation mit lokalen Datenspeichern als sog. Publikationsagenten, also jenen Einrichtungen, die weiterhin für Qualitätssicherung und die Pflege und Speicherung der Inhalte, sowie die Metadatenerzeugung zuständig sind. Die Datensätze selber verbleiben bei diesen lokalen Datenzentren, die TIB speichert die Metadaten und macht alle registrierten Inhalte über eine Datenbank suchbar.[60]

60 Brase (2004); Lautenschlager et al. (2005)

Für die Registrierung von Datensätzen wurde an der TIB ein Webservice eingerichtet. Komplementär wurden bei den Publikationsagenten entsprechende Klienten eingerichtet, die sowohl eine automatisierte als auch manuelle Registrierung ermöglichen. In allen Datenzentren sind die SOAP[61]-Klienten vollständig in die Archivierungsumgebung integriert, so dass zusätzlicher Arbeitsaufwand für die Registrierung entfällt. Mithilfe dieser Infrastruktur sind bisher problemlos mehrere hunderttausend DOI Namen registriert worden. Das System baut seitens der TIB auf dem XML-basierten Publishing-Framework COCOON von Apache auf. Dazu wurde COCOON um eine integrierte Webservice-Schnittstelle erweitert, wodurch die Anbindung von weiterer Software überflüssig wird. Die modulare Struktur des Systems erlaubt es, dieses auf einfache Weise auf alle weiteren Inhalte, die mit DOI Namen registriert werden, anzupassen.

DataCite

Seit Januar 2010 erfolgt die DOI-Registrierung an der TIB unter dem Namen „DataCite" in weltweiter Kooperation mit anderen Bibliotheken und Informationseinrichtungen. DataCite hat sich zum Ziel gesetzt, Wissenschaftlern den Zugang zu Forschungsdaten über das Internet zu erleichtern, die Akzeptanz von Forschungsdaten als eigenständige, zitierfähige wissenschaftliche Objekte zu steigern und somit die Einhaltung der Regeln guter wissenschaftlicher Praxis zu gewährleisten.
Partner aus acht Ländern haben sich unter der Leitung der TIB unter Dach von DataCite zusammengefunden: die British Library, das Technical Information Center of Denmark, die TU Delft Bibliothek aus den Niederlanden, das Canada Institute for Scientific and Technical Information (CISTI), die California Digital Library und die Purdue University aus den USA, die Bibliothek der ETH Zürich, das Institut de l'Information Scientifique et Technique (INIST) aus Frankreich, sowie aus Deutschland neben der TIB noch die ZB MED und das Leibniz-Institut für Sozialwissenschaften (GESIS).[62]

Status

Die DOI-Registrierung von Forschungsdaten ermöglicht eine elegante Verlinkung zwischen einem Wissenschaftlichen Artikel und den im Artikel analysier-

61 SOAP steht für Simple Object Access Protocol, ein Netzwerkprotokoll, mit dessen Hilfe Daten zwischen Systemen ausgetauscht werden können

62 http://www.datacite.org

ten Forschungsdaten. Artikel und Datensatz sind durch die DOI in gleicher Weise eigenständig zitierbar.
So wird beispielsweise der Datensatz:

Kuhlmann, H et al. (2009):
Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands.
doi:10.1594/PANGAEA.727522,

in folgendem Artikel verwendet:

Kuhlmann, Holger; Freudenthal, Tim; Helmke, Peer; Meggers, Helge (2004):
Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation.
Marine Geology, 207(1-4), 209-224,
doi:10.1016/j.margeo.2004.03.017

Diese Verlinkung wird auch bei der Anzeige des Artikels über das Portal „ScienceDirect“ dargestellt (Abbildung 3). Durch eine Kooperation des Datenzentrums „Publishing Network for Geoscientific & Environmental Data (PANGAEA)“ mit Elsevier wird bei jedem Artikel, der in ScienceDirect angezeigt wird automatisch geprüft, ob für diesen Artikel Forschungsdaten verfügbar sind, die mit einer DOI registriert wurden und ggf. ein Verweis direkt auf die Vorschauseite des Artikels platziert.

Abbildung 3: Anzeige eines Artikels in ScienceDirect mit Verweis auf die verfügbaren Forschungsdaten (Supplementary Data)

Mittlerweile hat die TIB ihr Angebot auch auf andere Inhaltsformen ausgeweitet.[63] Als Beispiele seien hier genannt:

- doi:10.1594/EURORAD/CASE.6634 in Kooperation mit dem European Congress for Radiology (ECR) wurden über 6.500 medizinische Fallstudien registriert.
- doi:10.2312/EGPGV/EGPGV06/027-034 in Kooperation mit der European Association for Computer Graphics (Eurographics) wurden

63 Weitere Informationen zu den Aufgaben der TIB als DOI-Registrierungsagentur und dem Nachweis von Forschungsdaten durch DOI-Namen sind auf den Internetseiten der TIB zu finden
http://www.tib-hannover.de/de/die-tib/doi-registrierungsagentur/ und
http://www.tib-hannover.de/de/spezialsammlungen/forschungsdaten/

über 300 Artikel (Graue Literatur) registriert.
• doi:10.1594/ecrystals.chem.soton.ac.uk/145 Gemeinsam mit dem Projekt eBank des UK Office for Library Networking wurden erstmals DOI Namen für Kristallstrukturen vergeben.
• doi:10.2314/CERN-THESIS-2007-001 in Kooperation mit dem CERN werden DOI Namen für Berichte und Dissertationen vergeben
• doi:10.2314/511535090 Seit Sommer 2007 vergibt die TIB auch DOI Namen für BMBF Forschungsberichte.
• doi:10.3207/2959859860 ist ein Beispiel für ein in Kooperation mit der ZB MED registrierten Wissenschaftlichen Film.

DOI-Namen und Langzeitarchivierung

Die Referenzierung von Ressourcen mit persistenten Identifiern ist ein wichtiger Bestandteil jedes Langzeitarchivierungskonzeptes. Der Identifier selber kann natürlich keine dauerhafte Verfügbarkeit sicherstellen, sondern stellt nur eine Technik dar, die in ein Gesamtkonzept eingebunden werden muss. Ein Vorteil der DOI ist hier sicherlich einerseits der zentrale Ansatz durch die überwachende Einrichtung der IDF, der die Einhaltung von Standards garantiert und andererseits die breite Verwendung der DOI im Verlagswesen, das an einer dauerhaften Verfügbarkeit naturgemäß interessiert ist. In sehr großen Zeiträumen gerechnet gibt es natürlich weder für die dauerhafte Existenz der IDF noch der CNRI eine Garantie. Allerdings ist die Technik des Handle Systems so ausgelegt, dass eine Registrierungsagentur jederzeit komplett selbstständig die Auflösbarkeit ihrer DOI-Namen sicherstellen kann.

Literatur

Brase, Jan (2004): *Using Digital Library Techniques - Registration of Scientific Primary Data.* Lecture Notes in Computer Science, 3232: 488-494.

International Organisation for Standardisation (ISO) (1997): ISO 690-2:1997 Information and documentation, TC 46/SC 9

Lautenschlager, M., Diepenbroek, M., Grobe, H., Klump, J. and Paliouras, E. (2005): *World Data Center Cluster „Earth System Research" - An Approach for a Common Data Infrastructure in Geosciences.* EOS, Transactions, American Geophysical Union, 86(52, Fall Meeting Suppl.): Abstract IN43C-02.

Uhlir, Paul F. (2003): *The Role of Scientific and Technical Data and Information in the Public Domain,* National Academic Press, Washington DC