

DIN ISO 24619

ICS 01.140.20

Einsprüche bis 2018-04-16

Entwurf

**Management von Sprachressourcen –
Identifikation und nachhaltiger Zugang (PISA) (ISO 24619:2011);
Text Deutsch und Englisch**

Language resource management –
Persistent identification and sustainable access (PISA) (ISO 24619:2011);
Text in German and English

Gestion des ressources langagières –
Identification et accès pérennes (ISO 24619:2011);
Texte en allemand et anglais

Anwendungswarnvermerk

Dieser Norm-Entwurf mit Erscheinungsdatum 2018-02-16 wird der Öffentlichkeit zur Prüfung und
Stellungnahme vorgelegt.

Weil die beabsichtigte Norm von der vorliegenden Fassung abweichen kann, ist die Anwendung dieses Entwurfs
besonders zu vereinbaren.

Stellungnahmen werden erbeten

- vorzugsweise online im Norm-Entwurfs-Portal von DIN unter www.din.de/go/entwuerfe bzw. für Norm-Entwürfe der DKE auch im Norm-Entwurfs-Portal der DKE unter www.entwuerfe.normenbibliothek.de, sofern dort wiedergegeben;
- oder als Datei per E-Mail an nat@din.de möglichst in Form einer Tabelle. Die Vorlage dieser Tabelle kann im Internet unter www.din.de/go/stellungnahmen-norm-entwuerfe oder für Stellungnahmen zu Norm-Entwürfen der DKE unter www.dke.de/stellungnahme abgerufen werden;
- oder in Papierform an den DIN-Normenausschuss Terminologie (NAT), 10772 Berlin, Burggrafenstr. 6, 10787 Berlin.

Die Empfänger dieses Norm-Entwurfs werden gebeten, mit ihren Kommentaren jegliche relevanten
Patentrechte, die sie kennen, mitzuteilen und unterstützende Dokumentationen zur Verfügung zu stellen.

Gesamtumfang 71 Seiten

DIN-Normenausschuss Terminologie (NAT)

Inhalt

	Seite
Nationales Vorwort.....	4
Nationaler Anhang NA (informativ) Literaturhinweise	4
Vorwort	5
Einleitung	6
1 Anwendungsbereich.....	7
2 Normative Verweisungen	7
3 Begriffe	7
3.1 Ressourcen.....	7
3.2 Identifikatoren	10
3.3 Rollen, Institutionen und Services	11
3.4 Aktionen.....	12
4 Hintergrund.....	13
5 Anforderungen für PID-Schemata und die Verwendung von PIDs.....	14
5.1 Allgemein.....	14
5.2 Anforderungen für PID-Schemata	14
5.2.1 Allgemein.....	14
5.2.2 Doubletten verwalten	14
5.2.3 Auf Metadaten für eine Ressource zugreifen	15
5.2.4 Sichere und verlässliche Verwaltung.....	15
5.2.5 Identifikatoren für eine Teilressource	15
5.2.6 Als URL dargestellte PIDs	15
5.3 Verwendung von PIDs.....	16
5.4 Informationen zur Zitation und zu persistenten Identifikatoren	16
5.5 Verweise auf Teile von Ressourcen	17
5.5.1 Allgemeines	17
5.5.2 Medien (mit Bezug auf die Zeit)	17
5.5.3 Textuelle Ressourcen.....	17
5.5.4 Metadaten-Verzeichnisse, Terminologien, Ontologien	17
5.6 Sammlungen	18
6 Weitergehende Anforderungen	18
6.1 Granularität von Identifikatoren.....	18
6.2 Empfehlungen.....	19
Anhang A (informativ) Unabhängige Ressourcen, zusammengestellte Ressourcen und Teile von Ressourcen.....	20
A.1 Übersicht.....	20
A.1.1 Allgemeines	20
A.1.2 Ressourcen.....	20
A.1.3 Ressourcen-Bundles.....	21
A.1.4 Ressourcensammlungen.....	21
A.1.5 Metadatenbeschreibungen	22
A.1.6 Bemerkungen zur Granularität.....	22
A.2 Gängige Vorgehensweisen, um auf Ressourcen im Internet zu verweisen	23
A.3 Systeme zur Auflösung persistenter Identifikatoren	24

A.4	Verweise auf Teile von Ressourcen	25
A.4.1	Allgemeines	25
A.4.2	Medien (mit Bezug auf die Zeit)	25
A.4.3	Textuelle Ressourcen.....	25
A.4.4	Wissensbasen.....	26
A.5	Verweis auf Sammlungen von Ressourcen	27
A.6	Angaben zur Zitation.....	27
A.6.1	Allgemeines	27
A.6.2	ISO 690:2010.....	28
A.6.3	Der APA-Zitationsstil.....	28
A.6.4	Handbuch der MLA.....	29
A.6.5	STD-DOI und DataCite.....	29
A.6.6	Ein Vorschlag zur Standardisierung von wissenschaftlichen Zitationen quantitativer Daten	29
Anhang B (informativ) Implementierungen für Systeme zur Verwaltung, Bearbeitung und Auflösung persistenter Identifikatoren		30
B.1	Persistente URL (PURL)	30
B.2	Handle-System (HS).....	30
B.3	Archival Resource Key (ARK).....	31
B.4	Weitere PID-Systeme	32
Anhang C (informativ) Abkürzungen.....		33
Literaturhinweise.....		35
Alphabetisches Stichwortverzeichnis.....		37

Nationales Vorwort

Dieses Dokument enthält die deutsche Übersetzung der Internationalen Norm ISO 24619:2011, die vom Technischen Komitee ISO/TC 37 „Terminology and other language and content resources“ erarbeitet wurde, dessen Sekretariat von SAC (China) gehalten wird.

Das zuständige nationale Normungsgremium ist der Arbeitsausschuss NA 105-00-06 AA „Sprachressourcen“ im DIN-Normenausschuss Terminologie (NAT).

Um Zweifelsfälle in der Übersetzung auszuschließen, ist die englische Originalfassung beigelegt. Die Nutzungsbedingungen für den deutschen Text des Norm Entwurfes gelten gleichermaßen auch für den englischen Text.

Für die in diesem Dokument zitierten internationalen Dokumente wird im Folgenden auf die entsprechenden deutschen Dokumente hingewiesen:

ISO 690:2010 siehe DIN ISO 690:2013-10

ISO 2108 siehe DIN ISO 2108

Nationaler Anhang NA (informativ)

Literaturhinweise

DIN ISO 690:2013-10, *Information und Dokumentation — Richtlinien für Titelangaben und Zitierung von Informationsressourcen (ISO 690:2010)*

DIN ISO 2108, *Information und Dokumentation — Internationale Standard-Buchnummer (ISBN)*

Vorwort

ISO (die Internationale Organisation für Normung) ist eine weltweite Vereinigung von Nationalen Normungsorganisationen (ISO-Mitgliedsorganisationen). Die Erstellung von Internationalen Normen wird normalerweise von ISO Technischen Komitees durchgeführt. Jede Mitgliedsorganisation, die Interesse an einem Thema hat, für welches ein Technisches Komitee gegründet wurde, hat das Recht, in diesem Komitee vertreten zu sein. Internationale Organisationen, staatlich und nicht-staatlich, in Liaison mit ISO, nehmen ebenfalls an der Arbeit teil. ISO arbeitet eng mit der Internationalen Elektrotechnischen Kommission (IEC) bei allen elektrotechnischen Themen zusammen.

Internationale Normen werden in Übereinstimmung mit den Gestaltungsregeln der ISO/IEC-Direktiven, Teil 2 erarbeitet.

Die Hauptaufgabe der Technischen Komitees besteht in dem Erarbeiten von Internationalen Normen. Die von den Technischen Komitees angenommenen Norm-Entwürfe werden den Mitgliedsorganisationen zur Umfrage zur Verfügung gestellt. Für eine Veröffentlichung als Internationale Norm wird eine Zustimmung von mindestens 75 % der Mitgliedsländer, die abgestimmt haben, benötigt.

Es wird auf die Möglichkeit hingewiesen, dass einige Elemente dieses Dokuments Patentrechte berühren können. ISO ist nicht dafür verantwortlich, einige oder alle diesbezüglichen Patentrechte zu identifizieren.

ISO 24619 wurde vom Technischen Komitee ISO/TC 37, *Terminology and other language and content resources*, Unterkomitee SC 4, *Language resource management*, erarbeitet.

Einleitung

Referenzen und Zitationen bilden wichtige Bestandteile von Dokumenten und wissenschaftlichen Veröffentlichungen. Sie werden üblicherweise von Autoren verwendet, um die Quellen ihrer Arbeit angemessen kenntlich zu machen bzw. um ihre Argumentation zu stützen. Zitationen enthalten im Allgemeinen Informationen, die es dem Leser erlauben, die mögliche Relevanz der zitierten Veröffentlichung festzustellen und diese eindeutig zu identifizieren. Bibliothekare und andere Fachleute sind in der Lage, das Dokument auf Basis der Verweisinformation mit Hilfe etablierter Verfahren aufzufinden.

Die zunehmende Verfügbarkeit von Dokumenten im World Wide Web hat dazu geführt, dass vermehrt Web-Adressen (URI [4]) bei Verweisen mit angegeben werden. Dadurch wird es möglich, auf die verwiesenen Dokumente direkt über Web-Browser oder andere Anzeigewerkzeuge zuzugreifen. Dies wurde bereits in Normen wie ISO 690 empfohlen, wobei dort stärker die Identifizierung von Ressourcen und ihrer Teile im Vordergrund steht, und weniger die Nachhaltigkeit des Zugangs. Mittlerweile wird die automatische oder interaktive Auswertung solcher Verweise immer häufiger erforderlich, was einen verlässlichen Zugang zu den referenzierten Ressourcen notwendig macht. Zugangsprobleme, die beim Umzug von Ressourcen entstehen, haben zur Verwendung von Systemen persistenter Identifikatoren (PID) geführt [23], [24]. Aktuelle Ansätze [18], [19], [24] behandeln das Problem des Ressourcen-Umzugs durch die Bereitstellung sogenannter Resolver-Dienste, die einen Ressourcen-Identifikator in die entsprechende aktuelle Adresse überführen. Diese Resolver-Dienste erlauben es außerdem, weitere Metadaten mit dem Identifikator zu verknüpfen. Elaborierte Ansätze wie die Digital Object Identifiers (DOI) [14] verwenden diese Möglichkeit zur Verwaltung zusätzlicher Informationen, etwa zu den Urheberrechten.

Beim Referenzieren und Zitieren von wissenschaftlichen Daten, wie einzelnen Ressourcen oder Datensätzen, ist der Einsatz persistenter Identifikatoren bislang weniger stark ausgeprägt. Persistente Identifikatoren sind jedoch auch in diesem Zusammenhang von großem Nutzen, da sie dem Leser einer Veröffentlichung bzw. dem Nutzer einer Wissensressource einen direkten Zugriff auf die wissenschaftlichen Primärdaten gestatten, auf die verwiesen wird. Verwendet man Verweise, um auf wissenschaftliche Daten, inklusive Sprachressourcen, zuzugreifen, dann ist es wichtig, auch auf Teile der Ressourcen verweisen und zugreifen zu können. Dies trifft besonders auf den Bereich der Sprachressourcen zu, da hier gewöhnlich mehrere Analyseebenen unterschiedlicher Differenziertheit auf ein und derselben Datenmenge oder Ressourcensammlung überlagert sind. Daher beschäftigt sich dieser Standard im Hinblick auf die Verwendung von PID-Systemen sowie die an sie zu stellenden Anforderungen insbesondere mit der Frage, wie diese Systeme effizient mit der Identifikation von Teilressourcen und dem Zugriff darauf umgehen können. Spezielle Empfehlungen zeigen auf, wie das Problem der unterschiedlichen Granularität bei der Erstellung von PIDs für Ressourcen und Ressourcensammlungen behandelt werden kann.

Ein weiterer Grund für die Verwendung von PID-Systemen zur Identifikation von Ressourcen in wissenschaftlichen Datensätzen ist dadurch gegeben, dass moderne Archive und Repositorien dazu übergehen, aufeinander bezogene komplexe Ressourcen, die über verschiedene Speicherorte verteilt sein können, miteinander zu verbinden. In solchen Fällen sind permanente Verknüpfungen unverzichtbar. Beispielsweise kann ein lexikalischer Eintrag in einem multimedialen Lexikon auf Bilder verweisen, die nicht notwendigerweise physikalisch im Lexikon vorhanden sind oder die sogar auf einer externen, von einer anderen Organisation verwalteten Web-Site zu finden sind. Die Verknüpfung zwischen dem lexikalischen Eintrag und dem entsprechenden Bild muss jedoch gültig bleiben, selbst wenn im Laufe der Zeit Server umziehen oder Dateien verschoben werden. Neuere Ansätze im Bereich der E-Sciences, die mit verteilten Diensten auf verteilten Ressourcen arbeiten, sind ebenfalls darauf angewiesen, dass sie für Dienste in transparenter Weise zugreifbar sind, unabhängig davon, wo sich letztere befinden oder von wem sie betrieben werden. Dies bedeutet auch, dass das Auflösen von Referenzen auf Ressourcen keinesfalls durch unnötige Abhängigkeiten beeinträchtigt werden sollte, die auf nicht nachhaltigen oder unzuverlässigen Diensten beruhen, ob solche Abhängigkeiten nun technischer oder organisatorischer Natur sind.

Eine weitere Komplikation bei der Aufgabe, PID-Dienste für die gesamte Gemeinschaft der Anbieter von Sprachressourcen bzw. -technologien verfügbar zu machen, besteht darin, auflösbare PIDs bereitzustellen, ohne dass den Ressourcen-Anbietern zusätzliche finanzielle Belastungen entstehen, die über die üblichen Aufwendungen für die Bereitstellung von Ressourcen im Internet hinausgehen.

1 Anwendungsbereich

Diese Internationale Norm spezifiziert Anforderungen an Verfahren für persistente Identifikatoren (PID) und für die Verwendung von PIDs als Referenzen auf und Zitationen von Sprachressourcen, sowohl in Veröffentlichungen als auch in den Sprachressourcen selbst. Beispiele für Sprachressourcen sind hier digitale Wörterbücher, sprachspezifische terminologische Ressourcen, Lexika für die maschinelle Übersetzung, annotierte multimediale und multimodale Korpora, Textkorpora, die z. B. mit morpho-syntaktischen Annotationen versehen sind, usw. Solche Ressourcen werden unter anderem von Computer-linguisten und angewandten Linguisten sowie von weiteren Experten aus der Wissensverwaltung erstellt.

Diese Internationale Norm beschäftigt sich auch mit Fragen der Persistenz und der Granularität der Referenzen auf solche Ressourcen, zunächst dadurch, dass persistente Referenzen durch ein PID-Verfahren umgesetzt werden; zusätzlich dadurch, dass Anforderungen an die hierfür verwendeten PID-Verfahren formuliert werden.

PID-Verfahren ermöglichen es auch, allgemeine Metadaten mit diesen Identifikatoren zu verbinden, die auch Informationen zur Zitation beinhalten können. Diese Internationale Norm gibt dabei die Mindestanforderungen für eine effiziente Verwendung von PIDs im Bereich der Sprachressourcen an und verweist auf verschiedene relevante Normen und gebräuchliche Verfahren, etwa: ISO 690 [16], APA [3], MLA [9] für Zitationen, ISO/IEC 21000-17, IETF RFC 5147, Annotea [2], Zeiträume [22], XPointer, um Teile zu adressieren, PURL [23], ARK [18], Handle System [24] und DOI [14].

2 Normative Verweisungen

Die folgenden zitierten Dokumente sind für die Anwendung dieses Dokuments erforderlich. Bei datierten Verweisungen gilt nur die in Bezug genommene Ausgabe. Bei undatierten Verweisungen gilt die letzte Ausgabe des in Bezug genommenen Dokuments (einschließlich aller Änderungen).

ISO 12620:2009, *Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources*

ISO/IEC 21000-17:2006, *Information technology — Multimedia framework (MPEG-21) — Part 17: Fragment Identification of MPEG Resources*

W3C 2003, *XPointer Framework*: [online] W3C Recommendation 25 March 2003 [viewed 2010-08-04]. Available from: <http://www.w3.org/TR/xptr-framework/>

WILDE, E. und DUERST, M. *URI Fragment Identifiers for the text/plain Media Type*, IETF RFC 5147, April 2008 [viewed 2010-12-22]. Available from: <http://www.rfc-editor.org/rfc5147.txt/>

3 Begriffe

Für die Anwendung dieses Dokuments gelten die folgenden Begriffe.

3.1 Ressourcen

3.1.1

Ressource

digitales Objekt, das über das Internet erreichbar ist und das eindeutig über eine *URI* (3.2.2) adressiert werden kann

ANMERKUNG 1 Vergleiche auch IETF RFC 3986.

ANMERKUNG 2 Im Zusammenhang mit dieser Internationalen Norm kann eine Ressource sich auch auf eine *Sprachressource* beziehen, für die es eine Online-Darstellung gibt.

ANMERKUNG 3 Eine Ressource kann verschiedene Darstellungen haben. In Abhängigkeit vom *PID-Schema* (3.2.5) wird die Identifikation einer spezifischen Repräsentation in einem Identifikator kenntlich gemacht (ARK, siehe B.3) oder der Kommunikation [8] zwischen dem *Web-Client* (3.3.8), der die aufgelöste PID verwendet, um die *Ressource* (3.1.1) herunterzuladen, und dem *Ressourcen-Server* (3.3.6) überlassen.

3.1.2

Sprachressource

digitale Ressource, die Informationen über eine oder mehrere Sprachen liefert

ANMERKUNG Sprachressourcen umfassen lexikographische, terminologische, morphosyntaktische, korpuspezifische und semantische Ressourcen sowie digitale Ressourcen, die benutzt werden, um linguistische Phänomene in Texten und multimedialen/multimodalen Aufnahmen zu untersuchen. Sie werden u. a. von Linguisten, Informationswissenschaftlern, Lexikographen und Terminologen erstellt und verwendet. Sie enthalten häufig viele kleine Teile, die zu größeren Sammlungen zusammengefasst werden, und dienen oft als Referenz, z.B. für standardisierte Terminologien und Glossare, die durch Normungsorganisationen und Regelsetzer wie ISO, IETF, W3C usw. herausgegeben werden.

3.1.3

komplexe Ressource

Ressource (3.1.1), die aus mehreren Teilen besteht, von denen auf jeden Teil einzeln zugegriffen werden kann

ANMERKUNG Eine komplexe Ressource kann zu einer virtuellen Ressource zusammengefasst werden, wenn ihre Teile über unterschiedliche *Repositorien* (3.1.6) verteilt sind.

3.1.4

Sammlung

Gruppierung einiger *Ressourcen* (3.1.1), auf die als Ganzes verwiesen werden soll

3.1.5

veröffentlichte Sammlung

Sammlung, die zielgerichtet erstellt wurde und die als unabhängige Einheit von einem *Archiv* (3.1.7) oder *Repositorium* (3.1.6) gepflegt wird und für die eine angemessene *Zitation* (3.1.16) verfügbar ist

3.1.6

digitales Repositorium Repositorium

Einrichtung, die einen zuverlässigen Zugang zu gepflegten digitalen *Ressourcen* (3.1.1) ermöglicht

3.1.7

Archiv

digitales Archiv

Repositorium (3.1.6) für die langfristige Speicherung von darin enthaltenen Daten

ANMERKUNG Häufig sind die Daten in digitalen Archiven auch online verfügbar, was die Notwendigkeit von *persistenten Identifikatoren* (3.2.4) unterstreicht.

3.1.8

virtuelle Ressourcensammlung

virtuelle Zusammenstellung getrennter, nicht schon anderweitig zusammenhängender Ressourcen zu einer *Sammlung* (3.1.4), die zu einem bestimmten Zweck zusammengestellt wurde und die durch eine einzelne *PID* (3.2.4) zusammen mit *Teil-Identifikatoren* (3.2.7) referenziert wird, um die Bestandteile der Sammlung zugänglich zu machen

ANMERKUNG Eine Bibliographie oder ein Index können eine einzelne PID zusammen mit Erweiterungen verwenden, um einen Zugang zu Bestandteilen von *Ressourcen* (3.1.1) zu ermöglichen, die für eine Monographie oder ein Projekt

verwendet werden, ohne die Dateien an einem bestimmten Ort zu speichern. Das heißt, dass die einzelnen Bestandteile an ihren ursprünglichen Speicherorten verbleiben, aber als Teil eines Ganzen referenziert werden.

3.1.9

Version

bestimmte Form oder Variante einer *Ressource* (3.1.1), die sich von anderen Ausprägungen der *Ressource* zumindest in einem Aspekt oder durch eine Information unterscheidet

ANMERKUNG Versionen werden häufig durch aufsteigende Folgen gekennzeichnet (z. B. Version 1, 2, etc.), aber die Versionskennzeichnung von dynamischen Ressourcen erfolgt häufig durch einen Zeitstempel.

3.1.10

Momentaufnahme

Kopie einer *Ressource* (3.1.1), die den Stand der *Ressource* oder Sammlung zu einem bestimmten Zeitpunkt darstellt

3.1.11

abstrakte Ressource

Ressource, die nicht über ein Netzwerk zugänglich ist, die aber durch eine *URI* (3.2.2) bezeichnet wird, normalerweise etwas wie eine Klasse oder Eigenschaft

ANMERKUNG Üblicherweise werden abstrakte Ressourcen mittels URIs identifiziert, z. B. in Ontologien, die in RDFS (RDF Schema) oder OWL (Web Ontology Language) ausgedrückt werden. Die Architektur des Internet macht es nicht erforderlich, dass Informationsressourcen durch solche URI abrufbar sind. Wenn ein Identifikator einer abstrakten *Ressource* nicht dazu gedacht ist, *aufgelöst zu werden* (3.4.1), wie es z. B. bei XML Namespace URIs der Fall sein kann, besteht nicht die Notwendigkeit, für so eine *Ressource* eine *PID* (3.2.4) zu vergeben.

3.1.12

Teilressource

Teil

identifizierbare, zugängliche Einheit in einer unabhängigen *Ressource* (3.1.1) oder in einem größeren Teil davon

ANMERKUNG Die Teile können in anderen Teilen eingebettet werden. In dynamischen Internetumgebungen ist das Bilden von Untereinheiten Änderungen unterworfen, die auch auf unterschiedliche Interpretationen zurückzuführen sind. Dies macht ein bestimmtes Maß an Nutzerentscheidungen notwendig, um solche Untereinheiten festzulegen und zu erkennen.

3.1.13

Fragment

Teil oder Teilmenge einer ursprünglichen *Ressource* (3.1.1), eine Perspektive auf eine Repräsentation einer ursprünglichen *Ressource* oder eine andere *Ressource*, die als Bestandteil einer *Ressource* definiert oder durch eine solche Repräsentation beschrieben wird

ANMERKUNG 1 Vergleiche auch IETF RFC 3986.

ANMERKUNG 2 In dieser Internationalen Norm wird die Benennung *Fragment* nur im Sinne von IETF RFC 3986 verwendet, wenn in einem Internet-Kontext eine *Client-Anwendung* (3.3.5) das *Fragment* aus einer *Ressource*, die es enthält, abrufen.

3.1.14

terminaler Bestandteil

Teil (3.1.12) einer *Ressource* (3.1.1), der nicht in kleinere Teile unterteilt ist

3.1.15

interner Bestandteil

Teil (3.1.12) einer *Ressource* (3.1.1), der sowohl in einer *Ressource* eingebettet als auch selbst unterteilt ist

3.1.16

Zitation

Informationsgegenstand, der einen Leser oder Nutzer von einer Ressource auf eine andere *Ressource* (3.1.1) verweist

3.1.17

Referenz

digitales Objekt, das auf Daten an anderer Stelle verweist

ANMERKUNG Auch wenn *Zitation* (3.1.16) und Referenz oft quasi synonym verwendet werden, gilt in dieser Internationalen Norm, dass Zitationen Angaben für Menschen beinhalten, während Referenzen den genauen Ort einer *Ressource* (3.1.1) angeben. Referenzen können maschinenlesbar und so eingerichtet sein, dass sie beim Vorliegen bestimmter Kriterien ausgeführt werden können.

3.1.18

Annotationsebene

separate Informationsebene, die Kommentare, Anmerkungen, Erläuterungen oder andere Arten externer Bemerkungen enthält, die einer *Ressource* (3.1.1) angefügt werden können

ANMERKUNG Beispielsweise können Diagramme oder Bilder durch zusätzliche Informationen angereichert werden, auch können Textkorpora entweder innerhalb der *Ressource* oder per Standoff-Annotation annotiert werden.

3.1.19

Standoff-Annotation

Annotation, die außerhalb einer annotierten *Ressource* gespeichert wird

3.2 Identifikatoren

3.2.1

Identifikator

digitaler Identifikator

Zeichenkette, die mit digitalen, nicht-digitalen oder abstrakten Einheiten verbunden ist, etwa Bücher, Bilder, Berichte, Metadatensätzen oder Ereignissen

3.2.2

URI

Uniform Resource Identifier

Zeichenkette in einer Syntax nach IETF RFC 3986, die verwendet wird, um eine *Ressource* (3.1.1) eindeutig zu benennen oder zu bestimmen

3.2.3

URI-Namensschema

oberste Ebene der URI

ANMERKUNG 1 Jedes Schema bringt seine eigenen Konventionen für *URIs* (3.2.2) mit.

ANMERKUNG 2 Typische URI-Schemata sind http, https, ftp, mailto usw. Sie werden bei der IANA registriert.

3.2.4

PID

persistenter Identifikator

eindeutiger *Identifikator* (3.2.1), der einen dauerhaften Zugang zu einem digitalen Objekt sicherstellt, unabhängig vom tatsächlichen Aufbewahrungsort oder vom aktuellen Besitzer

ANMERKUNG Eindeutig bedeutet in diesem Kontext, dass die PID nicht nochmals für eine andere *Ressource* ausgegeben wird. Jedoch kann die gleiche PID auf unterschiedliche Darstellungen oder *virtuelle Ressourcensammlungen* (3.1.8) der *Ressource* im Ermessen des Bereitstellers einer *Ressource* verweisen.

3.2.5

PID-Schema

Schema für die Zeichenketten der Identifikatoren [*PID* (3.2.4)-*Schema*] für digitale Objekte, die über das Web zugänglich sind, zusammen mit einem Verfahren, das es erlaubt, diese Identifikatoren in die derzeitige *URI* (3.1.1) des Objektes aufzulösen

ANMERKUNG 1 Ein PID-Schema im Sinne dieser Internationalen Norm erlaubt sowohl den Zugriff auf einzelne Objekte als auch auf *Teile* (3.1.12) und *Fragmente* (3.1.13), die in solchen Objekten enthalten sind. Ein PID-Schema kann auf vorhandene Protokolle zur Adressauflösung im Internet aufbauen oder aber die Interaktion mit Proxy-basierten Adressauflösungssystemen beinhalten.

ANMERKUNG 2 Ein PID-Schema im Sinne dieser Internationalen Norm erlaubt auch die Auflösung anderer Informationen, die mit der PID verbunden sind.

3.2.6

ausführbarer Identifikator

URI (3.2.2) mit einem *Identifikator* (3.2.1), der einer Ressource zugeordnet ist und so dargestellt ist, dass, wenn es in einem Web-Dokument angegeben ist und „angeklickt“ wird, der Browser auf die Ressource (3.2.1) umgeleitet wird; möglicherweise werden auch weitere Dienste hinsichtlich der Ressource bereitgestellt.

ANMERKUNG 1 Diese Funktion impliziert, dass die *URI* auf einen verwendbaren *Resolver-Proxy* (3.3.7) verweist.

ANMERKUNG 2 In einigen *PID-Schemata* (3.2.5) sind die *PIDs* (3.2.4) als *URIs* realisiert und damit automatisch ausführbar.

3.2.7

Identifikator für eine Teilressource

Teil-Identifikator

Zeichenkette, die sich auf eine *Teilressource* (3.1.12) bezieht, die innerhalb eines gegebenen Ressourcentyps auf spezifische Weise bestimmt werden kann (etwa durch Zeit in Medien, Bereich in einem Bild, Datensatz in einem Datenstrom usw.)

ANMERKUNG Teil-Identifikatoren im Sinne dieser Internationalen Norm sind für die Auflösung auf der Server-Seite gedacht, wohingegen *Identifikatoren für ein Fragment* (3.2.8) auf der Client-Seite verarbeitet werden.

3.2.8

Identifikator für ein Fragment

Identifikator (3.2.1), der verwendet wird, um auf einen *Teil* (3.1.12) einer *Ressource* (3.1.1) im Web-Kontext zu verweisen

ANMERKUNG 1 Vergleiche auch IETF RFC 3986.

ANMERKUNG 2 Ein Identifikator für ein Fragment, wie in IETF RFC 3986 definiert, wird durch das Vorhandensein des Doppelkreuz-Zeichen („#“) angezeigt und mit der *URI* (3.2.2) beendet. *Fragmente* (3.1.13) im Sinne dieses RFC werden aus der Ressource entnommen und durch die *Client-Anwendung* (3.3.5) aufgelöst.

ANMERKUNG 3 Vom W3C gibt es einen Regelvorschlag, wie Fragmente zu verarbeiten sind [27].

3.3 Rollen, Institutionen und Services

3.3.1

Archiveinrichtung

Einrichtung für die Verwaltung eines *digitalen Archivs* (3.1.7)

3.3.2

Bereitsteller einer Ressource

Organisation, die eine *Ressource* (3.1.1) online verfügbar macht

ANMERKUNG Eine Ressource kann auch ein Service sein.

3.3.3

Resolver

PID-Resolver

Anwendung, die einen *Identifikator* (3.2.1) in einen verwendbareren Identifikator überträgt, insbesondere eine Ressourcen-**PID** (3.2.4) in eine *URI* (3.2.2), wodurch eine Client-Anwendung an den Speicherort der *Ressource* (3.1.1) verwiesen wird

3.3.4

System zur Auflösung

System, das die Übermittlung eines *persistenten Identifikators* (3.2.4) zu einem Webdienst unterstützt, um aktuelle Informationen über das bezeichnete Objekt zu erhalten, z. B. den Speicherort (*URI*) (3.2.2) des Objekts oder der Metadaten

ANMERKUNG Das ganze System zur Auflösung kann als der „*PID-Resolver*“ (3.3.3) verstanden werden, häufig werden aber verschiedene Resolver oder Resolver-Dienste verwendet.

3.3.5

Client-Anwendung

Anwendung, die einen anderen Service verwendet, normalerweise auf einem anderen Computersystem

3.3.6

Ressourcen-Server

Computer, der als Antwort auf eine bestimmte Anfrage einer Client-Anwendung schließlich Zugang zu einem Objekt ermöglicht

3.3.7

Resolver-Proxy

HTTP Resolver-Proxy

Anwendung, die einen Service implementiert, der *als URL dargestellte* (3.4.3) *PIDs* (3.2.4) unterstützt, wodurch man auf Ressourcen oder auf weitere Informationen zur PID zugreifen kann

3.3.8

Web Client

Client-Anwendung, mit deren Hilfe man auf Ressourcen im Web mittels http-Protokoll zugreifen kann

3.4 Aktionen

3.4.1

Referenzen auflösen

auf einen Wert zugreifen, auf den durch eine *Referenz* (3.1.17) verwiesen wurde

ANMERKUNG Im Zusammenhang mit der Auflösung einer Referenz einer *URI* (3.2.4) ist damit gemeint, dass man eine Darstellung von der *Ressource* (3.1.1) erhält, auf die die URI zeigt.

3.4.2

auflösen

einen *Identifikator* (3.2.1) durch einen anderen Namen oder eine andere Adresse ersetzen, durch die eine Ressource zugänglich wird

ANMERKUNG Der Vorgang des AuflöSENS kann mehrere Schritte umfassen, um eine verwendbare Adresse einer *Ressource* zu bekommen.

3.4.3

einen Identifikator als URI darstellen

einen *Identifikator* (3.2.1) wählen, der als *URI* (3.2.4) verwendbar ist

ANMERKUNG So kann man beispielsweise einen *ausführbaren Identifikator* (3.2.6) erstellen.

4 Hintergrund

PIDs gibt es im Bereich verschiedener elektronischer Ressourcen, und die vorliegende Internationale Norm sagt nichts über diese Arten aus. Der Typ einer Ressource, die eine PID erhält, hat jedoch Konsequenzen für die Anforderungen, die man an eine PID hat. Ressourcen können in drei Klassen unterteilt werden:

- Ressourcen ohne weitere Abhängigkeiten, wie in Bild 1 gezeigt;
- ein beliebiger Teil solch einer Ressource, der näher bestimmt werden muss;
- eine Sammlung von Ressourcen, auf die als Ganzes verwiesen wird.

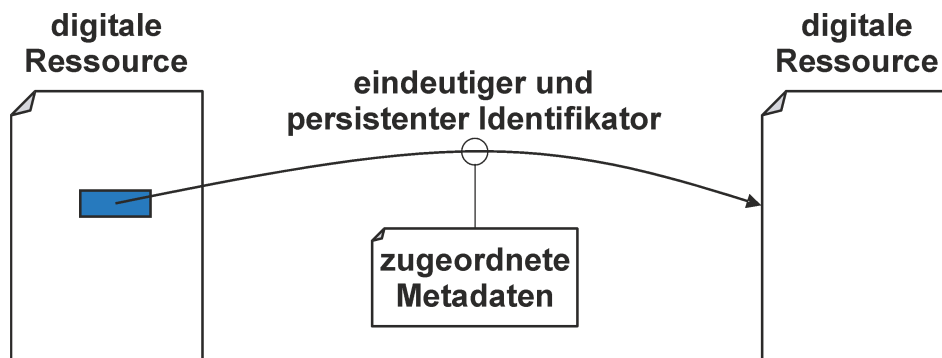


Bild 1 — Verwendung von PIDs, um von einer Ausgangsressource auf eine Zielressource zu verweisen

Diese Internationale Norm beschäftigt sich damit, wie man auf eine elektronische Ressource maschinenlesbar verweisen kann. In Bild 1 sieht man einen eindeutigen persistenten Identifikator in einer Ausgangsressource, der auf eine Zielressource verweist. Die PID kann mit verschiedenen Metadaten verknüpft werden.

Der Begriff der Ressource wird in diesem Zusammenhang sehr breit aufgefasst und die Art, wie auf sie Bezug genommen wird, hängt vom Zusammenhang ab. Beispielsweise kann eine Illustration zu einem Text entweder eine unabhängige Ressource sein, die ihre eigene eindeutige PID erhält, wodurch auf sie verwiesen werden kann, oder die Illustration kann in ein Dokument eingebunden werden, wodurch sie keinen Identifikator für sich alleine hat, da sie Teil eines Dokumentes ist. Außerdem ist es möglich, dass eine Referenz auf einen Teil der Illustration verweist. Eine Ressource kann dabei in einem Umfeld alleine verwendet werden und in einem anderen als Teil einer komplexen Ressource. Ein interner Bestandteil einer Ressource kann als terminaler Bestandteil betrachtet werden, aber eine Weiterverarbeitung in einer dynamischen Umgebung kann dazu führen, dass die Einheit Unterbereiche erhält, die selbst zugänglich sind. Diese Internationale Norm soll jeden dieser Fälle behandeln.

Im Fall einer komplexen Sprachressource sollten einige Ressourcen ihre eigenen persistenten Identifikatoren erhalten. Andere Ressourcen stellen Ressourcen dar, die andere enthalten, die selbst Teilressourcen beinhalten. In diesem Fall erhält die Ressource, die andere enthält, eine PID, und die Teile können durch Teil-Identifikatoren, die an diese PID angehängt werden, identifiziert werden. Diese Internationale Norm stellt Richtlinien zur Verfügung, wie für konkrete Ressourcen dabei vorgegangen werden sollte.

Diese Internationale Norm verwendet bestehende Normen und Verfahren für Formate für Identifikatoren von Teilressourcen und Fragmente, wo diese verfügbar sind, und sie stellt Richtlinien für die auf, wo bestehende Normen nicht ausreichen oder nicht angewendet werden können. Eine Diskussion bezüglich der Ressourcen-Typen, auf die sich diese Internationale Norm bezieht, befindet sich in Anhang A.

Hinsichtlich Sammlungen von Sprachressourcen betrachtet die Norm zwei Arten von Sammlungen:

- Ressourcensammlungen, die als komplexe Ressourcen in einer mehr oder weniger statischen Form archiviert werden, so dass die Definition der Sammlung als unabhängige Einheit von einem Archiv oder Repositorium festgelegt wird, die dann auch einen persistenten Identifikator für eine Sammlung vergeben. Die Archiveinrichtung ist also dafür verantwortlich, dass z. B. die Verbindung zwischen der PID und der Sammlung als ein Metadatensatz im Katalog eingetragen wird.
- Eine andere Art der Sammlung ist dergestalt, dass sie nicht als Sammlung von Archiveinrichtungen oder Erstellern erdacht wurde, aber sie ihren Status als komplexe Ressource durch Forschungsarbeiten oder andere Verwendungen erhält, die eine Nachvollziehbarkeit der Ressource erfordert, etwa die Verwendung in einer Monographie oder in einem Forschungsprojekt. Solche zielgerichtet erstellten Sammlungen können unter Umständen außerhalb des konkreten Arbeitsprojekts, für das sie erstellt wurden, irrelevant sein. Allerdings kann der Verweis von Forschungsdokumenten auf solche Sammlungen sehr lästig sein, wenn eine solche Sammlung aus hunderten individuellen Ressourcen besteht. Deshalb ist erforderlich, auf solche Sammlungen mit einer PID zu verweisen, die für alle Ressourcen und entsprechenden Metadaten steht, aus denen so eine Sammlung besteht. Natürlich kann so eine Art der Referenz nur dann verwendet werden, wenn es eine Ausprägung dieser Sammlung gibt.

5 Anforderungen für PID-Schemata und die Verwendung von PIDs

5.1 Allgemein

Verfügbare Normen und Verfahren zum Referenzieren und Zitieren, insbesondere im Bereich der Sprachressourcen, finden sich im Anhang A. Dieser Abschnitt bezieht sich zunächst auf die Anforderungen für das PID-Schema und anschließend auf die Anforderungen, wie PIDs als Referenz und zum Zitieren für Sprachressourcen verwendet werden.

5.2 Anforderungen für PID-Schemata

5.2.1 Allgemein

Ein PID-Schema im Sinne dieser Internationalen Norm muss Folgendes erlauben:

- a) Auflösung von einer einzelnen PID in mehrere URIs oder Services;
- b) Bezug und Zugang zu entsprechenden Metadaten;
- c) angemessene Sicherungsmaßnahmen, um böswillige oder versehentliche Änderungen an der Abbildung von PID auf URI und PID auf Metadaten zu verhindern;
- d) Verweis auf Teile einer Ressource (Identifikatoren für Teilressourcen, Fragmente oder beides);
- e) Darstellung von PIDs als URI, um Identifikatoren in Web-Dokumenten ausführbar zu machen, ohne Clients anpassen zu müssen.

5.2.2 Doubletten verwalten

Üblicherweise werden Ressourcen mehrfach abgelegt oder gespiegelt, so dass Kopien auf verschiedenen Ressourcen-Servern zur Datensicherheit oder zum schnelleren Zugang bereitliegen. Das PID-Schema sollte diese Arten der Doublettenbildung unterstützen, so dass mit einer PID mehrere URIs verbunden sein können.

5.2.3 Auf Metadaten für eine Ressource zugreifen

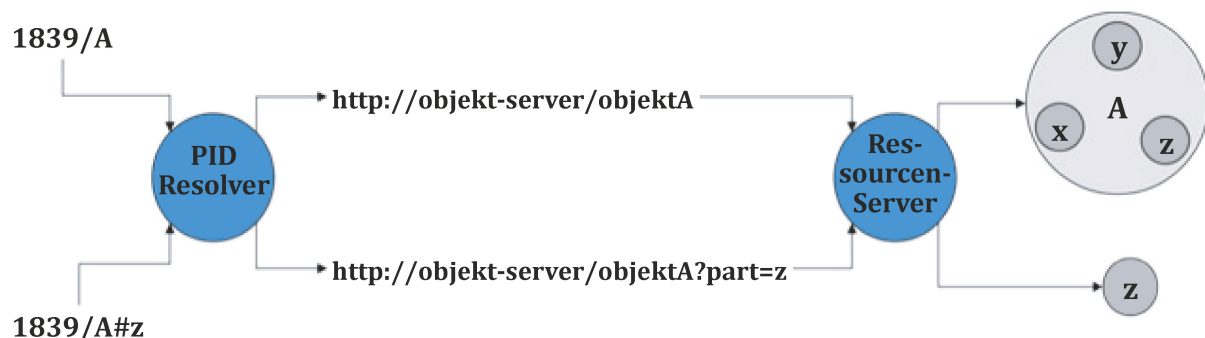
Neben der Abbildung auf eine URI wird das PID-Schema auch dazu genutzt, Metadaten, die zu einer Ressource gehören, zuverlässig der Ressource zuzuordnen. Auch wenn die vorliegende Internationale Norm keine Metadaten eines bestimmten Typs verlangt, wird doch erwartet, dass die PID zu mit der Ressource verbundenen Metadaten aufgelöst wird, die in einem XML-Format vorliegen, so dass andere Dienste diese Metadaten verwenden können.

5.2.4 Sichere und verlässliche Verwaltung

Das PID-Schema muss angemessene Sicherheitsmaßnahmen vorsehen, so dass nur der Besitzer oder Verwalter einer Ressource die Abbildung von PID/URI oder die zugehörigen Metadaten ändern kann.

5.2.5 Identifikatoren für eine Teilressource

Es ist wichtig, dass für jeden Teil einer Ressource, der adressiert werden kann, eine PID zur Verfügung steht, oder sogar, dass allgemein alle möglichen Optionen zur Segmentierung einer Ressource in ihre Teile zur Verfügung stehen. Daher müssen PID-Schemata ein System zur Vergabe von Identifikatoren für Teile oder Fragmente vorsehen, die zusammen mit der PID der Ressource benutzt werden. Da es nötig ist, dass man genau eine Zeichenkette verwendet, um eine Teilressource zu identifizieren, sollte die PID-Syntax die Verkettung einer PID mit Teil-Identifikatoren ermöglichen. Zum Beispiel sollte ein PID-Resolver konfigurierbar sein, so dass eine Kombination von PID und Teil-Identifikator in eine URI aufgelöst wird, die korrekt vom Ressourcen-Server interpretiert werden kann, um den angefragten Teil einer Ressource auszuliefern [siehe A.2 a)].



Eine komplexe Ressource „A“ mit Konstituenten x, y und z wird durch die PID „1839/A“ referenziert.

Der PID-Resolver übersetzt den Identifikator „1839/A“ in die URI `http://objekt-server/objektA`, was vom Ressourcen-Server so verstanden wird, dass das Objekt „A“ auszuliefern ist. Der Teil „z“ der Ressource „A“ wird dann durch die PID „1839/A#z“ angegeben. Damit die Verknüpfung durch den Ressourcen-Server leicht aufgelöst werden kann, sollte der PID-Resolver in der Lage sein, den Identifikator „1839/A#z“ in die URI `http://objekt-server/objektA?part=z` oder eine ähnliche Repräsentation zu übersetzen, die der Ressourcen-Server so interpretiert, dass er Teil „z“ des Objekts A ausliefert.

Bild 2 — Verarbeitung von Teil-Identifikatoren durch den PID-Resolver (am Beispiel des Handle-Servers)

5.2.6 Als URL dargestellte PIDs

Das PID-Schema sollte eine Proxy-Resolver-Implementierung bereitstellen, die in der Lage ist, PIDs aufzulösen, die als URL ausgeführt sind. Dadurch können Web-Clients solch einen Identifikator auflösen, indem sie das http-Schema verwenden, ohne dass sie spezielle Plugins oder Spezialsoftware benötigen. Bei einigen PID-Schemata ist die PID bereits eine HTTP-URI, in diesem Fall ist ein getrennter Proxy-Resolver nicht nötig.

5.3 Verwendung von PIDs

Die Zitation von digitalen Sprachressourcen, die über das Web zugänglich sind, sollte eine PID enthalten, die entweder zu einer URI, zu der Ressource oder zu einem Metadatensatz aufgelöst wird, der die Ressource beschreibt. Letztere Möglichkeit kann verwendet werden, falls eine Ressource nicht (sofort) verfügbar gemacht werden kann, oder für Sammlungen, wobei in diesem Fall der Metadatensatz, falls verfügbar, Identifikatoren für die einzelnen Bestandteile enthalten sollte.

Falls die PID zu einer URI für eine Ressource aufgelöst wird, sollte ein Metadatensatz in einem XML-Format und zusammen mit einem Schema, das zu der Ressource gehört, mit dem Identifikator verbunden und über das Resolver-System bereitgestellt werden. Zitationsinformationen sollten im Metadatensatz enthalten sein. Es gibt keine anderen Bedingungen, die für einen Metadatensatz spezifiziert sind, der mit einem Identifikator für eine Ressource verbunden ist.

In Web-Dokumenten sollte der Identifikator (z. B. ein Handle), der in Zitationen verwendet wird, nicht nur den Anforderungen der Syntax des PID-Schemata entsprechen, sondern auch als URI repräsentiert werden, so dass er in einem Webbrowser oder in einem anderen Dokumentbetrachter ausführbar wird. Als Beispiel für die Verwendung mit dem Handle-System (HS):

1839/00-0000-0000-0000-4 -> <http://hdl.handle.net/1839/00-0000-0000-0000-0000-4>

Die Resolver-Adresse <http://hdl.handle.net> zeigt auf den http-Resolver-Proxy, eine Anwendung, die http-Anfragen im Form von PIDs, die als URL dargestellt werden, als Eingabe entgegen nimmt, und die den „echten“ PID-Resolver verwendet, um den Client auf die Ressource weiterzuleiten. Beim Handle-System ist zum Beispiel der http-Resolver-Proxy entweder ein zentraler Resolver-Proxy des Handle-Systems oder ein Resolver-Proxy für das Handle-System, dessen Verfügbarkeit durch den Bereitsteller einer Ressource sichergestellt wird.

Beschränkungen auf Teile einer Ressource, analog zu Seiten oder Kapiteln, sollten, wenn nötig, unter Verwendung eines passenden Trennzeichens an die PID angehängt werden. Ein Beispiel ist ein Zeitabschnitt, auf den über das HS verwiesen wird:

1839/00-0000-0000-0000-4@time(100s,200s)

Diese Erweiterung bezieht sich auf einen Teil einer Audio-Datei (der Teil-Identifikator ist nicht standardisiert).

Falls ein passender Identifikator für ein Fragment für einen Ressourcen-Typ existiert, kann dieser Teil zu einer URI hinzugefügt werden, um einen ausführbaren Identifikator zu erstellen. Ein Beispiel im Handle Kontext:

[http://handle.net/1839/00-0000-0000-0000-0000-4?urlappend=#ffp\(track_ID=101\)*mp\(/~time\('npt','50'\)\)](http://handle.net/1839/00-0000-0000-0000-0000-4?urlappend=#ffp(track_ID=101)*mp(/~time('npt','50')))

Dieses Beispiel verwendet die besondere Funktion „urlappend“ des Handle-Proxy-Resolvers, um einen Identifikator für ein Fragment an das aufgelöste Handle anzuhängen.

5.4 Informationen zur Zitation und zu persistenten Identifikatoren

Bestehende Verfahren können beibehalten werden, falls die URIs durch PIDs ersetzt werden und die Beschreibungen für Teile von Dokumenten in Identifikatoren für Teile oder Fragmente umgewandelt werden. Verweise aus Web-Dokumenten müssen auch PIDs enthalten, die als URL dargestellt werden, falls die PID-Syntax nicht einem URI-Schema entspricht, das bei der IANA registriert ist [13].

5.5 Verweise auf Teile von Ressourcen

5.5.1 Allgemeines

Jeder Standard der ISO oder IETF zur Bestimmung eines Teils einer Ressource kann als Teil-Identifikator verwendet werden, wenn nichts anderes festgelegt wurde. Bei Ressourcen, für die kein solcher Standard besteht, ist es möglich, einen Text in der Zitation zu verwenden, der für Menschen lesbar ist, z. B.: 10 s bis 120 s für einen Zeitabschnitt. Dieser Ansatz funktioniert aber nicht für Computerprogramme.

Wenn man Identifikatoren für Teile oder Fragmente verwendet, um auf einen Teil einer Ressource zuzugreifen, ist es wichtig, den Unterschied zwischen Identifikatoren für Fragmente wie denen, die in IETF RFC 3986 beschrieben sind, und der Funktionalität eines passenden Ressourcen-Servers klarzustellen. Falls der Prozess, der die PID auflöst und eine URI zurückgibt, einschließlich einen Identifikator für ein Fragment, z. B. indem die folgende URI verwendet wird (mit Angaben für ein Fragment, wie für textuelle Medien nach IETF RFC 5147 vorgeschlagen ist):

`http://myserver/myresource#line=10,20`

so wird diese URI dazu führen, dass ein Webbrowser das ganze Dokument abrufen, der Browser würde dann den Teil des Dokuments von Zeile 10 bis 20 herausfiltern und dem Nutzer zeigen. Wenn es sich um ein kleines Dokument handelt, ist das angemessen. Ist das Fragment aber z. B. aus einer 2 GB großen Medien-Datei, muss man einen speziellen Ressourcen-Server und eine URI vergleichbar mit den Festlegungen von Teilen in Annodex [22] verwenden:

`http://videosever.com/videoA.anx?t=15.0/30.0`

Durch diese Notation wird der Ressourcen-Server angewiesen, nur das Video-Segment zwischen Sekunde 15 und Sekunde 30 auszuliefern.

Überarbeitungen dieser Internationalen Norm werden dazu führen, dass die Liste der verfügbaren Formate für Identifikatoren für Teile und Fragmente wachsen wird.

5.5.2 Medien (mit Bezug auf die Zeit)

ISO/IEC 21000-17 für Ressourcen in MPEG-21 muss für Medien und zeitlich alignierte Daten verwendet werden. Die Syntax von Annodex kann verwendet werden, um die Zeitintervalle in URI-Anfragen und Fragmenten [22] für alle passenden Medien darzustellen. Bei anderen Formaten hängt der Teil-Identifikator vom verwendeten Format ab.

5.5.3 Textuelle Ressourcen

Für textuelle Ressourcen in XML muss XPointer verwendet werden. IETF RFC 5147 muss für textuelle Dokumente ohne Formatierungen (Plain-Text) verwendet werden. Bei anderen Formaten hängt der Teil-Identifikator vom verwendeten Format ab.

5.5.4 Metadaten-Verzeichnisse, Terminologien, Ontologien

Nach ISO 12620:2009 muss jede Datenkategorie im ISO/TC 37 Datenkategorie-Repository (DCR) ihre eigene PID erhalten, die durch eine Verkettung des PID für das gesamte DCR und den spezifischen Datenkategorie-Identifikator gebildet wird. Diese PIDs werden als Permalink [30] gemäß den derzeitigen Verfahrensvorschlägen des World Wide Web Consortiums erstellt.

Teile eines RDF-Graphen können durch eine der RDF-Anfragesprachen angegeben werden, aber hierfür gibt es derzeit keine abschließenden Spezifikationen. Bei anderen Formaten hängt der Teil-Identifikator vom verwendeten Format ab.

5.6 Sammlungen

Bestehende „veröffentlichte“ Sammlungen müssen eine PID erhalten, die von demjenigen verwaltet wird, der die Sammlung verwaltet. Diese PID muss sich auf eine Beschreibung der Sammlung beziehen, die zum Beispiel ein Katalogeintrag oder eine bestimmte Metadatenbeschreibung sein kann. In virtuellen Sammlungen sollte die PID zu einer maschinenlesbaren Metadatenbeschreibung führen, die Zugriff auf die relevanten Informationen gewährt, insbesondere zu den in der Sammlung enthaltenen Ressourcen durch deren jeweilige PIDs.

6 Weitergehende Anforderungen

6.1 Granularität von Identifikatoren

Bezüglich der Granularität unterscheidet diese Internationale Norm zwischen der Identifizierung von Teilen und der Identifizierung von Fragmenten, wie in 5.5 beschrieben. Ein Identifikator für ein Fragment wird nach IETF RFC 3986 als ein optionaler Teil eines Verweises auf eine URI verstanden. Nach IETF RFC 3986 kann einer URI ein optionaler Identifikator für ein Fragment angefügt werden, wobei der Identifikator vom Rest der URI durch ein # (Doppelkreuz)-Zeichen getrennt ist. Der Separator ist kein Teil des Identifikators des Fragments.

Die URI mit dem Identifikator für ein Fragment kann von einer Anwendung verwendet werden, um eine bestimmte Ressource zu finden und auf sie zuzugreifen, wenn sie Teil einer primären Ressource oder in ihr eingebettet ist. Das Format des Identifikators des Fragments hängt von der Art der Ressource ab.

Die Funktion, einen Identifikator für ein Fragment zu interpretieren und die Referenz aufzulösen, ist eine Funktion eines Web-Clients. Daher muss die ganze Primärressource heruntergeladen werden, bevor die Client-Anwendung das benötigte Fragment auswählen kann. Da ein Identifikator für ein Fragment während des Abrufens nicht an andere Systeme weitergeleitet werden muss, verwenden einige Systeme in der Web-Architektur (z. B. Proxies) die Identifikatoren von Fragmenten nicht. Daher können http-Weiterleitungen für Fragmente keine Anwendung finden. Als Ausnahme gelten Identifikatoren für Fragmente von RDF-Dokumenten, die nicht auf Teile von Dokumenten verweisen, sondern eher auf ein Objekt in dem Dokument, das dadurch beschrieben wird, dass es einen Identifikator für ein Fragment hat. Daher erlaubt die Verwendung von Identifikatoren für Fragmente zusammen mit einer URI es einer Client-Anwendung, sich auf Teile einer Ressource zu beschränken, und zwar auf der Grundlage von Informationen, die die Anwendung selbst enthält.

Im Gegensatz zur Auflösung von Fragmenten bezieht sich die Verwendung von Teil-Identifikatoren wie „z“ in der URI

`http://myserver/myObjectService?part=z`

darauf, dass der entfernte Server den Teil auswählen kann und an die Client-Anwendung schickt.¹⁾

1) Diese strikte Trennung zwischen den Rollen von Web-Client und -Server beim Auflösen von Verweisen auf Fragmente kann sich ändern. Die Arbeitsgruppe für Fragmente von Medien hat einen W3C-Entwurf erstellt, worin vorgeschlagen wird, dass Web-Clients mit einem Server den Transport von Auszügen einer Ressource aushandeln können [27].

6.2 Empfehlungen

Diese Internationale Norm unterstützt verschiedene Ebenen der Granularität. Die folgenden Empfehlungen dienen dazu, die Effizienz zu erhöhen und Interoperabilität mit anderen Namens-Schemata zu verbessern.

- Falls es ein Schema für Identifikatoren für eine Art von Ressourcen gibt, wie zum Beispiel ISBN, dann sollte diese Granularitätsebene bewahrt werden, d. h. es sollten keine neuen PIDs ohne zwingenden Grund ausgegeben werden, z. B. für Kapitel. Kapitel würde man bevorzugter Weise über Teil-Identifikatoren zusammen mit der PID des Buches adressieren.
- Falls eine Ressource mit dem Inhalt einer ganzen Datei verbunden ist, sollte die Ressource eine eigene PID erhalten.
- Falls die Ressource eigenständig ist und auch außerhalb eines weiteren Kontextes verwendet werden kann, sollte die Ressource eine eigene PID erhalten.
- Falls eine Ressource unabhängig von einer sie enthaltenden Ressource zitierfähig sein sollte, sollte die Ressource eine eigene PID erhalten.

Diese Empfehlungen sind allerdings abhängig von den Anforderungen der Ressourcenersteller an die Granularitätsebenen, die in ihrer speziellen Ressourcen-Umgebung relevant sind.

Anhang A (informativ)

Unabhängige Ressourcen, zusammengestellte Ressourcen und Teile von Ressourcen

A.1 Übersicht

A.1.1 Allgemeines

In Wissenschaft und Industrie gibt es einen wachsenden Bedarf, auf digitale Sprachressourcen eindeutig verweisen zu können, ebenso auf Teile von Ressourcen und auf Sammlungen von Sprachressourcen. Dies ist nicht nur erwünscht, um Referenzen aus wissenschaftlichen Veröffentlichungen zu finden und zu überprüfen, sondern es wird auch mehr und mehr notwendig, verschiedene Typen von Verweisen zwischen Sprachressource oder Teilen von solchen Ressourcen zu verwalten.

A.1.2 Ressourcen

Eine Ressource ist alles, was identifiziert werden kann [7]. Trotz der Unbestimmtheit in dieser Beschreibung ist es für Forschende wichtig, linguistisch relevante Einheiten als kohärente Objekte eines Repositoriums zu bestimmen, sowohl für die interaktive Arbeit von Menschen als auch zunehmend maschinenlesbar. Solche Objekte können die Basis für verschiedene voneinander unabhängige Weiterverarbeitungen sein, d. h. dass diese Objekte unabhängig in größeren Kontexten verwendet werden. Oft kann ein solches Objekt als eine einzelne Datei in einem Dateisystem identifiziert werden, aber es ist auch möglich, dass es als Objekt innerhalb eines anderen (z. B. als Datenbankeintrag aus einer Datenbank) bereitgestellt wird. Solche Ressourcen können sehr unterschiedliche Formate haben, z. B.:

- eine digitale Videoaufnahme eines Interviews;
- eine Aufnahme eines Liedes;
- eine komplexe Annotation eines Kommunikationsvorgangs;
- ein Foto, das ein Sprechereignis zeigt;
- ein Lexikon einer bestimmten Sprache;
- eine grammatische Beschreibung;
- eine Aufnahme eines Eye-Trackers während eines Leseexperiments;
- eine Metadatenbeschreibung einer Ressource oder einer Ressourcensammlung;
- ein Dokument, das Texte, Fotos und anderes enthält.

Der Verwendungszweck einer solchen Ressource ist dem Ermessen des Erstellers überlassen. Einige Ressourcen bestehen aus einer Anzahl von Annotationsebenen, andere Ersteller stellen u. U. unterschiedliche Einheiten für jede Annotationsebene bereit, je nach wissenschaftlichen Anforderungen oder Erwägungen zur Datenverwaltung. Metadaten für viele Ressourcen können z. B. in einer relationalen Datenbank abgelegt werden. In einem solchen Fall gibt es eine „Metadatenbeschreibung“ nur als Ergebnis einer Anfrage. Solche Fälle umfassen nicht eine identifizierbare Ressource, da die referenzierte Ressource die ganze Datenbank ist. Das zeigt, dass es notwendig ist, Teile einer solchen Ressource zu referenzieren.

Wenn eindeutige und persistente Identifikatoren mit Ressourcen verbunden werden, ist es nötig, dass die Ressourcen selbst persistent sind. Das jeweilige Repositorium stellt eine Ressource mit einem eindeutigen Identifikator zur Verfügung um sicherzustellen, dass, wenn eine Referenz aufgelöst wird, der eigentliche Inhalt geliefert wird (auf Fragen des User-Interfaces, das im Laufe der Zeit u. U. gleiche Inhalte auf unterschiedliche Weise anzeigt, wird hier nicht eingegangen, auch wenn dies die Interpretation beeinflussen kann). Wenn der Inhalt einer Ressource sich verändert, unabhängig wodurch, entsteht demzufolge eine neue Ressource, weil eine neue Ausprägung entsteht, die sich von der vorherigen Version unterscheidet. Im Fall einer dynamischen Datenbank, die ständige Änderungen beinhaltet, werden Momentaufnahmen angelegt, auf die referenziert werden kann. Alle Beziehungen zwischen solchen Momentaufnahmen können mittels Metadaten ausgedrückt werden, z. B. durch qualifizierte Links oder/und durch Zitationsinformationen.

A.1.3 Ressourcen-Bundles

Manchmal machen Repositorien Angaben darüber, dass einige Ressourcen miteinander auf Grund von formalen Kriterien verwandt sind. Beispiele dafür sind:

- Audio- und Video-Aufnahmen, die zum gleichen Zeitpunkt für die gleiche Dauer erstellt wurden; dies kann Aufnahmen umfassen, die parallel zueinander erstellt wurden, zusammen mit Annotationen, die zu diesen Aufnahmen hergestellt wurden. Alle Ressourcen sind zur gleichen Zeit am selben Ort erstellt worden und sind daher eng verwandt. Nutzer möchten häufig auf diese Ressourcen als Einheit verweisen können;
- ein Lexikon, das auch Fotos oder andere Medien als Ressourcen zur Erweiterung lexikalischer Einträge einbindet. In diesem Fall kann es auch wünschenswert sein, diese verwandten Objekte als Bundle zu referenzieren;
- ein Textkorpus mit dazugehöriger Standoff-Annotation;
- ein Korpus von historischen Texten und dazugehörigen Faksimiles.

Im Allgemeinen sollten diese Bundles als Instanzen virtueller Ressourcensammlungen behandelt werden, was normalerweise durch einen gemeinsamen Metadatensatz erreicht wird, der wiederum auf die einzelnen Ressourcen verweist.

A.1.4 Ressourcensammlungen

Sammlungen sind beliebige Gruppierungen von beliebig vielen Ressourcen, die aufgrund einzelner Bedingungen zusammengefügt werden, auf die aber als Ganzes verwiesen wird. Beispiele solcher Sammlungen sind:

- das niederländische *Corpus Gesproken Nederlands (CGN)*, oder das *British National Corpus*, jeweils auf dem Publikationsstand eines bestimmten Termins;
- eine Sammlung aller Ressourcen, die eine bestimmte Sprache beschreiben und die in einem bestimmten Repositorium abgelegt sind;
- eine Sammlung von sonst nicht zusammenhängenden Ressourcen, die als Grundlage für eine Dissertation oder eine andere Monographie dient;

- eine Sammlung von Ressourcen, die zusammengefügt wurde, damit man sie zusammen in einem digitalen Archiv archivieren kann;²⁾
- Sammlungen von Teilen einer Ressource, z. B. Verweise auf mehrere Segmente einer Mediendatei, etc.

Die zugrundeliegende Beziehung zwischen den Ressourcen kann beliebig sein. Da solche Sammlungen einen formalen Status haben, kann es auch notwendig sein, auf die jeweilige Ausprägung zu verweisen, was üblicherweise über die Metadaten geschieht. Ressourcen-Bundles wie in A.1.3 sind Ressourcensammlungen, die linguistisch interessante Gemeinsamkeiten haben.

A.1.5 Metadatenbeschreibungen

Jede Ressource und Ressourcen-Sammlung sollte mit einer Metadatenbeschreibung verbunden sein, damit man von ihrer Existenz weiß. Dies liegt in der Optimierung der Archivierungssysteme und in der Langzeitarchivierung begründet; dabei wird zwischen den Objekten selbst und den möglicherweise vielen Kopien unterschieden, die in unterschiedlichen Repositorien vorliegen können. Metadatenbeschreibungen stehen dabei nicht nur für Ressourcen oder Sammlungen, sondern enthalten auch zusätzliche Informationen über diese. Falls es notwendig ist, auf die Metadatenbeschreibungen selbst zu verweisen, müssen sie als Objekte analog zu anderen Ressourcen behandelt werden. Oftmals werden Metadatensätze aber als Antwort auf eine Anfrage an eine komplexe Datenbank dynamisch generiert, was bedeutet, dass die Metadatenbeschreibung keine Ressource an sich, sondern Teil einer Ressource ist. Diese Internationale Norm berücksichtigt beides, da PIDs sowohl auf Ressourcen als auch auf Teile von Ressourcen verweisen können. Der Ressourcen-Server sollte so konfiguriert sein, dass er in beiden Fällen die gleichen Informationen ausgibt.

Im Fall der Sammlungen und Bundles verweisen die Metadatenbeschreibungen auf die eingebundenen Ressourcen, was auch rekursiv sein kann. Nicht Gegenstand dieser Internationalen Norm ist, festzulegen, wie die Metadatenbeschreibungen die Ressourcen auflisten müssen. Allerdings werden die Mechanismen angegeben, die verwendet werden, damit die Ressourcen mittels PID durch die Metadaten referenziert werden können.

Es gibt kein einheitliches Verfahren, wie die Versionierung von Metadaten durchgeführt wird. Einige Repositorien passen die Versionsnummer nicht an; andere ändern sie, wenn Informationen verändert werden. Einige ändern auch die Versionsnummer, wenn ein einzelnes Objekt, auf das verwiesen wird, verändert wird. Diese Internationale Norm gibt keine Herangehensweise an das Versionieren vor.

A.1.6 Bemerkungen zur Granularität

Die Unterscheidung zwischen Teilen einer Ressource, einer Ressource und einer Ressourcensammlung bezieht sich stark auf die Granularitätsebenen, die von einer Archiveinrichtung verwendet werden, welche das Repositorium verwaltet. Es gibt unterschiedliche Verfahren, die die unterschiedlichen Bedürfnisse und Umgebungen widerspiegeln können. Daher macht diese Internationale Norm keine weiteren Angaben zur Granularität. Die Richtlinien zur Granularität beziehen sich auf die Ebene der PIDs, die Sammlungen, Ressourcen und Ressourcen-Teilen zugewiesen werden und auch nur insoweit wie sie für die Performanz, Interoperabilität, Effizienz und zur Verwaltung erforderlich sind.

In vielen Projekten entsteht die Frage, welche Objekte geeignet sind, ihre eigene PID zu erhalten, und welche Objekte über Teil-Identifikatoren zusammen mit der PID der sie enthaltenden „größeren“ Ressource referenziert werden können. Performanz und die Möglichkeit, die Objekte zu verwalten, sind hier von

2) Das OAIS-Modell beschreibt verschiedene Arten von Paketen, die man verwenden kann, um gesammelte Ressourcen als identifizierbare Objekte in einem Archiv abzulegen. Darüber entscheidet der Ersteller eines Archivs auf der Grundlage des Inhalts oder aufgrund von Datenverwaltungsbedürfnissen. Der METS [20] Standard unterstützt z. B. den Begriff „Paket“ in diesem Sinne.