nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

13.2.2 DOI

Herausgeber:

Heike Neuroth
Hans Liegmann †
Achim Oßwald
Regine Scheffel
Mathias Jehn
Stefan Strathmann

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Im Auftrag von:

nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de

Kontakt:

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
neuroth@sub.uni-goettingen.de
Tel. +49 (0) 55 1 39 38 66

13.2.2 Der Digital Objekt Identifier (DOI) und die Verwendung zum Primärdaten-Management

Dr. 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)[23].
Der DOI-Name ist ein dauerhafter persistenter Identifier, der zur Zitierung und Verlinkung von elektronischen Ressourcen (Texte, aber Primärdaten 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[24]) 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-Sys-

23 http://www.doi.org/

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

tem 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 Vorteile 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-Namen 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 12/07 gibt es 8 DOI-Registrierungsagenturen, die teilweise kommerzielle, teilweise nicht-kommerzielle Ziele verfolgen. Bei den Agenturen handelt es sich um

- Copyright Agency Ltd[25], CrossRef[26], mEDRA[27], Nielsen BookData[28] und

25 http://www.copyright.com.au/
26 http://www.crossref.org/
27 http://www.medra.org/
28 http://www.nielsenbookdata.co.uk/

R.R. Bowker[29] als Vertreter des Verlagswesens,
- Wanfang Data Co., Ltd[30] als Agentur für den Chinesischen Markt,
- OPOCE (Office des publications EU)[31], dem Verlag der EU, der alle offiziellen Dokumente der EU registriert
- Technische Informationsbibliothek (TIB) als nicht-kommerzielle Agentur für Primärdaten 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 Primärdaten

Der Zugang zu wissenschaftlichen Primärdaten 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 Primärdaten eingebettet in einen singulären Forschungsprozess, ausgeführt von einer definierten Gruppe von Forschern, geprägt von einer linearen Wertschöpfungskette:

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

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 13.2.2.1), 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

29 http://www.bowker.com/
30 http://www.wanfangdata.com/
31 http://www.publications.eu.int/

zu Wissen gestaltet wird.

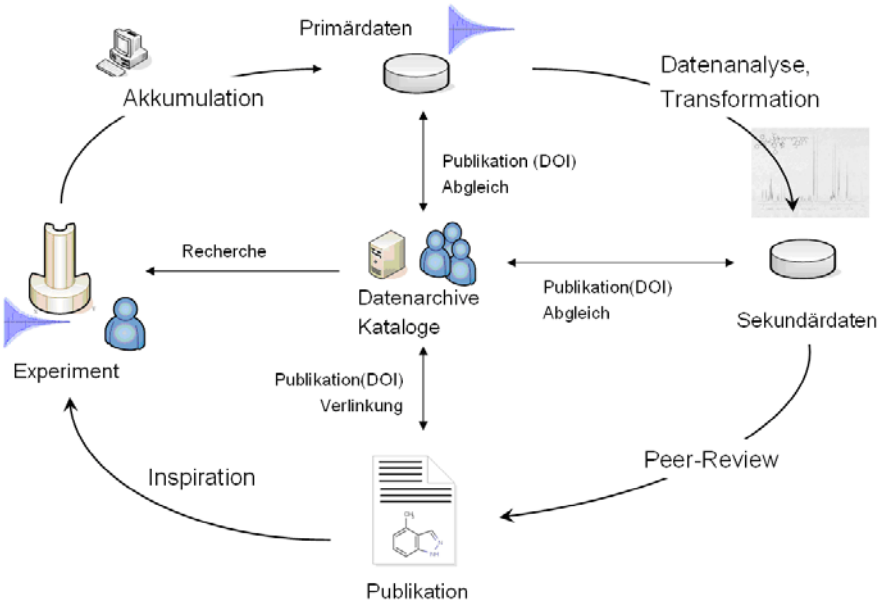

Abbildung 13.2.2.1: Ein idealer Zyklus der Information und des Wissens

Der DFG-Ausschuss „*Wissenschaftliche Literaturversorgungs- und Informationssysteme*“ hat 2004 ein Projekt[32] gestartet, um den Zugang zu wissenschaftlichen Primärdaten 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 Primärdatensätze 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 500.000 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.

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

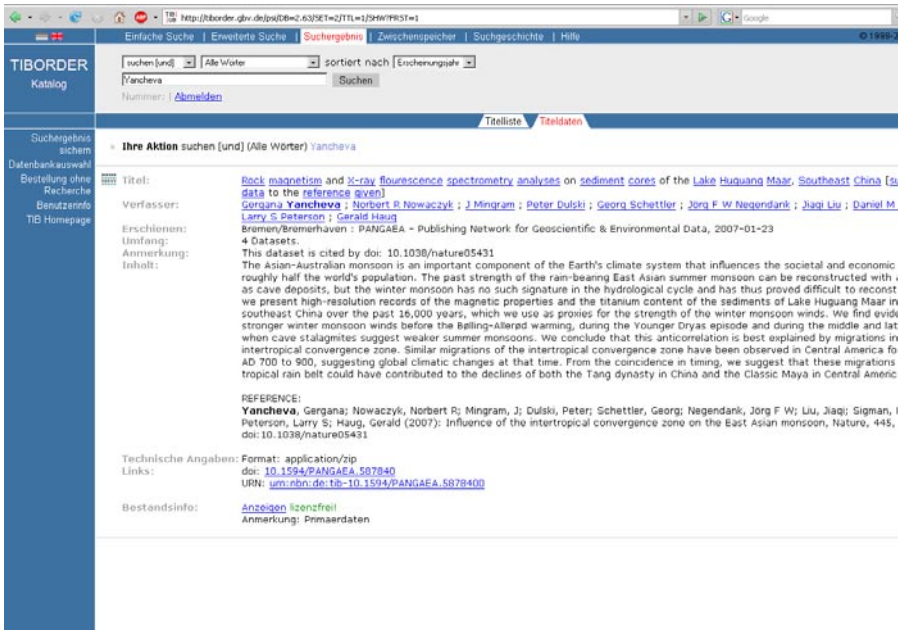

Abbildung 13.2.2.2: Anzeige eines Primärdatensatzes im Online-Katalog der TIB Hannover

Zusätzlich werden Sammlungen oder Auswertungen von Primärdatensätzen auch in den Katalog der TIB aufgenommen. Die Anzeige eines Primärdatensatzes im Katalog der TIB sehen sie in Abbildung 13.2.2..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. (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[33]-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

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

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.

Status

Die DOI-Registrierung von Primärdaten ermöglicht eine elegante Verlinkung zwischen einem Wissenschaftlichen Artikel und den im Artikel analysierten Primärdaten. Artikel und Datensatz sind durch die DOI in gleicher Weise eigenständig zitierbar.

So wird beispielsweise der Datensatz:
G.Yancheva, . R Nowaczyk et al (2007)
Rock magnetism and X-ray flourescence spectrometry analyses on sediment cores of the Lake Huguang Maar, Southeast China, PANGAEA
doi:10.1594/PANGAEA.587840

in folgendem Artikel zitiert.
G. Yancheva, N. R. Nowaczyk et al (2007)
Influence of the intertropical convergence zone on the East Asian monsoon
Nature 445, 74-77
doi:10.1038/nature05431

Mittlerweile hat die TIB ihr Angebot auch auf andere Inhaltsformen ausgeweitet.[34] Als Beispiele seien hier genannt:

- doi:10.1594/EURORAD/CASE.1113 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 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

34 Weitere Informationen zu den Aufgaben der TIB als DOI-Registrierungsagentur und dem Nachweis von Primärdaten 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/primaerdaten/

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-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, J., 2004. Using Digital Library Techniques - Registration of Scientific Primary Data. Lecture Notes in Computer Science, 3232: 488-494.

International Organisation for Standardisation (ISO). 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