

Konzept zur Zitierfähigkeit Wissenschaftlicher Primärdaten

Michael Lautenschlager

(Weltdatenzentrum für Klima, Max-Planck-Institut für Meteorologie, Hamburg)

Irina Sens

(Technische Informationsbibliothek und Universitätsbibliothek Hannover)

Beitrag eingereicht bei und akzeptiert von

"Information – Wissenschaft und Praxis"

Zusammenfassung

Im Rahmen einer DFG finanzierten CODATA Arbeitsgruppe wurden Defizite in der Verfügbarkeit wissenschaftlicher Daten, insbesondere zur interdisziplinären Datennutzung, identifiziert und die Ursachen analysiert. Ein Konzept zur Verbesserung der elektronischen Datenbereitstellung, basierend auf dem DOI, wird vorgestellt und Möglichkeiten zur Umsetzung aufgezeigt. (Abkürzungen im Text sind im Anhang aufgelistet.)

1. Einführung in die Problematik

Der Zugang zu geeigneten Daten ist eine grundlegende Voraussetzung für die wissenschaftliche Arbeit vor allem in den Naturwissenschaften. Deshalb ist es notwendig, bestehende und zum Teil auch neu aufkommende Einschränkungen bei der Datenverfügbarkeit zu vermindern. Vom International Council of Scientific Unions (ICSU) wurde dazu ein Committee on DATA for Science and Technology (CODATA) gegründet, das sich sowohl mit technischen als auch mit politischen Aspekten des

Datenzugangs befasst. CODATA hat unter anderem eine ad-hoc Gruppe gegründet, die sich mit dem Konflikt zwischen einer Verbesserung des Schutzes geistigen Eigentums (gefördert durch die World Intellectual Property rights Organisation, WIPO) und dem traditionell freien Zugriff auf Daten in einigen Wissenschaftsbereichen auseinandersetzt. In diesem Zusammenhang sind beispielsweise Entwicklungen wie der urheberrechtliche Schutz von Datenbanken (z.B. durch die Datenbank-Richtlinie der Europäischen Union) oder die Kommerzialisierung bisher staatlicher Datenproduzenten zu erwähnen, die letztendlich den Zugriff auf Daten durch finanzielle und administrative Hürden erschweren.

Defizite in der Bereitstellung von Daten insbesondere zur interdisziplinären Datennutzung waren vom deutschen CODATA Landesausschuss identifiziert worden. Er hat daraufhin gemeinsam mit dem DFG-Ausschuss "Wissenschaftliche Literaturversorgungs- und Informationssysteme" die Gründung einer Arbeitsgruppe¹ befürwortet, die dann von der DFG für ein Jahr gefördert wurde. Im Folgenden wird der Vorschlag dieser Gruppe zur Verbesserung des Zugangs zu wissenschaftlichen Daten dargestellt. Unter „Wissenschaftlichen Daten“ sollen hier Primärdaten zusammen mit ihrer Beschreibung (Metadaten) verstanden werden.

Beispielhaft seien an dieser Stelle Primärdaten aus dem Bereich der Klima- und Umweltforschung aufgeführt. Hier werden unterschieden große Datenmengen homogener Struktur, wie sie sich aus Klimamodellrechnungen und Satellitenmessungen ergeben, und kleiner Datenmengen inhomogener Struktur, wie sie bei Instrumentenbeobachtungen einzelner Messstationen erhoben werden. Schwierigkeiten in der interdisziplinären Datennutzung fokussieren sich bei den Modell- und Satellitendaten auf Datenvolumina im Terabyte-Bereich und die binären Datenformate, die zur Archivierung verwendet werden, und bei den Daten aus

¹ Mitglieder: Carola Kauhs (Bibliotheksleitung, Max-Planck-Institut für Meteorologie, Hamburg), Dr. Michael Lautenschlager (AG-Sprecher und Direktor Weltdatenzentrum für Klima; Max-Planck-Institut für Meteorologie, Hamburg), Dr. Manfred Reinke (Wissenschaftliche Informationssysteme, Stiftung Alfred-Wegener-Institut für Polar- und Meeresforschung, Bremerhaven), Prof. Dr. Gerhard Schneider (Leiter Rechenzentrum, Universität Freiburg), Dr. Irina Sens (Stellvertretende Leiterin, Technische Informationsbibliothek und Universitätsbibliothek Hannover), Dr. Uwe Ulbrich (Institut für Geophysik und Meteorologie der Universität zu Köln), Dr. Joachim Wächter (Leiter Daten- und Rechenzentrum, GeoForschungsZentrum Potsdam)

Instrumentenbeobachtungen auf die heterogenen Datenbeschreibungen und die Zersplitterung in viele, geographisch verteilte Archive.

Im wissenschaftlichen Bereich besteht zwar grundsätzlich Bereitschaft, Daten für eine interdisziplinäre Nutzung zur Verfügung zu stellen, aber es ist zur Zeit unüblich, dass die erforderliche Mehrarbeit für Aufbereitung, Kontextdokumentation und Qualitätssicherung im Wissenschaftsbetrieb anerkannt wird. Die klassische Form der Verbreitung wissenschaftlicher Ergebnisse ist ihre Veröffentlichung in Fachzeitschriften, normalerweise ohne Veröffentlichung der zugrunde liegenden Daten. Derartige Zeitschriftenartikel werden im "Citation Index" erfasst. Dieser Index wird zur Leistungsbewertung von Wissenschaftlern herangezogen. Datenveröffentlichungen werden darin bisher nicht berücksichtigt.

Projektdateien sind breit über Forschungsinstitute verstreut und werden von Wissenschaftlern erhoben und verwaltet. Aufgrund der fehlenden Anerkennung der mit der Aufbereitung verbundenen Arbeit sind Projektdateien häufig schlecht dokumentiert und somit schwer zugänglich sowie nicht langfristig gesichert. Große Datenbestände bleiben ungenutzt, da sie nur einen kleinen Kreis von Wissenschaftlern bekannt und zugänglich sind. Viele Primärdaten bleiben ungenutztes Rohmaterial.

Die Diskussion zur Fälschung wissenschaftlicher Ergebnisse führte zur Verabschiedung der Regeln guter wissenschaftlicher Praxis in den Wissenschaftseinrichtungen wie DFG, HGF, MPG und Universitäten. Die Regeln beinhalten auch Richtlinien für den Datenzugang. Primärdaten einer Veröffentlichung müssen mindestens 10 Jahre gespeichert und zugänglich sein, um eine Prüfung der Ergebnisse zu ermöglichen. Zwar werden diese Vorschriften im Regelfall eingehalten, aufgrund der damit verbundenen Zeitbelastung werden Daten aber normalerweise nur in Rohform archiviert und nicht in ihrer Feinstruktur aufgearbeitet und erschlossen.

Die wichtigsten Ziele eines neuen Umgangs mit Primärdaten sind also langfristige und allgemein zugängliche Speicherung. Durchsetzbar ist dies am besten über eine

persönliche Motivation der Wissenschaftler. Dies ließe sich nach Meinung der Arbeitsgruppe durch zwei Faktoren befördern: (1) Daten sind nach diesem Konzept nicht mehr ausschließlich Teil einer wissenschaftlichen Veröffentlichung, sondern besitzen eine eigenständige Identität. (2) Damit werden Primärdaten, ähnlich wie Zeitschriftenartikel zitierbar.

Mit der anerkannten Datenpublikation erhält ein Autor also eine zitierfähige Veröffentlichung. Zeitschriftenartikel, welche die Daten verwenden, verweisen auf die Datenpublikation. Umgekehrt kann auch von den publizierten Daten auf Artikel in Zeitschriften verwiesen werden, die den Datensatz verwenden. Die Publikation von Daten kann also sinngemäß in das bestehende System von wissenschaftlichen Veröffentlichungen und deren Zitierbarkeit eingebunden werden. Teile des Konzepts sind unter anderem im Bereich der Kristallographie und der Genetik verwirklicht. Die dort bestehenden und anerkannten Systeme der Datenpublikation werden durch die entsprechenden Fachzeitschriften und deren Verlage gefördert. Urheberrecht und Zugang zu den Daten sind allerdings eingeschränkt.

2. Identifikation von Daten über Persistent Identifier

Zur eindeutigen und zuverlässigen Identifizierung von Primärdaten ist ein Authentifizierungssystem – vergleichbar im gedruckten Bereich mit der ISBN oder im Datenbankbereich mit der CAS-Nummer – notwendig, sogenannte Persistent Identifiers, die den dauerhaften Zugriff auf digitale Objekte ermöglichen.

Für digitale Objekte gibt es dazu derzeit zwei verschiedene Lösungen – DOI und URN - die sich von der Grundidee und der technischen Realisierung wenig unterscheiden.

2.1 Digital Object Identifier

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. Gemanagt wird das System des DOI durch die 1998 gegründete International DOI Foundation (IDF).

Das IDF System besteht aus der „International DOI Foundation“ selbst, der eine Reihe von Registrationsagenturen („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 unter einem gemeinsamen Konzept zu verwalten. Diese gemeinsame Konzepte werden Anwendungsprofile („Application Profile (AP)“) genannt.

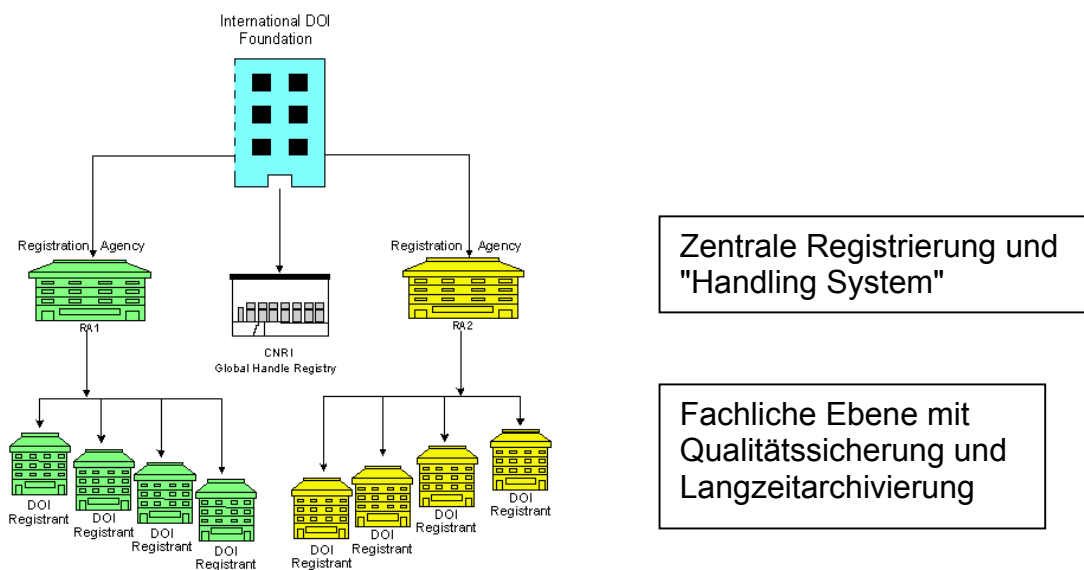


Abbildung: Aufbau des IDF (Quelle: DOI Handbook, URL: www.doi.org/hb.htm)

RAs haben unterschiedliche Anwendungsprofile und unterstützen mit definierten Serviceleistungen die Anbieter beim Einsatz des DOI. Die bekannteste und größte RA ist CrossRef es gibt derzeit sechs weitere (Stand: Juli 2003). Die RAs arbeiten inhaltlich unabhängig voneinander.

Das System des DOI kann für unterschiedliche Datentypen und Anwendungen genutzt werden. Häufig werden laienhaft DOI und CrossRef synonym verwendet. CrossRef ist die bisher größte Registrierungsagentur.², Sie unterstützt das Verlinken von Zitaten in elektronischen Verlagspublikationen über die Grenzen verschiedener Verlage hinaus Da die Nutzung wissenschaftlicher Primärdaten eine andere Anwendung darstellt, scheidet CrossRef als RA aus. Es sei ausdrücklich darauf

² CrossRef wurde Anfang 2000 als Konsortium von Zeitschriftenverlagen gegründet und umfasst heute mehr als 200 Herausgeber. Nähere Informationen finden sich unter www.crossref.org.

hingewiesen, dass die Verwendung von DOIs unabhängig von kommerziellen Verwertungen ist, sie aber generell ermöglicht.

Ein DOI ist eine eindeutige unintelligente und einfach strukturierte Zahlenfolge, die einer digitalen Einheit jeglicher Größe und Granularität sowie jedem Dateityp zugewiesen werden kann. Er bleibt während seiner Lebensdauer unverändert und sagt nichts über den Inhalt der Einheit aus. Er setzt sich aus zwei Komponenten zusammen, die als Präfix und Suffix bezeichnet und durch einen Schrägstrich voneinander getrennt werden. Den individuellen Teil im alphanumerischen Suffix bestimmt der Anbieter selbst. Bereits bestehende (beispielweise hausinterne) Identifikationssysteme können verwendet werden. Gewährleistet werden muss auf jeden Fall die Eindeutigkeit eines DOI in Kombination mit Präfix und Suffix. Eine Überprüfung der Eindeutigkeit findet bei der Eingabe des DOI in das zentrale Verzeichnis statt.

Geht eine mit einem DOI gekennzeichnete Einheit in andere Besitzverhältnisse über oder erhält sie eine neue Platzierung im Web, so sind gemäß der vertraglich geregelten DOI-Richtlinien die Verweise zu aktualisieren. Bei der Auflösung eines DOI durch das zentrale Verzeichnis werden dann später diese Änderungen wiedergegeben. Zu einem DOI gehören strukturierte, öffentlich zugängliche Metadaten. Ein nicht Mindestsatz ist von der IDF fest vorgegeben. Für einen Primärdaten-DOI werden weitere notwendige Metadaten bestimmt werden. Diese werden von der RA in ihrem Anwendungsprofil festgelegt. Der Zweck der Metadaten ist dabei das über einen DOI eindeutig identifizierte Objekt formal und inhaltlich zu beschreiben, da der DOI selbst keinerlei Informationen dazu enthält.

Ein DOI ist prinzipiell aus Präfix / Suffix aufgebaut: **10.1007/s102360100001**. Hinter dem Beispiel verbirgt sich ein Zeitschriftenartikel aus "Ocean Dynamics" mit dem Titel "Momentum transfer at the ocean–atmosphere interface: the wave basis for the inertial coupling approach" veröffentlicht von John A. T. Bye und Jörg-Olaf Wolff.

Wenn ein Nutzer einen DOI³ aufruft, wird eine Nachricht an das zentrale Verzeichnis gesendet, in dem die mit diesem DOI assoziierte aktuelle Adresse gespeichert ist. Diese Adresse wird dem Nutzer übermittelt und ermöglicht zum heutigen Zeitpunkt beispielsweise in einem Browser die Umleitung auf die Internet-Adresse des digitalen Objekts. Die Auflösung des DOI ist aber prinzipiell technologieunabhängig und damit nach heutiger Erkenntnis zukunftssicher.

2.2 Uniform Resource Name

Der gleichen Ansatz wie mit den DOI wird mit den Uniform Resource Names (URN) verfolgt. Die Federführung bei der Betreuung des URN-System liegt bei der Internet Engineering Task Force (IETF)⁴. Die IETF ist der Zweig des Internet Architecture Board (IAB), der für die Entwicklung der Protokolle, deren Implementierung und Standardisierung verantwortlich ist. Seit Januar 1986 hat sich die IETF in eine große, offene und internationale Gemeinschaft von Designern, Administratoren, Herstellern und Forschern entwickelt, die sich mit der Evolution der Internet-Architektur und der reibungslosen Operation des Internets beschäftigen.

Im Rahmen des Global-Info-Projektes CARMEN-AP4 „Persistent Identifiers and Metadata Management in Science“ wurde die Verwendung von URNs an der Deutschen Bibliothek eingeführt⁵, um Online-Dissertationen zu erfassen und zu referenzieren.

2.3 Präferiertes System

Generell sind beide Lösungen verwendbar. Die Bevorzugung der DOIs im Rahmen von CODATA beruht auf drei Gründen:

- Der existierende Einsatz von DOIs bei Crossref schafft eine einfache Vernetzung zu den Verlagspublikationen und damit eine potenziell höhere Akzeptanz bei den Wissenschaftlern.
- Das System ist länger etabliert und international verbreitet.
- Es besteht ein globaler Auflösungsmechanismus für DOIs, der zentral gepflegt und vorgehalten wird.

³ Ein zentrales Auflösungssystem (Resolver) ist unter <http://dx.doi.org/> zu finden.

⁴ <http://www.ietf.org/>

⁵ Ausführliche Infos unter http://www.bis.uni-oldenburg.de/carmen_ap4/index.html

3. Modell für die Umsetzung

3.1 Eine Registrationsagentur für wissenschaftliche Primärdaten

Eine nicht-kommerzielle Registrierungsagentur für wissenschaftliche Primärdaten soll die Möglichkeit für wissenschaftliche Datenzentren schaffen, sich für die Vergabe von DOIs für die Publikation der von ihm verwalteten wissenschaftlichen Primärdaten registrieren zu lassen (DOI Registrant). Nach einem erfolgreichen Vertragsabschluss mit der Registrationsagentur wird jedem Datenzentrum von dieser ein Präfix zugeteilt. Die Registrierungsagentur erhält zu einem publizierten Datensatz vom entsprechenden Datenzentrum den diesem Datensatz zugewiesenen DOI und die zugehörigen Datenbeschreibungen, die im Anwendungsprofil der Registrationsagentur definiert sind, und verwaltet sie mit Hilfe der Technologie des „Handle System®“. Voraussetzung für die Vergabe eines DOI Präfixes an ein Datenzentrum ist die Einhaltung von Qualitätsrichtlinien, die für alle DOI Registranten verbindlich im Anwendungsprofil der Registrationsagentur festgelegt werden. Es ist daher notwendig, eine eigene RA für wissenschaftliche Primärdaten einzurichten, um die Anforderungen der Wissenschaft umsetzen zu können. Der direkte Kontakt mit dem Wissenschaftler bzw. einem Institut findet auf der Ebene der Registranten statt. Hier ist die fachliche Ebene mit Qualitätsprüfung und Langzeitarchivierung angesiedelt.

Die Spezifikationen der Anforderungen, die an Datendokumentation und Archivierung für die DOI-Vergabe gestellt werden, sind abhängig von den einzelnen Fachdisziplinen. Datenzentren in diesem Zusammenhang sind nicht notwendigerweise klassische Rechenzentren, die Daten mit Hilfe eines entsprechenden Maschinenparks verwalten, es können auch Wissenschaftseinrichtungen oder Institutionen sein, die den Besitz von Daten verwalten, das technische Handling aber Dritten übertragen haben, damit aber nicht von der Verpflichtung entbunden sind, die Konsistenz der Daten und die vertraglichen Verpflichtungen bzgl. DOI gegenüber der Registrationsagentur einzuhalten und zu überwachen.

3.2 Exemplarische Implementierung

Die Definition des Anwendungsprofils "Scientific-and-Technical-Data DOI" und damit auch die Aufgaben der DOI-Registranten werden im Rahmen eines Implementierungsprojektes erfolgen. Die Arbeitsgruppe wird dieses Projekt auch weiterhin tragen und begleiten. Die Technische Informationsbibliothek (TIB⁶), Hannover, könnte die Aufgabe der zentralen Registrierungsagentur übernehmen, die in der Gruppe vorhandenen Langzeitarchive (Weltdatenzentrum Klima⁷, Weltdatenzentrum MARE⁸ und Datenzentrum des GFZ⁹ Potsdam) würden als DOI-Registranten auf Fachebene wirken. Dieses Konzept wurde auf der März 2002-Sitzung des DFG-Bibliotheksausschusses diskutiert und positiv bewertet. Dabei wurde der Implementierungsrahmen auf den Bereich Geodaten beschränkt, da hier die meisten Erfahrungen der Gruppe vorliegen und Langzeitdatenarchive zur Verfügung stehen.

Die Einrichtung einer RA ist mit Kosten verbunden. Dieses betrifft zum einen Mitgliedskosten im IDF, Franchisekosten zur Nutzung des Handle System® der „Corporation for National Research Initiatives (CNRI)“ sowie Personal- und Sachkosten für die Bereitstellung und Entwicklung des DOI Systems für wissenschaftliche Primärdaten. In der exemplarischen Implementierung sollen die inhaltlichen und wirtschaftlichen Randbedingungen des Aufbaues einer RA detailliert ermittelt und beschrieben werden. Hinzu kommen die Entwicklungen auf Registrantenseite zur Syntaxkontrolle, Sicherstellung der Langzeitarchivierung und Pflege des Systems.

Ziel der exemplarischen Implementierung ist, praktische Erfahrungen bei der Einführung und im Umgang mit dem System "Primärdaten- DOI" in der Verantwortung der Wissenschaften zu gewinnen, Best Practise Beispiele zu erarbeiten und Entscheidungsgrundlagen für eine breite Einführung im Bereich der Wissenschaft zu gewinnen. Der personelle und sachliche Aufwand, der bei der

⁶ URL: www.tib.uni-hannover.de

⁷ URL: www.mad.zmaw.de/wdcc

⁸ URL: www.wdc-mare.org

⁹ URL: dc.gfz-potsdam.de/dc/

Registrierungsagentur dauerhaft entstehen kann, muss während der prototypischen Implementierung ermittelt werden.

Parallel zum DOI-System soll das Konzept der URN beobachtet und mit der DDB abgestimmt werden, so dass eine Entscheidung getroffen werden kann, in welchem Umfang der URN Eingang in das Konzept zur Zitierung von Primärdaten finden kann.

3.3 Erfolgsaussichten der Einrichtung einer Registrierungsagentur

Für die Einrichtung einer RA für naturwissenschaftliche und technische Primärdaten (Scientific and technical data) ist die Anerkennung durch den IDF erforderlich. Erste Kontakte mit dem derzeitigen Direktor Paskin haben ergeben, dass eine grundsätzliche Bereitschaft zu Einrichtung einer entsprechenden RA von Seiten des IDF vorhanden ist.

4. Perspektiven

Die Entwicklung der Informations- und Kommunikationstechnologie schafft neue Möglichkeiten, mit Daten und Informationen umzugehen. Der Wandel bringt gerade für die einzelnen Wissenschaftsdisziplinen besondere Herausforderungen mit sich, denn die bisherigen Formen der Wissensgenerierung, -verteilung und -nutzung verändern sich drastisch. Ein zentraler Grundstein ist der Zugriff auf und die langfristige Verfügbarkeit von wissenschaftlichen Daten.

Die Umsetzung des vorgestellten Konzepts zur Publikation und Zitierfähigkeit von wissenschaftlichen Daten schafft attraktive Anreize für den Wissenschaftler seine Daten auch anderen Forschern und Projekten zur Verfügung zu stellen. Mittelfristig kann so die Verfügbarkeit von hochwertigen Daten bzw. Inhalten wesentlich verbessert und gesichert werden. Das vorgeschlagene Konzept zur Verbesserung der "Zitierfähigkeit wissenschaftlicher Primärdaten" bietet das Potential als Wissenschaftsinfrastruktur breit über Fachdisziplinen hinweg Anwendung zu finden und bildet einen entscheidenden Schritt zur Modernisierung wissenschaftlicher Informationsversorgung. Dabei ist die Anwendbarkeit nicht auf Deutschland beschränkt. Im internationalen Umfeld bieten CODATA und das "World Data Center"

System von ICSU Möglichkeiten zur Umsetzung. Auf europäischer Ebene kommen Infrastrukturprogramme der EU zur Anwendung und bieten sich Kooperationen mit Einrichtungen der Wissenschaftsförderung in Ländern der EU an.

Danksagung

Die Autoren bedanken sich bei der DFG für die Unterstützung der Arbeit, die Förderung der Arbeitsgruppe und die Finanzierung einer Pilotinstallation im Projektrahmen. Den Mitgliedern der Arbeitsgruppe danken wir für ihre intensive und konstruktive Arbeit während des knappen Jahres ihres Bestehens. Herrn R. Bertelmann, Bibliothek des Wissenschaftsparks Albert Einstein (gemeinsame Bibliothek des Geoforschungszentrums Potsdam, des Potsdam-Instituts für Klimafolgenforschung und der Forschungsstelle Potsdam des Alfred-Wegener-Instituts für Polar- und Meeresforschung), danken wir für Korrekturlesen und wertvolle Hinweise.

Abkürzungen

ANSI	American National Standards Institute
BMBF	Bundesministerium für Bildung und Forschung
CAS	Chemical Abstract Service
CNRI	Corporation for National Research initiatives
CODATA	Committee on Data for Science and Technology
DDB	Die Deutsche Bibliothek
DFG	Deutsche Forschungsgemeinschaft
DOI	Digital Object Identifier
GFZ	GeoForschungsZentrum Potsdam
HGF	Hermann von Helmholtz-Gemeinschaft Deutscher Forschungszentren
ICSU	International Council of Scientific Unions
IAB	Internet Architecture Board
IDF	International DOI Foundation
IETF	Internet Engineering Task Force
ISBN	International Standard Book Number
MPG	Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V.
NISO	National Information Standards Institute
PI	Persistent Identifiers
RA	Registrierungsagentur
TIB	Technische Informationsbibliothek
URN	Unified Ressource Locator
WDC	World Data Center
WIPO	World Intellectual Property Rights Organisation