

Eindeutige Autoren-Identifikation – (PAI – Persistent Author Identification) – Versuch einer Annäherung

Helge Steenweg

Im vorliegenden Beitrag wird die Notwendigkeit einer eindeutigen Autorenidentifikation (Persistent Author Identification - PAI) diskutiert und die bisher vorhandenen fachlichen, nationalen, überregionalen und kommerziellen Ansätze vorgestellt und erläutert. Es zeigt sich, dass die Entwicklung auf dem Sektor der Personennormdateien (authority records) mittels VIAF und ISNI auf einem guten Wege ist. Bei der notwendigen Zusammenführung der bisherigen Autoren-Identifikatoren in einem übergeordneten System könnte der neuen ORCID-Initiative große Bedeutung zukommen.

The necessity of an explicit author identification is discussed, and the already existing attempts in various disciplines, on national and international levels as well as commercial ideas are presented and commented. It can be shown that the developments in the field of authority records on the basis of VIAF and ISNY indicate the right direction. The necessary merge of existing author identity elements into to a new system with higher priority could be importantly.

Einleitung [1]

Beschäftigt man sich mit automatisierten Erfassungen von Publikationen oder der Verlinkung von Open Access (OA) Publikationen, so treten sehr schnell Probleme bei der Zusammenführung von Publikationen mittels Autorennamen auf. Dieses Problem ist nicht neu, schon bei jeder Katalogisierung gilt es, unterstützt von Normdatenbanken Zuordnungen zu treffen.

Der Name ist dabei ein Identifizierungsmerkmal (identifier - ID) für eine Person, hier im Speziellen für den Autor. Weitere zusätzliche Merkmale können Beruf, Wohnort, Arbeitsstätte, Kollegen (Co-Autoren) sein. Diese Merkmale werden benötigt, wenn der Name nicht eindeutig ist.

Als generelle Probleme bei Namen können auftreten:

1. Differierende Namensschreibweisen:
 - ⊙ Namen variieren (durch Eingabefehler, fehlende oder falsch interpretierte Richtlinien)

- ⊙ Schreibweise differiert (durch diakritische Zeichen, Akzente, Umlaute, - Probleme der „pre-Unicode-Ära“)
- ⊙ Wechselnde Nennung von Zweitvornamen bzw. Initialen (nicht immer angegeben, Zweit- und Drittornamen nicht immer benutzt)
- ⊙ Abkürzungen, Ehrennamen, Kosenamen (z.B.: Gandhi-Gandhiji)
- 2. Namenswechsel (Heirat)
- 3. Benutzung von Pseudonymen
- 4. elektronische Leseweise durch Worttrennungen oder eingestreute Formatierungszeichen (HTML-Code) gestört
- 5. falsche Zuordnungen von Publikationen durch Namensgleichheit (ev. lösbar durch Umfeldanalyse)
- 6. verschiedene Formate derselben Publikation (PDF, DOC, RTF, HTML etc.) auf verschiedenen Dokumentservern sind bei automatisierter Zusammenführung unterschiedliche Dokumente

Um welche Zahlenmengen geht es?

Eine exakte Zahl für die Ermittlung von Au-

toren-IDs zu benennen fällt extrem schwer, da bei allen Nennungen von Zahlen stets zu berücksichtigen ist, was unter dem jeweiligen Begriff zu fassen ist. Geht man von Zahlen der UNESCO in ihrem neuesten Wissenschaftsbericht [2] aus, so arbeiteten 2007 weltweit 7,2 Millionen Forscher [3]. Weltweit wurden 2008 knapp eine Millionen Publikationen veröffentlicht, eine Steigerungsrate von fast 35 Prozent gegenüber dem Jahre 2002 [4].

Die International Association of Scientific, Technical and Medical Publishers (STM) gibt (Stand Oktober 2009) an, dass sie Beiträge von insgesamt sieben Millionen Autoren von Zeitschriftenartikeln verzeichnet, mit einem Wachstum von 210.000 neuverzeichneten pro Jahr [5]. ISNI [6] schätzt, dass in der aufzubauenden Datenbank zirka 20 Millionen Einträge und 600.000 Neuzugänge pro Jahr verzeichnet werden.

Vorteile

Durch die Einführung einer eindeutigen Autorenidentifikation werden Verknüpfungen von Daten, insbesondere Zitationen und Referenzierungen möglich. Verknüpfungen zwischen Datenbeständen gestalten sich vom Aufwand her deutlich einfacher. Es wäre zudem möglich, Cluster (Friend-to-Friend) (Subject-to-Subject) sichtbar zu machen. Beziehungen zwischen Forschungsbeschreibungen und Forschungsergebnissen (Daten und/oder Publikationen) treten deutlicher hervor.

Über eine Autoren-ID ist es leichter, weitere Publikationen des Autors zu finden, Forschungsnetzwerke über Mitautoren zu erfassen und die Rezeption von Forschungsergebnissen nachzuvollziehen. Die Zusammenstellung eines Portfolios eines Forschers/Forscherin kann unabhängig von seinen in unterschiedlichen kommerziellen Systemen vorhandenen Publikationsdaten erstellt werden. Dabei benutzte Schlagwortwolken (tag clouds) fallen entsprechend exakter aus.

Mittels einer eindeutigen Autoren-ID lassen sich Daten in Forschungsdatenbanken, Forschungsinformationssystemen und Institutional Repositories deutlich leichter verknüpfen. Dubletten können einfacher erkannt und bearbeitet werden.

Eine übergeordnete, frei zugängliche Vergabe dieser IDs macht unabhängiger von kommerziellen Ansätzen dieser Art, die sich häufig abschotten. Auch kleinere Zeitschriften (OA) können über die ID in Zitationslisten einbezogen werden und erhalten dadurch auch für Autoren ein größeres Gewicht (Impact-Faktor). Insgesamt dürften die Zitationsrankings durch eindeutige Zuordnungen mittels Autoren-ID deutlich verbessert werden.

Nachteile

Ein Aspekt, der sicherlich zu berücksichtigen ist, gilt dem Datenschutz. Es muss gewährleistet sein, dass Autoren über ihre individualisierten Daten verfügen können, d.h. insbesondere Entscheidungsmöglichkeiten haben, ob und wie ihre Daten benutzt werden können.

Ein sicherlich anführbares Argument gegen eine umfassende Autorenidentifikation ist, dass in Zusammenhang mit obigem Argument eine Art „Gläserner Forscher“ entsteht – mit Folgen und Auswirkungen auf Dotierungen, Projektfinanzierungen und Image. Wie diese Folgen aussehen (positiv oder negativ) lässt sich allgemein nicht voraussehen und ist sicherlich individuell zu entscheiden.

Nicht von der Hand zu weisen ist, dass durch eine Autoren-ID bibliometrische Auswertungen einfacher werden. Ob Bibliometrie selbst und/oder manche verwandten Methoden innerhalb der Bibliometrie sinnvoll sind oder nicht, ist allerdings eine andere Frage. Mit einer Autoren-ID würden diese Fragen jedoch in jedem Fall genauer zu beantworten sein.

Verschiedene Wege

Es gibt verschiedene Möglichkeiten, wie man sich die Gestaltung von Datenbanken zur Autorenidentifikation vorstellen kann. Ein wichtiges Unterscheidungsmerkmal wäre die Art der Eintragung: Sollen Autoren Ihre Daten selbst eintragen oder sollen Daten automatisiert erfasst werden? Bei einer selbsteintragenden Community müssen Kontrollfunktionen bei der Authentifizierung vorhanden sein. Sollen Kollegen Datensätze Wiki-like ändern dürfen (Frage der Kontrolle, Fehleranfälligkeit).

Andererseits stellt sich die Frage bei automatisierten Daten, inwieweit die verwandten Algorithmen fehlertolerant sind und ob und wann diese Daten editiert werden? Darf der Autor selbst diese Daten editieren und verändern? Gibt es Probleme bei automatisiert erstellten Autorenverknüpfungen [7]?

Es gibt sicherlich weitere Gestaltungsmöglichkeiten, sicher auch Kombinationen – entscheidend für zukünftige Konzepte dürfte aber die Akzeptanz der Autoren („vertrauenswürdig“) sein, die ein solches System nutzen.

Fazit

Es sollte eine Kennzeichnung (numerisch oder alphanumerisch) existieren, die jeden Autor eindeutig identifiziert. Dazu muss eine übergreifende Datenbasis vorhanden sein, die „vertrauenswürdig“ für Autoren wie Leser ist.

Eindeutige Autoren-
Identifikation – (PAI –
Persistent Author
Identification) –
Versuch einer
Annäherung

⊙ Eindeutige Autoren-Identifikation – (PAI – Persistent Author Identification) – Versuch einer Annäherung

Existierende Ansätze

Für Autoren-Identifikations-Systeme gibt es erste Ansätze und für bestimmte Fachgebiete sogar gut funktionierende Lösungen, dennoch stellt die Festlegung, Administration und Bereitstellung einer weltweit eindeutigen Autoren-Identifikation ein globales Problem dar, dass nicht eine Institution oder ein wie auch immer aufgestellter internationaler Verlag oder Datenbankanbieter alleine lösen kann.

Ansätze fachlich

arXiv.org

Im Dokumentenserver arXiv finden sich Preprints aus den Fachgebieten Biologie, Informatik, Mathematik und Physik [8]. Seit dem Jahr 2005 wurden innerhalb von arXiv be-

reits Verknüpfungen mittels „Authority records“ vorgenommen, seit dem Jahr 2009 haben die arXiv-Verantwortlichen einen „public author identifier“ eingeführt [9]. Derzeit müssen die Autoren-Profile vom Autor selbst überprüft werden. Die Notation einer arXiv-Author-ID besteht aus einer kompletten URI, die über einen Webbrowser eingegeben, eine HTML-Seite der Publikationen ausgibt. Gleiches geschieht über einen RSS-Feed (Atom).

Mittels arXiv-Author-ID gibt es ebenso eine arXiv-Facebook Applikation, über die in Facebook eigene Artikel des Autors angeklickt werden können.

RePEC/AuthorClaim

RePEC (Research Papers in Economics) ist ein Zusammenschluss dezentral betriebener Datenbanken, in der Nachweise zu 600.000 Zeitschriftenaufsätze aus dem Bereich der Wirtschaftswissenschaften von zirka 26.000 Autoren [10] gespeichert sind. Autoren können sich selbst eintragen. RePEC ist ein offenes System ohne kommerzielle Vernetzung.

Auf der technischen Basis von RePEC wurde ein allgemeiner Registrierungs-Service für Autoren unter dem Namen AuthorClaim [11] geschaffen. Autoren müssen sich für die Erstellung eines Profils selbst anmelden.

Ansätze national

Nationale Personennormdateien, hier PND – GND

In fast allen Ländern existieren lokale Personennormdateien. In Deutschland sind in der PND (Personennamendatei) [12] individualisierte und nicht individualisierte Namen enthalten. Individualisiert werden die Datensätze durch Angaben zu Lebensdaten, Beruf etc [13]. Zur Zeit liegen von 3,6 Millionen Gesamtbestand etwa 1,8 Millionen als individualisierte Namenseinträge vor.

Die Daten der PND sollen zukünftig in die Gemeinsame Norm Datei (GND) [14] überführt werden. Insbesondere bei zeitgenössischen Autoren ist zu beobachten, dass die Nachweise und Individualisierungen häufig überarbeitungswürdig sind.

DAI

In Zusammenarbeit von SURF [15] und OCLC PICA startete bereits 2005 ein von kommerziellen Interessen losgelöstes Wissenschaftsprojekt in den Niederlanden mit zwei Pilotanwendungen in Groningen und Nijmegen, um in Forschungsinformationssystemen wie dem niederländischen Metis [16] Verknüpfungen zwischen Autoren und ihren Werken darstellen zu können. Auf der Basis des National Thesaurus Authornames

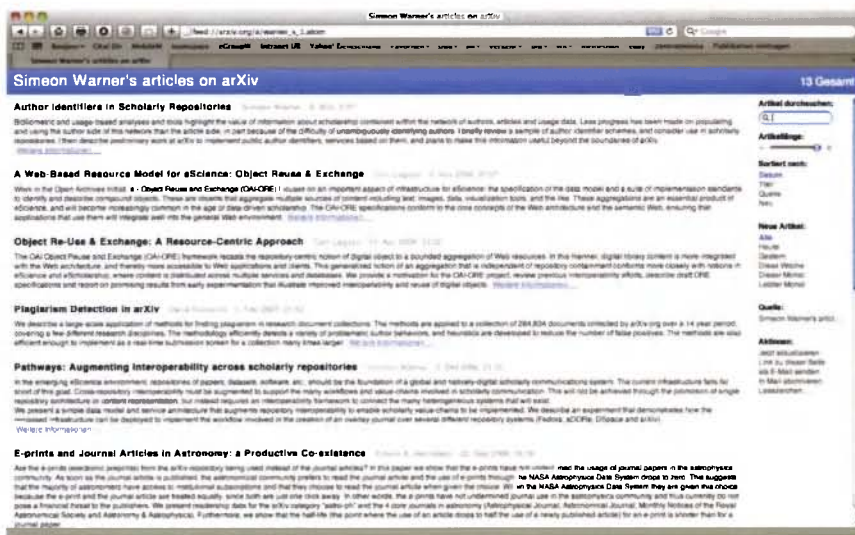


Abb. 1: arXiv – Author Identifier. Anzeige als RSS-Feed

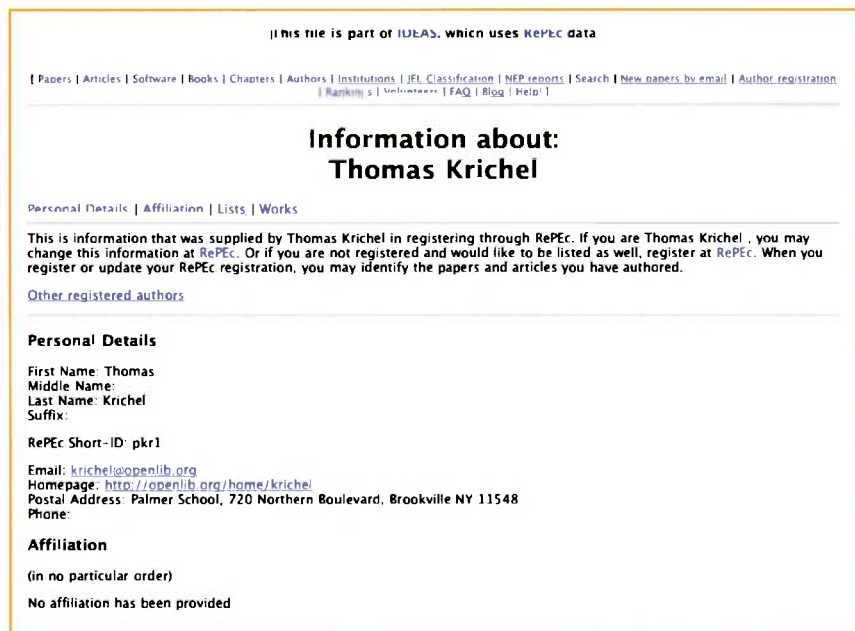


Abb. 2: RePEC Autorensicht



Abb. 3: PND – Suche

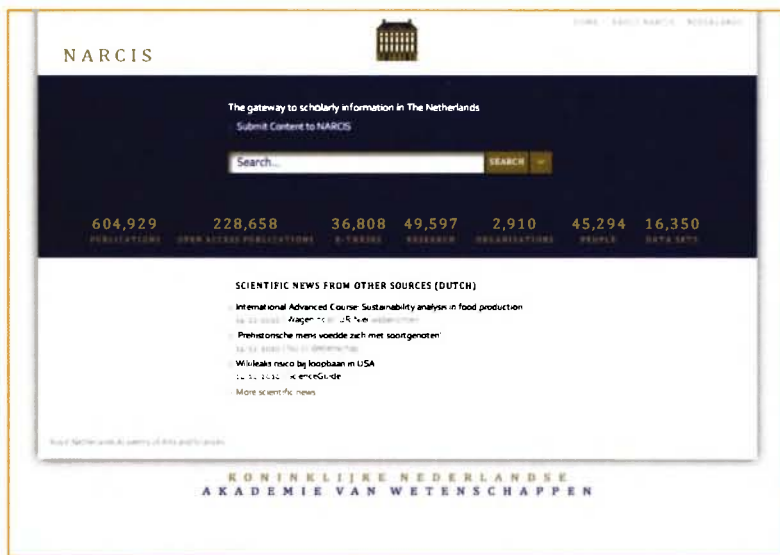


Abb. 4: NARCIS

Abb. 5: NARCIS Autorensicht

(NTA) – vergleichbar der deutschen PND – wurde die Digital Author Identification (DAI) entwickelt, die die Verknüpfung von personenbezogenen Daten mit PICA-Bibliotheks-Daten ermöglicht [17]. Im Jahre 2006 wurde DAI auf weitere 11 Universitäten und andere Forschungseinrichtungen ausgeweitet. Die niederländische Digital Author Identification soll kompatibel zu ISO 27729 (ISNI - International Standard Name Identifier) [18] sein.

Anwendung findet die DAI z.B. in NARCIS [19] (National Academic Research and Collaborations Information System), einem Forschungsinformationssystem, das über 600.000 Publikationen von 45.000 Autoren verzeichnet.

Es ist gedacht, um Forschung sichtbar zu machen, und bietet neben persönlichem Curriculum Vitae der Wissenschaftler auch Angaben zu Forschungsvorhaben und Publikationen, meist mit Zugriff auf OA-verfügbare Dokumente. Die Publikationen können nach Typ, Veröffentlichungszeitpunkt, Zugang (OA) oder Institution gefiltert werden.

Names-Projekt

Das Names-Projekt in England ist ein vom JISC (Joint Information Systems Committee) und der British Library gefördertes „Name Authority System“ [20] für britische Repositorien [21]. Es begann mit Phase I (Förderzeitraum 2007–2009) und wird in jedem Fall bis 2011 (Phase II) [22] mit der Schaffung eines Pilotsystems [23] fortgeführt.

In dem Pilotsystem sind derzeit (2010) etwa eine Million Namensprofile gespeichert, die aus einem Subset der ZETOC-Datenbank der British Library extrahiert wurden [24] und mit anderen Datenquellen angereichert werden. Die Daten können u.a. in HTML, MARC21 und RDF ausgegeben werden.

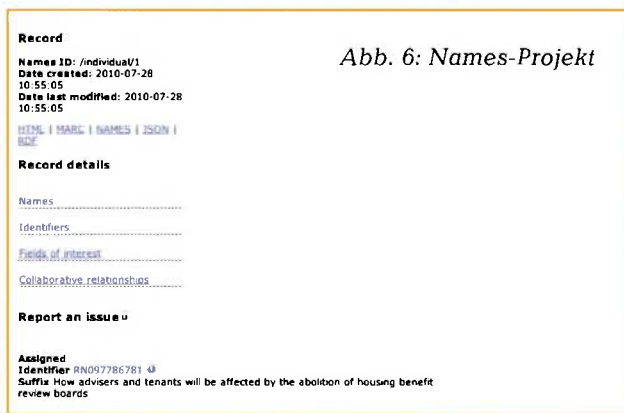
Ansätze übergeordnet/international

VIAF (Virtual International Authority File)

Gestartet als Gemeinschaftsprojekt von einigen Nationalbibliotheken (LOC-Library of Congress, DNB-Deutsche Nationalbibliothek, BNF-Bibliothèque Nationale de France) und OCLC



Abb. 6: Names-Projekt



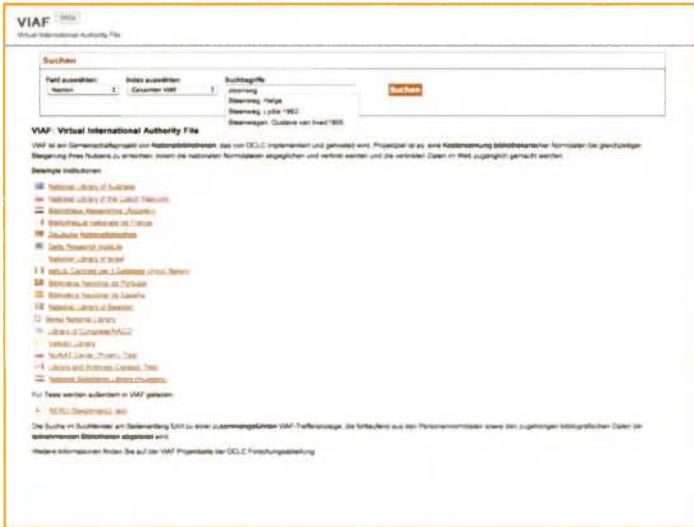


Abb. 7: VIAF-Suche

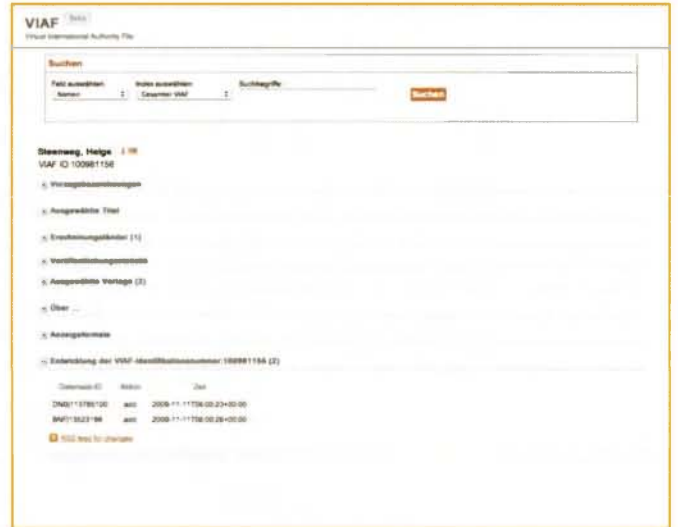


Abb. 8: VIAF-Detailldarstellung

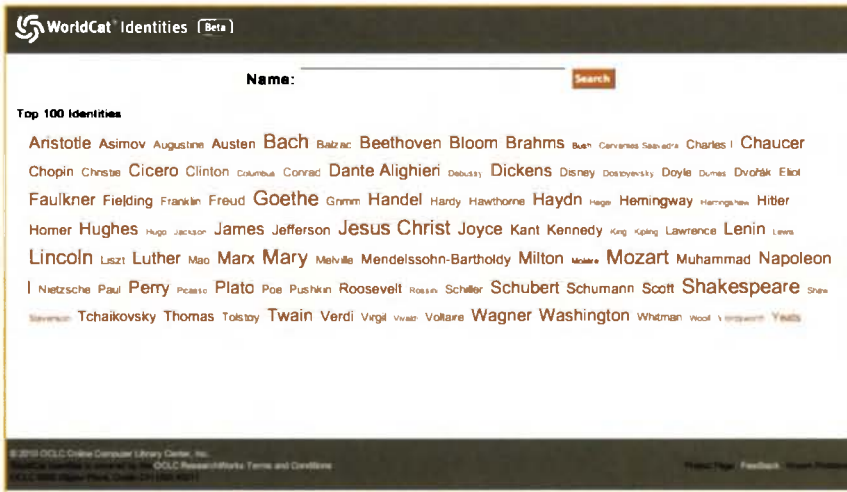


Abb. 9: WorldCat Identities – Autorensicht

mit der Zielsetzung, eine internationale Normdatenbank zu schaffen, umfasst VIAF heute Daten aus 18 Normdatenbanken von 15 Organisationen [25]. Benutzt man das Interface zur Suche, so erscheint die Darstellung in der Sprache des Nutzers [26].

Bei jedem Eintrag sind neben der VIAF-ID und den bibliographischen Angaben zu Publikationen grafisch die Quellen der nationalen Normdateien und ihre geographische Verteilung dargestellt.

Über die enthaltenen ISBNs wird der Publikationsverlauf des Autors [27] dargestellt. Die Datensätze lassen sich im MARC21-, UNIMARC- oder RDF-Format ausgeben.

WorldCat Identities

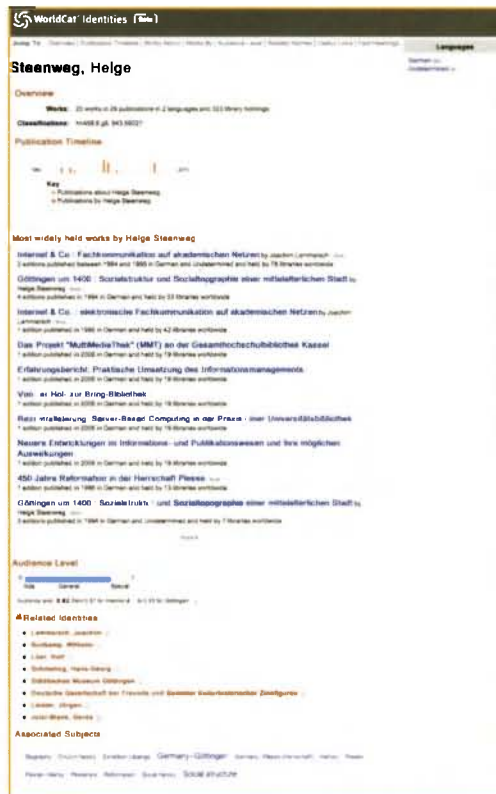
Die Forschungsabteilung von OCLC hat 2007 aus allen Namensnennungen im WorldCat die WorldCat Identities [28] ermittelt und abfragbar gemacht. Die Eingangsseite ist schlicht gehalten und nimmt heutige Nutzererwartungen auf. Die Webseite erhält neben einer Eingabemöglichkeit zur Suchabfrage eine Tagcloud mit den 100 meistabgefragten Namen (top 100 identities):

Die Identity-Nummer setzt sich offensichtlich in Anlehnung an das Library of Congress Name Authority File (LCNAF) zusammen:

lccn-n84-188342

Nach Eingabe eines Namens erhält man eine Liste von Personennamen, wobei auch hier die Größe des angezeigten Namens die Anzahl der Einträge charakterisieren soll. Neben direkten Übereinstimmungen mit dem gesuchten Namen werden auch Namen mit ähnlicher Schreibweise angezeigt, die bei einer ungenauen Suche angezeigt würden. Betrachtet man einen Autor, so kann

Abb. 10: WorldCat Identities – Autorensicht



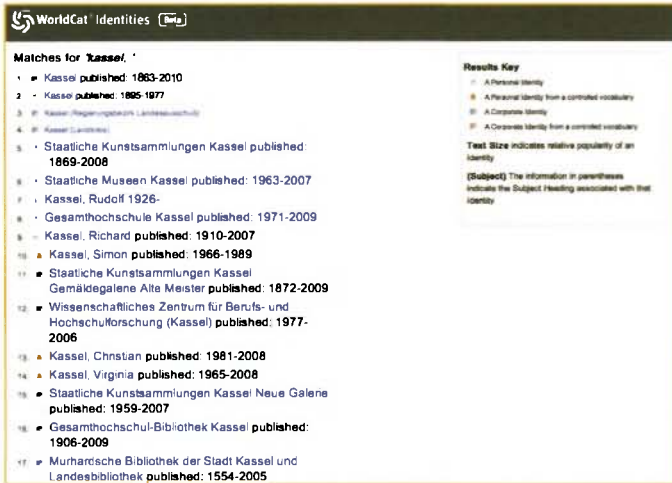


Abb. 11: WorldCat Identities – Suchergebnis



Abb. 12: ISNI - International Standard Name identifier

man sich in der Autorensicht neben einer Kurzstatistik (Publikationen, Sprachen, Bibliotheksnachweise) eine Klassifikationsnotierung, eine Zeitschiene der Publikationen samt detaillierter Auflistung und Co-Autorenschaften bzw. Beziehungen zu Herausgebern anzeigen lassen. Eine Tag-Cloud-Liste beschließt den Eintrag.

Die Daten in den WorldCat Identities sind offensichtlich nicht aktuell aus dem WorldCat übernommen, da dort bereits neuere Publikationen vorhanden sind, die nicht angezeigt werden. Die Zahl der in der Beschreibung des Systems genannten zirka 30 Millionen Namen ist dahingehend zu relativieren, dass neben Autoren auch Institutionen und Körperschaften erfasst sind.

ISNI (International Standard Name Identifier)

Als derzeitiger Draft ISO Standard (ISO 27729) soll der ISNI (International Standard Name Identifier) [29] die Urheber medialer Werke vor allem wegen der daran hängenden Rechte eindeutig bestimmen. ISNI soll nicht umfassende Informationen zum jeweiligen Urhebernamen beinhalten, allenfalls Links zu anderen Informationsquellen. Entstanden ist der Standardisierungsgedanke u.a. durch OCLC, die Bibliothèque Nationale de France und die British Library, die VIAF einbringen [30].

Neben natürlichen Personen können auch sonstige Rechteinhaber (z.B. Record Label) oder fiktive Personen (Peter Pan) eine ISNI bekommen. Schreibvarianten erhalten dieselbe ISNI (Beispiel: Günter Grass, Guenter Grass, Guenter Graß).

Transliterationen von Namen (z.B. aus dem Kyrillischen) erhalten dieselbe ISNI ebenso wie linguistische Namensvarianten (ISNI Beispiel: Pyotr/Peter). Pseudonyme werden als eigene ISNI geführt, können jedoch miteinander verlinkt werden [31].

Der Standard Identifier setzt sich aus 16 Dezimalzeichen zusammen; die letzte Ziffer ist Prüfziffer:

Beispiel: ISNI 1422 4586 3573 0476

Die ISNI-Datensätze sollen folgende Metadaten beinhalten:

- ⦿ Name
- ⦿ Datum und Ort der Geburt und/oder Tod,
- ⦿ Klasse (Musical, Audio-Visual, Literary)
- ⦿ Rolle (Autor, Künstler, Verleger)
- ⦿ Titel
- ⦿ URI/URL für nähere Informationen

Beispiele [32]:

ISNI	Vorname	Nachname	Klasse	Rolle
0000 1000 0000 1774	John	Smith	Literary	Author
0000 1000 0000 115x	Sam	Davis	Musical Work	Composer
0000 1000 0000 1264	West Side	Inc.	Literary	Publisher

Um eine ISNI zu erhalten, benötigt man die Hilfe einer Registration Agency (RAG), die ihrerseits mit der Registration Authority (RA) zusammenarbeitet, die die Datensätze pflegt. Die erste ISNI-Datenbank wird sich aus den Datenbanken der Gründungsmitglieder zusammensetzen. Schätzungen des ISNI-Board-of-Directors gehen davon aus, dass insgesamt bei 600000 Neuzugängen pro Jahr 20 Millionen Namenseinträge anfallen werden bei 600.000 Neuzugängen pro Jahr [33].

Ansätze kommerziell

Microsoft Academic Search (Beta)

Vergleichbar zu Google Scholar hat Microsoft ein akademisches Suchportal aufgebaut, das seit 2009 im Netz zugreifbar ist: Microsoft Academic Search [34]. Derzeit sind 7,4 Millionen Publikationen von 5,9 Millionen Autoren erfasst.

Abb. 13: Microsoft Academic Search – Autorensicht



In der frei zugänglichen Suchmaschine werden bei der Suche Publikations-Metadaten (teilweise mit Link zum Volltext) und Zitationen angezeigt. Die Suchergebnisse können gefiltert werden (Jahr, Art der Publikation). Verbindungen zwischen Wissenschaftlern können im sog. Visual Explorer visualisiert dargestellt werden [35].

Die Autorenprofile können editiert werden, werden aber editorisch überprüft. Vom Autor können bibliographische Daten (BibTeX-Format) hochgeladen werden, auch diese werden erst nach Prüfung verändert. Es sind derzeit knapp sechs Millionen Autoren verzeichnet, jedoch kann sich diese Zahl relativieren, da durch die teilweise automatisierte Übernahme und Erfassung auch Duplikate entstehen, die noch nicht bereinigt sind, wie z.B. bei der Suche nach Tim Berners-Lee ersichtlich ist:

ProQuest COS Scholaruniverse

Proquest bietet mit COS Scholaruniverse [36] Zugriff auf zirka 2,4 Millionen Autorenprofile und verknüpft über den Autorennamen Publikationslisten mit Autorenetzwerkinformationen (Co-Autorenschaften auf unterschiedlichen Leveln, Verknüpfung zu Arbeitskollegen über Institution) [37]. Proquest legt Wert auf die Qualität der Daten und überprüft offensichtlich über eine Editorfunktion die zusammengetragenen bzw. vom Autor editierten Daten [38].

Die Autorensicht bietet neben einer Publikationsliste (inklusive Co-Autoren-Verknüpfung) Angaben zur Arbeitsstätte, zu Titeln bzw. Abschlüssen und Fachgebietseinordnungen auch Informationen zur Anzahl von Co-Autoren und Arbeitskollegen [39].

ResearcherID

Researcher-ID [40] ist ein Produkt von Thomson Reuters [41] und bietet auf der Basis des Web-of-Science (WoS) ein Portal zu Wissenschaftlern und ihren Publikationen an. Die Abfrage erfolgt entweder über eine einfache Abfragemaske bzw. Interaktive Karten. Über die interaktive Landkarte kann die Anzahl von Einträgen pro Land abgefragt werden. Für Deutschland sind aktuell 3483 Researcher-IDs vergeben (zum Vergleich: USA 18299, China 10194, Großbritannien 7392, Spanien 4365, Frankreich 3046, Italien 2472) [42]. In der Autorensicht sind neben der Publikationsliste auch kurze Informationen zum Autor ersichtlich:

Die ResearcherID hat die Form: G-4394-2010, wobei vermutlich die letzten vier Ziffern das Eintragungsjahr bezeichnen. Der Anfangsbuchstabe steht jedoch nicht für das Land, in dem der Autor beheimatet ist. Über einen Reiter auf der Ergebnisseite kann man, sofern der Autor seine Daten für diese

Abb. 14: Microsoft Academic Search – Visualisierung von Co-Autorenschaften

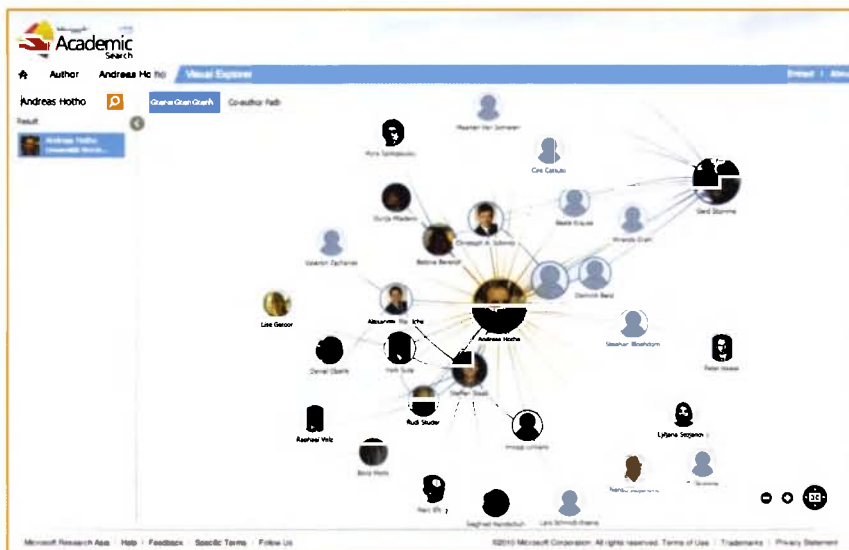


Abb. 15: Microsoft Academic Search – Forschungsumgebung

Eindeutige Autoren-Identifikation – (PAI – Persistent Author Identification) – Versuch einer Annäherung



Abb. 16: Proquest COS Scholaruniverse - Autoranzeige



Abb. 17: ResearcherID Autorensicht

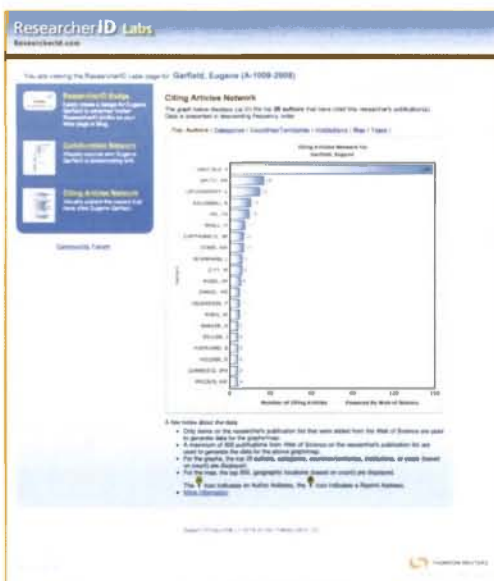


Abb. 18: ResearcherID Autorensicht

Art der Auswertung freigegeben hat, zu den sog. ResearcherID-Labs, einer Tool-Sammlung, die Analysen zur Forschungsumgebung ermöglicht, gelangen. Zum einen werden Forschungsbeziehungen des Autors visualisiert (nach Namen, Institutionen, kartografischer Darstellung oder Forschungsfach), zum anderen nach gleichen Kriterien die Zitationsbelege.

Mittlerweile kann sich ein Autor auch selbst registrieren („request an invitation“). Entsprechende Metadaten werden dann standardmäßig über WoS hinzugefügt, allerdings besteht auch die Möglichkeit, bibliographische Daten über eine Datei hochzuladen. Diese Datei muss dann im RIS-Format vorliegen, einem Dateiformat der Thomson Reuters eigenen Literaturverwaltungssoftware Endnote [43].

ScopusID

Innerhalb des Scopus-Systems [44] vergibt Elsevier die so genannte ScopusID (Scopus Author Identifier, eine „unique identifier number“ für Autoren, die in Scopus gelistet werden [45].

Daten werden diesem Identifier automatisiert zugeordnet mittels Algorithmen, wobei beim Matching weitere Identifizierungsmerkmale wie Beruf, Arbeitsstätte, Fachgebiet bzw. Co-Autorenschaften hinzugezogen werden. Sofern das automatisierte Matching keine 99-prozentige Übereinstimmung erbracht hat, erfolgt keine Verknüpfung. Laut Aussage des Herstellers können allerdings 95 Prozent der bearbeiteten Fälle automatisiert zugeordnet werden [46].

Namen und Namensabkürzungen wie „Lewis, M.“ „Lewis, Michael“ und „Lewis, M.L.“ werden dabei unter einer ID zusammengefasst. Neben diesen Initialauflösungen der Vornamen soll der automatische Matching-Algorithmus ebenfalls in der La-

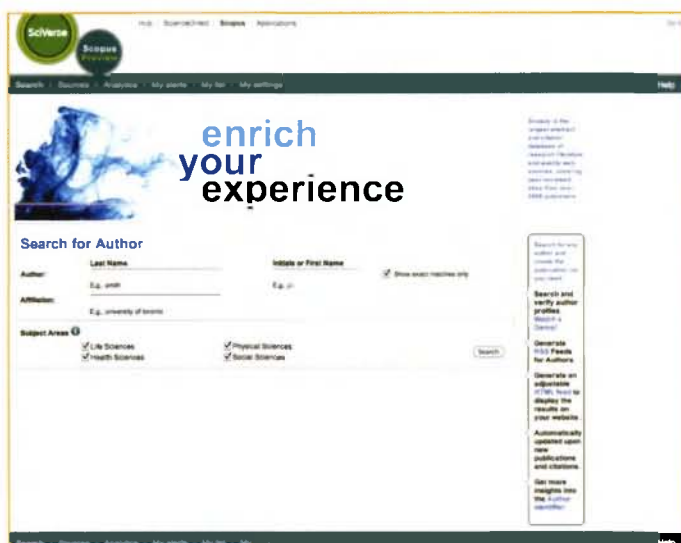


Abb. 19: Scopus

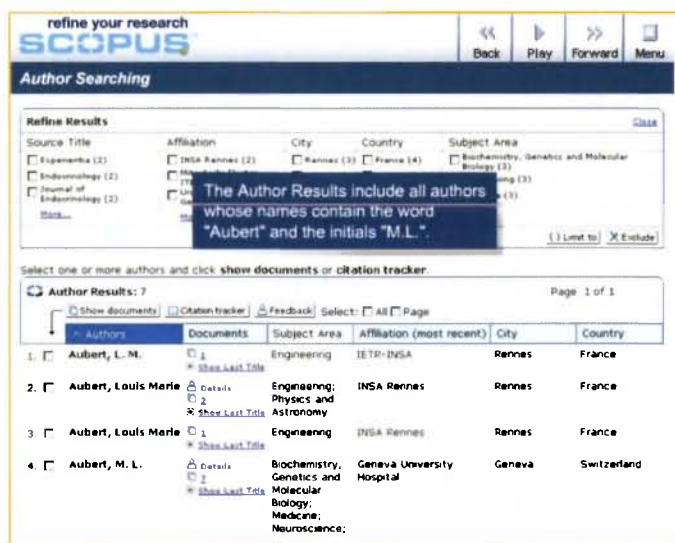


Abb. 20: Scopus – Suche

ge sein, internationale Namensvarianten, alternative Schreibweisen und Variationen des Vor- und Nachnamens zu erkennen und zu einer ID zusammenzuführen [47].

Tabelle der Kennungen

Stellt man – soweit möglich – die Darstellungsformen der Autoren-Identifikation in einer Tabelle dar, so wird schnell klar, dass die einzelnen Ansätze auch in der Form stark differieren.

Betrachtet man die jeweils genannten Autorenzahlen, so fällt auf, dass für viele Autoren bereits Identitäten vergeben wurden, wahrscheinlich haben viele Wissenschaftlern in mehreren Systemen Profile mit entsprechenden Duplikaten ihrer bibliographischen Metadaten.

Im Bereich der Normdatenbanken zeichnet sich ein Weg ab, der zu einer umfassenden universalen Personennormdatenbank mit eindeutiger ID führen kann. Bereits jetzt sind Daten von 15 nationalen Personennormdatenbanken sowie Testbestände von

drei weiteren Ländern in VIAF enthalten [48]. Metadaten aus VIAF sollen Bestandteil des ISNI werden.



Schwieriger gestaltet sich die Situation bei den Autoren-Identifikationen bestehender Wissenschaftsdatenbanken und Portale. Welche Möglichkeiten der Verknüpfung über eine allgemein gültige ID bestehen, zeigt deutlich das niederländische DAI-Beispiel mit dem Projekt NARCIS.

Im kommerziellen Sektor existierend viele verschiedene Ansätze. Um all diese freien wie kommerziellen Ansätze für eine Autorennormdatenbank wieder unter einen Hut zu

Identifikationsverfahren	ID-Darstellung	Anzahl
arXiv	arxiv.org—warner_s_1	Seit 2009 im Aufbau
RePEc	pkrl	26.000 Autoren
PND	113785100	3,6 Mio. Namen
DAI	info:eu-repo/dai/nl/109311094	NARCIS: 45.000 Autoren
Names Project	names/individual/97273	ca. 1 Mio. Namen
Viaf	100981156	2009: 10 Mio. Namen
WorldCat Identity	lccn-n84-188342	ca. 30 Mio. Namen
ISNI	1422 4586 3573 0476	Schätzung: 20 Mio. Namenseinträge/600.000 Neuzugänge pro Jahr
Microsoft Academic		
Search Beta	3338442 (?)	ca. 6 Mio. Autorennamen
Proquest COS		
Scholaruniverse	?	2,4 Mio Autorenprofile
ResearcherID:	G-4394-2010	
Scopus	7405627609	? (41 Mio records)

bringen, gilt es, möglichst viele Beteiligte an einen Tisch zu bekommen. Eine sich neuformierende Initiative, ORCID Inc., könnte hier weiterhelfen.

ORCID

ORCID, Inc. (Open Researcher and Contributor ID) bezeichnet sich als „non-profit“-Organisation, die sich zum Ziel gesetzt hat, das Problem der fehlenden Eindeutigkeit von Autorennamen in der Wissenschaft zu lösen und die unterschiedlichen Akteure im wissenschaftlichen Publikations- und Informationswesen zusammenzubringen [49]. Idealerweise soll ein allumfassendes Verzeichnis als De-facto-Standard das Problem der Autorentifizierung lösen.

Aus den Ergebnissen einer Online-Umfrage vom Oktober 2010 [50], ergibt sich, dass das positive Interesse an der ORCID-Initiative zwar mit knapp 40 Prozent insgesamt durchaus respektabel ist, aufgeschlüsselt nach Nutzergruppen jedoch bei den Verlegern 70 Prozent, im akademischen Umfeld jedoch nur knapp 30 Prozent beträgt [51]. Auch bei der Frage nach der Bereitschaft, für ORCID-Dienstleistungen zahlen zu wollen, ist die Abweichung erheblich. Der gemittelte Prozentsatz für die Zahlungswilligkeit liegt bei knapp 60 Prozent. Während die Verlegervertreter mit knapp 85 Prozent sehr deutlich diese Bereitschaft bejahen, beträgt der Prozentsatz der Befragten aus dem akademischen Umfeld nur mehr knapp 50 Prozent [52]. Sehr deutlich wird der Unterschied auch bei der Frage nach alternativen Finanzierungswegen. Hier scheinen gegenseitige Einschätzungen sehr deutlich: „Charge the publishers who are the ones most likely to benefit“ bzw. „Universities would clearly pay to have access to these data. Most of the institutions that we work with pay for data licenses“. Insgesamt votierte die Mehrheit für offenen und freien Zugang zu diesen Diensten [53].

Das Verzeichnis des Board of Directors gibt Hoffnung, dass diese Initiative trotz mancher gegensätzlicher Erwartung erfolgreich ist. Neben Universitäts- und Großforschungseinrichtungsvertretern und Vertretern von Forschungsförderern sind auch Verlage (u.a. Wiley Blackwell, Thomson Reuters, Elsevier) und OCLC vertreten [54]. Neben der Unterstützung durch Mitglieder und Sponsorengelder soll ein tragfähiges „business model“ erarbeitet werden, um das Ziel von ORCID zu erreichen. Ein erstes Beta-System soll im Laufe des Jahres 2011 zugänglich sein [55].

In den am 8. Dezember 2010 veröffentlichten zehn ORCID-Prinzipien [56] ist u.a. ausgesagt, dass Wissenschaftler ihr ORCID-Pro-



Abb. 21: Die ORCID-Initiative

fil unentgeltlich anlegen, ändern und pflegen können. Dies beinhaltet auch, dass sie entscheiden können, welche ihrer Daten öffentlich zugänglich und somit nutzbar sind. Alle freigegebenen Daten unterliegen der CC0 (Creative Commons Zero) Lizenz. Der Zugriff auf ORCID-Identifizierer und Profildaten soll zum Teil gebührenpflichtig werden; die Gebühren sollen allerdings nur zum langfristigen Unterhalt der nicht Gewinn orientierten Organisation verwandt werden. Die von ORCID entwickelte Software wird laut dieser Prinzipien unter Open-Source-Lizenz stehen.

Die Erwartungen in diese Initiative sind hoch, zumal einige wichtige Vertreter von DAI-Systemen an diesem Projekt beteiligt sind. Auf der Homepage von CrossRef sind diese Erwartungen wie folgt beschrieben:

Our members will be aware that CrossRef has been exploring the possibility of creating an "author DOI" or "contributor ID" system. In doing so, it has become clear that the issues and use-cases involved in identifying researchers span a broad collection of stakeholders including libraries, institutions, funders, publishers and, of course researchers themselves. In short, this is not primarily "a publisher problem." As such, we believe that the ORCID approach to creating an inclusive and open organization representing all the stakeholders in the scholarly communications process represents the best chance of creating a successful contributor identification system

For ten years the CrossRef DOI has allowed researchers, librarians and publishers to unambiguously identify published scholarly literature. We believe that the next step toward creating a robust and persistent scho-

⊙ Eindeutige Autoren-Identifikation – (PAI – Persistent Author Identification) – Versuch einer Annäherung

larly record is to develop a similar mechanism for unambiguously crediting researchers for their scholarly contributions. We look forward to working with the ORCID members to make this happen. [57]

Ausblick (PAI - Persistent Author Identification)

Wenn die ORCID-Initiative auf dem projektierten Weg bleibt und diesen konsequent umsetzt, wäre dies ein entscheidender Schritt hin zur Lösung des Identifikations-Problems bei Autoren. Aus den bisherigen ORCID-Prinzipien scheint die Einsicht hervorzugehen, dass alle Beteiligten erkannt haben, dass es dringend eine Konsolidierung aller bisherigen Ansätze dahingehend geben muss, dass eine unabhängige Institution eine frei zugängliche Konversionstabelle aller ID-Ansätze zusätzlich versehen mit einer eindeutigen persistenten, digitalen Autorenidentifikation (PAI) [58] effektiv nutzbar bereitstellt. Bei der weitergehenden Nutzung von Autorenprofilen wird man abwarten müssen, inwieweit das „business model“ der ORCID-Initiative tragfähige und möglicherweise bezahlbare Ansätze für Nutzer dieser Dienste anbietet.

In jedem Fall dürfte diese Initiative alle Beteiligten nicht davon entheben, die bisherigen Autorenprofile weiter zu pflegen bzw. auszuweiten. Insbesondere im nationalen Bereich bedeutet dies, dass man erhebliche Arbeitsleistung in die Aktualisierung der PND/GND stecken muss. Dies kann weiterhin verteilt geschehen, setzt aber auch voraus, dass entsprechende Änderungen am Datenmodell bzw. Konzept erfolgen. Hiermit hätte man ungeachtet kommerzialisierender Konzepte sich die Möglichkeit offengehalten, Autorenprofile wie bisher nutzen und somit vor allem für die zahlreichen Open-Access-Repositoryn die OA-Dokumente über die persistente Autoren-ID verknüpfen zu können. Auf diese Weise ließen sich Publikations-Metadaten aus kommerziellen Systemen mit entsprechenden OA-Publikationen in Institutionellen Repositoryn sinnvoll für alle Beteiligten verknüpfen.

Eine umfassende Zusammenstellung von Monographien, Zeitschriften- und Kongressbeiträgen sowie OA-Publikationen eines Autors wäre über diese Autorenidentifikation ebenso möglich wie die Kartierungen von Co-Autorenschaften. Der Informationsgehalt für die Wissenschaft würde sich erheblich erweitern.

Anmerkungen

- [1] Begriffe wie Autor, Forscher etc. bezeichnen im Folgenden stets männliche und weibliche Form. Auf eine Schreibweise „AutorInnen,ForscherInnen“ wird daher verzichtet.
- [2] Unesco Science Report 2010 – vgl. www.unesco.de/sciencereport2010.html. Eine deutsche Zusammenfassung findet sich unter: <http://www.unesco.de/fileadmin/medien/Dokumente/Wissenschaft/USR2010d.pdf>.
- [3] Ebda, Tabelle 2.
- [4] Ebda, Tabelle 3.
- [5] Vgl. http://www.isni.org/docs/isni_organisation_overview.pdf, S.2.
- [6] S.u.
- [7] Welche Probleme automatisiert erstellte Autorenverknüpfungen bei Google Scholar bewirken siehe bei: Peter Jacso, (2010) „Metadata mega mess in Google Scholar“, Online Information Review, Vol. 34 Iss: 1, pp.175 – 191. Vgl. auch <http://www.emeraldinsight.com/journals.htm?articleid=1840432&show=abstract>. – vgl. auch: <http://www.libraryjournal.com/article/CA6698580.html?q=jacso>.
- [8] <http://arxiv.org>.
- [9] Vgl. http://arxiv.org/help/author_identifiers.
- [10] Vgl. hierzu zuletzt Krichel, Thomas: DAI in: <http://www.dini.de/fileadmin/workshops/DINI-Helmholtz-2010/Krichel-AuthorIdentification.pdf>.
- [11] Vgl. <http://authorclaim.org>.
- [12] Vgl. <http://www.d-nb.de/standardisierung/normdateien/pnd.htm>.
- [13] Vgl. http://www.d-nb.de/standardisierung/normdateien/pnd_info.htm.
- [14] Vgl. <http://www.d-nb.de/standardisierung/normdateien/gnd.htm>.
- [15] <http://www.surffoundation.nl/nl/Pages/default.aspx>.
- [16] Zu Metis s. <http://aptest.uci.kun.nl/metis/service/Metisguide/index.htm>.
- [17] Informationen zur DAI, zur DAI-API und der DAI-extension im DAI-Wiki: <http://wiki.surffoundation.nl/display/standards/DAI>.
- [18] S.u.
- [19] Informationen unter: http://www.narcis.nl/content/pdf/narcisflyer_en.pdf; Zugang: <http://www.narcis.nl/?Language=en>.
- [20] Vgl. <http://mimas.ac.uk/news/2010/05/namesorting-smiths>.
- [21] Vgl. den Report zu Phase I des Names-Projektes vom Juli 2009: <http://names.mimas.ac.uk/documents/Names-phase-one-final-report.pdf>.
- [22] Vgl. die Folien von Amanda Hill: <http://www.sli-deshare.net/amandahill/a-rose-by-any-other-name>.
- [23] S. <http://names.mimas.ac.uk/> bzw. die Script-Version mit Vorschlagwesen beim Eintippen des Namens: <http://names.mimas.ac.uk/script-test-two>.
- [24] Vgl. http://names.mimas.ac.uk/documents/Names_stakeholder_feedback_2010.pdf.
- [25] Vgl. <http://viaf.org>. Hier ist auch der Zugriff auf die Beta-Version möglich.
- [26] Informationen unter: <http://www.oclc.org/research/activities/viaf> - vgl. auch www.oclc.org/research/activities/viaf/ala2006c.ppt.
- [27] Ähnlich den WorldCat-Identities (s.u.).
- [28] Vgl. <http://worldcat.org/identities/> bzw. <http://www.oclc.org/research/activities/identities/default.htm>.

- [29] Vgl. <http://www.isni.org>.
- [30] Vgl. http://www.isni.org/docs/isni_organisation_overview.pdf.
- [31] Vgl. http://www.isni.org/docs/isni_faq.pdf.
- [32] Beispiele aus: <http://www.isni.org/db.html>.
- [33] Vgl. http://www.isni.org/docs/isni_organisation_overview.pdf.
- [34] Vgl. <http://academic.research.microsoft.com>.
- [35] Vgl. <http://academic.research.microsoft.com/About/Help.htm>.
- [36] Informationen unter <http://www.refworks-cos.com/cos scholaruniverse>.
- [37] Vgl. <http://www.scholaruniverse.com/scholar-network-info.jsp>.
- [38] Vgl. <http://www.scholaruniverse.com/ScholarUniversePR7-08.pdf>.
- [39] Auf der Basis der Angaben der Arbeitsstätte.
- [40] <http://www.researcherid.com>; Informationen unter: <http://www.researcherid.com/resources/html/faq.html>.
- [41] <http://isiwebofknowledge.com/researcherid/>, Informationen im Factsheet: http://wokinfo.com/media/pdf/ResearcherIDFS_web.pdf.
- [42] Vgl. <http://www.researcherid.com/ViewProfile-Search.action>.
- [43] RIS kann man konvertieren über andere Literaturverwaltungsprogramme oder online mit Bookmarking- bzw. Bibliographieprogrammen wie zum Bsp. BibSonomy – www.bibsonomy.org – mit zahlreichen Import- und Export-Möglichkeiten.
- [44] Vgl. <http://www.scopus.com/home.url>.
- [45] Vgl. <http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier>.
- [46] Vgl. http://www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/pdf/br_author_identifier.pdf.
- [47] Vgl. http://www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/pdf/FAQ_Author_Identifier_09.pdf.
- [48] Vgl. die Auflistung auf der Homepage von VIAF, <http://www.viaf.org>.
- [49] Vgl. Presseerklärung vom 7. September 2010: <http://orcid.org/sites/default/files/ORCIDInc-Press.pdf>.
- [50] Mit Dank an Martin Fenner für den Hinweis – Es handelt sich um 223 ausgewertete Fragebögen, deren Auswertung nunmehr (seit 2. Dezember 2010) auch online verfügbar ist – vgl. <http://www.orcid.org/node/216>; <http://www.orcid.org/sites/default/files/ORCIDSurveyResults.pdf>.
- [51] <http://www.orcid.org/sites/default/files/ORCIDSurveyResults.pdf>, Folie 8.
- [52] <http://www.orcid.org/sites/default/files/ORCIDSurveyResults.pdf>, Folie 11.
- [53] <http://www.orcid.org/sites/default/files/ORCIDSurveyResults.pdf>, Folie 16ff.
- [54] Vgl. <http://www.orcid.org/board-of-directors>.
- [55] Vgl. <http://www.orcid.org/sites/default/files/ORCID Sponsorship2010.pdf>.
- [56] Vgl. <http://www.orcid.org/principles>.
- [57] Vgl. <http://www.crossref.org/01company/orcid.html>.
- [58] Auch unique author identification benannt.

Eindeutige Autoren-Identifikation – (PAI – Persistent Author Identification) – Versuch einer Annäherung

Anschrift des Verfassers:
Dr. Helge Steenweg
Stellv. Bibliotheksdirektor UB Kassel
Diagonale 10
34127 Kassel
steenweg@bibliothek.uni-kassel.de