OA-NETZWERK

Dienstepapier

Datenaggregations- und aufbereitungsdienst für digitale Inhalte wissenschaftlicher Repositorien in Deutschland

Stand: 19. April 2012

AutorInnen: Maxi Kindling, Sammy David, Julia Iwanowa

DOI: 10.5281/zenodo.1315346

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einordnung und Zielstellung

Dieses Papier beschreibt das im DFG-Projekt „Weiterentwicklung und Betrieb des Netzwerks zertifizierter Open-Access-Repositorien“ (OA-Netzwerk)[1] entwickelte Angebot eines OA-Dokumentenraums und der darauf basierenden Dienste. Die Dienste werden dabei nicht im Sinne einer technischen Dokumentation dargestellt, sondern durch ihre grundsätzliche Funktionalität und ihre Leistungen veranschaulicht. OA-Netzwerk wird als Projekt in zwei Projektphasen im Zeitraum von 2007 bis 2012 gemeinschaftlich durch die Humboldt-Universität zu Berlin (Computer- und Medienservice), die Niedersächsische Staats- und Universitätsbibliothek Göttingen und die Universität Osnabrück (Fachbereich für Mathematik und Informatik) durchgeführt. Das Ziel besteht darin, ein Netzwerk zertifizierter Open-Access-Repositorien in Deutschland aufzubauen und dadurch die Open-Access-Infrastruktur für die deutsche Forschung zu stärken. Grundlage dieser gemeinsamen Infrastruktur und der mit OA-Netzwerk verbundenen oder kooperierenden Projekte sind die Standards des DINI-Zertifikats[2].

Derzeit sind alle DINI-zertifizierten Repositorien in OA-Netzwerk eingebunden. Das DINI-Zertifikat für Dokumenten- und Publikationsservices der Deutschen Initiative für Netzwerkinformation e.V. (DINI) ist derzeit bereits in der dritten Generation (DINI-Zertifikat 2010) verfügbar. Die Ergebnisse von OA-Netzwerk stellen einen wichtigen Schritt zu einer verbesserten Informationsversorgung im Bereich von Open-Access-Publikationen dar und bilden die technologische Grundlage für zahlreiche bereits entwickelte oder noch in Planung befindliche Mehrwertdienste. Eine Weiterführung des Betriebs ist unter dem Gesichtspunkt der weiteren Verbreitung des Open-Access-Gedankens und einer Verbesserung der Nutzung der vorhandenen Infrastruktur zweckmäßig und wünschenswert. Anbieter übergreifender und auf den bereitgestellten Daten basierender Dienste im nationalen und internationalen Kontext profitieren davon ebenso unmittelbar wie Betreiber von Repositorien, die durch OA-Netzwerk eine wirkungsvolle Einbindung ihrer Bestände in Informationsnetzwerke und damit eine erhöhte Sichtbarkeit erreichen. Mittelbar dient OA-Netzwerk damit Wissenschaftlern—sowohl bei der Informationsrecherche als auch beim Publizieren eigener Ergebnisse.

[1] Siehe `http://www.dini.de/projekte/oa-netzwerk/`
[2] Siehe `http://www.dini.de/dini-zertifikat/`

2 Zusammenfassung der OA-Netzwerk-Dienste

Den Kern der von OA-Netzwerk bereitgestellten Dienste bilden die Zusammenführung, Aufbereitung und Bereitstellung von Daten, die in wissenschaftlichen digitalen Repositorien in Deutschland verteilt vorgehalten werden. Derzeit sind die Inhalte DINI-zertifizierter Repositorien erfasst, d. h. das Wachstum dieses Dokumentenraums ist an die Fortschritte der Zertifizierungsaktivitäten gebunden. Die Aggregation kann auf die Inhalte nicht zertifizierter Dokumentenserver ausgeweitet werden. Die aufbereiteten Daten sind sowohl über ein Webinterface für Endnutzer[3] als auch über standardisierte Schnittstellen zur weiteren automatischen Verarbeitung nutzbar. Die Suche über die Dokumente wird in andere Plattformen wie die Informationsplattform Open-Access.net integriert, die sich als Anlaufstelle für OA-Interessierte etabliert hat.[4]

OA-Netzwerk bietet eine gemeinsame und stets aktuelle Sicht auf die in den eingebundenen Repositorien veröffentlichten wissenschaftlichen Publikationen und stellt somit eine Vermittlungsschicht zwischen den verteilt realisierten institutionell oder fachlich ausgerichteten Repositorien einerseits und übergreifenden Diensten andererseits dar. OA-Netzwerk bildet die Datenbasis für Datenaggregatoren bzw. Anbieter von Mehrwertdiensten. Eine hohe Datenqualität wird dabei durch die Überprüfung, Harmonisierung, Anreicherung und Kontextualisierung der Originaldaten aus den Repositorien (Fachklassifikation, Volltextindexierung ggf. einschließlich OCR, Dubletten- und Ähnlichkeitserkennung, semantische Anreicherung) und durch die Beschränkung des Dokumentenraums auf DINI-zertifizierte Repositorien erreicht, die einen qualitativen Mindeststandard sichern. Der asynchrone Harvesting- und Aufbereitungsmechanismus[5] stellt eine hohe Performanz und Verfügbarkeit des Gesamtdienstes sicher. Dabei ist aufgrund kurzer Updateintervalle die Aktualität des Datenbestandes, also die Übereinstimmung mit den lokalen Datenbeständen in den jeweiligen Repositorien, gegeben. Die Qualität der bereitgestellten Daten sowie die Bereitstellung und Weiterentwicklung von Mehrwertdiensten und die Möglichkeiten der Integration extern entwickelter Dienste kennzeichnen das Diensteangebot von OA-Netzwerk und heben es von anderen Aggregationsdiensten wie BASE (Bielefeld Academic Search Engine)[6] ab. Die entwickelten Komponenten zur Aufbereitung der Daten können natürlich auch für andere Datenbestände genutzt werden, die über vergleichbare Schnittstellen verfügbar sind.

[3] Siehe `http://oansuche.open-access.net`
[4] Siehe `http://www.open-access.net/`
[5] Das Administrationstool OAN Admin (?) bietet die Möglichkeit, Einstellungen für das Harvesting sowie die Anreicherung von Objekten nach Bedarf zu konfigurieren, vgl.. . .
[6] Vgl. `http://base.ub.uni-bielefeld.de/`

3 Technisches Grundkonzept

Die Software zur Realisierung des Gesamtdienstes und der einzelnen Teildienste basiert auf den Prinzipien einer Serviceorientierten Architektur (SOA). Die Daten werden in ihren unterschiedlichen Verarbeitungsstufen in einer zentralen Datenbank gemäß einem einheitlichen Objektmodell gespeichert und durch eine gemeinsame REST-Schnittstelle zugänglich gemacht. Die unterschiedlichen Komponenten des OAN-Dienstes greifen entsprechend ihrer jeweiligen Zugriffsrechte lesend bzw. schreibend auf den Datenbestand zu. Dabei gilt der Grundsatz, dass durch Harmonisierung und Anreicherung Daten aus vorhergehenden Verarbeitungsstufen nicht ersetzt sondern ergänzt werden, sodass für jedes Datenobjekt einschließlich der ursprünglich durch das Harvesting eingestellten Fassung alle Versionen rekonstruierbar sind.

Als interne Dienste sind folgende Funktionen implementiert: a) Harvesting; b) Aggregation; c) Volltextanalyse; d) Klassifikation; e) Semantische Anreicherung. Grundsätzlich laufen die Dienste vollautomatisch. Zur Behandlung etwaiger Probleme im laufenden Betrieb und zur Verwaltung existiert eine Administrationsoberfläche. Die oben genannten Dienste sind so konzipiert, dass sie autonom und asynchron arbeiten. Die Abhängigkeit einzelner Dienste untereinander impliziert eine bestimmte Ausführungsreihenfolge bezüglich eines konkreten Datenobjekts, die durch angehängte Statusinformationen sichergestellt wird. Die konsequente Umsetzung der beschriebenen REST-Architektur ermöglicht eine verteilte Realisierung der Dienste. Derzeit laufen die Dienste bei unterschiedlichen Projektpartnern und greifen über die einheitliche REST-Schnittstelle auf den gemeinsamen Datenbestand zu.

3.1 Überblick über die OAN-Komponenten

3.1.1 Harvesting

Als Objekte im Sinne von OA-Netzwerk werden (wissenschaftliche) Publikationen bzw. Dokumente einschließlich aller verfügbaren Metadaten betrachtet. Der Ansatz folgt der internen Repräsentation der meisten Repositorien und greift die Logik des OAI-Protokolls (OAI-PMH) auf, auf dem der Datenabgleich zwischen den Repositorien und OA-Netzwerk basiert.

OA-Netzwerk sammelt die Metadaten aller öffentlich verfügbaren Dokumente der angeschlossenen Repositorien über OAI-PMH ein, speichert sie in Form des originären

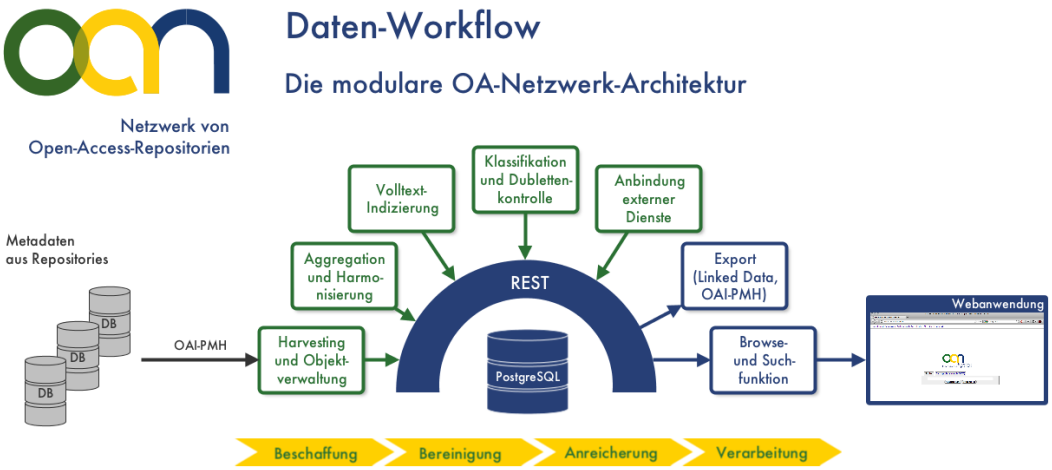

Abbildung 1: OA-Netzwerk: Datenworkflow, Stand 2012

XML-Datenstroms im Format DC Metadata Element Set[7] und überführt die einzelnen Datensätze in die interne Objektrepräsentation. Für jedes auf diese Weise neu angelegte bzw. durch Änderungen aufseiten der betreffenden lokalen Repositorien aktualisierte Objekt müssen die darauf aufsetzenden internen Dienste ausgeführt werden, bevor der vollständige Datensatz nach außen zur Verfügung gestellt wird. Das Aktualisierungsintervall für den Harvesting-Dienst beträgt derzeit 24 Stunden. Bestandteil des Dienstes ist außerdem ein automatischer Feedbackdienst, der beim Harvesting auftretende Fehler (z. B. durch fehlerhafte Zeichenkodierung, Protokollfehler, Übertragungsfehler oder der Nicht-Erreichbarkeit von Objekten) erkennt, klassifiziert und per E-Mail an die Betreiber der jeweils betroffenen Repositorien sowie an die OA-Netzwerk-Administratoren meldet.

3.1.2 Aggregation

Dieser Dienst wird auf neu erfasste bzw. aktualisierte Dokumente und Metadaten angewendet, die durch das Harvesting ermittelt worden sind. Im Wesentlichen führt der Aggregationsprozess zwei Schritte aus. Zum einen wird eine aktuelle und fehlerkorrigierte Kopie der Originaldaten angelegt. Zum anderen werden auf Basis dieser Kopie die Metadaten strukturiert in ein internes Metadatenformat (bislang wird das DC Element Set unterstützt) überführt. Aus den Metadaten werden einheitliche Formate (z. B. das

[7] Dublin Core DC Metadata Element Set v 1.1, siehe `dublincore.org/documents/dces`

Datumsformat, Angaben zur Sprache oder Publikationstyp) erzeugt. Es findet damit eine Homogenisierung der Metadaten statt, um eine einheitliche Struktur der Daten zu wahren.

Zu diesem Zweck wird bei der Aggregation ein Mapping-Prozess initiiert. Dieser korrigiert Feldinhalte aus den Rohmetadaten, die nicht den DINI-Kriterien entsprechen. Darunter fallen Angaben zum Dokumenttyp sowie DDC-Kategorien. Häufig verwendete Begriffe werden dabei den Kategorien entsprechend des „Gemeinsamen Vokabulars für Publikations- und Dokumenttypen“ nach den DINI-Kriterien zugeordnet. Auf dieser Basis ist das Browsing nach Dokumenttyp und Klassifikation in der Rechercheoberfläche umgesetzt.

Nachfolgende Dienste reichern die homogenisierten Daten mit zusätzlichen Informationen an.

3.1.3 Volltextanalyse

Der Dienst basiert nicht nur auf den über die OAI-Schnittstellen der beteiligten Repositorien eingesammelten Metadaten der Publikationen, sondern bezieht deren Volltexte mit ein. Nur mithilfe der vollständigen Publikationen sind weiterführende Anreicherungen und Kontextualisierungen der Datensätze (beispielsweise eine automatische Fachklassifikation) möglich.

Dazu wird in einem ersten Schritt die URL ermittelt, unter der die jeweilige Publikation tatsächlich verfügbar ist. Im einfachsten Fall kann dafür zwar unmittelbar die in den Metadaten enthaltene URL (kodiert im Feld `dc:identifier`) herangezogen werden. In den meisten Fällen führt diese allerdings nicht direkt zum Volltext, sondern beispielsweise zu einer Einstiegsseite des veröffentlichenden Repositoriums mit Metadaten und weiterführenden Informationen, die dann ihrerseits einen Link zur Publikation enthält. Diese URL automatisiert und zuverlässig zu ermitteln, ist die Voraussetzung dafür, auf das Dokument zugreifen zu können.

Die Volltextdatei[8], die aus rechtlichen Gründen nur temporär gespeichert wird, wird anschließend für die Volltextindexierung vorbereitet.

Um die Publikationen durchsuchbar und nach Kategorien auffindbar anzubieten, werden die Volltexte durch eine Suchmaschine indexiert, was derzeit mithilfe der Software ElasticSearch realisiert wird.[9] Anschließend wird bei fehlendem Metadatum automatisch die Sprache des Volltextes ermittelt.

[8] Derzeit wird ausschließlich das Format PDF unterstützt.
[9] Hierbei handelt es sich um Open Source Software, siehe `http://www.elasticsearch.org/`

3.1.4 Klassifikation

Ein weiterer Dienst zur Anreicherung bzw. Erzeugung fehlender Metainformationen ist der Klassifikator. Auf Ebene der Sachgruppen der Deutschen Nationalbibliografie (basierend auf DDC) wird mit dem Klassifikator die Zuordnung der Fachgebiete zu den Publikationen erzielt. Zur Bestimmung dieser DDC-Klassifikationsinformationen wird die „Automatic Classification Toolbox for Digital Libraries“ eingesetzt. Dieser Dienst wurde ursprünglich an der Universität Bielefeld entwickelt und stellt eine öffentliche Web-Service Schnittstelle zur Verfügung.

3.1.5 Semantische Anreicherung und Überführung in das standardübergreifende Datenmodell EDM

Das Ziel des Linked-Open-Data-Dienstes von OA-Netzwerk ist die Konvertierung, Kontextuallisierung, Verlinkung und Bereitstellung der geharvesten, aggregierten und homogenisierten bibliographischen Metadaten als Linked Open Data. Als Ergebnis wurde die OA-Netzwerk Datenbasis[10] als Linked Open Data veröffentlicht und für externe Nutzer zur Verfügung gestellt[11]. Die Daten sind frei verfügbar und unter CC0-Lizenz ohne Einschränkung weiternutzbar. Aus unserer Sicht wird diese radikale freie Datenveröffentlichung auf der einen Seite die Linked Open Data Community und auf der anderen Seite die Entwicklung und Implementierung neuer Software auf Basis dieser offenen Daten unterstützten.

Technische Grundlagen für die Semantische Anreicherung der OA-Netzwerk-Dienste sind Semantic Web-Standards des W3C-Konsortiums[12]; dazu zählen u. a. RDF[13], RDF-S[14] sowie OWL[15]. Hinzu kommt ein Standard der Open Archives Initiative[16] für die eindeutige Identifzierung von verteilten Web-Ressourcen, OAI-ORE [17]. Um eine Datenintegration und Kontextualisierung zwischen Europeana und OA-Netzwerk zu erreichen, wird zudem das Datenmodell von Europeana[18] eingebunden. Die Grundlage

[10] OA-Netzwerk als Linked Open Data, siehe `http://oanet.cms.hu-berlin.de/d2r-oan/`
[11] OA-Netzwerk bei The Data Hub, siehe `http://thedatahub.org/dataset/oanetzwerk`
[12] World Wide Web Consortium, siehe `http://www.w3.org/`
[13] Resource Description Framework, vgl. `http://www.w3.org/RDF/`
[14] RDF-Schema, vgl. `http://www.w3.org/TR/rdf-schema/`
[15] Web Ontology Language, vgl. `ttp://www.w3.org/TR/owl-features/`
[16] Vgl. `http://www.openarchives.org/`
[17] Open Archives Initiative Object Reuse and Exchange (OAI-ORE), vgl. `http://www.openarchives.org/ore/1.0/primer.html`
[18] Europeana Datenmodell (EDM), für weitere Details und technische Dokumentation siehe `http://pro.europeana.eu/web/guest/edm-documentation`

für die Implementierung der Semantic Web-Anwendungen von OA-Netzwerk bildet die D2RQ-Plattform[19] .

3.2 Administration

Die technisch-administrativen Aufgaben und Möglichkeiten können größtenteils bequem über die zentrale Webanwendung wahrgenommen werden. Neben der Steuerung und Zeitplanung der einzelnen Dienste bzw. der Dienstekette lässt sich auch der Ablauf dieser Dienste über entsprechende Monitoring-Dienste verfolgen. Ebenso lässt sich die Verfügbarkeit der Repositorien überwachen, indem täglich automatisch Anfragen an jedes Repository des Netzwerks gesendet werden. In einer Historie kann die Verfügbarkeit auch rückwirkend eingesehen werden.

Neue Repositorien lassen sich spielend einfach dem Netzwerk hinzufügen. Sollten Repositorien aufgenommen werden, die nicht dem DINI-Netzwerk angehören, lässt sich die Qualität neuer Repositorien vorab über den OAN-Validator-Dienst abprüfen. Prinzipiell ist das System robust und eine Neuaufnahme ist in jedem Fall möglich, unabhängig von der Punktbewertung des Validators.

Eine einfache Informationsübersicht über die enthaltenen Daten gibt Aufschluss über die Zusammentsetzung der Gesamtdaten. Es lässt sich beispielsweise leicht erkennen, wie die Zusammensetzung nach DDC-Kategorie oder Dokumenttyp gegliedert ist. Auch lassen sich nicht bearbeitbare Datensätze anzeigen.

3.3 Bereitstellung nach Außen

OA-Netzwerk bildet einen Datenknoten, der über standardisierte Schnittstellen den eigenen Datenbestand Dritten zur Verfügung stellen kann, die beispielsweise weitere Dienste entwickeln. OA-Netzwerk bietet unterschiedliche Möglichkeiten an, auf die bereitgestellten aggregierten und angereicherten Daten zuzugreifen. Drei grundsätzliche Mechanismen sind dabei zu unterscheiden.

Zum einen können der gesamte Bestand oder definierte Teile daraus abgerufen (z. B. nach fachlichen Kriterien bestimmt, wie Selektionsmöglichkeiten nach DNB-Sachgruppen, nach DC oder weiteren Kriterien) und durch einen asynchronen Aktualisierungsmechanismus regelmäßig abgeglichen werden (OAI-PMH). Dies wird etwa durch die Bereitstellung der OA-Netzwerk-Daten für das europäische Repositorien-Netzwerk DRIVER verfolgt. Weiterhin wird der OA-Netzwerk-Datenraum für das ebenfalls DFG-geförderte

[19] Für weitere Details und technische Dokumentation siehe `http://d2rq.org/`

Projekt OA-Fachrepositorien[20] bereitgestellt, das auf dieser Grundlage und einer fachspezifischen Auswahl den Datenfluss an nationale und internationale disziplinär ausgerichtete Dokumentenserver ermöglicht. Mithilfe der originären Auszeichnung jedes einzelnen Objekts mit dem OAI-Identifier können auch unveränderte Originaldaten gesammelt und weiterverwendet werden.

Zum anderen stellt OA-Netzwerk einen Suchdienst bereit, der die Implementierung synchroner Dienste (z. B. Metasuche) erlaubt; dies wird über SRU/W- bzw. REST-Schnittstellen realisiert. Die prototypische Implementierung einer zentralen, webbasierten Suchplattform demonstriert das mögliche Zusammenspiel von Volltext- und Metadatensuche sowie zusätzlichen Mehrwertdiensten wie Nutzungsdaten, RSS-Feeds und weitere.

Schließlich werden die vollständigen Daten auch in Form von Linked Open Data (LOD) bereitgestellt. Sie können damit in verschiedenen Kontexten des WWW eingebunden werden. Über eine Weboberfläche wird der gesamte Datenbestand von OA-Netzwerk im RDF-Format als Linked Open Data veröffentlicht. Die Daten können sowohl mit herkömmlichen Webbrowsern als auch mit Semantic Web-Browsern oder aber über eine SPARQL-Schnittstelle durchsucht werden.

3.4 Integration und Bereitstellung von Mehrwertdiensten und das Umfeld von OA-Netzwerk

OA-Netzwerk ist neben der dargestellten technischen Ebene auch organisatorisch in ein vielseitiges Umfeld von Projekten und Initiativen in den Bereichen Open Access und elektronisches Publizieren eingebunden. Hier ist insbesondere die Arbeitsgruppe „Elektronisches Publizieren“ von DINI zu erwähnen, für die REST-Schnittstellen bereitgestellt werden.

Über die Bereitstellung des Datenraums hinaus integriert OA-Netzwerk auch externe Mehrwertdienste. Während der Projektlaufzeit erfolgt die Kooperation mit Projekten aus dem Open-Access-Umfeld wie OA-Statistik (OAS)[21], Distributed Open Access Reference Citation Service (DOARC)[22] und OA-PlagiatSuche (OAPS)[23]. Die Zusammenarbeit mit OA-Statistik verfolgt dabei das Ziel, die Nutzungsdaten über OA-Dokumente auch in der Weboberfläche der OA-Netzwerk-Suche sichtbar zu machen. Das Pro-

[20] Siehe http://www.ub.uni-konstanz.de/bibliothek/projekte/open-access-fachrepositorien.html
[21] Siehe http://www.dini.de/projekte/oa-statistik/
[22] Siehe http://doarc.projects.isn-oldenburg.de/
[23] Siehe http://svn.ibr.cs.tu-bs.de/projects/oaps/

jekt OA-PlagiatSuche nutzt die OA-Netzwerk-Daten zum Abgleich der auf Plagiate untersuchten Dokumente. Ferner wird eine Suchschnittstelle für Plagiate auch auf der Suchoberfläche von OA-Netzwerk angeboten.[24] OA-Netzwerk ist offen für weitere Entwicklungen und die Integration weiterer Mehrwertdienste.[25] So gibt es enge Kooperationen mit der Informationsplattform Open-Access.net für Wissenschaftler und der Informationsplattform CARPET[26], die Werkzeuge für das elektronische Publizieren bereitstellt und damit für die Entwickler von Mehrwertdiensten von Interesse ist.

[24] Die neue Version der OAN-Suche wird Ende März online sein.
[25] Siehe `http://www.dini.de/ag/e-pub/`
[26] Siehe `http://www.carpet-project.net/`

4 Zusammenarbeit mit anderen Projekten

Der gemeinsame Daten- und Suchraum von OA-Netzwerk setzt sich aus den individuellen Beständen der angebundenen Repositorien zusammen. Die technischen Grundlagen dieses verbundenen Datenraums werden durch das DINI-Zertifikat gesetzt; ein harmonisierter bzw. standardisierter Bestand ist das Ergebnis. Dieser Bestand sowie die OAN-Infrastruktur stehen anderen Projekten offen, um darauf aufsetzend Mehrwertdienste zu entwickeln.

Ausgehend von diesem Ziel hat OA-Netzwerk während seiner Projektlaufzeit versucht, Kooperationen und Vernetzung auf organisatorischer und technischer Ebene zu initiieren. Die organisatorische Ausprägung dieser Bemühungen ist der mittlerweile auf neun Projekte angewachsene Kreis „OA-Projects“. Allen Projekten ist gemein, dass sie beispielhafte Lösungen für Aufgaben des elektronischen Publizierens zu entwickeln suchen und sich dem Open-Access-Gedankens verpflichtet fühlen. Die Mehrzahl von ihnen wurde initial von der Deutschen Forschungsgemeinschaft gefördert. Folgende Projekte arbeiten im OA-Projects-Kreis zusammen:

- OA-Netzwerk
- OA-Statistik
- OA-Fachrepositorien
- OA-Plagiatsuche
- OA-Policies
- DOARC (vormals OA-Citation)
- OJS[27]
- CARPET

[27] Open Journal Systems, in Zusammenarbeit mit dem Public Knowledge Project, siehe `http://pkp.sfu.ca/?q=ojs`

5 Übernahmeszenario

5.1 Übernahme der Dienste

Das OA-Netzwerk unterscheidet sich von anderen Diensteanbietern und Rechercheplattformen auf wissenschaftlichen Dokumentenräumen durch eine Qualitätssicherung der enthaltenen Dokumente innerhalb des OAN-Datenraumes. Zu diesem Zweck umfasst der OAN-Datenraum Repositorien, die mit dem DINI-Zertifikat ausgezeichnet sind und damit eine Reihe von Richtlinien erfüllen, die eine robuste, langfristige und reichhaltige Bereitstellung von sowohl Volltexten als auch Metadaten gewährleisten. Innerhalb des Projekts OA-Netzwerk sind eine Reihe von Diensten entwickelt worden, die—begonnen beim Harvesting—eine Reihe weiterer verarbeitender und datenanreichender Aufgaben erfüllen. Dabei sind sämtliche Dienste modular und im hohen Maße für den Einsatz innerhalb verteilter Systeme gedacht. Diese Dienste arbeiten innerhalb einer Kette, in der einige Dienste—besonders die datenanreichernden Dienste (z. B. DDC-Klassifikator, Spracherkennung)— aufbauend auf anderen Diensten arbeiten. Um das Potential des gesamten OAN-Systems auszuschöpfen, empfehlen wir den Einsatz der gesamten Dienstekette. Bei Bedarf ist es auch möglich, einzelne anreichernde Dienste aus der Kette zu entfernen, wenn deren Funktionalität durch einen anderen Dienst ersetzt oder gänzlich darauf verzichtet werden soll. Die Funktionalität der Dienste wurde im Einzelnen bereits im vorangehenden Teil dieses Papiers verdeutlicht. Eine Übernahme des OAN-Systems und der damit zusammenhängenden Software-Komponenten wird im nächsten Abschnitt näher erläutert.

5.2 Übernahme des Datenraums

Neben der Übernahme der einzelnen Dienste empfehlen wir auch die Übernahme des bestehenden OA-Netzwerk-Datenraumes, um den Gedanken an einen qualitativ hochwertigen Datenraum fortzuführen. Dabei wird die DINI weiterhin eine aktive Rolle spielen, indem das DINI-Zertifikat national weiter verbreitet wird und einen noch höheren Anklang und Unterstützungsgrad unter institutionellen und fachlichen Repositorien erhält und somit der Datenraum für zertifizierte Repositorien stetig wachsen kann. Zu diesem Zweck besteht über eine webbasierte Manageranwendung die Möglichkeit, neue Repositorien einzubinden. Prinzipiell ist es natürlich auch möglich, nichtzertifizierte Repositorien in den OAN-Datenaum aufzunehmen. Für die Integration eines nicht-DINI-zertifizierten Repositoriums besteht zur Wahrung einer selbstdefinierbaren Qualität der Metadaten die Möglichkeit, mittels eines Repository-Validators diesen

potentiellen Repositorienkandidaten auf Metadatensatzebene auf die Konformität der selbstgeforderten Regeln hin automatisiert zu überprüfen. Dadurch wird schnell ersichtlich, ob ein Repositorium den eigenen Qualitätsansprüchen genügt oder nicht. Der für diesen Zweck eingesetzte Validator entspricht dem weiterentwickelten Validator des OpenAIRE-Projekts (Stand Mai 2011). Mit dem Validator lässt sich die Analyse von Repositorien ebenfalls über die Manageranwendung steuern und erzielte Ergebnisse auswerten.

5.3 Bereitstellung des OAN-Datenraums für Projektpartner/Datenlieferanten

Da ein Teil der Datenanreicherungen und Mehrwertdienstinformationen technisch von Schwesterprojekten in das OA-Netzwerk eingespielt werden, ist bei einer Übernahme des OAN auch eine Öffnung des REST-Servers nach außen hin (WWW) erforderlich, um den von Partnerprojekten benötigten OAN-Datenraum über die Schnittstellen des REST-Servers auch weiterhin zur Verfügung zu stellen. Weiterhin werden die erzeugten Mehrwertinformationen über diese bereitgestellten Schnittstellen wieder in die OAN-Datenbasis eingespielt. Die Nutzung dieser Schnittstellen wird von einem automatisierten Workflow der Projektpartner bedient und sollte demnach auch langfristig bereitgestellt werden, um die statistische (bereitgestellt durch OA-Statistik) und Zitationsdaten (bereitgestellt durch DOARC) zu den Publikationen bzw. Metadaten im OA-Netzwerk regelmäßig entgegennehmen zu können.

5.4 Lieferumfang, Anforderungen an System und Personal

Die einzelnen Komponenten des OAN-Basisdienstes sind vornehmlich in Java und Ruby, vereinzelt auch in Perl, programmiert. Um eine Übernahme bestmöglich zu unterstützen, wird der kommentierte Quelltext sowie eine Dokumentation der Dienste und des Gesamtsystems geliefert. Des Weiteren werden folgende Dokumente mitgeliefert:

- Auflistung der Anforderungen an Soft- und Hardware
- Anleitung für Konfiguration und Inbetriebnahme
- Dokumentation der REST-Schnittstelle, über die einzelne OAN-Komponenten Daten austauschen und abgleichen
- Anleitung für Weiterentwicklung des OAN-Pakets

Als Ansprechpartner für technische Fragen werden Sammy David und Julia Iwanowa auch weiterhin zu Verfügung stehen. Darüber hinaus trägt die Arbeitsgruppe „Elektro-

nisches Publizieren“ in ihrer Arbeit an dem DINI-Zertifikat dafür Sorge, dass Qualitätskritierien für Repositorien auch zukünftig spezifiziert werden und so die Anforderungen neuen Entwicklungen entsprechend angepasst werden. Zudem wird über die DINI die Bekanntmachung des Dienstes weiter gefördert und ggf. der Pool an teilnehmenden Repositorien stetig ausgebaut.

Eine technische Übersicht über die einzelnen Komponenten, aus denen sich das Gesamtsystem OA-Netzwerk zusammensetzt, gibt die „Technische Kurzdokumentation“. Darin beschrieben sind auch technische Anforderungen an die Systemumgebung und der zu erwartende Aufwand für Aufbau und Regelbetrieb des OA-Netzwerks.