Open Access – Verfügbar ist noch nicht präsent

Verbesserung der Sichtbarkeit von Open Access Repositories durch die Bildung von Netzwerken

Robin Malitz | malitzro@cms.hu-berlin.de

Der Open-Access-Gedanke propagiert den freien Zugriff auf wissenschaftliche Texte – lesen und gelesen werden. Es steht außer Frage, dass das Veröffentlichen elektronischer Dokumente in Open Access Repositories diesen freien Zugriff gewährt. Doch genügt dieses, um gelesen zu werden? Sicher nicht. Impact kann nur haben, wer auch wahrgenommen wird. Sichtbarkeit heißt die wesentliche Herausforderung für Open Access und Vernetzung fördert diese auf verschiedene Weise.

Der Artikel skizziert die Hintergründe und grundlegenden Zusammenhänge, die sich aus dem Wechselspiel von Open-Archive-Welt und den etablierten Mechanismen des World Wide Web ergeben, und stellt u. a. die Bedeutung des auf die deutsche Repositories-Landschaft fokussierten Projektes „Open-Access-Netzwerk" im Kontext der Sichtbarkeit dar.

Suchmaschinen und das „Unsichtbare Netz"

Um der Fragestellung der Sichtbarkeit von Open-Access-Dokumenten nachzuspüren, beginnt man am besten aus der Sicht des Suchenden, der sich ins Netz begibt, um nach inhaltlich für ihn relevanten, wissenschaftlichen Volltexten zu recherchieren. Der erste Anlaufpunkt sind im einfachsten Falle die gängigen Suchmaschinen, die man in der Hoffnung mit Schlagworten füttert, dass sie das, was im Netz verfügbar ist, auch finden. Schnell wird jedoch klar: Das gezielte, fachspezifische Recherchieren nach online publizierten Dokumenten gerät zur mühseligen Goldwäscherei in den angezeigten Trefferseiten. Selbst vielversprechende Links führen nicht immer zum gewünschten Volltext.

Besser scheint da, wird direkt der Recherchezugang eines Open Access Repositories genutzt. Auch hier gibt es heutzutage in fast allen Standardlösungen an zentraler Stelle eine Suchmaske. Auf Knopfdruck erhält man eine Trefferliste, die ausschließlich aus wissenschaftlichen Dokumenten besteht, deren Volltexte man einsehen, herunterladen und ausdrucken kann. Mehr noch, häufig kann man viel detaillierter in einer Art Expertensuche gezielt nach Veröffentlichungen mit bestimmten Eigenschaften recherchieren. Hier unterscheiden sich die angebotenen Such-Möglichkeiten von Repository zu Repository. Die Vorteile dieser Recherchemöglichkeit liegen jedoch auf der Hand.

Eine Kehrseite gibt es auch hier. In der Regel kann nur die lokal archivierte Menge an Dokumenten gezielt durchsucht werden. Was ist jedoch mit Dokumenten, die an anderen Einrichtungen publiziert worden und nicht in dem gerade durchsuchten Repository verfügbar sind? Diese lassen sich im Letzteren dann nicht finden. Da es allein in Deutschland über einhundert wissenschaftliche Open Access Repositories gibt – DINI listet 138 verschiedene Webadressen auf – ist eine lokale Suche bei einzelnen Repositories weder sinnvoll noch komfortabel.

Wer sich nun geschlagen gibt und wieder die Suchmaschinen bemüht, dem fällt vielleicht sogar auf, dass einige in Repositories frei verfügbare Dokumente nicht gefunden werden. Das liegt daran, dass die Rechercheoberflächen der Repositories dynamische Webseiten sind, die die nimmermüden, automatischen Späher der Suchmaschinen, die Robots und Webcrawler, technologisch durch Software-Barrieren aussperren. Für die Suchmaschinen sind die archivierten Dokumente dadurch zumeist unsichtbar, weshalb dieses Phänomen als „Invisible Web" oder „Deep Web" bezeichnet wird. Auch wenn die Suchmaschinen immer leistungsfähiger werden, müssen sich die Betreiber von Repositories häufig Umwege ausdenken, um wenigstens die Volltexte der Dokumente – wenn auch nicht die dazugehörigen, für Recherchierende so wertvollen Metadaten – für die Suchmaschinen-Welt sichtbar zu machen. Verfügbar ist eben noch nicht präsent.

Von Spezialsuchdiensten zu Rechercheplattformen

Die Problematik ist bekannt: Schon früh erkannte man die Notwendigkeit, dass die in den Repositories gesammelten Publikationen Spezialsuchdiensten verfügbar gemacht werden müssen. Diese Spezialsuchdienste sollen für den Suchenden die Grenzen zwischen den Repositories verwischen und ihm dabei erlauben, möglichst viele der zu einem Dokument erfassten Metadaten in seine Suchanfrage einzubeziehen. Wird z. B. nach Dokumenten eines bestimmten Autors zu einem bestimmten Erscheinungsjahr gesucht, dann soll der Spezialsuchdienst alle passenden Dokumente als Treffer präsentieren, unabhängig davon, wo das Dokument tatsächlich archiviert worden ist. Solche Spezialsuchdienste gibt es bereits. Im Open Access Bereich bieten bislang Webangebote wie OAIster, Google Scholar, BASE, DRIVER und andere den Suchenden Recherche-Möglichkeiten.

Der Begriff „Spezialsuchdienst" ist bei den aktuellen Projekten und Angeboten eigentlich zu einseitig. Recherche durch die Option einer detaillierten Suchanfrage ist nur ein Mehrwert-Dienst von vielen, den der Recherchierende im Kontext wissenschaftlicher Publikationen benötigt. Hinzu kommen Nachweis-Dienste (Browsing), Alerting, Newsfeeds, Export von Metadaten in Literaturverwaltungsformate und viele mehr.

Heutzutage spricht man daher oft von Portalen oder Plattformen. Der erste Begriff betont, dass der Nutzerzugang zu ganz verschiedenen Mehrwertdiensten gebündelt angeboten wird, der zweite, dass das Angebot technologisch so gestaltet wurde, dass das bestehende System als Basis und Präsentationsoberfläche für weitere Mehrwertdienste genutzt werden kann.

Infrastruktur für hochwertige Mehrwertdienste

Repositoriesbetreiber, auf der anderen Seite, sind nun bestrebt, ihre Dokumente in Plattformen mit möglichst vielfältigen und hochwertigen Mehrwertdiensten anzubieten. Autoren publizieren vornehmlich dort, wo sie glauben, gelesen zu werden, und gelesen wird im Netz, wo man komfortabel recherchieren kann. Die Güte von Mehrwert-Diensten hängt unmittelbar vom Zusammenspiel von Plattformen und Repositories ab, von der Infrastruktur, die man dadurch erschafft. Daher lohnt es sich, zunächst einen Blick auf die technologischen Hintergründe zu werfen – sowohl auf Seiten der Plattformen als auch auf Seiten der Repositories.

Konzeptuell haben sich für die beiden Blickwinkel in der Fachwelt zwei Begriffe etabliert: Die Repositories, die in der Rolle als Datenlieferanten auftreten, werden als Data Provider bezeichnet. Das Gegenstück, die Plattformen, auf denen man Mehrwertdienste anbieten kann, stellen die Service Provider dar. Im Folgenden werden beide Begriffspaare analog benutzt.

Ganz allgemein betrachtet, müsste eine Plattform, also ein Service Provider, zunächst Zugriff auf die Daten aller Repositories haben, die für die Mehrwertdienste der Plattform als Data Provider auftreten sollen. Beginnt ein Service Provider erst nach der Suchanfrage eines Nutzers damit, die Daten der verschiedenen Repositories abzufragen und auszuwerten (ähnlich wie Cross-Search-Verfahren), würde eine Antwort des Dienstes an den Nutzer bei vielen teilnehmenden Repositories sehr lange dauern. Alternativ den Service Provider nur auf Basis eines einzigen, gigantischen Repositories anzubieten, das Kopien aller Daten und aller Volltext-Dateien vorrätig hält und diese parallel zu den Ursprungs-Repositories pflegt, ist jedoch sehr ineffizient und in größeren Dimensionen nicht mehr sinnvoll zu verwalten. In der Realität wurde daher ein praktischer Kompromiss gefunden: Die Volltexte der Dokumente verbleiben physisch in den Repositories, wo sie bei nachhaltiger Planung langzeitarchiviert werden können. Nur die Metadaten – darunter Hyperlinks zu den Volltexten – werden an die Service Provider übertragen, wo sie als Kopien gespeichert und in einer gemeinsamen Datenbasis für die Mehrwertdienste vorgehalten werden.

Es bleibt also die Frage, nach welchem Mechanismus die Repositories den Service Providern ihre Metadaten anbieten können. Der seit etlichen Jahren im Produktiveinsatz befindliche und den Kennern der Thematik hinlänglich bekannte Standard sind die so genannten OAI-PMH-2.0-Schnittstellen. Dies sind spezielle, maschinenlesbare Schnittstellen, auf die die Software der Service Provider automatisiert zugreifen kann, um sich Metadaten in verschiedenen XML-Formaten abzuholen. Um die Daten aktuell zu halten, ohne immer den gesamten Datenbestand kopieren zu müssen, sind die Schnittstellen so konstruiert, dass die abfragenden Programme im Normalfall nur Änderungen bezüglich des letzten Durchlaufs beziehen können. Diesen Mechanismus des inkrementellen Updates nennt man daher auch „Harvesting", zu Deutsch „Ernte".

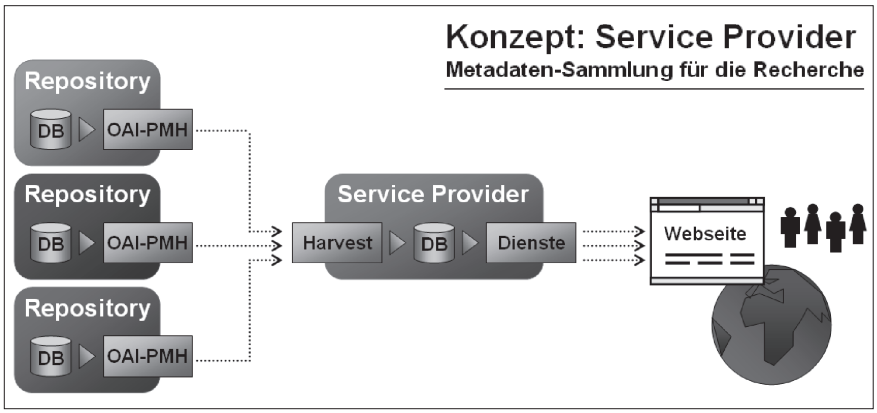

Abb. 1: Mittels des OAI-PMH 2.0 Harvesting-Mechanismus werden Metadaten in der Datenbank eines Service Providers hinterlegt, welcher daraufhin Dienste in einer zur Recherche bestimmten Webseite einer weltweiten Leserschaft anbieten kann.

Herausforderungen und Chancen eines Netzwerks

Die Landschaft der Open Access Repositories ist noch immer sehr heterogen. Im August 2007 startete daher das von der Deutschen Forschungsgemeinschaft geförderte Projekt „Open-Access-Netzwerk“. Dessen Ziel ist es, die Betreiber von deutschen Open Access Repositories zusammenzubringen und zu unterstützen. sollen so homogen zu einem gemeinsamen, virtuellen Datenraum zusammengefügt werden, dass der Nutzer der darauf angebotenen Mehrwert-Dienste deren ursprüngliche Verteiltheit nicht mehr wahrnimmt. Dazu müssen die eingesammelten Metadaten fehlerkorrigiert, harmonisiert, in Bezug zueinander gesetzt und mit weiteren, aus dem Kontext der akkumulierten Datenbasis bezogenen Daten angereichert werden. dem Nutzer direkt neue Sichten und Mehrwerte bieten oder die Datenqualität des Systems weiter erhöhen, wird die API für die REST-Schnittstelle offengelegt. Das Partnerprojekt „Open-Access-Statistik“ wird auf diese Art und Weise Nutzungsstatistiken einbringen und unter anderem auf der gemeinsamen, webbasierten Benutzeroberfläche anbieten. Weitere Projekte sind in Planung.

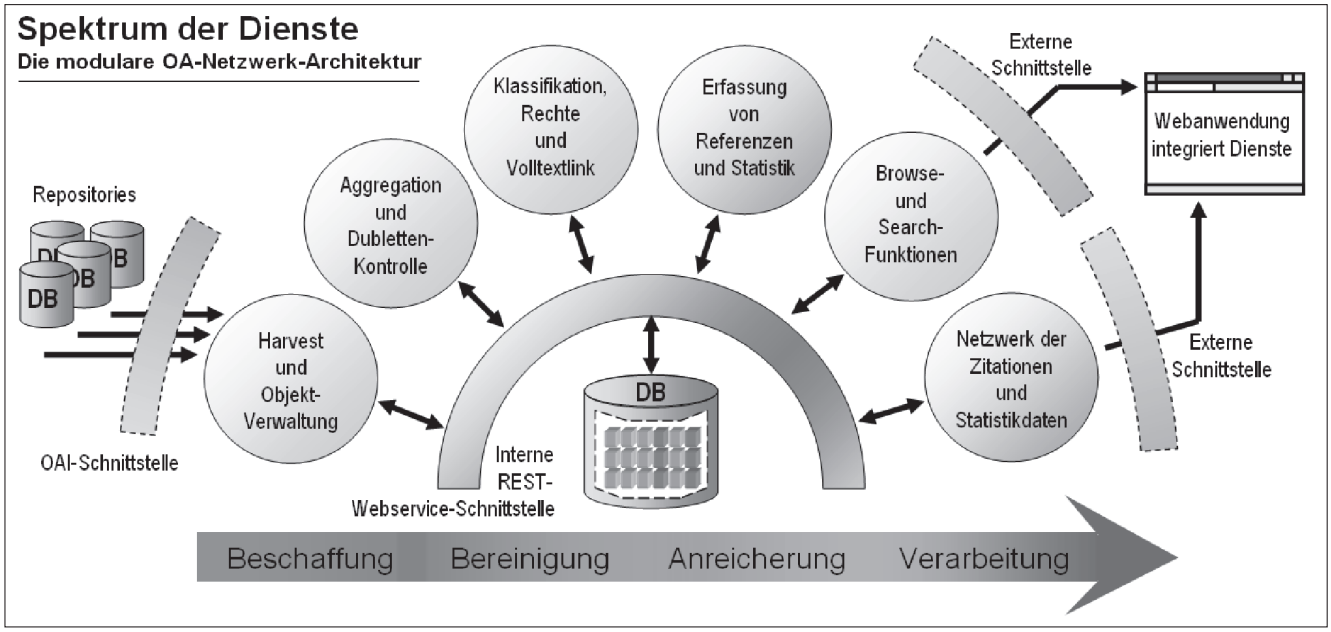

Abb. 2: Die Architektur des Service Providers von Open-Access-Netzwerk ist aus verteilt arbeitenden Modulen aufgebaut, die über eine Webservice-Schnittstelle kommunizieren, um die Metadaten schrittweise zu verbessern, anzureichern und darzubieten.

Auf der organisatorischen Ebene wird das Projekt die Betreiber von Repositories bei der DINI-Zertifizierung unterstützen und liefert Rückmeldung, wie ihre Güte und Leistungsfähigkeit als Data Provider erhöht werden kann. Auf der technischen Ebene stellt das Projekt einen eigenen Service Provider als Platt-form bereit, um neue Mehrwert-Dienste aufzunehmen und die Qualität der bestehenden Dienste zu verbessern. Letztlich dient dies alles dem Ziel, der wissenschaftlichen Gemeinschaft einen Anlaufpunkt zu geben, um zu lesen und gelesen zu werden, wo der in der Open Access Repositories gesicherte Anteil des deutschen Forschungsbeitrags sichtbar wird.

Die größte technische Herausforderung liegt in der Informationsintegration: Metadaten aus heterogenen Quellen

Im System des Service Providers von Open-Access-Netzwerk findet fortlaufend eine solche Wertschöpfungskette in Richtung besserer Metadaten statt. Eine Reihe verteilter, über einen REST-WebService an die zentrale Datenhaltung gekoppelter Dienste bereitet automatisch die durch Harvesting bezogenen Daten auf und reichert sie mit Mehrwertdaten an. Dabei werden unter anderem Klassifikationen hervorgehoben, über eine Analyse ähnliche Objekte zusammengeführt und der zur Volltext-Datei führende Link identifiziert. Die so aufbereitete Datenbasis ist durch eine webbasierte Nutzerschnittstelle für Suche und Browsing, Alerting-Dienste, News-Feeds und OAI-Export zugänglich und bietet verschiedenste Sichten auf den Datenbestand. Um langfristig neue oder verbesserte Dienste zu integrieren, die

Im Projektverlauf hat sich jedoch erneut bestätigt, dass auch mit findigen Algorithmen die integrierten Daten nur bis zu einem gewissen Grad qualitativ hochwertiger sein können als die Originaldaten in den Repositories. Information kann nicht aus nichts entstehen. Eine Beschränkung sind beispielsweise die mit dem angebotenen Metadatenformat abzubildenden Daten. Das ist als größter gemeinsamer Nenner „Dublin Core Simple“. Wie das Format auf Seiten der Repositories konkret angewendet wird, welche Fehler sich bereits bei der Erfassung einschleichen und wie alternative Formate zu integrieren sind – das offenbart viel Handlungsbedarf. Hier muss der technischen Weiterentwicklung eine organisatorische vorausgehen. Die zukünftigen Aufgaben müssen darauf abzielen, das entstehende Netzwerk

zu stärken und die Betreiber der Repositories noch enger in einer aktiven, deutschlandweiten Community zusammenzubringen.

Fazit

Open Access als Modell wissenschaftlicher Literaturversorgung kann langfristig nur dann die gesetzten Erwartungen erfüllen, wenn dem rechtlich-organisatorischen Rahmen des freien Zugriffs auf elektronische Publikationen auch ein entsprechender, infrastruktureller Überbau zur Seite gestellt wird. Erweiterbare Plattformen mit nutzerorientierten Mehrwert-Diensten und darin eingebundene Open Access Repositories mit möglichst hochwertigen Metadaten fördern durch gezielte Zusammenarbeit wechselseitig ihren Mehrwert und damit den Nutzungsanreiz des Open-Access-Weges, den sie der wissenschaftlichen Community bieten.

Literatur

[1] Webseiten der UB Bielefeld: http://www.ub.uni-bielefeld.de/biblio/search/help/invisibleweb.htm; Letzter Zugriff 01.02.2009

[2] DINI-Arbeitsgruppe Open Archives Initiative in Deutschland: *Elektronisches Publizieren an Hochschulen, Inhaltliche Gestaltung der OAI-Schnittstelle*. http://edoc.hu-berlin.de/series/dini-schriften/2-de/PDF/2-de.pdf; 2003

[3] Lagoze, Carl; Van de Sompel, Herbert: *The Making of the Open Archives Initiative Protocol for Metadata Harvesting*. http://public.lanl.gov/herbertv/papers/The%20Maling%20of%20the%20Open%20Archives%20Initiative.pdf; 2002

[4] Dobratz, Susanne; Scholze, Frank: *Qualitätssicherung durch das DINI-Zertifikat*. Erstveröffentlichung in ZfBB 54 (2007) 4–5, S.194–198; http://edoc.hu-berlin.de/oa/articles/rem4Ar8VXZ0CA/PDF/27LxBlVkDRfoA.pdf

[5] Webseiten des Open-Access-Netzwerk-Projektes bei DINI; http://www.dini.de/projekte/oa-netzwerk/; Letzter Zugriff 01.02.2009

[6] Scholze, Frank; Horstmann, Wolfram: *Infrastruktur und Netzwerke für Repositories*. http://www.zbw.eu/ueber_uns/projekte/vascoda/ws_2007-10-31/03_scholze_horstmann_kiel_so.pdf; 2007

[7] Prof. Schirmbacher, Peter; Severiens, Thomas: *Aufbau eines Netzwerkes von Open Access zertifizierten Repositories auf der Basis des DINI-Zertifikats*. DFG-DINI-Workshop „Förderung der wissenschaftlichen Informationslandschaft in Deutschland - Chancen und Strategien beim Aufbau vernetzter Repositories“; http://www.dini.de/fileadmin/workshops/oa-netzwerk-2008/Severiens_Schirmbacher.pdf; 2008