

ABI Technikfrage

Wie kommt das Repository in die Suchmaschine?



<http://doi.org/10.1515/abitech-2018-0017>

Beate Rajski ist Bibliothekarin an der TU Hamburg. Sie leitet in der Bibliothek die Abteilung Digitale Dienste und ist Referentin für Forschungsdaten. Aktuell ist sie im Rahmen des Programms „Hamburg Open Science“ der Hamburger Hochschulen Projektleiterin für das Forschungsdatenmanagement. Sie beantwortet die aktuelle ABI Technikfrage: Wie kommt das Repository in die Suchmaschine?

Viele wissenschaftliche Einrichtungen stellen Publikationen über ein eigenes Repository bereit. Die Publikationen sind meist mit guten Metadaten versehen, über einen persistenten Link verlässlich zitierbar und werden dauerhaft gespeichert. Genutzt werden sie aber erst, wenn sie dort angezeigt werden, wo Forschende suchen. Und das tun diese so gut wie nie im Repository selbst.

Repositories bedienen deshalb eine spezielle Schnittstelle: OAI-PMH („Protocol for Metadata Harvesting“ der Open Archive Initiative). Mit diesem Protokoll wird die Interoperabilität für Metadaten aus Repositories gewährleistet. Beliebige Dienste können über dieses Protokoll Publikationsdaten aus Repositories sammeln und aggregiert zur Verfügung stellen. Bekannte Anwendungen sind Base oder WorldCat.org. Zu Aggregatoren, die für ein Repository besonders relevant sind, nehmen die Betreiber des Repositories in der Regel direkten Kontakt auf, um die Aufnahme sicherzustellen.

Mindestens genauso wichtig ist der Eintrag des Repositories in das Directory of Open Access Repositories unter opendoar.org. Hier können Aggregatoren gezielt relevante Repositories finden und die dort hinterlegten technischen Informationen für die Nachnutzung auswerten. Das EU-Projekt OpenAIRE nutzt dies zum Beispiel für die Zusammenführung von Open Access Veröffentlichungen in der EU.

Trotzdem ist OAI-PMH ein Standard, der außerhalb der bibliothekarischen Welt so gut wie unbekannt ist. Für viele Forschende ist Google Scholar der erste Einstiegspunkt für die Suche nach wissenschaftlichen Artikeln. Google Scholar stellt nicht nur die Daten zu Tausenden von Verlagspublikationen bereit, sondern durchforstet systematisch die Webseiten wissenschaftlicher Einrichtungen auf der Suche nach Veröffentlichungen, die in den Index aufgenommen werden können. Daraus ergibt sich für die meisten Repositories die komfortable Situation,

dass die enthaltenen Publikationen mit kurzem Verzug auch in Google Scholar angezeigt werden.

Für eine optimale Darstellung in Google Scholar ist bei den Repositories Optimierungsaufwand notwendig. Dies beginnt mit der Sitemap, die als maschinenlesbares Inhaltsverzeichnis dient. Sie enthält Links auf alle Seiten, die die Suchmaschinen erfassen sollen. Dazu gehört auch die Angabe des letzten Änderungsdatums. Die Links aus der Sitemap führen zu den sogenannten Landing Pages. Dies sind die Seiten, die in einem Repository eine Veröffentlichung möglichst umfassend mit Metadaten beschreiben. Landing Pages können so unterschiedlich sein wie die Zielgruppen, für die sie entworfen wurden, eine Suchmaschine erwartet aber standardisierte Daten. Um dem Rechnung zu tragen, schreiben Repositories die beschreibenden Metadaten und den Link zum Volltext zusätzlich versteckt als „<meta>“-Tags in den maschinenlesbaren HTML-Header der Seite.

Google Scholar übernimmt diese Daten für seinen Index. Je vollständiger die Meta-Tags sind, desto verlässlicher kann die Suchmaschine verschiedene Quellen einer Veröffentlichung zusammenführen und im besten Fall zu einem kostenpflichtigen Artikel den Link zur Open Access Version aus einem Repository anbieten. Gleichzeitig werden auch Volltexte bis zu einer Größe von 5 MB in den Index aufgenommen und dadurch über Google Scholar auffindbar.

Autoreninformationen



Beate Rajski

Universitätsbibliothek der TU Hamburg
Denickestraße 22
21073 Hamburg
rajski@tuhh.de
<http://orcid.org/0000-0002-0555-2280>