
FIZ-OAI: Eine skalierbare Open-Source-Implementierung zur Aggregation und Dissemination von Metadaten mit OAI-PMH

Stefan Hofmann ^{1*}, Kerstin Soltau ¹, Michael Hoppe ¹, Felix Bach ¹

¹ FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur;

* Kontakt: stefan.hofmann@fiz-karlsruhe.de

Das Protocol for Metadata Harvesting OAI-PMH stammt aus dem Jahr 2001, ist aber immer noch weit verbreitet. Mit FIZ-OAI hat FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur¹ einen skalierenden, robusten und einfach zu wartenden OAI-Provider als Open-Source-Software entwickelt, um große Mengen an Metadaten aus unterschiedlichen Forschungsdateninfrastrukturen zur freien Nachnutzung verfügbar zu machen. Aggregatoren wie CrossRef², die DNB³, Europeana⁴, BASE⁵ oder B2Find⁶, aber auch viele weitere Partner nutzen diesen Service, um Metadaten abzurufen (zu „harvesten“). In diesem Beitrag stellen wir die Architektur und wesentlichen Funktionen von FIZ-OAI vor und skizzieren aktuelle Anwendungsfälle von Forschungs- und Publikationsumgebungen in verschiedenen wissenschaftlichen Disziplinen.

Keywords: OAI-PMH, Metadaten, Metadata Harvesting, Software, Open-Source

1 <https://www.fiz-karlsruhe.de>; *Besucht am 16. Mai 2025.*

2 <https://www.crossref.org/>; *Besucht am 16. Mai 2025.*

3 <https://www.dnb.de>; *Besucht am 16. Mai 2025.*

4 <https://www.europeana.eu/>; *Besucht am 16. Mai 2025.*

5 <https://www.base-search.net/>; *Besucht am 16. Mai 2025.*

6 <https://b2find.eudat.eu/>; *Besucht am 16. Mai 2025.*

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: <https://doi.org/10.11588/heibooks.1652.c23944> (CC BY-SA 4.0).

1 Hintergrund und Motivation

Die rasche Zunahme der Menge an digitalen Forschungsdaten und Publikationen (Bornmann und Mutz 2015) stellt die wissenschaftliche Informationslandschaft vor große Herausforderungen. Für eine effektive Nachnutzung, Archivierung und den interdisziplinären Austausch sind präzise, standardisierte Metadaten unverzichtbar, da sie Inhalte, Entstehungsumstände und wichtige Verfügbarkeitsinformationen (z. B. Lizenz, Speicherort) der Datensätze beschreiben. Nur so lassen sich Datensätze in wissenschaftlichen Repositorien systematisch finden, vergleichen und weiterverarbeiten (Baca 2016).

Neben der Speicherung großer Datenmengen rückt die Frage in den Fokus, wie die Daten nachhaltig bereitgestellt und automatisiert weiterverarbeitet werden können. Dabei hat sich seit mehr als zwei Jahrzehnten das Protocol for Metadata Harvesting (Lagoze und Van de Sompel 2001) der Open Archives Initiative (Lagoze u. a. 2002) bewährt. Dieses ermöglicht den Aufbau verteilter Dateninfrastrukturen, indem Metadaten verschiedener Dienste über eine standardisierte Schnittstelle zentral abgerufen werden können.

Obwohl das OAI-PMH-Protokoll bereits im Jahr 2001 eingeführt wurde und mittlerweile zahlreiche technische Alternativen, wie Linked Open Data⁷ oder ResourceSync⁸ verfügbar sind, findet es weiterhin breite Anwendung⁹ in fast 7.000 verfügbaren OAI-Providern. Diese fortwährende Nutzung lässt sich primär auf seine umfassende Community-Unterstützung, die HTTP-basierte Architektur und die präzise definierte Spezifikation¹⁰ zurückführen.

Stand 2019 betrieb FIZ Karlsruhe eine Vielzahl von Anwendungen (DDB, Thieme Connect, RADAR, zbMATH), deren Metadaten zur freien Nachnutzung verfügbar gemacht werden sollten. Es wurde ein einheitlicher OAI-Provider gesucht, welcher in allen Projekten eingesetzt und in die bestehenden Infrastrukturen integriert werden sollte. Aus den Projekten heraus wurden die folgenden Anforderungen an den Provider definiert:

- Einfache Installation und Betrieb über Docker-Container
- Langfristig wartbare Software auf Java-Basis, die aktuelle Sicherheitsstandards einhält
- Skalierbarkeit für Millionen von Metadatensätzen
- Vollständige OAI-PMH-Implementierung
- Open-Source zur Förderung einer aktiven Community

Die Open Archives Initiative listet auf ihrer Homepage über 30 frei verfügbare OAI-Provider¹¹ als potentielle Kandidaten. Die Wahl fiel auf das erweiterbare OAICat-Framework.

7 https://de.wikipedia.org/wiki/Linked_Open_Data; *Besucht am 16. Mai 2025.*

8 <https://www.openarchives.org/rs/toc>; *Besucht am 16. Mai 2025.*

9 <https://www.openarchives.org/Register/BrowseSites>; *Besucht am 16. Mai 2025.*

10 <https://www.openarchives.org/OAI/openarchivesprotocol.html>; *Besucht am 16. Mai 2025.*

11 <https://www.openarchives.org/pmh/tools/>; *Besucht am 16. Mai 2025.*

work¹². Es basiert auf Java, steht unter einer Open-Source-Lizenz und unterstützt eine flexible Datenhaltung, welche auch die Anbindung von clusterfähiger Datenbanken beinhaltet. Somit waren alle genannten Anforderungen erfüllt. Auf Basis von OAICat wurde bei FIZ Karlsruhe die Software FIZ-OAI entwickelt. Dabei handelt es sich um einen leistungsstarken und modular aufgebauten OAI-Provider, der in der Lage ist, Millionen von Metadatensätzen zu verwalten und bereitzustellen. Die Software steht unter einer Open-Source-Lizenz (Apache 2.0)¹³ frei zur Verfügung, so dass wissenschaftliche Einrichtungen einen eigenen OAI-Provider selbstständig und flexibel aufsetzen können.

2 Architektur

Aufbauend auf den geschilderten Zielsetzungen erläutern wir im Folgenden, wie sich diese in der konkreten Systemarchitektur widerspiegeln. Der nächste Abschnitt gibt detaillierte Einblicke in die technischen Komponenten und deren Zusammenspiel. Abbildung 1 veranschaulicht die Mehrschichten-Architektur von FIZ-OAI, bestehend aus einem Webserver (FIZ-OAI-Provider), einem Anwendungsserver (FIZ-OAI-Backend), einer Datenbank (Cassandra) und einem Suchindex (ElasticSearch / Apache Solr). Hinzu kommt noch eine datengebende (Loader) Komponente, welche das OAI-Backend mit Metadaten befüllt.

Der *FIZ-OAI-Provider*¹⁴ verarbeitet OAI-PMH-Anfragen und beruht auf dem *OAICat-Framework* von OCLC. Beim „Harvesten“ generiert der OAI-Provider für jedes angefragte OAI-PMH-Verb die passende Antwort durch Suchanfragen an das Backend. Die genauen Anfrage-/Antwortformate und Fehlerbehandlungen sind in der OAI-PMH Spezifikation¹⁵ zu finden.

Das *FIZ-OAI-Backend*¹⁶ bietet eine REST-API zur Speicherung und Suche von Metadaten, die in einem Apache Cassandra¹⁷ Datenbank-Cluster abgelegt werden. Cassandra ist eine verteilte NoSQL-Datenbank und ermöglicht linear skalierbare Lese- und Schreibleistungen für große Datenmengen. Um Suchanfragen von OAI-PMH-Verben wie ListIdentifiers/ListRecords effizient bearbeiten zu können, kommt wahlweise *ElasticSearch*¹⁸ oder *Apache Solr*¹⁹ zum Einsatz. Beide Suchplattformen können in einem Cluster betrieben werden und sind extrem leistungsfähig in der Verwaltung von Volltext- und Metadatenindizes.

12 <https://www.oclc.org/research/areas/data-science/oaicat.html>; *Besucht am 16. Mai 2025.*

13 <https://www.apache.org/licenses/LICENSE-2.0>; *Besucht am 16. Mai 2025.*

14 <https://github.com/ER-FIZKarlsruhe/fiz-oai-provider>; *Besucht am 16. Mai 2025.*

15 <https://www.openarchives.org/OAI/openarchivesprotocol.html#ProtocolMessages>; *Besucht am 16. Mai 2025.*

16 <https://github.com/ER-FIZKarlsruhe/fiz-oai-backend>; *Besucht am 16. Mai 2025.*

17 <https://cassandra.apache.org>; *Besucht am 16. Mai 2025.*

18 <https://www.elastic.co/de/elasticsearch>; *Besucht am 16. Mai 2025.*

19 <https://solr.apache.org/>; *Besucht am 16. Mai 2025.*

Eine datengebende Anwendung (*LOADER*), spricht über die REST-API mit dem Backend. Über HTTP werden die Metadaten in einem XML-Eingangsformat übermittelt. Alle Daten werden in Cassandra persistiert und die für OAI-PMH relevanten Felder (Identifier, Datum, Format, Set) für die Suche indexiert. Weiterhin kann das Backend über vom LOADER bereitgestellte XSLT-Transformationen (Crosswalks) das XML-Eingangsformat in weitere XML-Formate transformieren, um den Bedürfnissen unterschiedlicher Communities gerecht zu werden. Durch das Gruppieren von Datensätzen in Sets²⁰ können spezifische Metadaten-Kollektionen separat abgerufen werden.

Ein zentrales Ziel bestand darin, den gesamten Software-Stack via Docker-Container so aufzubereiten, dass er einfach installiert und gewartet werden kann, wodurch Betreiber rasch produktive Instanzen in lokalen Rechenzentren oder Cloud-Umgebungen einrichten können. Hierfür steht ein unterstützendes Projekt²¹ auf GitHub bereit. Das Projekt besteht aus einer Docker-Compose Datei, in welcher obige Komponenten zu einer funktionalen Einheit zusammengefügt werden. Mit Hilfe eines Install-Script, kann in wenigen Minuten eine FIZ-OAI Instanz aufgesetzt werden.

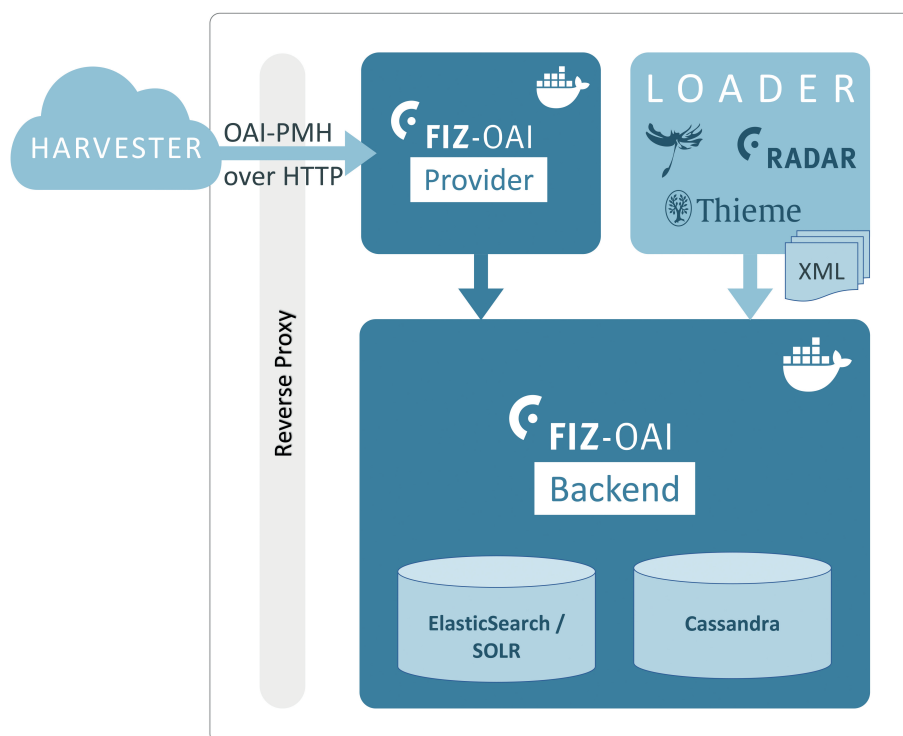


Abbildung 1: FIZ-OAI Architektur.

20 <https://www.openarchives.org/OAI/openarchivesprotocol.html#Set>; Besucht am 16. Mai 2025.

21 <https://github.com/ER-FIZKarlsruhe/fiz-oai-docker>; Besucht am 16. Mai 2025.

3 Anwendungsfälle

Auf dem OAI-Provider von *Thieme Connect*²² stehen aktuell über eine Million Metadatensätze in den Formaten NLM, JATS, MARC zur Verfügung, welche bevorzugt für Zeitschriftenartikel und Bücher verwendet werden. Dies ermöglicht die Bereitstellung von medizinischen Fachartikeln für Aggregatoren wie der Deutschen Nationalbibliothek.

Der OAI-Provider der *Deutschen Digitalen Bibliothek (DDB)*²³, verdeutlicht insbesondere die gute Skalierbarkeit des Systems. Mit über 50 Millionen Metadatensätzen, die in 3 Formaten (DDB, EDM, OAI-DC) abgerufen werden können, sowie 4.232 verfügbaren Sets, handelt es sich um die größte aktuell laufende FIZ-OAI Installation. Diese Formate sind spezifisch für Objekte wie Bilder oder Texte aus Kultur- und Wissenssammlungen.

Auch für Forschungsdatenrepositorien wie *RADAR*²⁴ und die geplanten Dienste in der Nationalen Forschungsdateninfrastruktur (NFDI) bietet FIZ-OAI einen stabilen, wartungsarmen Zugang zum Austausch von Metadaten. Beispielsweise wird gegenwärtig eine Instanz für *NFDI4Chem* aufgebaut, um chemiebezogene Forschungsdaten zentral auffindbar zu machen. Eine *zbMATH* Instanz mit Metadaten aus der Domäne Mathematik soll zeitnah über 4 Millionen Metadatensätze ausliefern.

Tabelle 1: Übersicht der Metadatenfelder.

Instanz	Dokumente	Formate	Harvester
DDB	~ 50.000.000	DDB, EDM	Europeana
Thieme	~ 1.000.000	NLM, JATS, MARC	TIB, DNB, ZB MED
RADAR	~ 2.000	RADAR, DATACITE	B2FIND, BASE, NFDI4Chem
zbMATH Open	~ 4.000.000	DATACITE	MaRDI, DataCite

Danksagungen

Wir danken allen Partnerinstanzen und „Harvestern“, die uns stets konstruktiv bei der Optimierung und Weiterentwicklung der Software zur Seite stehen. Wir laden ausdrücklich zur aktiven Mitwirkung und Weiterentwicklung von FIZ-OAI ein.

²² <https://www.thieme-connect.de/oai/OAIHandler?verb=Identify>; *Besucht am 16. Mai 2025.*

²³ <https://oai.deutsche-digitale-bibliothek.de/?verb=Identify>; *Besucht am 16. Mai 2025.*

²⁴ <https://www.radar-service.eu/oai/OAIHandler?verb=Identify>; *Besucht am 16. Mai 2025.*

Beiträge

Die Grundstruktur des vorliegenden Textes basiert auf einem initialen Entwurf, der mit Hilfe von ChatGPT generiert wurde.

Interessenkonflikte

Die Autor*innen erklären, dass keine Interessenkonflikte bestehen.

Literaturverzeichnis

- Baca, Murtha. 2016. *Metadata and the Web*. 3. Aufl. Herausgegeben von Murtha Baca. Los Angeles: Getty Publications. Besucht am 27. Oktober 2023. <https://www.getty.edu/publications/intrometadata/>.
- Bornmann, Lutz und Rüdiger Mutz. 2015. „Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references“. *Journal of the Association for Information Science and Technology* 66 (11): 2215–2222. <https://doi.org/10.1002/asi.23329>.
- Lagoze, Carl, Herbert Van de Sompel, Michael Nelson und Simeon Warner. 2002. *Open Archives Initiative Protocol for Metadata Harvesting Version 2.0*. Besucht am 21. Mai 2025. <https://www.openarchives.org/OAI/openarchivesprotocol.html>.
- Lagoze, Carl und Herbert Van de Sompel. 2001. „The open archives initiative: building a low-barrier interoperability framework“. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, 54–62. JCDL01. ACM. <https://doi.org/10.1145/379437.379449>.