

Friedrich Summann

Sichtbarkeit und Qualität von Repositorien – aus Sicht eines Service- Providers am Beispiel BASE

Handbuch Repositorienmanagement, Hg. v. Blumesberger et al., 2024, S. 521–547
<https://doi.org/10.25364/978390337423228>



Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz, ausgenommen von dieser Lizenz sind Abbildungen, Screenshots und Logos.

Friedrich Summann, Universitätsbibliothek Bielefeld, friedrich.summann@uni-bielefeld.de |
ORCID iD: 0000-0002-6297-3348

Zusammenfassung

BASE (Bielefeld Academic Search Engine) ist eine global agierende wissenschaftliche Suchmaschine mit Fokus auf institutionelle Repositorien und verwandte Datenquellen im akademischen Kontext. Für OAI-PMH Data Provider ist eine Erfassung via BASE ein Mosaikstein zur globalen Sichtbarkeit der Repositorien-Inhalte und der Quelle selbst. Auf der anderen Seite ist BASE als OAI-Service Provider auf die Qualität der bereitgestellten Schnittstellen (speziell OAI-PMH) angewiesen. Das gilt für die technische Erreichbarkeit und Stabilität, aber auch insbesondere für die gelieferten Metadaten, da diese die Grundlage für die qualitative Bearbeitung der Suchanfragen und der Ergebnisanzeige bilden. In dieser Hinsicht kann BASE als Spiegel der globalen Publikationslandschaft dienen und bietet daher mit Tools wie Quellenliste, OAI-PMH Validator und („Goldene Regeln“ / „Golden Rules“) Empfehlungen, Angebote und Informationen zur Analyse und Optimierung der Repositorien-Infrastruktur. Die Suchmaschine BASE ist seit 2004 produktiv und indexiert zur Zeit mehr als 11.400 Datenquellen auf globaler Basis. Erfasst werden aktuell (Januar 2024) 352 Millionen Publikationen und davon werden ca. zwei Drittel via OAI-PMH Harvesting von Repositorien abgeholt. Daher hat in diesem Kontext BASE mit seinen in der Praxis seit 2004 erworbenen umfassenden Erfahrungen eine gewichtige Rolle in der Repositorien-Community gespielt und hat darauf basierend bei der Abfassung von Guidelines (DRIVER Guidelines, OpenAire Guidelines, DINI-Zertifikat) und Vokabular-Vorbereitungen (COAR Vocabularies) mitgearbeitet. Ganz konkret ist diese Expertise eingeflossen in die sogenannten „Goldenen Regeln“ bei BASE, die grundlegende Empfehlungen eingängig und pragmatisch zusammenfassen und damit in effizienter Weise die Maßnahmen zur qualitativen Sichtbarkeit formulieren und unterstützen. Zahlreiche Kommunikationskanäle mit Repositorien-Manager:innen und -Communities (z. B. DSpace, Eprints, Opus, Goobi, MyCoRe, Invenio) haben zudem zur Optimierung von Schnittstellen beigetragen. Solche Bemühungen sind immer von gegenseitigem Nutzen: BASE ist nicht der einzige Service, der diese Schnittstelle nutzt. Daher kommt die Qualität der Quellen allen Mitgliedern der Repositorien-Landschaft zugute, während der Suchservice bei BASE selbst optimiert wird. Abschließend werden diese grundlegenden Punkte der im BASE-Rahmen entwickelten Qualitätsmerkmale an konkreten Beispielen vorgestellt und erläutert.

Schlagwörter: Repositorium; Zertifikat; BASE; Suchmaschine

Abstract

Visibility and Quality of Repositories – From the Perspective of a Service Provider, Using the Example of BASE

BASE (Bielefeld Academic Search Engine) is a global scientific search engine focusing on institutional repositories and related data sources in the academic context. For OAI-PMH Data Providers, the acquisition via BASE is a mosaic piece that contributes to the global visibility of the repository content and the source itself. At the same time, BASE as an OAI service provider is dependent on the quality of the interfaces provided (especially OAI-PMH). This applies to the technical accessibility and stability and is particularly true for the metadata provided, which form the basis for the qualitative processing of search queries and the display of results. In this respect, BASE can serve as a mirror of the global publication landscape and therefore offers recommendations and information for analysing and optimising the repository infrastructure with tools such as the source list, OAI-PMH Validator and the “Golden Rules” (“Goldene Regeln”). The BASE search engine has been productive since 2004 and currently indexes more than 11.400 data sources on a global basis. It currently (January 2024) covers 352 million publications, about two thirds of which are retrieved from repositories via OAI-PMH harvesting. BASE plays an important role in the repository community with its extensive experience and has contributed to the drafting of guidelines (DRIVER Guidelines, OpenAire Guidelines, DINI Certificate) and vocabulary preparations (COAR Vocabularies). This expertise has been incorporated into the so-called “Golden Rules” at BASE, which summarise basic recommendations in a catchy and pragmatic way and thus efficiently formulate and support the measures for qualitative visibility. Numerous communication channels with repository managers and communities (e.g. DSpace, Eprints, Opus, Goobi, MyCoRe, Invenio) have also contributed to the optimisation of interfaces. Such efforts are always of mutual benefit: BASE is not the only service using this interface. Thus, the quality of the sources benefits all members of the repository landscape, and the search service of BASE itself is optimised. At the end of the contribution, the above-mentioned quality features developed within the BASE framework will be presented and explained using specific examples.

Keywords: Repository; certificate; BASE; search engine

1. Einführung

BASE¹ ist eine global agierende wissenschaftliche Suchmaschine mit Fokus auf institutionelle Repositorien und verwandte Datenquellen im akademischen Kontext. Entstanden im Umfeld der OAI-Bewegung galt zudem dem Kontext von Open Access immer ein besonderes Augenmerk bei der Entwicklung.

BASE operiert überwiegend an der Schnittstelle von OAI-Service Providern und OAI-Data Providern. Dementsprechend sind die Randbedingungen dieser Kommunikation von besonderer Bedeutung für BASE und die darin erfassten Quellen. Für OAI-PMH Data Providern ist eine Erfassung via BASE ein relevanter Mosaikstein zur globalen Sichtbarkeit der Repositorien-Inhalte, aber auch des Repositoriums und der Institution, die es betreibt. Weitere Nachnutzungsdienste mit ähnlicher Ausrichtung im Hinblick auf Sichtbarkeit sind einerseits Verzeichnisse (Registries) auf regionaler, nationaler oder internationaler Ebene wie OpenDOAR² oder der Open Archives Initiative OAI³ und andererseits Aggregatoren und mit BASE vergleichbare Suchinstrumente wie OpenAIRE⁴ und CORE⁵.

Von der Seite des OAI Service Providers betrachtet ist BASE auf die Auffindbarkeit und, wenn diese erfolgt ist, auf die Qualität der bereitgestellten Schnittstellen (speziell OAI-PMH⁶) und der damit transportierten Inhalte in besonderer Weise angewiesen. Das gilt vordergründig für die technische Erreichbarkeit und Stabilität dieser Schnittstellen, weil damit zunächst die Grundlage für die Datenabholung geschaffen wird. Ein besonderer Fokus liegt auf Umfang und Qualität der gelieferten Metadaten, da diese die Basis für die Indexierung und damit für die spätere qualitative Bearbeitung der Suchanfragen und der Ergebnisanzeige bilden.

An dieser Stelle sollte erwähnt werden, dass vor dem Indexieren – und dieses Vorgehen gilt für alle Suchmaschinen – eine breite Palette von Aktivitäten zum Reparieren, Anpassen, Erweitern und Normalisieren der Metadaten stattfindet. Das beinhaltet bei BASE Fehlerbehebungen wie das Entfernen einzelner problematischer Datensätze, Zeichensatzkorrekturen und die Entfernung von Dateiformatfehlern. Die Umsetzung des jeweiligen Metadatenformats (insbesondere beim „unscharfen“ Dublin Core-Format) bedeutet potentiell immer die Behandlung proprietärer Festlegungen (nicht immer findet sich z. B. die URL zum Dokument im <dc:identifizier>-

1 Bielefeld Academic Search Engine: <https://www.base-search.net/>

2 Directory of Open Access Repositories: <https://v2.sherpa.ac.uk/opensoar/>

3 <https://www.openarchives.org/>

4 <https://www.openaire.eu/>

5 <https://core.ac.uk/>

6 Open Archives Initiative Protocol for Metadata Harvesting: <http://www.openarchives.org/pmh/>

Feld), die gegebenenfalls in diesem Schritt korrigiert werden. Bei der Normalisierung werden verschiedene Vokabulare (Standarddefinitionen, aber auch durch eigene Auswertungen abgeleitete Mapping-Tabellen) in festgelegte interne Datenschemata überführt. Mit Text-Mining-Methoden werden dafür geeignete Informationen identifiziert und separat abgelegt wie DOIs, ORCID-iDs und andere Identifier. In einem dreistufigen Verfahren wird der Open-Access-Status aus Angaben zum Repository, gegebenenfalls aus Set-Informationen, einem Open-Access-Status-Vokabular oder einer Lizenzangabe abgeleitet und in einem eigenen Feld abgelegt. Die breite Vielfalt der unterschiedlichen Festlegungen führt dazu, dass in BASE für jede Datenquelle ein individueller Pre-Processing-Schritt vorgenommen wird. Ergänzt werden dann bei jedem Satz Informationen, die aus den gesammelten Meta-Informationen zur Datenquelle erhoben werden. Dazu gehören Attribute wie Land und Kontinent und der interne Bezeichner der Quelle, so dass diese Aspekte auch im Zusammenhang mit dem Dokument abfragbar sind. Gleichzeitig können Herkunft und Kontext des Dokuments in den Ergebnislisten über die Information zur Datenquelle näher beschrieben werden. Ein beträchtlicher Teil dieser Vorarbeiten lässt sich in den Möglichkeiten der erweiterten Suchmaske, den angebotenen Facetten zur Suchverfeinerung und letztlich auch in der Ergebnisanzeige wiederfinden. So zeigt die nächste Abbildung die zur jeweiligen Datenquelle in der Trefferliste abrufbare Metainformation.

Datenlieferant:	GAMS - Geisteswissenschaftliches Asset Management System (Zentrum für Informationsmodellierung, Universität Graz) <i>GAMS - Humanities' Asset Management System (Austrian Centre for Digital Humanities, University of Graz)</i>   ✦ URL: http://gams.uni-graz.at/ ✦ Kontinent: Europa ✦ Land: at ✦ Breiten- und Längengrad: 47.077787 / 15.449913 (Google Maps OpenStreetMap) ✦ Anzahl der Dokumente: 34.896 ✦ Open Access: 34.896 (100%) ✦ Typ: Digitale Sammlung ✦ System: Fedora ✦ Datenlieferant in BASE seit: 2015-05-06 ✦ BASE URL: https://www.base-search.net/Search/Results?q=coll:ftunivgrazgams
-----------------	--

Abb. 1: BASE-Zusatzinformation zur Datenquelle in der Ergebnisliste

Mit der Gesamtheit der erfassten Datenquellen als Grundlage kann BASE als ein Spiegel der globalen Publikationslandschaft dienen und bietet mit Tools wie Quellenliste, OAI-PMH-Validator und den sogenannten „Goldene Regeln“-Empfehlungen⁷ („Golden Rules“) Service-Angebote und Informationen zur detaillierten Analyse der potentiellen sowie der bereits erfassten Datenquellen. Damit wird auch eine Grundlage zur Optimierung der Repositorien-Infrastruktur geliefert, die sowohl in der Startphase bei den vorbereitenden Aktivitäten als auch nach der Bereitstellung zum nachträglichen Feintuning Verwendung finden kann.

Um den Entwicklungen im Bereich der Repositorien Rechnung zu tragen und das Serviceangebot aktuell zu halten, ist BASE darauf angewiesen, neue Entwicklungen zu berücksichtigen. Zuletzt waren das Features wie das Open-Access Boosting (Open-Access-Publikationen werden in den Suchergebnissen höher gerankt) und die gesonderte Behandlung und Bereitstellung für die Suchanfragen von DOI, OA-Status, Lizenzangaben und auch ORCID-Identifizier. Die umfangreiche Ausweitung auf Crossref-Quellen⁸ für akademische Inhalte und die Anreicherung von Open-Access-Links via unpaywall⁹ zeigt auch, dass neue Schnittstellen genutzt werden, wenn diese wertvolle Zusatzinformationen liefern können.

2. BASE, Suchindex und globales Discovery-System

Die Suchmaschine BASE ist seit 2004 produktiv und indexiert 2024 mehr als 11.400 Datenquellen auf globaler Basis. Erfasst werden aktuell (Januar 2024) 352 Millionen Publikationen und davon werden ca. zwei Drittel via OAI-PMH-Harvesting von Repositorien abgeholt. BASE spielt mit seinen umfassenden Praxiserfahrungen¹⁰ eine gewichtige Rolle in der Repositorien-Community und hat auf dieser Grundlage auch bei der Abfassung von Guidelines (DRIVER Guidelines¹¹, OpenAIRE Guidelines¹², DINI-Zertifikat¹³) und Vokabular-Vorbereitungen (COAR Vocabularies¹⁴) mitgearbeitet. Ganz konkret bezogen auf die BASE-Anforderungen ist diese Expertise in die „Goldenen Regeln“ eingeflossen, die grundlegende Empfehlungen eingängig und praxisorientiert zusammenfassen und damit in effizienter und pragmatischer Weise die konkreten Maßnahmen zur qualitativen Sichtbarkeit formulieren und

7 https://www.base-search.net/about/de/faq_oai.php

8 <https://www.crossref.org/>

9 <https://unpaywall.org/>

10 Pieper, D.; Summann, F. (2015)

11 Vanderfeesten, M.; Summann, F.; Slabbertje, M. (2008)

12 <https://guidelines.openaire.eu/en/latest/>

13 Müller, U.; Scholze, F.; Vierkant, P. et al. (2019)

14 Controlled Vocabularies for Repositories: <https://vocabularies.coar-repositories.org/>

unterstützen. Zahlreiche Kommunikationskanäle mit einzelnen Repositorien-Managern:innen und auch -Communities (z. B. DSpace, EPrints, Opus, Goobi, MyCoRe, Invenio) haben zudem zur Optimierung der Schnittstellen dieser technischen Systeme beigetragen.

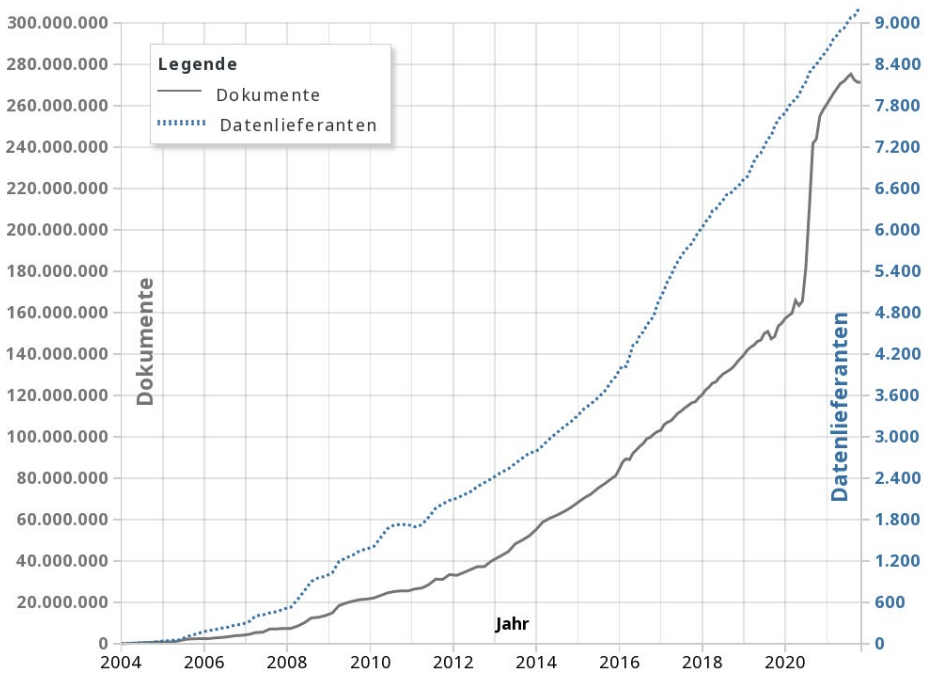


Abb. 2: Die Entwicklung der BASE-Suchmaschine seit 2004

3. Die Goldenen Regeln – aus Praxiserfahrungen kondensierte Empfehlungen

BASE indexiert grundsätzlich alle Quellen mit wissenschaftlichen Inhalten (Zeitschriften, Repositorien, Digitale Sammlungen, Forschungsdatenrepositorien etc.), die über eine OAI-Schnittstelle verfügen und die Metadaten über OAI-PMH liefern. Entsprechend ist die Ausgestaltung dieser Schnittstelle entscheidend, ob und wie die auf diesem Wege erfassten Dokumente optimal und vollständig in BASE präsentiert werden.

OAI-PMH (OAI Protocol for Metadata Harvesting) hat inzwischen eine mehr als 20-jährige Geschichte hinter sich, und natürlicherweise haben sich Probleme¹⁵ mit dem Protokoll herausgestellt und in mancher Hinsicht auch verstärkt. Ein Grundaspekt ist, dass das OAI-PMH-Protokoll nicht web-nativ ist. Das führt dazu, dass es von der technischen Web-Entwicklung entkoppelt ist und – noch ungünstiger – weniger Beachtung bei der Nutzung findet. So fällt es weit weniger auf, wenn es Probleme mit der Schnittstelle gibt. Dies könnte auch der wesentliche Grund sein, warum Google Scholar nach anfänglichen Experimenten mit OAI-PMH auf das Google-übliche Web-Crawling setzte. Der zweite große Problembereich ist die Anfälligkeit für Unterbrechungen und Ausfälle. Zwar erlaubt das Protokoll ein Wiederaufsetzen, was in der Praxis allerdings in den meisten Fällen wirkungslos ist und nur bei externen Verbindungsproblemen hilft. Oftmals hingegen sind die Ursachen interne Probleme (OAI-Meldung „Internal Error“) wie Zeichensatzprobleme einzelner Datensätze, was zur nicht überbrückbaren Wiederholung der Fehlersituation führt. Zur Verarbeitung größerer Datenmengen werden die Schnittstellenaufrufe aufgeteilt und sequentiell durchgeführt. Der Datenprovider liefert einen Resumption Token in seiner Antwort, der beim nächsten Aufruf als Parameter mitgeliefert wird und somit als Kontrollelement zur korrekten Abwicklung verwendet wird. Diese Technik zur Weiterschaltung erweist sich als unzuverlässig und vor allem im Fehlerfall als nicht umgehbar. Das hat zur Folge, dass ein immer wiederkehrender Absturz nur in Ausnahmefällen (z. B. durch Editierung des Resumption Tokens, was in manchen Fällen möglich ist) überwunden werden kann. Immerhin kommt eine andere Fehlerquelle recht selten vor – nämlich die Endlosschleife, bei der meist immer die gleiche Antwortdatei ohne ein reguläres Ende ausgegeben wird. Diese Situation ist immer mit der Gefahr verbunden, dass die Festplattenkapazitäten des Datenabrufers zur Neige gehen. Die Probleme haben sich durch die immer größer werdenden Repositorien und die damit verbundenen längeren Laufzeiten des Harvesting-Vorgangs deutlich verstärkt.

15 Tennant, R. (2004), S. 32.

Trotz der offensichtlichen Schwachpunkte des Protokolls haben sich die Weiterentwicklungen wie OAI-ORE¹⁶ und ResourceSync¹⁷ bisher nicht durchsetzen können. Aus dieser rückblickenden Sicht scheint es doch voreilig gewesen zu sein, die Option einer weiteren Protokollversion 3.0 aufzugeben. Da das Protokoll mit seiner Einfachheit durchaus in mancher Hinsicht besser ist als sein Ruf, hätte dieser Schritt ein pragmatischer Ansatz zur Behebung der gravierendsten Fehler sein können. Neben den technischen Problemen wäre die Festlegung eines alternativen Pflichtmetadatenformats mit weniger „Unschärfe“ als Dublin Core sicherlich eine weitere Option für eine wesentliche Verbesserung gewesen.

Unabhängig davon hat es sich für BASE als notwendig erwiesen, unablässig auf Entwicklungen im Bereich Repositorien zu reagieren und diese zu integrieren. Dazu gehörten die Erweiterung zusätzlicher Publikationsservertypen wie Digitale Sammlungen oder CRIS-Systeme, aufkommende neue Metadatenformate wie OAI DataCite (oai_datacite) und auch alternative Metadatenschnittstellen wie Crossref.

Vor diesem Hintergrund lassen sich Empfehlungen zu Einstellungsoptionen und Probleme mit OAI-Schnittstellen auf drei unterschiedliche Bereiche aufteilen:

- Protokollfehler und technische Kommunikationsprobleme
- Metadaten-Festlegungen und deren Defizite und Unschärfen
- Allgemeine Kommunikationsregeln

Im Folgenden sollen diese Punkte näher betrachtet werden. Die Betrachtung folgt im Wesentlichen den Goldenen Regeln, ausführliche Details sind dort zu finden.

3.1. OAI-Schnittstelle allgemein

Die OAI-Schnittstelle eines Repositoriums sollte frei zugänglich und stabil sein und konstant antworten. Die Anfrage nach ListRecords im Pflichtformat oai_dc liefert Ergebnisse, ohne dass es dabei zu einem Absturz oder einem Ausgabefehler kommt. In regelmäßigen Abständen sollte die Funktionsfähigkeit der OAI-Schnittstelle, z. B. per Browser, geprüft werden.

16 Open Archives Initiative Object Exchange and Reuse: <https://www.openarchives.org/ore/>

17 ResourceSync Framework Specification (ANSI/NISO Z39.99-2017): <http://www.openarchives.org/rs/1.1/resourcesync>

Laut Protokoll ist das Format oai_dc zwingend notwendig, was aus Effizienzgründen – alle optionalen Metadatenformate sind nur in eingeschränktem Umfang vertreten – zur Folge hat, dass meist oai_dc, also Dublin Core, für die Weiterverarbeitung benutzt wird.

Ein recht häufiger Fehler entsteht, wenn das gelieferte XML-Format der Datei formal nicht korrekt ist und dadurch die gesamte Datei nicht weiterverarbeitet werden kann. In diesem Fall kann die Datenstruktur nicht via Standard-Parser-Verfahren analysiert werden. Oftmals ist dieser Fehler verbunden mit falscher Zeichencodierung des Dokuments. Auch diese Eigenschaft kann durch Überprüfung der Schnittstelle im Browser betrachtet werden.

Mit Hilfe des sogenannten inkrementellen Harvestings, wie es im Protokoll über den Parameter ‚from‘ unterstützt wird, rufen Service Provider nur die Ergänzungen und Veränderungen der Metadaten ab einem durch den Parameter festgelegten Zeitpunkt auf Datensatzbasis ab. Bei der inzwischen sehr großen Zahl von Repositorien mit ihren teilweise hohen Satzzahlen ist es für OAI-Service Provider üblicherweise unverzichtbar, diese Technik zu verwenden, um das aufwändige periodische Komplett-Harvesting zu vermeiden. Dabei müssen die gelöschten Datensätze per Deleting Strategy ‚transient‘ oder ‚persistent‘ ausgeliefert werden. Immer wieder kommt es allerdings zu mehr oder weniger umfangreichen Abweichungen, was dazu führt, dass bei BASE periodisch ein sogenannter Refresh vorgenommen wird. Damit wird vermieden, dass neue Datensätze, Korrekturen oder Löschungen fehlen.

3.2. Allgemeine Konfigurationseinstellungen

Bei jeder ListRecords-Antwort einer OAI-Schnittstelle sollten im Idealfall 50 – 1.000 Datensätze ausgegeben werden. Der sogenannte Resumption Token am Ende einer OAI-PMH-Antwortdatei sollte funktionieren und die nächsten Datensätze ausliefern. Werden weniger als 50 Datensätze pro Seite ausgeliefert, führt dies zu entsprechend zahlreichen einzelnen Aufrufen mit entsprechend erhöhtem Kommunikationsaufwand. Mehr als 1.000 Datensätze pro Seite machen dagegen die gelieferten Dateien verhältnismäßig groß und erhöhen damit oftmals das Risiko von Abbrüchen beim Harvesten der Datensätze. Funktioniert der Resumption Token-Mechanismus nicht, ist eine vollständige Indexierung aller Datensätze des Repositoriums nicht möglich.

3.3. Inhalte und Metadatenqualität

Von besonderer Bedeutung für die Qualität der Suche und der Informationsbeschreibung der nachnutzenden Dienste ist die Qualität der gelieferten Metadaten. Insbesondere wird mit der Anwendung von bibliographischer Expertise und Datenpflege (data curation) die Grundlage geschaffen, die die Sichtbarkeit des Repositoriums und seiner Inhalte von automatischen Crawling-Techniken und der Extraktion differenzierter Felder von Webseiten qualitativ unterscheiden lässt. Das gilt insbesondere für die inhaltliche Zuordnung der Felder, aber auch für Umfang und Qualität der jeweiligen Feldinhalte.

3.4. Zeichenkodierung

Probleme gibt es immer wieder mit der Zeichenkodierung, sowohl im internen Datenfluss (was verschiedentlich zu systeminternen Abbrüchen bei der Datenbereitstellung via OAI-PMH führt) als auch beim ausgelieferten XML-Format, was zu Parser-Problemen führen kann. Daher sollten alle Inhalte in der OAI-Schnittstelle (insbesondere die textbasierten Felder Titel, Autor:innennamen, Abstracts) korrekt in UTF-8 kodiert sein. Andere Kodierungen oder doppelte UTF8-Kodierungen erzeugen Fehler in der Indexierung und als Folge auch Fehler bei der Darstellung von Suchergebnissen.

3.5. Vollständigkeit und Qualität der Metadaten

Protokoll-bedingt wird von Service Providern überwiegend das Standard- und OAI-PMH-Pflichtformat oai_dc (Dublin Core) verwendet, da jeder Data Provider es anbieten muss und es daher aus Effizienzgründen für alle Datenanbieter Verwendung finden kann. Als allgemeine Anforderung kann festgehalten werden, dass jeder Datensatz einer OAI-Schnittstelle möglichst vollständige Metadaten zu einem Dokument liefern und dabei nach Möglichkeit standardisierte Vokabulare verwenden sollte. Ganz wesentlich – und in gewisser Hinsicht ein K.o.-Kriterium – ist die Angabe einer funktionierenden URL, die auf das Dokument weist, üblicherweise in <dc:identifizier>. Nur dann ist möglich, aus den Suchergebnissen einen Link zum Repository und dem darin enthaltenen Dokument bereitzustellen.

Allgemein gilt: Je vollständiger die Metadaten sind, umso besser werden die zugehörigen Dokumente aus dem Repository in den Suchumgebungen der jeweiligen Service Provider auffindbar sein. Von besonderer Bedeutung ist die Verwendung von standardisierten Vokabularen, um Dokumente sprachübergreifend z. B. einem eindeutigen und korrekten Dokumenttyp oder Nachnutzungsrecht zuzuordnen.

Dublin Core ist das OAI-PMH-Pflichtdatenformat, was gewisse Vorteile für die universelle und effiziente Bereitstellung und Nutzung mit sich bringt. Erkauft wird das allerdings mit einer großen Unschärfe in der Verwendung der Felder und der Präsentation der Feldinhalte.

3.6. Hinweise zu einzelnen Metadaten-Feldern (Dublin Core)

Die folgende Auflistung zeigt eine Liste der Dublin-Core-Felder¹⁸ nach Relevanz (Werte sind Muss, Soll und Kann) aus Sicht der BASE-Suchmaschine:

Information	Element in oai_dc	Angabe
URL der Publikation	<dc:identifizier>	Muss
Titel	<dc:title>	Soll
Autor	<dc:creator>	Soll
Publikationstyp	<dc:type>	Soll
Erscheinungsdatum	<dc:date>	Soll
Sprache des Dokuments	<dc:language>	Soll
Zugriffs- und Nachnutzungsrechte	<dc:rights>	Soll
Quellenangabe/Zitation	<dc:source>	Soll
Sonstige an der Publikation Beteiligte	<dc:contributor>	Kann
Dateiformat	<dc:format>	Kann
Inhaltsbeschreibung	<dc:description>	Kann
Schlagwörter	<dc:subject>	Kann
Verlag	<dc:publisher>	Kann
Verwandte Dokumente	<dc:relation>	Kann
Inhaltliche Abgrenzung	<dc:coverage>	Kann

Konkret lässt sich im Hinblick auf die Dublin-Core-Felder festhalten:

Bereits weiter oben ist auf die elementare Notwendigkeit einer URL als Link zum Dokument hingewiesen worden. Es kommt durchaus vor, dass gar kein Link zum Dokument oder zur Landing Page angeliefert wird oder nur ein (lokaler) Identifier, mit dem sich in den meisten Fällen und mit einem gewissen Aufwand ein Link konstruieren ließe. Effizient sind solche individuellen Lösungen aber nicht. Ein weiteres, häufiger vorkommendes Problem ist es, dass im Feld <dc:identifizier> Links zum Dokument zu finden sind, die nicht funktionieren – oftmals verursacht durch eine Fehlkonfiguration im zugrundeliegenden Publikationssystem.

Mit Blick auf wissenschaftliche Publikationen und ihre Einordnung und Bewertung sind Informationen in den Feldern Titel <dc:title>, Autor:in (<dc:creator>, <dc:contributor>), Erscheinungsdatum (<dc:date>) und, wenn vorhanden, zur Zitatangabe insbesondere bei Zeitschriftentiteln (<dc:source>) von besonderer Relevanz und

18 Siehe https://www.base-search.net/about/de/faq_oai.php

begründen die Qualität der Metadaten und damit auch die Sichtbarkeit des einzelnen wissenschaftlichen Dokuments.

Persistente Identifier (z. B. DOI, ISBN, ISSN etc.) sollten in <dc:identifier> ausgegeben werden und sind für die Verlinkung von formal oder inhaltlich verbundenen Objekten und das Data Enrichment (in dem aus externen Quellen zusätzliche Daten zu dem über den Identifier zugeordneten Datenobjekt eingespielt werden) von besonderer Wichtigkeit.

Es sei darauf hingewiesen, dass Dublin Core-Felder grundsätzlich wiederholbar sind und daher keine Aufführung mehrerer Attribute (Autoren, Klassifikationscodes) in einem einzigen DC-Feld notwendig ist. Auf der anderen Seite kann bei Datenverbund-Strukturen (Attribute wie E-Mail-Adresse, ORCID-iD oder Affiliationsangaben bei Autor:innen) die Verwendung eines Separatorzeichens, meist ein Semikolon, zur Trennung der jeweiligen Abschnitte sinnvoll sein.

Vokabular-gestützte Felder erlauben eine inhaltsscharfe, generische und einheitliche Verwendung der auftretenden Terme in der Suche, z.B. durch trennscharfe Auswahlboxen und eine präzise Facettierung der Suchergebnisse. Erfahrungsgemäß enthalten solche Felder in der Praxis eine Vielzahl unterschiedlicher Terme, teilweise aus proprietären, oft lokal definierten Vokabularen, teilweise frei wählbar und häufig in verschiedenen Sprachen.

Das gilt ganz besonders für die Felder <dc:date> (Erscheinungsjahr), <dc:language> (Sprache), <dc:type> (Publikationstyp), <dc:rights> für die Lizenzangaben und den Open Access-Status.

Das Feld <dc:date> sollte nur einmal belegt sein. Wenn kein konkretes Erscheinungsdatum vorliegt, sollte geschätzt werden.

Sprache des Dokuments <dc:language>

Angaben zur Sprache eines Dokuments nach ISO 639¹⁹ (2- oder 3-Letter-Code) sollten im Feld <dc:language> zur Verfügung gestellt werden. Angaben zur Sprache werden in BASE für Dokumente ansonsten nicht oder fehlerhaft ausgegeben und die Einschränkung auf eine Sprache funktioniert für die Quelle nicht korrekt.

19 ISO 639 Code Tables: https://iso639-3.sil.org/code_tables/639/data

Zugriffs- und Nachnutzungsrechte <dc:rights>

Zugriffsrechte (Access-Status)

Im Feld <dc:rights> sollten Zugangsinformationen zum Volltext nach dem info-eu-repo-Access-Rights-Vokabular oder dem COAR-Access-Rights-Vokabular enthalten sein. Für die Endnutzer:innen ist die Information über den Zugang zu einem Dokument in der Trefferliste von besonderer Bedeutung. Stehen diese Angaben nicht oder nur unzureichend zur Verfügung, werden Informationen zum Zugang von Dokumenten unvollständig, gar nicht oder fehlerhaft ausgegeben und die Einschränkung auf bestimmte Zugangsarten funktioniert für die Quelle nicht richtig.

Nachnutzungsrechte (Lizenzen)

Den Autor:innen sollte die Möglichkeit angeboten werden, Dokumente unter eine Lizenz zu stellen. Dabei sollten möglichst weit verbreitete Lizenzen wie die Creative-Commons-Lizenzen zur Auswahl angeboten werden. Die entsprechende Lizenz sollte in der OAI-Schnittstelle in einem weiteren <dc:rights>-Feld angeliefert werden.

Stehen diese Angaben nicht oder nur unzureichend zur Verfügung, werden Informationen zur Nachnutzung von Dokumenten unvollständig, gar nicht oder fehlerhaft ausgegeben und die Einschränkung auf Nachnutzungsmöglichkeiten funktioniert für die jeweilige Quelle nicht korrekt.

Quellenangabe / Zitation <dc:source>

Angaben zur Quelle oder zur Zitation (insbesondere bei Artikeln, Titel, Band und Heft der Zeitschrift) stehen bevorzugt in <dc:source>. Dabei sollte insbesondere darauf geachtet werden, die ISSN der enthaltenden Zeitschrift anzugeben.

Diese Angaben ermöglichen es Nutzer:innen, die Dokumente besser zu finden und damit auch besser zitieren zu können.

3.7. Über BASE und OAI hinaus: Organisatorische Vorbereitungen

Kontaktpersonen

In den Identify-Angaben der OAI-Schnittstelle ist im Feld adminEmail eine E-Mail-Adresse anzugeben, über die eine Kontaktaufnahme zum technischen Betreiber der OAI-Schnittstelle möglich ist. Zudem ist es empfehlenswert, auf der Homepage des Repositoriums eine (funktionierende) E-Mail-Adresse zur Verfügung zu stellen, über die die direkte Kontaktaufnahme zum Betreiber gewährleistet ist. Nur wenn

die E-Mail-Adressen funktionieren und E-Mails gelesen und beantwortet werden, kann bei Problemen oder Fragen Kontakt mit dem/der Repository-Verantwortlichen aufgenommen werden.

Kontakt-Bereich mit aktiver E-Mail-Adresse

Die Adress-Information im Identify-Response ist auf die technischen Details der OAI-Schnittstelle ausgerichtet. Für allgemeine Fragen und für Endnutzer:innenanfragen sollte von der Startseite ein Kontakt-Bereich verlinkt werden (Kontakt/Contact). Dort ist die funktionierende E-Mail-Adresse des/der inhaltlichen und/oder technischen Verantwortlichen anzugeben. E-Mails, die an diese Adressen geschickt werden, sollten von den Verantwortlichen regelmäßig gelesen und beantwortet werden. Fehlt ein Kontakt-Bereich oder werden E-Mails nicht gelesen, ist eine Kontaktaufnahme kaum möglich, wenn es bei der Indexierung des Datenlieferanten zu Problemen kommt oder sich Rückfragen ergeben. Dies kann dazu führen, dass die betreffende Quelle nicht indexiert werden kann.

Web-Adresse des Repositoriums

Erfahrungsgemäß sollte die Startseite möglichst unter einer eigenen Subdomain (ohne Port und Unterverzeichnis) angeboten werden. Jede Änderung am Port oder einem Unterverzeichnis führt dazu, dass die Links zu dieser Quelle nicht mehr funktionieren, oft auch die damit verbundene sogenannte OAI base URL. Vermieden werden sollten auch Versionsnummern in der Subdomain oder Verzeichnisnamen (z.B. ojs3.domain.de oder ojs.domain.de/ojs-3/). Wie zuvor erwähnt, führt jede derartige Änderung an der URL dazu, dass Links auf die Quelle nicht mehr erreichbar sind.

Ändert sich etwas an der Internet-Adresse des Datenlieferanten (und sei es nur ein Zeichen), sollte möglichst von der alten Adresse eine Weiterleitung auf die neue gesetzt werden. Es sollte auch darauf geachtet werden, dass die OAI-Schnittstelle durch Setzen einer Weiterleitung weiterhin erreichbar ist.

Üblicherweise führen Fehler in der Erreichbarkeit dazu, dass diese Quellen von den Service Providern gelöscht werden.

Der Name des Repositoriums oder der Titel einer Zeitschrift sollte immer im Quelltext der Webseite an einer Stelle im Klartext zu finden sein, entweder im <title>, der Überschrift (<h1>) oder als Alternativtext eines Logos. Damit ist das Branding des Datenanbieters gewährleistet und zugleich wird mit dieser Information der wissenschaftliche Kontext transparenter.

Startseite auch in Englisch

Für das Repositorium sollte zumindest die Startseite auch in englischer Sprache angeboten werden. Die Repositorien-Community ist global ausgerichtet und durch einen Internetauftritt in englischer Sprache wird einem internationalen Publikum ein unkomplizierter Zugriff auf die Inhalte ermöglicht. Eine englischsprachige Startseite führt auch zu einer besseren Auffindbarkeit der Quelle und ihrer Inhalte in allen damit indexierenden Suchdiensten.

Informationen über grundlegende Änderungen beim Repositorium

Sollten sich der Name des Datenlieferanten oder die URL der OAI-Schnittstelle ändern (z. B. durch den Umzug auf ein anderes technisches System), sollte dieses in den einschlägigen Verzeichnissen mitgeteilt werden, im Falle von BASE über das zugehörige Kontaktformular.

Sichtbarkeit der Quellen und Indexierung in Suchmaschinen

Die Quelle sollte in den OAI- und Repositorien-Verzeichnissen angemeldet sein (z. B. OpenDOAR, ROAR²⁰, re3data²¹, openarchives.org, BASE, aber auch nationale und community-bezogene Auflistungen) und es sollte auf die Aktualisierung bei Änderungen der Angaben geachtet werden.

Auf diese Weise sind die Quelle und damit auch die Schnittstelle weltweit sichtbar und es ermöglicht gegebenenfalls allen interessierten Suchmaschinen die Indexierung von Dokumenten des Repositoriums.

Im Rahmen des Quellen-Discovery werden diese Verzeichnisse z. B. von BASE automatisch (via API oder durch Text Mining) in regelmäßigen Abständen gescannt und neue Einträge gefunden.

Die gute Auffindbarkeit einer Quelle in allgemeinen und wissenschaftlichen Suchmaschinen führt dazu, dass Dokumente aus der Quelle leichter zu finden sind und häufiger abgerufen und genutzt werden.

20 Registry of Open Access Repositories: <http://roar.eprints.org/>

21 Registry of Research Data Repositories: <https://www.re3data.org/>

4. Die globale Repositorienlandschaft – Analysen aus BASE-Sicht

Im Falle der Suchmaschine BASE liegen die Metadaten in der Rohfassung (nach dem Harvesting-Vorgang) im Datenspeicher, zugleich aber auch in normalisierter Form vor. So werden im Zuge der Normalisierung Datumsformate, Sprachangaben und Publikationstypen angepasst und Inhalte wie DOIs und ORCID-iDs extrahiert. Zudem werden Informationen aus internen und externen Datenquellen ergänzt. Dazu gehören Metainformationen zum Datenanbieter (insbesondere Ländercode, Bezeichnung und ROR-id) und von besonderer Relevanz die Open-Access-Status- und Linkinformationen von unpaywall. Diese Angaben werden vor dem Indexieren in einem speziellen Indexformat abgelegt und sind daher nach der Indexierung in transformierter Form im Index vorhanden und damit in komplexer Weise abfragbar. Dadurch sind direkte Analysen in den Daten möglich, die durch die umfassenden Suchmöglichkeiten im Index ergänzt und unterstützt werden können. Zusätzlich werden Metadaten zu den Repositorien selbst erfasst, um Hinweise zum wissenschaftlichen Kontext (Herkunft, publizierende Institution) insbesondere für die Endnutzer:innen, aber auch für Auswertungszwecke liefern zu können. Insgesamt können dadurch über Analysen umfangreiche Indikatoren²² für die Repositorienlandschaft auf Länderebene, aber auch zum einzelnen Repository selbst geliefert werden.

Da viele der Informationen zu definierten Zeitpunkten abgespeichert werden, können derartige deskriptive Darstellungen über längere Zeiträume vorgenommen werden und es ist damit möglich, Entwicklungen über Timeline-Analysen aufzuzeigen. Im Folgenden sollen einige Auswertungen vorgestellt werden, die Aufschlüsse über die globale und nationale Repositorienlandschaft mit besonderem Fokus auf Österreich erlauben und auf diese Weise auch vergleichende Analysen ermöglichen. Die Auswertungen basieren sämtlich auf den Indexdaten der BASE-Suchmaschine in Kombination mit den im BASE-Kontext erfassten Metadaten der BASE-Datenquellen.

BASE operiert global und von daher kann die Suchmaschine auch Zahlen und Übersichten in Form von Weltkarten ausgeben. So zeigt die nächste Abbildung die Anzahl der Dokumente der Publikationsserver für alle Länder der Erde mit einem Wertebereich und zugeordneter Farbskala von blau (0 = niedrig) bis rot (1.000.000

22 Summann, F.; Czerniak, A.; Schirrwagen, J.; Pieper, D. (2020). S. 35.

und mehr = hoch). Deutlich erkennbar sind die Schwerpunkte in Europa, Nordamerika und Ozeanien, aber auch die starke Rolle der Staaten in Südamerika, China, Indien und Indonesien.

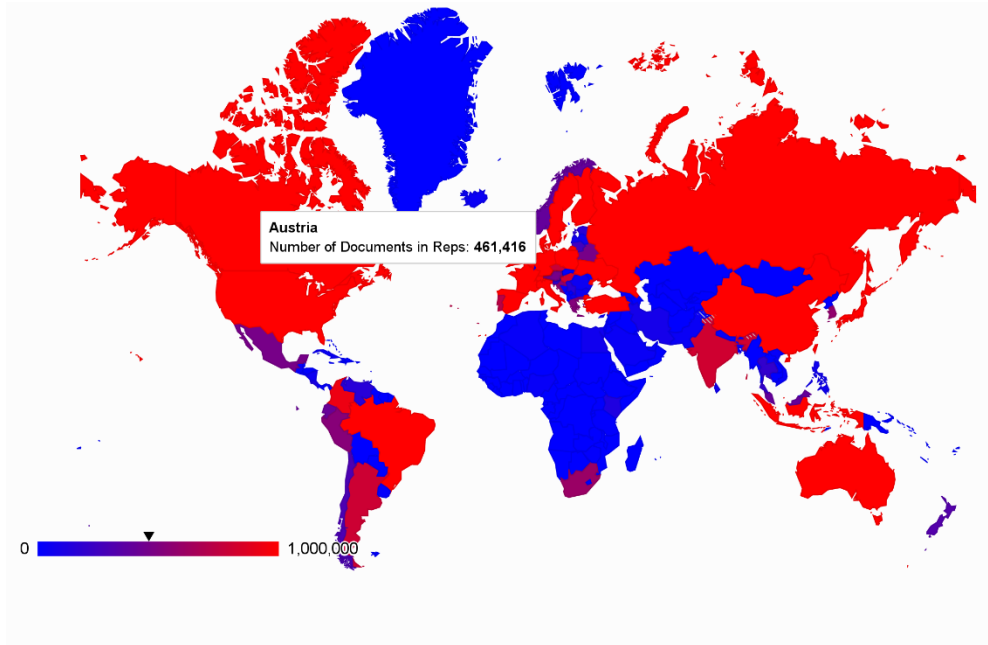


Abb. 3: Heatmap Anzahl Dokumente in Repositorien pro Land (Stand 1.10.2021)

Die folgende Zeitleistendarstellung zeigt die Anzahl neu indexierter Datenquellen pro Jahr seit 2004 für Europa, wobei man einige klare Tendenzen erkennen kann. Nach der ersten Ausbauphase bis 2008 folgt eine eher gleichmäßige Entwicklung, bis ab 2014 wieder eine Zunahme einsetzt. Das ist erklärbar einerseits durch die Ausweitung von Repositorien-Formen durch Digitale Sammlungen und Forschungsinformationssysteme und andererseits durch die zunehmende Integration von reinen bibliographischen Publikationsnachweisen ohne Volltextbereitstellung, wie es bei Publikationssystemen üblich ist.

Repository Development

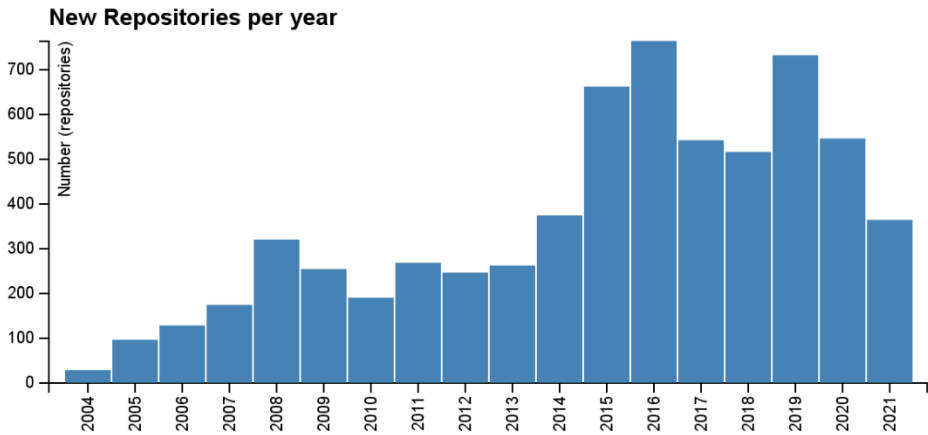


Abb. 4: Neu indexierte Datenquellen in BASE pro Jahr in Europa 2004 – 2021

Für Österreich zeigt sich im Vergleich eine zeitlich etwas andere Verteilung mit einigen Lücken in den betrachteten Jahren und einer zeitlichen Verschiebung der Anstiege.

Repository Development

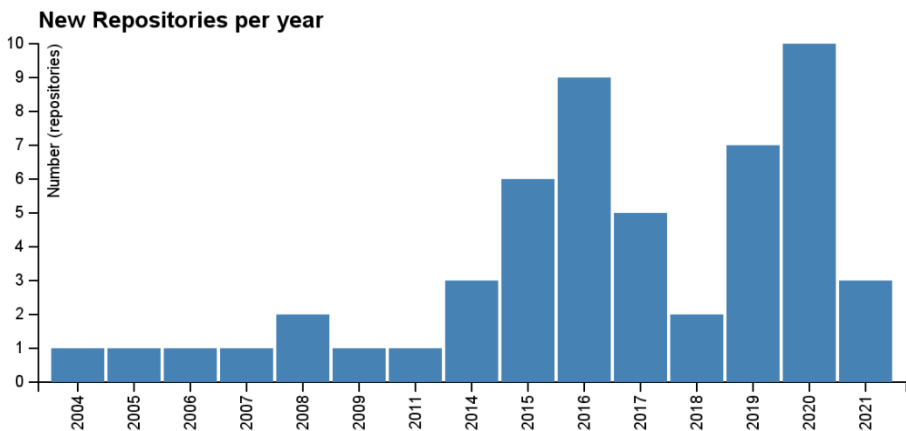


Abb. 5: Neu indexierte Datenquellen aus Österreich in BASE pro Jahr 2004 – 2021

Detaillierter in Bezug auf Anzahl der Repositorien und Anzahl der Dokumente zeigt die folgende Graphik die Entwicklung in Österreich für die jeweils kumulierte Anzahl der Repositorien und der darüber angebotenen Dokumente.

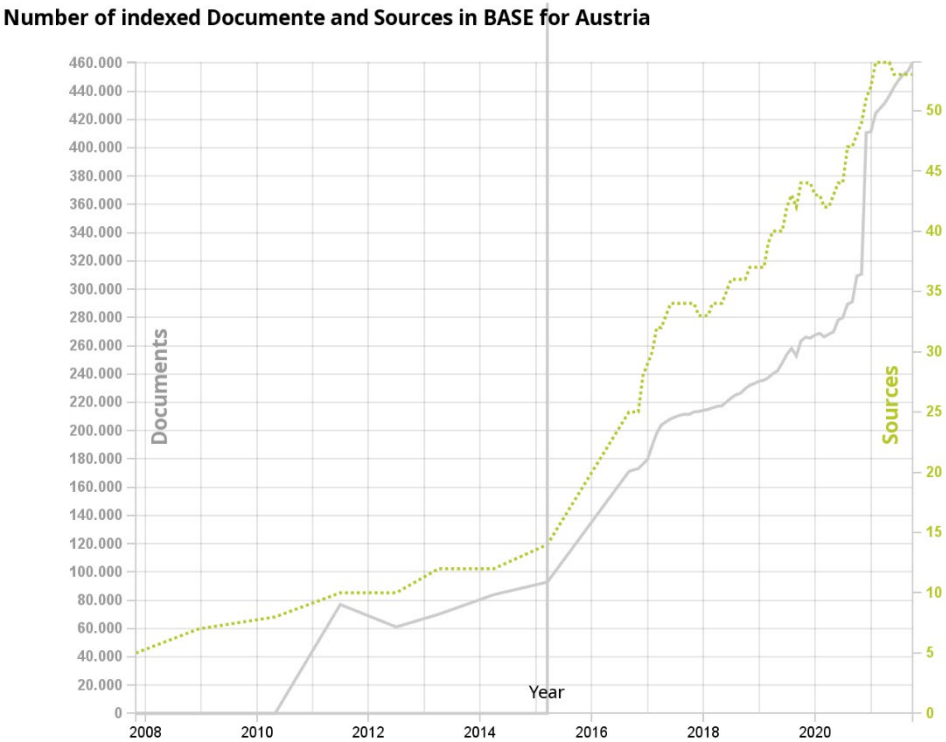
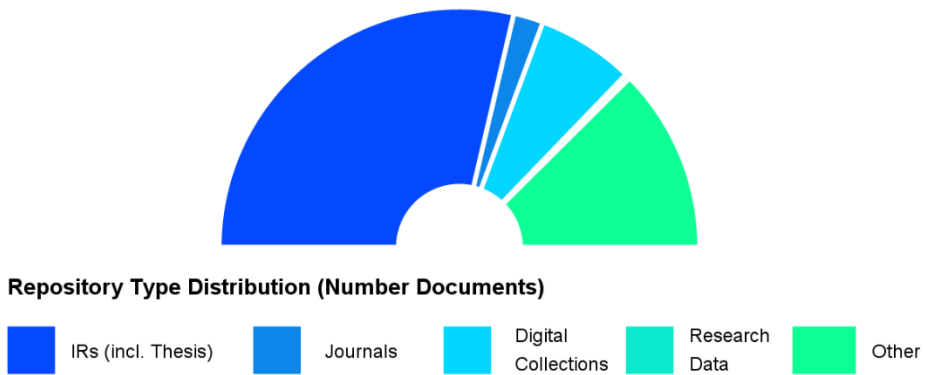


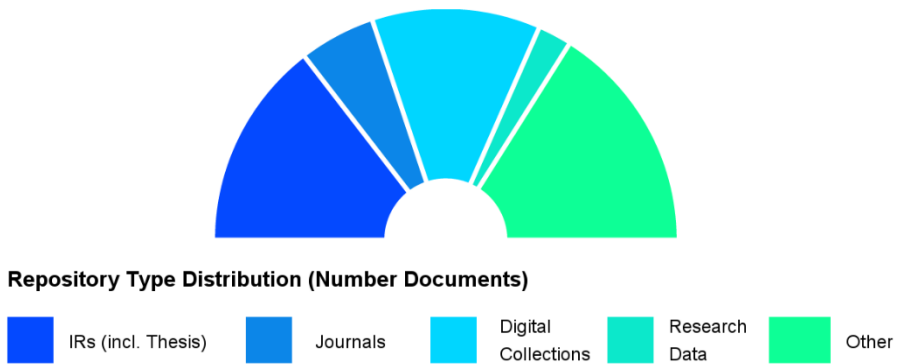
Abb. 6: Entwicklung Anzahl Datenquellen und Anzahl Dokumente im BASE Index aus Österreich

Eine grobe Analyse nach Repositorientyp für Österreich (Stand Oktober 2021) zeigt Abbildung 7. Offenbar liegt der Schwerpunkt auf institutionellen Repositorien, während Zeitschriften- und Forschungsdatenplattformen eine untergeordnete Rolle spielen.



**Abb. 7: Verteilung Anzahl Dokumente nach Repositorientyp in Österreich
(Stand 01.10.2021)**

In Deutschland sieht die Verteilung gleichmäßiger aus. Neben einem höheren Anteil von Zeitschriftenservern fällt auch die vergleichsweise hohe Anzahl von Forschungsdatenservern auf.



**Abb. 8: Verteilung Anzahl Dokumente nach Repositorientyp in Deutschland
(Stand 01.10.2021)**

BASE versteht sich als Suchmaschine im Kontext der Open-Access-Bewegung. Die nächste Karte zeigt eine Weltkarte mit dem Prozentanteil von Open-Access-Publikationen in BASE pro Land. Die Angaben beruhen auf den normalisierten Angaben in `<dc:rights>` zum Open-Access-Status und zeigen deutlich den Einfluss von Open-Access-Strategien etwa in Südamerika, Portugal und Spanien.

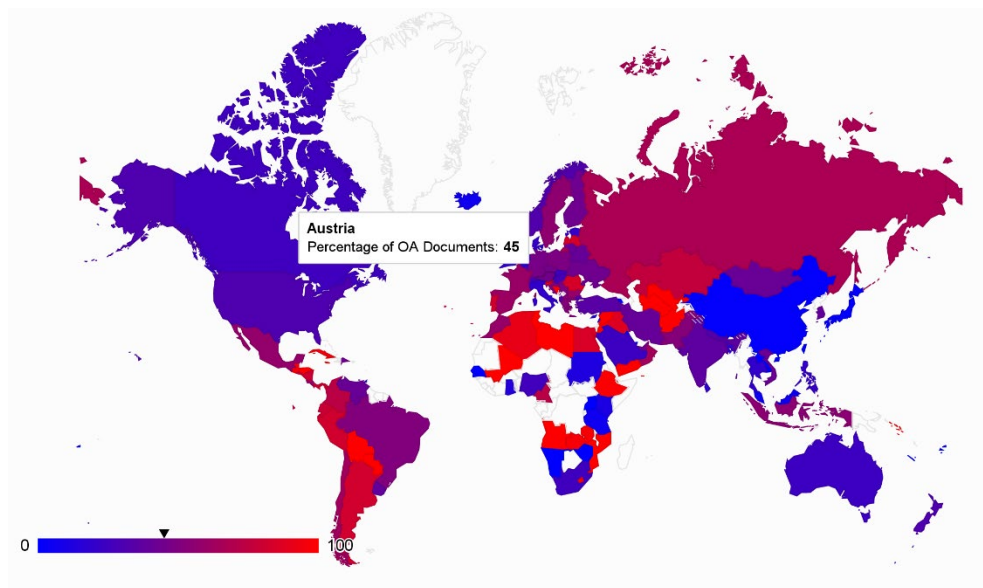


Abb. 9: Prozentualer OA-Anteil an Gesamtzahl Publikationen in BASE pro Land

In der nächsten Abbildung lässt sich der Unterschied zwischen der Schweiz und Österreich bei der Verteilung der Dokumente nach Publikationssprache erkennen. In Österreich liegen die Dokumente zu mehr als zwei Dritteln in deutscher Sprache vor, während in der Schweiz der Anteil für Publikationen in englischer Sprache mit 78 % deutlich höher ist. Die Statistiken basieren auf dem normalisierten `<dc:language>`-Feld.

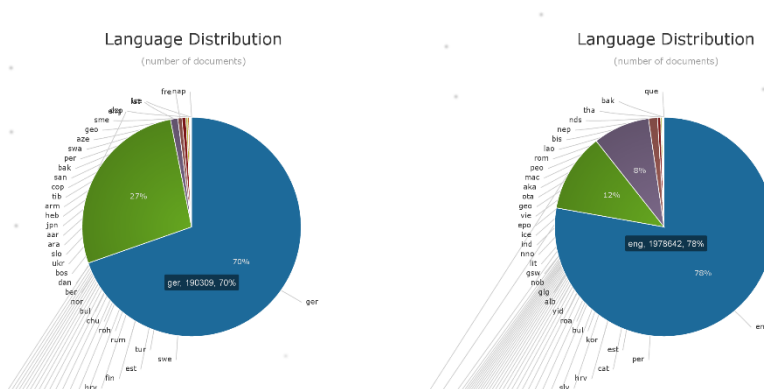
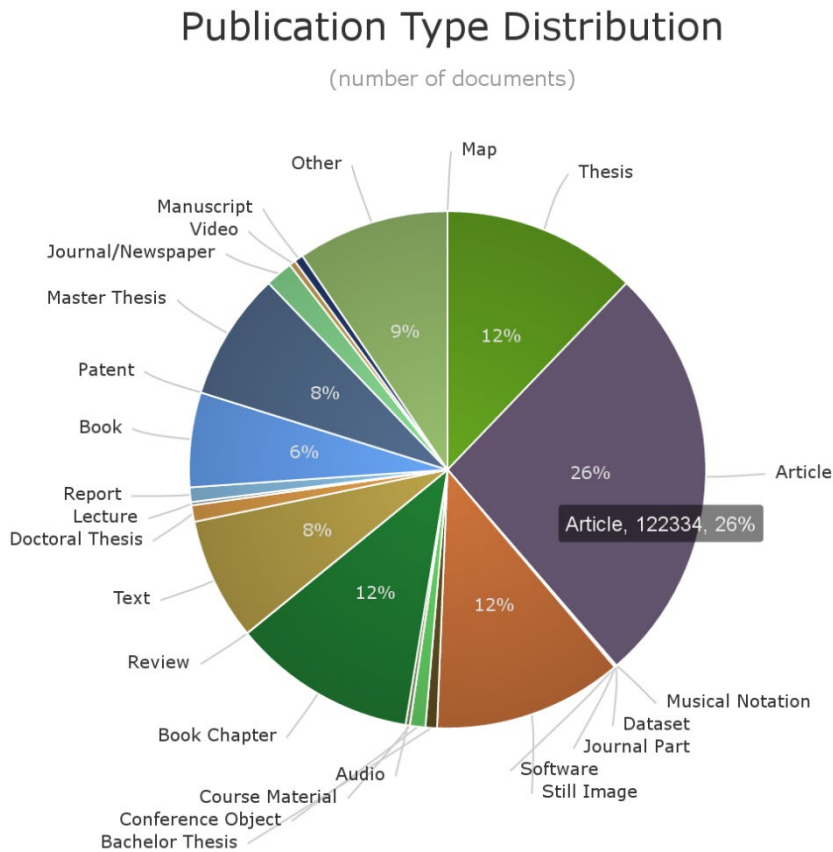


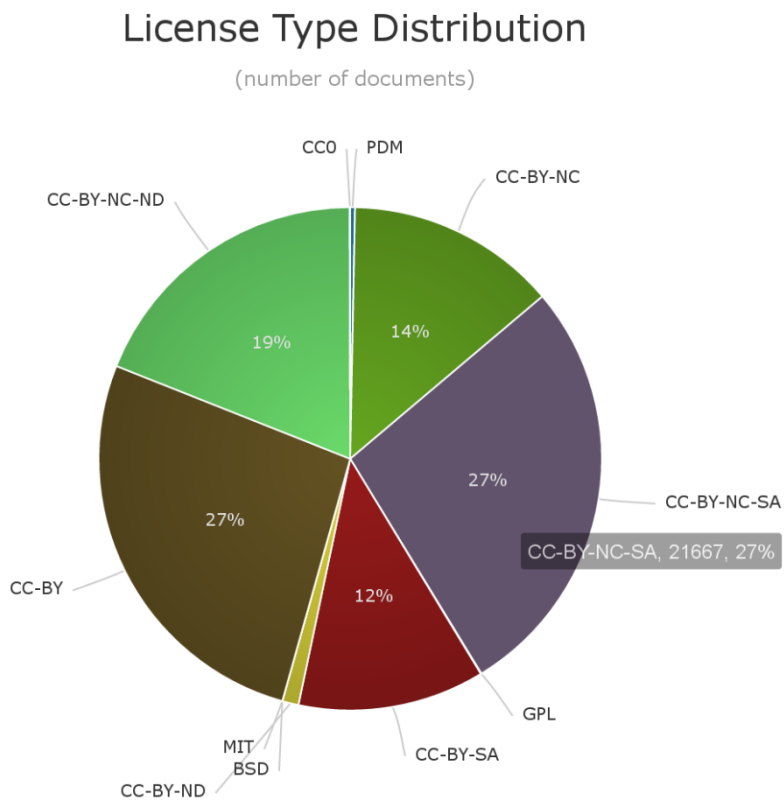
Abb. 10: Verteilung der Dokumente nach Sprache der Publikation (links Österreich, rechts Schweiz, Stand 01.10.2021)

Das nächste Diagramm zeigt die Verteilung der Publikationstypen in den Dokumenten der österreichischen Repositorien. Das dabei verwendete Vokabular ist in BASE aus einer statistischen Auswertung der vorkommenden Terme und ihrer Häufigkeit abgeleitet worden und findet sich auch in der erweiterten Suchmaske als Auswahl für den Suchaspekt „Dokumentart“ wieder.



**Abb. 11: Verteilung Dokumente nach Publikationstyp für Österreich
(Stand 01.10.2021)**

Zur Information über Nachnutzungsmöglichkeiten der Dokumente sind die Lizenzbedingungen relevant, die üblicherweise in <dc:rights> ausgegeben werden. Bei diesen Informationen ist der Umfang der verwandten Terme vergleichsweise stark normiert und daher relativ umfassend umzusetzen. Die nächste Graphik zeigt die Verteilung der verschiedenen Lizenzbedingungen und den Anteil solcher Metadatensätze an der Gesamtheit der Publikationen für Österreich im Oktober 2021.



Percentage Records with License Information = 17.14 %



**Abb. 12: Verteilung Lizenzbedingungen in österreichischen Repositorien in BASE
(Stand 01.10.2021)**

5. Fazit

Repositorien haben sich zu einer festen Größe in der globalen Publikations- und Informationsinfrastruktur entwickelt. Über die Jahre hat es immer wieder Entwicklungen und Veränderungen gegeben, auf die sich sowohl die Betreiber:innen als auch die Endnutzer:innen und Datenaggregator:innen einstellen mussten. Oftmals sind diese Entwicklungen regional unterschiedlich ausgefallen und vor allem setzen sie sich global mit zeitlichem Verzug in den Regionen durch. Deutlich lässt sich erkennen, wie bestimmte Veränderungen in den Repositorien-Schwerpunkten (insbesondere Nordamerika und Europa) ihren Anfang nehmen und sich dann mit beträchtlichem Zeitverzug in den regionalen Communities verbreiten. Das führt auch weiterhin dazu, dass die Zahl der Repositorien weltweit zunimmt, auch wenn es Verschiebungstendenzen, z. B. in Richtung Crossref, gibt. Für diese quantitative Zunahme waren und sind insbesondere zunehmend Nachweise und damit erweiterte Materialien aus Forschungsinformationssystemen und Plattformen mit Digitalisaten (in digitalen Sammlungen), Forschungsdaten und Open Educational Resources (OER) verantwortlich.

Aus diesen Tendenzen lässt sich ableiten, dass auch in der Zukunft die Kernmerkmale Sichtbarkeit, Metadatenqualität und Umfang der Nachweise und Publikationen grundlegende Parameter für die Bewertung von Repositorien bleiben. Daher lohnen sich auch aktuell und weiterhin Maßnahmen zur Verbesserung in diesen Bereichen. In diesen Punkten treffen sich die Interessen von Datenanbietern und Datenaggregatoren wie BASE. Für letztere kommt als weiteres Anforderungskriterium die Stabilität der Schnittstelle und die Transparenz der Systemeigenschaften hinzu, was sich insbesondere bei der Kommunikation von Änderungen auswirkt. Nach wie vor spielt im Hinblick auf die Schnittstellen zum Metadatenaustausch das OAI-PMH-Protokoll eine bedeutende Rolle und entsprechend lohnt sich ein genauer Blick auf die detaillierten Gegebenheiten der Nutzung. Letztlich transportieren die Schnittstellen das im System aufgebaute Qualitätsniveau und damit auch die Sichtbarkeit.

Bibliografie

- Müller, Uwe; Scholze, Frank; Vierkant, Paul et al. (2019): DINI-Zertifikat für Open-Access-Publikationsdienste. (DINI Schriften 3). Humboldt-Universität zu Berlin.
- Pieper, Dirk; Summann, Friedrich (2015): 10 Years of "Bielefeld Academic Search Engine" (BASE). Looking at the Past and Future of the World Wide Repository Landscape from a Service Providers Perspective. OR2015, 10th International Conference on Open Repositories, Indianapolis, Indiana. https://pub.uni-bielefeld.de/download/2766308/2766316/or2015_base_unibi.pdf (abgerufen am 13.10.2021)

- Summann, Friedrich; Czerniak, Andreas; Schirrwagen, Jochen; Pieper, Dirk (2020): Data Science Tools for Monitoring the Global Repository Eco-System and its Lines of Evolution. In: Publications 8 (2), p. 35. <https://doi.org/10.3390/publications8020035>
- Tennant, Roy (2004): Digital Libraries. Metadata's Bitter Harvest. In: Library Journal 129 (12), p. 32.
- Vanderfeesten, Maurice; Summann, Friedrich; Slabbertje, Martin (Hg.) (2008): DRIVER Guidelines 2.0. Guidelines for Content Providers – Exposing Textual Resources With OAI-PMH. https://pub.uni-bielefeld.de/download/2491610/2491613/DRIVER_Guidelines_v2_Final_2008-11-13.pdf (abgerufen am 05.01.2022)

Friedrich Summann war bis 2022 Leiter der LibTec-Abteilung (Bibliothekstechnologie und Wissensmanagement) der Universitätsbibliothek Bielefeld. Er arbeitet aktiv in zahlreichen Projekten im Kontext digitaler Informationsversorgung (u. a. BASE, ORCID-DE, Digital Humanities, GO:AL).