Tobias Beinert

Webarchivierung an der Bayerischen Staatsbibliothek

Web archiving at the Bayerische Staatsbibliothek

DOI 10.1515/bd-2017-0052

Zusammenfassung: Die Bayerische Staatsbibliothek sammelt und archiviert seit dem Jahr 2010 landeskundlich und wissenschaftlich relevante Websites. Der Artikel gibt einen Überblick über das Sammel- und Archivierungsprofil der Bayerischen Staatsbibliothek für Websites, die rechtlichen Grundlagen, den entwickelten Workflow sowie die Verzeichnung und Zugänglichmachung von archivierten Websites. Abschließend werden weitere, zukünftige Perspektiven dargestellt.

Schlüsselwörter: Webarchivierung, Langzeitarchivierung, Bayern

Abstract: The Bayerische Staatsbibliothek has been collecting and archiving websites dealing with regional studies and science since the year 2010. The article provides a survey of the collection and archiving profiles of the Bayerische Staatsbibliothek concerning websites, the legal basis, the workflow which has been developed as well as the registration and making available of websites in the archives. Finally, further perspectives for the future are presented.

Keywords: web archiving, long-term archiving, Bavaria

1 Entwicklung der Webarchivierung der Bayerischen Staatsbibliothek

Wie andere National-, Landes- und Regionalbibliotheken sowie Forschungseinrichtungen stand auch die Bayerische Staatsbibliothek spätestens seit Anfang der 2000er Jahre vor der Frage, wie digitale Veröffentlichungen gesammelt, verzeichnet, archiviert und bereitgestellt werden sollen, um diese zum Einen für die Wis-

Tobias Beinert: beinert@bsb-muenchen.de

senschaft von morgen zugänglich zu machen und zum Anderen eine zeitgemäße landeskundliche Überlieferung im 21. Jahrhundert zu ermöglichen.
Seit 2005 wird daher für die Langzeitarchivierung der mittlerweile mehr als 1,2 Millionen Retrodigitalisate, die im Münchener Digitalisierungszentrum (MDZ) oder von zahlreichen Kooperationspartnern und Dienstleistern produziert wurden sowie alle Formen von Netzpublikationen das Bibliothekarische Archivierungs- und Bereitstellungssystem (BABS) in Zusammenarbeit mit dem Leibniz-Rechenzentrum betrieben.[1] Das Langzeitarchivierungssystem Rosetta der Firma ExLibris wurde 2012 als zentrale Komponente in diese technisch-organisatorische Infrastruktur integriert.

Als Lösung für die Sammlung und Archivierung von Websites wurde nach einer eingehenden Evaluierung ab 2010 das gemeinsam von der British Library und der Nationalbibliothek von Neuseeland entwickelte und als Open Source nachnutzbare Web Curator Tool (WCT) durch das Münchener Digitalisierungszentrum an der Bayerischen Staatsbibliothek implementiert und in den Produktivbetrieb übernommen.[2] Das WCT verfügt über Module für die Genehmigungseinholung, die Erstellung von Archivkopien (Harvesting/Crawling), eine teilautomatisierte Qualitätskontrolle und die Archivierung und deckt damit den Workflow der selektiven Webarchivierung komplett ab. Die Langzeitarchivierung der mit dem WCT erstellten Archivcontainer im WARC (Web ARChive)-Dateiformat mit den dazugehörigen Metadaten erfolgt mit Rosetta; als Viewer ist die OpenWayback-Software zur Navigation, Auswahl und Anzeige der einzelnen Zeitschnitte im Einsatz.

2 Sammel- und Archivierungsprofil

Das Sammel- und Archivierungsprofil der Bayerischen Staatsbibliothek für Websites, wie auch für andere digitale Ressourcen, beruht erstens auf ihrer Rolle als zentrale Landes- und Archivbibliothek des Freistaats Bayern und zweitens auf ihrer Funktion als Forschungsbibliothek mit überregionalen Aufgaben in der

1 Vgl. Brantl, Markus/Ceynowa, Klaus et al: Digitale Langzeitarchivierung in Bayern. Vom explorativen Projekt zum nachhaltigen Modell, in: BIBLIOTHEK Forschung und Praxis. Band 35 (2011), Heft 1, Seiten 15–25, http://dx.doi.org/10.1515/bfup.2011.003.

2 Eine ausführliche Darstellung findet sich in: Beinert, Tobias/Hagenah, Ulrich/Kugler, Anna: Es war einmal eine Website … – Kooperative Webarchivierung in der Praxis, in: o-bib, Band 1 (2014), Nr. 1, S. 291–304, http://dx.doi.org/10.5282/o-bib/2014H1S291-304.

wissenschaftlichen Informationsversorgung.[3] In der Praxis ergibt sich daraus, dass die Websites von Ministerien, Behörden und Dienststellen des Freistaats als amtliche Veröffentlichungen ebenso zu sammeln sind wie Angebote, die die Geschichte, Gesellschaft und Kultur Bayerns zum Gegenstand haben (Bavarica). Die zweite Säule des Profils bilden Websites aus dem Bereich der traditionellen wissenschaftlichen Sammelschwerpunkte der Bayerischen Staatsbibliothek: Altertumswissenschaften, Geschichte, Musik und Osteuropa.
Während für die amtlichen Websites größtmögliche Vollständigkeit angestrebt wird, werden wissenschaftlich relevante Angebote und Bavarica derzeit selektiv gesammelt und archiviert. Fachlich einschlägig bewanderte Experten der Bayerischen Staatsbibliothek wählen anhand der sachlichen, regionalen und zeitlichen Kriterien des Sammel- und Archivierungsprofils für Websites[4] die zu archivierenden Websites aus dem In- und Ausland aus. Das dynamisch angelegte Profil definiert zudem, welche Arten von Webangeboten aus technischen und/oder rechtlichen Gründen von einer Archivierung ausgeschlossen sind.

Das in Landes- und Archivbibliotheken sowie auch in den ehemaligen Sondersammelgebietsbibliotheken für gedruckte Publikationen ehemals angestrebte Ziel einer größtmöglichen Vollständigkeit der Sammlungen ist für digitale Veröffentlichungen ganz generell, aber insbesondere im Kontext der Webarchivierung neu zu bewerten bzw. neu zu definieren. Neben der schieren Menge an Webangeboten, die bereits eine inhaltlich basierte Auswahlentscheidung, also ein selektives Vorgehen erforderlich machen, ergeben sich auch aus dem unvollständigen Rücklauf im Zuge des Genehmigungsverfahrens sowie den Einschränkungen, die aus der derzeit verfügbaren Technik resultieren, letztlich Lücken im Bestandsaufbau. Als neue Zielsetzung für das Sammeln und Archivieren von Netzpublikationen soll daher in Anknüpfung an die Überlegungen von Altenhöner und Schrimpf eine „repräsentative Vollständigkeit“ angestrebt werden, die über eine stärkere Einbindung von Fachcommunities sowie einer Verbesserung der gesellschaftlichen Rückkopplung (z. B. durch eine Beteiligung der Nutzerinnen und Nutzer am Bestandsaufbau) im Auswahlprozess erreicht werden kann.[5]

---

3 Eine Übersicht über den gesamten Bereich der Netzpublikationen in Bayern findet sich in: Balz, Nina/Schoger, Astrid: Bayern, in: Bibliotheksdienst. Band 47 (2013), Seiten 605–608, https://doi.org/10.1515/bd-2013-0065.

4 Vgl. Sammel- und Archivierungsprofil der Bayerischen Staatsbibliothek für Websites, verfügbar unter: https://www.babs-muenchen.de/content/DFG-Projekt_Webarchivierung/Sammel_und_Archivierungsprofil_Websites_BSB.pdf [Zugriff: 10.02.2017].

5 Vgl. Altenhöner, Reinhard, Schrimpf, Sabine: Lost in tradition? Systematische und technische Aspekte der Erwerbung von Internetpublikationen in Archivbibliotheken, in: Schüller-Zwierlein, André (Hrsg.): Diachrone Zugänglichkeit als Prozess. Kulturelle Überlieferung in systematischer

Grundsätzlich erweist sich die Webarchivierung – nicht nur für die Bayerische Staatsbibliothek – als ein wichtiger Baustein in der digitalen Bestandsentwicklung, um der hohen Flüchtigkeit von frei zugänglichen und wissenschaftlich relevanten Informationsressourcen adäquat zu begegnen und sie dauerhaft für die Wissenschaft und andere interessierte Nutzer zugänglich zu machen. Die Webarchivierung ist mittlerweile integraler Teil des digitalen Bestandsaufbaus des Hauses, der abteilungsübergreifend durch die Etablierung entsprechender Workflows sowie die Implementierung und den Betrieb neuer technischer Infrastrukturen in die Praxis umgesetzt werden konnte.

3 Rechtliche Situation in Bayern

Den rechtlichen Rahmen für die Archivierung von Websites durch die Bayerische Staatsbibliothek bilden derzeit zwei Bestimmungen: Während mit dem Erlass der Bayerischen Staatsregierung zur Abgabe amtlicher Veröffentlichungen bereits seit 2008 eine Regelung zu elektronischen Publikationen von staatlichen Stellen[6] existiert, ist die Novellierung des Pflichtstückegesetzes in Bayern im Hinblick auf die Sammlung, Archivierung und Zugänglichmachung von digitalen Medienwerken weiter ausstehend. Daraus ergibt sich, dass bislang für alle Websites, die nicht von Einrichtungen des Freistaats herausgegeben werden, ein aufwändiges Verfahren zur Einholung von Genehmigungen durch den Rechteinhaber durchgeführt werden muss. Die positiven Rücklaufquoten liegen hier durchschnittlich lediglich bei 25–30%; im Fall der Bavarica bei erfreulichen fast 50%. Hierbei wird der Bayerischen Staatsbibliothek neben den für die eigentliche Archivierung notwendigen Rechten für die Vervielfältigung und Bearbeitung auch das Recht zur öffentlichen Zugänglichmachung eingeräumt, so dass die Archivobjekte über die Katalog- bzw. Nachweissysteme der Bayerischen Staatsbibliothek erschlossen bzw. verzeichnet werden und über den Viewer OpenWayback frei im Web bereitgestellt werden können.

Eine Neufassung der Pflichtablieferung sollte für den Bereich der landeskundlichen Netzpublikationen explizit die im Kontext digitaler Medien urheberrechtlich relevanten Aspekte berücksichtigen. In einigen Bundesländern verfügen die in

Sicht. Berlin 2014, S. 297–328 und Beinert, Tobias/Schoger, Astrid: Vernachlässigte Pflicht oder Sammlung aus Leidenschaft? Zum Stand der Webarchivierung in deutschen Bibliotheken, in: Zeitschrift für Bibliothekswesen und Bibliographie, Band 62(2015), Heft 3/4, S. 172–183, S. 175 ff.

6 Vgl. Bekanntmachung der Bayerischen Staatsregierung vom 2. Dezember 2008 (Abgabe Bibliotheken – Abg-Bibl): Az.: B II 2-480-30, 22-40-WFK, AllMBL 16 (2008), S. 818–819.

den letzten Jahren novellierten Gesetze bereits über entsprechende Regelungen, möglicherweise sorgt zudem die sich derzeit in der Diskussion befindliche Erweiterung des Gesetzes über die Deutsche Nationalbibliothek für eine entsprechende rechtliche Klarstellung.[7] Nur wenn die Pflichtablieferung für digitale Medien auf eine solide rechtliche Basis gestellt wird, kann die Bayerische Staatsbibliothek ihren Auftrag als zentrale Landes- und Archivbibliothek für den Freistaat Bayern auch zukünftig vollumfänglich erfüllen.

Allerdings wären auch mit einer Anpassung der Pflichtgesetze auf Landes- bzw. Bundesebene noch nicht alle rechtlichen Fragen der Webarchivierung abschließend geklärt. Dies gilt insbesondere für Fragen des Persönlichkeitsrechts und des Datenschutzes.[8]

4 Workflow Webarchivierung

Die Auswahl der zu archivierenden Websites wird vor dem Start des Archivierungsprozesses in den Fachabteilungen getroffen. Nachdem für den Bereich der freien Netzpublikationen die klassischen bibliographisch-bibliothekarischen Nachweismittel (bisher) fehlen, muss das Auffinden bzw. die Identifizierung von Websites mit wissenschaftlicher oder landeskundlicher Relevanz für die Webarchivierung in der Regel durch ein gezieltes fachspezifisches Monitoring und Recherchen im Internet erfolgen. Die Mitarbeiterinnen und Mitarbeiter der Fachabteilungen nehmen die Verzeichnung und Erschließung von ausgewählten Websites in der institutionsübergreifenden Erschließungsdatenbank Academic Linkshare vor, anschließend wird auch im Web Curator Tool ein entsprechender Datensatz angelegt und eine Genehmigungsanfrage per Mail verschickt. Stimmt der Rechteinhaber der Langzeitarchivierung durch die Bayerische Staatsbibliothek zu, wird die Software zur Erstellung der Archivkopien und damit der Prozess des so genannten Crawling/Harvesting gestartet. Dieser Schritt wird halbjähr-

7 Vgl. Referentenentwurf des Bundesministeriums der Justiz und für Verbraucherschutz: Entwurf eines Gesetzes zur Angleichung an die aktuelle Erfordernisse der Wissensgesellschaft (Urheberrechts-Wissengesellschafts-Gesetz – UrhWissG), verfügbar unter: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RefE_UrhWissG.pdf?__blob=publicationFile&v=1 [Zugriff: 10.02.2017].

8 Vgl. Steinhauer, Eric W: Wissen ohne Zukunft, Der Rechtsrahmen der digitalen Langzeitarchivierung von Netzpublikationen, verfügbar unter: https://www.deutsche-digitale-bibliothek.de/content/ueber-uns/aktuelles/wissen-ohne-zukunft-der-rechtsrahmen-der-digitalen-langzeitarchivierung-von-netzpublikationen-ein-beitrag-von-eric-w-steinhauer [Zugriff: 10.02.2017].

lich automatisiert wiederholt. Die Crawler-Software verfolgt die Links innerhalb einer Website und lädt die dahinter liegenden Dateien herunter. Da in diesem Prozess aus technischen Gründen teilweise Dateien bzw. Inhalte nicht kopiert werden können, aber auch teilweise zu viel bzw. unerwünschte Daten eingesammelt werden, wird eine detaillierte intellektuelle Qualitätskontrolle durchgeführt. Dabei werden sowohl der erste als auch die fortlaufend eingehenden Zeitschnitte aus technischer und inhaltlicher Sicht überprüft. Um diesen Arbeitsschritt möglichst effizient zu gestalten, erfolgt er für alle Websites an zentraler Stelle im Münchener Digitalisierungszentrum durch entsprechend spezialisierte Mitarbeiterinnen. Er umfasst zunächst die Auswertung bestimmter Logdateien, bevor eine visuelle Kontrolle des gesamten Zeitschnitts vorgenommen wird. Die Qualität eines Zeitschnitts wird dabei anhand folgender Merkmale bewertet: Vollständigkeit der zentralen Inhalte, Konsistenz, Erhalt der Funktionalität, Erhalt des Look and Feel.[9] Erfüllen die Archivkopien dabei bestimmte Mindestanforderungen nicht, ist entweder eine Bearbeitung des Zeitschnitts oder die Wiederholung des Crawls mit geänderten Parametern notwendig, nur in seltenen Fällen muss von einer Archivierung ganz abgesehen werden. Nach erfolgreicher Qualitätskontrolle werden die einzelnen Zeitschnitte und entsprechende Metadaten in das Langzeitarchiv Rosetta übernommen, in diesem Zuge erfolgen auch eine technische Überprüfung der WARC-Dateien sowie Viruschecks und die Erstellung von Checksummen. Zudem wird von Rosetta für jeden Zeitschnitt ein Uniform Ressource Name (URNs) als persistenter Identifikator vergeben.

9 Eine ausführliche Darstellung findet sich in: Qualität und Prozessoptimierung bei der Langzeitarchivierung von Websites: Konzeptuelle Überlegungen zur Steuerung des Ressourceneinsatzes bei der selektiven Webarchivierung, verfügbar unter: https://www.babs-muenchen.de/content/DFG-Projekt_Webarchivierung/Webarchivierung_Qualitaet_und_Prozessoptimierung.pdf [Zugriff: 10.02.2017].

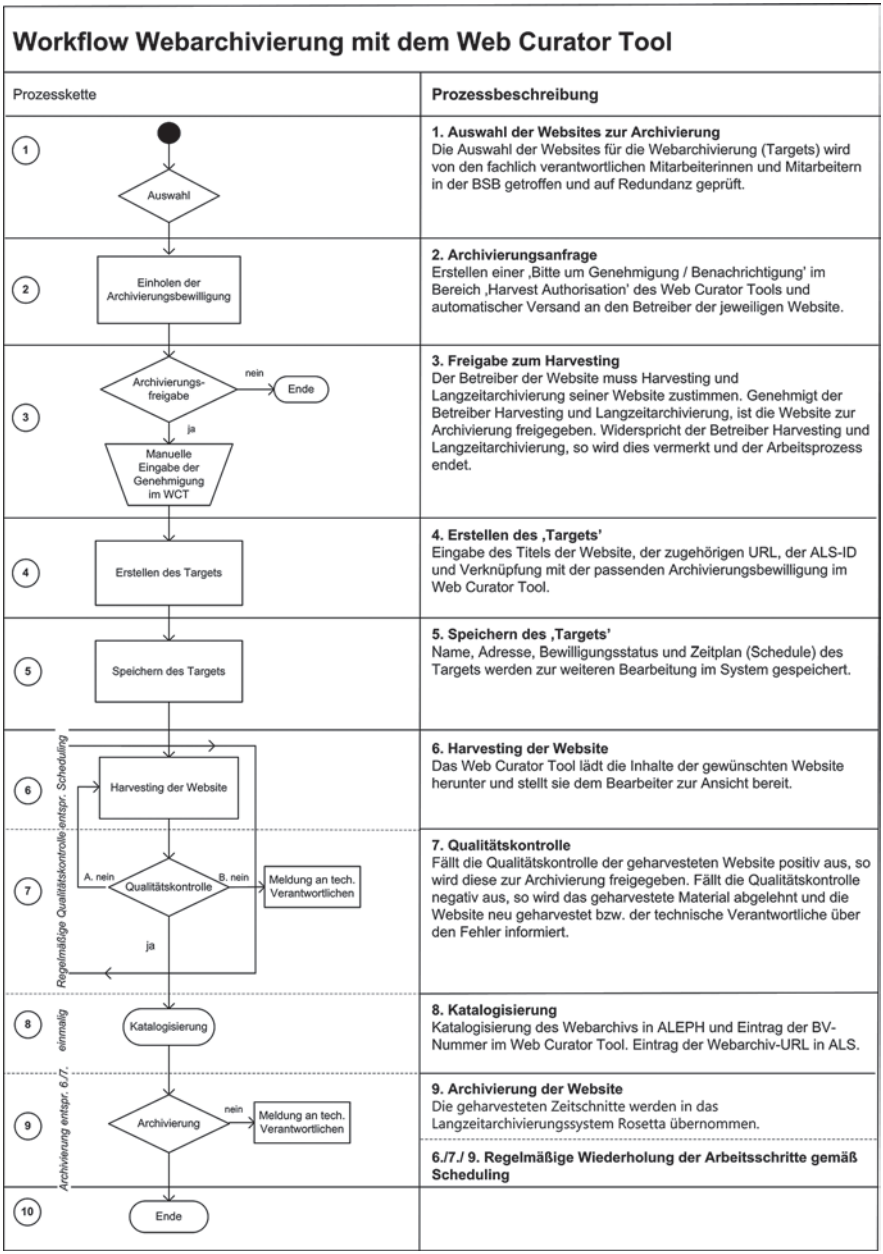

Abb. 1: Workflow Webarchivierung an der Bayerischen Staatsbibliothek.

Mittlerweile werden ca. 1.500 Websites regelmäßig zwei Mal im Jahr archiviert, bislang liegen ca. 8.000 Zeitschnitte vor. Davon entfallen 1.130 Zeitschnitte von 168 Websites auf die Ministerien und Behörden Bayerns, der Rest auf einschlägige als fachwissenschaftlich relevant eingestufte, thematische Websites (Stand: Januar 2017). Der zeitliche Aufwand für die Webarchivierung konnte im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts Langzeitarchivierung von Websites durch Gedächtnisinstitutionen im Rahmen einer Machbarkeitsstudie exemplarisch für das Vorgehen der BSB dokumentiert werden. Als Richtwert für den kompletten Prozess der erstmaligen Archivierung einer Website inklusive der Genehmigungseinholung und Qualitätskontrolle, können im Normalfall 20–35 Minuten veranschlagt werden, die fortlaufende Überprüfung der weiteren Zeitschnitte schlägt dann mit drei bis zehn Minuten zu Buche.[10]

5 Verzeichnung und öffentliche Zugänglichmachung

Alle archivierten Websites sind mit einer Titelaufnahme im Katalog des Bibliotheksverbunds Bayern nachgewiesen und werden über den OPACplus-Katalog der Bayerischen Staatsbibliothek für die Nutzer weltweit zugänglich gemacht. Nach dem Aufrufen des Katalogdatensatzes erscheint eine Übersichtsdarstellung aller archivierten Zeitschnitte im Viewer OpenWayback, was eine gezielte Navigation zum gewünschten Datum ermöglicht. Die Zeitschnitte sind durch ein entsprechendes Banner als Archivversion gekennzeichnet. Über das Banner kann auch ein Zitierhinweis mit einen Uniform Ressource Name (URN) für jeden Zeitschnitt aufgerufen werden, sodass eine dauerhafte Referenzierbarkeit gewährleistet ist. Die fachwissenschaftlichen Websites sind darüber hinaus auch in den Internetressourcenguides der Virtuellen Fachbibliotheken bzw. Fachinformationsdienste verzeichnet und von dort sowohl als Archivversion als auch via Link zum Originalangebot zugänglich. Websites mit Bezug zu Bayern werden auch in der Bayerischen Bibliographie nachgewiesen und sind über das landeskundliche Portal „Bayerische Landesbibliothek Online (BLO)“ zugänglich.

10 Vgl. Erfahrungsbericht: Retrospektive Langzeitarchivierung von in Academic Linkshare erschlossenen Internetressourcen, verfügbar unter: https://www.babs-muenchen.de/content/DFG-Projekt_Webarchivierung/Erfahrungsbericht_LZA_von_Internetressourcen.pdf [Zugriff: 10.02.2017].

Abb. 2: Zeitschnitt einer archivierten Website mit Zitierhinweis.

6 Zusammenarbeit mit der SUB Hamburg

Seit 2013 besteht eine Kooperation mit der Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky (SUB). Im Rahmen einer Projektförderung durch die Deutsche Forschungsgemeinschaft wurde das an der BSB bereits im Betrieb befindliche System für die Sammlung, Archivierung und Bereitstellung von Websites ausgebaut, um prototypisch die Nachnutzung durch eine andere Gedächtnis- und Forschungseinrichtung zu ermöglichen.[11] Nach einer Testphase nutzt die SUB seit 2015 die an der BSB betriebene technische Infrastruktur für ihre Webarchivierung von landeskundlichen Publikationen.[12]

11 Eine Darstellung zum Projekt Langzeitarchivierung von Websites durch Gedächtnisinstitutionen: Entwicklung eines Servicemodells auf Grundlage praktischer Erfahrungen findet sich unter: https://www.babs-muenchen.de/index.html?c=projekte_webarchivierung&l=. [Zugriff: 10.02.2017].

12 Eine ausführliche Darstellung findet sich im Beitrag von Ulrich Hagenah in diesem Heft.

7 Perspektiven

Gerade im Bereich der Langzeitarchivierung von amtlichen Websites ist für die Zukunft zu prüfen, ob sich hier durch die Webarchivierung neue Wege in der Sammelstrategie ergeben. Bis dato werden amtliche Veröffentlichungen in digitaler Form zum Teil als eigenständige Publikation erschlossen, verzeichnet und archiviert, aber in vielen Fällen auch als Teil eines Webangebots regelmäßig in das Archiv übernommen; in einigen Fällen bewahrt die Bayerische Staatsbibliothek auch noch eine parallel erscheinende gedruckte Version auf. Sofern die Vollständigkeit der jeweiligen Publikationen sowie eine entsprechende Auffindbarkeit innerhalb der archivierten Zeitschnitte sichergestellt ist, sollte hier perspektivisch auf ein redundantes Vorgehen verzichtet werden.
Durch eine verstärkte Abstimmung und Kooperation mit anderen Institutionen könnte der Bestand an archivierten Websites deutschlandweit gezielt ausgebaut werden. Während auf nationaler Ebene hier vor allem die Zusammenarbeit mit der Deutschen Nationalbibliothek eine wichtige Rolle spielt, könnte auf Landesebene der Ausbau der Aktivitäten durch eine stärkere Einbindung der wissenschaftlichen und regionalen Bibliotheken und staatlichen Archive erreicht werden.

Für den Bereich der wissenschaftlich relevanten Websites bieten schließlich die Fachinformationsdienste (FIDs) einen Anknüpfungspunkt. Nach der Beendigung der Förderung der Erschließung von Internetressourcen im Rahmen der Sondersammelgebiete bzw. der FIDs durch die Deutsche Forschungsgemeinschaft ist die Verzeichnung und die Archivierung von frei zugänglichen Netzpublikationen und Websites für die Fachwissenschaft nicht mehr sichergestellt: die FID-Richtlinien treffen keine Aussage zum Umgang und Erhalt von frei zugänglichen bzw. im Open Access publizierten Netzpublikationen und Websites. Hier könnte die erfolgreiche Inbetriebnahme bzw. Fortführung von institutionsübergreifenden Serviceangeboten eine Option darstellen, Websites im Kontext des Bestandsaufbaus umfassender zu berücksichtigen, um so den dauerhaften Zugang zu diesen Ressourcen für Forschung und Wissenschaft sicherzustellen.

Tobias Beinert
Bayerische Staatsbibliothek
Digitale Bibliothek und Bavarica
Ludwigstr. 16
80539 München
Deutschland
E-Mail: **beinert@bsb-muenchen.de**