Eine Strategie zur Förderung der digitalen Langzeitarchivierung

Georg Rehm

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH
Alt-Moabit 91c, 10559 Berlin
georg.rehm@dfki.de, http://georg-re.hm

20. Juli 2013

1 Einleitung und Überblick

Das Thema *digitale Langzeitarchivierung* wird in Deutschland zwar seit einigen Jahren diskutiert, eine tatsächliche operative Umsetzung durch die Implementierung entsprechender Infrastrukturen oder das Anbieten von Dienstleistungen steckt allerdings noch in den Kinderschuhen.[1] Über die Gründe kann vortrefflich spekuliert werden – vermutlich liegt ihr Kern in dem Umstand, dass es sich um ein Thema handelt, das inhärent eine Vielzahl verschiedener Gebiete involviert und somit nur interdisziplinär bearbeitet und umgesetzt werden kann. Die digitale Langzeitarchivierung berührt u. a. die *Informations- und Archivwissenschaft* (Umfang, Struktur und Inhalte von Metadatenvokabularen, Prinzipien der Archivierung, Strukturierung, Annotation und Verschlagwortung von Objekten etc.), die *Bibliothekswissenschaft* (bestmögliche Distribuierung von Informationen, Katalogisierungsprinzipien, Informationsversorgung), die *Rechtswissenschaft* (Urheberrechts- und Copyright-Fragen, z. B. im Hinblick auf das Kopieren und Archivieren digitaler Objekte aus dem Web), die *Kulturwissenschaft* (etwa bezüglich der Frage nach der Differenzierung zwischen zu archivierenden und nicht zu archivierenden Objekten; ab welcher Schöpfungshöhe soll ein digitales Objekt für die Nachwelt langfristig gespeichert werden?), die *Geschichtswissenschaft und Gedächtnisinstitutionen* (u. a. als Nutzer und Anbieter von Archiven digitaler Objekte), sogar die *Architektur* (wie sollte ein Rechenzentrum baulich beschaffen sein, um ein digitales Archiv für die nächsten 500 Jahre sicher beherbergen zu können und wo sollte es im besten Fall aufgebaut werden, um Naturkatastrophen möglichst zu vermeiden?) sowie die *Informatik und Informationstechnologie* (u. a. schnelle, redundant ausgelegte Netzwerke, Massenspeicher, Backups, Zugriffssysteme, APIs, Nutzeroberflächen und Anfragesprachen, Import und Export von Daten, Suchalgorithmen, Datenintegrität, Revisionssicherheit etc.). Je nach Typ und Natur der zu archivierenden digitalen Objekte kommen gegebenenfalls noch die domänenspezifischen *Fachwissenschaften* hinzu. Dazu später mehr.

Die verschiedenen Beiträge in diesem Band verdeutlichen, dass die digitale Langzeitarchivierung eine wichtige gesellschaftliche Aufgabe ist. Ohne eine Strategie und die Initiierung konkreter Maßnahmen besteht die Gefahr des unwiederbringlichen Verlustes einer Vielzahl digitaler und analoger Kulturgüter. Der Erarbeitung und speziell auch der Umsetzung einer derartigen Strategie steht jedoch die Realität gegenüber. Diese ist insbesondere gekennzeichnet von einer globalen Wirtschaftskrise, die sich auch auf die Bereitstellung von Fördergeldern

[1] Dieser Beitrag wird erscheinen im Abschlussbericht der 8. Initiative, „Nachhaltigkeit in der Digitalen Welt“, des Internet und Gesellschaft Co:llaboratory, siehe http://www.collaboratory.de/w/Initiative_Nachhaltigkeit_in_der_Digitalen_Welt. Der Autor bedankt sich bei Olga Chiarcos, Martin Ostrowski, Eric Steinhauer und Markus Stumpf für hilfreiche Kommentare und wertvolle Anregungen.

für Themengebiete auswirkt, die von Gesellschaft und Politik zwar als relevant und wichtig wahrgenommen werden, für die aber letztlich keine ausreichenden Mittel zur Implementierung bereitgestellt werden können. Die oben punktuell aufgeführten beteiligten Gebiete und die Bandbreite der zu bearbeitenden Fragestellungen illustrieren, wie umfangreich ein derartiges Mammutvorhaben tatsächlich wäre. Gerade in einer Zeit, in der über Konzepte wie das Recht auf Vergessenwerden[2] und den digitalen Radiergummi diskutiert wird, dürfte es eine Herausforderung sein, genügend Aufmerksamkeit zu bekommen und anschließend Ressourcen in Form von Fördermitteln für die digitale Langzeitarchivierung einzuwerben. Andererseits ist das Thema gerade jetzt mit den Spionageprogrammen „Prism" der National Security Agency (NSA) und „Tempora" des UK Government Communications Headquarters (GCHQ) wieder im Mittelpunkt einer global geführten öffentlichen Debatte angelangt. Dies stellt ein großes Risiko für die Fördersituation dar, kann aber auch mit einem positiven Effekt verbunden sein. Dieser Aspekt wird im letzten Abschnitt wieder aufgegriffen.

Auf Grund der Komplexität der Thematik, der immensen Höhe der zu erwartenden Kosten und des gewaltigen Umfangs der zu archivierenden Daten erscheint es zum derzeitigen Stand der Debatte unwahrscheinlich, dass in absehbarer Zeit eine nationale Strategie zur digitalen Langzeitarchivierung entwickelt und auch umgesetzt werden wird. Der drohende Verlust digitaler Kulturgüter ist selbstverständlich nicht auf die Bundesrepublik beschränkt. Ganz im Gegenteil, die digitale Langzeitarchivierung ist keine nationale, sondern eine internationale Herausforderung und Aufgabe. Die Diskussion sollte zusätzlich zur nationalen auch auf die internationale Ebene im Europaparlament und in die Europäischen Kommission verlagert und dort verstärkt werden – gleiches gilt für die UNESCO. Bis zur jüngsten Restrukturierung innerhalb der Europäischen Kommission wurden gelegentlich Projekte im Bereich „Cultural Preservation and Digital Heritage" gefördert und zwar mit einem Gesamtbudget von ca. 15-20 Mio. Euro pro Jahr.[3] Seit der erwähnten Umstrukturierung heißt die Abteilung nun jedoch „Creativity".[4] Entsprechend dieser Neuausrichtung hin zu digitaler Kreativität wurden die Mittel für digitale Nachhaltigkeit drastisch gekürzt. Mit anderen Worten: Das Thema Langzeitarchivierung besitzt derzeit und in den kommenden Jahren auf der Ebene der Europäischen Kommission keine Sichtbarkeit, weshalb nur eingeschränkte Ressourcen zur Verfügung stehen. Ein möglicher Weg, dennoch zu einer Förderung dieses wichtigen Themas zu gelangen, könnte es also sein, die digitale Langzeitarchivierung mit verwandten Themen zu verbinden, die derzeit mehr Visibilität genießen und auf diese Weise nach und nach auch die digitale Nachhaltigkeit wieder in den Fokus der Aufmerksamkeit zu rücken.

2 Digitale Langzeitarchivierung: Status Quo

Die digitale Langzeitarchivierung ist ein internationales Thema und eine internationale Herausforderung. In Deutschland wird das Thema nicht auf Bundes-, sondern in erster Linie auf Landesebene bearbeitet, so dass Expertise, Infrastrukturen, Prozesse, Lösungen, Erfahrungen und Best-Practice-Ansätze von Bundesland zu Bundesland und von Institution zu Institution stark variieren.

Die zentrale Initiative der letzten Jahre ist Nestor („Network of Expertise in long-term storage and availability of digital resources"), das „deutsche Kompetenznetzwerk zur digitalen Langzeitarchivierung".[5] In Nestor arbeiten Bibliotheken, Museen und Archive gemeinsam zum Thema Langzeitarchivierung und Langzeitverfügbarkeit digitaler Objekte. Nestor wurde in zwei Projektphasen (2003–2006, 2006–2009) mit insgesamt ca. 1,2 Mio. Euro vom BMBF gefördert.[6] Auch nach dem Ablauf der Förderung ist Nestor weiter aktiv und agiert als selbstständiger Kooperationsverbund mit Partnern wie z. B. der Bayerischen Staatsbibliothek, dem Bibliotheksservice-Zentrum Baden-Württemberg, der Deutschen Nationalbibliothek und dem Institut

[2] http://de.wikipedia.org/wiki/Recht_auf_Vergessenwerden
[3] http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-projects_en.html
[4] http://cordis.europa.eu/fp7/ict/creativity/creativity_en.html
[5] http://www.langzeitarchivierung.de
[6] http://www.langzeitarchivierung.de/Subsites/nestor/DE/Header/Ueberuns/Projektartikel.html

für Deutsche Sprache. Nach eigenen Angaben pflegt Nestor enge Kontakte zu entsprechenden Initiativen anderer Länder und beteiligt sich an europäischen und internationalen Projekten.

Nestor hat im Rahmen seiner Laufzeit und auch im Anschluss wertvolle Impulse für die Entwicklung eines allgemeinen Rahmens für das Thema sowie einer Strategie geliefert. Die zahlreichen Ergebnisse von Nestor flossen ein in Handbücher zur Langzeitarchivierung (Neuroth et al., 2010, 2012), drei DIN-Standards sowie einige Ratgeber und Broschüren.[7] Außerdem wurde ein Konzept für die Aus- und Weiterbildung entwickelt. Ebenfalls hat Nestor einen nicht zu unterschätzenden ersten Beitrag zum Aufbau einer Community rund um das Thema Langzeitarchivierung geliefert.

Konkrete technische Lösungen für die Langzeitarchivierung hat Nestor nicht hervorgebracht. Diese waren allerdings auch nicht das Ziel der Initiative, werden jedoch nichtsdestotrotz von den beteiligten Einrichtungen und Gedächtnisinstitutionen dringend und zwingend benötigt.

3 Gemeinsame Bedarfe – gemeinsame Lösungen

Derzeit verwenden interessierte oder das entsprechende Mandat besitzende Institutionen für die digitale Langzeitarchvierung eine heterogene Mischung verschiedenster Ansätze, Software-Werkzeuge, Datenbanken und Repository- sowie Content-Management-Systeme. Ein wesentlicher Grund für den aktuellen, in jeder Hinsicht von Heterogenität geprägten Stand besteht in der Tatsache, dass die digitale Langzeitarchivierung in Deutschland auf Grund der Kulturhoheit Ländersache ist und somit alle Länder und sogar die einzelnen Organisationen innerhalb der Länder ihre eigenen Wege verfolgen und eigene Lösungen erarbeiten, d. h. entweder selber Software entwickeln oder mal diese und mal jene Software einsetzen und dabei immer wieder den gleichen Problemen begegnen.

Der vorliegende Beitrag ruft dazu auf, nicht nur eine nationale, sondern eine internationale Strategie für digitale Langzeitarchivierung zu erarbeiten und diese, möglichst europaweit, als digitale Infrastruktur technisch umzusetzen und interessierten Einrichtungen und gegebenenfalls auch Unternehmen anzubieten. Zahlreiche Gründe sprechen für einen derartigen, europaweit koordinierten Ansatz.

Zunächst ist zwischen zentralen und dezentralen Aufgaben zu differenzieren. Die zentralen, immer wiederkehrenden Aufgaben beinhalten technische Aspekte der Speicherung und Verfügbarmachung von Daten, die Spezifikation abstrakter Metadatenschemata sowie die Bearbeitung rechtlicher Fragestellungen und Herausforderungen. Es ergibt keinerlei Sinn, dass z. B. dezentral arbeitende Gedächtnisinstitutionen das Rad immer wieder neu erfinden, Software für Repositorien evaluieren, Rechtsexperten beauftragen oder Untersuchungen verschiedener Query-Sprachen durchführen. Derartige abstrakte technische Aspekte sollten im besten Fall von einer gemeinsamen technischen Infrastruktur angeboten und bestmöglich unterstützt werden. Eine solche technische Infrastruktur kann beispielsweise konzeptualisiert werden als eine Art virtueller Cloud-Speicher, der über verabredete Programmierschnittstellen (APIs) zugreifbar ist. Aus Gründen der Ausfallsicherheit und Redundanz wäre es hilfreich, ein derartiges Cloud-Archiv als verteilten Speicher aufzubauen, an dessen Bereitstellung viele verschiedene Rechenzentren teilnehmen – auch über Ländergrenzen hinweg.

Der evidenteste Fall einer zentralen, maximal generischen Aufgabe betrifft die Archivierung von Webseiten: Hier reichte letztlich eine einzelne Einrichtung aus, die sämtliche Webseiten Deutschlands, Europas oder auch der Welt konstant traversiert, archiviert und langfristig verfügbar hält – im Idealfall auf mehrere, räumlich weit entfernte, redundant speichernde Datenzentren verteilt. An einer solchen Aufgabe könnte sich z. B. ein institutionalisiertes Konsortium aus mehreren europäischen Datenzentren beteiligen. Die Präzedenzfälle sind alle weltweit operierenden Suchmaschinen (Google, Bing), sowie das Internet Archive.[8] Wenn es um die Archivierung typischer digitaler Objekte mit einer generischen, sehr weit verbreiteten Struktur und etablierten Forma-

[7] http://www.langzeitarchivierung.de/Subsites/nestor/DE/Publikationen/publikationen_node.html
[8] http://www.google.com, http://www.bing.com, http://www.archive.org.

ten geht (Webseiten, Sounddateien, Filme etc.), können und müssen derartige zentrale Dienste Verwendung finden und zwar vollautomatisch. Die selektive, punktuell, auf Anfrage durchgeführte Archivierung einzelner Webseiten oder kleinerer Websites kann und darf keine dezentral durchzuführende Aufgabe kleinerer oder größerer Landesbibliotheken sein.

Zu den dezentralen Aufgaben gehören die regional-, institutions-, domänen- und objektspezifischen sowie alle fachwissenschaftlichen Aspekte; die Grenze zwischen zentralen und dezentralen sowie zwischen domänenunabhängigen und domänenabhängigen Aufgaben ist dabei nicht als trennscharf zu verstehen. Dezentrale Aufgaben sind etwa die Spezifizierung von Metadatenschemata für bestimmte, insgesamt eher seltene oder ungewöhnliche Typen analoger oder digitaler Objekte oder die Entwicklung hochgradig spezialisierter Scanner oder Archivierungsverfahren. Gedächtnisinstitutionen und vergleichbare Einrichtungen, die sich mit speziellen Domänen oder Fachwissenschaften beschäftigen, können wiederum im Hinblick auf ihre spezifischen Lösungen kooperieren und von Synergien profitieren. Je inhärent dezentraler und fachwissenschaftlicher die Archivierungsaufgabe, desto unwahrscheinlicher ist es, diese vollautomatisch durchführen zu können.

Die Planung und Umsetzung eines derartigen Vorhabens verursacht immense Kosten. Daher ist es ist zwingend notwendig, Synergien mit ähnlichen Initiativen zu identifizieren und diese bestmöglich zu nutzen. Nur wenn eine derartige kontinentweite Allianz aus Archiven, Bibliotheken, Universitäten, Administrationen, Museen und anderen Gedächtnisinstitutionen usw. gebildet wird, kann die Umsetzung eines derartigen sehr ambitionierten Plans in den Bereich des Machbaren rücken.

4 Technische Herausforderungen

Ein derartiges Vorhaben wird mit verschiedenen technischen Herausforderungen konfrontiert. Exemplarisch seien an dieser Stelle einige wenige genannt, die sich auf die Archivierung des World Wide Web beziehen: In den Anfangstagen und -jahren des WWW bestanden Webseiten aus statischen Dokumenten, die beim Aufruf aus dem Dateisystem des Webservers geladen und über das Netzwerk zur Darstellung an den Browser des Nutzers durchgereicht wurden. Derartige statische Dokumente sind heutzutage nur noch die Ausnahme, denn Dokumente bestehen typischerweise aus mehreren einzelnen Komponenten, Widgets, Mashups und Objekten, die aus verschiedenen Quellen stammen und rechtlich verschiedene Urheber besitzen. Nach dem Aufruf wird die Seite entweder vom Webserver oder vom Browser dynamisch zusammengesetzt; ein Aufruf aus Berlin kann zu einem anderen Dokument führen als ein Aufruf aus Köln, Hamburg, London, New York, Hong Kong oder Lima. Variationen können stattfinden in Bezug auf die Sprache, grafische Gestaltung, sowie Anzahl und Ausprägung dynamisch eingebetteter Objekte und auch auf die Frage, in welchen (anderen) Webservices der Nutzer gerade angemeldet ist (Stichworte: Personalisierung von Inhalten, personalisierte Werbung). Welche Variante, welche Sicht eines Dokuments kann oder soll archiviert werden? Ist diese Variante archivierungswürdiger als jene? Und: Ergibt das Archivieren eines derart emergenten Mediums überhaupt einen Sinn, wenn der individuelle Wahrnehmungs- und Erlebniskontext nicht zur Verfügung steht, wenn also z. B. nur einzelne Bestandteile einer Seite im Archiv vorliegen und andere nicht, wenn aus der dynamischen und interaktiven Webseite ein statistisches Dokument geworden ist, das zwar als solches betrachtet, mit dem aber nicht interagiert werden kann, weil Hyperlinks und eingebetteten Objekte nicht funktionieren? Im Hinblick auf genuine Webapplikationen, die nicht etwa wie digitale Dokumente wirken, sondern tatsächlich die Anmutung typischer Software-Anwendungen besitzen (z. B. Google Mail, Apple iWorks, Bing Maps etc.) verkompliziert sich dieses Problem um ein Vielfaches (ausführlich hierzu Rehm, 2007).

Doch nicht nur die Ebene der Inhalte, auch die Ebene von Software und Hardware spielen zentrale Rollen: Abhängig vom verwendeten Browser kann sich das Aussehen eines Dokuments verändern. Gerade die Anfangsjahre des World Wide Web waren gekennzeichnet von den „browser wars“, in denen insbesondere Netscape und Microsoft versuchten, die Konkurrenz durch neue, allerdings proprietäre Funktionen zu übertrumpfen,

um auf diese Weise mehr Marktanteile zu bekommen.[9] Während Standardisierungsbemühungen seitens des World Wide Web Consortiums (W3C) und anderer Organisationen die Situation hier mittlerweile deutlich entschärft haben, stellt sich dennoch die Frage, ob man nicht zusätzlich auch die Ebene der Software und auch der Hardware in den Archivierungsprozess einbeziehen sollte. Ist es ein authentisches Erlebnis, ein HTML-Dokument von 1995 mit einem aktuellen Browser wie etwa Chrome oder Firefox darzustellen und zu lesen? Oder sollte das Dokument nicht vielmehr mit der Software und im besten Fall auch der Hardware (inklusive typischer Internet-Bandbreite) seiner Zeit dargestellt werden? Der erste grafische Webbrowser überhaupt wurde von Tim Berners-Lee auf einem Rechner der Firma NeXT entwickelt und trug den Namen WorldWideWeb. Ist es die Aufgabe von Gedächtnisinstitutionen, ein authentisches Gesamterlebnis des Bedienens dieser längst nicht mehr verfügbaren Software auf einer seit vielen Jahren nicht mehr verfügbaren Hardware zu ermöglichen? Möglicherweise können Emulatoren und entsprechende Software-Archive einen entscheidenden Beitrag leisten, um derartige Zeitreisen zu realisieren. Gegebenenfalls könnte hier auch das Erhalten exemplarischer Hard- und Software sowie typischer Webseiten aus verschiedenen Epochen des World Wide Web und klar gekennzeichneten Designperioden bzw. mit unterschiedlichem Funktionsumfang ein möglicher Ansatz sein, wie er auch in der Welt der Dinge im Musealen verfolgt wird.

5 Synergien durch verwandte Initiativen

Bei den verschiedenen Vorträgen und Diskussionen im Rahmen der Initiative „Nachhaltigkeit in der Digitalen Welt“ des *Internet und Gesellschaft Co:llaboratory*[10] stellte sich deutlich heraus, dass verschiedene Entwicklungen der letzten Jahre, insbesondere auf der europäischen Ebene, nicht oder nur marginal bekannt waren. Dies betrifft nicht nur den aktuellen Stand z. B. frei verfügbarer Crawler-Technologien, sondern etwa auch die Bereiche Forschungsinfrastrukturen und Metadatenschemata zur Beschreibung digitaler Objekte.

Wie bereits in Abschnitt 3 erläutert wurde, ruft dieser Artikel dazu auf, eine internationale Strategie für digitale Langzeitarchivierung zu entwickeln, diese als europaweite Infrastruktur zu implementieren und interessierten Einrichtungen und Unternehmen als Service anzubieten. Umgesetzt werden kann dieser Vorschlag nur in enger Kooperation mit bestehenden Initiativen, die sich mit digitalen Forschungs-, Metadaten- und Kommunikationsinfrastrukturen beschäftigen. Hierzu zählen insbesondere langfristig angelegte Vorhaben wie Europeana, CLARIN und META-SHARE des europäischen Exzellenznetzwerks META-NET.[11]

Diesen Infrastrukturen, Angeboten und Systemen ist gemein, dass sie in ihren teils sehr umfangreichen Katalogen diverse Metadaten über digitale Objekte unterschiedlicher Art beinhalten, die online nach verschiedenen Kriterien durchsucht und recherchiert werden können (Rehm et al., 2008a,b). Eben dies ist eine der zentralen Kernfunktionen einer Plattform für digitale Nachhaltigkeit (Rehm et al., 2009, 2010) und auch für die digitale Langzeitarchivierung: Die Erfassung, Repräsentation und Speicherung von Metadaten über digitale Objekte und die Bereitstellung möglichst komfortabler Such- und Rechercheschnittstellen. Obwohl sich die oben genannten Plattformen in vielerlei Hinsicht unterscheiden, ist dies ihr gemeinsamer Kern.[12]

In allen genannten und verschiedenen weiteren Initiativen, die sich ebenfalls mit der Speicherung und Verfügbarmachung digitaler Objekte beschäftigen, stehen immer wieder die gleichen Fragestellungen im Mittelpunkt (Schmidt et al., 2006, Rehm et al., 2008c, Gavrilidou et al., 2012). Hierzu zählen unter anderem rechtliche Aspekte (Urheberrechtsfragen in Bezug auf die ursprünglichen Objekte, in Bezug auf ihre digitalen Entsprechungen und auch in Bezug auf die zugehörigen Metadaten, Lehmberg et al., 2007, 2008, Weitzmann et al.,

[9] http://en.wikipedia.org/wiki/Browser_wars

[10] http://www.collaboratory.de/w/Initiative_Nachhaltigkeit_in_der_Digitalen_Welt

[11] http://www.clarin.eu, http://www.europeana.eu, http://www.meta-share.eu, http://www.meta-net.eu.

[12] Dieser gemeinsame Kern wiederum wurde von verschiedenen Initiativen zum Anlass genommen, entsprechende Open Source-Technologien wie z. B. Fedora Commons zu entwickeln, die das Management, die Nachhaltigkeit und die Verlinkung beliebiger digitaler Inhalte erlauben, siehe http://www.fedora-commons.org

2010, Steinhauer, 2013), bestmögliche Suche in den Katalogen (Information Retrieval, intelligentes Querying, Vorschläge, Nutzermodelle etc.), Erarbeitung und Spezifikation von Metadatenschemata zur Beschreibung der digitalen Objekte der jeweiligen Domäne, Speicherung der Metadaten sowie der digitalen Objekte (zentrale vs. dezentrale Repositorien, Hochverfügbarkeit etc.) und Import bestehender Datenbestände. Die Bearbeitung derartiger Fragen findet typischerweise in fachwissenschaftlichen EU-geförderten Projektkonsortien statt. Die Community um digitale Langzeitarchivierung sollte mit den oben genannten und verschiedenen weiteren Initiativen in Dialog treten, um gemeinsame Themen und mögliche Synergien zu identifizieren, um im Anschluss strategische Allianzen zu schmieden.

Die oben genannten Initiativen sind Beispiele einiger Fachwissenschaften, die innerhalb geförderter Projekte Plattformen und Infrastrukturen entwickeln, um ihre Datensammlungen (Forschungsdaten, Primärdaten, Analysedaten, Mess- und Sensordaten etc.) sowie zugehörige Werkzeuge und Technologien in zentralen „one stop shop"-Repositorien zu sammeln und der jeweiligen Forschungs-Community anzubieten. CLARIN macht dies für die Computerlinguistik und verfolgt einen komplexen eScience-Ansatz, wohingegen META-NET (Rehm und Uszkoreit, 2011, 2013) und META-SHARE (Piperidis, 2012) die angewandt ausgerichtete und mit der Industrie eng kooperierende Sprachtechnologie adressieren und einen eher einfachen peer-to-peer-Ansatz präferieren, der verteilte Knoten vorsieht, die jeweils organisations-, einrichtungs- oder auch landesspezifische Metadatenkataloge beinhalten, die in regelmäßigen Abständen in einem gemeinsamen Katalog aggregiert werden. Vergleichbare Initiativen existieren in diversen weiteren Fachwissenschaften. Europeana wiederum ist eine deutlich umfangreicher finanzierte Initiative, die die Metadatenkataloge zahlreicher Museen und Gedächtnisinstitutionen zusammenführt, dabei aber mit dem Problem hochgradig heterogener Metadaten konfrontiert wird. Eine weitere relevante, interdisziplinäre Community ist aus dem Einsatz von Auszeichnungssprachen wie SGML und XML für Aufgaben in der Geisteswissenschaft hervorgegangen; bei den jährlich stattfindenden internationalen Konferenzen der „Digital Humanities"-Reihe werden mittlerweile auch neue Ansätze und Lösungen präsentiert, die für die digitale Nachhaltigkeit relevant sind.

Es zeigt sich also, dass derzeit die zentralen Impulse im mittelbaren Bereich der digitalen Langzeitarchivierung, die auch genügend Sichtbarkeit und Unterstützung und somit Förderung erhalten, primär aus den verschiedenen Fachwissenschaften mit ihren jeweils spezialisierten Initiativen wie z. B. CLARIN und META-SHARE stammen. Diese Wissenschaften erzeugen und analysieren digitale Objekte jeglicher Art, Komplexität und Größe. Die zentrale Sammlung und Bereitstellung der Daten hat unter anderem genuin wissenschaftliche Gründe, so dass interessierte Kollegen beispielsweise Experimente nachvollziehen können. Es geht dabei aber auch um die Beschleunigung des Forschungstransfers durch die Bereitstellung etwa von Sprachressourcen oder neuen Technologien, so dass interessierte Unternehmen von den neuesten Forschungsergebnissen profitieren können. Interessanterweise wird der Aspekt der Langzeitarchivierung in derartigen Initiativen zwar auch immer wieder diskutiert, er steht aber nicht im Mittelpunkt der Diskussionen, so dass im Falle einer langfristigen Kooperation zwischen den Experten für Langzeitarchivierung und den o. g. fachwissenschaftlichen Initiativen beiderseitig mit gewinnbringenden Synergien gerechnet werden kann.

6 Schlussfolgerungen und Empfehlungen

Das Thema *digitale Langzeitarchivierung* wird mit sehr hohen Kosten, in verschiedenen Bereichen einer unklaren Rechtslage und riesigen Datenmengen konfrontiert. Außenstehenden erscheint das Anliegen von Fürsprechern des Themas oftmals zwar wichtig, es wird aber nicht als dringend wahrgenommen, was dazu führt, dass die digitale Nachhaltigkeit gerne aufgeschoben bzw. ihre Priorität herabgestuft wird. Eine Lösung für dieses Dilemma kann und muss es sein, die digitale Langzeitarchivierung bzw. zentrale Aspekte der digitalen Langzeitarchivierung unter einem anderen Oberthema zu bearbeiten. Ein mögliches Oberthema könnte die Entwicklung einer europäischen Suchmaschine bzw. einer digitalen Infrastruktur mit zahlreichen integrierten Services für

Europa sein. Ein alternatives Thema könnte eine möglichst generische digitale Infrastruktur für eScience und digitale Forschungsdaten sein. Solche EU-Infrastrukturen könnten zahlreiche positive Strahleffekte ausüben und Innovation sowie Forschungstransfer fördern. Unter diesem Mantel könnte auch das Thema der digitalen Langzeitarchivierung bearbeitet werden.

Dieser Beitrag strebt den Versuch eines Brückenschlags an zwischen Experten für digitale Langzeitarchivierung und verschiedenen fachwissenschaftlich ausgerichteten Initiativen, die sich eher mit dem Aufbau domänenspezifischer Forschungsinfrastrukturen beschäftigen. Eine weitere, bereits erfolgreich auf europäischer Ebene tätige relevante Initiative in diesem Zusammenhang ist Europeana. Nicht nur Deutschland, sondern Europa benötigt eine gemeinsame Position und Strategie für digitale Langzeitarchivierung, die als kontinentweite Infrastruktur implementiert und interessierten Einrichtungen und ggf. auch Unternehmen angeboten werden kann. Der Artikel appelliert dazu, eine enge Kooperation mit den bestehenden Initiativen aufzubauen, die sich mit digitalen Forschungs-, Metadaten- und Kommunikationsinfrastrukturen beschäftigen.

Deutschland benötigt dringend eine Strategie zur digitalen Langzeitarchivierung. Sie sollte kleine, realistische, finanzierbare, umsetzbare Schritte enthalten. Die Planung der Strategie sollte auf Grund der zu erwartenden massiven Kosten für die Infrastruktur und Technologieentwicklung nicht isoliert erfolgen, sondern übergreifend von der Europäischen Kommission koordiniert werden, um z. B. eine gemeinsame technologische Infrastruktur zu entwickeln und bereit zu stellen (etwa nach dem Vorbild von Europeana). Die digitale Langzeitarchivierung sollte ein integraler Bestandteil der nächsten Version der digitalen Agenda der Europäischen Kommission werden. In „Digital Agenda for Europe“ (EC, 2010) tauchen die Begriffe „archive“ oder auch „sustainability“ in der hier gemeinten Lesart nicht ein einziges Mal auf. Die Entwicklung und Umsetzung einer nationalen bzw. internationalen Strategie zur digitalen Langzeitarchivierung birgt genügend Potenzial, um zahlreiche Strahleffekte in verschiedene Bereiche der Informationstechnologie zu generieren. Dabei ist es essentiell, eine Infrastruktur zu entwickeln oder die Entwicklung einer Infrastruktur fortzuführen und dabei vorhandene Technologielücken zu schließen, um auf diese Weise sowohl den europäischen Bürgerinnen und Bürgern als auch Kommunikation, Wirtschaft und Kultur wichtige und relevante neue Dienste anzubieten, die einen genuinen Mehrwert schaffen. Die Anforderungen, Methoden und Prinzipien, die für die Langzeitarchivierung wissenschaftlicher Forschungsdaten gelten, können letztlich auch für alle anderen Daten angewendet werden, seien es private Daten, Webseiten oder auch öffentliche Verwaltungsdaten, weshalb wiederum ein gemeinsames technologisches Fundament eingesetzt werden kann, das auf konkreten Ebenen spezifische Services anbietet. Die Entwicklung von META-SHARE (2010-2012, siehe z. B. Federmann et al., 2012) hätte beispielsweise durch eine verfügbare Basisinfrastruktur drastisch vereinfacht werden können, die Evaluation verschiedener verfügbarer Open-Source-Frameworks hätte entfallen können, ebenso große Teile der Entwicklung der Nutzerschnittstelle.

Kurzfristig ist der Einsatz bereits verfügbarer Infrastrukturen zu testen und für den Zweck der digitalen Langzeitarchivierung zu evaluieren. Falls dabei technologische oder konzeptuelle Lücken identifiziert werden, können diese gemeinsam mit den Fachwissenschaftlern geschlossen werden. Von Relevanz sind hierbei auch die jeweiligen Funktionsmerkmale der domänen- und fachwissenschaftsspezifischen Infrastrukturen: Wo liegen Gemeinsamkeiten vor, wo Unterschiede? Basierend auf derartigen Evaluationen kann die Funktionalität einer abstrakten und generischen Basisinfrastruktur konzipiert werden.

Essentiell ist der Aufbau einer belastbaren europaweiten Community und Technologieallianz rund um das Thema digitale Langzeitarchivierung, die in der Lage ist, mit einer Stimme zu sprechen. Hierbei können insbesondere Verbindungen zu Unternehmen, denen entsprechende digitale Services kostenpflichtig angeboten werden, mittelfristig für eine substantielle Gegenfinanzierung und auch für Visibilität auf der Ebene nationaler und internationaler Entscheidungsträger sorgen. Ebenfalls sind gute Verbindungen zu den diversen Initiativen rund um Forschungsinfrastrukturen zu etablieren. Operationalisiert werden können solche Verbindungen über gemeinsame Forschungsthemen wie z. B. Metadatenschemata, Semantic Web (ausführlich hierzu Sasaki, 2013), Linked Open Data bis hin zu Technologiekomponenten, die von beiden Seiten benötigt werden. Eines

dieser gemeinsamen Themen könnte eine vollautomatische und möglichst generische Beantwortung der kritischen Frage sein, welche digitalen Objekte archiviert und welche ignoriert werden sollen. Es erscheint hilfreich, diese Frage automatisch von einem transparenten und objektiven Algorithmus beantworten zu lassen, der jedoch auch Ausnahmen zulässt. Als Indizien hierfür könnten entsprechende positive Rückmeldungen („likes", „favourites", „bookmarks" etc.) der Nutzer in den verschiedenen sozialen Medien und die digitale Sichtbarkeit der Objekte sowie ihrer Urheber in den großen Suchmaschinen sowie auch in traditionellen Archiven und Bibliotheken dienen. Zu den großen technischen Herausforderungen zählt dabei später, in der Phase des Betrachtens, allerdings die adäquate Rekonstruktion des kommunikativen und situativen Kontextes (siehe Abschnitt 4).

Mittel- bis langfristig sollte gemeinsam mit den verschiedenen Initiativen der Fachwissenschaften und insbesondere auch mit Europeana eine generische, europaweite Cloud-Infrastruktur entwickelt und implementiert werden, die Archivierung und Retrieval beliebiger digitale Objekte ermöglicht, d. h. ebenfalls die digitale Langzeitarchivierung mit Services unterstützt, die sowohl von Gedächtnisinstitutionen, aber auch von interessierten anderen Organisationen und Unternehmen genutzt werden können. Eine derartige Infrastruktur sollte als inhärente und dringend benötigte Komponente für Nachhaltigkeit der digitalen Infrastruktur des digitalen Marktes konzeptualisiert und anschließend extern kommuniziert werden. Dieses Thema ist in jeder Hinsicht kompatibel mit den beiden neuen großen Förderprogrammen Connecting Europe Facility (EC, 2011) und Horizon 2020 (EC, 2012). Hierzu ist auf der europäischen Ebene hinreichend Sichtbarkeit für dieses Thema zu generieren, denn nur auf der Ebene der EU können mittel- bis langfristig ausreichende Mittel zur Verfügung gestellt werden, um zu einer tatsächlich benutzbaren Implementierung zu gelangen, die europaweit Verwendung finden kann. Auf Grund der Höhe der zu investierenden Kosten erscheint es nicht realistisch, dass eine rein nationale Anstrengung dies gewährleisten könnte, so dass eine europaweite digitale Infrastruktur anzustreben ist. Diese könnte bestehen aus mehreren verteilten Datenzentren, bei denen Langzeitarchivierung, Hochverfügbarkeit und Hochgeschwindigkeit zum Designprinzip erklärt werden. Im Zeitalter von Prism und Tempora kann bereits jetzt prognostiziert werden, dass die EU oder möglicherweise auch neutrale Länder wie z. B. die Schweiz Gegenmaßnahmen ergreifen werden und zwar durch die Förderung und Realisierung einer Gruppe vertrauenswürdiger digitaler Dienste, die den Bürgerinnen und Bürgern zur Verfügung gestellt werden. Abhörsichere Kommunikationsverfahren mit vollständiger Endpunkt-zu-Endpunkt-Verschlüsselung und Kontrollmechanismen wie z. B. Trust-Centern sind notwendig, um für maximale Transparenz und Sicherheit zu sorgen, was ebenfalls für das Anbieten von Diensten für die digitale Langzeitarchivierung an Unternehmen ein missionskritischer Faktor sein wird. Optimalerweise ist es nur eine Frage der Zeit, bis die Europäische Union realisiert, dass derartige Dienste einen inhärenten Teil der europäischen Infrastruktur darstellen und ebenso wie Brücken, Straßen und Glasfasernetze gefördert werden müssen. Mit der „Digital Component" des 11 Mrd. Euro umfassenden Programms Connecting Europe Facility werden derzeit die ersten Schritte geplant. Falls also in einigen Jahren EU-geförderte Cloud-Speicher und sonstige Cloud-Dienste entstehen werden, sollten die Initiativen für digitale Langzeitspeicherung hinreichend gut aufgestellt und in der Lage sein, davon zu profitieren, indem alle generischen, domänen- und fachspezifischen Infrastrukturen daran angekoppelt werden.

Mittel- bis langfristig kann die Community um die digitale Langzeitarchivierung auch entscheidende Impulse liefern für die seit Jahren schwelende, aber auch stagnierende Debatte um eine nationale und internationale Reform des Urheberrechts, da sie verschiedene Aspekte dieser Debatte berührt und streng genommen erst durch eine vollständige Überarbeitung eine solide Rechtsgrundlage erhalten kann (Steinhauer, 2013). Geschieht diese Überarbeitung nicht bzw. zu spät, wird Deutschland digital vom Rest Europas bzw. vom Rest der Welt abgehängt. Die notwendige Überarbeitung des Urheberrechts könnte auch dafür sorgen, dass der Umgang mit digitalen Massendaten in der Forschung (etwa im Bereich Information Retrieval oder Sprachtechnologie) vereinfacht bzw. überhaupt erst ermöglicht wird. Durch entsprechende Forschungsprogramme können neue Innovationen, Verfahren, Algorithmen, Produkte, Dienstleistungen und Spin-Offs entstehen.

Literatur

EC (2010): "A Digital Agenda for Europe". European Commission. http://ec.europa.eu/information_society/digital-agenda/publications/.

EC (2011): "Connecting Europe Facility: Commission adopts plan for 50 billion Euros boost to European networks". European Commission. http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1200.

EC (2012): "Horizon 2020: The Framework Programme for Research and Innovation". European Commission. http://ec.europa.eu/research/horizon2020/.

Federmann, Christian; Giannopoulou, Ioanna; Girardi, Christian; Hamon, Olivier; Mavroeidis, Dimitris; Minutoli, Salvatore und Schröder, Marc (2012): "META-SHARE V2: An Open Network of Repositories for Language Resources Including Data and Tools". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, S. 3300–3303.

Gavrilidou, Maria; Labropoulou, Penny; Desipri, Elina; Piperidis, Stelios; Papgeorgiou, Haris; Monachini, Monica und Frontini, Francesca (2012): "The META-SHARE Metadata Schema for the Description of Language Resource". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, S. 1090–1097.

Lehmberg, Timm; Chiarcos, Christian; Rehm, Georg und Witt, Andreas (2007): "Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten". In: *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, herausgegeben von Rehm, Georg; Witt, Andreas und Lemnitzer, Lothar, Tübingen: Gunter Narr, S. 93–102.

Lehmberg, Timm; Rehm, Georg; Witt, Andreas und Zimmermann, Felix (2008): "Digital Text Collections, Linguistic Research Data, and Mashups: Notes on the Legal Situation". *Library Trends* 57 (1): S. 52–71.

Neuroth, Heike; Oßwald, Achim; Scheffel, Regine; Strathmann, Stefan und Huth, Karsten (Herausgeber) (2010): *nestor Handbuch – Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. vwh. Nestor. Version 2.3. http://www.langzeitarchivierung.de/Subsites/nestor/DE/Publikationen/Handbuch/handbuch_node.html.

Neuroth, Heike; Strathmann, Stefan; Oßwald, Achim; Scheffel, Regine; Klump, Jens und Ludwig, Jens (Herausgeber) (2012): *Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme*. vwh. Nestor. Version 1.0. http://www.langzeitarchivierung.de/Subsites/nestor/DE/Publikationen/Handbuch/handbuch_node.html.

Piperidis, Stelios (2012): "The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, S. 36–42.

Rehm, Georg (2007): *Hypertextsorten: Definition – Struktur – Klassifikation*. Norderstedt: Books on Demand.

Rehm, Georg; Schonefeld, Oliver; Trippel, Thorsten und Witt, Andreas (2010): "Sustainability of Linguistic Resources Revisited". In: *Balisage 2010. International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML*. Montréal, Canada.

Rehm, Georg; Schonefeld, Oliver; Witt, Andreas; Chiarcos, Christian und Lehmberg, Timm (2008a): "SPLICR: A Sustainability Platform for Linguistic Corpora and Resources". In: *KONVENS 2008 (Konferenz zur Verarbeitung natürlicher Sprache) – Textressourcen und lexikalisches Wissen*, herausgegeben von Storrer, Angelika; Geyken, Alexander; Siebert, Alexander und Würzner, Kay-Michael. Berlin, S. 86–95.

Rehm, Georg; Schonefeld, Oliver; Witt, Andreas; Hinrichs, Erhard und Reis, Marga (2009): "Sustainability of Annotated Resources in Linguistics: A Web-Platform for Exploring, Querying and Distributing Linguistic Corpora and Other Resources". *Literary and Linguistic Computing* 24 (2): S. 193–210. Selected papers from Digital Humanities 2008.

Rehm, Georg; Schonefeld, Oliver; Witt, Andreas; Lehmberg, Timm; Chiarcos, Christian; Bechara, Hanan; Eishold, Florian; Evang, Kilian; Leshtanska, Magdalena; Savkov, Aleksandar und Stark, Matthias (2008b): "The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources". In: *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.

Rehm, Georg und Uszkoreit, Hans (2011): "Multilingual Europe: A challenge for lang tech". *MultiLingual* S. 51–52. Issue April/May.

Rehm, Georg und Uszkoreit, Hans (Herausgeber) (2013): *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg, New York, Dordrecht, London: Springer. Buy this book at springer.com or amazon.de.

Rehm, Georg; Witt, Andreas; Hinrichs, Erhard und Reis, Marga (2008c): "Sustainability of Annotated Resources in Linguistics". In: *Digital Humanities 2008*. ACH, ALLC, Oulu, Finland.

Sasaki, Felix (2013): "Nachhaltigkeit und das Semantic Web". In: *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*, herausgegeben von Klimpel, Paul und Keiper, Jürgen, iRights Media. In diesem Band.

Schmidt, Thomas; Chiarcos, Christian; Lehmberg, Timm; Rehm, Georg; Witt, Andreas und Hinrichs, Erhard (2006): "Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources". In: *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*. East Lansing, Michigan.

Steinhauer, Eric W. (2013): "Wissen ohne Zukunft? Der Rechtsrahmen der digitalen Langzeitarchivierung von Netzpublikationen". In: *Was bleibt? Nachhaltigkeit der Kultur in der digitalen Welt*, herausgegeben von Klimpel, Paul und Keiper, Jürgen, iRights Media. In diesem Band.

Weitzmann, John Hendrik; Rehm, Georg und Uszkoreit, Hans (2010): "Licensing and Sharing Language Resources: An Approach Inspired by Creative Commons and Open Science Data Movements". In: *Proceedings of the LREC 2010 Legal Issues for Sharing Language Resources: Constraints and Best Practices*, herausgegeben von Choukri, Khalid; DiPersio, Denise; Kupietz, Marc und Mapelli, Valérie. Malta.