# Der Vergänglichkeit des Internets entgegenwirken

## DIE PROJEKTE **LIWA** UND **ARCOMEM** SOLLEN WEB-ARCHIVE QUALITATIV VERBESSERN

**Das World Wide Web ist zu einem allgegenwärtigen, aber auch vergänglichen Medium geworden, in dem sich viele Aspekte unserer heutigen Gesellschaft spiegeln. Diese offene Kommunikationsplattform ermöglicht es jedem Einzelnen, seine Meinung vor einem breiten Publikum zu äußern und mit anderen zu diskutieren. Dadurch ist es zum ersten Mal in der Geschichte möglich, eine Vielfalt an Meinungen und Ansichten zu dokumentieren. Dieses unschätzbare Wissen für die Zukunft zu bewahren, ist eine der Aufgaben der Web-Archivierung.**

Das World Wide Web – kurz das Web – ist mittlerweile für viele Menschen ein fester Bestandteil des täglichen Lebens geworden. Es gestattet, aktuelle Informationen zu verschiedensten Themen abzurufen, aber sich auch selbst in Diskussionen einzubringen (zum Beispiel Diskussionsforen, Facebook, Twitter), eigene Arbeiten mit wenig Aufwand zu veröffentlichen (zum Beispiel Blogs, Online-Publishing) oder gemeinsam an Projekten zu arbeiten (zum Beispiel Open Source Software, Wikipedia, GuttenPlag). Das Web stellt somit eine unschätzbare Sammlung von Wissen über unsere heutige Gesellschaft dar und sollte deshalb genauso bewahrt werden, wie dieses bereits mit traditionellen Medien geschieht.

Allgemeine Aussagen wie »das Web vergisst nicht« sind durch zahlreiche Studien widerlegt worden. So wurde nachgewiesen, dass innerhalb eines Jahres mehr als 40% der Seiten im Web ihren Inhalt ändern und dass die Halbwertzeit einer Webseite bei circa zwei Jahren liegt. Das bedeutet, dass innerhalb von 24 Monaten die Hälfte der Seiten im Web verschwindet. Praktische Gründe für das Verschwinden von Seiten sind beispielsweise Änderungen der verwendeten Software, wodurch vorhandene Verweise (Links) zerstört werden, Eigentümerwechsel bei Domainnamen, aber auch Naturkatastrophen, die zu Serverausfällen führen.

Um trotz dieser hohen Dynamik des Webs seine Inhalte zumindest teilweise zu erhalten, sind in den letzten 15 Jahren weltweit verschiedene Initiativen ins Leben gerufen worden, die Technologien zum Sammeln von Webinhalten entwickeln. Neben dem Internet Archive (**http://www.archive.org/**) waren dies in den Anfängen hauptsächlich Nationalbibliotheken, die sich um diese Thematik gekümmert haben. Mittlerweile zeigen auch die Industrie und andere Einrichtungen (zum Beispiel Rundfunkanstalten, Parlamente, Ministerien) Interesse an der regelmäßigen Sammlung von Web-Inhalten.

Der Prozess zum Einsammeln der zu archivierenden Seiten (»web crawling«) umfasst im Wesentlichen die Auswahl der Inhalte, das Herunterladen der Seite mit anschließender Extraktion der enthaltenen Verweise, die Speicherung der Seiten im Archiv sowie die qualitative Begutachtung der gesammelten Inhalte durch den Archivar. Durch das wachsende Interesse an der Web-Archivierung ergeben sich auch neue Anforderungen an alle genannten Schritten bezüglich der Qualität und Nutzbarkeit der Ergebnisse.

Das Forschungszentrum L3S entwickelt deshalb zusammen mit internationalen Partnern aus Wissenschaft, Industrie und öffentlichen Einrichtungen in den zwei Europäischen Projekten LiWA und ARCOMEM neue innovative Lösungen, um diese Anforderungen zu lösen.

### Verbesserung der Archivqualität

Das Ziel des vom Forschungszentrum L3S koordinierten Projektes LiWA (Living Web Archives) war die qualitative Verbesserung von Web-Archiven auf den Gebieten der Archiv-Genauigkeit, Archiv-Kohärenz und der Langzeitinterpretierbarkeit von Archiven. Dazu wurde der weitverbreitete Open Source Crawler Heritrix (**https://web-archive.jira.com/wiki/display/Heritrix**) um verschiedene Module erweitert.

Eine Herausforderung beim Erreichen einer hohen Archiv-Genauigkeit, d.h. der Übereinstimmung des Archivinhalts mit der ursprünglichen Webseite, ist die zunehmende Verwendung von dynamischen Inhalten durch den Einsatz von JavaScript oder Flash. Waren in den Anfängen des Web Verweise (Links) von einer Webseite auf eine andere noch explizit in der Seite enthalten, werden diese heute häufig erst

in dem Moment dynamisch erzeugt, wenn der Benutzer einen Link anklickt. Mit dem im LiWA-Projekt entwickelten Ansatz des »Execution Based Crawling« werden alle Benutzerinteraktionen mit einer Seite simuliert, das heißt, es werden der Reihe nach alle Links und Schaltflächen virtuell gedrückt. Nach jedem Simulationsschritt wird die geänderte interne Repräsentation der Webseite analysiert und die enthalten Links extrahiert. Das Ergebnis dieser Extraktion wird dem Crawler als neuer Sammelauftrag übergeben.

Bei der Extraktion der Links stößt man auch schnell auf solche, die nicht für den Benutzer bestimmt sind, sondern gezielt zur Manipulation von Suchmaschinen eingestreut wurden. Da Crawler für die Web-Archivierung ähnlich wie Suchmaschinen arbeiten, können sie leicht in dieselben Fallen laufen und so größere Mengen an Spamseiten einsammeln. Die Identifikation und Reduktion von Spam in Archiven wurde in LiWA durch die Kombination von maschinellen Lernverfahren und Listen von erlaubten und nicht-erlaubten Webseiten gelöst. Der Archivar hat zudem die Möglichkeit, Web Spam in seinem Archiv zu markieren und somit den Spamfilter zu trainieren. Dieses Training und eine regelmäßige Anpassung der zu analysierenden Spameigenschaften sind notwendig, da sich die Methoden und Strategien der Spammer ebenfalls regelmäßig ändern.

Eine allgemeine Anforderung an Web Crawler ist, dass diese sich »freundlich« zu den Websites verhalten müssen, deren Inhalte sie einsammeln möchten. Das bedeutet, dass sie nur alle 1 bis 3 Sekunden eine Anfrage an den Webserver schicken sollten. Folglich kann das Einsammeln einer großen Website mehreren Stunden bis Tage dauern. In dieser Zeit können sich die Strukturen und Inhalte verändern, die wiederum zu fehlerhaften Verweisen und inkohärenten Inhalten im Archiv führen können. In LiWA wurden deshalb Methoden entwickelt um Seiten mit einer hohen Änderungsfrequenz zu identifizieren und dieses Wissen für die Planung des Crawling einzubeziehen.

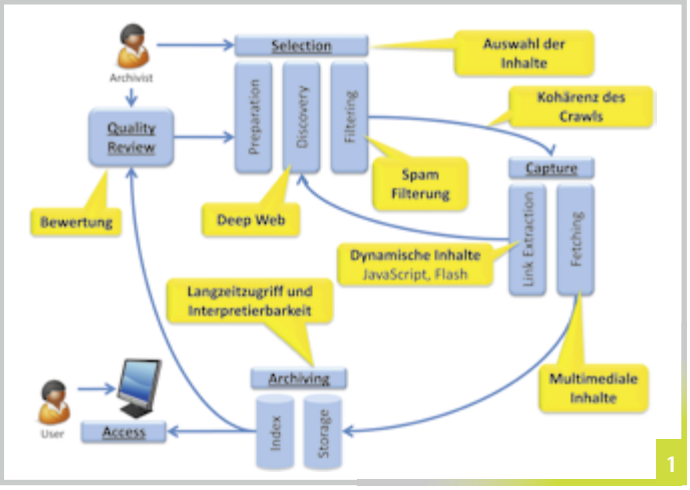


Abbildung 1
*Der Prozess der Webarchivierung mit ausgewählten Forschungsthemen*

### Nutzung des SocialWeb für die Web-Archivierung

Ein wichtiger Aspekt für die Web-Archivierung ist die Auswahl der Inhalte. Derzeit geschieht dieses durch den Archivar, der manuell eine Liste von Ausgangsseiten (»seed list«) erstellt, die durch den Web-Crawler gesammelt werden sollen. Eine Möglichkeit, den Crawler inhaltlich zu steuern, existiert derzeit aber noch nicht. Archivaren fällt es zudem zunehmend schwerer, manuell die relevanten Inhalte auszuwählen, da immer mehr Benutzer soziale Plattformen wie Blogs, Facebook etc. für die Publikation von Inhalten nutzen und herkömmliche Archivierungslösungen diese Plattformen nicht berücksichtigen. Hier setzt das im Januar 2011 gestartete europäische Projekt ARCOMEM (Archive Communities MEMories) an.

Ziel des ARCOMEM-Projektes ist die Umwandlung von statischen Webarchiven in eine Art »kollektives Gedächtnis«, das eng mit seinen Benutzern verbunden ist. Erreicht werden soll dies, indem nicht nur Inhalte von sozialen Plattformen gesammelt, sondern die Plattformen auch nach Hinweisen auf weitere relevante Inhalte analysiert werden sollen. Die Idee dahinter ist, dass Benutzer von sozialen Plattformen in ihren Beiträgen auch Verweise auf andere Diskussionen oder externe Quellen angeben, die der Benutzer als interessant empfunden hat. Je mehr Empfehlungen für eine Seite ausgesprochen werden, desto relevanter ist diese aus der Sicht der Öffentlichkeit. Der ARCOMEM-Webcrawler behandelt diese Seiten mit höherer Priorität als andere. Zusätzlich wird die Bedeutung einer Seite zum Sammelzeitpunkt dokumentiert.

Zur Entscheidung, ob eine Seite relevant für das Archiv ist, trägt aber nicht alleine ihre Empfehlung auf sozialen Plattformen bei. Der Archivar kann zusätzlich eine Liste von gewünschten Inhalten (zum Beispiel Personen, Orte, Ereignisse wie »Olympische Spiele«) und/oder Themen (zum Beispiel Finanzkrise) vorgeben.

**Dr.-Ing. Thomas Risse**
Jahrgang 1969, Stellvertretender Geschäftsführer des Forschungszentrums L3S, Wissenschaftlicher Leiter des ARCOMEM Projektes, Koordinator des LiWA Projektes. Kontakt: risse@l3s.de

Während des Crawls wird jede Seite bezüglich ihrer Relevanz anhand dieser Spezifikation überprüft. Wird eine Seite als nicht relevant eingestuft, so ist die Annahme, dass die enthaltenen Verweise ebenfalls auf nicht relevante Seite zeigen. Weiterführende Crawlaktivitäten in dieser Richtung werden deshalb an dieser Stelle eingestellt.

Um die spätere Nutzung der gesammelten Inhalte zu verbessern, werden diese automatisch einer Inhaltsanalyse unterzogen. Hierbei werden, unabhängig von der Crawlspezifikation, die vorkommenden Entitäten, Ereignisse und Themen sowie die Grundstimmung der Seite extrahiert und analysiert und zusammen mit der Seite gespeichert. Der Archivar oder die späteren Nutzer können dann über verschiedene Facetten gezielt auf Inhalte im Archiv zugreifen.

## Langzeitinterpretierbarkeit der Inhalte

Eine projektübergreifende Aktivität, in die das Forschungszentrum L3S insbesondere involviert ist, ist die Langzeitinterpretierbarkeit der Archivinhalte. Sprache und die Bedeutung von Begriffen und Konzepten beziehungsweise deren Wahrnehmung ändern sich über die Zeit. Wenn dem Benutzer des Archivs beispielsweise bekannt ist, dass die heutige Stadt St. Petersburg zeitweise auch die Namen Leningrad und Petrograd trug, kann er dieses in seinen Suchanfragen berücksichtigen. Für viele Begriffe geht dieses Wissen allerdings verloren und so können relevante Inhalte in digitalen Archiven nur schwer gefunden werden, da die Anfragen die »falschen« Begriffe verwenden. Suchanfragen können in diesem Fall automatisch mit der Hilfe von Thesauri (Synonym-Wörterbücher) ergänzt werden. Thesauri können aber nur einen Teil der sprachlichen und konzeptuellen Entwicklung abdecken. Das ist insbesondere auf sozialen Plattformen der Fall, wo die Benutzer relativ schnell neue Begriffe, Spitznamen oder Abkürzungen entwickeln. Aus diesem Grund arbeitet das Forschungszentrum L3S an Methoden, die die Beschreibung von Begriffen aus den Archivinhalten direkt extrahieren. Die Ergebnisse im Rahmen des LiWA Projektes haben gezeigt, dass die verwendeten Methoden zu Extraktion und Bündelung von Wörtern geeignet sind, um Begriffe zu beschreiben und diese über die Zeit zu verfolgen. Im Rahmen des ARCOMEM Projektes werden diese Technologien weiterentwickelt für den Einsatz mit Informationen von sozialen Plattformen.

Abbildung 2
*Das ARCOMEM-Team (v.l.n.r.): Gerhard Gossen, Dr. Thomas Risse, Nina Tahmasebi, Dr. Stefan Dietze, Elena Demidova, Gideon Zenz*

## Fazit

Das Thema Web Archivierung, obwohl schon seit 1997 betrieben, ist immer noch ein recht junges Thema, das weltweit nur von einer kleinen Gruppe engagierter Organisationen betrieben wird. Mit den Projekten LiWA und ARCOMEM werden verschiedenen Methoden zur qualitativen Verbesserung der Archive und zur Entlastung der Archivare entwickelt. Die weiterhin dynamische Entwicklung des Web macht es aber notwendig, auch in Zukunft weiterhin Forschung in Bereich Web Crawlings und Nutzung von Langzeitarchiven zu betreiben.

## Referenzen


- Internet Archive: **http://www.archive.org/**
- LiWA Projekt (EU IST 216267): **http://www.liwa-project.eu/**
- ARCOMEM Projekt (EU IST 270239): **http://www.arcomem.eu/**
- Nina Tahmasebi, Kai Niklas, Thomas Theuerkauf, Thomas Risse; Using Word Sense Discrimination on Historic Document Collections; In Proc. of the 10th Annual Joint Conference on Digital Libraries (JCDL 2010). ACM, New York, NY, 89–98.
- Thomas Risse, Stefan Dietze, Diana Maynard, Nina Tahmasebi, Wim Peters; Using Events for Content Appraisal and Selection in Web Archives; In Proc. of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011), Bonn, Germany, October 23, 2011.
- Ntoulas, A., Cho, J., Olston, C.: What's new on the web?: the evolution of the web from a search engine perspective. In: Proceedings of the 13th International Conference on World Wide Web 2004, pp. 1–12. ACM Press, New York, 2004.