Zeitschrift: Bulletin / Vereinigung Schweizerischer Hochschuldozenten = Association Suisse des Professeurs d'Université

Herausgeber: Vereinigung Schweizerischer Hochschuldozenten

Band: 30 (2004)

Heft: 1

Artikel: Zur Archivierung wissenschaftlicher Texte und verwandten Problemen

Autor: Koch, Hans

DOI: https://doi.org/10.5169/seals-894284

Nach einer erfolgreichen Entwicklung von eL-Angeboten sollen mittelfristig qualitativ hochstehende und zertifizierte eL-Ressourcen, insbesondere eigenständige *short* eL-Kurse, einzeln oder gebündelt über das produzierende Kompetenznetzwerk hinaus angeboten und genutzt werden können. Für die Verbesserung der dazu notwendigen Infrastruktur sowie die Einhaltung von rechtlichen Bestimmungen sind der SVC und die einzelnen Institutionen gefordert.

Das Bundesprogramm "Swiss Virtual Campus" hat in vielfacher Hinsicht Neuland betreten und tut dies weiter. Die ersten Erfahrungen zeigen auch in der Schweiz, dass man mit den neuen Technologien keine grenzenlosen Innovationsschübe im Bildungswesen erwarten darf. Schritte der Verbesserung bezüglich Interaktivität, Effizienz, Motivation, Lehr- und Lernerfolg, Vernetzung, Qualität und mittelfristig auch Kosten sind allerdings offensichtlich.

Dank

Der Verfasser dieser Veröffentlichung möchte es nicht unterlassen, den Mitgliedern des SVC Lenkungsausschusses, der SVC Koordinationsstelle, den Generalsekretariaten der Schweizerischen Hochschulkonferenz (SUK) und der Rektorenkonferenz der Schweizer Universitäten (CRUS) sowie dem Bundesamt für Bildung und Wissenschaft (BBW) und dem Bundesamt für Berufsbildung und Technologie (BBT) für die stets vorbildliche Zusammenarbeit zu danken.

***

Zur Archivierung wissenschaftlicher Texte und verwandten Problemen

Hans Koch

Das Hauptziel dieses Beitrags ist zu zeigen, dass (und wie) Netzwerke von elektronischen Archiven in der Zukunft eine wichtige Rolle spielen könnten in der akademischen Infrastruktur und Forschung, und dass man solche Archive mit bescheidenen Mitteln betreiben kann. Letzteres wird anhand eines konkreten Beispiels illustriert.

Als Mathematiker bin ich daran interessiert, dass Forschungsresultate sich schnell und ungehindert verbreiten und auffinden lassen. Vor dreizehn Jahren wurden Preprints von neuen Artikeln nur an gezielte Gruppen verschickt, und jedermann benützte dazu seine eigene Adressenliste. Aber die Idee von elektronischen Preprint-Archiven war schon in der Luft. Ich diskutierte diese Idee mit meinen Kollegen Rafael de la Llave und Charles Radin in Austin, und wir entschieden uns, ein solches Archiv in Mathematischer Physik, unserem Forschungsgebiet, aufzubauen.

Am 10. Juli 1991, nach ungefähr sechs Wochen Planen, Programmieren, und Testen, konnten wir die Geburt des Archives mp_arc öffentlich bekannt geben und die ersten vier Preprints archivieren. Nach meinem Wissen war dies das erste vollautomatische Preprint-Archiv, sicher in den Gebieten Mathematik und Physik. Kollegen aus allen Ecken der Welt konnten innert Sekunden ihre Preprints bei uns via Email mit Kennwort "archive paper" deponieren. Andere Kennwörter konnten benutzt werden, um Preprints, Abstracts, Indexe, und andere Informationen herunterzuladen, oder sich in eine öffentlichen Adressenliste einzutragen. Aus dieser Adressenliste entstand ein Jahr später das Abonnentenverzeichnis für den wöchentlichen Email-Versand der neusten Abstracts. Die Anzahl der gespeicherten Artikel war um diese Zeit auch gross genug, dass es sich lohnte, eine Suchmaschine zu starten. Andere Marksteine in der Entwicklung von mp_arc waren der Anschluss ans Web im April 1994 und die Inbetriebnahme des ersten Mirror-Sites im November 1995 in Genf. Gegenwärtig enthält mp_arc ungefähr 3600 Artikel, und es kommen pro Jahr etwa 500 neue dazu.

Die intensive Benutzung von Preprint-Archiven zeigt deutlich, dass diese einen wertvollen Dienst leisten. Eines der nächsten Ziele besteht darin, all diese Archive so zu vernetzen dass das Resultat mehr ist als nur die Summe seiner Bestandteile. Aber bevor ich mich diesem Thema zuwende, möchte ich etwas näher beschreiben, was beim Betreiben eines einzelnen Archives wichtig ist.

Einer der Kernpunkte ist natürlich die Absicherung der Daten. Elektronische Datenträger sind viel empfindlicher als Bücher, und es ist möglich, innerhalb von Sekunden Gigabytes von Daten zu verlieren. Die beste Absicherung bieten Redundanz und periodische Kontrolle. Eine übliche erste Stufe ist die logische Zusammenfassung von Festplatten (RAID), was vor den Folgen eines einzigen Festplatten-Crashes weitgehend schützt. Für mp_arc werden zudem die wichtigsten Daten laufend auf einer anderen Maschine gesichert. Zusätzlich werden jede Nacht von unserem Institut Absicherungen auf Band gemacht, und die verschiedenen Mirror-Sites kopieren die neu dazugekommenen Daten auf ihre Fileserver. Mirror-Sites sind extrem wichtig für ein elektronisches Archiv. Sie ermöglichen beispielsweise, dass nach einem Katastrophenfall das Archiv relativ schnell und komplett rekonstruiert werden kann. Eine altbekannte Weisheit besagt, dass viele Kopien ein Dokument schützen. In dieser Beziehung sind digitale Dokumente oft besser gesichert als gedruckte. (Aus ähnlichem Grund sind offene Dokumente sicherer als kopiergeschützte, und interessante sicherer als uninteressante.)

Ein anderes Problem besteht darin, die Integrität der Daten in einem Archiv oder Mirror zu gewährleisten. Es genügt nicht, Kopien der ursprünglichen Daten aufzubewahren, für den Fall dass die zirkulierenden Daten sich als beschädigt oder gefälscht herausstellen. Speichermaterialien werden unzuverlässig - im Falle von Bändern oder CDs innerhalb von ein paar Jahrzehnten. Eine gute Lösung besteht darin, die Daten regelmässig mit den letzten Sicherheitskopien zu vergleichen. Falls dies zu aufwendig ist, kann man Ähnliches mit Checksummen erreichen. (Mit modernen Checksummen wie MD5 ist die Wahrscheinlichkeit, dass zwei verschiedene Dateien dieselbe Checksumme haben, im wesentlichen Null.) Dieser Prozess lässt sich einfach automatisieren.

Langfristig stellt sich auch das Problem, dass Datenformate aus der Mode kommen oder veralten, und dass dadurch Dokumente für viele unlesbar werden. Gewisse Firmen beschleunigen diesen Prozess sogar absichtlich, damit die Kunden von ihrer Software abhängig bleiben, mit dem Einführen und periodischen "Verbessern" von firmeneigenen Formaten. Ein Beispiel sind die .doc files, die in der Regel nicht einmal der Autor vollständig entschlüsseln kann. Eine Lösung besteht darin, Artikel nur in gängigen Formaten zu akzeptieren, die zudem öffentlich und vollständig dokumentiert sind. Letzteres erlaubt es, auch in ferner Zukunft Artikel-Varianten in dann gängigen Formaten herzustellen. Eine Dokument-Variante ist eine Kombination des originalen Dokumentes, oder eines präzisen Verweises auf dieses Dokument, mit einer (möglicherweise nicht inhaltsgetreuen) umformatierten Version des originalen Dokumentes. Das Umformatieren sollte nur sehr selten nötig sein und keine besonderen Schwierigkeiten bereiten, falls man sich auf weit verbreitete und offene Formate beschränkt.

Preprint-Archive müssen auch eine bescheidene Art von Qualitätskontrolle durchführen, um zu vermeiden, dass sie mit Plunder überschwemmt werden. Für ein Archiv, welches sich auf ein gewisses Forschungsgebiet konzentriert und mit einer Gruppe von Forschern in diesem Gebiet eng verbunden ist, ist es relativ einfach, ein gewünschtes Profil zu erhalten. Im Falle vom mp_arc werden Artikel zwar direkt nach dem Hochladen in das Archiv aufgenommen und sichtbar, aber als Sicherheitsmassnahme sind diverse Aufnahmefrequenzen absichtlich beschränkt. Da wir Betreiber das Archiv täglich in unserer Arbeit benützen, fallen uns Artikel schnell auf, welche für das beabsichtigte Zielpublikum klarerweise uninteressant sind. Solche Artikel werden dann entfernt, und der Autor erhält eine nette Erklärung. Da dies bekannt ist, müssen nur etwa 2 % aller zugesandten Artikel von uns entfernt werden; das heisst der Arbeitsaufwand ist minimal. Für Serientäter wird einfach die Aufnahmefrequenz auf Null beschränkt.

Das Erstellen von Webseiten kann völlig automatisiert werden und ist eher trivial, falls man sich auf das Wesentliche beschränkt. Auch Suchmaschinen sind einfach und billig zu betreiben, da es für diesen Zweck gute und kostenlose Software gibt. Grosse Archive beschränken das Absuchen normalerweise auf Metadaten (Autoren, Titel, Jahr, Abstract), da Volltextsuchen bedeutend mehr Ressourcen benötigen. Aber kleinere Archive wie mp_arc lasten sogar mit Volltextsuche nicht einmal einen einzigen PC aus. Dies ist einer der Vorteile eines Netzwerkes von kleinen oder mittelgrossen Archiven, welche an Orten betrieben werden, wo die Ressourcen ohnehin vorhanden sind.

Es ist erstaunlich, wie wenig benötigt wird, um ein solches Archiv zu betreiben. Die Ausgaben für Festplattenspeicher und Magnetbänder (ohne Wiederbenützung) waren bei mp_arc letztes Jahr etwa hundert Dollars. Der Betrag verringert sich jedes Jahr, obwohl die Anzahl der gespeicherten Artikel ständig anwächst. Das Budget für Software ist Null, da wir nur kostenlose Programme benützen, und diese mit simplen hausgemachten Skripten koordinieren.

Die Rechenzeit ist wie erwähnt vernachlässigbar, und der Webserver unseres Institutes verkraftet den zusätzlichen Web-Verkehr von mp_arc ohne Anstrengung. Unterhaltungsarbeit fällt auch kaum an; das System läuft typischerweise monatelang ohne Aufsicht.

Das einzige, was wirklich Arbeit verursacht, ist das Implementieren von neuen Technologien und Diensten. Dies addiert sich zu schätzungsweise einer Woche pro Jahr für eine Person, falls man der Versuchung widerstehen kann, jeder vorübergehenden Mode zu folgen. Zum Glück sind solche Aufgaben nicht dringend; das heisst man kann sich den günstigsten Zeitpunkt dafür auswählen. Natürlich verlangt solche Arbeit etwas Expertise. Aber es wäre natürlich möglich, die nötige Software in einem gemeinsamen Projekt zu entwickeln, an dem sich mehrere Archive beteiligen. In jedem Fall gibt es unter Wissenschaftlern mehr als genug Leute, die ein Archiv in ihrem Forschungsgebiet betreiben könnten, und deren Institute die nötige Hardware schon haben.

Wer sich aber mehr Luxus wünscht, kann Archivdienste auch kaufen: Die jährliche Lizenz für eine (hoch polierte) kommerzielle Lösung einer führenden Firma kostet eine vier bis sechs-stellige Summe, je nach Umfang des Projektes.

Wie erwähnt existiert heute schon eine beachtliche Anzahl von Archiven für wissenschaftliche Texte. Eine Webseite der Amerikanischen Mathematischen Gesellschaft listet zum Beispiel 17 subjektorientierte Archive im Gebiet der Mathematik auf, 64 institutionelle Archive, und 5 Dachserver. Die Informatik scheint auch eine Kollektion von diversen Archiven zu besitzen. Aber kulturelle Unterschiede zwischen den verschiedenen wissenschaftlichen Disziplinen sind auch hier sichtbar. In der Physik besteht allgemein eine Tendenz in Richtung grosser Dachorganisationen. Dementsprechend dominieren heutzutage in diesem Gebiet grosse Archive, wie das ArXiv an der Cornell Universität, oder der CERN Document Server in Genf, welche von staatlichen Zuschüssen reichlich unterstützt werden. Die Benutzung von elektronischen Archiven hat sich besonders schnell in der Hochenergiephysik durchgesetzt, was wohl damit zusammenhängt, dass die meisten Ideen in diesem Fachgebiet relativ kurzlebig sind. In anderen Wissenschaften, vor allem denjenigen mit engeren Beziehungen zur Industrie, hat die Archiv-Idee weniger schnell Fuss gefasst, und kommerzielle Variationen dieses Konzeptes dominieren zum Teil die Szene (ChemWeb, BioMed, BioOne, HighWire, BePress, ...).

Der Vorteil von Web-zugänglichen Textsammlungen gegenüber Reihen von gefüllten Bücherregalen ist deutlich: bequeme Suchmöglichkeiten, schneller Zugriff von fast überall, und das einfache Verbinden von verwandten Dokumenten. Dazu kommen die bedeutend niedrigeren Kosten für die Produktion, Lagerung, und den Transport. Deshalb wird es kaum lange gehen, bis Fachzeitschriften praktisch nur noch in der Online-Version erhältlich sind. Die gedruckte Version auch noch zu abonnieren ist schon heute ein Luxus, und dieser Luxus wird verschwinden, sobald ihn sich niemand mehr leisten kann. Auch ältere Zeitschriftenbände werden bald kaum mehr aus den Regalen geholt, da diese von Verlegern und anderen Vereinigungen (wie JSTOR) schon systematisch digitalisiert werden.

Es ist klar, dass die Forschungsbibliothek, wie wir sie heute kennen, früher oder später verschwinden wird. Um zu vermeiden, dass die Wissenschaft in diesem Wandel leidet, ist es nötig, dass alle betroffenen Gruppen sich an einer Diskussion über Richtlinien und konkrete Schritte beteiligen. Leider steht dafür nicht allzuviel Zeit zur Verfügung, da sich die Auswahlmöglichkeiten unter dem finanziellen Druck der Grossverleger ständig verringern.

Eine der bedeutendsten Kräfte auf "unserer" Seite ist die *Scholarly Publishing and Academic Resources Coalition (SPARC)*, eine Allianz von Bibliotheken in (zurzeit) ungefähr 200 Forschungsinstitutionen. Die SPARC fördert Initiativen, welche den Konkurrenzkampf im Verlagswesen erweitern, und setzt sich für offenen Zugang zu akademischen Texten ein. Die Mitglieder-Institutionen verpflichten sich, via finanzielle Zusagen gewisse Fachzeitschriften zu unterstützen, deren Verleger in ein Partnerschaftsverhältnis mit SPARC eingetreten sind. Andere Aktivitäten werden durch Mitgliederbeiträge finanziert. Einige ermutigende Fortschritte sind schon erzielt worden, und ich nehme an, dass davon in anderen Beiträgen zu diesem Bulletin die Rede sein wird.

Die SPARC unterstützt unter anderem die *Open Archives Initiative (OAI)*, welche Bibliothekare, Informatiker, Verleger, Wissenschaftler, und Leute aus Universitätsverwaltungen zusammenbringt. Die OAI arbeitet an den technischen Fundamenten für ein weltweites Repository von frei zugänglichen Forschungsartikeln. Dies beinhaltet unter anderem die Entwicklung von Protokollen für die Kommunikation zwischen Datenanbietern (offenen Archiven) und Dienstanbietern (Harvesters), Spezifikationen für Metadaten, und Richtlinien für gewisse Software-Komponenten.

Ein offenes Archiv im Sinne der AOI gewährt freien Zugriff auf Metadaten, aber nicht unbedingt auf den Volltext oder andere Daten. Die Hoffnung ist natürlich dass die meisten Archive auch den Volltext offen anbieten werden, und dass die restlichen dann zunehmend unter Druck geraten dasselbe zu tun. Derzeit sind etwa hundert Archive entsprechend ausgerüstet und registriert, und es existieren schon mehrere - zum Teil experimentelle - Harvesters (ARC, OAIster, Celestial, TORII, ...).

Was folgt ist eine Skizze der möglichen Aufgaben von Archiven und Harvesters im Zeitalter der digitalen Bibliotheken. Eine zusätzliche Funktion, das "Markieren" von Dokumenten, wird anschliessend erwähnt.

- Die Funktion eines Archives besteht darin, die hereinkommenden Dokumente mit einem Kennsatz und Datum zu versehen, sie sicher aufzubewahren, und mit anderen Servers zu kommunizieren. Dazu muss unter anderem für jedes Dokument ein Metadatensatz hergestellt werden, welcher den Kennsatz dieses Dokumentes enthält, sowie Information über den Inhalt. Diese Metadaten, zusammen mit Daten über das Archiv selbst, werden anderen Servern auf Abfrage zur Verfügung gestellt. Da digitale Daten sehr einfach zu fälschen sind, muss ein Archiv auch dazu benützt werden können, schnell zu verifizieren, ob ein gespeichertes Dokument mit einer sich im Umlauf befindenden Kopie übereinstimmt. Zu diesem Zweck könnte zum Beispiel eine Checksumme in einer abrufbaren Metadaten-Komponente gespeichert werden. Eine andere Aufgabe, welche gewisse Archive erfüllen sollten, ist das Herstellen von Dokument-Varianten in gängigen Formaten. Kennsätze könnten zu diesem Zweck mit einem geeigneten Postfix versehen werden.
- Um sicherzugehen, dass ein gegebener Dokument-Kennsatz weltweit nur einmal benützt wird, sollte dieser unter anderem einen Archiv-Kennsatz enthalten, der bei einer beaufsichtigenden Agentur registriert worden ist. (Zwei Standardisierungsvorschläge für Dokument-Kennsätze sind der DOI and PII). Für eine gewisse Zeit hilft dies auch beim Auffinden eines Dokumentes. Um dies aber auch langfristig zu ermöglichen, sollte ein Verweis auf ein Dokument nicht nur dessen Kennsatz angeben, sondern noch zusätzliche Information über den Inhalt, entweder explizit oder via den Kontext, innerhalb dessen das Dokument zitiert wird. Für gewisse Dokumente werden entsprechende Ortungsmechanismen schon entwickelt und benutzt (OpenURL).
- Ein Harvester sammelt ausgewählte Daten von Archiven (oder nur Metadaten, wie zurzeit in den OAI Protokollen vorgesehen ist) und erhöht den Wert dieser Sammlung durch das Bereitstellen von Hilfsmitteln zum Absuchen, Zugreifen, oder Filtern dieser Daten. Dies sind im wesentlichen die minimalen Dienste einer digitalen Bibliothek. Als zusätzlicher Dienst könnten nützliche Indexlisten erstellt werden, möglicherweise mit Anmerkungen, in welchen Dokumentverweise nach gewissen Themen zusammengefasst sind. Diese Listen sind selber wieder Dokumente, die wertvoll genug sein können um in einem Archiv aufzubewahrt zu werden. Unter anderem könnten solche Indexlisten von speziellen Ortungs-Servern benützt werden, um ein gewünschtes Dokument zu finden.
- Eine andere Art von kombinierten Dokumenten erhält man, indem man einen Artikel (oder nur einen Verweis darauf) mit einer Marke oder Urkunde verbindet.
- Eine wissenschaftliche Fachzeitschrift ist im wesentlichen eine Sammlung solcher Marken-Artikel. Die Urkunde bescheinigt, dass der Artikel einen gewissen Bewertungsprozess erfolgreich bestanden hat. Bei einer elektronischen Fachzeitschrift spielt sich damit der ganze Prozess bis zum publizierbaren Artikel innerhalb der Wissenschaftsgemeinde ab. Der Vertrieb kann einfach von Archiven und Harvestern übernommen werden, und diese könnten zum Beispiel von wissenschaftlichen Institutionen betrieben werden, im selben Rahmen wie heute unsere Bibliotheken. Dann bleiben nur noch die Verwaltungskosten, welche möglicherweise beim Bewerten und Revidieren eines Artikels anfallen. Es scheint sinnvoll, dass diese vom Auftraggeber bezahlt werden, das heisst typischerweise von den Autoren oder deren Institutionen. Damit wird es möglich, wissenschaftliche Fachzeitschriften kostenlos anzubieten, wie es von den Autoren auch gewünscht wird.

  Ein solches Modell ist nicht nur möglich, wie hunderte von schon existierenden Beispielen zeigen (siehe DOAJ.org), sondern auf lange Sicht unvermeidlich. Was dem Prozess im Wege steht, ist die Tatsache, dass die Verleger von prestigeträchtigen Fachzeitschriften heute vom Konkurrenzkampf innerhalb der Wissenschaftsgemeinde prächtig profitieren können. Die Wissenschaftsgemeinde zahlt Riesensummen für Lizenzen, nur um Zugriff auf Artikel zu erhalten, welche ihre Mitglieder vorher an Aussenstehende verschenkt haben. Und im Prozess ruinieren wir unsere eigenen Bibliotheken. Diese wehren sich so gut sie können, aber gezwungenermassen eher mit kurzfristigen Notlösungen (Konsortien, LOCKSS, ...). Was wir brauchen, sind Schritte, welche unseren Forschungsartikeln langsam den finanziellen Wert entziehen. Dies heisst, dass wir direkt oder indirekt von Copyrights wegkommen müssen welche es uns verbieten, publizierte Artikel kostenlos in Archiven der Öffentlichkeit zugänglich zu machen.