



Die erste Seite der ersten Nummer der «Zürcher Zeitung» in maschinenlesbarer Form. Bis Mitte 2005 sollen alle 225 Jahre der NZZ für die elektronische Volltextsuche erschlossen sein. (Bild pd)

Zeitreisen mit der Suchmaschine

225 Jahrgänge der NZZ werden elektronisch erschlossen

Rund 1500 Mikrofilmrollen wurden vor einigen Monaten an das Fraunhofer-Institut für Medienkommunikation in Sankt Augustin geliefert. Das Fachwissen dieses Instituts, das sich auf Forschung und Entwicklung im Bereich der neuen digitalen Medien spezialisiert hat, kommt für einmal einem «alten» Medium zugute: Die Computerwissenschaftler werden die rund zwei Millionen NZZ-Seiten, deren Bilder auf den Filmrollen abgebildet sind, digitalisieren. Im Sommer soll die Arbeit beendet sein.

S. B. Das beigefarbene Buch sieht unscheinbar aus, doch hinter abgegriffenen, schiefen Deckeln lauert Leben. Gleich auf der ersten Seite geht es zur Sache: Graf d'Estaing, Vizeadmiral der französischen Flotte, erreicht mit 20 Kriegsschiffen und rund 3000 Soldaten die amerikanische Ostküste und beginnt mit der Belagerung der Stadt Savannah, wo sich britische Soldaten verschanzt haben. Zusammen mit den amerikanischen Verbündeten wagen die Franzosen den Angriff, Graf d'Estaing stellt sich an die Spitze seiner Truppen. Der Vorstoss wird abgewehrt, eine Schussverletzung zertrümmert dem Grafen das Knie, später zwingen aufkommende Herbststürme die Schiffe zur Heimkehr nach Frankreich.

Das unscheinbare Buch, in beigefarbenes Leder gebunden, lässt von aussen nicht erkennen, was es in sich birgt, nur gerade die vier Zahlen «1780» auf dem Buchrücken verweisen auf den Inhalt. Es liegt überraschend leicht in der Hand, das Papier fühlt sich weich an, fast wie Stoff. Und doch erweckt es Ehrfurcht: Das ist das Original, so also sah die «Neue Zürcher Zeitung» aus, als die Zeiten noch alt waren.

Grobmaschiges Netz

Die NZZ ist eine der ältesten Tageszeitungen der Welt, das ist bekannt. Weniger bekannt dürfte sein, dass die NZZ – vermutlich – die Zeitung mit den ältesten Findmitteln ist. Als Findmittel bezeichnen Archivare Werkzeuge für die inhaltliche Erschliessung von Information. Bei der Suche nach NZZ-Artikeln, die nach 1871 verfasst wurden, helfen Registerbände. Diese voluminösen Bücher decken jeweils nur ein paar Monate ab. Später kamen Karteikarten hinzu, die Artikel zu bestimmten Stichworten über die Buchdeckel der Registerbände hinaus auflisten. Schliesslich gibt es zu wichtigen Themen Dossiers, das sind Mäppchen, die ausgeschnittene Artikel versammeln. Die Stichwörter und Oberthemen legen ein grobmaschiges Netz über den Inhalt der Zeitung; wozu den Archivaren vergangener Zeiten kein Stichwort einfiel, das verschwindet im Dunkel der Geschichte. Zum Beispiel Sport: «Die NZZ berichtete schon sehr früh über Sport», erzählt Ruth Haener, Leiterin des NZZ-Archivs, «weil aber körperliche Betätigung in diesem hochgeistigen Hause lange gering geschätzt wurde, kommt die Frühgeschichte des Sports in den alten Registerbänden und Karteikärtchen kaum vor.»

Seit 1993 sind alle in der NZZ erschienenen Texte elektronisch archiviert. Immer wieder wurde der Wunsch geäussert, auch ältere Ausgaben, vielleicht auch gleich alle Ausgaben der NZZ zu digitalisieren. Doch noch im Jahr 2001 musste dieses Vorhaben aus Kostengründen ver-

worfen werden; es hätte mehrere Millionen Franken gekostet. Im Hinblick auf das 225-Jahr-Jubiläum wurde das Projekt erneut geprüft. Und siehe da: Der technische Fortschritt erlaubte Kosteneinsparungen. Das Fraunhofer-Institut für Medienkommunikation (IMK) in Sankt Augustin übernahm den Auftrag für 600 000 Euro.

2 000 000 Seiten

Die Kriegstaten des Grafen d'Estaing im Rahmen des amerikanischen Unabhängigkeitskrieges, der Untergang der «Titanic», ein Schiffsunglück auf dem Zürichsee, die Landung der Alliierten in der Normandie, die Mondlandung: Auf mehr als zwei Millionen Seiten reflektierte die NZZ während der vergangenen 225 Jahre das Weltgeschehen. In der zweiten Hälfte des 20. Jahrhunderts wurde begonnen, diese Seiten auf Mikrofilm zu kopieren. Rund 1500 Rollen 35-Millimeter-Film wurden belichtet.

Allerdings wurden die Seiten nicht immer in derselben Art und Weise abfotografiert. Bis 1960 wurden die Zeitungssseiten doppelseitig, als Buch gebunden, aufgenommen. Dadurch ergaben sich Verzerrungen. Manchmal kommt auch noch der Tisch, der das Buch trug, ins Bild. In jüngerer Zeit wurden die einzelnen Zeitungssseiten unter eine Glasplatte gelegt und vor einem schwarzen Hintergrund abgelichtet. Der Abstand der Kamera variiert, manchmal kommen die Zeitungsränder ausserhalb des Aufnahme Fensters zu liegen.

Bildpunkte zu Buchstaben

Bei der Digitalisierung werden in einem ersten Schritt die Filme eingescannt und in Bilddateien umgewandelt. Dann werden die Seitenränder ausfindig gemacht. Das ist – wegen der verschiedenen Methoden, mit denen die Seiten aufgenommen wurden – nicht immer einfach. Mit Hilfe von selbst entwickelter Software werden beim Fraunhofer-IMK Verzerrungen und Unschärfen entfernt. Dann gilt es in der Abfolge der Bilder, die von einem Film gewonnen wurden, die Titelseiten aufzuspüren, damit die Seiten zeitlich eingeordnet werden können. Schliesslich werden die Bilder im Tiff-Format gespeichert. Das Tiff-Format wurde unter anderem deshalb gewählt, weil es dank guter Dokumentation und weiter Verbreitung als De-facto-Standard gilt und mit grosser Wahrscheinlichkeit auch noch in ferner Zukunft gelesen werden kann. Auch benötigt es wenig Platz und erlaubt die Einbettung von Metadaten.

Der letzte wichtigste Verarbeitungsschritt ist die Verwandlung von Bildpunkten in Buchstaben. Dazu wird das von der russischen Software-Firma Abby entwickelte Texterkennungsprogramm (Optical Character Recognition, OCR) Finereader eingesetzt. Dieses Programm muss auch mit Frakturschriften zurechtkommen, die bei der NZZ bis 1946 verwendet wurden. Die Erkennungsgenauigkeit ist gemäss Stefan Eickeler, der beim IMK als Projektleiter die Digitalisierung der NZZ betreut, sehr hoch. Allerdings gebe es Seiten, bei denen Flecken die Erkennung erschweren; grosse und zum Teil noch ungelöste Schwierigkeiten bereiten Passagen, bei denen Fraktur- und Antiquaschriften gemischt vorkommen. Falsch erkannte Wörter könnten jedoch automatisch, mit Hilfe eines Wörterbuchs, korrigiert werden. Auch beeinträchtigten einzelne Fehler im elektronischen Text das Suchresultat kaum, so dass eine 100-prozentige Erkennungsgenauigkeit nicht angestrebt werden müsste.

Für die Digitalisierung verwendet das IMK einen hybriden Cluster mit 20 Rechnern unter Windows und Linux. Resultat der Arbeit ist eine XML-Datei, die mit dem Text auch noch Metadaten abspeichert, die einzelnen Absätzen Titel zuordnen und Auskunft geben können über typo-

Selbstverwaltung des Wissens

Online-Nachschlagewerk Wikipedia als Nachrichtendienst

Jimmy Wales hat etwas von einem Religionsgründer: Der bald 40-jährige Amerikaner mit Vollbart wirkt verständnisvoll bis gütig, wenn er mit seiner Gemeinde redet. Er wird nicht laut und hört sich geduldig Argumente von allen Seiten an, hat aber seine eigenen Auffassungen. «Jimbo», wie ihn seine Freunde nennen, ist einer der Gründer der Online-Enzyklopädie Wikipedia,¹ die am Samstag ihren vierten Geburtstag feiert. Wichtigstes Ziel des in Florida lebenden Netzbürgers ist es, gemeinsam mit Tausenden von freiwilligen Helfern die grösste und beste Wissenssammlung der Welt aufzubauen und das darin gespeicherte kollektive Wissen der Menschheit allen frei zugänglich zu machen.

Britannica und Brockhaus

Quantitativ hat Wales schon viel erreicht: «Wir haben über 1,2 Millionen Artikel in mehr als 200 Sprachen», sagt der Philanthrop. Insgesamt enthalte die exponentiell wachsende Wissensdatenbank über 130 Millionen Wörter. «Das sind mehr als in der Britannica und im Brockhaus zusammen», zieht der Präsident der hinter Wikipedia stehenden Stiftung Wikimedia einen Vergleich zu den wichtigsten Konkurrenten. Die Wikipedia-Seiten, die in Englisch derzeit 412 000 und in Deutsch 172 000 Einträge umfassen, sind im Web populärer als etwa die Homepage des US-Senders Fox News.

Das Erfolgsrezept hört auf den Namen «Open Content»: So wie bei Open-Source-Programmen zahlreiche Entwickler gemeinsam den Quellcode bearbeiten und Fehler ausbügeln, erstellen und verbessern die Wikipedia-Autoren und -Nutzer die Beiträge kooperativ in einer Online-Datenbank. Jeder kann bei dem zum Einsatz kommenden Basisprinzip einer Wiki-Software durch Druck auf einen «Edit»-Knopf Inhalte bearbeiten. Letztlich entscheidet aber laut Wales ein besonders aktiver Kreis von rund 2000 Nutzern, die bereits jeweils mehr als 100 000 Änderungen an Einträgen vorgenommen haben, über die Brauchbarkeit eines Enzyklopädie-Artikels. «Es ist eine verwirrende, aber praktikable Mischung aus Konsens, Demokratie, Aristokratie und Monarchie, die Wikipedia am Laufen hält», erläutert Wales das Prinzip. – Das ausgemachte Kollaborationsrezept will Wales nun auch auf die Nachrichtenpro-

duktion übertragen. Wikimedia² heisst das jüngste Kind der Gemeinschaft. Hier sollen tagesaktuelle Nachrichten durch eine Kooperation von Amateuren entstehen. Es handle sich um ein Experiment, schwächt Wales allzu hohe Erwartungen ab. Eine Art Abfallprodukt zur Verwertung der Energie, die bei Wikipedia aufgrund ihres zu ereignisbezogenen Charakters nicht berücksichtigt werden könne. Doch der Anspruch ist gross: Die Wikimedianer sollen laut ihrem Propheten die Voreingenommenheit eliminieren, die sich auch in Qualitätsblätter einschleicht.

Kritik wegen Unzuverlässigkeit

Wikimedia startet allerdings in einer Zeit, in der sich die Wikipedianer verstärkt mit dem Vorwurf der Unzuverlässigkeit auseinandersetzen müssen. Amerikanische Zeitungen verbreiten Warnungen von Bibliothekaren, gemäss denen es der Online-Enzyklopädie an der fachlichen Qualität der Darstellung mangle. Einige machten die Probe aufs Exempel und schmuggelten subtile Falschinformationen in vormals korrekte Beiträge. Nach einer Woche standen viele davon noch in der Datenbank. Jüngst meldete sich dann der Wikipedia-Mitbegründer, Larry Sanger, mit einem viel diskutierten Online-Beitrag kritisch zu Wort. Er glaubt, dass die anti-elitäre Haltung von Wales und seinem engsten Kreis die Missachtung von ausgewiesenen Experten der Wissensproduktion fördere und einen kritischen Umgang mit den publizierten Informationen verhindere.

Wales hält die Einführung eines Mechanismus zur Qualitätssicherung, der einzelnen Redaktoren mehr Macht verschaffen würde, jedoch «für unnötig, ja nicht einmal erstrebenswert». Alles müsse allein von der Basis her kommen. Es bedürfe keiner «neuen Erfindungen», erklärte er jüngst an einem Entwicklertreffen in Berlin, «damit wir eine Qualität erreichen, die traditionelle Publikationsmodelle möglicherweise sogar übertrifft». Seine Jünger halten bisher zu ihm. Es bedürfe schon einer Art Umsturz, um gewöhnliche Redaktionsprozesse einzuführen, glaubt auch Sanger. Aber vielleicht frisst die Wikipedia-Revolution ja doch noch ihre Kinder. *Stefan Krempf*

¹ wikipedia.org
² wikimedia.org

grafische Merkmale von Wörtern und ihre Position auf der Seite. Eine einzelne Seite benötigt in digitalisierter Form 4 MByte, das vollständige elektronische Archiv wird 10 TByte umfassen. Dafür musste im Intranet der NZZ durch die Anschaffung von grossen EMC-Speichersystemen Platz geschaffen werden; die Bereitstellung dieses Speicherplatzes kostete rund 300 000 Franken.

Ein Traum geht in Erfüllung

«Die grössten Herausforderungen», so berichtet Projektleiter Eickeler, «sind die Variationen bei der Mikroverfilmung und die grosse Anzahl der Zeitungssseiten. Es mussten sehr robuste Verfahren entwickelt werden, um die Extraktion der Zeitungssseiten bei den unterschiedlichen Bedingungen bei der Mikroverfilmung durchführen zu können. Aufgrund der Menge der Zeitungssseiten sind manuelle Korrekturen nicht möglich.» Diese Arbeit sei für Computerwissenschaftler «sehr interessant», denn nur bei umfangreichen Datenbeständen könnten «ausgereifte intelligente Verfahren» entwickelt werden. Die Mikrofilme mit den alten NZZ-Seiten wurden den Spezialisten des IMK im Frühsommer 2004 geliefert, im September 2005 soll die Arbeit der Digitalisierung abgeschlossen sein.

Für die Hüterin des beigefarbenen Buches, Ruth Haener, geht mit der Digitalisierung aller NZZ-Jahrgänge «ein Traum in Erfüllung»: «Die Jahre 1780 bis 1871 waren bisher nicht erschlossen, d. h., dass wir bei Anfragen diese Zeit betreffend die Zeitung durchblättern mussten. Mit der neuen Technik erübrigt sich diese ineffiziente Arbeitsweise. Wir sind auch nicht mehr auf das Gedächtnis einzelner Köpfe angewiesen, die wissen, wie früher Information erschlossen worden ist. Das macht uns freier. Die Information kann fliessen.» Heisst das, dass, wo es die Möglichkeit der Volltextsuche gibt, Archivare nicht mehr gefragt sind? «Die klassische Erschliessungsarbeit am Text ist schon heute nur noch ein verhältnismässig kleiner Teil unserer Arbeit. Gefordert sind wir etwa bei der Erschliessung von Bildern oder Filmen. Der Beruf ist für mich vielseitiger und interessanter geworden.»

Erholungskur im Entsäuerungsbad

Das beigefarbene Buch, in dem die ersten Ausgaben der NZZ aufgehoben sind, hat in jüngster Vergangenheit gelitten. Die erste Seite ist stark abgegriffen, unten, wo von den Kriegstaten des Grafen d'Estaing die Rede ist, sogar leicht eingerissen. Der 1780er Band war jüngst im Zusammenhang mit dem 225-Jahr-Jubiläum stark begehrt, viele Leute wollten mit eigenen Augen sehen, wie diese Zeitung in ihren Anfängen ausgesehen hat. Doch jetzt wird der Band restauriert, säurefrei verpackt und dunkel und bei geregelten klimatischen Bedingungen gelagert. In ein paar Jahren soll er ins Entsäuerungsbad. Das «Büchlein», das so viele Schicksale schildert und selber so viel erlebt hat, darf nun – dank der Computertechnik – zur Ruhe kommen.

Kurzmeldungen

Hersant-Gruppe lanciert Schweizer TV-Beilage. Das französische Medienhaus Hersant gibt seit kurzem in der Westschweiz eine neue TV-Beilage heraus. «TV Plus» wird den drei Zeitungen «La Côte», «L'Impartial» und «L'Express» beigelegt. *(sda)*

SAP als SwissT.net. Der Schweizer Automatiz-Pool (SAP) positioniert sich neu als Swiss Technology Network (swissT.net). Der 1976 gegründete Wirtschaftsverband vertritt rund 400 Mitgliedfirmen und befasst sich mit politischen und wirtschaftlichen Aspekten von Automation, Elektronik, Informatik, Medizin- und Energietechnik. Der Auftritt unter neuem Namen und mit neuer Corporate Identity erfolgt ab Februar. *S. B.*

Anzeige



Jano Berni (37), Kundenberater, bestellt seit 29. Oktober 2004 online. Er schrieb per E-Mail:

Letzte Woche habe ich Ihren Online-Shop zum ersten Mal benutzt. Gefreut hat mich Ihre Homepage, welche gut strukturiert und deswegen sehr benutzerfreundlich ist. Ebenso erfreulich war die pünktliche Auslieferung. Die Sahne auf den Erdbeeren war, dass der gelieferte Wein bereits gekühlt war und somit für das Abendessen mit Gästen perfekt verwendet werden konnte. Kompliment auf der ganzen Linie.



www.coop.ch **coop**
Online bestellt – nach Hause gebracht

Geld von der US-Regierung

Doppelrolle eines Kommentators

ras. Ein bekannter amerikanischer Radio- und Zeitungskommentator, Armstrong Williams, hat von der Regierung 240 000 Dollar erhalten. Dafür musste er die Bildungsinitiative «No Child Left Behind» unterstützen und promoten. Dies meldete kürzlich die Zeitung «USA Today», für welche Williams ebenfalls Kommentare schrieb. Gegenüber der Zeitung sagte Williams, er habe die Initiative unterstützen wollen, weil er von ihr überzeugt sei. Nicht zum ersten Mal versucht die US-Regierung, sich ins Mediensystem einzuschleichen. So liess sie im vergangenen Jahr Videofilme in der Machart einschlägiger journalistischer Berichte herstellen, um für die eigene Politik zu werben. Die Filme wurden den Fernsehstationen gratis zur Verfügung gestellt.