Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

THORSTEN KOCH

Verteilter Dokumentenspeicher

—

Erfahrungen mit den Kluwer-Daten des Friedrich-Althoff-Konsortiums

Unter Kennzeichen 08C5929 vom BMBF gefördert.

ZIB-Report 03-50 (Dezember 2003)

Verteilter Dokumentenspeicher*

—

Erfahrungen mit den Kluwer-Daten des Friedrich-Althoff-Konsortiums

Thorsten Koch

20. Dezember 2003

Zusammenfassung

Dieser Bericht beschreibt die Erfahrungen, die bei der Speicherung von 240.000 als PDF Dateien vorliegenden Artikeln des Kluwer-Verlages gewonnen wurden. Darüber hinaus werden einige Überlegungen zu Metadatenextraktion und Indizierung aufgezeigt.

Inhaltsverzeichnis

*Das Vorhaben wurde unter dem Kennzeichen 08C5929 vom BMBF gefördert.

1 Einleitung

Dieser Bericht beschreibt die Erfahrungen sowie einige Schlußfolgerungen, die im Rahmen der VDS-Vorstudie bei der Speicherung der vom Friedrich-Althoff-Konsortium lizenzierten Zeitschriften des Kluwer-Verlages gewonnen wurden.

Bei der verwendeten Hardware handelt es sich um einen Dell PowerEdge 2600 Server, bestückt mit zwei hyperthreadingfähigen 2,8 GHz Intel Xenon CPU's, 4 GB Hauptspeicher und insgesamt acht 140 GB Festplatten. Von diesen wurden sechs zu einem RAID-5 mit 628 GB Kapazität verbunden, das zur Speicherung der Daten eingesetzt wurde.

In Abschnitt 2 werden einige grundlegende Überlegungen zum Thema Metadaten aufgezeigt. Abschnitt 3 beschreibt die Erfahrungen mit den Kluwer-Daten. In Abschnitt 4 werden Gründe für eine Verwendung des Dateisystems zur Speicherung von Artikeln gezeigt. Abschließend werden in Abschnitt 5 noch die wesentlichen Erkenntnisse zusammengefaßt.

Im weiteren werden folgende Begriffe verwendet:

Ebene-1 Metadaten: (bei Artikeln) Autor, Titel, Journal, Volume, Ausgabe, Jahr, Seiten, usw.

Ebene-2 Metadaten: Im Artikel enthaltene Verweise auf andere Artikel (Literaturreferenzen) und Materialien.

Die Ebene-2 Metadaten sind von besonderer Bedeutung, weil erst sie es erlauben, eine Vernetzung der Artikel durchzuführen. Dies bildet neben der chronologischen Ordnung von Journalen ein zweites sehr wichtiges Ordnungskriterium.

2 Das Metadatenproblem

Die im Rahmen des Projekts zu speichernden Artikel werden zumindest anfänglich ausschließlich als Dateien in Adobes *Portable Document Format*[1] (PDF) vorliegen, da sich dieses Format bei den Verlagen zur Speicherung von Artikeln weitgehend durchgesetzt hat.

Obwohl sich Metadaten teilweise[2] in PDF Dateien einbetten lassen, ist dies in der Regel bei den im Projekt vorliegenden Artikeln nicht der Fall.

Leider hat sich bisher kein Standard für die Speicherung von Metadaten etabliert. Dies liegt zum Teil auch daran, daß die Metadaten bei den Verlagen als Teil des Dokumentenmanagements abfallen. Da hier sehr unterschiedliche Systeme im Einsatz sind, gibt es eine große Auswahl an verschiedenen

[1] `http://partners.adobe.com/asn/acrobat/sdk/public/docs/PDFReference15_v5.pdf`

[2] Die genauen Fähigkeiten hängen von der PDF Version ab und nehmen von Version zu Version zu.

Formaten für die Metadatenspeicherung, z.B. OA3, OX1, RDF um nur einige zu nennen.

Nachfolgend werden mehrere Herangehensweisen zur Lösung des Problems aufgezeigt und kurz ihre jeweiligen Vor- und Nachteile beschrieben:

2.1 Manuelle Erfassung

Die klassische Methode; die benötigten Metadaten werden manuell abgeschrieben.

2.1.1 Vorteile:

- Diese Methode wird schon seit mehr als hundert Jahren praktiziert und ist daher sehr gut verstanden. Alle Dokumente können problemlos erfaßt werden.
- Die notwendige Qualifikation der Durchführenden ist verhältnismäßig niedrig, und in den Bibliotheken ist eine große Zahl geschulter Mitarbeiter vorhanden.
- Mit systematischen Fehlern ist nicht zu rechnen.

2.1.2 Nachteile:

- Der geplante Umfang der Speicherung umfaßt mehrere Millionen Artikel, entsprechend groß ist der Aufwand.
- Der Aufwand läßt es aussichtslos erscheinen, auch Metadaten zweiten Grades zu erfassen.
- Da die entstehenden Fehler nicht systematisch sind, ist es schwer bzw. aufwendig, sie zu finden.

2.2 Import von Verlagsdaten

Angenommen:

1. Alle Verlage können Metadaten in irgendeinem Format liefern, und
2. eine festgelegte Untermenge von Feldern ist immer enthalten,

dann wäre es technisch möglich, die von den Verlagen gelieferten Metadaten automatisch in ein gemeinsames Format zu konvertieren.

2.2.1 Vorteile:

- Der Vorgang ist vollautomatisch.

2.2.2 Nachteile:

- Annahme (1) und (2) sind leider unzutreffend. Einige Verlage liefern keine Metadaten. Selbst bei gemeinsamen Feldern wie z.B. *Autor* gibt es keine einheitliche Schreibweise für den Inhalt (Vorname Name oder Name, Vorname oder nur Initialen bei Vornamen usw.)

- Die schiere Vielzahl der Verlage und der von ihnen eingesetzten Datenformate, hat einen erheblichen personellen Aufwand zur Folge, sowohl beim Erstellen der Konverter, bei deren Anwendung und Pflege, als auch bei der manuellen Nachbearbeitung von Problemfällen.

- Es gibt keinerlei Kontrolle über die Qualität der von den Verlagen gelieferten Daten.

- Das Verfahren führt dazu, daß nur die kleinste gemeinsame Teilmenge von Feldern verwendet werden kann. Dies schließt Ebene-2 Metadaten aus.

2.3 Extraktion aus den Artikeln

Es liegt in der Natur der Sache, daß viele Metadaten in den Artikeln selbst, sprich den PDF Dateien implizit enthalten sind. Daher besteht prinzipiell die Möglichkeit diese programmgesteuert aus den Dateien zu extrahieren.

Allerdings ist PDF ein Containerformat, d.h. die Tatsache, daß eine Datei im PDF Format vorliegt, erlaubt kaum eine Aussage über die Art der Speicherung.

- Dateien, die nur Standardschriften benutzen, sind üblicherweise in Text wandelbar, da die Kodierung der Standardschriften bekannt ist.

- Dateien mit nicht eingebetteten Schriften sind normalerweise gut in Text umzuwandeln, können aber, sofern man nicht im Besitz der entsprechenden Schriften ist, nicht korrekt angezeigt werden. Zur Archivierung ungeeignet.

- Dateien mit eingebetteten Schriften können nicht wieder in Text gewandelt werden, falls eine unbekannte Kodierung der Schriften vorliegt. Einzige Möglichkeit ist die Umwandlung in Rasterdaten, s.u. Sind die Schriften als Rasterdaten eingebettet, sind darüber hinaus die Möglichkeiten zur auflösungsunabhängigen Anzeige stark eingeschränkt.

- Rasterdaten (z.B. retrodigitalisiertes Material) sind nur mit Hilfe einer OCR (Optical Character Recognition) Software wieder in Text umwandelbar.

Grundsätzlich besteht eine sehr hohe Gefahr, bei der Rückumwandlung in Text Akzente und dergleichen zu verlieren, wenn diese nicht als eigenständiges Zeichen in der Schrift vorhanden sind.[3] Wie sich gezeigt hat, sind systematische Fehler bei der Rückumwandlung in Text möglich.

2.3.1 Vorteile:

- Der Vorgang ist vollautomatisch und liefert alle gewünschten Daten.

2.3.2 Nachteile:

- Derzeit ist die praktische Umsetzung noch ein offenes Forschungsgebiet. Einige Projekte in Teilbereichen waren erfolgreich, es ist aber nach wie vor unklar, in welchem Umfang Vollständigkeit bei der Erfassung der Daten erzielt werden kann.
- Der notwendige programmiertechnische Aufwand ist erheblich.

2.4 Verzicht auf Metadaten

Die vorhergehende Aufstellung zeigt, daß die Erfassung der Metadaten immer mit erheblichem Aufwand verbunden ist. Daher stellt sich die Frage:

Wozu brauchen wir die Metadaten und ist das den Aufwand wert?

oder, um es etwas weniger provokant zu formulieren:

Können wir unsere Ziele auch ohne Metadaten verwirklichen?

Eine Idee ist hierbei, eine Volltextindizierung der Texte vorzunehmen und dann darauf die Suche durchzuführen.

2.4.1 Vorteile:

- Der Vorgang ist vollautomatisch und mit verhältnismäßig geringem Aufwand durchführbar.

2.4.2 Nachteile:

- Das Fehlen von Metadaten erschwert die Einordnung der Suchergebnisse in eine Rangfolge (Ranking).
- Die vorliegenden Daten im PDF Format sind wesentlich unstrukturierter als HTML Daten, da es sich nur noch um reinen Text handelt und keine Markierungen mehr für Titel oder Verweise vorliegen.
- Die Vollständigkeit der Erfassung ist derzeit fraglich und schwer zu überprüfen.

[3] Komplexere Objekte, wie z.B. mathematische Formeln können nur unter immensem Aufwand, wenn überhaupt, wieder in eine Textform gebracht werden.

2.5 Schlußfolgerungen

Der vermutlich erfolgversprechendste Ansatz ist:

- Die Arbeit auf möglichst viele teilnehmende Institutionen zu verteilen und
- die Metadatenformate, für die große Volumina an Artikeln elektronisch vorliegen, automatisch zu konvertieren.
- Parallel dazu führt man eine Volltextindizierung unter Verwendung der Metadaten durch, was ein qualitativ hochwertiges Ranking erlaubt.

Man sollte aber bedenken, das dies einen beträchtlichen personellen und auch koordinatorischen Aufwand bedeutet.

3 Die Kluwer-Daten

Die Zeitschriften des Kluwer-Verlages liegen als PDF Dateien vor. Zusätzlich gibt es zu jedem Artikel noch eine Datei mit Ebene-1 Metadaten in einem verlagseigenen SGML Format[4].

Die initiale Übertragung der Daten erfolgte auf 15 DVD's. Nachfolgend wurde jeweils vom Verlag eine E-Mail Benachrichtigung geschickt, wenn weitere Artikel vorlagen. Diese konnten dann zusammengepackt als Datei im ZIP Format[5] innerhalb eines Monats vom File Server des Verlages per FTP abgeholt werden.

Dieses Verfahren hat zufriedenstellend funktioniert, es haben sich aber grundsätzliche Probleme gezeigt, auf die im weiteren detailliert eingegangen wird.

3.1 Datenvolumen

Währendes des Vorprojekts wurden ca. 240.000 Artikel gespeichert. Die Artikel verteilen sich dabei auf 775 Zeitschriften. Der Platzbedarf inklusive Ebene-1 Metadaten beträgt 74 GB. Die benötigte Zeit um alle Artikel einzulesen beträgt etwa 40 min.

Wie man in Tabelle 1 sehen kann, ist der Hauptteil der Artikel zwischen 1997 und heute angesiedelt, wobei aber nicht alle Zeitschriften in allen Jahrgängen vertreten sind.

Auffällig ist das starke Absinken der durchschnittlichen Artikelgröße ab 1997. Dies hat einen einfachen Grund; der Anteil der retrodigitalisierten Artikel nimmt ab, je neuer die Artikel sind. Da retrodigitalisierte (gescannte)

[4] Kluwer Academic Publishers Oasis Version 3

[5] `http://www.pkware.com/products/enterprise/white_papers/appnote.html`

Jahr	Zeit-schriften	Gesamt Artikel	Gesamt MB	ϕ MB/ Artikel	Artikel/ Zeitschrift
1990	1	19	32	1.7	
1991	1	28	43	1.5	
1992	2	56	63	1.1	
1993	2	73	110	1.5	
1994	2	71	92	1.3	
1995	2	63	92	1.5	
1996	1	23	16	0.7	
1997	439	24836	11108	0.4	57
1998	510	29986	12605	0.4	59
1999	422	22974	7594	0.3	54
2000	695	37317	8000	0.2	54
2001	694	47612	11387	0.2	69
2002	700	44912	13313	0.3	64
2003	609	28712	8564	0.3	47

Tabelle 1: Verteilung der Artikel über die Jahrgänge

Dokumente mehr Platz benötigen, als solche die originär elektronisch vorliegen, nimmt der durchschnittliche Platzbedarf auch mit den Jahren entsprechend ab. Die Verteilung der Artikelgrößen ist aus Tabelle 2 ersichtlich.

Durchschnitt	300	kB
Median	138	kB
Maximum	60.000	kB

Von kB	Bis kB	Anzahl
0	-49	36934
50	-99	53194
100	-199	61277
200	-399	41045
400	-999	29912
1000	-3999	14477
4000	-	1143

Tabelle 2: Artikelgrößen

Nach derzeitiger Schätzung, sind 10% der vorliegenden Artikel retrodigitalisiert. Das ist in sofern problematisch, als für eine Volltextindizierung, wie auch für eine Metadatenextraktion direkt aus dem Artikel, eine Umwandlung der Bilddaten zurück in Text erforderlich ist. Hierzu ist Schrifterkennungssoftware (OCR) nötig. Untersuchungen hierzu sind geplant, da noch unklar ist, mit welcher Genauigkeit[6] die Rückumwandlung möglich ist.

[6]Eine 99% Präzision bei der Buchstabenerkennung bedeutet mehr als 20 Fehler pro Seite.

3.2 Indizierung

Hier einige statistische Informationen, die für eine Aufwandsabschätzung einer Volltextindizierung von Nutzen sein könnten.

Tabelle 3 zeigt die Verteilung der Wortanzahlen in den Artikeln. Es wurde die Anzahl der Worte nach der Umwandlung in Text mittels `pdftotext`[7] gezählt. Dabei ist zu beachten, daß digitalisierte Artikel sehr niedrige Wortzahlen haben, da diese nicht umgewandelt werden können.

Worte gesamt:	1.1 Mrd.
Durchschnittlich je Artikel:	4700

Von Worte	Bis	Anzahl Artikel
0	-99	13214
100	-999	16471
1000	-2499	42622
2500	-4999	74538
5000	-9999	72009
10000	-19999	16671
20000	-	1279

Tabelle 3: Artikellängen in Worten

Für eine potentielle Indizierung wurden nur Worte gezählt, die mindestens aus drei Buchstaben bestehen. Das schließt Zahlen ebenso aus wie insbesondere “to” und “a”. Es verbleiben dann etwa 766 Mill. Worte. Davon sind 4,3 Mill. verschieden[8].

Von Häufigkeit	Bis	Anzahl Worte
	1	2.538.158
	2	505.003
	3	232.036
	4	145.551
	5	99.292
6	-9	221.406
10	-99	463.018
100	-999	110.484
1000	-9999	24.406
10000	-99999	6.216
100000	-	1.112

Tabelle 4: Worthäufigkeiten

[7] `http://www.foolabs.com/xpdf/`

[8] Dies beinhaltet allerdings ca. 100.000 Emailadressen und vermutlich nochmals ebensoviele Autorennamen. Darüber hinaus ca. 8000 Gensequenzen und eine Anzahl von Worten die durch eine falsche Textrückwandlung entstanden sind.

Tabelle 4 zeigt die Häufigkeitsverteilung von Worten. Wie man sieht, tritt etwa die Hälfte aller Worte nur einmal auf, während es etwa 1000 Worte gibt die mehr als 100.000 mal auftreten. Die Spitzenreiter sind wie erwartet "the", "and" und "for" die sechseinhalb, drei und eine Million mal auftreten. Tabelle 5 zeigt die Zahlen für die häufigsten Substantive.

Häufigkeit	Substantiv	Häufigkeit	Substantiv
1.374.390	time	736.298	cells
1.237.654	figure	729.734	function
1.236.167	data	717.888	values
1.076.197	results	710.149	group
1.007.954	system	690.851	cell
1.005.742	model	660.205	control
984.054	study	659.853	effect
972.013	number	658.034	order
962.900	table	654.996	water
940.262	case	654.085	university
924.177	analysis	647.472	value
819.398	fig	629.039	temperature
745.418	research		

Tabelle 5: Die 25 häufigsten Substantive

3.3 Keine Synchronisation

Zu keinem Zeitpunkt gab es vom Verlag Hinweise, welche Artikel vorhanden sein sollten. Dies macht es praktisch unmöglich festzustellen, ob alle lizenzierten Artikel vorhanden sind oder nicht. Erschwerend kommt hinzu, daß keinerlei Bestätigungen auf die E-Mail Benachrichtigungen erfolgt, d.h. sollte eine E-Mail verloren gehen, wird die entsprechende Datei nicht abgeholt und enthaltene Artikel fehlen. Da zwischenzeitlich kein Abgleich erfolgte, sind so entstandene Lücken nicht zu entdecken.

In einem Fall stellte sich im Nachhinein heraus, daß eine der übertragenen ZIP Dateien fehlerhaft[9] war. Da schon mehr als ein Monat vergangen war, gab es keine Möglichkeit die Datei nochmals zu übertragen. Auf Rückfrage teilte der Verlag mit, daß wenn *wir* mitteilen könnten, was uns fehle, diese Artikel nachgeliefert werden können. Offensichtlich besitzt der Verlag auch kein Wissen darüber, welche Artikel geliefert worden sind[10].

Da keinerlei Rückmeldungen oder Überprüfungen erfolgen, muß das hier verwendete Verfahren früher oder später dazu führen, daß ein Teil der Artikel

[9] Vermutlich aufgrund eines Übertragungsfehlers

[10] Inzwischen ist das Verfahren geändert worden, und jede Benachrichtigungsemail enthält eine Liste der zu erwartenden Dateien.

verloren geht. Dabei ist es nicht einmal immer möglich diesen Sachverhalt als solchen sicher festzustellen.

3.4 Keine Verifikation

Ebenfalls gibt es keinerlei Möglichkeit festzustellen, ob die von uns gespeicherten Dateien identisch zu denen des Verlages sind. Sollte an einer Stelle beim Vervielfältigen, Umkopieren, auch zwischen verschiedenen Medien, ein Datenfehler eintreten, ist dies nicht festzustellen. Da diese Art von Fehlern akkumulierend auftritt, wird sich mit der Zeit eine Verschlechterung der Qualität des Datenbestandes einstellen, wenn keine Abhilfe geschaffen wird. Daher berechnen wir nach Eintritt einer Datei in unser System eine MD5 [11] Prüfsumme, die es ermöglicht unbeabsichtigte Änderungen an dieser Datei mit sehr hoher Wahrscheinlichkeit aufzuspüren.

Sowohl für das Verifikations- wie für das Synchronisationsproblem wäre es eine einfache Lösung, wenn der Verlag für jede Zeitschrift eine Liste der zugehörigen Dateien und deren MD5 Prüfsumme vorhalten würde. Diese würde es ermöglichen, jederzeit den Datenbestand abzugleichen und zu verifizieren.

3.5 Konformität der PDF Dateien

Die Untersuchung der von Kluwer gelieferten Dateien zeigte, daß mehr als 100 verschiedene Versionen von PDF produzierenden Softwareprogrammen eingesetzt wurden.

Es stellte sich heraus, daß es schwer zu überprüfen ist, ob

1. ein als PDF vorliegender Artikel zulässiges der Spezifikation entsprechendes PDF ist.
2. der Inhalt optisch erkennbar ist.
3. der Inhalt in Text umgewandelt werden kann.

Punkt 1 resultiert aus der hohen Komplexität von PDF. Hier gibt es aber einige Werkzeuge, die geeignet scheinen eine hinreichende Prüfung durchzuführen. Inwieweit es möglich ist, verläßlich Verletzungen der Spezifikation zu entdecken, ist offen. Die Experimente haben vielfältige Fehlermeldungen diesbezüglich zu Tage gefördert. Hier Beispiele der von `pdftotext` gelieferten Fehlermeldungen:

`Copying of text from this document is not allowed.`
: Bei sechs Dokumenten war die Markierung gesetzt, die ein Extrahieren des Textes verbietet. Ebenso können PDF Dokumente als "nicht druckbar" markiert sein. Wiewohl sich diese Markierungen umgehen

[11] `http://www.ietf.org/rfc/rfc1321.txt`

lassen, kann man Dateien, die solche Markierungen tragen, mit Recht als *ungeeignet* zurückweisen.

`PDF file is damaged - attempting to reconstruct xref table`
: Bei 33 Dokumenten liegen strukturelle Fehler vor. Eine Stichprobe ergab, daß diese Dokumente auch mit dem Acrobat Reader nicht mehr korrekt gezeigt werden.

`Unknown compression method in flate stream`
: Da PDF ein Containerformat mit vielen Unterformaten ist, kann es immer wieder passieren, daß Inkompatibilitäten zwischen Werkzeugen auftreten. (3 Dokumente).

`PDF version 1.5 -- xpdf supports version 1.4`
: Da es keine Festlegung gibt, welche PDF Versionen verwendet werden dürfen, müssen alle Werkzeuge, die zu deren Verarbeitung eingesetzt werden, sowohl mit den ältesten als auch den neuesten Versionen zurechtkommen. Wobei man damit rechnen muß, daß unmittelbar nach Erscheinen einer neuen PDF Version, diese auch verwendet wird. (160 Dateien).

Insgesamt hatten 277 Dateien oder etwa 1 Promille derartige Probleme.[12]

Punkt 2 ist vermutlich nicht lösbar. Es ist möglich, innerhalb eines PDF Dokuments zu zeichnen, Rasterdaten einzublenden und sogar Programme ablaufen zu lassen, um den gewünschten optischen Eindruck zu erzielen. Automatisch festzustellen, ob Teile einer Seite vielleicht nicht sichtbar sind, ist kaum möglich.

Punkt 3 ist der schwerwiegendste. Wie schon angesprochen, ist es u.a. für eine Volltextindizierung notwendig, den reinen Text aus einem PDF Dokument zu extrahieren. Es gibt vielfältige Gründe, warum dies nicht möglich sein kann, insbesondere wären zu nennen: Unbekannte Schriftkodierungen und Rasterdaten (z.B. Scanns). Die Experimente mit den Daten haben auch

[12] Um noch einmal zu verdeutlichen, wie komplex das Problem ist, nachfolgend unkommentiert die übrigen Fehlermeldungen, die auftraten. Dabei ist nicht per se klar, ob das Problem bei der PDF Datei oder `pdftotext` liegt.

```
Bad annotation action
Bad image parameters
Dictionary key must be a name object
Arg #0 to 'BMC' operator is wrong type (string)
Unterminated string
No current point in lineto
Wrong number (2) of args to 'Tz' operator
Bad annotation destination
Couldn't find cidToUnicode file for the 'Adobe-Japan1' collection
Mismatch between font type and embedded font file
Missing 'endstream'
```

gezeigt, daß es zu fehlerhaften Extraktionen kommen kann, z.B. wenn ein Wortende nicht erkannt wird und dadurch Worte wie “aaffectinteractionwithvoltage” oder ’acetyltransferaseactivity’ entstehen. Dabei ist nicht erkennbar, ob es sich hier möglicherweise nur um eine (aus unserer Sicht eher unwichtige) Bildunterschrift handelt, oder wohlmöglich eine Überschrift.

3.6 Sonstige Fehler

Bei unseren Untersuchungen fanden wir innerhalb der Daten diverse Fehler und Inkonsistenzen. Insbesondere gibt es sowohl PDF Dateien von Artikeln für die keine Metadaten vorliegen, als auch Metadaten für die keine Artikel vorhanden sind. Fehler in den Metadaten, wie fehlendes Veröffentlichungsdatum, kommen vor. Dateinamen sind doppelt vergeben worden. Zusammen mit den o.a. genannten Problemen ist der Anteil von fehlerhaften Dateien aber gering und liegt vermutlich unter zwei Promille. Leider ist für den Projektzeitraum keinerlei Verfahren vereinbart worden, das hier eine Meldung und Korrektur von erkannten Fehlern vorsieht.

4 Artikelspeicherung im Dateisystem

Wir haben uns bewußt dafür entschieden, die Artikel im Dateisystem zu speichern. Folgende Vorteile ergeben sich:

- Einfach, da lediglich Dateien kopiert werden. Dies gilt zumindest solange, wie keine komplexen Managementfunktionen eingebunden werden.
- Schnell. Einen schnelleren Zugriff als direkt auf eine Datei im Dateisystem ist kaum machbar.
- Billig, da keine zusätzliche Software benötigt wird. Aktuelle Betriebssysteme bieten z.T. eine Auswahl verschiedener Dateisysteme an, die entsprechend den Anforderungen verwendet werden können.
- Pflegeleicht. Da die Daten statisch sind, können kaum Probleme auftreten, die Wartungsarbeiten am Dateisystem (z.B. Defragmentierung) erforderlich machen.
- Platz effizient. Der Overhead je Datei ist gering. Üblicherweise wird etwas die Hälfte eines Dateisystemblocks je Datei nicht genutzt.
- Vielseitiger Zugriff möglich (Direkt, FTP, HTTP, NFS, etc.). Es gibt kaum etwas, das nicht auf Dateisystem zugreifen kann.
- Sicher, vielseitige Backuplösungen verfügbar. Durch redundante Hardware wie RAID Systeme kann eine hohe Verfügbarkeit sichergestellt

werden. Backuplösungen für Dateisysteme sind in großer Zahl verfügbar. Das Neuerstellen oder Wiederherstellen eines nicht mehr operablen Systems ist mit — im Verhältnis zu anderen Systemen — geringem Aufwand möglich.

- Zukunftssicher, da hohe Migrierbarkeit. Dateisysteme wird es auf absehbare Zeit geben. Eine Migration von Daten auf die jeweils nächste Generation ist hier immer gewährleistet.

Die Nachteile sind nicht ganz klar, in wesentlichen fehlender Komfort und die Notwendigkeit, bestimmte Aufgaben selbst zu programmieren, die von DMS/CMS standardmäßig geleistet werden.

4.1 Zugriffskontrolle

Innerhalb des Vorprojekts waren hierzu keine Arbeiten geplant. Wir haben eine einfache IP-adressenbasierte Zugriffskontrolle implementiert, die problemlos funktioniert.

Nach dem vor kurzem bekannt gewordenen Zusammenbruch der Kontenintegrität bei Microsofts Passport System kann man nur feststellen, daß eine zentrale Zugriffskontrolle mit außerordentlich hohen Risiken behaftet ist. Abgesehen von den technischen Problemen, sind die Kosten für mögliche Schäden im Falle von Sicherheitslücken nicht tragbar. Hier ist in jedem Fall eine verteilte Lösung anzustreben.

5 Zusammenfassung

Die bisherigen Erfahrungen sind ausgesprochen positiv. Es ist gelungen, ein mittelgroßes Datenvolumen zu speichern und zu verarbeiten. Die erkannten Probleme liegen mehr im systematischen denn im technischen Bereich und sollten daher vorrangig bei dem Abschluß neuer Verträge berücksichtigt werden.

In Zukunft sollte bei der Datenlieferung von den Verlagen darauf geachtet werden, daß folgende Fragen beantwortet werden können:

- Vollständigkeit
  Was hätte geliefert werden sollen?
- Integrität
  Ist das, was wir gespeichert haben, unverändert?
- Konformität
  Entsprechen die Dateien der PDF Spezifikation?
- Berichtigung
  Wie werden erkannte Fehler gemeldet und behoben?

Die technischen Herausforderungen, die sich für den Fortgang des Projektes abzeichnen, liegen vor allem bei der Textrückwandlung und Metadatenextraktion. Wie extrahiert man aus den vorliegenden Daten die Metadaten der Ebene-1 und Ebene-2. Hierbei wird es interessant sein, den Vergleich mit den Datenlieferungen anderer Verlage durchzuführen.

1.7