EUCOR-Bibliotheksinformationen - Informations des bibliothèques: 14 (1999)

EVA - Volltextarchiv der Universitätsbibliothek Karlsruhe

Dr. Michael Mönnich (UB Karlsruhe)

Dipl. Inform. Günter Radestock (UB Karlsruhe)

Zusammenfassung:
Das Elektronische Volltextarchiv (EVA) der Universitätsbibliothek Karlsruhe dient der langfristigen Archivierung elektronischer Dokumente, die von Mitarbeitern der Universität publiziert werden. Die Basis bildet das Postscriptformat, aus dem automatisch ein strukturiertes Hypertext-Dokument (HTML), ein Volltextindex und ein Printformat erstellt und abgespeichert werden. EVA enthält über 1.000 Dokumente, darunter eine wachsende Zahl an Dissertationen. Im Aufsatz werden die Umsetzung in der Universität sowie die verwendeten Formate und Techniken dargestellt.

Elektronisches Publizieren heute

Fast alle Publikationen von Mitarbeitern der Universität Karlsruhe werden mit Textverarbeitungssystemen erstellt und liegen primär in digitaler Form vor. Somit sind die Voraussetzungen für das elektronische Publizieren und die Verbreitung dieser Texte über Datennetze - insbesondere über das Internet - gegeben. In der Tat bieten bereits zahlreiche Mitarbeiter der Universität über die WWW-Server von Instituten, Fakultäten oder privat ihre Publikationen im Internet an.

Bei näherem Betrachten werden dabei einige Probleme deutlich, welche die Benutzbarkeit der Publikationen zum Teil einschränken:

- Die Texte liegen in unterschiedlichen Datenformaten vor.
- Die Dokumente sind meist uneinheitlich oder gar nicht erschlossen.
- Die Dokumente sind meist nicht in Katalogen verzeichnet, sondern nur über Suchmaschinen u.ä. unzuverlässige Hilfsmittel auffindbar.
- Die Server, auf denen die Dokumente aufliegen, werden z.T. unzureichend gewartet, mit der Folge, daß hohe Ausfallzeiten und Datenverluste vorkommen.
- Die Adressen können sich ändern.
- Die Langzeitsicherung der Dokumente ist ungeklärt. Die Flüchtigkeit der elektronischen Dokumente birgt die Gefahr in sich, daß wichtige wissenschaftliche Erkenntnisse nach einigen Jahren nicht verfügbar sind, wenn nicht rechtzeitig Maßnahmen zur Archivierung getroffen werden.

Der konkrete Zugriff auf die elektronischen Dokumente gestaltet sich deshalb zur Zeit noch recht mühsam.

Volltextarchiv als Dienstleistung der Bibliothek

Traditionell versorgt die Universitätsbibliothek die Universität mit wissenschaftlicher Literatur aller Art. Dazu gehört in Zukunft auch die Bereitstellung elektronischer Dokumente. Zudem hat die Universitätsbibliothek als zentrale Archivbibliothek der Universität die

Verpflichtung, die langfristige Archivierung dieser Dokumente ebenso zu gewährleisten, wie es bei Printmedien üblich ist.

Daher wurde an der Universitätsbibliothek Karlsruhe damit begonnen, ein elektronisches Volltextarchiv (EVA) aufzubauen, das alle elektronischen Dokumente enthält, die in der Universität erzeugt werden. Ausgangsbasis hierfür ist das konventionelle Veröffentlichungsverzeichnis, das seit Berichtsjahr 1968 von der UB in Printform und seit 1989 als Datenbank

(*http://www.ubka.uni-karlsruhe.de/hylib/vv_suchmaske.html*) angeboten wird und alle Publikationen nachweist.

EVA stellt ein Konzept dar, wie elektronische Dokumente einheitlich präsentiert, umfassend recherchiert und langfristig archiviert werden können.

Im einzelnen bietet EVA:

- Zugriff auf die Texte direkt nach der Katalogrecherche
- Einfacher und komfortabler Zugriff auf die Dokumente (Bildschirmlesen und Ausdruck)
- Gute Recherchemöglichkeiten im Text der Dokumente
- Rund um die Uhr-Verfügbarkeit
- Sicheres Backup
- Sicherung der langfristigen Verfügbarkeit, Archivierung (gegebenenfalls mit Überführung in neues Datenformat)
- Sicherstellung der Authentizität der Dokumente

Aufbereitung der Daten

Um das Ziel der Einheitlichkeit zu gewährleisten, werden die Daten in der UB so aufbereitet, daß zum einen eine Volltext-Recherche möglich ist und zum anderen der Zugriff auf die Dokumente in einem einheitlichen Rahmen stattfindet. Zu diesem Zweck wird ausgehend vom Postscriptformat automatisch ein strukturiertes Hypertext-Dokument (HTML), ein Volltextindex und ein Printformat (Postscript oder GIF) erstellt und abgespeichert. In den Katalog wird ein Link auf die HTML-Datei gesetzt.

Verfügbarkeit und Archivierung

Um die Verfügbarkeit möglichst optimal zu gestalten, werden die Dokumente von der Originallokation kopiert und auf dem Server der UB redundant gehalten. Würden nur die Links auf die Instituts-, Fakultäts- und sonstigen Server gehalten, könnten die Ausfallzeiten dieser nachgeordneten Systeme die Verfügbarkeit beeinträchtigen. Die Sicherung geschieht über lokales Backup und Sicherung über das ADSM-System des Rechenzentrums der Universität. Die langfristige Archivierung geschieht dann auf der Basis dieser Dokumente.

Authentizität der Dokumente

Die Authentizität von Dokumenten, die auf dem EVA-Server aufliegen, gewährleistet die Universitätsbibliothek. Das Einbringen von neuen oder geänderten Dokumenten geschieht nur nach Absprache mit dem Autor.

Urheberrecht

Urheberrechtliche Probleme treten vor allem bei Texten auf, die in Zeitschriften, Kongreßbänden und Büchern erscheinen. Es gibt Verlage, die den Autoren die elektronische Parallelveröffentlichung von Aufsätzen, die in Printform erscheinen, verbieten. Eventuelle urheberrechtliche Fragen im Zusammenhang mit der elektronischen Verbreitung seiner Texte zu regeln, obliegt dem Autor. Jeden Einzelfall mit dem Verlag abzuklären, kann die UB nicht leisten. Sollten nach der Veröffentlichung auf dem Server der Bibliothek rechtliche Probleme auftauchen, so kann der Zugang zu den entsprechenden Dokumenten sehr schnell gesperrt werden.

Vorgehen

Im Januar 1997 wurden von der UB alle Einrichtungen der Universität über das Vorhaben EVA informiert und um Mitarbeit, d. h. die Bereitstellung von Dokumenten, gebeten. Außerdem können auch Papierdokumente in der UB eingescannt werden. Die Dokumente können als Postscriptfiles angeliefert werden. Die Anlieferung kann über Diskette oder direkt über FTP erfolgen (***ftp://ftp.ubka.uni-karlsruhe.de/pub/incoming/vvv/***).

Elektronische Dissertationen

Im Oktober 1999 enthielt EVA 1019 Dokumente. Darunter befinden sich 62 Dissertationen sowie sämtliche internen Berichte der Fakultät für Informatik ab 1993.

Das elektronische Publizieren von Dissertationen in EVA gestatten von 12 folgende 9 Fakultäten der Universität Karlsruhe:

- Fakultät für Mathematik
- Fakultät für Bio- und Geowissenschaften
- Fakultät für Geistes- und Sozialwissenschaften
- Fakultät für Architektur
- Fakultät für Bauingenieur- und Vermessungswissenschaften
- Fakultät für Maschinenbau
- Fakultät für Elektrotechnik und Informationstechnik
- Fakultät für Informatik
- Fakultät für Wirtschaftswissenschaften

Die Fakultät für Chemieingenieurwesen und Verfahrenstechnik genehmigt elektronische Dissertationen nur auf Antrag an den Fakultätsrat und mit Zustimmung des Betreuers.

Bei Veröffentlichung der Dissertation in elektronischer Form müssen die Doktoranden an die Universitätsbibliothek abliefern:

1. Die Dissertation in einer mit der UB abgestimmten elektronischen Version für den Volltextserver

2. Drei archivgeeignete gedruckte Exemplare der Dissertation für Archiv- und Bestandszwecke der Universitätsbibliothek

3. Zwei archivgeeignete gedruckte Exemplare der Dissertation als Pflichtexemplare für Die Deutsche Bibliothek. (Die elektronische Fassung wird von der Universitätsbibliothek an Die Deutsche Bibliothek übermittelt.)

4. Eine Bescheinigung über die Identität von gedruckten Exemplaren und elektronischer Form, ausgestellt vom Gutachter oder vom Doktoranden (Festlegung durch die Fakultät).

5. Gegebenenfalls eine Einverständniserklärung der Fakultät für die elektronische Publikation

6. Gegebenenfalls eine Bescheinigung des Gutachters über die inhaltliche Richtigkeit, wie bei einigen Promotionsordnungen vorgeschrieben.

Die Anzahl und Abgabeform der laut Promotionsordnung an Institut, Fakultät und Referenten abzuliefernden gedruckten Exemplare der Dissertation sind hiervon nicht berührt.
Im Falle des elektronischen Publizierens überträgt der Doktorand der Universität das Recht, im Rahmen der gesetzlichen Aufgaben der Universitätsbibliothek die Dissertation in Datennetzen zur Verfügung zu stellen. Die Dissertation verbleibt daher dauerhaft auf dem Server der Universitätsbibliothek. Durch eine spätere zusätzliche Verlagsveröffentlichung der Dissertation kann dieses Recht nicht eingeschränkt werden.

Zugriff auf die Dokumente

Der Zugriff auf die Dokumente erfolgt auf Basis des WWW über

- den OLIX-OPAC der UB, mit Recherchemöglichkeit nach Autor, Titelstichworten usw.
- die Metadaten im Veröffentlichungsverzeichnis
- den Zugriff auf einen Volltextindex aller Dokumente
- die Recherche im Volltext der einzelnen Dokumente
- Zugriff über einen hierarchischen Dateibaum, z.B. chronologische Liste aller Dokumente oder Dissertationen nach Fakultäten.

Bei der Volltextrecherche werden gefundene Treffer im Text markiert, bei mehreren Treffern kann direkt auf die nächste bzw. vorhergehende Fundstelle gesprungen werden.

Verwendete Technik

Formate

Funktional können drei verschiedene Formate unterschieden werden, die unterschiedliche Anforderungen erfüllen müssen: Anlieferungsformat, Archivierungsformat und Präsentationsformat. Da kein Format gefunden wurde, das alle Anforderungen erfüllt, werden von uns verschiedene Format verwendet, zwischen denen konvertiert wird, ausgehend von dem Archivierungsformat(en).

Anlieferungs- und Archivierungsformat

Postscript

Die Verwendung von Postscript als Anlieferungsformat minimiert einerseits den Zusatzaufwand der Autoren bei der Anlieferung und erlaubt andererseits eine weitgehend automatische Weiterverarbeitung an der Bibliothek.

Der Autor kann mit jeder gängigen Textverarbeitung eine Postscriptdatei über die Druckfunktion erstellen. Diese Postscriptdatei kann er dann mit seinem WWW-Browser auf dem FTP-Server der UB ablegen, ein schnelles Verfahren ohne große Fehleranfälligkeit.

Postscript hat sich seit seiner Vorstellung vor zwölf Jahren (eine lange Zeit im EDV-Bereich) als Sprache bzw. Dateiformat zur Beschreibung gedruckter Dokumente durchgesetzt. Es ist inzwischen so weit verbreitet, daß man davon ausgehen muß, daß auch zukünftige Hard- und Software den Standard Postscript unterstützen wird. Zu Archivierungszwecken kann eine Postscriptdatei ein gedrucktes Dokument ersetzen, da die Papierform durch Ausdrucken auf einem postscriptfähigen Drucker jederzeit fehlerfrei reproduziert werden kann. Durch Ablegen von Postscriptdateien auf einem ans Internet angeschlossenen Server wird der weltweite Zugriff auf die Dokumente ermöglicht.

PDF, Ursprungsformat

Der Autor hat die Möglichkeit, außer Postscript weitere Formate anzuliefern, die gespeichert

werden und dem Benutzer zum Download zur Verfügung stehen.

Aus der vom Autor gelieferten Postscriptdatei kann automatisch eine PDF-Datei generiert werden. Die automatisch generierte PDF-Datei enthält aber keine Verweise innerhalb des Dokumentes, kein klickbares Inhaltsverzeichnis. Falls der Benutzer mit Software arbeitet, die PDF-Dateien mit Inhaltsverzeichnis erzeugen kann, oder die Verweise mit der Acrobat-Software selbst in die PDF-Datei einfügen möchte, bieten wir ihm die Möglichkeit, auch eine PDF-Datei zu liefern, die anstelle der von uns erzeugten PDF-Datei verwendet wird. Eine ausschließliche Lieferung im PDF-Format ist leider nicht möglich, da die Weiterverarbeitung des PDF-Formats mit technischen Schwierigkeiten verbunden ist.

Eine Speicherung des Ursprungsformats (z.B. MSWord, LaTeX) ist ebenfalls sinnvoll. Im Ursprungsformat sind Informationen enthalten, die bei der Generierung von Postscript oder PDF verlorengehen (Überschriften, Indexmarkierungen und ähnliches). Diese Informationen könnten für einige Benutzer interessant sein, können aber auch bei der Verfügbarkeit von entsprechenden Konvertern für eine erneute Aufbereitung der Dokumente verwendet werden.

Formate zur Präsentation

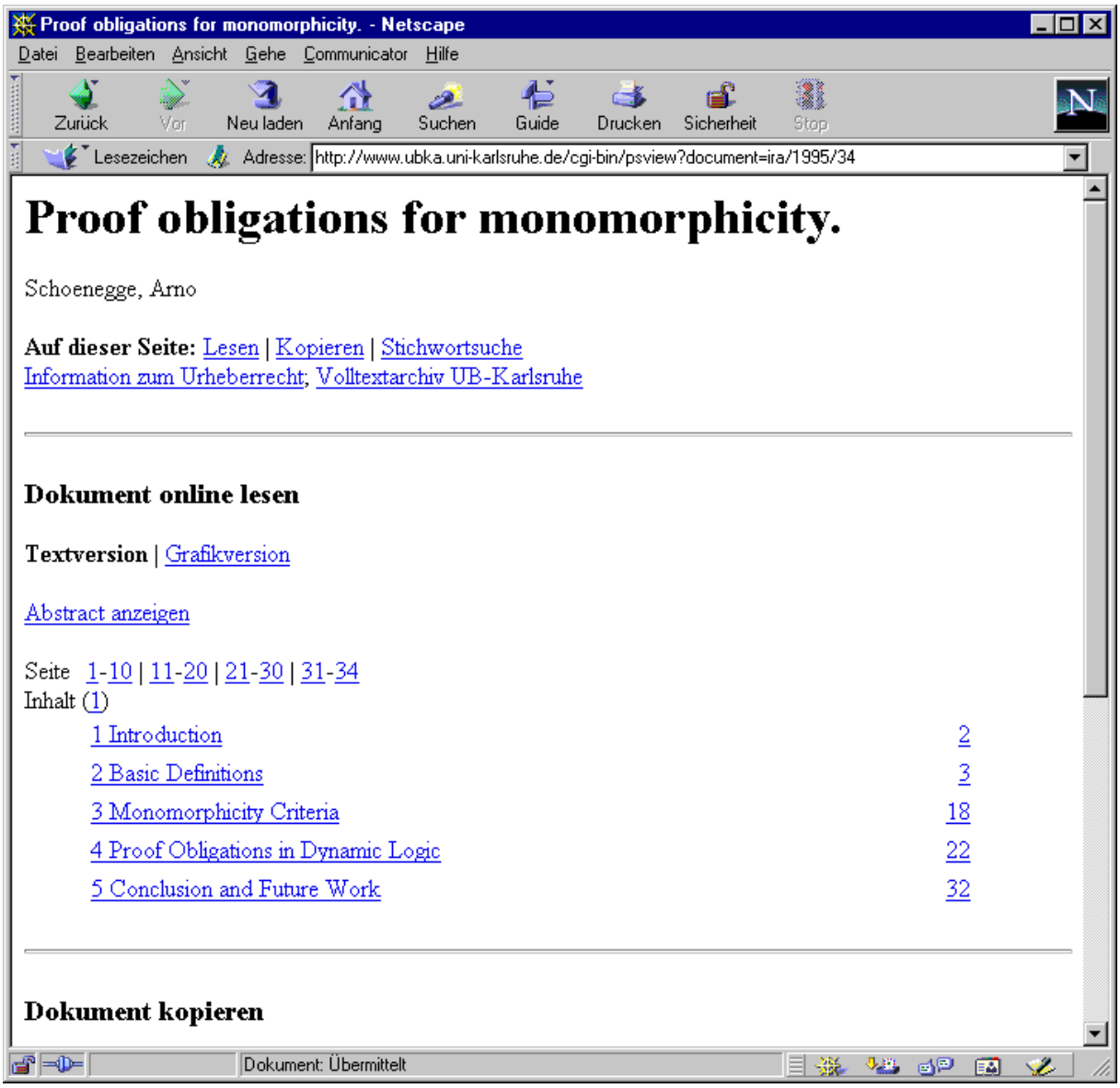

BILD 1: SCREENSHOT http://www.ubka.uni-karlsruhe.de/cgi-bin/psview?document=ira/1995/34

HTML, GIF

Da "rohe" Postscriptdateien unhandlich in der Benutzung sind, werden die Dokumente in EVA als (dynamische) HTML-Seiten präsentiert. Für die Akzeptanz von EVA bei den Nutzern ist es notwendig, daß ein Browsing am Bildschirm möglich ist, sowie eine Volltextsuche. Diese Seiten enthalten entweder einen Teil des Dokumenttextes oder die Abbildung einer Dokumentseite. Im HTML-Text kann der Nutzer sich schnell über den Inhalt des Dokumentes informieren, im Dokument navigieren und suchen. Zudem sind die HTML-Dokumente als reine ASCII-Texte relativ klein und daher schnell über Datennetze zu übertragen.

HTML bietet für die Darstellung von Formeln, Grafiken und Tabellen jedoch nur eingeschränkte Möglichkeiten. Dies ist insbesondere bei natur- und ingenieurwissenschaftlichen Texten ein gravierender Nachteil, der die Texte teilweise unleserlich macht. Deshalb besteht bei allen in EVA gespeicherten Dokumenten die

Möglichkeit, jederzeit von der Textversion des Dokumentes zur Abbildung einer Seite im Grafikformat GIF zu wechseln.

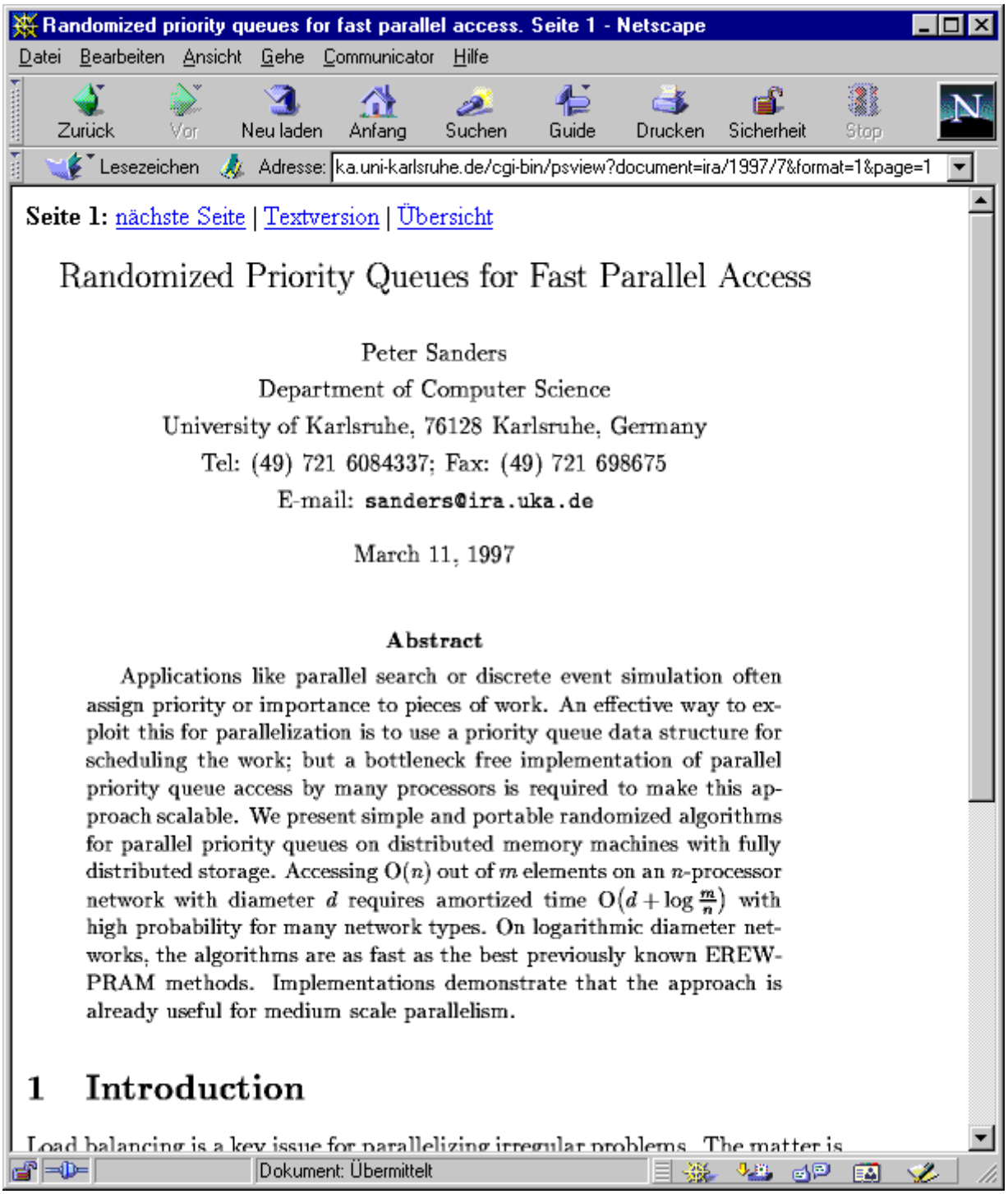

BILD 2: SCREENSHOT http://www.ubka.uni-karlsruhe.de/cgi-bin/psview?document=ira/1997/7&format=1&page=1

Postscript, PDF

Außer dem Ausdruck ist das direkte Lesen von Postscriptdateien am Bildschirm möglich. Dazu ist jedoch ein Postscriptinterpreter nötig, der auf dem Rechner des Lesers installiert werden muß. Ein solcher Postscriptinterpreter ist das frei verfügbare Ghostscript, das mit

Ghostview oder GSview eine brauchbare Benutzerschnittstelle zum Lesen von Dokumenten am Bildschirm verfügt.

Einfacher zu installieren und komfortabler zu bedienen ist der Acrobat-Reader von Adobe, den Entwicklern von Postscript. Der zum kostenlosen Download angebotene Acrobat-Reader kann Dateien im PDF Format anzeigen und ausdrucken. PDF-Dateien können automatisch aus Postscriptdateien erzeugt werden und weisen einige Vorzüge gegenüber Postscript auf:

Verwendete Werkzeuge

Zur Aufbereitung der Dokumente dient das Werkzeug *Pscript*, das an der UB entwickelt wurde und auf verschiedenen frei verfügbaren Software-Tools aufbaut. Zentraler Bestandteil von Pscript ist ein Postscript-nach-Text Konverter, der die Postscript-Eingabedatei mit Ghostscript interpretiert, und aus dem dabei erstellten Protokoll den Text und die Struktur des Dokumentes (Seiten, Absätze, Seitennummern, Inhaltsverzeichnis) rekonstruiert.

Die Grafikversionen der Dokumente werden ebenfalls mit Ghostscript erstellt. Die von Ghostscript erzeugten Bilddateien werden mit dem Grafikpaket NetPBM in das im Web gebräuchliche GIF-Format übersetzt. Durch die Erzeugung von Bilddateien in hoher Auflösung und nachträgliches Verkleinern wird dabei ein Antialias-Effekt bewirkt, d.h. die Zeichen des Textes erscheinen durch die Verwendung von Graustufen glatter, als wenn sie in niedriger Auflösung erzeugt worden wären.

Die Aufbereitung von Dokumenten mit Pscript erfolgt weitgehend automatisch, ist jedoch parametrisierbar. Die meisten Dokumente im Elektronischen Volltextarchiv werden mit den gleichen Standardparametern übersetzt, durch Anpassung der Parameter kann man auch ungewöhnliche Dokumente in EVA übernehmen. Bei der Erzeugung der Bilddateien kann beispielsweise die Auflösung, die Anzahl Farben (bzw. schwarz/weiß, Graustufen oder Farben) eingestellt werden, um neben "normalen" Dokumenten auch solche mit besonders kleinen Indices in Formeln oder solche mit farbigen Abbildungen lesbar anzuzeigen.

Inzwischen stehen auch Werkzeuge zum Einbringen von eingescannten Dokumenten in das Volltextarchiv zur Verfügung. Einscannen bietet sich vor allem bei Dokumenten an, zu denen (wegen ihres Alters) keine elektronische Fassung existiert oder bei denen diese nur schwer beschafft werden kann. Diese Dokumente werden zunächst mit dem LEA-System der UB-Karlsruhe (***http://lea.ubka.uni-karlsruhe.de/lea/***) gescannt und in das von EVA benötigte Format konvertiert∗ Die Acrobat-Software (Version 4.0) von Adobe erzeugt aus den gescannten Bildern PDF-Dateien, wobei der Text mit OCR-Verfahren rekonstruiert wird (bei sehr altem Schriftsatz in ungebräuchlichen Sprachen, z. B. Latein ist das problematisch; bei modernen Texten sind die Ergebnisse überraschend gut). Im so erzeugten EVA-Dokument können in einem weiteren Schritt durch manuelle Nachbearbeitung noch der Text korrigiert, Inhaltsverzeichniseinträge markiert und überflüssig gescannte Seitenränder abgeschnitten werden. Im Idealfall können die so bearbeiteten Dokumente kaum noch von den aus Postscript erzeugten unterschieden werden.

Beispiele für solche Dokumente finden sich unter
http://www.ubka.uni-karlsruhe.de/ausstellung/wasbleibt/

∗ Tangen, D.; Radestock, G.: Das Elektronische Aufsatzliefersystem LEA in: EUCOR-Bibliotheksinformationen 11/1997) S. 32-37.