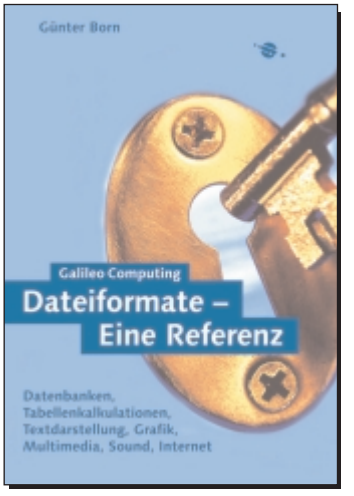
Dieses Kapitel stammt aus dem Buch

›Dateiformate – Eine Referenz‹

von Günter Born.

www.borncity.de



ISBN 3-934358-83-7
119,90 DM

Informationen zum Buch
mit Bestellmöglichkeit

www.galileocomputing.de

Galileo Computing

8 Word 97-Dateiformat (DOC)

Microsoft benutzt für Microsoft Word für Windows ein eigenes Dateiformat, welches in den verschiedenen Word-Versionen (6.0, Word 95 und Word 97) variiert. Nachfolgend finden Sie Informationen zum Aufbau des Dateiformats in Word 97.

Der Aufbau einer Word DOC-Datei

Microsoft Word ist als OLE2-Anwendung programmiert, die DOC-Dateien bestehen daher aus Streams, die über OLE2-Docfile-API-Aufrufe erstellt wurden. Eine Anwendung, die die Binärdaten aus den DOC-Dateien lesen möchte, sollte daher auch diese API-Aufrufe verwenden. Eine Word DOC-Datei besteht aus einem Main-Stream, einem Summary-Stream und keinem oder mehreren Objekt-Streams (enthalten die Daten von Objekten, die im Word-Dokument eingebettet sind). Objekt-Streams werden durch Word nicht ausgewertet, vielmehr verwendet das Programm die betreffenden OLE-Anwendungen zum Zugriff auf die Daten. Nachfolgend wird der Inhalt des Main-Streams beschrieben.

Struktur des Main-Streams (nicht komplex)

Word kann den Main-Stream in einem komplexen und in einem nicht komplexen Format hinterlegen. In der nicht komplexen Variante beginnt die Datei mit einem Word-Datei-Header (FIB = File Information Block). Daran schließen sich der eigentliche Text (Text, Fußnoten, Kopfzeilen) sowie die Formatierungsinformationen an. Die Textposition wird im File Information Block im Feld *fib.fcMin* (Offset 18H) hinterlegt. Die erste Datenstruktur (FKP) mit der Formatbeschreibung beginnt an einer 512-Byte-Grenze hinter dem letzten Textzeichen. Die restlichen FKP-Blöcke werden als 512-Byte-Seiten angehängt. Die FKP-Blöcke für Zeicheneigenschaften (CHP), Absatzeigenschaften (PAP) und LVCs werden in Word 97 abwechselnd (interleaved) gespeichert, während frühere Word-Versionen diese kontinuierlich ablegten. An die Formatierungsinformationen können sich die Daten für Bilder, eingebettete Objekte etc. anschließen.

File Information Block (in Word 1.0, 2.0 und 6.0)

Der Header einer WinWord-Datei besteht aus 384 (17FH) Bytes, gefolgt vom Textbereich. Der Text wird dabei mit ANSI-Zeichen gespeichert. Die folgende Tabelle zeigt die Struktur des WinWord-Headers.

Offset	Bytes	Bemerkungen
00H	2	Signatur 9B A5H (WinWord 1.0) DB A5H (WinWord 2.0) D0 CFH (WinWord 6.0)
02H	2	Version (Major)
04H	2	Version (Minor)
06H	2	Language Stamp
08H	2	nächste Seitennummer
0AH	1	Flag (Bit 2: 1 = komplexes Format)
0BH	1	Encryption (1 = Ja)
0CH	6	intern benutzt
12H	1	Plattform (0: Windows, 1: Mac)
13H	1	reserviert
14H	2	Zeichensatz (0: ANSI, 100H: Mac)
16H	2	interner Zeichensatz
18H	4	Offset auf 1. Zeichen im Text
1CH	4	Offset Ende Textbereich + 1
20H	4	Offset Dateiende
...		andere Zeiger

Tabelle 8.1 Struktur des WinWord-Headers

File Information Block (in Word 97)

In Word 97 wurde die Struktur des File Information Block (FIB) geändert, um Erweiterungen am Format besser durchführen zu können. Der Header einer DOC-Datei besteht aus 898 Bytes, gefolgt vom Textbereich. Die folgende Tabelle skizziert die Struktur des File Information Block (FIB).

Offset	Bytes	Bemerkungen
00H	2	Signatur D0 CFH (ab Word 6 bis Word 97)
02H	2	FIB-Version (ist immer größer >= 101 für alle Dokumente ab Word 6.0 und höher)
04H	2	Produktversion
06H	2	Language stamp (für lokalisierte Version)
08H	2	nächste Seitennummer

Offset	Bytes	Bemerkungen
0AH	2	Flag Bit 0: 1= Dokument ist eine Vorlage Bit 1: 1= Dokument ist ein Glossar Bit 2: 1= Dokument als Complex gesichert (Schnellspeicherung) Bit 3: 1= Dokument enthält mindestens ein Bild Bit 4–7: Zahl der Schnellspeicherungen des Dokuments Bit 8: 1= Dokument verschlüsselt Bit 9: 0= FIB bezieht sich auf Table-Stream "0Table" 1= FIB bezieht sich auf Table-Stream "1Table" Bit 10: 1= Datei mit Schreibschutz öffnen empfohlen Bit 11: 1= Schreibschutz gesetzt Bit 12: 1= Extended-Characters benutzt Bit 13: FloadOverride Bit 14: FarEast Bit 15: Crypto
0CH	2	*FibBack*: Datei kann mit Lesern höherer Versionen gelesen werden
0EH	4	*LKey*: File-Encrypted-Key (nur gültig bei verschlüsselten Dokumenten)
12H	1	Plattform (0: Windows, 1: Mac)
13H	1	Flags Bit 0: 1= Dokument zuletzt auf Mac gespeichert Bit 4: 1= Dokument mit Word 97 gesichert
14H	2	Standard-Characterset-ID; Zeichensatz (0: ANSI, 100H: Mac)
16H	2	*ChsTables*: Standard-Extended-Character-Set (Zeichensatz) ID für Texte in internen Datenstrukturen.
18H	4	*fcMin*: Datei-Offset auf 1. Zeichen im Text, in nicht komplexen DOC-Dateien läßt sich die Zeichenposition (caracter position, cp) gemäß folgender Formel in eine Dateiposition (file position, fc) umrechnen: fc = cp + fib.fcMin.
1CH	4	*FcMac*: Datei-Offset letztes Zeichen des Texts im Dokument-Stream + 1
20H	2	*csw*: Zahl der Shorts-Einträge; das Feld beginnt ab Offset 22H
22H	2	*wMagicCreated*: ID zur Identifikation des schreibenden Programms (0x6A62 = Word)
24H	2	*wMagicRevised*: ID letzte Dateimodifikation
26H	2	*wMagicCreatedPrivate*: private Daten
28H	2	*wMagicRevisedPrivate*: private Daten
2AH	2	---
2CH	2	unbenutzt
2EH	2	unbenutzt
30H	2	unbenutzt
32H	2	unbenutzt
34H	2	unbenutzt
36H	2	unbenutzt

Offset	Bytes	Bemerkungen
38H	2	unbenutzt
3AH	2	unbenutzt
3CH	2	Dokument mit Fernost-Version von Word geschrieben
3EH	2	*clw*: Zahl der »Longs«-Einträge, das Feld beginnt ab Offset 40H
40H	4	*cbMac*: Datei-Offset auf letztes geschriebenes Byte + 1 in der Datei
44H	4	*lProductCreated*: Build-Datum des Creators (10695 = 6. Jan. 1995)
48H	4	*lProductRevised*: Build-Datum der Anwendung, mit der die Datei zuletzt geschrieben wurde
4CH	4	*ccpText*: Länge Main-Dokument-Stream
50H	4	*ccpFtn*: Länge Fußnoten Sub-Dokument-Stream
54H	4	*ccpHdd*: Länge Header Sub-Dokument-Stream
58H	4	*ccpMcr*: Länge Makro Sub-Dokument-Stream
5CH	4	*ccpAtn*: Länge Anmerkungen (Annotation) Sub-Dokument-Stream
60H	4	*ccpEdn*: Länge Endnoten (Endnoten) Sub-Dokument-Stream
64H	4	*ccpTxbx*: Länge Textfeld (Textbox) Sub-Dokument-Stream
68H	4	*ccpHdrTxbx*: Länge Header-Textbox Sub-Dokument-Stream
6CH	4	*pnFbpChpFirst*: falls ungenügend Speicher zum Expandieren der *plcfbte*-Struktur vorhanden ist, wird *plcfbte* als verzeigerte Liste à 512-Byte beginnend ab *pn* gespeichert.
70H	4	*pnChpFirst*: Seitennummer der ersten Seite im Dokument, welches die CHPX FKP beinhaltet. Das Kürzel FKP steht für Formatted Disk Pages. CHPX gibt die Ausnahmen von Zeicheneigenschaften (Character Property Exception) an.
74H	4	*cpnBteChp*: Zahl der CHPX FKP innerhalb der Datei.
78H	4	*pnFbpPapFirst*: falls ungenügend freier Speicher zum Expandieren der *plcfbte*-Struktur, wird diese in 512-Byte-Blöcken gesichert. *Pn* zeigt auf den Anfang der Liste.
7CH	4	*pnPapFirst*: Seitennummer der ersten Seite, die PAPX FKP-Informationen enthält. (PAPX steht für Paragraph Property Exception)
80H	4	*cpnBtePap*: Zahl der in der Datei gespeicherten PAPX FKP
84H	4	*pnFbpLvcFirst*: falls ungenügend freier Speicher zum Expandieren der *plcfbte*-Struktur, wird diese in 512-Byte-Blöcken gesichert. *Pn* zeigt auf den Anfang der Liste.
88H	4	*pnLvcFirst*: Seitennummer der ersten Seite, die LVC FKP-Informationen enthält
8CH	4	*cpnBteLvc*: Zahl der in der Datei gespeicherten LVC FKP
90H	4	*FcIslandFirst*
94H	4	*FcIslandLim*
98H	2	*cfclcb:* Zahl der Felder im Feld mit den FC/LCB-Paaren (das Feld schließt sich ab Offset 9AH an
9AH	4	*fcStshfOrig:* Offset original Allocation für STSH im Table-Stream

Offset	Bytes	Bemerkungen
9EH	4	*lcbStshfOrig:* Bytes der originalen STSH-Allocation
A2H	4	Offset des STSH im Table-Stream
A6H	4	Bytes der aktuellen STSH-Allokation
AAH	4	Offset in Table-Stream mit Fußnotenreferenzen
AEH	4	Byte-Counter Fußnotenreferenz
B2H	4	Offset in Table-Stream mit Fußnotentext
B6H	4	Bytes in Fußnotentext (0 = keine Fußnote)
BAH	4	Offset in Table-Stream der Anmerkungen (ATRD = Annotation Table Reference Definition). Die Zeichenpositionen (CP) in diesem Datenbereich (PLC) definieren den Offset der Anmerkungsreferenzen im Hauptdokument.
BEH	4	Bytes der Anmerkungsreferenz (Annotation Reference PLC)
C2H	4	Offset in Table-Stream mit Anmerkungstexten (Annotation Text PLC)
C6H	4	Bytezähler für Anmerkungstext PLC (Annotation Text PLC)
CAH	4	Offset in Table-Stream der Section-Descriptor SED PLC
CEH	4	Bytezähler für Section-Descriptor PLC
D2H	4	unbenutzt
D6H	4	unbenutzt
DAH	4	Offset in Table-Stream der PHE PLC-Struktur (Paragraph Heights)
DEH	4	Bytezähler für Paragraph-Height PLC
E2H	4	Offset in Table-Stream der Glossar-String-Tabelle (Glossary String Table)
E6H	4	Bytes in der Glossar-String-Tabelle
EAH	4	Offset in Table-Stream des Glossary PLC
EEH	4	Bytes in Glossary PLC
F2H	4	Offset in Table-Stream des Header HDD PLC (Struktur mit Kopfzeilen)
F6H	4	Bytezähler des Header-PLC (Länge Struktur mit Kopfzeilen)
FAH	4	Offset in Table-Stream mit der Zeichenformatierungstabelle (Character Property Bin Table PLC). Beschreibt Textformatierung des Hauptdokuments und aller Unterdokumente.
FEH	4	Bytes in der Zeichenformatierungstabelle (Character Property Bin Table PLC)
102H	4	Offset in Table-Stream Absatzformatierung (Paragraph Property Bin Table PLC). Beschreibt Textformatierung des Hauptdokuments und aller Unterdokumente.
106H	4	Bytes in der Absatzformatierungstabelle
10AH	4	Offset in Table-Stream (reserviert, intern benutzt)
10EH	4	Zahl der reservierten Bytes für interne Benutzung
112H	4	Offset in Table-Stream mit Font-Informationen (STTBF)
116H	4	Bytes in der Struktur (STTBF)
11AH	4	Offset in Table-Stream mit Feldpositionen des Hauptdokuments (FLD PLC)

Offset	Bytes	Bemerkungen
11EH	4	Bytes in dieser Struktur
122H	4	Offset in Table-Stream mit Feldpositionen im Header-Unterdokument (FLD PLC)
126H	4	Bytes in dieser Struktur mit Feldpositionen im Header-Unterdokument
12AH	4	Offset in Table-Stream mit Feldpositionen im Fußnoten-Unterdokument (FLD PLC)
12EH	4	Bytes in dieser Struktur mit Feldpositionen im Fußnoten-Unterdokument
132H	4	Offset in Table-Stream mit Feldpositionen im Anmerkungs-Unterdokument (FLD PLC)
136H	4	Bytes in dieser Struktur mit Feldpositionen im Anmerkungs-Unterdokument
13AH	4	unbenutzt
13EH	4	unbenutzt
142H	4	Offset in Table-Stream mit Struktur (STTBF), die Bookmark-Namen des Hauptdokuments enthält
146H	4	Bytes in der Struktur
14AH	4	Offset in Table-Stream in Struktur (PLCF), die die Zeichenpositionen (Offsets) auf den Anfang der Bookmarks im Hauptdokument enthält
14EH	4	Länge
152H	4	Offset in Table-Stream in Struktur (PLCF), die die Zeichenpositionen (Offsets) auf das Ende der Bookmarks im Hauptdokument enthält
156H	4	Länge
15AH	4	Offset in Table-Stream in Struktur mit Makrobefehlen
15EH	4	Länge »undocumented Structur«
162H	4	unbenutzt
166H	4	unbenutzt
16AH	4	unbenutzt
16EH	4	---
172H	4	Offset in Table-Stream mit Druckertreiberdaten (Treibername, Port etc.)
176H	4	Länge Struktur Druckerdaten in Byte
17AH	4	Offset in Table-Stream mit Druckerumgebung (Portrait-Modus)
17EH	4	Länge Struktur Druckerdaten in Byte
182H	4	Offset in Table-Stream mit Druckerumgebung (Landscape-Modus)
186H	4	Länge Struktur Druckerdaten in Byte
18AH	4	Offset in Table-Stream mit Window Save State-Datenstrukturen
18EH	4	Länge der Struktur in Byte
192H	4	Offset in Table-Stream mit Dokumenteigenschaften-Struktur
196H	4	Länge der Struktur in Byte

Offset	Bytes	Bemerkungen
19AH	4	Offset in Table-Stream mit (associated) Strings (beschreiben Summary-Info und Pfade zu speziellen mit dem Dokument verbundenen Dokumenten)
19EH	4	Länge der Struktur in Byte
1A2H	4	Offset in Table-Stream auf die Struktur, in der Informationen über Complex-Files hinterlegt sind (Schnellspeicherungsmodus)
1A6H	4	Länge der Struktur in Byte (0 = Dokument nicht im Schnellspeicherungsmodus hinterlegt)
1AAH	4	unbenutzt
1AEH	4	unbenutzt
1B2H	4	Offset in Table-Stream mit dem Namen der Originaldatei
1B6H	4	Länge der Struktur in Byte
1BAH	4	Offset in Table-Stream mit Strings, die die Besitzer (owner) der Anmerkungen enthält
1BEH	4	Länge der Struktur in Byte
1C2H	4	Offset in Table-Stream in Struktur, die Bookmarks für das Anmerkungs-Subdokument enthält
1C6H	4	Länge der Struktur in Byte
1CAH	4	Unbenutzt
1CEH	4	unbenutzt
1D2H	4	unbenutzt
1D6H	4	unbenutzt
1DAH	4	Offset in Table-Stream in FSPA PCC-Struktur des Hauptdokuments (0 = Dokument enthält keine Office-Art-Objekte)
1DEH	4	Länge der Struktur in Byte
1E2H	4	Offset in Table-Stream in FSPA PCC-Struktur des Header-Dokuments (0 = Dokument enthält keine Office-Art-Objekte)
1E6H	4	Länge der Struktur in Byte
1EAH	4	Offset in Table-Stream in Bookmark First-Struktur des Anmerkungs-Subdokuments
1EEH	4	Länge der Struktur in Byte
1F2H	4	Offset in Table-Stream in Bookmark Last-Struktur des Anmerkungs-Subdokuments
1F6H	4	Länge der Struktur in Byte
1FAH	4	Offset in Table-Stream mit Print Merge State-Informationen
1FEH	4	Länge der Struktur in Byte
202H	4	Offset in Table-Stream mit Formularfeldern (enthält Strings für Drop-down-Controls)
206H	4	Länge der Struktur in Byte
20AH	4	Offset in Table-Stream mit Endnote-Referenzen

Offset	Bytes	Bemerkungen
20EH	4	Länge der Struktur in Byte
212H	4	Offset in Table-Stream in Struktur, die auf den Endnote-Text zeigt
216H	4	Länge der Struktur in Byte
21AH	4	Offset in Table-Stream in Struktur mit Feldpositionen in Endnote-Subdokumenten
21EH	4	Länge der Struktur in Byte
222H	4	unbenutzt
226H	4	unbenutzt
22AH	4	Offset in Table-Stream mit Office-Art-Objekt-Datentabelle
22EH	4	Länge der Struktur in Byte
232H	4	Offset in Table-Stream mit Kürzel für Autorennamen, die Änderungen am Dokument vorgenommen haben
236H	4	Länge
23AH	4	Offset in Table-Stream mit Beschriftungen (Caption Titles) des Dokuments
23EH	4	Länge der Struktur in Byte
242H	4	Offset in Table-Stream der Struktur, die Objektnamen und Indizes für Objekte, die eine automatische Beschriftung (Auto-Caption) erhalten.
246H	4	Länge der Struktur in Byte
24AH	4	Offset in Table-Stream der Struktur, die die Abschnitte (Boundaries) der Teildokumente in einem Masterdokument (Zentraldokument) beschreibt
24EH	4	Länge der Struktur in Byte
252H	4	Offset in Table-Stream der Struktur, die die Stati der Rechtschreibprüfung beschreibt
256H	4	Länge der Struktur in Byte
25AH	4	Offset in Table-Stream der PLCF-Struktur, die die Anfangspositionen des ersten Zeichens der einzelnen Text in einem Textfeld (Textbox) beschreibt
25EH	4	Länge der Struktur in Byte
262H	4	Offset in Table-Stream der Struktur, die Feldgrenzen (Field Boundaries) in Textfeldern (Textboxes) beschreibt
266H	4	Länge der Struktur in Byte
26AH	4	Offset in Table-Stream der PLCF-Struktur, die die Anfangspositionen der Zeichen in Header-Textbox-Unterdokumenten beschreibt
26EH	4	Länge der Struktur in Byte
272H	4	Offset in Table-Stream der FLD PLCF-Struktur, die Feldgrenzen (Field Boundaries) in der Header-Textbox beschreibt
276H	4	Länge der Struktur in Byte
27AH	4	Makrobenutzer
27EH	4	Länge der Struktur in Byte

Offset	Bytes	Bemerkungen
282H	4	Offset in Table-Stream mit eingebetteten True-Type-Fontdaten
286H	4	Länge der Struktur in Byte
27AH	4	unbenutzt
27EH	4	unbenutzt
292H	4	Offset in Table-Stream mit Struktur, die Seitenbeschreibungen im Haupttext enthält
296H	4	Länge der Struktur in Byte
29AH	4	Offset in Table-Stream mit Struktur, die Seitenumbrüche (Break Descriptors) im Haupttext enthält
29EH	4	Länge der Struktur in Byte
2A2H	4	Offset in Table-Stream mit Struktur, die Seitenbeschreibungen im Fußnotentext enthält
2A6H	4	Länge der Struktur in Byte
2AAH	4	Offset in Table-Stream mit Struktur, die Seitenumbrüche (Break Descriptors) im Fußnotentext enthält
2AEH	4	Länge der Struktur in Byte
2B2H	4	Offset in Table-Stream mit Struktur, die Seitenbeschreibungen im Endnotentext enthält
2B6H	4	Länge der Struktur in Byte
2BAH	4	Offset in Table-Stream mit Struktur, die Seitenumbrüche (Break Descriptors) im Endnotentext enthält
2BEH	4	Länge der Struktur in Byte
2C2H	4	Offset in Table-Stream mit STTBF-Struktur, die Feld-Schlüsselwörter (Field Keywords) enthält (nur in internationalen Word-Versionen benutzt)
2C6H	4	Länge der Struktur in Byte
2CAH	4	Offset in Table-Stream mit Struktur, die Mailer-Routing-Slip beschreibt
2CEH	4	Länge der Struktur in Byte
2D2H	4	Offset in Table-Stream mit Struktur, die die Benutzernamen enthält, die das Dokument in alternative Speicherorte gesichert haben
2D6H	4	Länge der Struktur in Byte
2DAH	4	Offset in Table-Stream mit Struktur, die Dateinamen enthält, die im Dokument referenziert werden
2DEH	4	Länge der Struktur in Byte
2E2H	4	Offset in Table-Stream mit List-Format-Informationen
2E6H	4	Länge der Struktur in Byte
2EAH	4	Offset in Table-Stream mit List-Format-Override-Informationen
2EEH	4	Länge der Struktur in Byte
2F2H	4	Offset in Table-Stream mit Textbox-Break-Table des Hauptdokuments

Offset	Bytes	Bemerkungen
2F6H	4	Länge der Struktur in Byte
2FAH	4	Offset in Table-Stream mit Textbox-Break-Table des Header-Dokuments
2FEH	4	Länge der Struktur in Byte
302H - 32AH	4	Zeiger in Main-Stream mit undokumentierten Undo-Daten 4-Byte-Zeiger auf Struktur 4-Byte-Länge der Struktur in Byte
332H	4	Offset in Table-Stream mit String-Tabelle der Style-Namen für Glossareinträge
336H	4	Länge der Struktur in Byte
33AH	4	Offset in Table-Stream mit undokumentierten Grammatik-Optionen
33EH	4	Länge der Struktur in Byte
342H	4	Offset in Table-Stream mit undokumentierten OCX-Daten
346H	4	Länge der Struktur in Byte
34AH	4	Offset in Table-Stream in die Character Property Bin-Table (Zeichenformatierung)
34EH	4	Länge der Struktur in Byte
352H	4	FILETIME (low Date/Time)
356H	4	FILETIME (low Date/Time)
35AH	4	Offset in Table-Stream in LVC PLCF-Struktur
35EH	4	Länge der Struktur in Byte
362H	4	Offset in Table-Stream in Auto-Summary-Struktur
366H	4	Länge der Struktur in Byte
36AH	4	Offset in Table-Stream mit Struktur für Grammatik-Prüfungsstatus
36EH	4	Länge der Struktur in Byte
372H	4	Offset in Table-Stream mit Listennamen-String-Tabelle
376H	4	Länge der Struktur in Byte
37AH	4	Offset in Table-Stream mit undokumentieren Undo-/Versionsdaten
37EH	4	Länge der Struktur in Byte

Tabelle 8.2 Struktur des Word 97 File Information Blocks

Die über die Zeiger im File Information Block referenzierten Datenstrukturen des Streams enthalten dann Informationen zur Formatierung des Textbereichs oder zur Verwaltung der Daten.

Aufbau des Textbereichs

Der Text beginnt in der Datei an der im File Information Block (FIB) im Feld *fcMin* angegebenen Position. Dies ist in der Regel die auf den FIB folgende nächste freie Adresse, die auf einer 128-Byte-Grenze liegt. Der Text selbst wird im ASCII-Code hinterlegt. Hierbei werden einige ASCII-Codes als Steuerzeichen interpretiert.

- Absatzwechsel werden durch den Code 13 (0DH) markiert.
- Feste Zeilenumbrüche werden durch den Code 11 (0BH) markiert.
- Feste Trennzeichen (Breaking Hyphens) besitzen den Code 45 (2DH). Bedingte Trennstriche (Non-required Hyphens) erhalten den Code 31 (1FH). Geschützte Bindestriche (Non-Breaking Hyphens) werden mit dem Code 30 (1EH) gespeichert.
- Geschützte Leerzeichen (Non-Breaking Spaces) erhalten den Code 160 (A0H) und normale Leerzeichen den Code 32 (20H).
- Seiten- und Abschnittswechsel erhalten den Code 12 (0CH), was dem Form-Feed-Zeichen entspricht. Die Unterscheidung zwischen Seitenwechseln und Abschnittswechseln erfolgt über einen Eintrag in der Abschnittstabelle (Section Table).
- Spaltenwechsel erhalten das Zeichen 14 (0EH), und Tabulatoren werden mit dem Code 9 (09H) hinterlegt.
- Feldmarkierungen werden mit dem Zeichen 19 (13H) eingeleitet und mit dem Zeichen 21 (15H) abgeschlossen. Der Feldseparator ist der Code 20 (14H).
- Das Zeichen 7 markiert das Ende einer Tabellenzelle oder einer Tabellenzeile. Die Unterscheidung zwischen Zelle und Zeile erfolgt über die Absatzeigenschaften (Paragraph Property).

Zusätzlich benutzt Word noch einige ASCII-Codes als spezielle Zeichen, falls die Zeicheneigenschaft des Zeichens auf »spezial« gesetzt wurde.

ASCII-Code	Spezialzeichen
0	Aktuelle Seitennummer
1	Bild
2	Fußnotenreferenz (automatisch numeriert)
3	Trennzeichen Fußnote
4	Fortsetzungszeichen Fußnote
5	Verweis auf Anmerkung
6	Zeilennummer
7	handgezeichnete Anmerkung (Pen Windows)
8	Zeichenobjekt
10	abgekürztes Datum »Die. 1. Dez. 2000"
11	Zeitangabe
12	aktuelle Abschnittsnummer
14	abgekürzte Wochenangabe (Mo, Die etc.)
15	Wochentag (»Montag«)
16	Tag, kurze Angabe (»9«)

ASCII-Code	Spezialzeichen
22	Stunde aktuelle Zeit (ohne führende Null)
23	Stunde aktuelle Zeit (ggf. mit führender Null)
24	Minute aktuelle Zeit (ohne führende Null)
25	Minute aktuelle Zeit (ggf. mit führender Null)
26	Sekunde aktuelle Zeit
27	AM/PM aktuelle Zeit
28	aktuelle Zeit (altes Format)
28	aktuelle Zeit (altes Format)
28	aktuelle Zeit (altes Format)
29	Datum mit vollem Monatsnamen
30	Kurzdatum (12.4. 00)
33	Kurzangabe Monat (12 für Dezember)
34	Jahr 4 Ziffern
35	Jahr 2 Ziffern
36	abgekürzter Monat (»Apr«)
37	Monat ausgeschrieben
38	Uhrzeit in Std:Min ohne führende 0 bei Std.
39	Datum lang
41	Hilfsfeld Print Merge

Tabelle 8.3 Spezialzeichen im Textbereich

Bei Abschnittswechseln wird gleichzeitig auch ein Absatzwechsel eingeleitet. Das letzte Zeichen im Dokument ist immer ein Absatzwechsel. Ist der Text nicht im Complex-Format gespeichert, wird die Lage des Text-Streams des Dokuments durch die Felder *fib.fcMin* im Header bis zu *fib.fcMac* angegeben. Im komplexen Format muß das Dokument durch Tabellen, die Textfragmente enthalten, zusammengesetzt werden (wird hier nicht behandelt).

Neben dem eigentlichen Dokumenttext finden sich im Textbereich auch Kopf- und Fußzeilen sowie Anmerkungen etc. Die Größe des Hauptdokuments, der Kopfzeilen, der Fußzeilen etc. wird im File Information Block ab Offset 4CH in eigenen Feldern (*fib.ccpText, fib.ccpFtn, fib.ccpHdr* etc.) angegeben. Zum Zugriff auf Fußnoten etc. müssen Sie einfach die entsprechenden Längenangaben zum Zeiger für den Textanfang hinzuaddieren.

Zeichen- und Absatzformatierung

Zeichen- und Absatzformate werden in Word-Dokumenten in einem komprimierten Format hinterlegt. Die in der DOC-Datei enthaltenen Daten entsprechen daher nicht den Formateigenschaften für Zeichen oder Absätze im Text, sondern geben die Abweichungen in der Formatierung zu einem Referenzformat an. Word benutzt Strukturen, um die Formatierung zu hinterlegen. Eine PAP (Paragraph Property) ist eine Datenstruktur, die die unkomprimierten Absatzformate enthält. Ähnliches gilt für eine CHP (Character Property). Jeder Absatz im Dokument erhält einen Standardsatz an Absatz- und Zeichenformaten über die Formatvorlage (Style-Sheet-Datenstruktur) zugewiesen.

Eine PAP wird in die komprimierte Form PAPX überführt, indem die aktuellen Absatzformate mit den Formaten in der Formatvorlage des Absatzes verglichen werden. Weicht eine Absatzeigenschaft von der Vorlage ab, wird diese in der Liste als *sprms*-Eintrag hinterlegt. Ein *sprm* ist eine Anweisung, mit der eine oder mehrere Eigenschaften eines Zeichens, eines Absatzes, einer Tabelle etc. verändert werden können. *Sprm* ist dabei ein 2-Byte-Opcode ab Offset 0, der die auszuführende Operation definiert. Bei Bedarf lassen sich für die Operation erforderliche Informationen als Parameter fester Länge anhängen (ab Offset 2). Es gibt *sprm*-Einträge fester und variabler Länge. Liegt eine variable Länge für *sprm* vor, wird die Zahl der Bytes im folgenden Parameter ab Offset 2 und der Parameter ab Offset 3 gespeichert.

Bei Zeichenformaten wird ein ähnlicher Ansatz gewählt. Zuerst muß das Zeichenformat der Standardvorlage bekannt sein. Dann werden die Abweisungen der CHPX (komprimierte Zeicheneigenschaften) benutzt, um die individuelle Formatierung zu ermitteln. Auch hier werden *sprms*-Strukturen zur Beschreibung der Zeichenformate eingesetzt. Ein CHPX FKP und ein PAPX FKP besitzen ähnliche Strukturen. Ein FKP ist eine 512-Byte-Datenstruktur, die als Seite in einer Word-Datei hinterlegt wurde.

Anmerkung: Die einzelnen Datenstrukturen in der Word-Datei sind recht komplex. Aus Platzgründen ist es an dieser Stelle nicht möglich, die gesamten Strukturen der Streams sowie die Algorithmen zur Bestimmung der Textpositionen zu beschreiben. Entwickler können die betreffenden Spezifikationen direkt von Microsoft anfordern (*officeff@microsoft.com*). Registrierte Leser finden auf der Webseite des Verlages einen Link, über den sich die Formatbeschreibungen direkt abrufen lassen (sowohl für Word 6.0 als auch für Word 97).