Digitalisierung

"Wer in der Zukunft leben will, muss in der Vergangenheit buchstabieren." (André Malraux)

1. Digitale Dokumente

Heutzutage wird der größte Teil der Medien digital erzeugt, angefangen von Textdokumenten, über Fotografien bis zu Musikstücken.

Digitale Dokumente besitzen gegenüber den klassischen analogen Dokumenten eine Reihe von Vorteilen:

- Möglichkeit der verlustfreien Kopie,
- sehr kompakte Speicherung,
- kein Verschleiß durch Benutzung oder Lagerung,
- einfachere Erschließung.

Bei klassischen Kopierverfahren wie dem Fotokopieren einer Textvorlage oder dem Kopieren eines Tonbandes ist die Kopie immer etwas schlechter als das Original. Eine digitale Kopie ist zu 100% mit dem Original identisch, kann also ihrerseits wieder als Vorlage für eine Kopie dienen.

Die Speicherung von digitalen Medien ist in der Regel auch sehr kompakt. Ein Buch wie die Bibel besteht aus etwa acht Millionen Buchstaben und wird in der Regel auf sehr dünnem Papier gedruckt. Auf einer CD oder einen USB-Stick von nicht einmal 5cm Länge lassen sich fast einhundert derartiger Dokumente gleichzeitig speichern, auf einer DVD sogar noch deutlich mehr.

Digitale Textdokumente lassen sich im Volltext durchsuchen, es gibt inzwischen auch Verfahren um Bilddokumente noch bestimmten Mustern zu durchsuchen. Dadurch lassen sich die Inhalte von digitalen Dokumenten auch ohne arbeitsaufwändige Verschlagwortung einigermaßen erschließen.

Für die Vorteile der digitalen Dokumente kauft man sich aber auch eine Reihe von Nachteilen bzw. Risiken ein:

- Die Träger der digitalen Information haben sich in den letzten Jahrzehnten oft und stark verändert,
- die Lebensdauer der Datenträger liegt nur in der Größenordnung von Jahren bzw. Jahrzehnten,
- die Dateiformate unterliegen einem ständigen Wandel.

Die abgebildete 5,25" Diskette aus dem Jahr 1986 enthält mehrere digital erzeugte Textdokumente aus dieser Zeit. Die Probleme lassen daran recht deutlich zeigen.
Das erste Problem besteht darin ein passendes Laufwerk für diese Disketten im 5,25" Format zu finden. Falls ein entsprechendes Laufwerk vorhanden ist besteht das nächste Problem das Dateiformat der Diskette zu lesen. Im Zusammenhang mit 5,25" Disketten gibst es sehr unterschiedliche Formate, z.B. für Amiga, Commodore oder IBM-PC. Die dargestellten Diskette wurde auf einem Commodore 8032 erstellt, ist auf einen aktuellen Rechner kaum lesbar, selbst wenn er über ein passendes Laufwerk verfügen sollte. Selbst wenn es gelingt das Dateisystem zu lesen, so bleibt immer noch das Problem des Dateiformates. Die heute üblichen Textverarbeitungsprogramme und ihre Dateiformate gab es 1986 noch nicht. Beim Einlesen der Dokumente von dieser Diskette geht vermutlich alles verloren, was nicht reiner Text ist. Das ist in diesem Fall nicht weiter schlimm, abgesehen von den Umlauten hatte man damals nur einfache ASCII-Zeichen in den Textdokumenten.

Selbst wenn die technischen Voraussetzungen für das Einlesen der Diskette vorhanden sind, so ist nicht sicher, ob die Daten noch erhalten sind oder die Magnetisierung in den Jahren gelitten hat. Bei optimalen Lagerbedingungen kann eine solche Diskette auch nach Jahrzehnten durchaus noch lesbar sein. Bei den moderneren Datenträgern wie CD und DVD bestehen erhebliche Bedenken hinsichtlich der Haltbarkeit, die hängt hier sehr stark von der Qualität der Rohlinge ab.

Wenn man sich bei der Digitalisierung der Risiken bewusst und zu einem regelmäßigen Umkopieren auf aktuelle Datenträger und Dateisysteme bereit ist, dann überwiegen die Vorteile die Nachteile erheblich.

2. Digitalisierung von Dokumenten

Die Digitalisierung auch von historischen Dokumenten ist eine der großen Aufgaben unserer Zeit. Die Digitalisierung macht die Dokumente der Allgemeinheit zugänglich, ohne dass ihnen bei der Nutzung ein Schaden entstehen kann.

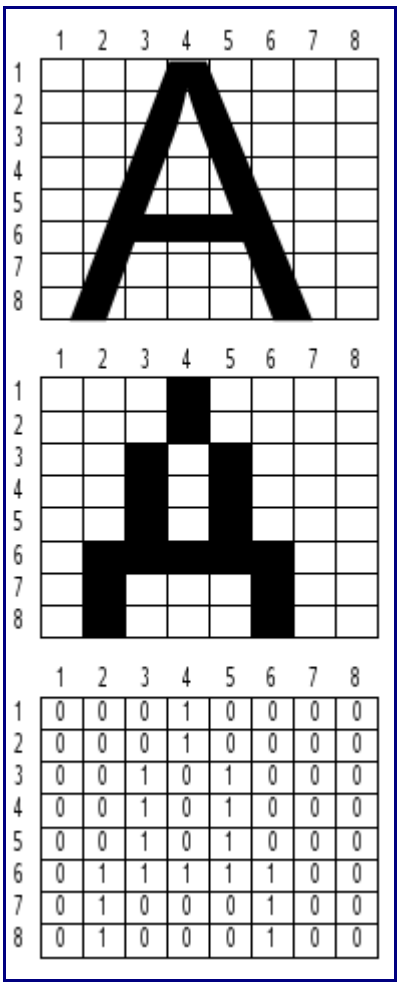
Eine gewisse Vorreiterrolle spielt hier die „Library of Congress“, die Bibliothek des Amerikanischen Kongresses. Zum zweihundertsten Geburtstag im Jahr 2000 wurden etwa fünf Millionen Dokumente zur amerikanischen Geschichte und Kultur digitalisiert und im World Wide Web online in Form einer "National Digital Library" (http://www.loc.gov) zur Verfügung gestellt. In Deutschland wäre das Projekt Gutenberg zu nennen, welches unter http://gutenberg.spiegel.de/ etwa 80.000 Dokumente digital zur Verfügung stellt.

Bei jeder Form von Digitalisierung muss auch beachtet werden, mit welchem Ziel die Digitalisierung erfolgt. Geht es primär darum das Dokument zu sichern, oder darum es zugänglich zu machen. Bei der Sicherung darf es zu keinerlei Informationsverlusten kommen, das Datenvolumen spielt dabei nur eine untergeordnete Rolle. Wenn es darum geht Dokumente zugänglich zu machen, eventuell sogar über das WWW, dann spielt das Datenvolumen eine große Rolle. Zu große Dokumente lassen sich im WWW kaum oder nur mit entsprechenden Kosten übertragen.

3. Digitalisierung von Bildern

Flache Originale, wie Fotos, Karten oder Grafiken lassen sich mit relativ geringem Aufwand digitalisieren. Dazu benötigt man eine Gerät, welches die Vorlage einlesen kann, entweder einen Scanner oder ggf. auch einen digitalen Fotoapparat.

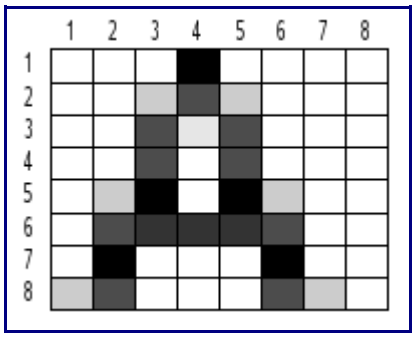
Ein Hochleistungs-Scanner



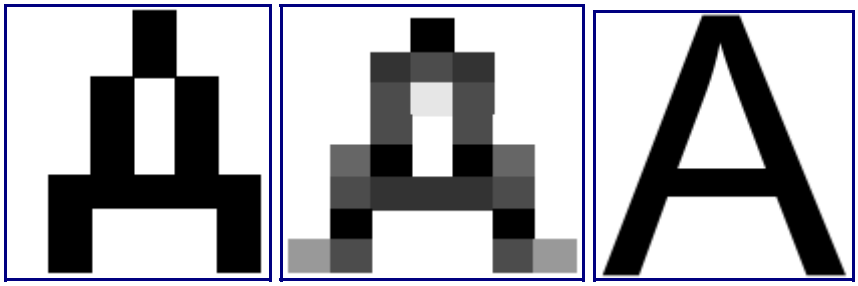
Der Scanner legt ein virtuelles Gitter über die Vorlage, hier den Buchstaben „A“, und wertet für jedes einzelne Feld in diesem Gitter aus, welche Farbe der entsprechende Bildpunkt besitzt.

Das Ergebnis für den einzelnen Bildpunkt hängt von den Einstellungen für die Farbtiefe ab. Wurde hier nur „Schwarz/Weiß“ ausgewählt, so hat der Scanner nur die Auswahl zwischen diesen beiden Farben. Für jedes einzelne Feld muss der Scanner also entscheiden, ob der Schwarzanteil einen Grenzwert überschreitet oder nicht. Felder die den Grenzwert nicht überschreiten werden als weiße Felder betrachtet.

Schwarze Bildpunkte werden dann mit einer „1“ kodiert und weiße Bildpunkte mit einer „0“.

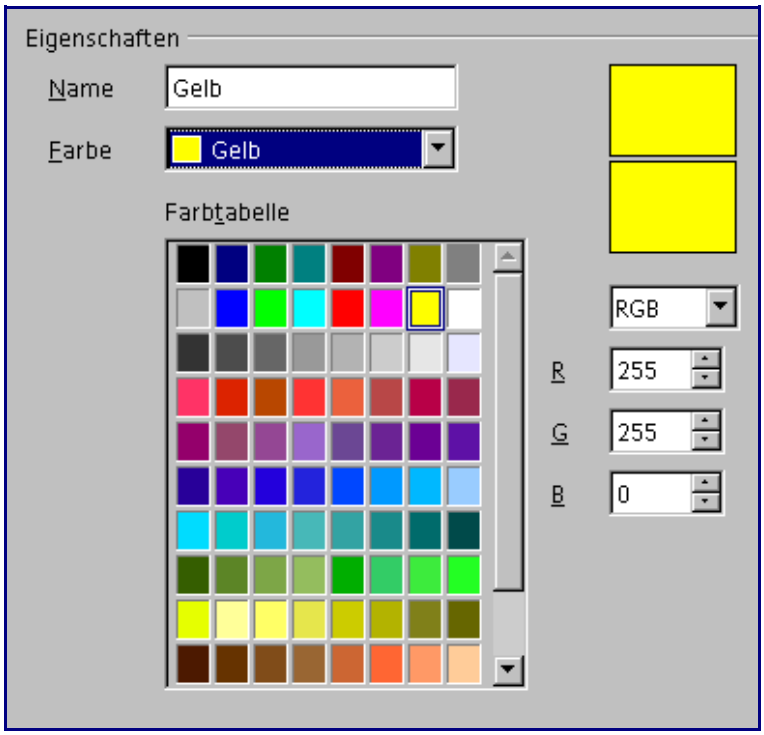


Bei Graustufen stehen ihm insgesamt 256 Abstufungen zwischen Schwarz und Weiß zur Verfügung. Schon das verringert bei Linienmustern die Treppeneffekte, die durch die leichte Unschärfe der Darstellung verringert werden. Dafür muss das Raster aber klein genug sein:

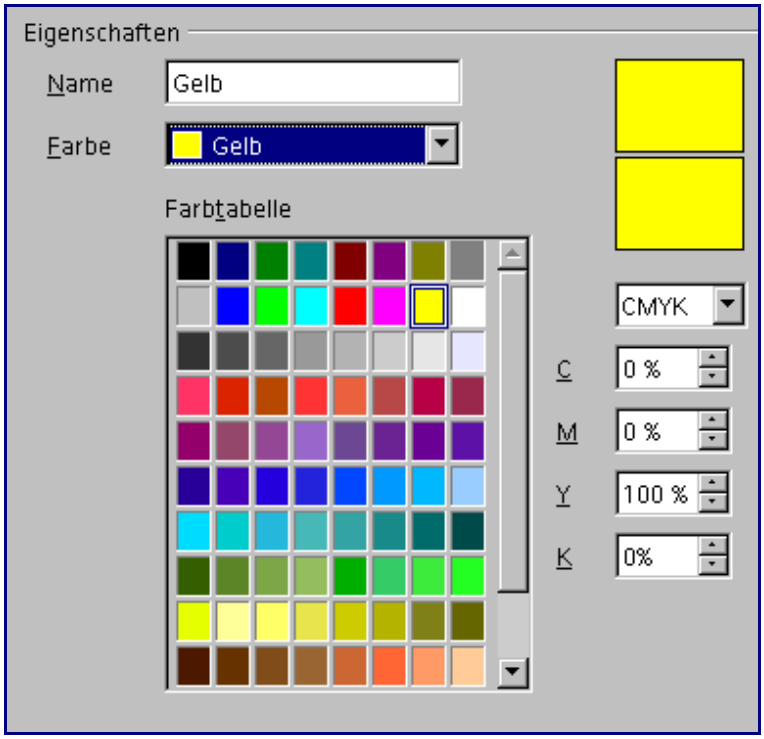


Bei Scans im Farbmodus sollte man 24 Bit Farbtiefe auswählen. Der Scanner wertet dann den Rot, den Grün und den Blauanteil des Farbpunktes getrennt aus. Für jede der drei Farben stehen ihm wieder 256 Abstufungen zur Verfügung. Der Scanner erzeugt dann für jeden Bildpunkt drei Byte an Informationen und kann 256*256*256 ≈ 16,8 Millionen Farben unterscheiden.

Die darstellbaren Farben lassen sich auf sehr unterschiedliche Arten beschreiben. Am PC am weitesten verbreitet ist das schon beschriebene RGB-Format, bei dem jeder Bildpunkt durch Mischung der Grundfarben Rot, Grün und Blau gekennzeichnet ist und das mit Werten zwischen 0 und 255 operiert.



Im Druckbereich weit verbreitet ist dagegen das Modell CMYK, das auf den Grundfarben Zyan, Magenta, Gelb und Schwarz basiert. Die Zahlenangaben werden als Prozentwert im Bereich zwischen 0 und 100 interpretiert.



Eine Umrechnung zwischen den beiden Formaten ist direkt möglich. Das im Beispiel dargestellte reine Gelb hat im RGB-Format die Werte (255, 255, 0) und im CMYK (0, 0, 100, 0).

Neben RGB und CMYK gibt es noch weitere Farbmodelle, die für spezielle Anwendungen eine Rolle spielen.

Neben der Farbtiefe ist die Auflösung beim Scannen eine wichtige Größe. Die Auflösung wird hier in der Regel in „dots per inch“ (dpi) angegeben, was soviel bedeutet wie „Punkte pro Zoll“ (1 Zoll = 2,54cm). Die dpi-Zahl gibt also an, wie viele Felder das virtuelle Gitter pro Zoll besitzt. Eine Angabe von 300 dpi besagt also, dass auf 2,54 cm jeweils 300 Gitterpunkte kommen.

Ein Wert von 300 dpi ist heute als Auflösung bei Druckern üblich, bessere Geräte kommen auf 600 dpi. In etwa dieser Größenordnung sollte eine Scanauflösung auch liegen. Aktuelle Scanner liefern sogar deutlich höhere Auflösungen, 2400 dpi sind keine Seltenheit, das erlaubt dann auch die Vergrößerung von Ausschnitten.

Man sollte die Scanauflösung niemals unnötig hoch wählen, sonst ergibt sich ein enormer Speicherbedarf. Beim Scannern einer kompletten DIN-A4 Seite (21cm x 29,7 cm) mit 300 dpi ergeben sich (21*300)/2,54=2.480 Punkte in der Breite und (28,7*300)/2,54=3.507 Punkte in der Höhe, insgesamt also 2.480 x 3.507 = 8.697.360 Bildpunkte. Bei einem Scan mit 24 Bit Farbtiefe ergibt sich so ein Speicherbedarf von etwa 25 MByte für eine Seite. Bei einer Steigerung der Scanner-Auflösung auf 600 dpi vervierfacht sich der Speicherbedarf schon auf 100 MByte.

4. Grafikformate

Beim Speichern einer Grafik, wie sie ein Scanner oder ein digitaler Fotoapparat liefert, ist es wichtig das richtige Dateiformat zu wählen. Die Zahl der verfügbaren Dateiformate ist riesig, es dürften über einhundert verschiedene Formate existieren.

Zum Glück spielen von diesen vielen Formaten nur einige eine Rolle. Ein gutes Dateiformat für die Archivierung sollte auf alle Fälle möglichst unabhängig von einem Betriebssystem sein und einen einfachen Datenaustausch über Betriebssystemgrenzen und Anwendungsgrenzen hinweg erlauben. Für viele Zwecke ist es auch nützlich, wenn das Dateiformat komprimiert ist, was den Bedarf an Speicherplatz und Übertragungszeit reduziert. Das Beispielbild benötigt bei 1024x768 Bildpunkten und 24Bit (3Byte) Farbtiefe 1024x768x3Byte =2359296Byte = 2304 kByte ohne Komprimierung.

volle Auflösung 1024x768 Punkte (166 kByte als .jpg)

Das „**Tagged Image File Format**“ (tiff oder auch tif) ist ein flexibles Dateiformat für Bilder und Zeichnungen. Viele digitale Fotoapparate liefern dieses Format, wenn höchste Qualität gefordert wird. Üblicherweise benötigt eine TIFF-Datei genau so viel Speicherplatz, wie die theoretische Berechnung ergibt. Ein Bild wie das nebenstehende mit einer Auflösung von 1024x768 Bildpunkten benötigt mit 24 Bit Farbtiefe ziemlich genau 2,3 MByte.

volle Auflösung 1024x768 Punkte (2304 kByte als .tif)

```
Hinweis: Manche Browser können Bilder im TIF-
Format nicht darstellen, normalerweise wird
dann ein Download des Bildes angeboten.
```

Speziell in digitalen Netzen sehr weit verbreitet war lange Zeit das „**Graphics Interchange Format**“ (gif).Dieses Dateiformat verfügt über eine verlustfreie Komprimierung, die den Speicherbedarf erheblich reduziert. Das Beispielbild belegt in diesem Format nur etwa 446 kByte, das ist eine Reduzierung um mehr als den Faktor 5. Leider erlaubt das Dateiformat keine 24-Bit sondern nur 8-Bit, so dass die Werte nicht ganz vergleichbar sind. Auf Webseiten sehr verbreitet ist das GIF-Format, aufgrund zweier spezieller Eigenschaften. Es unterstützt Animationen, hierzu werden mehrere Bilder in der gleichen Datei abgespeichert und wie bei einem Daumenkino nacheinander abgespielt. Weiter unterstützt es auch transparente Darstellungen, wobei eine Farbe des Bildes als transparent deklariert wird. Die erlaubt Grafiken, die so wirken, als ob sich nicht rechteckig wären, sondern rund oder gezackt, einfach weil der überflüssige Bereich durchsichtig ist. In den letzten Jahren hat die Bedeutung des Formates sehr nachgelassen, weil es Probleme mit einem Patent für das Kompressionsverfahren gab. Dieses Patent ist inzwischen abgelaufen, trotzdem wird oft lieber zum moderneren PNG-Format gegriffen.

volle Auflösung 1024x768 Punkte (446 kByte als .gif)

Das „**Portable Network Graphics Format**“ (png) wurde gezielt als Ersatz für das gif-Format geschaffen. Es wurde darauf geachtet, dass es keine patentrechtlichen Probleme gibt. Gedacht ist dieses Format vor allem für den Einsatz in Netzen wie dem Internet. Die Farbtiefe darf bis zu 48-Bit betragen und die Kompression ist ebenfalls verlustfrei. Die Dateigröße für die Beispielgrafik beträgt in diesem Format trotzdem nur etwa 709 kByte.

volle Auflösung 1024x768 Punkte (709 kByte als .png)

Vor allem im Zusammenhang mit digitalen Fotos weit verbreitet ist das jpg-Format der „**Joint Photographics Expert Group**“. Dieses Format ist vor allem auf eine hohe Komprimierung von Fotos ausgelegt. Dazu benutzt es eine verlustbehaftete Komprimierung, die die Unzulänglichkeiten des menschlichen Auges mit einbezieht. Das menschliche Auge kann nur relativ wenige Farben voneinander unterscheiden, vor allem wenn sie benachbart sind. Einfach ausgedrückt lässt das Kompressionsverfahren einfach alle Informationen weg, die dem Auge sowieso nicht auffallen. Beim erneuten Laden des Bildes lässt sich dieser Informationsverlust nicht mehr beseitigen. Man sollte daher Bilder die man bearbeiten möchte nicht in diesem Format speichern, da mit jeder neuen Speicherung erneut komprimiert wird, mit eventuell weiteren Informationsverlusten. Mit der Bei-

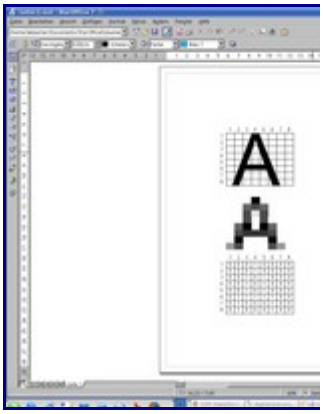
spielgrafik bringt es dieses Dateiformat auf eher enttäuschende 100 kByte. Das hängt damit zusammen, dass es sich um eine Grafik handelt und nicht eine Foto. Bei Grafiken liegen größere gleichfarbige Bereiche vor, hier haben die Kompressionsverfahren von png und gif ihre Stärken. Bei Fotos haben benachbarte Bildpunkte fast immer unterschiedliche Farben, hier ist das jpg-Format deutlich stärker als die anderen Formate. Die Kompressionsrate im jpg-Format ist wählbar, wobei mit höheren Kompressionsraten auch die Informationsverluste wachsen.

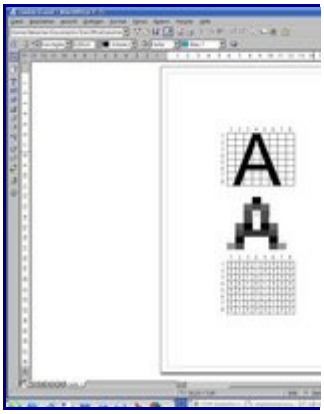
166 kByte vom Fototapparat

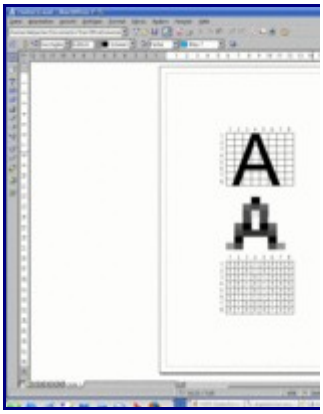
48 kByte bei 50% Qualität

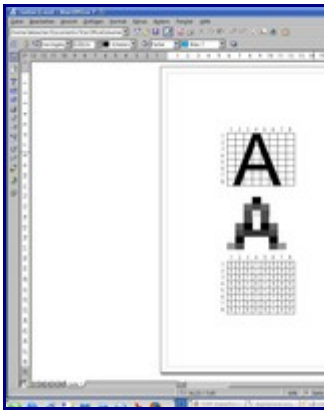
20 kByte bei 10% Qualität

Die Dateigröße hängt aber nicht nur vom jeweiligen Dateiformat ab, sonder auch von der Art der Vorlage. Fotos, wie im hier gewählten Beispiel, lassen sich nur schlecht in den Formaten GIF und PNG komprimieren. Für Fotos sollte man immer das JPG-Format wählen. Zeichnungen, vor allem am Computer erstellte, lassen sich dagegen effektiver im GIF oder PNG Format speichern. Als Beispiel hierfür soll eine Bildschirmkopie aus der Textverarbeitung dienen. Diese Abbildung ist 1000x1000 Punkte groß und belegt bei 3 Byte Farbtiefe als 1000*1000*3Byte=3.000.000 Byte = 2.930 kByte = 2,9 MByte.


3000 kByte als .tif


106 kByte als .jpg


91 kByte als .gif

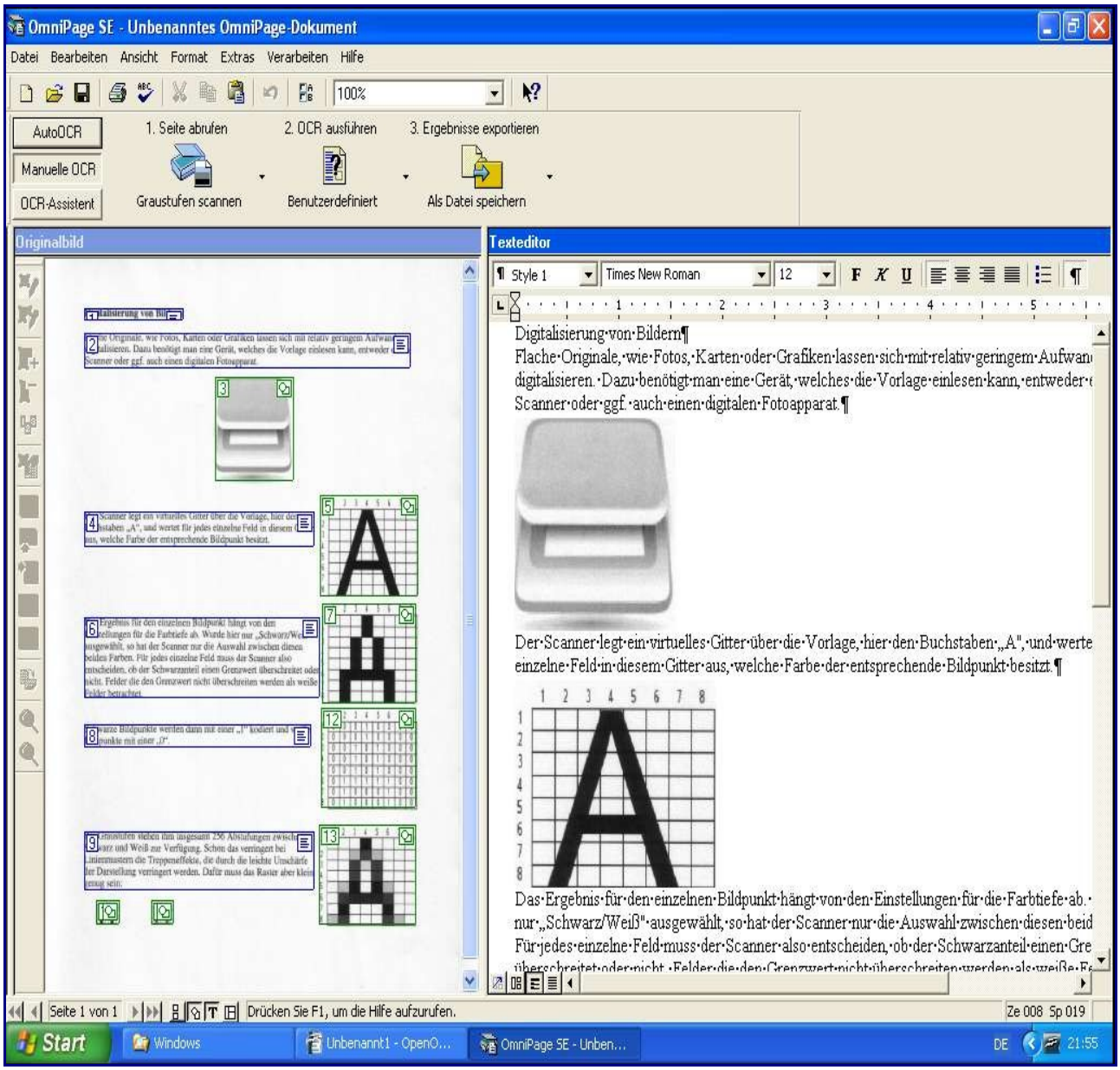

50 kByte als .png

5. Texterkennung

Auf einer normalen Textseite befinden sich üblicherweise zweitausend bis dreitausend Buchstaben, was einen entsprechenden Speicherbedarf ergibt. Will man diese Textseite nachträglich digitalisieren, so ergibt sich beim Scannen eine Speicherbedarf von 25 Mbyte und das Ergebnis ist „nur“ eine Grafik, kann also weder einfach bearbeitet noch durchsucht werden. Will man aus der Grafik wieder eine Textdatei erhalten, so braucht man eine Software zur „Optical Character Recognition“ (OCR) eine optische Texterkennung.

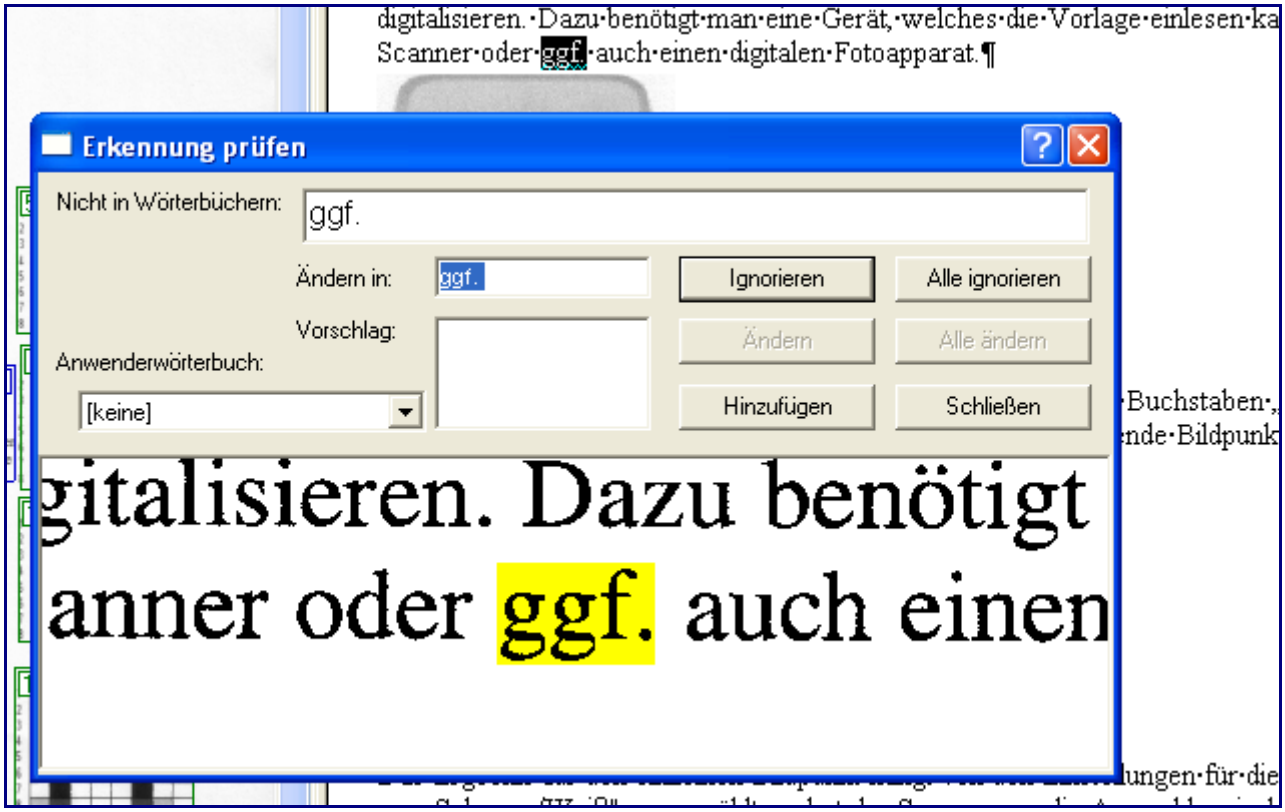
Derartige Programme können die Punktmuster analysieren, die der Scanner liefert und ihnen die ursprünglichen Buchstaben wieder zuordnen. Das funktioniert aber nur dann gut, wenn die Vorlage eine hohe Qualität hat und mit einer Maschine in einer einfachen Schrift erstellt wurde. Die Erkennung von Handschriften oder ungewöhnlichen Druck-Schriften bereitet den Programmen erhebliche Probleme.

Moderne Programme wie „Abby Finereader“ oder „Omnipage“ führen bei geeigneten Vorlagen zu erstaunlichen Ergebnissen.



Im linken Teil des Bildschirmausschnittes ist die Seite als Grafik zu sehen, so wie die Software sie eingelesen hat. Jeder Textbereich und jede Abbildung ist mit einem Rahmen versehen. Im rechten

Teil ist dann das Ergebnis der OCR zu sehen, ein Dokument welches mit jeder Textverarbeitung weiter verarbeitet werden kann.



Moderne OCR-Software erhöht die Erkennungs-Qualität durch die Benutzung von Wörterbüchern. Falls sie ein Wort nicht erkennt oder einfach nicht kennt, so fragt sie beim Benutzer nach und bietet ihm eine Korrekturmöglichkeit.

Eine Seite aus einer Zeitschrift oder einem Buch lässt sich in der Regel problemlos einlesen und es bleiben sogar Formatierungen erhalten. Die Zahl der Fehler liegt meist erheblich unter der Zahl der Fehler beim Abtippen der Vorlage.

abweichende Lizenzbestimmungen

Das vorliegende Dokument ist in Zusammenarbeit mit Volker Reissmann vom Staatsarchiv Hamburg entstanden. Die eingebundenen Fotos stammen von ihm und somit liegen die Rechte an den Fotos auch bei Volker Reissmann.