



doc|stream

Kostengünstige Altarchiverfassung am Beispiel von Patentschriften

Dieser Bericht über die Erfassung eines papierbasierten Altarchivs am Beispiel von Patentunterlagen ist unser Kompetenzbeweis als Dienstleister für die elektronische Erfassung von papierbasierten Altarchiven.

2.000 Aktenordner mit einem Gewicht von 6 t, mit 1,8 Mio. Seiten sowie einem Speichervolumen von 2 Terabyte, verarbeitet mit Stromkosten von 1.000 Euro sind nur einige der interessanten Kennzahlen unseres „Patent Capture Services“ (PCS).

Der Bericht beschreibt die erforderlichen Vorgehensschritte und schildert unerwartet aufgetretene Besonderheiten und deren Lösung sowie die Möglichkeiten von Outsourcing manueller Datenerfassung nach Indien.

Es ist aber auch ein Beispiel für die umfassenden Anforderungen an einen Scan-Dienstleister (Document Capture Services), der sich auf komplexe inhaltliche Erschließung, sprich „Automatische Indexierung“ spezialisiert hat. Für DocStream war das Projekt „Patent Capture Services (PCS)“ gleichzeitig ein hervorragendes Testbett für die in Entwicklung befindlichen Produkte der automatischen Interpretation von image- und textbasierten Dokumenten. DocStream erledigte diese Dokumentenerfassungs-Dienstleistung auf Basis eines weltweiten Technologiepartner-Netzwerkes und gibt damit ein Praxisbeispiel für die Möglichkeiten kostengünstiger globaler Kooperation auf dem Arbeitsgebiet der elektronischen Dokumentenerfassung. Nach dem Scannen der Patentakten in Deutschland erfolgten die Dokumentenklassifikation sowie die Informationsextraktion daraus auf Basis der von den Technologiepartnern beigetragenen Software-Arbeitspakete im Mix mit zugekauften Nutzungslizenzen marktgängiger Produkte. Die Aufgabe der manuellen Datenerfassung nicht automatisch lesbarer Informationen wurde von unserem Offshore-Partner erledigt.

Inhalt

Einleitung – Mit der Patentakte vom Keller ins Intranet.....	2
1. Aufgabenstellung und Anforderungen	3
2. Lösungskonzept und Vorgehen	9
3. Entwicklungsleistungen und Verarbeitungssystem	10
4. Arbeitsvorbereitung und Scannen.....	11
5. OCR-Verarbeitung, Indexierung und Datenextraktion.....	12
6. Offshore-Arbeiten.....	14
7. xml-Schema, DTD und xml-Zusammenführung.....	16

Mit der Patentakte vom Keller ins Intranet

Kostengünstige Altarchiverfassung zum Festpreis durch globale Kooperation

Die Micronas GmbH (Freiburg i. Br.) ist ein High-Tech-Unternehmen, welches im Jahr 2004 einen Umsatz von 624 Mio. Euro erzielte. Ihre ICs finden sich u.a. in den neuesten Generationen von Flachbildschirm-TV-Geräten aller bedeutenden Markenhersteller der Unterhaltungselektronik weltweit. Dem entsprechend wichtig ist es, dieses geballte Know-how durch Patente zu schützen, Einsprüchen zu entgegnen und das bereits in Patent-Schriftform gegossene Wissen zu verwalten. Dies ist die Aufgabe der Abteilung „Intellectual Property Rights Management“ der Micronas.

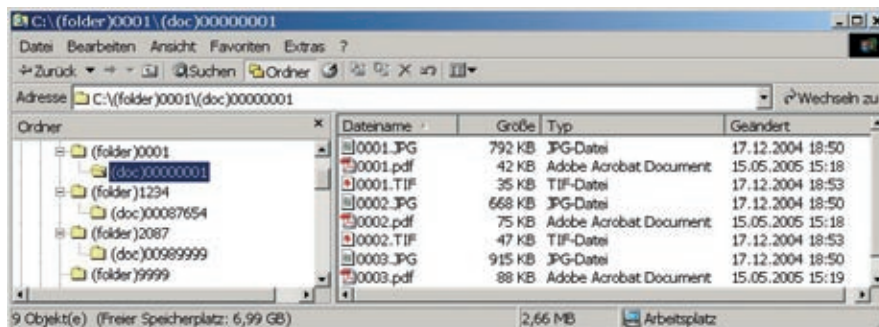
Micronas ist ein weltweit operierender Halbleiterentwickler und -hersteller, rund um den Globus forschen, entwickeln und produzieren z.Zt. 1.900 Mitarbeiter als eingespieltes Team. Das im Aufbau befindliche Patentrecherche-System wird deshalb via PC mit Internetbrowser für geschlossene Benutzergruppen weltweiten Zugriff erlauben, eben „Information at your Fingertips“ für alle.

Dabei wurde das papierbasierte Altarchiv von 2.000 Aktenordnern mit 100.000 Patentschriften, Gebrauchsmustern plus Patentliteratur nicht ausgespart. Es handelt sich um bereits veröffentlichte deutsche, europäische und US-Patente und PCT-Schriften mit unterschiedlichsten Layouts, im wesentlichen aus der Zeit von 1949 bis Anfang der 80-er Jahre. Nicht nur von historischem Interesse sind hier auch die firmeninternen Anmerkungen auf den Dokumenten, die dank Farbiges originalgetreu festgehalten werden.

Ein Musterbeispiel für anspruchsvolle Dokumentenerfassungs-Dienstleistung in Bild und Index.



1. Aufgabenstellung und Anforderungen



Das papierbasierte, historische Patentarchiv, das in elektronische Form überführt wurde, bestand aus:

- _ 1.600 Aktenordnern mit teilstrukturierten Dokumenten (16 Paletten je 100 Ordner)
- _ 400 Aktenordnern mit weitgehend unstrukturierten Dokumenten (4 Paletten je 100 Ordner)
- _ 50 Schubler Indexkarten


Die ca. 2.000 Aktenordner mit insgesamt 6.000 kg Gewicht, sollten aus Transportrisikogründen nicht komplett auf einmal, sondern in 10 Chargen von durchschnittlich 200 Ordnern zur Scanstelle transportiert werden.

Nach Rücksendung des Schriftgutes wurde dieses bis zur endgültigen Vernichtung in Kartons ausgelagert. In die nach der Verarbeitung nun frei gewordenen 3 Räume werden Micronas-Mitarbeiter einziehen.

Die wesentlichen technischen Projektvorgaben waren:

- _ Einsatz von Windows 2000/XP.
- _ Die Indexierung sollte in Form einer Kombination der in Verzeichnissen und Unterverzeichnissen enthaltenen Images mit einer xml-Struktur erfolgen, welche die Dateninhalte der Images wiedergibt.



N° 18,450  A.D. 1892

Date of Application, 14th Oct., 1892
Complete Specification Left, 19th May, 1893—Accepted, 5th Aug., 1893

PROVISIONAL SPECIFICATION.

Improvements in or applicable to Pianofortes, Organs, Harmoniums, and like Musical Instruments.

I, WILLIAM WILKENINGHAUS, of 13, Harnell Street, Falcon Square, in the City of London, Merchant, do hereby declare the nature of this invention to be as follows:—

This invention relates to the combination with pianofortes, organs, harmoniums and like keyed instruments of means whereby the act of playing will communicate to the performer a current of electricity.

The invention may be carried out in various ways, for example a battery may be placed in any convenient part of the instrument and be joined up to suitable conductors introduced into the face of the keys and assuming that one pole of the battery be connected to one half of the keys and the other pole to the remaining half, then as the instrument is played with both hands the circuit would be completed and a current would be passed through the body of the performer; or all the keys may be connected to one pole of the battery, and a suitable contact piece adapted to be applied to any convenient part of the person might be arranged in connection with the other pole of the battery.

Dated this 14th day of October 1892.

G. F. REDFERN & Co.,
4, South Street, Finsbury, London, Agents for the Applicant.

COMPLETE SPECIFICATION.

20 Improvements in or applicable to Pianofortes, Organs, Harmoniums, and like Musical Instruments.

I, WILLIAM WILKENINGHAUS, of 13, Harnell Street, Falcon Square, in the City of London, Merchant, do hereby declare the nature of this invention and in what manner the same is to be performed, to be particularly described and ascertained in and by the following statement:—

This invention relates to the combination with pianofortes, organs, harmoniums and like keyed instruments of means whereby the act of playing will communicate to the performer a current of electricity.

This invention may be carried out in various ways, for example, an electric battery may be placed in any convenient part of the instrument and be joined up to suitable conductors introduced into the face of the keys and assuming that one pole of the battery be connected to one half of the keys and the other pole to the remaining half, then as the instrument is played with both hands the circuit would be completed and a current would be passed through the body of the performer; or all the keys may be connected to one pole of the battery, and a suitable contact piece adapted to be applied to any convenient part of the person might be arranged in connection with the other pole of the battery.

To enable my invention to be fully understood I will describe the same by reference to the accompanying drawing, in which:—

40 Figure 1 is a sectional plan of a pianoforte having my improvements applied thereto;

Figure 2 is a side elevation partly in section on the line 2—2 Figure 1; and, Figure 3 is a section on the line 3—3 Figure 1;

[Price 8d.]

1

9  PATENTAMT.  ALBUMEN DES 3. SEPTEMBER 1890

Eigenschaft des Kaiserlichen Patentamt.

PATENTSCHRIFT

Doppel — № 113816 — reg. 10e3

KLASSE 48 A.

EDUARD MIES IN BÜDESHEIM, RHEINHESSEN.

Verfahren zum Niederschlagen von Metallen auf Aluminium.

Publiziert im Deutschen Reichs vom 16. December 1898 ab.

Das Verfahren vorliegender Erfindung bezieht sich auf die Gewinnung haltbarer Niederschläge von Metallen auf Aluminium.

Unter dem bekannten Verfahren, auf Aluminium Niederschläge anderer Metalle zu erzielen, findet sich keines, welches die wirklich directe Niederschlagsarbeit auf Aluminium zum Zweck hat; es soll z. B. erst Aluminium in Cyanquecklösungen vorbereitet werden, wodurch selbstverständlich eine Amalgamierung eintritt, und auf das Amalgam soll dann ein anderes Metall, z. B. Kupfer, niederschlagen werden. Dieses Verfahren ist für Demonstrationen geeignet, nicht aber für die Praxis. Metallniederschläge sollen auch einen großen Hitzegrad vertragen können, ohne den innigen Zusammenhang zu verlieren. Das trifft aber für obiges Verfahren nicht zu, da es bekannt ist, daß Amalgam beim Erhitzen das Quecksilber verliert; es sucht sich das Quecksilber beim Erhitzen irgendwo einen Weg, wodurch das auf dem Amalgam befindliche Metall unbedingt gelockert wird.

Bei einem andern Verfahren behandelt man Aluminium zuerst mit Kalilauge, ohne daß die Lauge abgeseigt wird, und soll dann direct der Metalniederschlag bewirkt werden. Die Nachteile eines solchen Verfahrens ergeben sich aus Folgendem:

Tascht man nämlich metallisches Aluminium in eine Lauge, z. B. Kalilauge, so bildet sich unter Wasseroxidation Kalialuminat oder Aluminiumoxydalkali nach der Formel:


$$Al_2 + 6 KOH = Al_2(O K)_3 + 3 H_2$$

Man ist somit gar nicht im Stande, auf dem metallischen Aluminium einen dauerhaften Niederschlag zu erzeugen, da die Lauge nicht abgewaschen werden soll und die ganze Metalloberfläche mit dem Aluminiumoxydalkali bedeckt ist. Bedeutet man ferner, daß noch Kalilauge von dem Aluminiumoxydalkali mechanisch zurückgehalten wird, so muß, da es nicht abgewaschen werden soll, unbedingt eine weitere Gussentwicklung stattfinden, welche ein festes Halten des Niederschlags auf dem Aluminium ausschließt. Scheint auch in manchen Fällen eine feste Verbindung des niederschlagenen Metalles auf Aluminium möglich, d. h. gleich nachdem der Überzug hergestellt, fest zu halten, dann tritt doch stets nach dem bisherigen Verfahren nach einiger Zeit ein Löslösen des Überzuges auf.

Diesen Uebelständen läßt das Verfahren vorliegender Erfindung ab, was von dem bisherigen Verfahren durchaus verschieden ist.

Der Erfinder hat in Erkenntnis der Mängel bisheriger Methoden sein Hauptaugenmerk darauf gerichtet, wirklich dauerhafte Niederschläge von Metallen auf Aluminium zu erzeugen, welche allen Einflüssen trotzen und allen Erwartungen, die man an galvanische Niederschläge zu stellen berechtigt ist, entsprechen. Bei diesem neuen Verfahren wird das metallische Aluminium, Blech, Draht etc. in einer siedenden, mit etwas Schwefelsäure angesäuerten Lösung von phosphorsaurer Natrium und schwefelsaurer Magnesia etwa 5 Minuten erhitzt, mit Wasser abgeseigt und

2

BUNDESREPUBLIK DEUTSCHLAND  DEUTSCHES PATENTAMT

Patentschrift DE 28 58 826 C 2

Reg. H 03 J 7/18 H 03 J 1/02 H 03 J 3/10

Aktenzeichen: P 28 58 826 C 2
Anmeldetag: 23. 12. 79
Offenlegungstag: 12. 7. 79
Veröffentlichungstag der Patentschrift: 12. 9. 98

Eintragsbuch-Nr. 1 2. DEZ. 1925

Innerhalb von 3 Monaten nach Veröffentlichung der Erteilung kann Einspruch erhoben werden

Uebersichtsprotokoll: 30.12.77 IT 8960A-77

Patentinhaber: Società Italiana per lo Sviluppo dell' Electronica S.I.S.V. S.p.A., Nove, Turin/Torino, IT

Vertreter: Eisenführ, Speiser & Partner, 28195 Bremen

Teil aus: P 28 58 042.1

Erfinder: Belsaoni, Pietro, Pinerolo, Turin/Torino, IT

Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:

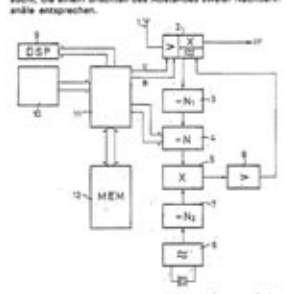
- DE 26 51 030 A1
- DE 28 48 682 A1
- Microcomputer tunes and holds frequencies, in US-Z.: Electronics, 1977, Sept., S. 46-48, 50;
- Ein Abstimmensystem für Fernsehgeräte in DE-Z.: Funk-Technik, 1976, Nr. 1/2, S. 4 u. 5;
- Ein digitales Abstimmensystem mit Frequenzsynthese in DE-Z.: Funk-Technik, 1977, Nr. 16, F & E 270, 272, 281, 282, 284;
- BELT, F. H. „Tuning in on Colors, US-Z.: Electronics World, April 1968 S. 34-36, 44-47;
- Picture-Tube as Fine-Tuning Indicators, in US-Z.: Electronics World, Dec. 1967, S. 16;

Schaltungsanordnung zum Abstimmen eines Fernsehgeräteeinpfängers

Schaltungsanordnung zur Abstimmung eines Fernsehgeräteeinpfängers mit Hilfe einer phasensensitiven Rückkopplung (PLI-System), mit

- einem Oszillator (2) variabler Frequenz, dessen Frequenz elektrisch steuerbar ist;
- einem frequenzstabilen Bezugsozillator (3) fester Frequenz;
- einer Phasen-Frequenz-Vergleichschaltung (3) zur Erzeugung eines elektrischen Feinabstimmens, das proportional zu einer Frequenzdifferenz ist und dem Oszillator (2) variabler Frequenz zugeführt wird, um dessen Frequenz zu steuern;
- einer Frequenz-Zählvorrichtung (4), die mit einer veränderlichen Zahl N programmierbar und mit der Vergleichschaltung (3) verbunden ist;
- einem Schreib-Lese-Speicher (12, RAM) als Speichervorrichtung, die Daten über längere Zeit zu speichern vermag;
- einer Eingabevorrichtung (13) mit Tasten zum Erzeugen elektrischer Steuerbefehle;
- Mitteln zur manuellen Feinabstimmung im Sinne einer Veränderung von normgemäßem Empfangsfrequenzen, welchen entsprechende Fernsehkanäle zugeordnet sind;
- einer Einrichtung (8) zum Anzeigen der Kanalnummer des momentan selektierten Fernsehkanals und
- einem Mikroprozessor (11), der die elektrischen Steuerbefehle von der Eingabevorrichtung (13) empfängt und die zugeordnete Kanalnummer an die Frequenz-Zählvorrichtung (4) leitet sowie auf der Anzeigevorrichtung (8) sichtbar macht,

dadurch gekennzeichnet, daß der Mikroprozessor (11) auf der Anzeigevorrichtung (8) ferner eine numerische Angabe über die Größe einer gewünschten Abweichung gegenüber einer festgelegten, standardisierten Sendefrequenz zusätzlich zu der Kanalnummer sichtbar macht, wobei sich die gewünschte Abweichung entweder bei einer manuell vorgegebenen, willkürlichen Veränderung oder bei einem automatischen Feinabstimmvorgang ergibt, in welchem der Mikroprozessor (11) alle Empfangsfrequenzen in Frequenzschritten absucht, die einem Bruchteil des Abstandes zweier Nachbarkanäle entsprechen.



Wurde Anmeldung

DE 28 58 826 C 2

3

Europäisches Patentamt
European Patent Office
Office européen des brevets

Publication number 0 181 189 B1

EUROPEAN PATENT SPECIFICATION

Date of publication of patent specification: 13.05.92
Application number: 8530901.8
Date of filing: 04.11.85

Intrinsische-Frist 1 3. FEB. 1993

Video signal processing system.

Priority: 05.11.84 US 668478
Date of publication of application: 14.05.86 Bulletin 86-20
Publication of the grant of the patent: 13.05.92 Bulletin 92-20
Designated Contracting States: DE FR GB IT

References cited: US-A-4 109 276; US-A-4 214 262; US-A-4 389 678

Proprietor: RCA Thomson Licensing Corporation
2 Independence Way
Princeton New Jersey 08540(US)

Inventor: den Hollander, Willem
Spitalstrasse 27
CH-8552 Schlieren(CH)
Inventor: Harlmeier, Werner Nikolaus
Bergrasse 12
CH-8554 Geroldswil(CH)

Representative: Pratt, Richard Wilson et al
London Patent Operation G.E. TECHNICAL SERVICES CO. INC. Burdett House 15/16
Buckingham Street
London WC2N 8DU(GB)

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid (Art. 99(1) European patent convention).

EP 0 181 189 B1

4

Abb 1 & 2: Zwei der ältesten Patente unseres PCS-Projektes

Abb 3: Beispiel einer moderneren Fassung eines Patendokumentes

Abb 4: Beispiel eines Europäischen Patendokumentes

Abb 5: Beispiel eines Suchberichtes (search report)

Abb 6: Beispiel einer Indexkarte

INTERNATIONAL SEARCH REPORT		International application No. PCT/IB 97/00509
A. CLASSIFICATION OF SUBJECT MATTER		
IPC6: H03F 3/45 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols)		
IPC6: H03F		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
SE,DK,FI,NO classes as above		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
WPIL, EDOC, JAPIO		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 4068184 A (ADEL ABDEL AZIZ AHMED), 10 January 1978 (10.01.78), column 1, line 1 - column 4, line 39, figure 1 --	1-8
A	US 3922614 A (VAN DE PLASSCHE), 25 November 1975 (25.11.75), figure 2, abstract --	1-8
A	EP 0380152 A1 (N.V. PHILIPS' GLOEILAMPENFABRIEKEN), 1 August 1990 (01.08.90), figure 2A, abstract -- -----	1
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input checked="" type="checkbox"/> See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "B" earlier document but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "A" document member of the same patent family		
Date of the actual completion of the international search		Date of mailing of the international search report
19 November 1997		20-11-1997
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86		Authorized officer Viktor Skoog Telephone No. +46 8 782 25 00
Form PCT/ISA/210 (second sheet) (July 1992)		

5

Patenthistorie

Das Layout von Patendokumenten hat sich über viele Jahre und je nach Herkunftsland gewandelt. Schriftform und Strukturmerkmale des ältesten Patents vom 14. Oktober 1892 (siehe Abb.) sind völlig unterschiedlich von dem eines heutigen Patendokumentes (siehe Abb.).

Die Konsequenzen daraus für eine automatische Dokumentenerfassung sind offensichtlich. Das Verarbeitungssystem muß an die „historischen Schichten“ der Patendokumente angepasst werden.

2 019 804	21a1-36/02
Sie.	23. 3.70 4.11.71
EW.:	
RECH. 1	PRÜF. 33174
DAS 9 181	LIZ. B. 1
unverändert erteilt 5/181	
z.gew., z.genom., vsgt., gel. 15 186	

6

Imageformate

Es wurde festgelegt, den von unseren Kodak Scannern i260 gelieferten "Multi-Stream-Output" zu nutzen und pro Dokumentseite zwei Imageformate zu erzeugen. Durch das gleichzeitige Scannen von Vorder- und Rückseite eines Blattes entstehen also 4 Images.

_ *Format 1* ist ein 24 bit Farbbild in JPEG-Format (*.jpg), das von der Kodak Capture Software in der Einstellung „Better Quality“ nur wenig komprimiert wurde; dieses „Quasi-Original“ (typische Speichergröße einer Textseite mit kleinen Fonts ist ca. 1 MB)

_ *Format 2* ist ein bitonales Image mit TIFF Gruppe 4 (2-dimensionaler Kompression), das mittels adaptiver Schwellwertbildung im Scanner erzeugt wurde; dies war unser „ICR/OCR-Image“ (typisch kleiner als 100 KB)

In einem nachgeschalteten Prozess wurde ein „operatives Image“ mittels eines von DocStream erstellten Stapelverarbeitungsprogrammes erzeugt (ähnlich dem Luratech Pdf-Compressor). Der dazu verwendete Luratech-SDK wurde um die Funktionalität erweitert, den Images Texte zu unterschieben, um diese durchsuchbar zu machen:

_ Dieses *Format 3* ist ein 24 bit Farbbild „JPM-PDF“, erzeugt mittels der "Luratech compound document compression" Algorithmen. Die typische Dateigröße eines komprimierten Farbbildes beträgt nur etwas mehr als 120 KB und liegt somit in der Größenordnung eines Schwarz-/Weiß-Bildes. Zur Anzeige und auch zum Durchsuchen dieses JPM-PDF Imageformates kann der kostenlos erhältliche Acrobat Reader 6.0 verwendet werden.

Anforderungen aus Sicht der Dokumentensuche (Retrieval)

Das Auffinden von Patentdokumenten (Retrieval) soll möglich sein:

Patentordner

Aufgrund des Patentordners in dem sich die Patente ursprünglich befunden haben

Patentamt

Mittels Suche über das zuständige Patentamt (z.B. DE = German Patent Office; EP = European Patent Office; US = Patent Office of the United States etc.)

Patent-ID

Aufgrund der Gruppenzugehörigkeit der Patent-ID (d.h. der unterschiedlichen Endungen)



1.6 Besondere Anforderungen

Folgende Überlegungen führten zu besonderen Projektanforderungen:

Weshalb hochqualitative Originale von 1 MB pro Seite?

Häufig bestehen Ängste, das Papierarchiv aufzulösen und dann wirklich zu verschreddern, weil unklar ist, welche Verarbeitungsmöglichkeiten die Zukunft birgt. Ein elektronisches Archiv parallel zum Papierarchiv zu halten, macht aber Sinn nur für eine Übergangszeit.

Unsere Lösung der Zukunftssicherheit besteht deshalb darin,

_ bereits heute sehr hochwertige JPG-Dokumentenimages in Farbe und in recht hoher Auflösung (300 dpi) aufzunehmen und mit geringer Kompression ins Archiv zu legen (ca. 1 MB/Seite).

_ Die JPG-Dokumentenimages in ein für heutige Übertragungsbandbreiten handhabbares, operativ verwendbares JPM-PDF Format (von Luratech) zu konvertieren (bezogen auf das JPG-Format ca. Kompressionsfaktor 8 -10 je nach Inhalt, also ca. 120 KB je Seite). Dieses JPM-PDF Image entspricht nicht nur neuester internationaler Norm, sondern lässt sich auch mit dem kostenlos erhältlichen Adobe Acrobat Reader 6.x Viewer betrachten.

_ Unterlegt mit hochqualitativen OCR-Ergebnissen entstehen hoch qualitative und perfekt durchsuchbare PDF-Images, darstellbar im kostenlosen Acrobat Reader 6.o.

Ergebnis: Ein zukunftssicheres „Archiv-Image im Tresor“ und „Operatives Image“ für heutigen praktischen Gebrauch.

Weshalb Farbbilder von Akten?

Viele der Patentdokumente enthalten handschriftliche, aber oft nur kontrastarm vorhandene Anmerkungen, Unterstreichungen und Farbmarkierungen, die bei einfachem Scannen in Schwarz-/Weiß verloren gingen. Damit diese Background-Informationen erhalten bleiben, wurden alle Dokumente mit einer Auflösung von 300 dpi in Farbe (24 bit Tiefe) gescannt.

Verglichen mit dem visuell flachen Eindruck bitonaler Images machen diese Farbbilder einen wirklich originalgetreuen Eindruck und sind eine Freude für die Augen der Bearbeiter, ein ergonomischer Aspekt, der nicht unterschätzt werden darf.

Farbimages sind die Zukunft, auch in der Elektronischen Dokumentenverarbeitung, ohne Frage. Die Welt ist bunt und wer benützt heute noch einen Fernseher in Schwarz-/Weiß? Die Kapazitäten von Speichermedien explodieren bei sinkenden Preisen und auch die Übertragungsbandbreiten steigen.

Weshalb eine so aufwendige Indexierung?

Patentdokumente bestehen aus Titelseiten, Textseiten, Abbildungsseiten, Leerseiten, Anhängen und Suchberichten (search reports). Die seitenweise Dokumentenklassifikation in diese 6 Klassen ist also die Weiterverarbeitungsbasis.

Hochqualitative OCR/ICR-Verarbeitung aller Seiten ist die Basisinformation der Freitext-Indizierung. Zusätzlich und im Gegensatz zu den anderen Seitenklassen erfordert die Titelseite wegen der dort enthaltenen bibliografischen Daten jedoch eine spezielle Form der automatischen Indizierung, nämlich das dynamische Suchen variabel angeordneter Lesezonen. Dies geschieht für die ausgesuchten Indexinformationsfelder über automatische Begriffs-extraktionen auf Basis der OCR-Ergebnisse.

Bei papierbasierter Verarbeitung bilden auch heute noch zusätzliche handschriftliche Vermerke das Bindeglied zur firmeninternen Fachklassifikation. Bei den vorliegenden Patenten waren diese kryptischen Indizes an irgend einem freien Fleck der Patenttitelseite aufgebracht, oftmals in Fließhandschrift („Arztschrift“). Hier versagt heutige Automatisierungstechnik, also eine Aufgabe für manuelle Erfassung.

Wozu durchsuchbare Images?

Bei der Recherche in der Patentdatenbank werden Treffer erzeugt und die relevanten Dokumente zum Download angeboten. Für die Bearbeitung der heruntergeladenen Patentdokumente ist natürlich eine automatische inhaltliche Suche wünschenswert anstatt alle Seiten eines Dokuments visuell durchzusehen. Das Erstellen durchsuchbarer PDF-Images geschieht durch Unterlegen mit OCR-Text.

Unsere Tests an älteren Patentdokumenten haben deutliche Schwächen der in der kostenpflichtigen Acrobat 6.0 Vollversion enthaltenen OCR-Funktion aufgedeckt, weshalb wir hier eine Eigenentwicklung betrieben haben. Mit unserer aktuellen Eigenentwicklung „d-ImageSearch“ sind wir in der Lage, jedwede und auch problematische Images mit Textinhalten durchsuchbar zu machen, und zwar in höchster Qualität. Diese PDF-Images können mit Acrobat Reader 6.0 angezeigt und durchsucht werden („Bearbeiten/Suchen“).



2. Lösungskonzept und Vorgehen

Grundlage für die Auftragserteilung an DocStream war der Nachweis, die Machbarkeit der Anforderungen an den ersten 200 Patentordnern zu zeigen, und zwar bei einem einzuhaltenden, sehr engen Kostenrahmen (Festpreis).

Hierzu haben wir folgende Maßnahmen ergriffen:

_ Die heutige Scannertechnik machte es möglich: Unsere Kodak-Scanner i 260 erfassen in einem Zug Vorder- und Rückseite und gleichzeitig in Farbe und Schwarz-/Weiß.

_ Zur Datensicherung in der Scanstelle und zur Datenübertragung in das Verarbeitungszentrum bzw. zu Micronas wurden externe Harddisks mit Firewire-Schnittstelle und einem Fassungsvermögen von 300 GB eingesetzt.

_ Zur inhaltlichen Verarbeitung wurde bei Doc-Stream ein spezielles automatisches Produktionssystem aufgesetzt, eine Mischung aus kombinierter geometrischer und statistischer Dokumentenklassifikation, strukturierter Formularlesung, und freier OCR-Verarbeitung.

_ Der veranschlagte und schließlich auch erreichte Gesamt-Datenbestand an Images und Indexdaten (inkl. Verarbeitungs-Overhead) beträgt 2 TeraByte. Die Datensicherung erfolgte auf LaCie Externspeichern mit Firewire-Schnittstelle.

_ Zur Validierung der manuell erfassten Daten wurde mit dem „PCS-Validator“ ein spezielles Tool für die Anzeige und Prüfung von Patentseiten erstellt.

_ Nach ausführlichen Produktionstests der Acrobat 6.0 Vollversion zur Erstellung durchsuchbarer PDF-Images wurde entschieden, statt dessen einer Eigenentwicklung den Vorzug zu geben, die es erlaubt, je nach Schriftgut verschiedene markt-gängige OCR-Engines einzusetzen. Zur Kompression der JPG-Farbimages haben wir uns für das Lura-tech-Kompressionsverfahren entschieden und diese JPM-PDF-Images durchsuchbar gemacht.

Als Dienstleister für anspruchsvolle Dokumentenerfassung sind wir frei, Komplettsysteme oder Softwaremodule, Kaufprodukte oder Eigenentwicklungen, so zu kombinieren und einzusetzen, wie es für unsere Produktionszwecke des jeweiligen zu erfassenden Schriftgutes optimal ist.

DocStream erledigte diesen ‚Document Capture Service‘ auf Basis eines weltweiten Technologie-partner-Netzwerkes und gibt damit ein Praxisbeispiel für die Möglichkeiten kostengünstiger globaler Kooperation auf dem Arbeitsgebiet der elektronischen Dokumentenerfassung. Und die manuelle Erfassung nicht maschinenlesbarer Schreibrift-Objekte haben wir unsere Offshore-Partner in Indien anvertraut.

Insgesamt also nicht nur eine „runde Sache“, sondern auch eine globale.



3. Entwicklungsleistungen und Verarbeitungssystem

Um die mit Micronas in einem gemeinsamen Workshop festgelegten, anspruchsvoll spezifizierten Kundenanforderungen und die Vorgaben des Festpreisangebotes überhaupt erfüllen zu können, mussten moderne Automatisierungstechnik kombiniert werden mit attraktiven Programmier-Stundensätzen und günstigen Kostenstrukturen von Offshore-Anbietern.

Die damit projektspezifisch konfigurierten Komponenten sowie Beratungs- und Entwicklungsleistungen sind:

_ Scan-Strecke, bestehend aus 2 Kodak-Scannern i260 mit Kodak Capture Software. Dabei manuelles Erfassen der Aktenrücken-Beschriftungen und Leistungsdatenerfassung

_ Eigene DocStream Workflow-Komponente zur automatischen Steuerung der Verarbeitungsstufen

_ MySQL-Datenbanksoftware 3.51 als zentrale Datenhaltung mit Tabellen für Imagenamen jeder Dokumentenseite (TIFF und JPG-Formate), zur Speicherung der Verarbeitungsergebnisse von Dokumentenklassifikation und -Extraktion sowie zur Buchführung der Verarbeitungsstadien

_ Einsatz der Finereader-Engine 7.0 zur OCR-Verarbeitung aller Textseiten in Maschinenschrift

_ Kompression der Farbbilder im JPG-Format nach dem Luratech-Verfahren plus eigen entwickelte Unterlegung mit den bereits gewonnenen OCR-Ergebnissen, um hochqualitativ durchsuchbare und hoch komprimierte JPM-PDF-Images zu generieren

_ Dokumentenklassifikation in 6 Klassen nach zwei Verfahren: Kombination der imagebasierten, geometrischen Dokumentenklassifikation (FormsRec 4.7.2) mit der eigen entwickelten textbasierten, statistischen Dokumentenklassifikation „d-Class“, für die spezielle Klassifikatoren berechnet wurden

_ Farbbild-gestützte Nachkorrekturstationen, erforderlich zur Validierung der 6 Dokumentenklassen vor allem bei problematischem Schriftgut

mit Artefakten (FormsRec 4.7.2)

_ Bei DocStream entwickelte automatische Suche von Textobjekten d.h. dynamisch gesuchter Indexfelder (eine Art semi-strukturierter Formularlesung mit geometriebasierten OCR-Ergebnissen)

_ FTP-Verbindung zur Übertragung von Patenttitelseiten (als komprimierte JPM-PDFs in Farbe) zu unserem Offshore-Partner (im Projektverlauf mehrere Gigabyte), um handschriftliche firmenspezifische Indizes manuell zu erfassen sowie Extraktionsergebnisse von Patenttitelseiten zu validieren

_ Eigenentwicklung einer „Image Cutter“-Software, um Dokumentenschnipsel anstatt von Komplettdokumenten zu übertragen

_ Rückerhalt der gezippten XML-Ergebnisse vom Offshore-Partner via eMail

_ Bereitstellen eines projektspezifisch entwickelten Validierungstools „PCS-Validator“ (xml-basierter Viewer für Patentimages) zum Gebrauch bei der Überprüfung strittiger Index-Bedeutungen durch Micronas-Experten

_ DocStream Entwicklung eines „xml-Export-Combiners“ zur Zusammenführung der xml-basierten Ergebnisse automatischer Informationsextraktion aus Patenttitelseiten mit den manuell erfassten Indizes bis auf Seitenebene

_ Beratungsleistungen bei Imagequalität, Schema und DTD für xml-Datenexport



4. Arbeitsvorbereitung und Scannen

Die Arbeitsvorgaben und –Methoden waren folgende und wurden durch den Scan-Master kontrolliert:

- _ Arbeitsvorbereitung: Dokumente einzeln dem Ordner entnehmen, Heftklammern entfernen, Seitensortierung überprüfen und unterschiedliche Formate ausrichten, Trennblätter einfügen und Dokumentenstapel zum Scannen bereitstellen
- _ Aktenrücken erfassen: Da der Nummernkreis, zu dem das Patentedokument gehört, wird abgeprüft wird, müssen die Angaben, welche Patentnummern in welchen Ordnern abgelegt sind, vom Aktenrücken manuell erfasst werden
- _ Durchführen des Scanvorganges (sowohl unter Verwendung der automatischen Zuführung als auch im Einzelfall durch manuelle blattweise Zuführung)
- _ Visuelle Qualitätsprüfung der gescannten Seiten am Monitor, überwachen automatischer Kontrollfunktionen von Scanner und Scansoftware.
- _ Überwachen der automatischen Verzeichniserzeugung und Datenspeicherung durch die Scansoftware
- _ Fehler-Behandlung im Falle von Papierstau, korrektes Wiederaufsetzen
- _ Protokollführung und Datensicherung für jede Schicht

Die Trennblätter waren so gestaltet, daß mit der Kodak Capture Software das Hochzählen und Neu-anlegen von Verzeichnissen vollautomatisch erfolgte.

Weitere Anforderungen an den Scanprozeß waren:

- _ Alle Seiten sollten als Einzelbilder abgelegt werden (also beispielsweise kein Multi-TIFF)
- _ Leere Seiten sollten nicht eliminiert werden, um aus rechtlichen Gründen keine „fehlenden Seitennummern“ zu erzeugen
- _ Farbimages sollten nicht total unkomprimiert gespeichert werden, da mit 24 MB zu groß

_ Im Falle auftretender Kopien von Patentedokumenten sollten diese mit einer ergänzenden Zusatz der Patent-ID verarbeitet werden

_ Die in den Originalordern ggfs. falsch einsortierten Patentedokumente sollten dort verbleiben und nicht umsortiert werden

_ Gelegentlich falsch einsortierte Nachforschungsunterlagen (search report, rapport de recherche Europeenne) sollten in der vorgefundenen Reihenfolge verbleiben

Zur Sicherstellung der korrekten Abarbeitung wurden schriftliche Arbeitsanweisungen erstellt für:

- Die Arbeitsvorbereitung
- Den Scanprozeß
- Das Rücksortieren und Bündeln der gescannten Patentedokumente
- Das Speichern und Sichern erfasster Images



5. OCR-Verarbeitung, Indexieren und Datenextraktion

Im Information Retrieval bedeutet Indexierung, eine Menge von Dokumenten so aufzubereiten, dass man schnell auf sie zugreifen kann. Wird ein Index erstellt, so wird indexiert – wird später etwas einem bestehenden Index hinzugefügt, wird es indiziert.

In Zusammenhang mit Images (einer Ansammlung von Bildpunkten / Pixel), auch Non Coded Information (NCI) genannt, ist mit Indexieren das Erzeugen von Codierter Information (CI) gemeint, mittels der man schnell auf Images zugreifen kann (beispielsweise unter Verwendung von Suchalgorithmen).

Es ist heute allgemein bekannt, daß die Indexierung durch automatische OCR / ICR-Verfahren niemals völlig fehlerfrei erfolgt. Deshalb müssen Leseergebnisse immer geprüft werden, entweder durch Plausibilitätsprüfungen oder durch Informationsabgleiche oder durch Einsatz von Korrekturpersonal. Oder man setzt bei der Suche (Information Retrieval) fehlertolerante Abfragetechniken ein.

Bei Patentedokumenten enthält die Titelseite die wichtigste Zusammenfassung relevanter Informationen und stand deshalb bei unserer Bearbeitung im Vordergrund, obwohl wir tatsächlich durchgängig alle Seiten mit OCR-Verfahren bearbeitet haben. Wir haben deshalb bei Titelseiten eine zusätzliche feldorientierte Informationsextraktion durchgeführt.

Dokumentenklassifikation

Nächste Aufgabe war, in einem Stapelverarbeitungsprozess die Seiten eines Patentedokumentes nach deren Klassen zu sortieren, also nach Titelseiten, Textseiten, Abbildungsseiten, Leerseiten plus bei den oft vorhandenen Anhänge zusätzlich in die Klassen Anlagen und Suchberichte.

Weiter war aus der Patenttitelseite auszulesen, ob es sich bei dem jeweiligen Patentedokument um

eine Offenlegungsschrift (synonym „Europäische Patentanmeldung“ bzw. „European Patent Application“ bzw. „Demande de brevet Europeen“), um eine Auslegeschrift oder um eine Patentschrift (synonym „Europäische Patentschrift“, bzw. „European Patent Specification“ bzw. „Fascicule de brevet Europeen“) handelt.

Üblicherweise werden hier imagebasierte, geometrischen Dokumentenklassifikationsverfahren eingesetzt, die auch sehr anschaulich sind. Man sieht das fragliche Schriftgut durch und setzt an denjenigen Dokumentregionen Lese- und Prüfzonen ein, wo Informationen stehen, die für die gesuchten Dokumentenklasse typisch sind. Bei gut bekannten und stetig wiederkehrenden Dokumentklassen funktioniert diese Art der Seitensortierung tatsächlich recht nett. Was aber, wenn das Schriftgut verkleinerte, verdrehte oder sehr gestörte Seiten enthält? Oder wenn, wie in unserem Falle, das Gesamtschriftgut derart umfangreich ist, daß man mit dem Trainieren der Vielzahl imagebasierter, geometrischer Dokumentenklassen nicht mehr nachkommt bzw. die Personalkosten horrend ansteigen?

Auch der Qualität wegen haben wir deshalb parallel unsere eigen entwickelte textbasierte, statistische Dokumentenklassifikation „d-Class“ eingesetzt, für die spezielle Klassifikatoren berechnet wurden, welche die Seitenarten identifizieren und sortieren. Die Eigenschaft von d-Class, das Layout eines Dokumentes völlig unberücksichtigt zu lassen, ist beispielsweise bei der automatischen Verarbeitung verkleinerter oder vergrößerter Dokumente nützlich, oder auch im Falle schlecht gescannter Images, bei denen Regionen des Dokuments bildlich ausgefallen sind. Auch Schieflagen von Dokumentenimages werden in weiten Grenzen toleriert, solange die OCR-Engine noch einigermaßen relevante Informationen generiert.



Diese Technik der statistischen Dokumentenklassifikation ist allerdings weit weniger anschaulich, aufwendig im Training, aber qualitativ von außerordentlich hoher Qualität.

Nur, welcher Scan-Dienstleister beherrscht diese Technik oder will sich diesen Aufwand leisten? Derzeit sicherlich ein Alleinstellungsmerkmal von DocStream GmbH.

Dokumentenextraktion

Die aus der Titelseite eines Patentedokumentes auszulesenden Daten waren folgende:

Schriftnr., Titel, Erfinder, Veröffentlichungsdatum, Erteilungsdatum, Abstract, Anmelder, Patentinhaber, Aktenzeichen, Anmeldetag, Prioritätsangaben, Prioritätsaktenzeichen, Prioritätsdatum, Prioritätsland, Benannte Vertragsstaaten, IPC, Beschreibung, Patentansprüche.

Auch hier wurden 2 konkurrierende bzw. ergänzende Verfahren eingesetzt:

_ Die imagebasierte Extraktion wurde vorwiegend für relativ statisch angeordnete Felder eingesetzt bzw. dynamische Feldsuche wurde regelbasiert vorgenommen.

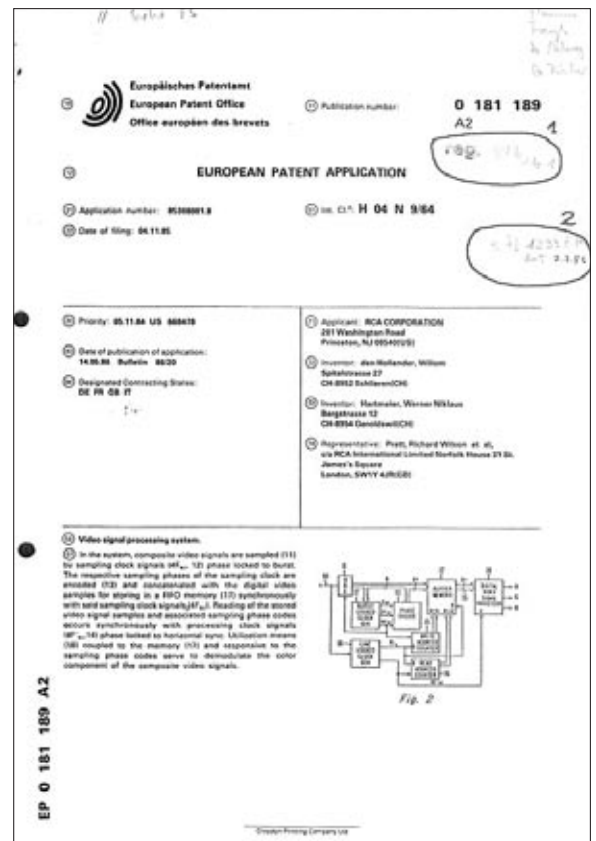
_ Die textbasierte Informationsextraktion erfolgte auf Basis der generell ermittelten OCR-Ergebnisse, für deren Textobjekte zuvor die Geometriedaten ermittelt worden waren.

Für alle genannten Felder (Indizes) wurden umfangreiche Plausibilitätsprüfungen durchgeführt.



6. Offshore-Arbeiten

Die durch unseren Offshore-Partner manuell zu erfassenden beiden Handschriftfelder aus ca. 80.000 Patentimages (je 90 bis 120 KB) waren folgende:



Beispiel 1



Feld 1 ist gekennzeichnet durch einen Stempel „reg.“

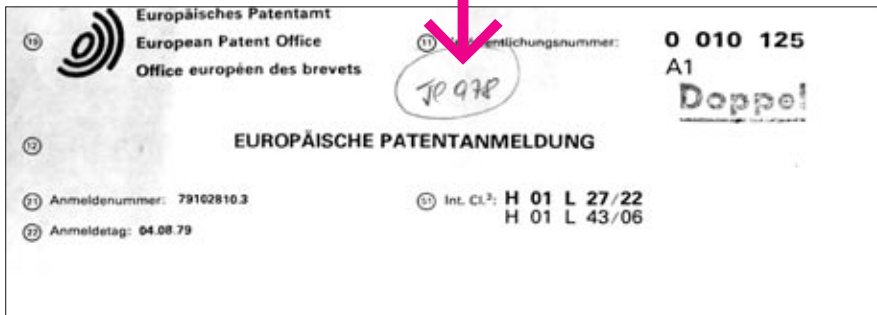
Feld 2 ist das auf der linken Seite des Dokumentes aufgebrauchte Feld.

Für beide Felder gelten ausführliche Plausibilitätsprüfungen. Die Ergebnisse wurden als xml-Datensatz übermittelt.

Beispiel 2



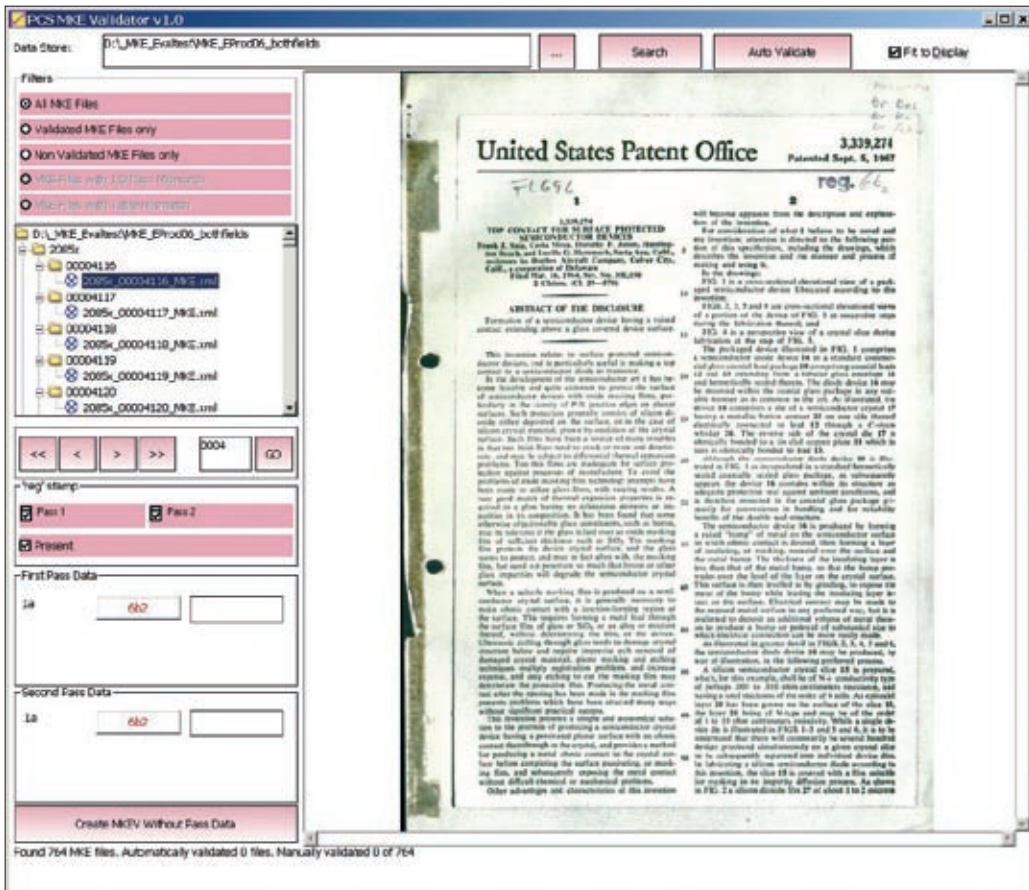
Beispiel 3:



Beispiel 4:



Die Überprüfung unklarer Angaben erfolgte mit Hilfe des von uns erstellten Software-Tools „PCS-Validator“:



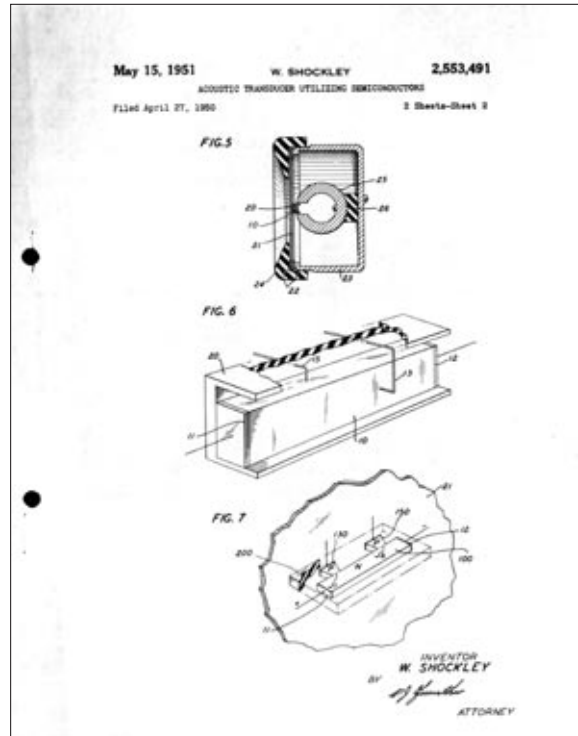
7. xml-Schema, DTD; xml-Zusammenführung

Dokumentenindex-Export

Die auf diese Weise klassifizierten und extrahierten Seiteninformationen wurden als xml-Ausgabedatensatz strukturiert.

Die xml-Datenströme der automatisch extrahierten plus der Offshore manuell erfassten Daten wurden mittels unseres „xml-Export-Combiners“ zusammengeführt.

Die Besonderheit war, mit Doppelausgaben von Patentdokumenten (z.B. mit Kopien) fachgerecht umzugehen. Hier haben wir uns entschlossen, nicht etwa solche Duplikate wegzuerwerfen, sondern multiple xml-Dateien zu erzeugen, indem logisch eine neue Dokumentenbezeichnung erzeugt wurde. Auf Basis der Bezeichnung der ersten Titelseite identischer Dokumente wurde für die weitere Titelseite die Dokumentennummer um eine Stelle beginnend mit der Ziffer 2 hochgezählt.



Beispiel für eines der Patente des weltbekannten Erfinders und Nobelpreisträgers William Shockley:

Beispiel für eines der „Horror-Patentdokumente“ für Dokumentenerfassung:

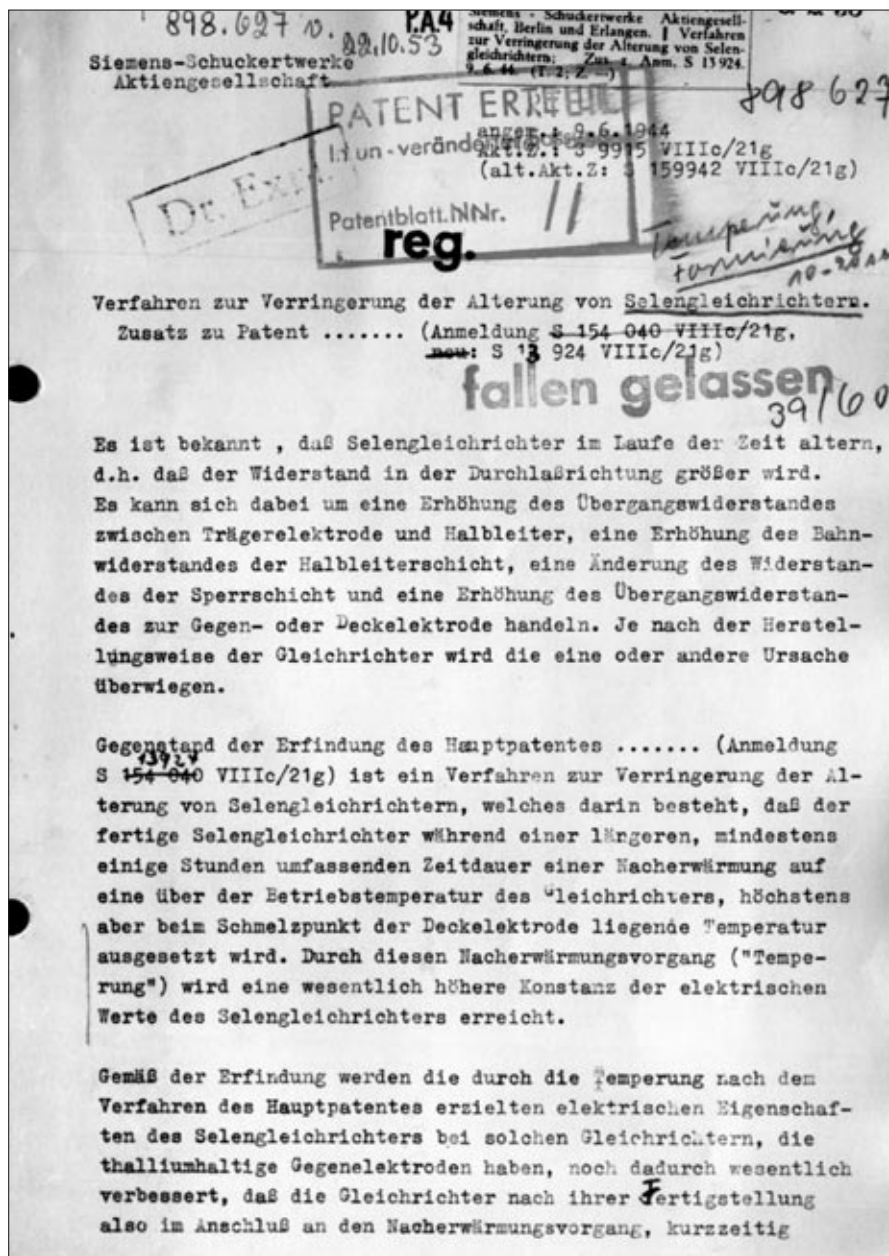


Abb.: Dr. Wulbrand Jahnke, Projektleiter Patentrecherche-system bei Micronas GmbH



Vorteil der elektronischen Patentakte ist die Möglichkeit der Zusammenarbeit im Team, zwischen verschiedenen Abteilungen und auch standortübergreifend.

Dr. Wulbrand Jahnke, Projektleiter Patentrecherche-system bei Micronas Intellectual Property Management: „Suchen in eingescannten Dokumenten? Mit DocStream ein Klacks“.

Unser Ausblick auf andere Erfassungsservices

Viele Altarchive enthalten Informationsschätze auf Papier. Leichter und schneller ist jedoch der Zugriff auf deren elektronische Form. Und natürlich auch Platz- und damit Lagerkosten sparender.

Zuvor muss das Altarchiv jedoch fachgerecht konvertiert werden, und das kann komplexer sein als zunächst vermutet, denn die Transformation papiergebundener Information in elektronische Form geht häufig nicht ohne „Bocksprünge“ ab. Dann ist nicht nur reichlich Erfahrung gefragt, sondern ein Dienstleister mit allen Tools & Tricks.

Über DocStream

Die DocStream GmbH wurde im März 2002 als Systemhaus und Dienstleister mit Schwerpunkt dokumentenzentrierter Anwendungen unterschiedlicher Branchen gegründet („Document-Input-Technologies“).

Als Softwarehaus für Projektlösungen der Dokumentenverarbeitung realisieren wir auf unserem Arbeitsgebiet kundenspezifische Projekte unter Einsatz führender Fremdprodukte als auch von Eigenentwicklungen.

Mit unserem Angebot als Dienstleister für Document-Capture Services übernehmen wir das Scannen und automatische Auswerten von Kundendokumenten, insbesondere aus Altarchiven. Für unsere angebotenen Scan-Services verwenden wir professionelle Dokumenten-Produktions-Scanner. Durch modernste Technik und professionelle Software erzielen wir eine optimale Imagequalität – die Voraussetzung für Weiterverarbeitungs- und Archivfunktionen. Mit unseren Produkten zur Dokumenten-Interpretation / Schriftenlesung und Dokumenten-Klassifikation (Sortierung) bieten wir die Kompetenz eines erfahrenen Spezialisten.

Weitere Informationen über die Firma und ihre Produkte erhalten Sie unter
www.docstream.de

Kontakt:

Werner G. Richter

Geschäftsführender Gesellschafter

Phone: +49-7551-9495890 | Fax: +49-7551-9495899

mailto: info@docstream.de

DocStream GmbH

Weinbergstr. 23c | 88662 Überlingen

Germany

Über Micronas

Micronas (SWX Swiss Exchange: MASN, Frankfurt: MNSN, Prime Standard Segment, TecDAX), ein weltweit operierender Halbleiterentwickler und -hersteller, ist ein führender Anbieter innovativer IC- und Sensor-Systemlösungen für die Bereiche Unterhaltungselektronik und Automobilelektronik. Als Marktführer bei innovativen, globalen TV-Systemlösungen bringt Micronas ihre Expertise in neue Märkte ein, die durch die Digitalisierung von Audio- und Video-Inhalten entstehen. Micronas zählt alle bedeutenden Markenhersteller der Unterhaltungselektronik weltweit zu ihren Kunden, viele davon in einer dauerhaften Partnerschaft, die auf den gemeinsamen Erfolg ausgerichtet ist. Sitz der Holding ist in Zürich (Schweiz), der operative Hauptsitz in Freiburg (Deutschland). Derzeit beschäftigt die Micronas-Gruppe rund 1900 Mitarbeiter. 2004 erzielte das Unternehmen einen Umsatz von CHF 963 Mio.

Weitere Informationen über die Gruppe und ihre Produkte erhalten Sie unter
www.micronas.com

copyright © DocStream GmbH



