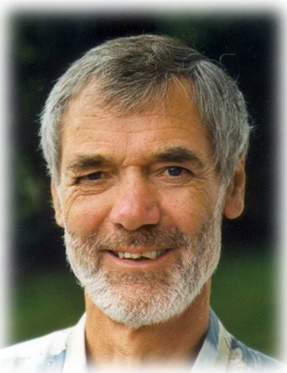


Das Heiderprojekt

Projektdokumentation der Digitalisierung der Heider-Regesten



Regal mit den Heider-Büchern im Oberösterreichischen Landesarchiv Linz



Ing. Sepp Asanger
4040 Linz
Langfeldstraße 11
sepp.asanger@aon.at

INHALTSVERZEICHNIS

1	Allgemeines	6
1.1	Ziel des Projektes	6
1.2	Kontrolldatei.....	6
1.3	Datentypen.....	6
1.3.1	Buchseiten.....	6
1.3.2	Scans	6
1.3.3	Textdateien.....	7
1.3.4	Excel-Dateien.....	7
1.4	Ordnerstruktur	7
1.5	Namensschema	7
1.6	Erweiterung des Namensschemas für Ordner und Dateien	8
1.7	Änderung des automatischen Dateinamens vom Buchscanner auf die neue Namensgebung	9
2	Allgemeines zum Heider-Projekt.....	11
2.1	Was ist der Heider-Index?	11
2.2	Josef Heider	11
2.3	Gemeinsames Projekt des Oberösterreichischen Landesarchivs und des Vereins Familia Austria	
(FA)	12	
3	Organisation der Heiderbücher.....	13
3.1	Bücher mit Einführung je Pfarre	13
3.2	Aufteilung der Heiderbücher	13
3.3	Alphabetischer Index.....	13
3.4	Seiten.....	14
3.5	Mengenangaben.....	14
4	Aufbau der Heider Heiratsregesten.....	15
4.1	Allgemeines	15
4.2	Paarweise Erfassung der Brautleute	15
4.3	Spalten	15
4.4	Erklärungen zu den Spalten.....	16
4.4.1	Tag	16
4.4.2	Name	16
4.4.3	Eltern, Beruf, Ort	16
4.4.4	Tom. (tomus).....	17
4.5	Schriftbild.....	17
5	Gesamtablauf	18
6	Digitalisierung der Heiderbücher am Buchscanner des OÖLA.....	19
6.1	Der Buchscanner	19
6.2	Ausrichtung des Buches auf der Buchwippe.....	19

6.3	Optische Verzerrung.....	20
6.4	Durchscheinendes Papier.....	20
7	Nachbearbeitung der digitalisierten Buchseiten	21
7.1	Seitennummern auf Vollständigkeit prüfen und korrigieren.....	21
7.2	Buchseiten ohne Seitennummer	21
7.3	Mögliche Fehler in der Nummerierung der Heider-Buchseiten und ihre Behandlung im Dateinamen.....	22
7.3.1	Auslassungen in der Seitennummerierung.....	22
7.3.2	Doppelvergabe von Seitennummern	22
7.3.3	Ziffernsturz.....	22
7.3.4	Falsche Hunderterstelle.....	22
7.3.5	Übersprungene Seitennummern (häufig in der Zehnerstelle)	23
7.3.6	Rücksprung in der Seitennummer	23
7.4	Seiten gerade richten	23
7.5	Seiten beschneiden	23
8	OCR-Schrifterkennung mit ABBYY FineReader	24
8.1	Verwendete Software	24
8.2	Optionen	24
8.2.1	Kein automatisches Starten des Lesens, Bildneigung.....	24
8.2.2	Speichern der Textdatei.....	25
8.2.3	Schriftart für Text.....	25
8.2.4	Zwischenspeichern der OCR-Ergebnisse.....	25
8.3	Wörterbuch.....	25
8.4	Öffnen der zu übersetzenden Scans (Buchseiten)	29
8.5	Bereich der Texterkennung	29
8.6	Seiten lesen.....	32
8.7	Bearbeiten des Textes.....	33
8.7.1	Seitennummer	33
8.7.2	Hochzeitsdatum.....	34
8.7.3	Name	34
8.7.4	Eltern, Beruf, Ort	34
8.7.5	Tomus.....	34
8.7.6	Falsche und doppelte Zeilen und Leerzeilen.....	35
8.7.7	Übertippte Zeichen und manuelle Korrekturen im Heider-Index	35
8.7.8	Bearbeitung rückgängig machen.....	36
8.7.9	Speicherung der Textseiten	36
9	Prüfung und Seitenformatierung mit dem Texteditor KEDIT	38
9.1	Allgemeines.....	38
9.2	Aufruf der Prüf- und Formatierungsfunktionen über Funktionstasten	38
9.3	Profile	39

9.3.1	Startposition des Elternblockes	39
9.3.2	Reihenfolge von tomus und pagina.....	39
9.3.3	tomus-Nummer der Kirchenbücher.....	39
9.4	Informationen zur Textseite.....	40
9.5	Trennen Namen- und Elternblock	41
9.6	Prüfungen und Bearbeitung einer Textseite mit F7- und F8-Taste	42
9.6.1	Festlegung der Startpositionen der vier Textbereiche.....	42
9.6.2	Prüfung der Seitennummer des Buches	43
9.6.3	Lesen der vorhergehenden Seite	44
9.6.4	Prüfung des Heiratsdatums	44
9.6.5	Prüfung der Namenszeilen	44
9.6.6	Prüfung der Elternzeilen	44
9.6.7	Prüfung der tomus pagina-Information	45
9.6.8	Nachtrag	47
9.6.9	Rückgängig machen von Textänderungen	48
9.6.10	Ungeklärte Probleme, Kommentar und Fehlerprotokoll.....	48
9.6.11	Anhang	49
9.7	Arbeiten mit dem KEDIT Texteditor	50
10	Übergabe der Daten für die weitere Bearbeitung	52
11	Abbildungsverzeichnis.....	53

1 Allgemeines

Diese Dokumentation soll es anderen Personen ermöglichen, sich in das Projekt einzuarbeiten und die Arbeit auf einem eigenen Computer teilweise oder zur Gänze zu übernehmen. Die Dokumentstruktur kann auch als Vorlage für ähnliche Projektdokumentationen dienen. Die Formatierung ist für beidseitigen Ausdruck ausgelegt.

Das Projekt ist zwar unter dem Namen Heider-Projekt bekannt geworden, weil die Idee von der Erfassung des Heider-Index seinen Ausgang nahm und das ist auch nach wie vor der Kern des Projektes. Im Zuge der Arbeiten erfuhr dieses Projekt eine allgemeinere Bedeutung, weshalb es so angelegt ist, dass es auch andere Quellen mit einschließen kann.

Die hier beschriebenen Strukturen und Verarbeitungen sind mit dem Projekt gewachsen und entsprechen dem gegenwärtigen Stand. Die Namenskonventionen und die Datenaufbereitung waren zu Beginn des Projektes noch nicht so weit entwickelt, weshalb Namen und Qualität der Dateien aus der Anfangszeit vom heutigen Stand abweichen können.

1.1 Ziel des Projektes

Ziel des Projektes ist die Erfassung aller brauchbaren Indizes zu Kirchenmatrikeln, die in Papierform oder in elektronischer Form im Oberösterreichischen Landesarchiv (OÖLA) zur Verfügung stehen und deren Einspeisung in die Datenbanken des Vereins Familia Austria (FA).

1.2 Kontrolldatei

Im OÖLA gibt es eine Auflistung aller im OÖLA verfügbaren Indizes. Diese Bestandsliste ist nach Pfarren geordnet und ist für das Projekt die Grundlage der Datenquellen. Sie wurde mit Details zu den Heiderbüchern wie Buchaufteilung, Seitenanzahl und erfasste Jahre ergänzt und als Excel Datei angelegt. Aus dieser Datei entstand eine Kontrolldatei für das Projekt, in dem die beteiligten Mitarbeiter den Arbeitsfortschritt bei der Erfassung der Datenquellen eintragen und Besonderheiten protokollieren. Bei Vergabe der Dateien an externe Mitarbeiter zur manuellen Kontrolle der Daten werden der Name der Personen sowie Datum der Aus- und Rückgabe eingetragen. Damit ist diese Kontrolldatei ein wichtiges Steuerinstrument für das Projekt und gibt einen Überblick über den aktuellen Stand.

1.3 Datentypen

1.3.1 Buchseiten

Die Indizes liegen in der Regel in gebundenen Büchern vor, in Ausnahmefällen sind die Daten auch elektronisch verfügbar.

1.3.2 Scans

Die Indizes werden nicht manuell abgeschrieben, sondern über einen Buchscanner erfasst, soweit sie nicht bereits elektronisch vorliegen. Diese Scans der Indexseiten sind in computertechnischem Sinn Bilder (oder Images) im jpg-Format. Eine Scan-Datei entspricht z.B. einer Seite eines Heider-Buches.

1.3.3 Textdateien

Mit spezieller Software werden die Texte der Scans in computer-lesbare Schrift übersetzt. So entstehen aus den jpg-Dateien Textdateien mit der Dateiendung `txt`. Eine Text-Datei entspricht einer Scan-Datei bzw. einer Buchseite.

1.3.4 Excel-Dateien

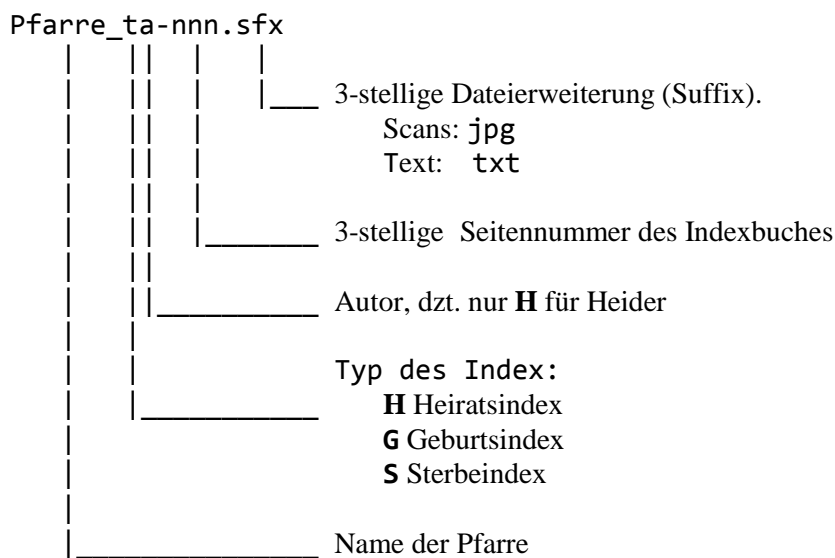
Bei der Verarbeitung aller Textdateien eines Index-Buches entsteht eine Excel-Datei, in der je Zeile ein Ereignis (Geburt, Heirat oder Tod) gespeichert ist. Eine Excel-Datei entspricht also einem ganzen Indexbuch. Vor der Einspeisung in die FA-Datenbank werden zu der Excel-Datei weitere Informationen hinzugefügt, damit die Daten den Anforderungen der FA-Datenbank entsprechen.

1.4 Ordnerstruktur

Je Pfarre gibt es einen Ordner (= Verzeichnis) mit dem Namen der Pfarre. In diesem Pfarrordner befinden sich weitere Unterordner mit den Dateien der Scans bzw. der Texte und die Excel-Dateien.

1.5 Namensschema

Die Namensgebung der Ordner und der Dateien erfolgt so, dass je nach Hierarchie die Zuordnung zur Pfarre, zum Autor und Art des Index und zu den Seiten eines Buches möglich ist, selbst wenn sich eine Datei oder ein ganzer Unterordner außerhalb des zugehörigen Verzeichnisses befindet. Dies wird ermöglicht durch das folgende Namensschema auf der untersten Ebene, also auf der Ebene der Dateien:



Ein Indexbuch kann mehrere hundert Seiten umfassen, daher ist die Seitennummer der Datei dreistellig mit führenden Nullen.

Je höher die Ebene, desto weniger Elemente enthält der Name. Als Beispiel dient der Heider-Geburtsindex der Pfarre Goisern:

Beschreibung	Schema	Beispiel
Pfarrordner	Pfarre	Bad Goisern
Ordner für Scandateien	Pfarre_ta_Scan	Goisern_GH_Scan
Ordner für Textdateien	Pfarre_ta_Text	Goisern_GH_Text

Scandatei, 23. Seite	Pfarre_ta-nnn.jpg	Goisern_GH-023.jpg
Textdatei, 23. Seite	Pfarre_ta-nnn.txt	Goisern_GH-023.txt
Zusammenfassung der Textdateien	Pfarre_ta_Buch.xls	Goisern_GH_Buch.xls
Ladedatei für die FA Datenbank	Pfarre_ta_DB.xls	Goisern_GH_DB.xls

Der volle Name der Pfarre kommt nur im Pfarrordner vor. In den Unterordnern und in den Dateien wird eine Kurzform verwendet (die aber oft mit der Langform identisch ist) damit der Name nicht zu lang wird. Dieses Namensschema ist zumindest für die Speicherung der Dateien auf meinem Computer gültig. Sobald die Daten zur Bearbeitung an Kollegen weitergegeben werden, mag sich an den Strukturen und Namen etwas ändern. Das Prinzip, dass jeder Datensatz immer eindeutig identifizierbar und auf seine Quelle rückführbar ist, bleibt aber bis in die FA-Datenbank erhalten.

1.6 Erweiterung des Namensschemas für Ordner und Dateien

Das bisher größte verarbeitete Indexbuch ist der Heiratsindex von Heider der Pfarre Ried in der Riedmark mit 419 Seiten. Ein Buch mit dieser Dicke ist schon sehr unhandlich. Die Gesamtseitenzahl des Heiratsindex von Gallneukirchen beträgt 856 Seiten. Heider hat bei derart großen Pfarren den Index auf mehrere Bücher aufgeteilt. Diese Aufteilung erfolgte auf zwei mögliche Arten: Entweder nach dem Buchstaben des Index oder nach dem Heiratsdatum.

Beispiel für Buchaufteilung nach Indexbuchstaben der Pfarre Altmünster:

Altmünster: 1. Buch Index A bis G,
2. Buch Index H bis Z

Beispiel für Buchaufteilung nach Datum der Pfarre Gallneukirchen:

Gallneukirchen: 1. Buch 1600 bis 1670
2. Buch 1671 bis 1724
3. Buch 1725 bis 1784

In diesen Fällen ist eine zusätzliche Unterscheidung in der Namensgebung der Ordner und der Dateien notwendig. Dazu wird zwischen dem Autor und der Seitennummer entweder ein Buchstabe für den ersten Indexbuchstaben oder das Jahr des Beginnes des Indexbuches eingeschoben.

Pfarre_ta_i_nnn.sfx
|
|_____ erster Indexbuchstabe des Buches

Pfarre_ta_jjjj_nnn.sfx
|
|_____ erstes Heiratsjahr des Buches

Beispiel für Buchaufteilung nach Indexbuchstaben der Pfarre Altmünster:

Typ	Beispiel
Pfarrordner	Altmünster
1. Ordner für Scandateien	Altmünster_HH_A_Scan
2. Ordner für Scandateien	Altmünster_HH_H_Scan
1. Ordner für Textdateien	Altmünster_HH_A_Text
2. Ordner für Textdateien	Altmünster_HH_H_Text
Scandatei der 23. Seite im 1. Heiratsindexbuch	Altmünster_HH_A-023.jpg

Scandatei der 482. Seite im 2. Heiratsindexbuch	Altmünster_HH_H-482.jpg
Textdatei der 23. Seite im 1. Heiratsindexbuch	Altmünster_HH_A-023.txt
Textdatei der 482. Seite im 2. Heiratsindexbuch	Altmünster_HH_H-482.txt

Beispiel für Buchaufteilung nach Datum der Pfarre Gallneukirchen:

Typ	Beispiel
Pfarrordner	Gallneukirchen
1. Ordner für Scandateien	Gallneukirchen_HH_1600_Scan
2. Ordner für Scandateien	Gallneukirchen_HH_1671_Scan
3. Ordner für Scandateien	Gallneukirchen_HH_1725_Scan
1. Ordner für Textdateien	Gallneukirchen_HH_1600_Text
2. Ordner für Textdateien	Gallneukirchen_HH_1671_Text
3. Ordner für Textdateien	Gallneukirchen_HH_1725_Text
Scandatei der 23. Seite im 1. Heiratsindexbuch	Gallneukirchen_HH_1600-023.jpg
Scandatei der 18. Seite im 2. Heiratsindexbuch	Gallneukirchen_HH_1671-018.jpg
Scandatei der 56. Seite im 3. Heiratsindexbuch	Gallneukirchen_HH_1725-056.jpg
Textdatei der 23. im 1. Heiratsindexbuch	Gallneukirchen_HH_1600-023.txt
Textdatei der 18. im 2. Heiratsindexbuch	Gallneukirchen_HH_1671-018.txt
Textdatei der 56. im 3. Heiratsindexbuch	Gallneukirchen_HH_1725-056.txt

Die hohe Seitennummer im 2. Heiratsbuch von Altmünster kommt dadurch zustande, weil bei Aufteilung eines Index auf mehrere Bücher die Seitennummer in den Folgebüchern weitergezählt wird, also nicht wieder bei 1 beginnt. Die Eindeutigkeit eines Dateinamens wäre daher auch dann gegeben, wenn es jeweils nur einen Ordner für die Scandateien und für die Textdateien gäbe, also die Untergliederung nach den Indexbuchstaben wegfiel. Ich habe jedoch bewusst die Gliederung der Heiderbücher in der elektronischen Speicherung beibehalten, damit die Beziehung Buch-Ordner erhalten bleibt. Außerdem hat Heider bei der Pfarre Gallneukirchen wieder eine andere Vorgehensweise gewählt, vermutlich um beim Geburtsindex Seitennummern über 999 zu vermeiden.

Beim Geburtsindex von Gallneukirchen mit ca. 1.600 Seiten muss selbst die oben geschilderte Aufteilung noch erweitert werden. Hier hat Heider seine Bücher sowohl nach Jahren und innerhalb der Jahre nach Buchstaben gegliedert, also eine Kombination aus beiden Methoden. Entsprechend muss auch in der Namensgebung Jahr und Indexbuchstabe kombiniert werden.

Beispiel für die 277. Scansseite des Indexbuches Geburten der Jahre 1601 bis 1670 für Indexbuchstaben A bis H:

Gallneukirchen_HH_1600A-277.jpg

1.7 Änderung des automatischen Dateinamens vom Buchscanner auf die neue Namensgebung

Der Buchscanner vergibt beim Einscannen für die Dateien eine automatische Dateinummer, die auf das vorbeschriebene Schema zu ändern ist. Das manuelle Ändern hunderter von Dateinamen wäre ein unvermeidbar hoher Aufwand und zudem fehleranfällig. Das Neu Nummerieren erfolgt daher per Software. Ich verwende dafür ACDSee, ein Programm zur Verwaltung von Bildern. Der gleichbleibende Text (z.B. Gallneukirchen_HH_1725-) wird fix definiert, die Dateinummer wird vom Programm automatisch vergeben. Durch fehlerhafte Nummerierung in den Heider-Büchern kommt es dabei allerdings zu Abweichungen, die durch Vergleich der Seitennummer mit der Da-

teinummer gefunden werden. Die automatisch vergebene Dateinummer muss dann gegebenenfalls angepasst werden (siehe Kapitel *Nachbearbeitung der digitalisierten Buchseiten auf Seite 21*). Das Ändern der Dateinamen ist nur für die Scans notwendig. Die bei der OCR-Schrifterkennung entstehenden Textdateien übernehmen den Namen der Scandatei, jedoch mit der Dateierweiterung txt.

2 Allgemeines zum Heider-Projekt

2.1 Was ist der Heider-Index?

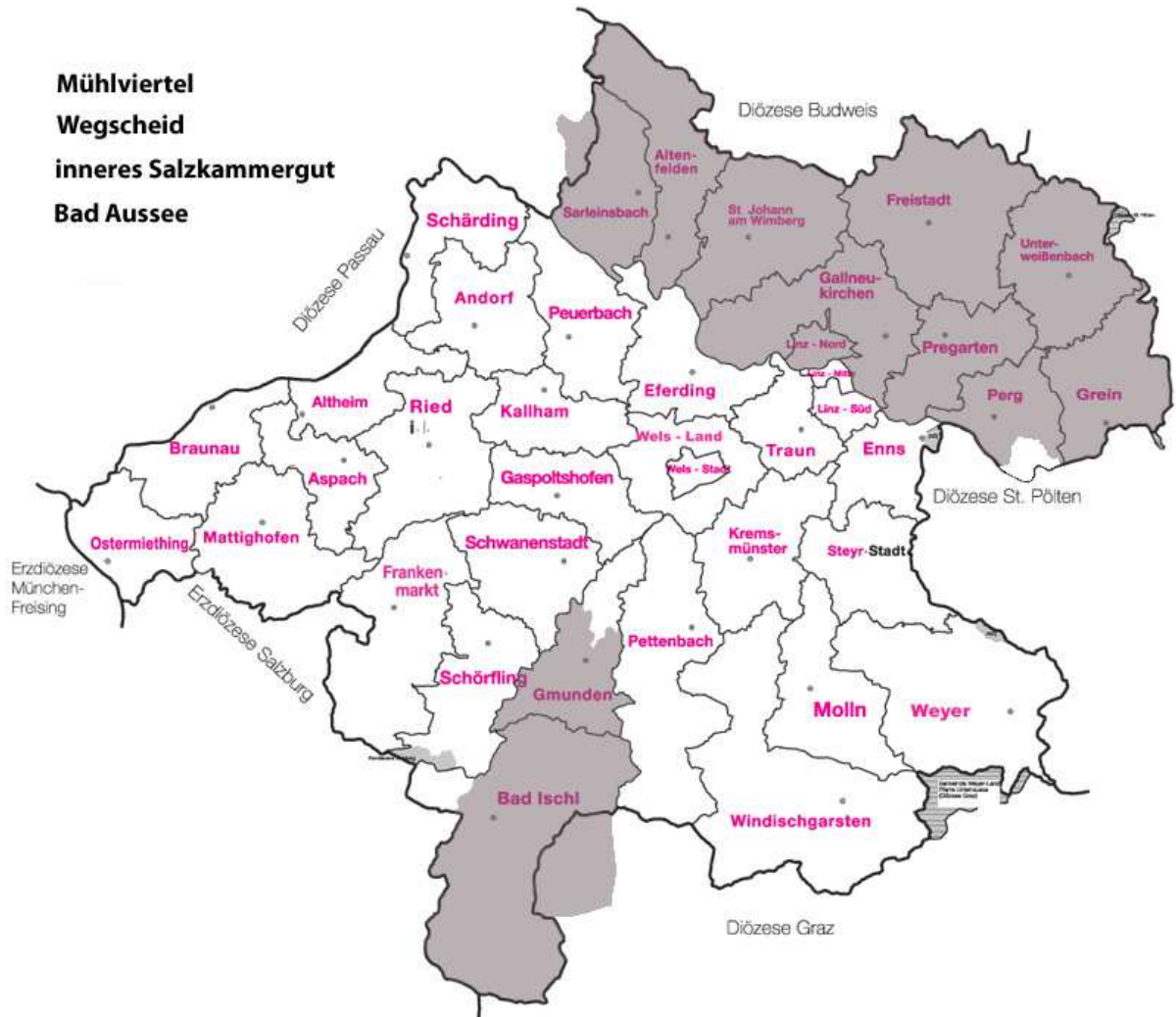


Abb. 1: Abdeckungsbereich des Heider-Index

Der Heider-Index oder richtiger die Heider-Regesten sind ein ausführlicher Index der Geburts-, Heirats- und Sterbematrikeln von deren Beginn bis mindestens 1784 aller Pfarren des Mühlviertels mit Ausnahme der Pfarre Mitterkirchen, der Pfarre Wegscheid in Bayern und der Pfarren des Salzkammergutes. Die Daten sind mit Schreibmaschine geschrieben und in Büchern gebunden. Ein kompletter Satz der Bücher befindet sich im Oberösterreichischen Landesarchiv (OÖLA), ein zweiter Satz befindet sich bei der heraldisch genealogischen Gesellschaft Adler in Wien. Jede Diözese bekam einen Satz für ihren Bereich und in den Pfarren befinden sich die Heider-Bücher der Pfarre. Benannt ist der Index nach Josef Heider, der dieses gewaltige Werk in den Jahren 1957 bis 1983 erstellt hat.

2.2 Josef Heider

Josef Karl Heider, geb. am 13.3.1903 in Wien-Ottakring, Professor, wissenschaftlicher Konsulent der OÖ Landesregierung, Ehrenmitglied der Gesellschaft „Adler“, war Versicherungsangestellter und Prokurist in Wien, Leopoldstadt. Josef Heider hatte Vorfahren im Mühlviertel und im Salz-

kammergut, die er erforschte. Während seiner Nachforschungen in den Pfarren gab ihm der Dechant von Pabneukirchen, ein entfernter Verwandter, die Anregung, zu den Pfarrmatriken Register zu erstellen (die es in den alten Büchern in der Regel nicht gab).

2.3 **Gemeinsames Projekt des Oberösterreichischen Landesarchivs und des Vereins Familia Austria (FA)**



Abb. 2: Sepp Asanger mit Dr. Gerhart Marckhgott, Leiter des OÖLA bei der Vertragsunterzeichnung

2009 wurde die Idee geboren, die Heiratsbücher des Heider-Index zu digitalisieren und die Daten in die damals schon bestehende Hochzeitsdatenbank der Familia Austria einzuspeisen. Nach technischer Prüfung der Machbarkeit kam es 2010 zu einer vertraglichen Vereinbarung zwischen dem OÖLA und FA. Entsprechend dieser Vereinbarung darf FA die Bücher auf einem der Buchscanner des OÖLA kostenlos digitalisieren. Als Gegenleistung überlässt FA die maschinenlesbaren Daten dem OÖLA zur Verwendung in dessen lokalen Netzwerk und im Benutzerservice des OÖLA.

3 Organisation der Heiderbücher

3.1 Bücher mit Einführung je Pfarre



Abb. 3: Heider-Indexbücher der Pfarren Liebenau bis Niederkappel

Die Heider-Bücher sind je Pfarre (nach der damaligen Pfarreinteilung) erstellt. Am Beginn jedes Tauf-, Heirats- und Sterbebuches werden die Bände der Kirchenbücher (tomus), die in der Regel mit römischen Ziffern bezeichnet sind, und der in ihnen enthaltene Zeitraum der Matrikeln angegeben. Lücken in den Kirchenbüchern und andere wichtige Hinweise auf den Zustand der Kirchenbücher, Auslagerungen in andere Bücher oder

Pfarren und Namenssynonyme finden sich dort neben der Angabe, wann J. Heider das jeweilige Buch erstellt hat. Oft sind auch Fotos der Kirche und eine geografische Darstellung eingeklebt. Im ersten Taufbuch gibt es meistens einen kurzen geschichtlichen Überblick zur Pfarre und manchmal auch Häuserverzeichnisse.

3.2 Aufteilung der Heiderbücher

Abhängig von der Größe der Pfarre gibt es mehr oder weniger Bücher in unterschiedlicher Zusammenfassung. Bei kleinen Pfarren sind Heirats- und Sterbeindex in einem gemeinsamen Buch zusammengefasst. Für die Pfarre Kefermarkt und bei den evangelischen Landhausmatrikeln von Linz gibt es für alle drei Kategorien nur ein Heider-Buch. Bei sehr großen Pfarren hingegen sind Tauf- und Heiratsindex auf mehrere Indexbücher aufgeteilt, weil die Bücher sonst zu dick geworden wären. Die Aufteilung der Bücher erfolgt entweder nach Buchstabengruppen oder nach Zeiträumen (siehe auch *Erweiterung des Namensschemas für Ordner und Dateien auf Seite 8*). In Hallstadt sind die Jahre 1602 bis 1784 nach Buchstaben aufgeteilt, die Jahre 1785 bis 1852 sind für alle drei Kategorien in einem Buch zusammengefasst. Es gibt also die unterschiedlichsten Kombinationen, die in der Kontrolldatei auch dargestellt werden. (Der Hinweis $\frac{1}{2}$ z.B. bei Heirat und bei Sterbe bedeutet, dass ein Buch zur Hälfte aus den Heiraten und zur Hälfte aus den Sterbeeintragungen besteht.)

3.3 Alphabetischer Index

Vor allem vor dem 18. Jahrhundert wurden bei bestimmten Buchstaben nicht klar differenziert. Das betrifft die Buchstaben B und P, C und K, D und T, F und V und I und J. Heider hat daher diese Buchstaben unter B, C, D, F bzw. I zusammengefasst. Andererseits hat er den sehr starken Buchstaben S auf Sch, St und S aufgeteilt.

3.4 Seiten

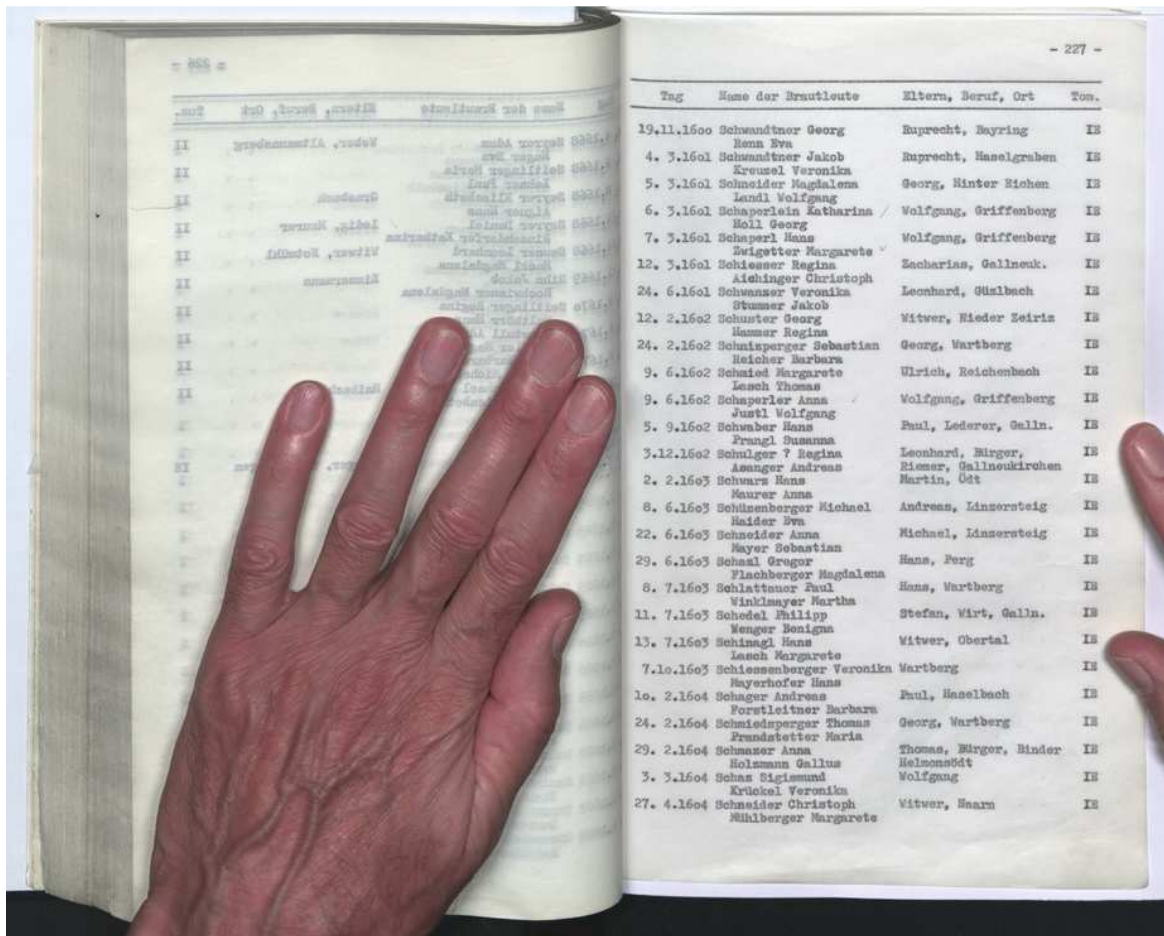


Abb. 4: geöffnetes Heider-Buch

Heider hat für seine Bücher ein sehr dünnes, durchscheinendes Papier verwendet und einseitig beschrieben. Die Seiten mit Indexeintragungen sind fortlaufend nummeriert. Allgemeine Seiten am Beginn eines Buches oder die Seiten mit dem Text Anhang am Ende eines Buches haben keine Seitennummer. Bei der Nummerierung sind dem Schreiber oder der Schreiberin häufig Fehler wie Ziffernstruz, Zehner- oder auch Hundertersprünge vor und zurück unterlaufen. Dadurch entstanden Lücken in der Seitennummerierung bzw. doppelte Seitennummern.

3.5 Mengenangaben

Auf dem Titelbild dieses Dokumentes ist das Regal im OÖLA zu sehen, in dem die Heider-Bücher aufbewahrt werden. Es sind die oberen drei Reihen mit 306 Indexbüchern, die, aneinander gereiht, 6,3 Meter ergeben mit mehr als 74.000 Seiten. In der vierten Reihe von oben sind noch einige weitere blau gebundene Bücher zu sehen. Es sind Aufstellungen von Familiennamen und deren Verbreitung, die ebenfalls Heider angelegt hat.

Die folgende Tabelle zeigt die Seitenanzahl nach Kategorie, wobei bei Geburten und Sterbe nur die Indexseiten lt. Heider ohne den nicht nummerierten einführenden Seiten mit Anmerkungen und Beschreibungen gezählt wurden.

Kategorie	Seiten
Geburt	33.640
Heirat	18.030
Sterbe	22.612
Summe	74.282

4 Aufbau der Heider Heiratsregesten

4.1 Allgemeines

Das Projekt ist grundsätzlich nicht auf die Heiratsbücher beschränkt. Auf Grund der besonderen Bedeutung der Heiratsbücher für die genealogische Forschung und auf Grund des enormen Umfanges wurde entschieden, mit den Heiratsbüchern von Josef Heider zu beginnen. Da der Heiratsindex der komplexeste aller 3 Indizes ist, wurde damit der technisch schwierigste Teil in Angriff genommen. Um die Beschreibung der durchzuführenden Arbeiten des Digitalisierungsprojektes zu verstehen, muss man den Aufbau des Heiratsindex im Detail kennen.

4.2 Paarweise Erfassung der Brautleute

Jedes Brautpaar kommt in der Regel doppelt vor, einmal unter dem Anfangsbuchstaben der Braut und unter dem Anfangsbuchstaben des Bräutigams. Damit ist sichergestellt, dass man nach beiden Namen suchen kann. In diesem Dokument nenne ich die zweite Indexeintragung "Gegeneintragung". Jede Indexeintragung hat mindestens 2 Zeilen. In der ersten Zeile steht der Name, der mit dem aktuellen Indexbuchstaben beginnt und darunter, etwas eingerückt, steht der zweite Name. Haben Braut und Bräutigam den gleichen Anfangsbuchstaben (bzw. fallen sie in dieselbe Buchstabenengruppe), dann folgen beide Doppelzeilen unmittelbar hintereinander: Braut-Bräutigam, Bräutigam-Braut, bzw. Bräutigam-Braut, Braut-Bräutigam, je nach alphabetischer Sortierfolge der Namen. Das sind dann also mindestens 4 Zeilen.

- 38 -

Tag	Name	Eltern, Beruf, Ort	Tom.
27. 2.1770	Peer Maria Gimbs Franz	fil. ill. Mathias Peer M.: Elisabeth Kain	C
15. 1.1770	Pirglehner Maria Klackl Josef	Jakob - Maria Bauer, Frankenmarkt	C
30. 4.1770	Prem Maria Aschauer Johann	fil. ill. Johann Prem M.: Maria Steyrer	C
28. 5.1770	Pomberger Franz Pögner Eva	Ferdinand - Elisabeth Messerschmied	C
28. 5.1770	Pögner Eva Pomberger Franz	Josef - Magdalena Messerschmied	C
15.10.1770	Puz Thomas Kain Maria	Andreas - Eva auf der Wiesen	C
19.11.1770	Peer Elisabeth	Paul - Elisabeth	C

Abb. 5: Heiratspaare, Pomberger Franz und Pögner Eva kommen zwei Mal in umgedrehter Reihenfolge vor

4.3 Spalten

Der Heiratsindex beinhaltet sehr viel mehr Informationen als ein gewöhnlicher Index mit Namen und Datum (weswegen man auch von Regesten spricht). Die Inhalte sind in 4 Spalten mit den folgenden Überschriften organisiert:

Tag (Heiratsdatum)

Name der Brautleute

Eltern, Beruf, Ort**Tom.** (Buch und Seite)

Die Inhalte in diesen Spalten können zum Teil sehr variieren und müssen richtig interpretiert werden. Wegen des begrenzten Platzes im A\$-Hochformat sind die Spalten nicht hundertprozentig abgegrenzt, sondern können teilweise in die Nachbarspalte hineinragen. Welche Daten tatsächlich bei einer Indexeintragung enthalten sind hängt in erster Linie von der Quelle ab. In den frühen Heiratsbüchern waren die Matrikeln noch sehr rudimentär und entsprechend wenige Daten konnte Heider in den Index übernehmen. Andererseits hat Heider z.B. die Namen der Trauzeugen nicht übernommen, die in den Matrikeln immer vorkommen.

4.4 Erklärungen zu den Spalten**4.4.1 Tag**

Heiratsdatum im Format `tt.mm.jjjj`, steht immer am Beginn (Spalte 1) der ersten Zeile einer Indexeintragung.

Varianten: nur Jahr oder nur Jahr und Monat oder Jahr + Matrikelnummer statt eines Datums.

Das Datum, sofern vollständig angegeben, sollte innerhalb eines Indexbuchstabens aufsteigend sein. Eine Abweichung von dieser Regel kann entstehen, wenn innerhalb kurzer Zeit zwei Personen mit gleichem Anfangsbuchstaben geheiratet haben und wenn der Name mit der niedrigeren Sortierfolge ein späteres Heiratsdatum hat.

Beispiel: Abel heiratet 5 Tage nach Angerer. Durch die Sortierfolge wird Abel vor Angerer gereiht, obwohl er ein späteres Heiratsdatum hat.

Ein Datum vor 1600 und nach 1890 kommt normalerweise nicht vor und wird bei der automatischen Prüfung als Fehler angezeigt.

Vereinzelt steht statt des Heiratsdatums nur das Jahr auf Position 1 und danach die Matrikelnummer in der Form `jjjj Nr nnn`. Fallweise gibt es statt eines konkreten Datums Angaben wie Fasching oder Ostern.

4.4.2 Name

In der ersten Zeile stehen Familienname und Vorname der Person, die mit dem aktuellen Indexbuchstaben beginnt. Der Name beginnt in der 12. Position der Zeile, außer, das Heiratsdatum beinhaltet Jahr und Matrikelnummer. Diese Angabe ist um eine Stelle länger als eine normale Datumsangabe, daher beginnt in diesem Fall der Name auf Position 13. Familienname und Vorname der zweiten Person stehen in der Zeile darunter und beginnen auf Position 14.

4.4.3 Eltern, Beruf, Ort

Die Angaben in dieser Spalte beziehen sich immer auf die Person, die in der ersten Zeile steht. Die Angaben können die Eltern betreffen, den verstorbenen Ehemann einer Braut (erkennbar durch den Zusatz Witwe nach ...) oder den Bräutigam (meist wenn der Bräutigam Witwer ist), in Ausnahmefällen auch die Braut (als Witwe). Oft findet man in dieser Spalte auch Altersangaben die in den späteren Matrikeln vermerkt wurden. Die Altersangabe ist dann mit dem Text Alter, Jahre, alt, J.a. etc. gekennzeichnet.

Die verschiedenen Angaben sind durch Beistriche getrennt, z.B. `Witwer, Weber, Aigen`. Bei Platznot wurden aber oft die Leerstellen weggelassen: `Ratsbürger, Fleischhacker`. Sehr oft ist ein langer Begriff abgekürzt oder abgeteilt und in der nächsten Zeile fortgesetzt:

`Witwer, Gemainbraumei-
ster`

4.4.4 Tom. (tomus)

In der ersten Zeile der Indexeintragung steht hier immer die Nummer des Kirchenbuches, die Seitennummer fehlt häufig. Manchmal fehlt auch die Nummer des Kirchenbuches. In diesem Fall handelt es sich um einen Fehler beim Erstellen der Indexeintragung, der aber leicht erkenn- und korrigierbar ist.

Die normale Form ist `tomus pagina` (sofern die Seitennummer angegeben ist). Tomus muss vom Elterntext und von der pagina durch mindestens eine Leerstelle getrennt sein. Manchmal ist die Reihenfolge umgedreht: `pagina/tomus` (dann oft, aber nicht immer, durch einen Schrägstrich getrennt). Die Reihenfolge kann mitten in einer Buchseite wechseln. Normalerweise stehen diese Angaben in der ersten Zeile der Indexeintragung. Da die Buchseiten im Hochformat in der Breite sehr begrenzt sind, wurde bei manchen Heider-Büchern die tomus-Angabe ganz rechts geschrieben und die pagina in der 2. Zeile darunter. Eine besondere Form ist die Angabe der Seitennummer in der Elternspalte mit dem Zusatz Seite oder S. Diese Form ist extrem unübersichtlich.

Tomus und pagina sollten innerhalb eines Indexbuchstaben aufsteigend sein. Eine krasse Abweichung von dieser Regel entsteht bei den Kirchenbüchern, in denen die Matrikeleintragungen nach Ortschaften gegliedert sind. Hier beginnen innerhalb desselben Buchstabens die Matrikeln je Ortsteil datumsmäßig wieder von vorne. Dadurch springen die Seitennummern im Index, der ja nach dem Anfangsbuchstaben des Namens sortiert ist, ständig auf und ab und entziehen sich damit einer Kontrolle.

4.5 Schriftbild

Die Indexeintragungen wurden mit Schreibmaschine geschrieben. Im Laufe der Jahre kamen mindestens zwei verschiedene Schreibmaschinen zum Einsatz, wobei die Qualität der Buchstaben besser wurde und damit auch die Erkennungsrate bei der maschinellen Schrifterkennung. Um die Tastatur einfach zu halten gab es damals für die Ziffern 1 und 0 keine eigene Tasten bzw. Typen. Für die 1 wurde das l verwendet, für die 0 ein O und nur aus dem Zusammenhang ist erkennbar, ob es sich um Buchstaben oder Ziffern handelt. Die Qualität der Schrift hängt auch von der Anzahl der erstellten Durchschläge ab und von der Verwendungsdauer des Farbbandes bzw. der zwischengelegten Pauspapiere.

5 Gesamtablauf

Um die Daten der maschineschriebenen Seiten der Heiderbücher in eine der Datenbanken von FA zu übernehmen, sind eine Reihe von mehr oder minder aufwändigen Schritten notwendig. Besonderen Wert haben wir dabei auf größtmögliche Genauigkeit gelegt.

Ein Grundsatz war, die Daten in der Originalform zu übernehmen. Dabei ergaben sich aber einige Schwierigkeiten. So hat J. Heider bereits viele Namen ‚standardisiert‘. Das erleichtert zwar das Suchen nach Namen in der Datenbank, die in der originalen Schreibweise sonst schwer zu finden wären, ist aber doch eine Verfälschung der Originalquelle. Andererseits hat Heider Ortsnamen in der Originalschreibweise übernommen, die der Matrikelschreiber falsch aufgeschrieben hat. Als Beispiel sei hier der Ortsname Tragwein angeführt, der im Mühlviertel oft als Tragein ausgesprochen und manchmal auch so geschrieben wurde. In eindeutigen Fällen wurde im Hinblick auf die Datenbankabfrage der korrekte Name in die Datenbank übernommen, in diesem Beispiel also Tragwein.

Folgende Arbeitsschritte sind für den gesamten Ablauf erforderlich:

1. Digitalisieren der Heiderbücher im OÖLA
2. Nachbearbeiten der digitalisierten Bilder
3. OCR Schrifterkennung mit ABBYY Fine Reader und manuelle Korrekturen
4. Formatierung und Prüfung der Seiten mit PC Editor und weitere Korrekturen
5. Zuordnung der Index-Daten zu Excel-Spalten
6. Manuelle Prüfung durch Korrekturleser: Excel-Datei versus Bild-Dateien
7. Verschmelzung (Verheiratung) der Paare auf 1 Zeile
8. Ergänzung der Heider-Hochzeitsdaten mit überregionalen Informationen
9. Einspeisung in die DB

In dieser Dokumentation werden die Arbeitsschritte 1 bis 4 beschrieben, für die ich zuständig bin.

6 Digitalisierung der Heiderbücher am Buchscanner des OÖLA

6.1 Der Buchscanner



Abb. 6: Buchscanner im OÖLA

Der Buchscanner hat eine in der Mitte geteilte Auflagefläche, deren beiden Hälften sich in der Höhe gegeneinander verschieben lassen (Buchwippe). Damit kann ein möglichst flaches Auflegen des Buches erreicht werden, auch wenn die linke und die rechte geöffnete Buchhälfte unterschiedlich dick sind. Das verhindert jedoch nicht die Aufwölbung der Seiten in der Buchmitte. Diese Aufwölbung führt zu einer Verzerrung der eingescannten Seite und in weiterer Folge zu Problemen bei der Texterkennung. Man muss daher versuchen, mit den Fingern die zu scannende Seite möglichst zu straffen, ohne dabei Daten zu verdecken. Je dicker ein Buch und je schmaler der Seitenrand, desto schwerer gelingt dies. In besonders schwierigen Fällen verwendete ich mit Erlaubnis des OÖLA eine Glasplatte. Damit kann auch eine wegen des dünnen Papierses stark gewellte Buchseite geglättet werden. Man muss auch darauf achten, dass das Buch möglichst gerade aufliegt und die Seite durch das Straffen mit den Fingern nicht schräg verzogen wird. In *Abb. 4: geöffnetes Heider-Buch auf Seite 14* sind die Schwierigkeiten gut zu erkennen.

Ideal wäre es, wenn zumindest bei dicken Büchern die linke Buchhälfte 70° bis 80° aufgestellt wäre, weil dadurch die rechte Buchhälfte flacher auf dem Tisch läge. Dazu wäre aber eine spezielle Vorrichtung notwendig.

Am Scanner gibt es einige Einstellmöglichkeiten über Tasten am vorderen Rand des Buchscanners. Als Auflösung wähle ich 300 dpi. Wichtig ist, nur die rechte Buchhälfte beim Scannen zu erfassen (die linke Seite ist immer leer). Die automatische Optimierung durch den Scanner, die Verzerrungen und andere Störungen ausgleichen sollte, hat sich nicht bewährt.

6.2 Ausrichtung des Buches auf der Buchwippe

Man kann das Buch an der Teilungsfuge der Auflagefläche einigermaßen mittig platzieren. Sonst steht aber kein Hilfsmittel zur Verfügung, um eine gleichbleibende Positionierung des Buches sicherzustellen, die dann bei der OCR-Schrifterkennung helfen würde. Wichtig ist, dass die Seiten möglichst gerade erfasst werden. Schiefe Seiten führen bei der Schrifterkennung zu gravierenden Problemen. Leider sind bei einigen neu gebundenen Heiderbüchern die Seiten schief gebunden, so dass das Ausrichten des Buches parallel zur Teilungsfuge der Buchwippe nur begrenzt hilft. Es ist dann extrem schwierig, die erforderliche Verdrehung des Buches abzuschätzen, um einen geraden Scan der Seite zu erhalten. Ich habe einige Zeit mit einer selbst gebauten Vorrichtung experimentiert, die fest auf der rechten Buchwippe aufliegt und eine drehbare Unterlage mit einem Anschlag hat, auf der das Buch aufliegt. Damit war es wesentlich leichter, eine gleichbleibende Position des Buches auf dem Tisch einzuhalten und das Buch nötigenfalls etwas zu drehen, um einen möglichst geraden Scan zu erreichen.

Das Ergebnis des Scans wird am Bildschirm angezeigt. Häufig muss die Position des Buches oder die Seite nachjustiert und der Scan wiederholt werden. Erst wenn der Scan passt, wird die digitalisierte Seite auf einen Stick gespeichert.

6.3 Optische Verzerrung

Der Buchscanner hat keinen Scanschlitten wie ein Flachbettscanner, der parallel über die Buchseite fährt. Vielmehr befindet sich der Scanner etwa 1 Meter über dem Tisch und tastet mit einer Schwenkbewegung die Seite ab. Dabei ändern sich laufend Winkel und Abstand zur Buchseite, wodurch eine optische Verzerrung entsteht. Die geringste Verzerrung ergibt sich, wenn das Buch am hinteren Rand des Tisches liegt. Je weiter vorne das Buch liegt, desto größer ist die Verzerrung. Es ist daher zu empfehlen, das Buch etwas 2 cm vom hinteren Tischrand entfernt zu platzieren.

6.4 Durchscheinendes Papier

Das Durchscheinen der Schrift von der nächsten Seite durch das dünne Papier wird durch Einlegen eines weißen Blattes verhindert. Dieses Blatt sollte jedoch nur geringfügig über den Buchrand hinausragen, da der Scanner die überschüssige weiße Fläche mit erfasst und das Bild damit unnötig vergrößert. In *Abb. 4: geöffnetes Heider-Buch auf Seite 14* kann man das Durchscheinen der Maschinenschrift auf der linken Buchhälfte und das rechts eingelegte Zwischenblatt sehen.

Besonderheiten, die beim Einscannen eines Buches auffallen wie z.B. fehlende Seiten oder Fehler in der Seitennummerierung des Heiderbuches werden notiert, damit man bei der Nacharbeit daheim nicht irrtümlich meint, Seiten ausgelassen zu haben.

7 Nachbearbeitung der digitalisierten Buchseiten

7.1 Seitennummern auf Vollständigkeit prüfen und korrigieren

Jede Buchseite wird beim Scanvorgang mit einer automatisch vergebenen fortlaufenden Nummer (= Dateiname) auf einem Stick gespeichert. Für die weitere Verarbeitung ist es wichtig, dass über den Dateinamen direkt die Pfarre, die Indexart und die Seitennummer des Heiderbuches gefunden werden kann. Das ist durch diese Automatik jedoch nicht gewährleistet, weil

1. der Aufbau des Dateinamens völlig anders ist,
2. in den Heiderbüchern nur Seiten mit Indexdaten eine Seitennummer haben, nicht aber die allgemeinen Seiten am Beginn des Buches bzw. Seiten mit dem Text *Nachtrag*, die ja ebenfalls eingescannt werden,
3. die Seitennummerierung im Heider-Buch oft fehlerhaft ist.

7.2 Buchseiten ohne Seitennummer

Im Heiderbuch beginnt die erste Indexseite mit der Seitennummer -1-. Nichtindexseiten mit den allgemeinen Angaben zu den Kirchenbüchern haben keine Nummer. Damit die Dateien dieser Seiten ebenfalls eindeutig nummeriert sind, werden diese allgemeinen Seiten am Beginn eines Buches mit 000 und den Großbuchstaben A, B, C, D, etc. nummeriert.

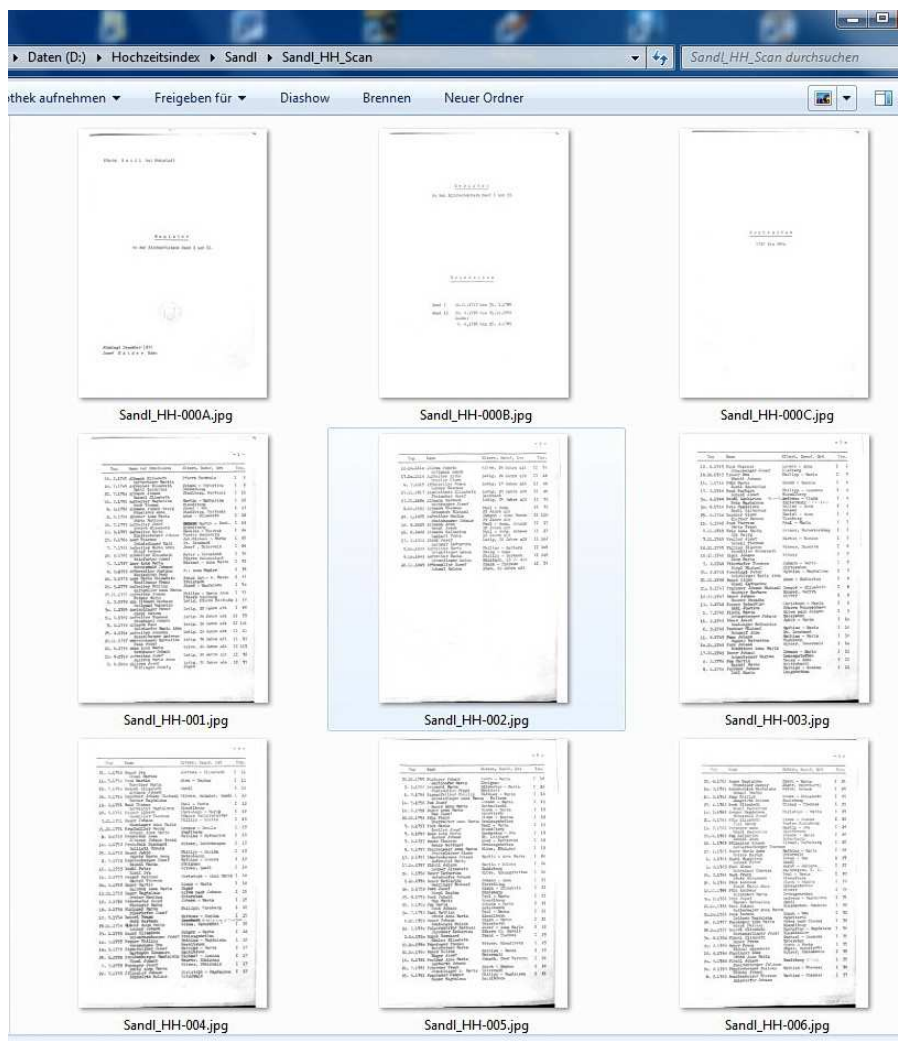


Abb. 7: Dateinamen der Scandateien

Dateien mit dem Text **Anhang** am Ende eines Buches sind ebenfalls nicht nummeriert und erhalten die Nummer der unmittelbar vorangegangenen Buchseite plus A, z.B. 095A (wenn -95- die vorhergehende Buchseite ist).

Ein Großbuchstabe in der Dateinummer ist immer ein Hinweis dafür, dass diese Nummer nur für den Dateinamen vergeben wurde und in dieser Form im Heider-Buch nicht vorkommt. Hingegen kommen im Heider-Buch sehr wohl Kleinbuchstaben in der Seitennummer vor, nämlich dann, wenn Heider eine Seitennummer nachträglich zwischen zwei Seiten eingefügt hat. Eine Nummer – 095a – wäre ein Beispiel für eine solche Nummer. Solche Seitennummern werden dann im Dateinamen ebenfalls mit dem Kleinbuchstaben übernommen (hier 095a).

7.3 Mögliche Fehler in der Nummerierung der Heider-Buchseiten und ihre Behandlung im Dateinamen

Fehlerhafte Nummerierungen werden, sobald sie bei der Bearbeitung entdeckt werden, in der Kontrolldatei protokolliert, um den Umstand für die weiteren Bearbeiter zu dokumentieren. (siehe Kapitel *Kontrolldatei auf Seite 6*).

7.3.1 Auslassungen in der Seitennummerierung

Wurden im Heider-Buch Seitennummern durch Überspringen ausgelassen, so ist dieser Umstand in der Kontrolldatei vermerkt. In den Dateinamen gibt es die identische Auslassung, Datei- und Seitennummer stimmen also überein.

Seitennummer: -68-, -70-
Dateinummer: 068, 070

7.3.2 Doppelvergabe von Seitennummern

Nachdem der Dateiname innerhalb eines Verzeichnisses eindeutig sein muss, darf dieselbe Dateinummer nicht doppelt vorkommen. Doppelte Seitennummern bekommen daher im Dateinamen fortlaufende Großbuchstaben angehängt (der Großbuchstabe ist das Zeichen dafür, dass die Nummer in dieser Form im Heiderbuch nicht vorkommt).

Seitennummer: -68-, -69-, -69-, -70-
Dateinummer: 068, 069A, 069B, 070,

7.3.3 Ziffernsturz

Ist die Seitennummer im Buch durch einen Ziffernsturz verdreht, wird die Nummer im Dateinamen berichtigt.

Seitennummer: -68-, -96-, -70-
Dateinummer: 068, 069, 070

7.3.4 Falsche Hunderterstelle

Manchmal wird über ein oder mehrere Seiten die Hunderterstelle falsch geschrieben. Wie beim Ziffernsturz wird die Nummer im Dateinamen berichtigt.

Seitennummer: -213-, -214-, -115-, -116-, -217-,
Dateinummer: 213, 214, 215, 216, 217,

7.3.5 Übersprungene Seitennummern (häufig in der Zehnerstelle)

Durch Überspringen von Nummern entstehen Lücken in der Nummerierung. Diese Lücken werden wie die Auslassungen behandelt.

Seitennummer: -63-, -64-, -75-, -76-,
Dateinummer: 063, 064, 075, 076,

7.3.6 Rücksprung in der Seitennummer

Wie bei der Doppelvergabe entstehen hier doppelte Seitennummern. In der Dateinummer bleibt die letzte richtige Seitennummer stehen, Doppelnummern werden durch Anhängen von Großbuchstaben vermieden. Damit ist über den Dateinamen sichergestellt, dass die Sortierreihenfolge der Scans richtig bleibt.

Seitennummer: -63-, -64-, -55-, -56-, -57-, -58-, -59-, -60-,
-61-, -62-, -63-, -64-, -65-,
Dateinummer: 063, 064, 064A, 064B, 064C, 064D, 064E, 064F,
064G, 064H, 064I, 064J, 065,

7.4 Seiten gerade richten

Es wurde schon erwähnt, dass die Buchseiten möglichst gerade eingescannt werden sollen. Ist eine Seite schief, so entstehen am Zeilenbeginn unerwünschte Leerzeichen, die bei der OCR Schrifterkennung zu einer Verschiebung des Zeilenbeginns führen. Bei allem Bemühen gelingt das aber am Buchscanner nur ungenügend, weshalb die Scandateien nachträglich so ausgerichtet werden müssen, dass die Zeilen vertikal genau untereinander stehen. Dafür gibt es üblicherweise in den Programmen zum Verwalten der Bilder, z.B. in ACDSee, eine entsprechende Funktion. Da die Schiefstellung von Bild zu Bild verschieden ist, kann diese Korrektur nur Bild für Bild gemacht werden. (Siehe *Abb. 14: schiefer Scan auf Seite 30*).

7.5 Seiten beschneiden

Oft entsteht beim Einscannen ein relativ großer weißer Rand. Dieser nimmt dann im Textfenster der OCR Schrifterkennung viel Platz weg. In solchen Fällen ist es hilfreich, die Scans zu beschneiden. Das Beschneiden soll aber nicht zu knapp erfolgen, weil sonst das Aufziehen und Anpassen des Rahmens für die Schrifterkennung (damit wird der Textbereich markiert, der übersetzt werden soll) schwierig wird. Ein Rand von 8 bis 10 mm ist eine praktikable Größe. (Siehe *Abb. 15: Erkennungsrahmen gerade, aber rechts zu knapp auf Seite 31*).

8 OCR-Schrifterkennung mit ABBYY FineReader

8.1 Verwendete Software

Als OCR-Software (Optical Character Recognition, optische Buchstabenerkennung) wird dzt. der ABBYY FineReader 9.0 Professional Edition von Adobe eingesetzt. Ende 2014 habe ich die Version 12 versuchsweise probiert. Die bestehenden Probleme mit Version 9.0 (Wörterbuch) wurden aber nicht verbessert. Außerdem ergab sich ein neues Problem, weil einstellige Leerzeichen manchmal durch eine Leerstelle (wie bisher), manchmal aber auch durch ein Tabulatorzeichen ersetzt werden. Damit verliert eine Textzeile völlig seine Form und das Ergebnis war unbrauchbar. Das Problem konnte mit der Supportstelle von Abbyy weder erklärt noch gelöst werden.

Die Software versucht, aus der Bilddatei (die digitalisierten Buchseiten sind computertechnisch nur Bilder) die Texte zu erkennen und damit für den Computer lesbar zu machen. Aus den Bilddateien entstehen dadurch Textdateien, die die Grundlage für die weiteren Verarbeitungsschritte sind. Die Erkennungsquote der Texterkennung hängt wesentlich von der Schrift- und von der Scanqualität ab.

8.2 Optionen

Für ein erfolgreiches Arbeiten muss ABBYY FineReader an die speziellen Bedürfnisse angepasst werden. Dies geschieht vor allem mit der Funktion *Extras/Optionen...* Hier sollen die wichtigsten Optionen erläutert werden.

8.2.1 Kein automatisches Starten des Lesens, Bildneigung

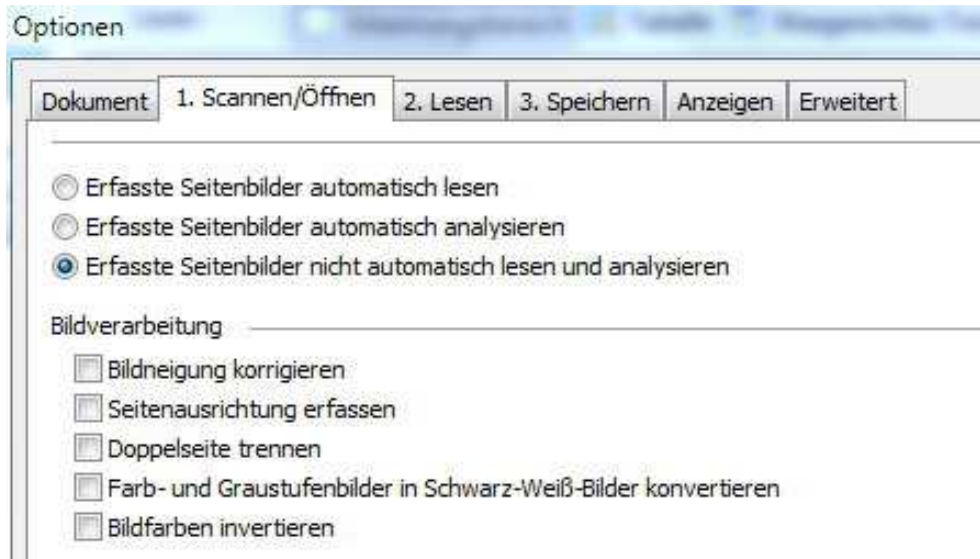


Abb. 8: Start-Option

Im Reiter *1. Scannen/Öffnen* kann aus drei Möglichkeiten ausgewählt werden, wann der Prozess der Texterkennung beginnen soll. Hier ist die Variante *Erfasste Seitenbilder nicht automatisch lesen...* zu wählen. Andernfalls beginnt der Prozess bereits unmittelbar nach dem Laden der Scans, also noch bevor der Erkennungsbereich manuell definiert werden kann.

Die Option *Bildneigung korrigieren* sollte nicht ausgewählt werden. FineReader richtet die Bilder waagrecht aus. Da durch Verzerrungen der Seite die Zeilen manchmal schief sind, werden die Zei-

len zwar waagrecht, aber der linke Rand wird dabei schief. Entscheidend ist jedoch, dass die Zeilen senkrecht untereinander stehen, was durch das Korrigieren der Bildneigung verhindert wird.

8.2.2 Speichern der Textdatei

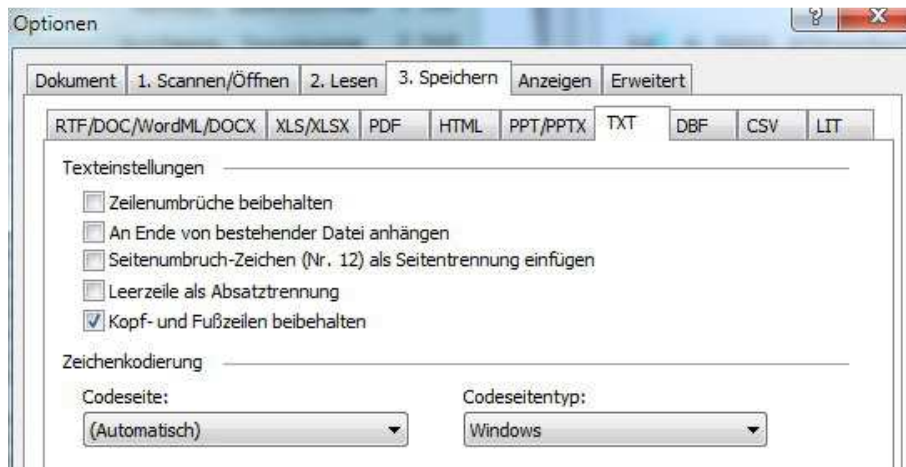


Abb. 9: Optionen für Speichern

Der übersetzte Text kann in verschiedenen Formaten abgespeichert werden (RTF, XLS, HTML, PDF, ...) Wir verwenden beim Reiter **3. Speichern** das Format *TXT*.

Kopf- und Fußzeile beibehalten ist auszuwählen, weil FineReader die Seitennummer zu einer Kopfzeile macht, die natürlich mit dem Text mitgespeichert werden muss.

8.2.3 Schriftart für Text

Die von Heider verwendeten Schreibmaschinen sind vom Typ monospace, d.h. jedes Zeichen hat gleiche Buchstabenbreite. Im Reiter *Anzeigen* muss daher für die Textanzeige ebenfalls eine monospace-Schriftart gewählt werden, weil sonst die fixe Spalteneinteilung optisch verloren geht. Ich habe deshalb beim Reiter *Anzeigen* die Schriftart *Consolas* ausgewählt.

8.2.4 Zwischenspeichern der OCR-Ergebnisse

Im Reiter *Erweitert* gibt es das Kästchen *Beim Start zuletzt geöffnetes FineReader Dokument öffnen*. Ein Häkchen bewirkt, dass nach Unterbrechung der Arbeit beim nächsten Start des FineReaders der Zustand zum Zeitpunkt des Abbrechens wieder hergestellt wird. Man kann also das Bearbeiten der Seiten jederzeit unterbrechen und beim nächsten Start mit der zuletzt bearbeiteten Seite weiter arbeiten.

8.3 Wörterbuch

FineReader kennzeichnet Wörter, die ihm unbekannt sind mit einer roten Unterstreichung, um auf diese Weise auf Erkennungs- oder Rechtschreibfehler aufmerksam zu machen. FineReader stellt dafür eine Reihe von verschiedenen Sprachen mit integrierten Wörterbüchern zur Verfügung. Da es sich bei den Kirchenmatrikeln fast ausschließlich um Eigennamen und alte Berufsbezeichnungen handelt, ist die Verwendung einer normalen Sprache mit seinem internen Wörterbuch und die Bildung von Flexionen wie Mehrzahl, Fall, Zeit etc. sinnlos bzw. sogar hinderlich. Sinnvoll ist nur ein Wörterbuch, in dem die als korrekt eingestuft Namen gespeichert werden. Man muss also zuerst ein eigenes, leeres Wörterbuch anlegen, in diese Namen und Begriffe bei ihrem ersten Auftreten gespeichert werden. Da mich das Anlegen eines solchen Wörterbuches viele Tage und unzählige

Mails mit der Supportgruppe von FineReader gekostet hat, soll der ganze Vorgang hier im Detail beschrieben werden.

Nach Installation des FineReaders ist in der Regel als Dokumentsprache Deutsch aktiviert. Es könnte auch z.B. Englisch oder Französisch sein. Alle diese Sprachen haben ihre Wörterbuchunterstützung integriert und ein Benutzerwörterbuch kann nicht ausgewählt werden. Das ist erst möglich durch Wahl einer benutzerdefinierten Sprache.

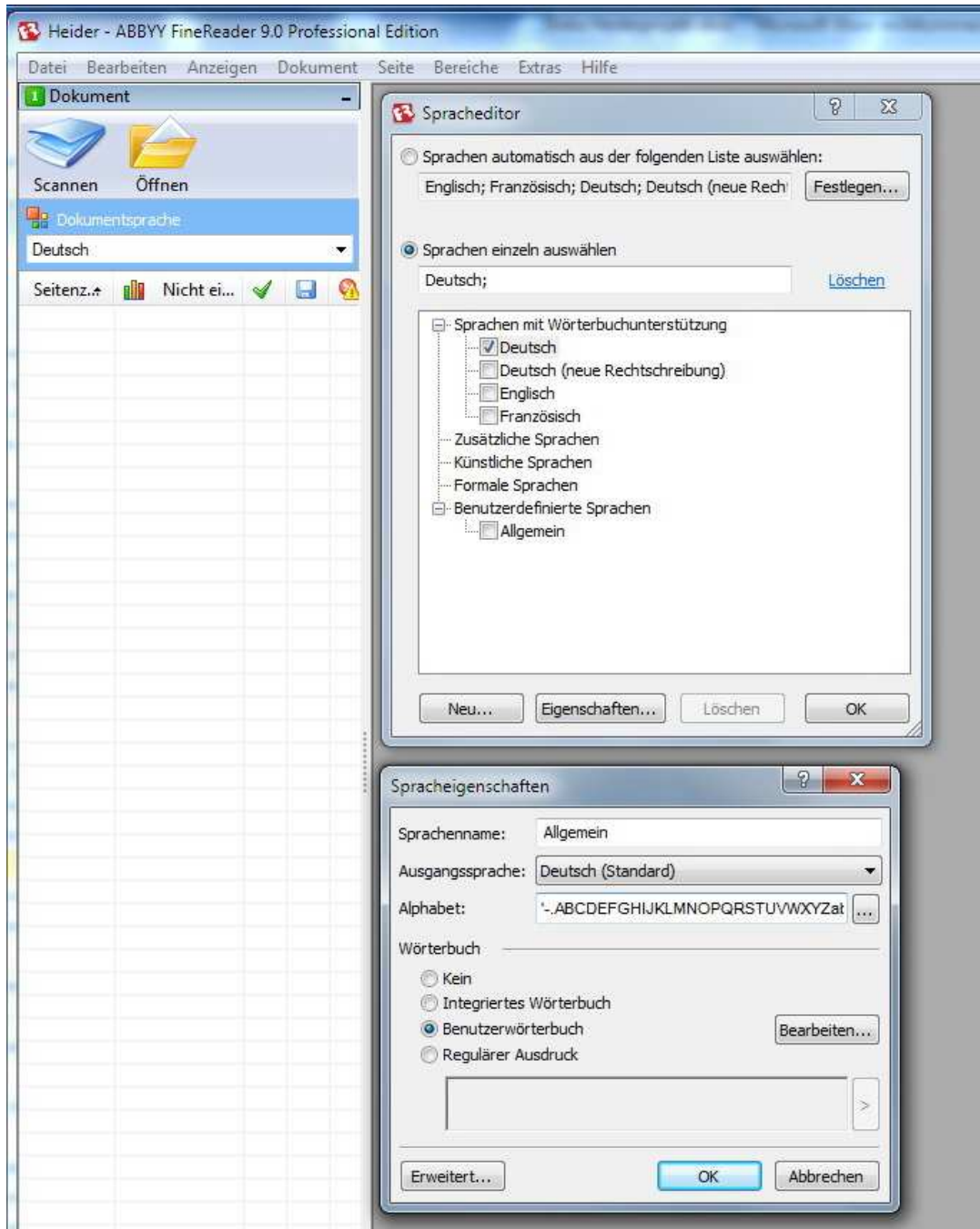


Abb. 10: Anlegen einer Benutzerdefinierten Sprache mit Benutzerwörterbuch

Im Dokumentfenster ist noch Deutsch als Dokumentsprache zu sehen. Mit *Extras/Spracheditor...* kommt man in das Fenster, in dem Benutzerdefinierte Sprachen angelegt werden können. In obiger Abbildung ist bereits die Sprache *Allgemein* angelegt. Diese Sprache entstand durch Klicken auf *Neu...*, wodurch man in das nächst Fenster kommt. Als Sprachname habe ich deshalb *Allgemein* gewählt, weil das zugehörige Benutzerwörterbuch denselben Namen *Allgemein* bekommt und da-

mit in der alphabetisch sortierten Liste der Wörterbücher ganz oben erscheint. Zunächst muss aber noch die Ausgangssprache gewählt werden, die in unserem speziellen Fall allerdings keine Bedeutung hat, und als Wörterbuch ist Benutzerwörterbuch auszuwählen.

Im Dokumentfenster kann nun als Sprache *Allgemein* ausgewählt werden. *Mit Extras/Wörterbuch anzeigen...* kommt man an die schon erwähnte Liste der Wörterbücher, in dem wir das noch leere Wörterbuch *Allgemein* ganz oben finden. Am unteren Rand des Fensters muss das Kästchen für *Microsoft Word-Wörterbuch verwenden* abgewählt werden.

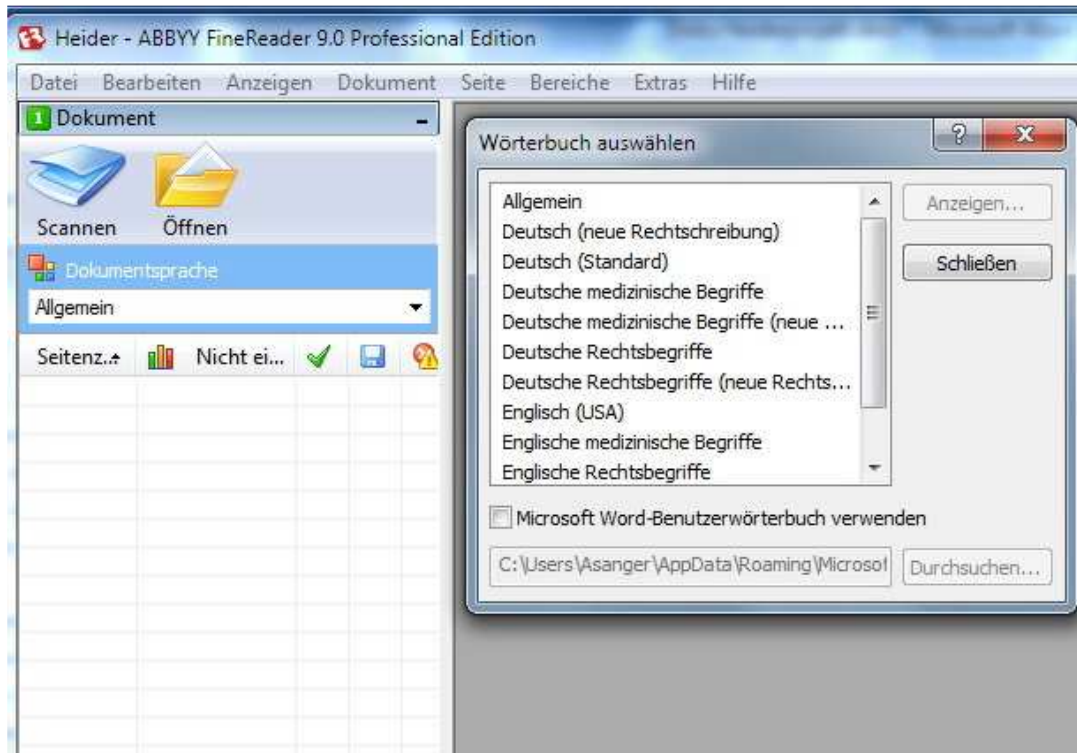


Abb. 11: Auswählen der Dokumentsprache und Anzeigen des Benutzerwörterbuches

Beim Auftreten neuer Begriffe ist zu überlegen, ob sie in das Benutzerwörterbuch übernommen werden sollen. Bei Standardnamen und -berufen ist das eindeutig. So wird man z.B. sehr bald die wichtigsten Vornamen und Berufe erfasst haben. Bei Familien- und Ortsnamen gibt es aber oft sehr viele Varianten und es ist manchmal nicht leicht zu entscheiden, ob eine Variante nur ein Schreibfehler in der Matrikel oder bei Heider ist. Auch sollte man prüfen, ob eine Variante vielleicht nur ein einmaliger Ausreißer ist oder doch öfter vorkommt. Je mehr Varianten im Wörterbuch gespeichert sind, desto größer die Gefahr, dass fehlerhaft geschriebene Namen nicht als Fehler angezeigt werden. Werden umgekehrt zu viele Begriffe durch die rote Unterstreichung markiert, besteht wieder die Gefahr, dass die Markierungen durch ihre Vielzahl nicht mehr ins Auge springen und übersehen werden.

Namen sind oft sehr ortsspezifisch und kommen dann nur in bestimmten Pfarren vor. Es ist daher sinnvoll, die gespeicherten Begriffe von Zeit zu Zeit zu überprüfen und ausgefallene Begriffe und Schreibweisen wieder aus dem Wörterbuch zu löschen und damit dessen Umfang zu reduzieren. Leider werden die Begriffe im Fenster des Wörterbuches mit einem relativ kleinen Font angezeigt, was das Bearbeiten des Wörterbuches unnötig erschwert. Außerdem ist das Aufsuchen eines Begriffes nur durch Verschieben des seitlichen Schiebers möglich und nicht z.B. durch Eingabe der ersten Buchstaben. Man kann sich daher nur sehr mühsam an den gesuchten Begriff herantasten. Wenn man umfangreichere Änderungen am Wörterbuch vornehmen möchte, ist es daher einfacher, dieses zuerst zu exportieren. Beim Exportieren wird eine Datei mit der dreistelligen Erweiterung txt

gespeichert, die man extern mit einem Editor eigener Wahl bearbeiten kann. Das veränderte Wörterbuch kann mit der Importfunktion wieder zurückgeladen werden. Zuvor muss aber das Benutzerwörterbuch im FineReader gelöscht werden, weil beim Importieren beide Wörterbücher zusammengemischt werden. Dadurch scheinen die Begriffe, die man extern gelöscht hat, nach dem Importieren wieder auf, wenn man sie nicht auch aus dem FineReader entfernt hat.

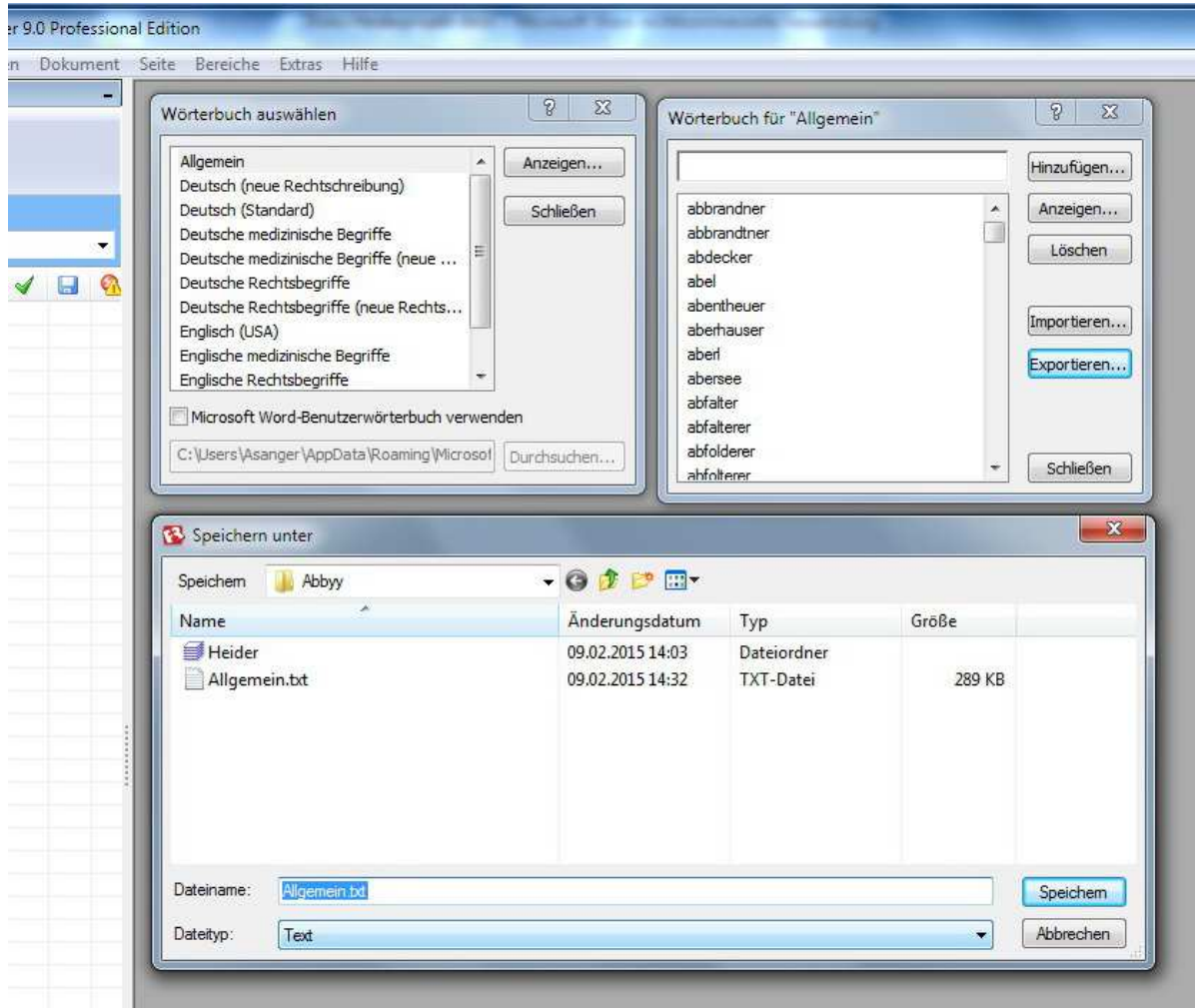


Abb. 12: Exportieren und importieren des Benutzerwörterbuches

Achtung: Nach der Funktion *Lesen* des FineReaders sind erwartungsgemäß Begriffe, die nicht im Benutzerwörterbuch enthalten sind, im Textfenster rot unterstrichen. Sobald ein solcher Begriff durch Anklicken mit der rechten Maustaste dem Wörterbuch hinzugefügt wird, verschwindet der rote Unterstrich. Dieses an sich logische Verhalten ändert sich, wenn man z.B. ein neues Wörterbuch anlegt. Ich habe stundenlang mit einem neuen Wörterbuch und allen möglichen Einstellungen experimentiert, aber die Rechtschreibprüfung wurde nicht mehr aktiv. **Bei grundlegenden Eingriffen muss die Lese-Funktion neu durchgeführt werden, damit die Rechtschreibung wieder aktiv wird.**

Ein Mangel ist die Speicherung der Begriffe ohne Großschreibung. Wird z.B. der Name Georg ins Wörterbuch aufgenommen, steht dort nur georg. Dadurch sind für den FineReader georg und Georg gültige Schreibweisen.

8.4 Öffnen der zu übersetzenden Scans (Buchseiten)



Abb. 13: Dokument-Fenster

Mit der Funktion *Öffnen* können aus dem Verzeichnis, in dem die Scans enthalten sind, eine Seite oder mehrere Seiten geöffnet werden. Für mich hat es sich als praktisch erwiesen, immer einen Zehnerblock von Scans zu öffnen, z.B. die Seiten 020 bis 029. Im Fenster *1 Dokument* sieht man die Anzahl der geöffneten Seiten und den Verarbeitungsfortschritt.

Das Beispiel zeigt das Erkennungsergebnis der ersten 5 geöffneten Scan-Dateien an und welche der Dateien nach der Bearbeitung schon gespeichert wurden (Diskettensymbol).

8.5 Bereich der Texterkennung

ABBYY FineReader kann auf eine Textstruktur eingehen. So können beispielsweise spaltenorientierte Texte auch spaltenweise bearbeitet werden. Wie wir gesehen haben, ist der Heider-Index in 4 Spalten gegliedert. Weil die Spaltengrenzen jedoch oft nicht sauber eingehalten werden, ist eine automatische spaltenweise Texterkennung nicht sinnvoll. Es wird vielmehr die ganze Buchseite bzw. der ganze Scan als ein Textbereich definiert und verarbeitet. Das zeilenweise Aufbereiten der übersetzten Texte in die einzelnen Spalten erfolgt erst nachträglich manuell und mit einem eigenen Texteditor.

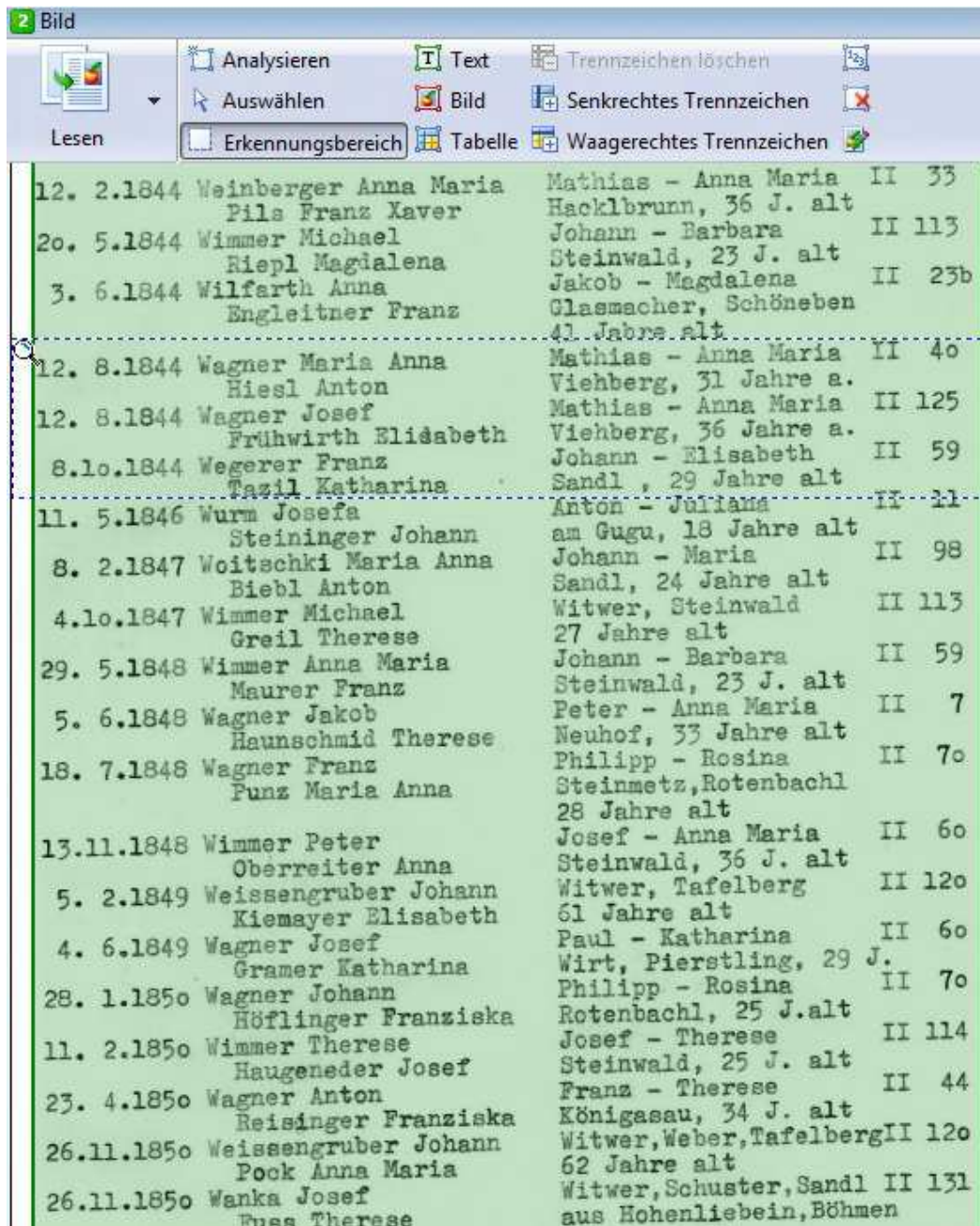
Der Erkennungsbereich kann im Fenster *2 Bild* auf 2 Arten definiert werden:

- Klicken auf das grüne Symbol für *Text* und aufziehen des Texterkennungsrahmens mit der Maus
- Laden eines vordefinierten Bereiches unter *Bereiche/Bereichsvorlage laden...* Vorher muss ein passender Rahmen mit der Funktion *Bereichsvorlage speichern...* gespeichert worden sein.

Nach Aufziehen bzw. nach Laden des Bereichsrahmens muss dieser in der Regel noch genau justiert werden. Dabei ist darauf zu achten, dass

- alle Texte innerhalb des Bereichsrahmens liegen,
- der Rahmen möglichst genau am linken Rand, also beim Datum, beginnt. Ist dort der Abstand zwischen Rahmen und erstem Zeichen zu groß, fügt FineReader unerwünschte Leerzeichen ein. Hier kann man auch gut erkennen, ob die Zeilen sauber untereinander stehen.

In folgenden Beispiel ist der Scan nicht sauber gerade ausgerichtet, daher wird der Abstand des Heiratsdatums vom Texterkennungsrahmen in der unteren Seitenhälfte immer größer, was schließlich zur Einfügung einer unerwünschten Leerstelle vor dem Datum führt. Mit der Option *Bildneigung korrigieren* im Reiter 1. *Scannen/Öffnen* werden zwar die Zeilen horizontal ausgerichtet, dadurch geht aber die vertikale Ausrichtung wieder verloren.



12.	2.1844	Weinberger Anna Maria Pils Franz Xaver	Mathias - Anna Maria Hacklbrunn, 36 J. alt	II 33
20.	5.1844	Wimmer Michael Riepl Magdalena	Johann - Barbara Steinwald, 23 J. alt	II 113
3.	6.1844	Wilfarth Anna Engleitner Franz	Jakob - Magdalena Glasmacher, Schöneben 41 Jahre alt	II 23b
12.	8.1844	Wagner Maria Anna Hiesl Anton	Mathias - Anna Maria Viehberg, 31 Jahre a.	II 40
12.	8.1844	Wagner Josef Frühwirth Elisabeth	Mathias - Anna Maria Viehberg, 36 Jahre a.	II 125
	8.10.1844	Wegerer Franz Tasil Katharina	Johann - Elisabeth Sandl, 29 Jahre alt	II 59
11.	5.1846	Wurm Josefa Steininger Johann	Anton - Juliana am Gugu, 18 Jahre alt	II 21
8.	2.1847	Woitschki Maria Anna Biebl Anton	Johann - Maria Sandl, 24 Jahre alt	II 98
4.	10.1847	Wimmer Michael Greil Therese	Witwer, Steinwald 27 Jahre alt	II 113
29.	5.1848	Wimmer Anna Maria Maurer Franz	Johann - Barbara Steinwald, 23 J. alt	II 59
5.	6.1848	Wagner Jakob Haunschmid Therese	Peter - Anna Maria Neuhof, 33 Jahre alt	II 7
18.	7.1848	Wagner Franz Furz Maria Anna	Philipp - Rosina Steinmetz, Rotenbachl 28 Jahre alt	II 70
13.	11.1848	Wimmer Peter Oberreiter Anna	Josef - Anna Maria Steinwald, 36 J. alt	II 60
5.	2.1849	Weissengruber Johann Kiemayer Elisabeth	Witwer, Tafelberg 61 Jahre alt	II 120
4.	6.1849	Wagner Josef Gramer Katharina	Paul - Katharina Wirt, Pierstling, 29 J.	II 60
28.	1.1850	Wagner Johann Höflinger Franziska	Philipp - Rosina Rotenbachl, 25 J. alt	II 70
11.	2.1850	Wimmer Therese Haugeneder Josef	Josef - Therese Steinwald, 25 J. alt	II 114
23.	4.1850	Wagner Anton Reisinger Franziska	Franz - Therese Königasau, 34 J. alt	II 44
26.	11.1850	Weissengruber Johann Pock Anna Maria	Witwer, Weber, Tafelberg 62 Jahre alt	II 120
26.	11.1850	Wanka Josef Fuss Therese	Witwer, Schuster, Sandl aus Hohenliebein, Böhmen	II 131

Abb. 14: schiefer Scan

Im nächsten Beispiel ist der Scan zwar gerade gerichtet, der Rahmen rechts aber etwas knapp an den Text gerückt, sodass in der 3. Eintragung das b bei der Seitennummer (23b) nicht mehr richtig

erkannt wird. Generell sind bei diesen beiden Beispielen die linken und rechten Ränder durch das Beschneiden des Bildes etwas zu knapp, wodurch zum Verschieben des Erkennungsrahmens sehr wenig Spielraum bleibt.

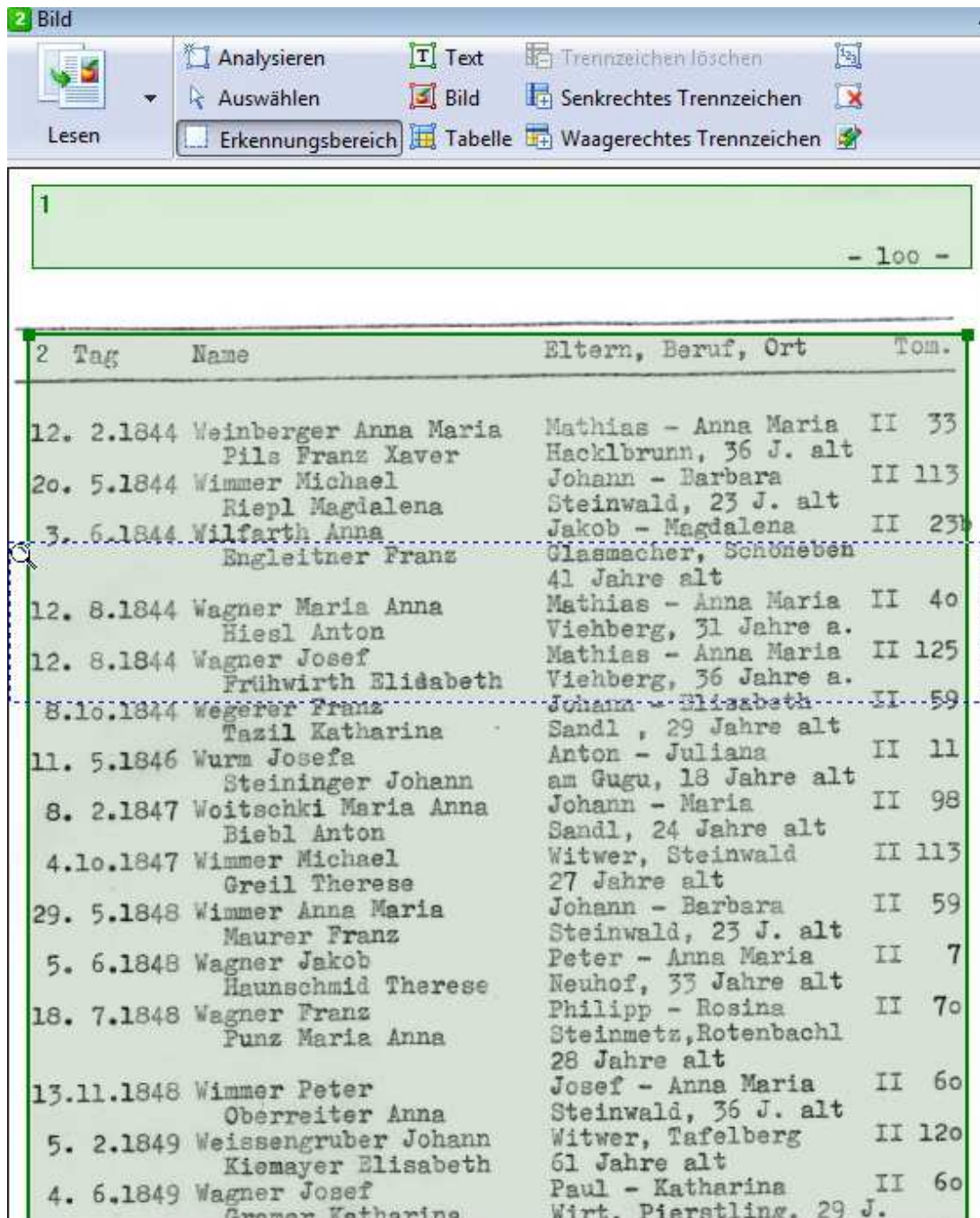


Abb. 15: Erkennungsrahmen gerade, aber rechts zu knapp

Die genaue Einstellung des Rahmens ist nur auf der linken Seite wichtig. Rechts, oben und unten können freie Bereiche bleiben. Graue Linien vom Seitenrand des Heider-Buches oder Abbilder von Fingern, mit denen die Seite beim Scannen gestrafft wird, sollten jedoch möglichst außerhalb des Rahmens liegen, weil sie sonst Störung verursachen.

Der Rahmen kann als Ganzes durch Gedrückt-halten der Taste *Steuerung* und gleichzeitiges Bewegen der Maus verschoben werden. Die Ränder können aber auch einzeln verschoben werden wie bei einem normalen Fenster. Vor allem bei einem vordefinierten Auswahlrahmen sollte man nicht

vergessen, im Fenster *2 Bild* ganz unten zu kontrollieren, ob alle Zeilen innerhalb des grünen Textfeldes liegen. Manche Buchseiten können länger sein als üblich und dann liegen die letzten Zeilen möglicherweise nicht mehr innerhalb des Rahmens und würden bei der Texterkennung verloren gehen.

Der Erkennungsrahmen lässt sich desto genauer justieren, je größer das Bild-Fenster ist. Ich maximiere daher das Bild-Fenster mit der F6-Taste. Man muss dann zwar mehr herumschrollen, um die Ränder zu kontrollieren, aber mit dem Mausrad geht das ja sehr flott und der Mehraufwand lohnt sich. Nach der Justierung des Rahmens schalte ich mit der F8-Taste das Fenster *3 Text* um.

Manchmal beginnt das Datum um eine Stelle zu weit links. Hier ist es besser, den Rahmen so zu ziehen, dass diese eine Stelle aus dem Rahmen herausfällt. Damit wird vermieden, dass bei allen anderen richtigen Zeilen eine falsche Leerstelle eingefügt wird. Bei der späteren Kontrolle hat ein solches Datum ein ungültiges Format und wird als falsch markiert. Durch Einfügen des abgeschnittenen Zeichens wird das Problem rasch behoben.

Tag	Name	Eltern; Beruf; Ort	Tom;
2. 2.1616	Egger Hieronimus	Christoph, Bürger	I 540
	Kagerer Anna	Weber, Sarleinsbach	
26. 4.1616	Eitlesberger Barbara	Sebastian, Ruzersdorf	I 541
	Grosshaupt Zacharias		
9.11.1616	Eberl Paul	Hans	I 542
	Lichtmüller Sabina		
5.1619	Erlpauer Barbara	Mathias, zu Erl	I 547
	Schöllinger Ambros		
14.10.1619	Ertl Christina	Peter - Katharina	I 549
	Muschler Leonhard	Bürger, Wegscheid	
11. 2.1620	Erhard Maria	zu Kasten	I 550
	Pauernfeind Wolfgang		
2. 7.1621	Elmansberger Moriz	Witwer, Hohenschlag	I 559
	Barbara		
26. 2.1623	Eckschlager Barbara	Gabriel - Elisabeth	I 569
	Gözendorfer Simon	aus der Zwettl	
12. 2.1624	Eder Maria	Sigmund - Susanna	I 572
	Altendorfer Thomas	Altenfelden	
17. 5.1624	Eder Georg	August - Apollinia	I 574
	Wiesinger Regina	Hagenhüsl	
23. 1.1625	Elmansberger Maria	Leonhard - Katharina	I 576
	Kembtner Mathias	Elmansberg	
26. 5.1626	Emersdorfer Andreas	Stefan - Barbara	I 584
	Kemeter Katharina	Hohenschlag	
15. 4.1627	Eder Benedikt	August - Apollonia	I 585
	Gruber Katharina	Hanghof	
10. 6.1627	Eidenberger Stefan	Witwer, i.d.Zellerin	I 587
	Gramesreiter Katharina		
15. 7.1627	Ederdorfer Stefan	Michael - Katharina	I 591
	Anna		

Abb. 16: Tag außerhalb des Erkennungsrahmens

8.6 Seiten lesen

Durch Klick auf die Funktion *Lesen/Seite lesen* wird die OCR Schrifterkennung für die ausgewählten Seiten gestartet. Im Fenster *1 Dokument* kann man den Fortschritt des OCR-Prozesses mitverfolgen. Je bearbeiteter Seite wird die Gesamtzahl der Zeichen angezeigt. Die unsicheren Zeichen werden in absoluter Zahl und als Prozentsatz angezeigt. Nach Ende des OCR-Prozesses werden Seiten durch einen Klick im Fenster *3 Text* angezeigt. Dabei sind die unsicheren Zeichen blau unterlegt, unbekannte Wörter sind rot unterstrichen. Ein Prozentsatz von 1 bis 3 % unsichere Zeichen ist ein guter Wert. Es sind dann zwar viele Zeichen blau unterlegt, die meisten sind aber trotzdem richtig und viele Zeichen werden bei der späteren Prüfung mit dem Texteditor automatisch

umgewandelt und brauchen daher hier nicht korrigiert zu werden. Viele Umwandlungsfehler ergeben sich in der Überschriftszeile. Nachdem diese aber zur Gänze gelöscht wird, sind sie ohne Belang.

8.7 Bearbeiten des Textes

Für das Bearbeiten des Textes sollte unbedingt das Zoomfenster eingeschaltet sein. Es befindet sich standardmäßig im unteren Teil des Bildschirms und wird über *Anzeigen/Zoomfenster/Zoomfenster anzeigen* aktiviert. Sobald im Bild- oder Textfenster ein Zeichen angeklickt wird, wird der Originaltext im Zoomfenster vergrößert dargestellt. Damit kann der Originaltext mit dem Ergebnis der Umwandlung sofort verglichen werden. Handschriftliche Korrekturen, die ja berücksichtigt werden sollen, sind oft nur im Bild-Fenster gut zu lesen. In solchen Fällen schalte ich mit der F6-Taste auf das Bildfenster um und kehre danach mit der F8-Taste wieder zum Textfenster zurück.

Da im nächsten Verarbeitungsschritt viele Umwandlungsfehler per Programm berichtigt werden, ist deren manuelle Korrektur hier nicht erforderlich. Für die gesamte Textseite betrifft das folgende Zeichen:

- -, ~, ~ -
- ; :

In verschiedenen Textblöcken werden zusätzlich spezifische Korrekturen durch das nachfolgende Prüfprogramm gemacht, auf die im jeweiligen Kapitel hingewiesen wird. Auch diese Umwandlungsfehler des FineReaders brauchen daher hier nicht manuell korrigiert zu werden.

Fehlende Leerzeichen zur Worttrennung kommen häufig vor, z.B. nach einem abgekürzten Vornamen wie A.Maria oder nur um Platz zu sparen. Fehlende Leerzeichen sind für die spätere maschinelle Verarbeitung in den Programmen von Dr. P. Haas nicht notwendig. Es entstehen aber durch zusammengezogene Worte ungültige Begriffe, die FineReader rot unterstreicht. Ich füge daher fehlende Leerzeichen immer händisch ein, was den Text übersichtlicher macht und die visuelle Kontrolle erleichtert.

Zusätzliche Leerzeichen entstehen vor allem durch die OCR-Umwandlung selbst, weil es auch bei Monospace-Fonts schwierig ist, die genaue Anzahl von Leerzeichen bei langen Abfolgen derselben in den Scans zu ermitteln. Das führt bei den letzten beiden Textblöcken häufig zu Verschiebungen und die Textblöcke beginnen dann nicht mehr exakt an der vorgesehenen Spalte. Mehrfache Leerzeichen zwischen Wörtern werden später automatisch entfernt und brauchen daher nicht korrigiert zu werden.

Als erstes sind die Zeilen nach der Seitennummer bis zur ersten Indexeintragung zu löschen. Die Überschriftszeile vom Heider-Buch und die Leerzeilen werden nicht gebraucht. In der nun folgenden Prüfung der Textseite gehe ich blockweise vor, d. h. ich konzentriere mich zuerst nur auf die Spalte mit dem Hochzeitsdatum, dann auf die Namen, danach auf die Spalte für Eltern etc. und zuletzt auf die Angaben für tomus/pagina. Die erste Prüfung betrifft jedoch die Seitennummer.

8.7.1 Seitennummer

Vor und nach der Seitennummer sollte ein - sein. Manchmal wird dieses Zeichen in *, —, ~ oder ~ umgewandelt. Diese Fehler werden später automatisch korrigiert. Auch ein fehlender – wird akzeptiert. Die Zeichen o, O, c und © werden später auf 0 umgewandelt, die Zeichen l, i und I werden auf 1 umgewandelt. Alle diese Fälle bedürfen daher keiner händischen Korrektur.

8.7.2 Hochzeitsdatum

Wichtig ist hier die Stellengenauigkeit. Der Tag muss immer auf der Position 1 beginnen (bzw. auf der Position 2 bei einstelligem Tag), Monat auf Position 3 (bzw. 4) und das Jahr auf Position 7. Der Trennpunkt zwischen Tag, Monat und Jahr wird sehr oft falsch umgewandelt. Da der Trennpunkt später automatisch gesetzt wird, ist es ohne Belang, welches Zeichen dort steht. Allerdings kann es irritieren, wenn statt des Punktes eine Ziffer oder ein Buchstabe steht. In den Spalten für Tag, Monat und Jahr werden die Zeichen o, 0, c und @ später auf 0 umgewandelt, die Zeichen l, L, i und I werden auf 1 umgewandelt, H auf 11. Alle diese Fälle bedürfen daher keiner händischen Korrektur.

Bei schlechter Schriftqualität wird die 3 oft als 8 oder auch als 5 interpretiert. Ich prüfe daher die 3 vor allem bei schlechter Schriftqualität auch dann, wenn sie nicht blau markiert ist. Auch die Ziffer 0 ist sehr verdächtig, denn normalerweise wird dafür das o verwendet. Das als Ziffer 0 umgewandelte Zeichen hingegen ist im Scan häufig eine 8.

Das Hochzeitsdatum sollte nicht absteigend sein. Diese Prüfung wird im nachfolgenden Verarbeitungsschritt mit dem Prüfprogramm von KEDIT durchgeführt.

8.7.3 Name

Der Name in der ersten Zeile muss immer auf der Position 12 beginnen, außer in den wenigen Fällen, in denen statt des Datums das Jahr mit nachfolgender Matrikelnummer steht. Dort beginnt der Name in Position 13. Die Folgezeilen beginnen auf Position 14. Gültige, rot unterstrichene Begriffe werden mit der rechten Maustaste in das Wörterbuch übernommen, sofern man sich dafür entscheidet. Offensichtliche Tippfehler und Buchstabenverdrehungen sind zu korrigieren, ebenso Namen mit kleinem Anfangsbuchstaben. Steht ein Begriff, z.B. Beruf oder Ort irrtümlich in der Namensspalte, so ist er in die nächste Spalte zu verschieben. Sind Familien- und Vornamen in umgekehrter Reihenfolge angegeben, so sind sie umzudrehen.

8.7.4 Eltern, Beruf, Ort

Leerzeichen fehlen hier häufig, um Platz zu sparen. Die Vornamen der Eltern sollten mit einem – und je einem Leerzeichen davor und danach getrennt sein. Der Beistrich ist ein Trennzeichen z.B. zwischen Stand, Beruf und Ort und ist daher wegen der späteren Analyse der Inhalte wichtig. Falsch umgewandelte Beistriche sind daher zu korrigieren. Abgekürzte Namen ergänze ich nur in eindeutigen Fällen, z.B. Schwertbg in Schwertberg.

8.7.5 Tomus

Tomus, die Nummer des Buches (in der Regel eine römische Ziffer, manchmal aber auch Groß- und Kleinbuchstaben wie A, B, ...), ist zwingend, die Seitennummer fehlt in vielen Heider-Indizes. In der späteren Verarbeitung werden folgende Zeichen per Programm korrigiert:

l, 1, i, X	I
Y, y, f, 7, v, T, ¥	V
G, 0, c	C
3, ß	B
H	II

Bei der Seitennummer werden später folgende automatische Korrekturen gemacht:

o, 0, @	0
I., i. 1	1

H

11

Da in der Seitennummer Kleinbuchstaben vorkommen können, wird c **nicht** in 0 umgewandelt.

Tomus und pagina sollten aufsteigend sein. Diese Prüfung wird im nachfolgenden Verarbeitungsschritt mit dem Prüfprogramm von KEDIT durchgeführt.

8.7.6 Falsche und doppelte Zeilen und Leerzeilen

Manchmal sind Zeilen im Heider-Buch mit dem Wort falsch gekennzeichnet, weil beim Schreiben ein Fehler erkannt wurde. Solche Zeilen sind zur Gänze zu löschen. Das gilt auch für doppelte Eintragungen, wenn sie identisch sind. Leerzeilen werden später automatisch gelöscht und bedürfen daher keiner Aktion.

8.7.7 Übertippte Zeichen und manuelle Korrekturen im Heider-Index

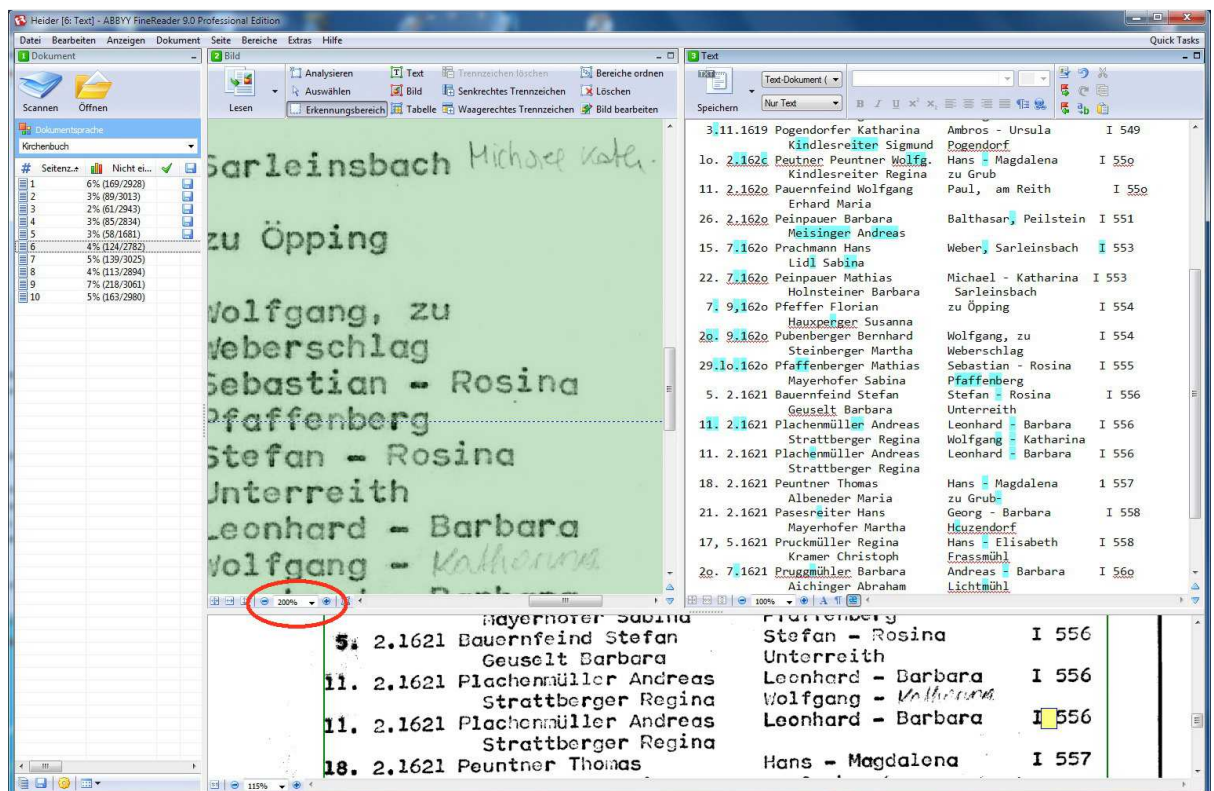


Abb. 17: händische Korrekturen im Heider-Index

Häufig findet man auf den maschinengeschriebenen Seiten der Heider-Bücher händische Eintragungen. Das sind Korrekturen von Forschern, die Fehler gefunden haben. Wir gehen davon aus, dass diese Korrekturen zu Recht angebracht wurden und übernehmen diese. Da sie mit Bleistift und oft sehr zart geschrieben sind, sind sie nur schwer oder gar nicht lesbar. Hier hilft es, das Fenster 1 *Bild* zu vergrößern. Diese Methode hilft auch in den Fällen, in denen Texte übertippt wurden und nicht ganz klar ist, welcher Buchstabe gemeint ist.

8.7.8 Bearbeitung rückgängig machen

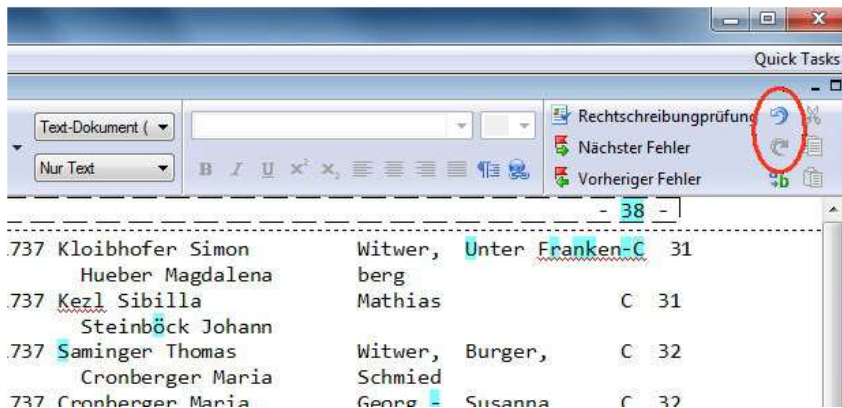


Abb. 18: OCR Textänderungen rückgängig machen

Änderungen des Textes können durch Klick auf die Icons rückgängig gemacht bzw. wieder aktiviert werden.

8.7.9 Speicherung der Textseiten



Abb. 19: Textdatei speichern

Die aus den Bilddateien und die aus ihnen entstandenen, in vorbeschriebene Weise korrigierten Textdateien haben den identischen Dateinamen. Lediglich die Dateierweiterung (die letzten drei Stellen hinter dem Punkt des Dateinamens) ist nicht **jpg** sondern **txt**. Damit ist der direkte Zusammenhang zwischen der Seite des Heider-Buches, des Scans und der Textseite gegeben. Die erforderliche

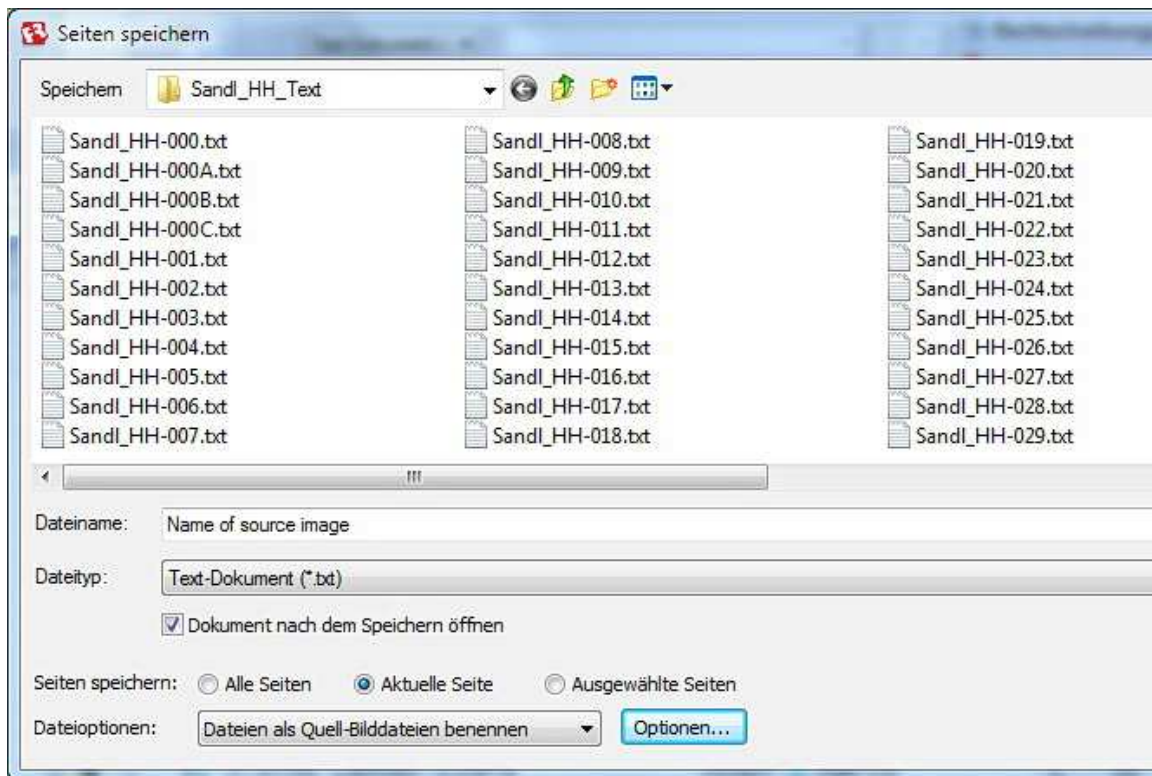


Abb. 20: Textdatei speichern und Texteditor aufrufen

derliche Einstellmöglichkeit zur Übernahme des Dateinamens findet man aber nicht unter den Optionen beim Reiter *Speichern*, wie man vermuten würde, sondern erst im Fenster *Seite speichern*. Dieses Fenster öffnet sich, sobald man die Texterkennung einer Seite abgeschlossen hat und im Fenster *3 Text* des FineReaders auf die Ikone *Speichern* klickt. In dem sich dann öffnenden Fenster muss bei Dateioptionen *Dateien als Quell-Bilddateien benennen* ausgewählt werden. Der Dateityp ist dort bereits **.txt* eingestellt.

In diesem Fenster befindet sich auch das Kästchen für *Dokument nach dem Speichern öffnen*. Hier ist ein Häkchen zu setzen, wenn man unmittelbar nach dem Speichern die Textseite mit einem Texteditor zur weiteren Bearbeitung aufrufen möchte, bevor man zur Korrektur der nächsten umgewandelten Seite geht. Diese Arbeitsweise ist deshalb empfehlenswert, weil beim Korrigieren einer Textseite manchmal Unklarheiten auftreten, die man nicht sofort klären kann oder will. Um dennoch nicht auf die spätere Klärung zu vergessen, ist bei der anschließenden Bearbeitung mit dem Texteditor ein entsprechender Kommentar anzubringen.

9 Prüfung und Seitenformatierung mit dem Texteditor KEDIT

9.1 Allgemeines

Am Anfang meiner Arbeit mit den umgewandelten Scan-Dateien habe ich die Texte ausschließlich manuell geprüft. Bald ergab sich jedoch die Frage, ob Prüfungen wie aufsteigendes Datum oder Seitennummer und saubere Formatierung durch exakte Trennung der Textblöcke nicht durch Einsatz geeigneter Software entscheidend verbessert und beschleunigt werden könnte. Auf der Suche nach einem programmierbaren PC-Texteditor stieß ich dann auf KEDIT, den ich in einer Testversion aus dem Internet auf meinen PC lud. Es stellte sich dann heraus, dass diese Testversion für meine Zwecke ausreichte und außerdem gab es sehr gute Unterstützung bei technischen Fragen, die sich bald ergaben.

KEDIT ist ein Texteditor, mit dem man sowohl zeilen- als auch blockweise Texte manipulieren kann und mit dessen umfangreicher Makrosprache eigene Programme erstellt werden können. Mit zunehmender Kenntnis der Möglichkeiten dieser Makrosprache wurden die Prüf- und Formatierungsprogramme immer umfangreicher und führten zu einer erheblichen Qualitätssteigerung der bearbeiteten Textseiten, die zu einer deutlichen Beschleunigung in der weiteren Verarbeitung führte. Die Programme sind mit vielen Kommentaren versehen, die das Verstehen der Programmlogik erleichtern helfen.

Nach Verfügbarkeit dieser programmierten Prüfungen stellte ich fest, wie viele Fehler man bei einer rein manuellen Prüfung übersieht, denn das Kontrollieren ist eine stereotype und zugleich für die Augen anstrengende Arbeit, bei der man schnell ermüdet.

9.2 Aufruf der Prüf- und Formatierungsfunktionen über Funktionstasten

Es gibt insgesamt 10 von mir erstellte Funktionen, die über die Funktionstasten F2 bis F11 aufgerufen werden.

- F2** Ein-/Ausschalten des bottom Toolbars,
siehe *Anhang* auf Seite 49 und *Arbeiten mit dem KEDIT Texteditor* auf Seite 50.
- F3** Beliebigen Begriff in den Dateien des aktuellen Verzeichnisses suchen.
Manchmal erinnert man sich an einen Begriff, den man nachträglich ändern möchte, weiß aber nicht mehr in welcher Textdatei oder in welchen Textdateien er vorkommt. Hier hilft die Suche nach diesem Begriff durch alle Textdateien des aktuellen Verzeichnisses. Mit den Pfeiltasten links und rechts kann man die Suchrichtung bestimmen.
- F4** Nach nächster Datei mit einem Kommentar suchen.
Damit kann man rasch prüfen, ob es noch Indexeintragungen gibt, die einer weiteren Nachprüfung z.B. im Original bedürfen. Mit den Pfeiltasten links und rechts kann man die Suchrichtung bestimmen.
- F5** XEDIT-Profile ändern.
Damit können Variable auf die Werte gesetzt werden, wie sie für die Textprüfung benötigt werden (siehe *Profile* auf Seite 39).
- F6** Elternblock um 4 Stellen nach rechts verschieben (siehe *Trennen Namen- und Elternblock* auf Seite 41)
- F7** Tomus-Block 4 Stellen nach rechts verschieben und anschließende Formalprüfung wie unter F8
- F8** Formalprüfung einer Textseite
- F9** Gegeneintragung zu einer Indexeintragung suchen (siehe *Ungeklärte Probleme, Kommentar und Fehlerprotokoll* auf Seite 48)

- F10** vorwärtsblättern im aktuellen Verzeichnis
F11 rückwärtsblättern im aktuellen Verzeichnis

9.3 Profile

Beim Start von KEDIT wird als erstes ein Profile durchlaufen, in dem bestimmte Programmeinstellungen gesetzt werden. In diesem Profile sind auch Werte definiert, die für die Prüfung einer Textseite benötigt werden und die sich von Buch zu Buch ändern können (siehe folgende Beschreibung). Vor Beginn der Arbeit mit einem Buch müssen diese Werte richtig eingestellt werden. Mit der Taste F5 wird das Profile zum Ändern der Werte aufgerufen. Sie werden ab dem nächsten Start von KEDIT wirksam.

9.3.1 Startposition des Elternblockes

Die Startposition des Elternblockes schwankt zwischen Position 35 und 37. Nachdem die konkrete Position ermittelt wurde muss sie in der Variablen *Elternblock* im Profile definiert werden:

```
'editv set Elternblock 36'
```

9.3.2 Reihenfolge von tomus und pagina

In den Heider-Büchern gibt es drei Varianten, wie tomus und pagina angeordnet sein können. In der Variablen TomSequence wird die aktuelle Reihenfolge definiert. Im Profile sieht dies wie folgt aus:

```
* Die Reihenfolge von Tomus und Pagina ist in manchen Büchern verkehrt
* und wird hier definiert
'editv set TomSequence C'      /* Pagina unterhalb von Tom */
'editv set TomSequence A'      /* Reihenfolge: Tom-Pag (Normalfall) */
*'editv set TomSequence B'     /* Reihenfolge: Pag-Tom */
```

(Mit einem * am Zeilenbeginn wird ein Makrobefehl zu einem Kommentar und damit unwirksam.)

9.3.3 tomus-Nummer der Kirchenbücher

Die Anzahl der von Heider erfassten Heiratsbücher einer Pfarre hängt von der Größe der Pfarre ab. Deshalb gibt es unterschiedlich viele Heiratsbücher, die Heider mit meist römischen Ziffern bezeichnet hat. Alle erlaubten tomus-Nummern bzw. Bezeichnungen, die in einem Heider-Buch vorkommen dürfen, sind in der Variablen AllTom definiert. Beispiel für Heiratsbücher I bis IV:

```
AllTom='I II III IV'
```

Die Nummern sind ein- oder mehrstellige Zeichen bzw. Begriffe, die durch eine Leerstelle getrennt sind. Hier sind z.B. I und IV erlaubte tomus-Bezeichnungen, V hingegen wäre nicht erlaubt und führt zu eine Fehlermeldung.

9.5 Trennen Namen- und Elternblock

Mein erster Blick gilt der Abgrenzung von Namen- und Elternblock (Achtung auf den unteren Teil der Seite, falls sie nicht zur Gänze im Fenster *2 Bild* sichtbar ist!). Gibt es Namen, die in den Elternblock hineinragen, werden alle Zeilen, deren Namenstext vom Elterntext durch mindestens ein Leerzeichen getrennt sind, ab dem Elternblock (minus einer Toleranzstelle) durch Drücken der F6-Taste um 4 Stellen nach rechts verschoben. Die in den Elternblock hineinragenden Namenstexte bleiben hingegen unverändert. Durch nochmaliges Drücken der F6-Taste kann der Prozess so lange wiederholt werden, bis die beiden Blöcke sauber getrennt sind. Manuelle Nacharbeit in den nicht

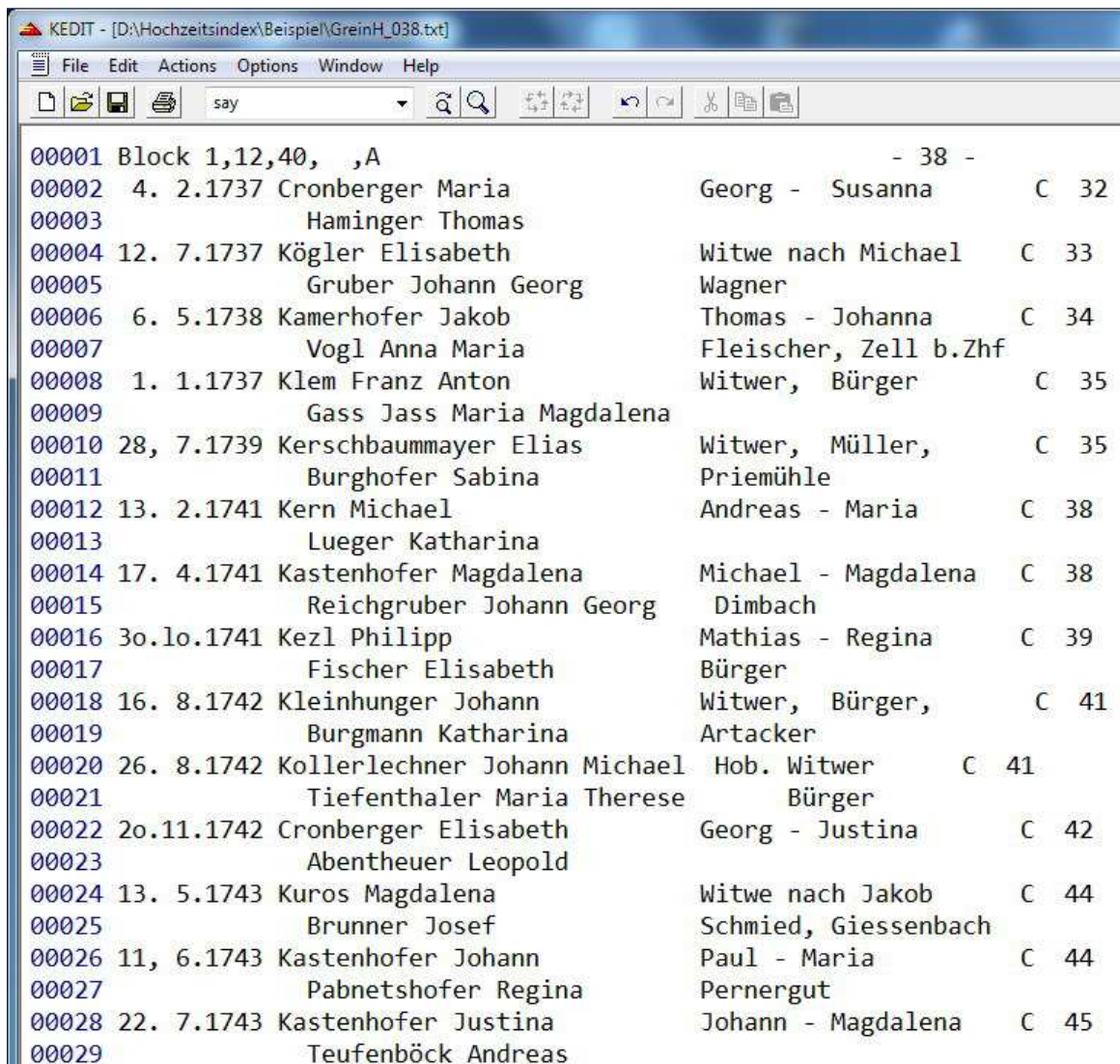


Abb. 22: Namen- und Elternblock getrennt

verschobenen Texten ist meist notwendig, weil bei den nicht verschobenen Zeilen die toms-Information gegenüber den rechtsverschobenen Texten zu weit links steht.

Auf welcher Position der Elternblock beginnt, ist in der Variablen *Elternblock* im Profile definiert. In unserem Beispiel ist das die Position 36. Die aktuelle Positionsangabe für den Elternblock der konkreten Textdatei, die sich durch das Bearbeiten verschiebt, wird in der Blockinformation der ersten Zeile der Textseite (Block) gespeichert.

Im Beispiel sehen wir 4 Namenszeilen, die in den Elternblock hineinragen. Nach Drücken von F6 sind Namens- und Elternblock bereits genügend weit auseinander gerückt. Bei Kollerlechner

Johann Michael in Zeile 00020 muss die tomus-Angabe noch händisch nach rechts verschoben werden. In Zeile 00001 sehen wir die aktuelle Elternposition (40) für diese Textdatei.

9.6 Prüfungen und Bearbeitung einer Textseite mit F7- und F8-Taste

9.6.1 Festlegung der Startpositionen der vier Textbereiche

Bei der Prüfung der gesamten Textseite muss das Prüfprogramm die 4 verschiedenen Textblöcke unterscheiden können, denn die inhaltliche Prüfung ist je Block sehr unterschiedlich. Beim Datums- und Namensblock gibt es da kein Problem, denn ihre Positionen sind mit 1 und 12 (bzw. 13) wohl definiert. Auch der Elternblock ist kein Problem, denn seine Startposition ist im Profile definiert und mit der Funktion F6 neu positioniert. Nur der tomus-Block kann Schwierigkeiten bereiten, denn sein Beginn ist sowohl durch hineinragende Elterntexte als auch durch Textverschiebungen durch die OCR-Umwandlung unsicher. Es kann auch nicht davon ausgegangen werden, dass der tomus-Block mit einem in der Variablen `AllTom` definierten Begriff beginnt, weil die tomus-Zeichen manchmal fehlen oder falsch übersetzt sind oder weil bei einer `pagina/tomus`-Reihung zuerst die Seitennummer steht. Die einfachste Lösung besteht meist darin, den Cursor an der Position zu platzieren, an der die am weitest links stehende tomus-Nummer (bzw. Seitennummer bei Reihenfolge `pagina/tomus`) beginnt, bevor man mit der F8-Taste die Prüfung zum ersten Mal startet. Ausschlaggebend ist hier ausschließlich die Position des Cursors und nicht die Zeile, in der er sich befindet. Ab diesem Zeitpunkt ist die Startposition des tomus-Blockes bekannt und die Position des Cursors ist bei weiteren Prüfaufrufen nicht mehr relevant.

Ist zwischen Elternblock und tomus-Block etwas Leerraum (vor allem bei den Matrikeln des 17. Jahrhunderts gibt es oft nur wenige Informationen und dadurch kurze Indexeintragungen), genügt es, den Cursor irgendwo innerhalb dieses Leerraumes zu platzieren.

Ragt ein Text in der **ersten** Zeile des Elternblockes über die Position, an der man den Cursor setzt, hinaus, wird der in den tomus-Bereich hineinragende Text als tomus-Information interpretiert, was zu einem falschen Ergebnis führt. In diesem Fall muss manuell eingegriffen werden, indem man den Elterntext kürzt, z.B. durch Löschen mehrfacher Leerstellen oder durch Linksverschieben des Elterntextes bis max. eine Stelle vor der Startposition des Elterntextes (eine Stelle wird toleriert) oder durch Verschieben von tomus-Informationen nach rechts. Beim Erstmaligen Drücken der F8-Taste wird der tomus-Block um 2 Stellen nach rechts verschoben, um ihn deutlich vom Elternblock abzugrenzen. Jeder weitere Aufruf mit F8 verschiebt den tomus-Block nicht mehr. Im Unterschied dazu verschiebt die F7-Funktion den tomus-Block jedes Mal um 4 Stellen nach rechts.

Eine zweite Möglichkeit ist, die F7-Taste zu drücken. Das setzt aber voraus, dass bei der Form `tomus pagina` die tomus-Information in allen Zeilen nur gültige Zeichen enthält und tomus durch mindestens eine Leerstelle vom Elterntext getrennt ist oder dass bei der Form `pagina/tomus` ein Schrägstrich gesetzt ist. Die F7-Funktion orientiert sich nämlich entweder an gültiger tomus-Information wie in der Variablen `AllTom` definiert oder am Schrägstrich und verschiebt den tomus-Block um 4 Stellen nach rechts. Bei der F7-Funktion muss der Cursor in der Nähe der tomus-Information positioniert sein, sofern es sich um den ersten Prüflauf handelt (danach ist ja die Startposition von tomus bekannt und die Cursor-Position ist nicht mehr relevant). Die Suche nach tomus `pagina` bzw. `pagina/tomus` beginnt 3 Stellen weiter links von der Cursorposition, d.h. es gibt eine Toleranzgrenze von 3 Stellen. Auf Grund dieser Arbeitsweise kann die F7-Funktion eine Mischung aus beiden tomus-Formen innerhalb derselben Seite erkennen und bearbeiten.

Elterntexte, die ab der zweiten Zeile einer Indexeintragung in den tomus-Bereich hineinragen, beeinträchtigen nicht die korrekte Bearbeitung der tomus-Information. Trotzdem sollen Eltern- und tomus-Block durch weitere F7-Aufrufe so lange auseinander gezogen werden, bis sie saubere getrennt sind, weil das die Übersichtlichkeit der Seite verbessert. Die aktuellen Startpositionen des Eltern- und des tom-Blockes werden bei jeder Manipulation in der Block-Information in der ersten Zeile der Datei gespeichert.

Ob beim ersten Prüflauf mit der F7-Taste oder mit der F8-Taste gearbeitet wird, hängt von der Blockaufteilung und der Qualität der OCR-Erkennung ab. Sind z.B. alle oder zumindest fast alle tomus-Werte richtig übersetzt, ist man mit der F7-Taste oft besser bedient, weil der Cursor nicht so genau platziert werden muss. Das Beispiel in *Abb. 23: F7-Funktion versus F8-Funktion auf Seite 43* eignet sich gut für die F7-Funktion. Sind die Erkennungen der tomus-Werte hingegen sehr schlecht, müssten alle falschen Zeichen berichtigt werden, bevor man mit der F7-Funktion startet.

ID	Date	Name	Partner	Page
00000		*** Top of File ***		
00001		- 5 -		
00002	29.10.1662	Asanger Eva	Mathias - Anna	II 62
00003		Wegerbauer Elias	Tobretshofen	
00004	19. 6.1663	Azesberger Paul	Simon ~ Eva	II 65
00005		Schlagintweit Johann	zu Hörborg	
00006	31. 1.1664	Asanger Maria	Mathias - Anna	II 67
00007		Heiter Michael	Dobretshofen	
00008	25- 2.1664	Azesberger Eva	Hans - Barbara	II 63
00009		Schwarzpauer Adam		
00010	3. 3.1666	Asanger Maria	Georg - Maria	II 80
00011		Fuxberger Michael	am Gattern	
00012	4m 7.1666	Auer Maria	Johann - Maria	II 81
00013		Mandl Casper	Hünerbach	
00014	13. 9.1666	Aichinger Helene	Adam - Magdalena	II 83
00015		Resch Wolfgang	Zimmermann, Ohnersdorf	
00016	17. 2.1667	Azesperger Sabina	Michael - Susanna	II 84
00017		Püringer Thomas	Kickingeredt	
00018	3.10.1667	A(Ö)schedl Maria	Sebastian - Maria	II 88
00019		Rottperger Johann	Galleiten	
00020	14. 5.1668	Azesberger Elias	Simon - Eva	II 89
00021		Hörezeder Magdalena	zu Henberg	
00022	2. 9.1668	Altendorfer Georg	Tobias - Maria	II 90
00023		Mauracher Katharina	Altendorf	
00024	3. 3.1669	Amesdorfer Maria	Mathias - Magdalena	II 94
00025		Scharinger Simon	Amesdorf	
00026	25. 6.1669	Aichinger Maria	Tobias - Sara	II 96
00027		Craillin Leonhard	Bürger, Bäcker	
00028	6.11.1669	Aumüller Katharina	Mathias - Margarete	II 97
00029		Reisinger Johann	Bistum Salzburg	
00030	11. 2.1670	Azesberger Barbara	Michael - Ursula	II 98
00031		Höglinger Stefan	Kickingeredt	
00032	16. 2.167c	Azesberger Johanna	Simon - Eva	II 99
00033		Magauer Josef	zu Hörberg	
00034	19. 5.1671	Aiglesberger Adam	Witwer, Mayerhof	II 107

Abb. 23: F7-Funktion versus F8-Funktion

Hier ist die F8-Funktion von Vorteil, weil die üblichen Erkennungsfehler in der tomus-Spalte vom Programm automatisch korrigiert werden. Die Prüfung der 4 Blöcke ist bei beiden Methoden identisch, weil die F7-Funktion nach der formalen Aufbereitung der Daten die F8-Funktion aufruft, für die dann schon die genauen Startpositionen der Textblöcke bekannt ist.

9.6.2 Prüfung der Seitennummer des Buches

Fehlerhafte Zeichen werden laut 8.7.1 *Seitennummer* auf Seite 33 korrigiert. Die Seitennummer wird mit der Dateinummer verglichen. Stimmen diese nicht überein, wird eine Fehlermeldung generiert.

Normalerweise müssen Buchseite und die Nummer im Dateinamen übereinstimmen. Abweichung gibt es bei fehlerhafter Seitenummerierung im Heider-Buch, siehe 7.3 *Mögliche Fehler in der Nummerierung der Heider-Buchseiten und ihre Behandlung im Dateinamen* auf Seite 22. Die Seitennummer wird überm `tomus`-Block in der ersten Zeile der Textdatei platziert.

9.6.3 Lesen der vorhergehenden Seite

Für manche der folgenden Prüfungen ist die letzte Indexeintragung der vorhergehenden Seite mit dem gleichen Indexbuchstaben maßgeblich. Vor Beginn wird daher diese Eintragung gesucht. Wird keine vorhergehende Seite gefunden, weil z.B. die aktuelle Seite die erste des Indexbuchstabens ist, wird durch eine Nachricht darauf hingewiesen.

9.6.4 Prüfung des Heiratsdatums

Das Format des Heiratsdatums und Varianten sind im Kapitel *Tag* auf Seite 16 beschrieben. Fehlerhafte Zeichen werden laut *Hochzeitsdatum* auf Seite 34 8.7.1 korrigiert. Ist das Datum absteigend, wird eine Fehlermeldung ausgegeben und die Zeile gelb markiert. Zwischen Tag und Monat wird je ein Trennpunkt gesetzt.

Handelt es sich um das erste Datum der Seite, erfolgt die Prüfung auf nicht absteigend mit dem letzten Datum der vorhergehenden Seite.

9.6.5 Prüfung der Namenszeilen

Der erste Name steht in der ersten Zeile auf Position 12, außer im Heiratsdatum stehen Jahr und die Matrikelnummer. In diesem Fall beginnt der Name auf Position 13. Der Anfangsbuchstabe muss ein Großbuchstabe sein und mit dem Indexbuchstaben der Seite übereinstimmen. Vor dem Namen kann ein `v.`, `von`, `de`, oder `(` stehen. Der Anfangsbuchstabe steht dann entsprechend weiter rechts. Bei der Prüfung auf Übereinstimmung ist das spezielle Alphabet im Heider-Index zu beachten (siehe *Alphabetischer Index* Seite 13). Eine Nichtübereinstimmung wird durch eine Fehlermeldung und durch gelbe Einfärbung der Zeile angezeigt. Nichtübereinstimmung kann entweder durch einen Fehler in der OCR-Umwandlung entstehen, der dann zu korrigieren ist, oder durch eine fehlerhafte Eintragung im Indexbuch. Es kommt relativ häufig vor, dass nach der richtigen Indexeintragung auch gleich die Gegeneintragung, also die zweite Indexeintragung des Partners bzw. der Partnerin mit dem (hoffentlich) gleichen Heiratsdatum, gemacht wurde, obwohl der zweite Name einen anderen Anfangsbuchstaben hat. In diesem Fall bleibt diese zweite Indexeintragung stehen.

Der zweite Name der Indexeintragung (und ggf. Folgezeilen) beginnt (beginnen) auf Position 14. Sind ab der zweiten Zeile die ersten 13 Stellen nicht leer, erfolgt eine Fehlermeldung und die Zeile wird gelb eingefärbt. In der Regel ist hier nur der Text um 1 Stelle nach links gerutscht und leicht zu korrigieren. Überflüssige Leerstellen in den Texten der Namenszeilen werden vom Programm entfernt.

9.6.6 Prüfung der Elternzeilen

In der Elternzeile kann es bereits zu kleinen Verschiebungen des Textes kommen, sodass der erste Buchstabe schon von der definierten Position des Elterntextes stehen kann. Deshalb wird hier max eine Stelle sozusagen als Toleranz akzeptiert, die nicht händisch korrigiert zu werden braucht. Beginnt der Elterntext weiter rechts, werden die führenden Leerstellen eliminiert, genauso wie mehrfache Leerstellen zwischen den Textteilen. Nach erfolgter Prüfung sind die Elterntexte genau an der vorgesehenen Position ausgerichtet. Besonderes Augenmerk sollte auf Punkt und Beistrich gelegt werden, weil diese öfters durch FineReader falsch umgewandelt werden. Der Punkt hat in der Regel die Bedeutung einer Abkürzung, der Beistrich trennt unterschiedliche Begriffe. Die Vornamen der Eltern sollten immer durch einen Bindestrich getrennt sein.

9.6.7 Prüfung der tomus pagina-Information

Aufgrund der vielen Variationsmöglichkeiten ist die Prüfung und Aufbereitung dieser Information besonders aufwändig. Die verschiedenen Varianten wurden bereits im Kapitel *Tom. (tomus)* auf Seite 17 besprochen. Welche Zeichen im Rahmen dieser Prüfung automatisch korrigiert werden, ist im Kapitel *Tomus* auf Seite 34 erklärt.

Wie beim Hochzeitsdatum werden tomus und pagina auf nicht absteigende Reihenfolge geprüft. Ein absteigender tomus ist in der Regel ein Fehler, vereinzelt auftretende absteigende pagina sind meist ebenfalls Fehler. Gehäuft absteigende pagina können bei den Hochzeitsmatrikeln auftreten, deren Eintragungen nach Ortschaften gegliedert sind (siehe Kapitel *Tom. (tomus)* auf Seite 17). Es wird dann oft eine ganze Liste mit Fehlermeldungen und entsprechend viele gelb eingefärbte Zeilen angezeigt, eine Überprüfung auf Richtigkeit ist nicht realistisch machbar. Hier prüfe ich nur, ob unter den vielen Hinweisen auf falsche pagina auch andere Meldungen vorkommen. Diese fallen optisch leicht auf. In dem folgenden Beispiel ist das gut zu sehen: Von den 10 Fehlermeldungen betreffen 9 die absteigende Nummerierung. Nur in der Zeile 00008 beginnt der Name mit p statt P und die anderslautende Fehlermeldung ist gut im Meldungsblock zu erkennen.

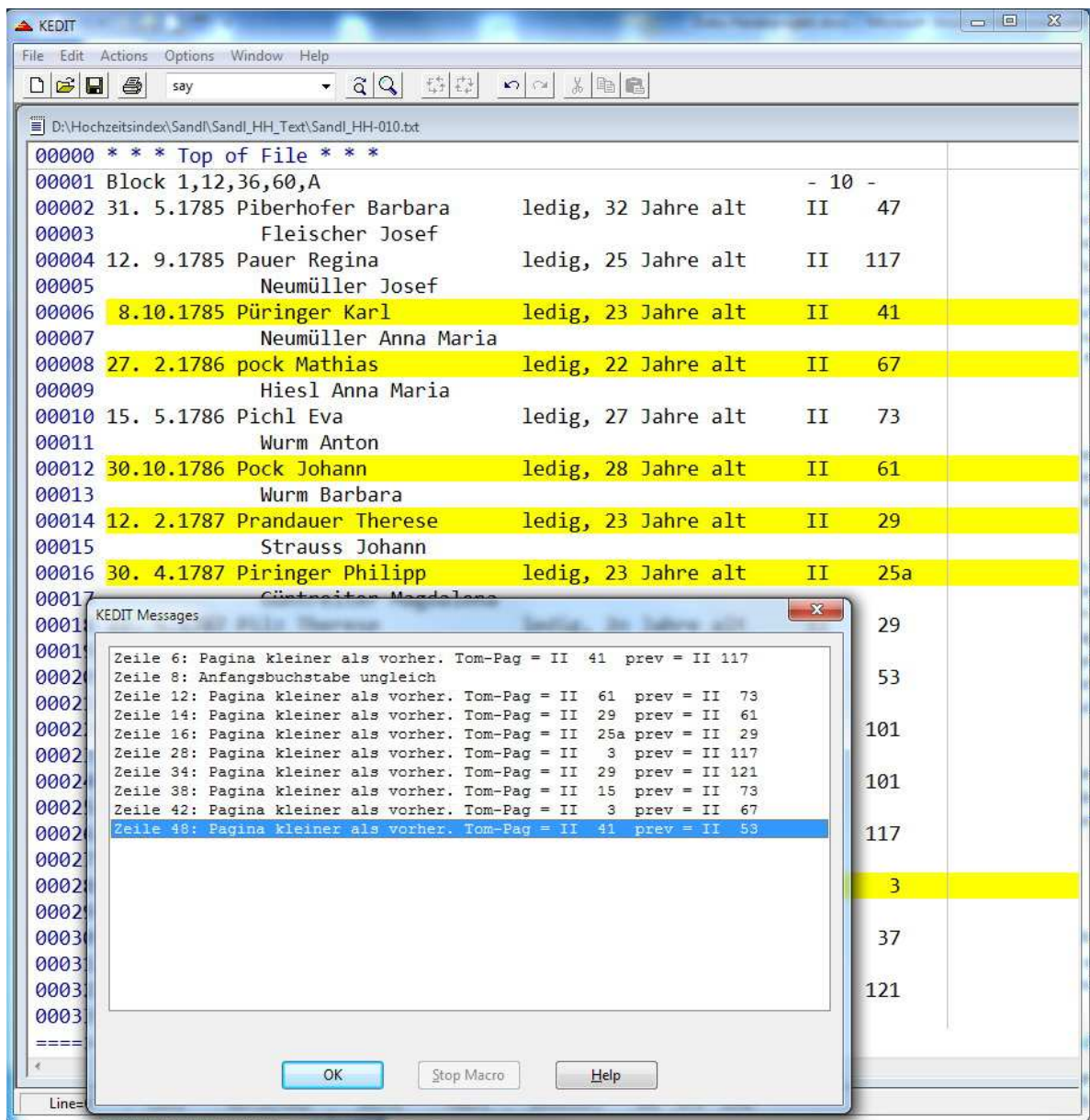


Abb. 24: Prüfung tomus pagina auf nicht absteigende Nummerierung

Die Anzeige der vielen gelb eingefärbten Zeilen wegen absteigender pagina können durch Aufruf des Programmes `xp` (exclude pagina) in der Befehlszeile von KEDIT unterdrückt werden. Dadurch ist die Anzeige der sonstigen Fehler deutlicher erkennbar.

Gleichgültig, welche Variation von `tomus` und `pagina` vorkommt, das Programm analysiert die Informationen und bringt sie in die Normalform, also `tomus pagina`, und schreibt sie sauber untereinander, wodurch die Übersichtlichkeit und Überprüfbarkeit der Textseite wesentlich verbessert wird, wie folgendes Beispiel veranschaulichen soll.

- 171 -

Tag	Name der Brautleute	Eltern, Beruf, Ort	Tom.
29. 1.1685	Singhofer Michael Mitterbacher Eva	Witwer, Koglleiten	II
10. 7.1701	Springer Jakob Moser Rosina	Hofgärtner, Waldhausen	II
20. 8.1702	Singhofer Hans Pichler Eva	Schneider, Dandorf	II
9. 7.1709	Sinhofer Barbara Ihrendorfer Johann	Michael - Sara Kefermühl	II
7. 9.1722	Singhofer Magdalena Örlinger Andreas	Urban - Barbara Öllergut	III
26.10.1744	Sperl Eva Maria Zöchbauer Johann Michael	Bernhard - Regina Bürger, Seiler, St.Flor.	IV
5. 5.1761	Seber Mathias Diessenböck Maria	Witwer, Schneider, Stangl	IV
14. 2.1774	Sauereis Andreas Preulechner Therese	Friedrich	IV
11. 6.1776	Spalt Maria Therese Hiezinger Josef	Jakob	IV
12. 5.1777	Sauereis Maria Leitner Philipp	Friedrich, Halseck	IV
30. 4.1798	Satzl Katharina Hiesböck Philipp	34 Jahre, Baumgartenberg Geisberg	V 134
23.11.1802	Sauereis Kaspar Klamhofer Rosina	22 Jahre, Priehtsberg	VI 289
3. 9.1811	Salzmann Franz Achleitner Magdalena	22 Jahre, Lederer Münzbach	VI 4
28. 4.1823	Sonnleitner Katharina Humerbichler Josef	Ledig, Dandorf	VI 66
21. 2.1829	Sauereis Josef Fröschl Elisabeth	Kaspar - Rosina 22 Jahre, Priehtsberg	VI 292
6. 9.1830	Sauereis Rosina Moser Jakob	Witwe nach Kaspar 45 Jahre, Priehtsberg	VI 293
12. 5.1833	Satzl Magdalena	Georg - Regina	VI

Abb. 25: Buch-Scan mit pagina teilweise unterhalb von tomus

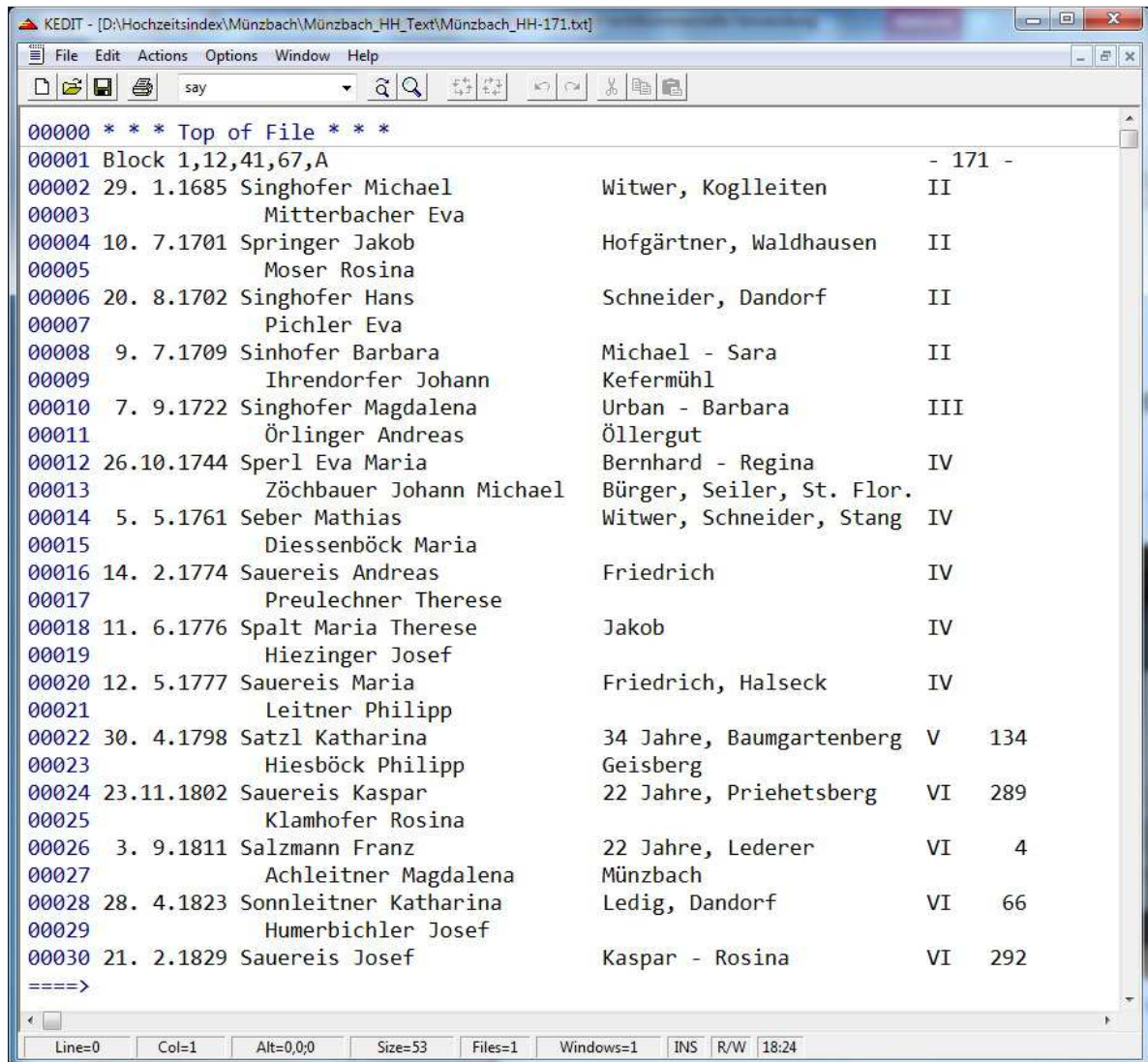


Abb. 26: aufbereitete Textseite

9.6.8 Nachtrag

Auf der letzten Seite eines Buchstabens steht zumeist ein Nachtrag, gekennzeichnet durch den Text **Nachtrag**: Die danach folgenden Indexeintragungen werden auf gleiche Weise wie oben beschrieben geprüft. Datums- und tomus-Prüfungen beginnen also wieder von vorne.

9.6.9 Rückgängig machen von Textänderungen

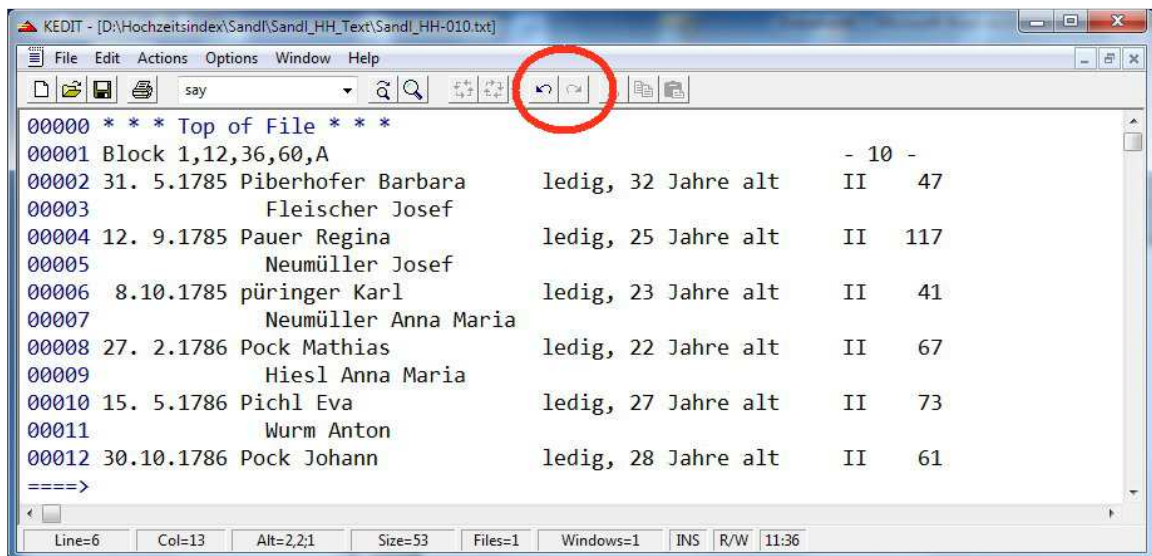


Abb. 27: Rückgängig machen von Änderungen in KEDIT

Auch im KEDIT können Textänderungen, die manuell oder über ein Prüfprogramm gemacht wurden, wieder rückgängig gemacht werden.

9.6.10 Ungeklärte Probleme, Kommentar und Fehlerprotokoll

Durch die Prüfungen werden viele Fehler oder zumindest Ungereimtheiten entdeckt. Die meisten lassen sich sofort klären und beheben, manche lassen sich durch Prüfung der Gegeneintragung klären und manche nur durch Nachschau in der Primärquelle, also im Matrikelbuch oder auch gar nicht.

Die Gegeneintragung kann man in den Fällen leicht finden, die alphabetisch vor dem aktuellen Indexbuchstaben liegen, also bereits als Textdatei gespeichert sind. Platziert man den Cursor in der ersten Zeile der Indexeintragung und betätigt die F9-Taste, versucht das Programm, die Gegenseite zu finden und zeigt die gefundene Seite neben der Ausgangsseite an. Nun kann man sofort erkennen, ob z.B. das Datum, die Namen oder die tomus pagina Informationen übereinstimmen oder ob bei einer der beiden Indexeintragungen ein offensichtlicher Fehler vorliegt. Nach Korrektur ist das Fenster der gefundenen Gegeneintragung zu schließen.

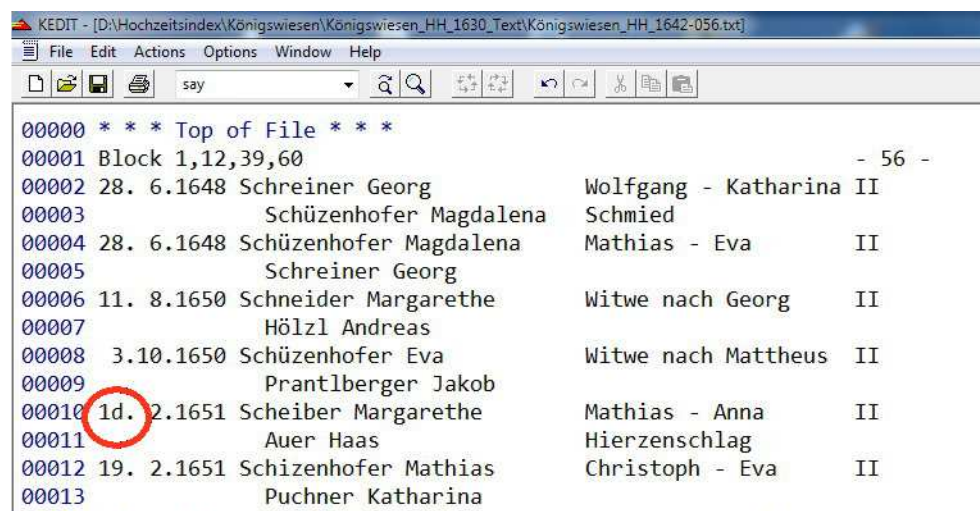


Abb. 28: Dieses falsche Tagesdatum lässt sich durch die Gegeneintragung korrigieren

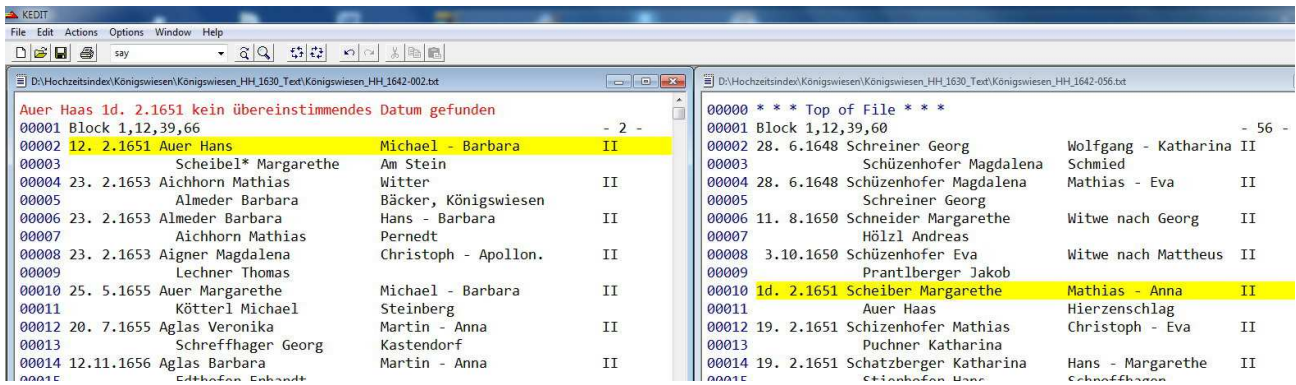


Abb. 29: links ist die Gegeneintragung mit einem Hinweis auf nicht übereinstimmendes Datum

Obwohl das Datum wegen des Tippfehlers nicht exakt übereinstimmen kann, wird die Gegeneintragung gefunden, gelb markiert und neben dem Ausgangsfenster angezeigt. Hier kann also sehr schnell geklärt werden, dass es sich um den Tag 12 handelt.

Meistens ist das Auffinden eines Fehlers aufwändiger, sei es, weil die gesuchte Gegeneintragung noch nicht als Textdatei vorliegt, sei es, weil beide Eintragungen identisch sind und daher keinen Hinweis auf einen Fehler liefern oder weil die Nachschau im Kirchenbuch erforderlich ist. Da die Kirchenbücher übers Internet online verfügbar sind, wäre das zwar sofort möglich, ist aber relativ mühsam und zeitaufwändig. Ich ziehe es in diesen Fällen daher vor, mit einem Kommentar die Unklarheit zu dokumentieren. Der Kommentar ist in die Zeile zu schreiben, in der das Problem auftritt und zwar hinter der *tomus pagina* Information. Die Position ist durch eine senkrechte Linie gekennzeichnet, die vom prüfenden Programm (Funktion F7 oder F8) erzeugt wurde.

Erst nach vollständiger Bearbeitung eines Heider-Buches arbeite ich die gesammelten Kommentare sequentiell ab. Als erstes prüfe ich jene Eintragungen, die durch Aufrufen der Gegeneintragung leicht zu korrigieren sind. Geklärte Fehler vermerke ich als Kommentar, damit bei der späteren manuellen Kontrolle klar ist, dass hier ein Wert korrigiert wurde. Für die übrig gebliebenen Eintragungen lege ich eine Datei mit der Nummer 000 an, in die ich diese Zeilen kopiere. Bei Datumsfehlern kopiere ich auch die vorhergehende und die nachfolgende Indexeintragung mit, denn der Datumsfehler liegt nicht notwendiger Weise in der Eintragung, in der der Fehler entdeckt wurde. Jeder kopierten Eintragung stelle ich die Seitennummer voran, damit ich später wieder direkt zur Textdatei finde (siehe *Abb. 30: Fehlerprotokoll eines Hochzeitsbuches auf Seite 50*).

Die Eintragungen in dieser Datei speichere ich nach aufsteigendem Heiratsdatum. Dadurch kann ich dann beim Abarbeiten der fehlerhaften bzw. unklaren Fälle die Kirchenmatrikeln, die ich übers Internet aufrufe, sequentiell abarbeiten. Das ist erheblich schneller, als wenn man bei jedem auftretenden Fall im Kirchenbuch vor- und zurückblättern oder das Kirchenbuch immer wieder wechseln muss. Bei umfangreicheren Fehlerprotokollen kläre ich die Fälle im OÖLA, weil dort das Aufsuchen der Kirchenbücher bzw. der Seiten viel schneller geht als übers Internet.

9.6.11 Anhang

Am Ende eines Heider-Buches gibt es meist eine nicht nummerierte Buchseite mit dem Text *Anhang* und danach ein oder mehrere Seiten mit Eintragungen, die in den Matrikelbüchern fehlerhaft, unleserlich oder unvollständig sind und die Heider nicht richtig zuordnen konnte. Bei deren Überprüfung sind fast alle Zeilen markiert, weil hier die Prüfung des Anfangsbuchstabens einen Fehler erzeugt und diese Prüfung im Grunde unsinnig ist. Diese Fehlermeldungen sind daher zu ignorieren. Mit Aufruf des Programmes *xí* (exclude initial) wird die Prüfung des Anfangsbuchstabens unterdrückt und damit die Fehlerliste übersichtlicher.

```

KEDIT - [D:\Hochzeitsprojekt\Heider Projekt\doku\Sandl_HH-000.txt]
File Edit Actions Options Window Help
say
00000 * * * Top of File * * *
00001 Pfarre Sandl bei Freistadt
00002
00003 ***** 095
00004 23. 6.1777 Wegerer/Wagner Franz      Martin - Katharina      I      55      Wagner?
00005                Hackermüller A. Maria    Königsau
00006
00007 ***** 091
00008 38. 7.1817 Ulrich Viktoria            ledig, 25 Jahre alt     II     28      Tag?
00009                Klein Martin
00010
00011 ***** 065
00012 Seite neu einscannen
00013
00014 ***** 062
00015 11. 2.1810 Oeblreiter Josef           ledig, 23 Jahre alt     II     16      Oberreiter?
00016                Bauer Katharina
00017
00018 ***** 052
00019 1. 7.1816 Ledermüller Anna Maria      Witwe, 45 Jahre alt     II     78
00020                Riesenhuber Lorenz
00021 19. 8.1818 Lambart Therese            Witwe, 55 Jahre alt     II     107     1816?
00022                Rosenmarin Wenzl
00023 10. 2.1817 Lehner Anna Maria          ledig, 30 Jahre alt     II     16
00024                Etz Josef
00025
00026 ***** 051
00027 10. 9.1768 Liebel Biebl Maria         Peter - Eva             I      35      Biebl Bubl?
00028                Undasch Josef              Sandl
00029 24.10.1768 Liegler Katharina           Paul - Katharina        I      36
00030                Hackermüller Michael    Müller, böhm. Schanz
00031 4.12.1770 Langecker Josef             Mathias - Anna Maria    I      40
00032                Poleder Susanna          Glasmacher zu Ehren-
00033                                reichstal
00034 27. 5.1771 Liebel Biebl Elisabeth     Peter - Eva             I      41      Biebl Bubl?
00035                Schaubberger Johann    Sandl
00036
00037 ***** 039
=====
Line=0 Col=1 Alt=0,0,0 Size=58 Files=1 Windows=1 INS R/W 16:46

```

Abb. 30: Fehlerprotokoll eines Hochzeitsbuches

9.7 Arbeiten mit dem KEDIT Texteditor

KEDIT ist ein mächtiges Werkzeug zum Bearbeiten von Textdateien am PC. Das Bearbeiten kann manuell durch direktes Verändern der Texte, kopieren, einfügen, löschen von Zeilen und Blöcken mit Editor-Befehlen erfolgt oder mit selbst entwickelten Programmen, die mit einer eigenen Makrosprache erstellt wurden. Sowohl das direkte Arbeiten mit dem Editor als auch das Erstellen von Programmen mit der Makrosprache erfordert das Einarbeiten in die Materie. An dieser Stelle sollen nur einige Hinweise für das Arbeiten mit KEDIT gegeben werden. Für das Erlernen der Editor-Funktionen und der Programmiersprache stehen ein Benutzer- und ein Referenz-Manual in Englischer Sprache zur Verfügung.

Neben der direkten Bearbeitung von Texten können mittels Befehlen umfangreichere Manipulationen vorgenommen werden. Diese Befehle werden im Befehlsfeld, in der Regel am unteren Ende des KEDIT-Fensters, eingegeben. Einige dieser Befehle können durch Drücken der F2-Taste als Icons am unteren Fensterrand eingeblendet werden (toolbar). Nach Auswahl des Textteiles - auch blockweise möglich – kann dieser Textteil über diese Icons manipuliert werden, sodass man die dafür notwendigen KEDIT-Befehl gar nicht kennen muss. Durch nochmaliges Drücken der F2-Taste wird die toolbar ausgeschaltet und die Zeile steht damit wieder zur Textdarstellung zur Verfügung. Das Beispiel unten zeigt einen Textblock, der um eine Stelle nach rechts verschoben ist. Nach Auswahl des Textblockes und klicken des Symbols mit dem Pfeil nach links wird der ganze Block um eine Stelle nach links verschoben.

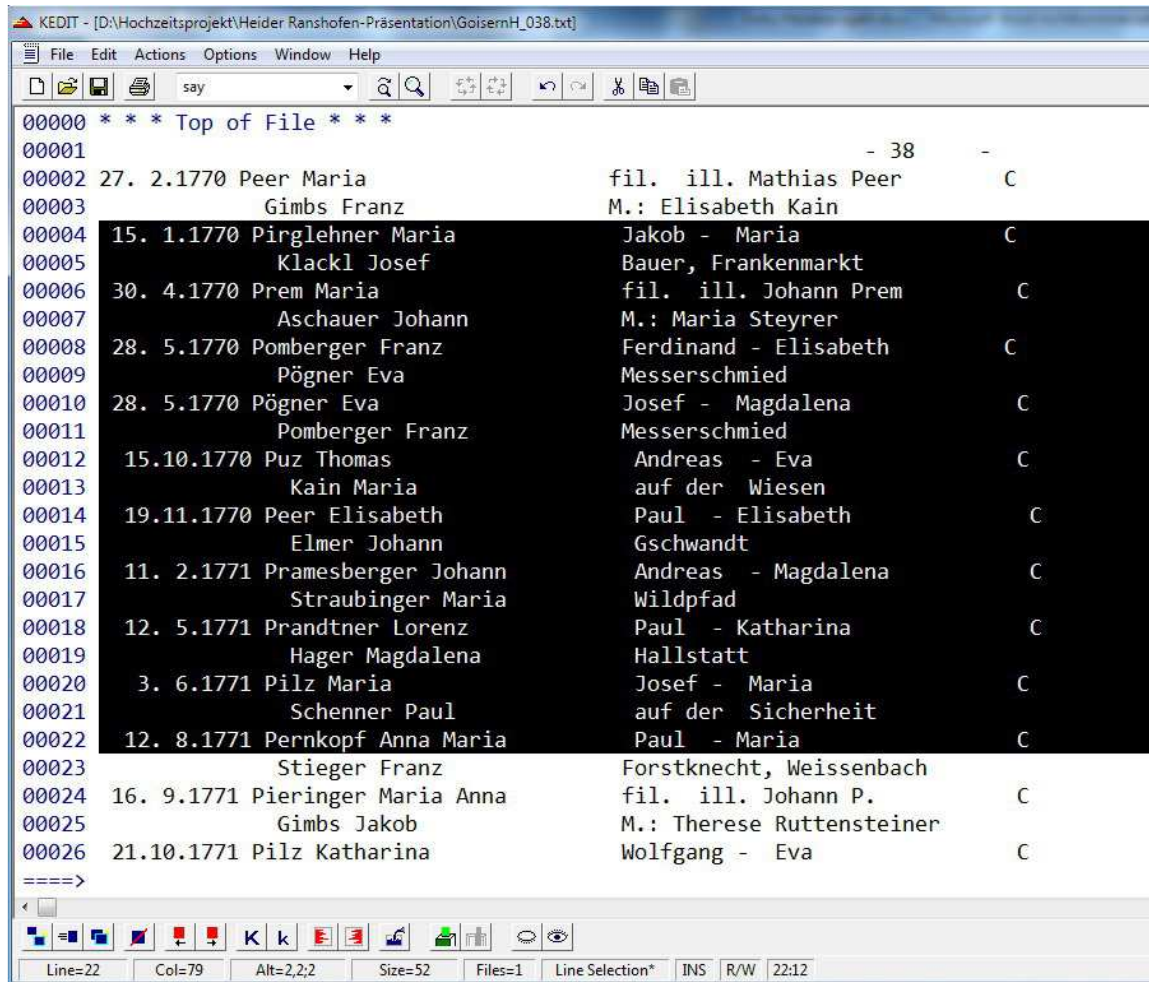


Abb. 31: KEDIT bottom toolbar

Sehr praktisch sind die Zeilenbefehle, die im Feld mit der 5-stelligen Zeilennummer eingegeben werden, z.B. *c* zum Kopieren eine Zeile oder *d3* zum Löschen von 3 Zeilen oder *mm* zum Verschieben eines ganzen Blockes. Durchgeführt wird die Aktion durch Drücken der F12-Taste.

10 Übergabe der Daten für die weitere Bearbeitung

Zu Beginn des Projektes erfolgte der Datenaustausch über Stick oder CD. Durch die seither dramatisch gestiegenen Kapazitäten des www erfolgt der Datenaustausch heute über das Netzwerk. Bei Dropbox habe ich im Laufe der Zeit einen kostenlosen Speicher von 8,75 GB angesammelt. In diesem Speicher habe ich den Ordner Heider Projekt angelegt, auf den alle Projektmitarbeiter Zugriff haben. Bei Zuteilung eines Heider-Buches zur manuellen Prüfung an einen externen Mitarbeiter bekommt dieser einen Link zu den Scandateien des zu prüfenden Buches. Nach Abschluss der Prüfung und Einspeisung der Daten in die FA-Datenbank werden die nicht mehr benötigten Verzeichnisse mit ihren Dateien aus der Dropbox gelöscht.

11 Abbildungsverzeichnis

Abb. 1: Abdeckungsbereich des Heider-Index.....	11
Abb. 2: Sepp Asanger mit Dr. Gerhart Marckhgott, Leiter des OÖLA bei der Vertragsunterzeichnung	12
Abb. 3: Heider-Indexbücher der Pfarren Liebenau bis Niederkappel	13
Abb. 4: geöffnetes Heider-Buch.....	14
Abb. 5: Heiratspaare, Pomberger Franz und Pögner Eva kommen zwei Mal in umgedrehter Reihenfolge vor	15
Abb. 6: Buchscanner im OÖLA	19
Abb. 7: Dateinamen der Scandateien	21
Abb. 8: Start-Option	24
Abb. 9: Optionen für Speichern.....	25
Abb. 10: Anlegen einer Benutzerdefinierten Sprache mit Benutzerwörterbuch	26
Abb. 11: Auswählen der Dokumentsprache und Anzeigen des Benutzerwörterbuches.....	27
Abb. 12: Exportieren und importieren des Benutzerwörterbuches.....	28
Abb. 13: Dokument-Fenster	29
Abb. 14: schiefer Scan.....	30
Abb. 15: Erkennungsrahmen gerade, aber rechts zu knapp.....	31
Abb. 16: Tag außerhalb des Erkennungsrahmens	32
Abb. 17: händische Korrekturen im Heider-Index	35
Abb. 18: OCR Textänderungen rückgängig machen.....	36
Abb. 19: Textdatei speichern.....	36
Abb. 20: Textdatei speichern und Texteditor aufrufen.....	37
Abb. 21: Namen- und Elternblock laufen ineinander.....	40
Abb. 22: Namen- und Elternblock getrennt.....	41
Abb. 23: F7-Funktion versus F8-Funktion	43
Abb. 24: Prüfung tomus - pagina auf nicht absteigende Nummern.....	45
Abb. 25: Buch-Scan mit pagina teilweise unterhalb von tomus.....	46
Abb. 26: aufbereitete Textseite.....	47
Abb. 27: Rückgängig machen von Änderungen in KEDIT.....	48
Abb. 28: Dieses falsche Tagesdatum lässt sich durch die Gegeneintragung korrigieren	48
Abb. 29: links ist die Gegeneintragung mit einem Hinweis auf nicht übereinstimmendes Datum	49
Abb. 30: Fehlerprotokoll eines Hochzeitsbuches	50
Abb. 31: KEDIT bottom toolbar.....	51