Kalkulierbare Dubletten ??

Calculating bibliographic Duplicates
Les doubles bibliographiques calculables

Harald Jele

Im Rahmen dieses abgeschlossenen Projekts konnte gezeigt werden, dass unter bestimmten Bedingungen beim Laden von Titeldaten in Online-Kataloge die üblichen Methoden der Dublettenerkennung durch statistische Abschätzungen ersetzt werden können.
Dabei wird aufgrund von Sicherheits- und Genauigkeitsannahmen die Größe einer Stichprobe errechnet, diese aus der Gesamtmenge nach einem Zufallsverfahren gewählt und händisch (intellektuell) geprüft. Das Ergebnis dieser Prüfung wird in weiterer Folge als ein gültiges für die Gesamtmenge angesehen.

In this project we show that under certain conditions the checking for duplicates while loading bibliographic data into an online catalog can be replaced by a statistical estimation. Due to the selected significance level (of 90%) and tolerance bounds (10%) the minimum of the needed sample size is calculated, chosen from the population at random and verified by hand.
The result of this examination is regarded in further consequence as a valid result for the total quantity.

Dans le cadre de ce projet qui a été réalisé, on a pu constater qu'à certaines conditions lors du chargement des dates descriptives bibliographiques dans le catalogue en-ligne, les méthodes usitées permettant d'identifier les descriptions doubles pourraient être remplacées par un système d'évaluation statistique.
Lors de ce procédé on calcule la taille d'un échantillon pris au hasard en tenant compte du niveau de signification et des limites de tolérance, cet échantillon est choisi parmi l'ensemble des données par une méthode au hasard et contrôlé à la main (intellectuellement). Le résultat de cet examen sera valable pour la totalité des dates par la suite.

1 Einleitung – Rahmenbedingungen

Vor einigen Jahren befanden wir uns in der allseits bekannten Situation, dass einer unserer Mitarbeiter in ein erweitertes Tätigkeitsfeld mit eingebunden werden konnte, dessen spezifische Qualifikationen besonders in einem Gebiet ausgeprägt waren, die uns mit seiner absehbaren Pensionierung verloren gehen sollten.
Da es sich bei bei den angesprochenen Qualifikationen – wie die Projektbeschreibung noch zeigt – wesentlich um Fertigkeiten[1] handelte, die auch durch effizientes Wissensmanagement innerhalb der Einrichtung nicht „konservierbar" waren, wurde die Idee geboren, diese nicht mehr alltäglichen Fähigkeiten für einen Teil der ihm im Betrieb „verbleibenden Zeit" zu nutzen.

Eine Möglichkeit, diese Qualifikationen für den weiteren Projektverlauf zu nutzen, war die Einbindung des Mitarbeiters in die Überarbeitung eines handgeschriebenen Bestandsverzeichnisses der Bibliothek. Konkret handelte es sich dabei um jenes Inventar, das die Signaturen `1-26177` ausweist.
Für die Art der Überarbeitung wurde vereinbart, dass die Ergebnisse des Projekts u.a. die Herstellung einer elektronischen Form des Ausgangsmaterials darstellen sollen; eine andere – wie z.B. eine rein maschinenschriftliche – Form wurde ausgeschlossen.[2]
Ein Teilziel dieses Projekts lag somit in der Retrokatalo-

[1] wie das flüssige Lesen der Deutschen Kurrentschrift oder der selbstverständliche Umgang mit Titelangaben in altgriechischer und lateinischer Sprache

[2] es sollten in diesem Sinn aus den Einträgen des Bandkatalogs keine konventionellen Katalogkarten hergestellt werden. Eine solche Vorgehensweise wäre wohl zudem nicht unbedingt zeitgemäß

gisierung eines spezifischen Bestandsteils in den Online-Katalog.
Die in dem betreffenden Inventar verzeichneten Druckwerke sind vielfach deutsche Übersetzungen griechischer und lateinischer Klassiker überwiegend des 18. und 19. Jhdts. sowie jene Werke in deutscher Sprache, die wohl zu den meistverlegten aus dieser Zeit zählen. Die neuesten Ausgaben darunter datieren jedenfalls in die Zeit vor 1925.

Zu bedenken waren dabei allerdings auch eine Reihe von „Wenn und Aber“, die in Summe letztlich überwiegend aus Zeitmangel verhinderten, dass das Projekt vollständig durchgeführt oder gar abgeschlossen werden konnte:

- Bildschirmarbeit, die evtl. durch diesen Kollegen geleistet werden sollte, konnte nicht in dessen Arbeitsweise eingeplant sein, denn eine spezifische Sehbehinderung und die daraus resultierte Befreiung von Arbeiten am Computer ließ eine solche nicht zu
- unterstützende Methoden der modernen Katalogisierung[3], die durch Online-Systeme typischerweise geleistet werden, waren dementsprechend nicht einsetzbar und konnten nur außerhalb seines Einsatzgebietes (nachträglich) verwendet werden
- eine möglichst kurze Einarbeitungszeit in die Erfassung von bibliographischen Informationen unter Berücksichtigung des Regelwerkes „RAK-WB“ auf der Basis von MAB2 musste garantiert sein – anderenfalls rechtfertigte die erbrachte Leistung keinesfalls den Umfang des Ergebnisses

2 Lösungsansatz

Die vorgegebenen Rahmenbedingungen machten die Entwicklung und den Einsatz sehr unkonventioneller Methoden notwendig.

Diese fußten letztlich alle in der Entscheidung, dass der Bearbeiter[4] in „möglichst gewohnter Form“ sämtliche notwendigen Informationen „auf Papier“ erfasst und diese an eine weitere Person übergibt, deren Wissen um die weitere Verarbeitung im Idealfall sehr gering sein können sollte.[5]
Die Erfassung der (bibliographischen und bestandsrelevanten) Daten geschah zwar in sehr konventioneller Art und Weise; trotzdem wurde dabei darauf geachtet, dass diese zumindest in einer spezifischen, strukturierten Form durchgeführt wird, die nachfolgend möglichst fehlerfrei maschineninterpretierbar ist.

Anschließend wurden die so erstellten Blätter durch weitere Verarbeitung mit Methoden der OCR (Optical Character Recognition) in computerlesbare Textfiles umgewandelt.
Diese wiederum wurden durch entsprechende Programme geprüft und zu passenden Datensatzstrukturen aufbereitet, sodass sie in den Verbundkatalog und anschließend durch Replikation in den lokalen Bibliothekskatalog online – und im Arbeitsgang möglichst unterbrechungsfrei – geladen werden konnten.

Beim Laden dieser Datensätze wurde zur Minimierung des Aufwandes bewusst darauf verzichtet, eine maschinell gesteuerte oder gar intellektuell durchgeführte Dublettenkontrolle anzuwenden.
Im Vorfeld des Projekts sowie im laufenden Betrieb wurde jedoch durch Bildung von Stichproben überprüft, ob eine Abschätzung der Menge an zu erwartenden Dubletten möglich ist.

Die hier zu bewältigende Aufgabenstellung war letztlich zu zeigen, ob durch ein relativ einfach zu handhabendes statistisches Verfahren – unter den hier angegebenen Rahmenbedingungen – nützliche Aussagen über den zu erwartenden Fehler getätigt werden können, wenn angestrebt wird, eine bestimmte Anzahl von Titeldubletten aus einer vorgegebenen Menge an zu verarbeitenden Datensätzen nicht zu überschreiten.
In *Abb.1* ist der vollständige Arbeitsablauf schematisch dargestellt.

3 Die Stichprobe

Die von uns gewählte Vorgehensweise zielte im Wesentlichen ja darauf ab, Datensätze auf der Basis „strukturierter Texte“ aufzubereiten und diese in weiterer Folge in das Bibliothekssystem zu laden, ohne zu überprüfen, ob diese bereits im System vorhanden sind.

Wenn also ein zu ladender Titeldatensatz bereits vor dem Laden im System vorhanden war, so hätten wir einen doppelten Eintrag zu ein und demselben Titel erzeugt. Unter der Prämisse, dass solche Dubletten zu vermeiden sind bzw. keinesfalls bewusst erzeugt werden

[3] dies meint vor allem eine „formale Prüfung“, der Einsatz von elektronischen „Schablonen“ („Templates“), automationsunterstützte Abbildung von Titel-Hierarchien, die sich bei Einsatz von RAK-WB ergeben und entsprechend Reihen- oder Bandangaben mit den zugehörigen Stücktiteln durch „Verlinkung“ wiedergeben etc.

[4] der hier ausschließlich in seiner Rolle als Experte zur regelkonformen Informationserschließung gesehen wird – und dem unter den in den Rahmenbedingungen genannten Prämissen – ausschließlich bereits vertraute Arbeitsweisen geboten werden müssen

[5] in unserem Fall kam dafür eine kurz angelernte Person zum Einsatz, deren Anlernzeit im Bereich von wenigen Stunden lag

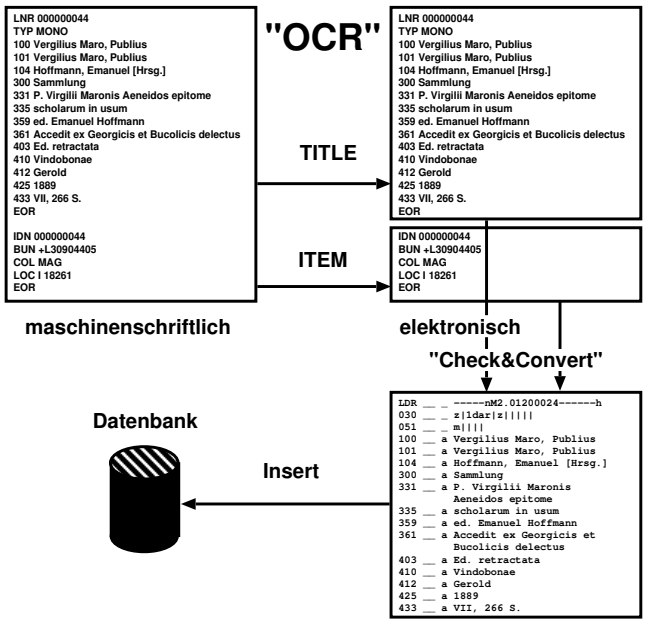

Abbildung 1: Schematische Darstellung des Arbeitsablaufs

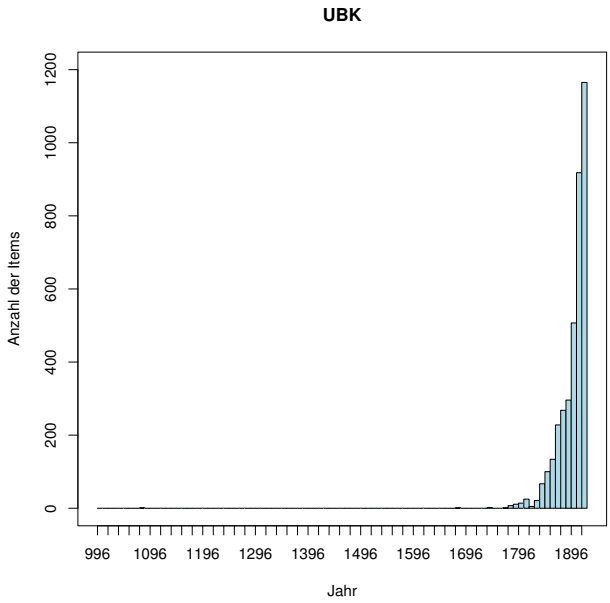

Abbildung 2: Lokalsystem der Universität Klagenfurt: Verteilungsstruktur der zählbaren Titel pro Jahr im Bereich `996-1925`. Zählung vom `Nov. 2000`

dürfen, galt es, im Vorfeld der Titelerfassung zu entscheiden, ob eine „eher große“ oder eine „eher kleine“ Menge an bereits vorhandenen Titeln zu erwarten ist.[6]

Diese Abschätzung wurde im hier beschriebenen Projekt zudem wesentlich davon beeinflusst, dass Titel nicht direkt in das lokale Bibliothekssystem geladen, sondern dass diese in einem ersten Schritt – entsprechend den üblichen Arbeitsabläufen des Österreichischen Bibliothekenverbundes – in das Verbundsystem eingebracht wurden.
Einmal im Verbundsystem gespeichert, wurden diese durch einen Kopiervorgang (Replikation) in das lokale System übernommen.

Für die Bestimmung der zu erwartenden Titeldubletten war daher notwendig, beide Systeme zu betrachten.

Zur Sichtung der Bestandszahlen wurde zu Beginn die Anzahl der bereits in beide Systeme eingebrachten Titel der Erscheinungsjahre `996-1925` ermittelt.[7]
In Schritten von `5` Jahren wurde die Anzahl der vorhandenen Titel maschinell gezählt und die daraus resultierende Verteilungskurve graphisch aufbereitet.[8]

Für das letztlich gewählte statistische Verfahren ist die Gesamtmenge der gezählten Titel nicht wesentlich.
Der Vergleich zwischen den *Abb.2* und *3* zeigt, dass für den gesamten, betrachteten Jahresbereich die Menge der auffindbaren Titel pro Jahreszahl im Verbundsystem wesentlich größer ist als im Lokalsystem der Universität Klagenfurt. Wesentlich für die Bestimmung des Stichprobenumfangs (bei ausgezählten Werten) ist für unser Verfahren jedoch, dass die beiden Systeme eine sehr ähnliche (eigentlich: die gleiche) Verteilungsstruktur aufweisen. Vor allem der für unser Vorhaben wesentliche Betrachtungszeitraum von `1796-1925` zeigt bei genauerer Betrachtung passende Verteilungswerte.[9]

Aus diesem Grund kann der Stichprobenumfang für beide Systeme gleich gewählt werden – und es ist nicht weiter nötig, die Stichproben entsprechend einer speziellen Strategie zu ziehen.

Für die Berechnung bzw. Schätzung des Stichprobenumfangs wählten wir jenen Ansatz, der z.B. in Sachs (1992, S.444) referiert wird. Dabei wird im Wesentlichen davon ausgegangen, dass bei gezählten Werten der Mindestumfang einer Stichprobe nicht von der zu untersuchenden Grundgesamtheit abhängt.[10]
Vielmehr ist davon auszugehen, dass Werte für die „Sicherheit“ der zu erwartenden Ergebnisse (bei uns: `90%`)

[6] in der Ermittlung der zu erwartenden Mengen lag somit auch die weitere Entscheidung, ob ein solcher Weg gangbar – also von uns im Verbund vertretbar – ist

[7] an dieser Stelle ist darauf hinzuweisen, dass die ermittelte Anzahl der eingebrachten Titel, die auf Erscheinungsjahre bis `1496` hinweisen, in vielen Fällen auf Tippfehler zurückzuführen ist. Dies stört in weiterer Folge jedoch weder die Ergebnisse noch die gewählte Vorgehensweise

[8] zu beachten ist, dass die hier ausgewertete Zählung vom `Nov. 2000` stammt. Auf dieser Zählung beruhen auch sämtliche daraus abgeleiteten oder in Folge berechneten Werte

[9] da in diesem Text nicht alle Abbildungen wiedergegeben werden können, sind diese unter folgendem Link vollständig einzusehen
`http://www.uni-klu.ac.at/ub/ub-edv/projekte/bock/graphiken/alle_abbildungen.pdf`

[10] dieser Ansatz findet sich u.a. auch in Dürr & Mayer (1987, S.136-137)

sowie Werte für die Genauigkeit der durchzuführenden Messung (bei uns: 10±5%) angenommen werden müssen.

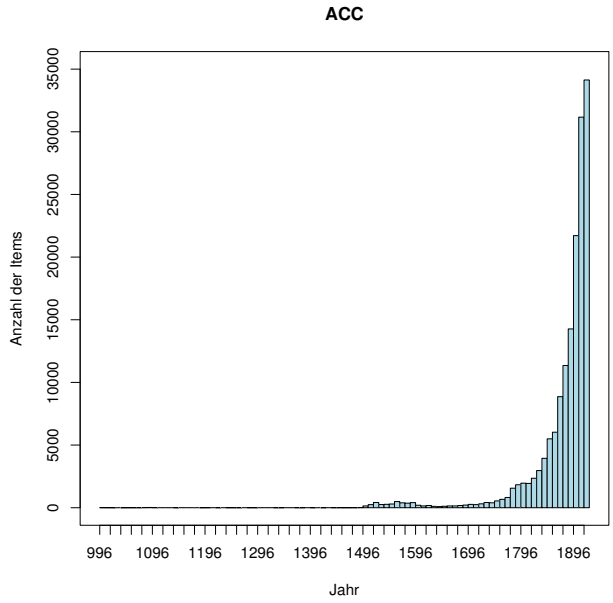

Abbildung 3: Bibliothekssystem des Österreichischen Bibliothekenverbundes: Verteilungsstruktur der zählbaren Titel pro Jahr im Bereich 996-1925. Zählung vom Nov. 2000

Entsprechend dieser Vorgaben errechnet sich der Stichprobenumfang wie folgt:

1. aus der Sicherheit von 90% ergibt sich $\Phi(z) = (\gamma + 1)/2 = 0.95 \rightarrow z = 1.645$
2. bei Einhaltung einer Genauigkeit von $p = h \pm \varepsilon\% = 10 \pm 5\%$
3. diese Werte sind anzuwenden auf

$$n = \frac{z^2 \cdot h \cdot (1-h)}{\varepsilon^2} = \frac{1.645^2 \cdot 0.1 \cdot 0.9}{0.05^2} = 97.39 \approx 100$$

Der nach oben abgeschätzte Wert sagt aus, dass aus den zu überprüfenden Signaturen 1-26177 100 ausgewählt werden müssen.

Durch ein Zufallsverfahren wurden diese (vgl. *Abb.4*) ermittelt – und die zugehörigen Titel anschließend sowohl im Verbundkatalog als auch im lokalen Katalog auf deren Vorhandensein geprüft.

Anhand der Stichprobenliste zeigte sich, dass bei der Erstellung des händischen Inventars nicht alle Signaturen durchgehend verwendet wurden: einige fehlten, bzw. waren kleinere Lücken zu erkennen.
Aus diesem Grund wurde für den Fall entschieden, die jeweils nächsthöhere, tatsächlich verwendete Nummer zur Stichprobe heranzuziehen.

80	3649	8540	13031	20396
678	3814	8642	14361	20764
743	3821	9003	14538	20824
758	4015	9082	14628	20941
773	4046	9567	14954	21439
969	4495	9964	15330	21662
1003	4602	9966	15571	21809
1170	5344	10046	15854	22124
1304	5449	10382	15884	22150
1539	5501	10747	15921	22430
1580	5518	10870	16007	23573
1694	5744	11774	16541	23722
1937	6151	11934	16693	23803
2350	6377	12441	17572	24151
2492	6716	12473	18390	24317
2776	6961	12500	18457	24440
2982	7154	12530	18503	24798
3201	8131	12532	18539	24957
3373	8298	12564	19177	24967
3593	8505	12940	19680	25898

Abbildung 4: Die durch ein Zufallsverfahren ausgewählten Signaturen im errechneten Umfang von 100 aus der Menge 1-26177

Die Überprüfung zeigte, dass von den 100 überprüften Titeln sieben (7) im Verbundkatalog und keiner im lokalen Katalog vorhanden waren. Das heißt, dass beim Laden dieser 100, zufällig ausgewählten Katalogdatensätze (in den Verbundkatalog) 7% Dubletten erzeugt würden.

Zur manuellen Dubletten-Prüfung der Titel sind jedoch noch folgende Beobachtungen anzumerken:
Aus den bibliographischen Beschreibungen, die im Verbundkatalog in sieben Fällen aufgefunden wurden, konnte der – obwohl sehr geschulte – Bearbeiter nicht immer eindeutig feststellen, ob der im Katalog nachgewiesene Titel seinem Werk vollständig entspricht oder (bloß) ein sehr ähnliches Werk anführt. Das hat in einem Fall dazu geführt, dass wir keine Entscheidung herbeiführen konnten, ohne das entsprechende Werk aus der nachweisenden Bibliothek zu bestellen. In zwei weiteren Fällen konnten wir uns darauf verständigen, im Zweifelsfall den bibliographischen Nachweis als Dublette anzusehen.
Die Gründe dafür lagen jedenfalls in den allermeisten Fällen in der für diese Zeit typischen Gestaltung der Werk-Titelseiten: Die für uns wesentlichen bibliographischen Informationen zur Dubletten-Kontrolle fehlten zuweilen oder waren nicht immer eindeutig interpretierbar.

Aus dieser Erkenntnis ergab sich für uns überdies die Haltung, dass wahrscheinlich viele der von uns produzierten Dubletten sich als solche nur in jenen Fällen zeigen, in denen der bibliographische Datensatz mit dem entsprechenden Werk direkt verglichen werden kann.[11]

[11] also vor allem nur unter jenen Umständen, die von Bibliothekaren liebevoll als „Autopsie“ bezeichnet werden :-) Da eine solche Titel-Autopsie durch Benutzer und Bearbeiter im üblichen Fall nicht (systematisch) anzunehmen ist, ist

4 Durchführung

Die „Datenerfassung“ geschah in diesem Fall – wie bereits angedeutet – mit einer (noch nicht ausgemusterten) Kugelkopf-Speicher-Schreibmaschine auf normalem, weißen Papier im Format A4.

```
LNR   000000044
TYP   MONO
100   Vergilius Maro, Publius
101   Virgilius Maro, Publius
104   Hoffmann, Emanuel [Hrsg.]
300   Sammlung
331   P. Virgilii Maronis Aeneidos epitome
335   scholarum in usum
359   ed. Emanuel Hoffmann
361   Accedit ex Georgicis et Bucolicis delectus
403   Ed. retractata
410   Vindobonae
412   Gerold
425   1889
433   VII, 266 S.
EOR

IDN   000000044
BUN   +L30904405
COL   MAG
LOC   I 18261
EOR
```

Abbildung 5: Monographische Titelaufnahme mit Schreibmaschine

Auf diesem wurden im oberen Bereich die bibliographischen Informationen – ausgerichtet in zwei Spalten – geschrieben:

- die linke Spalte beinhaltete im Wesentlichen die Kennung der MAB2-Kategorien, -Indikatoren und -Teilfelder (dafür konnten im Verlauf des Projekts auf der Basis von Erfahrungswerten Vorlagen erstellt werden, die – ähnlich einer Bildschirm-Schablone – zeitsparend gefüllt wurden).
  Neben den üblichen Kategorien wurden zur Vereinfachung der weiteren Verarbeitung Kennungen vermerkt, die im Zuge der Umwandlung in Datensatzstrukturen jedoch wieder entfernt oder einfach missachtet wurden.
  So wurde z.B. mit der Kennung *LNR* die Datensatznummer vor der Textumwandlung mittels OCR vermerkt, die in weiterer Folge (nach der erfolgreichen Verarbeitung) bedeutungslos wurde,
  die Kennung *TYP* (mit den zulässigen Werten „MONO“ für monographische Katalogisate und
  „HIER“ für hierarchisch gegliederte Titelaufnahmen) vermerkte den Datensatztyp, anhand dessen die Vollständigkeit und Korrektheit der vorhandenen Kategorien geprüft wurde
  sowie die Kennung *EOR*, die das Ende einer Datenaufnahme anzeigte (vgl.*Abb. 5 und 6*)
- in der rechten Spalte wurden die zugehörigen Kataloginformationen (=die bibliographischen Angaben) in der auch für einen Online-Katalogisierer üblichen Schreibweise (Notation) vermerkt

```
LNR   000000048
TYP   HIER
100   Homerus
104   Pauly, Franz [Hrsg.]
105   Pauly, Franciscus
108   Wotke, Karl [Bearb.]
109   Wotke, Carolus
304   Odyssea
331   Homeri Odysseae epitome
335   in usum scholarum
359   ed. Franciscus Pauly
410   Lipsiae u.a.
412   Freytag u.a.
451   Bibliotheca scriptorum graecorum et
451   romanum : A, Scriptores graeci
501   Nebent.: Odysseia. - Überw. in griech. Schr.
530   2. Odysseae lib. XIII - XXIV. - Ed. 4. corr.
530   quam cur. Carolus Wotke, 1887. - XI, 165 S.
EOR

IDN   000000048
BUN   +L30904806
COL   MAG
LOC   I 18256,A,4,2,2
EOR
```

Abbildung 6: Katalogisat mit Bandangabe auf Papier

Im unteren Blattbereich wurden die spezifischen Exemplardaten vermerkt. Deren Kennungen sind in ähnlicher Weise wie jene der bibliographischen Einträge zu interpretieren; ihre spezifische Bezeichnungsform wurde von uns (willkürlich) frei gewählt:

- die Kategorienkennung *IDN* beinhaltete den (laufenden) Zählerwert des zugehörigen Titels und bildete somit die Referenz zwischen den einzelnen Exemplaren sowie dem entsprechenden Titel
- mit *BUN* wurde der im Buch durch den Bearbeiter eingeklebte Barcode wiedergegeben
- der Inhalt der Kategorie *COL* repräsentierte den Standort des Werkes
- die Kennung *LOC* wies die Signatur des Werkes aus

Die anschließende Verarbeitung erfolgte programmgesteuert und – zumindest im Fall, dass keine Fehler in den zu verarbeitenden Daten zu beanstanden waren – darauffolgend unterbrechungsfrei.
Das heißt, dass für die dafür eingeschulte Bearbeiterin die Anwendung eines „Ein-Knopf-Verfahrens“[12] aus-

die Anzahl der durch dieses Verfahren subjektiv wahrnehmbaren Titel-Dubletten zudem als wesentlich geringer anzusehen als die Anzahl der tatsächlich produzierten und objektiv messbaren

[12] tatsächlich konnte ein solches „Ein-Knopf-Verfahren“ auch real – und nicht nur sinngemäß – umgesetzt werden: durch den am Scanner stirnseitig markant angebrachten Start-Knopf konnte zugleich der Programmstart für sämtliche, nachfolgende Programmschritte ausgelöst werden

reichte, um aus den maschinenschriftlichen Vorlagen fertige Datensätze im Bibliothekssystem anzulegen.

Jedes maschinengeschriebene Blatt wurde im ersten Programmschritt gescannt. Anschließend wurde mittels gängiger OCR-Software die daraus entstandene Bild-Datei in eine Text-Datei umgewandelt. Aufgrund der Homogenität der Vorlagen-Schrift konnten bereits nach zehn Scan-Durchgängen praktisch fehlerfreie Ergebnisse erzielt werden. Des weiteren waren im ganzen Projektverlauf keine Fehler durch die Bild-Textumwandlung feststellbar, sodass die übliche OCR-Fehlerrate in diesem Fall völlig vernachlässigt werden konnte.

Jede Textdatei wurde sofort weiter verarbeitet, d.h., die Erstellung der einzelnen Datensätze aus den gescannten Informationen wurde nicht stapelartig[13], sondern ad hoc verfolgt.

In einem weiteren Schritt wurden die Inhalte der aktuellen Textdatei nach formalen Gesichtspunkten geprüft. Auf der Grundlage der Programmiersprache Perl realisierten wir einen einfachen, sequentiell gesteuerten Parser. Mit diesem wurden – entsprechend unseren Vorgaben – sowohl die maschinengeschriebenen Feldangaben als auch die formale Korrektheit der Feldinhalte geprüft. Zudem mussten dabei jene definierten Abhängigkeiten zwischen einzelnen Kategorien ermittelt (und im Fehlerfall angezeigt) werden, die durch die kombinierte Anwendung von MAB2 und RAK-WB vorgegeben sind.[14]

Neben der Fehlerkontrolle wurden beim Parsing[15] der Texte erste Schritte zur Datenaufbereitung unternommen. Da mit moderner Hardware der Zeitaufwand für diese Programmschritte im Bereich von Milli-Sekunden nicht mehr wirklich messbar und außerdem subjektiv nicht erkennbar ist, wurde entschieden, sämtliche Schritte der weiteren Verarbeitung (die Ladevorgänge in das Bibliothekssystem ausgenommen) in jedem Fall durchzuführen; nämlich auch dann, wenn schon beim Parsing-

Record label	LDR	_	_	-----nM2.01200024------h
Enc. det. rec.	030	_	_	z|1dar|z|||||
Spec.det.n-ser	051	_	_	m||||
1. Person a.f.	100	_	a	Virgilius Maro, Publius
Cross ref. 100	101	_	a	Virgilius Maro, Publius
2. Person a.f.	104	_	a	Hoffmann, Emanuel [Hrsg.]
Collection	300	_	a	Sammlung
Tit.prop.orig.	331	_	a	P. Virgilii Maronis Aeneidos epitome
Furth.tit.inf.	335	_	a	scholarum in usum
Author	359	_	a	ed. Emanuel Hoffmann
Added works	361	_	a	Accedit ex Georgicis et Bucolicis delectus
Ed. stat. orig.	403	_	a	Ed. retractata
Publ.pl.1.publ.	410	_	a	Vindobonae
Name 1. publ.	412	_	a	Gerold
Years of publ.	425	_	a	1889
Pagination	433	_	a	VII, 266 S.

Abbildung 7: Beispiel eines generierten Datensatzes zum Laden einer monographischen Titelaufnahme

Record label	LDR	_	_	-----nM2.01200024------h
Enc. det. rec.	030	_	_	z|1dar|z|||||
Spec.det.n-ser	051	_	_	m||||
1. Person a.f.	100	_	a	Homerus
2. Person a.f.	104	_	a	Pauly, Franz [Hrsg.]
Cross ref. 104	105	_	a	Pauly, Franciscus
3. Person a.f.	108	_	a	Wotke, Karl [Bearb.]
Cross ref. 108	109	_	a	Wotke, Carolus
Uniform title	304	_	a	Odyssea
Tit.prop.orig.	331	_	a	Homeri Odysseae epitome
Furth.tit.inf.	335	_	a	in usum scholarum
Author	359	_	a	ed. Franciscus Pauly
Publ.pl.1.publ.	410	_	a	Lipsiae u.a.
Name 1. publ.	412	_	a	Freytag u.a.
1.Coll.tit.ori.	451	_	a	Bibliotheca scriptorum graecorum et romanum : A, Scriptores graeci
Footnotes	501	_	a	Nebent.: Odysseia. - Überw. in griech. Schr.
Tit. ref.	530	_	a	2. Odysseae lib. XIII - XXIV. - Ed. 4. corr. quam cur. Carolus Wotke, 1887. - XI, 165 S.

Abbildung 8: Beispiel eines generierten Datensatzes zum Laden einer Titelaufnahme mit Bandangabe

Prozess Fehler erkennbar waren.[16]
Die hier mitgeführte Aufbereitung beinhaltete Schritte wie z.B. die Eliminierung überflüssiger Leerzeichen oder die Ersetzung jener Sonderzeichen, die auf dem Kugelkopf der Schreibmaschine nicht vorhanden waren und deshalb durch Ersetzungszeichen angemerkt werden mussten.[17]

Jenen Textdateien, die aufgrund der Prüfungen soweit korrekt erschienen, dass sie in weiterer Folge ins Bibliothekssystem geladen werden konnten wurden anschließend jene „kodierten Angaben“ hinzugefügt, die Da-

[13] „stapelartig“ meint hier: in einem klassischen Batch-Verfahren

[14] erinnert sei an dieser Stelle überdies auf die Bedeutung der bereits genannten, den bibliographischen Beschreibungen und Bestandsangaben hinzugefügten Hilfskategorien MONO und HIER, aus denen bei der formalen Prüfung von Kategorienabhängigkeiten weiters darauf geschlossen werden konnte, welche Kategorien mindestens vohanden sein müssen. Die beschreibbaren Abhängigkeiten werden in diesem Text nicht weiter erläutert; sie sind jedoch aus den Programmquellen vollständig ermittelbar
vgl. `http://www.uni-klu.ac.at/groups/ub/ub-edv/projekte/bock/code/bock.pl.txt`

[15] dessen Aufgabe sich im Grunde ja darauf beschränkte, die maschinengeschriebene, blattorientierte, externe Textrepräsentation in eine Perl-interne Hash-orientierte Datenstruktur überzuführen, die in weiterer Folge besser programmtechnisch zu verwalten war

[16] dadurch konnten in vielen Fällen auch bessere (umfangreichere und aussagekräftigere) Fehlerprotokolle für den Bearbeiter erstellt werden

[17] in unserem Fall wurde z.B. das Zeichen „|“ als Ersetzungszeichen für ein „[“ ausgewählt bzw. ein Zeichenpaar, das aus der Kombination von „|“... und ...„|“ bestand wurde anschließend durch ein Zeichenpaar bestehend aus „[“... und ...„]“ ersetzt

tensätze entsprechend den Vorgaben von MAB2 in maschinenlesbaren Verfahren besser interpretierbar gestalten. Dazu gehören neben den Informationen zum Leader (`LDR`) u.a. die Angaben innerhalb der Kategorien `030` und `051`.[18]

Nachdem Titel- und Exemplarinformationen im vorhandenen Bibliothekssystem durch unterschiedliche Routinen zu laden sind, wurden diese in verschiedene Datensätze aufgespalten und entsprechend den günstigsten Ladevorgaben getrennt gespeichert.

Als Ladeformat kam das systemeigene Templateformat zur Anwendung, das im Client der Bibliothekssoftware auch dazu verwendet wird, „Schablonen" für die Katalogisierung vorzubereiten, mit denen in der Routinearbeit wiederkehrende Tätigkeiten abgekürzt werden.
Da diese dem Client direkt in seinem Programmverzeichnis zugänglich sind, wurde dieses Verzeichnis so eingerichtet, dass sämtliche Prüf- und Schreibroutinen dort Schreib- bzw. Lesezugriff haben.
Nach der Prüfung und Aufbereitung einer Textdatei in das entsprechende Format waren die einzelnen Datensätze bis zum Prozess des Ladens in diesem Verzeichnis abgelegt.[19]

Die eigentlichen Ladeskripts wurden mit einer einfachen Skriptsprache realisiert, die Zugang zu den wichtigsten Basisfunktionen des Betriebssystems bietet. In unserem Fall wurde dies mittels dem Programm „MacroExpress" realisiert.[20]

Die Reihenfolge der Speicherung erfolgte nach den im Verbundsystem des Österreichischen Bibliothekenverbundes zur Zeit üblichen Vorgaben:

1. Titelspeicherung im Verbund
2. Kopieren des Titels in das Lokalsystem und Speicherung dort
3. Verknüpfung eventuell vorhandener, lokaler Bestandsangaben mit den lokal gespeicherten Titeldatensätzen (dabei werden zudem jene Bestandsangaben in das Verbundsystem repliziert, die für eine konsistente, verbundweit gültige Bestandsanzeige – wie z.B. die Zeitschriftenbestände aus `Kat.200` – notwendig sind)
4. Verknüpfung der Exemplare mit den lokal gespeicherten Titeldatensätzen (dabei werden überdies rudimentäre Exemplarangaben in das Verbundsystem repliziert)

Der Einsatz der eben beschriebenen Methode zum Laden von Daten hatte des weiteren den Vorteil, dass dem Laden nicht das übliche Indizieren „extra" nachfolgen musste.
Für das System wurden Daten ja in einer Weise geladen, als wären besonders fleissige Bearbeiter am Werk – und entsprechend dem Verhalten, dass neu eingebrachte Daten über die genutzten Bearbeiter-Schnittstellen ohnehin vom System indiziert werden, musste auf eine gesonderte Indizierung nicht weiter Rücksicht genommen werden.

5 Zusammenfassung

Die Methode, mit statistischen Verfahren einen möglichen „Fehler" beim Datenladen (hier: das Produzieren von Titel-Dubletten) abzuschätzen, hat sich als brauchbar erwiesen.
Der alternative Ansatz, ein maschinelles oder intellektuell unterstütztes Prüfverfahren anzuwenden, ist im Gegensatz dazu als wesentlich teurer und zeitaufwändiger anzusehen – und (wie im Text angesprochen) auch maschinell sehr schwierig umzusetzen, ohne dabei eine allzu große Fehlerrate zu akzeptieren (vgl. dazu z.B. auch Jele (2001, S.65-66)).

Dies kann als ein wesentliches und letztlich auch praktikables Ergebnis des hier vorgestellten, eher unkonventionellen Projektansatzes gesehen werden, bei dem ein handgeschriebenes Bibliotheksinventar über den Umweg (auf Papier) neu erstellter, maschinengeschriebener Kataloginformationen in den Bestand des Online-Katalogs des Österreichischen Bibliothekenverbundes „geladen" wurden.
Der letztlich größte Zeitaufwand war für den Scan-Vorgang zu reservieren, da jede A4-Seite vollständig gescannt werden musste. D.h., die Anschaffung eines schnell arbeitenden Einzugscanners im Einsatz mit geringer Auflösung führt dazu, den Zeitaufwand bei der elektronischen Verarbeitung möglichst zu minimieren. Die weiteren Programmschritte wie das OCR, die Prüfung und Aufbereitung der Daten sowie das Laden in die Systeme stellen einen eher unwesentlichen Zeitfaktor da.

6 Literaturverzeichnis

6.1 Gedruckte Quellen

Dürr, Walter & Mayer, Horst (1987): Wahrscheinlichkeitsrechnung und schließende Statistik. 2., vollst.

[18] Beispiele dazu sind in *Abb.7* und *8* zu erkennen

[19] nach dem Laden wurden diese zwar nicht gelöscht, aber aus Gründen geringerer Fehleranfälligkeit in ein definiertes Zielverzeichnis verschoben

[20] vgl. dazu `http://www.macroexpress.com`
Eine Lösung mittels „WSH" (=Windows-Scripting-Host) oder „VBA" (=Visual Basic for Applications) kam in diesem Fall nicht in Frage, da der Client der Bibliothekssoftware zu beiden keine programmierbare Schnittstelle (API, Application Programmable Interface) bietet

durchges. und verb. Auflage. Hanser, München
Jele, Harald (2001): Informationstechnologien in Bibliotheken. Oldenbourg, München
Sachs, Lothar (1992): Angewandte Statistik. Anwendung statistischer Methoden. Siebente, völlig neu bearbeitete Auflage. Springer, Berlin

6.2 Online-Quellen

`http://www.macroexpress.com` : Programmpaket, das zum Scripting des Clients der Bibliothekssoftware verwendet wurde
`http://www.uni-klu.ac.at/ub/ub-edv/projekte/bock/code/bock.pl.txt` : der von uns eingesetzte Perl-Parser
`http://www.uni-klu.ac.at/ub/ub-edv/projekte/bock/graphiken/alle_abbildungen.pdf` : Verzeichnis der Abbildungen

1

Dr. Harald Jele ist Leiter der Abteilung EDV-Administration und -Entwicklung der Universitätsbibliothek Klagenfurt

Adresse:
Universität Klagenfurt
Universitätsstraße 65-67
9020 Klagenfurt, Österreich
Fax: 0043-463-2700-9599
E-Mail:harald.jele@uni-klu.ac.at