Konversion digitaler Findmittel

„Fruchtbar und weit umfassend ist das Gebiet der Geschichte“ (Schiller).

In diesem Aufsatz geht es um ein Software-Werkzeug, das als Hilfsmittel für die Konversion archivischer Findmittel eingesetzt wurde.

In der Zeit von Februar 2006 bis Dezember 2008 wurden im Generallandesarchiv Karlsruhe in einem Konversionsprojekt 191 Bestandsdateien mit insgesamt 126.131 Titelaufnahmen aus verschiedenen Quellsystemen in das aktuelle Zielsystem übernommen.

Findmittelkonversion - Übersicht

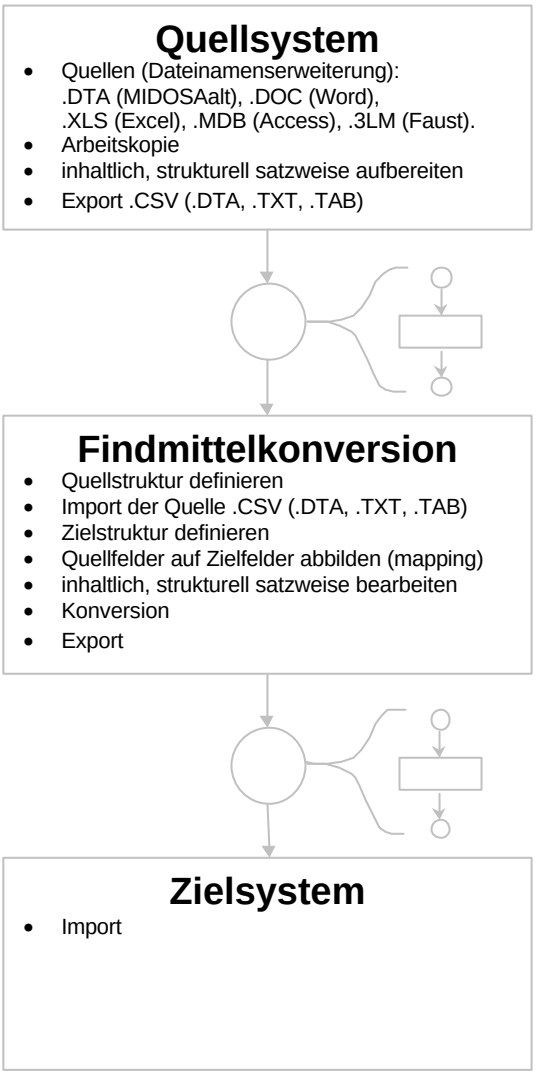
Sie sehen rechts in der Abbildung ungefähr den Verlauf einer Findmittelkonversion.

In der Praxis gibt es „Zwischenfälle“, die hier nicht besonders erwähnt sind.

Ein Großteil der Arbeit besteht in der Aufbereitung der Quelldaten.

Indem die Quelldaten in ein einheitliches Format gebracht werden, hat man einen besseren Zugang zur inhaltlichen Bearbeitung der Quelldaten, zum Beispiel für die Behandlung historischer Schreibweisen, zur Korrektur von Schreibfehlern.

Vor und nach dem Import in das Zielsystem sind wesentliche zusätzliche Arbeitsschritte erforderlich.



Relationen und Archive

Archive sind Aufbewahrungsorte für öffentliche **Dokumente**.

Der mengenmäßige Umfang aller Dokumente in einem Archiv liegt in Größenordnungen von wenigen Metern (Privatarchive) bis mehrere Kilometer (Stadt-, Landes- und Staatsarchive).

Zum einzelnen Dokument gelangt man daher in der Regel nur durch geeignete **Findmittel**, die kennzeichnende Inhalte der Dokumente in übersichtlicher und in knapper Form wiedergeben. Solche Findmittel werden beim Erschließen, meistens zeitnah bei Neuzugang von Dokumentenbeständen hergestellt.

Drei Verfahrensweisen zur Herstellung eines Findbuchs in Reihenfolge ihres historischen Erscheinens:

1. Von Hand direkt in ein Buch geschrieben.
2. Von Hand per Schreibmaschine auf einzelne Blätter getippt und schließlich als Buch gebunden.
3. Von Hand per Tastatur getippt, dabei automatisch auf maschinenlesbare Speicher übertragen, schließlich automatisch auf Papier gedruckt und als Buch gebunden und mit Mitteln der Datenkommunikation einem dezentralen Benutzerkreis zugänglich gemacht.

Traditionelles Findmittel zur lokalen Benutzung im **Repertoriensaal** eines Archivs ist auch heute, trotz Existenz von Hightech-Produkten, nach wie vor das Findbuch. Ein typisches Findbuch unserer Zeit besteht aus: Titelblatt, Inhaltsverzeichnis, Vorwort, Abkürzungsverzeichnis, **Titelaufnahmen**, Ortsindex, Personenindex, Sachindex, Konkordanzen.

Mit Verbreitung der PC's in den 80-er Jahren des 20-ten Jahrhunderts verschwinden Schreibmaschinen rasch vom Markt. Mit dem PC, der Hardware, wurden Softwareprodukte für die verschiedensten Zwecke angeboten, darunter auch früh Textverarbeitungsprogramme und Datenbanken.

Die Notwendigkeit einer **Findmittelkonversion** ergibt sich zwar prinzipiell aus einem technischen Wandel und dem Bedürfnis oder der Anforderung, die beste Technik zu nutzen, beispielsweise um Qualität und Kosten für die Pflege der Findmittel im Rahmen zu halten und um gesellschaftlichen Entwicklungen zu folgen und dadurch mitzugestalten: aber eben der technische Wandel ist vor allem erst ein Charakteristikum der neuen Technik, das heißt, erst mit Einführung der PC's und den Entwicklungen im Bereich der Software, die seitdem zur Herstellung der Findmittel eingesetzt werden.

Im Gegenzug dazu ist allerdings auch charakteristisch, daß bereits zu Zeiten, da Findbücher von Hand geschrieben wurden, also weit vor dem Aufkommen der PC-betriebenen relationalen Datenbanken, daß die **Titelaufnahmen**, die den Kern der Findbücher ausmachen, bereits eine relationale, tabellarische Struktur haben. Jeder Eintrag, ein n-Tupel, existiert darin in der Regel genau einmal.

Die Vorstellung einer Titelaufnahme als n-Tupel, also einem Vektor aus den n Elementen e_1 bis e_n, wird formal durch die Definitionsgleichung $T = (e_1, e_2, e_3,.., e_n)$ ausgedrückt und verbal durch die Aussage „eine Titelaufnahme T besteht aus den Elementen e_1 bis e_n". Die Bezeichnungen e_1 bis e_n sind hier Abkürzungen oder **Namen** für **Inhalte** jeweils unterschiedlicher **Bedeutungen**. Die Betonung liegt hier auf „unterschiedlicher Bedeutung".

Bedeutungen B und **Namen** A stehen formal in der Beziehung: B **ist ein** A und die Umkehrung A **ist ein** B ist zumindest im Kontext einer Titelaufnahme bzw. eines Dokumentenbestandes definitiv gültig!

Ein bestimmter Sachverhalt, eine bestimmte Bedeutung wird in einem bestimmten Kontext durch einen bestimmten Namen benannt: ein Name wird anstelle seiner Bedeutung benutzt.

Beispiel: einfache Titelaufnahmen vier verschiedener Bestände mit den Namen ihrer Elemente.

n	Titelaufnahme T_1	Titelaufnahme T_2	Titelaufnahme T_3	Titelaufnahme T_4
1	Bestellnummer	Laufende Nummer	Bestellnummer	Laufende Nummer
2	Betreff	Inhalt	Laufende Nummer	Alte Signatur
3	Entstehungszeit	Laufzeit	Umfang	Inhalt
4			Laufzeit	Laufzeit
5			Titel	
6			Enthält	
7			Darin	
8			Provenienz	
9			Vorsignaturen	
10			Sperrdatum	

Die oben formulierte Beziehung zwischen Bedeutung und Name ist die Begründung für die (extreme) Forderungen, daß hier erstens, für verschiedene Bedeutungen keine gleichen Namen und zweitens, für ein- und denselben Sachverhalt keine verschiedenen Namen benutzt werden sollten.

Bei der Findmittelkonversion hat man es typischerweise mit zwei verschiedenen Kontexten T_q ,T_z zu tun, das heißt, mit einer Quellstruktur T_q deren Inhalte in eine Zielstruktur T_z zu übertragen sind. Dabei werden Elemente des Quellvektors T_q auf Elemente des Zielvektors T_z abgebildet.

Dazu einige Konstellationen, die bei der Findmittelkonversion zu beachten sind:

a) Alle Elemente der Quell- und Zielvektoren werden abgebildet.

b) Einige Elemente des Quell- oder des Zielvektors werden nicht abgebildet.

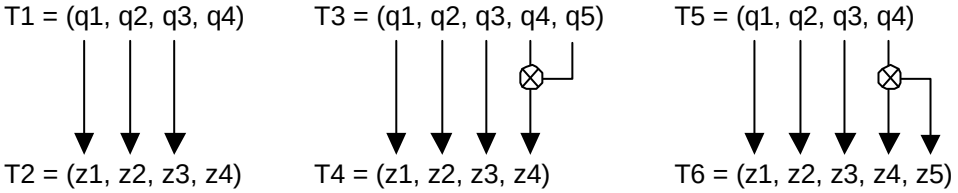
c) Quell- und Zielvektor haben die gleiche Anzahl Elemente.

d) Quell- und Zielvektor haben unterschiedliche Anzahl Elemente.

e) Jedes Element des Quellvektors wird auf genau ein Element des Zielvektors abgebildet und auf jedes Element des Zielvektors wird genau ein Element des Quellvektors abgebildet. Dies ist die einfachste Konstellation: eine eineindeutige, vollständige Abbildung oder 1-zu-1 mapping.

f) Jedes Element des Quellvektors wird zwar auf genau ein Element des Zielvektors abgebildet, aber auf mindestens ein Element des Zielvektors werden mindestens zwei Elemente des Quellvektors abgebildet. Dies ist eine Konstellation bei der Inhalte zusammengeführt werden.

g) Mindestens ein Element des Quellvektors wird auf mindestens zwei Elemente des Zielvektors abgebildet, aber auf jedes Element des Zielvektors wird genau ein Element des Quellvektors abgebildet. Dies ist eine Konstellation bei der Inhalte getrennt werden.



Damit wäre die Aufgabe der Findmittelkonversion bereits gelöst, wären da nicht die Inhalte, von denen alle oder die meisten im Kontext der Quellstruktur T_q anders dargestellt werden als im Kontext der Zielstruktur T_z.

Zunächst auch hier ein kurzer theoretischer Blick auf die Beziehungen zwischen Inhalt, Bedeutung und Name.

Inhalte C und **Bedeutungen** B stehen formal in der Beziehung: C **ist ein** B. Die Umkehrung B **ist ein** C gilt hier nur im Kontext der jeweiligen Bedeutung, aber in der Regel schon nicht mehr im Kontext einer Titelaufnahme, wenn darin verschiedene Bedeutungen durch formal identische Inhalte repräsentiert werden: man kann von einem Inhalt nicht eindeutig auf seine Bedeutung schließen.

Inhalte C und **Namen** A stehen damit im Kontext ein- und derselben Titelaufnahme formal in der Beziehung: C **ist ein** A, oder verbal: ein bestimmter Inhalt C ist ein durch einen bestimmten Namen A benennbares „Objekt".

Beispiel: Beziehungen zwischen Inhalt, Name und Bedeutung

> 19.2.1595 **ist ein** Geburtsdatum. Ein Geburtsdatum **ist ein** bestimmter, durch eine Urkunde belegter Kalendertag, an dem ein bestimmter Mensch geboren wurde.
>
> 17.1.1516 **ist ein** Vertragsdatum. Ein Vertragsdatum **ist ein** bestimmter, durch eine Urkunde belegter Kalendertag, an dem die Wirksamkeit eines bestimmten Vertrages eingetreten ist.
>
> „Otto von Gemmingen d.Ä., Erbregelung, 3 Sg. anh." **ist ein** Titel. Ein Titel **ist eine** Kurzfassung des Inhalts einer bestimmten Pergamenturkunde.

Die Inhalte einer Titelaufnahme $T = (e_1, e_2, e_3,.., e_n)$ werden formal ebenfalls als n-Tupel durch die Definitionsgleichung $C_i = (c_{1i}, c_{2i}, c_{3i},..., c_{ni})$ mit i = 1 bis m ausgedrückt, oder verbal: „der i-te Eintrag einer Titelaufnahme T besteht aus den Inhalten $c_{1i}, c_{2i}, c_{3i},..., c_{ni}$". Damit Selektionen und Änderungen von Inhalten eindeutig möglich sind, darf jedes der m Elemente höchstens einmal vorkommen. In einer Datenbank ist dann die Menge aller C_i zusammen mit T eine Datenbanktabelle oder Relation R und ein C_i ein Datensatz dieser Relation.

Ein uralter Trick, um irgendwelche „Objekte" eines x-beliebigen Kontextes eindeutig zu bezeichnen, besteht darin, daß man diese „Objekte" numeriert: beginnend mit Null erhält jedes neue Objekt die nächst größere ganze Zahl. Zahlen entfernter Objekte werden in dem gegebenen Kontext nie mehr zur Numerierung verwendet. Nach diesem Prinzip werden die Schlüssel der Datensätze einer Datenbanktabelle in der Regel automatisch gebildet. Auch Bestell- und Ordnungsnummern können nach diesem Prinzip vergeben werden. Eine lückenlose Numerierung kann zur Identifikation fehlender und mehrfach vorhandener Bestell- und Ordnungsnummern eines Bestandes und als Hilfe bei der Fehlerkorrektur zum Beispiel von Signaturen eingesetzt werden.

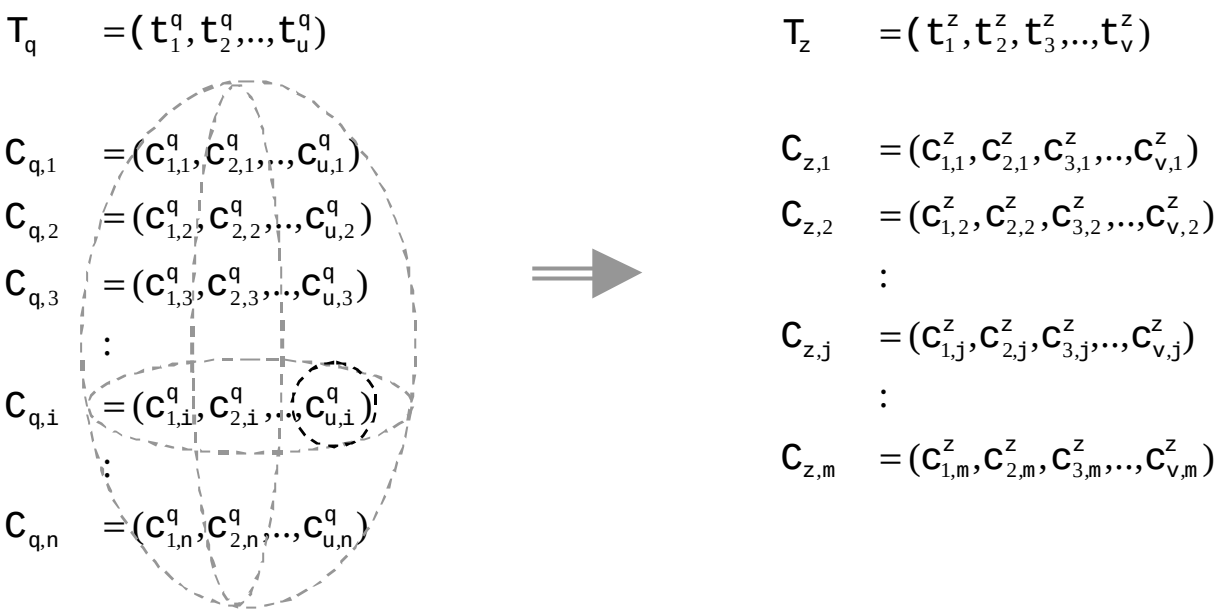
Inhalte von Titelaufnahmen sind meistens Zeichenfolgen in aktueller deutscher Sprache oder in anderen Sprachen, außerdem können auch graphische Repräsentationen, wie Skizzen, Zeichnungen, Bilder vorkommen. Dagegen wird man Ton und Film in einem Findbuch kaum antreffen, im digitalen Findmittel dagegen sehr wohl.

Beim PC-Einsatz bestehen alle Zeichenfolgen aus den Zeichen eines geeigneten PC Zeichensatzes. Alle graphischen und audio-visuellen Inhalte sind hier, als Ganzes betrachtet, kompakte, binärcodierte Objekte.

Im Mittelpunkt dieser Betrachtung der Findmittelkonversion stehen vor allem die Inhalte der Titelaufnahmen, die durch Zeichenfolgen darstellbar sind und die einem somit beim Lesen eines Findbuchs und bei der Ein- und Ausgabe am PC begegnen können.

Zum Begriff der Bedeutung (Semantik) kommt schließlich noch der Begriff der Repräsentationsform (Syntax und Pragmatik) des Inhalts.

Bei der Übertragung der Inhalte, das heißt, bei der Findmittelkonversion im engeren Sinn, werden **Methoden** eingesetzt, welche die gegebenen Repräsentationsformen einer Quellstruktur T_q mit ihren Inhalten $C_{q,i}$ in eine Zielstruktur T_z mit Ihren Inhalten $C_{z,j}$ überführen.



Diese Methoden können auf vier Ebenen ansetzen:

(1) die gesamte **Tabelle** (Zeichensatz-Konvertierung)
(2) eine einzelne **Spalte** oder mehrere Spalten (oben beschriebene Konstellationen)
(3) eine einzelne **Zeile** oder mehrere Zeilen (Behandlung von Fortsetzungsdatensätzen)
(4) ein einzelnes **Element** oder mehrerer einzelne Elemente (Fehlerkorrektur)

Auf Basis dieser Überlegungen ist die eigentliche Konversion weitgehend automatisch durchführbar.

Dazu abschließend einige Aspekte der Titelbildung, die bei der Findmittelkonversion eine Rolle spielen (Arbeitsnotizen, Checkliste):

Identität zwischen den Bezeichnungen auf dem Objekt (Archivalie) und in der Datenbank
Schreibfehler, Computerzeichensatz und Schriftzeichen, besonders historische, Fehlende Information

Verwendung von Leerzeichen

Verwendung von Satzzeichen: Punkt, Komma, Strichpunkt
Aufzählung, Reihung, Und, Oder, Stufung, Relation, Ist-ein, Gehört-zu, Anfang und Ende, Satzende

Verwendung von Klammern: runde, eckige, geschweifte

Verwendung von Anführungszeichen: doppelte und einfache

Verwendung von Abkürzungen

Rechtschreibung, historische Schreibweise

Editorielle Texte, Eintragungen, Kennzeichnungen, Hinweise
Funktion ?

Suchen und Ersetzen
Achtung!

Vielen Dank für Ihre Aufmerksamkeit!

◆