Ist automatische Normierung möglich?

Klaus Lepsky, Institut für Informationswissenschaft der Fachhochschule Köln

1. Einleitung

Normierung ist allgemein ein nützliches Instrument der formalen und inhaltlichen Dokument- und Medienbeschreibung. Aus diesem Grund werden in der bibliothekarischen Formal- und Inhaltserschließung zentrale Beschreibungselemente über sog. Normdateien kontrolliert, die über die Festlegung von Ansetzungsformen die einheitliche Beschreibung sichern, gleichzeitig durch die Bereitstellung von Nicht-Ansetzungsformen (Verweisungsformen) die Suche auch mit nicht bevorzugten Schreibweisen unterstützen (Synonymen). Normierungselemente in der bibliothekarischen Formalerschließung sind Verfassernamen, Körperschaften, in der Inhaltserschließung sind es Schlagwörter. Die gemeinsame Verwendung der Normdateien in Verbindung mit einem einheitlichen Erfassungsstandard (RAK) bzw. einem Quasi-Erschließungsstandard (RSWK) erleichtert die Datenübernahme und führt zu verlässlichen Erschließungs und Katalogumgebungen.

Für die Erschließung nicht-textlicher Objekte hat sich eine derartige Rahmenumgebung bislang nicht entwickelt. Objektdokumentation im musealen Bereich und Bilddokumentation in der Kunstgeschichte erfolgen nach jeweils lokalen Richtlinien, ein Rückgriff auf gemeinsame Erschließungsressourcen (z.B. Normdateien) ist nicht möglich, weil diese entweder nicht existieren oder existierende nicht allgemein genutzt werden.[1] Der Wunsch nach „normenden Instanzen“, mindestens aber nach einer Verständigung auf gemeinsame Standards wächst, allerdings ist es höchst unwahrscheinlich, dass sich die große Zahl sehr heterogener Erschließungswelten in ein gemeinsames Konzept bringen lässt. Realistischer ist es, von der existierenden Vielfalt auszugehen und Anstrengungen zu unternehmen, die Vielfalt nicht zum Problem werden zu lassen.[2] Dies bedeutet in erster Linie, Versuche zu unternehmen, unterschiedliche Beschreibungsdaten, die aber das Gleiche meinen, mit maschineller Hilfe zusammenzubringen. Dieser Beitrag versucht, für derartige Ansätze die Möglichkeiten und Grenzen des automatisch Machbaren aufzuzeigen.

1 Es gibt natürlich kunstgeschichtliche Normdateien bzw. Ressourcen, die eine Nutzung als Normdatei erlauben: The Union List of Artist Names (ULAN), The Art & Architecture Thesaurus (AAT), The Getty Thesaurus of Geographic Names (TGN), alle hrsg. vom Getty Research Institute, Los Angeles (www.getty.edu/research/conducting_research/vocabularies); Allgemeines Künstlerlexikon (AKL), K.G. Saur. Weitere Informationen zur Thematik in Lebrecht, Heike: Methoden und Probleme der Bilderschließung. Köln: Fakultät für Informations- und Kommunikationswissenschaften, 2003. 90 S. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; Bd.42) (www.fbi.fh-koeln.de/institut/papers/kabi/band.php?key=53).

2 So auch Krause, Jürgen: Konkretes zur These, die Standardisierung von der Heterogenität her zu denken. In: Zeitschrift für Bibliothekswesen und Bibliographie. 51(2004) H.2, S.76-89.

2. Normierung formaler Merkmale

Dieses Bild von Canaletto wird in der Bilddatenbank „Die virtuelle Galerie der 25.000 Meisterwerke“ folgendermaßen beschrieben:

```
Canaletto (I): Ansicht von Dresden, Der Neumarkt in Dresden vom
Jüdischen
Friedhof aus, mit Frauenkirche und Altstädter Wache.

1749-1751, Öl auf Leinwand, 136 × 236 cm.
Dresden, Gemäldegalerie.
Kommentar: Vedutenmalerei, Stadtlandschaft.
Land: Italien und Deutschland.
Stil: Venezianische Malerei des 18. Jahrhunderts.[3]
```

Unter dem Aspekt der Normierung ist v.a. die Bezeichnung „Canaletto (I)“ auffällig, die schließen lässt, dass es weitere Canalettos in der Datenbank gibt, die aber natürlich alle voneinander unterschieden werden. Unabhängig von der Bilddatei existiert eine Künstlerdatei, in der sich zu Canaletto (I) folgender Eintrag findet:

```
Canaletto (I)
eigentlich: Bernardo Bellotto
* 30.01.1721 Venedig
† 17.11.1780 Warschau
Wirkungsorte: Venedig, Warschau, Dresden[4]
```

Daneben existiert ein

```
Canaletto (II)
eigentlich: Giovanni Antonio Canal
Maler, Radierer, Zeichner
* 17.10.1697 Venedig
```

3 Bilddatenbank: Canaletto (I), S. 11. Die virtuelle Galerie der 25.000 Meisterwerke, S. 3895 (c) 2004 Zweitausendeins, Frankfurt am Main.

4 Bilddatenbank: Canaletto (I), S. 1. Die virtuelle Galerie der 25.000 Meisterwerke, S. 3885 (c) 2004 Zweitausendeins, Frankfurt am Main.

```
† 10.04.1768 Venedig
Wirkungsorte: Venedig, London
```
[5]

Die Vorzüge der Normierung von Künstlernamen für eine Bilddatenbank liegen auf der Hand und entsprechen zunächst den allgemeinen Zielsetzungen einer formalen Erschließung[6]:

- *Zusammenführung von Gleichem*: Die Normierung und hier auch Individualisierung von Künstlernamen erlaubt die Zusammenführung aller Werke eines Künstlers für die Suche. Die Festlegung einer Ansetzungsform („Canaletto“) für den Künstlernamen mit gleichzeitiger Erfassung alternativer Schreibweisen erlaubt bei Einbindung dieser Informationen in den Suchindex eine erfolgreiche Suche für alle Namensvarianten eines Künstlers.
- *Trennung von Verschiedenem*: Nur die Individualisierung in Verbindung mit der Normierung erlaubt die getrennte Zuweisung von Werken zu Künstlern mit identischer Namensform. Erfolgt keine Individualisierung, werden Werke unterschiedlicher Künstler unter nur einer Namensform zusammengefasst, also z.B. unter Canaletto.

Die Methode, mit der eine Normierung und Individualisierung zu erreichen ist, kann natürlich nur die intellektuelle Entnahme aus dem Dokument bzw. intellektuelle Ermittlung auf der Basis des Dokuments sein. Für diese Ermittlung ist das Wissen nötig, dass es überhaupt zwei Personen mit identischer Namensform gibt und dass daher, wie in unserem Fall, für die identische Namensform unterschiedliche alternative Namensformen existieren, von denen auf die bekanntere Ansetzungsform zu verweisen ist.

Hilfsmittel für diese Normierung formaler Beschreibungsmerkmale sind Normdateien, die sprachlich kontrollierte Beschreibungselemente zur Verfügung stellen, insbesondere sind dies Normdateien für Personen-/Künstlernamen, Werktitel und Körperschaften.[7]

3. Normierung inhaltlicher Merkmale

Ähnlich wie die formale Erschließung dient die Inhaltserschließung der Zusammenführung von Dokumenten, hier allerdings der Zusammenführung hinsichtlich gleicher Inhalte bzw. Themen. Eine derart konsistente Beschreibung von Bildinhalten kann dabei nur über die Verwendung von Dokumentationssprachen erreicht werden. Schlagwortsprache, Thesaurus und Klassifikation bilden den methodischen Rahmen, innerhalb dessen inhaltliche Beschreibungen kontrolliert erstellt werden können.

Erreicht werden kann eine inhaltliche Beschreibung nur durch die intellektuelle Inhaltsanalyse. Vor allem Bilder entziehen sich durch die potenzielle Mehrschichtigkeit ihres Inhalts einer automatischen Analyse: stellt Canalettos Bild einen „öffentli-

5 Bilddatenbank: Canaletto (II), S. 1. Die virtuelle Galerie der 25.000 Meisterwerke, S. 3947 (c) 2004 Zweitausendeins, Frankfurt am Main

6 Vgl. auch Eversberg, B.: Zur Theorie der Bibliothekskataloge und Suchmaschinen. In: Die Bibliothek zwischen Autor und Leser: 92 Deutscher Bibliothekartag in Augsburg 2002. Hrsg.: H. Benkert u.a. Frankfurt a.M.: Klostermann, 2003. (Zeitschrift für Bibliothekswesen und Bibliographie: Sonderh.84); Eversberg, B.: Was sollen Bibliothekskataloge. http://www.allegro-c.de/formate/gz-1.htm.

7 Allg. bibliothekarische Normdateien: Personennamendatei (PND), Gemeinsame Körperschaftsdatei (GKD), Schlagwortnormdatei (SWD), die kooperativ gepflegt werden und über Die Deutsche Bibliothek bezogen werden können. Dort (www.ddb.de) auch weitere Informationen zu den Normdateien.

chen Platz mit Häusern und Kirche“ oder den „Neumarkt in Dresden mit Frauenkirche und Altstädter Wache“ dar?[8] Die schon bekannte Bildbeschreibung ignoriert den ersten Aspekt völlig, obwohl er natürlich im Bild vorhanden ist, stattdessen beschreibt sie den Inhalt als „Stadtlandschaft“ und ordnet das Bild der „Venezianischen Malerei des 18. Jahrhunderts zu“. Letzteres ist dem Bild nur mit erheblicher Vorkenntnis zu entnehmen und hängt ganz wesentlich von der Berücksichtigung einiger „außerbildlicher“ Rahmenfaktoren ab (wie z.B. der Kenntnis spezifischer Stilmerkmale, des Künstlernamens und der Echtheit des Bildes).

Im Gegensatz zur Normierung in der formalen Erschließung ist Normierung in der inhaltlichen Erschließung begriffliche und sprachliche Normierung. Begriffliche Normierung bewirkt ein Zusammenführen auf thematischer Ebene, also etwa die Zusammenführung aller Bilder gleichen Inhalts unter einem Schlagwort. Um dies zu erreichen, ist es bei der Erstellung der Dokumentationssprache „Schlagwortmethode“ erforderlich, genau diesen Inhalt „auf den Begriff zu bringen“ und gleichzeitig von Inhalten, von denen unterschieden werden soll abzugrenzen. Sprachliche Normierung meint die Festlegung eines bestimmten Schlagworts als Vorzugsbenennung mit Relationierung aller nicht verwendenten (aber theoretisch denkbaren) Formen, d.h. Schaffung von Verweisungsformen.

Normierungsinstrumente in der inhaltlichen Erschließung und Dokumentationssprache selbst sind nicht strikt voneinander zu trennen. Die Klassifikation „ist“ das Instrument zur begrifflichen Normierung, deren Register häufig sprachlich normiert. Der Thesaurus stellt sprachlich normierte und begrifflich relationierte Deskriptoren zur Beschreibung von Inhalten bereit. Die Schlagwortmethode verwendet ebenfalls kontrolliertes Vokabular, um Inhalte begrifflich zusammenzuführen. Die Erstellung all dieser Erschließungselemente ist intellektuell aufwändig und ihre Anwendung in der Erschließung ist es ebenfalls.

4. Möglichkeiten automatischer Verfahren

Angesichts der erheblichen Rolle, die intellektuelle Anteile an der formalen und inhaltlichen Erschließung und, damit einhergehend, der Normierung formaler und inhaltlicher Beschreibungen haben, ist nur schwer vorstellbar, dass Verfahren des automatischen Indexierens einen Beitrag hierzu leisten können. Automatische Indexierung ist in der Tat zunächst lediglich ein Sammelbegriff für alle Verfahren, die aus Dokumenten Stichwörter entnehmen und auf dieser Basis den Dokumenten geeignete Indexterme zuweisen.[9] Dabei umfasst der von linguistisch und statistisch arbeitenden Verfahren erreichte Funktionsumfang die folgenden Teilfunktionen:

Linguistische Funktionalität:

- Erzeugung von grammatikalischen Grundformen:
  Bildern → Bild
- Zerlegung von Komposita im Deutschen:
  Bilderschließung → Bild, Erschließung
- Bildung von Wortableitungen (bevorzugt adjektivische auf substantivische Form):
  mittelalterlich → Mittelalter

8 Es sind natürlich durchaus mehr als drei Ebenen denkbar. Vgl. zur allg. Problematik der Bilderschließung auch Lebrecht, Methoden und Probleme der Bilderschließung, 2003.

9 Vgl. zur Funktionalität der automatischen Indexierung: Lepsky, Klaus: Automatische Indexierung zur Erschließung deutschsprachiger Dokumente. In: nfd Information - Wissenschaft und Praxis. 50(1999) H.6, S.325-330.

- Erkennen von Mehrwortgruppen, festen Wendungen:
- *mittelalterliches Tafelbild → Tafelbild, mittelalterlich*
- Relationierung von Synonymen (bzw. hierarchischen Beziehungen)
  Madonnenbild → Marienbild

Statistische Funktionalität:

- Bereitstellung gewichteter Indexterme (z. B. für ein Relevance Ranking von Trefferlisten):
  Bild {1.97}; Bilderschließung {3.75}
- Zuteilung von kontrolliertem Vokabular auf der Basis statistischer Analyse von Worthäufigkeiten in Dokumenten;
- Automatische Klassifizierung von Dokumenten, d.h. Zuteilung von Notationen existierender Klassifikationen zu Dokumenten;
- Clustering von Dokumentkollektionen, d.h. Erkennen von thematisch ähnlichen Dokumenten.

Auf der Basis dieses Funktionsspektrums können Dokumente um zusätzliche, z.B. sprachlich normierte Indexterme angereichert werden, um die Retrievalmöglichkeiten im Sinne einer Recall-Erhöhung zu verbessern. Insbesondere die lexikalische Zuteilung von relationiertem Vokabular (unter Ausnutzung vorhandener terminologischer Quellen (z.B. Normdateien, Thesauri)) nach zuvor erfolgter grammatikalischer Vereinheitlichung führt zu einem starken Anstieg zusätzlicher, nützlicher Sucheinstiege.[10]

Der Einsatz statistischer Verfahren ist im wesentlichen konzentriert auf das Ranking von Treffermengen, statistisch basierte Zuweisungsverfahren von Dekriptoren eines Thesaurus oder Notationen von Klassifikationssystemen sind weitaus seltener und fast ausnahmungslos gebunden an enge Rahmenbedingungen für die Dokumentenkollektion, z.B. starke fachliche Einschränkung oder das Vorhandensein von Volltexten bzw. Volltextbestandteilen.[11] Es ist offensichtlich, dass sich diese Einsatzmöglichkeiten – die durchaus auch als Einsatzgrenzen betrachtet werden sollten – nicht mit den Ansprüchen an eine „automatische Normierung“ in Einklang bringen lassen. Für den potenziellen Nutzen der automatischen Indexierung für die Zwecke einer Normierung ist daher zunächst zu klären, was genau Normierung im hier gebrauchten Kontext meinen kann.

Bislang wurde nicht zwischen den beiden Konzepten Normierung und Homogenität unterschieden. Die Betrachtung von Formal- und Inhaltserschließung ließ selbstverständlich erscheinen, dass das Ziel der Normierung nur Homogenität der Erschließung bedeuten kann. Andererseits ist Homogenität ein typisches Katalogmerkmal als Instrument für das Zusammenführen von Gleichem. Es bleibt zu fragen, inwieweit dies eine Zielsetzung ist, die in Retrievalumgebungen eine gleich große Rolle spielt und, darüber hinaus, ob dieses Ziel allein durch Normierung, also durch Homogenität der Erschließung erreicht werden kann.

10 Gödert, Winfried, Lepsky, Klaus: Semantische Umfeldsuche im Information Retrieval. In: Zeitschrift für Bibliothekswesen und Bibliographie. 45 (1998) H. 4, S. 401-423.

11 Vgl. für einen Überblick zur automatischen Klassifizierung: Oberhauser, Otto: Automatisches Klassifizieren. Entwicklungsstand - Methodik – Anwendungsbereiche. Mit einem Vorwort von Winfried Gödert. Frankfurt u.a. 2005. Zur statistisch basierten Deskriptorzuteilung sind die Ergebnisse des AIR/PHYS-Projekts noch immer aktuell: Fuhr, N.; Knorz, G.; Lustig, G.; Schwandtner, M.; Biebricher, P.: Entwicklung und Anwendung des automatischen Indexierungssystems AIR/PHYS. In: Nachrichten für Dokumentation. 39 (1988), S. 135-143.

Eine Rangliste der Recherchemöglichkeiten in unterschiedlich charakterisierten Kollektionen und Suchszenarien sieht, sortiert nach absteigendem Sucherfolg, etwa folgendermaßen aus:

a) kontrollierte Suche (Thesaurus, Registerbegriffe einer Klassifikation) auf homogene Erschließung;

b) freie Suche auf homogene Erschließung;

c) freie Suche auf heterogen (oder gar nicht) erschlossene Kollektion.

Der Sucherfolg für die kontrollierte Suche auf eine homogene Erschließung ist garantiert, weil hier das im Katalogmodell wirksame Prinzip von Such- und Erschließungsseite unterstützt wird. Die freie Suche auf eine homogene Erschließung ist allein abhängig vom Treffen eines Erschließungsmerkmals; gelingt dies, ist der Sucherfolg genauso hoch wie bei a), wird kein Erschließungsmerkmal getroffen, misslingt die Recherche oder liefert Teilmengen aufgrund zufällig getroffener Beschreibungsmerkmale wie z.B. Titelstichwörter. Die Situation für die freie Suche auf eine heterogen erschlossene Kollektion ist schließlich die unbefriedigendste, denn selbst das Treffen eines Erschließungsmerkmals mit dem Suchbegriff liefert nur eine mehr oder weniger große Teilmenge als Resultat.

Selbstverständlich sind alle Suchkonstellationen in der Praxis vertreten, wobei die Problemsituationen in b) und c) in den seltensten Fällen bewusst herbeigeführt werden. Häufig sind sie die Konsequenz aus technischen Rahmenbedingungen (z.B. fehlende Einbindungsmöglichkeiten für genormtes Vokabular in die Suche) oder das Ergebnis des Zusammenführens unterschiedlich erschlossener Bestände (z.B. in größeren Verbundsystemen). Klar ist aber zumeist, dass eine Veränderung der Situation hin zum Idealszenario a) als Lösungsmöglichkeit für die existierenden Probleme wegen entweder zu großer Dokumentmengen und/oder zu geringer personeller Ressourcen nicht in Frage kommt.

Damit bleibt für realistische Möglichkeiten der Verbesserung der Situation nur der Einsatz automatischer Hilfsmittel. Natürlich darf auch hier nicht erwartet werden, dass aus einer Ausgangslage wie in b) oder c) eine Verbesserung in Richtung von a) möglich ist. Die Idee der automatischen Indexierung ist vielmehr, die Situation c), also die freie Suche auf eine heterogen erschlossene Kollektion vom Ergebnis her „erträglicher“ zu machen. Was ist damit gemeint?

Formale Erschließung

Wesentlicher Zugriffspunkt in der formalen Erschließung von Kunstwerken ist der Künstlername, für den es häufig eine Vielzahl von möglichen Varianten gibt. Retrievaltechnisch entsteht das Problem, dass eine fehlende Normierung unterschiedlicher Namensformen das Finden der Gesamtheit aller Objekte des Künstlers verhindert. Zu lösen ist dies entweder durch die automatische Generierung der Vorzugsform auf der Basis der erkannten Variante (Zielvorstellung Homogenität (A)) oder durch die Generierung >aller< Varianten (Zielvorstellung Heterogenität (B)).

```
A) DaVinci Leonardo     ⇘
   Leonard de Vinci     ⇨   Leonardo da Vinci
   Leonardus Vincius    ⇗

B) Lionardo da Vinci    ⇨   Leonardo da Vinci
                        ⇘   DaVinci Leonardo
                        ⇘   Leonard de Vinci
                            etc.
```

Technisch wird bei beiden Verfahren auf die Erkennung von Mehrwortgruppen im Rahmen der automatischen Indexierung zurückgegriffen. Voraussetzung für eine Abbildung von erkannten Mehrwortgruppen auf eine oder mehrere Namensformen ist das Vorhandensein einer lexikalischen Quelle, die die Beziehungen synonym zu verwendender Namensformen enthält, d.h. eine Quelle für die Normierung ist (z.B. die Personennamendatei, Künstlernamen der Schlagwortnormdatei). Systeme zur automatischen Indexierung des Deutschen verfügen über eine solche Fähigkeit zur Erkennung von Mehrwortgruppen ebenso wie über die Fähigkeit zur Integration unterschiedlicher Wörterbuchquellen für derartige Erkennungsläufe.[12]

Inhaltliche Erschließung

Ähnlich stellt sich die Situation für die inhaltliche Erschließung dar: auch hier gibt es für einen Sachverhalt in der Regel unterschiedliche Bezeichungen, auch hier ist es relativ häufig eine Mehrwortgruppe, die für die Beschreibung verwendet wird. Zielsetzungen können wiederum Homogenität oder Heterogenität sein:

```
A) Madonnenbild (Kunst)  ↘
   Maria (Kunst)         ⇨  Marienbild
   Madonna               ↗

B) Kunst / Maria (Motiv)    ⇨  Marienbild
                            ↘  Maria (Kunst)
                            ↘  Madonnenbild (Kunst)
                               etc.
```

Weitere, bislang noch ungenannte Voraussetzung für eine automatische Indexierung im hier gezeigten Umfang ist das Vorhandensein einer Objektbeschreibung, der die für die Relationierung verwendeten Terme entnommen werden können. Dies muss mindestens eine formale Beschreibung sein, besser natürlich zusätzlich eine inhaltliche Beschreibung. Im Gegensatz zur automatischen Indexierung textbasierter Dokumente kann in der Objektdokumentation nicht das Dokument selbst (bzw. die in ihm enthaltene Gesamtheit der Terme) für die automatische Verarbeitung herangezogen werden. Für die automatische Generierung alternativer Bezeichungen ist es dabei letztlich unerheblich, welcher Kategorie die entnommenen Terme entstammen.

Es ist wichtig zu sehen, dass die Entscheidung für Verfahrensweg A) oder B) wesentlich davon abhängt, ob prinzipiell eine Erschließungs- bzw. Retrievalsituation besteht, die Homogenität unterstützt, oder z. B. eine Situation vorliegt, die mangels inhaltlicher oder gemeinsamer inhaltlicher Erschließung große Heterogenität aufweist. B) ist im Ergebnis völlig unabhängig vom Vorhandensein normierender Instanzen, dadurch dass jeder Term auf alle bekannten Varianten abgebildet wird, d. h. alle Suchen mit einer der Varianten zum gleichen Ergebnis führen. Verfahren A) fordert in der Suche die Eingabe der Vorzugsbenennung oder muss alternativ in der Suche, etwa durch Einbindung des Thesaurus, von Alternativen auf die bevorzugte Form verweisen. Damit ist die Existenz eines (d. h. einheitlich für alle Dokumente zu verwendenden) Vokabulars Voraussetzung des Verfahrens.

Noch einmal: Die Vorzüge des automatischen Verfahrens liegen v.a. darin, dass die Terme, die benötigt werden, aus allen Kategorien stammen können, dass die Anforderungen an die Erschließung der Dokumente extrem niedrig sind und dass vorhandene Normierungsvokabularien verwendet werden können unabhängig davon, ob sie für die zu verarbeitende Dokumentkollektion verwendet wurden oder werden.

12 Dies gilt z.B. für die Indexierungen IDX (http://www.dfki.de/lt/idx.php), Extrakt (http://www.textec.de/) und Autindex (http://www.iai.uni-sb.de/iaide/de/prod_autindex.htm).

Aus der vorliegenden Dokumentbeschreibung zum Canaletto-Bild würden durch automatische Indexierung unter Berücksichtigung des vollständigen Funktionsumfangs folgende für die Suche zu verwendenden Terme generiert werden könnnen:

```
Canaletto (I): Ansicht von Dresden, Der Neumarkt in Dresden vom
Jüdischen Friedhof aus, mit Frauenkirche und Altstädter Wache.
```

- Zuweisung alternativer Namensformen aus einer Normierungsquelle:
  `Canaletto (I)` ⇨ `Bernardo Bellotto; Bellotto, Bernardo`
- Bereitstellung von Substantiven als Indexterm; falls „Ansicht von Dresden“ in einer Normierungsquelle Werktitel ist, könnte dieser ebenfalls indentifiziert werden:
  `"Ansicht von Dresden"` ⇨ `Dresden, Ansicht`
  `"Neumarkt in Dresden"` ⇨ `Dresden, Neumarkt`
- Bereitstellung von sprachlich standardisierten Grundformen als Indexterme; Erkennung der Mehrwortgruppe „Jüdischer Friedhof“, Invertierung der Mehrwortgruppe:
  `"Jüdischen Friedhof"` ⇨ `Jüdischer Friedhof;Friedhof,jüdischer`
  `Frauenkirche` ⇨ `[Dresden,] Frauenkirche`
- Bereitstellung von sprachlich standardisierten Grundformen als Indexterme; Erkennung der Mehrwortgruppe „Altstädter Wache“:
  `"Altstädter Wache"` ⇨ `Altstädter Wache;Altstadt;Wache;[Dresden]`

Insgesamt sind mit den gewonnenen Indextermen Merkmale der formalen und inhaltlichen Dokumentbeschreibung automatisch zugewiesen worden, die zum Teil genormten Vokabularien entstammen. Dadurch ist keinesfalls eine genormte Beschreibung entstanden, es stehen lediglich zusätzliche Zugriffsmöglichkeiten auf das Dokument zur Verfügung, ein legitimes Ziel im Information Retrieval.

5. Bedingungen, Thesen, Probleme

Verfahren zur automatischen Indexierung eignen sich nicht für jede Dokumentkollektion und jeden Einsatzzweck. Sinnvoll ist ihr Einsatz aber unbedingt dort, wo Kollektionen aus unterschiedlich erschlossenen Teilkollektionen bestehen – ein Zustand, der bei allgemein starker Retrokonvertierungstätigkeit, häufig in Verbindung mit Verbundlösungen, in Bibliotheken bereits zum Standard geworden ist. In derartigen Kollektionen stehen unterschiedliche Erschließungssysteme nebeneinander und meist sind erhebliche Teile der Kollektion nur notdürftig (also lediglich formal) erschlossen. Konversion und Bestandskumulation sorgen dabei zumeist für Kollektionsgrößen, für die intellektuelle Verfahren nicht mehr in Betracht kommen.

Für solche heterogen und/oder schwach erschlossenen Kollektionen ist die automatische Indexierung die einzige Möglichkeit, die Retrievalbedingungen entscheidend zu verbessern. Stehen darüber hinaus für die Zwecke der automatischen Indexierung kontrollierte Terminologien zur Verfügung, lassen sich, wie gezeigt, über die Defizite fehlender Erschließung hinaus auch die durch Heterogenität hervorgerufenen Probleme bis zu einem gewissen Grad bewältigen. Voraussetzung dafür ist die Existenz von Objektbeschreibungen, die automatisch indexiert werden können, wobei deren Qualität die Qualität der automatischen Indexierung direkt bestimmt. Voraussetzung für die automatische Generierung von kontrollierten Erschließungsmerkmalen ist die Existenz von umfangreichen Terminologien (Normdateien, Thesauri), die fachlich orientiert sein müssen, jedoch nicht für die Erschließung der zu indexierenden Kollektion verwendet worden sein müssen.

Aus diesen Voraussetzungen und aus den skizzierten Möglichkeiten automatischer Indexierungsverfahren lassen sich zusammenfassend einige Thesen für die Erschließungspraxis im Bereich der Objektdokumentation ableiten:

- Es ist wichtiger, Objekte >überhaupt< formal und inhaltlich zu beschreiben als Objekte >normiert< zu beschreiben.

  Interessanterweise kreisen die meisten Ansätze zur Lösung einer unbefriedigenden Erschließungssituation in der Objektdokumentation um die Entwicklung und Gestaltung „einheitlicher" Verfahren zur Normierung. Darin wird die Vorstellung deutlich, Heterogenität könne allein durch Normierung, d.h. Homogenität begegnet werden. Alternativen, insbesondere die Alternative, die Heterogenität zu akzeptieren und Wege zu suchen, das Retrieval auf heterogene Daten zu verbessern, liegen außerhalb des aktuellen Diskussionsfokus. Ließe man sich auf ein solches Modell ein, würde recht schnell deutlich werden, dass sich die Ergebnisse, die sich durch automatische Indexierung erzielen lassen, mit der Qualität der vorhandenen Dokumentbeschreibungen sehr leicht steigern lassen. Überspitzt formuliert: Ziel der Erschließung sollte nicht die normierte Beschreibung sein, Ziel sollte eine Beschreibung sein, die den Möglichkeiten der automatischen Indexierung entgegenkommt und damit dem Retrieval dient.

  Erfolgreiches Information Retrieval ist direkt abhängig von dem Vorhandensein einer ausreichenden Menge an Zugriffsmöglichkeiten auf ein Dokument, d. h. von der Zahl brauchbarer Indexterme zu einem Dokument. Konventionelle, an Katalogfunktionen orientierte Erschließungsinstrumente stellen Indexterme nicht in ausreichender Zahl zur Verfügung, weil sie bemüht sind, die Dokumentinformationen möglichst stark zu verdichten. Dies kommt den Suchmöglichkeiten in listen- und registergestützten Umgebungen entgegen, verhindert aber einen ausreichenden Recall beim Retrieval auf solche Daten. Die Angleichung der Bedingungen des Katalogs an die Gegebenheiten eines Retrievals ist nur bedingt möglich, weil beide Prinzipien einander widerstreitende Ziele verfolgen: aus Sicht des Retrievals sind katalogorientierte Dokumentbeschreibungen zu informationsarm, aus Sicht des Katalogs sind retrievalorientierte Dokumentbeschreibungen zu wenig strukturiert und kontrolliert bzw. normiert. So kann es etwa aus Retrievalgründen sehr zweckmäßig sein, statt nur der Ansetzungsform (für Namen, Schlagwort etc.) auch die Nicht-Ansetzungsformen als Indexterme zur Verfügung zu stellen, denn die Einbindung normierten Vokabulars und dazugehöriger Verweisungsstrukturen ist in Retrievalsystemen nur bedingt möglich.[13]

  Die Anpassung der Indexterm-Situation an die Bedingungen des Information Retrieval ist die Hauptmotivation für eine automatische Indexierung. Dies bedeutet, dass die in 4. genannten Funktionalitäten nicht nur zu einer Verbesserung der Retrievalfähigkeit bereits vorhandener Terme in der Dokumentbeschreibung führen, sondern insb. durch die Kompositumzerlegung, die Bildung von Wortableitungen und natürlich die Einbindung von relationierten genormten Vokabularien die Zahl der Indexterme signifikant erhöhen. Dies gelingt umso besser, je größer die Zahl der Terme im Dokument ist, d.h. je mehr Aufsatzpunkte die Indexierung hat.

- Normdateien und Thesauri aufzubauen und zu pflegen ist wichtiger als ihre Verwendung in der intellektuellen Erschließung.

  Der einfachste Weg, die Ergebnisse einer automatischen Indexierung zu verbessern, besteht darin, die ihr zur Verfügung stehenden terminologischen

13 Übrigens auch in Katalogen, denn sobald im Katalog eine Mischsuche auf normiertes Vokabular und nicht-normietes Vokabular angeboten wird (z.B. also auf einen Basic Index aus Schlagwörtern und Titelstichwörtern), ist die Nutzung der Verweisungsstrukturen nicht mehr möglich.

Ressourcen zu vergrößern. Aufbau und Pflege von genormten Terminologien sollten, wenn über ihren Einsatz für eine automatische Indexierung entschieden ist, ganz bewusst im Hinblick auf diesen Einsatz weiter entwickelt werden. Das bedeutet z. B., dass die konkrete Art und Weise der Ansetzung im Thesaurus eher unbedeutend ist, wichtiger ist die Verfügbarkeit möglichst vieler äquivalenter Bezeichnungen.[14]

- Es ist leichter, mit den Mitteln des Retrievals eine heterogene Erschließungssituation zu bewältigen als Homogenität in der Erschließung zu erreichen.

  Der vermutlich schwerste Schritt hin zu dem hier skizzierten Erschließungsmodell ist die Aufgabe des Ideals einer homogenen Erschließung. Das große Beharrungsvermögen dieser Idealvorstellung lässt sich z. B. in der bibliothekarischen Inhaltserschließung mit den „Regeln für den Schlagwortkatalog“ seit Jahren beobachten. Obwohl die Zahl der nicht nach den RSWK erschlossenen Dokumente in den Verbundkatalogen durch Retrokonversion großer Alttitelbestände kontinuierlich gestiegen ist, damit natürlich auch die Heterogenität stark angewachsen ist, wird das Modell der einheitlichen und normierten intellektuellen Erschließung ideenlos fortgeschrieben. Dabei wäre es nicht nur möglich, sondern auch sinnvoll, die Schlagwortnormdatei zu einem für die automatischen Indexierung geeigneten Instrument weiter zu gestalten, um so für die Gesamtbestände deutlich bessere Retrievalbedingungen zu erzielen.

Ein realistisches Erschließungsmodell für die Objektdokumentation besteht aus drei Komponenten:

- einer *intellektuellen Erschließung von Objekten*, die Objekte ausführlich (im Sinne von nicht eng) verbal beschreibt[15] und damit eine ausreichende Termbasis für die automatische Indexierung liefert;
- dem *systematischen Auf- und Ausbau terminologischer Ressourcen* (u.a. auch Normdateien) für die automatische Indexierung;
- dem *Einsatz einer automatischen Indexierung* für die Anreicherung und sprachliche Normierung der Objektbeschreibungen (aufbauend auf den von Normdateien bereitgestellten Relationen).

Verzicht auf die intellektuelle Erschließung ist in der Objektdokumentation anders als in der Dokumenterschließung nicht möglich, weil nicht-sprachliche Objekte bzw. Objektrepräsentationen nicht retrievalfähig sind.[16] Darüber hinaus entziehen sich Kunstwerke durch die Existenz mehrschichtiger Beschreibungsebenen dem inhaltlichen „auf den Begriff bringen“, verlangen bereits in der Inhaltsanalyse eine starke

14 Der Nutzen der Schlagwortnormdatei für die automatische Indexierung ist u.a. deshalb eingeschränkt, weil die Ansetzungsformen in der SWD (erst recht natürlich die RSWK-Ketten, die aber im Zusammenhang mit automatischer Indexierung keine Bedeutung haben) zwar schlagwortkatalog- und listentauglich sind, allerdings nur eingeschränkt retrievaltauglich (Homonymenzusätze, Ansetzungsketten, geringe Zahl an ausgewiesenen Synonymbeziehungen etc.).

15 „Ausführlich“ steht hier nicht für eine dem Abstract ähnliche Form der textlichen Beschreibung sondern für eine mit unterschiedlichen Kateogrien/Aspekten arbeitende schlagwort-ähnliche Erschließungsmethode, die auf extreme Informationsverdichtung ebenso verzichtet wie auf völlige Terminologiekontrolle, d.h. auch über eine freie Komponente verfügt. Die Zusammenführung von freien Erschließungselementen und ggf. vorhandenen kontrollierten Vokabularien kann dann wiederum über die automatische Indexierung erfolgen.

16 Echtes „Bildretrieval“, d.h. Retrieval nach visuellen Merkmalen, ist noch keine Alternative zur sprachlichen Objektbeschreibung und wird es zumindest für den Bereich der kunstgeschichtlichen Bilderschließung auch nicht werden können, vgl. Lebrecht, Methoden und Probleme der Bilderschließung, 2003, Kapitel 3.2.

intellektuelle Auseinandersetzung mit dem Objekt. Dennoch sollte die intellektuelle Erschließung durch den Verzicht auf streng normierende Rahmenbedingungen insgesamt weniger aufwändig sein, sehr häufig sollte es eben auch möglich sein, bereits vorliegende Beschreibungen zu nutzen.

Entscheidend für die Qualität des hier beschriebenen Erschließungsansatzes ist die Verfügbarkeit der für die automatische Indexierung zu verwendenden terminologischen Quellen. Neben der Einbeziehung zentral gepflegter Instrumente wie die Normdateien ist v.a. die Integration spezieller Erschließungsvokabularien von Interesse, die bislang nur lokal eingesetzt wurden. Auf- und Ausbau der terminologischen Basis ist zweifellos arbeitsaufwändig, muss allerdings nur einmal geleistet werden. Die kontinuierliche Pflege vorhandener Terminologie ist mit relativ geringem Aufwand verbunden. Die automatische Indexierung selbst schließlich erfodert lediglich finanziellen und organisatorischen Aufwand zum Zeitpunkt der Einführung. Dafür erhält man dann allerdings ein System, dessen Leistungsumfang immer wieder abgerufen werden kann, d.h. auch bereits indexierte Kollektionen können bei deutlichen Vokabularverbesserungen einfach erneut automatisch indexiert werden, das Ergebnis der Indexierung damit auch für zurückliegende Bestände verbessert werden – ein unschätzbarer Vorzug gegenüber allen katalogorientierten Erschliessungsmodellen, die bei Umstellungen zu Katalogbrüchen führen.

Abschließend noch einmal zurück zur Ausgangsfrage: Ist automatische Normierung möglich?

Eine der möglichen Antworten wäre wohl: vielleicht, aber eigentlich ist automatische Normierung gar nicht sinnvoll. Sinvoller ist es, dafür zu sorgen, dass mit erheblichem Aufwand hergestellte Objektbeschreibungen unter Retrievalbedingungen gesucht und gefunden werden können. Dazu bedarf es nicht der normierten Objektbeschreibung, dazu bedarf es vielmehr des unvoreingenommenen und abgestimmten Einsatzes aller heute zur Verfügung stehenden Erschließungs- und Retrievalinstrumente.

Prof. Dr. Klaus Lepsky
Institut für Informationswissenschaft, Fachhochschule Köln
Claudiusstraße 1, 50678 Köln
http://www.fh-koeln.de
klaus.lepsky@fh-koeln.de