Technology
Arts Sciences
TH Köln

Institut für Informationswissenschaft

Fakultät für Informations- und Kommunikations-wissenschaften

Prof. Dr. Klaus Lepsky
klaus.lepsky@th-koeln.de

Informationserschließung – Strukturierte Dokumentbeschreibung [DIS 07]

Tutorial

März 2024

Modulinhalte

Das Modul „Informationserschließung – Strukturierte Dokumentbeschreibung (DIS 07)“ schließt an das Modul „Informationserschließung und Wissensorganisation (DIS 02)“ an und setzt die dort behandelten Inhalte voraus. Es behandelt die Prinzipien einer formalen und inhaltlichen Dokumentbeschreibung für heterogene Dokumenttypen. Durch die selbstständige Analyse gängiger Datenformate, die Erweiterung und Anpassung eigener Kategorienschemata und die praktische Konzeptionierung und Durchführung von Datenimporten wird ein vertieftes Verständnis von Datenstrukturen für bibliografische Referenzdaten erreicht.

Dieses Tutorial enthält die praktische Aufgabenstellung für das Modul DIS 07 „Informationserschließung und strukturierte Dokumentbeschreibung“ als Basis für ein Selbststudium der Lehrinhalte.

Das Inhaltsverzeichnis auf der folgenden Seite enthält die Teilaufgaben des Programms, die in der genannten Reihenfolge zu bearbeiten sind.

Inhaltsverzeichnis

1 Einrichten der bibliografischen Datenbank „literatur.dbm“ (Midos 6) 4
- 1.1 Arbeitsumgebung einrichten . . . 4
- 1.2 Datenbank „literatur.dbm“ öffnen . . . 4
- 1.3 Lektüre . . . 4
- 1.4 Datenbank kennenlernen . . . 4
- 1.5 Arbeitsergebnis . . . 5

2 Erfassen der formalen Dokumentdaten für die zu ergänzenden Dokumente (Midos 6) 5
- 2.1 Lektüre . . . 5
- 2.2 Dokumenterfassung . . . 6
- 2.3 Arbeitsergebnis . . . 6

3 Inhaltliche Erschließung durch Wortlisten und/oder aspektorientierte Thesauri (Midos 6, Midos-Thesaurus) 6
- 3.1 Inhaltliche Erschließung . . . 7
- 3.2 Arbeitsergebnis . . . 8

4 Automatische Schlagwortvergabe als interaktive, dokumentbezogene Vorgehensweise (Midos 6) 8
- 4.1 Lektüre . . . 8
- 4.2 Automatische Schlagwortvergabe . . . 8
- 4.3 Analyse der Ergebnisse . . . 8
- 4.4 Arbeitsergebnis . . . 9

5 Import von Fremddaten („fremddaten.bib“); Anpassung der Datenstruktur durch globales Suchen und Ersetzen; Harmonisierung der Datenbank (Notepad++, Midos 6) 9
- 5.1 Lektüre . . . 9
- 5.2 Bearbeiten der Fremddaten . . . 9
- 5.3 Import der Fremddaten . . . 10
- 5.4 Arbeitsergebnis . . . 11

6 Erstellen eines Ausgabeformates zur Anzeige der Dokumentbeschreibungen am Bildschirm und innerhalb einer Bibliografie (Midos 6) 11
- 6.1 Lektüre . . . 11
- 6.2 Erstellen eines Ausgabeformats . . . 11
- 6.3 Arbeitsergebnis . . . 12

7 Erstellen einer Bibliografie mit formalen und sachlichen Registern (Midos 6) 12
7.1 Lektüre 12
7.2 Erstellen einer Bibliografie 12
7.3 Arbeitsergebnis 13

1 Einrichten der bibliografischen Datenbank „literatur.dbm“ (Midos 6)

Für die Arbeit an der praktischen Aufgabenstellung wird eine bereits eingerichtete Datenbank mit bibliografischen Referenzdaten zur Verfügung gestellt. Diese ist herunterzuladen, mit *Midos* einzurichten und vorzubereiten für die nachfolgenden Aufgaben.

Arbeitsaufwand: 4 h

1.1 Arbeitsumgebung einrichten

Laden Sie den Ordner „gln-daten.zip“ herunter. Entpacken Sie die Archivdatei. Speichern Sie den Ordner in einem Verzeichnis, dessen **vollständiger Pfadname keine Leerzeichen und keine Sonderzeichen** enthält.

1.2 Datenbank „literatur.dbm“ öffnen

Starten Sie *Midos* und öffnen Sie über „Datenbank – Öffnen“ die Datenbank „literatur.dbm“ im Verzeichnis „gln-daten/literatur“.

1.3 Lektüre

Lesen Sie zur Einführung in die Thematik die Abschnitte 3.1 und 3.2 des Kapitels 3 des Buchs „Informationserschließung und Automatisches Indexieren“.

Beachten Sie: Das im Kapitel 3.1 beschriebene Einrichten zusätzlicher Kategorien ist nicht nötig; diese sind bereits Bestandteil der Datenbeschreibung der Datenbank „literatur.dbm“. Die im Text genannten Thesauri sind wie beschrieben mit der Datenbank zu verknüpfen, alle Datensätze von „literatur.dbm“ sind allerdings bereits erschlossen.

1.4 Datenbank kennenlernen

Machen Sie sich mit der Datenbank und ihrer *Datenbeschreibung* vertraut. Untersuchen Sie, welche unterschiedlichen *Dokumenttypen* in der Datenbank vorhanden sind und welche Felder der Datenbeschreibung für diese jeweils besetzt sind (vgl. Folie 20 des Vorlesungsskriptes). Die Art und Weise der Erfassung der Daten für bibliografische Objekte erfolgt in direkter Abhängigkeit vom jeweiligen Dokumenttyp des Objekts.

Erstellen Sie (mindestens) für die Felder „Dokumenttyp“, „Sprache“, „Form“ und „Sparte/Anwendungsfeld“ eine *Wortliste* im *Datenbankeditor*.

Verknüpfen Sie die vier Thesauri „deskr.mth“, „objekt.mth“, „wissfach.mth“ und „geo.mth“ mit der Datenbank. Kopieren Sie dazu die vier Thesaurusdateien aus dem Verzeichnis „gln-daten/thesauri“ in das Verzeichnis der Datenbank „gln-daten/literatur“. Aktivieren Sie die Thesaurusverknüpfung im *Midos*-Datenbankeditor über den Dialog „Optionen – Thesaurus“ (vgl. Folie 45 des Vorlesungsskriptes).

1.5 Arbeitsergebnis

Die fertig eingerichtete Datenbank „literatur.dbm“, deren Datenbeschreibung bzw. Struktur bekannt ist. Die Datenbank ist mit vier Thesauri verknüpft. Die in der Datenbank enthaltenen unterschiedlichen Dokumenttypen sind bekannt.

2 Erfassen der formalen Dokumentdaten für die zu ergänzenden Dokumente (Midos 6)

Die Datenbank „literatur.dbm“ soll um selbst erfasste Dokumente ergänzt werden. Dazu werden insgesamt 17 PDF-Dateien zur Verfügung gestellt, die Informationen über die zu erfassenden Dokumente enthalten. Wichtiger Grundsatz für die Erfassung der Dokumente ist die Herstellung einer einheitlichen Dokumentkollektion, d. h. die Erfassung soll so erfolgen, dass die Gesamtkollektion homogen bleibt. Dazu ist die genaue Berücksichtigung der in der Datenbank verwendeten Erfassungsprinzipien nötig (vgl. Abschnitt 1). Da diese dokumenttypisch sind, sind die zu erfassenden Vorlagen den Dokumenttypen der Datenbank zuzuordnen. Die Retrieval-Qualität einer Datenbank mit bibliografischen Referenzdaten hängt dabei entscheidend von der einheitlichen, sorgfältigen und korrekten Erfassungspraxis ab.

Arbeitsaufwand: 16 h

2.1 Lektüre

Lesen Sie zur Einführung in die Thematik die Abschnitte 3.5 und 3.6.1 des Kapitels 3.

2.2 Dokumenterfassung

Erfassen Sie für die 17 PDF-Dateien im Verzeichnis „gln-daten/titelseiten“ bibliografische Datensätze in der Datenbank „literatur.dbm“. Orientieren Sie sich dabei an den bereits vorhandenen 100 Dokumenten der Datenbank. Verwenden Sie zur Unterstützung die Datenbank „Literatur zur Informationserschließung“[1] sowie die Folien 2-20 des Vorlesungsskriptes.

Beachten Sie: Es gibt zwei PDF-Vorlagen, für die jeweils mehr als ein Datensatz zu erfassen ist. Bei der Vorlage „isko-2008.pdf“ handelt es sich um einen Tagungsband (ein sog. „Sammelwerk“), in dem zwei enthaltene Beiträge/Aufsätze markiert sind; hier sind sowohl der Tagungsband also auch die beiden Aufsätze als Dokumente zu erfassen. Bei der Vorlage „pellegrini-2006“ handelt es sich ebenfalls um ein Sammelwerk; hier ist analog zu verfahren. Den 17 Vorlagen entsprechen daher 21 Datensätze der Datenbank.

Die Vorlagen enthalten alle für die Erfassung benötigten Daten; ein Ermitteln zusätzlicher Daten ist nicht verlangt. Falls auf den Vorlagen Informationen enthalten sind, die als Zusammenfassung bzw. Abstract verstanden werden können, sind diese ebenfalls zu erfassen.

2.3 Arbeitsergebnis

Eine Datenbank „literatur.dbm“ mit insgesamt 121 Datensätzen.

3 Inhaltliche Erschließung durch Wortlisten und/oder aspektorientierte Thesauri (Midos 6, Midos-Thesaurus)

Die selbst erstellten Dokumentbeschreibungen sollen inhaltlich erschlossen werden. Dafür wird ein aspektdifferenziertes Erschließungskonzept mit vier Thesauri und mehreren Wortlisten verwendet.

Arbeitsaufwand: 8 h

[1] Die Datenbank ist durch ein Passwort geschützt: user: „midos“, pw: „retrieval“; alternativer Zugang: Literatur zur Informationserschließung (neu)

3.1 Inhaltliche Erschließung

Erschließen Sie die 21 selbst erfassten Datensätze mit den vier bereits mit der Datenbank verknüpften Thesauri „deskr.mth“, „objekt.mth“, „wissfach.mth“ und „geo.mth“. Orientieren Sie sich dabei an dem in den Folien 35-48 des Vorlesungsskriptes dargestellten Erschließungskonzept.

Beachten Sie: Jedem Dokument ist mindestens ein *Allgemeiner Sachdeskriptor* aus dem Thesaurus „deskr.mth“ zuzuweisen. Teilen Sie die Deskriptoren nach dem Prinzip des koextensiven Erschließens zu, d. h. die Spezifität der zugeteilten Deskriptoren sollte der Spezifität des Dokumenteninhalts entsprechen. Alle weiteren Erschließungsmerkmale werden nur dann zugeteilt, wenn es einen **inhaltlichen Bezug** des Dokuments zu diesen gibt. Insgesamt sind folgende Felder der inhaltlichen Erschließung zu berücksichtigen:

- *Allgemeiner Sachdeskriptor*: beschreibt den Inhalt des Dokuments mit allgemeinen Sachdeskriptoren, z. B. „Information retrieval“ oder „Formalerschließung“; die Deskriptoren stammen aus dem Thesaurus „deskr.mth“;
- *Objekt*: wird nur berücksichtigt, wenn der Inhalt des Dokuments eine Beziehung zu Produkten, Verfahren oder Regelwerken aufweist, die einen Eigennamen haben, z. B. „Google“ oder „DDC“; die Deskriptoren stammen aus dem Thesaurus „objekt.mth“;
- *Wissenschaftsfach*: wird nur berücksichtigt, wenn der Inhalt des Dokuments eine Beziehung zu einem Wissenschaftsfach aufweist, z. B. „Mathematik“; die fachliche Herkunft des Dokuments ist dabei ohne Belang; die Deskriptoren stammen aus dem Thesaurus „wissfach.mth“;
- *Geografika*: wird nur berücksichtigt, wenn der Inhalt des Dokuments eine Beziehung zu einem Land oder Ort hat, sich mit diesem explizit beschäftigt, z. B „Ungarn“; Tagungsorte oder Verlagsorte sind ohne Belang; die Deskriptoren stammen aus dem Thesaurus „geo.mth“;
- *Behandelte Form*: wird nur berücksichtigt, wenn der Inhalt des Dokuments eine Beziehung zu einer Medienform hat, sich explizit mit einer solchen beschäftigt, z. B. „Videos“; die Erschließungsmerkmale stammen aus der Wortliste „Behandelte Form“;
- *Sparte/Anwendungsfeld*: wird nur berücksichtigt, wenn der Inhalt des Dokuments eine Beziehung zu einem Anwendungsfeld oder einer Berufssparte hat, z. B. „Bibliotheken“; die Erschließungsmerkmale stammen aus der Wortliste „Sparte/Anwendungsfeld“.

3.2 Arbeitsergebnis

Eine Datenbank „literatur.dbm“ mit insgesamt 121 inhaltlich erschlossenen Datensätzen.

4 Automatische Schlagwortvergabe als interaktive, dokumentbezogene Vorgehensweise (Midos 6)

Die 21 selbst erstellten Datensätze sollen mit der *Midos*-Funktion „Automatische Schlagwortvergabe“ automatisch erschlossen werden.

Arbeitsaufwand: 2 h

4.1 Lektüre

Lesen Sie zur Einführung in die Thematik Abschnitt 3.4 des Kapitels 3.

4.2 Automatische Schlagwortvergabe

Führen Sie für die 21 selbst erstellten Datensätze eine „Automatische Schlagwortvergabe“ im *Datenbankeditor* durch. Verwenden Sie für die Funktion die vorbereitete *Positivliste* „auto-ws.wtx“ aus dem Verzeichnis „gln-daten/wortlisten“. Orientieren Sie sich bei der Vorgehensweise an der Beschreibung in Abschnitt 3.4 des Kapitels 3 und den Folien 49-52 des Vorlesungsskriptes.

Beachten Sie: Die Leistungsfähigkeit der „Automatischen Schlagwortvergabe“ hängt ab von den im Einstellungsdialog festgelegten „Suchfeldern“. Hier sollten ausschließlich Felder ausgewählt werden, die einen *inhaltlichen* Bezug zum Dokument, z. B. „Titel“ oder „Abstract“, haben.

4.3 Analyse der Ergebnisse

Untersuchen Sie für ausgewählte Schlagwörter die Gründe für ihre Erzeugung. Überlegen Sie, welche Terme im Dokument zu den automatisch erzeugten Schlagwörtern geführt haben und welche Rolle dabei jeweils die Synonymliste „auto-sw.txt“ und die Funktionalität der Automatischen Schlagwortvergabe gespielt haben.

4.4 Arbeitsergebnis

Eine Datenbank „literatur.dbm“ mit 121 intellektuell und automatisch erschlossenen Datensätzen.

5 Import von Fremddaten („fremddaten.bib“); Anpassung der Datenstruktur durch globales Suchen und Ersetzen; Harmonisierung der Datenbank (Notepad++, Midos 6)

Die Datenbank soll durch den Import von 400 bibliografischen Datensätzen aus einem Fremdformat erweitert werden. Dafür ist es nötig, die Fremddaten mit einem Texteditor (*Notepad++*) zu bearbeiten und an die Datenbankstruktur von „literatur.dbm“ anzupassen.

Arbeitsaufwand: 16 h

5.1 Lektüre

Lesen Sie zur Einführung in die Thematik Abschnitt 3.8 bis Abschnitt 3.8.5 des Kapitels 3.

5.2 Bearbeiten der Fremddaten

Öffnen Sie die Datei „fremddaten.bib“ aus dem Verzeichnis „gln-daten/fremddaten“ mit dem Texteditor *Notepad++*. Analysieren Sie das *Midos-Speicherformat* für die Datensätze der Datenbank „literatur.dbm“. Wandeln Sie die Datei „fremddaten.bib“ durch *Suche-und-Ersetze*-Operationen in *Notepad++* in eine Datei im *Midos-Speicherformat* um. Orientieren Sie sich bei den benötigten *Suche-und-Ersetze*-Operationen an der Darstellung in den Folien 53-70 des Vorlesungsskriptes.

Beachten Sie: Die Umwandlung der Datei „fremddaten.bib“ in das *Midos-Speicherformat* erfordert mit *Notepad++* zahlreiche Schritte, inkl. der Verwendung von teilweise komplexeren *RegExp*-Befehlen. Es ist daher zweckmäßig, neue Ersetzungsbefehle zunächst zu probieren, bevor sie auf die gesamte Datei angewendet werden. Das Ergebnis jeder Ersetzung sollte gründlich geprüft werden. Erfolgreiche Operationen sollten zwischengespeichert werden.

Durch Funktionsänderungen in *Notepad++* sind einige der *Suche-und-Ersetze*-Operationen inzwischen einfacher als im Buch beschrieben. Die Folien des Vorlesungsskriptes zeigen die neuere, vereinfachte Variante.

Midos verarbeitet für die Datenbank „literatur.dbm“ nur Daten im sog. *ANSI*-Zeichensatz. Fremddaten im *UTF-8*-Zeichensatz müssen (spätestens) vor dem Import in *Midos* in den *ANSI*-Zeichensatz umgewandelt werden. Die Umwandlung ist beispielsweise mit *Notepad++* über die Funktion „Konvertierung – Konvertiere zu ANSI“ möglich.

Die Datei „fremddaten.bib“ liegt in einem modifizierten *BibTex*-Format vor. Es gibt einige Felder, die nicht Bestandteil der Datenbeschreibung von „literatur.dbm“ sind. Diese Felder sollen ebenfalls importiert werden. Die Datenbeschreibung ist um diese Felder zu ergänzen.

Für zeilenübergreifende *Suche-und-Ersetze*-Operationen ist es hilfreich, auch die normalerweise nicht sichtbaren Sonderzeichen am Zeilenumbruch zu sehen. In *Notepad++* ist dies über „Ansicht – Nicht druckbare Zeichen – Zeilenende anzeigen“ möglich.

Zeilenübergreifende *Suche-und-Ersetze*-Operationen lassen sich in *Notepad++* über *Regular Expressions* mit den Zeichen „\r“ für „return“ und „\n“ für „newline“ realisieren.

Das *Midos-Speicherformat* beendet einen Datensatz jeweils mit drei „&&&“. Vor dem Import müssen diese unbedingt eingefügt werden.

Beachten Sie: Die Datei „fremddaten.bib“ enthält Felder und Daten, die in der Datenbank „literatur.dbm“ nicht vorhanden sind. Untersuchen Sie diese Inhalte. Falls es sich um Daten der inhaltlichen Erschließung handelt, sollen sie importiert werden. Dazu ist die Datenbeschreibung von „literatur.dbm“ um die entsprechenden Felder zu erweitern.

5.3 Import der Fremddaten

Importieren Sie die umgewandelte Datei „fremddaten.bib“ in die Datenbank „literatur.dbm“.

Beachten Sie: Die in das *Midos-Speicherformat* umgewandelte Datei „fremddaten.bib“ muss nicht im eigentlichen Sinne „importiert“ werden, weil sie bereits dem Internformat von *Midos* entspricht. Wenn Sie die Datei „fremddaten.bib“ umbenennen in „fremddaten.dbm“ und diese Datei in das Verzeichnis der Datenbank „literatur.dbm“ kopieren, können Sie die Fremddatei über „Datenbank – Öffnen“ in *Midos* als neue Datenbank öffnen.

Nach der erfolgten Überprüfung des Inhalts der Datenbank „fremddaten.dbm“ lässt sich diese über die Funktion „Datenbank – Datei ergänzen mit“ um die Datensätze

der Datenbank „literatur.dbm“ ergänzen (oder umgekehrt: „literatur.dbm“ ergänzen mit „fremddaten.dbm“).

5.4 Arbeitsergebnis

Eine Datenbank „literatur.dbm“ mit 521 Datensätzen.

6 Erstellen eines Ausgabeformates zur Anzeige der Dokumentbeschreibungen am Bildschirm und innerhalb einer Bibliografie (Midos 6)

Für die Datensätze der Datenbank soll ein Ausgabeformat erstellt werden, das für Referenzen in einem Literaturverzeichnis geeignet ist.

Arbeitsaufwand: 10 h

6.1 Lektüre

Lesen Sie zur Einführung in die Thematik den Abschnitt 3.3 des Kapitels 3.

6.2 Erstellen eines Ausgabeformats

Erstellen Sie ein *Midos*-Ausgabeformat mit folgenden Eigenschaften:

- grundsätzliche Eignung für Referenzen in einem Literaturverzeichnis einer wissenschaftlichen Arbeit;
- Darstellung aller für Referenzen in einem Literaturverzeichnis benötigten formalen Daten, entweder in einem standardisierten Format (z. B. *APA*, *Harvard*, *DIN 1505*) oder im montierten Ausgabeformat der Datenbank Literatur zur Informationserschließung;
- Darstellung der inhaltlichen Erschließungsdaten für alle Dokumente, d. h.:
  - alle Thesaurus-Kategorien,
  - alle importierten Inhaltserschließungsfelder der Fremddaten,
  - Automatische Schlagwörter für die selbst erstellten Dokumentbeschreibungen,
  - und vorhandene Abstracts.

Orientieren Sie sich bei der Vorgehensweise an der Beschreibung in Abschnitt 3.3 des Kapitels 3 und der Darstellung in den Folien 71-83 des Vorlesungsskriptes.

Beachten Sie: Neben dem montierten Ausgabeformat mit den o. g. Eigenschaften muss es eine kategorisierte Vollanzeige geben, die **alle** Felder der Datenbeschreibung ausgibt.

6.3 Arbeitsergebnis

Zwei Ausgabeformate für die Datensätze der Datenbank „literatur.dbm“: eine kategorisierte Vollanzeige mit allen Feldern der Datenbeschreibung, ein montiertes Ausgabeformat mit Eignung für Referenzen in einem Literaturverzeichnis.

7 Erstellen einer Bibliografie mit formalen und sachlichen Registern (Midos 6)

Für die Datensätze der Datenbank soll eine Bibliografie erstellt werden, die alle Datensätze in einem montierten Ausgabeformat (vgl. Abschnitt 6) enthält und Suchmöglichkeiten über Register anbietet.

Arbeitsaufwand: 4 h

7.1 Lektüre

Lesen Sie zur Einführung in die Thematik den Abschnitt 3.11 des Kapitels 3.

7.2 Erstellen einer Bibliografie

Erstellen Sie eine Bibliografie mit *Midos* im *rtf*-Format mit folgenden Eigenschaften:

- Ausgabe aller Datensätze der Datenbank im selbst erstellten Ausgabeformat (vgl. Abschnitt 6) mit einheitlicher und durchgehender Sortierung der Referenzen im Primärteil;
- mindestens drei Register für:
  - alle Deskriptoren (Allgemeine Sachdeskriptoren, Objekte, Geografika);
  - Automatische Schlagwörter;
  - Personen;

Öffnen Sie die Bibliografie im *rtf*-Format mit *Word* oder einer anderen Textverarbeitung und speichern sie im *pdf*-Format ab.

Orientieren Sie sich bei der Vorgehensweise an der Beschreibung in Abschnitt 3.11 des Kapitels 3 und der Darstellung in den Folien 85-90 des Vorlesungsskriptes.

7.3 Arbeitsergebnis

Eine Bibliografie im *pdf*-Format mit einem Primärteil, der durchgängig sortiert alle Datensätze der Datenbank „literatur.dbm“ im montierten Ausgabeformat enthält (vgl. 6) und Registern, die eine Suche in der Bibliografie ermöglichen.