Sandra Hamm, Kurt Schneider

Automatische Erschließung von Universitätsdissertationen

Neue Perspektiven für die Formalerschließung in der Deutschen Nationalbibliothek

Ziel: Unterstützung und Beschleunigung der Formalerschließung

Seit über 40 Jahren arbeiten Bibliothekarinnen und Bibliothekare an der Automatisierung von Erschließungsprozessen. Von Anfang an war dabei ein zentrales Ziel, die Arbeit der formalen Erschließung zu unterstützen und zu beschleunigen, um eventuell vorhandene Bearbeitungsrückstände schneller abbauen oder eine Vielzahl bislang unzureichend erschlossener Bestandsgruppen überhaupt erst erschließen zu können[1]. Trotz vielfältiger Forschungsansätze und praktischer Versuche sind bis heute jedoch nur wenige regelbasierte Systeme im Einsatz, die Bibliotheken bei der Erstellung einfacher Titelaufnahmen erfolgreich unterstützen[2].

Automatische Katalogisierung gedruckter Hochschulschriften

Seit Mai 2014 setzt die Deutsche Nationalbibliothek erstmals ein automatisches Verfahren zur Erschließung gedruckter Monografien im Produktivbetrieb ein. Dabei werden die wichtigsten bibliografischen Informationen aus digitalisierten Titelseiten von Hochschulschriften, von denen die Deutsche Nationalbibliothek in den letzten Jahren im Durchschnitt rund 14.000 Exemplare pro Jahr erhalten hat und die über eine weitgehend einheitliche Struktur verfügen, extrahiert und automatisch in die entsprechenden Felder der Titelaufnahme übernommen. Die Bearbeiterinnen und Bearbeiter bewerten diese Veränderung im Arbeitsprozess als große Unterstützung; die Bearbeitungszeit je Publikation konnte verkürzt werden.

Motivation

Seit Jahren verzeichnet die Deutsche Nationalbibliothek einen stark wachsenden Zugang an Publikationen, 2013 waren es insgesamt rund eine Million Medieneinheiten, davon rund 370.000 Netzpublikationen[3]. Trotz des stetigen Anstiegs insbesondere im Bereich der digitalen Medien ist ein Rückgang bei den Printpublikationen nicht in Sicht. Um den wachsenden Zugang bei begrenzten Personalressourcen weiterhin bewältigen zu können, sollen automatische Verfahren insbesondere im Erschließungskontext verstärkt eingesetzt werden. Dabei liegt der Fokus nicht nur auf den digitalen Medien; auch für gedruckt vorliegende Publikationen sollen neue Erschließungsmethoden erprobt und implementiert werden, damit die Bibliothek auch in Zukunft ihren gesetzlichen Auftrag anforderungsgerecht erfüllen kann[4].

Erfüllung des gesetzlichen Auftrags trotz steigender Zugangszahlen

Kooperation

Kooperation mit Universität Innsbruck

In Kooperation mit der Universität Innsbruck hat die Deutsche Nationalbibliothek das Pilotprojekt »Halbautomatische Formalerschließung von Universitätsdissertationen« initiiert. Der Grundgedanke dabei war, die im Rahmen des EU-Projektes IMPACT in Innsbruck entwickelte Software zur Strukturerkennung[5] in einer weiterentwickelten Version in den seit 2008 von der Deutschen Nationalbibliothek betriebenen Workflow zur Kataloganreicherung einzubinden. In dessen Kontext werden die Inhaltsverzeichnisse aller neu eingehenden Bücher laufend digitalisiert. Der Arbeitsablauf beim Scannen sollte nahezu unverändert bleiben: außer den Inhaltsverzeichnissen sollten lediglich die Titelseiten von Hochschulschriften zusätzlich digitalisiert werden.

Software

Strukturerkennung für Titelseiten

Die von der Universität Innsbruck, Gruppe Digitalisierung und Elektronische Archivierung, entwickelte Software »Functional Extension Parser« beziehungsweise »Title Page Parser« analysiert logische Strukturen in Dokumenten[6]. Sie ist modular auf-

gebaut. Basis sind XML-Dateien mit Koordinaten- und Styleinformationen, die im Rahmen der Digitalisierung von z. B. Titelseiten generiert werden. Auf diese extrahierten OCR-XML-Fakten können verschiedene Regelsets angewendet werden.

Extraktion bibliografischer Daten

Im Rahmen des Projektes wurde ein bereits vorhandenes Regelset speziell für Titelseiten von Dissertationen weiterentwickelt, um folgende bibliografische Angaben automatisch erkennen zu können:

- Verfasser
- Titel
- Ort, Hochschule, Dissertation/Habilitationsschrift, Promotionsjahr (Hochschulschriftenvermerk)
- Erscheinungsjahr
- Format

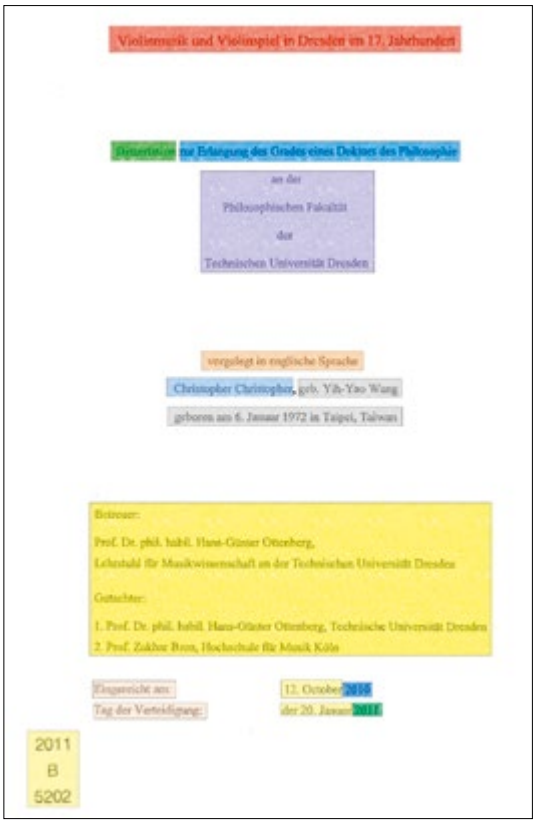
Regelerstellung

Grundlage für die Regelerstellung bildeten dabei Muster, Regelmäßigkeiten und Zuordnungsmerkmale innerhalb der Titelseiten einer repräsentativen Auswahl von mehreren hundert Hochschulschriften, wie z. B.

- einleitende Wendungen, die auf den Verfasser deuten (»vorgelegt von«, »presented by« u. Ä.),
- Attribute für den Titel wie u. a. Position und Schrift,
- Indizien für Promotions- und Erscheinungsjahr,
- nicht auszuwertende, aber häufig vorkommende Phrasen (»zur Erlangung des Grades«, »aus dem Institut/Fachbereich«)

sowie diverse sonstige Regeldefinitionen.

Darüber hinaus wurden als Wörterbuch alle deutschen Hochschulen mit Promotionsrecht, die Personendatensätze aus der Gemeinsamen Normdatei (GND)[7] sowie eine Zusammenstellung potenziell vorkommender akademischer Grade in der Datenbank des Title Page Parsers hinterlegt.



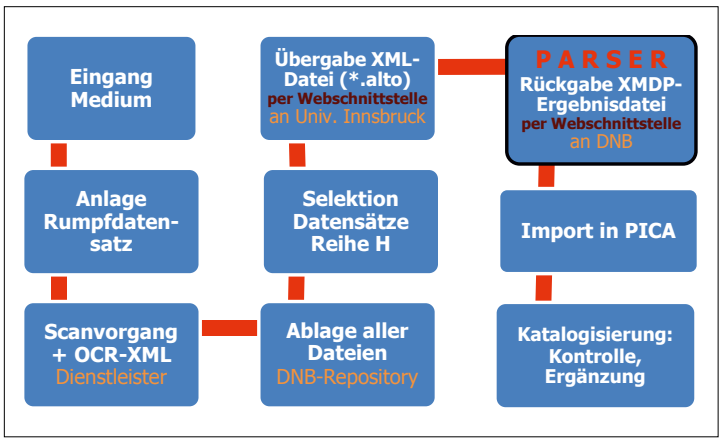
Beispiel für Strukturanalyse (Quelle: Universität Innsbruck)

Pretest

Im Rahmen eines Pretests wurden zunächst eintausend digitalisierte Titelseiten manuell annotiert (siehe Abbildung). Die dadurch entstandene »Ground Truth« – das gewünschte Extraktionsresultat – wurde danach für den Vergleich mit den durch den Parser automatisch erkannten Metadaten herangezogen. Mittels dieser Evaluationsmethode konnte für die oben genannten bibliografischen Angaben eine durchschnittliche Erkennungsquote von 92 Prozent F-Measure (Soll-Ist-Vergleich) ermittelt werden.

Aufgrund der sehr zufriedenstellenden Testergebnisse sollte die Strukturerkennungssoftware auch im Produktivbetrieb der Deutschen Nationalbibliothek für Publikationen der Bibliografiereihe H (Hochschulschriften) erprobt werden. Der Title Page Parser wurde dabei im Remote-Betrieb in Innsbruck betrieben; der Austausch der Daten mit der Deutschen Nationalbibliothek erfolgte via Webschnittstelle.

Workflow



Workflow

Anlegen von Rumpfdaten-sätzen

Nach Eingang der Hochschulschrift findet deren Akzessionierung durch Mitarbeiterinnen und Mitarbeiter der Deutschen Nationalbibliothek statt. Hierbei wird ein sogenannter Rumpfdatensatz mit voreingestellter Reihe H-Codierung angelegt, der außer dem Sprachcode lediglich die Akzessionsnummer zur Identifizierung der Publikation im Geschäftsgang beinhaltet.

Dieser Arbeitsschritt erfolgt stapelweise und dauert weniger als eine Minute je Hochschulschrift.

```
Set 1 | Setgröße 128 | Datensatz 1 | PPN 1058561480 | Format DA

Eingabe: 1249:22-09-14 Änderung: 1249:22-09-14 09:55:26 Status:
1249:22-09-14
0500 Aaa
0599 14-09-22 : f
0600 rh
0701 /a/F-2014-125424#2
1500 /1ger
2240 DNB:1058561480
4715 =u $=c 04=d DNB=e 1
[0292 ] frankfurt dnb <101b>
7001 22-09-14 : a
7800 299904490
7900 22-09-14 09:55:26.000
8100 F-2014-125424
```

Rumpfdatensatz in PICA: nur Reihe H-Codierung (0600) und Akzessionsnummer (0701/8100), zzgl. Sprachcode (1500)

Digitalisierung der Titelseiten

Im Anschluss an die Akzessionierung digitalisiert ein Scandienstleister außer den Inhaltsverzeichnissen aller eingehenden Monografien bei Hochschulschriften auch die Titelseiten und stellt neben anderen Ausgabeformaten die dabei produzierten OCR-Ergebnisse auch im ALTO-XML-Format[8] bereit.

Die Dateien werden via OAI jede Nacht vom Dienstleisterserver eingesammelt und im Repository der Deutschen Nationalbibliothek gespeichert. Durch programmbasierte Selektion erfolgt die Übergabe der ALTO-XML-Dateien von Hochschulschriften über eine Webschnittstelle an die Universität Innsbruck.

Ergebnis-übernahme in Katalogisie-rungsmaske

Nach Anwendung des Title Page Parsers werden die Ergebnisse, d. h. die je Titelseite extrahierten bibliografischen Informationen, ebenfalls via Webschnittstelle als XMetaDissPlus-Files[9] an die Deutsche Nationalbibliothek zurückgeliefert und hier durch Konversion direkt in die entsprechenden PICA-Felder der jeweiligen Katalogisate übernommen.

Alle diese Prozesse nach dem Scanvorgang erfolgen vollautomatisch.

Link zu diesem Datensatz	[illegible]
Titel/Bezeichnung	[illegible]
Person(en)	[illegible]
Erscheinungsjahr	[illegible]
Umfang/Format	[illegible]
Hochschulschrift	[illegible]
Sprache(n)	[illegible]
Weiterführende Informationen	[illegible]
Frankfurt	[illegible]
Leipzig	[illegible]

Kataloganzeige maschinell gewonnener Formaldaten (die Titelseite ist via Link »Inhaltsverzeichnis« zugreifbar)

Die maschinell erstellte Titelaufnahme wird unmittelbar nach Übermittlung der Daten aus Innsbruck im Katalog der Deutschen Nationalbibliothek angezeigt, auch wenn sie in Teilen noch unvollständig oder eventuell fehlerbehaftet ist. Durch die Verlinkung der gescannten Titelseite und des Inhaltsverzeichnisses können Katalogbenutzer weitere detaillierte Informationen zur Publikation direkt abrufen. Insbesondere bei Titeln mit mathematischen oder chemischen Formeln, die in einer Titelaufnahme generell nicht vorlagengemäß dargestellt werden können, erweist sich der unmittelbare Zugriff auf die digitalisierte Titelseite von Vorteil.

Nachbear-beitung per Autopsie

In einem abschließenden Bearbeitungsschritt werden die automatisch erzeugten Titeldaten durch Katalogisiererinnen und Katalogisierer überprüft, gegebenenfalls korrigiert, um noch fehlende Angaben ergänzt (Seitenzahl, Illustrationen, Normda-

tenverknüpfung) und für die Anzeige in der Deutschen Nationalbibliografie freigegeben.

Ergebnisse

In der Zeit von Mai bis Juli 2014 wurden mit dem beschriebenen Verfahren im Produktivbetrieb der Deutschen Nationalbibliothek rund 2.500 Hochschulschriften bearbeitet und zusätzlich die Ergebnisse der automatischen Erschließung von rund 500 Hochschulschriften im Rahmen der Projektevaluierung einer detaillierten Prüfung unterzogen. In die Evaluation nicht einbezogen wurden einige wenige Dissertationen, deren bibliografische Informationen nicht vollständig auf der Haupttitelseite, sondern auf mehreren Seiten verteilt zu finden waren.

Hohe Erkennungsquoten

Die Erkennungsquote für Verfasserangaben lag bei 85,5 Prozent, für die Titelerkennung bei 83 Prozent. Dabei ist zu berücksichtigen, dass auch dann positiv bewertet wurde, wenn außer den korrekten Verfasser- und Titelinformationen noch überschüssige Textteile wie z. B. die Berufsbezeichnung im Namen oder Lehrstuhlangaben im Titel mit übernommen wurden. Dadurch sind zwar Löschoperationen im Rahmen der per Autopsie erfolgenden Nachbearbeitung erforderlich, diese sind hier jedoch weitaus weniger aufwendig zu realisieren als die manuelle Eingabe von Titelinformationen im Rahmen der konventionellen Katalogisierung.

Die automatische Erkennung des Promotionsjahres sowie die Unterscheidung zwischen Dissertationen und Habilitationsschriften erfolgte mit über 98 Prozent nahezu fehlerfrei. Die Erkennungsquote für die Hochschule inklusive Ortsangabe lag bei 76,7 Prozent. Das Erscheinungsjahr wurde in 94,6 Prozent der überprüften Publikationen korrekt erkannt.

Die durchschnittliche Erkennungsquote bezogen auf alle ausgewerteten Felder betrug insgesamt 89,3 Prozent und lag damit im Produktivbetrieb nur geringfügig niedriger als im vorab durchgeführten Testszenario.

Bis zu ein Drittel weniger Bearbeitungszeit

Die Bearbeitungsdauer je Hochschulschrift konnte insgesamt reduziert werden. Bei Vorliegen guter Extraktionsergebnisse kann für die Erfassung der Formaldaten im Vergleich zur konventionellen Katalogisierung bis zu ein Drittel der Bearbeitungszeit eingespart werden. Infolgedessen stieg im betrachteten Zeitraum die monatliche Bearbeitungsmenge um rund 25 Prozent an. Auch die Mitarbeiterinnen und Mitarbeiter im Katalogisierungsteam, deren Haltung zum neuen Verfahren vor Projektbeginn zum Teil noch etwas skeptisch war, beurteilten die neue Bearbeitungsweise und den damit erreichten höheren Durchsatz insgesamt als sehr positiv. Insbesondere bei sehr langen und komplizierten Titeln, die gerade im Hochschulschriftenbereich oft vorkommen, wurde die Bereitstellung automatisch generierter Titelinformationen als unmittelbar spürbare Arbeitserleichterung erfahren.

Spürbare Arbeitserleichterung

Fazit

Die Unterstützung der Formalkatalogisierung und die Erhöhung des Katalogisierungsdurchsatzes sind durch Verfahren der automatischen Erschließung nicht nur möglich, sondern gerade für die Bearbeitung großer Mengen relativ gleichartig gestalteter Publikationen auch sinnvoll und hilfreich.

Erfolgreicher Einsatz des Verfahrens

Der im Rahmen des Projektes entwickelte Workflow hat sich in der Praxis bewährt.

Der eingesetzte Parser birgt noch Optimierungspotenzial, beispielsweise sind die Ergebnisse der oben erwähnten Ground Truth bisher noch nicht in die Regeldefinition eingeflossen (u. a. Reihenfolgewahrscheinlichkeit der Angaben), was zu weiteren Verbesserungen etwa durch maschinelles Lernen führen könnte; auch darf vermutet werden, dass sich auf der Basis größerer Datenmengen die Ergebnisse noch verbessern ließen. Darüber hinaus könnte die Integration zusätzlicher Wörterbücher oder der Einbau linguistischer Analysekomponenten zur Steigerung der Erkennungsquoten beitragen.

Eine Weiterentwicklung des Workflows und der eingesetzten Software könnte vor allem dann lohnend sein, wenn es gelingen würde, das hier zunächst für Hochschulschriften eingesetzte Verfahren auch auf andere Publikationstypen mit hoher Erscheinungsdichte zu übertragen, sofern keine qualifizierten Fremddaten aus Quellen Dritter zur Verfügung stehen. Denkbar sind vorrangig außerhalb des Verlagsbuchhandels erscheinende Printpublikationen wie z. B. Kongressschriften und Forschungsberichte,

Übertragung auf weitere Publikationstypen geplant

aber auch die stark wachsende und heute schon sehr hohe Zahl frei zugänglicher Netzpublikationen, die sich noch nicht im Bestand der Deutschen Nationalbibliothek befinden. Will man diese objektspezifisch erschließen, sind alleine schon aus Ressourcengründen andere als automatisch unterstützte Erschließungsverfahren kaum vorstellbar. Erste Schritte zur Adaption der erprobten Methoden sind daher in Vorbereitung.

Anmerkungen

1 Weibel, Stuart; Oskins, Michael; Vizine-Goetz, Diane: Automated title page cataloging : a feasibility study, in: Information Processing and Management, 25 (1989) 2, pp. 187–203
Davies, Roy: Expert systems and cataloguing, in: The application of expert systems in libraries and information centres. – London : Bowker-Saur, 1992, S. 133–166
De Silva, Sharon M.: A review of expert systems in library and information science, in: Malaysian Journal of Library and Information Science, 2 (1997) 2, S. 57–92

2 Eine Ausnahme ist das von der Bonner Firma ImageWare Components GmbH angebotene Paket zur teilautomatisierten Erschließung von Zeitschriften und Periodika: <http://www.imageware.de/de/Loesungen_SW/C-3/>

3 Jahresbericht der Deutschen Nationalbibliothek 2013, S. 40: <http://files.dnb.de/jahresbericht2013>
(Die Daten für 2014 lagen bei Redaktionsschluss noch nicht vor.)

4 Strategische Prioritäten 2013–2016 der Deutschen Nationalbibliothek, S. 8: <http://d-nb.info/1050432266/34>

5 Functional Extension Parser, auch Title Page Parser:
<http://www.impact-project.eu/uploads/media/IMPACT_D-EE_4.3_Functional_Extension_Parser.pdf>

6 Die Gruppe Digitalisierung und Elektronische Archivierung (DEA) des Instituts für Germanistik an der Universität Innsbruck gewann im Jahr 2013 den INEX-Wettbewerb für die automatisierte Erkennung struktureller Metadaten, s. a. Doucet, Antoine [u. a.]: Overview of the ICDAR 2013 competition on book structure extraction, in: Proceedings of the Twelfth International Conference on Document Analysis and Recognition (ICDAR 2013), S. 1438–1443

7 Gemeinsame Normdatei (GND): <http://www.dnb.de/gnd>

8 »ALTO« steht für »Analyzed Layout und Text Object«, ein XML-Schema zur Beschreibung von Layout und Inhalt digitalisierter Textquellen: <http://www.loc.gov/standards/alto/>

9 Bei »XMetaDissPlus« (XMDP) handelt es sich um einen Metadatenstandard zur Ablieferung von Hochschulschriften: <http://www.dnb.de/DE/Standardisierung/Metadaten/xMetadissPlus.html>