



## Texterfassung und OCR

Das Thema Texterfassung taucht regelmäßig dann auf, wenn ältere Dokumente wieder veröffentlicht werden sollen.

Folgende Szenarien sind denkbar:

- es gibt nur noch ein Buch oder einen Ausdruck
- die Daten der Voraufgabe existieren zwar noch, sind aber für die Weiterverarbeitung nicht brauchbar. Zum Beispiel könnten ausschließlich Postscript Daten vorhanden sein.
- ein Autor liefert ein handschriftliches Manuskript
- posthume Veröffentlichungen von Vorlesungen

Um Ihnen die Umsetzung solcher Projekte zu ermöglichen, haben wir verschiedene Varianten erarbeitet.

Diese reichen von

- einfachem Abschreiben
- über Doppelerfassung und anschließendem Vergleich
- bis zum Scannen und zur Erfassung mit OCR

Welche Variante für Sie die Beste ist, lässt sich pauschal nicht sagen. Hier ist immer ein Test notwendig.

Diese Tests führen wir sehr kurzfristig durch, damit Sie schnell entscheiden können, ob dieses Projekt realisiert werden kann.

### Was ist OCR?

OCR steht für *Optical character recognition*, zu deutsch Optische Zeichenerkennung. Hierbei wird in der Regel ein Text gescannt und dieser dann durch ein spezielles Programm verarbeitet. Als Ergebnis wird meist eine MS-Word-Datei gewünscht, es sind aber auch andere Datei-Formate möglich. Sprechen Sie uns diesbezüglich an.

### Problematische Texte

Als problematisch für die OCR-Erfassung können folgende Aspekte gesehen werden:

- mathematische und chemische Formeln
- Frakturschrift
- sehr holzhaltiges Papier
- ...

### Beispiel für eine Erfassung mit OCR

Original als Grafik

Im Taieri-Fluß in Neuseeland wurden, auf geschichtete Weise durch randomisierte Wahl von Wasserläufen ähnlicher Dimension, drei Nebenflüsse (acht Teileinzugsgebiete) mit 198 Standorten ausgewählt. Besondere Sorgfalt wurde darauf verwendet, nicht der Versuchung zu erliegen, Standorte mit einfachem Zugang zu wählen (in der Nähe von Straßen oder Brücken), da eine solche Wahl möglicherweise zu einem systematischen statistischen Fehler der Resultate geführt hätte. Die Standorte wurden nach vier Gesichtspunkten des Fischbesatzes klassifiziert: (1) keine Fische, (2) nur *Galaxias*, (3) nur Forellen oder (4) sowohl *Galaxias* als auch Forellen vorhanden. Für jeden Standort wurde eine Reihe von physikalischen Variablen gemessen (Wassertiefe, Fließgeschwindigkeit, Phosphorkonzentration im fließenden Wasser, prozentualer Anteil des Bachbettes, der aus Kieselsteinen besteht etc.). Ein statistisches Verfahren, die *mul-*

**Die Populations-  
ebene – die  
Europäische  
Forelle und  
die Verbreitung  
einheimischer  
Fische**

umgewandelt in Text

Im Taieri-Fluß in Neuseeland wurden, auf geschichtete Weise durch randomisierte Wahl von Wasserläufen ähnlicher Dimension, drei Nebenflüsse (acht Teileinzugsgebiete) mit 198 Standorten ausgewählt. Besondere Sorgfalt wurde darauf verwendet, nicht der Versuchung zu erliegen, Standorte mit einfachem Zugang zu wählen (in der Nähe von Straßen oder Brücken), da eine solche Wahl möglicherweise zu einem systematischen statistischen Fehler der Resultate geführt hätte. Die Standorte wurden nach vier Gesichtspunkten des Fischbesatzes klassifiziert: (1) keine Fische, (2) nur *Galaxias*, (3) nur Forellen oder (4) sowohl *Galaxias* als auch Forellen vorhanden. Für jeden Standort wurde eine Reihe von physikalischen Variablen gemessen (Wassertiefe, Fließgeschwindigkeit, Phosphorkonzentration im fließenden Wasser, prozentualer Anteil des Bachbettes, der aus Kieseln besteht etc.). Ein statistisches Verfahren, die *multiple Diskriminanzanalyse*, wurde dann benutzt, um festzustellen, ob und in welchen physikalischen Parametern sich die verschiedenen Standorttypen voneinander unterscheiden. Mittelwerte und Standardfehler dieser Schlüsselparameter des Lebensraumes sind in Tabelle 1.1 wiedergegeben.

Die Populations-  
ebene – die  
Europäische  
Forelle und  
die Verbreitung  
einheimischer  
Fische

Forellen traten fast immer unterhalb von Wasserfällen auf, die groß genug waren, um eine stromaufwärts gerichtete Wanderung zu verhindern. Sie waren vorwiegend in niedrigen Höhenlagen verbreitet, weil sich jene Standorte, die keine Wasserfälle stromabwärts hatten, fast ausschließlich in niedrigen Höhenstufen fanden. Standorte, an denen *Galaxias* vorkam (oder gar keine Fische), lagen immer stromaufwärts von einem oder mehreren großen Wasserfällen. Die

**Tabelle 1.1** Mittelwerte und Standardfehler (in Klammern) für wichtige Diskriminanz-Variablen, die Standorttypen (An- bzw. Abwesenheit von Fischarten) an 198 Standorten des Taieri-Flusses bestimmten (aus Townsend & Crowl, 1991)

Standorttyp	Anzahl der Standorte	Anzahl der Wasserfälle flußabwärts	Variable*	
			Meereshöhe (Meter über dem Meeresspiegel)	Prozentualer Anteil des aus Kieseln bestehenden Flußbettes
Kein Fisch	54	4,37 (0,64)	339 (31)	15,8 (2,3)
	74	0,42 (0,05)	224 (28)	48,9 (2,4)

-----  
Da-TeX - Gerd Blumenstein | Erich-Zeigner-Allee 69-73 | D-04229 Leipzig

Telefon: +49 (0) 341 9 26 12 77 | Fax: + 49 (0) 341 9 26 12 78