



Web Genres

Marcel Kamlot



Genre

- Gruppierung anhand von lose definierten Kriterien
- **Film** (Drama, Thriller), **Musik** (Rap, Klassik), **Buch** (Krimi, Sachbuch), **Reden** (Ankündigung, Dankesrede)...
- Charakteristika:
 - **Inhalt**
 - **Form**
 - **Funktion**
- Text-Genres: nicht einheitlich standardisiert
 - Brown corpus: 15 Kategorien
 - British National Corpus: 70 Klassen

ID	File	Genre
A16	ca16	news
B02	cb02	editorial
C17	cc17	reviews
D12	cd12	religion
E36	ce36	hobbies
F25	cf25	lore
G22	cg22	belles_lettres
H15	ch15	government
J17	cj19	learned
K04	ck04	fiction
L13	cl13	mystery
M01	cm01	science_fiction
N14	cn15	adventure
P12	cp12	romance
R06	cr06	humor



Inhalt

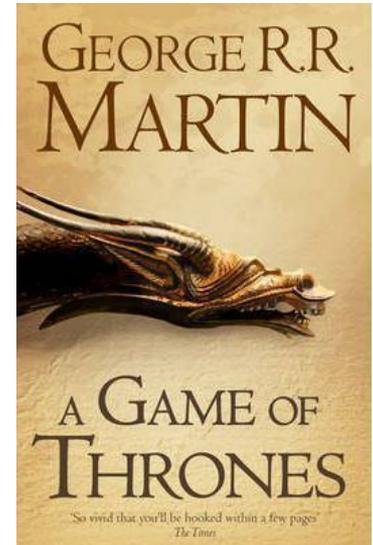
- Fiktion
- Geschichte
- Personen
- Handlung
- Orte
-

Form

- Sätze
- Viele Seiten
- Kapitel
- Vorwort, Prolog, Epilog...
- Handlungsbogen
- ...

Funktion

- Unterhaltung
- ...



Roman



Inhalt

- Wörter
- Definitionen
- Regeln
- ...

Form

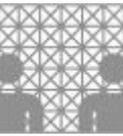
- [Wort] - [Erklärung]
- ...

Funktion

- Lehre
- Wörter definieren
- ...



Wörter- buch



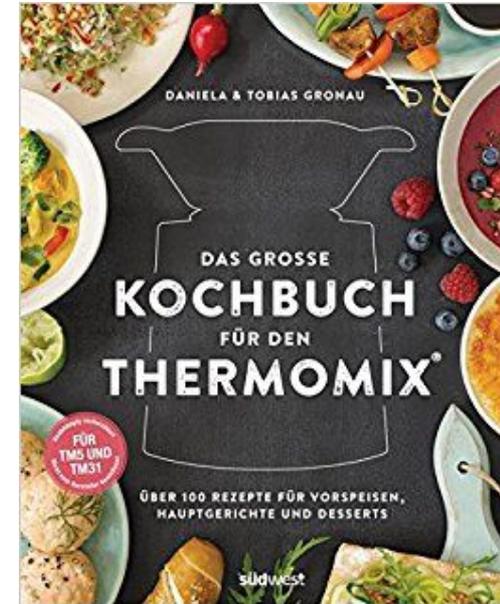
Inhalt



Form



Funktion



Koch-
buch



Leseritdel eatunirobe anumqua quepliuia

Illquat bitundae dolorum estio nequis idiore molorempore, in remvo lupta conse ximporis

Agnam acceritis sit quam ipsid mo- sam is dolore nisi expe ga plabor aut- rendus nos sequeae modelon et, nis alia pro derewelic idender icsosed quat doloria plabipiwet quereere sa. Sequius etio et labge ga ide ma quo- magratio besturmagus etewee, nis veremo quatinis nestibus manctat a doluptatem que pligula cusam. Ut uterem que dolo. Optama sita nian.

Estio idemest il ex doluptam que volupta volorest at laer. Conactis parum quis nihil exere magnibi laborio te ped . Eos explaut emals doloretore ni doloremp teptatur as. Cupictur meclissi si queuo blab- oratiuiv a collutatus. Ipsandae qui- ae nim rerla veri. Evendae molur sum is elecac. Harionecus et lab- kif quam, vilaces ut reicagewee.



Schmerzmittel



Zwischenzeit - 20 Min.



100 kcal, 42 g KH, 10 g Ei, 10 g Fett



Schmerzmittel



Zwischenzeit - 20 Min.



100 kcal, 42 g KH, 10 g Ei, 10 g Fett



Ceratemeniper estoriatis conoparu

Gaet as digende litios sita sit additionserum nimpore invellar



Schmerzmittel



Zwischenzeit - 20 Min.



100 kcal, 42 g KH, 10 g Ei, 10 g Fett

FÜR 2 PERSONEN

100 g Kollenchinfrucht



Gianononus nemeteaciet kuwia

Doluptas destentisio dolorit dolut exelcil lor aborro toremuisir



Schmerzmittel



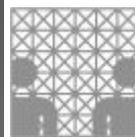
Zwischenzeit - 20 Min.



100 kcal, 42 g KH, 10 g Ei, 10 g Fett

FÜR 3 PERSONEN

100 g Kollenchinfrucht

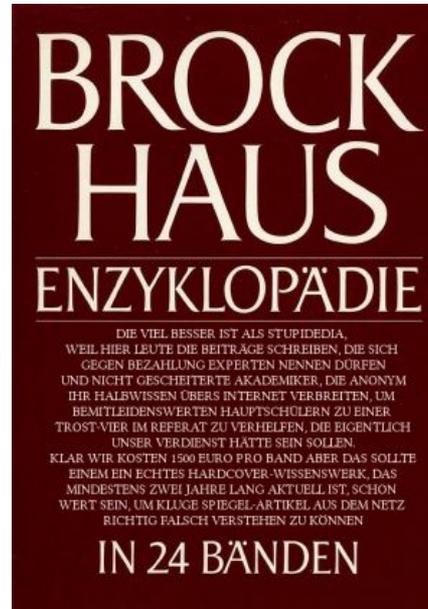
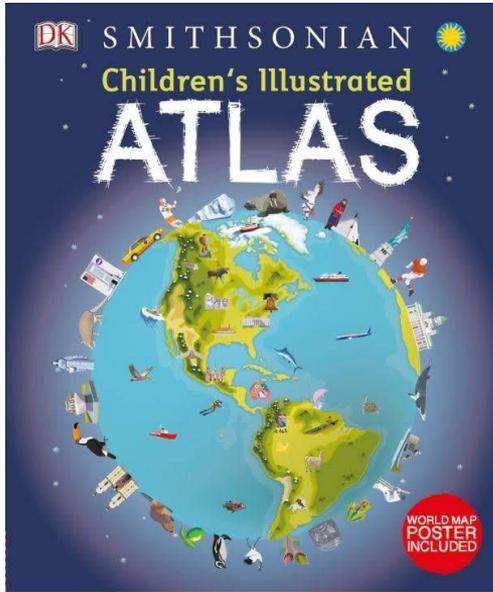




Inhalt

Form

Funktion





Genres im Web

- Oft angelehnt an Papier-Formen
 - Zeitungen und Magazine
 - Erleichtern Eingewöhnung
 - Schnelleres Finden
 - Bessere Orientierung
- Innovative Informations-Strukturen jetzt möglich
- Web-Genres: Informationsräume, die keine Papier-Äquivalente haben, und gemeinsame Form, Eigenschaften und Organisation haben (A. Dillon and B. Gushrowski, 2000)
 - Persönliche Homepage



Web Genres

- Beispiele:
 - Persönliche Homepage
 - FAQ
 - Blog
 - Suchmaschine
- Anwendung:
 - Informationsgewinnung
 - Ausfiltern irrelevanter Dokumente
 - Informationsextraktion
 - Identifizieren relevanter Teile
 - Metadaten automatisch extrahieren
 - Suche in bestimmten Teilen



Genre Tagging

Beispiel: Genre Tagging, Rehm et al. 2008

- 50 Webseiten, 7 **menschliche** Annotatoren, **keine** definierten Labels
- Ergebnis: Viele verschiedene Labels
-

Three Pollutants and an Emission: A Playbill for the Multipollutant Legislative Debate

Consistency: High

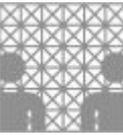
- Participant 1: News **article**
- Participant 2: **Article**/commentary
- Participant 3: **Article**
- Participant 4: Feature
- Participant 5: A newsletter **article**
- Participant 6: News **article**
- Participant 7: Journalistic

Save your settings:

- ▶ Create a Brookings Account
- ▶ If you already have an account, please login
- + My Topics Edit
- + My Programs Edit
- + My Experts Edit
- + Saved Pages Edit

Related Topics

- ▶ Environment



Automatisches Tagging

- Features:
 - Häufige Wörter, Funktionswörter
 - Zeichensetzung
 - POS-Trigramme
 - Links
 - Seitenstruktur, HTML
 - Wortlisten



Anwendung

- Inhalt extrahieren
- Spam-Identifizierung
- Suchfilter
- Email-Sortierung



Danke!
Fragen?