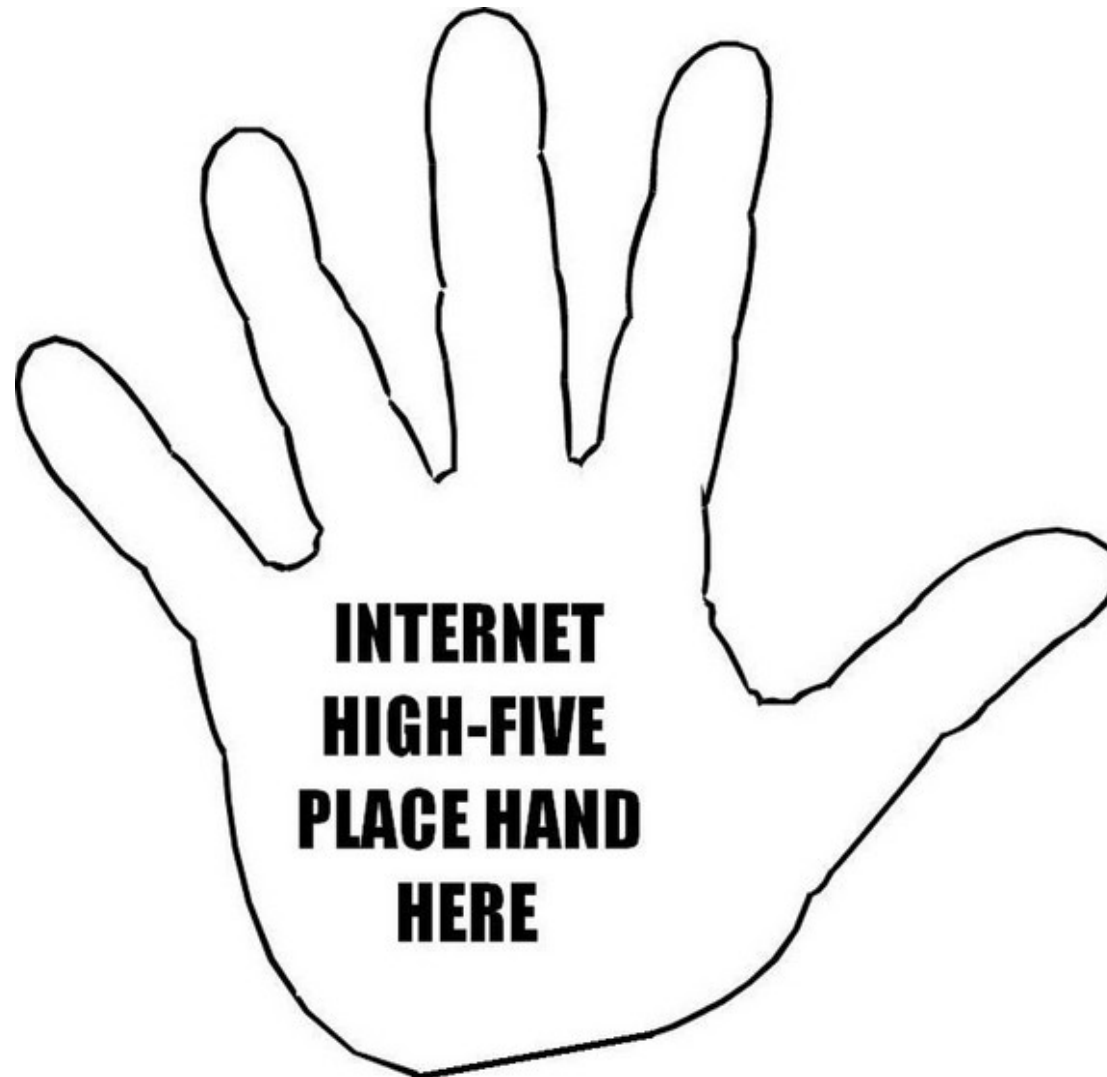


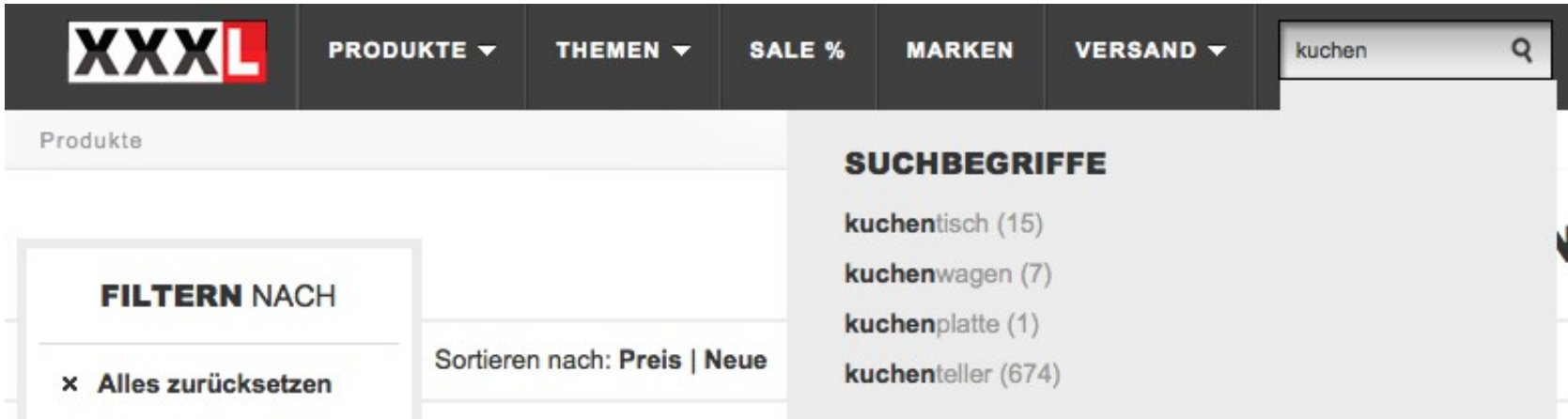
Drupal Camp Essen 2013

Deutsche und mehrsprachige Volltextsuche  
mit Apache Solr

Markus Kalkbrenner (mkalkbrenner)



# Noch besser mit Autovervollständigung



The screenshot shows the top navigation bar of an e-commerce site. The navigation bar includes the 'XXXL' logo, menu items for 'PRODUKTE', 'THEMEN', 'SALE %', 'MARKEN', and 'VERSAND', and a search bar containing the text 'kuchen'. Below the search bar, a dropdown menu is open, displaying search suggestions under the heading 'SUCHBEGRIFFE'. The suggestions are: 'kuchentisch (15)', 'kuchenwagen (7)', 'kuchenplatte (1)', and 'kuchenteller (674)'. To the left of the search bar, there is a 'Produkte' section with a 'FILTERN NACH' button, a 'Sortieren nach: Preis | Neue' dropdown, and a 'x Alles zurücksetzen' button.

Autocomplete von „kuchen“: Kuchenplatte, Kuchenteller, ...

PREISFRAGE: Was ist ein „Kuchenwagen“?

# Kuchenvagen = Küchenwaage! WTF?

XXXL PRODUKTE THEMEN SALE % MARKEN VERSAND Ich suche... (0) (0) (0)

Produkte

## 7 TREFFER FÜR „KUCHENWAGEN“

Sortieren nach: Preis | Neue

Produkte 1 bis 7

### FILTERN NACH

× Alles zurücksetzen

#### SUCHE

kuchenvagen

LOS

#### KATEGORIE

Küchen, Essen, Haushalt (7)

#### FARBE

Weiß (4)

Grau/Silber (2)

Rot (1)

#### PREIS

10 €

bis

40 €

#### MARKE

LEIFHEIT (1)



KÜCHENWAAGE

€ 19,99

Sale %



KÜCHENWAAGE

€ 34,99



KÜCHENWAAGE

€ 9,99



KÜCHENWAAGE

€ 24,95



KÜCHENWAAGE



KÜCHENWAAGE



KÜCHENWAAGE

# 22240 Kartoffel Rezepte vs. 7738 Kartoffeln Rezepte

**kochbar** powered by  Newsletter bestellen

Rezepte Videos Themen Shopping Forum

## 22240 Kartoffel Rezepte

Suche:  

**Ähnliche Rezepte:**

- Kartoffel Gratin
- Kartoffel Auflauf
- Überbackene Kartoffel
- Zucchini Kartoffel
- Kartoffel Ei
- Kartoffel Gemüse Auflauf
- Kartoffel Zucchini Auflauf
- Kartoffel Lauch Suppe

**MIT diesen Zutaten:**

- Creme Fraiche
- Käse Überbacken
- Kräuterquark
- Joghurt
- Kräutersoße
- Quark

**OHNE diese Zutaten:**

- Fleisch
- Ei Fleisch
- Vorkochen

**Kategorie:**

- Als Hauptspeise
- Glutenfrei
- Ohne Weizen
- Zum Abendessen
- Zum Mittagessen

1 2 3 4 ... »

Bild	Name	Wertung	Aufrufe seit	Dauer	Schwierigkeit	eingestellt von
	Pflanzenkartoffel	★★★★★ 14 Stimmen 5,00	1647 seit 25.02.2011	40 Min.	leicht	asya
	KARTOFFEL-ZUCCHINI SUPPE	★★★★★ 11 Stimmen 4,82	9712 seit 15.08.2010	45 Min.	leicht	Wolfgang49
	Kartoffel, Sellerie, Apfel Roesti	★★★★★ 4 Stimmen 5,00	2559 seit 09.05.2009	30 Min.	leicht	Meikew

**kochbar** powered by  Newsletter bestellen

Rezepte Videos Themen Shopping Forum

## 7738 Kartoffeln Rezepte

Suche:  

**Ähnliche Rezepte:**

- Rosmarin Kartoffeln
- Gefüllte Kartoffeln
- Paprika Kartoffeln
- Senf Kartoffeln
- Kartoffeln Möhren Eintopf
- Gebackene Kartoffeln
- Kretanische Kartoffeln
- Knoblauch Kartoffeln

**MIT diesen Zutaten:**

- Quark
- Möhren Untereinander
- Blumenkohl
- Thunfisch
- Räucherlachs
- Hering

**OHNE diese Zutaten:**

- Ei
- Fleisch
- Fett
- Folie Grillen
- Ei Vegetarisch
- Zucker

**Kategorie:**

- Als Hauptspeise
- Glutenfrei
- Laktosefrei
- Ohne Weizen
- Zum Mittagessen

1 2 3 4 ... »

Bild	Name	Wertung	Aufrufe seit	Dauer	Schwierigkeit	eingestellt von
	Lachsfilet mit Dillsauce, grünen Bohnen und Kartoffeln	★★★★★ 13 Stimmen 5,00	2366 seit 20.01.2012	k.A.	leicht	MausVoh
	Spargeln mit Soffritto-Kartoffeln und Mayo-Mousseline	★★★★★ 16 Stimmen 5,00	666 seit 17.04.2011	100 Min.	leicht	marcos
	Mediterrane Kartoffeln mit Parmesan Lachs	★★★★★ 23 Stimmen 5,00	652 seit 12.08.2012	k.A.	leicht	Pastapabst

# 97 Erdäpfel Rezepte



The screenshot shows the 'kochbar' website interface. At the top, there is a navigation bar with 'Rezepte', 'Videos', 'Themen', 'Shopping', and 'Forum'. Below this, a search bar contains 'erdäpfel'. To the right of the search bar, there are filters for 'Ähnliche Rezepte' (Similar Recipes) including 'Erdäpfel Gurkensalat', 'Erdäpfel Gulasch', 'Süsse Erdäpfel', 'Erdäpfel Suppe', 'Schloß Erdäpfel', 'Erdäpfel Auflauf', 'Eingebrannte Erdäpfel', and 'Erdäpfel Gemüse'. Below the search bar, there are sections for 'MIT diesen Zutaten:' (With these ingredients) and 'OHNE diese Zutaten:' (Without these ingredients). The 'MIT diesen Zutaten:' section has a search box and a checkbox for 'Gurken'. The 'OHNE diese Zutaten:' section has a search box and checkboxes for 'Als Hauptspeise', 'Glutenfrei', 'Laktosefrei', 'Ohne Weizen', and 'Zum Mittagessen'. Below these sections, there is a pagination bar showing '1 2 3 4 ... > >'. The main content area displays a table of search results with columns for 'Bild', 'Name', 'Wertung', 'Aufrufe seit', 'Dauer', 'Schwierigkeit', and 'eingestellt von'.

Bild	Name	Wertung	Aufrufe seit	Dauer	Schwierigkeit	eingestellt von
	Dinkel-Erdäpfel Brot	★★★★★ 5 Stimmen 5,00	904 seit 23.08.2010	k.A.	leicht + Zutaten	Kolibris
	Schweineschnitzel Wiener Art mit steirischem ...	★★★★★ 1 Stimme 5,00	897 seit 17.01.2011	70 Min.	leicht + Zutaten	dppd
	Chilli Kräuterofenkartoffeln	★★★★★ 45 Stimmen 4,91	258 seit 29.01.2011	15 Min.	leicht + Zutaten	moniundpeter

Ich bin gemein ;-)

# “Apache Solr Search Integration“ is awesome!\*

\* ... Wenn man eine englischsprachige Webseite betreibt :-)

Out of the Box erzeugen die Drupalmodule „Apache Solr Search Integration“ und „Search API“ genau derartige Effekte auf nicht englischsprachigen Seiten.

Warum eigentlich?





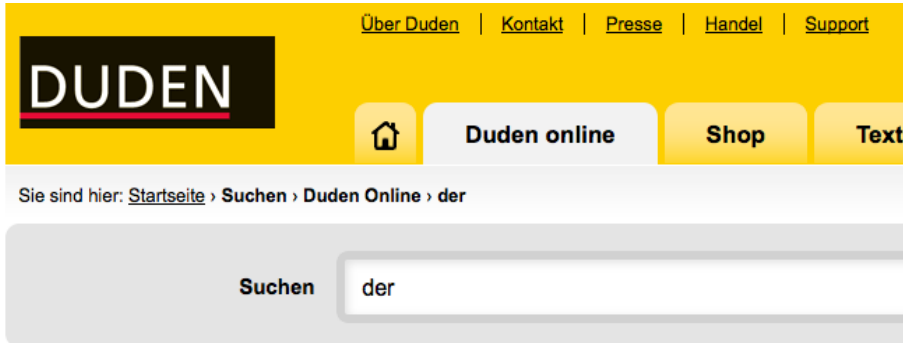
- Stoppwörter (stop words)
- Wortstammbildung (stemming)
- Wortzusammensetzungen (compound words)
- Rechtschreibprüfung (spell checker)
- Synonymlisten (synonyms)
- ...

=> Sämtliche Konfigurationen sind auf englischsprachige Inhalte ausgerichtet!



- Grundsätzlich eine gute Idee:
  - Wörter nicht in den Suchindex aufnehmen, die zu unspezifisch sind.
- Typische Stoppwörter sind z.B. bestimmte Artikel:
  - Englisch: „the“
  - Deutsch: „der“, „die“, „das“
- Stoppwörter unterscheiden sich von Sprache zu Sprache!

# Auch die Domäne ist wichtig!



The screenshot shows the Duden website's search interface. At the top, there is a navigation bar with links for 'Über Duden', 'Kontakt', 'Presse', 'Handel', and 'Support'. The Duden logo is prominently displayed on the left. Below the navigation bar, there are buttons for 'Duden online', 'Shop', and 'Text'. A breadcrumb trail indicates the current location: 'Sie sind hier: Startseite > Suchen > Duden Online > der'. A search bar contains the text 'der' and a 'Suchen' button.

## Suchergebnisse

Die Suche nach "der" lieferte 135019 Treffer.

[Duden online \(133765\)](#)

[Shop \(443\)](#)

[Sprachwissen \(811\)](#)

### der

Pronomen - 1a. in Beispielen wie »ein Stuhl, ...1b. (anstelle eines Genitivattributs) in Beispielen ...2. in Beispielen wie »der sich ...

[Zum vollständigen Artikel klicken Sie hier.](#)

### der

Pronomen - 1a. immer betont; 1b. nicht betont; anstelle eines Genitivattributs; 2a. unmittelbar hinweisend

[Zum vollständigen Artikel klicken Sie hier.](#)

### der

- Die Stoppwortliste hängt nicht nur von der Sprache, sondern auch vom Inhalt und dem Zielpublikum ab, der so genannten Domäne.
- Auf duden.de ist „der“ korrekterweise kein Stoppwort.

# Getrennte Suchindexe für jede Domäne wären besser!

Über Duden | Kontakt | Presse | Handel | Support

# DUDEN

Home Duden online Shop Text


Sie sind hier: [Startseite](#) > [Suchen](#) > [Shop](#) > [der](#)

Suchen


## Suchergebnisse

Die Suche nach "der" lieferte 135019 Treffer.

[Duden online \(133765\)](#) [Shop \(443\)](#) [Sprachwissen \(811\)](#)



**Duden - Lexikon der Familiennamen**  
Mit bekannten Namensträgerinnen und -trägern.  
Sofort lieferbar  
**12.95 €**



**Duden - Lexikon der Vornamen**  
5., völlig neu bearbeitete Auflage  
Sofort lieferbar  
**8.99 €**

Duden - Lexikon der Vornamen

Über Duden | Kontakt | Presse | Handel | Support

# DUDEN

Home Duden online Shop Text

Sie sind hier: [Startseite](#) > [Suchen](#) > [Sprachwissen](#) > [der](#)

Suchen

## Suchergebnisse

Die Suche nach "der" lieferte 135019 Treffer.

[Duden online \(133765\)](#) [Shop \(443\)](#) [Sprachwissen \(811\)](#)

### Sprachratgeber: Der Anlagenvermerk

Für alle, die an einer soliden Anlagenberatung interessiert sind, haben wir die wichtigsten Informationen zusammengestellt. Erfahren Sie, wie **der** Anlagenvermerk in **der** geschäftlichen Korrespondenz gestaltet werden sollte. Kleinere Gegenstände oder schriftliche Unterlagen wie Zeugnisse, Arbeitsproben oder Formulare, die Briefen beigefügt sind, werden Anlagen genannt. Nach den Empfehlungen **der** DIN 5008 steht **der** Anlagenvermerk, also **der** Hinweis, dass ein Schreiben eine Beilage enthält, mit mindestens drei Leerzeilen Abstand unter dem Gruß oder **der** Firmenbezeichnung. Falls **der** Unterzeichner maschinenschriftlich angegeben wird, schließt sich **der** Anlagenvermerk nach einer Leerzeile an. Ist wenig Platz vorhanden,

### Aktuelle Meldung: Journalistenpreis der Hauptstadt der deutschen Sprache

# Mama, stirb?



- Identische Wörter haben in verschiedenen Sprachen unterschiedliche Bedeutungen, z.B. „die“:
  - Deutsch: „die“
  - Englisch: „sterben“, aber auch „Würfel“ oder „Chip“ usw.
  - Niederländisch: „jene“
  - ...

=> Das nennt man „False Friends“.
- Identische „Buchstabenkombinationen“ sind in einer Sprache Stoppwörter, in einer anderen nicht!
- Eine einzelne große sprachübergreifende Stoppwortliste ist nicht möglich!

- Die Reduktion eines Suchbegriffs auf seinen Wortstamm ermöglicht das Finden von passenden Inhalten unabhängig von der Flexion des Suchbegriffs, z.B. Singular oder Plural:
  - Tomate => tomat
  - Tomaten => tomat
- Aber: Die Algorithmen zur Wortstammbildung unterscheiden sich von Sprache zu Sprache!
- Für einige Sprachen gibt es gar keinen Algorithmus!
- Der englische „Default Stemmer“ führt bei allen anderen Sprachen zu Fehlern!

- Algorithmus für deutsche Sprache:
  - Tomate => tomat
  - Tomaten => tomat
  - tomato => tomato
  - tomatoes => tomato
- Algorithmus für englische Sprache:
  - tomato => tomato
  - tomatoes => tomato
  - Tomate => tomat
  - Tomaten => tomaten



- Wortstammbildung hilft nicht immer.
    - Deutsch:  
Kartoffel, Kartoffeln
    - Englisch:  
goose, geese  
mouse, mice
- ⇒ Derartige Pluralformen müssen in sprachspezifische Synonymlisten eingetragen werden!

- In Abhängigkeit von der Domäne der Webseite kann es sinnvoll sein, einzelne Wörter von der Wortstammbildung oder der Anwendung von Stoppwörtern aus zu schließen.
- Produkt- oder Markennamen beinhalten oft einen Plural oder ein Stoppwort:
  - Drupal Gardens
  - Die Ärzte
  - The Who
  - Pittsburgh Steelers

- Typisch für die deutsche Sprache: Wörter aus zusammengesetzten Substantiven, z.B. „Dampfschiffahrt“.
- Je nach Anwendungsfall wünscht man sich einen Suchtreffer bei folgenden Suchbegriffen:
  - Dampf
  - Dampfschiff
  - Schiff
  - Schiffahrt
  - Fahrt
- Solr bietet dafür den CompoundWordFilter. Die Standardkonfiguration der Drupalmodule unterstützt diesen aber nicht!

- Begriff „Rechtschreibprüfung“ ist eigentlich falsch, besser wäre „Wortkatalog“
- Ein Katalog ist die Grundlage für verschiedene Features:
  - „Did you mean ...“
  - Autovervollständigung
- Diese Kataloge sind fraglos sprachabhängig!
- Die Rechtschreibprüfung berücksichtigt Stoppwörter!  
=> Vermutlich ist eine dedizierte Stoppwortliste sinnvoll
- Kataloge können statische Listen sein. Default bei Drupal ist eine dynamische Generierung aus dem Content!  
=> Das kann bei User-generated-Content peinlich werden ;-)

- In der Drupalstandardkonfiguration wird z.B. aus einem „Ä“ ein „A“:
  - Stoppwörter mit Umlauten funktionieren nicht!
  - Die Wortstammbildung verursacht „Fehler“:
    - Kuchen => kuch
    - Küche => kuch
    - Küchen => kuch
- Abschalten des Filters ist keine Lösung, sonst bekommt man andere Probleme, z.B. mit „Crème fraîche“
- Besser: In jeder Sprache die „eigenen“ diakritischen Zeichen gesondert behandeln und die „fremden“ verwerfen.

# War es das jetzt endlich?

- Nein, es ist noch (viel) komplexer:
  - Word Delimiter
  - Length Filter
  - Lower Case
  - Elevator
  - Ranking und Boosts
  - Multisite
  - N-Gram
  - ...
- **Und mehrsprachige Webseiten?**



- Die Solr-Dokumentation oder ein Buch lesen und den Index perfekt auf den eigenen Anwendungsfall hin konfigurieren.
- Jemanden beauftragen, der sich damit auskennt ;-)
- Mithilfe der folgenden Dupalmodule brauchbare Ergebnisse Out-of-the-box erzielen und die wichtigsten Einstellungen innerhalb von Drupal verfügbar haben:
  - Apache Solr Config Generator
    - Inkl. Apache Solr Advanced Settings
    - Inkl. Apache Solr Text Files
  - Apache Solr Multilingual
    - Apache Solr Multilingual Config Generator



- Übrigens: Ohne Apache Solr Multilingual kann die Apache Solr Search Integration Inhalte, die per Entity Translation übersetzt werden, gar nicht verarbeiten!

- drupal.org:
  - mkalkbrenner: <https://drupal.org/user/124705>
  - [https://drupal.org/project/apachesolr\\_multilingual](https://drupal.org/project/apachesolr_multilingual)
  - [https://drupal.org/project/apachesolr\\_confgen](https://drupal.org/project/apachesolr_confgen)
- twitter: <https://twitter.com/mkalkbrenner>
- github: <https://github.com/mkalkbrenner>
- Apache Solr Multilingual in Action:
  - <https://pgsbox.de/>
  - <https://pgsbox.com/>
  - <https://pgsbox.nl/>