Drupal City Berlin 2011

Language-Specific and Multilingual Full-Text Searching

Markus Kalkbrenner (mkalkbrenner)

bio·logis

cocomore

“Apache Solr Search Integration“ is awesome!*

* ... if you're running an English website :-(

What's the Problem?

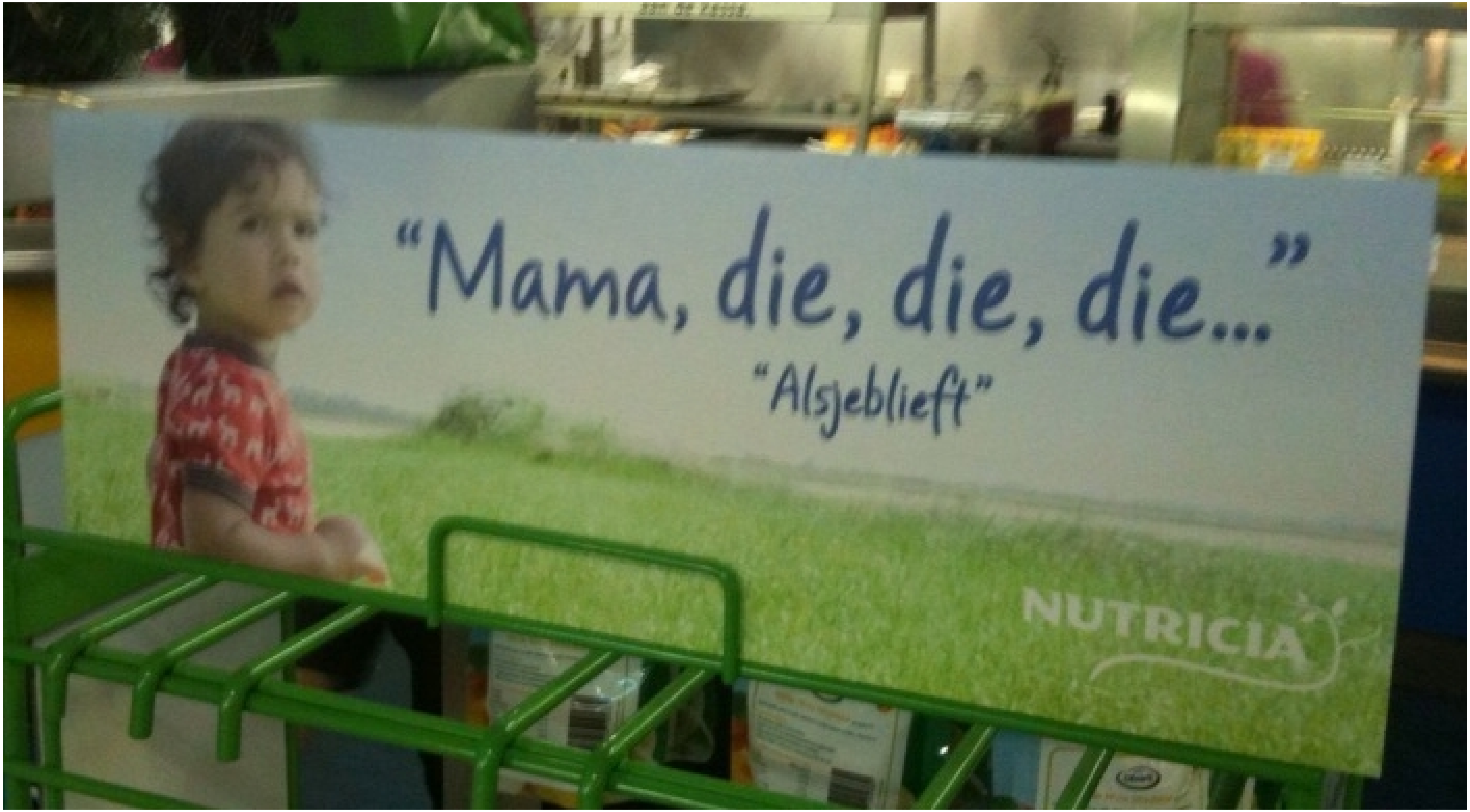
- Stop-words
- Stemming
- Compound words
- Spell checking
- Synonyms
- …

=> everything is preconfigured for English

Stop-words

- Exclude words from your index that are too unspecific
- Definite articles are typical stop-words:
  - English: „the“
  - German: „der“, „die“, „das“
- Stop words have to be language specific
  (False Friends)
- Stop words also depend on the purpose of your site!

Mummy, that one, that one, that one ...



bio·logis

cocomore

Stemming

- Reducing a word to its stem enables the user to find content, independent of the keyword's inflection, e.g. singular or plural
    - `tomato   => tomato`
    - `tomatoes => tomato`
- The stemming algorithm differs from language to language!
- For some languages there's no stemmer!

German Stemming

- English stemming:

```
tomato   => tomato
tomatoes => tomato
Tomate   => tomat
Tomaten  => tomaten
```

- German stemming:

```
Tomate   => tomat
Tomaten  => tomat
tomato   => tomato
tomatoes => tomato
```

Synonyms

- In some cases, stemming does not solve a problem.
  - English:
    `goose, geese`
    `mouse, mice`
  - German:
    `Kartoffel, Kartoffeln`

=> provide language-specific synonyms

Protected Words

- Depending on the purpose of your site, it makes sense to prevent stemming of some words.
- Often product names or brands contain a plural form or stop words:
  - `drupal gardens`
  - `Pittsburgh Steelers`
  - `The Who`
- Protected words might be language specific.

Compound Words

- In German, „Dampfschifffahrt“ should be found if you search for
  - Dampf
  - Dampfschiff
  - Schiff
  - Schifffahrt
- Solr provides a CompoundWordFilter, Drupal Apache Solr Search Integration does not.

Spell Checking

- “Did you mean ...”
- Backend for auto-complete
- No doubt that spell checking should be language-specific.
- Spell checking uses stop-words, too.

ISOLatin1Accent

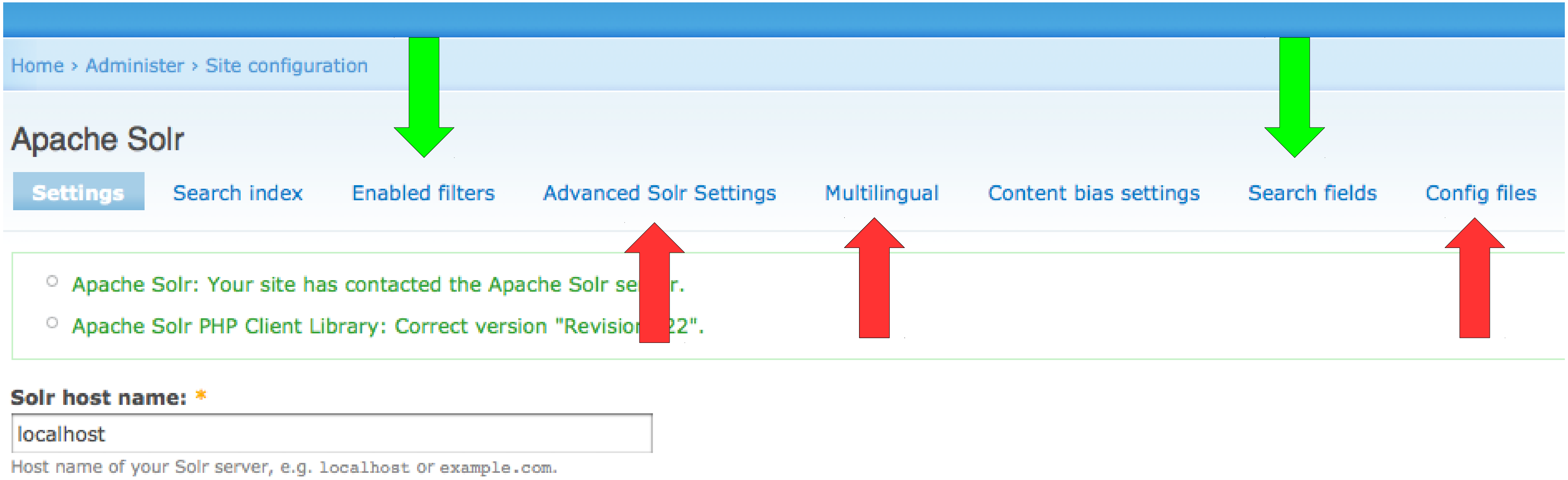
- By default, „Ä“ becomes „A“
- Take this into account when defining your stop words, synonyms, etc.
- German stemmer in combination with ISOLatin1AccentFilter:
  - Kuchen => kuch
  - Küche   => kuch
  - Küchen => kuch

Apache Solr Multilingual Module

- Configures Solr according to your site's language without hacking XML-config-files
- Offers additional advanced configuration options
- Handles multiple languages in one Solr index
- Provides better language facet
- Implements (basic) „CLIR“ - Cross Language Information Retrieval

Reliable Setup for Drupal 6

- Apache Solr Multilingual 6.x-2.0-beta1
- Apache Solr Search Integration 6.x-2.0-beta5



Home › Administer › Site configuration

Apache Solr

- Apache Solr: Your site has contacted the Apache Solr se r.
- Apache Solr PHP Client Library: Correct version "Revision 22".

Solr host name: *

localhost

Host name of your Solr server, e.g. localhost or example.com.

Drupal 7?

- Translation process changed dramatically compared to Drupal 6
- There's more than one way, fields in core, i18n, …
- Field-translation has been “fixed/reverted/removed” from 7.0 to 7.8
- Apache Solr Search Integration, itself, has been a moving target

=> we stopped developing, but now it seems that things stabilize now

Contributing

- Code ;-)
- Default stop word lists, synonym lists, compound word lists
- Language-specific default configurations for all the advanced options not mentioned in this session

Questions?

- Apache Solr Search Integration
  http://drupal.org/project/apachesolr
- Apache Solr Multilingual
  http://drupal.org/project/apachesolr_multilingual
- bio.logis GmbH
  http://bio.logis.de