

Crawling von Datenschutzerklärung-Historien

Zwischenpräsentation

Alexander Prull, Jörn-Henning Daug, Simon Kaleschke

Universität Leipzig

16. Dezember 2016

Gliederung

- 1 Projektbeschreibung
- 2 Lösungsansatz
- 3 Softwarearchitektur
- 4 Ergebnisse

Gliederung

- 1 Projektbeschreibung
- 2 Lösungsansatz
- 3 Softwarearchitektur
- 4 Ergebnisse

Projektbeschreibung

Motivation

- Aktuell geltende DSEs analysieren.
- Die Entwicklungsgeschichte von DSEs betrachten.
- Trends und Veränderungen beobachten.

Aufgaben

- DSEs (täglich) extrahieren.
- Diese geeignet anzeigen.
- Unterschiede über Zeit darstellen.

DSE = Datenschutzerklärung

Gliederung

- 1 Projektbeschreibung
- 2 Lösungsansatz**
- 3 Softwarearchitektur
- 4 Ergebnisse

Lösungsansatz

Aufteilung

- Extraktion: Ruby, XPath, SQLite.
- Backend: Java, REST-Services.
- Frontend: HTML5, CSS3, JavaScript, AngularJS, Bootstrap.

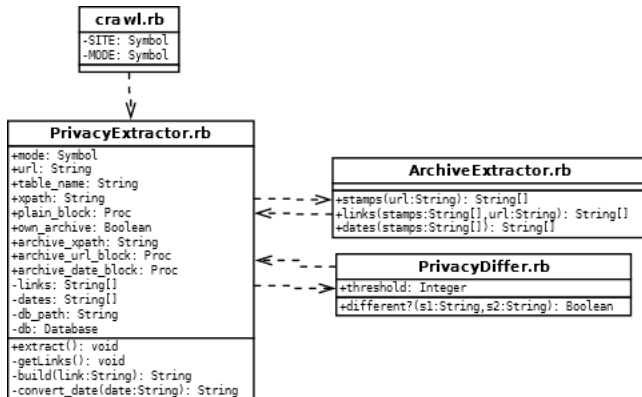
Arbeitspakete

- Recherche (100 %)
- Grundgerüst (100 %)
- Feinschliff (80 %)

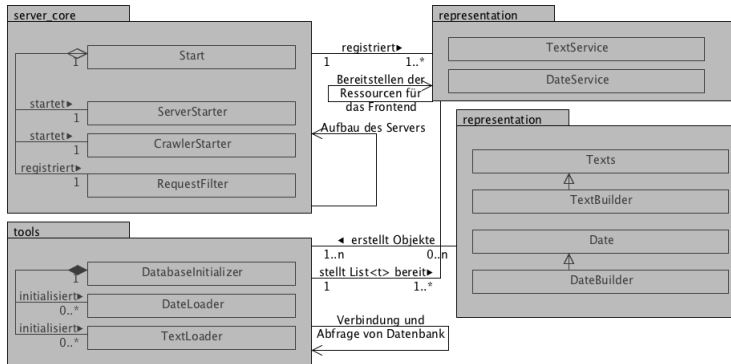
Gliederung

- 1 Projektbeschreibung
- 2 Lösungsansatz
- 3 Softwarearchitektur**
- 4 Ergebnisse

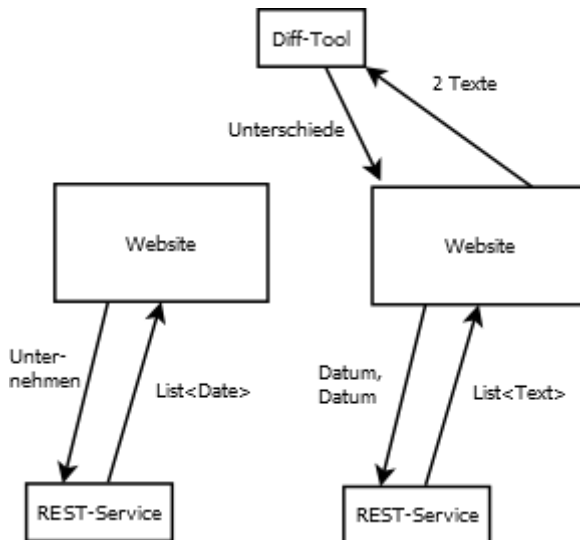
Extraktion



Backend



Frontend



Datenbankstruktur

Die SQLite-Datenbank hat folgende Spalten:

Spalte	Bedeutung
ID	Primärschlüssel.
SYSTEM_DATE	Datum, an dem die Datenschutzhistorie in dem aktuellen Zustand war.
DISPLAY_DATE	Gleiches Datum in benutzerfreundlicher Schreibweise.
LINK	Hyperlink, wo die extrahierte Historie zu finden ist.
CONTENT	Der extrahierte Plaintext.

Gliederung

- 1 Projektbeschreibung
- 2 Lösungsansatz
- 3 Softwarearchitektur
- 4 Ergebnisse**

Übersicht gecrawlte Webseiten

Firma	Archiv	Versionen	Zeitspanne	Qualität
Alternate	✗	7	08/2014 - 07/2016	0.9
Amorelie	✗	11	01/2013 - 10/2016	0.7
Apple	✗	12	09/2014 - 09/2016	1.0
Burgerking	✗	2	02/2015 - 12/2016	0.7
Edeka	✗	4	11/2014 - 10/2016	0.7
Google	✓	22	06/1999 - 01/2017	0.8
Microsoft	✗	7	02/2016 - 01/2017	0.9
Payback	✗	10	09/2011 - 10/2016	0.9
Paypal	✗	2	04/2014 - 11/2016	0.9
RocketbeansTV	✗	4	10/2014 - 09/2016	1.0

Übersicht gecrawlte Webseiten

Firma	Archiv	Versionen	Zeitspanne	Qualität
Steam	✗	3	09/2012 - 11/2016	0.8
Subway	✗	3	05/2016 - 10/2016	0.9
Süddeutsche	✗	1	04/2015 - 04/2015	0.2
Trivago	✗	1	04/2016 - 04/2016	0.7
Twitter	✓	10	05/2007 - 01/2016	0.8
Uni Leipzig	✗	1	06/2013 - 06/2013	0.5
Vine	✗	5	03/2013 - 01/2017	1.0
WhatsApp	✓	2	07/2012 - 01/2017	0.9
Wikimedia	✓	4	06/2006 - 06/2014	1.0
Zalando	✗	3	09/2010 - 11/2012	0.9