Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI

Bundesamt für Kultur BAK
Schweizerische Nationalbibliothek NB

Workshop Webarchiv Schweiz, Schweizerische Nationalbibliothek, 25.11.2015

Harvesting und Qualitätskontrolle

Harvesting Umgebung

Crawler:
Heritrix 3.2.2

Open Wayback 2.1

Webkit:
PhantomJS
CasperJS

SIP ID	Seed URL	Status	VQI	Crawl Start	URL Queue	URL Download	Size	Queued / Downloaded	HTTP Status Codes	Actions
193677	http://www.eurexchange.com	RUNNING		2015-11-12 09:30	148'818	107'554	9.1 GB			Actions
195354	http://cc-lenzerheide.ch	RUNNING		2015-11-17 07:00	354	126	2.0 MB			Actions
195392	http://www.ppur.org	RUNNING		2015-11-17 07:00	1'398	354	9.5 MB			Actions
195393	http://www.mat-ou-brillant.ch	RUNNING		2015-11-17 07:00	450	102	703.2 kB			Actions
195394	http://www.bibliotheca.com	RUNNING		2015-11-17 07:00	388	63	244.9 kB			Actions

Die Crawl Jobs werden über ein Webinterface überwacht. Es können bis zu 20 Harvesting-Aufträge gleichzeitig laufen. Durch das Pausieren eines Crawls kann das Zwischenresultat bereits in der Wayback Machine geprüft und mit Regeln verbessert werden.

Herausforderungen beim Harvesting

- Schnelle technologische Entwicklung im Web versus Tools zur Webarchivierung
- Dynamische Funktionen
- Wayback Probleme
  - Elemente wie z.B. Bilder werden nicht angezeigt
  - Navigation ist nicht bedienbar
- Flash
- Bildergalerien
- Social Media (Facebook, Youtube etc.)

Visual Quality Index

Zur Unterstützung der Qualitätskontrolle hat die NB einen Visual Quality Index implementiert.

Von allen Crawls werden beim Harvesting automatisch zwei Screenshot-Vergleiche erstellt.

1. Der Vergleich der Live Website mit dem aktuellen Crawl
2. Der Vergleich der letzten archivierten Version in Access mit dem aktuellen Crawl

176073	http://www.abbaye-stmaurice.ch	QS	1691 16193
176107	http://www.tma.ethz.ch	QS	788 133
176058	http://www.naturama.ch	QS	5990 21251
176049	http://www.museumaarau.ch	QS	276 1866

Die Übereinstimmung der Screenshots ergibt einen Vergleichswert, der mit den Ampelfarben grün / orange / rot optisch dargestellt wird und die Qualitätskontrolle vereinfacht.

Workshop Archives Web Suisse, Bibliothèque nationale suisse, 25.11.2015

Harvesting et contrôle de qualité

Environnement de harvesting

Crawler:
Heritrix 3.2.2

Open Wayback 2.1

Webkit:
PhantomJS
CasperJS

Les crawl jobs sont surveillés via une interface web. Jusqu'à 20 commandes de harvesting peuvent tourner en même temps. En mettant un crawl en pause, on peut déjà contrôler le résultat intermédiaire dans la Wayback Machine et l'améliorer au moyen de règles.

Défis lors du harvesting

- Rapidité de l'évolution technologique dans le web versus outils pour l'archivage du web
- Fonctions dynamiques
- Problèmes de la Wayback
  - Des éléments comme par ex. des images ne sont pas affichés
  - La navigation n'est pas utilisable
- Flash
- Galeries d'images
- Médias sociaux (Facebook, Youtube etc.)

Visual Quality Index

La BN a implémenté un Visual Quality Index pour soutenir le contrôle de qualité.

Pour tous les crawls, deux captures d'écran sont comparées automatiquement au moment du harvesting.

1. Comparaison du site web live avec le crawl actuel
2. Comparaison de la dernière version archivée dans Access avec le crawl actuel

La concordance des captures d'écran donne une valeur comparative qui est représentée par les couleurs vert / orange / rouge et qui simplifie le contrôle de qualité.