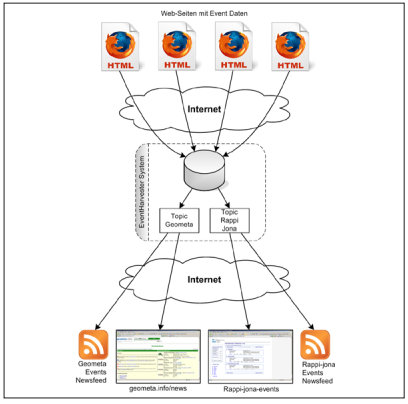
Gian-Reto Alig

EventHarvester

Event Web Scraping-System

Diplomand	Gian-Reto Alig
Examinator	Prof. Stefan F. Keller
Experte	Claude Eisenhut, Eisenhut Informatik AG, Burgdorf BE
Themengebiet	Internet-Technologien und -Anwendungen

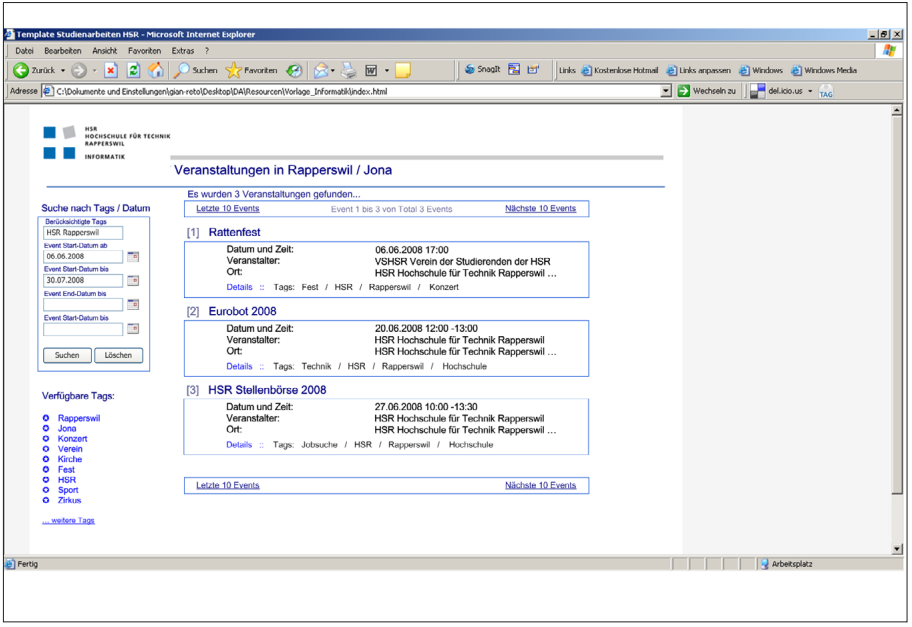


Arbeitsweise EventHarvester

Ausgangslage: Die Daten für Anlässe im Internet zu finden kann ziemlich schwierig sein, da es sehr wenig Web-Seiten gibt, die diese Anlässe gebündelt auflisten und da diese meist nur für sehr spezifische Themengebiete verfügbar sind (z.B. tilllate.ch für Ausgangs-Events). Oft findet man einen Anlass nur auf den Verbands- oder Vereinsportalen, sodass das Eintragen aller relevanten Daten in die persönliche Agenda zu einem mühseligen Prozess wird.

Lösungsansatz: Es wäre möglich, die relevanten Web-Seiten automatisch nach Anlass-Daten zu durchforsten (dieser Vorgang wird Web Scraping genannt) und die gefundenen Daten auf einem Event-Portal darzustellen, wodurch sich ein interessierter Benutzer nur noch eine Adresse merken muss. Dies würde zudem ein automatisches Eintragen der Anlässe ermöglichen, während die meisten heutigen Event-Portale auf einem manuellen Eintragssystem basieren.

Realisierung: Bei der Evaluierung wurde das «web-harvest»-System wegen der Integrierbarkeit in ein eigenes Java-Projekt ausgewählt. Es wurde ein flexibles Web-Scraping-System aufgebaut, das über eine Reihe von Konfigurationsdatenbank-Tabellen gesteuert wird, wobei die



Screenshot der Seite Rappi-Jona-events

gefundenen Anlässe in einer anderen Tabelle in derselben Datenbank gespeichert werden. Ein Admin-Web-GUI erlaubt das Konfigurieren des Systems über das Netz. Es wurde eine JSP-Erweiterung für die News-Seite von Geometa.info geschrieben sowie ein zusätzliches JSP-GUI entwickelt, dem alle Anlässe zum Thema Rapperswil-Jona zugeordnet wurden.

Ausblick: Der EventHarvester hat noch viel Entwicklungspotential. Die Ortsangaben können durch Verwendung eines Geo-Services so aufbereitet werden, dass die Events auf einer Karte angezeigt werden können. Oder es wäre auch möglich, dass eine Crawler-Komponente das Netz nach neuen Seiten für den Web-Scraping-Vorgang durchsucht. Viele Punkte der Konfiguration könnten durch ein intelligenteres System hinfällig werden. Mehr Informationen unter: http://<domain>/rappi_jona_events