

Exposé zur Studienarbeit

# Implementierung einer Suche für Outbreak Database

27. Juli 2008

Autor: Marco Eckstein

E-Mail: eckstein@informatik.hu-berlin.de

Betreuer: Prof. Dr. Ulf Leser, Dipl.-Inf. Michael Behnke (Charité)

## Hintergrund

Bei der Suche nach Literatur zu nosokomialen Ausbrüchen<sup>1</sup> von Krankheitserregern in Online-Verzeichnissen medizinischer Literatur wie PubMed[1] erhält man eine Vielzahl von Artikeln, von denen nur ein Bruchteil tatsächliche Ausbrüche beschreiben, so dass eine aufwändige Auslese der relevanten Artikel erfolgen muss. Um den Artikeln die wesentlichen Kriterien eines Ausbruchs, z. B. Ausbruchsursache, Übertragungsweg und getroffene Maßnahmen, zu entnehmen, muss außerdem jeder Artikel ausgewertet werden.

Um diese Probleme zu lösen und damit Ärzten, Epidemiologen und weiterem medizinischem Personal die Arbeit zu erleichtern, wurde Outbreak Database[2][3] entwickelt, eine Datenbank die eine Menge von Outbreak-Artikeln enthält. Dabei beschreibt ein Outbreak-Artikel einen nosokomialen Ausbruch und enthält die wesentlichen Angaben dazu in strukturierter Form, sowie einen Verweis auf den Original-Artikel, so dass gezielte Suchen durchgeführt und Ergebnisse übersichtlich dargestellt werden können. Die Outbreak-Artikel werden von Ärzten mit Fachwissen in Infektionsepidemiologie erstellt und aktualisiert. Bisher sind mehr als 2000 Ausbrüche mit über 200 verschiedenen Erregern (Spezies) aus den Jahren 1962 bis 2006 beschrieben, und weitere werden folgen.

Initiatoren des Projektes sind das Institut für Medizinische Mikrobiologie und Krankenhaushygiene der Medizinischen Hochschule Hannover, die Firma Schülke & Mayr und das Institut für Hygiene und Umweltmedizin[4] der Charité – Universitätsmedizin Berlin. Es gab bereits eine Vorgängerversion von Outbreak Database, diese Version ist jedoch nicht mehr online und wird nicht weiterentwickelt. Am Institut für Hygiene und Umweltmedizin wird Outbreak Database z. Zt. neu implementiert; der Datenbestand wird aber weiterverwendet.

## Projektrahmen

Im Rahmen der Neuimplementierung sind bisher entwickelt worden bzw. befinden sich in Entwicklung:

- XML Schema für die Outbreak-Artikel, aufbauend auf dem alten Datenbestand
- Relationales Datenbankschema, passend zum XML Schema
- Komponenten zur Datenkonvertierung:
  - Konvertierung der Daten des alten Datenbestandes in XML-Dateien, die dem XML Schema entsprechen.
  - Konvertierung der XML-Dateien in Datensätze der Datenbank und umgekehrt
  - Konvertierung der XML-Dateien in PDF und HTML-Dokumente sowie in Übersichtstabellen (HTML)

---

<sup>1</sup>Nosokomiale Ausbrüche sind Ausbrüche in medizinischen Einrichtungen wie z. B. Krankenhäusern.

- Eine Website als GUI für die Suche und die Anzeige der Ergebnisse.

Außerdem wurde eine einfache Volltextsuche auf den Daten mit Lucene.Net[5][6] implementiert. Hierbei können ähnliche Suchanfragen gestellt werden, wie es in Websuchmaschinen üblich ist, z. B.: *SP:Bacillus OR MRSA* (findet Outbreak-Artikel die als Spezies *Bacillus* oder irgendwo *MRSA* enthalten)

## Ziele

Die bisherige Implementierung der web-basierten Suche erfüllt jedoch nicht alle Anforderungen, denn zum einen gibt es weitere bezüglich der Funktionalität:

- Es soll nicht nur Volltextsuche möglich sein, sondern auch exakte Suche („ist gleich“) und Bereichsanfragen („ist kleiner als“). Die verschiedenen Sucharten sollen kombinierbar sein, also z. B. Volltextsuche mit exakter Suche.
- Bei einigen Feldern in einem Outbreak-Artikel, die mehrfach vorkommen (also Listen von Werten) spielt die Position des Wertes eine Rolle. Z. B. ist der erste Erreger in einer Liste von Erregern der wichtigste an diesem Ausbruch beteiligte. Dies soll beim Ranking der Suchergebnisse berücksichtigt werden.
- Es sollen deskriptive statistische Auswertungen ermöglicht werden: Die Ergebnisse sollen nach vorhandenen Werten gruppiert und die Anzahl der Treffer pro Gruppe soll ermittelt werden (GROUP-BY).

Zum anderen sind folgende Anforderungen bezüglich der Nutzbarkeit zu berücksichtigen:

- Da die Zielgruppe des Projektes medizinisches Personal ist, muss die Suche einfach und intuitiv zu bedienen sein. Folglich ist z. B. SQL keine geeignete Anfragesprache für den Endnutzer.
- Es gibt pro Outbreak-Artikel bis zu 75 unterschiedlich benannte Felder mit (größtenteils textuellen) Werten. Diese Felder sind in den XML-Dateien hierarchisch angeordnet, und die Darstellung für den Endnutzer als PDF oder HTML lehnt sich daran an. Es ist wünschenswert, dem Benutzer sich ebenfalls an dieser Struktur anlehrende Werkzeuge zu bieten, die ihm eine gute Orientierung ermöglichen, in welchen Feldern er suchen kann.
- Außerdem ist es bei vielen Feldern sinnvoll, dem Nutzer anzuzeigen, welche Werte im Datenbestand vorkommen, wie z. B. eine Liste von Ländern in denen Ausbrüche stattfanden. Dies soll Tipparbeit ersparen und „falsche“ Anfragen verhindern helfen, wie z. B. die Suche nach *England*, wenn Ausbrüche die dort stattfanden unter *UK* eingeordnet sind. Bei anderen Feldern mit z. B. Wahrheitswerten oder Datumsangaben soll der Nutzer ebenfalls unterstützt werden, so dass er nicht rätseln muss, was genau er eingeben muss um ein gewünschtes Ergebnis zu erhalten.

Ziel der Studienarbeit ist es, die obengenannten Anforderungen umzusetzen:

- Es soll eine Nutzerschnittstelle für eine erweiterte Suche entworfen und (zumindest) ein Prototyp implementiert werden.
- Es soll die eigentliche Suchfunktionalität implementiert werden. Das Ergebnis ist eine Liste von Paaren der Form (*Outbreak-Artikel-ID*, *Ranking*), wobei das *Ranking* anzeigt, wie gut der Outbreak-Artikel zur Suchanfrage passt.

## Vorgehen

Zunächst wird untersucht werden, welche Möglichkeiten ASP.NET 3.5 und AJAX für die Gestaltung einer Suchschnittstelle bieten, und es werden parallel dazu Ideen entwickelt werden, wie diese Möglichkeiten optimal eingesetzt werden können, um die Nutzbarkeit der

Schnittstelle zu erhöhen. Darauf aufbauend erfolgt zunächst eine Designstudie (d. h. ein statischer Entwurf der Schnittstelle). Diese Studie wird mit relevanten Anwendern besprochen und das Feedback wird in eine neue Version eingearbeitet. Schließlich wird die Schnittstelle implementiert (mindestens als Prototyp). Bei der Schnittstelle wird es sich im Wesentlichen um ein langes Formular handeln; u. U. wird an einigen Stellen ein iterativer Sucheinschränkungsprozess stattfinden.

Dann müssen die Möglichkeiten der Werkzeuge ausgewertet werden, die für die Implementierung der Suchfunktionalität in Frage kommen. Diese Werkzeuge sind Microsoft SQL Server 2005[7] und Lucene.Net.

Schließlich wird die eigentliche Suchfunktionalität implementiert werden.

Oberste Priorität haben die Implementierung des Formulars, das den Nutzer bei der Eingabe sinnvoller Werte für die Volltextsuche unterstützt, sowie die Entwicklung eines geeigneten Rankings der Ergebnisse.

Als Programmiersprache wird C# verwendet, ggf. auch JavaScript.

## Literatur und Links

[1] PubMed <http://www.ncbi.nlm.nih.gov/pubmed/>

[2] Outbreak Database <http://www.outbreak-database.com/>

[3] Irina Zuschneid, Sabine Stamm-Balderjahn, Sonja Hansen, Katrin Groneberg, Michael Behnke, Henning Rüden, Petra Gastmeier: „Outbreak-database. Eine Datenbank für Ausbrüche in medizinischen Einrichtungen“. <http://www.dghm.org/texte/Outbreak%20database.pdf> (Eine etwas veraltete aber recht ausführliche Beschreibung von Outbreak Database)

[4] Institut für Hygiene und Umweltmedizin <http://www.charite.de/krankenhaushygiene/>

[5] Lucene.Net <http://incubator.apache.org/lucene.net/>

[6] Otis Gospodnetic, Erik Hatcher: „Lucene in Action“ Manning 2005.

[7] Robert Vieira: „Professional SQL Server 2005 Programming“ Wiley 2007.

(Die Hyperlinks wurden am 29.06.2008 aufgerufen.)