

Kundenprojekt Web-Technologien  
Institut für Informatik, Freie Universität Berlin

---

**Entwicklung einer Software zum Editieren, Anwenden und  
Auswerten von Hearst Pattern für die neofonie GmbH**

# **Angebot**

**09.05.2007**

**Version 1.2**

**Die Projektgruppe des Kundenprojekts Web-Technologien des Instituts der Informatik der FU Berlin (im folgenden Auftragnehmer), vertreten durch Marco Jeschke und Miao Wang, unterbreitet der neofonie Technologieentwicklung und Informationsmanagement GmbH (im folgenden Auftraggeber) folgendes Angebot:**

## 1 Einleitung

Der Auftraggeber wünscht sich die Entwicklung einer Software zum Editieren, Anwenden und Auswerten von Hearst Pattern. Ziel ist es mit dieser Software vorhandene Hearst Pattern auszuwerten und neue zu generieren. Dabei soll das Verfahren mittels Hearst Pattern die automatische Erzeugung von Ontologien und Extraktion von Assoziationen ermöglichen. Desweiteren sollen die Ergebnisse eine Hilfestellung geben zukünftige Suchmaschinen mit semantischen Zusammenhängen anzureichern und so netzbasierte Informationssysteme zu verbessern.

## 2 Angebotsumfang

### 2.1 Beschreibung der Aufgabe

Die Erzeugung der Hearst Pattern wird in einem webbasierten Editor realisiert. Im Folgenden werden die Hearst Pattern über einem großen Corpus (einer oder mehrerer Suchmaschinen o.ä.) abgefragt. Im Fokus steht dann die Auswertung der Ergebnisse des Corpus, welche nach Signifikanz und Häufigkeit sortiert werden. Die Ergebnisse werden durch eine Weboberfläche visualisiert, um eine differenziertere Bewertung der Hearst Pattern zu ermöglichen. Die Pattern können wahlweise in deutscher oder englischer Sprache formuliert werden.

Zudem sollen erstellte Pattern abgespeichert und erneut eingeladen werden können. Alle möglichen Einstellungen bezüglich Schwellwerte zur Berechnung der Signifikanz, Konfigurationsmöglichkeiten und Kardinalitäten sollen einstellbar sein.

Die Software wird in Java implementiert und beinhaltet eine offene Architektur zwecks Modularität und Erweiterbarkeit. Vor allem sollen neue Suchmaschinen einfach eingefügt werden können. Neben der Funktionalitäten der webbasierten GUI, soll jene Funktionalitäten auch über eine definierte Java-API verfügbar sein, sodass eine spätere Erweiterung für das automatische Lernen der Pattern ermöglicht wird. Diese API soll unabhängig von den verwendeten Suchmaschinen sein. Anhand exemplarischer Lexika soll die prinzipielle Eignung der Software aufgezeigt werden. Zu der lauffähigen Version der Software wird zudem eine Dokumentation angefertigt, die aus Architektur und Übersichtsdokumentation, Code- und Benutzerdokumentation besteht.

### 2.2 Bezugsdokumente

- Präsentationsfolien zu den Anforderungen der neofonie GmbH (Projekt Hearst Pattern) und die Präsentation vom 25.04.2007
- Scrum Product Backlog der neofonie GmbH (siehe Anhang)

## 2.3 Zu erbringende Leistungen

Im Rahmen dieses Angebots wird der Auftragnehmer den Entwurf und die Implementierung der Software in folgenden Arbeitspaketen erbringen:

- Arbeitspaket Architektur
  - Entwicklung einer modularen Architektur des Gesamtsystems
  - Definition der Java-API Schnittstelle
  - Anbindung an APIs, Datenbank und Dateisystem
- Arbeitspaket GUI
  - Erstellung eines webbasierten Editors zur Erstellung von Hearst Pattern
  - Abspeichern und Laden von abgespeicherten Pattern
  - Webbasierte Visualisierung der Suchergebnisse und Instanzen möglicher Ergebnisse
  - Implementierung der Schnittstellen der Client API
- Arbeitspaket Pattern
  - Implementierung der Schnittstelle zu externen Suchmaschinen (direkte Unterstützung von Google über die Google API und der neofonie search :engine)
  - Implementierung der Verwaltung der Pattern
  - Entwicklung von Testpattern und exemplarischer Lexika
  - Entwicklung von Algorithmen für die Phrasenerkennung
- Arbeitspaket Algorithmen
  - Implementierung von Filter zur Extraktion der Ergebnisse
  - Implementierung von Algorithmen zur Bestimmung der Häufigkeitsverteilungen, der Signifikanzauswertung und der Gewichtung der Ergebnisse der Pattern
  - Aufbereiten der Ergebnisse für die Visualisierung
  - softwareseitige Qualitätsauswertung bzgl. Information-Retrieval Kriterien (precision, recall, f-measure) gegen einen Gold-Standard
- Arbeitspaket Integration und Test
  - Erstellung von Unittests mit JUnit, mit denen die Software getestet werden kann
  - Durchführung der Testläufe und Qualitätsprüfung anhand gängiger Information Retrieval Gütekriterien (precision, recall, f-measure)
  - Evaluation des Systems bzgl. der exemplarischen Lexika
  - Integration der verschiedenen Bestandteile
  - Überwachen der Architektur und der Java-API Schnittstelle
- Arbeitspaket Dokumentation
  - Überwachung der Kommentierung des Codes
  - Erstellung der Architekturdokumentation
  - Erstellung der Übersichtsdokumentation
  - Erstellung des Benutzerhandbuchs in deutscher und englischer Sprache

## 2.4 Zu liefernde Ergebnisse

Geliefert wird eine CD mit folgendem Inhalt:

- lauffähige Version der Software
- exemplarische Lexika: (Assoziationen Politiker -> Thema und Person -> Geburtsdatum)
- Quellcode mit kommentierten Methodenköpfen, Modulen und Schnittstellen
- Benutzerhandbuch in deutscher und englischer Sprache im PDF-Format

Zudem wird eine lauffähige Version der Software zum Abnahmezeitpunkt auf einem Testserver zugänglich sein.

## 3 Rahmenbedingungen

### 3.1 Abgrenzungskriterien

Die Software garantiert die Erstellung und Auswertung von Hearst Pattern, jedoch keine Garantien über die Zuverlässigkeit oder Qualität der Suchergebnisse durch die angebundenen Suchmaschinen.

Die Software gewährt die Lauffähigkeit auf einem Standardsystem. Die Integration des entwickelten Systems in eine vorhandene Umgebung beim Auftraggeber, sowie Wartung, Betrieb und Weiterentwicklung der Software sind nicht Bestandteil dieses Angebots. Es wird keine Garantie gegeben, dass die Software problemlos in die Softwarelandschaft des Auftraggebers eingegliedert werden kann.

Eine Schulung über die Dokumentation hinaus wird nicht durchgeführt.

Die Performanz der Software ist nicht Kriterium der Abnahme. Sie wird insoweit optimiert, wie das zeitliche Budget nach Fertigstellung der Hauptfunktionalität dies zulässt.

### 3.2 Mitwirkungspflicht des Auftraggebers

Der Auftraggeber stellt alle notwendigen Informationen und Dokumente kostenlos und zeitnah zur Verfügung und wird gemeinsam mit dem Auftragnehmer alles Notwendige unternehmen, um das Projekt gemeinsam und erfolgreich abschließen zu können.

Ferner stellt der Auftraggeber Inputlisten für die exemplarischen Lexika bereit, sowie seine Suchmaschine für Abfragen des Systems zur Verfügung, einschließlich jeglicher Informationen und Dokumentation diesbezüglich.

### 3.3 Besondere Regelungen zur Zusammenarbeit oder zum Projektverlauf

Benötigte Informationen und Anfragen sollten vom Auftraggeber innerhalb von 3 Werktagen bereitgestellt oder beantwortet werden, andernfalls obliegt es der Projektleitung im gemeinsamen Interesse zu entscheiden. Sollte dies nicht möglich sein, verzögert sich die Fertigstellung des Projekts um genau diese Zeitspanne (bis zur Beantwortung/Bereitstellung) und der Auftraggeber hat die zusätzlichen Kosten zu tragen.

Die Kommunikation zwischen Auftraggeber und Auftragnehmer erfolgt immer mit Kenntnis der Projektleitung.

Vom Auftraggeber geforderte Anforderungsänderungen bedürfen der Schriftform. Anforderungsänderungen nach Vertragsabschluss können nach eingehender Prüfung abgelehnt oder

nach Möglichkeit mit in das Projekt eingebunden werden. Zusätzlich anfallende Kosten übernimmt der Auftraggeber.

Das entstehende Produkt wird unter der BSD-Lizenz (Berkeley Software Distribution) veröffentlicht.

### 3.4 Zeitraum und Termine

- 16.05.2007 – Meilenstein 1
  - Präsentation erster Konzepte zur Architektur
  - Präsentation eines Click-Dummys zur Demonstration der Usability
  - Konzepte bzgl. der Signifikanz-Auswertung
- 06.06.2007 – Meilenstein 2
  - Erste lauffähige Version als Prototyp
- 18.07.2007 – Meilenstein 3 und Endabnahme
  - Vorführung der vollfunktionsfähigen Software mit Testdaten
  - Abgabe der zu liefernden Ergebnisse

## 4 Kommerzielle Regelungen

### 4.1 Preis

Die Personalplanung (Anzahl der Personen pro Paket x Personenstunden) und die daraus resultierende Summe der Personenstunden sind nachfolgend aufgliedert nach Meilensteinen und Arbeitspaketen:

	Meilenstein 1	Meilenstein 2	Meilenstein 3	Personenstunden
Dauer	1 Woche	3 Wochen	6 Wochen	10 Wochen
Architektur	5 x 4h	4 x 12h	3 x 24h	140 h
GUI	5 x 4h	4 x 12h	2 x 24h	116 h
Pattern	4 x 4h	3 x 12h	4 x 24h	148 h
Algorithmen	4 x 4h	3 x 12h	4,5 x 24h	160 h
Integration/Test	0 x 4h	2 x 12h	3 x 24h	96 h
Dokumentation	0 x 4h	2 x 12h	1,5 x 24h	60 h
Projektleitung	2 x 4h	2 x 12h	2 x 24h	80 h
<b>Summe</b>	<b>20 x 4h</b>	<b>20 x 12h</b>	<b>20 x 24h</b>	<b>800 h</b>

Es ergibt sich ein gesamter Personalaufwand von 800 Personenstunden. Dies entspricht 20 Personen à 4 SWS über einen Projektzeitraum von 10 Wochen vom Projektstart am 09.05.2007 bis zum Projektende am 18.07.2007.

## 4.2 Zahlungsweise

Nach Abnahme des Projektes ist die Bezahlung je eines Scheins à 4 SWS für jeden Projektteilnehmer, der erfolgreich am Projekt mitgewirkt hat, fällig. Diese ist zahlbar, innerhalb von 6 Wochen nach Beendigung des Projekts.

## 4.3 Lieferung und Abnahmeregelung

Das Projekt wird abgenommen, wenn:

- die Software die Erstellung, das Editieren, Abspeichern und Einladen von Hearst Pattern unterstützt
- eingegebene Hearst Pattern ausgewertet und die Ergebnisse visualisiert werden
- die Java-API das gewünschte Verhalten aufweist
- alle möglichen Schwellwerte und Einstellungsoptionen einstellbar sind
- alle Teilergebnisse der zuliefernden Ergebnisse (siehe Punkt 2.4) vorliegen

Die Antwortzeiten der Software sind kein Abnahmekriterium.

Alle Ergebnisse und Teilergebnisse werden, gemäß der festgelegten Termine (siehe 3.4), spätestens mit einer Frist von einer Woche, abgenommen. Nach Ablauf der Frist gilt das Teilergebnis auch ohne Rückmeldung als abgenommen.

Abgenommene Teilergebnisse sind nicht mehr Bestandteil der Endabnahme.

Nach Abnahme aller Teillieferschritte gilt das Projekt als abgenommen und beendet. Die Abnahme darf nicht aus trivialen Gründen abgelehnt werden.

## 5 Rechtliche Aspekte

Für direkte oder indirekte Schäden oder Datenverluste, die durch die Benutzung dieser Software oder dem Scheitern des Projektes entstehen, übernimmt der Auftragnehmer keine Haftung.

## 6 Nebenabsprachen

---

### Auftrag erteilt.

Berlin,

---

Ort, Datum

---

Unterschrift Auftraggeber

---

Unterschrift Auftraggeber

---

Unterschrift Auftragnehmer

---

Unterschrift Auftragnehmer

**Anlagen:** Scrum Product Backlog der neofonie GmbH