Prof. Dr. Gerhard Reinelt
Dipl.-Inform. Meinrad Funke
Institut für Angew. Mathematik
Universität Heidelberg

Software–Praktikum Informatik

Gültigkeitstests von HTML-Links

Kontext In steigendem Maße werden Informationen über das *World Wide Web* ausgetauscht. Oft wird es sogar mit dem Begriff "Datenautobahn" gleichgesetzt. Leider gibt es auf dieser Autobahn einige Sackgassen und fehlende Abfahrten. Denn nicht immer verweist jeder Link auf ein gültiges Ziel. Und auch der umgekehrte Fall, nicht erreichbare Seiten, kann vorkommen, auch wenn er naturgemäß nach außen selten auffällt.

Aufgabe Im Rahmen des Software–Praktikums soll ein Werkzeug entwickelt werden, mit dessen Hilfe man Links, auf die kein Zugriff möglich ist, und nicht eingetragene Seiten entdecken kann. Vorgegeben werden dabei eine Startseite, Rechner auf denen rekursiv weitergesucht werden soll, und Verzeichnisse, in denen alle Dokumente erreicht werden sollen.
Außerdem ist eine Liste von bereits getesteten Links zu berücksichtigen, um den Aufwand eines weiteren Durchlaufs nach einer Korrektur zu vermindern.
Die Resultate sind in geeigneter Form zu sammeln. Dabei sollen die folgenden Punkte berücksichtigt werden

- Dateinamen und Zeilen in den entsprechenden Seiten bei fehlgeschlagenen Zugriffen (gesammelt bei mehreren Einträgen)
- Absolute Pfade nicht eingetragener Dateien

Programmintern sollten, soweit möglich, eindeutige (absolute) Namen verwendet werden. Nach außen sind die Namen, die in den Dokumenten und Dateien vorgegeben sind, maßgebend.

Voraussetzungen Für dieses Projekt sind Kenntnisse von `HTML`, `C` und einer `Unix`–Umgebung erforderlich. Gegebenenfalls müssen sie vor Beginn oder parallel erworben werden.

Ansprechpartner Interessenten melden sich bitte bei Meinrad Funke, Zi. 221, INF 294.