Daten

Es liegen verschiedene Datensätze vor. Auf dieser Seite werden sie kurz beschrieben. Bitte entpackt sie gegebenfalls in eure Working Copy, aber checkt die lokal entpackten Daten nicht ein, weil sie unter Umständen sehr groß werden.

CD Tupel

Die ca. 1,8 Mio. CD-Tupel wurden von FreeDB http://www.freedb.org/ extrahiert. Jeder Datensatz enthält Informationen über eine CD, z.B. Künstler, CD-Titel, Genre, Erscheiningsjahr, etc. Die Titel der Tracks wurde auf 3 beschränkt, um die Datenmenge nicht zu groß werden zu lassen.

Adress-Tupel

Der Adress-Datensatz enthält ca. 1. Mio. Tupel mit Personen- und Kontaktdaten. Der Datensatz darf nicht weitergegeben werden und die Nutzung des Datensatzes ist ausschließlich innerhalb des Seminars erlaubt.

Bilder

Dieser Datensatz enthält knapp 10000 recht kleine Bilder. Er heißt "Corel 10000". Es sind jeweils ca. 100 Bilder der gleichen Domäne, wobei es auch grafisch exakte Duplikate zu geben scheint. Man könnte auch ein kleineres Fenster über einige der Bilder schieben, dann hat man mehr ähnliche Segmente.

DBpedia Abstracts

Enthält die Short Abstracts aller 466.444 von DBpedia erfassten deutschen Wikipedia-Artikel als CSV-Datei. Short Abstract bezeichnet dabei eine verkürzte Version des ersten Absatzes eines Artikels. Quelle: DBpedia Downloads

EPCs

Dieser Datensatz enthält 604 Event-driven-Process-Chain-Diagramme. Diese müssen unbedingt im HPI bleiben und dürfen nicht weitergegeben werden. Jedes Diagramm ist ein Graph aus Ereignissen, Funktionen und Konnektoren (und Kanten). Diese sind in XML angegeben. Zum Betrachten eignet sich ProM.