Dokumentation

Vorverarbeitung

Die Daten sind gegeben als ca. 130.000 einzelner XML-Dokumente, die neben den für uns relevanten Daten noch weitere Informationen enthalten.
Zunächst müssen diese Daten zu einer einzelnen für uns brauchbaren Datei zusammengefasst werden.
Zunächst werden die Dateien mit **prepare_xml.py** zusammengefasst, das Ergebnis ist eine große Datei, die alle Informationen der ursprünglichen Dateien enthält.

Die resultierende Datei **all.xml** wird mit XSLT (siehe **info_extraction.xsl**) transformiert in den Index, der unsere Datenbasis darstellt.

Als Vorbereitung für das Auslesen der biographischen Informationen müssen die Namen der Herausgebern mit Ressourcen in DBpedia verknüpft werden. Zunächst werden die Herausgeber dazu mit XQuery (siehe **names.xquery**) ausgelesen. Die Namen werden dann in DBpedia nachgeschlagen. Nur ein kleiner Teil lässt sich so eindeutig zuordnen. Daher haben wir einige der häufig vorkommenden Herausgeber manuell zugeordnet. Um eine Zählung durchzuführen haben wir **count.py** geschrieben. Die Zuordnungen werden in den Index geschrieben. Dies alles ist in **dbpedia_name_lookup.py** zusammengefasst.

Ablauf der Abfrage

Auf der Startseite (**website/index.html**) erscheinen Suchfelder für den Titel, das Datum, den Ort und den Herausgeber. Wenn eine Suche gestartet wird wird der Index anhand dieser Daten gefiltert und die Ergebnisse auf der Ergebnisseite angezeigt.

Wenn der Herausgeber eine Verknüpfung zu DBpedia hat, ist ein Name anklickbar. Klickt man darauf, werden seine biographischen Informationen beim Server angefragt. Dieser liest sie mit SPARQL aus DBpedia aus. Mit REST und Ajax wird das Ergebnis ans Frontend gesendet. Dort wird es auf der Website angezeigt.

Informationen zu den Erstellen

Damit auch Mikroformate verwendet werden, haben wir eine Seite (**website/ueber_uns.html**) in der unsere Namen als vcards hinterlegt sind.