Aufgabe 1B: XML-Verarbeitung & XSLT

Erstellung eines Content Scrapers mittels XSLT

Das Fotoportal *PhotonPainter* möchte die Daten des befreundeten Portals *Stock Associates* nutzen. Leider verfügt dieses über keine Webservice-Schnittstelle und kann derzeit auch keinen Programmierer entbehren, der sich um diese Aufgabe kümmert. Als provisorische Lösung haben die Entwickler von *PhotonPainter* die vorhandene Katalog-Druckfunktion von *Stock Associates* ausgemacht. Da diese den Datenbestand des Portals nach einem festen Schema ausgibt, kann ein sogenannter Content Scraper eingesetzt werden, der diesen Katalog einliest, parst und in die Datenbank von *PhotonEmitter* eingliedert.

Ihre Aufgabe ist es, diesen Content Scraper mit dem Arbeitstitel *PhotonCollector* zu entwickeln. Er wird von *PhotonPainter* in regelmäßigen Abständen als Hintergrund-Task aufgerufen, um den Datenbestand von *Stock Associates* zu ermitteln, die Performance ist also nur zweitrangig. Das Ziel von *PhotonCollector* ist es, die Daten in einem Format bereitzustellen, das vom REST-Backend *PhotonEmitter* weiterverarbeiten kann. Dies beinhaltet das Entfernen des XHTML-spezifischen Codes und von Daten, die nur für *Stock Associates* zutreffen, wie beispielsweise die Kosten eines Bildes. Anschließend sollen die Daten in das Backend eingespielt werden; dabei sind Duplikate vorher zu entfernen. Mittels eines XSL-Stylesheets sollen die Daten direkt in das für das Backend erforderliche XML-Datenformat umgewandelt werden. Das Einfügen ins Backend erfordert dabei zwei Schritte: zunächst den Upload der Binärdatei und anschließend das Hinzufügen der Metadaten.

Tipp: Konzentrieren Sie sich zunächst zu Testzwecken auf die Verarbeitung eines einzelnen (des ersten) Eintrags. Anschließend können Sie Ihr XSL-Stylesheet so erweitern, dass es das gesamte Eingabedokument transformiert, so dass dieses für den Transfer nur noch „zerschnitten“ werden muss. Nutzen Sie zur Duplikaterkennung die IDs der Einträge und verwenden Sie einen Zwischenspeicher.

Das Ergebnis muss folgende Anforderungen erfüllen:

- Die Anwendung muss aus zwei nacheinander folgenden Schritten bestehen. Zunächst muss das fremde Datenformat in das *PhotonEmitter*-eigene umgewandelt werden und in einem zweiten Schritt sind die Einträge zu prüfen und gegebenenfalls ins Backend einzufügen.
- Verwenden Sie nur ein einziges XSL-Stylesheet für die Transformation.
- Ein Logfile ist zu erzeugen, das den Start und das Ende eines Suchlaufs sowie die hinzugefügten Einträge beinhalten soll.
- Das Programm soll über den Aufruf einer frei konfigurierbaren URL initiiert werden können.
- *Alle verwendeten Bestandteile müssen entweder Public Domain sein oder die erforderlichen Nutzungsrechte müssen schriftlich vorliegen.*

Abzugeben sind:

- XSL-Stylesheet (`transform.xsl`)
- Quellen (z.B. als gepacktes Eclipse-Projekt)
- Ausführbare Webanwendung als WAR-Archiv (`photoncollector.war`)