Aufgabe 1B: XML-Verarbeitung & XSLT

Erstellung eines Content Scrapers mittels DOM und 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 XHTML-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. *PhotonCollector* soll dabei (1) die XHTML-Quelldatei (Druck-Katalogseite von *Stock Associates*) einlesen, (2) die Daten via XSLT-Stylesheet in das XML-Zielformat des *PhotonEmitters* umwandeln, und (3) die Daten über die Web-Service-Schnittstelle *PhotonEmitter* einfügen. Dabei werden zunächst die Binärdatei hochgeladen und in Folge die Metadaten hinzugefügt.

Das Ergebnis muss folgende Anforderungen erfüllen:

- Die Aufgabenteile müssen in der Anwendung klar erkennbar sein, das beinhaltet insbesondere das **Kommentieren** des Quellcodes.
- Es ist ein **Logfile** zu erzeugen welches Informationen über den Ablauf sowie Informationen über die verarbeiteten Fotos enthält.
- Verwenden sie nur ein einziges **XSL-Stylesheet** für die Transformation.
- **Prüfen** sie vor dem Einfügen in den Webservice ob das Foto schon im Web Service vorhanden ist (Duplikaterkennung) und laden sie nur neue Fotos hoch.
- 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`)

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.