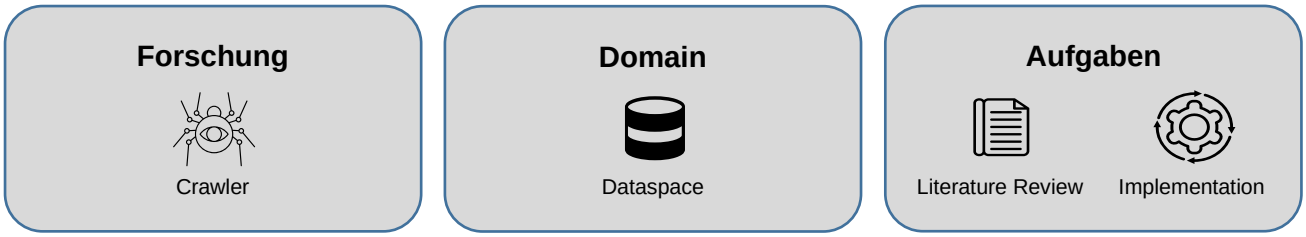
----- English version below -----

Ausschreibung Masterarbeit

Der Einsatz von Webcrawlern zur automatisierten Datenaggregation und Qualitätsüberwachung in Dataspaces: Potenziale, Herausforderungen und Implementierungsstrategien



Ausgangslage

Mit der steigenden Anzahl an Datenquellen im Internet und der zunehmenden Vernetzung von Informationsressourcen gewinnt das Konzept von Dataspaces an Bedeutung. Dataspaces bieten eine flexible und kollaborative Umgebung zur Verwaltung, Aggregation und Analyse von Daten aus verschiedenen Quellen. Der Bedarf, relevante Datenquellen zu identifizieren, kontinuierlich zu aktualisieren und zu überwachen, wird dabei immer wichtiger. Hier setzen Webcrawler an, die systematisch und automatisiert Daten sammeln, um die Datenbasis in einem Dataspace anzureichern und ihre Qualität zu sichern. Das Projekt untersucht das Potenzial von Webcrawlern zur effizienten Datenintegration und -qualitätssicherung in Dataspaces, wobei verschiedene Aspekte wie die automatisierte Datenaggregation, Metadatenextraktion, Such- und Indizierungsfunktionen, die Entdeckung neuer Datenquellen sowie die Qualitätsüberwachung im Mittelpunkt stehen.

Problemstellung

Dataspaces verlangen eine hohe Datenqualität, Aktualität und vollständige Verfügbarkeit der Datenquellen. Wie können Webcrawler zur kontinuierlichen Identifikation, Metadatenextraktion und Qualitätsüberwachung im Dataspace eingesetzt werden, und welche Herausforderungen ergeben sich dabei?

Vorgehensweise und Erwartete Ergebnisse

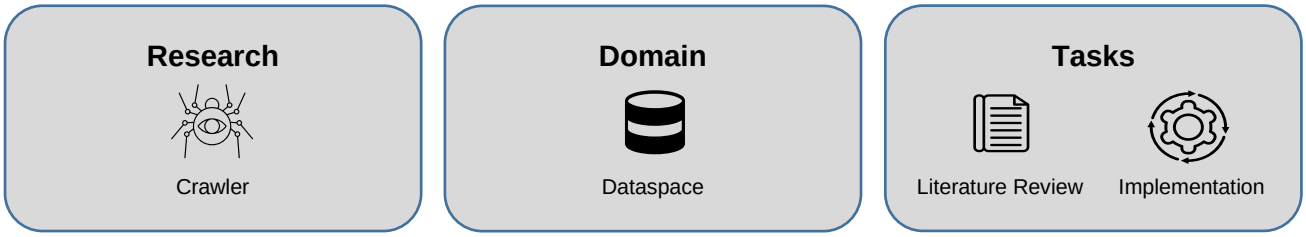
Um die Rolle von Webcrawlern in Dataspaces umfassend zu untersuchen, wird die Arbeit mit einer Literaturrecherche zur aktuellen Forschung und Praxis im Bereich Webcrawler und Dataspaces beginnen. Anschließend wird ein Crawler-Modell entwickelt, das für die spezifischen Anforderungen eines Dataspaces optimiert ist. Dies umfasst die Konfiguration des Crawlers zur Datenaggregation, Metadatenextraktion, Indizierung und Qualitätsüberwachung. Erwartet wird ein funktionsfähiger Webcrawler-Prototyp, der die Datenaggregation und -überwachung in einem Dataspace signifikant verbessert und so eine wertvolle Grundlage für die Entwicklung zukünftiger Dataspace-Anwendungen darstellt.

Ansprechpartner

Florian Hölken | **E-Mail:** hoelken@uni-wuppertal.de

Bachelor- / Master thesis

The use of web crawlers for automated data aggregation and quality monitoring in dataspaces: potentials, challenges and implementation strategies



Initial Situation

With the growing number of data sources on the Internet and the increasing networking of information resources, the concept of dataspaces is gaining in importance. Dataspaces offer a flexible and collaborative environment for managing, aggregating and analyzing data from different sources. The need to identify, continuously update and monitor relevant data sources is becoming increasingly important. This is where web crawlers come in, systematically and automatically collecting data to enrich the database in a dataspace and ensure its quality. The project investigates the potential of web crawlers for efficient data integration and quality assurance in dataspaces, focusing on various aspects such as automated data aggregation, metadata extraction, search and indexing functions, the discovery of new data sources and quality monitoring.

Problem Definition

Dataspaces require high data quality, up-to-dateness and complete availability of data sources. How can web crawlers be used for continuous identification, metadata extraction and quality monitoring in the dataspace, and what challenges does this pose?

Procedure and Expected Results

In order to comprehensively examine the role of web crawlers in dataspaces, the work will begin with a literature review of current research and practice in the field of web crawlers and dataspaces. Subsequently, a crawler model will be developed that is optimized for the specific requirements of a dataspace. This includes the configuration of the crawler for data aggregation, metadata extraction, indexing and quality monitoring. The model will be implemented and validated in several test phases to check its effectiveness and efficiency in the dataspace.

Contact Person

Florian Hölken | **E-Mail:** hoelken@uni-wuppertal.de