**Technisches Thema:**

Open Grid Service Architecture – Data Access and Integration (OGSA-DAI)
Eine Middleware-Platform zur Integration von Datenressourcen

Das technische Thema behandelt „OGSA-DAI“ (http://www.ogsadai.org.uk/). OGSA-DAI ist eine Middleware-Plattform für die Integration von Daten aus verschiedenen Quellen. Diese Middleware-Plattform bietet eine spezielle Unterstützung für das Arbeiten mit unterschiedlichen Datenressourcen, wie z.B. Relationale- oder XML Datenbanken und erlaubt einen Zugriff auf diese Ressourcen über Web Services.

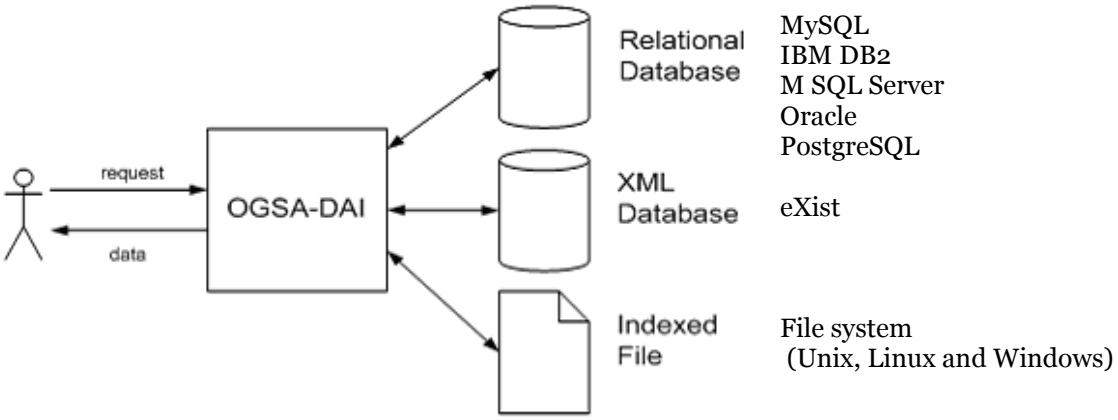


**Abbildung 1:** Einheitlicher Datenzugriff über OGSA-DAI

Abbildung 1 verdeutlicht das Ziel von OGSA-DAI ein einheitliches Interface für den Zugriff und die Integration von Daten aus verschiedenen Ressourcen zur Verfügung zu stellen. OGSA-DAI bietet einen transparenten Weg für Abfragen, Updates, Transformationen und das Bereitstellen von Daten via Web Services. Dabei werden die Komplexität der heterogenen Datenbanken und deren verteilte physikalische Lage vor dem User versteckt. Dies ermöglicht dem User auf unterschiedliche Datenquellen über ein Interface zuzugreifen.

Funktionalität von OGSA-DAI

- Unterstützung von Schnittstellen und DBMS
- Zugriff auf verschiedene Datenressourcen über ein Interface
- Komponenten zur Unterstützung von Abfragen, Transformationen und Bereitstellen von Daten
- Toolkit zur Entwicklung von Client-Anwendungen

OGSA-DAI wird als Einstimmung für das Anwendungsprojekt: „Daten Services für Bio-Daten“ behandelt, da es in diesem Projekt zum Einsatz kommen wird. Das technische Projekt soll einen Einblick in die Architektur und Funktionsweise von OGSA-DAI geben. Ziel ist es herauszufinden, wie und in welchem Ausmaß uns OGSA-DAI bei der Realisierung des Anwendungsprojektes unterstützen kann. Einen möglichen Anwendungsfall zeigt Abbildung 2.

In Abbildung 2 wird ein möglicher OGSA-DAI Workflow dargestellt. Gebündelte Funktionalitäten werden in OGSA-DAI zu so genannten „Aktivitäten“ zusammengefasst. Die grün bzw. gelb gefärbten Kästchen stellen solche Aktivitäten dar. Aktivitäten erwarten einen Input und stellen wiederum einen Output zur Verfügung, welcher dann als Input in die darauf folgende Aktivität einfließt. Eine Aneinanderreihung dieser Art ergibt in OGSA-DAI einen Workflow, der entsprechend der Spezifikation abgearbeitet wird. Abbildung 2 zeigt einen solchen Workflow, wo zuerst mittels der Aktivität „SQLQuery“ zwei unabhängige SQL Queries auf zwei verschiedene Datenressourcen abgesetzt werden. Die Aktivität „TupleMergeJoin“ bekommt die beiden Ergebnisse dieser Abfragen als Input und verarbeitet diesen Input. Der Ergebnisjoin dieser Verarbeitung wird als Ouptput an die nachfolgenden Aktivitäten weitergereicht. Die weiteren Aktivitäten führen Transformationen auf die Resultsets aus und letztlich wird das Ergebnis der Bearbeitung an den Aufrufer zurückgegeben.

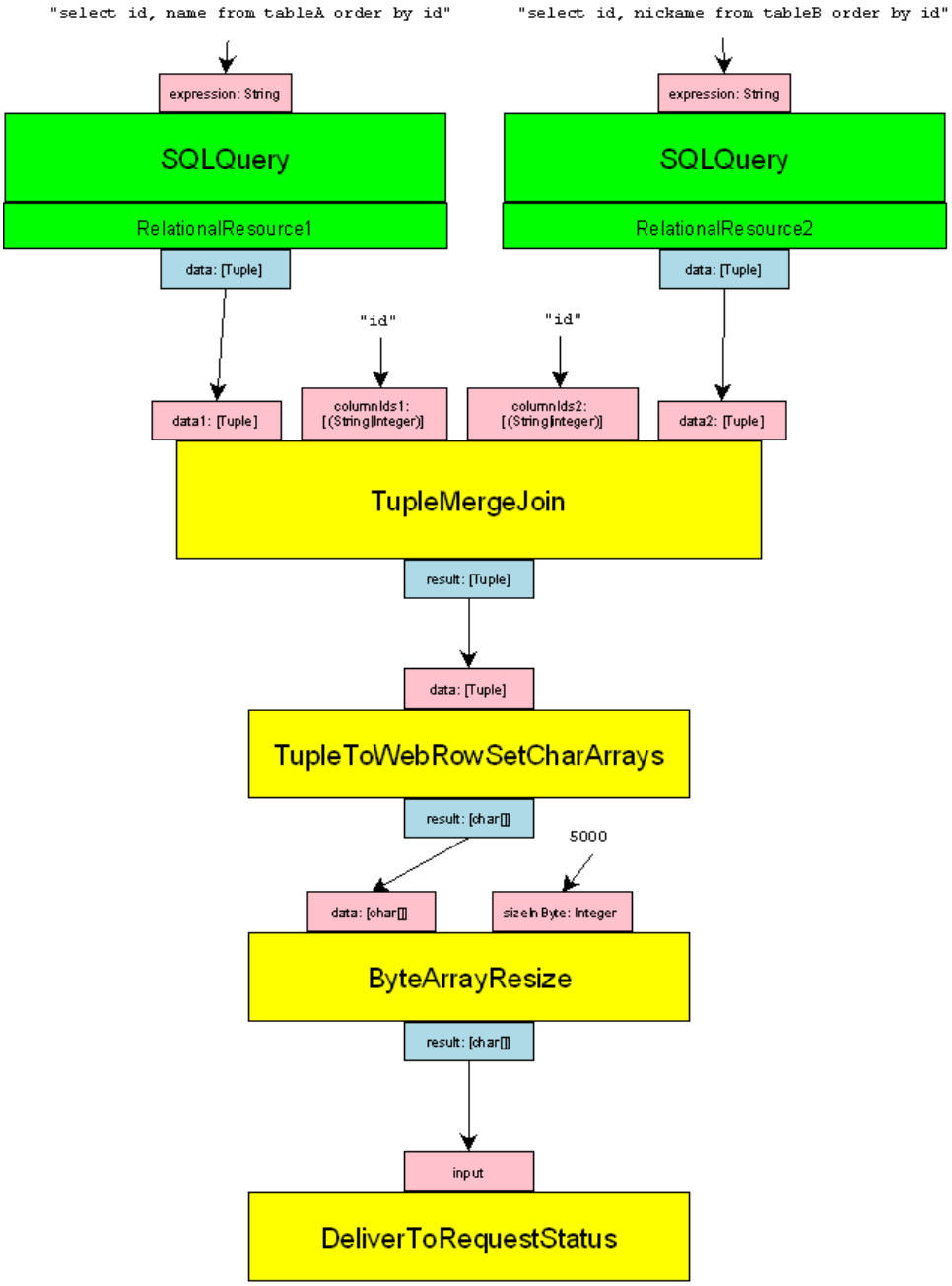


**Abbildung 2:** OGSA-DAI Workflow für einen Join von zwei Datenressourcen

**Inhaltliches Projekt:** Daten Services für Bio-Daten

Ziel dieses Projektes ist es ein Web-Service für den Zugriff auf verteilte Datenbanken zu gewährleisten. Im Rahmen des Projektes werden zwei Datenbanken (siehe Abb.3) durch dieses Service miteinander integriert. Die zu integrierenden Datenbanken sind einerseits eine Access-Datenbank mit pathologischen Daten von Patienten und andererseits, onkologische Daten von Patienten. Beide Datenbanken wurden von Herrn Konrad Stark in anonymisierter Form zur Verfügung gestellt. Über einen eindeutigen Schlüssel können Datensätze verschiedener Datenbanken miteinander assoziiert werden. Das System soll flexibel gestaltet werden, sodass eine einfache Integration auch von mehr als zwei Datenbanken erlaubt ist.

Das Service umfasst folgende Funktionalität:

- Feststellen der verschiedenen Daten Services
- Feststellen der verschiedenen Attribute, die für Abfragen relevant sind
- Formulieren und Ausführen von Abfragen
- Ergebnis zurückliefern

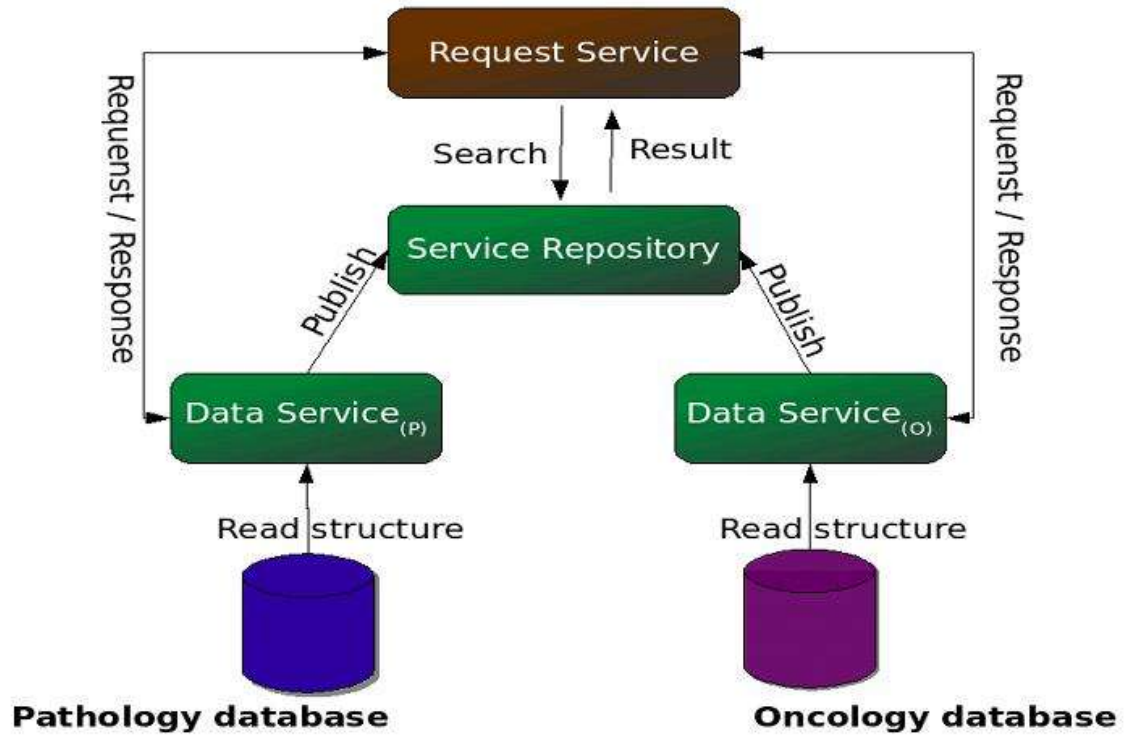


**Abbildung 3** Biomedizinisches Datenservice

## Konzeption

Eine zentrale Rolle bei der Lösung der Aufgabenstellung spielt OGSA-DAI. Die relevanten Datenbanken, wie z.B. die Pathologie-DB oder die Onkologie-DB werden in einem Service Repository als Daten Serxvices publiziert. Mit Hilfe von OGSA-DAI kann nun auf diese Daten Services zugegriffen werden und verschiedene Operationen können darauf ausgeführt werden. Sofern OGSA-DAI eine Datenressource unterstützt, braucht diese lediglich im Service Repository publiziert werden und ist somit integriert. Dabei ist es irrelevant wo die verschiedenen Datenbanken liegen,

solange eine gültige URL und die notwendigen Login-Daten bereitgestellt wurden. Über OGSA-DAI ist es nun möglich herauszufinden welche Ressourcen publiziert wurden. Außerdem bietet OGSA-DAI einen einheitlichen Zugriff auf die Datenbanken an.

Der Zugriff auf die unterschiedlichen Datenquellen ist demnach sehr gut und einheitlich möglich, sofern die Datenressource von OGSA-DAI unterstützt wird. Nun stellt sich die Frage, wie denn verschiedene Tabellen aus unterschiedlichen Datenbanken miteinander gejoint werden können. Ein Ansatz wäre z.B. alle relevanten Tabellen samt Daten aus den verteilten Datenbanken in eine lokale („temporäre") Datenbank zu laden und dann den Join auszuführen. Nach der erfolgreichen Durchführung des Joins werden die temporären Daten in der lokalen Datenbank wieder gelöscht. Alternativ könnte ein OGSA-DAI Workflow für den Join der verteilten Tabellen eingesetzt werden. Fraglich ist jedoch, ob durch die bereitgestellten OGSA-DAI Aktivitäten die geforderte Flexibilität hinsichtlich der Aufgabenspezifikation erfüllt werden kann.

**Probleme**

- Bei der Umsetzung der Aufgabenstellung gibt es eine Reihe von Problemen vor allem mit den unterschiedlichen Datenressourcen. Der Import der Testdaten stellt das erste Problem dar. Daten aus einer Postgre-Datenbank unter einer Unix-Umgebung können nicht ohne weiteres in eine Postgre-Datenbank unter einer Windows-Umgebung eingespielt werden. Das kann an den Testdaten selbst liegen, oder aber auch an der Migration der Daten von Linux zu Windows.

  Lösungsansatz*:*
  *Skript zum Migrieren der Daten oder Einsatz einer Linux Umgebung*

- OGSA-DAI unterstützt Postgre-DB, MySQL-DB, MS SQL Server-DB, etc., jedoch findet sich keine Access-DB Unterstützung wieder.

  Lösungsansatz*:*
  *Schreiben eines Supports für eine Access-DB unter OGSA-DAI*

- Der Zugriff auf die verschiedenen Datenressourcen ist zwar einheitlich, jedoch weisen die Datenressourcen unterschiedliches Verhalten auf. Die OGSA-DAI Aktivität „.GetAvailableTables" liefert das erwünschte Ergebnis, wenn die Aktivität auf beispielsweise eine MySQL-DB ausgeführt wird. Bei der Ausführung auf eine Postgre-DB werden jedoch nicht die normalen Tabellen, sondern Postgre-spezifische Tabellen zurückgeliefert.

  Lösungsansatz:
  *Schreiben einer eigenen Aktivität, die das gewünschte gleiche Verhalten auf alle unterstützten Datenquellen sicherstellt.*

**Anwendungsfall**
Ein Anwendungsfall kann demnach wie folgt aussehen: Ein User erfährt über ein Web-Interface welche Datenbanken durch dieses Service integriert werden und auf welche Attribute dieser Datenbanken er Abfragen absetzen kann. Der User wählt nun beispielsweise zwei Attribute aus der Pathologie-DB und zwei Attribute aus der

Onkologie-DB aus. Das Service generiert nun mit Hilfe eines eindeutigen Schlüssels eine Abfrage und setzt diese auf die beiden verteilten Datenbanken ab. Als Ergebnis bekommt der User eine Tabelle mit den vier gewünschten Attributen zurück.

**Verwendete Technologien/Tools:**
Bei diesem Projekt kommt es zum Einsatz einer Reihe von Technologien & Tools und dies führt in diesem Fall leider zu ein paar Abhängigkeiten. OGSA-DAI benötigt Java 1.4 oder Java 1.5, kann jedoch nicht mit Java 1.6 verwendet werden. Eine weitere Restriktion gibt es bezüglich Apache Tomcat, hier wird ausdrücklich die Version 5.0 von OGSA-DAI erwartet.

- Access-Datenbank
- Apache Ant 1.7
- Apache Tomcat 5.0 (für OGSA-DAI Services)
- Apache Tomcat 6.0 (für Client-Applikation mit JSF)
- Java 1.5
- Java Server Faces (JSF)
- MySQL-Datenbank
- NetBeans DIE 6.1 RC2
- OGSA-DAI (http://www.ogsadai.org.uk/)
- Postgre-Datenbank