

# Integrierte Informationsverwaltung für die Lebenswissenschaften mit der Information Workbench

Tobias Mathäß  
fluid Operations  
Altrottstr. 31  
69190 Walldorf, Germany  
tobias.mathaess@fluidops.com

Peter Haase  
fluid Operations  
Altrottstr. 31  
69190 Walldorf, Germany  
peter.haase@fluidops.com

**Abstract:** Forscher in den Lebenswissenschaften sind täglich mit den Problemen im Umgang mit riesigen Mengen heterogener und verteilter Daten konfrontiert. Oftmals ist es nicht möglich, komplexe Informationsbedürfnisse mit Hilfe von nur einer einzigen Datenquelle zu beantworten, sondern nur durch Kombination von Wissen aus verschiedenen Datenquellen. In diesem Beitrag stellen wir die Information Workbench vor: eine kollaborative Informations-Management-Plattform, welche die Integration heterogener strukturierter und unstrukturierter Daten unterstützt, eine einheitliche Sicht auf diese Daten bietet und somit quellübergreifende Suche über den gesamten Datenbestand ermöglicht. Die flexible und erweiterbare Benutzeroberfläche der Information Workbench erlaubt die datenabhängige Anzeige zusätzlicher Informationen zu gespeicherten Datenelementen. In diesem Beitrag demonstrieren wir den Einsatz der realisierten Ansätze, insbesondere hinsichtlich Datenintegration, Suchfunktionalität und Datenpräsentation in der Domäne der Lebenswissenschaften.

## 1 Motivation

Die Verwaltung, Speicherung und Verarbeitung digitaler Daten in den Lebenswissenschaften stellt Entwickler von Informations-Management-Systemen vor große Probleme. Die Daten, welche von Wissenschaftlern im Rahmen von Experimenten erzeugt, sind in hohem Maße heterogen. Arbeiten mehrere Wissenschaftler gemeinsam mit einem solchen System und erzeugen diese die anfallenden Daten nicht nach gemeinsamem Schema trägt dies weiter zur Heterogenität des Datenbestandes bei.

Hinzu kommt die Tatsache, dass auch öffentlich zugängliche Datensätze selten nach einem gemeinsamen Schema entworfen und erzeugt wurden. Es wird ein Werkzeug benötigt, welches eine gemeinsame Sicht auf die Daten ermöglicht. Dies gibt Forschern die Möglichkeit, Daten aus verschiedenen Datenquellen gleichzeitig und integriert zu nutzen. Diese Integration sollte transparent für die Benutzer sein, d.h. die Interaktion mit den integrierten Daten soll genau so von statten gehen wie mit einem einzigen Datensatz.

Die Information Workbench ist in der Lage, große Datenbestände zu verwalten, unabhängig davon ob diese Daten in strukturierter oder unstrukturierter Form vorliegen. Sie erlaubt den Benutzern, Daten zu verändern oder neu zu erzeugen, bereitet relevante Informationen sinnvoll auf und gibt dem Benutzer die Möglichkeit, den Datenbestand zu explorieren, also sich von einem Datenelement aus zu verwandten Elementen durch den Datengraphen zu bewegen. Eines der wichtigsten Merkmale der Information Workbench ist die Unterstützung für verschiedene Arten der Suche. Je nach Art des Informationsbedürfnisses und nach Fähigkeiten des Benutzers bieten sich bestimmte Suchparadigmen an. Das Spektrum der umgesetzten Arten der Suche reicht von einer Volltext-Suche über die strukturierten und unstrukturierten Daten

über formular-gestützte Suche, die den Benutzer bei der Formulierung strukturierter Anfragen unterstützt, bis hin zu Anfragen in einer strukturierten Anfragesprache.

Für die Datenintegration und Wissensrepräsentation baut die Information Workbench auf den Standards des Semantic Web und Linked Data auf, welche wir im Folgenden kurz einführen.

## 2 Grundlagen

In diesem Abschnitt stellen wir grundlegende Konzepte wie das Resource Description Framework (RDF) – das Standard-Datenmodell des Semantic Web – und die Linking Open Data Initiative (LOD) vor.

**Das Resource Description Framework** Das im Kontext des Semantic Web wichtigste Datenmodell ist das Resource Description Framework (RDF), welches vom World Wide Web Consortium (W3C) standardisiert wurde<sup>1</sup>. Ein RDF-Datenbestand besteht aus so genannten Statements. Ein Statement besteht aus einem Subjekt, einem Prädikat und einem Objekt. Das Subjekt ist hierbei ein Datenelement, eine so genannte Entität, das Objekt ist entweder eine weitere Entität, welche mit dem Subjekt in Beziehung steht, oder ein das Subjekt beschreibender Datenwert. Das Prädikat beschreibt die Art der Beziehung, in der Subjekt und Objekt stehen. Subjekt und Objekt eines solchen Statements können als Knoten im Datengraphen angesehen werden, das Prädikat als die verbindende Kante. Da die gleiche URI in verschiedenen Statements als Subjekt oder als Objekt benutzt werden kann definiert eine Menge von Statements einen gerichteten Datengraphen. Ein solcher Datengraph kann mit Hilfe von formalen Anfragesprachen angefragt werden.

**Linked Open Data** Die Idee hinter Linked Open Data (LOD) ist die Möglichkeit der Vernetzung verschiedener, eigenständiger RDF-Datensätze, indem URIs aus anderen Datensätzen benutzt werden. Betrachtet man zwei solcher Datensätze aus einer integrierten Sicht, so ergänzen sich diese zu einem gesamtheitlichen Bild. Auf Basis dieser Idee ist es möglich, bestehende Daten um bestimmte Aspekte zu erweitern. Enthält beispielsweise ein Datensatz Informationen über chemische Reaktionen sowie die an den Reaktionen beteiligten chemischen Verbindungen und ein weiterer Datensatz Detailinformationen über diese chemischen Verbindungen, so gibt eine integrierte Sicht auf beide Datensätze dem Benutzer die Möglichkeit, für eine chemische Verbindung sowohl alle Reaktionen, an denen diese Verbindung beteiligt ist, als auch weitere Informationen wie Schmelzpunkt, Molekülmasse oder Toxizität gleichzeitig einzusehen und zu analysieren.

**Bio2RDF** Das Bio2RDF-Projekt[BNT<sup>+</sup>08] hat zum Ziel, Daten die in der Domäne der Bioinformatik erzeugt wurden in das RDF-Format zu überführen, um sie dann als Teil des LOD-Projektes zu veröffentlichen. Hierzu bietet Bio2RDF Werkzeuge an, die es Wissenschaftlern ermöglichen, ihre Daten mit bereits existierenden Bio2RDF-Daten abzugleichen, und in die bestehende Linked Open Data Umwelt einzupflegen.

## 3 Szenario

In diesem Abschnitt sollen anhand eines Beispielszenarios die Anforderungen an ein Informationsmanagement-System, das Einsatz in der Domäne der Lebenswissenschaften findet, erläutert werden.

Ein Wissenschaftler in einem Pharmaunternehmen, der auf der Suche nach einem Heilmittel oder einem Impfstoff gegen das HIV-Virus ist, versucht die HIV-Infektion

---

<sup>1</sup><http://www.w3.org/TR/WD-rdf-syntax>

zu unterbinden, indem er den Prozess der Ansteckung unterbricht. Der Forscher versucht dies zu erreichen, indem er dem Organismus für eine für die Ansteckung notwendige Reaktion notwendige Katalysatoren entzieht.

Um eine Übersichtsseite über den HIV-Infektionsprozess sowie alle in dessen Verlauf stattfindenden chemischen Reaktionen zu erhalten, muss dem Wissenschaftler die Möglichkeit gegeben werden, eine strukturierte Anfrage auszuwerten. Da dem Benutzer die genaue Struktur der geladenen Daten unter Umständen nicht bekannt ist, muss er bei der Formulierung dieser Anfrage unterstützt werden. Der Wissenschaftler findet so heraus, dass sehr früh im Prozess eine Umformung der Ribonukleinsäure (RNA) stattfindet. Nachforschungen über diese Reaktion ergeben, dass das Protein Xeroderma Pigmentosum B (XPB) ein für die Reaktion notwendiger Katalysator ist. Durch Experimente verifiziert der Wissenschaftler, dass der Gesamtprozess durch Entzug von XPB unterbrochen werden kann. Er notiert eine Beschreibung der Durchführung seiner Experimente sowie deren Ergebnisse und speichert sie in der Datenbank, sodass andere Mitarbeiter diese einsehen können, um das gleiche Experiment nicht wiederholen zu müssen.

Nun versucht der Wissenschaftler weitere Informationen über das Protein XPB zu finden. Hierzu müssen Volltextsuch-Anfragen unterstützt werden, die ihm bei einer Suche nach dem Schlüsselwort "XPB" Ergebnisse sowohl aus den strukturierten als auch den unstrukturierten Teilen der Daten liefern. Zusätzlich sollen weitere Informationen über Medikamente, welche dieses Protein enthalten, und eine Liste der Reaktionen, an denen XPB beteiligt ist, geliefert werden. Der Forscher findet so heraus, dass XPB für viele im menschlichen Körper ablaufende DNA- und RNA-Reparatur-Vorgänge notwendig ist. Er sucht mit Hilfe eines Suchformulars nach chemische Verbindungen mit ähnlichen chemischen Eigenschaften wie XPB, die als Ersatzstoff in diesen Reparatur-Prozessen dienen könnten. Das Ergebnis dieser Anfrage wird mit Hilfe von Diagrammen visualisiert, was einen direkten analytischen Vergleich der gefundenen Verbindungen erlaubt. Da die Menge der so gefundenen Stoffe sehr groß ist, muss dem Benutzer die Möglichkeit gegeben werden, diese Menge auf eine Teilmenge einzuschränken, die gemeinsame strukturelle Eigenschaften haben.

Der Wissenschaftler findet auch auf diesem Wege keine geeigneten Ersatzstoffe. Da die voraussichtlichen Nebenwirkungen eines Entzuges von XPB aus dem System zu gravierend wären muss der Wissenschaftler ausgehend von der Übersicht über alle Reaktionen im Rahmen des Infektionsprozesses von neuem mit der Suche nach aussichtsreichen Reaktionen, welche unterbunden werden könnten, beginnen.

Zusammenfassend stellt die Suche nach Informationen zur Lösung von Problemen in den Lebenswissenschaften hohe Anforderungen an ein Informations-Verwaltungs-System in den Bereichen der transparenten Datenintegration, der Datenquellen übergreifenden Suche und Analyse sowie der Datenmanipulation und -annotation.

## 4 Information Workbench

Die Information Workbench ist eine Plattform für die kollaborative Verarbeitung von Informationen. Dabei werden insbesondere die folgenden Prozesse in der Interaktion mit den Informationen unterstützt:

- Integration von heterogenen und verteilten Informationsquellen,
- Erzeugung von Informationen durch den Endnutzer, z.B. in Form von wiki-basierter Dokumentation and Annotation,
- Browsing und Navigation über die aggregierten Informationen,
- Visualisierung von und Interaktion mit den Informationen über eine Vielzahl von Widgets,
- Integrierte Suche und Exploration,
- Verwaltung von Provenance, d.h. Daten über die Herkunft der Informationen.

Abbildung 1: Die Entität Koffein

In diesem Beitrag sollen die Fähigkeiten der Information Workbench hinsichtlich der Datenintegration und der verschiedenen Arten der Suche und Exploration näher erläutert werden.

#### 4.1 Datenintegration

Die Information Workbench ist in der Lage, sehr große Mengen von Daten zu integrieren. Neben der Möglichkeit der zentralisierten Integration, bei der die Datensätze in eine (lokale) Datenbank geladen werden, ist es alternativ möglich, auf verteilte Datenquellen föderiert zuzugreifen. Dies hat den Vorteil, dass neue Datenquellen einfach integriert und wieder entfernt werden können. Die restlichen Datenquellen bleiben davon unberührt.

Eine dritte Möglichkeit ist das Einbinden von entfernten Datenquellen über einen SPARQL-Endpunkt. Dies stellt den leichtgewichtigen Integrationsmechanismus dar, da lokal keinerlei Veränderungen vorzunehmen sind, und das Einbinden und wieder Entfernen aus der Datenbank-Föderation somit sehr schnell und einfach geht. Die Information Workbench erlaubt beliebige Kombinationen dieser drei Integrationsmöglichkeiten. Eine Performance-Analyse der verschiedenen Möglichkeiten der Datenintegration wurde in [HMZ10] durchgeführt. Sowohl zur lokalen Datenspeicherung als auch für die Föderierung wird das Sesame-Framework benutzt<sup>2</sup>.

Abbildung 1 zeigt exemplarisch die Detailseite für die Verbindung Koffein. Sie stellt in verschiedenen Widgets die aggregierten Daten aus den Datenquellen integriert dar: Der integrierte Datengraph zeigt strukturierte Daten und Relationen zwischen den Entitäten, ein (semantisches) Wiki-Widget ermöglicht Zugriff auf semi- und unstrukturierte Informationen. Widgets für chemische Verbindungen zeigen automatisch die chemische Struktur der Substanz. Die Auswahl geeigneter Widgets erfolgt automatisch in Abhängigkeit vom Typ der Entität.

#### 4.2 Suchparadigmen

Eine der größten Stärken der Information Workbench ist die Unterstützung sehr unterschiedlicher Suchparadigmen für verschiedene Arten von Informationsbedürfnissen, angefangen von einfacher Volltextsuche über Formular gestützte Suche bis hin zur Unterstützung von SPARQL als strukturierte Anfragesprache [TMH10]. Im

<sup>2</sup><http://www.openrdf.org>

Folgenden sollen die Suchparadigmen detailliert vorgestellt und hinsichtlich ihrer Eignung für verschiedene Informationsbedürfnisse bewertet werden.

**Hybride Suche:** Die Information Workbench unterstützt Volltextsuche sowohl auf den strukturierten RDF-Daten, als auch auf den unstrukturierten textuellen Daten, welche für jede Entität gespeichert werden können. Die Volltextsuche arbeitet Entitätszentriert, d.h. Ergebnisse einer Volltextsuche sind Entitäten, bei denen die gesuchten Wörter in die Entität beschreibenden Attributen oder im assoziierten unstrukturierten Teil der Daten vorkommen. Die Ergebnisse werden nach Relevanz bezüglich der Suchwörter geordnet. Desweiteren hat der Benutzer die Möglichkeit, das Ergebnis einer solchen Suche mit Hilfe dynamisch aus den strukturierten Daten erzeugter facetierter Suche weiter zu verfeinern. Volltextsuche bietet einen sehr einfachen Zugang zu den Daten, auch für Benutzer die mit analytischen und strukturierten Suchverfahren wenig vertraut sind, oder keinerlei Kenntnis über die Struktur und das Schema der unterliegenden Daten haben. Daher kann die Volltextsuche ein geeigneter Einstiegspunkt beim Arbeiten mit neuen, unbekanntem Daten sein, um sich einen schnellen Überblick zu verschaffen und um Einstiegspunkte in den Datengraphen zu finden, von denen aus die Daten weiter exploriert werden können.

Für die Realisierung der Volltextsuche wurde die Apache Lucene Bibliothek, eine Java-basierte Volltext-Engine, benutzt<sup>3</sup>. Sowohl der RDF-Graph als auch alle Wiki-Seiten werden indexiert und mit einer Entität assoziiert.

Ein Beispiel hierfür wäre die Suche nach einem sehr allgemeinen Schlüsselwort wie "drug", um sich einen schnellen Überblick über Medikamente und verwandte Daten zu verschaffen. Der Benutzer kann sich dann einige Ergebnisse anschauen, und sich so einen genaueren Überblick über die verfügbaren Daten und deren Struktur verschaffen, um dann in der Lage zu sein, präzisere Anfragen zu formulieren, oder direkt, unterstützt durch facetiierte Suche, die Ergebnismenge auf für ihn interessante Teile der Gesamtmenge einzuschränken.

**Expressive Suche ohne Schemakennnisse:** Die Information Workbench unterstützt die RDF-Anfragesprache SPARQL in vollem Umfang. SPARQL ist eine deklarative Anfragesprache, welche in ihrer Struktur Anfragesprachen anderer Datenmodelle, wie zum Beispiel SQL, ähnelt. Diese Art der Suche lässt eine in höchstem Maße präzise Formulierung von Anfragen zu, welche zu sehr exakten Ergebnissen führen. Allerdings kann vom Endnutzer nicht erwartet werden, dass er Erfahrung im Umgang mit komplexen Anfragesprachen hat. Desweiteren ist für das Formulieren korrekter SPARQL-Anfragen detailliertes Wissen über das Schema der Daten notwendig. Auch dies ist bei Benutzern von Informationsmanagement Systemen nicht zwingend der Fall. Daher bietet die Information Workbench verschiedene Möglichkeiten, um den Benutzer bei der Formulierung solcher Anfragen zu unterstützen. Diese reichen von einer formular-gestützten Suche, welche dazu geeignet ist, häufige und somit wahrscheinliche Arten von Anfragen zu formulieren, bis hin zur Interpretation von Schlüsselwörtern, bei der versucht wird eine unstrukturierte Eingabe des Benutzers in eine konkrete strukturierte Anfrage zu übersetzen und diese auszuwerten. Hierzu wird ein Strukturindex verwaltet, welcher es ermöglicht Schlüsselwörter auf Knoten des RDF-Graphen zu matchen und durch eine Graph-Exploration verbindende Subgraphen zu finden, welche Datenelemente für alle Schlüsselwörter enthalten. Theoretische Grundlagen dieses Verfahrens können in [TWRC09] nachgelesen werden.

Auch bei diesen Arten der Suche ermöglicht die Information Workbench zusätzlich das Verfeinern der gefundenen Ergebnisse durch facetiierte Suche.

Ein Beispiel hierfür ist die Suche nach allen Entitäten vom Typ "Compound", zusammen mit ihrem Schmelzpunkt und ihrer chemischen Formel. Der Benutzer hat verschiedene Möglichkeiten um zum gesuchten Ergebnis zu gelangen: Entweder er

---

<sup>3</sup><http://lucene.apache.org/java/docs>

formuliert selbst eine SPARQL-Anfrage, oder er erzeugt mit Hilfe eines Suchformulars durch Eingabe des Typs "Compound", eine Liste aller chemischen Verbindungen und wählt zusätzlich aus der Liste der Eigenschaften die angezeigt werden sollen die gewünschten Attribute "meltingPoint" und "chemicalFormula", oder er gibt die Anfrage "compound meltingpoint formula" ein und lässt diese als strukturierte Anfrage interpretieren. All diese Wege führen zum gleichen Anfrageergebnis, welches dann weiter verfeinert werden kann, beispielsweise auf Elemente, deren Schmelzpunkt in einem vorgegebenen Intervall liegen.

**Suche nach chemischen Verbindungen:** Speziell für Anwendungen in Chemie-affinen Domänen wie den Lebenswissenschaften wurde die Information Workbench mit einem Suchinterface für chemische Verbindungen ausgestattet. Hierbei hat der Benutzer die Möglichkeit, ein Molekül zu zeichnen und nach diesem Molekül in der Datenbank zu suchen. Hierzu wird die SMILES-Codierung der gezeichneten Struktur berechnet und mit Einträgen in der Datenbank verglichen.

Dieses Interface ist sehr speziell und domänenspezifisch. Es findet genau dann Anwendung, wenn dem Benutzer die Struktur einer gesuchten chemischen Verbindung bekannt ist, er aber deren Namen nicht kennt. Eine Suche nach Sub- oder Superstrukturen wäre an dieser Stelle ebenfalls denkbar.

## 5 Schluss

In diesem Beitrag haben wir die Information Workbench als Plattform für die Verwaltung von Informationen in den Lebenswissenschaften vorgestellt, der Fokus lag dabei auf Aspekten der Informationsintegration und neuen Paradigmen der Suche.

Zur Zeit wird die Information Workbench in Fallstudien von Forschungsprojekten (z.B. CollabCloud<sup>4</sup>) eingesetzt. Um die Verbreitung zu erhöhen und einfache Erweiterungen zuzulassen, wird der Kern der Information Workbench als Open Source Projekt entwickelt<sup>5</sup>. Die Erweiterung des Kerns ist über ein SDK einfach möglich, so können neue Datenquellen über zusätzliche Provider oder neue Widgets zur Visualisierung und Interaktion integriert werden. Die Entwicklung von kommerziellen Erweiterungen und Produkten ist möglich und geplant. Ein öffentlicher Demonstrator der Information Workbench mit freien Datensätzen aus den Lebenswissenschaften ist zugänglich unter [http://iwb.fluidops.com/index\\_ls.html](http://iwb.fluidops.com/index_ls.html).

## Literatur

- [BNT<sup>+</sup>08] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault und J. Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716, 2008.
- [HMZ10] Peter Haase, Tobias Mathäß und Michael Ziller. An Evaluation of Approaches to Federated Query Processing over Linked Data. In *To appear at the I-SEMANTICS 2010*, September 2010.
- [TMH10] Thanh Tran, Tobias Mathäß und Peter Haase. Usability of Keyword-Driven Schema-Agnostic Search. In *Proceedings of the 7th Extended Semantic Web Conference, ESWC 2010*, Seiten 349–364, 2010.
- [TWRC09] Thanh Tran, Haofen Wang, Sebastian Rudolph und Philipp Cimiano. Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data. In *ICDE*, Seiten 405–416. IEEE, 2009.

---

<sup>4</sup><http://www.collabcloud.de>

<sup>5</sup><http://code.google.com/p/iwb>