arbido

2011/3 Elektronische Bibliothek Schweiz

Belmonte Javier,

Ricci Fabio,

Schneider René,

RODIN – benutzerdefinierte Zusammenstellung von Informationsquellen unter Verwendung bibliographischer und enzyklopädischer Ontologien

Das Projekt RODIN (= ROUe D'INformation) ist seiner Definition nach ein anwendungsorientiertes Forschungsprojekt und hat das Ziel, eine alternative Portalidee zu realisieren, die die benutzerdefinierte Suche in heterogenen Informationsquellen erlaubt. Die Grundidee des alternativen Ansatzes fusst auf dem Einsatz von Semantic-Web-Technologie.

1. Einleitung

Für ein besseres Verständnis des Projekts ist weniger das Akronym selbst als vielmehr dessen Auflösung «Roue d'Information», also Wissensrad, sowie dessen wissenschaftshistorische Vorläufer hilfreich. Diese finden sich – in einer konkreten, physikalischen Ausprägung – im Bücherrad von Agos- tino Ramelli (1531–1600) sowie – in einer abstrakten-ideellen Ausprägung – in der Ars Magna von Raimundus Lullus (1232–1316).

Beide Werkzeuge verfolgten das Ziel, das Wachstum an verfügbarer Information handhabbar zu gestalten, indem sie versuchten, der Aufsplitterung des Wissens einen neuen systematischen Rahmen zu geben. Sowohl Ramelli als auch Lullus war bewusst, dass die geschlossene und begrenzte Welt des Mittelalters an etlichen Stellen aufgebrochen war, dass sie einen erheblichen qualitativen und quantitativen Zulauf erhielt und in ihrer geschlossenen Einheitlichkeit gefährdet war. So ist Ramellis Bücherrad als Werkzeug zu verstehen, dass der heraufziehenden Wissensexplosion Rechnung trägt, wohingegen Lullus den sowohl post-medievalen als auch prä-rinascimentalen Versuch unternahm, alle Gegensätzlichkeiten unter einem einheitlichen und allumfassenden Dach systematisch zu erfassen. Er kann damit zu Recht als Vorläufer der für die Gegenwart des Webs so wichtigen Ontologien und als Urahn des Semantic Web angesehenwerden.

Dem Denken von Ramelli und Lullus liegt ein Problem zugrunde, das für RODIN eine tragende Rolle spielt: Wie lassen sich unterschiedliche Informationsquellen und divergierendes Wissen miteinander verbinden; worin bestehen ihre Zusammenhänge; wie lassen sie sich in ihrer Singularität darstellen, ohne dass der Blick für das Ganze verloren geht?

2. Konzept

Diese Fragen sollen auch in RODIN beantwortet werden. Aus Gründen der Anschaulichkeit werden sie in ihren Grundlagen im Folgenden an den Mock-Ups (das sind, vereinfacht gesagt, die ersten graphischen Entwürfe) für die mobile Version von RODIN illustriert, die kurz vor dem Zeitpunkt dieser Veröffentlichung erstellt wurden. Auch wenn die Entwicklung von RODIN zu diesem Zeitpunkt weit vorangeschritten ist und ein webfähiger, erster Prototyp bereits zur Verfügung steht, bieten sich zur besseren Verständlichkeit der Systemgrundlagen die Mock-Ups der mobilen Version an, da sie den Entwickler dazu zwingen, das System und dessen Benutzerschnittstelle auf die Grundfunktionalitäten zu reduzieren.

Im Layout der Startseite wird ersichtlich, dass für RODIN letztlich drei Grundkomponenten von Bedeutung sind: an zentraler und oberer Stelle und von daher in mittiger Position die Suche, genauer die Suche in Widgets und Ontologien, die jeweils rechts und links positioniert sind.

Grundsätzlich wird bei der Gesamtkonzeption von RODIN und RODIN-mobile davon ausgegangen, dass jede Informationsrecherche, insbesondere die wissenschaftliche Suche nach Information, für gewöhnlich *mit* bzw. *in* einer zwar variablen und dynamisch veränderbaren, aber limitierten Anzahl von Informationsquellen stattfindet. Diese weichen je nach Benutzer voneinander ab, d.h., dass das System an vorderster Stelle dem Benutzer die Möglichkeit geben muss, diese einzelnen Suchquellen selbst zu definieren, quasi zu abonnieren. Diese Möglichkeit wird über die Funktionalität der Widgets gewährleistet.

Des Weiteren verlangt insbesondere die wissenschaftliche Recherche nach einer Möglichkeit, eine begonnene Suche kontextuell zu erweitern, zu vertiefen oder auszuweiten, also verwendete Suchbegriffe oder gefundene Dokumente in einen begrifflichen bzw. semantischen Kontext zu setzen. Diese Vorgehensweise wird über die Funktionalität der Ontologien gewährleistet.

3. Widgets

Der Benutzer kann also sowohl aus ei- ner Reihe von Suchquellen, (in der Desktop-Version von RODIN haben diese die Form von Widgets) als auch aus einer Auswahl von Ontologien seine individuelle Suchumgebung gestalten.

Dazu muss er zunächst die Informationsquellen auswählen, die er für die Recherche verwenden möchte, und diese in einem sogenannten Informationsuniversum zusammenstellen bzw. aggregieren. Im englischsprachigen Kontext und verwandten Systemen wie etwa Netvibes oder iGoogle wird in diesem Zusammenhang häufig auch von Dashboards gesprochen. Bei den einzelnen Komponenten dieser Dashboards kann es sich um Suchmaschinen, Informationsportale, RSS-Feeds und Ähnliches handeln, im Kontext von RODIN muss es sich aber um Informationsquellen mit Suchfenstern handeln, die über ihre Schnittstellen mit RODIN verknüpft sind.

Hinter der Funktionalität der Widgets verbirgt sich also nichts anderes als eine etwas anders geartete und differenziert visualisierte Meta-Suche. Eine Besonderheit liegt jedoch darin, dass eine Hauptintention des Projekts RODIN darin besteht, dem Benutzer eine Möglichkeit zu bieten, die Applikationen der Einzelprojekte von e-lib.ch: Elektronische Bibliothek Schweiz und deren Suchmaschinen mit anderen Webangeboten, die für ihn von Interesse sind, in einer Suchumgebung zu kumulieren, ohne direkt auf eine Indexierung oder ein Harvesting angewiesen zu sein. Dabei hat der Benutzer zudem die Möglichkeit, die Suchumgebungen bzw. sein Informationsuniversum über Reiter in einer Projekt- oder Dossierstruktur zu organisieren und projekt- oder arbeitsbereichsorientiert auszugestalten und dann – von einer einfachen Suchmaske aus – seine Suchen zu starten.

Nach jeder Suche werden dann die Ergebnisse der einzelnen Suchmaschinen ausgegeben. In der Desktopversion geschieht dies separat innerhalb der einzelnen Widgets, in der mobilen Version soll dies in einer fusionierten und dedoublierten Ergebnisliste geschehen. Der Benutzer hat dabei die Möglichkeit, über eine Tastenfunktion auf dem Display genauere Angaben über das Dokument und die dahinter befindliche Informationsquelle zu erhalten.

4. Ontologien

Simultan zu dieser einfachen Meta-Suche wird für jeden Suchbegriff jedoch auch eine Suche innerhalb einer oder mehrerer Ontologien angestossen. Dabei handelt es sich ausschliesslich um On- tologien, die auf Grundlage bibliographischer Daten, also aus Thesauri oder Taxonomien der Sacherschliessung unter Verwendung des SKOS-(= Simple Knowledge Organization System)Datenmodells, zur Verfügung stehen. Das SKOS-Datenformat hat sich in den vergangenen Jahren als Quasi-Standard herausgebildet, und mittlerweile stehen eine ganze Reihe wichtiger Datensätze in dieser Form zur Verfügung. An vorderster Stelle sind dabei die Library of Congress Subject Headings (LoCSH) oder die – für RODIN besonders interessanten und in das bestehende System integrierten – Datensätze des Standardthesaurus Wirtschaftswissenschaft (STW) oder DBPedia (einer im Semantic-Web-Format vorliegenden Untermenge von Wikipedia) zu nennen. Diese können analog zu den einfachen Suchquellen in RODIN vom Benutzer spezifiziert werden.

Diese Einbindung bedeutet konkret, dass nach semantischen Erweiterungen des Suchbegriffs gesucht wird. Dies umfasst Begriffe, die in den Ontologien unter den Kategorien «Broader», «Narrower» und «Related» einander zugeordnet sind, also Hypernyme, Hyponyme und Synonyme.

Das Auffinden dieser Suchbegriffe ermöglicht es dem Benutzer, seine Suche gezielt weiterzuführen, um sie – je nach Intention – zu erweitern bzw. zu extensivieren (Broader), zu vertiefen bzw. zu intensivieren (Narrower) oder auszuweiten bzw. zu expandieren (Related).

Dazu wählt der Besucher die Terme aus der Ontologie aus, die seine Suche in eine ihm relevant erscheinende Richtung lenken. Die gewählten Terme erscheinen in einer Brotkrümel-Navigationsleiste, von der die Terme wieder gelöscht werden und von der aus die Suche auch wieder angestossen werden kann. Als Brotkrümel-Navigation wird dabei ein Navigationselement verstanden, das dem Benutzer als Orientierung seinen Navigationskontext anzeigt. In der Desktop-Version hat der Benutzer darüber hinaus die Möglichkeit, durch einen rechten Mausklick einzelne Wörter aus den Ergebnissen zur Erweiterung der Suche in die Brotkrümel-Leiste hinzuzufügen.

Eine letzte Möglichkeit zur Vertiefung der Suche besteht darin, ein ganzes Element der Ergebnisliste für eine neue Suche auszuwählen. In diesem Fall wird – teils unter Verwendung eines eigens entwickelten Algorithmus, teils unter Verwendung von DBPedia Spotlight – überprüft, welche Schlagworte aus den Ontologien zu den Termen des Einzelergebnisses passen, und diese werden dem Benutzer auf der Brotkrümel-Leiste zur erneuten Suche vorgeschlagen. Auch hier hat der Benutzer die Möglichkeit, einzelne Begriffe zu löschen.

5. Architektur

Im Folgenden sollen einzelne technische Besonderheiten erläutert werden, die für ein genaueres Verständnis der Systemarchitektur von Wichtigkeit sind. Diese betreffen, wie in den vorangegangenen Abschnitten, die Anbindung der Widgets sowie die Einbindung der Ontologien.

Zur Realisierung von RODIN wurde Portaneo, ein auf PHP/AJAX basierendes Open-Source-Widgetportal eingesetzt und erweitert. Zur Implementierung des Rahmenwerks wurden PHP/AJAX auf einem Apache Server verwendet. Zur Anbindung der Widgets wurde in RODIN ein Rahmenwerk geschaffen, das es erlaubt, ausgehend von einer Suchquelle und den vorhandenen Abfragemöglichkeiten (RSS, SRU, API) sowie Output-Formaten (wie etwa XML, HTML, RDF u.Ä.) einen homogenen Ergebnis-Stream über eine eigene RODIN-Datenbank abzulegen, die Ergebnisse im Widget in homogener Form zu präsentieren bzw. bei Bedarf zu einem späteren Zeitpunkt effizient wieder abzurufen. Um wichtige Such- bzw. Filterungsaspekte einer Datenquelle zu berücksichtigen, erlaubt das Rahmen- werk im Widget die Gestaltung benutzerspezifischer Sucheinstellungen, die sich auf die Ergebnisdokumente auswirken.

Bei der Anbindung der Suchquellen mittels Widgets wird auf das Vorhandensein von RESTful APIs grosser Wert gelegt, obwohl derzeit von einem flächendeckenden Einsatz solcher APIs leider nicht ausgegangen werden kann. Das Widget-Rahmenwerk bietet deshalb eine Reihe von Werkzeugen an, die es gestatten, die unterschiedlichen Formate in ein eigenes Format zu verarbeiten.

Für die Anbindung einer Ontologie müssen – wie bei einer über ein Widget eingebundenen Suchquelle – deren Format und die damit verbundenen Abfragemöglichkeiten beachtet werden. Bei der Anbindung muss weiterhin berücksichtigt werden, ob die Ontologie auf dem Web über einen SPARQL-Endpoint oder lediglich als Datei verfügbar ist. Für RODIN wurden zunächst Ontologien im RDF-Format eingebunden: STW und DBPedia. Gemeinsam ist diesen Ontologien, dass sie in SPARQL abgefragt werden können. Dabei ist es nicht unbedingt notwendig, dass eine Ontologie in RDF vorhanden ist, sofern sie in SKOS spezifiziert ist.

Die Verwendung einer in SKOS spezifizierten Ontologie wird beschränkt auf folgende wenige, aber klare Fälle: Zu einem Suchbegriff bestehend aus einem einfachen oder zusammengesetzten Wort oder aus mehreren Wörtern werden die dazu in einer Relation («Related», «Broader» bzw. «Narrower») stehenden Termen gesucht. Die Eingabe für diese Suche besteht aus dem Inhalt des RODIN-Suchfelds zuzüglich einschränkender Terme aus der Brotkrümel-Leiste.

Zum schnellen Wiederauffinden der Terme, besonders bei grossen Ontologien, muss gewährleistet werden, dass die Terme und Relationen in einer adäquaten Speicher- und Ausführungsumgebung indexiert zur Verfügung stehen. Dazu eignen sich u.a. Triple-Stores (d.h. eine datenbankbasierte Umgebung, die über einer Abfragesprache – etwa SPARQL – erlaubt, an den Inhalt einer Ontologie heranzukommen), bei denen die Einzelkomponenten der Ontologie in RDF als klassische Triple bestehend aus Subjekt, Prädikat und Objekt dargestellt werden. Diese Form der Darstellung wird «semistrukturiert» genannt; sie erlaubt eine schnelle, SQL-ähnliche Wiederauffindung mittels SPARQL in einem SPARQL-Endpoint.

Bei der Anbindung der STW-Ontologie wurde ein Local Triple Store mittels ARC (d.h. eines auf PHP basierenden Rahmenwerks zur Verwaltung und Abfrage von RDF-Informationen mittels eines eigenen SPARQL-Endpoints) eingerichtet, der alle Informationen aus der STW-Ontologie indexiert zur Verfügung stellt. Da DBPedia über einen eigenen (Remote) Triple Store samt SPARQL-Endpoint verfügt, wurde jede Ontologie-Suchanfrage über ihren Remote Endpoint abgewickelt.

Die Ergebnisse aus den Suchanfragen beider Ontologien können dann nach ihren Relationen geordnet («Broader», «Narrower» oder «Related») dem Benutzer in einem Dialog präsentiert werden, aus dem dieser einen oder mehrere Terme auswählen und über die RODIN-Brotkrümel-Leiste zur Verfeinerung der Suche verwenden kann.

Die Software selbst kann dabei, entsprechend der allgemeinen Gepflogenheiten, in zwei Anwendungskontexten eingesetzt werden. Entweder in einer sogenannten «out-of-the-box» oder als «customized version»:

So wird RODIN zunächst in der während der Projektzeit entwickelten Form «out-of-the-box» als allgemeines Webportal zur Verfügung stehen. Anhand dieses allgemeinen Portals kann das Potential der dahinter befindlichen Algorithmen getestet und auf seine Verwendungsfähigkeit überprüft werden. Da der Quell-Code auf einer Open-Source-Plattform bereitgestellt wird, ergibt sich gleichzeitig die Möglichkeit, einzelne Module herauszulösen und in bestehende Systeme komplementär einzubinden.

Andererseits kann die gesamte Software auf einzelne Informationsportale zugeschnitten werden («customized version») und auf den Websites dieser Portale jene Widgets und Ontologien integrieren, die für den jeweiligen Wissensbereich (bspw. Geschichts- oder Wirtschaftwissenschaften) von Interesse sind.

6. Zusammenfassung und Ausblick

Wie erläutert, vereint RODIN letztlich zwei Suchstrategien: eine einfache, aber benutzerdefinierte Meta-Suche sowie eine ontologiegesteuerte fortgeschrittene Suche, die auf bibliographischen und enzyklopädischen Ontologien beruht. Der hohe Innovationscharakter und die damit verbundene Komplexität der Benutzerschnittstelle sowie der dahinter befindlichen Algorithmik setzen unter Umständen eine gewisse Einarbeitungszeit oder eine Schulung voraus. RODIN versteht sich von daher als Werkzeug für Informationsspezialisten und Benutzer, die ein differenziertes und differenzierendes System zur explorativen Recherche benötigen, das ihnen als Komplement zu herkömmlichen «einfachen» Suchmaschinen dienen kann.

Daraus folgt, dass neben dem Benutzer der Informationsspezialist eine herausragende Rolle spielen wird:

a) im Zusammenstellen, Lizenzieren und Publizieren von Widgets,

b) in der Schulung der Endbenutzer bei der Verwendung von RODIN und

c) in der Bereitstellung von Ontologien, die das Resultat seiner Arbeit in der Sacherschliessung darstellen.

Gerade der letzte Punkt unterstreicht den nicht zu unterschätzenden Beitrag, den die Sacherschliessung zur Weiterentwicklung des Webs bieten kann.

Javier Belmonte

HEG Genève

Fabio Ricci

HEG Genève

René Schneider

HEG Genève

Abstract

Français

Le projet RODIN (ROUe D'INformation) de la HEG Genève, filière Information documentaire, est un projet de recherche orienté sur l'application des technologies. Il vise à réaliser l'idée d'un portail alternatif qui permettrait aux utilisateurs de rechercher dans des sources d'informations hétérogènes. L'idée de base de cette approche alternative repose sur l'utilisation des technologies du Web sémantique. Elle est délimitée par les deux stratégies conventionnelles de récupération des données, c'est-à-dire d'une part par l'idée d'une indexation des moteurs de recherche et d'autre part par les systèmes Harvesting. Ces deux stratégies ne jouent aucun rôle dans le contexte du projet RODIN, mais plutôt la combinaison des sources d'information et des flux d'information, définie par l'utilisateur, ainsi que l'affinement des résultats de recherche traditionnelle au moyen d'ontologies bibliographiques.

Vers la fin de l'année 2011, RODIN sera disponible comme prototype test capable de fonctionner. Le code du programme sera rendu accessible au public sur une plateforme open source en vue de son intégration et de son développement. L'accès au système à des fins de test peut d'ici-là être demandé aux auteurs de l'article. Après la première phase du projet – sous condition d'un financement ultérieur – le système pourra être équipé d'une infrastructure robuste qui permettrait une utilisation simultanée par un grand nombre d'utilisateurs. En outre, RODIN devrait être disponible dans une version pour mobile, celle présentée dans cette publication.