



**Ulrich Wirth M.A.,**  
*wissenschaftlicher*  
*Dokumentar,*  
*Leiter der Euro-*  
*Schulen Trier, Höhere*  
*Berufsfachschule*  
*für Medizinische*  
*Dokumentations-*  
*assistenten*  
*E-Mail:*  
*wirth.ulrich@*  
*es.trier.eso.de,*  
*www.euroschulen-*  
*trier.de*

## Warum der Datenpool wieder zum Planschbecken werden muss

**Semantic-Web-Technologien gestalten die Recherche nach medizinischen Fachinformationen effizienter**

**W**eil die Schere zwischen Information und Redundanz insbesondere im Internet immer weiter aufzugehen scheint, wird es zunehmend mühsamer, mit dem geringstmöglichen zeitlichen wie ökonomischen Aufwand aktuell benötigte Informationen zu finden. Medizin und Life Sciences stellen da keine Ausnahme dar. Die Suche nach relevanten medizinischen Fachinformationen ist aufwendig und verursacht mitunter Kosten, die eigentlich unnötig sind. Abhilfe schaffen intelligente Recherche-Tools, die sich Semantic-Web-Technologien bedienen. Wie Recherche erleichtert wird, welche Möglichkeiten und Einsparpotentiale sich eröffnen, wenn Datenbanken wie Medline mit Text-Mining-Verfahren kombiniert werden, zeigt der folgende Beitrag am Beispiel des SsynX Knowledge Dashboard 2.0.

### **Das Aschenputtel-Konzept: Informationsvermittlung in Zeiten von Informationsflut**

»Wir informieren uns zu Tode.« [1] Noch vor dem Durchbruch des Internets warnte der Kommunikationswissenschaftler Neil Postman vor den negativen Begleiterscheinungen des medialen Fortschritts, einer Flut an Informationen ohne Erklärungen und Informationswert. Doch anstatt in einer Mischung aus Befreiungsschlag und Verweigerungshaltung buchstäblich den Stecker zu ziehen, muss diese Informationsüberflutung aktiv bewältigt und damit gemanagt werden, da Informationen und Wissen einen wesentlichen Faktor innerhalb der wirtschaftlichen Wertschöpfungskette darstellen. Selektion tut daher not, wobei diese Erkenntnis nicht neu ist: »Die guten ins Töpfchen, die schlechten ins Kröpfchen«, lautet ein Rezept im Märchen. Übersetzen wir gut mit relevant, schlecht mit irrelevant, verwandelt sich die Handlungsanweisung der Brüder Grimm in einen Hauptsatz der dualistischen Orientierungsstrategie modernen Information Retrievals. Dieser ist aktueller denn je. Zu den Recherchestrategien gehören nach wie vor auch Strategien zur Informationsvermeidung. Die Hauptaufgabe von Informationsvermittlern [2] besteht darin, im Data Smog die irrelevante Asche von den relevanten Linsen zu trennen.

Spätestens das Internet hat den Selektionsdruck noch erheblich vergrößert. Informationen sind allgegenwärtig, rein quantitativ betrachtet entwickelt sich die Zahl der

Informationen exponentiell. Die kognitive Verarbeitungskapazität des Menschen hat mit dieser Entwicklung indessen nicht Schritt halten können [3]. Alle Informationen verarbeiten zu wollen, überforderte jeden einzelnen Konsumenten, Aschenputtel müsste ihre Tauben sozusagen rund um die Uhr im Dreischichtsystem beschäftigen. Postmans Prophezeiung hat sich damit sogar übererfüllt: Zwar ist die Informationsfülle einer gefährlichen Informationsflut gewichen; doch was uns an der Schwelle zum dritten Jahrtausend droht, ist mehr: der Informationale Overkill.

### **Land unter im Cyberspace**

Um im Bild der katastrophischen Wassermetaphorik zu bleiben, könnte man diesen Zustand mit »Seenot in Permanenz« umschreiben. Rettung erhofft man sich von einem »intelligenteren« Internet: Mit dem Semantic Web ist sozusagen eine Cyberspace-taugliche Schwimmweste in der Erprobung. Das Internet muss sich flächendeckend vom bloßen Datenspeicher zum semantischen Netz weiterentwickeln, in dem Inhalte automatisch vorselektiert, thematisch aufbereitet sowie schnell und vollständig bereitgestellt sind. Von solchen wissensbasierten Systemen sind wir nicht mehr weit entfernt. Neue Methoden der Informatik und Computerlinguistik wie z.B. Text Mining basierte Anwendungen können schon heute für den nötigen informationellen Mehrwert sorgen, indem sie z.B. die in einer Datenbank angesammelten Informationen effizient und nutzerorientiert mittels semantisch aussagekräftiger Informationen strukturieren [4]. Dadurch unterstützen sie den Informationsmanager bei der Veredelung von Informationen im »Rohzustand« in handlungsrelevante Produkte. Wissen wird wieder überschaubar.

### **Wissensmanagement in Medizin und Life Sciences**

Auch in Medizin und den Life Sciences ist Wissensmanagement geschäftskritisch, da Wissen wesentlicher Bestandteil entsprechender Marktleistungen ist. Nicht nur in Universitätskliniken und Krankenhäusern, sondern auch in medizinisch-wissenschaftlichen Instituten, im öffentlichen Gesundheitswesen oder in der chemisch-pharmazeutischen Industrie werden valide Informationen



zu medizinischen und flankierenden Sachgebieten möglichst zeitnah zum Informationsbedürfnis benötigt. Dabei sitzen die Kliniker wie auch die wissenschaftlich tätigen Mediziner, Pharmazeuten und Naturwissenschaftler im selben Boot (das auf demselben Datenpool vor sich hin dümpelt), wenngleich sich die Bedürfnisse beider Gruppen beträchtlich unterscheiden:

Der klinisch tätige Arzt legt den Fokus primär auf medizinische Behandlungsformen, moderne Therapiekonzepte und Fallberichte, also auf relevante Informationen, die er unmittelbar zur Ausübung seines Berufes benötigt, bisweilen sehr schnell, denkt man an Notfälle oder die ambulante Behandlung von chronisch kranken Patienten [5]. Im Vordergrund steht also ärztliches Handeln und damit die medizinische Versorgung – weniger die Ergebnisse der Grundlagenforschung [6].

Genau darauf legt aber der Wissenschaftler sein Augenmerk, und er ist brennend an den aktuellen Entwicklungen in seinem Forschungsfeld interessiert. Um Ideen und Trends, die sich in der eigenen Branche entwickeln, überhaupt identifizieren zu können, muss er sich regelmäßig und damit wiederholt einen möglichst vollständigen Überblick über ein Themengebiet verschaffen. Trotz dieser Unterschiede: Beide Anspruchsgruppen können ohne aktuelle, relevante und qualitativ hochwertige Informationen nicht auf zuverlässiger Grundlage arbeiten. Eine weitere Gemeinsamkeit besteht in der Erwartungshaltung hinsichtlich Precision und Recall, den Gütemaßen des Information Retrievals, die bei einer Suche im medizinischen Umfeld ungleich höher sind, als dies in anderen Branchen der Fall sein dürfte [7].

### **Der Medizinische Dokumentar als Informationsvermittler**

Hier kommen die Informationsspezialisten ins Spiel. Die Informationsvermittlung ist ein grundlegender Baustein der Wissensarbeit von Medizinischen Dokumentaren [8]. Als interne oder externe Dienstleister recherchieren sie in Fakten- und Literaturdatenbanken, im Internet und in gedruckten Quellen. Sie strukturieren selbst die komplexesten Sachverhalte, beschaffen und erschließen medizinische Informationen. Falls es der Auftraggeber wünscht, bereiten sie die Ergebnisse auch grafisch oder in Form von Dossiers auf und präsentieren diese.

Dabei ist der Beruf des Dokumentars, ebenso wie vergleichbare Betätigungsfelder innerhalb der IuK-Branche, aktuell von starken Veränderungen geprägt: Kosten- und Zeitdruck sind hoch, die Konkurrenz schläft nicht, sondern lauert, Kunden werden immer anspruchsvoller. Um diesen Ansprüchen des Marktes gerecht zu werden, sind u.a. Maßnahmen des Wissensmanagements erforderlich.

Damit nicht genug, droht den klassischen Informationsstellen die »Entinstitutionalisierung« [9]. Durch die direkte Verfügbarkeit von netzfähigen elektronischen Ressourcen fühlen sich Benutzer nicht länger genötigt, ihren Informationsbedarf durch Information Professionals abzudecken. Warum Geld ausgeben? Selbst ist der Mann im Internetzeitalter! Viele glauben nunmehr, selbst recherchieren zu können, und insbesondere die sich ständig optimierende Suchmaschinentechologie hat dazu beigetragen, Recherchedienstleistungen vom Sockel zu stoßen. Doch geht es hier weniger um Google und Konsorten: Die Informationsbranche selbst hat sogar an ihrem eigenen Ast gesägt, weil sie auf die veränderten Recherchebedürfnisse reagiert und entsprechende Produkte auf den Markt gebracht hat, die intuitiv und mit beinahe schlafwandlerischer Sicherheit von jedermann zu bedienen sind. Ob Genios, Lexis-Nexis oder DIMDI SmartSearch – Datenbankrecherchen erfordern immer seltener ausgefuchste Retrievalstrategien, sondern bereiten dem Suchenden auch mit »trial and error« Erfolgsmomente.

Informationsdienstleistungen haben daher mit einem gewissen Rechtfertigungszwang zu kämpfen. Dennoch gilt: Information Retrieval in medizinischen Wissensbeständen ist nicht trivial und gehört in professionelle Hände. Ein konkretes Beispiel verdeutlicht nun, wie anspruchsvoll Informationsvermittlung sein kann.

### **Fallbeispiel: Vorbereitung eines Symposiums**

Stellen Sie sich einmal folgendes Szenario vor: Sie arbeiten in einer Klinik, die in einem interdisziplinären Netzwerk zusammen mit wissenschaftlichen Institutionen, universitären Fachbereichen und Partnern aus der Industrie neuartige Medizintechnologie entwickelt und klinisch-wissenschaftlich forscht. In regelmäßigen Abständen sollen Symposien stattfinden, auf denen die neuesten Entwicklungen im Arbeitsgebiet vorgestellt werden. Da Sie in der Fachinformationsstelle beschäftigt sind, liegt es nahe, Sie in die Vorbereitung mit einzubinden. Sie werden also damit betraut, alle Informationsdienstleistungen für die Auftaktveranstaltung zu erstellen. Konkret obliegt Ihnen die Erstellung des Tagungsprogramms. Sie sind verantwortlich für die Identifizierung und Akquise von ausgewiesenen Experten in den relevanten Sachgebieten, denn statt eines »Call for Papers«-Mailings sollen gezielt solche Redner für einen Beitrag gewonnen werden, deren Konferenz- und Seminarthemen trendsetzend sind. Davon erhofft man sich eine attraktive Veranstaltung. Ein ehrgeiziges Unterfangen, das mit möglichst wenig Aufwand erreicht werden soll, dabei rechtzeitig und schnell, damit den Experten ausreichend Zeit für die Abfassung eines wirklich spannenden Vortrags bleibt. Viel



Internetsuchmaschinen	Datenbanken
zu viele Treffer irgendwie relevanter Dokumente	höhere Relevanz durch schlagwort- und thesaurusbasiertes Retrieval
Ergebnis-Ranking nach Relevanz	strukturierte Darstellung der Suchergebnisse nach auswählbaren Kriterien (Relevanz, Erscheinungsjahr, Titel etc.)
keine Kategorisierung von Ergebnissen	Kategorisierung von Treffermengen z.B. nach Autoren, Themen, Erscheinungsjahren, Publikationstypen etc.
beschränkte Unterstützung der Boole'schen Logik zumeist einfache Benutzeroberfläche »quick and dirty«-Suche	volle Unterstützung der Boole'schen Logik teils komplexe und unkomfortable Benutzeroberfläche unter Umständen sind datenbankspezifische Recherchekennnisse bzw. Kenntnisse von Retrievalsprachen notwendig
Sprachproblematik bei Stichwort- bzw. Volltextretrieval: Suchergebnisse nur in der Sprache, in der die Suchanfrage formuliert wurde	sprachübergreifende Zusammenstellung der Suchergebnisse durch schlagwort- und thesaurusbasierte Suchen
keine oder eingeschränkte Suche nach Wortstämmen möglich	vielfältige Möglichkeiten der Trunkierung
i.d.R. keine Lemmatisierung (z.B. findet die Suche »Gehirn« nicht »Gehirne«)	oftmals Lemmatisierung
i. d. R. keine Abstandsoperatoren	Abstandsoperatoren
Unübersichtlichkeit: Suchwörter werden in der Ergebnisanzeige nicht hervorgehoben	Highlights von Suchbegriffen ermöglicht Übersichtlichkeit
keine Mindestqualität der Quellen erfordert zeitraubende eigene Qualitätsprüfung	qualitätsgesicherte Treffer, da viele Informationen aus der Fachpresse stammen
Inhalte sind nicht frei zugänglich oder können nicht gefunden werden, weil sie dynamisch aufgrund von Anfragen aus Datenbanken generiert werden (»Deep Web«-Problematik)	Datenbankinhalte sind sichtbare bzw. zugängliche Bestandteile des »Deep Web«
keine Kosten	Kosten
stärker unsystematische Suche	systematische Suche

**Tabelle 1:**  
**Rechercherelevante**  
**Merkmale von**  
**Internetsuchmaschinen**  
**und kommerziellen**  
**Datenbanken**

Vergnügen bei dieser sicherlich anspruchsvollen Recherche, die nebenbei so ziemlich alle der skizzierten Informationsbedürfnisse von Klinikern, Forschern und obendrein des Marketings vereint.

Der Researcher wird zunächst versuchen, die vom Auftraggeber gestellten Fragen (verbale Suchanfrage) in die spezifische Terminologie des Internets umzusetzen (formale Suchanfrage). Die Auftraggeber im Netzwerk spielen dabei die tragende Rolle, weil sie Forschungsfelder, Namen von Kollegen und einschlägige Fachbeiträge kennen – alles lohnende Einstiegspunkte in eine Recherche. Dem Presearch-Interview kommt demnach größte Bedeutung zu. Sodann wird der Informationsprofi seine Kenntnisse unterschiedlicher konventioneller und elektronischer Recherchetechniken einsetzen, um die gewünschten Informationen gezielt herauszufiltern. Welche elektronischen Möglichkeiten hätte er?

### Kommerzielle Internetsuchdienste

Im Web 1.0 funktioniert Recherche nach wie vor mit Bezeichnungen, nicht mit Begriffen. Wenn der Informationsvermittler alle relevanten Suchbegriffe kennt, im fraglichen Arbeitsgebiet ein ausgewiesener Fachmann ist und

genau spezifizierte Fragestellungen hat, ergibt auch eine Recherche in thematisch unspezifischen Suchmaschinen schnell hochrelevante Ergebnisse. Wenn nicht, dann nicht, was auf Kosten der Relevanz geht.

Wer ein breiter angelegtes Thema recherchiert – Stichwort Interdisziplinarität –, mit der spezifischen Fachterminologie nicht richtig vertraut ist, nur vage Vorstellungen von den relevanten Fragestellungen hat und sich darüber hinaus noch informieren möchte, was links und rechts des Kernthemas an verwandten Fragestellungen existiert, könnte auch auf ein systematisch geordnetes Verzeichnis ausweichen wollen.

Hinzu kommt die »Deep Web«-Problematik [10]. Darunter versteht man den gewissermaßen »unsichtbaren« Teil des World Wide Web, der von den Web Crawlern der Internetsuchmaschinen nicht indiziert werden kann. Das sind z.B. Inhalte, die nicht frei zugänglich sind oder dynamisch aufgrund von Anfragen aus Datenbanken generiert werden. Die Größe des »Deep Web« kann nur geschätzt werden. Es ist aber davon auszugehen, dass gerade im wissenschaftlichen Bereich der Informationsgehalt von Fachdatenbanken um ein Vielfaches größer ist als der des »Visible Web« [11].

### Gesundheitsportale

Stichwort Spezifität: Hier böte sich möglicherweise eine Recherche in einem Gesundheitsportal an [12]. Ein solches Portal fasst Informationen aus diversen angebotenen Quellen zum Themenbereich Gesundheit zusammen und ermöglicht den Nutzern mittels Browser die Recherche über Volltextsuche oder Navigation von Verzeichnisstrukturen. Oftmals erfordern die angebotenen Quellsysteme im Weiteren jeweils wieder eine spezielle Suche. Der Suchaufwand ist hoch, jedoch ist die Treffermenge durch die spezifische Filterung im Vorfeld sinnvoll eingegrenzt.

### Meta- und Clustersuchmaschinen

Einen schnellen Überblick gestatten Meta- bzw. Clustersuchmaschinen. Metasuchsysteme ermöglichen, mit nur einer Anfrage verschiedene Informationsressourcen gleichzeitig zu durchsuchen, ohne dass der Nutzer die Suchmaschine wechseln muss. Insbesondere Clustersuchmaschinen wie Surfswax oder Clusty sind interessant, weil sie die Suche nach so genannten »Konzepten« zulassen und damit auf Umwegen eine Recherche nach Begriffen ermöglichen. Unter »Konzept« versteht man die Erschließung eines breiter angelegten Themenfeldes. So wären z.B. »Alcohol-Induced Disorders« oder der dazugehörige MeSH-Code »C21.739.100.087« Konzepte zu »Alcoholic Liver Disease«,



aber auch »Insulin Resistance« oder »Hepatitis«, die nicht erhöhtem Alkoholkonsum anzukreiden sind. Denkbar sind hier zwei Szenarien. Die Suchmaschine nimmt dem Researcher die Arbeit ab, sich für die vielen unterschiedlichen Aspekte der Fragestellung einzelne begriffliche Konzepte und dazu passende, verwandte Suchwörter auszudenken. Dies ist etwa bei SurfWax der Fall. Ein anderer Zugang wäre das so genannte Clustering, bei dem anhand der Suchergebnisse thematische Kategorien gebildet und mit kurzen Überschriften versehen werden. Das ist die Vorgehensweise von Clusty.

### **Kommerzielle Datenbanken**

Blieben aus der Fülle der Recherchezugänge die Datenbanken übrig. Die in PubMed verfügbaren Datenbanken stellen für Medizin und Life Sciences einen umfassenden Datenpool dar [13]. Mit nur geringfügigen Einschränkungen kann dieser als vollständige Dokumentation der abgeschlossenen und eingeschränkt auch der aktuell laufenden wissenschaftlichen Arbeiten angesehen werden. Was im Sinne des Fallbeispiels relevant wäre, müsste folglich in PubMed zu finden sein.

In der Tabelle 1 auf Seite 162 werden die Retrievalmöglichkeiten von Internetsuchmaschinen und kommerziellen Datenbanken miteinander verglichen. Der Vergleich zeigt, welche Vorteile kommerzielle Datenbanken gegenüber Internetsuchdiensten (auch bei ausgefeilten Recherchestrategien und größtmöglichem technischem Support) bieten und dass die Vorteile trotz Kosten überwiegen. Ob sich durch die Datenbankrecherche aber die im Fallbeispiel erwähnten Trendsetter identifizieren lassen, ist fraglich, zumindest aufwendig und zeitraubend und verlangt vom Rechercheur selbst Kennerschaft.

### **Führen alle Wege nach Rom?**

Zusammenfassend lässt sich sagen, dass man über die genannten Suchzugänge oder eine Kombination aus diesen früher oder später entweder zu durchaus brauchbaren Näherungslösungen oder den gewünschten Suchergebnissen selbst gelangt. Alle diese Wege führen also irgendwie nach Rom oder zumindest in die Vororte Roms – und das irgendwann: früher oder eben später. Das erscheint doppelt problematisch. Da Zeit Geld ist, ist früher meistens besser, von unnötigen Kosten ganz zu schweigen. Und wer sich damit begnügt, nur die Vororte Roms erreicht zu haben, hat Rom selbst genau genommen verfehlt. Damit wären wir letztlich bei den Einflussgrößen Zeit und Qualität angelangt, neben Precision und Recall die ökonomischen Kriterien, mit denen eine Recherche bewertet werden muss.

Je nach Schwierigkeitsgrad der Suchanfrage helfen klassische Suchstrategien demnach nur bedingt weiter oder gestalten sich aufwendig. Hier kommen semantische Suchlösungen ins Spiel, die Recherche erleichtern, neue Möglichkeiten eröffnen und darüber hinaus Einsparpotentiale bergen. Bevor wir uns konkret der Recherche im SyynX Knowledge Dashboard 2.0 widmen, kurz einige Bemerkungen zur begrifflichen Klärung.

### **Was ist semantisch am Semantic Web?**

Semantische Suchlösungen verfolgen das Konzept, verschiedene Ressourcen (Dokumente, Bilder, Menschen, Konzepte etc.) semantisch zu verknüpfen [14]. Konsequenz angewandt, bedeutet das nicht weniger als den Übergang vom gegenwärtigen Web mit seinen einfachen Hyperlinks zu einem semantisch angereicherten Web, welchem Schritt für Schritt Bedeutung hinzugefügt wird, indem Beziehungen zwischen den verteilt gelagerten Ressourcen geknüpft werden, deren einzelne kontextuellen Bezüge bereits im gegenwärtigen Web angelegt sind.

Diese Beziehungen können von intelligenten Suchmaschinen mit Hilfe von semantischen und statistischen Verfahren lesbar gemacht werden, so dass sie den Textinhalt automatisch »verstehen«, wodurch Automatisierung auf der Ebene der Semantik ermöglicht wird. Dadurch können z.B. Suchergebnisse auch solche Dokumente einschließen, die den ursprünglichen Suchterm zwar nicht beinhalten, im Sinne der Suchanfrage aber dennoch relevant sind. Ähnlich wie schon bei den Cluster-Suchmaschinen können zudem verwandte Themengruppen angezeigt werden, wodurch dem Rechercheur die Navigation in den möglichen Ergebnismengen erleichtert wird. Bei einer Suche nach »Fatty Liver, Alcoholic« würde sowohl auf »Fatty Liver« als auch auf verwandte Konzepte wie »Digestive System Diseases« oder »Diabetes Mellitus« verwiesen werden – ähnlich wie in einem Thesaurus mit seinen Ober- und Unterbegriffen sowie verwandten Begriffen. Dadurch lassen sich Suchanfragen derart verfeinern, so dass die Treffermengen auf wenige, hochrelevante Dokumente beschränkt werden, die bei herkömmlichen Suchen erst mühsam in umfangreichen Ergebnislisten hätten identifiziert werden müssen.

### **PubMed revisited – Semantische Suche mit dem SyynX Knowledge Dashboard 2.0**

Wie erwähnt, lässt sich PubMed als annähernd vollständige Dokumentation von abgeschlossenen wie aktuell laufenden wissenschaftlichen Arbeiten betrachten. Was im Sinne einer x-beliebigen medizinischen Suchanfrage

The screenshot shows the SynnX Knowledge Dashboard 2.0 interface. At the top, the subject is set to 'Diabetes'. Below the search bar, there are filters for 'MESH' (Disease Management, Type 2 Diabetes Mellitus) and 'PubType'. A list of 8 publications is displayed, including titles like 'Benefits of information technology-enabled diabetes management' and 'Metabolic syndrome: Diagnosis, potential markers and management-an update'. On the right, there are sections for 'Profiles', 'Experts', 'Journals', 'PubType', and 'Locations'. A 'MeSH concepts (by semantic types)' sidebar lists various categories such as Disorders, Chronic Disease, Weight Loss, Cardiovascular Diseases, Proteinuria, Hyperglycemia, Obesity, Insulin Resistance, Disease, Stroke Volume, Diabetes Complications, Hypertension, Diabetic Retinopathy, Syndrome, Heart Failure, Congestive Heart Failure, Diabetes Mellitus, Vascular Diseases, Weight Gain, Diabetic Nephropathies, Diabetes Mellitus, Type 1, Prediabetic State, and Kidney Failure, Chronic. Below this, 'Chemicals & Drugs' lists Hypoglycemic Agents, Losartan, Hemoglobin, and Recombinant Insulin.

This screenshot provides a detailed view of 'Kidney Diseases' within the SynnX Knowledge Dashboard 2.0. The 'Browse MeSH' section shows 'Kidney Diseases (3033)', 'Kidney Diseases, Cystic (45)', and 'Polycystic Kidney Diseases (130)'. A 'Publication Timeline' chart shows activity from 1995 to 2013. Below the timeline, there are three columns of expert profiles: 'Currently active authors publishing since 2003' (e.g., Lewy AS, Mogensen CE), '> 200 publications' (e.g., Parving HB, Klein EE), and 'All complete overview' (e.g., Bakris GL, McCullough PA). Each profile includes the author's name and a count of publications.

**Abb. 1 (oben):** Treffermenge einer Suchanfrage nach »Disease Management« im Kontext von Diabetes und vertikaler Spezifikation nach »Diabetes Mellitus, Type 2«

**Abb. 2 (unten):** Potentielle Experten zum Thema »Kidney Diseases«

also relevant wäre, müsste sich mit an Sicherheit grenzender Wahrscheinlichkeit in PubMed finden.

Damit die Suche eine semantische wird, muss PubMed als Datenbasis also wie eben geschildert semantisch angereichert werden. Genau diesen Schritt vollzieht das SynnX Knowledge Dashboard 2.0 und kombiniert dabei die automatisierte Analyse großer Textmengen mit schnellen und vielfältigen Zugriffsmöglichkeiten. Die technische Basis für das Knowledge Dashboard bildet die Collexis Finger-Print-Technologie, eine ontologiebasierte Indexierungs- und Retrievallösung, die eigens für den Medizin- und Life-Sciences-Bereich entwickelt wurde [15]. Mit dieser Technik werden wissenschaftliche Publikationen, z.B. PubMed-Abstracts, in so genannte Fingerprints verwandelt, die die ontologiebasiert identifizierten Fachbegriffe und Bewertungsrelationen für jeden dieser Begriffe enthalten. Bei der Indexierung und in den Anwendungen kommen verschiedene Ontologien zum Einsatz, z.B. die Medical Subject Headings (MeSH), ein Thesaurus zur Sacherschließung von Büchern und Zeitschriftenartikeln in Medizin und Biowissenschaften, und der weltweit größte Metathe-

saurs Unified Medical Language System (UMLS). Dabei handelt es sich um mehr als nur ein einfaches Upgrade für PubMed. Was bringt das für eine Recherche, oder anders gefragt: Was kann das Dashboard, was PubMed nicht kann?

## Literaturrecherche nach Konzepten

Der klassische Rechercheanstieg ist der themenbezogene. Abb. 1 zeigt das Ergebnis einer Literaturrecherche nach dem Suchbegriff »Disease Management« im Kontext von Diabetes mit weiterer Spezifizierung nach »Diabetes Mellitus, Type 2«. Anschließend können wie gewohnt die dazugehörigen Abstracts aus PubMed aufgerufen werden.

Ein Mehrwert des Dashboards besteht in der übersichtlichen Anzeige aller biomedizinischen Konzepte, die in den 1.198 Treffern vorkommen. Die Aktivierung des Reiters »Profiles« lässt nun die vertikale Recherche innerhalb der Treffermenge zu, und zwar nach dem normierten Vokabular des MeSH-Thesaurus (»Disorders«, »Chemicals & Drugs«, »Procedures«, »Anatomy« und »Physiology«) sowie nach »Chemicals«. Dies kann einerseits als Inhaltsangabe verstanden werden (wir erinnern uns: Diese Angabe basiert aus der ontologiebasierten Auswertung der Abstracts), lässt andererseits nun aber das Durchsuchen von großen Textmengen zu.

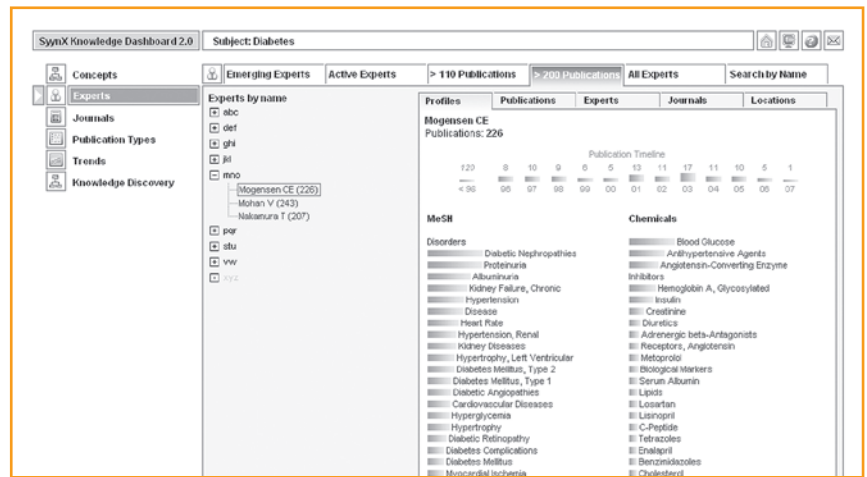
Wird nun unter »Procedures« z.B. der Suchbegriff »National Health Program« aktiviert, öffnet sich ein neues Fenster, welches die Suche feiner spezifiziert. Bei einer herkömmlichen Benutzeroberfläche hätte der Nutzer seine Suchanfrage selbständig reformulieren müssen und weitere Suchbegriffe umständlich händisch in die Suchmaske eingeben müssen, um in die Tiefe zu gehen. Das Browsen durch Konzepte spart also wertvolle Recherchezeit und damit Geld.

Im nächsten Schritt ließe sich z.B. herausfinden, in welchen Fachorganen publiziert wird (»Journals«), um welche Dokumenttypen es sich handelt (»PubType«) und an welchen Orten der Welt zu diesen Themen gearbeitet wird (»Locations«).

## Expertenprofile – Identifikation von Expertise

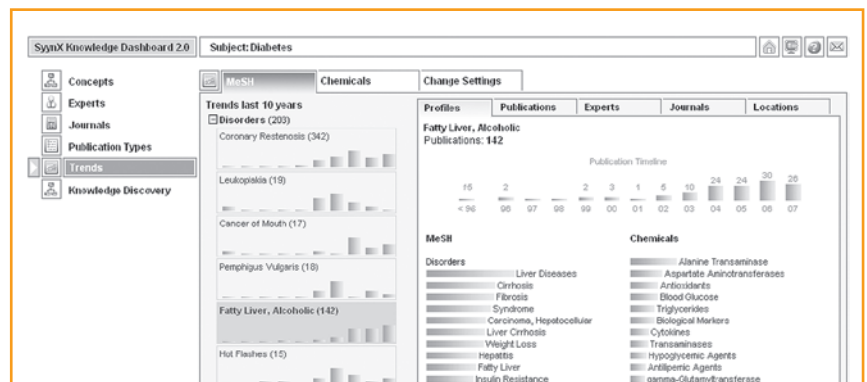
Bei der Umwandlung der PubMed-Abstracts in Fingerprints passiert aber noch mehr: Aus den Abstracts und den Meta-Informationen werden Expertenprofile generiert, die überblicken lassen, wer mit wem und zu welcher Zeit weltweit zum Gegenstand der Recherche gearbeitet hat. In der aus den Fingerprints generierten Navigation können nun schnell einschlägige Veröffentlichungen und

deren Autoren identifiziert werden. Im Sinne des eingangs skizzierten Fallbeispiels, der Vorbereitung eines Symposiums, hätte man damit also potentielle Redner eingekreist. Abb. 2 zeigt z.B. Experten zum Thema »Kidney Diseases« im Zusammenhang mit Diabetes; Abb. 3 zeigt das Profil eines Autors, aus dem deutlich hervorgeht, worüber er geforscht und was er publiziert hat. Hier könnte man kritisch einwenden, dass Vielschreiber durch statistische Verfahren automatisch zu Experten werden. Doch handelt es sich nicht um beliebige elektronische Quellen, sondern um Fachjournale mit eigener Qualitätssicherung: In der Regel sind alle eingereichten Artikel vor der Veröffentlichung einem strengen »Peer Review«-Verfahren unterzogen worden.



## Trends

Welche Fachbegriffe werden zunehmend häufig im Kontext des Dashboard-Themas genannt? Der Bereich »Trends« lässt nach wenigen Klicks erkennen, was gerade in der Szene angesagt ist bzw. in den vergangenen zehn Jahren angesagt war. Beispielsweise erfreut sich im Zusammenhang mit Diabetes die alkoholische Fettleber (Steatosis hepatis) als medizinisches Forschungs- und damit publizistisches Betätigungsfeld größter Beliebtheit: Die statistische Auswertung der MeSH-Konzepte ergibt, dass die alkoholische Fettleber seit 2003 in den Fokus der Forschung gerückt und geblieben ist und 2007 bereits 26 Publikationen erschienen sind (Abb. 4). Über die Ansicht der einzelnen Reiter können jetzt wieder einschlägige Publikationen, Experten etc. betrachtet werden. Durch automatisierte Inhaltsanalyse wird dem Researcher also enorme Arbeit abgenommen.



## Knowledge Discovery: Entdecke die Möglichkeiten

Die Möglichkeiten des »Knowledge Discovery« erleichtern Recherchen insbesondere zu Fragestellungen von wissenschaftlich tätigen Medizinerinnen und Naturwissenschaftlern. Dort wird aufbereitet, welche Begriffe erstmals in einer Publikation zum interessierenden Thema genannt werden und welche Fachbegriffe voraussichtlich mit dem Thema in Zukunft gemeinsame Relevanz entwickeln werden. SynX selbst erklärt dies so: »Die Basis für die Analysen bildet ein Vektorraum aus fünf Millionen Konzeptvektoren abgeleitet aus den PubMed-Abstracts seit 1996. Dieser Dokumentenraum wird durch Analyse der Konzeptassoziationen (gemeinsames Auftreten in mindestens einer Publikation) in einen Assoziationsraum gewandelt. Dieser Vektorraum, bestehend aus ca. 35.000 Dimensionen (den benutzten MeSH-Konzepten) und 60 Millionen

(direkten) Assoziationen, kann nun dafür genutzt werden, um mittels Ähnlichkeitsanalyse neue Assoziationen zu finden, die in dieser Kombination so noch nicht in der Literatur gebildet wurden. Durch Interpolation der Ähnlichkeitsfaktoren über einen Zeitstrahl von zehn Jahren können feinste Bewegungen in diesem Raum analysiert werden und Wahrscheinlichkeiten für entstehende Assoziationen berechnet werden. Das Ergebnis sind gewichtete Vorschlagslisten von potentiellen (also indirekten) Assoziationen, die es einem Experten erlauben, z.B. Substanzen passend zu einem bestimmten Symptom zu entdecken oder eventuelle gewünschte oder unerwünschte Nebenwirkungen einer Substanz früher in seine Überlegungen einzubeziehen als es ohne diese Analyse möglich wäre.« [16] Dieses Feature dürfte für die Forschung von allergrößtem Interesse sein und hilft z.B. auch dabei, Arbeitsgebiete zu identifizieren, die zukünftig »hip« sein werden. Dadurch können Themen und damit potentielle Märkte von morgen bereits heute besetzt werden.

## Fazit

In diesem Beitrag wurde gezeigt, was die semantische Suche mit automatisierten Wissensnetzen leisten kann. Anhand eines Fallbeispiels wurden zunächst

Abb. 3 (oben):  
Beispiel für ein  
Expertenprofil

Abb. 4 (unten):  
Beispiel für einen  
Forschungstrend

Recherchewerkzeuge vorgestellt. Dabei ist deutlich geworden, dass von den genannten Recherchertools Datenbanklösungen hinsichtlich Precision und Recall die größten Vorteile bieten, sowohl für gängige als auch für nicht mehr ganz so gängige Suchanfragen (Synonyme, Übersetzungen oder die Suche mit thematischen Verknüpfungen). Doch auch Datenbankrecherchen leiden unter Limitationen, die ihrerseits von intelligenten Verfahren wie Semantic-Web-Technologien mit ihren Möglichkeiten thematischer Verknüpfungen und der schlussfolgernden Suche nach impliziten Informationen überwunden werden können.

Das SsynX Knowledge Dashboard 2.0 bietet Zugang zu professionellem und qualitätsgesichertem Know-how, indem es alle Publikationen zu einem Themengebiet bereitstellt. Dies tut eine herkömmliche Datenbankrecherche via PubMed auch, doch vereinfacht und beschleunigt

das Dashboard Recherchen über die Möglichkeit konzeptionellen Browsens (vertikales Suchen ohne Eingabe von Suchtermen). Dadurch sinkt der Rechercheaufwand, was einen deutlichen Zeitgewinn mit sich bringt, wodurch überflüssige Kosten vermieden werden. Darüber hinaus wirken sich Ergonomie (eine Oberfläche für alles) und Visualisierungsmöglichkeiten erleichternd auf die Strukturierung komplexer Sachverhalte aus.

Die Möglichkeiten der Trendanalyse, das Erkennen von Zukunftsthemen, die Identifikation von Experten – das alles sozusagen per Knopfdruck – sind zusätzliche Vorteile, die auch äußerst erfahrene Information Professionals überzeugen dürften.

Durch semantische Technologien wird der unübersichtliche Datenpool wieder zu einem gefahrlosen Planschbeken, in dem keine Informationsflut mehr herrscht. Freuen wir uns auf die Möglichkeiten der Zukunft [17]. ■

### Literatur

- [1] Siehe Postman N: Wir informieren uns zu Tode. In: Die Zeit 41 (2. Oktober 1992), S. 61f.
- [2] Die feminine Variante ist bei der Verwendung maskuliner Termini impliziert. Die genutzten Begriffe sind Funktionsbegriffe und werden nicht geschlechtsspezifisch differenziert.
- [3] Vgl. Lewis R L, Vasishth S, Van Dyke J A: Computational Principles of Working Memory in Sentence Comprehension. In: Trends in Cognitive Sciences 10 (2006), S. 447-454. Online im Internet: <http://www.ling.uni-potsdam.de/~vasishth/Papers/tics.pdf> [Stand: 2007-11-02].
- [4] Vgl. Heyer G, Quasthoff U, Wittig T: Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. Herdecke und Bochum 2006.
- [5] Vgl. Holzinger A, Geierhofer R, Errath M: Semantische Informationsextraktion in medizinischen Informationssystemen. In: Informatik-Spektrum 30 (2007), S. 69-78, S. 70.
- [6] Vgl. Lenz R u.a.: Informationsintegration in Gesundheitsversorgungssystemen. In: Informatik-Spektrum 28 (2005), S. 105-119.
- [7] Vgl. Anm. 5, S. 69.
- [8] Vgl. Thorn K: Ein spannendes Arbeitsfeld – Medizinische Dokumentation. In: Deutsche Zeitschrift für Klinische Forschung 7/8 (2007), S. 74-78, S. 76f. Medizinische Dokumentare wollen wir hier synonym für alle Beschäftigten im Berufsfeld Medizinische Dokumentation verwenden. Die derzeitige Ausschreibungspraxis in Stellenannoncen legitimiert dies. Vgl. Wirth U: Ergebnisse der Online-Befragung »Weiterbildungsbarometer Medizinische Dokumentation« im Kontext von Target Marketing. In: Information in Wissenschaft, Bildung und Wirtschaft. Proceedings zur 29. Online-Tagung der DGI / 59. Jahrestagung der DGI. Frankfurt am Main, 10. bis 12. Oktober 2007. Hg. von Marlies Ockenfeld. Frankfurt am Main 2007, S. 201-220, S. 209.
- [9] Vgl. Seeger T: Entwicklung der Fachinformation und -kommunikation. In: Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. 5., völlig neu gefasste Ausgabe. Bd. 1. München 2004, S. 21-36, S. 33.
- [10] Vgl. Kostädt P: Einsatz von Suchmaschinen für den Zugriff auf das Deep Web. In: Content. Proceedings zur 28. Online-Tagung der DGI / 58. Jahrestagung der DGI. Frankfurt am Main, 4. bis 6. Oktober 2006. Hg. von Marlies Ockenfeld. Frankfurt am Main 2006, S. 39-44, S. 39.
- [11] Die neueste Untersuchung ist 2001 erschienen und gibt an, dass das Invisible Web etwa 400- bis 550-mal größer sei als das Visible Web. Vgl. Bergman M: The Deep Web. Surfacing Hidden Value. In: The Journal of Electronic Publishing 7 (2001). Online im Internet: <http://www.press.umich.edu/jep/07-01/bergman.html> [Stand: 2007-11-02].
- [12] Hier eine kleine Auswahl an konventionellen Gesundheitsportalen: [www.lifeline.de](http://www.lifeline.de), [www.onmeda.de](http://www.onmeda.de), [www.medinfo.de](http://www.medinfo.de), [www.medizin-forum.de](http://www.medizin-forum.de) und [www.careum-explorer.de](http://www.careum-explorer.de).
- [13] Kamphusmann T u.a.: German Medical Knowledge Guide – Prozessorientierte automatisierte Analyse großer Textkorpora. Online im Internet: <http://www.egms.de/en/meetings/gmds2006/06gmds301.shtml> [Stand: 2007-11-02].
- [14] Das Folgende nach Berners-Lee T, Miller E: Das Semantische Web hebt ab. Online im Internet: <http://www.w3c.de/Misc/tbl-em-sw.html> [Stand: 2007-11-02] und Pellegrini T, Blumauer A (Hg.): Semantic Web. Wege zur vernetzten Wissensgesellschaft. Berlin, Heidelberg und New York 2006.
- [15] Vgl. Herzog C, Diwersy M, Schmidt M: SsynX Knowledge Cockpit und SsynX Clinical Consult Application – passgenaue Wissenslösungen für Klinik und Forschung auf der Basis der Fingerprint-Textmining-Technologien. Online im Internet: <http://www.gmds2006.de/Abstracts/120.pdf> [Stand: 2007-11-02].
- [16] Siehe ebd.
- [17] Kostenloses und unverbindliches Plansch unter: <http://www.ssynx.de/produkte6.asp>.