Pfade im Informationsdschungel

Wer die verschlungenen Wege des Stoffwechsels erforscht, benötigt Orientierungshilfe. Die Datenbank SABIO-RK hilft mit allerlei Finessen der Informatik, benötigte Daten in der Flut an Publikationen zu finden.

Von Wolfgang Müller

Allein vor dem Rechner sitzend, versunken in einer abstrakten Welt aus Bits und Bytes – das ist das Bild, das sich viele von der Arbeit des Informatikers machen. Tatsächlich sieht die Realität oft anders aus. So unterstützt die HITS-Gruppe »Scientific Databases and Visualization« (SDBV) Systembiologen durch die Einrichtung und Pflege spezieller Datenbanken. Das erfordert interdisziplinäre Zusammenarbeit und regen Austausch mit den Nutzern.

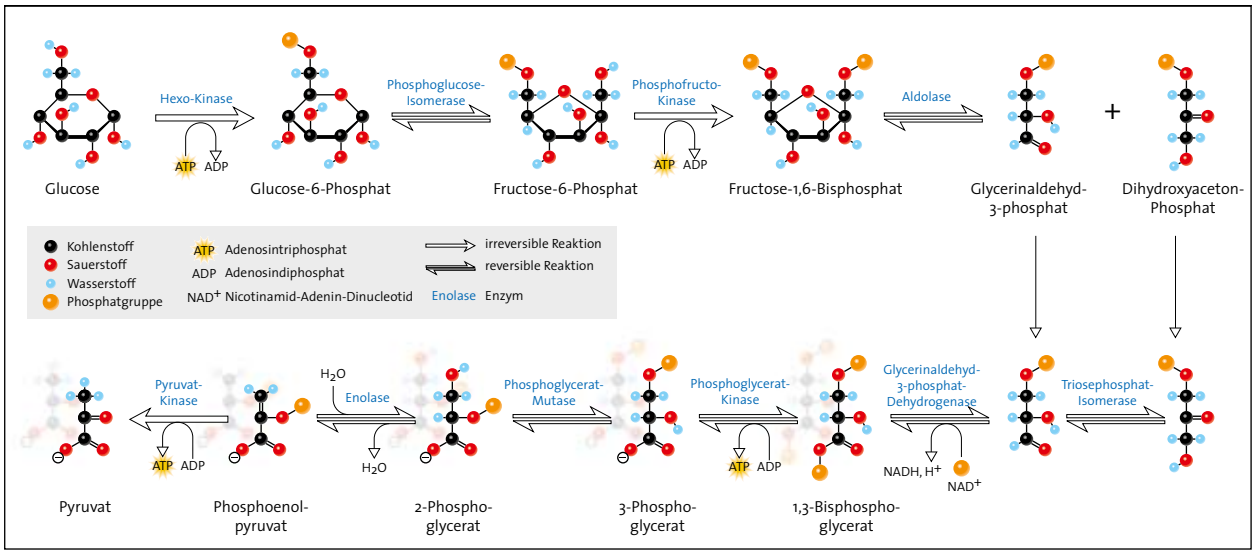
Systembiologen betrachten Vorgänge in lebenden Organismen nicht isoliert, sondern in größeren Zusammenhängen. Da sich hierbei schnell zu viele Informationen für einen einzigen Kopf anhäufen, arbeiten diese Forscher disziplinübergreifend zusammen. Während Experimentatoren sich zum Beispiel intensiv mit der Messung von Vorgängen innerhalb der Zelle befassen, haben Theoretiker etwa Stoffwechselketten und deren Kombinationen im Blick. Sie versuchen die zu Grunde liegenden biochemischen Prozesse in mathematischen Modellen zu formulieren, um nicht allein das »Wer reagiert mit wem?« zu beantworten, sondern auch Fragen wie »Wie schnell läuft die Reaktionskette bei den gegebenen äußeren Bedingungen ab?«. Solche kinetischen Modelle sind Differenzialgleichungen, die beispielsweise die zeitliche Veränderung der Glukosekonzentration und der durch den Abbau des Moleküls entstehenden Produkte widerspiegeln (unter anderem ATP und ADP, die als Energieträger im Körper fungieren). Über Koeffizienten lassen sich diese Gleichungen an die Temperatur, den pH-Wert und andere Parameter anpassen.

Wie überall in der Wissenschaft folgt der Erkenntnisgewinn dem immer gleichen Schema: Auf der Basis bereits publizierter Forschungsergebnisse entsteht eine Hypothese, die experimentell überprüft wird; die Analyse der Messergebnisse begründet dann ein Modell dessen, was im Experiment passiert ist. Alle gewonnenen Informationen werden schließlich publiziert und speisen wiederum neue Theorien und Experimente.

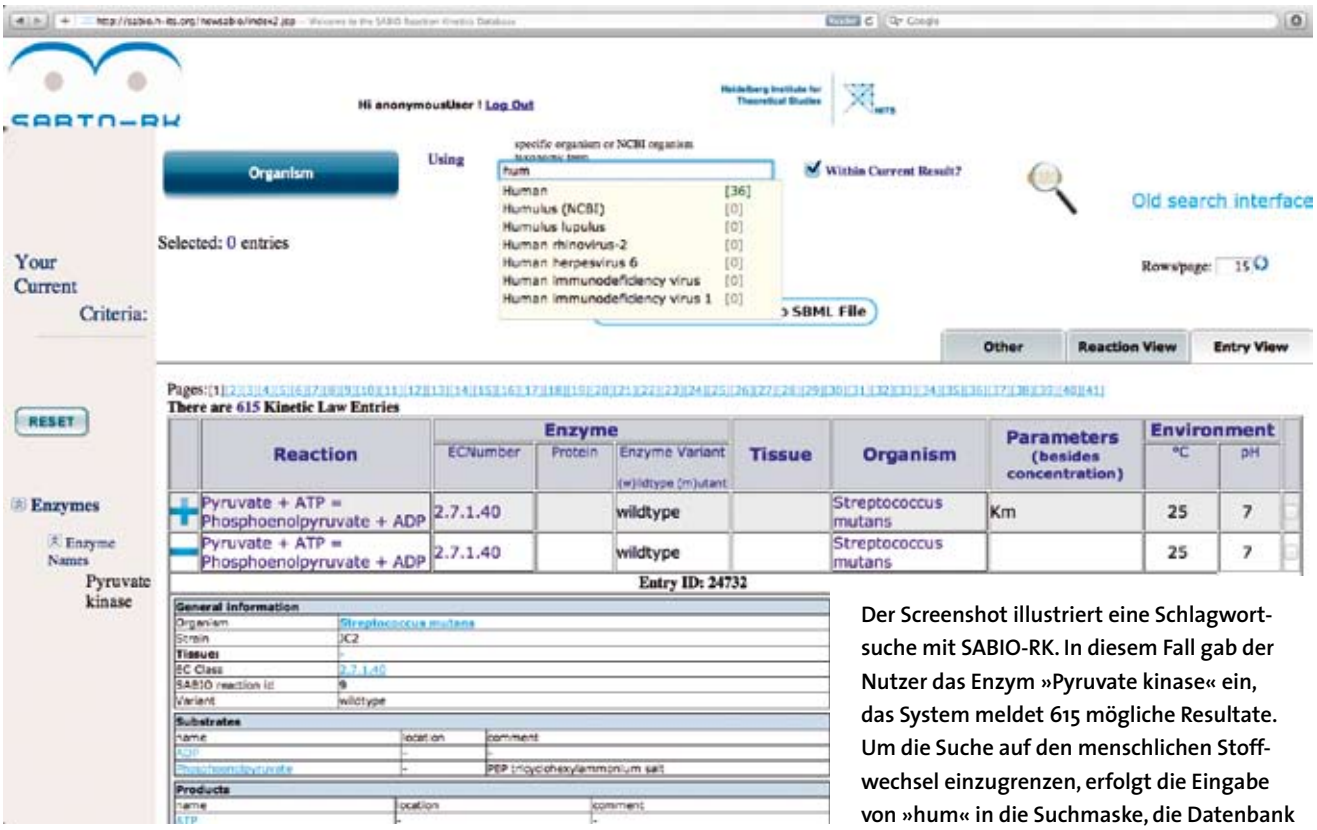
Und gerade an dieser Stelle helfen Datenbanken. Denn der Austausch über gedruckte Journale ist nicht nur langsam, es fällt Wissen-

Reaktionsketten im Visier der Forscher

Beim Glukosestoffwechsel wird das Zuckermolekül in einer Reaktionskette zu Pyruvat umgesetzt; es entstehen außerdem die Energieträger Adenosindiphosphat und Adenosintriphosphat (ADP und ATP). Immer wieder greifen dabei Enzyme wie die Hexokinase ein und katalysieren einen Zwischenschritt. Wie schnell aber laufen die Reaktionen ab, welchen Einfluss haben die Stoffkonzentrationen und die Umgebungsbedingungen? Solche Zusatzinformationen zu Stoffwechselwegen enthält die Datenbank SABIO-RK.



SPEKTRUM DER WISSENSCHAFT / ART FOR SCIENCE



Der Screenshot illustriert eine Schlagwortsuche mit SABIO-RK. In diesem Fall gab der Nutzer das Enzym »Pyruvate kinase« ein, das System meldet 615 mögliche Resultate. Um die Suche auf den menschlichen Stoffwechsel einzugrenzen, erfolgt die Eingabe von »hum« in die Suchmaske, die Datenbank liefert dazu eine Reihe von Vorschlägen.

schaftlern auch zunehmend schwerer, aus der gesamten Flut an Informationen nur die das jeweilige Thema betreffenden herauszufiltern. Aktuell verzeichnet PubMed, eine der wichtigsten Publikationsverzeichnisse für die Medizin, allein zur Leber – dem zentralen Organ des Stoffwechsels – mehr als 700 000 Veröffentlichungen. Um die jeweils relevanten zu ermitteln und daraus die für eine bestimmte Fragestellung wichtigen Daten zu entnehmen, benötigt ein Forscher die Unterstützung der elektronischen Medien.

Hierzu hat unsere Gruppe die Datenbank SABIO-RK *(System for the Analysis of Biochemical Pathways – Reaction Kinetics)* entwickelt. Wie es der Name andeutet, enthält sie von uns aufbereitete Angaben zu Stoffwechselwegen. So genannte Biokuratoren wählen zunächst potenziell nützliche Artikel anhand der Zusammenfassungen in PubMed aus. Hilfskräfte lesen diese Publikationen und geben die daraus entnommenen Daten zunächst in eine nichtöffentliche Version der Datenbank ein. Nun kommen wieder die Biokuratoren zum Zuge, die zum einen darauf achten, dass Gleiches gleich gespeichert wird. Dies betrifft sowohl die formale Struktur der Informationen als auch die Detailtiefe eventueller zusätzlicher Kommentare. Zum anderen sollte Gleiches auch gleich bezeichnet sein.

Über eine Suchmaske mit geeigneten Filtern kann ein Nutzer auf die Datenbank zugreifen – etwa nach Reaktionen suchen, an denen bestimmte Moleküle beteiligt sind. Die Informationen werden zudem auf Wunsch als SBML-Dateien ausgegeben, also in der Systems Biology Markup Language, einem international standardisierten Dateiformat der systembiologischen Modellierung. Ferner gibt es Verknüpfungen zu anderen Datensammlungen: So kann man sich mit einem Klick bei ChEBI *(Chemical Entities of Biological Interest)*, einer Datenbank, die am European Bioinformatics Institute in Hinxton (England) entwickelt und gepflegt wird, weitere Informationen zu einem Reaktionspartner holen.

Problematische Vielfalt der Namen

Für diese Arbeit benötigen wir mehr als die Expertise in der Informatik. Es genügt nicht zu wissen, wie Nutzer in einer Datenbank suchen und wie man sie dabei optimal unterstützen kann. Wir müssen auch verstehen, wie Systembiologen Daten gewinnen und einsetzen. Zudem ist SABIO-RK zwar einerseits eine Webanwendung, die wie ein Ingenieurserzeugnis geplant und gebaut werden muss. Darum herum ranken sich aber andererseits auch interessante Forschungsthemen.

So sind die Namen der reagierenden Stoffe oft nicht eindeutig, was die Forderung, Gleiches gleich zu benennen, zu einer anspruchsvollen Aufgabe macht. Beispielsweise bezeichnen das deutsche »Wasser« und die chemische Formel H_2O die gleiche Substanz. Für das englische *water* listet die Datenbank ChEBI nicht weniger als 14 Synonyme auf.

Auch die IUPAC, eine internationale Organisation, die regelt, wie chemische Verbindungen zu bezeichnen sind, lässt hier viel Spielraum. Ein Beispiel aus dem Glukosestoffwechsel: Glyceraldehyd-3-Phosphat, das korrekt auch als 3-Phosphoglyceraldehyd geschrieben werden kann, denn die standardisierte Nomenklatur erlaubt die Umstellung von Namensteilen.

Eine Vereinheitlichung ist bereits Teil der Kuratierung. So darf es schon bei der Eingabe nur entweder Glucose oder Glukose geben. Genauer gesagt, speichern wir nicht einen Textnamen, sondern die standardisierten Be-

zeichner der ChEBI: Der Glukose entspricht dort der Identifikator ChEBI:17234, der eindeutig und sprachunabhängig ist. Um eine derartige Umsetzung in einen standardisierten Bezeichner schon bei der Eingabe von Suchbegriffen durch die Nutzer zu unterstützen, lassen sich gängige Verfahren der Sprachverarbeitung wie Stemming-Algorithmen leider nicht einsetzen. Diese bilden Worte auf einen gemeinsamen Wortstamm ab, könnten beispielsweise für »gehst« und »geht« die Basis »geh« finden.

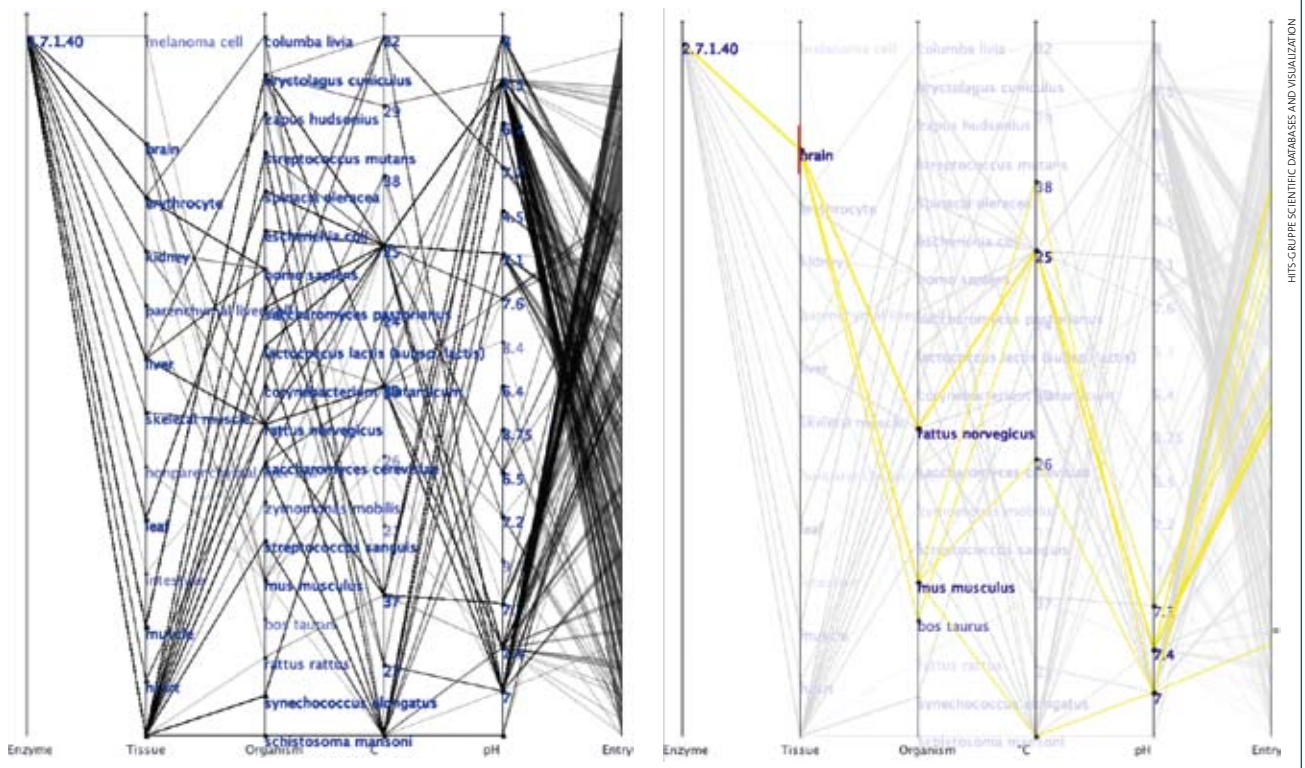
In langjähriger Zusammenarbeit mit der Gruppe von Uwe Reyle an der Universität Stuttgart entstanden zwei neue Verfahren zur Namen-Normalisierung. Das eine folgt einem morphologischen Ansatz, untersucht also die Form des Wortes. Dazu müssen wir jeden Namen in Wortbestandteile zerlegen. Diese werden sortiert, manche durch andere ersetzt. Die einzelnen Schritte sind jeweils so gewählt, dass Wörter gleichen Sinns auf gleiche künstliche Wörter abgebildet werden. Beispielsweise entfernt dieses Verfahren in den IUPAC-konformen englischen Bezeichnungen *1-butanol* und *butan-1-ol* die Bindestriche, sortiert die Wortbestandteile und kommt in beiden Fällen zu dem identischen Ergebnis *1butanol*.

Der zweite Ansatz hingegen beschäftigt sich mit dem Sinn der Wörter, ist also semantischer Natur. Dieser Algorithmus übersetzt Molekülbezeichnungen in chemische Strukturformeln und käme damit im Beispielfall ebenfalls zu dem Ergebnis, dass die zwei verschiedenen Wörter identische chemische Strukturen bezeichnen. Zwar wird aus dem Problem, korrekte Wort-Transformationsregeln zu suchen, nun die Aufgabe, Molekülnamen korrekt in Strukturen umzusetzen. Doch kann der semantische Ansatz viel mehr, vermag sogar mit Überbegriffen umzugehen: Sucht man etwa nach einer Reaktion eines Alkohols mit einem anderen Molekül, wäre der semantische Ansatz der Namen-Normalisierung im Vorteil, einerlei um welchen Alkohol es sich handelt; ein morphologischer Ansatz müsste hierzu stark erweitert werden.

Wir verfolgen deshalb beide Verfahren parallel. Die morphologische Methode steht kurz vor dem Einsatz, die semantische ist davon noch weiter entfernt. In der aktuellen Implementierung arbeitet sie auch deutlich langsamer. Als Anbieter einer Dienstleistung müssen wir uns fragen, mit welchem Aufwand wir

Stöbern in der Datenbank mit parallelen Koordinaten

Die Datenbank SABIO-RK entspricht einem vieldimensionalen Raum. So genannte parallele Koordinaten ermöglichen dennoch eine intuitive Herangehensweise. Das Beispiel beschränkt die Suche auf sechs Dimensionen: Enzym, Gewebetyp, Organismus, Umgebungstemperatur, pH-Wert und den Eintrag in der Datenbank (Entry-ID). Fragt man nach Reaktionen, die durch eine Pyruvatkinase katalysiert werden, ergibt sich das linke Bild. Offenbar wären Daten für verschiedene Zelltypen wie Melanome oder Erythrozyten abrufbar (linke Grafik), im Fokus der Suche stehen aber nur Informationen zu Gehirnzellen. Markieren des Kreuzungspunkts »Brain Tissue« lässt nicht relevante Linien verblassen. Auf einen Blick sieht der Nutzer nun beispielsweise, dass er Messergebnisse für Experimente an Ratten *(Rattus norwegicus)* abrufen kann (rechte Grafik). Sofern er sich aber für die Kinetik des Enzyms bei 26 Grad Celsius interessiert, müsste er auf Messungen an Mäusen *(Mus musculus)* zurückgreifen.



wie viel Resultat im Sinne unserer Nutzer bekommen. Gibt es Problemstellungen, bei denen er gern mehrere Sekunden wartet, bis ein Ergebnis vorliegt? Die Diskussion ist noch offen. Klar ist aber, dass wir mit beiden Ansätzen nicht immer richtigliegen. Ihr einziger Zweck ist es, den vom Nutzer gewählten Bezeichner einmalig in einen standardisierten umzusetzen. Intern wird dann nur noch dieser verwendet.

Wer in SABIO-RK nach bestimmten Reaktanten sucht, profitiert von dieser Namen-Normalisierung, denn er muss sich keine Gedanken darum machen, wie die Substanz oder das Enzym in den für ihn wichtigen Publikationen bezeichnet wurde. Viele Nutzer bringen aber weniger Vorwissen über unsere Datensammlung mit. Prinzipbedingt enthält sie nur einen sehr kleinen, aber gut gewählten und relevanten Teil der veröffentlichten reaktionskinetischen Daten.

Das Mantra der Datenbanksuche

Wir wollen einem Biologen ermöglichen, schnell herauszufinden, ob die für ihn wichtigen Daten in unserer Sammlung vorhanden sind. Das lässt sich vielleicht mit einem Kunden vergleichen, der ein Kaufhaus betritt, um eine ihm passende Hose zu finden. Vielleicht ist schon die Größe klar, aber andere Merkmale kommen erst bei der Suche selbst in den Sinn (Jeans, blau, elastischer Stoff).

Die meisten der heutzutage gebräuchlichen Such-Interfaces ermöglichen das, was die Computerwissenschaftler Ben Shneiderman und Catherine Plaisant von der University of Maryland *fact finding* und *extended fact finding* nennen. Hier geht es im Wesentlichen darum, bereits vorhandenes Wissen zu ergänzen wie: Es gibt eine Naturkonstante c, die Lichtgeschwindigkeit im Vakuum. Wie groß ist diese? Oder im Kontext der Systembiologie: Wie stark bindet ein bestimmtes Enzym bei einem pH-Wert von fünf an sein Substrat?

Dem Kunden im Kaufhaus wäre damit nicht geholfen, da er nur sehr vage Vorgaben machen kann. Dennoch wird er das gewünschte Produkt finden, da Einkaufszentren Anhaltspunkte zur Navigation geben. So wird er die Parfümerie- oder Süßwarenabteilung ignorieren, die richtige Etage für die Herrenoberbekleidung ansteuern, dort zu den Hosenständern gelangen, eine Grobauswahl anhand der Größen treffen und so fort.

Diesem Vorgehen entsprechen so genannte explorative Suchansätze. Sie sollen einem Nutzer sozusagen ein Gefühl für die Datensammlung vermitteln. Shneiderman hat dazu 1996 sein Visual Information Seeking Mantra für die Datenbankprogrammierung formuliert: *»Overview first, zoom and filter, details on demand«* – zunächst gilt es, einen Überblick zu vermitteln, dann immer näher heranzugehen und Daten herauszufiltern, schließlich Details bei Bedarf anzuzeigen.

Das bekannte Webprogramm Google Maps ist ein schönes Beispiel für eine gelungene Realisierung dieses Mantras. Nehmen wir an, Sie möchten einen abgelegenen Campingplatz in Südfrankreich finden, dann würden Sie von der Weltkugel ausgehend nach Südfrankreich zoomen, dort besonders grüne Regionen und darin wiederum nach Campingplätzen suchen. Erst dann kommen Details: Wie heißt der Ort, wie der Campingplatz, wie haben Nutzer ihn bewertet? Und all das geht so schnell vonstatten, dass kaum jemand bemerkt, wie die Satellitenkarte mit jedem Zoom nachgeladen wird.

Dass diese Technik auch für die Wissenschaft taugt, beweist SubtiPathways, eine Entwicklung der Universität Göttingen. Anstatt der Landkarten präsentiert es Stoffwechselwege; an manchen Orten darauf sind weitere Informationen hinterlegt. Dieser Ansatz eignet sich aber nur für Datensätze mit maximal fünf Dimensionen. Volker Springels Sternsimulationen (siehe den Beitrag S. 10) sind dafür ein beeindruckendes Beispiel: Verschiedene Energiedichten im dreidimensionalen Raum werden durch Helligkeiten als vierte Dimension dargestellt, deren zeitliche Veränderung erweitert die Simulation um eine fünfte Achse.

Stoffwechselpfade sind aber vielschichtiger: Es ist wichtig, an welchem Organismus, welchem Gewebe und welchem Zelltyp eine Messung durchgeführt wurde; oft sind vier bis fünf verschiedene Moleküle an einer Reaktion beteiligt, die sich je nach den äußeren Bedingungen unterschiedlich verhalten. Um einen solchen Prozess auf unseren dreidimensionalen Anschauungsraum abzubilden – und dann Techniken wie in Google Maps einzusetzen –, müssten wir Dimensionen weglassen.

Für deren Auswahl gibt es zwar durchaus Verfahren, wünschenswert im Sinne der explorativen Suche wäre aber nach wie vor, die Gesamtheit intuitiv erfassen zu können. Die von uns favorisierte Lösung sind so genannte parallele Koordinaten. Hier werden Punkte in hochdimensionalen Räumen nicht auf den dreidimensionalen Anschauungsraum projiziert, sondern als miteinander verknüpfte Linienzüge dargestellt (siehe Kasten linke Seite). Dabei entspricht jeder dieser Züge einer Dimension, daher die Bezeichnung des Verfahrens: Die Koordinatenachsen werden parallel zueinander gestellt, ein Punkt im vieldimensionalen Raum auf einen Linienzug in einer Fläche abgerollt.

Das Grundverfahren wurde bereits im ausgehenden 19. Jahrhundert entwickelt und seitdem auf verschiedene Problemstellungen angepasst. Seine Anwendung in Suchszenarien ist dennoch keineswegs trivial, da viele Fragen zu beantworten sind: Wie sollte man die Achsen anordnen, wie Kreuzungen besser kenntlich machen, wie dem Nutzer die Auswahl der ihn interessierenden Bereiche erleichtern? Für all diese Fragestellungen gibt es generelle und auch für das Problem angepasste Antworten.

Die beschriebenen Techniken – die Namen-Normalisierung ebenso wie die explorative Suche – haben zum Ziel, dem Nutzer einen Überblick zu geben und in ihm Erwartungen an Suchresultate zu wecken, die das System dann auch erfüllen kann. Um dies gut machen zu können, müssen wir erfahren, wie die Nutzer arbeiten, sowie ihre Wünsche und ihre Prioritäten kennen. Wir müssen als interdisziplinär arbeitende Gruppe in der Lage sein, Vorschläge zu machen. Dabei sind unsere Kuratoren wichtig, die aus der Biologie und Biochemie kommen und die Rolle der Nutzer übernehmen können, gleichzeitig aber auch informatisches Verständnis haben. Für viele andere Fragestellungen hingegen ist die direkte Zusammenarbeit mit Systembiologen außerhalb der Gruppe unerlässlich.

DER AUTOR

Wolfgang Müller studierte Experimentalphysik in Konstanz und parallel dazu Informatik an der Fernuniversität Hagen. Er habilitierte sich an der Universität Bamberg mit einer Arbeit zur Suche in selbstorganisierten verteilten Systemen. Seit 2009 leitet er die SDBV-Gruppe, die 1999 am EML Research, dem Vorgänger des HITS, von Isabel Rojas gegründet wurde.