Es war einmal ...

... ein „L“.

Von der Geburtsstunde des Internets berichtet eine Anekdote, die heute – gut vierzig Jahre später – beinahe wie ein „Märchen aus uralten Zeiten“ klingt: Am 29. Oktober 1969 stellte Professor Leonard Kleinrock von der University of California eine Verbindung zwischen dem Computer IMP1 in Los Angeles und einem baugleichen, 500 km entfernten Gerät in Stanford her. Ein Mitarbeiter meldete sich mit dem Wort *Login* an. Nach dem ersten Buchstaben fragte er per Telefon nach: „Habt ihr das L?“ „Ja, wir haben das L“, war die erlösende Antwort. Dieser historische Moment gilt als Beginn des Internetzeitalters. Dass die Verbindung schon beim dritten Buchstaben wieder zusammenbrach, war für den Siegeszug der neuen Technologie letztlich ohne Bedeutung.

Heute tauschen dank dieser Pioniertat einige hundert Millionen Menschen in aller Welt Billionen von Buchstaben pro Sekunde über das Internet aus. Von diesem unglaublichen Datenstrom profitiert auch die Medizin, sei es bei der Auswertung riesiger Gendatenbanken (s. S. 9) oder der Fernüberwachung herzkranker Patienten (Titelgeschichte).

Inzwischen steht bereits die dritte Internetgeneration ins Haus, die erstmals verspricht, nicht nur Buchstabenfolgen zu übertragen, sondern auch ein semantisches Verständnis der Inhalte von Webseiten zu vermitteln. Auf gut Deutsch: Eine Suchmaschine soll ein Wort wie *Herzinsuffizienz* nicht einfach als Folge von Zeichen, sondern als Synonym für eine potenziell tödliche Erkrankung verstehen.

Dr. Lutz Maicher von der Universität Leipzig stellt die technologischen Grundlagen dieser geradezu revolutionären Entwicklung vor, Dres. Volker Stümpflen und Michael Greeff vom Helmholtz-Zentrum München zeigen, wie man Tausende von wissenschaftlichen Arbeiten vollautomatisch lesen und auf hohem Niveau semantisch analysieren lassen kann.

Unglaublich, dass das alles vor nur wenigen Jahrzehnten mit einem einzigen „L“ begann.

gh

Die dritte Internet-Generation Web 3.0

Daten beginnen, Geschichten zu erzählen

Das Web 1.0 war starr, man konnte darin nur lesen, was Programmierer hineingestellt hatten. Im Web 2.0 kann dagegen jeder mitmachen; *Facebook*, *Wikipedia* und *Youtube* sind typische „Kinder" dieser zweiten Generation. Nun gehen die Entwickler noch einen Schritt weiter: Neue Formate und Suchtechniken sollen dem Internet ein Verständnis für seine eigenen Inhalte verschaffen.

Wird auf einer Seite im Internet eine Liste von Krankenhäusern mit all ihren Chefärzten veröffentlicht, so ist diese Information für den Leser lesbar und hilfreich. Schwierig wird es, wenn man zu allen Medizinern auch die Adresse der Sekretariate finden möchte. Noch mehr Handarbeit ist gefordert, sucht man alle Chefärzte im Umkreis von 30 Kilometern.

Eine Software kann solche Informationen aus dem HTML-Quelltext einer Webseite nicht automatisch verarbeiten, denn die Zeichenkette „St. Benedikt-Spital" oder „Prof. Dr. Lastenträger" hat für sie keine Bedeutung. Ein Programm versteht diese Buchstabenreihe genauso wenig wie ein Deutscher die Chat-Nachrichten japanischer Investmentbanker.

Alle Hoffnungen richten sich deshalb auf das *Web of Data*, auch *Semantic Web* oder *Web 3.0* genannt. Darin werden alle Informationen so veröffentlicht, dass sie sowohl für den Menschen als auch für Maschinen verständlich sind.

Die technologische Basis hierfür ist, dass jedes „Ding", über das Fakten gespeichert und veröffentlicht werden sollen, eine global gültige Kennung in Form einer typischen Internet-Adresse erhält. Dieser URI (*Uniform Resource Identifier*) beginnt wie eine Internetadresse (URL) mit *http://* und steht zwischen spitzen Klammern. Nach dem RDF-Standard (*Resource Description Framework*) bildet man aus solchen Adressen kurze Sätze, bestehend aus Subjekt, Prädikat und Objekt. So bedeutet beispielsweise < *http://lastentraeger.de*> <*http://example.org/work/head_physician*> <*http://benedikt-spital.de*>, dass Prof. Lastenträger ein Chefarzt am Benedikt-Krankenhaus ist.

Solche URI-Folgen sind sozusagen die„Kleber" des Web of Data, mit denen alle darin enthaltenen Fakten automatisch zusammengeführt werden; es entsteht eine 360°-Sicht auf beliebige Themen, also ein gewaltiges Netz an Informationen. Mit „Microformats" werden im Text stehende Namen, Orte oder Produkte durch den URI und bestimmte Fakten im Hintergrund ergänzt – für die Nutzer unsichtbar, aber für die Maschinen verwertbar.

Jedes „Ding" im Netz erhält eine global gültige Kennung.

Doch welche Vorteile bringt das für den Betreiber einer Webseite? Warum sollte er Aufwand dafür betreiben, die Informationen nicht nur menschen- sondern auch maschinenlesbar zu veröffentlichen? Weil er dann von den großen Suchmaschinen besser gefunden wird. Diese fangen nämlich allmählich an, solche semantischen Informationen zu „lieben".

Produkte, die mit Microformats auf Webseiten zusätzlich beschrieben werden – und somit auch eindeutig identifizierbar sind – erhalten ein gutes Ranking, denn für die Suchmaschine ist es nun ein Leichtes, dem Nutzer weitere Webseiten vorzuschlagen, die Informationen zu exakt demselben Produkt enthalten. Die strukturiert vorliegende Information, zum Beispiel den Verkaufspreis, muss die Suchmaschine nicht mehr „raten", denn sie ist ja maschinenlesbar veröffentlicht. Dadurch

erhalten Internetanwender reichhaltigere Suchergebnisse, sogenannte *Rich Snippets*.

Was bei großen Onlineshops oder Literaturdatenbanken bereits Standard ist, findet auch in anderen Suchbereichen immer stärkere Verbreitung: die Facettierung. Dabei werden außer der Ergebnisliste einer Stichwortsuche zusätzlich verschiedene Filter angeboten. Bei Produkten sind dies zum Beispiel Hersteller, Preiskategorie oder Material. Der Suchende kann so die Liste der relevanten Ergebnisse schnell auf das Wesentliche konzentrieren.

Zahlreiche im Netz veröffentlichte Informationen besitzen „geografische Relevanz", was vor allem an der zunehmenden Verbreitung von internetfähigen, mobilen Endgeräten liegt. Sie können deshalb durch ihren Zugriff auf das Netz mit Hilfe von GPS räumlich und zeitlich erfasst werden. Semantische Informationen, die durch solche Geodaten angereichert wurden, erhalten eine global gültige Verankerung. Sind also Längen- und Breitengrad des „St.-Benedikt-Spitals" bekannt, so können problemlos konfessionelle Krankenhäuser oder niedergelassene Internisten im Umkreis von 50 km gesucht werden.

Gerade ortsbezogene Dienstleistungen und die Personalisierung der Suche profitieren enorm von geografischen Bezügen der Daten. Mit semantischen Technologien sind diese leicht zu veröffentlichen und allein durch das Wissen darüber, dass zwei Dinge eine räumliche Nähe haben, ergeben sich zahllose nützliche Anwendungen, wenn man erst einmal darüber nachdenkt.

Daten sind trocken, sie müssen interpretiert werden. Geschickte Visualisierungen, die große Datenmengen kompakt zusammenfassen, ermöglichen ein „*making data to tell a story*". In sogenannten Mashups werden strukturierte, semantische Daten genutzt, aggregiert auf elektronischen Karten dargestellt, in thematischen Zeitleisten zusammengefasst, oder in verschiedenen numerischen Diagrammen nutzbar gemacht. Grundvoraussetzung in allen Fällen ist, dass qualitative Daten maschinenlesbar veröffentlicht wurden.

Rasant entwickelt sich parallel dazu der Bereich des sogenannten *Data Journalism*. Die Enthüllungen von *Wikileaks* und die Internetberichte aus den arabischen Krisenregionen haben gezeigt, welche ungeheure politische Kraft in der Veröffentlichung und Nutzung strukturierter Daten liegt. Die neuen Journalisten nutzen diese Macht, indem sie verschiedene Quellen kombinieren, systematisch nach verborgenen Zusammenhängen suchen, diese statistisch analysieren und visuell aufbereiten. So decken Datenjournalisten mit zunehmend leistungsfähigeren Werkzeugen Zusammenhänge auf, die der Öffentlichkeit bisher unbekannt waren.

Die Grundlagen sind geschaffen, nun steht das Tor zum Web 3.0 offen.

Es ist offensichtlich: Semantische Informationen eröffnen ungeahnte Möglichkeiten, aber für die praktische Nutzung sind die meisten heutigen Webseiten nur unzureichend strukturiert. Im Vergleich zu den „*dirty data streams*" der Vergangenheit müssen erst einmal kristallklare Datenströme geschaffen werden, und das kann nur gelingen, wenn jeder, der Daten besitzt, auch für deren saubere Identität sorgt. Der Weg ist noch weit, denn selbst Seiten mit dezidiert semantischen Daten haben nicht immer die nötige Qualität oder nutzen lokale URIs, die eine globale Vernetzung verhindern. Ist das Web 3.0 also nur eine schöne Vision von einigen Träumern und Technikfreaks? Mitnichten.

Das Web of Data wächst rasant und hat inzwischen nationale und weltpolitische Dimensionen. Die britische Regierung startet, unterstützt vom Erfinder des Internets, Sir Tim Berners-Lee, eine Transparenzinitiative zur Publikation aller öffentlich relevanten Daten, die Weltbank stellt eine Programmierschnittstelle (API, *application programming interface*) zur Verfügung, mit der ihre Daten umfassend genutzt werden können, und das deutsche Wissenschaftsministerium legte 2006 mit THESEUS ein „Leuchtturmprojekt" für semantische Suchmaschinen auf, für das 100 Millionen Euro zur Verfügung stehen. Die Grundlagen sind geschaffen; nun steht das Tor zum Web 3.0 für alle weit offen.

Der Vater der Idee

Der wohl bedeutendste Protagonist der Web-3.0-Idee ist der britische Informatiker Sir Tim Berners-Lee. Weltberühmt wurde er als „Vater des *World-Wide Web*", kurz www, für das er 1989 den ersten Browser entwickelte. Weiterhin ist er der Erfinder der *Hypertext Markup Language* (HTML), in der der Quelltext jeder Website geschrieben ist.
Seine grundlegenden Arbeiten führte Berners-Lee am europäischen Kernforschungszentrum CERN in Genf durch, heute ist er Professor am Massachusetts Institute of Technology (MIT) in Boston. Als Direktor des World-Wide-Web-Konsortiums (W3W) definierte er erstmals das semantische Web als „*a web of data that can be processed directly and indirectly by machines.*" Viele der hier dargestellten Konzepte gehen auf seine Anregungen zurück.

Weiterführende Literatur:

Mike Loukides: What is Data Science? The future belongs to the companies and people that turn data into products. O'Reilly (2010).

Dr. Lutz Maicher
Universität Leipzig
maicher@informatik.uni-leipzig.de