Thesen zur Disputatio von Jürgen Hermes 21.12.2011

These 1: Virtuelle Forschungsumgebungen erweitern die Möglichkeiten des wissenschaftlichen Austauschs bis hin zu einer problemlosen Reproduktion von Ergebnissen.

Empirisch ausgerichtete Wissenschaften beruhen essentiell auf der Reproduzierbarkeit der Ergebnisse ihrer Experimente. Obwohl die Wiederholung wissenschaftlicher Forschung eine so zentrale Rolle spielt, wird sie nur sehr selten wirklich praktiziert. Das liegt vor allem auch daran, dass selten Artikel zur Veröffentlichung in renommierten Zeitschriften angenommen werden, die keine neuen Erkenntnisse liefern, sondern lediglich bekannte Ergebnisse bestätigen. Ein zweiter Grund ist die Tatsache, dass Artikel in Fachzeitschriften in den seltensten Fällen tatsächlich alle relevanten Informationen enthalten, mit denen sich die Ergebnisse reproduzieren lassen.

Speziell in den Wissenschaften, die ihre Forschung *in silico* betreiben, kann der wissenschaftliche Austausch durch die digitale Weitergabe von Daten, Software und Ergebnissen massiv ausgebaut werden. Hierfür eignet sich besonders eine spezielle Form von virtuellen Forschungsumgebungen, die neben einem standardisierten Zugriff auf Rohdaten auch die komfortable Erstellung und Weitergabe von Versuchsanordnungen (Workflows) über diesen Daten gewähren können. Das in meiner Dissertation vorgestellte und von mir mitentwickelte Tesla (Text Engineering Software Laboratory) ist eine solche Forschungsumgebung, die die Reproduktion von Ergebnissen textprozessierender Experimente direkt und nachhaltig sicherstellt.

These 2: Die Herausarbeitung generischer Eigenschaften von Forschungsgegenständen heterogener Wissenschaftsbereiche ermöglicht den Transfer wissenschaftlicher Methoden und Werkzeuge.

Als Claude Shannon 1949 einen Informationsbegriff schuf, der von semantischen Inhalten bereingt war, legte er die Grundlage für eine allgemeine mathematische Theorie der Kommunikation. Diese allgemeine Kommunikationstheorie wurde von vielen, teilweise sehr unterschiedlichen Forschungsbereichen als Fundament verstanden, auf dem sie ihre eigenen Anwendungsbereiche konstituierten. Zu diesen Forschungsbereichen zählten die Kryptologie, die Sprachwissenschaft und die damals in den Kinderschuhen steckende Computerlinguistik genauso wie die Genetik, die mit der Entdeckung der Doppelhelixstruktur durch Watson und Crick 1953 in das Zentrum des öffentlichen Interesses gerückt war. Letztlich führten alle diese Forschungsrichtungen ihre Analysen auf eine Codierungs-/Decodierungsproblematik zurück, konnten aber damit – mit Ausnahme der Kryptologie – keine durchschlagenden Erfolge erzielen. In dieser Zeit wurden aber auch eine Reihe von Metaphern für biologische Einheiten und Vorgänge geprägt (*Entschlüsselung*, *Interpretation* und *Übersetzung* des genetischen *Codes*, DNA-*Buchstaben*, *Buch des Lebens* etc.), die zum Teil auch heute noch die wissenschaftliche Kommunikation bestimmen. Wenn man den Begriff *Text* weit genug definiert – als Sequenz diskreter Einheiten – so lässt sich auch die Abfolge von Basen auf der DNA als Text begreifen. Texte sind im Shannon'schen Sinn Träger von Information und dienen der Übermittlung von Botschaften. Das Spektrum der Informationsmenge in den Texten reicht dabei von minimaler Information (wenn immer das gleiche Zeichen wiederholt wird) bis zur maximalen Information (wenn die Zeichenfolge rein zufällig ist). Der interessante Bereich liegt dabei zwischen diesen beiden Polen, wo sich Muster (spezielle, wiederholte Abfolgen von Zeichenkombinationen) und Zufälligkeit überschneiden. Die Detektion von Mustern kann über ganz unterschiedliche Methoden geleistet werden. Vor allem die Bioinformatik hat sich in jüngerer Zeit damit beschäftigt, Muster effektiv in großen Textmengen aufzuspüren. Diese Methoden wurden teilweise auch schon auf natürlichsprachliche Daten übertragen. Ein für die Analyse unterschiedlicher Textsorten offenes System textprozessierender Komponenten kann den Austausch zwischen verschiedenen textprozessierenden Wissenschaften dabei zweckdienlich unterstützen.

These 3: Die Sprachtechnologie trägt dazu bei, dass sich das Internet wandelt von einem Medium, in dem gesucht wird, zu einem Medium, das antwortet.

Das Jahr 2011 wartete mit gleich zwei Überraschungen auf, welche die Fortschritte, die mittlerweile in der Sprachtechnologie erzielt wurden, eindrucksvoll dokumentieren: Zunächst ließ IBM im Frühjahr eine Maschine mit dem Namen Watson im Spiel Jeopardy gegen die bis dahin erfolgreichsten menschlichen Spieler antreten. Obwohl es in dem Spiel um die Beantwortung verklausulierter, natürlichsprachlicher Fragen geht, gewann Watson. Im Herbst brachte dann Apple sein neues iPhone auf dem Markt, das mit einer "persönlichen Assistentin", genannt Siri, ausgestattet ist, die Spracherkennung und Dialogsteuerung in einem bisher nicht gekanntem Maße beherrscht.

Seit die Suchmaschine Google 1998 online ging, die seither einen Großteil des Datenzugangs im Netz steuert, werden Spekulationen darüber angestellt, durch welche Technik sie wohl abgelöst werden wird. Oberflächlich gesehen scheint sich die Suchmaschine seit ihrem Start tatsächlich kaum verändert zu haben – die Benutzeroberfläche ist die gleiche geblieben, das Ranking der Suchergebnisse wird kontinuierlich marginal angepasst und – abgesehen von elementaren NLP-Techniken wie Stemming und Überprüfung der Rechtschreibung – hat sich nichts daran geändert, dass die Suche auf einem relativ simplen Abgleich von Zeichenketten beruht.

Allerdings hat sich das Netz, auf das Google Zugriff bietet, in den letzten 14 Jahren erheblich gewandelt. Es ist um ein Vielfaches größer geworden und seine Inhalte werden inzwischen zu einem beträchtlichen Teil von seinen Nutzern selbst generiert (Web 2.0). Soziale Netzwerke sorgen inzwischen dafür, dass persönliche Empfehlungen teilweise an die Stelle des Suchens treten. Hauptstoßrichtung der Veränderung des WWW bleibt aber die Etablierung eines Semantic Web. In diesem sollen natürlichsprachliche Informationen, aus denen das WWW zum ganz überwiegenden Teil besteht, eindeutige Beschreibungen ihrer Bedeutung zugeordnet werden, so dass die Inhalte nicht nur von Menschen, sondern prinzipiell auch von Maschinen verarbeitet werden können. Damit könnte eine Maschine in die Lage versetzt werden, auf Anfragen nicht mit einer Fülle von Verweisen auf Dokumente, sondern mit einer eindeutigen Antwort zu reagieren. IBM und Apple haben gezeigt, dass man heute schon durchaus im Stande ist, solche Antwortmaschinen zu bauen.