28 Datafication

"There are known knowns; there are things we know that we know.
There are known unknowns; that is to say, there are things that we now know we don't know.
But there are also unknown unknowns – there are things we do not know we don't know."
United States Secretary of Defense, Donald Rumsfeld

Dieses berühmt-berüchtigte Zitat des US-Verteidigungsministers (er hat es im Zusammenhang mit den angeblichen Massenvernichtungswaffen Sadam Husseins im Irak-Krieg verwendet) ist in die Geschichte der Informatik eingegangen. Es zeigt den Spannungsbogen zwischen Information und Informiertheit auf, wie er im HI-Axiom grundgelegt ist.
Allerdings hat Mr. Rumsfeld einen wesentlichen Aspekt verschwiegen. Mag sein, dass er die Gesetze der Kombinatorik nicht kennt. Aber viel wahrscheinlicher ist, dass er absichtlich den Fall der „unknown knowns" nicht erwähnt hat, weil dieser Fall das zentrale Problem der Datengenerierung (Datafication) darstellt.
Datafication heißt einerseits, breit angelegte automatisierte Datenerfassung aber auch andererseits die totale Datenunterfütterung des sozialen Diskurses und des modernen Lebens.
Geht man davon aus, dass jedes Wissen (the knowns) auf Information beruht und berücksichtigt man dabei, dass es unbekanntes Wissen (the unknown) gibt und diese beiden Phänomene unmittelbar zusammenhängen, ergeben sich 4 Kombinationsmöglichkeiten:

1. the known knowns, also das gewusste Wissen
2. the known unknowns, also das Wissen um das Nichtgewusste
3. the unknown unknowns, also das Nichtwissen des Nichtgewussten
4. the unknown knowns, also das nichtgewusste aber trotzdem vorhandene Wissen.

In der modernen Datenverarbeitung kommen alle vier Varianten vor.
Das erste Wissen (Fall 1) ist das was wir bewusst als Wissen bezeichnen. Wir haben die Daten die wir brauchen und verstehen sie anzuwenden und einzuordnen.
Im Fall 2 wissen wir wohl genau, dass es Informationen geben muss. Wir haben nur keinen Zugriff auf diese. Das ist der zentrale Geschäftsfall für Recherchen, Spionage und Geheimdienstarbeit aber auch Forschung und Journalistentätigkeit. Oft ist es möglich, die Informationsmenge festzustellen ohne die einzelnen Informationen zu kennen. In der theoretischen Physik und der Informatik werden Informationsmengen und deren Veränderung berechnet ohne die Information selbst zu erfassen. So kann man den Speicherinhalt einer Festplatte errechnen ohne noch zu wissen was später darauf steht. Jedes Defragmentierungsprogramm arbeitet nach diesen Prinzipien. Auch das Entropiegesetz in der Physik arbeitet mit diesem Grundsatz.
Im Fall 3 wissen wir nicht, dass es etwas zu wissen gäbe, also suchen wir erst gar nicht nach Informationen. Wir wissen nicht, was wir alles nicht wissen. Ein altes ungelöstes und unlösbares philosophisches Problem.
Im Fall 4 haben wir die Daten (Informationen) aber wir wissen nichts davon. Wir haben einfach noch nicht die notwendigen Fragen aufgeworfen. Fall 4 wird zunehmend der typische Zustand von Big Data. Daten, die automatisch gesammelt werden ohne dass jemand eine spezielle Fragestellung formuliert hat, fallen in diese Kategorie. Der NASA-Wissenschaftler David Wolpert hat schon vor Jahren die Theorie der Grenzen der Datenverarbeitung untersucht, das IHI hat darüber mehrfach berichtet. Datafication ist ein Prozess, der schon im 18ten Jahrhundert von Forschern wie Alexander von Humboldt begonnen wurde, indem systematisch alles was messbar ist, gemessen und aufgezeichnet wurde ohne schon zu wissen, was man später mit den erhobenen Daten anstellen könnte. In den letzten drei Jahren allein wurden jedoch so viele Daten aufgezeichnet wie in der gesamten Menschheitsgeschichte vorher. Das macht Datafication so interessant und wichtig.

So sammeln beispielsweise Videokameras im öffentlichen Raum unendlich viele visuelle Daten die nie jemand ansieht. Oder Sensoren in Industrieanlagen zeichnen Daten laufend auf, die aber nur bei Überschreitung von Kontrollparametern vom Verarbeitungssystem ausgelesen werden. Täglich werden hunderttausende Fotos auf Facebook oder Flickr hochgeladen die nur wenige Menschen ansehen.
Scannerkassen in Supermärkten sammeln Daten von jeder Ware die am Laufband vorbeikommt und verwenden nur einen Bruchteil der Informationen die dabei gesammelt werden.
Jedes Auto ist bereits ein fahrendes Computersystem, das gigantische Datenmengen generiert, die nur darauf warten weiterverarbeitet zu werden. Noch fehlen zwar die geeigneten Tools aber Firmen wie Google oder Splunk arbeiten hart an der Exploitation dieser wertvollen Rohdaten.
In jedem Gegenstand den wir täglich verwenden, stecken riesige Mengen von Informationen von denen wir gar nicht wissen, dass es sie gibt. Wir haben (besitzen) das Wissen (the known), weil es ja in unserem Besitz steht aber wir wissen nicht, dass wir es haben. Oder wissen wir wirklich, was gerade in unserem Handy vorgeht? Welche Komponente gerade Informationen sammelt und weiterleitet? Beispielsweise ob das Mikro oder die Kamera offen ist oder die Position bestimmt und gemeldet wird? Oder welche chemische Struktur unser Essen hat? Es wird zwar von irgendjemand oder irgendetwas die Information „gewusst“ aber nicht von uns – the unknown known.
Die Summe aller dieser vielen einzelnen Datengeneratoren ergibt den Gesamt-Prozess der Datafizierung. Eine Datenlawine. Ein Datenmeer. Oder wie Peter Weibel sagte: „Wir sind Daten“. Die Daten werden „wirklicher“ als die Wirklichkeit. Das konnte der verstorbene Senator Edward Kennedy auf einem amerikanischen(!) Flughafen persönlich testen als ihm der Mitflug verweigert wurde, weil das Computersystem auf Basis von Daten befand er sei ein potentieller Terrorist.
Es ist daher bei einiger Überlegung leicht zu verstehen, dass in diesen „unknown knowns“ ein höchst brisantes Potential steckt, das für viele mächtige Player äußerst verlockend ist. Die sind auch gar nicht interessiert, dass der Betroffene weiß, was es an Wissen über ihn gibt. Die bewusste Pflege des „unknown known“ ist daher ein beliebtes Spiel aller Machtträger wie Behörden, Konzerne oder Eliten. Es ist das Spiel mit der Geheimhaltung.

Es gibt aber auch eine Kehrseite der Datafizierung. Je mehr Daten gesammelt aber nicht verarbeitet werden, desto grösser wird der Anteil des „unkown known“ gegenüber dem „known known“. Oft wissen die Geheimdienste selbst gar nicht mehr was sie alles wissen. Die Überschussdaten werden zum Rohstoff künftiger Datenverarbeitung. Die rasante Verbilligung der Speichermedien begünstigt Strategien der Vorratsdatenspeicherung und erhöht die Lust an der Auswertung. Das kann durchaus auch ein Treiber für die Branchenkonjunktur im IT-Bereich werden.

50. IHI Bericht, 14.11.2013