

# Visualiseringer af den kollektive intelligens på World Wide Web

Speciale, IT-Universitetet 2008

Linje: Design, Kommunikation og Medier

Udarbejdet af:

Mads-Peter Eusebius Jakobsen (mpej@itu.dk)

Vejleder:

Leif Bloch Rasmussen (lbr.caict@cbs.dk)

# INDHOLDSFORTEGNELSE

<b>INDHOLDSFORTEGNELSE.....</b>	<b>1</b>
<b>1 INDLEDNING.....</b>	<b>3</b>
1.1 KOMMUNIKATION UDEN GRÆNSER.....	3
1.2 PROBLEMFELT.....	4
1.3 LÆSEVEJLEDNING.....	5
<b>2 METODE OG VIDENSKABSTEORI.....</b>	<b>8</b>
2.1 ABDUKTIV VIDENSKAB.....	9
2.2 ABDUKTION I DETTE PROJEKT.....	14
2.3 ØVRIGE VIDENSKABSTEORETISKE OVERVEJELSER.....	14
<b>3 WEB 2.0 OG WISDOM OF CROWDS.....</b>	<b>15</b>
3.2 WISDOM OF CROWDS.....	18
3.3 WEB PARADIGMER.....	23
3.4 DEN GLOBALE HJERNE.....	25
3.5 KOLLEKTIV INTELLIGENS OG AMATØRISME.....	28
3.6 OPSAMLING.....	32
<b>4 SEMANTIC SOFTWARE PÅ NETTET.....</b>	<b>33</b>
4.1 USTRUKTURERET, SEMISTRUKTURERET, STRUKTURERET.....	33
4.2 TEXT MINING.....	35
4.3 SEMANTIC WEB.....	38
4.4 DATA BEHØVER IKKE AT VÆRE INDHOLD.....	40
4.5 SEMANTISK ANALYSE.....	41
4.6 MENNESKER, MASKINER OG BEGREBER.....	43
<b>5 MENINGER OG BEGREBER.....</b>	<b>45</b>
5.1 BEGREBSPSYKOLOGIEN.....	45
5.2 DEN KLASSISKE TILGANG.....	46
5.3 PROTOTYPETILGANGEN.....	48
5.4 EKSEMPLARTILGANGEN.....	50
5.5 FORHOLDET MELLEMLIK EKSEMPLARER OG PROTOTYPER.....	51
5.6 VIDENSTILGANGEN.....	52
5.7 OPSAMLING PÅ MENING OG KATEGORISERING.....	54
<b>6 MENING OG SEMANTIC SOFTWARE.....</b>	<b>55</b>
6.1 TEGN OG BETYDNING.....	55
6.2 SEMANTIC SOFTWARE PÅ NETTET.....	58
6.3 SEMANTIC SOFTWARE I FORHOLD TIL MENNESKELIG KATEGORISERING.....	60
6.4 OPSUMMERING.....	61
6.5 HVAD KAN VI BRUGE SEMANTIC SOFTWARE TIL?.....	62
6.6 "METASEMANTIC" SOFTWARE.....	69
6.7 AGGREGATOR FOR WISDOM OF CROWDS.....	70

6.8	MANGEL PÅ OVERBLIK.....	70
<b>7</b>	<b>VISUALISERING .....</b>	<b>72</b>
7.1	HVAD ER EN VISUALISERING.....	72
7.2	VISUALISERING AF INFORMATION .....	73
7.3	VISUALISERINGENS DIALEKTISKE FORHOLD .....	79
7.4	NATURLIGE OG SYMBOLSKE TEGN .....	83
7.5	EKSEMPLER PÅ VISUALISERINGER .....	85
7.6	ERFARINGER FRA ANALYSER AF VISUALISERINGSEKSEMPLER.....	98
7.7	KONKRETISERING AF ERFARINGER .....	100
<b>8</b>	<b>FORSLAG TIL SEMANTISK VISUALISERING.....</b>	<b>101</b>
8.1	VISUALISERING AF DEN GLOBALE INTELLIGENS .....	101
8.2	ABDUKTIV TILGANG TIL VISUALISERINGER .....	101
8.3	SCENARIOE .....	102
8.4	VISIONER FOR VISUALISERINGEN.....	103
8.5	UDFORMNING .....	105
8.6	ANVENDELSE AF SEMANTISKE VISUALISERINGER.....	110
<b>9</b>	<b>ANTI-KONKLUSION .....</b>	<b>114</b>
9.1	OPSUMMERING .....	114
9.2	FREMTIDIGT ARBEJDE.....	115
9.3	AFSLUTNING .....	117
<b>10</b>	<b>LITTERATUR.....</b>	<b>118</b>
10.1	PRIMÆR LITTERATUR .....	118
10.2	SEKUNDÆR LITTERATUR.....	119
10.3	ANDRE VÆRKER .....	120

## Abstract

Through an abductive analysis of how the collective intelligence as represented on the World Wide Web can be harvested this project explores the potential use of visualizations based on semantically analyzed data. Several semantic software technologies, mainly Semantic software and Natural Language Processing, are evaluated against the theories of psychology of concepts. A generic design for semantic visualizations is proposed and the usage of a such explored. Natural Language Processing proves to be a promising technology for semantic analysis of data, but it is still under development and is not yet mature for general use on www. However, once developed further, the technology in conjunction with visualizations has a great potential for use in marketing, politics and in the general harvest of the collective intelligence or as an aggregator for the "Wisdom of crowds".

# 1 INDLEDNING

---

*If I'd asked people what they wanted, they would have said a faster horse.*

---

HENRY FORD

Henry Fords berømte citat bygger på en opfattelse af mennesker, der ikke tænker innovativt, og som ikke forstår deres egne behov. Det er en hyldest til designeren, som i sidste ende ved bedre end dem, der designes for. Citatet er berømt, fordi det sætter fokus på genialiteten hos individet – det sætter fokus på originalen. Og så er det symbol på en tid, hvor den eneste måde at komme frem i verden, var med den unikke, gode idé. En tid, hvor individet var klogere end masserne.

Med det andet årtusindeskiftes informations- og kommunikationsteknologi er der dog kommet nye boller på suppen. Udviklingen af World Wide Web<sup>1</sup> har betydet, at enhver bruger af internettet har mulighed for at ytre sin holdning, dele sin viden eller blot sine tanker. Gennem blogs, kommentering, debatfora og online communities er WWW ikke længere forbeholdt de få med teknisk viden, men de mange med noget på hjerte.

## 1.1 KOMMUNIKATION UDEN GRÆNSER

Vi er allerede begyndt at se eksempler på, hvordan masserne er begyndt at arbejde sammen for at skabe produkter, intet individ ville kunne have skabt. Wikipedia er et klassisk eksempel på, hvad mennesker kan opnå i fællesskab. Wikipedia er et online leksikon skabt af almindelige, ulønnede brugere. Enhver kan når som helst rette hvad som helst i leksikonet – og netop fordi enhver kan rette det, får fejl ikke lov til at stå ret længe. Wikipedia er et enestående opslagsværk med knap 2,5 millioner artikler<sup>2</sup> - og det er endda kun for den engelsksprogede del, som udgør mindre end en tredjedel af det samlede, multinationale Wikipedia<sup>3</sup>.

Wikipedia er blot et enkelt eksempel på, hvordan brugere kan skabe fantastiske ting, når mange mennesker giver en hånd med. Det er den kollektive indsats, der har gjort Wikipedia til et enestående opslagsværk. Men Wikipedia er også skabt til formålet – det ligger i selve programmets præmis, at det er brugerstyret. Der, hvor den virkelige udfordring ligger, er i at udnytte de brugergenererede data, der endnu ikke indgår i et kollektivt virke, som f.eks. et samlet forum eller en organisation.

En af visionerne bag udtryk som global intelligens eller Wisdom of crowds er, at mennesker kan opnå mere, end vi umiddelbart forestiller os, hvis vi i stedet for at tænke individualistisk tænker

---

<sup>1</sup> World Wide Web og internettet er ikke det samme. Internettet består af en række kommunikationsteknologier, heriblandt protokollen http, som World Wide Web benytter sig. World Wide Web er der hvor vi finder alle hjemmesider, søgemaskiner, blogs osv. Andre internetteknologier er eksempelvis ftp, udp, e-mail, usenet mv. Da jeg i dette projekt ikke behandler andre internetteknologier end WWW, er det i denne kontekst jeg benytter ordet internet.

<sup>2</sup> Der findes 2.445.762 artikler i Wikipedia pr. d. 8. Juli 2008.  
<http://en.wikipedia.org/wiki/Special:Statistics>

<sup>3</sup> Alle sprogversioner af Wikipedia kombineret udgør pr. 1. Marts 2008 9.790.407 artikler.  
[http://en.wikipedia.org/wiki/Wikipedia:Multilingual\\_statistics](http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics)

mere kollektivt. Det skal ikke forstås som, at vi alle skal droppe vores daglige liv og underkaste os den globale hjerne. Tværtimod skal vi som individer hverken tænke eller handle anderledes, end vi allerede gør. Potentialer ligger i den teknologi, i de værktøjer, der lader kommunikationen bryde tid/sted-begrænsningerne. En form for kommunikation, der for alvor er ved at blive allemandseje. Jeg vil i et senere kapitel uddybe begreberne global intelligens og Wisdom of crowds.

I modsætning til Wikipedia er det meste brugergenerede indhold dog ikke sat i system. Set ud fra et teknisk synspunkt er det svært at finde, kategorisere og anvende de enorme mængder af data, som brugerne producerer. Dette projekt handler dels om, hvad der skal til for at kunne få adgang til de mange brugergenerede data, og dels hvorvidt sådanne data overhovedet kan bruges til noget.

## 1.2 PROBLEMFELT

Udgangspunktet for dette projekt var oprindeligt at undersøge, hvordan man kunne udnytte de store mængder af brugergenererede data, der er tilgængelige på World Wide Web, i forbindelse med løsninger af politiske problemer. Altså en form for inddragelse af borgere i politiske sager gennem webbaseret kommunikation. Problemformuleringen lød da: "*Hvordan kan man gennem webbaseret kommunikation skabe inddragelse af brugere i politiske sammenhænge?*".

Det viser sig dog, at der forud for en sådan undersøgelse ligger en teknisk og designmæssigt udfordrende opgave, der skal løses, nemlig hvordan man i det hele taget kan drage nytte af store mængder af brugergenereret data. Idet den oprindelige problemformulering ikke afdækker, hvilke *politiske sammenhænge* der er tale om, eller hvem *man* er, har jeg i stedet valgt at undersøge en mere grundlæggende problemstilling, nemlig hvordan vi overhovedet kan få adgang til de data der genereres af brugerne, og hvordan en sådan adgang til data kan udnyttes.

Der bliver skrevet millioner af blogs verden over, og ifølge Dansk Statistik blogger 9%<sup>4</sup> af den danske befolkning.<sup>5</sup> De store mængder af data er et resultat af, at det er blevet lettere for internetbrugere at bidrage med indhold på WWW. Begreber som Wisdom of crowds, global intelligens og kollektiv handling er begyndt at dukke op for at beskrive de kræfter, der er på spil, når så mange brugere går sammen om at løfte forskellige opgaver. Wikipedia.org er et eksempel på denne *kollektive intelligens*, hvor tusindvis af brugere, uden kapital, uden konkret organisering og uden at få løn for det har skabt verdens største opslagsværk.<sup>6</sup> Når jeg i dette projekt bruger begrebet *kollektiv intelligens*, er der ikke tale om intelligens i almindelig forstand; den måles ikke i en intelligenskvotient. Den kollektive intelligens er et udtryk for en ressource, der er et resultat af store mængder af individers handlinger, meninger, viden eller deltagelse.

Men teknologien til at håndtere så store mængder af semantisk data har endnu ikke haft sit store gennembrud.

---

<sup>4</sup> Det skal understreges, at Dansk Statistiks undersøgelse er blevet kritiseret af flere. Eksempelvis mener blog-sitet Overskrift.dk, at tallet reelt er i nærheden af det halve.

<http://blog.overskrift.dk/2008/06/30/9-af-danskerne-blogger-agtigt/>

<sup>5</sup> Se Dansk Statistik: <http://www.statistikbanken.dk/statbank5a/SelectVarVal/Define.asp?Maintable=BEBRIT04&PLanguage=0>.

<sup>6</sup> Jeg vil vende tilbage til begreberne Wisdom of crowds og kollektiv intelligens i kapitel 3.

Semantic software er software, der i en eller anden grad behandler data som meningsbærende information. Da internettet var ungt, var alle søgemaskiner udelukkende baserede på nøgleordssøgning, hvilket ville sige, at der i princippet er større forskel på begreberne 'penge' og 'valuta', end der er på 'vind' og 'kvinde'. Sådanne søgemaskiner er ikke semantiske. På en moderne, semantisk søgemaskine vil jeg kunne finde det samme indhold på nettet ved at søge på 'søde hundeunger' og 'nuttede hvalpe', til trods for at jeg ikke bruger de samme nøgleord. Årsagen er, at en semantisk analyse af de to sætninger er meget tæt relaterede – de er næsten synonyme. Semantic software viser med andre ord et enormt potentiale for at kunne analysere og behandle store mængder af indhold skabt af brugere.

Jeg vil i dette projekt undersøge, hvordan semantic software kan bruges til at skabe overblik og drage nytte af den kollektive stemme, som mængden af brugergenereret indhold udgør. Dette giver anledning til følgende reviderede problemformulering:

### *1.2.1 PROBLEMFOMULERING*

**Hvordan kan semantic software udnyttes til at høste den kollektive intelligens, som brugere i fællesskab udtrykker på World Wide Web?**

Jeg vil besvare ovenstående problemformulering ved at besvare følgende underspørgsmål:

- Hvordan kommer den kollektive intelligens til udtryk på World Wide Web?
- Hvordan kan semantic software, herunder Semantic web, behandle brugergenererede data?
- Hvordan kan visualiseringer af data baseret på semantisk analyse anvendes i forbindelse med udnyttelse af brugergenererede data?

Jeg vil i projektet ikke bruge en konkret case, men i stedet foretage en mere abstrakt undersøgelse, der beskriver teknologierne, udviklingen og evt. muligheder for udnyttelse af disse. I kraft af projektets udgangspunkt, inddragelse af borgere gennem webbaseret kommunikation, bruges der igennem projektet en række eksempler, der tager udgangspunkt i politik, demokrati og borgerinddragelse, men jeg vil også bruge andre eksempler, som eksempelvis brug af semantic software i forbindelse med markedsføring og produktudvikling. Dette projekt vil således være en abduktiv analyse af, hvordan semantic software og visualiseringer kan bruges i forbindelse med anvendelse af brugergenereret data.

## 1.3 LÆSEVEJLEDNING

Forfatteren og Semantic web-pioneren Nova Spivack<sup>7</sup> beskriver internettets udvikling som en bølgende bevægelse mellem fokus på teknologisk udvikling og fokus på anvendelse og brugerinvolvering. Den teknologiske udvikling består af en række trin: Først udvikles værktøjet, dernæst lærer mennesker at bruge det. Den ekstensive brug af værktøjet stiller nye krav til teknologien, hvorfor værktøjet må videreudvikles, hvorefter det nye værktøj igen skal gøres

---

<sup>7</sup> Nova Spivack, f. 1969, USA, har været med til at stifte EarthWeb, Radar Networks og Lucid Ventures. Nova Spivack er en af udviklerne bag Twine.com.

anvendeligt for de almindelige brugere. Og således svinger pendulet frem og tilbage mellem et teknologisk og et brugerorienteret fokus.<sup>8</sup>

De første computere blev ikke skabt for forbrugernes skyld. De var komplicerede stykker værktøj designet til at løse dedikerede opgaver, og det krævede eksperter at bruge dem og at få dem til at fungere. Fokus var på teknologien.

I små skridt udviklede markedet sig, og på et tidspunkt blev det muligt at lancere den personlige computer. Der skete skift i fokus fra det teknologiske til det brugerorienterede, således at f.eks. ansatte med en anelse træning kunne lære at bruge computerne. Der skete en udvikling inden for brugervenligheden på computere, og med grafiske brugerflader og overskuelige menu-systemer blev det nemmere for almindelige forbrugere at være med.

En sideløbende udvikling øgede dog fokus for nye tekniske udviklinger: Computere i netværk. Da internettet bliver en realitet, er det udelukkende de mennesker, der har kendskab til filservere og html, der kan bidrage med indhold til internettet. Alle andre brugere kan kun 'kigge med'.

Nova Spivack beskriver begrebet *Web 2.0* som pendulets sving tilbage mod en brugerorienteret tilgang til teknologien. Med ny udnyttelse af eksisterende webteknologi og med et fokus på brugervenlighed, er det pludselig muligt for enhver webbruger at skabe sit eget indhold eller at kommentere på andres. Efter en kort videnskabsteoretisk redegørelse for denne rapport vil jeg i **kapitel 3** beskrive, hvad der kendetegner denne såkaldte anden generation af World Wide Web. Jeg diskuterer, hvorvidt Web 2.0 kan anskues som et paradigme, og hvilke visioner og idealer der ligger bag dette. I denne sammenhæng vil jeg beskrive begrebet Wisdom of crowds sammen med forskellige eksempler på kollektiv intelligens.

I **kapitel 4** lader vi pendulet svinge endnu engang. Web 2.0 efterlader sig enorme mængder af brugergenereret materiale. De omfattende data bliver for komplekse for de herskende teknologier, og nye metoder til at håndtere indhold bliver nødvendige. Semantic web er en række softwareteknologier, der forsøger at lære maskiner at forstå menneskelig tekst, således at vi får bedre værktøjer til at søge, behandle og analysere brugergenereret indhold. Semantic web er en form for semantic software. I dette kapitel beskriver jeg Semantic web på et design- og brugsorienteret niveau<sup>9</sup>, men jeg vil også give en kort gennemgang af forskellige andre semantic software-typer.

Semantic software behandler indhold ved at analysere meningen med teksten gennem begreber. Men hvad er mening, og hvordan bruger mennesker begreber til at forstå kategorier? **Kapitel 5** tager udgangspunkt i begrebspsykologien for at beskrive, hvordan mennesker benytter sig af begreber, og hvordan mening knyttes til begreber.

Trådene samles i **kapitel 6** hvor teorien om begreber sammenknyttes med visionerne for semantic software, og jeg diskuterer hvorledes semantic software kan spille en rolle for vores hverdag i fremtiden. Jeg sætter Semantic web i forhold til menneskelig begrebsforståelse (som

---

<sup>8</sup> Spivack 2007d

<sup>9</sup> jeg vil ikke diskutere algoritmer i denne rapport, og jeg vil kun bruge ganske få bider simplificeret kode, til at forklare den grundlæggende teknologi

præsenteret i kapitel 5) og jeg præsenterer en række eksempler på udnyttelse af semantic software.

Uanset hvordan man udnytter semantic software i forbindelse med World Wide Web, vil der være et behov for at kunne overskue de resultater, der skabes. Jeg præsenterer i **kapitel 7** en række metoder til brug i forbindelse med visualisering af information, og gennem analyse af en række eksisterende visualiseringer præsenterer jeg nogle tommelfingerregler, der bør anvendes i design af visualiseringer baseret på semantiske analyser. Disse vil i **kapitel 8** munde ud i et eksempel på, hvordan sådanne visualiseringer kan designes, for at visualisere kollektiv intelligens.

Som jeg vil uddybe i følgende kapitel, udgør denne rapport en abduktiv undersøgelse, hvorfor det ikke vil være passende med en traditionel konklusion. Derfor afsluttes rapporten i en anti-konklusion i **kapitel 9**, hvor jeg opsummerer rapporten og foreslår hvorledes der kan arbejdes videre med semantiske visualiseringer baseret på brugergeneret indhold fra World Wide Web.

Bemærk at kildelisten er delt i tre dele. **Den primære litteratur** er litteratur der bruges helt konkret i projektet. **Den sekundære litteratur** benyttes ikke direkte i rapporten, men har på anden vis spillet en rolle for projektet, men som der ikke nødvendigvis henvises til i rapporten. Til sidst er der en række **Andre værker**. Disse værker er ikke akademiske tekster, men kan eksempelvis være henvisninger til nogle af de kunstværker der analyseres i kapitel 7 eller andre værker der nævnes i rapporten.

## 2 METODE OG VIDENSKABSTEORI

Det er en lang tradition for at arbejde induktivt eller deduktivt i forbindelse med akademisk arbejde. Årsagen hertil er, at det er de eneste måder at præsentere sikker eller sand viden på. Induktion og deduktion kan eksemplificeres således:

**Induktion**                    *Enkeltilfælde: "Vores bønner er fra plasticsækken."  
Resultat: "Vores bønner er hvide."  
Regel: "Alle bønner fra plasticsækken er hvide."<sup>10</sup>*

**Deduktion**                    *Regel: "Alle bønner fra plasticsækken er hvide."  
Enkeltilfælde: "Vores bønner er fra plasticsækken."  
Resultat: "Vores bønner er hvide."<sup>11</sup>*

Man kan sige at hvor induktionen tager udgangspunkt i et eller flere enkeltilfælde og forsøger at sige noget generelt, arbejder deduktion den anden vej og bruger generel viden til at sige noget om enkeltilfældet. Idet induktionen tager udgangspunkt i reelle forsøg eller konkrete eksempler, kan man sige, at induktionen producerer sikker viden. Deduktion er logisk betinget og vil i højere grad antage præmisserne for et resultat for at være sande og kan derfor siges at producere sand viden.

Dette speciale adskiller sig fra størstedelen af akademisk litteratur ved hverken at arbejde induktivt eller deduktivt, men derimod abduktivt. Set i en videnskabsteoretisk optik er abduktionen ikke til for at finde de endelige resultater, men for at skabe grobunden for de idéer, der med induktion og deduktion evt. skal eller bør forskes i. Søren Kjørups<sup>12</sup> eksempel på abduktion lyder således:

**Abduktion**                    *Resultat: "Vores bønner er hvide."  
Regel: "Alle bønner fra plasticsækken er hvide."  
Enkeltilfælde: "Vores bønner er fra plasticsækken."<sup>13</sup>*

Som den opmærksomme læser vil bemærke, er ovenstående eksempel ikke logisk, og kan ingenlunde antages at være sandt. Ikke desto mindre er det en slutningsform vi benytter os af i vores dagligdag. Et andet eksempel på abduktion kunne lyde:

**Abduktion**                    *Resultat: "Bønnerne på gulvet ved siden er plasticsækken er hvide."  
Regel: "Små ting der ligger ved siden af plasticsække, er fra plasticsækken"  
Enkeltilfælde: "De hvide bønner på gulvet er faldet ud af plasticsækken."*

Det interessante ved eksemplet ovenfor - og abduktion generelt - er, at slutningen ikke udelukkende baseres på de variable, der er direkte indlejret i slutningen. Men med vores

---

<sup>10</sup> Kjørup 2002, s. 48

<sup>11</sup> Kjørup 2002, s. 48

<sup>12</sup> Søren Kjørup, f. 1943, mag.art i filosofi, professor ved Roskilde Universitetscenter

<sup>13</sup> Kjørup 2002, s. 48

dagligdagsviden og vores erfaringer fra lignende tilfælde kan vi danne os antagelser om, hvordan de hvide bønner kan være havnet på gulvet. Vi har måske erfaret, at kaffebønner også har en tendens til at ligge og flyde omkring plasticsække, der er fyldt med kaffebønner.

Men ved vi, om bønnerne er faldet ud af sækken? Kan vi danne sikker eller endda sand viden? Nej, det kan vi ikke. Så vidt vi ved, kan de hvide bønner stamme fra en helt anden sæk eller de kan være lagt der, blot for at vi skulle undre os over det. Er det sandsynligt at bønnerne er faldet ud af posen? Ja, det er det. Er det værd at finde ud af, hvorvidt de faktisk er faldet ud af posen? Lad os sige, at vi har et problem med løst liggende hvide bønner på vores gulv, og vi ønsker at sætte en stopper for det. Havde vi kun haft induktionen og deduktionen at holde os til, havde vi ikke med rimelighed kunnet træffe et fornuftigt valg. Men fordi vi mennesker også tænker abduktivt, er vi i stand til at fremsætte en mulig hypotese, at de hvide bønner sandsynligvis stammer fra sækken med de hvide bønner. Derfor har vi et grundlag for efterfølgende at kunne undersøge, hvorvidt bønnerne har en tendens til at falde ud af plasticsække, og gennem deduktion og induktion at lave en mere grundig analyse af problemet.

Lader vi bønnerne ligge, står vi altså med tre hovedtraditioner inden for videnskabsteorien; den induktive, som frembringer sikker viden, den deduktive, som bringer sand viden, og til sidst den abduktive, som skaber ny, mulig viden. Sidstnævnte, som denne rapport indskrives i, producerer ikke nødvendigvis hverken sikker eller sand viden, men kan ud fra en videnskabsteoretisk synsvinkel være interessant at studere nærmere. I det følgende afsnit vil jeg derfor beskrive, hvordan man kan udføre videnskab ud fra en abduktiv synsvinkel.

## 2.1 ABDUKTIV VIDENSKAB

Dette afsnit er hovedsageligt baseret på et appendiks om Peirce<sup>14</sup> fra Lise Schrøders Ph.d.-afhandling, "Tegn, fortælling og metaforståelse – et perspektiv på informationsdesign og bygningsdata"<sup>15</sup>. Charles Sanders Peirce anses for at være en af semiotikkens fædre. Hans livsværk er dog mere omfattende end som så. Peirce studerer forholdet mellem verden og den, der oplever den. Peirce er ikke i tvivl om, hvorvidt den rigtige verden eksisterer eller ej – det mener han, at den gør – men han er usikker på, om vi nogensinde oplever den direkte. Hans studier af tegnet, semiotikken, er udtryk for hans grundholdning, at fortolkningen af omverdenen er afgørende for vores billede af verden.

*I henhold til Peirces tankegang lever vi i to verdener – en faktisk og en forestillet, den forestillede indre verden og kendsgjerningernes ydre ...*

SCHRØDER 2005, s. 324

Peirce arbejder med en tredeling i forbindelse med oplevelsen af verden. Han kalder subjektet for førsteheden, objektet for andetheden og fortolkningen for tredieheden.

---

<sup>14</sup> Charles Sanders Peirce, 1839-1914, amerikansk filosof, pragmatiker og systematiker, og er blandt andet kendt for at have grundlagt semiotikken.

<sup>15</sup> Schrøder 2005

Førstehed	Andethed	Tredjehed
Subjekt	Objekt	Fortolkning
Opleveren	Genstand for det oplevede	Oplevelsen

FIGUR 1 – PEIRCES TREDELING

Det første udgør kvaliteter, mens kendsgerninger, handling og aktualitet hører til det andet. Det tredje omfatter love og tanker og forstås som det, der forbinder det første og det andet.

Et lille eksempel: En mand ser et egetræ, der minder ham om sit barndomshjem, hvor et lignende træ var plantet. Mandens oplevende jeg, som oplever træet og sine minder, er hvad Peirce kalder *førsteheden*. Det konkrete egetræ, dets stamme og blade er *andetheden*. *Tredjeheden* er den fortolkningsramme, der tillader manden at opleve træet, men er også det, der gør, at kun han vil mindes netop sit barndomshjem. Kun fordi han kender idéen om et træ, genkender han det som et træ, og kun fordi han voksede op et sted med et lignende egetræ, mindes han sin barndom.

Peirces tredeling bruges på tre niveauer, som igen svarer til det første, det andet og det tredje.

Det første niveau vedrører repræsentamen, dvs. selve tegnet:

**Kvali-tegnet eller tone**

Ved tone er der tale om rene kvaliteter (f.eks. følelsen af rød)

**Sin-tegnet eller token**

Token refererer til enkeltstående tegn (f.eks. et specifikt, konkret vejskilt)

**Legi-tegnet eller type**

Typer er repræsentationer for klasser af tegn (f.eks. 'trafikale vejskilte' som begreb).

Andet niveau afspejler forholdet mellem repræsentamen og objektet<sup>16</sup>:

**Ikonet**

Ved at ligne det, der henvises til (f.eks. et maleri, der ligner det, der afbildes)

**Indeks**

Ved at have en faktisk forbindelse (f.eks. fodspor i sneen)

**Symbol**

Ved at have en vedtaget betydning (f.eks. piktogrammer)

<sup>16</sup> Tredjeheden på andet niveau, er netop genstanden for semiotikken, og jeg vil kort vende tilbage til disse i kapitel 6.1

På det tredje niveau finder vi forholdet mellem repræsentamen og interpretanten:

**Rheme, term eller det åbne**

F.eks. et begreb

**Dicent eller det singulære**

F.eks. et udsagn

**Argument eller det formale**

F.eks. en lov

Indsætter vi tredelingen i begge dimensioner i et skema, får vi følgende:

		FØRSTEHED	ANDETHED	TREDJEHED
FØRSTEHED	<i>Repræsentamen</i>	Kvali-tegn	Sin-tegn	Legi-tegn
ANDETHED	<i>Repræsentamen/objekt-relation</i>	Ikon	Indeks	Symbol
TREDJEHED	<i>Repræsentamen/interpretant-relation</i>	Rheme	Dicent	Argument

FIGUR 2 - TEGNETS FÆNOMOLOGISKE OG ONTOLOGISKE KATEGORIER (SCHRØDER 2005)

Begynder vi at kombinere disse niveauer, får vi 27 kombinationer – 27 klasser af tegn. Ud af disse mente Peirce, at der var ti, der gav god mening:

- 1) Åben, ikonisk tone
- 2) Åben, ikonisk token
- 3) Åben, ikonisk type
- 4) Åben, indeksikalsk token
- 5) Åben, indeksikalsk type
- 6) Åben, symbolsk type
- 7) Singulær, indeksikalsk token
- 8) Singulær, indeksikalsk type
- 9) Singulær, symbolsk type
- 10) Formel, symbolsk type

Peirces tegnklasser kan anskues som kategoriseringsklasser.<sup>17</sup> Gary Shank & Donald J. Cunningham argumenterer for, at selvom nytteværdien af et sådant sæt kategoriseringsklasser ikke er stor, kan man ved at se på dem som modeller for *udledning* (jf. eng. inference) få ny brugbar viden ud af dem. Shank & Cunningham vælger at se de første seks klasser som en familie

<sup>17</sup> Shank 1996

af abduktionsmodeller, klasserne 7-9 som induktive og den sidste som deduktiv. Schrøder refererer til Søren Briers model, som også knytter abduktion med førsteheden (rhemet, det åbne).<sup>18</sup>

Jeg vil i dette projekt arbejde abduktivt ud fra Shank & Cunninghams seks modeller for udledning, der er baseret på Peirces tegnlære. Jeg vil i det følgende gennemgå de seks klasser.

### 2.1.1 ÅBEN, IKONISK TONE (ELLER INDFALD/IDÉ)

Som navnet antyder, er dette en meget usikker form for udledning. Det er muligheden for en mulig lighed. Som Shank & Cunningham selv udtrykker det:

*A more concrete way to characterize this type of reasoning is to describe it as reasoning in order to determine the possibility that our initial observations might serve as omens for possible evidence*

SHANK & CUNNINGHAM 1996, s. 3

Der er tale om meget subjektive udledninger, der ikke baserer sig på bevidste valg. Her er der ikke tale om bevisførelse, men nærmere en undersøgelse af, hvorvidt ens observationer kan føre til bevisdannelse.

Denne type udledning kan eksempelvis komme i spil, når man leder efter sine nøgler. Man kan måske ikke forklare hvorfor man leder i køkkenet, men det betyder ikke, at det er en dårlig idé at kigge der alligevel.

### 2.1.2 ÅBEN, IKONISK TOKEN (ELLER SYMPTOM)

*Here we have the case where we are trying to decide whether or not some actual observation has enough properties to be considered as some case.*

SHANK & CUNNINGHAM 1996, s. 3

Det er her vurderingen af, hvorvidt en eller flere egenskaber kan siges at være en del af et større, generelt fænomen, som kræver nærmere undersøgelse. Et eksempel herpå er symptomer på sygdom; en hoste kan være et tegn på, at man er ved at blive forkølet, men der kan også være tale om halsbetændelse eller at der simpelthen er røget noget i den gale hals. Vi udleder noget af tegnene, hvilket giver os anledning til at undersøge andre tegn (f.eks. andre symptomer).

### 2.1.3 ÅBEN, IKONISK TYPE (ELLER METAFOR/ANALOGI)

Shank & Cunningham beskriver denne type af abduktiv udledning som det at manipulere med ligheden for at opdage en mulig regel.<sup>19</sup>

Denne udledning er en udforskning af potentielle regler for sammenhæng ved at sidestille med mere eller mindre lignende emner. Formålet kan f.eks. være at få et nyt syn på et sæt regler.

---

<sup>18</sup> Brier og Shank er dog uenige om andetheden og tredjeheden. Hvor Brier knytter deduktion til andetheden, argumenterer Shank for, at induktion er den logiske stillingtagen til virkeligheden (det singulære), hvor deduktion er af mere formel karakter (jf. symbolets tredjehed). Denne uenighed har dog ingen konkret påvirkning på dette projekt, da dette projekt udarbejdes abduktivt.

<sup>19</sup> Shank 1996, s. 4

F.eks. kan man nå til nye konklusioner ved at anskue interaktionen mellem computer og bruger som et græsk teaterstykke, som f.eks. Brenda Laurel gør det i "Computers as theatre"<sup>20</sup>. Et andet eksempel er at bruge sine erfaringer fra brætspil som Matador eller Ludo, når man spiller andre brætspil, som f.eks. Trivial Pursuit.

Shank & Cunningham fremhæver i nedenstående citat, at det kan være relevant at ligestille emner, der ikke umiddelbart har noget til fælles, idet det kan lede til nye forståelser af det, der undersøges.

*Notice that there is no requirement that the juxtapositions make explicit sense. We can get away with this due to the fact that human beings are compelled to render any juxtaposition meaningful, or at least as meaningful as possible. In fact, juxtapositions which are arbitrary can be quite useful, in that they can lead to new areas of insight and understanding.*

SHANK & CUNNINGHAM 1996, s. 4

#### 2.1.4 ÅBEN, INDEKSIKALSK TOKEN (ELLER LEDETRÅD)

*This type of inference deals with possible evidence.*

SHANK & CUNNINGHAM 1996, s. 4

Denne form for abduktiv udledning nærmer sig deduktiv udledning, idet der her er tale om en form for bevisførelse. Ledetråden er dog ikke nødvendigvis entydig, og der er derfor tale om *muligheden* for beviser. Dette er detektivarbejdet. Kan læbestiften på cigaretten påvise, at gerningsmanden var en kvinde?

#### 2.1.5 ÅBEN, INDEKSIKALSK TYPE (ELLER DIAGNOSE/SCENARIE)

*This type of inference involves the formation of a possible rule based on available evidence.*

SHANK & CUNNINGHAM 1996, s. 4

Denne type af abduktiv inferens kan ses som en direkte forlængelse af den åbne, indeksikalske token. Her samles mulige beviser til et scenarie eller en diagnose, der forbinder de forskellige ledetråde.

*It is the act of reasoning that also finishes off the detection process by the creation of plausible scenarios from the body of clues.*

SHANK & CUNNINGHAM 1996, s. 4

Hvis vi fortsætter detektiv-analogien, kan vi anskue dette som detektivens arbejde med at samle trådene og formulere mordgældens scenarie.

---

<sup>20</sup> Laurel sammenligner en brugssituation med en computer med et handlingsforløb, der kan sættes på en spændingskurve og som udløser khartasis, f.eks. når man trykker 'Print' efter et endt arbejde.

### 2.1.6 ÅBEN, SYMBOLSK TYPE (ELLER FORKLARING)

Dette er forsøget på at danne en formel regel ved at udarbejde en troværdig forklaring. Shank & Cunningham understreger, at en god forklaring ikke blot forklarer det åbenlyse men belyser det, som før var uklart, eller som ikke 'passede ind'.<sup>21</sup> Der er altså stadig tale om abduktiv udledning, og ikke en deduktiv, logisk slutning.

I forbindelse med vores mordsags-eksempel, er dette advokatens bevisfremførelse i sagen.

## 2.2 ABDUKTION I DETTE PROJEKT

Jeg vil som sagt arbejde abduktivt i dette projekt. Den abduktive fremgangsmetode betyder, at projektet ikke vil have en entydig konklusion, men nærmere angive en mulig fremtid for semantic software og visualiseringer af analyser baseret på sådanne.

Jeg har i dette kapitel gennemgået seks forskellige abduktive modeller, der er udledt af Peirces kategoriseringsklasser. Af disse modeller vil jeg hovedsageligt benytte mig af metaforer, diagnoser og forklaringer. Metaforer bruges eksempelvis ved at benytte tidligere udviklingsmønstre til at forsøge at forudsige en fremtidig udvikling inden for webteknologi. Diagnoser og forklaringer bruges i rapporten til at tegne et billede af den nyeste udvikling inden for semantic software og gennem eksisterende viden om f.eks. begrebspsykologien, at forsøge at beskrive nogle af de problemstillinger, der kan forventes at opstå som følge af en stigende brug af semantic software.

## 2.3 ØVRIGE VIDENSKABSTEORETISKE OVERVEJELSER

I rapporten udledes naturligvis også enkelte deduktive slutninger –en rapport uden logiske sammenhænge ville formentlig være komplet uforståeligt.

Da genstandsfeltet for dette projekt, semantic software, endnu kun er sparsomt implementeret i kontrollerede miljøer på WWW, og da den kommende internetudvikling ikke er til at tage og føle på, har det ikke været oplagt at lave empiri eller induktive slutninger. Rapportens sidste del om visualiseringer lægger dog op til yderligere forskning og empiriske undersøgelser af brugeres perception og tolkning af visuelle repræsentationer af store mængder af semantiske data.

Meget af den litteratur, der vedrører den nyeste forskning og den nyeste udvikling på World Wide Web, er endnu ikke beskrevet akademisk. Mange af diskussionerne og litteraturen er derimod tilgængelig på web-mediet selv, hvorfor der bruges adskillige kilder fra WWW, eksempelvis fra diverse frontfigurer, forfattere og kritikeres blogs, hjemmesider eller optagelser af foredrag med disse.

---

<sup>21</sup> Shank & Cunningham 1996, s. 4

### 3 WEB 2.0 OG WISDOM OF CROWDS

Udtrykket Web 2.0 antyder måske, at WWW er noget, der er centralt styret, og som fra tid til anden bliver udgivet i en ny version. Men man hører aldrig om Web 1.0, og Web 2.0 er ikke et produkt. Så hvor stammer udtrykket fra?

Tim O'Reilly er blevet tildelt æren for at have opfundet udtrykket 'Web 2.0'. Han beskriver selv begrebet som at være opstået på en konference-brainstorm mellem forlaget O'Reilly og konferencebureauet MediaLive International. Diskussionen forinden havde handlet om, hvordan enkelte IT-virksomheder havde overlevet dot-com-bølgen i 2001. Enorme mængder af virksomheder var med til at oprette webshops og stod klar til at erobre det nye online-marked, men kun få overlevede. O'Reilly og MediaLive identificerede en række egenskaber for disse overlevende, og kaldte disse tendenser for Web 2.0.<sup>22</sup>

Tim O'Reillys gennemgår i sin artikel 'What is Web 2.0?'<sup>23</sup> de tendenser, som konferencedeltagerne mener tegner Web 2.0. Web 2.0 er altså ikke blot en enkelt tendens, men flere forskellige måder at anskue internettet på. Tim O'Reilly beskriver følgende fire tendenser:

- Internettet som platform
- World Wide Web skaber og bruger fællesskaber
- Data er værdi
- Konstant beta

Jeg vil følgende gennemgå disse fire tendenser.

#### 3.1.1 INTERNETTET SOM PLATFORM

O'Reilly ser ikke kun internettet som en platform for publikationer, men som den grundlæggende platform for software. Han beskriver internettet som en konkurrent til softwareudgivelser i traditionel forstand, hvor man køber et produkt i en æske. Software kan på internettet distribueres løbende og uafhængigt af fysisk distribuerede produkter. Der er ikke noget krav til, at du går ned i forretningen og køber den nyeste version og installerer den på din computer – derimod går du online og bruger den *webservice*, du ønsker, som altid vil være tilgængelig i den nyeste version.

*This time, though, the clash isn't between a platform and an application, but between two platforms, each with a radically different business model: On the one side, a single software provider, whose massive installed base and tightly integrated operating system and APIs give control over the programming paradigm; on the other, a system without an owner, tied together by a set of protocols, open standards and agreements for cooperation.*

O'REILLY 2005, s. 4

Internettet anskues hos O'Reilly ikke som et katalog over information, men som en platform for services og data. Som eksempel bruger han Google, som forstod at levere en enkel service uden overflødig distribuering af softwarepakker:

---

<sup>22</sup> Konferencens resultater er gengivet i O'Reilly 2005

<sup>23</sup> O'Reilly 2005

*No scheduled software releases, just continuous improvement. No licensing or sale, just usage.*

O'REILLY 2005, s. 3.

Webservicen Writely - der senere blev opkøbt af Google og i dag kendes som Google Docs<sup>24</sup> - er eksempelvis en onlineversion af en type program, som forbrugere tidligere måtte gå ned i en butik og købe og installere på sin computer, nemlig et tekstbehandlingsprogram. Webservices som Writely er ofte mindre og lettere programmer end dem, man traditionelt køber i butikken. Til gengæld er de tilgængelige, uanset hvilken computer man sidder ved, dvs. at man har adgang til sine dokumenter, der hvor man er. Programmet skal aldrig opgraderes eller opdateres - du sidder altid med den nyeste version af programmet, som bliver udviklet løbende.

### 3.1.2 WORLD WIDE WEB SKABER OG BRUGER FÆLLESSKABER

I artiklen fremhæver Tim O'Reilly også World Wide Webs evne til at samle og aggregere intelligens. Påstanden er, at visse ydelser bliver bedre des flere brugere, systemet har. Dette gælder især for de sociale services<sup>25</sup> på World Wide Web som f.eks. Facebook<sup>26</sup>, LinkedIn<sup>27</sup> eller Last.fm<sup>28</sup>, hvor det er bl.a. er forbindelsen til andre brugere, der skaber værdien i systemerne.

Last.fm er en slags online radio, hvis spilleliste tilpasses den enkelte bruger, og det er et godt eksempel på, hvordan mængden af brugere skaber forbedret kvalitet for hver enkelt bruger. Man kan bruge Last.fm på flere forskellige måder - f.eks. ved at indtaste eller vælge en bestemt gruppe eller kunstner, hvorefter Last.fm vil afspille musik, der relaterer sig til valget. En anden metode er at lade Last.fm 'optage', hvilke sange der afspilles i din medieafspiller på computeren og ud fra disse afspilninger lave en statistik på, hvilken type musik du kan lide. Det, der er det interessante ved Last.fm, er, at de data, hvormed de individualiserede spillelister genereres, kommer fra andre brugere. Når en bruger har installeret Last.fm på sin computer, bidrager han til en stor database med titler på de numre, der findes på computeren, og data om hvor ofte numrene bliver spillet. Ved at sammenligne forskellige profilers musikmag afspilles musik, som andre profiler, der ligner brugerens profil, godt kunne lide. Princippet er det samme som i Amazon.coms 'Users who bought this book, also bought ...'.

Fænomenet kaldes Collaborative Filtering<sup>29</sup> (CF), og bruges mange steder på nettet, hvor brugere har profiler. CF har sin styrke i, at den usynligt og automatisk kan skabe relevante anbefalinger for den enkelte bruger. Brugeren skal således ikke først identificere sine ønsker

---

<sup>24</sup> <http://docs.google.com>

<sup>25</sup> Clay Shirky definerer kort og godt social software således: "It's software that supports group interaction.", Shirky 2003, s.1

<sup>26</sup> <http://www.facebook.com> - Facebook er en social webservice, hvor hver bruger har en profil med en række 'applikationer' der knytter sig til denne profil. På Facebook forbindes profiler med hinanden ved at to profiler kan være venner eller være i samme gruppe, f.eks. efter studiested.

<sup>27</sup> <http://www.linkedin.com> - LinkedIn er en social webservice med fokus på professionelle forbindelser og profilering. På LinkedIn har hver bruger connections til andre brugere. Man har på denne måde mulighed for at kortlægge sit netværk - og udvide det, ved at se sine forbindelsers netværk.

<sup>28</sup> <http://www.last.fm> - Last.fm er en online musik-radio med dynamiske radiostationer. Hver bruger har mulighed for at få afspillet en radiokanal der er specifik for dem, baseret på en analyse af deres musikstil. Analysen kan være baseret på forskellige data, f.eks. den musik brugeren har liggende på sin computer eller deres bærbare musikafspiller.

<sup>29</sup> Ikke at forveksle med Information Filtering (IF), som tager udgangspunkt i et predefineret datasæt, og ikke data indsamlet om andre brugere. Se evt. Good et al. 2005.

eller interesser, idet analysen af den enkelte profil baseres på den enkelte brugers konkrete brug af webservicen. Collaborative Filtering fungerer ud fra en forudsætning om, at folk, der ligner hinanden på en række punkter, også vil ligne hinanden på øvrige punkter. Til gengæld kræves der et forholdsvist højt antal brugere, før systemet giver mening.<sup>30</sup>

Collaborative Filtering er blot et af flere eksempler på, hvordan fællesskaber og et højt antal brugere kan skabe værdi. En anden effekt af de store mængder af brugere er evnen til at filtrere viden. Et eksempel på dette er Wikipedia, et online leksikon, hvor alle brugere har mulighed for at rette ethvert opslag. Det overraskende ved Wikipedia er, at kvaliteten af leksikonet er utroligt høj og faktisk sammenlignelig med et leksikon skrevet af 'professionelle' forfattere.<sup>31</sup> O'Reilly påpeger, at mængden af indlæg i f.eks. blogs også udgør en lignende filtrering af viden, hvor folk kommenterer og diskuterer viden og fakta sammen med holdninger.

*If it were merely an amplifier, blogging would be uninteresting. But like Wikipedia, blogging harnesses collective intelligence as a kind of filter"*

O'REILLY 2005, s. 9.

O'Reilly nævner i denne sammenhæng James Surowieckis begreb *Wisdom of crowds*, hvilket jeg vil vende tilbage til senere i dette kapitel.

### 3.1.3 DATA ER VÆRDI

Hvor det tidligere har været producenter af softwarepakker eller hardwareløsninger, der har tjent penge i IT-industrien, spår O'Reilly data til at være det, der bærer den store værdi. De virksomheder, der vil komme til at tjene penge på data, er dels dem, der forstår at indsamle de rigtige data og dels dem, der forstår at anvende data i gode webservices.

*In many cases, where there is significant cost to create the data, there may be an opportunity for an Intel Inside style play, with a single source for the data. In others, the winner will be the company that first reaches critical mass via user aggregation, and turns that aggregated data into a system service.*

O'REILLY 2005, s. 11

Et godt eksempel herpå er de virksomheder, der har brugt mange penge på at affotografere hele Danmark fra flyvemaskiner og derpå sat det sammen til et fotografisk kort. Men også de firmaer, der tilbyder, at man kan slå en adresse op og finde den på kortet – og ikke blot se et kort, men se et luftfoto af den adresse, man søger. Et andet eksempel er Amazon.com, som dels indsamler data om brugernes klikstier og varehistorik, og dels tilbyder et anbefalingssystem<sup>32</sup>, der anbefaler brugeren varer baseret på, hvad andre brugere, hvis adfærd på hjemmesiden ligner denne brugers, har købt.

---

<sup>30</sup> Good et al. 2005, s. 2f

<sup>31</sup> Tidsskriftet Nature har forsøgt at sammenligne Britannica og Wikipedia. Den enkle undersøgelse viser, at der er overraskende mange fejl i Britannica og at der hverken er stor forskel i antallet af alvorlige (ud af de undersøgte opslag, blev der fundet fire alvorlige fejl i begge opslagsværker) eller i forkerte fakta/vildledende beskrivelser (162 i Wikipedia mod 123 i Britannica) i de to opslagsværker. Se Giles 2005.

<sup>32</sup> Amazon.com benytter sig også af Collaborative Filtering

Der er en klar forbindelse mellem opfattelserne af at såvel mængden af data som antallet af brugere har betydning for visse typer af websites værdi. Disse to tendenser kombineres ofte, således at brugerne er med til at skabe data. Dette sker f.eks. på det førnævnte Last.fm, hvor systemet bliver bedre til at skabe brugerspecifikke radiostationer, når der er mange brugere, der med hver deres musikstil præger og raffinerer systemet.

#### 3.1.4 KONSTANT BETA

Som nævnt i det første punkt, internettet som platform, har webservices en klar fordel over software, der skal installeres, derved at det altid er tilgængeligt i den nyeste version. Dette betyder ofte, at webservices konstant er under udarbejdelse, da der løbende laves rettelser og forbedringer. Dette står i kontrast til klassisk computersoftware, der udgives i færdige versioner, såkaldte *releases*. Før et stykke software bliver udgivet, er det ofte tilgængeligt i en betaversion, hvilket vil sige en næsten færdig version, der testes af en forholdsvis stor gruppe mennesker, som afrapportere evt. fejl, som så rettes i den endelige udgivelse. Konstant beta (eller *perpetual beta*, som O'Reilly selv kalder det) dækker over, at der løbende modtages feedback fra brugerne, og at programmet derfor løbende forbedres.

Perpetual beta er ikke blot et begreb, der kan bruges på webservices. Langt de fleste softwareprogrammer kan i dag downloades fra internettet og opdateres gennem en internetforbindelse. Såvel styresystemer som nicheprogrammer, såsom internetbrowsere og programmer til afspilning af medier, opdateres jævnligt gennem opdateringsservices indbygget i softwaren, og nye versioner af softwaren tilbydes jævnligt producenterne hjemmesider. Internettet har med andre ord frigjort softwaren fra de fysiske udgivelser.

Feedback spiller en enorm rolle i denne sammenhæng. Med perpetual beta, software der aldrig bliver udgivet som 'færdig software', er alle brugere beta-testere. Udviklerne har på denne måde enorme mængder af data at finde fejl ud fra, og det gør det nemmere at tilrette produktet. Inspirationen kommer muligvis fra Open Source-communityet, hvor softwaren udvikles af brugerne selv.

*The open source dictum, "release early and release often" in fact has morphed into an even more radical position, "the perpetual beta," in which the product is developed in the open, with new features slipstreamed in on a monthly, weekly, or even daily basis. It's no accident that services such as Gmail, Google Maps, Flickr, del.icio.us, and the like may be expected to bear a "Beta" logo for years at a time.*

O'REILLY, 2005 s. 17

## 3.2 WISDOM OF CROWDS

I sin bog, *The Wisdom of Crowds – Why the Many Are Smarter Than the Few*, forklarer James Surowiecki en lang række eksempler, der illustrerer hvordan store grupper under rette forhold kan være bedre til at løse problemer end enkeltindivider. Selvom Surowiecki ikke beskriver kollektiv intelligens på nettet, er Wisdom of crowds-begrebet dækkende for menneskers evner til – ofte autonomt – at løse komplicerede opgaver gennem kollektiv tænkning.

### 3.2.1 DEN GENNEMSNI TLIGE PRÆCISION

Han indleder med at beskrive et dyrskue i West England, hvor man i starten af det forrige århundrede kunne vinde en okse, hvis man kunne gætte, hvad den ville veje efter partering. Den, der kom tættest på det rigtige bud, ville vinde. Det rigtige svar, som de mange bondemænd prøvede at gætte, var 1198 pund. Hvad det tætteste bud var, melder historien desværre ikke noget om. Men tager man alle de individuelle bud og fandt gennemsnittet, ville man finde hele gruppens bud.

*That number represented, you could say, the collective wisdom of the Plymouth crowd. If the crowd were a single person, that was how much it would have guessed the ox weighed.*

SUROWIECKI 2004, INTRODUCTION S. XIII

Gruppens bud (baseret på gennemsnittet af 787 bud) var 1197 pund. Altså et næsten perfekt bud.

Som sagt ved vi i dag ikke, om der var nogen af deltagerne, der kom tættere på. Men lad os for eksemplets skyld sige, at en af 787 deltagere, vinderen af konkurrencen, rent faktisk gættede rigtigt – at der var en enkelt person, der vidste bedre end gruppen. Dette individ ville måske endda være i stand til at gætte rigtig hver eneste gang: en sand ekspert i at vurdere, hvad okser vejer efter partering.<sup>33</sup>

Det interessante for alle os andre, os ikke-eksperter, er så, hvordan man nemmest får en god vurdering af, hvad oxen vejer efter partering? At spørge en stor gruppe mennesker, der hver giver deres egen personlige, umiddelbare vurdering? Eller at søge alle deltagerne igennem for at finde den ene sande ekspert blandt de 787 mennesker, der virkede mest kvalificeret til at vurdere oksens vægt, og stole på, at man havde fundet den rigtige?

Wisdom of crowds udmærker sig altså ikke nødvendigvis ved, at en gruppe af mennesker vil være bedre til at vurdere en sag end en ekspert på et givent område. Men det kan ofte være den nemmeste måde at få et kvalificeret bud på.

Nu kan man argumentere for, at det at vurdere kvægs vægt ikke nødvendigvis er en af vor æras største udfordringer, og at eksemplets kompleksitet er forholdsvist beskedent, til trods for gruppens overraskende præcision. Sagens kerne er naturligvis, om vi kan bruge gruppers kollektive visdom til noget konstruktivt?

Surowiecki, begejstret som han ellers er for Wisdom of crowds, påpeger, at den kollektive visdom langt fra altid vil overstige individers. Eksempelvis er der ikke umiddelbart noget, der tyder på, at grupper er velegnede til at få idéer og udvikle dem.

*Invention may still be an individual enterprise . . . , but selecting among inventions is a collective one.*

SUROWIECKI 2004, S. 59

---

<sup>33</sup> Surowiecki 2004, Introduction s. XI f

Til gengæld er de enormt gode til at udvælge de bedste. Han sammenligner Wisdom of crowds med evolution på speed – evnen til at imitere og udvælge de bedste idéer gør den evolutionistiske, naturlige udvælgelse overflødig.

*Used well, imitation is a powerful tool for spreading good ideas fast – whether they be in culture, business, sports, or the art of wheat eating. At its best, you can see it as a way of speeding up the evolutionary process – the community can become more fit without the usual need for multiple generations of genetic winnowing.*

SUROWIECKI 2004, s. 59

Denne form for udvælgelse er eksempelvis den der gør sig gældende, når en type af produkt sælger bedre end andre typer. Desværre er denne naturlige udvikling dog også påvirket af andre elementer, f.eks. reklamer og falsk information, der kan resultere i såkaldte informationskaskader.

### 3.2.2 IMITATION OG INFORMATIONSKASKADER

Informationskaskader er det problem, som Surowiecki fremhæver som det mest fremtrædende i forbindelse med Wisdom of crowds. Informationskaskader er i bund og grund en ukontrolleret aggregering, der er styret af, hvornår hændelser sker.

Under optimale forhold vil en god idé sprede sig som ringe i vandet. En person, der får en god idé, viser andre sin idé, og efter at have vurderet idéen viser disse personer idéen til flere mennesker. Men nogle idéer sælger sig selv uden egentlig at være gode idéer. Surowiecki bruger et eksempel<sup>34</sup>, hvor en Amerikaner ved navn George Geddes ved et besøg i Canada havde opdaget, at planke-veje (dvs. veje bygget af planker) ikke blot gjorde det nemmere at rejse med kærer og hestevogne, men at det ligefrem var mere skånsomt: Man kunne simpelthen spare penge på sine hestevogne. Så var det egentlig et ganske enkelt regnestykke: Hvad koster plankevejene og hvad sparer alle de rejsende i udgifter til køretøjer? Plankevejspioneren Geddes kalkulerede sig frem til, at en plankevej ville kunne holde i otte år, og dermed være en rentabel forretning. Så i 1846 begyndte han at bygge veje. Vejene var en stor succes, og den behagelige form for transport betød, at der inden for 10 år var 352 plankevejeentreprenører i New York alene. Problemet var bare, at idéen om plankevejene var bedre end den udregning, der lå til grund for estimatet på otte års holdbarhed.

*Plank roads did not last the eight years Geddes had promised (let alone the twelve years that other enthusiasts had suggested). As Klein and Majewski show, the roads' actual life span was closer to four years . . .*

SUROWIECKI 2004, s. 52

Menneskemængden kan altså sommetider være mindre vis. Problemet i ovenstående tilfælde er, at ingen forholdt sig kritisk til vejenes holdbarhed, før de begyndte at producere dem. Den blinde tillid til den gode idé, som kunderne tog så varmt imod, endte derfor med at koste mange

---

<sup>34</sup> Surowiecki 2004, s. 51f

af de nyetablerede virksomheder deres levebrød. I 1880, bare 34 år efter, de første plankeveje var blevet bygget, var nærmest alle plankeveje forladte og forsømte.<sup>35</sup>

Kilden til overstående problematiske udvikling er ifølge Surowiecki et spørgsmål om, at beslutninger ikke træffes kollektivt men af individer over tid. Mange kender situationen, at man bruger andre individers valg til at træffe en beslutning, f.eks. ved at vælge at spise i den restaurant, der har flest besøgende (til trods for, at vi ikke ved, om restaurantens besøgende har truffet valget på samme baggrund). Problemet opstår altså, når folk vælger at gøre som de andre uden at forholde sig kritisk til deres egen imitation andres idéer.

### 3.2.3 GRUNDTVILKÅR FOR WISDOM OF CROWDS

Wisdom of crowds opstår mange gange af sig selv. Oftest i forbindelse med kognitions-, koordinations- og kooperationsproblemer, som Surowiecki anser for at være de felter, hvor Wisdom of crowds kan gøre nytte.

**Kognitionsproblemer** er problemer, der kan løses ved at tænke sig frem til et svar. Det kan f.eks. være at finde en løsning på et komplekst problem eller at gætte vægten på en okse.

**Koordinationsproblemer** er hvad vi løser, når vi går ned ad Strøget uden at støde frontalt ind i andre mennesker, eller når vi skaber uskrevne regler, som f.eks. at den, der først sætter sig på en plads i toget, har ret til at sidde på denne plads.

**Kooperationsproblemer** er dem, der tit kæmpes med i det politiske system, hvor et helt samfund skal enes. Kooperations- og koordinationsproblemer har overlap (er kødannelse i supermarkedet samarbejde eller blot koordinering?), men Surowiecki nævner primært problemer, hvor målet for samarbejdet i en eller anden forstand har en håndgribelig værdi.

Udover disse tre grundområder opstiller Surowiecki også fire betingelser for, at Wisdom of crowds bliver en succes<sup>36</sup>:

1. **Diversitet.** Det er ikke blot et spørgsmål om, at man skal have mange med i en gruppe, før de kan træffe gode valg. De skal også have forskellige meninger. Jo flere meninger, des bedre bliver det endelige resultat.
2. **Uafhængighed.** Hvis folk ikke er i stand til at tænke selv, vil de blot gøre som alle andre, og så er det kun et spørgsmål om, at den, der handler først, sætter trenden for alle andre (dvs. informationskaskader).
3. **Decentralisering.** Hvis folk alle kommer fra den samme organisation, hvis de alle er ledet af den samme leder, eller hvis deres handlinger er internt afhængige i gruppen, er der større sandsynlighed for, at de vil mene det samme eller lade sig inspirere af hinanden. Surowiecki nævner det ikke selv, men man kan sige, at mangel på decentralisering ikke fordrer uafhængighed og diversitet, hvilket netop er de to første punkter.
4. **Aggregering**<sup>37</sup>. Når alle folk har en stemme, kan de nemt komme til at overdøve hinanden. Der er brug for en eller anden form for system, der kan samle folks stemme.

---

<sup>35</sup> Surowiecki 2004, s. 52

<sup>36</sup> Disse fire betingelser er oversat fra engelsk: Diversity, independence, decentralization og aggregation.

### 3.2.4 WISDOM OF CROWDS OG DEMOKRATI

Det er oplagt at bruge demokrati som et eksempel på Wisdom of crowds. Surowiecki diskuterer dette synspunkt, og han hælder til den holdning, at demokratiet ikke er et udtryk for Wisdom of crowds. Han begrundelse er, at en demokratisk afstemning, som de har dem i USA, ikke umiddelbart løser hverken et kognitions-, koordinations- eller kooperationsproblem. Et af problemerne er, at vælgerne ikke stemmer for at løse et problem, men for at udpege en ekspert, som skal tage stilling for dem. Herudover er det problematisk, at så få amerikanere er oplyst om politiske forhold. Dette kan skyldes, at de ikke bliver aktivt involveret i beslutningsprocesserne, men blot vælger deres repræsentant i det politiske system. Surowiecki nævner som eksempel en undersøgelse, der har vist, at amerikanerne tror, at for hver \$1 der går på u-landshjælp, går der \$3 på det amerikanske forsvar, hvor forholdet faktisk er \$1 til \$19. Amerikanerne mener generelt, at USA bruger for mange penge u-landshjælp. Men som Surowiecki skriver:

*One reason may be that, as another University of Maryland poll shows, Americans think the United States spends 24 percent of its annual budget on foreign aid. The reality is that it spends less than 1 percent.*

SUROWIECKI 2004, s. 266

Surowiecki diskuterer også, hvorvidt den enkelte borger stemmer for sit eget bedstes skyld – eller for hele landets skyld. Som Churchill erkender han, at demokratiet formentlig er den 'mindst-slemme' styringsform.<sup>38</sup> Han så dog hellere en form for demokrati, hvor repræsentativt udvalgte borgere mødtes for at diskutere politiske problemstillinger. I sådan et forløb ville der være mulighed for at orientere borgerne om politiske problemstillinger på en fyldestgørende måde, således at borgerne kan tage stilling ud fra et bedre grundlag. Surowiecki diskuterer ikke desto mindre ikke, hvem der skal fremlægge denne viden, eller hvordan man sørger for at den ikke bliver partisk.<sup>39</sup>

### 3.2.5 KRITIK AF SUROWIECKIS BETINGELSER

Desværre virker Surowieckis grundbetingelser ikke særligt solide. De tre grundområder for Wisdom of crowds' virke er umiddelbart meget brede. Og med de skeptiske briller på ville man endda kunne sige, at kognitions-, koordinations og kooperationsproblemer tilsammen dækker over nærmest ethvert tænkeligt problem (selvom Surowiecki ikke selv mener, at politiske problemer som sådan er en del af disse områder og demokrati ikke viser Wisdom of crowds fra sin ypperligste side<sup>40</sup>).

Herudover er det også svært at tænke sig scenarier, hvor uafhængige, decentraliserede og forskelligartede menneskers holdninger skal kunne aggregeres. Det kan nærmest kun lade sig gøre, uden at mængdens individer er klar over deres rolle i mængden. I eksemplet, hvor en række dyrskuegæster skulle gætte vægten på en okse, er det svært at forestille sig, at der har

---

<sup>37</sup> Jeg bruger ordet aggregering (at aggregerer, eller en aggregator), i den forstand, at adskillige kilder samles og får et nyt, fælles udtryk. Man kan således sige at en afstemning aggregerer de enkelte deltagers stemmer, eller at en nyhedstjeneste der indsamler nyheder fra forskellige medier og samler dem på en hjemmeside fungerer som aggregator.

<sup>38</sup> Surowiecki 2004, s. 270

<sup>39</sup> Surowiecki 2004, s. 259ff

<sup>40</sup> Surowiecki, 2004, s. 259ff

været mange forskellige typer af mennesker: De kommer formentlig alle fra den samme by, og de fleste af dem har nok haft deres erhverv inden for landbruget. Ydermere er de alle placeret det fysisk samme sted under konkurrencen og har muligvis kunnet aflure hinandens svar. Et eksempel, der ikke stammer fra Surowiecki selv, men som ofte fremføres som et strålende eksempel på Wisdom of crowds, nemlig Wikipedia.org, udmærker sig ved, at folk ikke kommer fra det samme sted eller er ansat til at lave det samme. Der er altså gode kår for både decentraliseringen og diversiteten. Men idet stort set alle artikler er skabt og derpå tilrettet, er stort set alle artikler ikke skabt uafhængigt, men derimod i samspillet mellem individer.

Man skal dog være opmærksom på, at sammenarbejdet mellem mennesker ikke nødvendigvis er lutter lagkage. Clay Shirky<sup>41</sup>, en forfatter, kritiker og professor fra NYU, nævner i 'Here Comes Everybody' begrebet, *Tragedy of the Commons*. Shirky illustrerer Tragedy of the Commons med et lille samfund af fårehyrder, som hver har deres land med græs og hver deres gruppe får. Målet for enhver fårehyrde er at få sine får så fede så muligt for så lav en omkostning som muligt. Den optimale løsning er, at de alle bliver enige om at lade deres får græsse samtidig og kun på deres eget græs. Hvis din nabo vælger at lade sine får græsse længere tid, bliver du nødt til også at lade dine får græsse længere, så du kan holde øje med, at din nabos får ikke græsser på dit græs og dermed tærer på dine ressourcer. Med andre ord fungerer alt fint, så længe der er tillid mellem fårehyrderne. Men når der er tillid mellem alle fårehyrder er det nemt for enhver fårehyrde at udnytte de andres marker.<sup>42</sup>

Tragedy of the Commons handler om forholdet mellem personlige og sociale udbytter. Summen af udbytte er størst, hvis alle samarbejder, men hvis bare én snyder, får snyderen meget ud af dette, mens fællesskabet lider. Det er derfor, skattesystemet er sat i system, og det er derfor, den kolde krigs oprustninger ingen ende ville tage.

Såvel informationskaskader som Tragedy of the Commons er begreber, man bliver nødt til at være opmærksomme på, når man forsøger at udnytte menneskemængder i kollektive opgaver. Informationskaskader kan være svære at stoppe, idet det ofte er manglen på den rigtige viden, der tillader de falske drømmebilleder at eksistere længe. Endnu en gang kan man bruge Wikipedia som eksempel på, at når der er tilstrækkeligt mange brugere, og tilstrækkeligt engagerede brugere, stilles der høje krav til kvaliteten af bidrag, hvilket på sigt minimerer risikoen for informationskaskader. Også i forbindelse med Tragedy of the Commons, er det nemmere for individer at snyde gruppen, hvis der kun er ganske få brugere. Er der tilstrækkeligt mange brugere, og tilstrækkeligt mange der føler sig snydt, spredes ordet og 'dårlige brugere' fryses ude. Man kan med andre ord sige, at et system bliver sikrere, des flere brugere der af systemet.

### 3.3 WEB PARADIGMER

Ser vi på begrebet *Web 2.0*, kan det virke paradoksalt, at et af grundelementerne i Web 2.0-tankegangen er, at der ikke længere er brug for versioner – at alting altid opdateres løbende. Som O'Reilly også viser i sin artikel, er der dog heller ikke tale om 'et nyt internet', men ganske enkelt en række tendenser som kendetegner de succesrige foretagender på internettet. Man bør

---

<sup>41</sup> Clay Shirky, f. 1964, adjunct på New York University, konsulent og forfatter

<sup>42</sup> Shirky 2008, s. 51f

derfor være forsigtig, når man bruger termen *Web 2.0*, da det ikke dækker over noget entydigt, men over flere forskellige udviklinger på internettet.

Kan Web 2.0 siges at være et paradigme?

Den klassiske forståelse af paradigmer stammer fra videnskabsteoretikeren Thomas Kuhn og definerer paradigmet som :

*universelt anerkendte videnskabelige præstationer som for en tid angiver eksemplariske problemer og løsninger for et samfund af aktive forskere.*

KJØRUP 2000, s. 105

Støtter man sig op af Kuhns paradigmebegreb, får vi svært ved at anskue Web 2.0 som et paradigme; primært fordi Kuhns paradigmebegreb dækker over det elitære videnskabelige samfunds<sup>43</sup> forbilleder – hvilket man ikke kan sige er gældende for Web 2.0-tankegangen. Web 2.0 er ikke udviklet af videnskabsmænd, men er – efter O'Reillys billede – et udtryk for de tendenser, som overlevede IT-boblens brist i 2001. Ud fra en streng videnskabshistorisk tilgang er der altså ikke tale om et paradigme.

Kjørup påpeger dog, at der i Kuhns paradigmebegreb ligger to forskellige forståelser af paradigmet; en smal og en bred. Den smalle er den, som præsenteres i citatet ovenfor og handler om det eksplicitte forbillede. Den brede forståelse af paradigmet har dog en mere generel karakter:

*I den brede betydning, derimod, er et paradigme hele den generelle tænkemåde som deles af medlemmerne af en bestemt videnskabelig skoledannelse og som er eksemplificeret i forbilledet.*

KJØRUP 2000, s. 107

Denne brede forståelse eksemplificeres i en såkaldt *faglig matrix*, der deler paradigmet op i symbolske generalisationer, modeller, værdier og forbilleder. Selvom der i udviklingen af internettet strengt taget ikke er tale om en videnskabelig skoledannelse, så meget som en naturlig udvikling, er det alligevel nærliggende at anskue Web 2.0 som et paradigme. Som det er anskueliggjort i O'Reillys artikel, foreligger der både klare værdier og modeller i Web 2.0-paradigmet. Værdier såsom forståelsen af, at data og services omkring data har en grundlæggende værdi på internettet. Og modeller såsom at kvaliteten stiger med antallet af brugere. Forbilledet for Web 2.0 kunne præsenteres som tilgængeligheden, dvs. evnen til at bruge internettet som en platform for brugerne, således at brugerne kan udtrykke sig på internettet uden at have teknisk viden. Der er med andre ord samhørighed mellem den fremstilling, O'Reilly giver af Web 2.0 og Kuhns brede paradigmebegreb.

Men anskuer vi Web 2.0 som et paradigme, melder der sig et akut spørgsmål: Hvorledes forholder Web 2.0 sig til andre 'web-paradigmer', og hvilke? Paradigmer forstås ikke som

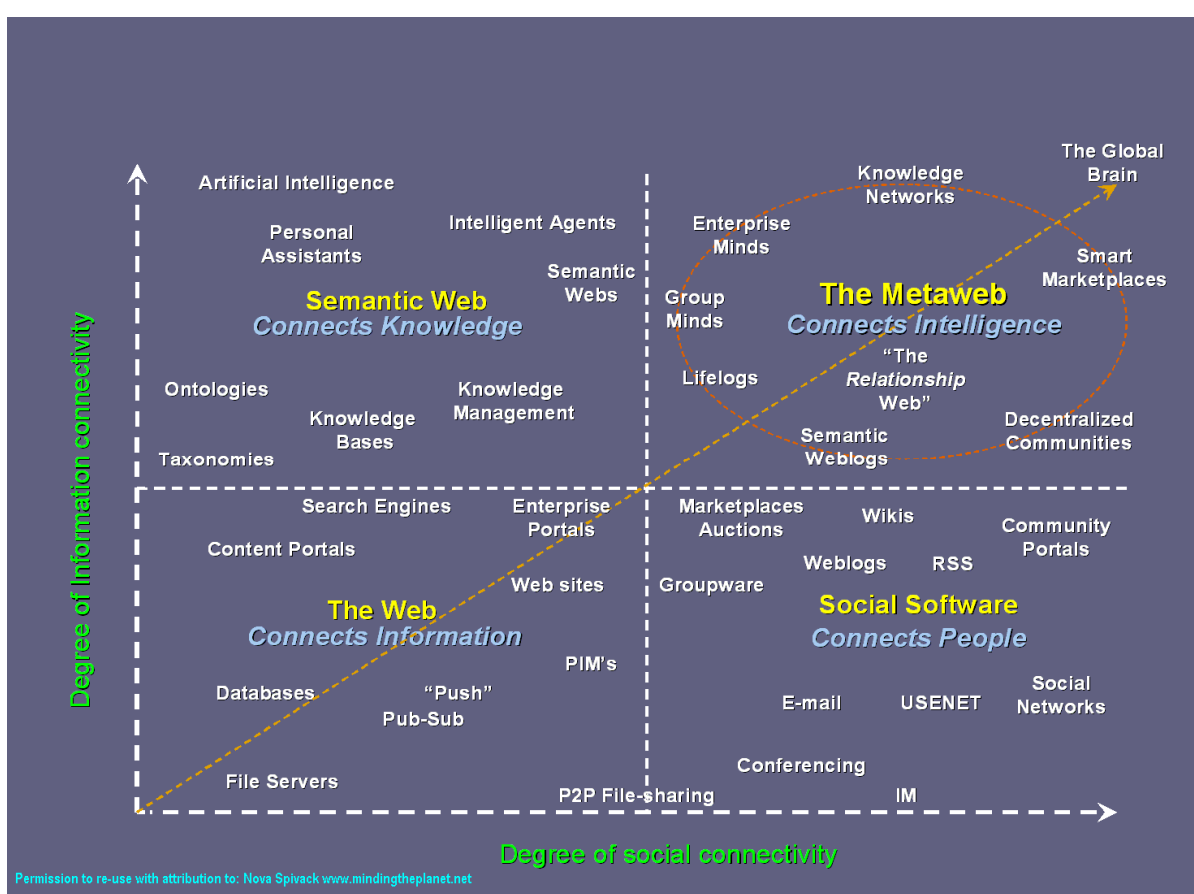
---

<sup>43</sup> Kjørup beskriver Kuhns opfattelse af *det videnskabelige samfund* som meget snæver: En gruppe på højst hundrede videnskabsmænd.

udviklinger, men som brud<sup>44</sup>, så hvad er det, Web 2.0-paradigmet bekæmper? Den amerikanske semantic-web-pioner Nova Spivack har et bud på et sådan paradigmeforløb.

### 3.4 DEN GLOBALE HJERNE

Nova Spivack er direktør i og grundlægger af firmaet Radar Networks, som står bag den endnu ikke offentlige webservice Twine. Twine har fået megen opmærksomhed på nettet, fordi det er en webservice, der kombinerer flere forskellige webteknologier, heriblandt teknologier som er meget nye i dag, men som har et enormt potentiale, og som forventes at blive en standard på WWW. Han iagttager to bevægelser i udviklingen af internettet: Bedre evne til at sammenkæde information og bedre evne til at sammenkæde mennesker. Man kan sige, at der er tale om en teknologisk og en social udvikling gennem web-teknologi.<sup>45</sup>



FIGUR 3 – NOVA SPIVACKS FIGUR OVER UDVIKLINGEN GENNEM FORBEDRING AF SOCIALE FORBINDELSER OG FORBINDELSEN AF INFORMATION

Spivack tegner et billede af, hvordan kommunikationsteknologi hjælper menneskeheden med at tage store spring i udviklingen. Som den første store begivenhed nævner han opfindelsen af

<sup>44</sup> Kjølrup 2000, s. 115

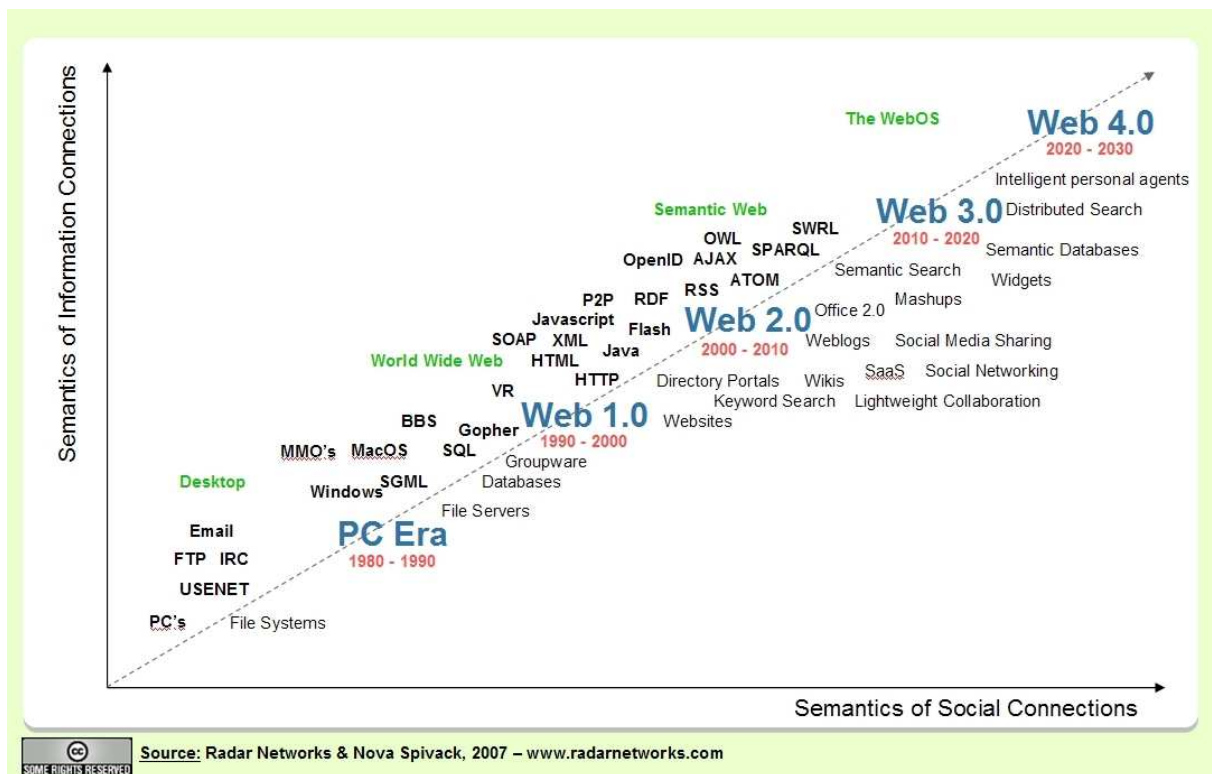
<sup>45</sup> Spivack 2007a

trykken, som tillod masseproducering af viden. En viden, der dog stadig er begrænset til et fysisk medie. Med opfindelsen af de elektroniske medier bliver viden pludselig nemmere at dele og at søge i.<sup>46</sup> Og nu har vi internettet, som tillader enhver at bidrage med viden – et projekt, der måske på sigt vil blive til 'den globale hjerne'<sup>47</sup>.

### 3.4.1 WEB-GENERATIONER

Spivack deler den elektroniske revolution ind i generationer – adskilt efter årtier.<sup>48</sup> Således taler Spivack altså ikke om Web 2.0 som en betegnelse for en række begivenheder, men ganske enkelt om WWW fra år 2000 til år 2010. Ved at opdele web-generationerne efter årstal mener Spivack altså at kunne undgå, at der hele tiden er nye interessenter, der tildeler sig selv æren for at være ham, der fandt på web X.0-begrebet.

Det er naturligvis en enklere og på mange måder mere uproblematisk måde at håndtere navngivningerne af paradigmerne på, men ikke desto mindre kan det gøre det svært at definere, hvad der tilhøre hvilken generation. Sat lidt på spidsen kan man sige, at det svarer til at påstå, at høj-renæssancen sluttede ved midnatstid d. 31. december 1619. Til Spivacks forsvar skal det naturligvis nævnes, at han ikke selv kalder web-årtierne for paradigmer, men for generationer. Uanset hvad de kaldes, kan man diskutere relevansen af en paradigme/generations-skellen, hvis de ikke er kendetegnede ved deres karakteristika men blot deres datoer.



FIGUR 4 – NOVA SPIVACKS FIGUR OVER WEB-GENERATIONERNE

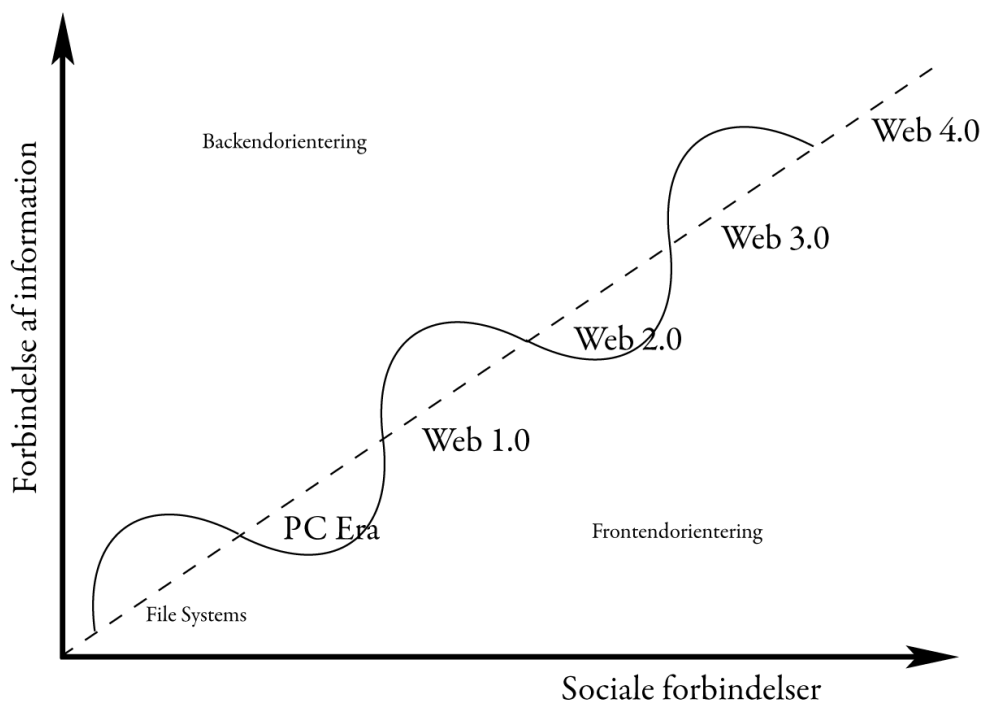
<sup>46</sup> Spivack 2007b, s. 4

<sup>47</sup> Spivack 2007b, s. 6

<sup>48</sup> Spivack 2007d, s.1

Spivack begynder ikke sit generations-forløb med Web 1.0, men med PC-æraen, der forløber fra 1980 til 1990. Dette er årene hvor PC'en får sit gennembrud. Herefter følger Web 1.0, 1990-2000, hvor WWW bliver etableret og får sit gennembrud. WWW fungerer på sin vis som en slags filserver, og der er fortrinsvis meget fokus på backend<sup>49</sup> – man skal f.eks. kunne HTML og have kendskab til, hvordan indholdet bliver publiceret på en webserver for at kunne tilføje indhold til internettet. Med Web 2.0, 2000-2010, bliver fokus flyttet fra backend til frontend; med øget kendskab til teknologien og nye værktøjer kan enhver oprette sin egen hjemmeside, blog eller deltage i fora og sociale netværk.

Med Web 3.0, 2010-2020, vil vi ifølge Spivack vende tilbage til at være mere fokuserede på backend. I WWW 3. generation vil der være øget fokus på at gøre internettet 'smart'. Der er mange data på nettet efter Web 2.0, men mange af disse data er ustrukturerede eller semistrukturerede data, som kan være svære at håndtere. Web 3.0 handler om at gøre internettet semantisk, utvetydigt og forbundet. Først med Web 4.0 vil fokus vende tilbage til frontenden, hvor Spivack forudser, at der vil dukke mange såkaldte *agenter* op, der gennem kunstig intelligens vil kunne navigere det semantiske web for brugeren og finde de informationer, der har relevans for den enkelte brugers interesse.



FIGUR 5 – NOVA SPIVACKS FIGUR HVOR SKIFTENE MELLEM FRONT- OG BACKENDORIENTERINGEN ER MARKERET

<sup>49</sup> Backend og frontend kan sammenlignes med en teaterscene, hvor backstage (backend) er der hvor teknikerne og skuespillerne forbereder og kontrollerer begivenhederne, hvor selve scenen (frontend) er det som publikum ser. På en hjemmeside kan backenden f.eks. være et CMS og frontenden den hjemmeside, som alle de besøgende ser.

Som nævnt anskuer Nova Spivack udviklingen inden for web i to forskellige retninger. Den ene beskriver udviklingen i evnen til at forbinde informationer over tid – den anden evnen til at lave sociale forbindelser over tid.<sup>50</sup> Med de to retninger for udviklingen - evnen til at forbinde informationer og evnen til at lave sociale forbindelser - bliver det naturligt at forvente, at det næste skridt er en automatisering, der gør det muligt for informationer at finde frem til de relevante modtagere: intelligent data.

### 3.5 KOLLEKTIV INTELLIGENS OG AMATØRISME

Surowiecki og Spivack tegner tilsammen en optimistisk fremtid for menneskeheden, hvor fællesskaber og menneskemængder tilsammen kan skabe og dele informationer i en grad, det aldrig tidligere har været muligt. Gennem teknologiske fremskridt fjernes der flere og flere af de forhindringer, der tidligere har gjort det svært at dele, forbinde og udbrede viden. Samtidig bliver kommunikationen uafhængig af tid og sted, og enhver kan kommunikere med enhver anden uden fysiske forhindringer. Offentlig diskussion, kommentering og kritik minimerer mængden af fejlbehæftet eller dårlig information, og med flere brugere stiger kvaliteten af indholdet.

Men både Surowieckis og Spivacks fremtidsscenarie kan virke jubeloptimistisk i deres fremstilling af den kollektive intelligens.. For selvom der med internetmediets frigørelse fra fysiske begrænsninger skabes uanede muligheder og udnyttelsesværdi, skabes der også problemer, som bør diskuteres.

#### 3.5.1 EKSPERTER VERSUS FOLKET

I dag har den akademiske verden en form for autoritet hvad angår viden. En historiker har eksempelvis en uddannelse, der sikrer et vist (om end altid begrænset) indblik i fagrelaterede emner, såsom kildekritik og en lang række historiske begivenheder. En hovedopgave og en eksamen skal endvidere bestås, før den studerende bliver anerkendt som historiker. Uddannelsen er sat i system, og samfundet anerkender uddannede menneskers faglighed. Hvis folket bliver i tvivl om et historisk emne, stoler vi på eksperternes, historikernes, udlægning af, hvad der faktisk er sket.

I dag har eksperterne en totalitær magt over viden. Det betyder, at fakta ikke kan diskuteres, og selv hvis flertallet i et samfund er overbeviste om, at et er tilfældet, er det i sidste ende de elitære eksperter, der beslutter hvad der er sandheden. Forfatterinden Ninka-Bernadette Mauritson har udgivet bogen *Kernesund familie – sådan!*, der beskriver erfaringer med kostvaner, der har beriget hendes og andres familier. Blandt rådene er, at man bør undgå gluten og at drikke mælk – det har hun selv og andre erfaringer med frigør familien fra en række skavanker. Uddannede eksperter har dog råbt vagt i gevær og udnævnt kostrådene til at være direkte usunde. Om det er familiernes egne oplevelser på egen krop eller eksperternes forskning, der har 'mest ret', ved vi ikke.

Følger vi Surowieckis beskrivelse af the Wisdom of crowds kan autoriteten fratages eksperten, og enhver kan have en holdning til, hvad der er sandheden. Viden er ikke længere elitær, men allemandseje, og det er – på næsten demokratisk vis – flertallet, der bestemmer, hvad der er

---

<sup>50</sup> Spivack 2007c

sandheden. Og her har vi netop problematikken: Hvis et samfund anerkender, at sandheden er, hvad flertallet i samfundet synes, vil fakta ikke længere have den tilstræbte objektivitet, som man målsætter i den akademiske verden. Viden vil blive en social konstruktion, hvor fakta bestemmes mellem mennesker, og ikke mellem mennesker og verden. Holdninger og fakta kan blive blandet sammen, og medierne får pludselig en uoverskuelig og meget konkret magt. Selvom hele verden blev overbevist om, at Holocaust aldrig fandt sted, ændrer det ikke ved det faktum, at det fandt sted.

Det er klart, at viden til enhver tid kan diskuteres – og selv uddannede fagfolks udtalelser bør ikke anskues for at være ubetinget sandhed. Ikke desto mindre er der gennem den systematiske uddannelse af fagfolk en form for kvalificering af den mindste fællesnævner. På WWW kan enhver udtale sig om ethvert emne, uden nødvendigvis at være kvalificeret. En håndværker kan i princippet give 'gode råd' om, hvilken medicin man bør tage, når man har ondt i ryggen – hvilket ikke nødvendigvis er de samme råd som en uddannet læge vil give. Omvendt betyder den enorme mængde af brugere på WWW, at dårlig information, f.eks. råd der gør mere skade end gavn, ofte vil blive identificeret og kommenteret. Men der er ingen garanti for, at et 'dårligt råd' vil blive identificeret – det kan med andre ord være svært at identificere troværdigheden på den viden, der er tilgængelig på WWW.

I sidste ende er det en diskussion om, hvorvidt man anskuer sandheden eller livsanskuelsen for at være noget objektivt identificerbart, eller om det er en social konstruktion. Hvorvidt fakta er menneskets forsøg på at aflæse og beskrive en faktisk eksisterende verden gennem deres sanseindtryk (fænomenologisk), eller om fakta er et forhandlingsresultat mellem menneskers italesættelse af sociale konstruktioner (diskursiv socialkonstruktivisme), er en diskussion, jeg ikke vil dykke nærmere ned i i dette projekt.

### *3.5.2 AMATØRENS PLATFORM*

Begrebet amatør kan have flere betydninger. Amatør bruges ofte om en ikke-professionel, i den forstand at amatøreren ikke har tilstrækkelig kompetence til at kunne tjene penge på sin interesse. Omvendt kan amatøreren også tegnes som det interesserede individ, der udfører et stykke arbejde ud fra en indre drift og ikke pga. en lønseddel. Amatørbegrebet dækker altså på en og samme tid over den ikke nødvendigvis kvalificerede og den nødvendigvis engagerede. Igen kan Wikipedia bruges som et eksempel på, hvad der kan lade sig gøre med engagerede amatører. Det er bemærkelsesværdigt, at Wikipedia, som har flere artikler end en encyclopædi i klassisk forstand, ikke har nogen kommerciel konkurrent i samme størrelsesorden.<sup>51</sup> Er det et symptom på en øvre grænse for hvad der kan organiseres i en virksomhed, eller er der bare ikke nogen der har prøvet?

Gennem Web 2.0 er det blevet muligt for enhver entusiast at ytre sig på internettet. Enhver, der har en holdning, har en stemme. Dette har resulteret i fantastiske produkter, såsom Wikipedia, men også i mere eller mindre ligegyldige blogs om morgenkaffe, bleskift og lokumspoesi. De store mængder af publiceret stof på World Wide Web er altså på en og samme tid en epoke i frigørelsen af information, men muligvis også en markant sænkning af den gennemsnitlige kvalitet for publiceret indhold. I den ikke-virtuelle verden skal et produkt gennem mange led, før

---

<sup>51</sup> Den engelsksprogede version af Wikipedia har i dag mere end 20 gange så mange (og generelt længere) artikler end Britannica.

det kan publiceres. En bog, der ikke er udgivet, skal gennem forlagets nåleøje og muligvis gennem mange gennemlæsninger og gennemskrivninger, før den bliver accepteret og udgivet. Dette står i kontrast til publiceringen af data på WWW, der igennem Web 2.0 kan foretages af enhver med den mest basale it-kundskab. På internettet bliver et blogindlæg muligvis ikke engang læst igennem en enkelt gang, før det publiceres. Udvælgelse og kvalificering er ikke tilstedeværende på WWW.

Og dog: For selvom alt i princippet er tilgængeligt for enhver på WWW, er nogle ting alligevel mere tilgængelige end andre. Et af de mest benyttede værktøjer til at finde indhold på WWW er at benytte sig af en søgemaskine, som f.eks. Google. Google opprioriterer visse hjemmesider, ikke baseret på deres indhold, men baseret på hvilke og hvor mange hjemmesider der i øvrigt henviser til den pågældende hjemmeside. En side, der slet ikke linkes til, vil således ikke få en særlig god placering i Googles søgninger og vil således ikke være særligt tilgængelig. På den måde foregår der en konstant filtrering på WWW, hvor visse hjemmesider tilskrives en særlig kvalitet, baseret på, hvordan der linkes mellem dem. Som nævnt tidligere i dette kapitel er der i Web 2.0-løsninger en tendens til, at kvaliteten øges sammen med antallet af brugere.

Bagsiden af medaljen er, at på de områder af WWW, hvor der ikke er mange brugere, vil kvaliteten derfor ikke altid være høj. Internettet har medført en opblomstring inden for mange nicheområder. Folk, der før ikke har kunnet finde hinanden eller har haft svært ved at dele deres interesse pga. geografiske afstande, har fået nye muligheder for at mødes og diskutere deres lidenskaber og interesser. I mange tilfælde betyder denne niche-tendens dog, at der frem for store grupper med mange interesserenter er mange små grupper med få interessenter. Dette giver ikke store muligheder for den form for filtrering, der ligger til grund for Wisdom of crowds.

### *3.5.3 NETDEMOKRATI OG REPRÆSENTABILITET*

Anskuer vi internettet som en slags virtuelt demokrati, støder vi ind i et problem, som ikke er diskuteret hos hverken Spivack eller Surowiecki: Er internettets brugere repræsentanter for virkelige menneskegrupper? I den vestlige verden har de fleste adgang til WWW, og det virker derfor oplagt at anskue internettet som 'informationens frigørere'. Men internettets manglende tilgængelighed glemmes. Først og fremmest findes der mange lande, som ikke har mulighed for at tage del i den kollektive intelligens, som der opstår på nettet. Internettet er ikke i samme grad tilgængelig for borgere i mange fattigere lande – og nogle lande, her tænker jeg især på Kina, er udsat for censur i en grad, der får begrebet 'Den globale hjerne' til at lyde som en vittighed. Selv i de lande, hvor de fleste borgere har adgang til internettet fra deres eget hjem, f.eks. Danmark, er der stadig store grupper der ikke har mulighed for at deltage på lige fod med de aktive internetbrugere; der findes eksempelvis en stor gruppe af ældre mennesker, der ikke har lært at bruge en computer. Og ser vi igen på den gruppe af mennesker, der faktisk bruger internettet, vil der kun være en begrænset del af denne gruppe, der selv bidrager med indhold.

Jacob Nielsen skriver i et af sine indlæg på sin hjemmeside, useit.com, at der er et 90-9-1-forhold mellem internetbruger. 90 procent af alle brugerne er dem, der kun kigger, og som aldrig deltager aktivt. 9 procent deltager fra tid til anden, men er hovedsageligt tilskuer. Kun den sidste procent er jævnlige aktive internetbrugere.<sup>52</sup>

---

<sup>52</sup> Nielsen, 2006

*1% of users participate a lot and **account for most contributions**: it can seem as if they don't have lives because they often post just minutes after whatever event they're commenting on occurs.*

JAKOB NIELSEN 2006

Spørgsmålet er, hvor meget vi ved om denne ene procent. Er denne ene procent repræsentativ for de øvrige internetbrugere – eller for den øvrige befolkning? Eller er vi blot ved at udnævne en ny elite – nemlig de aktive internetamatører?

#### 3.5.4 TEKNOLOGISKE SPRING

Jeg har i dette kapitel fremstillet både drømmebilleder og skrækscenarier for internettets kommende udvikling og dets betydning for brugerne. Jeg har forsøgt at præsentere visioner og kritik på deres egne præmisser og diskuteret dem løbende. På den ene side tegner der sig en lys fremtid, hvor information slippes fri i en grad, der er menneskeheden hidtil ukendt – og på den anden side venter der os en mørk, fordummende fremtid, hvor det gennemsnitlige bliver idealet, og hvor fakta som vi kender det i dag bliver en saga blot. Hvor en bog tidligere skulle gennem redaktører, faktatjek og korrekturlæsning, er det i dag blevet nemt at publicere tekst. Fra at indebære en overvejelse over udgifter, indtjening og opsættelse af succeskriterier er der i dag ingen begrænsning for publiceringen.

*The future presented by the internet is the mass amateurization of publishing and a switch from "Why publish this?" to "Why not?"*

SHIRKY 2008, s. 60

Clay Shirky sammenligner den periode, vi står i nu, med opfindelsen af trykken. Dengang blev bøger skrevet af munke, som med stor nøjsomlighed reproducerede den ene side af teksterne efter den anden. Da trykken blev opfundet, fik de en konkurrent – men de havde ikke forventet en konkurrent, der ville fratage dem deres erhverv.<sup>53</sup> Så på bekostning af de lærde munkes erhverv og den nytte, det var for dem at transskribere og den gamle tradition, overtog trykken munkenes arbejde.<sup>54</sup> Masseproduktion af tekst blev billigere, og større dele af verden lærte at læse og skrive, dele viden og producere nye tekster, der kunne masseproduceres i trykken. Shirky illustrerer med ovenstående eksempel, hvordan et teknologisk fremskridt sjældent sker uden omkostninger – men at gevinsten kan overskygge omkostningerne i en sådan grad, at man kan leve med det.

Selvom Shirkys eksempel virker ganske uskadeligt (hvilken rolle spiller en gruppe munkes erhverv i forhold til frigørelsen af viden?), er der flere eksempler fra verdenshistorien, hvor store teknologiske fremskridt er sket på bekostning af noget dyrebart. Eksempelvis medførte etableringen af atomkraftværker store mængder af forholdsvis ren energi til en lav pris – men indebar atombomberne der ramte Japan og ulykken i Tjernobyl.

---

<sup>53</sup> Shirky 2008, s. 66ff

<sup>54</sup> Udover udkonkurreringen af munkenes transskriberinger, er det ikke utænkeligt, at nogle har fundet den trykte bog mere ensformig og kvalitetsfattig end de håndskrevne (eller malede) værker, som munkene udførte med kunstnerisk snilde.

Selvom Web 2.0 næppe kan sammenlignes med kernesplætning, er der grund til at overveje konsekvenserne af de situationer, der kan fremkomme af sociale medier.

### 3.6 OPSAMLING

Jeg har i dette kapitel fremstillet forskellige syn på, hvad Web 2.0 er, samt diskuteret en række af de bevægelser, der kan siges at kendetegne Web 2.0.

Web 2.0 kan ikke defineres entydigt men er en betegnelse, der dækker over en lang række tendenser. Blandt de mest essentielle er den øgede mulighed for, at brugere kan bidrage med indhold uden kendskab til HTML eller andre tekniske platforme. Med den aktive brugerinddragelse følger social software, dvs. software der giver mulighed for social interaktion mellem brugere. Dette realiseres i f.eks. blogs, wikier, kommentarmuligheder og tagging<sup>55</sup>.

World Wide Web kan ansues som en platform, der i højere grad end trykken kan frigøre information, og det er blevet muligt at publicere næsten gratis og uden de klassiske filtreringer, der er et resultat af omkostninger og styring af mere klassiske publiceringsformer. Med Web 2.0 er brugeren sat i centrum, og enhver, der har internetadgang, har mulighed for at bidrage til de konstant voksende mængder af tilgængeligt indhold.

De mange aktive internetbrugere og den endeløse produktion af indhold finder nye måder at organisere sig på. Surowieckis begreb The Wisdom of the Crowds dækker over massernes evne til at skabe utrolige resultater, og WWW har på flere punkter vist potentiale for at kunne aggregere individers bidrag. Kritikere af Web 2.0-tendenserne frygter dog en amatørisering af medieindhold, som måske nok skaber mere indhold, men som sænker den gennemsnitlige kvalitet af indholdet som resultat af manglen på kulturel og systematisk filtrering, som vi kender det fra det fri marked.

Internettet er fyldt med indhold, meninger og idéer. Men mængderne er overvældene, og det er svært at skabe sig et overblik over indholdet. Hvis man ønsker at trække data ud af de enorme mængder af indhold, skal der ny tænkning på bordet – dels for at trække data ud, og dels for at evaluere brugbarheden af sådanne data. Det følgende kapitel handler om de teknologier, der findes og er ved at blive udviklet til at behandle data på nettet som mening. Grundidéen er at lære maskiner at forstå, hvad mennesker skriver, således at de kan udføre nogle af de opgaver, som vi mennesker ikke har lyst til at bruge tid på. Denne form for meningsdannende systemer kaldes med en fællesbetegnelse *semantic software* – og en samling af nogle af disse teknologier, kaldet *Semantic web*, er allerede ved at blive implementeret.

---

<sup>55</sup> Tag er engelsk, og betyder egentlig mærkat. Tag bruges både om koder, der omkredser et indhold, f.eks. **<b>dette står med fed</b>**, men også i forbindelse med tagging. I forbindelse med tagging er tags mærkater der beskriver et eller andet form for indhold.

## 4 SEMANTIC SOFTWARE PÅ NETTET

Jeg har i forrige kapitel beskrevet Web 2.0 som et paradigme og diskuteret Wisdom of crowds i forhold til www. Web 2.0-paradigmet kan anskues som en reaktion på en backendorienteret Web-tradition, hvor det primært var dem med teknisk know-how, der havde mulighed for at bidrage med indhold. I dag har alle med en internetadgang mulighed for at bidrage med indhold, hvilket betyder at de eksisterende teknologier til at administrere data på www bliver presset til deres yderste.

Jeg vil i dette kapitel præsentere forskellige eksisterende former for semantic software. Semantic software er et felt, der bliver forsket meget i, da det vil åbne op for helt nye måder at bruge internettet på. En af semantic softwareteknologierne kaldes for Semantic web. Dette kapitel handler fortrinsvis om Semantic web, men for at forstå dets svagheder og styrker vil jeg indledningsvis kort beskrive andre former for semantic software.

### 4.1 USTRUKTURERET, SEMISTRUKTURERET, STRUKTURERET

Internettet er fyldt med data. Man kan lidt forenklet sige at internettet udelukkende består af data og forbindelserne mellem dem. Men data er ikke bare data. Data med forskellige formål præsenteres på forskellig vis. En web-butik, der sælger cd'er, gemmer data om hver enkelt cd i en database, så man kan søge på titel eller kunstner, mens en blogger skriver en anmeldelse af en cd i formateret tekst. Begge disse typer af data har deres fordele. I databasen kan man nemt sammenligne forskellige cd'er, gruppere dem, søge på specifikke kriterier osv. Bloggeren behøver dog ikke at vide noget om relationelle databaser eller databaseforespørgsler for at skrive om sin nyindkøbte cd; læseren behøver heller ikke at slå op i databasen, men læser blot om cd'en i den kontekst, som bloggeren har defineret. I databasen er hver cd defineret ved at have nogle karakteristika, såsom titel, kunstner, udgiver, udgivelsesår, antal numre og spilletid, hvilket gør det muligt at sammenligne og genbruge informationer, der allerede ligger i databasen (f.eks. at den samme kunstner har lavet fire forskellige albums). På bloggen kan forfatteren nøjes med at skrive 'Radioheads nye album', hvilket muligvis vil være tilstrækkeligt, for at hendes læsere forstår hvilket album, der er tale om. Herudover giver databasen ikke mulighed for at beskrive, hvordan bandets udvikling mod elektronika-genren fungerer.

Strukturerede data				Ustrukturerede data
<b>id</b>	<b>navn</b>	<b>årstal</b>	<b>måned</b>	<p>Det omtalte projekt, er et kunstprojekt, hvor en pakke med et kraftig GPS-apparat er blevet sendt verden rundt med fragtselskabet DHL. Pakken er ikke bare blevet fragtet i en lige linje fra a til b, men har fulgt nogle meget præcise rutebeskrivelser for at tilbagelægge en meget kringlet sti. Stien kunne kun opfanges af GPS'en, som har tracket hele stien.</p>
1	Brian	1976	08	
2	Mette	1977	11	
3	Tine	1979	03	
4	Jørgen	1977	07	
5	Polle	1981	03	
6	Mads	1982	10	

FIGUR 6 – STRUKTUREREDE OG USTRUKTUREREDE DATA

Der findes ud fra en datalogisk synsvinkel to typer data, der svarer til ovenstående beskrivelse. Der er tale om strukturerede og ustrukturerede data. De to begreber beskriver ikke kvaliteten af data, men den måde, de er ordnet på. Hvor strukturerede data er defineret af metadata<sup>56</sup> og sat i system, er ustrukturerede data teknisk set ikke-ordnede. Ustrukturerede data kan i princippet være en veldisponeret, velskrevet tekst i 'menneskesprog', som f.eks. en akademisk tekst eller en roman. Teksten er altså ikke ustruktureret i den forstand, at det virker ustruktureret, når man læser det, men i den forstand at en maskine, der ikke umiddelbart kan forstå menneskesprog, har svært ved at adskille informationerne fra hinanden i teksten.

*Machine intelligence is a problem for text mining. Natural language has developed to help humans communicate with one another and record information.*

WEIGUO ET AL. 2006, s. 78

Blandt andre typer af ustrukturerede data finder vi eksempelvis stikord, billeder, video og musik. Selvom man på den ene side kan tale om, at data enten er strukturerede eller ej, kan man alligevel finde en vis form for struktur i meget ustrukturerede data. Meget ustrukturerede data besidder nemlig en slags implicit metadata. I f.eks. et leksikon er hvert emne defineret af et opslagsord og en beskrivelse, og hvert af opslagsordene er sat i alfabetisk rækkefølge. Denne form for strukturering af ellers ustrukturerede data kaldes for semistrukturerede data, men anses typisk for at være ustrukturerede data i forhold til f.eks. data i en database. Og desværre er de to typer af data ikke umiddelbart mulige at forbinde. Man kan med andre ord sige, at der

<sup>56</sup> Metadata er data der beskriver data. Metadata kan i princippet være det resume der står bag på en bog, men det kan også være de informationer der angives i en kildehenvisning eller ISBN-nummeret. Data kan dog også være af mere teknisk karakter, f.eks. beskrivelse af datatyper (en tekst, en dato, en webadresse, et kreditkortnummer, etc.).

findes to hovedgrupper: struktureret og ustrukturerede data, og den ustrukturerede data *kan* være semistruktureret. Det afgørende punkt for, hvorvidt data kan kaldes struktureret, er, at de relevante data er umiddelbart tilgængelige for maskinel aflæsning.<sup>57</sup>

Det er ikke umiddelbart muligt at bruge de mange fordele fra strukturerede data på semi- eller ustrukturerede data. Problemet ligger i, at til trods for, at mennesker ikke synes, det er svært at danne mening fra læst tekst, ligger maskin-intelligensen så langt fra den menneskelige måde at tænke på, at det er en umådeligt svær opgave for en maskine at afkode det ustrukturerede indhold.

Som ordet lægger op til, vil der dog være visse typer af data man kan udtrække fra semistrukturerede data, takket været metadata. En typisk HTML-side har f.eks. både en titel, en overskrift og i visse tilfælde også nøgleord, information om forfatter, og angivelser af hvor hvert tekst-afsnit starter og slutter. Et eksempel på systemer, der forsøger at aflæse HTML er søgemaskiner, der automatisk gennemlæser sider for f.eks. titel og overskrifter, mens den forsøger at finde nøgleord fra teksten (det kan f.eks. være antallet af forekomster af ord i teksten). Ud fra disse forsøger den at bestemme, hvilke nøgleord der bedst beskriver teksten. Der findes mange problemer med maskinaflæsning, der kan gøre det svært at lave en høj kvalitet i aflæsningen.

*Humans are able to distinguish and apply linguistic patterns to text, overcoming obstacles (such as slang, spelling variations, and contextual meaning). Computers do not handle them easily.*

WEIGUO ET AL. 2006, s. 78

Man kan sige, at den ideelle maskinelle aflæsning ville være i stand til at forstå teksten med samme præcision som et menneske, men med samme hastighed og overblik som en computer.  
58

## 4.2 TEXT MINING

Inden for maskinel aflæsning af data er *text mining* et af de mest diskuterede emner. Text mining går ud på at kunne aflæse data fra rå tekst (eller naturligt, skriftligt sprog) og udlede data, der kan bruges struktureret. Man kan sige, at da text mining i høj grad handler om fortolkning af data, står det i modsætning til data mining, der henter konkrete data i deres eksisterende form (f.eks. at hente valutakurser fra [www.valuta.dk](http://www.valuta.dk)). Text mining bruges i en vis grad af søgemaskiner men kan også bruges, hvis store mængder af rå tekst skal transformeres til strukturerede data. Ønsker man f.eks. at oversætte teksten fra en kogebog til en database med opskrifter, hvor man kan søge specifikt på ingredienser eller fedt-indhold, kræves det, at hver opskrift omdannes fra en mængde tekst til en struktureret samling af ingredienser, mængdemål og beskrivelser.

Weiguo et al. opstiller en række eksempler på, hvilke teknologier der typisk indgår i text mining. Der gøres dog rede for, at dette ikke nødvendigvis er de eneste eller mest korrekte teknologier at bruge inden for dette felt, der er i konstant udvikling.

---

<sup>57</sup> Allen et. al 2004

<sup>58</sup> Weiguo et al. 2006, s. 78

#### 4.2.1 UDTRÆK AF INFORMATION

Et system, der skal aflæse tekst, skal kunne udtrække informationer. Weiguo et al. bruger følgende tekst-eksempel:

---

*“Area relatives of a man being held hostage in Iraq waited for word about him Saturday as militants threatened to decapitate him, another American, and a Brit unless demands were met within 48 hours.”*

---

Ud af sådan et teksteksempel, skal en maskinel aflæsning f.eks. kunne identificere en række mennesker, de to amerikanere, en brite og amerikanerens familie; et land, Irak; og et tidspunkt, lørdag.<sup>59</sup>

#### 4.2.2 EMNE-TRACKING

Emne tracking giver mulighed for at identificere (og abonnere på) artikler om et særligt emne. Særligt relevant for disse teknologier er, at mange emner kan lyde, som om de handler om det samme - men i virkeligheden ligge meget langt fra hinanden. Weiguo et al. bruger eksemplet med, hvordan det kan være svært for et system at adskille to emner, der har ord til fælles, f.eks. emnerne *text mining* og *mining for minerals*.

Emne-tracking kan bruges til at følge et givent emne, således at man altid er opdateret på den seneste viden. Weiguo et al. nævner et scenarie, hvor en læge kan følge den seneste forskning inden for behandlingen af netop de sygdomme hans patienter har.<sup>60</sup>

#### 4.2.3 REFERERING

Med muligheden for at udtrække informationer fra teksten giver det god mening at udforske mulighederne for at kunne skabe automatiske resumeer af tekster. Dette kan enten foregå ved at konstruere tekst maskinelt, eller ved at genbruge visse sætninger fra original-teksten. Det er også oplagt at sammenkoble sådanne resumeer med emne-tracking.<sup>61</sup>

#### 4.2.4 KATEGORISERING

En af de mest enkle måder at behandle ustrukturerede data på er kategorisering. Kategorisering sker inden for prædefinerede kategorier, hvor hvert dokument vurderes og placeres i en kategori. Dette kan f.eks. ske ved ordoptælling, identifikation af nøgleord og opslag af synonymmer. Kategorisering giver mulighed for at knytte tekster, der ligner hinanden, sammen. Et typisk problem indenfor kategorisering er dog, at der ikke tages højde for semantik, og at f.eks. *surfe* både kan være noget man gør på havet og på WWW.<sup>62</sup>

---

<sup>59</sup> Weiguo et al. 2006, s. 78

<sup>60</sup> Weiguo et al. 2006, s. 78

<sup>61</sup> Weiguo et al. 2006, s. 79

<sup>62</sup> Weiguo et al. 2006, s. 79

#### 4.2.5 DOCUMENT CLUSTERING

I modsætning til kategoriseringer foregår document clustering uden prædefinerede kategorier. I stedet sammenlignes hvert enkelt dokument med alle andre dokumenter og grupperes med de dokumenter, det ligner.

*[. . .] it differs from categorization in that documents are clustered on the fly instead of through predefined topics.*

WEIGUO ET AL. 2006, s. 78

Som med kategoriseringer er der også her en fare for, at emner, der ikke er semantisk forbundne, ligestilles, fordi nøgleord fra to forskellige dokumenter kan være ens, men handle om noget vidt forskelligt.

#### 4.2.6 CONCEPT LINKAGE

Concept linkage minder om Document clustering, i den forstand at den forbinder artikler med hinanden baseret på indhold. Men hvor Document clustering knytter siderne sammen på et meget overordnet plan, f.eks. baseret på de ord, der går igen flest gange, baserer Concept linkage sig på at forbinde enkelte detaljer. Det kan f.eks. være nyttigt inden for den medicinske verden, hvor Concept linkage evner at forbinde f.eks. sygdomme baseret på forskellige symptomer – hvilket kan være svært at overskue for mennesker. Concept linkage giver med andre ord mulighed for at skabe sammenhænge, som mennesker måske ville overse eller ikke ville have mulighed for at finde.<sup>63</sup>

#### 4.2.7 INFORMATION VISUALIZATION

Gennem visualisering skabes der mulighed for at præsentere data med sammenhænge mellem emnerne.

*Visual text mining, or information visualization, puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching.*

WEIGUO ET AL. 2006, s. 80

Weiguo et al. beskriver dette felt ganske overfladisk, men jeg vil vende tilbage til emnet i kapitel 7.

#### 4.2.8 QUESTION ANSWERING

Question answering er en form forespørgselsteknik, der giver mulighed for at nuancere søgninger. Den bagvedliggende tanke er, at søgninger præciseres gennem sproglige formuleringer. Eksempelvis røber spørgsmålet "Hvem er Crusty?", at Crusty er en person.<sup>64</sup>

#### 4.2.9 TEXT MINING OG UDVIKLINGEN

Der sker meget inden for feltet text mining. Der bliver arbejdet på bedre fortolkning af indhold, og selvom man udviklingsmæssigt stadig er langt fra at have maskiner, der kan kommunikere gnidningsfrit med mennesker i naturligt sprog, betyder teknologier som dem, jeg her har ridset

---

<sup>63</sup> Weiguo et al. 2006, s. 79f

<sup>64</sup> Weiguo et al. 2006, s. 81

op, at vi hele tiden kommer tættere på noget, der er brugbart. For selvom maskiner stadig ikke kan læse en tekst og danne den samme fortolkning, som et menneske ville kunne, er spørgsmålet, hvor præcis en forståelse computeren har brug for at have, før at vi mennesker kan udnytte maskinens evner til at behandle enorme mængder af data på meget kort tid. Eksempelvis har Google, som øjensynligt ikke bruger nogen form for semantisk analyse af indholdet, vist sig at være et glimrende værktøj til at lokalisere sider på internettet.

### 4.3 SEMANTIC WEB

Tim Berners-Lee er anerkendt for at være skaberen af World Wide Web.<sup>65</sup> Berners-Lee har siden været aktiv med at skabe visioner for World Wide Web og var i 2001 medforfatter til artiklen *The Semantic Web*<sup>66</sup>. Artiklen beskriver et scenarie, hvor World Wide Web har fået en ny anvendelse, hvor såkaldte agenter (små, personlige, intelligente stykker software) hjælper brugeren med at løse hverdagsudfordringer. Grundidéen er at opmarkere indhold på internettet, således at det bliver nemmere at lave maskinel aflæsning.

*The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.*

BERNERS-LEE ET AL. 2001, S. 3

Det er vigtigt at forstå, at Semantic web ikke er en afløser af det nuværende World Wide Web, men at det er en udvikling af dette. Der er altså ikke tale om, at det eksisterende materiale ikke vil være tilgængeligt længere – men hvis teknologien tager til, vil der være flere og flere services, f.eks. søgemaskiner, der benytter sig af Semantic web-teknologierne, hvorfor det vil være attraktivt for enhver at benytte sig af Semantic web-teknologi.

Med Semantic web vil internettet blive mere struktureret. Det vil f.eks. sige, at det bliver muligt at søge mere præcist og mere omfattende. Visionen er, at indhold og webservices på nettet i langt højere grad skal kunne tale sammen. Det kan f.eks. være, at man ønsker at købe en Volvo V50. I dag vil man først skulle finde nogle forhandlere eller måske, hvis man er heldig, en service, der samler priserne fra flere forskellige forhandlere på samme side, hvorfra man så skulle finde den pågældende bil. Med Semantic web ville du i stedet bruge en agent, et stykke software, der gør arbejdet for dig. Du indtaster blot, at du vil *købe* en *Volvo V50*. Agenten vil så kunne finde forhandlere, priser, afstand til forhandlere. Med ganske få klik vil du også få en rejsebeskrivelse til den forhandler, du vælger, kunne slå telefonnummer op eller finde andre brugeres holdninger til netop denne forhandler. Alt sammen fordi agenten, maskinen, er i stand til at afkode, *hvad* det er der skrives om, og *hvor* der skrives om det.

Men hvordan kan maskinen vide, hvad forskellen er på at sælge eller købe en bil? Eller forskellen på butikkens og værkstedets telefonnummer? Eller om prisen er angivet i danske kroner eller euro? Og med eller uden moms? Med Semantic web forsøger man at håndtere dette problem med opmarkering af indhold.

---

<sup>65</sup> Abbate 2000, s. 214

<sup>66</sup> Berners-Lee et al. 2001

### 4.3.1 OPMARKERING AF INDHOLD, RDF OG OWL

Semantic web handler i bund grund om at opmarkere indhold, så det bliver lettere for maskiner at forstå indhold. Der går formentlig mange år, før maskiner vil kunne læse naturligt sprog lige så godt som mennesker. Semantic web forsøger at hjælpe maskinerne på vej, ved at forklare teksten løbende.

I et konkret (og forenklet) eksempel, hvor vi har en hjemmeside med en opskrift på brownies, kan man benytte opmarkering af indhold på forskellig vis. Der kan f.eks. være tale om layout-mæssig opmarkering, som vi kender det fra HTML<sup>67</sup>. Hvis man ønsker at angive, at noget skal være titlen på en side (den som vises i browservinduet), angives dette i HTML som f.eks. `<title>Opskrift på brownies</title>`. En overskrift (her 'overskrift 1', det øverste niveau i overskrifterne) har ligeledes et *tag*, der ser sådan ud: `<h1>Lækker chokoladecake på 20 minutter</h1>`.

I forbindelse med Semantic web, taler man derimod ikke om layout-mæssige opmarkeringer, men om semantisk opmarkering. På den måde kan chokolade opmarkeres som `<ingrediens>chokolade</ingrediens>`<sup>68</sup>, hvilket semantisk adskiller sig fra `<slik>chokolade</slik>`. Problemet med denne form for opmarkering er dog, som Nova Spivack gør opmærksom på<sup>69</sup>, at den software der skal aflæse denne form for opmarkering ligeledes skal vide, at *ingrediens* og *slik* har to forskellige semantiske betydninger.

Når man taler om Semantic web, taler man typisk om de to teknologier RDF og OWL<sup>70</sup>. Man kan sige, at RDF er den grundlæggende teknologi med meget basale muligheder for semantisk opmarkering af tekst og en begrænset ontologi (f.eks. markering af adresser eller bredde- og længdegrader), hvor OWL er en videreudvikling af RDF og har et langt større ordforråd. Det er vigtigt at skelne mellem Semantic web og semantic software. Semantic software er software, der bruger semantisk analyse af indhold, mens Semantic web dækker over en række specifikke web-teknologier, bl.a. RDF/OWL, som forventes at tegne fremtiden for det semantiske World Wide Web.

Nova Spivack eksemplificerer OWLs funktion, ved at tale om sportsvognen Lotus. Problemet med Lotus er, at det er et ord, der bruges om mange forskellige ting. F.eks. er Lotus et softwareprodukt, navnet på en softwareproducent, en blomst og en sportsvogn. Her kommer ontologien, OWL, ind i billedet og kan gøre det klart for softwaren, hvad vi mener, når vi skriver `<sportscar>Lotus</sportscar>`.

*An ontology defines facts such as "a sportscar is a kind of car," "a car is a ground vehicle," "a car is a product," "a car is a device," "a sportscar is a recreational or competitive vehicle," etc.*

SPIVACK 2007B, s. 3

---

<sup>67</sup> HTML står for Hyper Text Modeling Language.

<sup>68</sup> Jeg vil gerne understrege, at RDF og OWL-opmarkeringerne ikke er lige så enkle tags som dem jeg præsenterer i denne rapport, men princippet er det samme.

<sup>69</sup> Spivack 2007b, s. 3

<sup>70</sup> RDF og OWL er W3C-standarder og står for Ressource Definition Framework og Web Ontology Language.

Et af de spørgsmål, der naturligvis melder sig i forbindelse med Semantic web, er: hvem skal opmarkere teksten? Som jeg beskrev i foregående kapitel, er der med Web 2.0 sket en bevægelse væk fra det tekniske, og det at publicere er ikke længere begrænset til eksperterne. Men hvis al tekst fremover skal opmarkeres, således at hvert betydningsfuldt ord i teksten skal opmarkeres efter RDF- eller OWL-standarden, vil det frarøve amatørernes mulighed for at bidrage til Semantic web. Til trods for at World Wide Web, som vi kender det i dag, ikke vil forsvinde, kan man dog forestille sig, at mange ydelser og indhold vil blive populære gennem Semantic web teknologier og dermed tage opmærksomhed fra det indhold, der ikke er opmærkeret.

#### 4.3.2 MASKINERNE FORSTÅR INGENTING

Det er vigtigt at forstå, at det med opmarkering af indhold ikke gør det muligt for en maskine at forstå, hvad en tekst handler om. Uanset hvor mange tags, vi påhæfter hvert ord, betyder det ikke, at et stykke semantisk software med et bliver i stand til at forstå farven rød eller vide, hvordan en Volvo og en Boeing adskiller sig fra hinanden. Men med opmarkeringer kan man standardisere indhold og på den måde nemmere lave programmer, der leder efter forskellige forståelser i indholdet. Hvis vi alle kan blive enige om, at tagget `<price-dkr-inkl>pris</price-dkr-inkl>` betyder, at prisen er angivet i danske kroner inkl. moms, bliver det meget nemmere at lave software, der kan sammenligne priser fra forskellige butikker.

Det Semantiske web har altså ikke til formål at gøre maskiner bedre til at forstå naturligt sprog, men at benytte enkle metoder til at gøre det muligt at standardisere udtryk, så de kan tolkes uden for sammenhæng og behandles som strukturerede data.

#### 4.4 DATA BEHØVER IKKE AT VÆRE INDHOLD

Jeg har kort beskrevet forskellige former for Semantic software, og de metoder der anvendes til at analysere data. Men data behøver ikke nødvendigvis at være brugergeneret indhold. I forrige kapitel var jeg kort inde på begrebet Collaborative Filtering (CF) som er et godt eksempel på data, der ikke relaterer sig til indholdets semantik, men til brugen af indholdet.

Med CF konstruerer man en række profiler, f.eks. baseret på brugeres færden på et websted. Disse profiler kan sammenlignes og bruges til at anbefale eller vejlede brugerne baseret på deres profil. Her er et konstrueret eksempel, hvor en række brugere har besøgt forskellige sider på det samme websted:

	Side 1	Side 2	Side 3	Side 4	Side 5
Bruger A	x	x	x		x
Bruger B	x	x	x	x	
Bruger C	x		x		x
Bruger D			x	x	

FIGUR 7 - EKSEMPEL PÅ COLLABORATIVE FILTERING

I skemaet ovenfor har brugerne A og B tre sider tilfælles. Dette kan være et tilfælde, at brugerne har besøgt de tre samme sider, men det kan også skyldes, at brugerne A og B deler interesser. Det er derfor ikke utænkeligt, at bruger A kunne have interesse i side 4, ligesom bruger B kunne

have interesse i side 5 (eller bruger C kunne have interesse i side 2). Nu er dette et meget forenklet eksempel med meget få brugere og meget få sider. Amazon, som har millioner af brugere og millioner af produkter, bruger netop denne metode med stor succes til at anbefale andre kunder produkter, som kunder med samme interesser har kigget på; "Customers who bought this item also bought". Det interessante er, at CF systemet er fuldstændig automatisk, og at brugeren aldrig behøver at indtaste informationer såsom interesser. Brugers adfærd på sitet afslører hans interesser og gør det muligt at målrette, i Amazon.coms tilfælde, markedsføring til brugeren. Musik-servicen Last.FM fungerer på lignende vis ved at sammenligne brugers musiksmag og afspille musik

Collaborative Filtering er et således et eksempel på, hvordan data på nettet ikke nødvendigvis knytter sig til det konkrete indhold.

Sådanne systemer har både fordele og ulemper. En af de største fordele er, at det er en meget enkel metode, og at der ikke er behov for kunstig intelligens eller for at opmarkere indhold. Systemet virker – på bedste Wisdom of crowds-maner - bedst, når der er mange brugere. Med mange brugere bliver systemet forholdsvis robust og kan bruges på alt fra informationsarkitektur til markedsføring og anbefalingssystemer.

En af ulemperne ved denne type af systemer er, at de altid vil være baseret på en algoritme, og denne algoritme kan være bedre eller dårligere til at introducere nye emner. F.eks. vil det ofte være problematisk, når man introducerer et nyt emne. Når Amazon.com f.eks. sælger en nyudgivet bog, vil den fra starten ikke blive anbefalet til andre på baggrund af deres profiler, da den endnu ikke optræder i nogen profiler. For at bogen fremtræder i en profil, skal en bruger først have fundet bogen på egen hånd. Og hvis anbefalingerne baseres på tusindvis af profiler, skal tusindvis af brugere derfor finde bogen egenhændigt, før den vil blive anbefalet andre. Det er derfor ret svært at indføre nye emner til et CF-system. Et fuldstændig nøgternt CF-anbefalingssystem vil altså have en tendens til at anbefale de emner, der generelt er mest populære, hvilket vil gøre dem endnu mere populære, og det bliver sværere for andre emner at konkurrere.

#### 4.5 SEMANTISK ANALYSE

Udover Semantic web-løsningen, hvor mennesker opmarkerer tekst for at kunne bruge indholdet som strukturerede data, findes der også andre former for semantic software. Semantic software er en fællesbetegnelse for alle typer af software, der i en eller anden grad behandler semantik, dvs. at maskinen kan fortolke, forstå eller behandle indhold baseret på dets indhold. I forhold til Semantic web er det da oplagt at stille spørgsmålet om, hvorvidt det overhovedet er semantic software, idet maskinerne reelt ikke har en forståelse af indholdet. Men det har de alligevel, fordi man igennem ontologier forsøger at definere en lang række ord, og fordi opmærksomheden og den maskinelle tolkning af ordene har en semantisk karakter.

Nova Spivack, som selv arbejder med semantic software og Semantic web, definerer begreberne således:

*It's pretty simple actually. Any app that can understand language to some degree could be labeled as "semantic." [ . . . ] But for an application to be considered part of the "Semantic*

*Web" it has to support a set of open standards defined by the W3C, including at the very least RDF, and potentially also OWL and SPARQL.*

SPIVACK 2008A, s. 1

I den forstand er selv tekstbehandlingsprogrammer med stavekontrol, søgemaskiner, som lokaliserer nøgleord på indholdssider eller retkomma.dk, der forsøger at sætte komma i danske sætninger ud fra grammatiske regler, alle eksempler på semantic software. Nova Spivack beskriver, hvordan de mest populære søgemaskiner i dag benytter sig af Natural Language Search i behandlingen af nøgleord til at finde resultater.

*Companies working specifically on natural language search are making use of semantics, but at the word-level only. They use networks of words that are linked to synonyms, antonyms, homonyms and other variations.*

SPIVACK 2007D, s. 2

Et af nøglebegreberne inden for semantisk analyse er Natural Language Processing (NLP). NLP er en fællesbetegnelse for en række metoder, der bruges til at analysere naturlige sprog (som f.eks. dansk og engelsk i modsætning til HTML eller relationel algebra). Analysen kan have mange forskellige formål, f.eks. at lokalisere sætninger og sætte punktummer i tekster, eller at lokalisere navne eller steder fra teksten.<sup>71</sup>

Med muligheden for at identificere f.eks. ordtyper og sætningskonstruktioner bliver det også nemmere at lave semantiske tolkninger på indholdet. F.eks. gives der mulighed for at slå synonymer op, og ved at lave databaser over hvilke typer af ord, der ofte sættes sammen, er der mulighed for at skabe selv-lærende systemer, der kan tolke på forskellige forståelser af ord, der kan have flere betydninger. I princippet åbnes der også for muligheden for meget bedre og mere præcise oversættelser.

Man kan sige, at natural language processing står for de dele af semantic software, der beskæftiger sig med rå tekst – i modsætning til f.eks. RDF/OWL, som kræver, at mennesker først har opmarkeret teksten.

Der har været stigende fokus på *information retrieval* og NLP de sidste ti år – formentlig takket være internettet og de overvældende mængder af digital tekst.<sup>72</sup> World Wide Web har medført nye behov for at kunne lokalisere ikke blot de sider, man leder efter, men også den viden, man leder efter på sider, man ikke vidste fandtes. Man er kommet meget længere i udviklingen af metoder til maskinel analyse af tekst og udvinding af data, end da man først begyndte at interessere sig for feltet i 1960'erne. Ikke desto mindre kan maskiner til dags dato ikke aflæse tekst problemfrit. Ikke mindst på grund af nogle af de barrierer, der er svære at programmere sig ud af. Eksempelvis kan det være svært at håndtere sprogs udvikling – at der bruges slang, og

---

<sup>71</sup> Tang et al. (2005) bruger forskellige lignende metoder, til at analysere og "rengøre" e-mails, således at de kan bruges til f.eks. Text mining.

<sup>72</sup> Sebastiani 2002 og Weiguo et al. 2006

at visse udtryk forvanskes eller decideret ændrer betydning, som f.eks. de problematiske pendulord<sup>73</sup>.

#### 4.5.1 TWINE

Nova Spivack er som tidligere nævnt direktør for firmaet Radarnetworks, som bl.a. laver Twine.com. Twine er ikke blevet offentliggjort endnu, men de har løftet sløret for nogle af de funktioner, der ligger til grund for systemet. Grundtanken er, at man kan lave samlinger bestående af hjemmesider, videoer, kontakter, wikipediaopslag mm. på baggrund af deres semantiske betydning. Twine benytter sig af brugernes samordninger af indhold til at analysere indholdet og lave Semantic web-opmarkeringer af indholdet.

*In this approach, the meaning of the information can be explicitly coded into the information just as HTML codes are added into content today -- and this can be done by people or software, and even by communities. Once this meaning -- or semantics -- is explicitly encoded into content, it can then be re-used by other applications to make sense of the content. It's worth noting that explicit semantics in content can also help natural language processing apps, as well as apps that don't understand natural language.*

SPIVACK 2007D, s. 2F

Som Spivack skriver, kan opmarkeringen altså skabes af brugere eller communities, men også af software. Ligeledes kan opmarkeringen hjælpe med at udføre NLP. Karikeret kan man tale om en hermeneutisk cirkel, hvor opmarkeringen af indhold kan ske på baggrund af NLP, men hvor NLP også kan drage nytte af eksisterende opmarkeringer. For at den hermeneutiske cirkel ikke skal blive til en ond cirkel, virker det oplagt at inddrage brugeres brug og kategorisering af indholdet til at påvirke den maskinelle læsning af teksterne. Hvorvidt Twines metode med at skabe en brugerdreven og communitybaseret tankegang til at skabe kategorier, som brugerne selv kommer indhold ind i til at opmarkere indholdet, bliver en succes, ved vi ikke endnu – men umiddelbart virker det som en god tilgang.

#### 4.6 MENNESKER, MASKINER OG BEGREBER

Succeskriteriet for en maskinel, semantisk aflæsning er i dag ikke, at maskinerne har en korrekt forståelse af indholdet (maskinen behøver ikke at vide, hvordan rød ser ud), men at den er i stand til at gruppere indhold efter kriterier, der giver mening for mennesker (maskinen skal kunne kategorisere rød som en farve). Man kan sige, at der er tale om en skelnen mellem en forståelse på et erkendelsesniveau og på et kommunikationsniveau. For at kunne udnytte internettet som en database er det måske i første omgang ikke nødvendigt, at den maskinelle aflæsning af tekst skaber en forståelse på et erkendelsesniveau, men blot på et kommunikationsniveau. En fortolkning på et kommunikationsniveau kræver dog en forståelse af begreber, der svarer til menneskers måde at forstå begreber på.

Med en forståelse for, hvordan mennesker forstår og bruger begreber, vil vi bedre kunne forstå, hvordan semantic software kan bruges. Jeg vil derfor i det følgende kapitel beskrive begrebspsykologiens forståelse af, hvordan mennesker bruger begreber til at beskrive kategorier. Gennem beskrivelsen af hvordan mennesker forstår og bruger begreber, er vi bedret

---

<sup>73</sup> F.eks. ordet *bjørnetjeneste*, der ofte ses i betydningen af at gøre nogen en stor tjeneste, hvor det oprindeligt betyder at forsøge at gøre nogen en tjeneste, der i sidste ende gør mere skade end gavn.

rustet til at kunne vurdere hvilke semantiske teknologier, der bedst egner sig til at analysere semantikken i indhold skabt af mennesker.

## 5 MENINGER OG BEGREBER

Hvilken forståelse af begreber og meningsdannelse ligger til grund for den semantiske software og ikke mindst Semantic web? For at kunne forstå, hvad mening, semantik og begreber er – dels i menneskeverdenen og dels i maskinverdenen – er det relevant at studere, hvordan mennesker benytter sig af begreber. Dette kapitel handler udelukkende om begrebpsykologi, dvs. menneskers brug af begreber, og først i kapitel 6 vil jeg diskutere, hvordan vi kan bruge maskiner til at lave semantiske analyser, der giver mening i menneskenes begrebsverden. Med den forståelse i baghånden kan man nemlig bedre udvikle systemer til at overskueliggøre/anskuelligøre indholdet af diskussioner på sociale fora på nettet.

Dette kapitel er hovedsageligt baseret på forfatteren Gregory L. Murphys bog *The Big Book of Concepts*<sup>74</sup>, som udmærker sig ved at være en opsummering af de sidste ca. 50 års teorier, forskning og empiriske undersøgelser inden for feltet *psychology of concepts*, hvilket jeg her kalder *begrebpsykologi*<sup>75</sup>. Murphys referering af feltet præsenterer de forskellige forskningsresultater og teoretiske udlægninger på loyal vis, og dels ved at han løbende perspektiverer resultaterne med andre forskeres kritik samt nyere forskningsresultater.

### 5.1 BEGREBSPSYKOLOGIEN

Når jeg i denne rapport skriver om mening, er det ikke som i "Hvad mener du om Poul? Kan du lide ham?", men nærmere som i "Hvilken mening ligger bag ordet lykke?". Jeg skriver om mening, som forfatteren Gregory L. Murphy skriver om *concepts*. På den måde er *begreber* og *mening* meget tæt beslægtede. Jeg bruger ordet begreb om de ord, vi bruger, f.eks. hund, og mening om den betydning, vi tillægger hund, f.eks. at det er et pattedyr, har egenskaber som en hund og ligner andre hunde (en noget solipsistisk tolkning, men som det vil vise sig, er det at definere emner sværere end som så). Murphy har sammenfattet de sidste hundrede års psykologiske forskning inden for feltet *begreber*, og refererer i 'The Big Book of Concepts' utallige teorier og psykologiske eksperimenter, der er udført for at få et mere nuanceret billede af menneskers evne til at forstå og kommunikere om verden omkring sig.<sup>76</sup> Den mening vi, tillægger et begreb, er typisk udtryk for en mængde, en klasse af ting. Når vi taler om *hunde*, taler vi ikke nødvendigvis om en specifik hund, men muligvis om selve klassen *hunde*. Man kan altså sige, at begreber dækker over klasser af ting, som har noget tilfælles. Murphy beskriver forskellen på et begreb (concept) og kategori (category) således:

*In general, I try to use the word concepts to talk about mental representations of classes of things, and categories to talk about the classes themselves.*

MURPHY 2004, s. 5

Opfattelsen af, hvordan mennesker forstår begreber, har udviklet sig enormt inden for de sidste halvtreds år. De gamle grækere diskuterede, hvordan begreber defineres, og Sokrates hvad der udgjorde forskellige begrebers *araté*, f.eks. hvad der gør et bord til et særlig godt bord. Først i

---

<sup>74</sup> Murphy 2004

<sup>75</sup> Jeg skelner i mit ordvalg af *begrebpsykologi* (for *psychology of concepts*) mellem de psykologiske studier og det mere overordnede og filosofiske felt *begrebsteori*. Sidstnævnte felt vil jeg ikke komme ind på i denne rapport.

<sup>76</sup> Murphy 2004

1970'erne bliver der gjort endeligt op med, at det er definitioner, der ligger til grund for menneskers forståelse af begreber. Herefter er der opstået flere forskellige teorier om, hvordan mennesker lærer og bruger begreber. Heriblandt er der to, prototype- og eksemplartilgangen, der står særligt stærkt, og en enkelt, videnstilgangen, der bygger videre på disse to. Ingen af teorierne kan forklare alle fænomener fyldestgørende, men tilsammen giver de et dækkende billede af menneskers brug af begreber til at danne mening. Murphy skriver i sit konkluderende kapitel:

*In comparing them, it must be understood that there is no full-scale theory coming from any of these approaches that encompasses all the topics that I have discussed. That is, there is no specific prototype theory, for example, instantiated in a computational model or just a set of assumptions, that provides an account of induction, infant concepts, typicality, the basic level, word meaning, and so on.*

MURPHY 2004, s. 483F

Selvom jeg i dette kapitel hverken vil komme ind på spædbørns begreber eller mange af de andre forskningsemner inden for begrebspsykologien, finder jeg det stadig relevant at opridse de forskellige hovedteorier inden for feltet.

Idet prototype- og eksemplartilgangene er reaktioner på det, som Murphy kalder 'Classical view' (hvad jeg kalder den klassiske tilgang), vil jeg indledningsvist opridse denne. Man kan sige, at hvor den klassiske tilgang er udtryk for et paradigme, er prototype- og eksemplartilgangene udtryk for et andet.

## 5.2 DEN KLASSISKE TILGANG

Murphy kalder prototype-, eksemplar-, og videnstilgangene for 'approaches', hvor den klassiske tilgang bliver navngivet 'classical view'. Denne skelnen skyldes formentlig, at de psykologiske studier om begreber først for alvor tager fart i sidste halvdel af det tyvende århundrede, selvom man siden oldtiden har filosoferet over, hvad begreber er.<sup>77</sup> Den klassiske tilgang er ifølge Murphy et udtryk for en definatorisk forståelse af begreber, som man finder i begrebsteoretiske værker fra før 1970:

*... a reading of the most cited work on concepts written prior to 1970 reveals its assumption of definitions. I should emphasize that these writers did not always explicitly say, "I have a definitional theory of concepts." Rather, they took such an approach for granted and then went about making proposals for how people learned concepts... from experience.*

MURPHY 2004, s. 12

Den klassiske tilgang til begreber tager altså udgangspunkt i, at mennesker forstår begreber ud fra definitioner. Men hvad er en definition?

Definitioner virker måske som den mest logiske måde at adskille forskellige begreber fra hinanden på. Både stole og borde kan have fire ben og en vandret flade, men borde har en større flade, som er velegnet til at placere ting på, og er højere end stole. Det er altså ikke nok at sige, at et bord er et møbel med ben, da det ikke udelukker alle andre møbler med ben, heriblandt stole,

---

<sup>77</sup> Ibid., s. 11

sofaer og det lille skab, der står på fire ben i min stue. Man taler om *nødvendigheder* og *tilstrækkeligheder*.<sup>78</sup> For at noget kan kaldes et bord, er det en nødvendighed at det har ben – og det er tilstrækkeligt, at man kan bruge det til at sætte ting på det, eller: Alle borde har ben, og alt hvad man kan sætte ting på, kan kaldes et bord. Men som den opmærksomme læser måske har bemærket, er jeg allerede ved at presse mig selv op i et hjørne med disse definitioner. Man kan f.eks. forestille sig borde, der har meder i stedet for ben. Og borde er ikke det eneste, man sætter ting på. Jeg har f.eks. et fjernsyn stående på en lille skænk, som har fire ben. Jeg har altså noget stående på skænken, og skænken har fire ben. Det gør dog ikke skænken til et bord. For at en definition er en gyldig definition, skal den dække alle de egenskaber, som de definerede emner har til fælles, samt udelukke alle emner, der ikke er en del af definitionen.<sup>79</sup>

Definitionstilgangen medfører også meget strenge klasser. Strengt logisk vil et emne enten høre under en given definition – eller også vil den ikke. Hvis en Viggo ikke passer på definitionen af en hund, så er Viggo ikke en hund. Viggo kan ikke være 'lidt hund' ligesom min skænk ikke kan være 'lidt bord'.

Murphy opsamler den klassiske tilgang således:

*First, concepts are mentally represented as definitions. A definition provides characteristics that are a) necessary and b) jointly sufficient for membership in that category. Second, the classical view argues that every object is either in or not in the category, with no in-between cases.*

MURPHY 2004, s. 15

### 5.2.1 KRITIK AF DEN KLASSISKE TILGANG

Et af problemerne med den klassiske tilgang er, at det er enormt svært at skabe gyldige definitioner. Murphy citerer en af kritikerne, K. L. Smoke, for at have skrevet, at "if any concepts have ever been formed in such a fashion, they are very few in number. We confess our inability to think of a single one".<sup>80</sup> Man kan sige, at hvis folk, der studerer begreber, har svært ved at definere dem, så har alle andre mennesker det formentlig også, jf. mine utilstrækkelige forsøg ovenfor på at definere et bord. Man skulle tro, at hvis det er definitioner, der styrer menneskers evne til at skelne tomater fra agurker, at mennesket så havde adgang til disse definitioner. Da Pluto mistede sin status som planet, var en af degraderingens modstandsbevægelses argumenter også netop, at der ikke findes en entydig definition af en planet.<sup>81</sup>

Udover at mennesker ikke umiddelbart virker bevidste om de definitioner, som ifølge den klassiske tilgang er grundlæggende i forståelsen af begreber, er der heller ikke noget, der tyder på, at vi bruger begreber definatorisk. Mennesker er i stand til at 'bøje' forståelsen af ord, således at de passer på definitionen. Murphy skriver, at kategorier med fordel besidder en 'fuzziness', altså en uklarhed eller bøjelighed, der f.eks. gør, at vi ikke nægter folk i kørestol adgang til en park, blot fordi der står 'køretøjer forbudt' ved indgangen. En undersøgelse viste på samme

---

<sup>78</sup> Ibid., s. 12

<sup>79</sup> Ibid., s. 13

<sup>80</sup> Ibid., s. 13

<sup>81</sup> Ibid., s. 18

måde, at det lykkedes mange mennesker at leve med tvivlen om, hvorvidt en oliven er en frugt<sup>82</sup>, uden at være i tvivl om, hvad en oliven er, hvordan den smager og bruges i madlavning. Problemerne tårner sig op, når man forsøger at inddele alle begreber i firskårne kategorier, hvor der ikke er plads til *mellemting*. Der er ført forsøg, der undersøgte, hvorvidt folk anså køkkenvaske for at være et køkkenredskab eller ej. 55 procent mente at det var et køkkenredskab og 45 procent mente, at det ikke var. Der er med andre ord ikke enighed om kategoriseringerne.<sup>83</sup> Med de definerede kategorier vil der også være kategorier, der underordner sig andre kategorier (i et hierarkisk system). Eksempelvis hører kategorien stole til i kategorien møbler. Et eksempel fra Murphy beskriver, hvordan folk opfatter stole som møbler, og bilsæder som stole, men ikke nødvendigvis bilsæder som møbler.<sup>84</sup>

Den klassiske forståelse af begreber trængte med andre ord til at blive udfordret. Frem til 1970'erne var den klassiske tilgang den gængse. På meget kort tid blev tilgangen lagt på hylden, og to alternativer steg op fra dens aske: Prototypetilgangen og eksempelartilgangen.

### 5.3 PROTOTYPETILGANGEN

Prototypetilgangen gør op med den definitionsbaserede, klassiske tilgang. Murphy beskriver Eleanor Rosch som værende den, der giver den klassiske tilgang dødsstødet i 1970'erne.<sup>85</sup> I hendes kritik beskriver hun en ny måde at tænke på, som har en prototype som grundlag for begreberne. Grundidéen er, at individet har et ideelt billede, der er styrende for hver kategori, og som udgør grundlaget for hver eneste kategorisering.<sup>86</sup> Jeg kunne eksempelvis have et billede af en Golden Retriever som min prototype for kategorien *hund*. Denne for mig meget typiske hund vil så være sammenligningsgrundlaget for enhver kategorisering, jeg gør af hunde.

Det er dog problematisk at forestille sig, at prototyper skulle kunne udgøres af et *bedste eksempel*. Findes der eksempelvis en 'ideel fugl', som sammenfatter alle tænkelige fugle?<sup>87</sup> Det er svært at forestille sig, at man med en spurv som fugleprototype bevæger sig ned ad gaden og får øje på en struds og straks tænker: "Aha! Dette er en fugl!". Problemet er, at *det bedste eksempel* ikke varetager de enkelte træk, der gør en fugl til en fugl eller en stol til en stol – den forsøger at samle alle egenskaber i en og samme prototype, et *bedste eksempel*.

Prototypetilgangen har derfor bevæget sig fra at være prototypen som *det bedste eksempel* til at være en *akkumuleret repræsentation*. Dvs. at frem for at forestille sig en enkelt repræsentant fra hver kategori, der er styrende for hele kategorien, kan man forestille sig en række egenskaber, der tilsammen udgør en prototype. Men hver egenskab spiller ikke lige store roller. Først og fremmest kan egenskaberne rangordnes – når emner kategoriseres, er nogle egenskaber ganske

---

<sup>82</sup> En undersøgelse foretaget af McCloskey og Glucksberg (1978) bestod i at stille en række mennesker spørgsmål som 'Er en hund et dyr?' eller 'Er en oliven en frugt?'. Efter nogle uger blev forsøget gentaget med de samme undersøgelsesdeltagere. Resultatet viste, at med nogle typer af spørgsmål, f.eks. hvorvidt en oliven er en frugt, ændrede en stor procentdel holdning, og svarede det modsatte af hvad de havde svaret uger forinden. (Murphy 2004, s. 20)

<sup>83</sup> Murphy 2004, s. 20.

<sup>84</sup> Ibid., s. 27

<sup>85</sup> Ibid., s. 16

<sup>86</sup> Sådan blev Rosch i hvert fald læst – hun har selv opponeret mod denne fortolkning, da hun ikke ønskede at skabe forestillinger om hvordan begreberne faktisk blev opfattet. Derimod ville hun bevise og beskrive, at typikalitet var et vigtigt emne for menneskers kategoriseringer. (Murphy 2004, s. 41f)

<sup>87</sup> Murphy 2004, s. 42

enkelt vigtigere end andre. Murphy bruger våben som eksempel. Våben er meget ofte lavet af metal. Men denne egenskab er ikke lige så afgørende som våbenets evne til at gøre skade. Et våben bliver ikke mere et våben af at være lavet af mere metal, men bliver det i en vis forstand af at være bedre til at gøre skade.<sup>88</sup>

Herudover kan listens egenskaber være indbyrdes relaterede. En fugl kan have egenskaben 'rødt hoved' og det kan have egenskaberne 'blåt hoved' eller 'flerfarvet hoved'. Men det kan ikke have dem alle tre på samme tid, ligesom et bord ikke kan have fire ben og have seks ben på samme tid. Den ene udelukker de andre, fordi egenskaberne er relaterede.<sup>89</sup>

Ser vi en hund, kan vi altså genkende den som en hund, ikke nødvendigvis fordi den ligner en golden retriever, men fordi den har en lang række af de vigtigste hunde-træk, f.eks. to ører, en snude, tungen ud af munden og logrende hale. Emner, der er essentielle for at adskille den fra andre emner, rangerer højt på egenskabslisten. At halen logrer, og at den har tungen ud af munden, viser mig hurtigt, at der ikke er tale om en kat. Herudover vil træk, der er typiske for hunde, rangere højt. F.eks. vil egenskaber som 'har pels' og 'har to ører', rangere højt (selvom nogle hunde ikke har), mens 'har et øre' vil rangere lavt (selvom nogle hunde er ude for uheld og derfor kun har et øre).<sup>90</sup>

Med den akkumulerede repræsentation kommer vi uden om de problemer med eksklusivitet, vi så i den klassiske tilgang. Eksempelvis kan vores bilsæde godt være en stol, og en stol kan godt være et møbel, uden at et bilsæde nødvendigvis er et møbel. Årsagen er, at der ikke er en ekskluderende definition på *bilsæde*, men derimod en række egenskaber, hvoraf mange går igen i prototypen for en *stol*. En stol kan igen være beslægtet til et *møbel*, uden at bilsædet har nogen vigtige egenskaber tilfælles med *møbel*.

*Thus, concept A can be similar to concept B, and B can be similar to C, and yet A may not be very similar to C. This can happen when the features that A and B share are not the same as the features that B and C share.*

MURPHY 2004, s. 45

Prototypetilgangen er forholdsvist løst defineret. Murphy kritiserer mange af teoretikerne for ikke at være præcise nok i deres beskrivelse af, hvorvidt de mener, prototyper er *bedste eksempler* eller *akkumulerede repræsentationer*.

*In fact, many statements about prototypes in the literature are somewhat vague, making it unclear exactly what the writer is referring to a single best example? a feature list? if a feature list, determined how? This lack of specificity in much writing about prototype theory has allowed its critics to make up their own prototype models to some degree.*

MURPHY 2004, s. 45

---

<sup>88</sup> Ibid., S. 42f

<sup>89</sup> Ibid., s. 47f

<sup>90</sup> Ibid., s. 43

Herudover virker det også problematisk, hvordan prototyper dannes? Er det første møde med en hund styrende for individets prototypedannelse, eller udvikles det løbende? Der er i teorien uklarhed om, hvordan evt. egenskabslistes udbygges og nuanceres.

Sammen med prototypetilgangen opstod også eksemplartilgangen, som ligeledes er en reaktion mod den klassiske tilgang. Eksemplartilgangen har et meget klart bud på, hvordan menneskers evne til at udvikle eller nuancere begreber udbygges.

#### 5.4 EKSEMPLARTILGANGEN

Eksemplartilgangen har det til fælles med prototypetilgangen, at den ikke mener, begreber opstår på baggrund af definitioner. Begreber er udtryk for vores interaktion med omverden og vores sanseerfaringer. På den måde opstår begreber i gentagne møder med ting, der ligner hinanden. Når et menneske har set tilstrækkelig mange hunde, dannes der et indtryk baseret på summen af alle de møder med hunde, individet har haft, hvorigennem der dannes en slags akkumuleret indtryk, der udgør begrebet hund.

*According to exemplar models of categorization, people decide what a new item is by accessing already known exemplars and relying on them.*

MURPHY 2004, s. 80

Eksemplartilgangen adskiller sig således markant fra prototypetilgangen i opfattelsen af, at individet ikke genkender en labrador på at have fire poter, blød pels og ingen knurhår, men fordi den simpelthen ligner alle tidligere oplevede hunde.

*That is, one's concept of dogs is not a definition that includes all dogs, nor is it a list of features that are found to greater or lesser degrees in dogs. Instead, a person's concept of dogs is the set of dogs that the person remembers. In some sense, there is no real concepts (as normally conceived of), because there is no summary representation that stands for all dogs. However, as we shall see, this view can account for behaviors that in the past have been explained by summary representations.*

MURPHY 2004, s. 49

I modsætning til prototypetilgangen er eksemplartilgangen altså ikke i samme grad statisk. Hver eneste gang, man møder et nyt eksemplar, vil det påvirke ens mængde af eksemplarer. Dette har dog både styrker og svagheder.

Styrken ligger tydeligvis i evnen til at kunne genkende nye eksemplarer hurtigt, ved at have set forskellige typer inden for den samme kategori. Har en person set en masse forskellige hunde, har personen en stor referenceramme, der gør det nemmere at identificere en ny hunderace som en hund. Og ikke blot det: Ved at se eksemplarer i forskellig kontekst, bliver man også bedre til at identificere. Murphy eksemplificerer med oplevelsen af en hund på bagsædet af en bil. Måske er man ikke vant til at se hunde på bagsæder af biler, men når først man har set det et par gange, får man *bedre adgang* til sit begreb hund, når man ser en sådan på bagsædet af en bil.<sup>91</sup>

---

<sup>91</sup> Ibid., s. 88

Svagheden ved, at ethvert møde med et eksemplar påvirker den samlede referenceramme, man bruger til at identificere med, er at hjernen ikke skelner mellem mødet med eksisterende og nye eksemplarer. Studier, først af Nosofsky og senere af Barsalou, Huttenlocher og Lamberts, viser, at mennesker ikke skelner mellem nye og eksisterende eksemplarer, når de tillærer sig kategorier. Det vil sige, at hvis jeg i hele mit liv har mødt hunde et tusinde gange og de nihundrede af dem var min søsters golden retriever, så vil jeg have en overrepræsentation af golden retrievere i min referenceramme af hunde.<sup>92</sup>

Kritikere af eksemplartilgangen påpeger det usandsynlige i, at mennesker er i stand til at huske tusindvis af sanseindtryk hver gang, man identificerer et eksemplar. I realiteten ville det nemlig svare til, at hver gang jeg ser en kat, er det ikke kun de katte-eksemplarer, jeg har mødt i mit liv, jeg skal sammenligne mit sanseindtryk med, men også alle andre type-eksemplarer (f.eks. hunde), for at beregne hvilken kategori, det pågældende eksemplar bedst kategoriseres som.

Et andet kritikpunkt af eksemplartilgangen er, at vi faktisk ikke altid kategoriserer i forhold til vores sanseindtryk. F.eks. ligner hvaler og hajer hinanden meget, men de kategoriseres forskelligt.

## 5.5 FORHOLDET MELLEMLIK EKSEMPLARER OG PROTOTYPER

I forlængelse af denne svaghed er det værd at overveje, hvordan eksemplar- og prototypetilgangen relaterer sig til hinanden. En ting, de har til fælles, er et lighedskriterium. Uanset, om man taler om eksemplar eller prototypetilgangen, er lighed bærende for kategoriseringerne. Ved prototypetilgangen er det et spørgsmål om ligheder på egenskaber, hvor det ved eksemplartilgangen nærmere er et spørgsmål om lighed i sanseindtryk. Men hvad sker der f.eks., når man ser en lama for første gang? Lamaen vil da være det eneste eksemplar og i en tænkt forstand udgøre begrebet *lama*. Når man ser endnu en lama, vil man formentlig begynde at finde tilstrækkelige lighedstræk mellem de to dyr, at man kan danne en egentlig kategori for begrebet *lama*. Murphy beskriver, hvorledes de fleste teoretikere er enige om, at det er på denne måde, kategorier bliver til – ved enkeltstående møder med eksemplarer. Spørgsmålet er så, hvad der sker, hvis man har mødt tusindvis af lamaer. Husker man dem alle hver især? Eller har man dannet sig en definition af, hvad en lama er, som tilhængere af den klassiske tilgang vil påstå? Eller skabes der en slags egenskabsliste (fire ben, udstående øjne, pels, ildelugtende spyt osv.), som tilhængere af prototypetilgangen mener? Der er ikke noget klart svar, og det er netop dette, teoretikerne diskuterer. Der er dog noget, der tyder på, at eksemplarer spiller en stor rolle i forbindelse med indlæring af kategorier.<sup>93</sup>

Hverken prototype- eller eksemplartilgangen er nemme at undersøge, fordi kategoriseringen foregår autonomt, og fordi mennesker er så gode til at kategorisere, at det foregår meget hurtigt. Når vi genkender en hund, er vi os således ikke bevidste om, hvorvidt vi benytter os af en prototypemodel, eller om vi ubevidst løber en lang række tidligere møder med lignende dyr igennem, før vi kan kategorisere, hvad vi oplever som en hund.

At eksemplarer og prototyper spiller forskellige roller i forbindelse med læring af begreber, er dog blevet påvist gennem forsøg, hvor forsøgspersoner skal lære forskellige kategorier med

---

<sup>92</sup> Ibid., s. 58ff

<sup>93</sup> Ibid., s. 51

forskellige mængder af eksemplarer. I forsøg, hvor der bruges få eksemplarer, har forsøgspersoner tendens til at huske hvert enkelt eksemplar, når de en uges tid efter at være blevet præsenteret for eksemplarerne, skal kategorisere dem. Er der derimod mange eksemplarer, har forsøgspersonerne en tendens til at kategorisere efter overordnede egenskaber (som i en prototypes egenskabsliste) fra hver kategori. Man kan med andre ord forestille sig, at de konkrete eksemplarer spiller en stor rolle i den tidlige indlæring af kategorier, men at det er nemmere at huske prototyper på sigt, især når det gælder større datasæt.<sup>94</sup>

## 5.6 VIDENSTILGANGEN

Som tidligere nævnt er prototype- og eksemplartilgangene reaktioner på den såkaldte klassiske tilgang. De to tilgange har vist sig ikke nødvendigvis at være to adskilte tilgange, men at der er tale om to forskellige metoder, som mennesker bruger sammen i forbindelse med brug af begreber og kategorier. Videnstilgangen adskiller sig fra disse to tilgange ved ikke at være en reaktion på den klassiske tilgang, men nærmere en overbygning på prototype- og eksemplartilgangen.

Grundtanken i videnstilgangen er mere holistisk, end den er i de prototype- og eksemplartilgangene, som hver især beskæftiger sig med den meget afgrænsede opgave at danne begreber ud fra kategorier. Videnstilgangen ser nemlig ikke processen som en afgrænset proces, men som en del af det at lære gennem livet.

*We do not learn concepts in isolation from everything else (as is the case in many psychology experiments); rather, we learn them as part of our overall understanding of the world around us.*

MURPHY 2004, s. 60

Tanken er, at når vi identificerer noget, bruger vi ikke kun sammenligninger med tidligere oplevelser eller idealiserede mentale repræsentationer, men også den *viden*, vi har. Hører vi f.eks. en lyd fra noget, der rammer et vindue, og ved vi, at der ligger sne udenfor, vil vi ikke være længe om at komme til den konklusion, at der er tale om en at snebold har ramt ruden, til trods for at vi hverken har set snebolden eller den ballademager der har kastet den. Vi bruger vores viden om sne, vinduesruder og ballademagere til at skabe en forståelse af hvad der er sket – og til at identificere hvad det var der lavede lyden. Det er også vores viden om biologi der gør, at vi ikke kategorisere hvaler og hajer sammen. De er to forskellige arter, den ene et pattedyr, den anden en stor fisk.

Murphy bruger et eksempel, hvor et barn identificerer et lille, uldent, gråt dyr der svømmer i halen på en stor, hvid gås som gåsens unge, uden nogensinde før at have set en gæsling. Barnet bruger sin (om end sparsomme) biologiske viden til at ræsonnere, at små dyr, der følger tæt efter stort dyr, kan være deres afkom, til trods for den sparsomme lighed.<sup>95</sup>

Et andet eksempel på hvordan vores viden om emner påvirker vores kategorisering, er i forbindelse med light-produkter. Light-produkter er en kategori, hvor der er en meget snæver

---

<sup>94</sup> Ibid., s. 55ff

<sup>95</sup> Ibid., s. 61

lighed mellem de forskellige eksemplarer. Light-rullepølse, Minimum rugbrød og minimælk har typisk blot et lavt fedt- eller sukkerindhold til fælles og har på den måde ikke de samme åbenlyse ligheder, som man finder i andre fødevarer-kategorier som f.eks. mælkeprodukter, oksekød eller brød. Det er ikke vores sanseindtryk eller konkrete egenskaber, disse produkter har til fælles, men vores viden om deres lave indhold af fedt eller sukker, der forbinder dem.<sup>96</sup>

Murphy påpeger fire områder<sup>97</sup>, hvor viden spiller ind i menneskers kategorisering:

1. Viden bruges, hvor den er relevant. Når barnet, der for første gang oplever gæslingerne svømme bag gåsen, er det ikke tidspunktet på dagen, luftens temperatur eller antallet af siv i søen, der afgør barnets tolkning, men nærmere det, at gæslingerne følger tæt efter deres mor, at de er mindre end deres mor, at der ikke er tale om et helt andet type dyr osv.
2. Viden gør det nemmere at huske begreber. Når mennesker ikke forstår et emne i en kontekst, er det sværere at huske. Det er nemmere at genkende en flyver på, at den har vinger, fordi de spiller en essentiel rolle i flyets primære funktion, end det er at huske en tilfældig talrække, som man ikke har nogen viden om og derfor ikke kan placere i en kontekst.
3. Viden kan påvirke dine kategoriseringer direkte. Havde barnet ikke set gæslingernes mor, og havde barnet ikke haft viden om, at alle dyr får børn, og at børn tit følger deres mor, havde barnet formentlig haft svært ved at kategorisere gæslingerne som værende af same race som gåsen.
4. Viden kan være grundlaget for udledninger om en kategori. Murphy bruger et eksempel, hvor vi ser et pattedyr for første gang, men vi identificerer det hurtigt som et pattedyr og som en unge, fordi den holder sig til sin mor. Ser vi denne unge begynde at plage sin mor, vil vi hurtigt få tanken, at den formentlig er sulten og gerne vil patte. Denne tolkning sker til trods for, at vi aldrig har set pattedyret før og intet ved om netop dette pattedyrs egenskaber og handlingsmønstre. Men på baggrund af vores generelle viden er vi i stand til at forestille os, hvad barnet er ude efter.

Det, der adskiller videnstilgangen markant fra prototype- og eksemplartilgangene, er, at den foreslår, at mennesker bruger deres eksplicite viden i forbindelse med kategorisering. Hvor forskere er enige om, at brugen af prototyper og eksemplarer foregår implicit, kan den viden, vi er bevidste om, at vi har, påvirke den kategorisering, vi laver.<sup>98</sup>

Murphy mener, det er usandsynligt, at mennesker bruger deres eksplicite viden, når de til dagligt orienterer sig. Man tænker hverken over biologi eller stamtavler, når man identificerer en hund på vej til bageren, ligesom man ikke behøver at tænke på, hvor den vej, man går på, stammer fra. I andre tilfælde vil det dog være nødvendigt at trække på sin eksplicite viden for at

---

<sup>96</sup> Ibid., s. 63f

<sup>97</sup> Ibid., s. 146f

<sup>98</sup> Implicit viden kan forstås som den viden vi ikke tænker over vi har. F.eks. tænker du lige nu ikke over, at du kender alle bogstaverne i alfabetet eller din evne til at læse – du læser bare. Den eksplicite viden er den du er bevidst om at du har. Det du ved, du ved. F.eks. de refleksioner du har, mens du læser denne tekst. Et andet eksempel er at cykle: Vi bruger vores implicite viden til at holde balancen og bevæge os fremad, mens det er vores eksplicite viden, der hjælper os med at finde vej.

forstå en kategorisering, som f.eks. i tilfældet med hvaler og hajer, eller snebolden der rammer ruden.

*More generally, if you have to categorize something based on minimal perceptual information (e.g., a noise at the window), you might naturally rely almost entirely on background knowledge to arrive at an answer (a bird might have flown up to the window and then flown away; a dog couldn't get up there and disappear so suddenly).*

MURPHY 2004, s. 168

## 5.7 OPSAMLING PÅ MENING OG KATEGORISERING

Som Murphy selv skriver, er begrebsdannelse og kategoriseringsprocessen ikke nemme at finde rundt i.

*On the proposal I am making, people attempt to form prototypes as part of a larger knowledge structure when they learn concepts. But at the same time, they remember exemplars and these memories may influence them in a variety of ways. In short, concepts are a mess.*

MURPHY 2004, s. 492

Det er svært at undersøge, hvordan folk kategoriserer emner, fordi forsøg altid må foretages isoleret og typisk med meget forenklede forsøg. Ikke desto mindre er der blevet udført tilstrækkelig mange forsøg til, at der har udformet sig teorier, der forklarer de fleste af de mekanismer, man har identificeret. Ingen enkeltstående teori kan give forklaring på alt. Men hvor den ene teori fejler, sejrer den anden, og i samspillet mellem prototype-, eksemplar- og videnstilgangen får vi et dækkende billede af, hvilke mekanismer der er i spil, når mennesker kategoriserer.

Der er noget, der tyder på, at brugen af eksemplarer spiller en afgørende rolle, når nye kategorier opstår, men at mennesker bruger prototyper til at huske fællestræk og dermed ikke er afhængige af at kunne huske et enormt grundlag af eksemplarer for at kunne kategorisere.

Foruden brugen af prototyper og eksemplarer har det vist sig, at viden også kan spille en afgørende rolle for menneskers kategorisering af emner – dels i forbindelse med kategoriseringer, der beror på eksplicit viden om de enkelte emner (f.eks. i forbindelse med slankeprodukter eller haj/hval distinktionen), dels i forbindelse med at huske kategoriseringer (f.eks. at en flyver kan genkendes på at have vinger, som netop er en af de egenskaber, der gør det muligt at flyve), men også til at forstå kategorier (som f.eks. at små dyr, der plager større dyr, kan være dyrebørn, der er sultne).

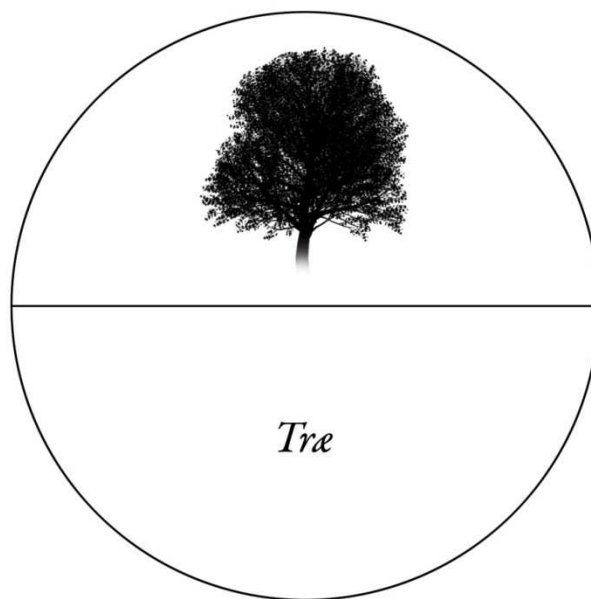
I det følgende afsnit vil jeg igen tale om internetteknologier og analysere, hvordan menneskets kategoriseringsmetoder kan sammenlignes med semantic software og Semantic web.

## 6 MENING OG SEMANTIC SOFTWARE

Med en forståelse af hvordan mennesker bruger begreber til at beskrive verden, er det relevant at se på, hvordan mening anvendes og kan forestilles at blive anvendt i semantic software. Jeg vil i dette kapitel beskrive forholdet mellem menneskelig og softwarebaseret tolkning af mening og efter en kort opsummering beskrive forskellige tænkelige scenarier for, hvordan semantic software kan tænkes at gøre en forskel for menneskers dagligdag. Ved at anskueliggøre semantic softwares mulige roller, er der et grundlag for at kunne udarbejde designforslag til en visualisering baseret på semantiske analyser.

### 6.1 TEGN OG BETYDNING

Når man taler om *semantics* - som i semantic software - tales der om *mening*. Semantikken<sup>99</sup>, der beskæftiger sig med betydning, forbindes ofte med semiotikken<sup>100</sup>, som er læren om tegn. Den del af semiotikken, der beskæftiger sig med sproget, kaldes som regel for semiologien, jf. Ferdinand de Saussures forelæsninger om almen lingvistik<sup>101</sup>. Ferdinand de Saussure opdeler tegnet i et indhold og et udtryk. Eksempelvis er det talte ord 'træ' et tegn, der som udtryk (det akustiske billede – det der kommer ud af afsenderens mund) bliver fortolket af en modtager, der danner en forståelse eller mening med udtrykket.



FIGUR 8 – SAUSSURE OPDELER TEGNET I ET INDHOLD OG ET UDTRYK.

---

<sup>99</sup> Af græsk, semantikos, betydning

<sup>100</sup> Af græsk, semion, tegn

<sup>101</sup> Saussure 1991. Teksten er ikke Saussures egen formulering, men en bearbejdning af noter fra hans forelæsninger. Saussure skrev aldrig selv om semiologien, men er gennem hans forelæsninger blevet anerkendt som frontfigur inden for den sproglige semiotik: semiologien.

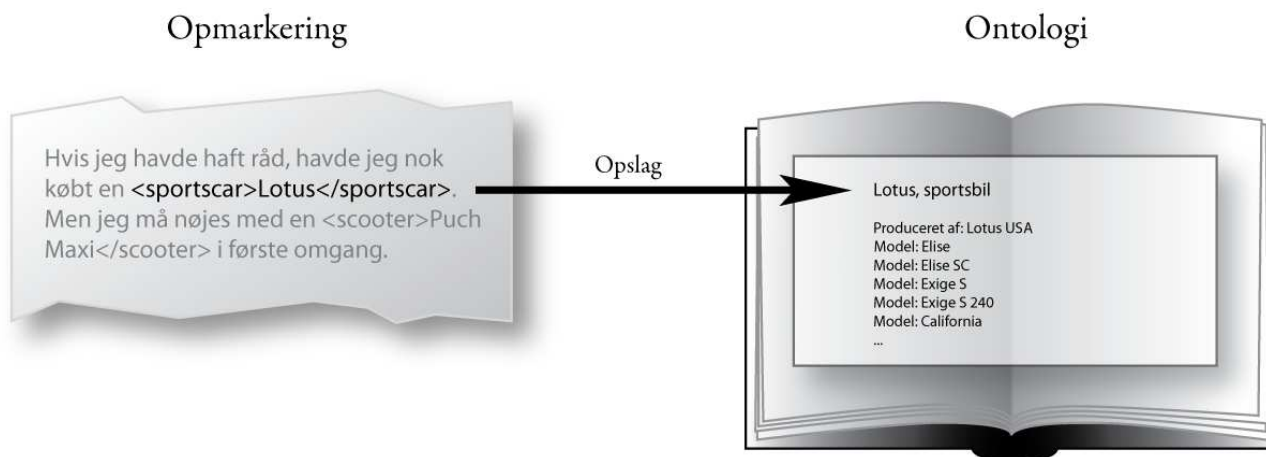
Tegnet som sådan består derfor dels af det udtryk, som afsenderen skaber, og dels af den betydning, som afsenderen forsøger at overlevere til modtageren. Man kan sige, at det semantiske ligger i forholdet mellem udtryk og indhold: Hvad betyder dette tegn? I forlængelse af diskussionen af, hvordan begreber og kategorier skabes (jf. Murphy 2004), er det oplagt at drage en parallel mellem semiologien og begrebspsykologien. Men hvor begrebspsykologien er baseret på empiriske data og beskæftiger sig med forståelsen og opfattelsen af begreber, handler semiologien om *kommunikationen*, der sker ved brug af tegn (f.eks. et begreb, der bliver udtalt). Semiologen interesserer sig egentlig ikke for, hvordan begreber kategoriseres eller forstås i forhold til hinanden, men nærmere til det udtryk og det indhold, der er indlejret i de tegn, vi benytter os af, når vi kommunikerer. Til trods for, at Saussure døde i starten af det tyvende århundrede<sup>102</sup> og derfor ikke har været vidne til kritikken mod den klassiske tilgang fra begrebspsykologien, er der ikke umiddelbart noget, der tyder på, at Saussure ville anskue begreber for at være bestemt af definitioner. Tværtimod gør Saussure sig umage med ikke at sige, at det er begreber, der udveksles, når to mennesker taler sammen. Det er en proces, hvor en idé udformer sig som et akustisk lydbillede (altså de lyde, der frembringes af afsenderen), opfattes af en modtager (som hører lydene) og derefter danner en idé ud fra indtrykket. Havde Saussure været tilhænger af den klassiske tilgang, havde han formentlig nøjedes med at forklare, hvordan den ene kan overføre sin idé om et bord ved at sige *bord* til den, der lytter. Men som Saussure påpeger, er der altså tale om, at den idé, der bliver til en formulering og siden hen et akustisk indtryk hos en modtager, ikke nødvendigvis resulterer i samme idé, som afsenderen havde forestillet sig. Sagt på en lidt anden måde kan det sagtens tænkes, at både Flemming og Ulla ville vælge at være fuldstændig enige om, hvilke møbler hos møbelhandleren der er borde. Men når Ulla siger bord, tænker hun på et stort fyrretræsbord, som det, hendes forældre købte, da hun var barn, hvorimod Flemming tænker på sit lille nye skrivebord fra IKEA.<sup>103</sup>

Denne tankegang bliver yderst relevant i diskussionen om, hvordan Semantic web er semantisk. Semantic web består i bund grund af en række udtryk, som bruges som meningsgivende metadata på eksisterende indhold, og en ontologi, som er en slags opslagsværk over betydninger, som de forskellige udtryk kan 'pege på'. Når jeg opmarkerer min tekst <sportscar>Lotus</sportscar>, ved teksten ikke selv, hvad der menes – den er jo trods alt stadig blot en tekst. Men ved et opslag i en ontologi vil man finde ud af, at <sportscar> er af typen bil, og at man i denne kategori finder Lamborghini, Ferrari og Porsche. Det er altså først i opslaget i ontologien, at der faktisk tilføjes mening til teksten. Opmarkeringen af teksten tilføjer med andre ord en detaljerigdom til udtrykket, der gør det nemmere for en maskine at tildele et indhold (dvs. en betydning eller mening).

---

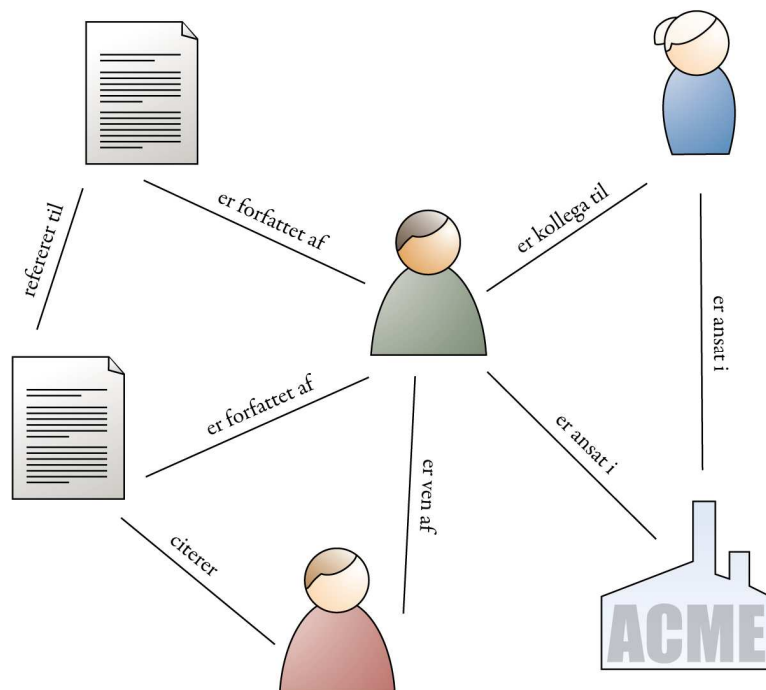
<sup>102</sup> Ligeså gjorde Charles Sanders Peirce, som anses for at være en af semiotikkens frontfigurer. Hvor Ferdinand de Saussure beskæftigede sig med semiologien, den lingvistisk orienterede del af semiotikken, er Peirce blevet anerkendt for sit enorme arbejde inden for hvad man kan kalde den brede semiotik, der i en vis forstand anskuer hele verden (eller perceptionen af denne) som tegn.

<sup>103</sup> Saussure 1991, s. 410f



FIGUR 9 - FORHOLDET MELLEM OPMARKERING OG OPSLAG I ONTOLOGIEN.

For at opmarkeringen skal give mening, skal der altså opmarkeres *i forhold* til en ontologi, som f.eks. OWL. Det giver ikke meget mening at opmarkere `<kagekrymmel>krokant</kagekrymmel>` hvis ontologien ikke har et opslag for `<kagekrymmel>`. Ontologien ikke må misforstås af nogen. Den bør derfor være konsistent og ikke være tvetydig eller foranderlig. Hvis ontologien er konsekvent, giver det i langt højere grad end i dag mulighed for at tyde, hvad enkelte sider på internettet handler om, hvem der har lavet dem, hvem de, der har lavet dem, er osv. Der vil være mulighed for at forbinde alt på nettet i en grad som ikke tidligere har været muligt.



FIGUR 10 - SEMANTIC WEB FORBINDER INFORMATIONER, MENNESKER, STEDER, VIRKSOMHEDER OSV.

Det er netop visionen med Semantic web, at man ikke skal være i tvivl om, hvem der er tale om, når der står 'Hans Jensen', ligesom at søgemaskinen skal være i stand til at skelne mellem softwareprogrammet Lotus og racerbilen Lotus. Det hele afhænger af en systematiseret og konsekvent måde at opmarkere tekst på, samt en konsekvent, utvetydig og uforanderlig ontologi som opslagsbase.

Men netop dette kan vise sig at være en svaghed for Semantic web, eftersom sproget både kan være tvetydigt og foranderligt. Det interessante ved ontologi-tankegangen er, at man faktisk forsøger at benytte sig af en definatorisk forståelse af begreber, netop som i den klassiske tilgang, som de fleste forskere inden for begrebspsykologien ikke mener afspejler den måde, mennesker faktisk forstår begreber på, da der ikke findes naturlige definitioner (se afsnit om den klassiske tilgang).

## 6.2 SEMANTIC SOFTWARE PÅ NETTET

Men hvad er da alternativerne? Findes der overhovedet modne teknologier som alternativer? Som jeg redegør for i kapitel 4, findes der en lang række teknologier, der passer under begrebet semantic software. Problemet med teknologier som NLP og de metoder, jeg har opridset i forbindelse med text mining, er, at der ikke er tradition for at bruge dem i stor skala, som f.eks. generelt på internettet. Metoderne bruges ofte i forbindelse med afgrænsede datasæt, som nemmere kan styres og overvåges. Et eksempel beskriver, hvordan indsamlede data fra biologiforskere blev scannet og lagt i en database gennem en automatiseret proces, der var i stand til at sammenkæde relaterede forskningsresultater.<sup>104</sup> Andre studier har gjort sig forsøg på at rydde op i manglende tegnsætning, slåfejl, for mange mellemrum og afstande mv. i e-mails.<sup>105</sup> Ingen har dog vist eksempler på behandling af de nærmest endeløse mængder af data, der er tilgængelig på nettet.

Der er dog også en lang række udfordringer, der gør det sværere at behandle tekst på nettet end i et kontrolleret miljø. Meget af den tekst, der findes på nettet, er, hvad man kan kalde 'uren tekst'. Det er ikke blevet korrekturlæst, og der kan være mange både stave- og grammatikfejl. Det kan f.eks. være svært for et automatiseret system at analysere en sætning, hvor der ikke er sat punktummer, eller hvor ingen af udsagnsordene er bøjet korrekt. En anden problemstilling er, at der ikke er nogen standard for sprog på internettet, så det vil altså også være nødvendigt for systemet selv at gætte, hvilket sprog der skrives på. Dette er måske ikke noget problem i længere tekster, men i meget korte tekster kan der måske være utilstrækkelige detaljer til at kunne bestemme sproget. Herudover er der heller ikke noget til hinder for, at man som bruger skriver på sit eget sprog i et forum, der ellers primært fungerer på et andet sprog. På WWW er der også en tendens til brug af slang og udvikling af sproget. Et eksempel herpå er den 'nørdede' jargon, hvor man intentionelt staver visse ord forkert og erstatter bogstaver med tal (f.eks. *teh* i stedet for *the*, *haz* i stedet for *have/has*, *l33d* i stedet for *elite*<sup>106</sup> osv.). Dette er blot nogle af de problemstillinger, forskere arbejder på at løse for at kunne benytte sig af kompetent software, der kan få maskiner til at forstå eller afkode menneskeligt sprog.

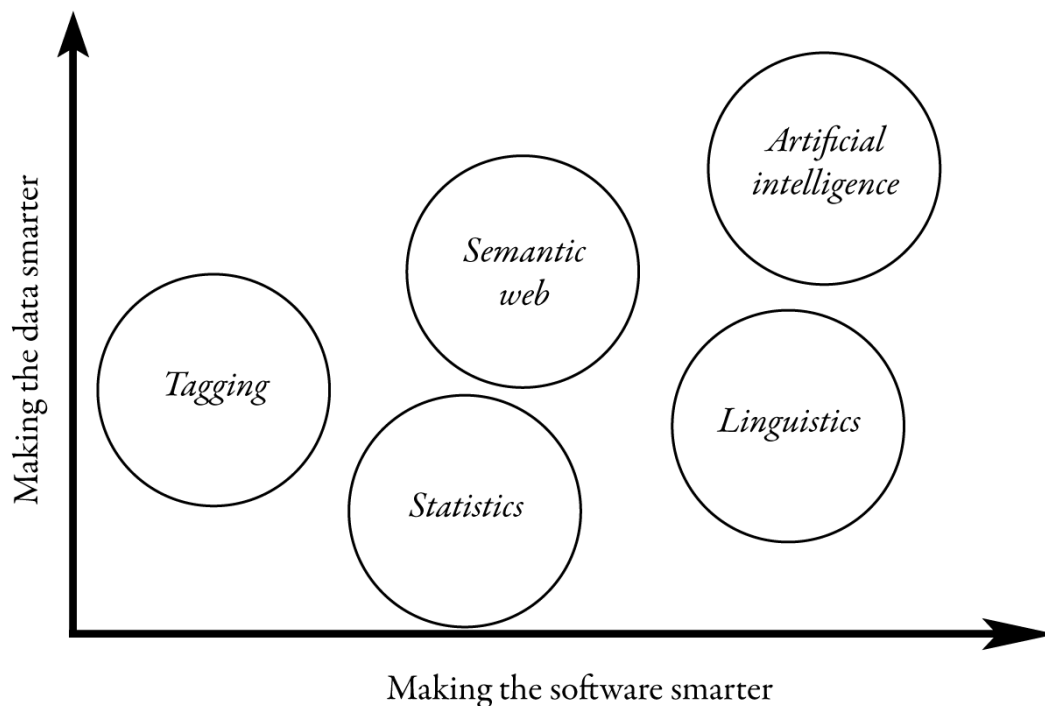
---

<sup>104</sup> Weiguo et al. 2006

<sup>105</sup> Tang et al. 2005

<sup>106</sup> Elite kan også udtales L (el) - eed (i't), hvilket bliver til l-eed. E erstattes med 3 og bliver således til l33d.

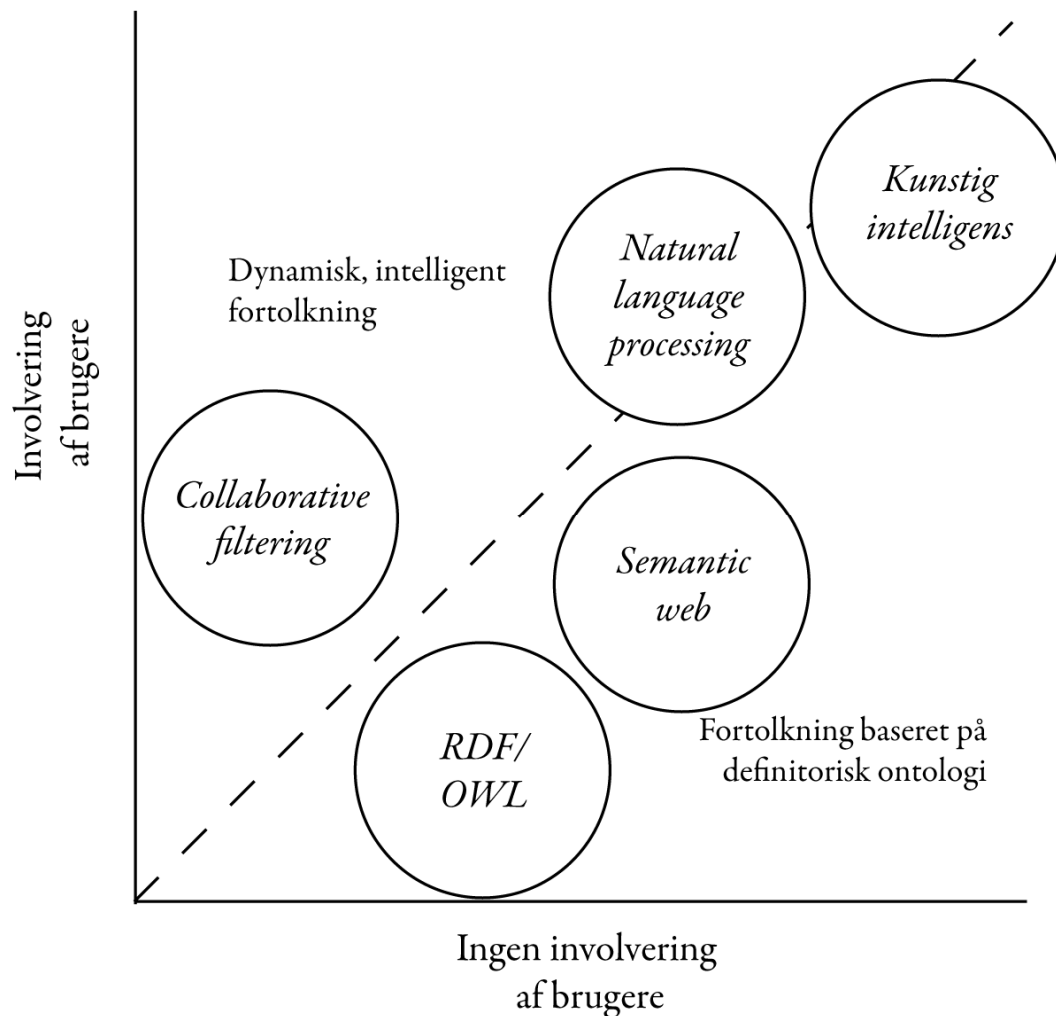
I bund og grund er der tale om to udviklinger inden for udvinding af data fra naturligt sprog. Den ene bevægelse handler om at gøre data 'klogere'; den anden at gøre softwaren 'klogere'. At gøre data klogere vil i denne sammenhæng sige at tilføje metadata til data, der ikke bare gør indholdet, men meningen med indholdet, generelt tilgængeligt. Det er, hvad man forsøger at opnå ved at opmarkere tekst med semantiske tags. Den anden bevægelse, at gøre softwaren klogere, handler om at gøre softwaren uafhængig af de data, den skal fortolke, således at den på automatisk vis kan fortolke f.eks. en tekst, uden at teksten på forhånd er opmarkeret. Her skal der trækkes på NLP og kunstig intelligens.<sup>107</sup>



FIGUR 11 – KLOGERE DATA OG KLOGERE SOFTWARE - FRIT EFTER SPIVACKS FIGUR.

At gøre data klogere kan f.eks. betyde at supplere data med metadata. F.eks. kunne metadata fortælle, at en pågældende side på nettet var en opskrift, hvilke ingredienser der var i, hvor mange kalorier der var i, hvem der havde lavet opskriften, hvilke bøger eller blade opskriften havde været i osv. At gøre softwaren klogere vil sige at finde de samme informationer, men uden at den oprindelige data er tilknyttet metadata. Softwaren skal så selv finde ud af, at det er en opskrift, identificere ingredienser og mængder, forfatter osv. På den ene side har man decentraliseret data – på den anden side har man centraliseret software. Nova Spivack mener, at Semantic Web fungerer som et kompromis mellem de to. Denne tolkning kan forstås således: På den ene side foregår der en forholdsvis enkel opmarkering af indholdet (som i princippet kan udføres automatiseret baseret på anden semantic software), men på den anden side holdes opmarkeringen op imod en ontologi, som er centraliseret; den er fælles for alle opmarkerede sider (som opfylder RDF/OWL-standarderne).

<sup>107</sup> Spivack 2008b



FIGUR 12 - INVOLVERING AF BRUGERE

Spivacks diskussion om klogere data vs. klogere software kan også beskrives som et spørgsmål om at inddrage brugere eller ej. Inddrages brugere løbende, som i en form for Collaborative Filtering, får man et system, der konstant opdaterer sig selv, men hvor mening er flygtig. Inddrager man ikke brugerne, og benytter sig af en statisk ontologi, forstener man derimod sproget, og tillader ikke ny anvendelse af det eksisterende sprog.

### 6.3 SEMANTIC SOFTWARE I FORHOLD TIL MENNESKELIG KATEGORISERING

En anden måde at beskrive semantic software på er ved at sammenligne med de metoder, mennesker bruger til at kategorisere.

Som jeg har nævnt tidligere, besidder Semantic web en definatorisk kategoriforståelse. Der findes en fast liste over, hvilke biler der falder under kategorien `<sportscar>`. Når der kommer en ny racerbil, skal ontologien opdateres, og definitionen på `<sportscar>`-kategorien bliver på den måde ændret. Som tidligere nævnt giver den definatoriske tilgang nogle problemer. Hvordan håndteres emner, der er en mellemting mellem to kategorier? Og hvad sker der i de tilfælde, hvor folk slår op under grøntsager for at finde tomater (hvor de burde have kigget under frugt,

da tomaten botanisk set er et bær)? Problemet med den definatoriske tilgang er, at et emne ikke umiddelbart kan være i to kategorier på samme tid.

Et af de problemer, forskere (bl.a. Eleanor Rosch) fandt frem til i deres kritik af den klassiske tilgang, var netop også, at mennesker ikke forstår emner i definerede kategorier.<sup>108</sup> I den klassiske tilgang tilhører noget enten en kategori, eller også tilhører det en anden. Det viser sig dog, at mennesker er bedre til at kategorisere ting, der er typiske for en kategori, end emner der er atypiske eller næsten tilhører en anden kategori. Dette vises helt konkret af undersøgelser af *false relatedness* foretaget af Smith & Medin (1981): Det viser sig, at det er sværere at svare på spørgsmålet 'er en hund en kat?' end på spørgsmålet 'er en hund et bjerg?'. Prototypetilgangen forklarer denne form for typikalitet med, at en hund og en kat har flere egenskaber til fælles end hunde og bjerge – derfor er de tættere på hinanden som begreber. Eksemplartilgangen, hvor mennesker benytter sig af erfarede møder med eksemplarer, forklarer det med, at vores møder med hunde ligner vores møder med katte – men ikke vores møder med bjerge.

OWL, den ontologiske teknologi, som Semantic web bygger på, er hierarkisk opbygget og afspejler derfor ikke menneskers måder at opfatte kategorier på og er ikke i stand til at repræsentere menneskers opfattelse af typikalitet. Men er dette overhovedet et problem? Den hierarkiske opbygning er meget enkel, og denne enkelhed har mange fordele. Den gør systemet nemmere at sætte sig ind, nemmere at udvikle software til, og de fleste mennesker har en god forståelse af det hierarkiske system. Så hvorfor udfordre dette? Det er en diskussion mellem, hvorvidt man ønsker et lettilgængeligt, decentralt system, som alle kan få adgang til, eller om man vil have et dynamisk centralt styret system, der er kompliceret, men mere korrekt, og i stand til at tilpasse sig internetbrugernes brug af sproget. Går man i den ene retning, risikerer man at lave et system, der med tiden bliver mindre og mindre præcist, mens sproget udvikler sig, og som ikke afspejler menneskers måde at kategorisere på. Går man i den anden retning, risikerer man at lave et system, der er meget tungt, centralt styret og enormt dyrt at udvikle. Indtil videre ser det ud til, at der er mest fokus på Semantic web-tankegangen, med en fast ontologi og en afhængighed af opmærkede data.<sup>109</sup>

## 6.4 OPSUMMERING

Lad os lave en kort mellemlanding: Der findes grundlæggende to metoder til at udtrække data fra naturligt sprog.

Den ene er centraliseret, afhænger i høj grad af maskinkraft, der gennem avancerede algoritmer analyserer tekst for at udtrække 'mening'. Når det endelig lykkes at skabe sådan et system, vil det være uafhængigt af, hvilken type af tekst, det får leveret, og det vil kunne udtrække data fra det. I bedste fald er der tale om en form for kunstig intelligens, der kan opdage sproglige udviklinger såsom slang og tilpasse sin egen algoritme. Metoden er tung og teknologi-afhængig, og selvom man er kommet meget langt, er denne form ikke en realitet på så stort og ukontrolleret et område som WWW. Med den rivende udvikling er det dog ikke usandsynligt, at man i en nær fremtid vil få systemer, der faktisk er brugbare.

---

<sup>108</sup> Murphy 2004, s. 35

<sup>109</sup> For en liste over aktivitet på området, henvises der til <http://www.w3.org/2001/sw/>. Enkelte projekter, eksempelvis Powerset baserer sig bl.a. på NLP og ikke på Semantic web.

Den anden metode tager udgangspunkt i at opmarkere data og skabe systemer, der forstår opmærkede data. Dette betyder væsentligt lettere applikationer, der ikke er afhængige af kunne tyde teksten fra bunden; de kan nøjes med at tyde metadata. Metoden kræver, at opmærkningen er konsekvent, f.eks. som i Semantic web-løsningen, hvor metadata fungerer i sammenhold til en ontologi, en database over mening. Denne løsning vil kunne have en effekt på data, der er opmærket. Systemer som Twine.com forsøger at løse dette problem ved at blande brugerinvolvering og automatiseret opmærkning af indhold. På grund af den manglende robusthed i forhold til naturligt sprog har Semantic web formentlig sin største fordel inden for kategorier, hvor der ikke sker store forandringer i definitionerne på emner, f.eks. i forbindelse med produktkataloger. Det er f.eks. oplagt at bruge opmærkning i forbindelse med CDer, således at man kan sammenligne priser eller lave store kataloger over musikudgivelser ved at forbinde eksisterende databaser. Andre eksempler på solide kategorier er bredde- og længdegrader, datoer og måleenheder.

Da begge teknologiske tendenser stadig er under udvikling, er det umuligt at konkludere, om en af tendenserne har en større fremtid end en anden. Semantic web-tankegangen, der bygger på opmærkninger af tekst og en vision om 'World Wide Web som database', kan reelt implementeres i dag, men den har nogle ulemper i forhold til naturligt sprog og vil være stærkest inden for meget solide kategorier, f.eks. i forbindelse med produkter. Og til trods for, at teknologier, der i højere grad baseres på behandling af naturligt sprog og kunstig intelligens tegner en lovende fremtid, er teknologierne i dag ikke modne til at kunne udnyttes på World Wide Web.

## 6.5 HVAD KAN VI BRUGE SEMANTIC SOFTWARE TIL?

Jeg har tidligere beskrevet, hvordan man kan anskue udviklingen inden for de personlige computere og World Wide Web som et pendul, der svinger mellem backend- og frontendorienterede teknologier. Lad mig kort resumere tendensen: Fra et backendorienteret Web 1.0, hvor der blev udviklet teknologier, der i højere grad lod brugere deltage aktivt på World Wide Web, svingede pendulet med Web 2.0 tilbage til at fokusere på frontend. Teknologierne var på plads, og der blev i stedet fokus på at udnytte teknologierne og gøre dem brugervenlige. I løbet af kort tid eksploderede mængden af indhold på internettet, og med den stigende mængde af indhold bliver det sværere og sværere at finde indhold gennem de traditionelle måder at søge informationer på. Pendulet er nu ved at have opsamlet tilstrækkelig energi til at svinge tilbage igen med et behov for nye metoder til at behandle de enorme mængder af data. Nova Spivack beskriver dette som en 'opgradering af webbens ryggrad'. Med et øget fokus på backenden vil man altså arbejde på at gøre data lettere tilgængeligt og at kunne sammenkæde data i en grad, der ikke tidligere har været mulig.

Ser vi på de teknologier, jeg har ridset op indtil videre, kan vi identificere en udvikling, der går i to retninger. Dels ønskes der at påføre World Wide Web nogle af egenskaberne fra database-systemer for at kunne udnytte nogle af de åbenlyse fordele, der er ved sådanne<sup>110</sup>. Og dels at kunne afkode meninger fra tekster, således at man bedre kan skabe muligheder for at gruppere,

---

<sup>110</sup> Databaser er nemme og hurtige at søge i, og er en relationel database skruet korrekt sammen, er der gode muligheder for at lave meget specifikke søgninger, der kombinerer forskellige kriterier i relationel algebra, f.eks. SQL. SPARQL er en del af Semantic web og er et forsøg på at lave en forespørgselsorienteret sprog til at udtrække data fra Semantic web-services.

sammenligne og navigere i stofområder baseret på deres mening og ikke blot nøgleord, sådan som man gør gennem søgemaskiner i dag.

Der bliver lagt mange kræfter i at implementere vor tids teknologier og i at forske i nye, der vil være endnu bedre til at analysere semantisk indhold. Bevægelsen mod mere semantiske behandlinger af indhold er nærværende, og det er kun et spørgsmål om tid, før semantisk software bliver hverdagsværktøj. Formentlig vil det være indlejret i software og andre IT-produkter, vi benytter, og de fleste brugere vil formentlig ikke opdage, at det er der – blot at de har lettere adgang til det materiale eller de data, de søger. Det store spørgsmål er så, hvornår front-/backendpendulet igen svinger frem og lader brugerne blive medskabere af nye former for data, og hvordan det kommer til at påvirke den måde vi kommunikerer på.

Jeg vil i dette kapitel komme med et par bud på, hvad Semantisk software kan bruges til, når teknologierne bliver modne til det. Eksemplerne tager udgangspunkt i, at teknologier som NLP er blevet mere anvendelige på WWW.

### 6.5.1 PERSONLIGE AGENTER

Som jeg allerede har nævnt, er både Spivack og O'reilly inde på, at hvis det bliver muligt for software at skabe mening med indhold på internettet, vil det åbne op for muligheden at lave intelligent software, der kan assistere mennesker. Nova Spivack påpeger, at man bør udnytte computere til at udføre de opgaver, som computere i princippet kan udføre hurtigere og bedre end mennesker. Han mener, at mennesker skal have mere plads til at gøre det, de gør bedst, ved at maskiner gør de ting, som mennesker ikke er gode til, nemlig alle de kedelige ting såsom at finde dem internetside der har den information vi leder efter, at finde den hurtigste rute til lufthavnen, at indsamle priser på ferier eller at lave en telefonliste over medarbejderne i afdelingen. Spivack mener derfor, man burde tale om 'artificial stupidity' frem for 'artificial intelligence'.<sup>111</sup>

O'Reillys vision for semantic web beskriver også scenarier, hvor mennesker skal udføre en opgave, men hvor personlige agenter hjælper brugeren med at udføre opgaven nemmest muligt. Skal en person f.eks. besøge sin syge mor på hospitalet, så vil en personlig agent kunne finde den korteste vej, regne ud, om det prismæssigt og tidsmæssigt bedre kan betale sig at tage offentlige transportmidler end at køre i bil, tage højde for vejarbejde, myldretid osv. Alt sammen muligt, fordi agenten har adgang til det relevante indhold på nettet. Skulle man udføre det samme arbejde selv, skal man først og fremmest vide, hvilke hjemmesider man skal besøge, indsamle data ud fra et scenarie ('jeg kører i bil derhen', eller 'jeg vil med det offentlige') og sammenkoble de mange informationer til en handlingsplan ('jeg tager bilen, kører ad ringvejen og drejer af før vandtårnet' eller 'tag 14'eren til Trianglen og skift bus'). Agenten kan, såfremt den har adgang til dataene (hvilket er formålet med Semantic web), opstille de forskellige scenarier og præsentere en række scenarier for brugeren. Brugeren står altså tilbage med den egentlige opgave 'at besøge mor på hospitalet' og skal ikke bekymre sig om at 'finde vej til hospitalet'.

Et af de problemer, man kan forestille sig i forbindelse med denne form for intelligente agenter, er, at der ofte spiller flere variabler ind, end hvad man kan finde af data på nettet, og at det kan være svært at identificere disse variabler. Eksempelvis kan det være, at jeg principielt synes, det

---

<sup>111</sup> Spivack 2008b

er miljøsvineri at køre i bil, eller at jeg ikke kan lide at køre i bus, fordi jeg ikke kan høre radio og synge med på alle sangene. Det kan også være, at den intelligente agent ikke er i stand til at indhente information om, at Langebro er oppe, og det derfor bedre kan betale sig at køre over Knippelsbro.

Man kan frygte, at denne automatiserede hjælp ender med at gøre os en bjørnetjeneste i den forstand, at hvis vi bliver for afhængige af denne form for automatiserede tjenester, får vi måske sværere ved at klare os uden. På samme måde, som man ikke træner sin hjerne i at regne ved at benytte sig af en lommeregner, eller som nogen frygter, at man forringer sin stedssans ved at bruge GPS-assistance, kan man frygte, at jo mere hjælp vi får til at navigere rundt i vores dagligdag, des mere afhængige af teknologien bliver vi.<sup>112</sup>

Intelligente agenter vil principielt kunne være en realitet, så snart Semantic web er udbredt nok til, at de relevante data er til rådighed. Det vil da blot være et spørgsmål om at definere, hvilke data der er relevante i hvilke situationer (vejkort, afstande, vejarbejde, myldretid, benzinpriser osv. i forbindelse med kørselsvejledninger eller kunstnernavne, albumtitler, genre, spillelængde, medie, pris, told, forsendelse, moms osv. i forbindelse med sammenligning af priser på musik).

### 6.5.2 *MARKEDSUNDERSØGELSER*

I en tid, hvor det bliver mere og mere almindeligt at dele sine meninger med omverdenen på nettet, kan man forestille sig, at de, der er mest interesserede i at få fingrene i semantisk software, er store selskaber, der sælger produkter til forbrugerne, og firmaer, der laver markedsundersøgelser. Med muligheden for at automatisere processen at gennemlæse utallige brugeres kommentarer om sit produkt, bliver det muligt at få ny viden om sine produkter og forbrugernes holdning til dem.<sup>113</sup>

Gennem text mining og natural language processing åbner der sig muligheder for løbende at overvåge store mængder af online medier som f.eks. blogs, sociale netværkstjenester, fora og wikier.

En af de helt store fordele ved dette er, at man undgår de problemer, man normalt møder ved at stille forbrugere konkrete, formulerede spørgsmål, da man meget nemt kommer til at påvirke dem gennem formuleringen af spørgsmålet. En anden fordel er, at det ikke er producenten (eller et analysefirma på vegne af producenten), der sætter rammen for indsamlingen af data. I stedet for at man kontakter forbrugeren, og beder dem 'finde på en mening', er der derimod tale om en mere naturlig reaktion fra forbrugerne, der af egen fri vilje publicerer deres holdning til et produkt på WWW.

Man kan sige, at denne form for undersøgelser afviger fra traditionelle markedsundersøgelser, da online medier kan udtrykke både personlige holdninger som dem, man kan forsøge at indfange ved mere traditionelle metoder som telefoninterview, spørgeskemaer og fokusgrupper, og de kan være meningsskabende. Man kan lidt karikeret sige, at mange af de holdningsudvekslinger, der førhen blev udvekslet over hegnet, er gået online, og det er i langt

---

<sup>112</sup> Denne tankegang stammer helt tilbage fra Cyberpunkens fødsel med McLuhan i 1964 (*Understanding Media: The Extension of Man*), hvor den paradoksale tanke om, at enhver udvidelse af det menneskelige sansesystem (f.eks. gennem teknologiske værktøj) også er en amputation.

<sup>113</sup> Denne type af overvågning foregår allerede i dag, dog kun på nøgleordsbasis.

højere grad end tidligere blevet muligt for enhver interessent at lytte med. Der har været en del eksempler på virksomheder, der har lidt store tab på grund af de sociale mediers evne til at sprede informationer hurtigt.<sup>114</sup> Hvis producenterne har et værktøj, der giver en finger på pulsen, vil de have et værktøj til at rette op på deres produkt eller lancere kommunikation, der kan afklare misforståelser eller gøre indrømmelser, før en sag evt. eskalerer.

Funktionaliteten af sådanne udtræk kunne styrkes gennem en emne-tracking eller document clustering (eller en kombination af disse). Emne-tracking giver producenterne mulighed for at følge netop deres produkt eller konkurrerende produkter, men også firmanavn, eventuelle berømte ansatte eller begivenheder, som sponsoreres af firmaet. Document clustering giver mulighed for at skabe en form for overblik over de indsamlede data og giver også mulighed for at lave kvantificerbare data ud fra det overvågede indhold, eksempelvis hvor mange brugere omtaler vores produkt i en positiv sammenhæng sammenlignet med, hvor mange der omtaler det i en negativ, eller hvor mange anmelder vores produkt, hvor mange bruger det som eksempel på en tendens, og hvor mange nævner det i forbindelse med noget helt andet?

Denne form for markedsundersøgelser virker som en guldgrube for producenterne. Men kan der også være problemer i forholdet til denne form for text mining? Et potentielt problem med at overvåge sociale medier som f.eks. blogs, social networking services og personlige hjemmesider er, at selvom disse mediers brugere formentlig er bevidste om, at enhver har adgang til at se eller læse deres producerede indhold, er det ikke sikkert, at de er bevidste om, at de muligvis bliver systematisk overvåget.

### 6.5.3 PERSONLIGE PROFILER

I science fiction filmen *Minority Report* bliver hovedpersonen identificeret på gaden af skærme, der viser reklamer netop for ham. Reklamerne tiltaler ham ved navn, og der bliver præsenteret markedsføring, der passer på netop ham. Reklamerne ved med andre ord noget om den, de henvender sig til, og reklameselskabet kan eksempelvis have kendskab til hans forbrugsvaner og interesser, og kan bruge denne viden aktivt i markedsføringen.

En af visionerne bag Semantic web-tankegangen er at kunne forbinde informationer og mennesker i større grad end tidligere. I denne forstand vil det være oplagt at behandle individer som en slags begreber, således at det bliver muligt at identificere, at to bøger er skrevet af den samme forfatter og ikke blot af to forskellige personer med samme navn. Det vil kræve en form for samling af offentlige profiler, hvor hvert enkelt individ kan identificeres utvetydigt. En sådan type profil vil være oplagt i forbindelse med nyheder, forfatterskab, ejerskab, copyright, relationer mv. Spørgsmålet er blot, hvem der skal administrere disse. Er det op til det enkelte individ at vedligeholde sin egen profil, eller kan et nyhedsmedie f.eks. tilknytte sin historie til en profil, således at en nyhed om lystfiskeren Bill Gates ikke forveksles med Microsoft-stifteren Bill Gates? En anden mulighed er, at profilen ikke knytter sig til personen, men til personens funktioner. Således kunne lystfiskeren og IT-ikonet være en og samme person i virkeligheden, men på internettet have to forskellige identiteter: en som lystfisker og en som stifteren af Microsoft. Hvis virksomheder kan tilknytte de informationer de har lyst til, til ens profil, kan man frygte at der vil opstå problematiske situationer; eksempelvis at et firma ønsker at påhæfte

---

<sup>114</sup> Et eksempel er den berømte video fra youtube.com, hvor en cykellås fra en af verdens største producenter bliver åbnet på 36 sekunder med en almindelig kuglepen.

informationer en profil, som personen ikke ønsker, bliver knyttet til sig, eksempelvis at man bruger sin fritid på lystfiskeri.

Allerede i dag ser vi denne type problemer i forbindelse med social networking-servicen Facebook. Da Facebook startede, var det muligt for enhver at lægge billeder op af andre mennesker og udpege, hvem der var på billedet. Lagde jeg f.eks. et billede op af min kusine, og havde jeg 'tagget' billedet med hendes navn, ville man kunne klikke på hendes navn og gå direkte til hendes profil. Herudover vil billedet også fremgå på hendes profil. Det kan man for så vidt stadigvæk gøre i dag, men i dag har den pågældende mulighed for at sige 'dette billede er ikke af mig'. Du kan med andre ord stå i en situation, hvor du bliver identificeret på et billede, som du måske ikke ønskede skulle offentliggøres eller knyttes sammen med din profil. Principielt er der ikke noget i vejen for, at andre brugere kan lægge billeder ind på Facebook og sige, at de er af dig, selvom billedet forestiller en helt anden.

Skræmmescenariet i forbindelse med en semantisk tolkning af dig som person, der vil være offentligt tilgængelig på WWW, er, at enhver kan knytte en hvilken som helst information til din person, uden at du i princippet har mulighed for at udvælge, hvad der forbindes med din person. Det virker måske umiddelbart usandsynligt, at det nogensinde vil komme så vidt, men tendensen til offentlige profiler og den manglende kontrol over, hvordan man repræsenteres på nettet, er nærværende.

Uanset hvilken metode, man forestiller sig, vil sådanne profiler formentlig kunne bruges til en form for markedsføring, hvor et individs profilering (gennem en eller flere profiler) vil kunne bruges aktivt i målretningen af markedsføring. Om man forestiller sig et scenarie som fra science fiction-filmen *Minority Report*, hvor reklamemedierne hele tiden har indblik i dine indkøbsvaner og markedsfører produkter for dig på åben gade, der måske kan afsløre noget om dine interesser (måske er jeg ikke interesseret i at hele verden får kendskab til min svaghed for salt lakrids), eller om man forestiller sig, at man har en eller flere profiler på internettet, som i princippet er offentligt tilgængelige, og som kan bruges til at lave direct marketing, er i bund og grund to sider af samme sag. Scenariet er, at der løbende skabes en profil over ens eksistens, som for omverdenen skaber en *mening* om, hvem man er.

Tanken om at forbinde mennesker og data i langt højere grad virker på overfladen som en god idé, fordi det kan sikre, at vi ved, hvem der menes, når der skrives om en specifik person. Men tanken om at have en unik, offentligt tilgængelig profil, der afslører personlige detaljer om ens liv, virker derimod mindre tillokkende.

#### 6.5.4 *BORGERINDDRAGELSE*

Jeg indledte denne rapport med at beskrive, hvordan World Wide Web i dag udgør et overflødhedshorn af holdninger, meninger og information udarbejdet af brugere. Med adgang til alle disse data og mulighed for at overskue dem, bliver der nye muligheder for at undersøge borgernes idéer og holdninger til politiske emner. Med teknologier som NLP, text mining, emne-tracking, udtræk af information, referering og document clustering bliver det muligt at identificere og overskue, hvad givne tekster handler om. På længere sigt vil det måske endda blive muligt at analysere, hvorvidt teksterne er positivt eller negativt stemt over for det emne, de beskriver, eller lokalisere konkrete forslag til forbedringer.

Bliver det muligt at undersøge tekster på WWW for holdninger og meninger i forbindelse med politiske emner, vil det være oplagt at undersøge mulighederne for at bruge disse data som indspark eller inspiration i det politiske arbejde, hvad end det er på et kommunalt, nationalt eller globalt niveau. Det interessante ved denne form for borgerinddragelse er, at den, ligesom jeg nævnte i afsnittet om markedsanalyser, ikke er en kommunikationsform, hvor nogen går ud og spørger folket eller sætter rammer for hvad der kan diskuteres. Omvendt er det brugerne (her borgerne) der selv kommer på banen med de emner, der optager dem. Det vil sige, at de borgere, der har en holdning om et særligt emne, vil være dem, der sætter dagsordenen for den politiske debat. Det er så at sige deres ytringer på eget initiativ. Dette står i modsætning til f.eks. demokratiske afstemninger, hvor alle stemmeberettigede borgere bliver bedt om at stemme på specifikke forslag.

Denne tankegang åbner op for en helt ny måde at tænke demokratisk på. En udtømmende diskussion af, hvordan denne form for inddragelse af borgere kan og bør bruges, vil du ikke finde i denne rapport. De semantiske teknologier er endnu ikke modne til denne form for indholdsanalyse, og det vil ikke give god mening at diskutere emnet i detaljen, før vi kender de teknologier, der ligger til grund for tankegangen. Ikke desto mindre vil jeg i det følgende beskrive en række af de problemstillinger, man bør være opmærksomme på, når teknologierne har udviklet sig til at være brugbare til denne form for dataudtræk i den skala, der her er tale om.

#### **6.5.4.1 Skriver brugere hvad de mener?**

Det første problem, jeg vil nævne i denne sammenhæng, er ikke af teknologisk karakter, men nærmere et filosofisk spørgsmål. Ved folk, hvad de skriver? Dette skal naturligvis ikke forstås som et spørgsmål om, hvorvidt folk ikke ved, hvordan man bruger det skriftlige sprog, eller at de skriver uden at tænke sig om. Nærmere handler det om, hvorvidt sproget og vores eksplicite viden i realiteten dækker over, hvad vi mener og føler. Hvis jeg skriver, at 'bolden er blå', og modtageren ved, hvilken bold der er tale om, vil det være forventeligt, at modtageren forstår budskabet inden for en vis margin (vi antager at, der enighed om, hvilken bold, der er tale om, og at det ikke er relevant, hvilken nuance af blå der menes). Skriver jeg derimod 'det føles af december', skal der en del mere viden til at tolke, hvorvidt jeg mener 'mørk som december måned', 'julestemning' eller 'eksamensstress'. Det kræver kendskab til, hvem *jeg* er, og hvad jeg normalt forbinder med december. Mit sprog er ganske enkelt ikke nuanceret nok (eller jeg har ikke de rette evner) til at beskrive den følelse, jeg har. Derfor tyr jeg til metaforer, som jeg antager, at den, jeg taler til, forstår i den kontekst, vi nu engang er i.

Mennesker er meget gode til at tolke meninger til trods for sproglige fejl. Enhver, der har forsøgt at transskribere en samtale, vil kunne skrive under på, at det er de færreste mennesker, der er i stand til at tale fuldstændig grammatisk korrekt, og som aldrig bruger forkerte ord om det, der tales om. Men vi bemærker det ikke til hverdag, fordi vi er gode til at abstrahere fra ligegyldige fejl, og vi kan udfylde hullerne, når folk bruger de forkerte ord. Spørgsmålet er, hvorvidt menneskers sprog i virkeligheden kun i ringe grad er i stand til at beskrive deres holdninger og meninger på en meningsfuld måde, som kan tolkes entydigt gennem maskinel aflæsning.

Diskussionen er af meget filosofisk karakter og handler i sidste ende om, hvorvidt vi mennesker sætter vores idéer i tale, og om hvorvidt vores sprog er et komplekst nok værktøj til at udtrykke

vores følelser, meninger og idéer. Under alle omstændigheder bør det vurderes, hvorvidt en maskinel aflæsning af brugergenereret indhold kan antages at præsentere borgernes egentlige holdninger og meninger, og hvorvidt en tolkning af disse kan bruges direkte i politiske beslutningsprocesser.

#### **6.5.4.2 Mener brugere hvad de skriver?**

Kan man virkelig regne med, at folk mener det, de skriver? Eller formuleret på en anden måde: Kan det tænkes, at brugere på nettet kan finde på at argumentere for noget, de ikke nødvendigvis selv står inde for? Her tænker jeg ikke blot på, at folk bliver inspireret til at diskutere emner, de ikke ville have diskuteret, hvis de ikke var blevet eksponeret i medierne, men mere på provokation/reaktions-forhold. F.eks. kan det tænkes, at nogle brugere påtager sig rollen som djævlens advokat for at nuancere eller udforske et argument – dette betyder ikke nødvendigvis, at de selv mener, at den argumentation, de fører, støtter den ideologi, de ønsker at følge. Der kan med andre ord være en mangel på sammenhæng mellem, hvad folk skriver, de mener, og hvad de faktisk mener.

#### **6.5.4.3 Ved brugere hvad de mener?**

Det sidste problem vedrører for så vidt ikke kun borgerinddragelse på WWW, men derimod den demokratiske styreform. Som jeg nævnte i afsnit 3.2.4 om Wisdom of crowds, er det ikke oplagt at anskue det demokratiske system som løsning på hverken kognitions-, kooperations- eller koordinationsproblemer. Problemet er, at det demokratiske system ikke tager udgangspunkt i at bruge folket til at finde den bedst mulige løsning, men spørger hver enkelt person om deres specifikke interesse. I et repræsentativt demokrati er der indført endnu en instans, nemlig de Folketingsvalgte politikere, som individerne stemmer ind i Folketinget for at varetage ens interesser eller ideologier. På den måde gives der bedre plads til at holde en politisk ideologi, idet det er udvalgte politikere, og ikke hvert enkelt individ, der fører politik.

Man bør derfor overveje, hvorvidt man vil benytte sig af borgerinddragelse i forlængelse af den nuværende demokratiske form, hvor borgerne afgiver deres stemmer ved at generere indhold baseret på deres ønsker, eller om man bør udnytte teknologien til at lade semantic software agere som en slags aggregator, der muliggør løsning af kognitions- eller kooperationsproblemer. Wisdom of crowds har vist sig at være nyttigt til at løse en lang række af problemtyper – de har dog alle det til fælles, at de involverede ikke handler ud fra spekulation om eget bedste, men om at løse et konkret problem. For at undgå tragedy of the commons (se afsnit 3.2.5), hvor individer udnytter fællesskabets sårbarheder for personlig vinding, bør man overveje, hvorledes borgerinddragelse anvendes politisk.

#### **6.5.4.4 To modeller for borgerinddragelse med WWW**

I ovenstående optik er der to oplagte modeller for at udnytte internettet i forbindelse med borgerinddragelse.

Den ene model tager udgangspunkt i, at der allerede findes enorme mængder af data genereret af brugere, og målet er at lokalisere det indhold, der giver udtryk for politiske holdninger, eller som tager stilling til politiske diskussioner, at gøre dem overskuelige og bruge disse som inspiration til politiske diskussioner. Der er altså ikke tale om en 'de fleste brugere på nettet mener at...', men nærmere en 'har vi overvejset dette synspunkt i vores diskussion af dette

emne?'. Her fungerer de eksisterende data som en katalysator, der kan fordre nye diskussioner eller nye måder at behandle eksisterende problemstillinger på. Denne tilgang til data giver rig mulighed for, at der eksperimenteres med præsentation af data, og at sammensætte data (mash-up) på nye måder. Kan det f.eks. tænkes, at der er en sammenhæng i diskussionerne mellem krig i Mellemøsten og miljøpolitik i Danmark? Eller hvordan er forholdet mellem diskussion af trafikssikkerhed og forskningsmidler til universiteterne? Gennem document clustering kan man samordne artikler efter forskellige kriterier og på den måde opdage nye sammenhænge.

Den anden model vil ikke i samme grad afhænge af eksisterende indhold, men vil være en måde at udnytte Wisdom of crowds til at løse kognitionsproblemer. Den konkrete fremgangsmetode skulle her være at lave et forum for diskussioner om specifikke politiske problematikker. WWW kan her registrere alle indlæg og skabe en historik. Desværre udgør et forum alene ikke en særlig god aggregator, idet indholdet blot lægger statisk i en database. Det er her, semantic software kommer ind i billedet – gennem referering og udtræk af information vil det ikke længere være uoverskueligt at skulle finde ud af hvad tusindvis af brugere kollektivt mener om et emne. Her er der tale om en langt højere grad af udnyttelse af Wisdom of crowds end det forrige scenarie, idet masserne her arbejder specifikt på at løse konkrete problemer. Udfordringerne her er naturligvis at få folk til at deltage, og i at få de rette vækstvilkår for Wisdom of crowds (diversity, independence og decentrelization). Problemer, der kan opstå, kunne f.eks. være, at brugerne ligner hinanden for meget (f.eks. ved at bo i samme by, have samme uddannelse eller politiske orientering) eller at det primært er personer, der i forvejen er politisk engagerede, der deltager i diskussionerne.

## 6.6 "METASEMANTIC" SOFTWARE

Ud over de ovenstående eksempler på, hvordan semantic software kan ændre den måde, vi bruger internettet på, har semantic software også potentiale til at gøre anden semantic software og kunstig intelligens endnu bedre. Jeg har tidligere i denne rapport stillet mig kritisk over for semantic webs brug af ontologier med definitioner af begreber. Denne kritik bunder i definitioners manglende fleksibilitet og evne til at tilpasse sig det konstant udviklende naturlige sprog.

Ved at udnytte de forskellige semantic software-teknologier sammen, er der dog mulighed for, at der kan skabes en form for synergi, hvor den ene teknologi hjælper den anden. Man kan forestille sig at man benytter sig af en ontologi som udgangspunkt (som i Semantic web), men at man benytter NLP og kunstig intelligens til løbende at opdatere denne ontologi. Altså en form for dynamisk ontologi. Dette er for så vidt en god idé, men et system, der kan tilpasse sin forståelse af tekster ved at læse tekster, kan måske læres forkerte betydninger af begreber, eller den kan få forkert viden om grammatik. I den virkelige verden reagerer vi måske, når vi hører nogen bruge ordet *bjørnetjeneste* om en stor tjeneste. Men hvis tilstrækkeligt mange mennesker bruger begrebet i den betydning på nettet, vil det være den betydning, der er rigtigt for den semantiske software. Der er med andre ord mulighed for, at sproget udvikler sig endnu hurtigere eller endda bliver flygtigt

Den helt klare fordel ved den dynamiske ontologi er naturligvis, at den ikke medfører de samme problemer som den klassiske, definatoriske ontologi. Det vil ikke længere være et problem, at et bilsæde er en slags stol, men ikke en slags møbel, til trods for at stole er en slags møbler. Det vil

heller ikke være noget problem, når der kommer nye ord ind i sproget, eller når eksisterende ord skifter betydning, udtryk eller stavemåde.

Omvendt er der en farer for, at en semantisk analyse af en tekst ændres i takt med at ontologien udvikles.

## 6.7 AGGREGATOR FOR WISDOM OF CROWDS

Web 2.0 har åbnet for en verden af brugergeneret indhold, og begreber som Wisdom of crowds og global intelligens dukker op i takt med, at mennesker bliver bedre til at samarbejde på tværs af tid og rum. Wikipedia bliver ofte brugt som det gode eksempel på, hvordan mennesker kan opnå utrolige resultater ved at løse store opgaver sammen. En af de store udfordringer, vi står overfor i dag, er at udnytte de store mængder af indhold, der er repræsenteret på nettet. Blot at lave nøgleordssøgninger slår ikke til. Dels fordi nøgleordssøgninger begrænser søgeresultater til hits baseret på enkelte ord, og ikke på disse ords mening. *Hund* og *vovse* har i den forstand lige så meget til fælles som *skruetrækker* og *badminton*. For at kunne dykke ned i de store mængder af data og behandle dem som sådanne, må de på en eller anden måde aggregeres.

Semantic software har potentiale som aggregator for Wisdom of crowds. Det skal forstås således, at Wisdom of crowds, som nævnt i afsnit 3.2.3, afhænger af diversitet, decentralisering og uafhængighed samt aggregering. De første tre vilkår angår forholdet mellem individerne i en mængde af mennesker, hvor aggregering i nogle tilfælde er noget, der kommer af sig selv, dog typisk inden for koordinations- og kooperationsproblemer (f.eks. at mennesker i mange kulturer selv danner kø i butikker og lignende, eller at vi ikke støder ind i hinanden, når vi går på en befærdet gågade). Når det gælder kognitionsproblemer, dvs. problemer, hvor menneskemængder sammen skal løse et problem, der kræver refleksion, eller finde svar på et spørgsmål, sker aggregeringen ikke automatisk. Der kræves en form for system. Jeg vil endnu engang hive eksemplet med dyrskuedeltagerne, der skulle gætte vægten på en okse, frem (se afsnit 0). Her var det gennem aggregeringen, dvs. indsamlingen af deltageres svar og udregningen af gennemsnittet, der kunne oversætte de mange deltageres bud til et kollektivt bud på oksens vægt.

Hvis vi i højere grad end i dag får mulighed for at overskue de enorme mængder af brugergeneret indhold på WWW, og bliver vi i stand til at måle trends, tendenser eller til at gruppere, analysere og sammenstille data skabt af mennesker, kan vi måske udnytte den globale hjerne i en grad, vi ikke kan forestille os i dag. Det kan lyde utopisk, at vi med et slag kan få indblik i web-brugernes idealer, drømme og håb, men teknologien er faktisk lige om hjørnet.

Den store udfordring bliver at udnytte den.

## 6.8 MANGEL PÅ OVERBLIK

Hvad end det gælder borgerinddragelse, markedsundersøgelser eller noget helt tredje, er en af de store udfordringer i forbindelse med disse enorme mængder af data at kunne overskue disse data. I mange tilfælde vil man måske gennem software kunne producere præcis de produkter, man ønsker sig af dataene, som f.eks. en forbedret informationsarkitektur (som amazon.coms brug af collaborative filtering). Andre gange vil mennesker gerne selv have indblik i data og have muligheden for at udforske dem på egen hånd, det kunne f.eks. være i søgen efter forbindelser

mellem politiske debatter. I sådanne tilfælde vil der være et behov for at kunne overskue disse data.

I denne forbindelse er det min påstand, at visualiseringer kan komme til at spille meget stor rolle. Visualiseringer giver mulighed for at aflæse store mængder af data på forholdsvis kort tid. Og veludvalgte visualiseringer giver mulighed for at gennemskue både helheden og detaljen og giver mulighed for at lægge en tolkning over data, således at man hurtigt kan aflæse den viden, der er relevant for brugeren.

I det følgende kapitel vil jeg præsentere forskellige synspunkter for design af visualiseringer, dels med teoretiske og praktiske eksempler. Til sidst vil jeg komme med et bud på, hvordan man kan forestille sig visualiseringer brugt i forbindelse med semantic software.

## 7 VISUALISERING

### 7.1 HVAD ER EN VISUALISERING

Når jeg i denne rapport bruger ordet 'visualisering', er det ment som en form for kondenseret, visuel repræsentation af information. Et landkort er i den forstand en visualisering af, hvordan landet ser ud (set fra oven). Landkortet bygger på data om afstande, højder, vegetation, veje, kysternes form og meget mere. Landkortet giver hurtigt brugeren indblik i en masse data, som ville være uoverskuelige i skrevne eller matematiske beskrivelser.

FIGUR 13 – ET LANDKORT UDEN NATURLIGE TEGN, SOM F.EKS. SKOVE, VEJE OG HØJDEFORSKELLE  
INDTEGNET.

På samme måde er grafer visualiseringer, til trods for at de *ikke ligner*. Grafen giver os også mulighed for at overskue f.eks. udviklingen af et målepunkt (f.eks. en aktiekurs), eller hvordan noget fordeles på forskellige områder (f.eks. et cirkeldiagram). Grafen har den egenskab, ligesom mange andre former for visualiseringer, at den er god til at give et overblik og fortælle den rigtige historie. Aktiekursen fortæller os ikke, hvorvidt der er sket et fald på netop 0,3

procentpoint, men den viser os, at der er sket et lille fald. Visualiseringen er altså sjældent god til at fremvise præcisionen i data, men den fortæller os, det vi har brug for at vide, f.eks. for at kunne handle på det – om vi skal købe aktien eller ej.



FIGUR 14 – EN AKTIEKURS.

## 7.2 VISUALISERING AF INFORMATION

Edward Rold Tufte er kendt for at være en visualiseringens 'grand old man'. Han er professor på Yale University i bl.a. statistik, informationsdesign og design af brugergrænseflader. I hans bog om visualiseringer, *Envisioning information*<sup>115</sup>, er hans budskab, at det aldrig er data eller information, der er rodet, blot måden hvorpå de vises.

*Confusion and cluttering are failures of design, not attributes of information.*

Tufte 1990, s. 53

Idéen bag hans påstand er naturligvis, at data i sig selv ikke kan være 'rodede', da det er vores måde at præsentere data på, der skaber forståelsen af data. Hvis præsentationen er god, vil vi nemt kunne sætte os ind i data og ikke finde dem rodede – hvis de ikke er, finder vi dem voldsomt uoverskuelige og måske endda ubrugelige. Forestil dig en telefonbog, der er ordnet efter telefonnumrenes numeriske rækkefølge frem for efter navnet bag nummeret. Det ville nærmest udelukkende være brugbart i den situation, hvor vi skal finde ud af, hvem der gemmer sig bag et telefonnummer. I stort set alle andre situationer vil en sådan telefonbog være fuldstændig meningsløs og ubrugelig.

Tufte gennemgår en række grundelementer i visualiseringer, der kan hjælpe designeren af en visualisering med at skabe klarhed uden at gå på kompromis med kvaliteten af data.

---

<sup>115</sup> Tufte 2004

### 7.2.1 MIKRO- OG MAKRODESIGN

Et af de emner, Tufte tager op, er mikro- og makrodesign og forholdet mellem dem. Særligt når der er tale om meget komplekse visualiseringer, der er baseret på mange data, kan det være svært at ordne disse data på en måde, så de både giver god mening både i detaljen og i helheden.

FIGUR 15 – SENLIS SET FRA OVEN (TUFTE 2004, S. 38).

Paradokset, som Tufte behandler, kredser om menneskers evne til at håndtere uanede mængder af kompleksitet i hverdagen og den, ifølge Tufte, fejlagtige opfattelse, at minimalisme gør ting nemmere at forstå. Som eksempel på dette viser han et luftfotografi af den gamle franske by Senlis. Vi har som beskuerer af fotografiet ingen problemer med at afkode billedets helhed, bygninger, træer, veje, og endda bemærker man hurtigt den interessante måde, hvorpå vejene danner cirkler rundt om byens centrum. Tufte beskriver, at fotoet har så mange detaljer, at *"digitizing these images for computers requires  $10^6$  to  $10^8$  bits"*<sup>116</sup>.

---

<sup>116</sup> Tufte 1990, s. 38

FIGUR 16 – ISOMETRISK BYKORT OVER MANHATAN, AF MICHEL ETIENNE OG LOUIS BRETEZ  
(TUFTE 2004, S. 37).

Ovenstående er et veldesignet isometrisk kort over Manhattan, som glimrer ved at have indtegnet utallige detaljer som vinduerne på skyskraberne (man kan faktisk tælle etagerne) og endda træer og markiserne på gadeplanet, såvel som vejnavne og navne på bygninger – uden at blive uoverskuelig. Det interessante ved kortet er, at kortet faktisk ville være ganske anvendeligt som bykort, men samtidig beskriver visuelle indtryk fra det sted kortet portrætter. Og det er måske endda mere overskueligt end et regulært, enkelt bykort. Det, som viser sig at være interessant, er, at det ikke er mængden af detaljer, men muligheden for at skelne de forskellige dele af billedet fra hinanden, der er afgørende for, om vi kan overskue billedet. Det, der falder en i øjnene ved kortet, er, at selvom bygninger og fortove er meget detaljerige, er selve gaderne fuldstændig tomme. Kontrasten mellem detaljerigheden og de rene gader gør det nemt at overskue den overordnede gadeplan.

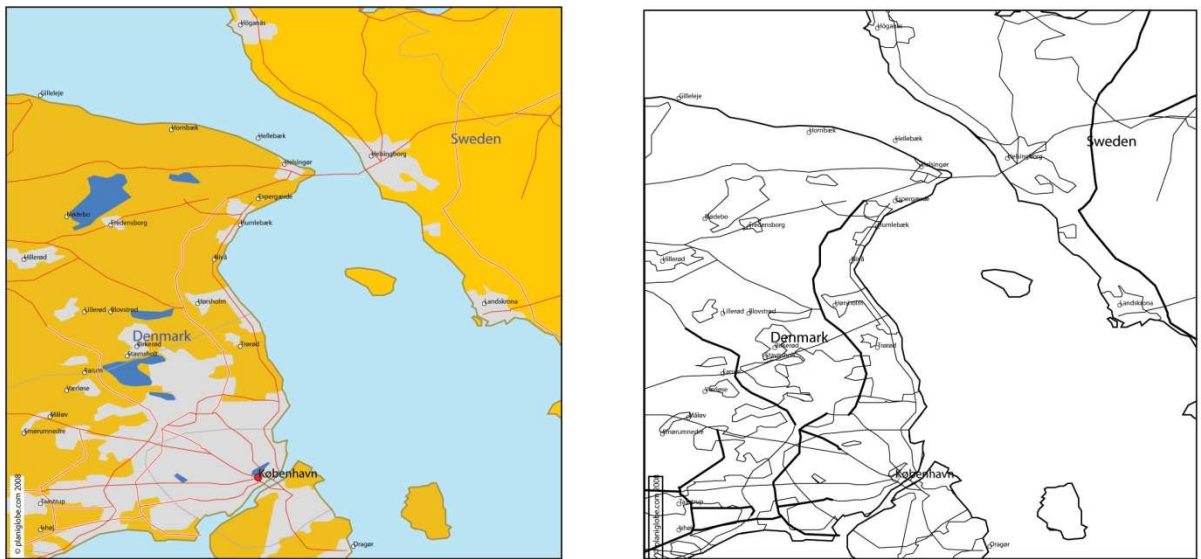
*Simplicity of reading derives from the context of detailed and complex information, properly arranged. A most unconventional design strategy is revealed: to clarify, add detail.*

TUFTE 1990, s. 37

Grundlæggende kan man sige, at for at meget komplicerede visualiseringer fungerer optimalt, skal de både give mening på et mikro- og et makroplan. Som på luftfotografiet over Senlis skal vi både kunne se byen som helhed og den enkelte bygning. Som det vil vise sig senere i dette kapitel, fungerer mange visualiseringer kun på det ene eller det andet plan, hvor resultatet enten bliver, at man ikke kan se skoven for bare træer, eller at man ikke kan se træerne, fordi skoven er i vejen.

### 7.2.2 ADSKILLELSE OG GESTALTER

Ikke meget ulig mikro-/makrotankegangen er det også vigtigt at kunne adskille forskellige datasæt fra hinanden i visualiseringer. Igen viser landkortet sig at være et godt eksempel. Landkort indeholder ofte mange forskellige typer af data, som f.eks. landenes konturer, angivelse af, hvor der er vegetation, hvor der er bjerge, hvor højt eller lavt landskabet er, hvilke byer der ligger hvor, hvordan veje forbinder disse osv. Adskiller man ikke disse lag, får man et ganske uoverskueligt kort, der gør det svært at skille de forskellige datasæt ad. I nedenstående eksempel viser to kort de samme data, men hvor de forskellige typer af data er adskilt gennem brug af farver og kanter på det andet kort, er de på det første kort alle sammen angivet på samme vis, nemlig med en tynd konturstreg.



FIGUR 17 – BEMÆRK HVOR MEGET LETTERE DET ER AT AFKODE KORTET TIL VENSTRE, TAKKET VÆRET ADSKILLELSE GENNEM FARVER..

Tufte bruger også nodepapir som et godt eksempel på, hvordan brug af lag kan benyttes til at gøre noderne mere læsbare. Ved at bruge meget tynde streger til de horisontale streger, hvorpå de tunge, fede noder hviler, er det let at adskille noder fra stregerne på nodepapiret.<sup>117</sup>

I denne sammenhæng er det vigtigt at forstå lidt om, hvordan mennesker opfatter omverdenen. Hvorfor er det nemmere at adskille plamager i forskellige farver end plamager, der blot er defineret af et omrids?

Gestaltpsykologien<sup>118</sup> beskæftiger sig med, hvordan mennesker opfatter omverdenen, ikke blot som sanseindtryk, men som fortolkede indtryk, som vi forstår som entiteter. Jeg ser ikke blot

<sup>117</sup> Tufte 1990, s. 59

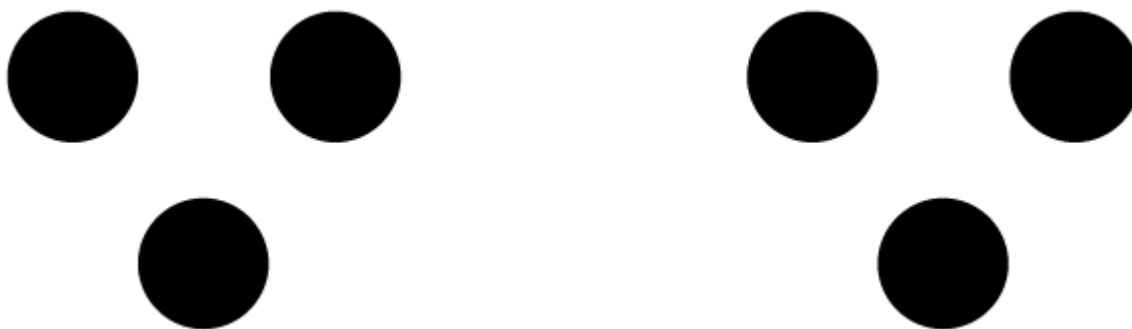
<sup>118</sup> Gestalt betyder egentlig form eller skikkelse. Gestaltpsykologien er en reaktion på behaviorismen, som brød med tanken om at psykologi kun handlede om studie i sind og sjæl og i stedet forsøgte at forklare menneskers reaktioner på stimuli. Behaviorismen er et forsøg på en mere naturvidenskabelig tilgang til de

reflekteret lys, farver og skygger på mit skrivebord; jeg ser min kuglepen, som jeg kan række ud efter og samle op. Jeg forstår den som en ting. Denne forståelse af omverdenen virker måske som en selvfølge, men strengt taget baserer vores opfattelse af omverdenen sig på sanseindtryk fortolket i hjernen. Det er med andre ord vores hjerne, der omdanner sanseindtrykkene til fortolkninger af fysiske objekter.

*[...] når [stimuli] opfattes i dets strenge betydning, er det almindeligvis ikke "en stimulus", der fremkalder en reaktion. Hvad synet f.eks. angår, er organismen tilbøjelig til at reagere på millioner af stimuli; og det første reaktionsstadium er organisation inden for et tilsvarende stort felt.*

KÖHLER 1972, s. 143

Köhler, som (sideløbende med Max Wertheimer og Kurt Koffka) anses for at være en af gestaltpsykologiens fædre, beskriver hvordan mennesker (og for den sags skyld dyr) har en tendens til at opfatte sanseindtryk i grupperinger. Nedenfor vil langt de fleste f.eks. se to grupper på tre prikker, og ikke blot seks prikker.



FIGUR 18 – TRE PRIKKER TO GANGE, ELLER BARE SEKS PRIKKER?.

Og grupperinger binder sig ikke kun til synssansen.

*Hvis jeg med korte mellemrum banker tre gange i bordet og efter et sekunds forløb gentager denne banken osv., vil folk opleve lydrækker i tid. Fysisk set er alle disse lyde naturligvis uafhængige foreteelser. [...] De grupperinger der faktisk høres er derfor eksempler på psykologisk, og ifølge isomorfistiske tese, også fysiologisk organisation.*

KÖHLER 1972, s. 133

---

psykologiske studier. Gestaltpsykologien fortsætter denne tradition, men det anerkendes at mennesket er mere komplekst en stimuli/repsions-tankegangen lægger op til. I stedet anskuer gestaltpsykologien vores møde med omverden som millioner og atter millioner af stimuli, der i hjernen bliver fortolket. Det er denne fortolkning af omverden der ligger til grund for gestaltpsykologiens anskuelse af menneskets psyke. Gestaltpsykologien blev genvundet terræn i forbindelse med design af brugergrænseflader og de såkaldte gestaltlove bruges ofte i forbindelse med design af sådanne grænseflader. (Köhler 1972, kap 1. og 2.)

Med eksempler som, at cirkler stadig opfattes som cirkler, selvom de ses fra vinkler, hvor de strengt taget bliver til ellipser<sup>119</sup>, eller at vi ikke tror, at mennesker, der bevæger sig væk fra os, faktisk bliver mindre og mindre<sup>120</sup>, understreger Köhler, at selvom mennesker afhænger af sanseindtryk, er det ikke selve sanseindtrykkene, men vores fortolkning af dem, der skaber forståelsen af verden omkring os.

Grupperinger er blot en af de effekter, som gestaltpsykologerne fandt. Senere hen har man formuleret de såkaldte gestaltlove, som er en række tommelfingerregler for, hvordan mennesker opfatter visuelle konstellationer. Disse 'love' benyttes især i designet af brugergrænseflader og i webdesign. Der er ikke generel enighed om, hvor mange gestaltlove, der findes. Jeg vil heller ikke gennemgå en liste over dem her, men blandt de typiske regler findes bl.a. reglen om nærhed, som formentlig er udledt af grupperingseffekten, som siger, at elementer, der er tæt på hinanden, opfattes som at høre sammen. Andre klassiske eksempler er reglen om forgrund og baggrund (jf. Rubens vase nedenfor), reglen om lukkethed (jf. figuren med tre cirkeldele og måske en trekant) og reglen om, at figurer, der ligner hinanden, har et samhørighedsforhold.<sup>121</sup>



FIGUR 19 – RUBENS VASE, TIL VENSTRE, BRUGES SOM EKSEMPEL PÅ FORGRUND OG BAGGRUND. TIL HØJRE ER DER TRE FIGURER – MEN SER VI FIRE? REGLEN OM LUKKETHED TEGNER DE USYNLIGE STREGER.

Tufte beskriver i forbindelse med *lag og adskillelse* tendenser, der svarer til de resultater, som gestaltpsykologerne fandt frem til. Tufte understreger vigtigheden i, at man ikke tilføjer elementer til sine designs, som man ikke intentionelt har tilføjet. Det lyder måske som vrøvl, men Tufte anbefaling bygger på hans pointe om, at hver eneste gang, man præsenterer to visuelle elementer, tilføjer man mindst tre: nemlig de to, man har tilføjet, og det forhold, der er mellem disse elementer. Tufte benytter sig af følgende huskeregel: "1+1=3".<sup>122</sup> Et enkelt

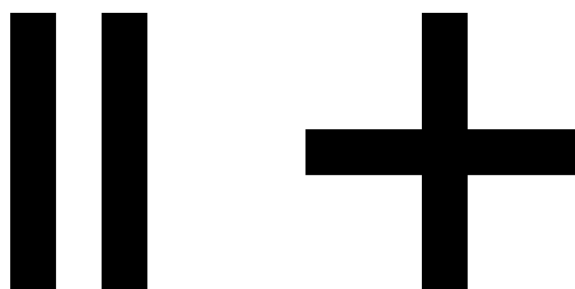
<sup>119</sup> Köhler 1972, s. 78

<sup>120</sup> Köhler 1972, s. 81

<sup>121</sup> Chang et al. 2002

<sup>122</sup> Tufte 1990, s. 61f

eksempel på Tufte's pointe kan vises ved at tegne to parallelle streger (se Figur 20). Nok viser illustrationen de to streger, men kan vi ikke også se en streg mellem de to streger? Og anstrenger vi os, kan vi måske også fornemme den kasse, der omslutter de to eller tre streger? I eksemplet til højre er der igen tale om to streger, men vel også et kors? Og ser vi også et hvidt kvadrat i hvert hjørne? Måske deler stregerne et stort kvadrat på midten? Stort set enhver tegning viser flere elementer end de streger, der udgør tegningen. "*Most of the time, that surplus visual activity is non-information, noise and clutter.*"<sup>123</sup>



FIGUR 20 – HVER GANG VI TILFØJER TO ELEMENTER, TILFØJER VI REELT MINDST TRE.

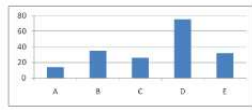
### 7.3 VISUALISERINGENS DIALEKTISKE FORHOLD

Visualiseringer kan siges at have et dialektisk forhold mellem kyniske data og sanselighed. På den ene side er visualiseringer ofte mere eller mindre mekaniske afbildninger af *hårde data*, dvs. data der er fuldstændig kvantificerbare. I denne gruppe finder vi tidtabeller, aktiekurser og statistiske grafikker. På den anden side har vi den anden ekstrem, den kunstneriske afbildning. Asger Jorns malerier kan i en vis grad siges at være visualiseringer af hans sjæleliv eller hans opfattelse af et emne – emner, der er svære at gøre kvantificerbare. Til gengæld besidder de en rigdom i deres udtryk, som visualiseringer fra den anden ende af skalaen har svært ved at hamle op med. Visualiseringen ligger i langt højere grad op til en individuel fortolkning, end de mere kvantificerbare visualiseringer, men også en mere sanselig og måske mere intuitiv oplevelse.

---

<sup>123</sup> Tufte 1990, s. 61

id	navn	årstal	måned
1	Brian	1976	08
2	Mette	1977	11
3	Tine	1979	03
4	Jørgen	1977	07
5	Polle	1981	03
6	Mads	1982	10



Datanær

Udtryksfuld

FIGUR 21 – DATANÆRHEDENS OG UDTRYKSFULDHEDENS DIALEKTISKE FORHOLD. FRA VENSTRE MOD HØJRE SER VI ET SKEMA, ET SØJLEDIAGRAM, ET KUNSTVÆRK BASERET PÅ ANALYSE AF LYDBØLGER OG LÆNGST MOD HØJRE, ET ABSTRAKT MALERI.

Skalaen bevæger sig altså fra den enkle visualisering, der lægger sig tæt op ad de data, den er baseret på og til de mere komplekse visualiseringer, der er af mere subjektiv karakter, som er mangetydige, og som har en detaljerighed. Imellem disse to ekstremer har vi en lang række visualiseringer, der benytter sig af elementer fra begge verdener. Det geografiske kort er et godt eksempel på en visualisering, der er baseret på hårde data (bredde- og længdegrader, afstande, type af land mv.) og som også benytter sig af detaljerigdomme i afbildningen af det, den repræsenterer. Korttegnere kan f.eks. male bølger på vandet, for at vise, at det er vand eller grønne områder for skovområder. Det er på den måde hårde data, der ligger til grund for kortet. Men med hjælp fra de mere detaljerede udtryksformer bliver kortet nemmere at aflæse. Man kunne f.eks. vælge at indtegne skovområder på et isopleth-kort, men dette ville medføre, at kortet ville være sværere at aflæse.

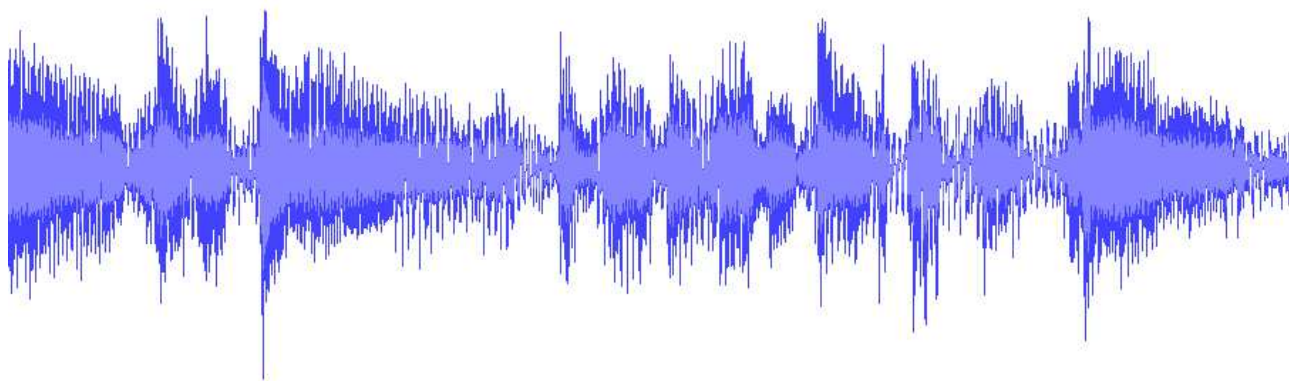
Hvis man anskuer *data* som værende noget utvetydigt og præcist (hvilket i praksis som regel vil betyde, at det skal være kvantificerbart), kan man sige, at visualiseringer til venstre i modellen ovenfor vil kunne repræsentere data i en høj kvalitet. Kvalitet forstås her som, at der ikke bliver indgået kompromis med præsentationen af de originale data, således at der forenkles eller generaliseres. Laver man en god graf over f.eks. en aktiekurs ud fra et skema, der beskriver aktiens udvikling over tid, og beder en anden om at lave den omvendte opgave ud fra denne graf, altså at lave et skema ud fra grafen, vil sidstnævnte skema være identisk med udgangspunktet (såfremt der ikke er lavet nogen fejl). Kvaliteten af data i visualiseringen, dvs. grafen, er høj, da den fuldstændig afspejler dataenes udgangspunkt.

Omvendt kan man sige, at kvaliteten af præsentationen af data vil være tilsvarende lav i den mere udtryksfulde del af spektret, hvor data ikke er det essentielle. Et abstrakt kunstværk eller et pressefoto har sjældent til formål at repræsentere det, jeg kalder *hårde data*, men nærmere at fremprovokere følelser, diskussioner, spørgsmål eller fordre til handling.<sup>124</sup> At visualiseringer i denne del af spektret ikke er gode til at repræsentere hårde data, gør dem ikke

<sup>124</sup> Man kan diskutere hvorvidt abstrakte malerier og pressefotos er lige 'udtryksfulde' i denne forstand. Nok er der flere aflæselige data i et pressefoto end i et abstrakt maleri, men hvad der placerer dem i samme side af spektret er deres funktion: at de har til formål at kommunikere følelser eller holdninger og ikke blot repræsentere data.

nødvendigt til dårlige visualiseringer. Man bør blot være sig bevidst om formålet med ens visualisering, før man vælger sit udtryk; det mere datanære eller det mere udtryksfulde.

Med denne dialektik anskues visualiseringer som enten datanære eller udtryksfulde. Det er klart, at langt fra alle visualiseringer er enten det ene eller det andet. Stort set enhver visualisering vil have dele fra begge verdener. Ikke desto mindre kan der være meget stor forskel på emner fra samme del af spektret. Kriteriet for reproducerbarhed, jf. eksemplet med skemaet der laves til en graf og tilbage til et identisk skema, kræver f.eks. andet end blot en høj kvalitet af præsentationen af data. Aktiekursen, som blev brugt som eksempel, glimrer ved sin enkelhed. Enhver, der forstår grundprincipperne i en graf, der bevæger sig over tid i den første akse og en variabel værdi i den anden, vil med lethed kunne aflæse værdierne for aktien uden større kendskab til børsmarkedet eller aktieværdier. Et andet eksempel på en datanær visualisering, der derimod ikke glimrer i sin enkelhed, er visualiseringen af lydbølger. Til trods for, at modellen bygger på samme grundelementer som aktiekursen (to akser, en for tid og en for en anden variabel), er lydbølgerne, der som i nedenstående tilfælde er skabt af en digitalisering af en lydoptagelse, ikke mulige at afkode. Med en smule viden om lydbølger kan vi vurdere, hvorvidt et område har en højere volumen end et andet. Men vi kan ikke reproducere lyden blot ved at kigge på visualiseringen eller på en anden vis reproducere de data, der ligger til grund for visualiseringen. Det er naturligvis dels, fordi lydbølgeanalyse er væsentlig mere kompleks end analysen af en aktiekurs, og dels fordi grundprincipperne for en akties pris er nemmere at sætte sig ind i end lydbølgernes episoder, frekvens, volumen mv.



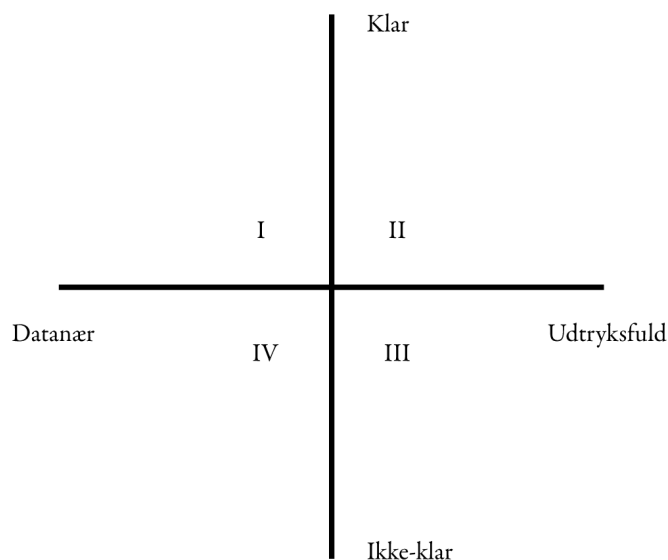
FIGUR 22 – LYDBØLGER SOM REPRÆSENTERET I ET STYKKE LYDREDIGERINGS SOFTWARE.

Jeg vil kalde dette begreb klarhed og indfører endnu en dialektik til vores model, der behandler forholdet mellem klar og ikke-klar<sup>125</sup>. Den tolkning, der tillægges en visualisering, vil dog altid være individuel. Det er derfor ikke et spørgsmål om, hvorvidt forskellige mennesker vil vurdere

---

<sup>125</sup> Jeg vælger ikke at benytte termen uklar, da det antyder en fejlagtighed eller mislykkethed. Manglen på klarhed kan være et bevidst valg, og bør opfattes på lige fod med klarhed og ikke forbindes med noget negativt.

om de skal købe eller sælge deres aktier efter at have set aktiekursen, eller hvad de synes om det, der bliver sagt på den lydoptagelse, der repræsenteres som lydbølger.



FIGUR 23 - VISUALISERINGSMODELLEN.

Et andet eksempel på klar/ikke-klar-dialektikken kunne tage udgangspunkt i de udtryksfulde visualiseringer. Et abstrakt maleri har en ikke-klarhed over sig, der i praksis betyder, at forskellige mennesker vil danne sig forskellige forståelser af værket. Et fotografi, som jo udmærket kan være udtryksfuldt, er dog mere klart i den forstand, at det kan antages, at de fleste mennesker vil have den samme tolkning af, *hvad der er på billedet*. Som med aktiekursen beskriver denne model ikke, hvorvidt folk vil få de samme følelsesmæssige reaktioner eller handle på samme måde, når de oplever visualiseringen, men blot gennemskueligheden og karakteren af visualiseringen.

Hermed har vi skabt en model til at beskrive visualiseringer ud fra to kriterier, nemlig deres datanærhed kontra deres udtryksfuldhed samt hvorvidt de er klare eller ikke klare. Jeg har valgt at kalde modellen for visualiseringsmodellen. Visualiseringsmodellen inddeles i fire kvadranter. I den **første kvadrant** finder vi visualiseringer, der ligger meget tæt på data, og som kan karakteriseres som klare. I denne kvadrant finder vi typer som skemaer<sup>126</sup> og enkle grafer. Enkle geografiske kort vil også kunne findes i denne kvadrant. Den **anden kvadrant** dækker over emner, der er klare, men som har et stærkt udtryk. Det kan være emner som grafer uden enhedsbeskrivelser, et fotorealistic maleri, detaljerige kort med kunstneriske tilføjelser eller de

<sup>126</sup> Skemaer er - til trods for i høj grad at være baseret på tekst - visualiseringer i den forstand at data på en meningsfuld måde i et todimensionelt rum.

så kaldte kartogrammer<sup>127</sup>, som jeg vil vende tilbage til. Den **tredje kvadrant** beskriver visualiseringer, der på en og samme tid er udtryksfulde og ikke-klare. Disse visualiseringer vil i naturens sag sjældent skulle bruges til at aflæse præcise data, men er af mere emotiv eller kunstnerisk karakter. Her kan der f.eks. være tale om farverige visualiseringer af musik, abstrakt maleri eller visualiseringer, der er visuelt dragende, men hvorfra der ikke kan aflæses data. Visualiserede lydbølger, DNA-kort, noder mv. hører til i **fjerde kvadrant**, som er kendetegnet ved at være tæt forbundet med de data, de repræsenterer, men som det kræver særlig viden eller værktøj for at kunne aflæse.

#### 7.4 NATURLIGE OG SYMBOLSKE TEGN

I kapitel 6 nævnte jeg kort semiotikken og Ferdinand de Saussures semiologi. Charles Sanders Peirce anses for at være en af semiotikkens fædre, og han brugte i modsætning til Saussure et meget bredere tegnbegreb, hvor alt i princippet er tegn: Høj temperatur er et tegn på feber, bunker af løv er et tegn på efterår, et symbol på døren til toilettet er et mande- eller kvindetegn, en halvnøgen mand med pile i kroppen er et tegn for den hellige Sebastian osv.

Grundlæggende deler Peirce tegnene i to kategorier, de indeksikalske og de kommunikative. De indeksikalske tegn er tegn *på* noget (spor i sneen, der er tegn på, at nogen har gået der, eller vejrhansens retning, der er et tegn på vindens retning), hvor de kommunikative tegn er tegn *for* noget (ordet *træ* er et tegn for begrebet træ, en vielsesring er et tegn for ægteskab eller et maleri af en hund er et tegn for den hund, der er afmalet). Man kan yderligere opdele de kommunikative tegn i to kategorier, nemlig de symbolske og de ikonografiske. De symbolske tegn er de konventionelle tegn, vi benytter, når vi kommunikerer, f.eks. er vores sprog opbygget af symbolske tegn. Symboler er tegn, hvis mening vi har vedtaget. Omvendt er det med de ikonografiske tegn, som er tegn for noget i kraft af deres lighed med det, de er tegn for. Eksempelvis er en tegning af en ko et tegn for ko, fordi vi kan genkende den som en ko. Der findes dog også mellemting mellem de symbolske og de naturlige tegn. Eksempelvis kan man sige, at de fleste piktogrammer i en vis forstand er afledt af naturlige tegn, men at de har fået en mere symbolsk karakter. Der er eksempelvis forholdsvist langt fra en smiley til et regulært ansigt. Ikke desto mindre forstår vi det som et ansigt og er i stand til at forstå nye smileys som andre ansigtsudtryk.

---

<sup>127</sup> Kartogrammer er geografiske kort, hvor de geografiske egenskaber forvrænges, således at størrelsen på de aftegnede lande illustrerer data. Du kan finde eksempler på dette på <http://www.worldmapper.org/>.



:)	:-)	;o)	>:^)	c",
----	-----	-----	------	-----

FIGUR 24 - EN SMILENDE DRENG OG EN RÆKKE SMILEYS

Man kan sige, at hvor der ikke er meget tvivl om betydningen bag symbolske tegn (gennem deres karakter af at være vedtagne) og omvendt med de naturlige tegn: de er tegn på, hvad end modtageren synes, de ligner. Et dollartegn eller ordet *bil* vil til en vis grænse forstås mere eller mindre entydigt blandt de fleste mennesker, mens følelsen, der bliver fremprovokeret af at se et udtryksfuldt maleri, kan antages at variere fra person til person.

I forhold til skalaen mellem de datanære og de udtryksfulde visualiseringer vil vi ikke finde nogen naturlige tegn blandt de datanære, men derimod vil de symbolske tegn være dominerende. Blandt de mere udtryksfulde visualiseringer, særligt i anden kvadrant, vil vi derimod se flere naturlige tegn, som f.eks. de grønne afbildninger af skove og tegnede bølger på de geografiske landkort.

Hvorvidt man skal lave visualiseringer af den ene eller anden slags, datanære eller udtryksfulde, må være en diskussion, der vedrører designet af hver enkel visualisering – og ikke en generel diskussion. Men man bør være bevist om, at bevæger man sig for langt i den udtryksfulde retning, så medfører det mangetydighed og subjektiv fortolkning, og bevæger man sig for langt i den anden retning, udnytter man ikke mange af visualiseringernes fordele; data forbliver uigennemskuelige og uoverskuelige. Når det gælder noget som f.eks. aktiekurser, er nærheden til data vigtig. Det er ikke nok for investoren at vide, om aktien er gået op eller ned samlet set i løbet af det sidste år; der kræves flere detaljer, f.eks. hvor meget den svinger, hvor meget den er steget eller faldet inden for de sidste timer osv. Omvendt er det med angivelsen af skovområder på landkortet. Det vigtigste er at vide at 'her ligger en skov', og ikke hvilken farve, bladene har, eller hvilke typer træer, der findes i denne skov. Men ved at angive skovområder som grønne (og

måske med antydningen af trækrønernes konturer), gør man det let at se, at der er tale om et skovområde og ikke en ørken eller by.

Man bør i designet af visualiseringer derfor overveje, hvor 'symbolske' eller 'naturlige' tegn det er optimalt at bruge i den pågældende situation. Man må finde den gyldne middelvej, der er relevant for den pågældende opgave.

## 7.5 EKSEMPLER PÅ VISUALISERINGER

Uanset om man laver visualisering baseret på semantic software, om man laver visualiseringer til tryk eller til interaktivt brug, bør man overveje, hvordan man bruger de udtryksmidler, man har til rådighed. Jeg har i dette kapitel introduceret en model til analyse af visualiseringer, som jeg følgende vil benytte, mens jeg gennemgår en række eksempler på visualiseringer. Jeg vil i min analyse af disse eksempler tage udgangspunkt i deres umiddelbare funktioner som visualiseringer, dvs. hvad udtrykker visualiseringen – såvel æstetisk som funktionelt.

Jeg har udvalgt nogle visualiseringer, der på en eller anden måde er baseret på en semantisk analyse, eller som leger med de semantiske tolkninger.

### 7.5.1 CHRIS JORDANS KUNSTVÆRKER

Chris Jordan arbejder ikke med visualisering i en traditionel forstand. Hans visualiseringer er ikke skabt for at præsentere data – de har ikke en praktisk funktion. Derimod er de skabt som individuelle stykker kommunikation, der fortæller noget om samfund og om forholdet mellem stort og småt. Chris Jordan er kunstner. Hans værker illustrerer tal, og ud af de visualiseringer, der her vil blive gennemgået, udgør visualiseringerne de mindst 'semantiske', i den forstand at de ikke baseres på en konkret semantisk analyse, men derimod leger med den semantik, der ligger i fortolkningen af tal.

1.140.000. En-million-et-hundrede-fyrre-tusinde.  $1,14 \times 10^6$ . Uanset hvordan man skriver tallet, er det svært at forstå, hvor mange det egentlig er. Tallet henviser ifølge Jordan til antallet af brune papirposer, der bliver brugt i amerikanske supermarkeder hver eneste time. For at forklare, hvor mange papirposer det i virkeligheden er, vælger Jordan at vise tilskuerne det i værket *Paper bags*.<sup>128</sup>

---

<sup>128</sup> Jordan 2007

FIGUR 25 –PAPER BAGS (CHIRS JORDAN 2007)

Når man som tilskuer ser de store mængder af papirposer, begynder det at gå op for en, hvor utroligt mange, der i virkeligheden er tale om. Selvom vores indtryk af, hvor mange poser der er på billedet, måske ikke er lige så nøjagtigt som det skrevne tal, forstår vi bedre mængden, når vi ser den visualiseret.

FIGUR 26 -PAPER BAGS, DETALJE (CHIRS JORDAN 2007)

På mikroplanet kan vi identificere hver enkelt papirpose – men vi kan ikke tælle dem. På makroplan forsvinder den enkelte papirpose, som forvandles til et tyndt lag i, hvad der virker som uendeligt høje bunker. Bunkernes slanke stigning kan måske lede tankerne hen på træstammer, som i øvrigt er kilden til det materiale, som papirposen er lavet af. Og det er netop her, at Jordans værker bliver særligt interessante. For selvom der reelt blot vises et enormt antal papirposer, styres vores tanker hen mod det enorme forbrug i den vestlige verden og den byrde, der lægges på naturen. Der er over 50 'træstammer' på billedet, svarende til en lille skov. Som tilskuer behøver jeg ingen viden om, hvor mange papirposer der kan laves af et enkelt træ. For uanset om det er en over- eller underdrivelse, er det tydeligt for enhver, at forbruget af papirposer alene udgør en belastning for naturen.

Tanken perspektiveres yderligere, når man overvejer, hvilke diskussioner der har været om, at papirposer er langt mere miljøvenlige end plastikposer. Eller når man tænker på, hvad de varer, som har fyldt alle papirposerne, har betydet for miljøet.

Man kan anskue Jordans værk som at være lidt af en omvendt graf. Han tager udgangspunkt i et enkelt tal, udforsker dette tals betydning og visualiserer det – ikke ved at koge ned, men ved at forstørre. Det omvendte ligger i, at man frem for at vælge den mest minimalistiske og mest overskuelige repræsentation (en graf eller et tal) vælger at vise kompleksiteten. På mikroplan er Jordans værker nærmest ulæselige. Man kan kun se de enkelte papirposer. Men på makroplan forsvinder papirposen og kun billedet af træstammerne står tilbage. Det er altså i skiftet mellem makro- og mikroplan, at hans budskab fremstår.

Men hvordan forholder visualiseringen sig i forhold til visualiseringsmodellen? Kan den placeres konkret i en kvadrant? Jeg stoler på, at Jordan ikke har snydt, og at man, hvis man tæller efter, faktisk vil finde 1,14 millioner papirposer. På mikroplanet er visualiseringen af papirposerne i en forstand meget datanær. Men i kraft af sin enorme kompleksitet, i kraft af at vi ikke kan tælle

de mange poser, er visualiseringen ikke klar. Den befinder sig i den fjerde kvadrant. Omvendt er billedet af træstammerne, visualiseringen på makroplan, blot en slags maleri af en skov med nøgne træstammer. Det er nemt at genkende, dvs. klart, men er ikke umiddelbart forbundet med nogen konkrete data. På makroplan placerer visualiseringen sig altså i anden kvadrant. Det interessante – og meget vellykkede – ved Jordans visualiseringer er, at de som sådan ikke er interessante på hverken mikro- eller makroplan. Men skiftet mellem de to planer er dybt fascinerende og som regel ganske tankevækkende.

### 7.5.2 WE FEEL FINE

*We Feel Fine* er et andet kunstnerisk projekt, som adskiller sig markant fra Chris Jordans værker. Dels fordi det ikke er et statisk værk, men et stykke software der løbende indsamler ny viden til sine visualiseringer, og dels fordi der er tale om en interaktiv visualisering. Jonathan Harris er manden bag softwaren, og i en optagelse af hans fremlæggelse af projektet på en TED-konference<sup>129</sup> beskriver han selv den grundlæggende funktionalitet i systemet således:

*This is a piece that every two or three minutes scans the worlds newly posted blog entries for occurrences of the phrases 'I feel' or 'I am feeling'.*

HARRIS 2007

Softwaren indsamler disse data, og forsøger at finde ud af, hvad tid blogindlægget blev offentliggjort, i hvilken by og nogle informationer om den person, der har offentliggjort indlægget, såsom køn og alder. Denne indsamling af information er dog kun den ene del af softwaren. Den anden del af softwaren er præsentationen af de data, som systemet har fundet frem til.

I korte træk kan man sige at *We Feel Fine* er et forholdsvist enkelt stykke semantisk software, der finder særlige nøgleord ('feel', 'feeling') og gemmer disse data sammen med nogle metadata. *We Feel Fine* kan siges at være et tidligt eksempel på visualisering af WWW-brugernes tanker. Præsenteret i en visualisering, der ægger til udforskning af hver enkelt sætning, sættes der fokus på værdien af hver enkelt individs verden. Ved at have valgt *følelser* som udgangspunkt skabes der et udtryk, der tager bloggeren seriøst. Den kaotiske, store mængde af følelser opsamlet af systemet minder brugeren af *We Feel Fine* om hvor mange individer, der lever og har følelser i verden. *We Feel Fine* kan næppe kaldes *Wisdom of crowds*, da der ikke adresseres et konkret problem. Alligevel trækker systemet på de følelser og holdninger, der kommer til udtryk af tusindvis af brugere, og repræsenterer i den forstand en form for kollektivt udtryk.

---

<sup>129</sup> TED står for Technology, Entertainment og Design og er en årlig konference (afholdt i USA - Long Beach, California), hvis formål er at sprede gode idéer og give inspiration til nye idéer.

FIGUR 27 –SCREENDUMP FRA WE FEEL FINE, MADNESS (HARRIS 2007)

Præsentationen af de data, der er indsamlet af We Feel Fine, er som sagt ikke statisk, men adskillige levende billeder. Den første visualisering, man mødes med, er et kaos af små prikker i varierende størrelse og farve, der virrer rundt på en sort flade. De små prikker reager på musen og har tendens til at samle sig om markøren. Ellers bevæger de sig blot kaotisk rundt. Hver enkelt prik repræsenterer en følelse, og klikker man på en prik, kan man se den sætning, som følelsen indgår i.

*The color of each particle corresponds to the type of feeling inside. So that happy, positive feelings are brightly colored and sad negative feelings are darkly colored. The diameter of each dot represents the length of the sentence.*

HARRIS 2007

Mens systemet loader, kommer der en besked om, at systemet er ved at hente følelser fra de sidste par timer. Vi får altså at vide, at de følelser, vi kigger på lige nu, er fra den nærmeste fortid – næsten synkrone. På den måde kan man sige, at visualiseringen forsøger at give et overblik over, hvilke følelser der lige nu eksisterer blandt engelsktalende bloggere.

Ser vi isoleret på visualiseringens mikroplan, kan vi ved at klikke på hver enkelt partikel få adgang til detaljerede informationer om hver følelse, såsom hvilken følelse der er registreret,

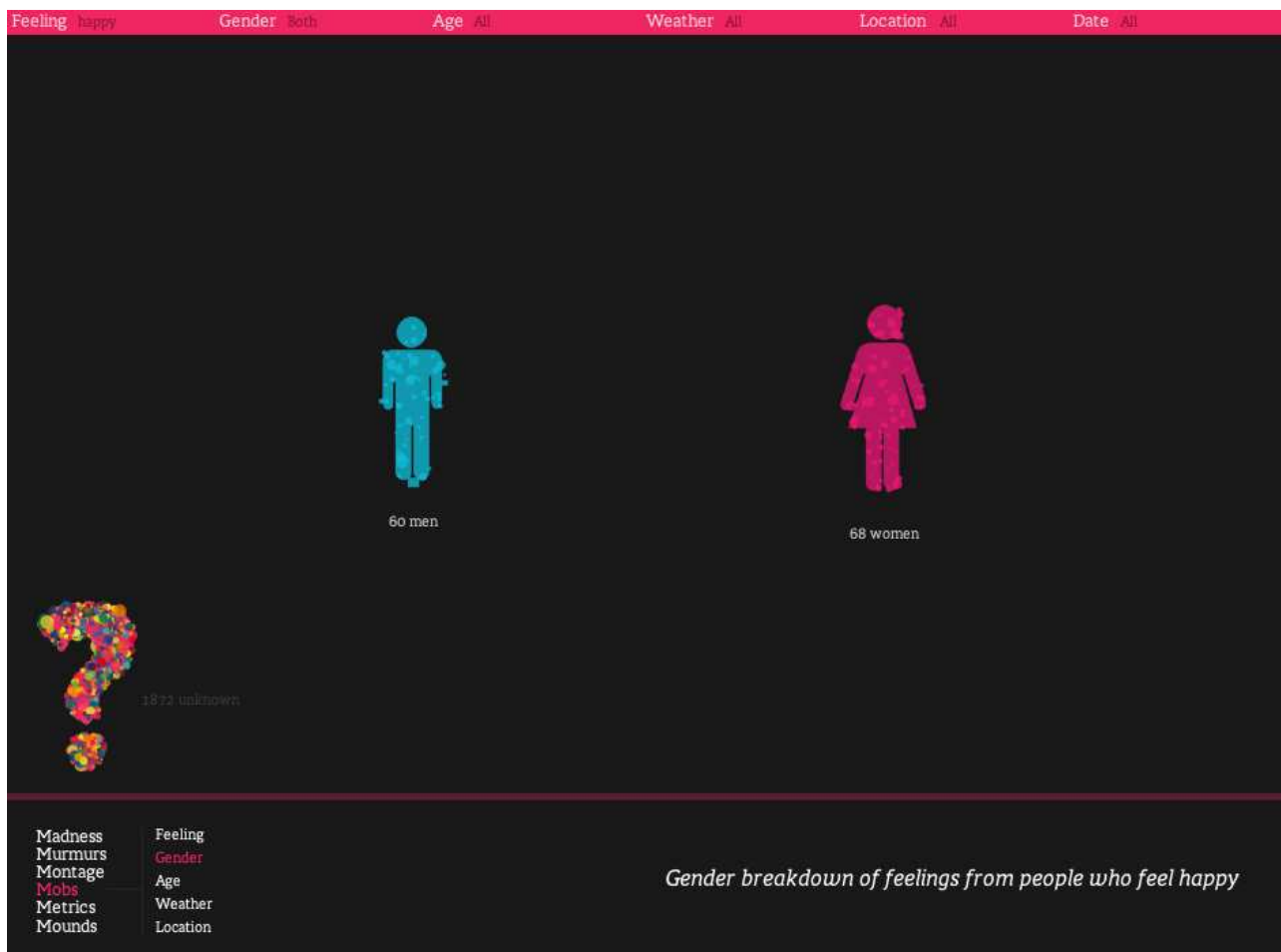
hvilken sætning den indgår i, og hvornår den er registreret. Vi kan endda få adgang til den blog, hvorfra sætningen stammer.

På makroplanet kan vi i princippet få et overblik over, hvor mange følelser der er registreret, hvor mange positive i forhold til hvor mange negative, og også hvor mange lange indlæg i forhold til hvor mange korte (altså i hvor meget tekst indlæggene indeholder). Men i praksis er det umuligt at aflæse disse data direkte. Visualiseringen lægger i høj grad op til udforskning, og til at man interagerer med visualiseringen, men som datarepræsentation er klarheden meget lav, særligt på makroplanet. Da man også kan sige, at vurderingen af, hvorvidt de enkelte ord er positive eller negative og i hvilken grad, i høj grad er subjektiv, kan man ikke sige at visualiseringen er særligt datanær på makroplanet.

We Feel Fine tilbyder dog en lang række alternative visninger, som giver mulighed for at foretage forskellige sorteringer af partiklerne. Man kan f.eks. vælge at se, hvor mange 'følelser' der er ytret af dels mænd og kvinder, i regnvejrs og solskinsvejr<sup>130</sup>, hvilke følelser der er flest af lige nu, og hvordan de følelser, der bliver vist lige nu, varierer fra gennemsnittet af de følelser, der registreres af systemet. Herudover kan man foretage forespørgsler på nogle få parametre, så man f.eks. kan se, hvad mænd i gruppen 20-29 år fra Ohio har følt, mens det har regnet. Gennem disse visninger stiger klarheden af visualiseringen markant.

---

<sup>130</sup> I det tilfælde hvor We Feel Fine kan finde ud af hvornår blogindlægget er skrevet, og hvor bloggeren skriver fra, er det også muligt at finde ud af hvordan vejret har været mens indlægget blev skrevet.



FIGUR 28 –SCREENDUMP FRA WE FEEL FINE, MOBS (HARRIS 2007)

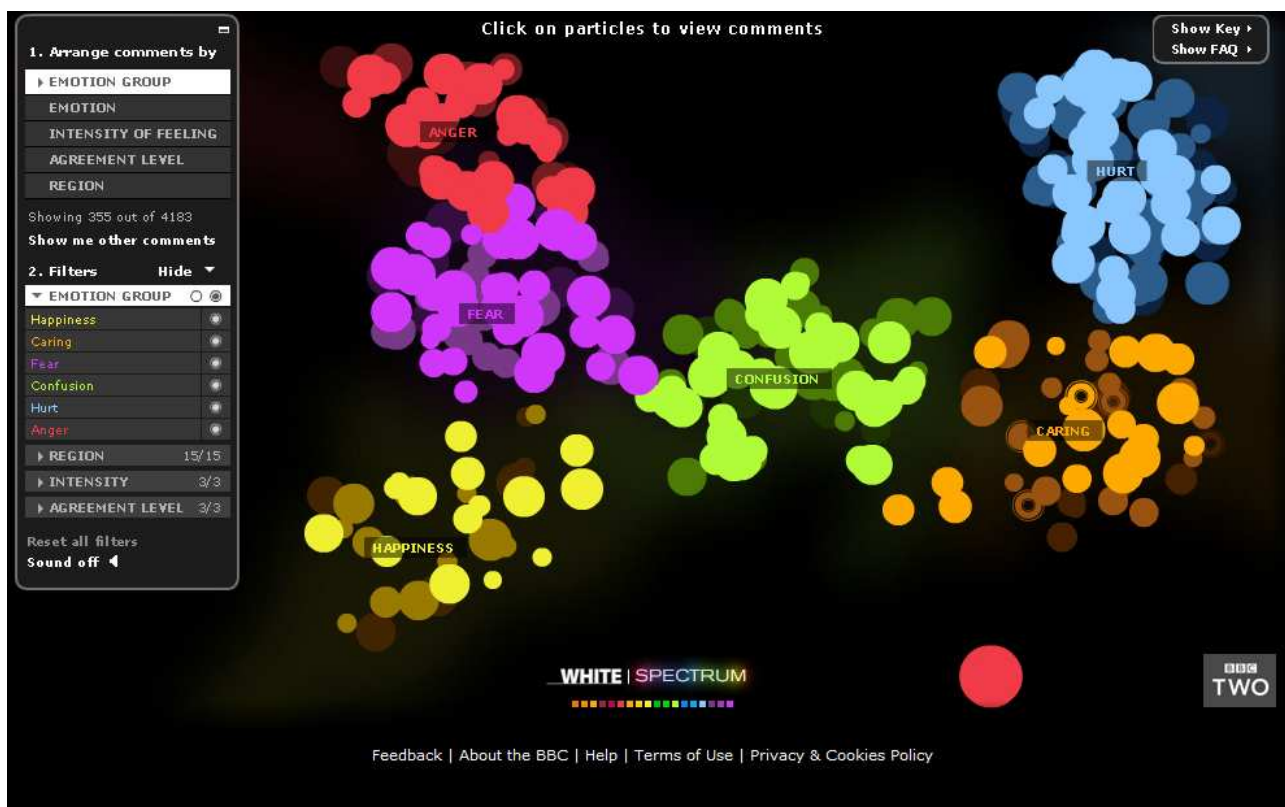
På makroplanet kan vi pludselig indlæse en masse detaljer om de data, vi får vist. Det, der derimod kan virke overraskende, er, at så snart at partiklerne bliver sat i sådanne systemer, f.eks. ved at vi adskiller mande- og kvindedefølelserne, så ophører hver enkelt partikels selvstændighed og bliver i stedet en del af en gruppe. Det virker umiddelbart mere tillokkende at undersøge partiklerne, når de farer rundt i et kaos og ikke umiddelbart er mulige at tolke på (før end man klikker på dem og finder ud af, hvad der ligger bag).

Denne visualisering er naturligvis et enkeltstående eksempel, og der kan ikke generaliseres ud fra dette. Ikke desto mindre lader det til at være tilfældet i netop dette værk, at når makroplanet bliver for klart og enkelt (f.eks. "Så mange kvinder i forhold til mænd" eller "Så mange i regnvejrs i forhold til solskinsvejrs") nedtones den eksplorative effekt, som i øvrigt virker godt, når hver enkelt element, hver partikel, farer kaotisk rundt og har sit eget liv.

### 7.5.3 WHITE / SPECTRUM

*White / Spectrum* er en visualisering af en del af BBCs projekt for involvering af brugere, kaldet *BBC News Have Your Say*. Debatten, som visualiseres, er en del af BBC 2s tema om den hvide arbejderklasse, og visualiseringen har en form, der minder om den fra *We Feel Fine*. I stedet for

at lede efter sætninger baseret på verber (som *I feel*) tages der udgangspunkt i adjektiver (som *suffered, resentful, warm, satisfied*), som dog kaldes *emotions*.



FIGUR 29 –BBC TWO'S WHITE | SPECTRUM.

Som i *We Feel Fine* er hver sætning repræsenteret af en partikel, der har en farve og en størrelse. Som udgangspunkt er partiklerne inddelt i såkaldte *emotion groups*, blandt andre *fear, caring, og anger*. Hver partikel er farvet efter sådanne følelsesgrupper, og de bevæger sig kaotisk rundt i nærheden af de andre partikler fra samme gruppe. Man har som bruger dog også mulighed for at få vist hver enkelt *emotion*, at få rangordnet følelserne efter intensitet eller grad af enighed (*disagree, neutral* eller *agree*). Farven på hver partikel udtrykker her, hvilken følelsesgruppe partiklerne tilhører, mens intensiteten af farven viser enighedsgraden, og størrelsen på partiklen illustrerer følelsens intensitet.

Det er desværre svært at gennemskue, hvordan BBC udvælger og behandler sine data, og der er desværre noget, der tyder på, at den semantiske behandling af debatindlæggene i høj grad er baseret på nøgleordssøgninger. F.eks. kan man under følelsesgruppen *Care* finde følgende indlæg:

---

*When you say that in Victorian Britain we had low wages, smoggy streets, crime, polluted rivers, no benefit safety nor free health care doesn't that sound like what Britain is becoming once again?*

---

Close this comment

Added Thursday 6th March 2008 at 4:49pm

when you say that in victorian britain we had low wages, smoggy streets, crime, polluted rivers, no benefit safety nor free health **care** doesn't that sound like what britain is becoming once again?

Logicman, Stevenage

View full comment

Vote

Agree Neutral Disagree

WHITE | SPECTRUM

BBC TWO

Feedback | About the BBC | Help | Terms of Use | Privacy & Cookies Policy

FIGUR 30 -ET CITAT FRA WHITE | SPECTRUM, GRUPPERET UNDER 'CARE'.

Brugeren nævner tydeligvis *health care* i sit indlæg, og det er et glimrende eksempel på, at bare fordi et adjektiv (hvilket ordet ikke engang er i denne sammenhæng) indgår i et tekststykke, beskriver dette adjektiv ikke nødvendigvis den sindsstemning, forfatteren har været i, mens teksten blev forfattet, sådan som White | Spectrum ellers umiddelbart lægger op til i sin ordning af partiklerne.

We Feel Fine og White | Spectrum skal dog undskyldes af, at teknologien, som jeg også tidligere har nævnt, på nuværende tidspunkt ikke er moden til at kunne lave avancerede nok semantiske analyser af ustruktureret tekst. Ser vi bort fra, at systemet ikke reelt fungerer efter hensigten, i hvert fald ikke den hensigt, der kommunikerer gennem White | Spectrums brugergrænseflade, og i stedet forestiller os, at den visualisering, vi ser, rent faktisk dækker over brugernes sindsstemninger (eller i det mindste den stemning eller holdning, der kommer til udtryk i debatindlæggene), kan vi se på systemet som et bud på, hvordan visualiseringer af semantiske analyser kan tage sig ud i fremtiden.

Det første, der da springer i øjnene, er inddelingen i følelsesgrupper. Denne opdeling er et bud på, hvordan man kan skelne mellem forskellige indlæg. At opdele eller gruppere indlæggene på en eller anden måde virker oplagt, idet det kan hjælpe folk med at overskue store mængder af data. I *We Feel Fine* så vi, hvordan partiklernes tilfældige færden henover skærmen fik helhedsbilledet til at virke kaotisk (selvom det også bidrog med et fokus på den enkelte partikel). Ved at opdele i grupper bliver det nemmere at forholde sig til data, idet grupperingerne byder på en fortolkningsramme. Om følelser er den rette fortolkningsramme vil naturligvis afhænge af det enkelte projekt, men idéen er spændende nok. Man bør som designer være sig bevidst, hvilken gruppering (og dermed fortolkningsramme) der egner sig bedst til hvilket projekt. I tilfældet med følelser lægges der op til en præsentation af: Hvilke følelser frembringer diskussionen om den hvide arbejderklasse?<sup>131</sup> Andre grupperinger kan på samme vis give anledning til andre spørgsmål. Eksempelvis kunne en gruppering efter emne illustrere, hvilke emner der fylder mest i diskussionen af den britiske, hvide arbejderklasse. *White | Spectrum* ville være et oplagt værktøj til at aggregere *Wisdom of crowds*, men idet der lægges vægt på følelser og ikke på holdninger eller emner, og da der i visualiseringen kun tages udgangspunkt i udsnit af alle indlæg, er det umuligt at benytte systemet aktivt i problemløsningen. Derimod lægger systemet op til en udforskning af, hvordan de forskellige følelser kommer til udtryk, og dermed at man sætter sig ind i de forskellige diskussionsemner.

I forhold til visualiseringsmodellen, jeg tidligere introducerede, placerer denne visualisering sig i den klare del af modellen, idet vi nemt kan se, hvad de forskellige grupperinger viser. Omvendt er data enormt upræcise, både hvad angår indsamlingen, og hvad angår repræsentationen, og visualiseringen kan derfor ikke siges at være datanær. Visualiseringen hører derfor primært hjemme i anden kvadrant.

#### *7.5.4 VISUALIZING THE BIBLE*

Chris Harrisons projekt, at visualisere Biblen, er endnu et forsøg på at visualisere tekst. Harrisons tilgang har dog en mere matematisk karakter end *We Feel Fine* og *White | Spectrum*, og den er måske nærmere beslægtet med Chris Jordans værker. Harrison har lavet en række visualiseringer, der tager udgangspunkt i specifikke typer af data fra Biblen, hovedsageligt interne referencer, bynavne og navne på personer. Disse har han efterfølgende forsøgt at forbinde på forskellig vis. Det mest berømte af værkerne fra hans projekt er visualiseringen af interne referencer.

---

<sup>131</sup> Desværre kan man ikke få svar på dette spørgsmål i selve applikationen, da *White | Spectrum* ikke viser alle indlæg ved hver visning, men blot et mindre uddrag. Derfor er der altid ca. lige mange partikler for hver følelsesgruppe. På en kort introskærm kan man dog se hvor mange indlæg og hvor mange procent der er i hver følelsesgruppe.

FIGUR 31 -VISUALIZING THE BIBLE (HARRISON 2008)

På en horisontal akse ses hver enkelt kapitel fra Biblen, repræsenteret ved en søjle, der angiver kapitlets længde. De enkelte kapitler er vist i den rækkefølge, de fremgår i Biblen, og grupperet efter den bog, de kommer fra. Adskillelsen af hver enkelt bog er skabt ved at lave bøjerne henholdsvis grå og hvide. Oven på søjlerne ses en masse (faktisk 63.779) halvcirkler, der hver forbinder to kapitler. Disse buer repræsenterer en intern reference. Længden af buen påvirker farven, og skaber hvad Harrison selv kalder en regnbueeffekt.

Til trods for, at det er svært at aflæse, hvilke kapitler der faktisk er forbundne, og til trods for, at denne viden om interne referencer i Biblen er relevant for de færreste, har visualiseringen alligevel en dragende effekt. Komplexiteten giver anledning til udforskning. Ude af kontekst er billedet skabt af visualiseringen ganske flot. Kombinationen af de gentagne buer, farverne og den uoverskuelige kompleksitet skaber en spændende komposition. I den forstand kan visualiseringen siges at have et stærkt udtryk. På den anden side har visualiseringen også den kvalitet, at hvis man evner at følge de enkelte streger, og kan man overskue de mange kapitler i bunden, vil man faktisk kunne aflæse, hvilke kapitler der indeholder krydsreferencer til andre kapitler. Igen er der altså forskel på, om man anskuer visualiseringen på et mikro- eller et makroplan.

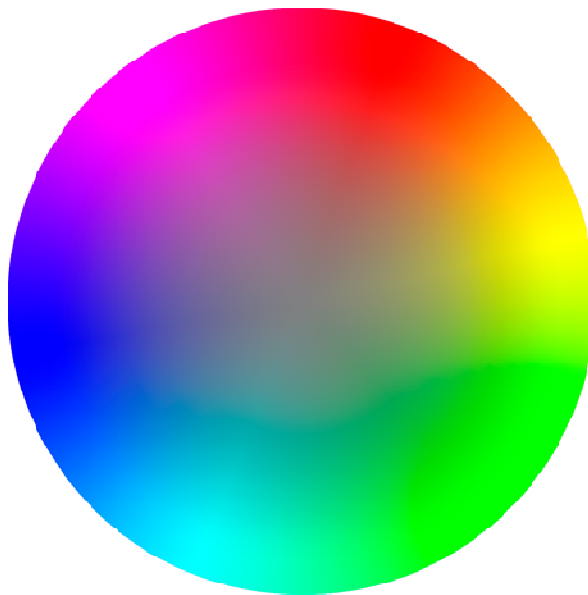
Forestiller vi os, at vi fik til opgave at finde ud af, hvilke krydshenvisninger der findes fra et givent kapitel, eller bliver vi bedt om at finde ud af, hvor mange krydshenvisninger der findes i et kapitel, står vi overfor en forholdsvis uoverskuelig opgave, hvor det formentlig ville være lettere at slå op i Biblen og se efter. At skulle finde det rette kapitel i visualiseringen, at skulle finde de

rette buer og følge dem til vejs ende uden at fare vild i virvaret af buer ville være en svær opgave. I den forstand kan visualiseringen siges at være ikke-klar.

Igen må vi spørge os selv, hvilken overordnet funktionalitet denne visualisering tilbyder. Svaret må umiddelbart ligge i de informationer, vi nemt kan aflæse ud af visualiseringen: Der er enormt mange interne krydshenvisninger i Biblen, og der ser ikke ud til at være nogen regler for afstanden mellem de kapitler, der refereres. Et forholdsvist enkelt budskab i en meget flot indpakning. Man kan sige, at visualiseringens makroplan, som typisk vil være modtagerens grundlag for førstehåndsindtrykket, skaber en forventning til visualiseringen. I dette tilfælde kunne man med ganske få midler have skabt et mere udfordrende førstehåndsindtryk, eksempelvis ved at gøre buerne en anelse bredere ved at lade dem forstærke hinanden hvor de krydser hinanden. Desuden kunne man have gjort dem retningsbestemte. Således ville det blive mere tydeligt hvilke kapitler, der f.eks. er de mest refererede i Biblen, samt hvilke kapitler der refererer mest til andre.

#### 7.5.5 SEMANTIK OM PERCEPTION

Den sidste visualisering, jeg vil tage frem her, er en, der ikke tager udgangspunkt i tekst, men som stadig kan siges at være en form for semantisk software. Systemet er i virkeligheden en slags undersøgelse, der tager udgangspunkt i brugernes navngivning af farver. En lang række brugere er blevet bedt om at navngive 10 tilfældige farver. De farver, som brugerne bliver bedt om at navngive, har alle en teknisk værdi, hvorfor de kan placeres i en farvecirkel.<sup>132</sup>



FIGUR 32 -FARVECIRKLEN

Når man sammenknytter information om farve og navngivning, kan man lave en farvecirkel med navne.

---

<sup>132</sup> Dolores Labs 2008

FIGUR 33 -THE AMAZING COLORWHEEL (DOLORES LABS, 2008)

Denne visualisering viser, hvilke ord der knyttes til hvilke farver, og illustrerer således, hvordan hvad der er 'royal blue' for en person, kan være såvel 'purple' som 'blue' for en anden. Ved at drage nytte af Wisdom of crowds fungerer visualiseringen som aggregator for det kognitive problem: Hvad hedder de forskellige farver?

Visualiseringen fungerer godt på både mikro- og makroplanet. På makroplanet fortæller den tilskueren, at vi bruger de samme ord til at beskrive forskellige farver, og forskellige ord kan beskrive den samme farve. På mikroplanet kan vi finde ud af, hvilke ord der bedst beskriver enkelte farver. Visualiseringen virker meget enkel, men anskuer vi data som f.eks. et skema eller

anden 'rå data', vil vi ikke kunne danne os et lige så godt overblik. Visualiseringen hjælper altså modtageren med at forstå data.

Til trods for, at visualiseringen kunne være mere datanær ved at vise tekniske beskrivelser af farverne, f.eks. gennem RGB-farvekoder<sup>133</sup>, er visualiseringen mere datanær, end den er udtryksfuld. Klarheden er til gengæld ganske høj. Til trods for, at vi ikke kan se, hvilken nøjagtig værdi de forskellige navne repræsenterer, kan vi se farven og se, hvordan farven ser ud i forhold til andre farver i farvecirklen. Så selvom visualiseringen vil have en forholdsvis central placering i visualiseringsmodellen, må den trods alt tilhøre første kvadrant mere end de andre kvadranter.

## 7.6 ERFARINGER FRA ANALYSER AF VISUALISERINGSEKSEMPLER

Efter at have gennemgået en række forskellige visualiseringer, der alle i en eller anden grad visualiserer semantiske analyser, er der dukket nogle generelle tendenser for semantiske visualiseringer op.

### 7.6.1 UOVERSKUELIGHED VERSUS GRUPPERING

Vi så med *We Feel Fine*, hvordan det uoverskuelige kaos i højere grad opfordrer til at udforske den enkelte partikel, end når partiklerne tilsammen tegner en figur, eksempelvis når partiklerne tegner et mandligt og kvindeligt piktogram, eller når partiklerne mimer en sol, skyer eller regn.

Hvorvidt makroplanet skal være rent kaos, eller om der kan eksistere grupperinger ved visualiseringer med mange emner, står dog ikke klart ud fra denne observation. I *White | Spectrum* virker det f.eks. fint at gruppere partiklerne i små uformelige klynger.

Man kan forestille sig, at partiklerne bliver mindre oplagte at udforske, når de bliver til dele af et naturligt eller genkendeligt ikonografisk tegn, som f.eks. et piktogram. Omvendt bruger Chris Jordan også genkendelige makrobilleder, f.eks. ved at lade tusindvis af cigaretpakker udgøre et maleri. Chris Jordan benytter sig dog, i modsætning til f.eks. *We Feel Fine*, af så mange partikler (her cigaretpakker), at man skal zoome meget langt ind på billedet for at se cigaretpakkerne. Og har man zoomet så langt ind, kan man ikke længere se maleriet. Man kan sige, at Chris Jordan formår at adskille sine informationer, jf. Tufes lag og adskillelse – altså evnen til at kunne adskille forskellige typer af informationer fra hinanden, ved at benytte forskellige visuelle kvaliteter eller lag.

---

<sup>133</sup> RGB-farvekoder deler beskrivelsen i farver op i tre dele: Rød, grøn og blå. Hver variabel kan have værdi mellem 0 og 255, hvor 0 er den laveste lysstyrke. Eksempelvis vil 0-0-0 give sort, 255-255-255 give hvid, 255-0-0 give rød og 255-0-255 give lilla.

FIGUR 34 –SKULL WITH CIGARETTE, TOTAL (TV.) OG DETALJE (TH.) (CHRIS JORDAN, 2007)

Når man designer visualiseringer, bør man derfor overveje, hvor vigtigt det er, at man på makroplan benytter sig af genkendelige figurer, og i så fald om man kan adskille makro- og mikrodesignet tilstrækkeligt fra hinanden, således at det ene plan ikke tager fokus fra det andet.

### *7.6.2 MAKROPLANET AFSLØRER VISUALISERINGENS FORMÅL OG FUNKTION*

I eksemplet med Harris' visualisering af Biblen så vi, hvorledes visualiseringen i højere grad lagde op til en illustration af kompleksitet end en overskueliggørelse af et komplekst felt.

Makroplanet er det første, man møder, når man ser en visualisering. Det er først, når brugeren vælger at 'dykke ned' i data, at man opdager et evt. mikroplan. Makroplanet bør derfor kommunikere, hvad brugeren kan forvente af visualiseringen.

### *7.6.3 GRUPPERINGER FORDRER LÆSNING EFTER GRUPPER*

Når emner grupperes, lægges der op til en læsning af emnerne efter disse grupper. Således lægger White | Spectrums gruppering efter følelsesgruppe op til, at man bruger visualiseringen til at sammenligne de følelser, der kommer til udtryk i de enkelte indlæg. Forskellige grupperinger kan således benyttes til at skabe forskellige tilgange til det materiale, der visualiseres.

#### *7.6.4 KOMPLEKSITET ER IKKE ET PROBLEM*

Tuftes kommentar om, at data ikke i sig selv kan være uoverskuelige, men at de kan være uoverskueligt præsenteret, er ikke blevet dementeret gennem analyserne. Tværtimod er der noget, der tyder på, at selv avancerede former for visualiseringer, som f.eks. dem der afdækker følelser holdninger, eller dem der skal vise, hvor mange 1,14 millioner er, faktisk kan være vellykkede. At kvaliteten af data i forbindelse med de semantiske analyser ikke ser ud til at kunne leve op til visualiseringernes formål, er en sag, der blev diskuteret i et tidligere afsnit af denne rapport.

Med korrekt brug af lag og adskillelse ser der ud til at være meget langt til grænser for, hvor komplekse visualiseringer der kan skabes.

#### *7.6.5 OVERSKUELIGHED OVER KOMPLEKSITET*

Der findes mange forsøg på at visualisere komplekse forhold. Et typisk problem for sådanne visualiseringer er, at de ender med at vise uoverskueligheden frem for at overskueliggøre.

Kompleksitet er et udtryk, der gentages i mange visualiseringer over komplekse datasæt. Det at udtrykke kompleksiteten har naturligvis sin egen værdi og kan også være et argument i sig selv. Det virker dog umiddelbart som om, at mange designere af visualiseringer ønsker at eksplicitere kompleksiteten af de emner, de portrætterer, frem for at arbejde med at gøre dem overskuelige eller værd at udforske.

Det primære mål for visualiseringer vil ofte (men ikke nødvendigvis) være at skabe et overblik over et datasæt, der ikke ellers er let tilgængeligt. Heri ligger der en udfordring i at overskueliggøre det uoverskuelige, og det er her, visualiseringens force kommer på banen. F.eks. ved at udnytte visuelle og interaktive elementer som størrelser, nærhed, former, farver, hastigheder, bevægelsesmønstre, lysintensitet, klarhed osv. har designeren en lang række værktøj, der kan udnyttes til at skabe lag og adskillelse. Ved samtidig at overveje forholdet mellem mikro- og makroplan er der også mulighed for at skabe forståelser af indholdet dels på overordnet akkumuleret plan og på et konkret eksempelniveau.

### **7.7 KONKRETISERING AF ERFARINGER**

Jeg har i dette kapitel præsenteret forskellige tilgange og værktøj til at designe visualiseringer, hvilket har resulteret i en række konkrete erfaringer. Disse erfaringer vil jeg drage nytte af i det følgende kapitel, hvor jeg vil præsentere et forslag til et design af en konkret visualisering.

## 8 FORSLAG TIL SEMANTISK VISUALISERING

Jeg har i forrige kapitel gennemgået teoretiske tilgange til visualisering af information, præsenteret en model for analyse af visualiseringer samt analyseret en række visualiseringer, der i en eller anden udstrækning bygger på semantiske data eller leger med vores opfattelse af mening. I dette afsnit vil jeg præsentere et forslag til en visualisering, der vil kunne bruges i forbindelse med borgerinddragelse, og som ligeledes tager udgangspunkt i en semantisk fortolkning af data.

### 8.1 VISUALISERING AF DEN GLOBALE INTELLIGENS

Med Web 2.0 er der skabt en grobund for Wisdom of crowds og den globale intelligens. Mens semantic software bliver bedre og bedre, må vi overveje, hvordan vi kan høste frugterne af World Wide Webs øgede muligheder for kommunikation og deling af viden. Mængderne af viden, holdninger og idéer er enorme og uoverskuelige. For at kunne aggregere de mange nedfældede tanker har vi derfor brug for et system, der kan hjælpe os med at anskue de eksisterende data uden at forringe, ødelægge eller simplificere dem. Visualiseringer er ikke magiske og tilbyder ikke en direkte læsning af Wisdom of crowds – svaret dumper ikke ned i turbanen. Men med visualiseringer gives der mulighed for på ny vis at skabe overblik og indblik i forholdet mellem data; og med muligheden for at ændre på de parametre, der udgør forholdene mellem de forskellige data, dukker der nye muligheder for fortolkning op.

Formålet med en visualisering er at skabe mulighed for at overskue enorme mængder af indhold, for på den måde at kunne sammenkoble de mange unikke, værdifulde stykker data, der udgør helheden. Men en visualisering vil altid være baseret på en fortolkning af data, og visualiseringen i sig selv vil ikke udgøre et referat af den globale intelligens. Men med muligheden for at styre visualiseringen og ved at ændre på fortolkningsrammen kan det indhold, der er skabt af internetbrugerne, få en ny betydning og tilsammen udgøre helheder, der giver os nye forståelser af de emner, der behandles.

Målet med en visualisering er ikke at fjerne meningen fra den enkelte brugers producerede indhold, men at forstå sammenhænge og sammenkoblinger mellem data, samtidig med at interessen for det individuelle indhold øges. Man kan sige, at det at skabe en foranderlig fortolkningsramme for de visualiserede data lægger op til en mere abduktiv og eksplorativ tilgang til de data, man ønsker at lære at kende. Derved understreger man også over for brugeren af visualiseringen, at tekst ikke blot er tegn for konkrete fænomener, men at indholdet fortolkes i en kontekst.

### 8.2 ABDUKTIV TILGANG TIL VISUALISERINGER

I rapportens videnskabsteoretiske kapitel redegjorde jeg for de seks abduktive modeller for udledning. Disse modeller kan også benyttes til at beskrive, hvorledes en visualisering kan bruges til at udlede mulige tolkninger af de semantiske data.

#### 8.2.1 *DE ÅBNE, IKONISKE MODELLER*

Den åbne, ikoniske tone vil særligt gøre sig gældende i det tidlige møde med en visualisering, hvor brugeren skal orientere sig og danne sin egen fortolkning af visualiseringens repræsentation af data. Man kan lidt nørklet sige, at visualiseringens modtager danner en

fortolkning af visualiseringens fortolkning af data. Skulle der i denne tidlige tolkning af visualiseringen opstå overraskelser eller uventede slutninger, kan dette lede til nye forståelser af de repræsenterede emner. Denne type slutning kan eksempelvis opstå i systemer som White | Spectrum, hvor en række emner grupperes under følelsen 'caring'. Vores tænkte modtager finder i denne gruppe et indlæg om unge kriminelle og drager derefter parallellen mellem *caring* og *de unge kriminelle* og får tanken, at de unge kriminelle føler sig svigtede.

Den åbne, ikoniske token vil derimod dreje sig om et mere overordnet billede. Ved denne form for slutning har brugeren allerede skabt sig en forståelse af repræsentationen og forsøger nu at lave en konkret fortolkning på data. Dette kan f.eks. være opdagelsen af, at enkelte grupper er væsentligt større end andre. Det er ikke i denne form for abduktiv udledning, at der udledes en konkret forklaring på symptomet – symptomet identificeres blot.

Den åbne, ikoniske type vil være mulig at fremkalde i muligheden for at vise metaforer direkte i visualiseringen. Eksempelvis ved at benytte visualiseringens makroplan på denne vis, såsom Chris Jordan gør det i mange af sine visualiseringer.

### 8.2.2 DE ÅBNE, INDEKSIKALSKE MODELLER

Idet der er tale om en visuel, symbolsk repræsentation af data, vil de åbne, indeksikalske udledningsmodeller ikke være oplagte at fremhæve. De to modeller (de åbne, indeksikalske token og de åbne, indeksikalske typer) knytter sig til indeksikalske forhold og vil i forbindelse med oplevelsen af en visualisering nærmere gå på, hvordan det kausale forhold mellem data og visuel repræsentation forholder sig.

### 8.2.3 DEN ÅBNE, SYMBOLKSE TYPE

Den åbne, symbolske type (også kaldet *forklaringen*) beror på en form for analyse af den præsenterede visualisering og forsøger at forklare elementer, der ikke giver sig selv. Det kan bl.a. være en form for konklusion på en samling observerede symptomer (dvs. de åbne, ikoniske token), eksempelvis at en stor gruppe indlæg om *unge kriminelle* og en stor gruppe indlæg om *mindre løn til skolelærere*, der hver især observeres som symptomer, bliver til en forklaring, der forbinder de to symptomer: At der er flere unge kriminelle, fordi lærerne ikke får nok i løn og derfor svigter eleverne.

## 8.3 SCENARIE

Formålet med at lave en visualisering i forbindelse med borgerinddragelse er at udnytte internettets evne til at få brugere til at deltage aktivt, ytre deres holdninger og diskutere emner. For overskuelighedens skyld vil jeg i dette tilfælde antage at datasættet stammer fra et modereret debatforum, stillet til rådighed af en offentlig instans, eksempelvis en kommune. Debatforummet skal have til formål at diskutere politiske emner; emner der kan være præsenteret af så vel borgere som byplanlæggere eller politikere. Der vil således være tale om en Web 2.0-løsning, hvor alle brugere har mulighed for at deltage på lige plan.

Er forummet en succes, vil mange brugere dele deres holdninger eller idéer om emner, og der vil foreligge et grundlag at lave semantiske analyser på. Her antages det, at teknologierne inden for semantisk software, særligt inden for NLP, text mining og evt. Semantic web er blevet tilstrækkeligt robuste til, at man kan lave meningsfulde fortolkninger af brugernes indlæg og

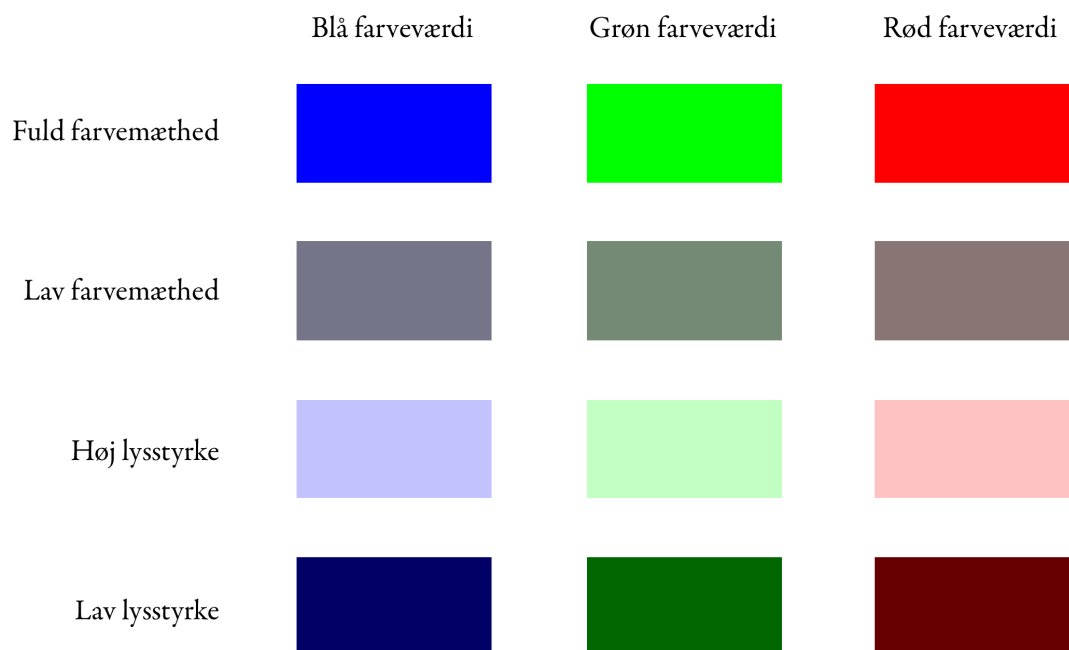
eksempelvis udtrække data for, hvilket emne de diskuterer, eller deres holdning (såfremt en sådan er ytret).

#### 8.4 VISIONER FOR VISUALISERINGEN

Et centralt spørgsmål for visualiseringen er, hvor den skal placeres i den model, jeg præsenterede tidligere i dette kapitel. Som udgangspunkt er det en god idé at stræbe efter klarhed, men da vi taler om semantik og begreber, findes der ingen sandheder. Som det blev diskuteret i kapitel 5, er begreber ikke definatoriske. At lave en visualisering, der definerer begreber entydigt, vil derfor være misvisende. Omvendt kan man sige, at visualiseringen ikke har nogen funktion, hvis den ikke kan aflæses. Der må derfor være tale om en balance mellem, at visualiseringen skal være intuitiv at aflæse, og at den ikke skal definere begreberne.

Vi vil opdage, at den samme diskussion gør sig gældende på visualiseringsmodellens anden akse, hvor visualiseringen på den ene side ikke må blive for datanær, igen pga. begrebernes *fuzziness*, og på den anden side stadig skal give mening. Flere af de visualiseringer, jeg har gennemgået i dette kapitel, har vist sig at have forskellige datanærheder på mikro- og makroplan. Det giver umiddelbart god mening, at man på et makroplan primært får et overblik og får lyst til at dykke ned i data, men at man på mikroplanet kan gennemskue, hvilke data der helt præcist ligger til grund for visualiseringen. Det vil med andre ord sige, at visualiseringen med fordel kan være udtryksfuld på makroplan og datanær på mikroplanet. Denne måde at håndtere overblik og detalje benytter sig af Tuftes ideal om lag og adskillelse. Det er dog vigtigt, at overblikket og detaljen er så adskilte, at de ikke blandes sammen. Dette kan opnås som hos Chris Jordan ved at sørge for, at detaljerne er så små dele af helheden, at man ikke kan se den enkelte detalje, når man ser overblikket, og at man ikke kan se helheden, når man ser detaljen.

Min visualisering vil bygge videre på nogle af de tanker, der ligger bag *We Feel Fine* og *White | Spectrum*, i den forstand at der tages udgangspunkt i, at hvert enkelt indlæg fra debatforummet repræsenteres som en partikel. En sådan partikel kan have en række databærende egenskaber. Vi har set dette i begrænset udformning blandt de tidligere nævnte visualiseringer, hvor partiklernes størrelse eller farve f.eks. bar en værdi, der beskrev indholdet af den tekst, som partiklen beskrev. Farver alene består eksempelvis af flere værdier, og en partikel kan således have en farveværdi, mæthed og en lysstyrke.



FIGUR 35 – ØVERST SER VI TRE FORSKELLIGE FARVE. NEDAD SER VI DE TRE FARVER MED NEDSAT FARVEMÆTHED, HØJERE LYSSTYRKE OG LAVERE LYSTTYRKE.

Udover farven på partiklerne er der også mulighed for at ændre på partiklens størrelse og form, den hastighed, den bevæger sig med, samt de mønstre, den bevæger sig i (eksempelvis i kurver, i cirkler, i zigzag). Herudover kan man benytte sig af grupperinger, jf. gestaltpsykologiens regel om, at emner, der er tæt på hinanden, opfattes som relaterede. Kombinerer vi grupperinger og farver, åbner der sig også en ny mulighed: at lægge baggrundsfarver på relaterede emner for at sige noget om gruppen som helhed.

Som tidligere nævnt sætter visualiseringens makroniveau en læsningsramme for de data, der præsenteres. For at undgå at være partisk eller at skabe en låst fortolkning af data, er det oplagt at arbejde med forskellige fortolkningsrammer for data. Dette kan eksempelvis foregå som på We Feel Fine, hvor der gives mulighed for at vælge at få data repræsenteret efter forskellige kriterier, f.eks. opdelt efter forfatternes køn. For ikke at miste visualiseringens kvantitative fordele er det vigtigt, at man ikke gør som i White | Spectrum og kun anvender tilfældige dele af datasættet. Det er vigtigt, at det kun er fortolkningsrammen, der udskiftes, når man vælger en ny måde at få data visualiseret på.

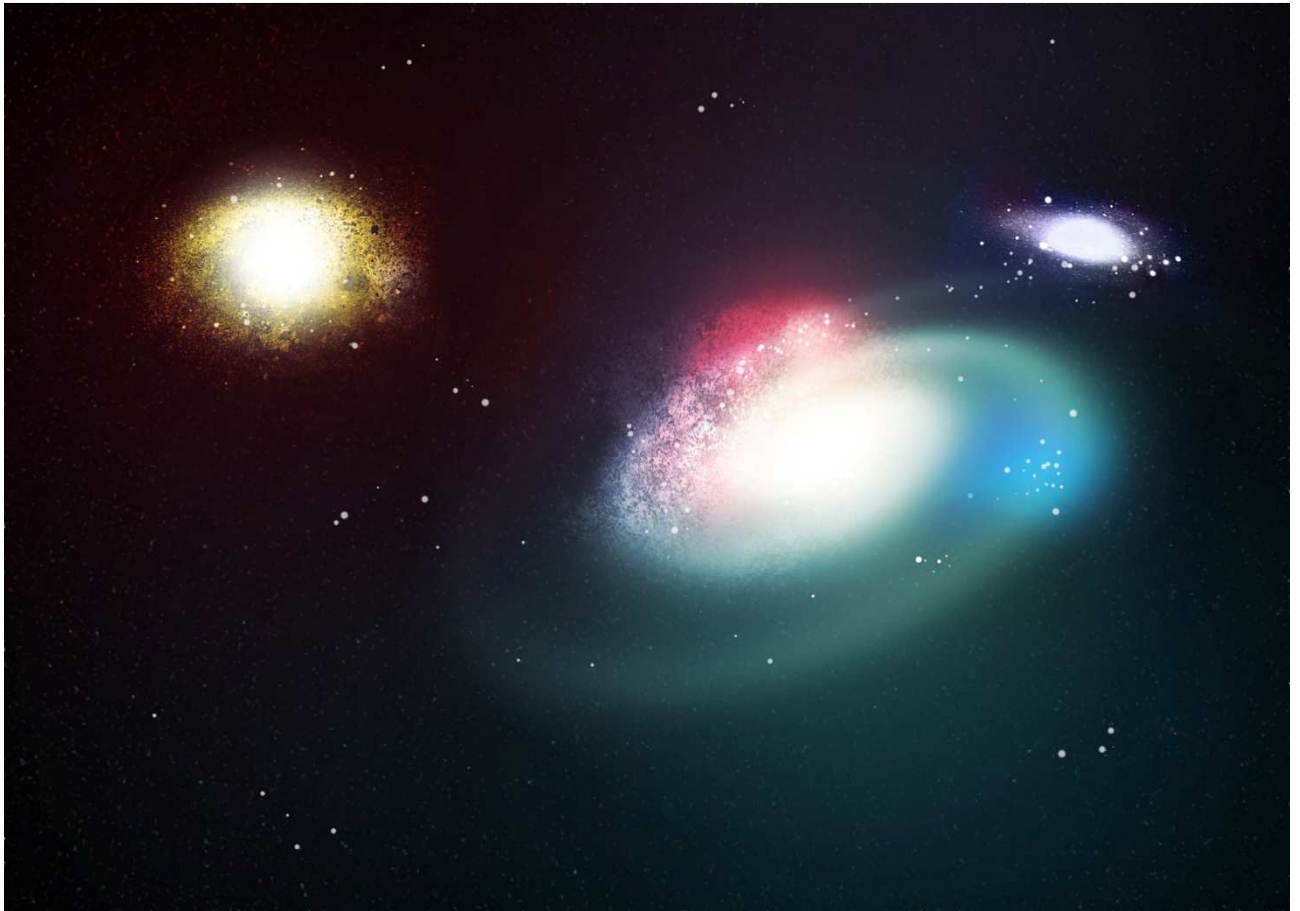
Det er også en mulighed i bedste Web 2.0-stil at blande datakilder, således at man f.eks. bruger debatindlæg sammen med geografiske kort (hvilke geografiske områder bliver diskuteret) eller sammen med data om mediedækning af relevante emner.

## 8.5 UDFORMNING

Udformningen vil være en dynamisk, interaktiv visualisering, hvor der er mulighed for at se data repræsenteret på forskellig vis.

Som i We Feel Fine har jeg her valgt et neutralt udgangspunkt, hvor data repræsenteres som partikler, der har en række egenskaber, og som har en form for semantisk tilknytning til andre emner. Tilknytningen til andre emner sker i grupperinger. Diskussioner baseret på tekstlige indlæg udformer sig typisk i en form for tråde. Jeg finder det oplagt at bibeholde viden om indlæggenes forhold til hinanden og lade disse afspejle sig i grupperingerne. Således har vi for det første en gruppering efter semantisk emne, så diskussioner om Danmarks investering i militær i en eller anden grad er forbundet til diskussionen om Danmarks deltagelse i krigen i Irak, men ikke særligt tæt forbundet med diskussionen om de studerendes understøttelse. Disse grupperinger skal ske ud fra en rent sproglig forbindelse mellem indlæggene. For det andet har vi en gruppering, der afspejler specifikke diskussioners udvikling. Således vil alle indlæg i en diskussion om Danmarks deltagelse i Irak-krigen være forbundet. Jeg har her valgt at bruge en upartisk analogi til at beskrive disse grupperinger: Universet. Med udgangspunkt i den analogi bliver semantiske grupperinger illustreret som galakser, hvor trådene i galaksen illustrer trådene i diskussionerne. Nogle galakser vil bære præg af at have enkelte lange tråde, andre vil være mere kaotiske med mange små tråde. Galakser kan smelte sammen eller ligge tættere eller længere fra hinanden. Det er også muligt, at enkelte semantiske grupperinger er så små, at der i visualiseringen ikke bliver skabt en galakse, men at der blot flyder enkelte ensomme stjerner rundt i universet.

Det skal understreges, at eksemplerne i dette afsnit ikke er baseret på reelle data, men er tænkte visualiseringer, hvis formål udelukkende er at give et indtryk af visualiseringens udformning.

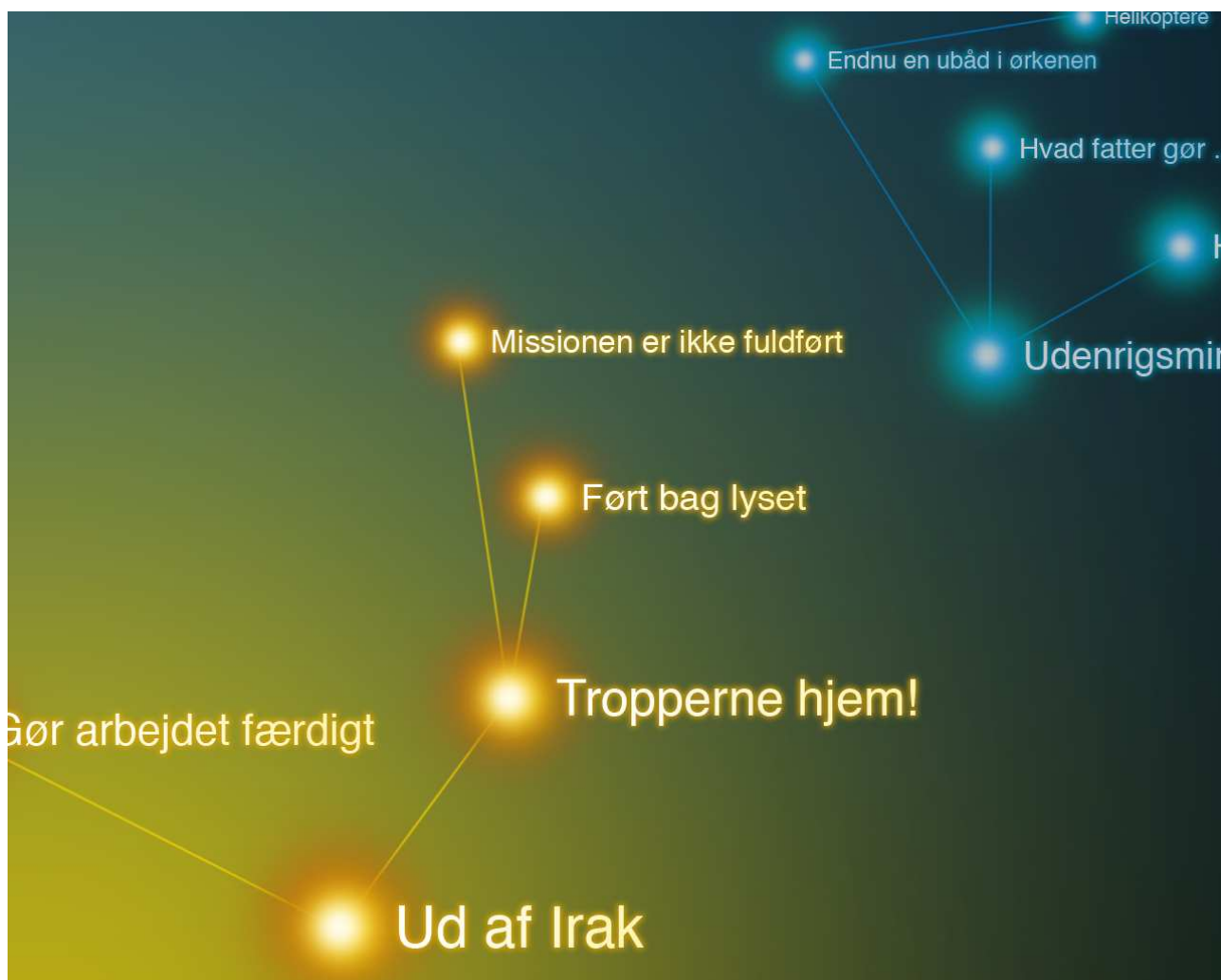


FIGUR 36 -VISUALISERING PÅ MAKROPLAN: HELHEDSBILLEDET LIGNER ET FOTO AF UNIVERSET.

Hver enkelt partikel vil knap nok være synlig på makroniveauet. Enkelte områder vil være oplyste med samme farve i hver 'galakse', som resultat af, at der lige nu eller for ganske nyligt har været meget aktivitet på netop dette område. Intensiteten af farverne i området illustrerer antallet af nye indlæg i det pågældende område, og farven illustrerer forskellige typer af diskussioner (er der f.eks. mere end to, der diskuterer, er der et enkelt indlæg pr. deltager i diskussionen som i en afstemning eller er det alle mod alle?). På den måde tages der i denne visualisering (i modsætning til f.eks. White | Spectrum) højde for, at et indlæg indgår i en sammenhæng.

Der vil være mulighed for at zoome ind på visualiseringen, således at man kan bevæge sig ned på et mikroplan. På mikroplanet vil hver enkelt partikel have en farve og en størrelse. Størrelsen vil afhænge af indlæggets popularitet, eksempelvis målt på antal læsninger eller svar. Helt nye indlæg vil repræsenteres med klare farver, hvor mørke partikler repræsenterer ældre indlæg. Farven vil, som på makroplanet, blive styret af den diskussion, indlægget indgår som en del af. Sammenhængen til de øvrige indlæg i diskussionen angives ved en tynd streg. Da hver partikel afgiver en lille smule lys, skabes der nogle effekter. Først og fremmest vil populære emner som sagt være større lyse mere op. Samtidig vil partikler, der er i nærheden af hinanden forstærke

lyset for hinanden og dermed skabe mere lys. På den måde vil dels enkelte partikler og dels meget aktive diskussioner producere meget lys og på den måde være synlige 'på afstand'.



FIGUR 37 -VISUALISERING PÅ MIKROPLAN: HER ER DE ENKELTE INDLÆG TYDELIGT MARKERET MED DERES TITEL.

For at forstærke effekten af hver enkel partikels identitet er det vigtigt ikke at repræsentere partiklerne som dele af større tegn. Det ville være oplagt at bruge et solsystem som metafor på mikroplanet, men dels for at undgå, at nogle indlæg skal være centrum for andre indlæg (at ingen indlæg bliver *Solen*) og dels for ikke at lade partiklerne degradere til del-enheder af et udtryk, har jeg fravalgt at benytte solsystemet som metafor på mikroplanet. Med linierne, der forbinder hver enkelt indlæg i diskussionen, kan man sige, at der skabes en slags stjernetegn, men da det vil være et tilfælde, hvis sådanne konstellationer skulle ligne reelle stjernetegn, anser jeg ikke dette for at være til fare for hver enkelt partikels identitet.

For enkelhedens skyld vil jeg i denne visualisering ikke benytte mig af forskellige former for partikler, men i stedet holde mig til en enkelt figur. Der er dog i princippet ingen grænser for, hvor mange forskellige typer af figurer, man kan bruge til f.eks. at illustrere forskellige emner,

brugere, sprog eller skrivestil. Ligeledes er det også, som tidligere nævnt, muligt at tildele partiklerne liv gennem bevægelse. I denne visualisering har jeg dog valgt ikke at benytte mig af bevægelse for overskuelighedens skyld.

### 8.5.1 DYNAMISK GRUPPERING

Jeg har tidligere været inde på, at semantiske visualiseringer nødvendigvis ikke kan beskrive en sandhed, idet de begreber, vi bruger, besidder en *fuzziness*, jf. Murphys eksempler med, hvorvidt et bilsæde er et møbel, eller om tomater er grøntsager og/eller frugter. Det er derfor også vigtigt, at man ikke lader sin visualisering blive 'for statisk', og at der er mulighed for at bruge visualiseringen på forskellige måder. Dette kan f.eks. foregå ved at lægge forskellige fortolkninger ned over visualiseringens måde at repræsentere data på, eller ved at manipulere med repræsentationen ved f.eks. at fremhæve særlige emner.

Fortolkningsrammer kan indføres ved at krydse forskellige former for semantiske grupperinger, eksempelvis gennem prædefinerede kategorier. I forbindelse med visualiseringen af dette politiske debatforum kunne dette eksempelvis foregå ved at gruppere indlæggene i forhold til de politiske grupperinger, der findes i det danske politiske system - f.eks. i forhold til ministerier. I forhold til univers-metaforen kunne galakserne således repræsentere ministerier, og hver enkelt diskussion eller indlæg ville således tilhøre et ministerium mere end et andet.

Ved denne form for sidestilling af emner efter specifikke kriterier, kunne det betyde, at emner, der normalt ikke virker tæt relaterede, såsom militære investeringer og SU til de studerende, pludselig blive meget tæt forbundne, fordi de begge er en del af Finansloven. Her kan man naturligvis overveje, hvorvidt det kan være relevant at inddrage fagekspertes, til at bidrage med idéer til grupperinger (i forhold til politik kan det eksempelvis være ministerielle, fløj-mæssige, økonomiske eller andre former for grupperinger) i samspil med amatørers grupperinger (som kunne være mindre fagspecifikke, f.eks. i stil med White | Spectrum, der er grupperet efter følelser).

Med den rette semantiske teknologi er der i og for sig intet til hinder for at bruge et hvilken som helst grupperingsgrundlag – endda et skabt af brugeren selv, således at brugeren selv kan gå på opdagelse i, hvordan forskellige måder at gruppere indholdet på giver forskellige resultater. Man kan så at sige definere sine egne prototyper for kategorierne. Denne tilgang, at man selv kan skabe fortolkningsrammen for visualiseringen af de data, der bliver vist, udtrykker i sig selv et budskab til brugeren af systemet om, at betydning ikke blot er betydning, men at ethvert tegn forstås i en kontekst.

### 8.5.2 DYNAMISK FILTRERING

Man kan også forestille sig andre former for interaktion med visualiseringen, eksempelvis gennem filtrering. Filtrering adskiller sig fra gruppering ved ikke at lave opdelinger af indholdet, men i stedet udvælge enkelte data og fremhæve disse.

Filtrering kan f.eks. foregå ved, at man definerer et udvælgelseskriterium og fremhæver de elementer, der lever op til dette kriterium. Det er eksempelvis en filtrering, der gør, at man i We Feel Fine kan få vist, hvad kvinder i aldersgruppen 30-39 har skrevet af følelser, mens det var regnvejrs. De kriterier, der er defineret i We Feel Fine, er prædefinerede. Dvs. at programmøren af systemet har udvalgt nogle kriterier, som brugeren har mulighed for at til- eller fravælge på

det repræsenterede datasæt. Dette er i dag en nødvendighed, da de færreste brugere er i stand til selv at formulere forespørgsler, som softwaren kan forstå. De fleste databaser er i dag baseret på SQL<sup>134</sup>, og at søge i sådan en database kræver enten, at der er udarbejdet en brugergrænseflade, hvor brugeren kan til- og fravælge kriterier, der stemmer overens med de datarækker, der findes i SQL-databasen, eller indgående kendskab til SQL-sproget.

Såfremt udviklingen af semantic software tillader en bedre fortolkning af naturligt sprog, giver det nye muligheder for at benytte sig af database-forespørgsler uden brug af SQL-statements. Det vil da kun forudsætte, at brugeren ved, hvilke data der er til rådighed at søge i. Findes der f.eks. data om køn og alder, vil en bruger kunne nøjes med at skrive *"Giv mig alle de indlæg, der handler om udenrigspolitik, skrevet af mænd i alderen 40-65 år"*. I dag vil det kræve kendskab til databasens opbygning at forbinde de rigtige tabeller og udføre en sætning i stil med *"SELECT title, text, url FROM post, author WHERE post.uid = author.id AND author.sex = 'male' AND author.age <= 65 AND author.age >= 40 AND post.text LIKE '%udenrigspolitik%'"*<sup>135</sup>.

At give brugerne denne magt til frit at kunne søge i datasættet giver naturligvis en række brugervenlighedsudfordringer, såsom hvordan man viser brugerne, hvilke typer af data der kan filtreres for, eller hvorvidt brugerens opfattelse af sin egen forespørgsel stemmer overens med systemets tolkning. Og hvordan sikres det, at der ikke opstår misforståelser? Jeg vil ikke diskutere dette nærmere i dette projekt, men blot angive muligheden for, at brugerne får større mulighed for at behandle data frit uden at have kendskab til teknisk formulerede forespørgsler.

En alternativ mulighed for dynamisk filtrering er at fabrikere en menu, hvor alle muligheder for filtrering fremgår, og hvor man kan vælge at skabe sin egen filtrering ved at kombinere de forskellige muligheder. Metoden svarer grundlæggende til den i We Feel Fine. Det kunne dog være interessant at visualisere filtreringen, således at filtreringsprocessen blev mere synlig. Dette kunne f.eks. ske ved, at man kunne slå filtreringen til og fra og se, hvordan de bortfiltrerede emner falmer eller genvinder deres lysstyrke. På den måde får man mulighed for at skabe et overblik over, hvilke og hvor mange partikler der filtreres fra. Ligeledes kunne det være interessant at arbejde med filtreringen som et værktøj, man bruger til at manipulere med visualiseringen direkte. F.eks. som en magnet, der opsamler alle de filtrerede partikler og samler dem i en særlig gruppe. Dette giver helt nye muligheder for at udforske data grafisk og sammenligne mængder og grupperinger.

### 8.5.3 MASHUPS

I forbindelse med Web 2.0 er mashups blevet en mere almindelig måde at behandle data på. Grundprincippet bag mashups er at blande to forskellige datakilder. Det kan f.eks. være at ordne antallet af forlovelser med statistik over temperatur, men det kan også være at blande to forskellige sange sammen til en enkelt sang (i Danmark også kendt som bastardhits). Mashups er dog ofte blandingen af geografiske data og andre typer af data, visualiseret på et kort.

---

<sup>134</sup> Structured Query Language, et forespørgselsprog brugt i forbindelse med relationelle databaser.

<sup>135</sup> En sådan SQL-sætning vil også kun finde sætninger hvor ordet 'udenrigspolitik' faktisk indgår i.

Eksempelvis et kort, der viser, hvor der er begået kriminalitet i Chicago<sup>136</sup> eller et kartogram, der manipulerer landenes reelle størrelser for at angive deres betydning på forskellige områder<sup>137</sup>.

Denne form for tankegang kunne være interessant at inkorporere i visualiseringen af borgeres diskussioner og dermed skabe nye former for fortolkninger på data. Eksempler på sådanne typer af mashups kunne f.eks. være i forbindelse med kort (hvor er indlægget skrevet, eller hvilken lokalitet handler de om?), men det er i virkeligheden kun fantasien, der sætter grænser. Det eneste krav er, at de to datakilder, man blander, på en eller anden måde kan forbindes. De indlæg, der ligger til grund for en visualisering som denne, vil typisk indeholde en del metadata, som f.eks. tidspunktet for hvornår indlægget er blevet postet, hvor langt indlægget er, hvor mange andre indlæg, forfatteren har lavet, og gennem semantisk analyse af indholdet kan endnu flere data findes frem: hvad handler indlægget om, hvor mange indlæg om samme emne findes der, nævnes der nogen navne, steder, tidspunkter, firmaer eller partier osv. Således kan man, på de indlæg hvor det fremgår, lave en mashup, der sammenligner, hvor ofte de forskellige politiske partier nævnes i forhold til de stemmer, de får, eller man kan måske finde sammenhænge mellem emner, man ikke havde forventet.

At lave mashups er en metode til at se på data med nye øjne, og de er derfor oplagte at tænke ind i det system, der skal skabe debat eller overvejelser om politik.

## 8.6 ANVENDELSE AF SEMANTISKE VISUALISERINGER

Det designforslag til en visualisering, jeg har præsenteret i rapporten, er forholdsvis generisk – den kan bruges på mange forskellige former for datasæt og med mange forskellige formål. Dette skyldes bl.a., at den er designet med det formål at være et værktøj til at udforske data ved at ændre fortolkningsrammen. En ændring af fortolkningsrammen kan dels foregå gennem en filtrering og dels gennem at gruppere ud fra bestemte kriterier. Derfor kan visualiseringen bruges til mange ting. Jeg vil i dette afsnit komme med en række visioner for, hvordan denne type af visualiseringer kan udnyttes konkret. I enkelte af disse eksempler kan man dog med fordel udforme yderligere funktionalitet.

### 8.6.1 ALTERNATIV BRUGERGRÆNSEFLADE WEB 2.0 LØSNINGER

Hvis vi starter på det mindst teknologikrævende niveau, kan man forestille sig, at visualiseringer baseret på semantiske analyser kan benyttes som en slags alternativ navigation til hjemmesider med store mængder brugergenereret indhold, som f.eks. debatfora eller blogs. I dag er der de såkaldte tag- eller search-clouds, der i en minimalistisk visualisering viser, hvilke tags eller søgeord der er mest brugt på sitet. Disse lider dog under de samme svagheder som nøgleordssøgninger, nemlig at ord som frakke, jakke, vinterjakke, sommerjakke, cottoncoat og andet overtøj vil blive præsenteret som lige så forskellige emner som *hund* og *invasion*, mens *Lotus* vil dække over både bilproducenter, blomster og softwarefirmaer. Ved at anvende en semantisk baseret visualisering vil det være muligt at lave grupperinger efter relevante semantiske kategorier, samtidig med at man har mulighed for at se størrelsesforholdene mellem de forskellige grupper. Denne form for visualiseringer vil ikke blot kunne hjælpe brugere med at finde indhold af den type, de søger, men også kunne give brugeren et overordnet indtryk af, hvad

---

<sup>136</sup> <http://chicago.everyblock.com/crime/>

<sup>137</sup> <http://www.worldmapper.org/>

indholdet på den givne hjemmeside er: Hvad diskuteres der på dette site? Hvilke emner er mest diskuteret? Og hvad rører der sig lige nu?

### *8.6.2 STATISTISK VÆRKTØJ TIL SYSTEMADMINISTRATORER AF WEB 2.0-LØSNINGER*

Ved at tilføje visualiseringen værktøj, som f.eks. muligheden for at se præcis hvor mange indlæg der findes om særlige emner og med muligheden for at indføre en tidsmæssigt perspektiv, gives der også mulighed for at udnytte denne form for visualisering som et statistisk værktøj for administratorer af webstedet. Det vil ikke blot være muligt at se, hvilke områder der diskuteres mest (og dermed kan antages at være gode indsatsområder), men også hvordan diskussioner udvikler sig. Denne form for indblik i brugernes aktivitet kan være værdifuld, når man skal udvikle sit site, for at få flere besøgende eller flere deltagende.

Som administratorværktøj kan semantic software også hjælpe med at gøre arbejdsbyrden ved en meget almindelig type opgave væsentlig mindre, nemlig moderation. Som regel, når man administrerer et website med meget brugergenereret indhold, er der et behov for at moderere indholdet, således at brugerne ikke udsættes for ubehagelige indlæg, chikane, mobning eller henvendelser fra eksempelvis pædofile. Ved hjælp af semantic software er det nemmere at lokalisere ubehageligheder uden at begrænse sig til at lede efter visse nøgleord. I en visualisering kan sådanne emner lyse op med en markant farve eller tildeles et iøjnefaldende bevægelsesmønster, således at det fanger administratorens blik. Administratoren kan derpå se nærmere på sagen og vurdere, om der er behov for moderation.

### *8.6.3 I FORBINDELSE MED BLOG-OVERVÅGNING ELLER MARKEDSUNDERSØGELSER*

Blog-overvågning er et emne, der får mere og mere opmærksomhed, da flere og flere virksomheder anser det for at være et behov at vide, hvad brugere skriver om deres produkter. Blogovervågning foregår i dag typisk ved, at man har en database over blogs, som løbende opdateres. Blogs'ne på listen gennemkøres jævnlige for at lede efter særlige nøgleord, f.eks. produktnavne eller virksomhedsnavne. I systemet We Feel Fine var nøgleordene "I feel" eller "I am feeling". Endnu en gang er der tale om et eksempel, hvor de eksisterende systemer er afhængige af nøgleord frem for at benytte sig af en semantisk analyse. Desuden er det svært at se, hvordan emner er sammenkædet, eller hvordan de indlæg, der f.eks. handler om det produkt, man ønsker at overvåge, relaterer sig til andre indlæg.

### *8.6.4 LØSNING AF POLITISKE PROBLEMER*

De ovenstående løsninger udnytter kun en brøkdel af det potentiale, der ligger i at kunne visualisere semantiske analyser på store mængder brugergenererede data. Ingen af dem drager for alvor nytte af den kollektive intelligens. Man kan med Surowieckis begreber sige, at ovenstående løsninger hverken løser kognitions, kooperations eller koordineringsproblemer, men at de nærmere fungerer som en slags statistisk overblik og overvågning af store datasæt.

En måde at udnytte denne form for visualiseringer på, er at bruge dem i forbindelse med kognitionsproblemer. I min beskrivelse af forslaget til udarbejdelsen af en visualisering tog jeg udgangspunkt i et debatforum for politiske emner. Som jeg har skrevet tidligere, tilbyder visualiseringer og semantiske analyser ingen nemme svar. Det er ikke som i en gammel tegnefilm, hvor en stor computer får en masse input og efter at have arbejdet på tallene spytter en lille strimmel ud med svaret på universets gåde. Til gengæld kan visualiseringen af

semantiske data vise sig at være et godt værktøj til at undersøge brugernes stemmer og holdninger, at sammenligne mange indlæg på en gang og at gruppere indlæg efter kriterier, som brugeren af visualiseringen selv bestemmer. I forrige kapitel nævnte jeg, hvordan denne form for undersøgelse af data kan siges at have en abduktiv karakter. Der er altså ikke tale om, at visualiseringen konkluderer, hvad den bedst tænkelige løsning er, men nærmere at den giver brugeren mulighed for at betragte den kollektive stemme fra forskellige vinkler.

En kritisk røst vil måske undre sig over, om det er nødvendigt med en sådan visualisering - om ikke det er tilstrækkeligt blot at kigge på de rå data? Hvis vi alligevel har den semantiske analyse, som vi her antager er temmelig god, kan vi da ikke blot tælle sammen: Hvor mange mener det ene og hvor mange mener noget andet? Her må vi blot stille os selv det spørgsmål, hvad det er for en type data, vi ønsker at repræsentere. Hvis vi blot ønsker at vide, hvor mange der skriver om et emne, og hvor mange der skriver om et andet, så jo, så er visualiseringen fuldstændigt unødvendig. Men hvis vi ønsker at kunne overskue forholdet mellem de semantiske emner, og hvis vi ønsker at bibeholde den *fuzziness*, der findes, når vi taler om begreber, så bliver vi også nødt til at forholde os til, at indlæg ikke *enten* tilhører det ene emne eller det andet, men at det relaterer sig til *alle* de forskellige emner. Skal vi overskue rå data, hvor de enkelte indlæg, ud over at være relateret til et eller flere emner, også skal vise os 'hvor tæt knyttet' indlægget er til de forskellige emner, bliver der hurtigt så mange variable, at de bliver svære at overskue i en tabel eller som talrækker. Visuelt er det dog ingen sag at repræsentere disse mængder af variable, og det er væsentligt nemmere at overskue dem.

#### 8.6.5 INDBLIK I KULTURELLE FORSKELLE

Visualiseringen giver også mulighed for at aggregere emner med henblik på at løse kooperationsproblemer. Ved at benytte globale datasæt, og med muligheden for semantiske analyser, der går på tværs af sprog, vil det måske i højere grad end i dag blive muligt at identificere kulturelle forskelle, eksempelvis hvad angår spørgsmålet om, hvilke emner, der er mest diskuteret i forskellige kulturer. Her er der altså tale om en slags etnografisk vinkel på de emner, der diskuteres. Muligheden for at identificere kulturelle forskelle vil måske give anledning til at adressere sådanne forhold.

Igen bør det pointeres, at visualiseringen i sig selv ikke adresserer koordinationsproblemer, men at den kan fungere som en aggregator, der kan illustrere kollektive stemmer og behandle dem gennem filtreringer og grupperinger. Det kan eksempelvis være interessant at anskue alle brugergenererede indlæg på WWW - i sig selv en meget stor opgave - og visualisere disse uafhængigt af sprog og nationalitet. Herudover kan man så indføre værktøj, eksempelvis give forskellige verdensdele, forskellige religioner, politiske fløje el. lign. forskellige farver, og dermed foretage en global analyse gennem brugernes kommunikation.

#### 8.6.6 UDVIKLING AF PRODUKTER

Jeg har nu nævnt forslag til, hvordan visualiseringen af semantisk analyseret data kan bistå til at afhjælpe både kognitions og kooperationsproblemer. Jeg vil derfor også komme med et enkelt eksempel på, hvordan visualiseringer kan tænkes at bruges i forbindelse med koordinationsproblemer.

Ved at undersøge, hvordan brugere omtaler et firmas produkter, og ved at lave fora og kommentarmuligheder, der lader forbrugere og andre brugere bidrage med forslag til udvikling af de produkter, der produceres, er der mulighed for at udnytte den kollektive intelligens direkte i sin produktudvikling. Allerede i dag benytter mange firmaer sig af denne tankegang, men idet der ikke er mulighed for at aggregere de enkelte bidrag, har de ikke mulighed for at aflæse bidragene som en kollektiv intelligens, og er i stedet nødt til at tage stilling til hvert enkelt bidrag. En visualisering vil her kunne skabe det overblik.

For at en sådan løsning skal give mening, kræves det naturligvis store mængder af input fra brugerne, som kan være svært at få fat i, med mindre man er producent af et meget omtalt produkt. Denne form for løsning henvender sig derfor primært til meget store virksomheder.

## 9 ANTI-KONKLUSION

Da dette projekt tilskriver sig en abduktiv videnskabsteoretisk tilgang, vil det være misvisende at kalde det sidste afsnit i rapporten for en konklusion. Konkluderes skal der heller ikke i en klassisk forstand. Kapitlet består derfor af en opsummering samt en perspektivering med opfordring til yderligere arbejde med visualiseringer baseret på semantiske data.

### 9.1 OPSUMMERING

Først og fremmest argumenterer James Surowiecki for, at der må eksistere nogle grundvilkår for kollektiv intelligens, eller *Wisdom of crowds*, som han kalder det. Disse grundvilkår handler grundlæggende om, at de individer, der tilsammen udgør 'crowds', skal være så alsidige og uafhængige af hinanden som muligt. De bør ikke indgå i et fælles system og bør ikke have mulighed for at påvirke hinanden. Når disse grundvilkår er på plads, mangler der dog stadig et enkelt element: En aggregator, der giver mulighed for at samle de individuelle bidrag til en samlet stemme.

World wide web tilbyder både diversitet, decentralisering og uafhængighed i den forstand, at enhver med internetadgang har mulighed for at bidrage med indhold. Ingen af de tre vilkår kan garanteres i online-verden. Idet langt det meste indhold er offentligt tilgængeligt, kan flere brugere have set det samme indhold og på den måde blive påvirket i en bestemt retning. Men takket været world wide webs uoverskuelighed og størrelse spiller sådanne forhold en stærkt begrænset rolle i forhold til andre variable, som f.eks. påvirkning fra tv eller aviser. Web 2.0-paradigmet har altså medført en god grobund for kollektiv intelligens, men tilbyder ikke umiddelbart en løsning til at aggregere de store mængder af viden. Fokus har ligget på, hvad Spivack kalder *social connectivity* og *frontend* frem for *information connectivity* og *backend*.

Semantic software menes af mange at spille en afgørende rolle på nettet i fremtiden, og nogle kalder ligefrem Semantic web for Web 3.0. Muligheden for at behandle indhold semantisk giver nye muligheder for at behandle de store mængder af brugergenereret indhold på nettet og kan meget vel tænkes at kunne bruges i forbindelse med aggregering af *Wisdom of crowds*. En grov beskrivelse af semantic software kan opdeles i to lejre; en der forsøger at gøre data klogere, og en der forsøger at gøre softwaren klogere. Man kan naturligvis ikke gøre data klogere, men udtrykket er et billede på, at man tilknytter data tilstrækkelige mængder af metadata, således at software kan behandle ustrukturerede data som var det struktureret. Hvor Semantic web er opbygget omkring en ontologi, en slags definitionsordbog, og ved at opmarkere data med metadata, giver man teksten mening i forhold til ontologien. Det er således i sidste ende denne ontologi, der styrer meningen af den tekst, der er opmarkeret. Omvendt benytter teknologier som Natural Language Processing sig af grammatiske analyser kombineret med en form for kunstig intelligens og erfaringer med andre sætninger til at analysere sætningers sammensætninger og danner en mening ud fra teksten baseret på en avanceret algoritme. Man kan sige, at hvor den ene metode (den metadataafhængige) benytter sig af, hvad begrebspsykologien kalder den klassiske tilgang, benytter den anden (den intelligente) sig af en eksemplartilgang, hvor øget erfaring med semantisk analyse forbedrer systemet løbende. Den metadataafhængige fremgangsmåde har den fordel, at den teknisk er nem at implementere, men den kræver enorm vedligeholdelse og en ny måde at producere indhold på. Herudover kan det være problematisk at have en semantisk analyse, der baseres på en begrebsforståelse svarende

til den klassiske tilgang, da denne ikke stemmer overens med menneskers begrebsforståelse. Den softwareafhængige form for semantic software har en bedre tilgang til begreber, men er krævende hvad angår systemressourcer, teknologien er ikke tilgængelig for enhver og er - til trods for store fremskridt - ikke en moden teknologi, der kan anvendes på så store, ukontrollerede datasæt som f.eks. world wide web. Selvom udviklingen retter sig mod bedre og mere sikre semantiske softwaresystemer, er vi der ikke endnu. Der er dog ingen tvivl om, at der hele tiden sker forbedringer, og det er sandsynligt, at vi inden for en nær fremtid vil kunne se semantiske webservices baseret på en form for Natural Language Processing.

Uanset hvilken teknologi, der bliver dominerende inden for semantic software, er det relevant at stille spørgsmål til, hvordan teknologien skal bruges, og hvordan vi behandler de data, der bliver produceret af semantic software. En af de største problemstillinger er, hvordan vi præsenterer data, der kan fortolkes subjektivt uden at gøre vold på dem. Jeg foreslår, at man ser på brugen af visualiseringer for at skabe en repræsentation af data, hvor fortolkningsrammen er foranderlig. I modsætning til en kvantificering af data kan man med visualiseringer vise tvetydigheder og *fuzziness*, og på den måde være tro mod data. Visualiseringer evner at gøre komplekse datasæt mere overskuelige, og de giver mulighed for at udforske data. Ved at designe sin visualisering på både et makro- og et mikroniveau gives der også mulighed for at beskue de samme data fra forskellige perspektiver, eksempelvis hvordan de enkelte emner relaterer sig til hinanden, og hvor mange emner der handler om de samme ting.

Jeg har præsenteret et forslag til en visualisering, der udnytter dette mikro/makro-forhold, og som balancerer mellem en datanærhed og en udtryksfuldhed, samtidig med at den præsenterer data på en forholdsvis gennemskuelig måde. Jeg har perspektiveret denne visualisering i forhold til brug af visualiseringer generelt, som en metode til at høste og udnytte og kollektiv intelligens som det kommer til udtryk på World Wide Web.

Det naturlige sprog vil altid være genstand for en fortolkning, der er afhængig af kontekst, og den bedste måde at repræsentere denne form for data er derfor at gøre det i et miljø, hvor der kan eksperimenteres med kontekst, og dermed danne nye forståelser af det repræsenterede. I denne sammenhæng er jeg kommet frem til, at man ikke umiddelbart kan bruge semantic software til at skabe konkrete løsninger eller endegyldige svar, men at man kan bruge det som et værktøj til at overskue store mængder af brugergenererede data, hvilket på abduktiv vis kan føre til bedre forståelser og muligvis skabe bedre vilkår for at kunne træffe gode beslutninger.

## 9.2 FREMTIDIGT ARBEJDE

Teknologierne inden for semantic software er stadig under udvikling - og faktisk har den langt igen. Derfor er det til stadighed usikkert, hvilke teknologier man bør satse på. På den korte bane kan man forvente, at Semantic web vil vinde frem, da den er billig, offentlig og forholdsvis nem at implementere. På længere sigt har de grammatiske analyser kombineret med kunstig intelligens en række fordele, der langt overstiger Semantic webs styrker, bla. ved at være stærkere på områder som skalerbarhed, fleksibilitet og selvvedligeholdelse.

Uanset hvordan det semantiske felt udvikler sig, er der dog en række undersøgelser, der ville være relevante at tage fat på allerede i dag. Dels fordi resultatet af sådanne undersøgelser har en generel værdi, som ikke nødvendigvis kun er anvendelig inden for visualiseringer baseret på

semantiske analyse, men som meget vel kan vise sig at være generelt anvendelige på visualiseringer af store datasæt. Her tænker jeg særligt på brugerundersøgelser, og i første omgang vil jeg pege på følgende to forskningsområder:

### 9.2.1 HVORDAN FORSTÅR BRUGERNE VISUALISERINGER

Jeg har i dette projekt abduktivt behandlet problemerne med at visualisere store mængder af data. Men hvordan forstår brugerne visualiseringer, der ikke er baseret på 'rå data', såsom tal? Kan brugerne afkode visualiseringernes *fuzziness*, og hvilke tolkninger lægger de over hhv. mikro- og makroplan?

Herunder er det også relevant at undersøge, hvor mange typer af variable brugere kan håndtere: Er der en øvre grænse for, hvor mange forskellige visuelle udtryk man kan påhæfte værdier, såsom farveværdi, mæthed, lysstyrke, form, bevægelse, baggrundsfarve, størrelse, placering osv.? Det bør også undersøges, om der er visse former for visuelle udtryk, der passer særligt godt på enkelte variable. Gestaltpsikologien beskriver nogle sådanne regelmæssigheder, eksempelvis at elementer, der placeres i nærheden af hinanden, opfattes som at være grupperede eller have noget tilfælles. Men hvilken form for fællesskab illustreres bedst ved en sådan nærhed? Semantisk lighed (som *hunde* overfor *hvalpe*) eller kausal lighed (som når et indlæg er et svar på et andet)?

### 9.2.2 HVORDAN FORSTÅR BRUGERNE SEMANTIC SOFTWARE

Jeg vil foreslå, at der forskes i brugeres forståelse af eksempelvis semantiske søgemaskiner og semantiske data. Nøgleordssøgninger har sine svagheder men er trods alt en meget transparent metode: Det er nemt for brugeren at gennemskue, at man søger på et eller flere ord, der indgår i den tekst, man leder efter. Men hvordan forholder det sig med semantic software? Kræver det en større forståelse af teknikken bag søgemaskinen (eller den form for semantic software, der nu engang er tale om) at benytte sig af den?

### 9.2.3 KOMBINATION AF VISUALISERINGER MED ANDRE TYPER AF SERVICES

Visualiseringer er kun en enkelt blandt mange muligheder, når det gælder om at udnytte semantic software på World Wide Web. Jeg har i rapporten nævnt enkelte andre af disse muligheder, bl.a. de intelligente assistenter, de såkaldte *personlige agenter*. Jeg har i rapporten skrevet en del om at kunne forbinde forskellige typer af data, og enkelte steder om muligheden for at forbinde mennesker. Web 2.0 har vist sig at gøre både de datamæssige og de sociale forbindelser stærkere, og Semantic web bygger videre på disse potentialer for WWW. Man bør dog ikke glemme, at denne nye semantiske teknologi også åbner op for en helt ny verden af forbindelser: nemlig forbindelsen af services, dvs. muligheden for at forbinde software på internettet. Det kan eksempelvis tænkes, at en service, der skal finde den billigste forhandler af et produkt, også kan finde den korteste rute til denne forhandler eller andre brugeres erfaringer med at handle hos netop den forhandler.

I denne sammenhæng er det oplagt at undersøge, hvorvidt visualiseringer kan spille en rolle i forbindelsen mellem webservices. Kan man f.eks. forestille sig, at det var nemmere at overskue forhandlere, priser, afstande og brugerbedømmelser i en visualisering, hvor priser udgør en akse, afstande en anden og hvor farven på forhandlerens navn præsenterer brugernes bedømmelse af denne forhandler (grøn for positiv og rød for negativ)? Og kan man forestille sig

en dataafhængig visualiseringsmodel, som brugeren selv definerer? Altså at man i stedet for at have en fast visualisering over produkter og afstande til forhandlere havde et visualiseringsværktøj, som brugeren selv kontrollerede? Det kunne f.eks. implementeres ved, at brugeren angav et søgekriterium, lad os sige titlen på en CD, og derefter valgte en række variabler: forhandler, pris, afstand fra mit hjem og bedømmelse. Herefter ville brugeren få visualiseret netop dette og hurtigt få et overblik over markedet og nemt vælge en forhandler med en god balance mellem pris, bedømmelse og afstand fra hjemmet.

Dette er blot et eksempel på den form for tænkning, der er plads til, når en ny teknologi vinder indpas. En undersøgelse af dette felt vil formentlig have en abduktiv karakter men vil kunne resultere i konkrete, innovative interaktionsdesigns, som ville skulle testes på brugere.

#### 9.2.4 NETDEMOKRATI

I kapitel 3 nævnte jeg, at 9 procent af webbrugerne kan forventes at deltage aktivt på nettet fra tid til anden, og kun 1 procent er flittige deltagere af f.eks. debatfora, blogs eller anden brugerdeltagelse på World Wide Web. Særligt i forbindelse med borgerinddragelse på internettet er det relevant at undersøge, hvordan disse i alt 10 procent af brugerne repræsenterer samfundet. Er der tale om en repræsentativ gruppe, eller er det en helt bestemt type brugere?

En undersøgelse af internetbrugere, brugeres deltagelse på World Wide Web, og hvorvidt man kan antage, at de aktive webbrugere er repræsentative for større grupper (eksempelvis alle danskere eller alle med stemmeret), ville bero på dels induktive og deduktive undersøgelser.

### 9.3 AFSLUTNING

Pendulet er i fuld fart mod et mere backend-orienteret World Wide Web, hvor fokus i denne omgang er at blive bedre til at forbinde både data og mennesker. Men pendulet vil svinge tilbage igen, og når det sker, vil brugerne endnu engang mærke et teknologisk spring. Jeg har præsenteret et forslag til, hvordan de data, der høstes gennem brugen af semantic software, kan visualiseres, således at brugerne i langt højere grad kan få adgang til og drage nytte af data, som man før primært havde adgang til via nøgleordssøgninger og tagging. Men dynamiske visualiseringer af data er kun en enkelt blandt mange kommende services, der har potentiale i forbindelse med semantic software. Det afgørende spørgsmål er, hvordan vi vælger at bruge sådanne værktøjer, og om vi evner at udvikle services, der for alvor kan sætte pendulet i svingning til brugernes fordel.

# 10 LITTERATUR

## 10.1 PRIMÆR LITTERATUR

**Abbate, Jane (2000):** *Inventing the internet*, s. 181-258: kap. 6. *Popularizing the internet*. MIT Press

**Chang, Dempsey; Dooley, Laurence; Tuovinen, Juhani E. (2002):** *Gestalt Theory in Visual Screen Design – A New Look at an Old Subject*, ACM International Conference Proceeding Series; Vol. 26, <http://portal.acm.org/citation.cfm?id=820062>

**Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001):** *The Semantic Web*. ScientificAmerican.com <http://www.sciam.com/article.cfm?id=the-semantic-web>

**Kjørup, Søren (2000):** *Menneskevidenskaberne – Problemer og traditioner i humanioras videnskabsteori*. Roskilde Universitets Forlag

**Kjørup, Søren (2002):** *Semiotik*. Roskilde Universitetsforlag (Samfundslitteratur), Frederiksberg

**Köhler, Wolfgang (1972):** *Getaltpsykologi*. Stjernebøgernes Kulturbibliotek, Vintens forlag

**Morville, Peter & Rosenfeld, Louis (2007):** *Information for the World Wide Web – third edition*. O'Reilly

**Murphy, Gregory L. (2004):** *The Big Book of Concepts*. A Bradford book, MIT Press

**Nathaniel Good, J. Ben Schafer, Joseph A. Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, and John Riedl (1999):** *Combining Collaborative Filtering with Personal Agents for Better Recommendations*. <http://www.cs.ubc.ca/~conati/532b/papers/good-konstan-aaai-99.pdf>

**O'Reilly, Tim (2005):** *What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software*. <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>

**Saussure, Ferdinand de (1991):** *Forelæsninger om almen lingvistik i: Det videnskabelige perspektiv. Videnskabsteoretiske tekster* (side 405-443). Schmidt, L.-H. (red.). Akademisk Forlag.

**Schrøder, Lise (2005):** *Appeniks 2: Peirces Semiotik i: Tegn fortælling og metaforståelse – et perspektiv på informationsdesign og bygningsdata*. Upubliceret ph.d.-afhandling. Aalborg Universitet. [www.plan.aau.dk/~lisesch](http://www.plan.aau.dk/~lisesch)

**Shank, Gary & Cunningham, Donald J. (1996):** *Modeling the Six Modes of Peircean Abduction for Educational Purposes* <http://www.cs.indiana.edu/event/maics96/Proceedings/shank.html>

**Shirky, Clay (2008):** *Here Comes Everybody – the power of organizing without organizations*, Allen Lane

**Spivack, Nova (2003):** *The Birth of "The Metaweb" -- The Next Big Thing -- What We are All Really Building*

[http://novaspivack.typepad.com/nova\\_spivacks\\_weblog/2003/12/the\\_birth\\_of\\_th.html](http://novaspivack.typepad.com/nova_spivacks_weblog/2003/12/the_birth_of_th.html)

**Spivack, Nova (2004):** *The Metaweb is Coming... See this Diagram...*

[http://novaspivack.typepad.com/nova\\_spivacks\\_weblog/2004/03/the\\_metaweb\\_is\\_.html](http://novaspivack.typepad.com/nova_spivacks_weblog/2004/03/the_metaweb_is_.html)

**Spivack, Nova (2007a):** *How the WebOS evolves?*

[http://novaspivack.typepad.com/nova\\_spivacks\\_weblog/2007/02/steps\\_towards\\_a.html](http://novaspivack.typepad.com/nova_spivacks_weblog/2007/02/steps_towards_a.html)

**Spivack, Nova (2007b):** *Harnessing the collective intelligence of the World-Wide Web,*

[http://novaspivack.typepad.com/nova\\_spivacks\\_weblog/2008/04/great-collectiv.html](http://novaspivack.typepad.com/nova_spivacks_weblog/2008/04/great-collectiv.html)

**Spivack, Nova (2007c):** *Video of Twine.com Founder Nova Spivack,* <http://digitalnow-resources.blogspot.com/2008/04/video-of-twinecom-founder-nova-spivack.html>

**Spivack, Nova (2007d):** *Web 3.0 Roundup: Radar Networks, Powerset, Metaweb and Others...*

[http://novaspivack.typepad.com/nova\\_spivacks\\_weblog/2007/02/web\\_30\\_roundup\\_.html](http://novaspivack.typepad.com/nova_spivacks_weblog/2007/02/web_30_roundup_.html)

**Spivack, Nova (2008a):** *On the Difference Between "Semantic" and "Semantic Web",*

[http://novaspivack.typepad.com/nova\\_spivacks\\_weblog/2008/05/on-the-differen.html](http://novaspivack.typepad.com/nova_spivacks_weblog/2008/05/on-the-differen.html)

**Spivack, Nova (2008b):** *Video of my Presentation at The Next Web 2008 Conference,*

[http://novaspivack.typepad.com/nova\\_spivacks\\_weblog/2008/06/video-of-my-pre.html](http://novaspivack.typepad.com/nova_spivacks_weblog/2008/06/video-of-my-pre.html)

**Surowiecki, James (2004):** *The Wisdom of the Crowds – Why the Many Are Smarter Than the Few.* Abacus

**Tufte, Edward R. (1990):** *Envisioning Information.* Graphic Press

**Weiguo Fan, Linda & Wallace, Stephanie Rich (2006):** *Tapping the power of text mining,* Communications of the ACM, September 2006/Vol. 49, No. 9, s. 76-82

## 10.2 SEKUNDÆR LITTERATUR

**Allen, James (red.) (2004):** *Challenges in Information Retrieval and Language Modeling,*

<http://wwwhome.ewi.utwente.nl/~hiemstra/papers/ir-challenges.pdf>

**Giles, Jim (2005):** *Internet encyclopaedias go head to head.* Nature, 438(7070):900–901, December 2005

**Hearst, Marti (2003):** *What Is Text Mining?* <http://people.ischool.berkeley.edu/~hearst/text-mining.html>

**Heie, Niels (1999):** *At tale til øjet – en bog om grafisk retorik,* Grafisk Litteratur

**Kjørup, Søren (1982):** *Hvorfor smiler Mona Lisa – en bog om billeder og deres brug.* Gjellerup

**Mark Maybury (2001):** *Human Language Technologies for Knowledge Management: Challenges and Opportunities,* <http://portal.acm.org/citation.cfm?id=1118222>

**Nallapati, Ramesh (2003):** *Semantic language models for topic detection and tracking*, <http://dx.doi.org/10.3115/1073416.1073417>

**Nielsen, Jakob (2006):** Participation Inequality: Encouraging More Users to Contribute, [http://www.useit.com/alertbox/participation\\_inequality.html](http://www.useit.com/alertbox/participation_inequality.html)

**Shirky, Clay (2003):** *A Group Is Its Own Worst Enemy*. [http://www.shirky.com/writings/group\\_enemy.html](http://www.shirky.com/writings/group_enemy.html)

**Tang, Jie; Hang Li, Yunbo Cao; Tang, Zhaohui (2005):** *Email Data Cleaning*, KDD '05 - Conference on Knowledge Discovery in Data, <http://doi.acm.org/10.1145/1081870.1081926>

**Sebastiani, Fabrizio (2002):** Machine Learning in Automated Text Categorization, <http://doi.acm.org/10.1145/505282.505283>

### 10.3 ANDRE VÆRKER

**BBC 2 (2007):** *White | Spectrum*, <http://www.bbc.co.uk/white/spectrum.shtml>

**Dolores Labs (2008):** *The Amazing Color Wheel!*, <http://assets.doloreslabs.com/jobs/colors/explorer/>

**Harris, Jonathan (2007):** *The Web's Secret Stories*, videooptagelse af Harris tale ved TED Talks marts 2007, [http://www.ted.com/index.php/talks/jonathan\\_harris\\_tells\\_the\\_web\\_s\\_secret\\_stories.html](http://www.ted.com/index.php/talks/jonathan_harris_tells_the_web_s_secret_stories.html) - Selve *We Feel Fine* kan opleves på <http://www.wefeelfine.org/>

**Harrison, Chris (2008):** *Visualizing The Bible*, <http://www.chrisharrison.net/projects/bibleviz/index.html>

**Jordan, Chris (2007):** Kunstværket *Paper bags*, <http://www.chrisjordan.com/>

**Newman, Mark; Dorling, Danny; Barford, Anna; Wheeler, Ben; Pritchard, John, Allsopp, Graham:** *Worldmapper*, <http://www.worldmapper.org/index.html>

**Spielberg, Steven (2002):** Filmen *Minority Report*, 20<sup>th</sup> Century Fox

**Wikipedia:** <http://www.wikipedia.org>