Datakvalitet i virksomheden

Karsten Boye Rasmussen

Resumé

Der findes mange fejlagtige data ude i virksomhederne fx: »Fejl i hver femte danskers pensionsopsparing« (DR.dk, 16. feb. 2009). I bogen »Information Revolution« kategoriseres virksomheder efter deres stadie på en »Information Evolution Model« (Davis, 2006, s. 13), hvor niveauerne angiver den stigende strategiske betydning af information. Denne artikel illustrerer gennem en mindre undersøgelse, hvorledes nogle danske virksomheder placerer sig i den igangværende evolution gennem opmærksomhed over for udfordringer, årsager og følger samt den organisatoriske forankring af virksomhedernes processer om datakvalitet.

Virksomheden kan opnå væsentlige fordele gennem veltilrettelagte data-processer for virksomhedens kerneydelser med datakvalitet for øje og kan dermed udvikle sig på »Information Evolution Model« og sikre sig et bedre strategisk grundlag. Forståelsen af datakvalitet i virksomheden skærpes ved mere omfattende tilgange med udfoldning af dimensioner og begreber til belysning af, hvad der udgør god datakvalitet. Artiklens regler er simple og generelle. Virksomhedens opnåelse af ekstra styrke vil ligge i implementeringen.

Indledning

Spørgsmålet om datakvalitet får yderligere relevans i sammenhæng med informationsbølgen og virksomhedernes begrundede frygt for »data overload«, når informationsmængden vokser eksponentielt (Liautaud, 2001, s. 45). Nyligt har det været fremført, at lagringskapaciteten ikke stiger med samme hast (Berman, 2008). Måske kan vi alligevel ikke bevare alt! Den betragtning har givet yderligere næring til vigtigheden af at kunne håndtere, skabe, udvælge og opbevare data af høj kvalitet. Virksomhedernes beslutninger er drevet af data – vi taler om »faktabaserede beslutninger«.

Begrebet data er jordbundet og »nørdet«. Der er ikke meget poesi at hente der! Bob Dylan nævner ikke »data« i sine sange, mens »det ukendte« forekommer flere steder. I virksomhedernes prosaiske verden er det tiltrækkende »det nye« – som leverer muligheder – men ikke »det ukendte« – som leverer farer. Vi agerer for at kende mere til det, der før var ukendt. Kendskabet opstår gennem data og anvendelsen af data; og virksomhederne er helt opmærksomme på det. Begreber og

programmer som »Total Quality Management« og »Six Sigma« er eksempler på vigtigheden af data. Det illustreres yderligere gennem virksomhedernes deltagelse i seminarer, workshops og konferencer om datakvalitet. I forbindelse med nogle seminarer om datakvalitet afholdt af SAS Institute A/S i november 2008 og januar 2009 blev indsamlet oplysninger fra de deltagende firmaer. Datamaterialet er venligst stillet til rådighed af SAS Institute A/S. Skemaet fungerede som seminartilmelding, og datamaterialet er her reduceret til de tilfælde, hvor der ud over de direkte tilmeldingsoplysninger blev afgivet mindst et svar i de i skemaet angivne kategorier. Der er ikke tale om en for Danmark repræsentativ undersøgelse. De deltagende firmaer i undersøgelsen må generelt karakteriseres som værende store, idet 70 pct. af firmaerne har angivet en årlig omsætning på over 1 mia. kr. Dette skyldes dog nok, at SAS software er udbredt i de store firmaer, ikke at det kun er de store firmaer, som oplever problemer omkring datakvalitet.

Begribelsen af datakvalitet illustreres gennem forskellige tilgange med flere dimensioner og forskellige begreber for datakvalitet. Tilgangene (Wang & Strong, 1996) kan opdeles i tre: 1) Den intuitive tilgang, som er pragmatisk og kan karakteriseres som hovedsagelig a-teoretisk, idet målet hverken synes at have teoretisk baggrund eller indeholde en intention om udvikling af teori om datakvalitet. 2) Den empiriske tilgang, som er induktiv, ved at teori eller kategorisering uddrages fra empiriske observationer. Empirien er typisk opnået gennem interview med personer ansvarlige for data i virksomheder. 3) Den teoretiske tilgang, kan også kaldes ontologisk eller deduktiv, ved at konsekvenser uddrages på grundlag af teoretiske grundlæggende antagelser om data i informationssystemer og afbildningen over for den virkelige verden.

Data og beslutninger

Anvendelsen af data bygger grundlæggende på troen om rationalitet og objektivitet. Vi indsamler data for ikke at famle i mørket og handle i uvidenhed. Information om kunden, produktet og aftaler om udvekslingen (fx pris) er nødvendige oplysninger, for at den afsluttende transaktion mellem kunden og virksomheden kan foretages. Internt i virksomheden findes tilsvarende en informationsopsamling, idet kæden mellem virksomhed og kunder og leverandører kan paralleliseres til relationerne mellem virksomhedens afdelinger, grupper og forretningsområder. De interne informationssystemers dataopsamling kan anvendes til mere end kontrol og opgørelser. Gennem dataanalyser vil de faktiske virksomhedsprocesser kunne blotlægges og effektiviseres. I fremstillingen her anvendes kunder og leverandører til illustration.

Adfærden op til den afsluttende transaktion mellem kunde og virksomhed registreres som handlinger foretaget af kunden, hertil kommer generelt beskrivelse af relationen til kunden i form af den historiske dataopsamling og fx yderligere segmentplacering af kunden, alt dette giver virksomheden – om ikke vished – så en viden og en formindskelse af usikkerheden, som er forbundet med virksomhedens ageren på markedet. Baggrunden er altså, at vi gennem anvendelsen af et informationssystem – som indeholder data – ønsker at opnå en større sikkerhed i vores

ageren eller handlen. For datasikkerhed opstilles ofte tre væsentlige karakteristika: konfidentialitet, integritet og tilgængelighed (Pfleeger, 2003, s. 10). Disse karakteristika ønskes opfyldt i passende mængde for alle informationssystemer. Informationssystemet har flere interagerende dele software, hardware, grænseflader, organisationen, etc. med relation til datasikkerhed. Når der fokuseres på datakvalitet i informationssystemerne, vil karakteristika for datasikkerhed også dukke op.

Anvendelsen af data sker både i det operative system og også i et system beregnet til generalisering og læring fra opsamlingen af de mange operative registreringer ved udvælgelse og overførsel til et »data warehouse« (Rasmussen, 2001). Over for beslutningerne i det operative system, der fx tages over for den enkelte kunde, kan stilles et »data warehouse« som »a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions« (Inmon, 1996, s. 33). En anden ekspert vedrørende data warehouse har på samme tid udtrykt: »The data warehouse provides access to consistent organizational data that can be combined for query, analysis, and presentation of published data with a quality that will act as a driver of business reengineering« (Kimball, 1996, s. xxiii-xxv). Udnyttelsen af data warehouse finder sted inden for adskillige områder i organisationen; anvendelsen er ikke blot reserveret en afdeling for kompliceret analyse. Hovedsageligt anvendes et data warehouse til rapportering, herunder ledelsesinformation og visuel præsentation, generel statistik om virksomheden, analyse med henblik på forklaring og data mining med henblik på handlen. Yderligere kan virksomhedens data warehouse have en tilbagevirkende indflydelse gennem vedvarende udvikling og forbedring af virksomhedens operative system.

Virksomhedens beslutninger kan være operationelle eller strategiske, og begge dele træffes på grundlag af data. Uanset datas formål kan begrebet om datakvalitet tilgås på tre måder (Wang & Strong, 1996), som i kort form udgår fra henholdsvis anvendelighed, oplevelse og ontologi.

Anvendelighed – »Fitness for use«

I den pragmatiske tilgang hoppes direkte til datas anvendelighed. Det pragmatiske syn på datakvalitet kommer til udtryk gennem udtrykket »fitness for use« (Bruckner & Schiefer, 2000, s. 36; Wang & Strong, 1996). Er data passende for »datakonsumenten«? (Strong et al., 1997, s. 104). Det pragmatiske findes også understreget direkte i begrebet »pragmatic information quality« (English, 1999, s. 151). Imidlertid, selvom »fitness for use« åbenlyst efterspørges, så gives ingen anvisning på, hvorledes en sådan »fitness« skal angives, måles eller opnås. »Fitness for use« er et slogan ligesom »All the news that's fit to print« fra New York Times. Begrebet har indbygget en selvopfyldende cirkularitet. Hvis nyheden er trykt i New York Times er den »fit« – var den »unfit«, ville den ikke være blevet trykt! Den subjektive beslutning om at trykke nyheden er dermed udslagsgivende.

Datakvalitet kan findes omgærdet af en intuitiv opfattelse af begrebet. Ofte henvises blot til »garbage in, garbage out« (GIGO), som om det er et selvindlysende

udsagn, der gør al argumentation unødvendig (Levitin & Redman, 1998; Berg & Heagele, 1997).

I den pragmatiske anskuelse af brugen af data, kan man også finde forfattere, som drejer fokus til brugerne. Den erkendte »fitness« har dermed en relativitet, som det fx fremgår af citatet: »The single most significant source of error in data analysis is misapplication of data that would be reasonably accurate in the right context« (Levitin & Redman (1998, s. 94) citerer Loebl fra 1990). Her kan siges, at være tale om en »fejl 40«; altså at årsagen til fejl skal findes 40 cm fra skærmen. Fejl ligger altså ikke i data selv, men i subjektet.

Netop fordi data både i det operative system og i data warehouse bringes ind i forskellige sammenhænge (applikationer) og benyttes af mange forskellige brugere (Tayi & Ballou, 1998), så er relativitet og subjektivitet vigtige i den videre undersøgelse. Relativiteten kan stilles på spidsen ved at spørge: »Er uanvendte data data uden kvalitet?« Nok ikke, men de er uden nærværende værdi. Et mere relevant billede kan opnås ved at se på »data life cycle«.

Værdi – i form af penge og med tanke på bundlinjen – er udgangspunktet i forretningspressens anbefalinger om datakvalitet (Kapochunas, 2002; Redman, 2004). Her fortælles ledere om vigtigheden af høj datakvalitet. Under alle omstændigheder er niveauet af datakvalitet absolut værd at diskutere nærmere, idet opnåelse af en højere datakvalitet klart har omkostninger; og disse omkostninger må sammenlignes med den forventede opnåede værdi af den højere datakvalitet. Dette regnestykke er dog særdeles vanskeligt at afslutte, bl.a. fordi de opnåede værdier kan ligge langt ude i fremtiden. Der er lanceret sådanne beregningsformler for at opnå en optimal allokering af ressourcer til forøgelse af kvaliteten af data-samlinger (Ballou & Tayi, 1989). Yderligere har andre udført undersøgelser omkring datamanagement, som resulterer i en modenhedsmodel med det højeste trin som en næsten fuldkommen sikkerhed på udfaldet (D'Angelo & Troy, 2000, s. 43). Hertil knyttede forfatterne dog den kommentar, at der antagelig ikke ville være noget kommercielt marked for denne grad af ydelse i form af datakvalitet. Tilsvarende har andre observeret, at i nogle tilfælde har anskaffelsen af data langt oversteget den totale opnåede gevinst (Trull, 1966, s. 276). Det kan jo sådan set ikke overraske, at der ikke nødvendigvis er en fortsat sammenhæng mellem omkostningerne og gevinsten. Det interessante er, om der ved undersøgelse af forholdet vil vise sig nogle områder, der kan siges at være af fundamental betydning for virksomhedens drift og overlevelse.

Eksakte tal, som giver indtryk af egentlige målinger af datakvalitet, findes ofte i forbindelse med andelen af virksomheder, som har oplevet problemer i forbindelse med datakvalitet. Et studie fra PricewaterhouseCoopers viste, at 75 pct. af 599 undersøgte virksomheder havde gennemlevet økonomiske problemer på baggrund af mangelfulde data (Computerworld.com, December 17, 2001). Den udtrykte prævalens kombineres ofte med yderligere udtryk for voldsomheden af »sygdommen« fx gennem citater som »about 14 % of the potential taxes due are not collected« (Wat-

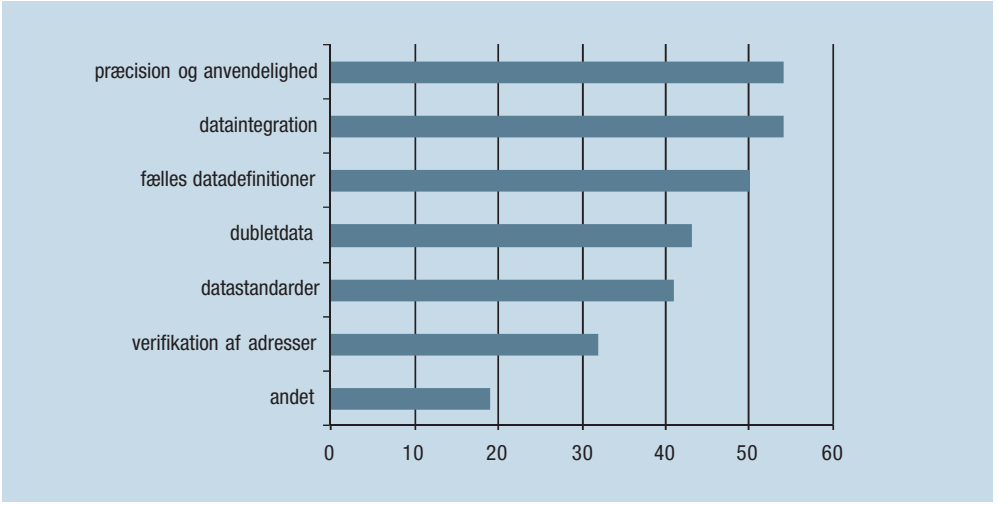
son et al., 2002, s. 496). Disse tilgange findes i adskillige bøger og artikler inden for området; ofte indeholdende adskillige sider blot med overskrifter fra virksomhedskatastrofer på baggrund af datakvalitet (English, 1999, s. 7-10; Huang et al., 1999, s. 2). Det kan antages, at alle virksomheder har oplevet problemer med datakvaliteten – ellers kan man fristes til at antage, at virksomhedens data slet ikke bruges! Utilstrækkelig datakvalitet præsenteres også med sammenregninger som udtrykker tab på samfundsmæssigt niveau: »Poor data quality costs the typical company at least ten percent (10 %) of revenue; twenty percent (20 %) is probably a better estimate« (Redman, 2004). I 1998 estimerede Redman, at tabet generelt vil omfatte omkring 8-12 pct. af overskuddet.

Der forekommer også publikation af fejlrater for det enkelte datafelt, fx at fejlraten skulle være 1-5 pct. (Redman, 1998). Undersøgelser viser, at mellem 20 og 40 pct. af regneark er fejlbehæftede (Davenport & Harris, 2007), hvilket specielt bør bemærkes i mindre firmaer, som i vid udstrækning baserer deres anvendelse af it på regneark. Metrikkerne for den anvendelsesorienterede datakvalitet omfatter således ofte et gæt eller estimat af dårlig datakvalitet i form af 1) omfanget eller udbredelsen af problemet i virksomhederne, 2) den økonomiske effekt og 3) proportionen af fejl i data, som grundlæggende er årsagen til disse dårligdomme.

Udfordret af datakvalitet

I forbindelse med de omtalte SAS-seminarer om datakvalitet besvarede nogle af firmaerne nogle spørgsmål omhandlende datakvalitet. De blev her bl.a. stillet spørgsmålet »Hvilke udfordringer står din afdeling over for?«, som resulterede i følgende procentfordeling for de angivne svarkategorier:

Figur 1. Udfordringer i forbindelse med datakvalitet. Procent. (n= 46, flere svar)



Manglende præcision og anvendelighed ligger højest. Det er spørgsmålet, om data er »fit«. Fra engelsk kan »accuracy« og »precision« begge oversættes til både »nøjagtighed« og »præcision«, men der kan skelnes. Nøjagtighed kan forbindes

til begrebet validitet, dvs. til spørgsmålet, om vi måler det rigtige. Er lønsummen fx et tilstrækkelig godt udtryk for udgiftssiden af virksomheden? Det afhænger af virksomheden! Præcision kan tilsvarende forbindes med reliabilitet, dvs. hvor stor en variation ville vi stå med, hvis vi kunne foretage målingen flere gange. Ofte forbinder vi automatisk præcision med tal og med det tal, der slynges ud. Er lønsummen fx angivet som 12 Mkr. vil vi forvente +/- 1 Mkr. som vores præcision. Var der angivet 12,1 Mkr. ville vi nok forvente 0,1 Mkr. som det mulige udsving. På strategisk niveau er vi interesserede i høj validitet og kan leve med nogle udsving, mens virksomheden på det operationelle niveau skal have styr også på ørerne. Det illustrerer, hvorledes data skal passe til sammenhængen. Mere end halvdelen af virksomhederne er ude for, at data ikke passer til anvendelserne. For visse data er der tale om binære valg. Som Fredsbevægelsen skrev på »bumper stickers« i 80'erne: »One nuclear bomb could ruin your whole day!«.

Manglende dataintegration ligger næsten tilsvarende højt, og i sammenhæng hermed ligger problemerne med, at virksomheden savner fælles datadefinitioner og -standarder. Dernæst døjer virksomhederne med redundans i form af delvise dubletter og også manglende verifikation af adresser. Problemer med dublerede adresser er et af de områder, hvor det typisk er virksomhedens kunder, der opdager problemerne omkring datakvalitet. Den manglende oprydning og sikkerhed giver indtryk af, at virksomheden ikke har styr på sine oplysninger, og det kan negativt påvirke det image, kunderne har af virksomheden. Det er unødvendigt, idet software (fx Dataflux) kan assistere virksomheden med oprydning i registrene. Tilsvarende bør virksomheder også sikre sig, at adressaten må kontaktes med reklamemateriale (»robinsonvask«).

Eksemplerne i litteraturen ovenfor implicerer en datafejlsmetrik, som ligger i nærheden af begreberne om akkurathed og præcision. Men ved at være netop implicit undgår den intuitive tilgang at udfordre dette begreb. Selvom nogle dimensioner bliver præsenteret (akkurathed, aktualitet, komplethed, konsistens), sker det på trods af en manglende beskrevet metodologi for, hvorledes forfatterne er nået frem til disse kvalitetsdimensioner (English, 1999, s. 141-154; Fox et al., 1994, s. 13-17).

Empirisk funderet datakvalitet leder til produktion af metadata

Det er forventningen om brugen af data, der ligger bag den intuitive tilgang. En videreudvikling herfra foretager en systematisk tilgang til brugernes evaluering. En sådan central empirisk undersøgelse om datakvalitet findes hos Wang og Strong (1996), som yderligere trækker på metoder fra en berømt tidligere undersøgelse af brugeres evaluering af informationssystemer (DeLone & McLean, 1992). Systemudviklingsparadigmer som »The Unified Process« har ligeledes fremhævet brugerens perspektiv gennem en tilgang, der er brugsdrevet eller »use-case driven« (Jacobson et al., 1999), og inden for systemudvikling findes også begrebet »quality attributes« (Larman, 2001, s. 42).

Wang and Strong (1996) udfører deres eksplorative empiriske studie af datakvalitet fra et brugerperspektiv ved at overføre marketingsmetoder, således at data betragtes som et produkt og brugeren som konsument af data. I et survey blev datakonsumenterne bedt om at vurdere vigtigheden af en lang række attributter vedrørende datakvalitet. De pågældende mangfoldige attributter var opstået gennem et tidligere survey, der tillod en stor iderigdom mht. angivelsen af selve attributterne. På basis af den kvantitative scoring af attributterne grupperedes disse ved hjælp af faktoranalyse til et noget mindre antal dimensioner for datakvalitet. Dimensionerne blev efterfølgende yderligere kondenseret til fire resulterende kategorier som en begrebsmæssig ramme for opfattelse af datakvalitet:

Figur 2. Fire begrebsmæssige kategorier som resultat af den empiriske tilgang

Indre	at dataværdierne samstemmer med den faktiske eller sande værdi
Kontekstuel	at data er anvendelige til brugerens opgave
Fremstilling	at data udtrykkes på en forståelig og klar måde
Tilgængelige	at data foreligger og at data er opnåelige

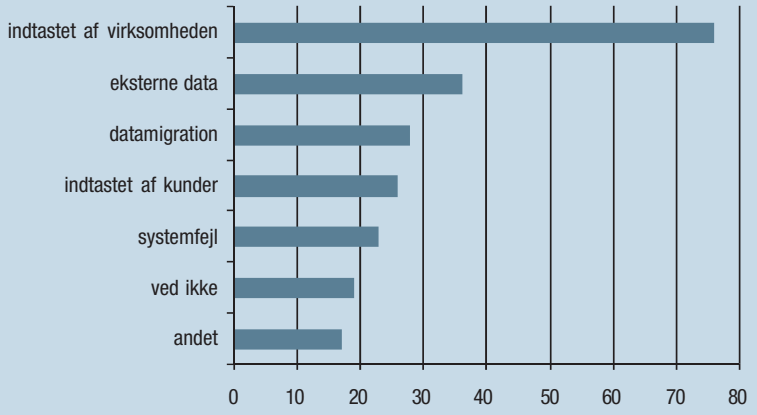
Kilde: Wang & Strong, 1996, s. 20.

Med en tilbagevenden til datasikkerhed kan begreberne: konfidentialitet, og specielt integritet og tilgængelighed, ses som parallelle begreber til den første og sidste kategori ovenfor. Fremstillingskategorien henleder opmærksomheden på, at selvom data er akkurate, brugbare og tilgængelige, så skal dette også fremstilles for brugeren. Data i høj kvalitet er mere end udfyldte felter i tabeller. Metadata med den indgående beskrivelse af data er en nødvendig forudsætning for, at brugeren kan fremfinde, evaluere og udnytte data. Dokumentationen af data (Rasmussen, 2000) er et vigtigt, men oftest upopulært og udskudt efterfølgende arbejde. Produktion af data har ligesom produktion af software stadig reminiscenser af gamle dages programmører med deres udtryk som »rigtige programmører dokumenterer ikke« og insisteren på, at deres 370/Asssembler er »selv-dokumenterende kode«.

Opsamling af regler for gode data

I virkelighedens og virksomhedernes verden ligger interessen for forbedring i høj grad på tilblivelsen af data og specielt med et fokus på tilblivelsen af fejlene i data. I undersøgelsen fra seminaret om datakvalitet spørges til »de primære årsager til eventuelle fejl i data«. Formuleringen »fejl i data« kan føre til en koncentration omkring den indre kategori af datakvalitet.

Figur 3. Årsager til fejl i data. Procent. (n= 46, flere svar)



Besvarelserne har i høj grad også fokus på det indre af virksomheden. Det er indtastningen af data i virksomheden selv, som ses som hovedårsagen til datafejl. I mange virksomheder sker der en gentagen indtastning af data. Her er et område, hvor virksomhederne virkelig kan forbedre deres processer gennem en integration (helst fuldstændig) af systemerne. Regel nummer 1: »Kun 1 gang dataindtastning!«

Kategorierne datamigration og/eller systemfejl vil tilsammen give 43 pct. og lægger således også en ekstra vægt på de interne processer i virksomheden. Derefter følger eksterne data som fejlkilde og til sidst indtastning foretaget af kunder. Da seminar-undersøgelsen ikke blotlægger mængdeforholdet mellem virksomhedens egne data og eksterne data eller mellem selvindtastede og kundeindtastede data, kan fejlårsagernes betydning ikke endeligt fastslås. Generelt opfattes kundeindtastede data som havende færre fejl, idet kunden har en direkte interesse i korrektheden af de indtastede data, fx angivelse af korrekt produkt og korrekt adresse i forbindelse med et køb. Regel nummer 2: »Lad kunderne taste«. Samme ræsonnement kan anvendes over for de eksterne data med beskrivelser, der automatisk bør følge med relationen til en leverandør. Ligesom kunden vil også leverandøren have stor interesse i at forsyne virksomheden med de korrekte data. Så regel 2 kan generaliseres til: »Lad de andre taste!«.

Når »de andre« skal foretage indtastning, er det, fordi at ved placering af dataindtastning på datafangststedet, hvor oplysningerne giver mening, kan fejl umiddelbart bemærkes og korrigeres. Mening kan også læres; så indtastere kan også oplæres i jobbet, det er ikke meningsløst arbejde. Hvis data skal kontrolleres, så er det datakontrollanten, der kan danne mening og se fejl, og kontrollanten bør så også have tilstrækkeligt ansvar til at ændre datafejl, der findes i systemet.

Selv hvor data giver god mening for indtasteren, kan der gøres skrivefejl, tastes i et forkert felt, etc. Validering af data kan fange mange ufrivillige og rigtig dumme

fejl – som man irriteret synes burde være fanget – fx: værdier uden for acceptabelt område, inkonsistens og logiske checks, inkl. kontrol for tomt felt eller for udfyldt felt, etc. (Welbrock, 1998, s. 144; Kubiak, 2008, s. 62). Validering bør typisk også foretage opslag til eksterne data, fx masterdata i virksomheden, hvor det fx undersøges, at kunden findes i forvejen, at produktet er registreret etc. »Peg & Klik«-systemer sikrer mod en del fejlangivelser, som fx ikke-eksisterende varenumre. Derfor bør 3. regel være: »Lav regler for data!«.

Det er ikke ukendt, at hvor virksomheder kæmper med en langtrukken tilretning af systemer – fx hvor overbelastning af it-afdelingen skaber en kø for systemvedligehold i form af nødvendige forbedringer – kan det føre til megen kreativitet hos ansatte, der har brug for løsninger her og nu. Det sker fx ved, at datafelter bruges til andre formål i afdelingerne. I senere led af datakæden kan en »extract-transform-load« (ETL) til et data warehouse resultere i problemer omkring datakvalitet fx hvor en 3. linje i adressefeltet blev opfattet som overflødigt, og derfor (mis) brugt til en fornuftig, men fejlanbragt kategorisering eller kommentar (Watson, 2001, s. 8).

Teoretisk funderet datakvalitet

Hvad enten man tilgår datakvalitet fra et intuitiv eller empirisk perspektiv, er der risiko for ikke at opsamle samtlige grundlæggende karakteristika omfattet af begrebet datakvalitet. Den intuitive tilgang kan åbenlyst være skævvredet af personlige idiosynkrasier og enkelthændelser, men det er også muligt, at en empirisk undersøgelse vil udtrække bredt udbredte misopfattelser og fordomme blandt databrugere. For fuldt ud at forstå datakvalitet er en ontologisk eller teoretisk tilgang således relevant, idet der dermed sikres en fuldere forståelse af udfaldsrummet med de fundamentale mulige kategorier. Et vigtigt bidrag kommer fra synspunktet med det overordnede mål at opnå viden om design af systemer med data af høj kvalitet.

Wand og Wang (1996) ser data som tilhørende et informationssystem og informationssystemet som en repræsentation af den virkelige verden. Data i informationssystemet er tilstande, og brugeren fortolker data fra informationssystemet som et udsagn om tilstande i den virkelige verden, men det er også muligt for brugeren at foretage direkte observation af disse samme tilstande i den virkelige verden. Videnskabsteoretisk må den direkte objektive observation tilskrives en gren af positivisme. Tilgangen findes også anvendt i en tidligere artikel om datakvalitet (Fox et al., 1994), hvor der identificeres tre kvalitetsemner: »those related to the quality of the model or view, those related to the quality of data values themselves, and those related to the quality of data representation and recording«. To af medforfatterne til 1994-artiklen undersøger det følgende år dette »conceptual view« (Levitin & Redman, 1995).

Den basale forståelse er, at »verden består af ting, der har egenskaber« (Wand & Wang, 1996). Med systemudvikling som en drivende kraft i undersøgelsen forekommer det naturligt at tænke i retning af en objekt-orienteret systemudviklingstermi-

nologi, hvor objekter (ting) har identitet og tilstande (egenskaber). Handlinger og love i artiklen kan tilsvarende betragtes som operationer og betingelser i den objekt-orienterede terminologi. Informationssystemet er selv en ting og har tilstande. Afbildningen mellem informationssystemet og systemet kaldet »den virkelige verden« leder til nogle repræsentationskategorier for kombinationer af tilstande mellem de to. Forfatterne Wand og Wang beskriver afbildningerne med grafiske eksempler; de er her transformeret til en opstilling baseret på multiplicitet i relationen mellem den virkelige verden (VV-systemet) og informationssystemet (IS).

Figur 4. Repræsentation af de mulige tilstande mellem den virkelige verden (VV) og informationssystemet (IS).

Repræsentations-kategori	Multiplicitet VV : IS	Forklaring
Korrekt	1:1 1:n	Den korrekte repræsentation eksisterer, når en tilstand i informationssystemet kan afbildes til en enkelt tilstand i den virkelige verden. Der kan forekomme redundans i form af overflødige tilstande i informationssystemet.
Ukomplet	1:0	Ukomplet repræsentation forekommer, når en tilstand i den virkelige verden ikke har en repræsentation i informationssystemet. Afbildningen er ikke udtømmende. Informationen savnes i IS.
Flertydig	n:1	Flertydig repræsentation forekommer, når mere end en tilstand i den virkelige verden dækkes af en tilstand i informationssystemet. Denne situation udelukker den korrekte inverse afbildning til den virkelige verden.
Meningsløs	0:1	Meningsløs repræsentation forekommer, når en tilstand i informationssystemet ikke kan afbildes til en tilstand i den virkelige verden. Vi har løsrevne data, men kender ikke deres forbindelse til den virkelige verden.

Kilde: Wand & Wang, 1996.

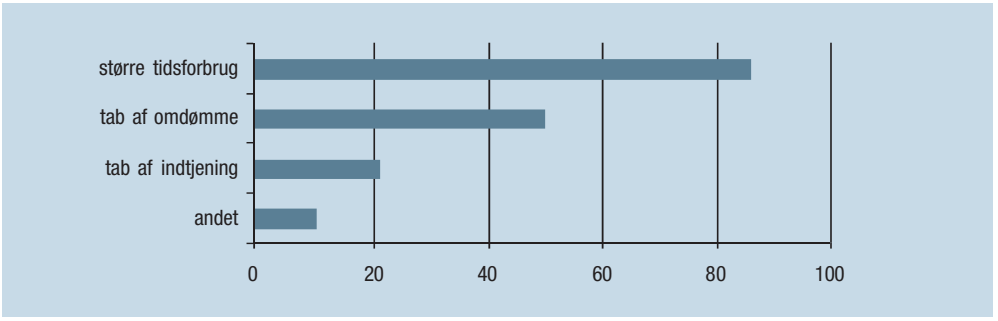
Wand og Wang deducerer således dimensioner for datakvalitet ud fra de mulige relationer mellem tilstande i den virkelige verden og informationssystemet: ukomplethed, flertydighed og meningsløshed. Den perfekte kategori (»korrekt«) inkluderer den åbenlyse 1:1-afbildning, men også 1:n, idet det er acceptabelt, at der findes flere tilstande i informationssystemet, som repræsenterer den samme tilstand i den virkelige verden. Med 1:n er der en overrepræsentation i informationssystemet, en slags redundans. Men hvis redundansen (fx på længere sigt) ikke er fuldt overensstemmende, vil det skabe problemer, fx ved de omtalte dubletter, hvor der findes flere (næsten) identiske adresser for den samme person eller firma. Det er yderligere åbenlyst, at ethvert informationssystem grundlæggende må være ukomplet, da der altid vil være tilstande i den virkelige verden, som ikke korresponderer med tilstande i informationssystemet. Ukomplethed skal for at have indhold forstås i sammenhæng med forventningen til informationssystemet, således at begrebet dækker manglen af de data, som burde være i systemet. Forklaring på eksistensen af meningsløshed kan fx være et resultat af et afbrudt link, hvorved forklarende metadata er bortkommet, således at relationen til den virkelige verden ikke længere kan etableres.

Mens den teoretiske tilgang på den ene side tilfører diskussionerne om manglende datakvalitet en tiltrængt stringens og også en fornyelse, så bliver datakvalitet kun groft behandlet. Således ligger den dikotome betragtning om data som værende enten korrekte eller fejlbehæftede langt fra de mere nuancerede tidligere betragtninger med inddragning af både »accuracy« og »precision« og validitet og reliabititet.

Betydningen af datakvalitet og forbedringer

I forbindelse med seminarerne om datakvalitet spurgtes deltagerne om »de mest betydningsfulde følger« af dårlig datakvalitet for virksomheden. Næsten alle nævner det ekstra tidsforbrug, der opstår i forbindelse med behandlingen af data.

Figur 5. Følger af manglende datakvalitet. Procent. (n= 46, flere svar)



»Tid er penge«, og der nævnes også direkte »tab af indtjening« samt under kategorien »andet« igen yderligere økonomiske forhold. Oven i disse omkostninger kommer yderligere kategorien »tab af omdømme«, som er vanskelig at opgøre på økonomisk vis, men som antages at have stor betydning for virksomheden.

Den måde, virksomheden opdager, at data ikke har tilstrækkelig kvalitet, forekommer fx gennem kuldsejlede projekter: »Data quality initiatives have long languished in the shadow of sexier projects. But thanks to failed CRM and ERP efforts, compliance violations, costly supply chain inefficiencies and more, that's starting to change.« (Gilhooly, 2005). Som det fremgår af citatet, investeres der nu ivrigere i systemer til sikring af datakvalitet. Der findes også i nyhedsmedierne adskillige spektakulære eksempler på omfattende spild. Manglende metadata og standardisering var baggrunden for, at »Mars Climate Orbiter« i 1999 gik ind i et katastrofalt kredsløb 100 km tættere på Mars end beregnet. Nogle – men ikke alle! – havde udført beregninger i det metriske system (Kubiak, 2008).

Systemer til forbedring af datakvalitet kan være dele af den extract-transform-load (ETL) proces, som importerer data til virksomhedens data warehouse. ETL-software kan således assistere ved at identificere felter, tabeller, områder etc., hvor der forekommer problemer med datakvaliteten, typisk identificeret som manglende værdier og usandsynlige værdier. Men hvor den normale »transform« for et data warehouse i høj grad etablerer en intern konsistens – som ikke findes, og ikke kan forventes at findes, i de mange adskilte og selvstændigt udviklede operative

systemer, der ofte føder data til data warehouse – så vil den fremadskuende virksomhed ved alvorlige problemer omkring datakvalitet søge at ændre de grundlæggende operative data-genererende systemer og processer i organisationen.

Spørgeskemaet ved seminaret om datakvalitet indeholdt også spørgsmål om »I hvilken grad er definitioner på nøglebegreber og datastandarder fælles i virksomheden?«. Her blev besvareren hypotetisk stillet samme spørgsmål gældende for fremtiden (»om 2 år«). Besvarelserne viser forventning om en bevægelse fra opmærksomhed hos enkeltpersoner, over afdelingsbaserede definitioner til gennemgående definitioner i hele virksomheden; hvilket illustrerer evolutionen nævnt i indledningen.

Relationerne mellem vurderingen af de enkelte datas kvalitet og inddragning af flere dimensioner med henblik på udarbejdelse af en samlende datakvalitet er ikke velundersøgte. I den intuitive tilgang sås eksempler med anvendelse af gennemsnit: at fx gennemsnitligt er 1 pct. af data forkerte. Men hvad betyder det? Almindeligvis betyder det ikke, at dataværdierne er 1 pct. forkerte. Hvis der var tale om præcision, ville fx forventes, at værdien oplyses som 99, hvor den korrekt skulle have været 100. Men udsagnet dækker snarere, at 1 pct. af datafelterne indeholder en ukorrekt værdi eller eventuelt ingen værdi. Men kan det så forventes, at beregninger vil være korrekte inden for +/- 1 pct.? Slet ikke. For det første vil fejlresultatet afhænge af typen af beregninger, men derudover kan en enkelt dataværdi være så ukorrekt, at kalkulationerne bliver helt ubrugelige.

Selve anvendelsen af begrebet dimension for datakvalitet implicerer, at dimensioner i datakvalitet ikke kan substitueres. Til vurdering af datakvalitet forslår nogle forfattere (Parssian et al., 2004) udregning af en kombineret score for datakvalitet for at kunne vælge mellem alternative kilder. Men opmærksomheden bør være rettet mod denne ikke-substituerbarhed for dimensioner i datakvalitet – en bedre score på en dimension kan ikke udligne manglen på en anden. Fx er det åbenlyst, at det har lille betydning, at data indeholder de eksakte sande værdier (akkurathed), hvis data ikke kan skaffes (tilgængelighed). Eller, at data er perfekt beskrevet og forståelige (fremstilling), hvis data ikke passer til opgaven (kontekstuel). Dimensionerne må være til stede og kan ikke erstatte og udligne hinanden.

Opmærksomhed over for dimensionerne i datakvalitet vil nødvendigvis føre til opdagelse af flere fejl. Men: »There is no harm in being sometimes wrong – especially if one is promptly found out« (Keynes). En fortolkning for virksomhederne kan ses som en 4. regel for aktivitet: »Find fejlene!«.

Summary
Many examples exist of poor data quality in enterprises. In the book »Information Revolution« enterprises are categorised according to their stage in the »Information Evolution Model« (Davis, 2006, p. 13), where the levels indicate the growing strategic importance of information. A small survey used in this article illustrates the positions of some Danish enterprises in the ongoing evolution through attention to challenges, causes and consequences and the organisational anchoring of business processes in data quality.
The enterprise can achieve significant benefits through sequenced data processes for core services in terms of data quality, and they can be developed in the »Information Evolution Model«, ensuring improved strategic decisions. The understanding of data quality in the enterprise is enhanced through more comprehensive approaches to the unfolding of dimensions and concepts to illustrate what constitutes good quality data. The rules given in the article are simple and general: Corporate strength lies in the implementation.

Referencer
Ballou, Donald P. & Tayi, Giri Kumar: Methodology for Allocating Resources for Data Quality Enhancement. Communications of the ACM, 32:3 s. 320-329, 1989.
Berg, Dennis & Heagele, Christopher: Improving Data Quality: A Management Perspective and Model. s. 85-99. Building, Using, and Managing the Data Warehouse (eds. Barquin,R.C. & Edelsterin, H.C.), Prentice Hall, 1997.
Berman, Francine: Got Data? A Guide to Data Preservation in the Information Age. Communications of the ACM. 12/08 Vol. 51 s. 50-56, 2008.
Bruckner, Robert M. & Schiefer, Josef: Using Portfolio Theory for Automatically Processing Information about Data Quality in Data Warehouse Environments. Advances in Information Systems, s. 34-43, 2000.
D'Angelo, John & Troy, Bob: Integrated data management improves return on investment. Oil & Gas Journal,July 31 s.40-44, 2000.
Davenport, Thomas H. & Harris, Jeanne: Competing Analytics. The New Science of Winning. Harvard Business School Press, Boston, MA, 2007.
Davis, Jim; Miller, Gloria J. & Russel, Allan: Information Revolution. Using the information evolution model to grow your business. John Wiley & Sons, Hoboken, NJ, 2006.
DeLone, William H. & McLean, Ephraim R.: Information Systems Success: The Quest for the Dependent Variable. Information Systems Research, 3:1 s. 60-95, 1992.
English, Larry P.: Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. John Wiley & Sons, 1999.
Fox, Christopher, Levitin, Anany V. & Redman, Thomas C.: The notion of data and its quality dimension. Information Processing & Management, Vol. 30:1 s. 9-19, 1994.
Gilhooly, Kym: Dirty Data Blights the Bottom Line. Computerworld.com, November 07, 2005.
Huang,K.-T., Lee, Yang W. & Wang,R.Y.: Quality Information and Knowledge. Prentice Hall, 1999.
Inmon, William H.: Building the Data Warehouse (2.ed.). John Wiley & Sons, 1996.
Jacobson, Ivar; Booch, Grady & Rumbaugh, James: The Unified Software Development Process. Addison-Wesley, Reading, MA, 1999.
Kapochunas, Andrew: ROI begins with improved data quality. Target Marketing, 25:7 s. 58, 2002.
Kimball, Ralph: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, 1996.
Kubiak, T.M.: Data Dependability. Qualilty Progress, June 2008; 41-6 s. 61-64, 2008.
Larman, Craig: Applying UML and Patterns (2.ed.). Prentice Hall, 2001.
Levitin, Anany V. & Redman, Thomas C.: Quality dimensions of a conceptual view. Information Processing & Management, 31:1 s. 81-88, 1995.
Levitin, Anany V. & Redman, Thomas C.: Data as a resource: Properties, implications, and prescriptions. Sloan Management Review, 40:1 Fall s. 89-101, 1998.
Liautaud, Bernard: e-Business Intelligence. Turning Information into Knowledge into Profit, McGraw-Hill, New York, NY, 2001.
Parssian, Amir; Sarkar, Sumit & Jacob, Varghese S.: Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Management Science, 50:7 s. 967-982, 2004.
Pfleeger, Charles P.: Security in Computing, 3rd ed.. Prentice-Hall, New Jersey, 2003.
Rasmussen, Karsten Boye: Datadokumentation. Metadata for samfundsvidenskabelige undersøgelser. Odense Universitets Forlag, Odense, 2000.

Rasmussen, Karsten Boye: Data mining – er der guld i virksomhedens data? Ledelse & Erhvervsøkonomi, 2001:2 s. 91-107, 2001.
Rasmussen, Karsten Boye: General Approaches to Data Quality and Internet-generated Data. Online Research Methods, The SAGE handbook of (red. Nigel Fielding, Raymond M. Lee, Grant Blank) London, Sage, s. 79-96, 2008.
Redman, Thomas C.: The impact of poor data quality on the typical enterprise. Communications of the ACM, 41:2 s. 79-82, 1998.
Redman, Thomas C.: Data: An Unfolding Quality Disaster. DMReview (www.dmreview.com). August 2004. s. 1-5, 2004.
Strong, Diane M.; Lee, Yang W. & Wang,R.Y.: Data quality in context. Communications of the ACM, 40:5 s. 103-110, 1997.
Tayi, Giri Kumar & Ballou, Donald P.: Examining data quality. Communications of the ACM, 41:2 s. 54-57, 1998.
Trull, Samuel G.: Some Factors Involved in Determining Total Decision Success. Management Science Series B, Managerial, 12:6, Series B, Managerial B270-B280, 1966.
Wand,Yair & Wang, Richard Y.: Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39:11 s. 86-95, 1996.
Wang,R.Y. & Strong, Diane M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12:4 s. 5-24, 1996.
Watson, Hugh J.: Recent developments in data warehousing. Communications of the Association for Information Systems, 2001:8, s. 1-25, 2001.
Watson, Hugh J.; Goodhue, Dale L. & Wixom, Barbara H.: The benefits of data warehousing: why some organizations realize exceptional payoffs. Information & Management, 39:6 s. 491-502, 2002.
Welbrock, Peter R.: Strategic Data Warehousing Principles using SAS Software, SAS Institute, 1998.