Kvalitetsmåling av ansattdata

Jan-Erik Hagen

Masteroppgave
Master i informasjonssikkerhet
30 ECTS
Institutt for informatikk og medieteknikk
Høgskolen i Gjøvik, 2005

Masterprogrammet i informasjonssikkerhet
har blitt kjørt i samarbeid med
Kungliga Tekniska högskolan (KTH),
Stockholm, Sverige

Institutt for
informatikk og medieteknikk
Høgskolen i Gjøvik
Postboks 191
2802 Gjøvik

Department of Computer Science
and Media Technology
Gjøvik University College
Box 191
N-2802 Gjøvik
Norway

Sammendrag

Denne rapporten gir oss kunnskap om hvordan forbedre eller bekrefte bedriftens kvalitet på ansattdata i et personalmiljø (HR miljø). Integritet er et av flere egenskaper innen informasjonssikkerhet og innebærer sikkerhet for at informasjonen og informasjonsbehandlingen er fullstendig, nøyaktig og gyldig og et resultat av autoriserte og kontrollerte handlinger. I denne sammenheng er både den fysiske og konseptuelle integritet ved ansattdata et sentralt emne.
Det er mye oppmerksomhet rundt den manglende datakvalitet i dagens bedrifter. Mange har bekymringsverdig dårlig kvalitet på sine forretningsdata. Denne oppgaven stiller da spørsmål om hvordan vi skal kunne definere kvalitetsnivået på de interne, administrative data som eksempelvis ansattdata. Denne type data er av spesiell interesse i sammenheng med informasjonssikkerhet da dette ofte legger grunnlaget for korrekt autorisering. Vi må være sikre på at en ansatt faktisk er en ansatt og at denne er knyttet til en gyldig organisatorisk enhet. Korrekt autorisering er fundamentalt for hele sikkerhetssystemet. Datakvalitet og informasjonssikkerhet henger nøye sammen.
For å finne dagens status i industrien vedrørende måling av datakvalitet ble følgende forhold undersøkt:

- Hvordan oppfattes datakvalitetsbegrepet i bedriftene i dag?
- Hvordan måle datakvalitet i HR miljø?
- Hvilke metrikker (målemetoder) er sentrale?
- Hvordan praktisere metrikker

Rapporten viser at måling av datakvalitet i HR miljø ikke er vanlig og ingen måler systematisk. Undersøkelser viser også at det er få formaliserte rutiner for rapportering av kvalitetssvikt og at dokumentasjon av forretningsregler (metadata) har lite fokus. Vi kan ikke si hva som er feil ved våre data når vi ikke vet hva som er rett og det er våre metadata som forteller hva som er rett. Relatert til dagens behov er det foreslått et sett med aktuelle målemetoder (metrikker), som måler nivå på:

- Trygg behandling av HR data
- Kompletthet ved informasjonsprodukter
- Ansattes relasjoner til organisatoriske forhold

Det er foreslått et rammeverk for måling av kvalitet via bruk av generelle metoder og teknikker og hvordan dette kan benyttes i et databasemiljø. Det er vist at databasetrigger er egnet for realisering av kvalitetsmåling av forretningsregler i et lavtransaksjonsmiljø, som i et HR miljø.

Abstract

This master thesis examine the possibility to measure data quality in a human relation (HR) environment. Integrity is one of several characteristics of information security and imply that information safety and handling are complete, accurate and valid and a consequence of authorized and controlled actions. Integrity by the physical and conceptual level in a employee data environment, are of interest in this context.
Poor quality data is the root of many issues of business importance that dominate the headlines. Several international surveys state that most of the businesses should be worrying about their data quality level. Most of the surveys deals with CRM data and thus concerns the business production data. This thesis describes a framework for measure the business own administrative data, i.e. employee data. This kind of data is of special interest concerning information security and is the foundation of correct authorisation. Data quality and information security are in this way tightly connected.
To be able to investigate data quality measuring, following items are discussed:

- How the business make use of data quality measuring.
- How to detect error in employee data in a HR environment.
- Data quality relevance for employee data.
- How metrics are put into practice.

An investigation, carried out through surveys and interviews, was conducted. The result show that there are few formalized methods for measuring data quality in the area of HR administration. There is little focus on documentation of business rules. There is no chance to tell what is wrong if you do not know what is right and it is the metadata that tell us what is right. So it is very important to document the business rules, i.e. to be able to automate data quality measures.
This thesis suggest some measurement methods (metrics) as:

- Ensure that Employee data is used and stored safely
- Completeness of Information Product (a context dependent collection of distinct data)
- Control of the connection between actual employees and the organizational units

A framework for measuring in HR data is outlined and it shall contribute to ensure a high level of data quality.

Executive Summary

‘Kvalitetsmåling av ansattdata’ er satt som tittel på denne masteroppgaven i informasjonssikkerhet. Her undersøkes om man kan kvantifisere mål på kvalitet for data i et HR miljø.
Vi ser hvordan informasjonssikkerhet kobles sammen med datakvalitet, blant annet basert på anerkjente forskningsaktiviteter innen området og egne undersøkelser. Datasikkerhet gjelder også kvaliteten på lagrede data. Det som står i et register skal være korrekt, i henhold til bedriftens forretningsregler og brukerens forventninger.
I et stadig mer dynamisk arbeidsliv hvor bedrifter migrerer og de ansatte oftere skifter organisatorisk tilhørighet, er administrering av datakvalitet en viktig del av det totale sikkerhetsbildet.
I denne oppgaven har 19 HR personer i 8 større norske selskap besvart et utfyllende skjema med 43 spørsmål om temaet datakvalitet. Tema som ble behandlet var:

- Hvordan man oppfatter datakvalitet i de forskjellige bedrifter
- Hvordan man måler, avslører og rapporterer feil i sine data i dag
- Hva man mener er viktige egenskaper ved datakvalitet

Dybdeintervju er gjort med ytterligere 5 personer for å supplere spørreundersøkelsen. Flere andre personer er konsultert ved kortere intervjuer.

Hva avdekker undersøkelsene?

Noe av det viktigste som fremkommer er at ingen måler kvalitet på en systematisk måte og ingen kan gi et klart svar på hvor god datakvaliteten er. Dette er slett ikke unikt for disse selskapene, flere større internasjonale undersøkelser viser akkurat det samme.
Av andre områder kan nevnes at kun halvparten av de spurte mente å ha en politikk for å rapportere feil i datakvalitet. De data som er av de mer sentrale data i et informasjonssikkerhetsperspektiv er de data som HR eier. Det gjelder data som viser hvem som faktisk er ansatt i bedriften og hvilke avdeling denne tilhører. Det er i undersøkelsen vist at kun 50% av de spurte mener bedriften har en vel definert politikk for dataeierskap.

Hvordan måle datakvalitet?

Med dette som bakgrunn presenterer denne oppgaven forslag til et rammeverk for måling av datakvalitet for ansattdata og en praktisk tilnærming er vist. Vi ser hvordan målinger kan strategisk plasseres i organisasjonen og hvilke teknologiske metoder som kan benyttes. Begrepet informasjonsprodukt er sentralt. Det er en sammensetning av enkelte dataelementer beregnet benyttet til et gitt formål som eksempelvis telefonkatalog, adgangssystem osv. Når vi benytter målinger kan vi ta hensyn til viktigheten av det enkelte element i informasjonsproduktet. På denne måten kan vi måle kun det som forretningen på forhånd mener er viktig å måle og ikke noe annet. Slik får vi også en bedre kost nytte av målingen. Det er ikke formålstjenlig å måle alt av data i bedriften, det blir for kostbart.

Hvordan sette oppgaven i sammenheng med virksomheten?

Måling av datakvalitet gir ikke så mye om man ikke setter dette inn i en overordnet strategi for kvalitetsstyring. Denne oppgaven viser hvordan selve målingene og rammen rundt disse kan administreres, men for å iverksette kvalitetsforbedringstiltak må også totale

kvalitetsforbedringsprosesser vurderes.

Hva kan vi erfare av denne oppgaven?

Det er viktig å etablere en felles forståelse for datakvalitet i HR organisasjonen. Ta eierskap over ansattdata som inngår i et sikkerhetsmessig fundament og etablere rutiner for å kontrollere og styre kvaliteten av disse. Etabler datakvalitet som et konkret begrep og gjør det kjent. Gi organisasjonen tid og opplæring til å få kontroll på de viktigste informasjonsproduktene og lag metrikker for disse. Koordiner kvalitetsarbeid med organisasjonens øvrige aktiviteter innen informasjonssikkerhet. Vedlikehold rutiner og kompetanse.

Forord

Denne masteroppgaven er et resultat av mange forhold.

For det første har jeg fått grundig innføring i de teoretiske områdene ved HiG av forelesere som har framført sitt budskap på en inspirerende måte og bidratt til å sette meg posisjon til å skrive denne oppgaven. God veiledning ble under oppgaveskrivingen gitt av Jan Arild Audestad. Min arbeidsgiver, Telenor, har gitt meg tid og ressurser til å investere i ny og nyttig kunnskap.
Jeg fått anledning til å jobbe sammen med to meget inspirerende personer under hele studiet, Tore og Tone. De har delt sin kunnskap og erfaring med en som trengte det til tider. Tenk på CIA var noe av det første jeg hørte.
Meget god hjelp fra ungene har jeg også fått. Fredrik, har vært en utrolig god diskusjonspartner og korrekturleser og Camilla har gitt befriende tanker når skrivingen har holdt på å ta helt overhånd. Sist men ikke minst er det en der hjemme som fortjener roser for sin tålmodighet, takk til Inger.

Uten bistand fra dere som svarte på spørreundersøkelsen, intervjuene og stilte opp i telefonmøter når det var noe jeg lurte på, ville dette ikke latt seg gjennomføre. Gode kollegaer har gitt verdifulle tilbakemeldinger. Ingen nevnt, ingen glemt heter det. Biblioteket ved HiG har vært gode å ha. Takk for god service.

Til slutt har jeg også kunne høste relevant erfaring fra eget arbeid innen informasjonsteknologiske metoder og teknologier fra mange års arbeid i private og offentlige institusjoner.

Viten kan man meddele men ikke visdom (ukjent)

Jan-Erik Hagen, Lillehammer juni 2005

Ord og begrepsforklaring

Begrep	Forklaring
CPRS	Country Population Registration System (folkeregister).
CRM	Customer Relationship Management
Databasetrigger	Prosedyre som automatisk aktiviseres ved endring av data i databasen.
Dbms	Data Base Management System (databasesystemet)
Dfd	Data Flow Diagram, viser dataflyt i et system.
ETL	Ekstrahering, Transformering og lasting av data i et Datavarehus.
Fremmednøkkel	Begrep for håndtering av integritet i rdbms.
Fuzzy	En utvidelse av Boolsk logikk som omhandler delvis sannhet.
IEEE	Institute of Electrical and Electronics Engineers.
IPMAP	Informasjonsproduktkart
IP	Informasjonsprodukt
ISO17799	'Beste praksis' for administrering av informasjonssikkerhet.
ISO9000	Standard som omfatter administrering av kvalitet.
ITIL	Information Technology Infrastructure Library.
Informasjons-produkt	Samling av dataelementer tiltenkt et spesielt formål. Se side 5
Integritet	I databasen betyr det: Nøyaktighet, korrekthet og gyldighet.
MIT	Massachusetts Institute of Technology.
Mastring	Samling av sentrale data som benyttes av flere avdelinger.
Metadata	Informasjon om data.
Metode	Fremgangsmåte for å løse et problem og komme til ny erkjennelse.
Metrikk	Se definisjon på side 15
Mindmap	Tankekart. En teknikk for å tegne sammenhenger.
NGE	Nyttegradselement. Attributt i Rolles_IP tabell (figur E.2)
NULL	Karakteriserer en variabel i et rdbms som ikke har noen verdi.
Pareto prinsippet	Kjent som 80-20 regelen. Mønsteret bak 80-20 prinsippet ble oppdaget allerede i 1897 av den italienske økonomen Vilfredo Pareto.
Rdbms	Relational Data Base Management System.
Sarbanes-Oxley	Public Company Accounting Reform and Investor Protection Act - SOX. Amerikansk lov om finansiell rapportering av selskap notert på amerikansk børs.
Sekvensdiagram	Beskriver kronologisk sekvens i et system.
Soundex	Fonetisk algoritme.
Sox	Se: Sarbanes-Oxley
Sql	Structured Query Language.
TDQM	Total Data Quality Management program, ledet av R.Wang i MIT.
Transaksjon	En samling operasjoner i databasen som utføres på en kontrollert og fullstendig måte.
Trigger	Se definisjon på side 13
UML	Unified Modeling Language. Et modellering- og spesifikasjonsspråk.
View	En beskrivelse av en virtuell tabell i en database.
eTOM	enhanced Telecom Operations Map.

Innhold

Figurer

Tabeller

1 Introduksjon

1.1 Emne

Denne oppgaven omfatter måling av datakvalitet i et miljø som forvalter personaldata (HR miljø). Vi ser dette i sammenheng med hva det betyr for informasjonssikkerheten samt en målemetode som kan være egnet for å kvantifisere kvalitet i denne sammenheng.

1.2 Problembeskrivelse

Hvordan er kvaliteten på data om de ansatte i bedriften og bruken av disse når selskaper har så dårlig kontroll på sine forretningsdata som undersøkelser i oppgaven viser til? Følgen ved ikke å ha tilstrekkelig kontroll av ansattdata kan være ukontrollerte hendelser ved fratreden, vanskelig å implementere felles tilgangskontroller, manglende grunnlag for effektiv inventarkontroll, unøyaktige telefon- og adresselister osv. Dårlig datakvalitet for ansattdata motarbeider en effektiv konsernomfattende kontrollert autorisering, noe som er et sentralt element i bedriftens sikkerhetssystem.
Mange bedrifter ønsker å få bedre kontroll på datakvaliteten. I henhold til en global undersøkelse utført av PricewaterhouseCoopers(PWC)[1], uttrykte 75% av selskapene som deltok i undersøkelsen at dårlig datakvalitet påvirket dem finansielt og 33% ble tvunget til å forsinke eller forkaste nye system. Senere undersøkelser viser også at mye gjenstår for å forbedre kvaliteten på de operasjonelle data[2].
Vi må kunne regne med at dette også gjelder for HR data. Endring i konserns selskapsstruktur samler ulike teknologiske og kulturelle miljø og utfordrer datakvaliteten for ansattdata. Hvilke muligheter har vi for å kontrollere datakvalitet ved ansattdata og kan vi finne en effektiv måte å måle denne kvaliteten?

1.3 Motivering og gevinstpotensiale

For å kunne utøve en effektiv ressursstyring i et HR miljø, må ansattdata være troverdige, korrekte og tilgjengelige. Ansattdata må være relatert til gyldige enheter i organisasjonen. Dette er kritisk når ansattdata i HR miljøet knyttes til bedriftens autoriseringsprosesser.
Hvordan er sammenhengen mellom datakvalitet og informasjonssikkerhet? Det er alment akseptert å betegne informasjonssikkerhet som samling av egenskapene konfidensialitet, tilgjengelighet og integritet. Mange vil hevde at informasjonssikkerhet dreier seg om beskyttelsestiltak mot en intelligent angriper. Andre vil være mer generelle og hevde at integritet er å sikre at informasjon ikke blir endret eller ødelagt på en uautorisert måte og at informasjon er i overensstemmelse med virkeligheten og konsistent. Noen hevder at dataintegritet ikke må forveksles med datakvalitet som er knyttet til riktigheten av de opplysninger som er formidlet og at dataintegritet kan være i behold selv om opplysningene objektivt sett er uriktige, dersom det var disse opplysningene avsenderen faktisk sendte[3].
Daler m.fl.[4] sier det slik: *Integritet i forbindelse med informasjonssikkerhet er at informasjonen og informasjonsbehandlingen er fullstendig, nøyaktig og gyldig, og et resultat av autoriserte og kontrollerte aktiviteter*. Bing [5] hevder at datasikkerhet ikke bare gjelder uautorisert tilgang og data på avveie. Det gjelder også kvaliteten på lagrede data, at det som står i et register er korrekt.

Når vi tenker informasjonssikkerhet, fortoner det seg noe underlig at kvalitet på data ikke kontrolleres på en mer formalisert metode når vi vet hvor mye fokus målinger på forretningssiden har i form av periodiske salgstall, kundetilfredshet, leveransepresisjon, logistikk, diverse presisjonsmål osv. Vi trenger indikatorer som kan gi gode indikasjoner vedrørerende flere av de viktige sidene ved en virksomhet, der i blant HR data. Data skal ha en korrekt fremstilling og disse skal ikke endres underveis i produksjon av det endelige informasjonsproduktet.

Data har først verdi når de oppfyller visse krav til kvalitet.

Data som flyter i et selskaps systemer er ikke alltid av en slik kvalitet som man kunne ønske seg. Dette vises i flere undersøkelser som vi tidligere har sett. Data kan bli oppdatert flere steder i et systemkompleks internt i et selskap og inkonsistens i data kan oppstå. Dette kan være årsaken til administrative operasjoner som sentralisert autorisering kompliseres eller blir umulig pga det leverte informasjonsproduktets kvalitetssvikt. Kvalitet ved informasjonsproduktet knyttet til ansattdata er viktig for flere aktører:

1. System- og informasjonseiere som bruker ansattdata i administrative prosesser.
2. Verdikjedeansvarlige for sikkerhetstjenere i bedriften ønsker å verifisere rett dataintegritet.
3. Kontroll- og regulatoriskemyndigheter (interne eller eksterne) vil vite om grunnlaget for den sikkerhetsmessige godkjenningen av informasjonssystemet (*dokumentasjon av data*) forsatt er gyldig.

God datakvalitet er et direkte eller indirekte krav fra blant andre myndighetene ved behandling av visse typer informasjon. Disse krav kan være nedfelt i lover og forskrifter som regulering av personopplysninger [6], forretningsspesielle lover eller rapporteringskrav som i Sarbanes-Oxley. God datakvalitet kan også være i henhold til standarder som ISO17799, ISO9000 eller forskjellige rammeverk selskapene er bundet til som eTOM eller ITIL.

1.4 Forskningsspørsmål

Vi er interessert i å undersøke motivasjon og mulighet til å måle datakvalitet. Data skal i hovedsak være en del av bedriftsinterne administrative data og helst inkludert i det informasjons-sikkerhetsmessige fundament. Ansattdata er data som kommer under denne kategoriseringen. Innledningsvis fremsettes en arbeidshypotese for datakvalitet i denne sammenheng som hevder: *Selskap som er bevisst begrepet datakvalitet, har godt definerte kilder til originale data, flere objektive kriterier enn subjektive og opererer i stabile og kontrollerte omgivelser, har tilrettelagt fundamentet for god datakvalitet og hvordan dette kan måles*. Her skal vi komme frem til svar på det sentrale og viktigste spørsmålet: *Hvordan kan vi måle og effektivt vurdere, kvaliteten for ansattdata*? I denne sammenheng ønsker vi å kunne besvare disse spørsmålene:

1. Hvordan oppfattes og forvaltes datakvalitet innen HR området i dag.
2. Hvordan måle kvalitet for IP som inngår i HR området?
3. Hvilke metrikker er sentrale for å måle kvaliteten i ansattdata?
4. Hvordan kan metrikker benyttes i praksis?

Er disse spørsmål besvart tidligere og i så fall hvor tilfredsstillende?

2 Relatert arbeid

Informasjonsforvaltning og datakvalitet er bredt omtalt både i og utenfor de faglige miljøene og begrepet benyttes i mange forskjellige sammenhenger[7][8][9]. Først ser vi generelt på hva som relateres til denne oppgaven for deretter å se spesielt på de forskjellige forskningsspørsmålene. Vi kan si at det er rammeverk og metoder for måling av data som er fremherskende, uten at det er gjort mye av faktiske metrikker og målinger innen HR området. Datakvalitet er omtalt i mange sammenhenger og i mange fora. Begrepet er også sentralt i flere lover i Norge såvel som i USA.

2.1 Generelt

2.1.1 Forskningsområder

Et rammeverk for datakvalitetforskning ble utarbeidet i 1995[10] og belyser blant annet kvalitet relatert til distribusjon. Det hevdes å være mange likheter mellom informasjon produksjon og fysisk, materialisert produksjon. Her fastsettes en del begreper innen rammeverk for datakvalitet og relaterer temaet til ISO 9000 og fysisk produkt fremstilling. Av dette defineres rammeverket for analyse av datakvalitet, og som består av syv elementer: 'Management responsibilities', 'Operation and Assurance costs', 'Research and Development', 'Production', 'Distribution','Personnell Management' og 'Legal Function'. Størst fokus har området 'Research and Development' og av de minst fokuserte er 'Distribution' og 'Personnell Management'. Rammeverket berører i liten grad temaet 'måling av datakvalitet'.
Wang med flere hevder i [10] at det ultimate forskningsspørsmål er å forsikre at leveranser av et dataprodukt lever opp til de kvalitetskrav kunden hevder.

2.1.2 Metrikker og rammeverk

Metrikker og bruk av disse er viktig i denne rapporten, og på dette området er det en god del arbeid som er utført. Carson[8] peker på datakvalitet og manglende internasjonal standardisering. Dokumentet har opphav i finansielt/statistisk miljø. Det har tildels en generell vinkling, setter datakvalitet i et internasjonalt perspektiv og forsøker å sette datakvalitet i et rammeverk. Dokumentet sier ikke noe om hvordan man faktisk kan måle datakvalitet men skisserer forslag til utforming av metrikker. Interessant er det internasjonale perspektivet. Ikke alle nasjonaliteter legger vekt på de samme kvalitetsdimensjoner, men uansett nasjon framtrer nøyaktighet ('accurate') og tidsriktighet ('timely') som viktige områder. Loshin[11] beskriver hvordan en måler datakvalitet ved praktisk tilnærminger og bruk av statistisk kontroll ved bruk av absolutt kontroll. Bruk av kontrollkart (Se kap 2.1.6) er sentralt og benyttes i verdikjeder hvor numeriske verdier kan måles.

2.1.3 Forbedringsprosesser

Når vi behandler datakvalitet er et viktig tema kvalitetsforbedringsprosesser. Dette har ikke hovedfokus i denne oppgaven, men nevnes fordi det er så tett knyttet til måling av datakvalitet. Uten endringsrutiner er målinger lite verdt i denne sammenheng.
Wang m.fl. arbeider med et forskningsprosjekt ved MIT innen 'Total Data Quality Management' (TDQM)[9]. TDQM beskriver et rammeverk og tar for seg definisjon, måling, analysering og forbedring av et informasjonsprodukt (IP). De oppsummerer en rekke egne og andres

forskningsrapporter, som gjelder datakvalitet. IP er et sentralt tema i denne sammenheng. Nøkkelen til måling av IP hevdes å ligge i utvikling av kvalitetsmetrikker for nøyaktighet, tidsriktighet, kompletthet og konsistens. Videre er det viktig å betrakte forretningsreglene til de aktuelle data og ta dette med ved kvalitetsmåling. TDQM har følgende forskningsfokus:

- Definisjon av datakvalitet
- Datakvalitetens påvirkning på forretningen
- Forbedring av datakvalitet

De har også vinklet forskningen mot implementasjon av datakvalitet i relasjonsmodellen på et konseptuelt nivå via 'Entity Relationship' (ER) modeller. Blant annet beskrives 'Polygen model', 'Attribute Based Model'[12][9] og utvidet relasjonsalgebra som tiltak på å inkludere datakvalitetselementer som en utvidelse av relasjonsmodellen. TDQM omhandler et viktig område som må tas i betraktning.

2.1.4 Subjektiv oppfatning

Datakvalitet kan oppfattes subjektivt og objektivt. Barbara Maxwell [13] fokuserer primært på sluttbrukerens subjektive oppfatning av hva datakvalitet er uten å komme inn på behovet for objektive, stabile kriterier for data kvalitet. Flere andre er også inne på dette og mener gode data er i henhold til hvordan brukerne definerer kvaliteten, blant annet Huh [7].
Maxwell påpeker behovet for å bedre kvalitet for HR data under 'Personnel Management' området. Her belyses for eksempel dårlig kvalitet som forskjellige verdier i navn for samme person men med samme menneskelige betydning (J-E Hagen, J E Hagen, Jan E. Hagen, ...) hvor alt er tekniske sannheter men hva er korrekt? Tre områder nevnes for datakvalitetsforbedringer:

- Data eierskap og organisasjon blir understreket som viktig.
- Presisjonsnivå (nøyaktighet) er et bevegelig mål?
- Prosedyremessige forhold (trening og kommunikasjon)

Maxwell påpeker HR funksjonens utall av forskjellige måter å behandle data på, dynamikken ved regler og lover samt problemet med å bestemme hva som er 'rett syn på saken'. Fokus må være på brukerens krav til data og som møter organisatoriske og forretningsmessige krav og ikke på teoretiske verdier som forteller hva som er absolutt rett eller galt med data. Her er det et poeng og som flere har nevnt: det er ikke formålstjenlig å søke den absolutte sannhet. Det blir for kostbart. Andre har også referert til samme problemstilling og benyttet Pareto prinsippet. Disse betraktninger er interessante sett i forhold til 'Managing Data Quality in Dynamic Decision Environments'[14](beskrevet senere).

2.1.5 Dimensjonering

Det er mange måter å definere datakvalitet på. Datakvalitet hevdes å være et multidimensjonalt konsept [14][15]. Dimensjonene som nevnes i [8] er: Integritet, konseptuell konsistens, nøyaktighet, nytteverdi/nyttighetsgrad ('serviceability') og tilgjengelighet Huh [7] trekker frem omtrent de samme egenskapene og legger til: Nøyaktighet relatert til opprinnelig datakilde, kompletthet i datasettet, konsistens ved presentasjon, ikke konflikt med andre relaterte datasett gyldig og ajourført.
I rammeverket som beskrives av Carson i 'Data Quality Assessment Framework' [8], settes det opp forslag til klassifikasjon av disse dimensjonene med følgende struktur: kvalitetsdimensjoner, elementer i disse dimensjonene og indikatorer. Pipino med flere [16] definerer et større sett med

dimensjoner og inkluderer blant annet rykte, forståelighet, objektivt ('unbiased'), troverdighet, representasjon, mm. De tar også med 'security' som en egen dimensjon i mening kontrollert data aksess.

2.1.6 Granskningsmetoder

Dårlige data har en oppførsel som virus, hevdes det av Huh [7]. Et dårlig dataelement spres ukontrollert til forskjellige forretningsprosesser som behandler data videre og sprer dårlige data igjen til andre prosesser. Til slutt kan dette medføre for eksempel feil i et beslutningsgrunnlag. Datakvalitet må sikres allerede ved kilden. Man må unngå å behandle dette i verdikjedene, da det kan koste mer å rette opp skadene som dårlig datakvalitet kan medføre. Huh fokuserer på prosessene som introduserer, editerer og transformerer data. Den eneste praktiske måten for å forhindre dårlige data er først og fremst å hindre disse data i å komme inn i systemets database. Granskningsmetodene skjer via stikkprøver og sporing.
Granskning av data i databaser utføres for å finne allerede eksisterende feil og estimere feilverdier. Data sporing utføres for å finne hvor data feiler og forhindre senere feil. Datasporing blir av Huh benyttet i forbindelse med sammensetting av data fra datafødsel til de er lagret i en 'master' database. Stikkprøver foreslås som inspeksjonsmåte fremfor distinkte målinger, bortsett fra de prosesser hvor input er basert på postnivå og ikke satsvis behandling. Stikkprøvene merker de data som overvåkes, uten å nevne noen spesiell måte å gjøre dette på. For å identifisere feil i input-prosessene hevdes det å være nødvendig å relatere seg til den virkelige verden. Mer om granskningsmetoder er nevnt av Theuwissen m.fl.[17]. De har i sitt foredrag pekt på måter å måle datakvalitet i databasesystemer:

- Standardiserte oppslagstabeller
- Frekvenstabeller
- Inkompatibele kombinasjoner

Måling kan skje på verdikjede nivå. En måler da de prosessene som har relasjon til produktivitetsmål. Slike mål kan gå på volumavvik og tidsavvik. Loshin([11]) beskriver statistisk prosesskontroll ved bruk av kontrollkart ('Control Chart') som har relasjoner til Shewart[1] og statistisk prosesskontroll. Disse testene baseres er beregnet på å måle data i en prosess-strøm.

2.1.7 Informasjonsprodukter

I denne oppgaven er vi interessert i kvalitet ved et informasjonsprodukt (IP). Flere har vurdert og definert IP-begrepet [14] og beskriver konstruksjonselementer for IPMAP[2]. Det er laget en rapport[18] som omhandler hvordan kvalitetskarakteristikker kan operasjonaliseres og relateres til informasjonsprodukter. Her beskrives kun teorier og fokuserer relasjonsalgebra via tupler og relasjoner. Dette bidrar til å bestemme hva som er gode data innen et relasjonsbasert miljø. Følgende metrikker er foreslått: nøyaktighet i en relasjon, kompletthet i en relasjon og ikke medlemskap i en relasjon.
I [14] sammenlignes modelleringsteknikker for det rammeverk som beskrives men UML er ikke tatt i betraktning. Nevner at DFD[3] kan supplere men ikke erstatte IPMAP og at DFD er prosess-sentrerte modeller. Et IP er komplett dersom alle nødvendige dataelementer er tilstede. Beregning av total kvalitet i det enkelte dataelement[14] beskrives som A_x, hvor A_i er nøyaktigheten ved dataelement i og brukerens oppfattelse av nøyaktighet er a_i (et tall mellom 0-1). Nøyaktigheten i

[1] Walter Shewhart - Av mange betegnet som opphavsmann til 'Total Quality Management'
[2] Information Product map
[3] Data Flow Diagram

punktet x er da:

$$A_x = [\sum_{i=1,n}(a_i * A_i)]/[\sum_{i=1,n}(A_i)]$$

Kompletthet i dataelement kan betraktes som summen av produktets kompletthet i element (C_i) og nødvendighet (c_i) som gir kompletthet av dataelement i et gitt punkt C_x:

$$C_x = \sum_{i=1,n}(c_i * C_i)/\sum_{i=1,n}(c_i)$$

Total kvalitet i produktet ved et hvilket som helst trinn x i IPMAP sier man er en addisjon av dimensjonene tidsriktig, nøyaktighet og kompletthet.

2.1.8 Lover og standarder

Begrepet datakvalitet er også lovfestet i nasjonale og viktige utenlandske lover. Sentralt her hjemme har vi personopplysningsloven [6] som omhandler kontroll med bruk og spredning av personopplysninger. Formålet med loven er å beskytte den enkelte mot at personvernet blir krenket gjennom behandling av personopplysninger og sier blant annet noe om hvilke opplysninger som kan behandles og måten opplysningene behandles på. Loven setter også krav til kvalitet. Loven skal bidra til at personopplysninger blir behandlet i samsvar med grunnleggende personvernhensyn, herunder behovet for personlig integritet, privatlivets fred og tilstrekkelig kvalitet på personopplysninger.
Sarbanes-Oxley[19] representerer viktig amerikansk lovfesting innen finansielt område for alle selskap som er notert på den amerikanske børs (som Telenor). Loven stiller krav til selskaper notert på amerikansk børs om kontroll og kvalitet på data/informasjon som gjelder finansielle forhold herunder oversikt over personalressurser og deres organisatoriske tilhørighet. Loven skal også sikre at personer ikke opererer utenfor sine rettigheter [4]
ISO 17799 (BS 7799) berører indirekte datakvalitet via 'Compliance' (tilpasset IT systemer til for eksempel lovverk og forskrifter).
'Department of Commerce'[20]: *Provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies.* Flere andre departement har også lignende definisjoner[21].
Interessant å legge merke til hvordan de forskjellige lover og standarder betegner graden av kvalitet, eksempelvis *tilstrekkelig*, *maksimert* og så videre.

2.2 Kunnskap direkte relatert til forskningsspørsmål

Hvilken kunnskap finnes allerede for de konkrete spørsmål som reises i denne oppgaven?

2.2.1 Forståelse av datakvalitet

Hvordan oppfattes datakvalitetsbegrepet og måling av dette i industrien? Foruten de referanser som allerede er henvist i dette kapitlet er det ikke mye som er funnet direkte relatert dette spørsmålet. Hva status er i dag i norsk industri vedrørende måling av datakvalitet generelt og i HR området spesielt er det også funnet lite informasjon om. Noe informasjon er å finne i 'Sjekkliste for datakvalitet i informasjonssystemer'. Denne er laget i regi av Den Norske Dataforening [22].

2.2.2 Hvordan måle kvalitet

Hvordan måle kvalitet for IP som inngår i HR området? Det er ikke mye som er funnet på dette området gjennom litteratursøk og som gjelder HR området. Av mer generell karakter, men relatert

[4] Den norske **aksjeloven** - Lov av 13. juni 1997 nr. 44, legger også føringer på kontroll og kvalitet.

til temaet, har Wang m.fl.[9] undersøkt automatisk bedømmelse av datakvalitet og beskriver her bruk av kvalitetsindikatorer og kvalitetsparametre. Når vi vurderer måling av datakvalitet for ansattdata hvor det kan være flere kilder involvert, er det gjort en del undersøkelser som beskriver slike sammenhenger, som 'The Polygen Model', 'Data Source Tagging Problem'[9], 'The Attribute Based Model'[9][12] og det som nevnes i 'Record Matching and the Object Identity Problem'[23].
Måling av kvalitet er nok noe omstridt, inkludert HR data. Dette blir som tidligere nevnt uttrykt i Maxwell sitt innlegg om kvalitet ved data i HR informasjonssystemer[13].

Hvor gode er våre data? For å besvare spørsmålet må vi ha funksjonelle metrikker for datakvalitet. Wang m.fl foreslår metode for å komme frem til generelle metrikker [16] innen forholdstall, minimum eller maksimum operasjoner og gjennomsnitt. Metrikkgrupper er studert av Umar m.fl.[24]. Her nevnes grupper som omhandler: Data, applikasjon og plattform. Metrikker er foreslått i generelle termer og klassifisert i prioritet, målemetode (verktøy), frekvens, kost, valgbarhet (må/kan måles). I en artikkel som er skrevet av sentrale personer i TDQM [12], er attributtbasert datakvalitet fokusert. Det slås fast at 'return of investment' (ROI) ved absolutt korrekte data kan være ulønnsom og at man kanskje skal vurdere *graderinger av troverdigheten* til attributter. For å få dette til innføres 'attributt-tagging'. Det er ikke nevnt hvordan denne teknikken kan overføres fra et rent relasjonsteoretisk konsept som rapporten omfatter, til praktisk nytte i et aktuelle miljø. Det er verdt å merke seg at forfatterne modellerer kvalitetsegenskapene inn i den forretningsorienterte modellen og hevder at disse to modellperspektivene må behandles som en atomisk enhet. Det betyr om man endrer på noen av attributtene i den forretningsmessige modell må også kvalitetsegenskapene endres.

2.2.3 Hvilke metrikker er sentrale

Lite er funnet ved litteratursøk utenom det som er nevnt generelt under 'metrikker og rammeverk' (kap. 2.1.2), men en generell studie omkring dette tema er utført i [16] ved å sammenligne subjektive og objektive målinger og analysere avviket ved disse to målemetodene. Prinsipper nedfelles for å hjelpe til å utvikle metrikker for datakvalitet.

3 Metodevalg

I denne oppgaven fremlegges et skriftlig sluttprodukt. Det skal inneholde både en teoretisk betraktning av begrepet datakvalitet og data som beskriver hva forskjellige selskaper mener bidrar til datakvalitet ved ansattdata. Videre gjennomføres en tolking av resultater fra undersøkelser i lys av det teorigrunnlaget som er lagt fram, og eget bidrag til å få bedre forståelse for datakvalitet ved ansattdata og eventuelt finne løsning på hvordan måle dette.
For å kunne finne frem til svar i denne oppgaven har jeg benyttet spørreundersøkelse, intervjuer, litteratursøk og eksempel på en praktisk implementering (programkoding).

3.1 Spørreundersøkelse

En av metodene som er blitt benyttet er spørreundersøkelse (se side 57). Dette skal gi et fundament for forståelse av dagens situasjon ved datakvalitetsbegrepet, hvordan dette praktiseres og hvordan miljøet for dette er i forskjellige bedrifter. Det er også tatt med noen utenlandske selskap for å eventuelt sammenligne karakteristikker mellom norske og utenlandske miljø.
Undersøkelsen består av et sett spørsmål som hver har fått et unikt nummer (Qnn). Disse numrene benyttes til referanser så det senere er enklere å relatere data ved analyser og konklusjoner.
Momenter ved undersøkelsen er som følger:

1. Hva mener man om begrepet datakvalitet
2. Bruk av målinger ved datakvalitet
3. Vektlegging av forskjellige dimensjoner ved begrepet datakvalitet
4. Kilde til ansattdata

For å komme frem til et fundament for spørreundersøkelsen ble det laget en samvirkemodell for datakvalitet (figur 3.1). En samvirkemodell løser opp kravene i forhold til en kausalmodell. De faktorer som påvirker datakvaliteten er gruppert i forhold som kan ha betydning. Modellen sier ingenting om de innbyrdes forhold mellom egenskapene, bare at de virker sammen og gir antatte resultater. Modellen ble basert på relatert arbeid samt forhåndsundersøkelser som ble foretatt i forskjellige firma.

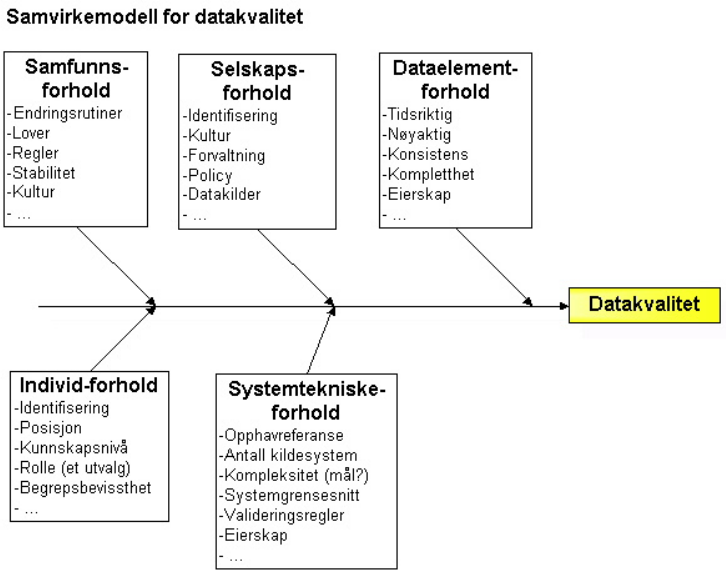
3.2 Intervjuer

Intervju ble benyttet for å kunne utdype problemstillinger etter at spørreundersøkelsen var gjennomført. Intervjuet ble basert på en enkel guide (se side 71) hvor det på forhånd var listet momenter som var viktig å belyse. Det var også viktig at alle intervjuede personer ble konfrontert med omtrent de samme problemstillingene.

3.3 Litteratursøk

Litteratursøk foregikk via studering av et utvalg bøker, spesifikke søk på Internett, bistand fra biblioteket ved høyskolen samt søk i anerkjente databaser som:

- CiteSeer. Referanser til artikler, proceedings, papers etc. og som inneholder siterings-funksjon.



Figur 1: Spørreundersøkelsen ble laget på grunnlag av en samvirkemodell for datakvalitet. En slik modell løser opp kravene i forhold til en kausalmodell. De faktorer som påvirker datakvaliteten er gruppert i forhold som kan ha betydning. Modellen sier ingenting om de innbyrdes forhold mellom egenskapene og er basert på relatert arbeid samt forhåndsundersøkelser som ble foretatt i forskjellige firma.

- ISI Web of Science. Dette er en portal for bl.a. Science Citation Index og er en del av Web of Knowledge.
- SpringerLINK. Referanser til artikler fra flere hundre elektroniske tidsskrift og serien Lecture Notes in Computer Science (LNCS) fra Springer-Verlag. Dette er en fulltekstdatabase.
- IEEE. Kilde for vitenskaplig litteratur innen datamaskin- og elektronikk-teknologi.

Litteratursøk er i denne oppgaven en vesentlig del av veien frem til økt kunnskap for å bedre forstå de ideer som allerede er fremsatt om datakvalitet og måling av dette. Forutsetning for denne metode er at det faktisk finnes relevant litteratur. Relevant informasjon er ordnet i egen bibliografisk database[1] som er tilgjengelig for interesserte. Denne strukturingen ble uunnværlig etter hvert som referanselisten vokste. Et av mine utgangspunkt for litteratursøk er hentet fra [10]. Her framstilles relativt aktuell forskning på området datakvalitet og fremtidige forskningsretninger innen temaet.

3.4 Eksperimentering

For å kunne sannsynliggjøre en måte å implementere måling av ansattdata benyttes et eksperiment. Eksperimentet baseres på en modell for bruk av databasetrigger i et HR miljø med hjelp av relevante datastrukturer for informasjonsprodukter og organisasjon.

3.5 Forskningsspørsmål og metodevalg

Hvilke metoder er passende å benytte i forhold til de forskningsspørsmål som er reist? Som innledet i 1.4, er det fremsatt en hypotese for datakvalitet og satt opp et sett med forskningsspørsmål. De *variabler* som fremkommer i denne hypotese vil være sentrale punkter

[1] JabRef version 1.6, Public License

i det videre arbeid og et utgangspunkt for å lage samvirkemodellen (3.1). Under listes forskningsspørsmålene med tilhørende metodebeskrivelse.

3.5.1 Hvordan oppfattes datakvalitetsbegrepet?

For å framskaffe kunnskap om situasjonen i dag, er spørreundersøkelsen en sentral komponent i denne oppgaven. Det ble innledningsvis foretatt intervjuer av noen aktuelle personer i eget HR miljø for å få grunnlagskunnskap før spørreundersøkelsen ble konkretisert.
Innledningsvis undersøkte jeg hvilke selskap som kunne være aktuelle å be om å bli med i en undersøkelse ved hjelp av enkel systematisk utvelgelse. Kriteriene er hovedsakelig norskeide selskap som har eierinteresse i utlandet. En henvendelse til Norges Bank og Statistisk Sentralbyrå om bistand til utvelgelse ga ikke resultater da disse ikke kunne utlevere data om enkeltforetak i henhold til statistikkloven. Til hjelp i dette arbeidet benyttet jeg da [25] som en kilde. Det ble så laget et informasjonsskriv og en presentasjon av oppgaven ble sendt utvalgte selskap etter først å ha avtalt dette med en representant for selskapet pr telefon. For å utarbeide spørreundersøkelsen er det nyttig å benytte flytdiagrammer og å følge forslag til fremdrift som antydet i [26]. I denne oppgaven forenklet jeg dette litt og et såkalt tankekart, også benevnt MindMap, ble benyttet istedet. Hvilket svar som var ønsket ga grunnlag for spørsmålene.
For å komme frem til spørreskjemaet, baserte jeg meg på innledende undersøkelser i forprosjektet, noen flere kvalitative undersøkelser og benyttet de variable som var fremkommet. På dette tidspunkt ble en klyngeutvelgelse benyttet [27] da det nå er kjent hvilke selskap og roller i selskapet som blir med i utvelgelsen. Klyngeutvelgelse er ressursmessig billigere å gjennomføre. Laget så en enkel samvirkemodell og utformer spørreskjemaet etter denne modellen.
En kvantitativ spørreundersøkelse vil også kunne anta subjektive retninger som for en kvalitativ undersøkelse. Spørsmålsstilling, rekkefølge, innelukking av valgalternativer, teoretiske variable som skal konverteres til måltall osv. Disse forhold ble forsøkt tatt i betraktning ved utarbeidelse av spørreskjemaet.

Etterfølgende intervjuer

For å få noe mer dybdekunnskap og belyst de mer kvalitative sider ved spørreundersøkelsen, ble det utført intervju av noen utvalgte personer. Utvalget ble foretatt på bakgrunn av firma, rolle i firma og funksjonell/teknisk funksjon i forbindelse med bruk av HR data.

3.5.2 Hvordan måle kvalitet for IP?

For å avklare spørsmålet om hvordan måle kvalitet for IP som inngår i HR området, ble det benyttet data fra spørreundersøkelse, intervju og litteratursøk. Litteratursøket var en vesentlig del av metoden for å komme frem til et fundament for å kunne svare på dette forskningsspørsmålet. Intervjuet ble også en viktig faktor i denne forbindelse.

3.5.3 Hvilke metrikker er sentrale?

Hvordan finne ut hvilke metrikker som er sentrale for å måle kvaliteten i ansattdata? For å komme frem til svar på dette spørsmålet ble det hentet data fra spørreundersøkelsen og intervjuer samt litteratursøk. Dette ble gjort for å finne ut av hva som er alment akseptert for måling av datakvalitet generelt, og hva som er spesielt i HR og ansattdata.

3.5.4 Hvordan praktisere metrikker?

Her ble også spørreundersøkelse,intervju og litteratursøk ble benyttet for å finne et mulig utgangspunkt. Basert på innhentet kunnskap ble det valgt en praktisk metode for å konkretisere

muligheten ved å måle et IP med en sentral metrikk i et alment benyttet miljø som i et relasjonsdatabasemiljø (rdbms). Det ble utarbeidet en modell som viser mulig struktur og en implementering som viser praktisk anvendelse. Dette siste er nødvendig da det i flere sammenheng er påpekt problemer med en slik implementering sett i forhold til uleselig kode og ytelse[28]. Metodene som ble brukt her ble basert på generelle analyse- og datamodelleringsteknikker.

4 Generelt om triggere og måling

Trigger er et begrep som benyttes i forbindelse med databasesystemer og ble definert i SQL:1999 (som vi også kjenner som SQL3). Noen databasesystemer har hatt dette i mange år, andre har det ennå ikke implementert. Dette er en spesiell form for lagrede prosedyrer som aktiviseres ved insert, delete, update av en gitt tabell i en database. Triggere benyttes oftest til å sikre referanseintegritet og er tenk benyttet i et utvidet integritetsbegrep i denne oppgave. Det er viktig å være klar over svake og sterke sider ved denne teknologien. Blant de sider som er verdt å nevne er mulighet for sentralisert kontroll og logikk. Ved bruk av trigger kan viktig kontroll implementeres et sentralt sted og ikke spres i brukergrensesnitt eller omkring i applikasjonslaget. Dette medfører også mindre kostnader enn om denne kontrollen skulle implementeres i distribuerte funksjoner i applikasjonslaget. Triggere kan håndheve restriksjoner som er mye mer komplekse enn deklarative implementeringer og de kan operere på kolonnebasis og sammenligne tilstand før og etter en endring som i en 'hva hvis analyse'.
Lagrede prosedyrer er en samling SQL setninger og valgfrie setninger for kontrollflyt som lagres under et navn i databasen. Denne kan kalles med parametre. Mer om triggere i [29],[30] og de enkelte referansehåndbøker for databasesystemene.

Server-baserte teknikker for å redusere nettverkstrafikk

Triggere, lagrede prosedyrer og views kan benyttes for å redusere nettverkstrafikk. Dagens beregningskraft på servere er stor i forhold til alminnelige kommunikasjonsnettverk. Sammen med en sentralisert plassering av logisk og fysisk håndtering av integritet og reduksjon av nettverksbelastning, vil det kunne være positivt å betrakte triggere også i en sammenheng hvor målinger av datakvalitet diskuteres.
Det er også flere motivasjonsfaktorer som taler for bruk av triggere, som eksempelvis håndtering av forretningsregler, noe som er vanskelig med kun bruk av deklarative utsagn i databaseskjemaet. Deklarative utsagn håndterer kun referanseintegritet mellom entiteter og bruk av NULL/ NOT NULL. Et annet moment er den sentraliserte kontroll som kan opprettholdes og som er spesielt viktig med så kalt 'mastring' av data. Mastring betyr her et sentralt lagringspunkt for data, om dette sentrale punktet er fysisk eller virtuelt har ikke betydning i denne sammenheng.

4.1 Triggerytelse

Kjøres alt på server ser vi at trigger-tillegget i tid er betydelig i det aktuelle testmiljøet. Men som vi ser av målinger utført og vist i vedlegg F, tabell 22, vil tidstillegget med trigger benyttet i en nettverksbasert kommunikasjon, være marginalt. Om dette er akseptabelt avhenger av brukerens oppfatning av responstid. Testene i 22 og 21 er gjort i et testmiljø uten annen trafikk. Nettverket var basert på 10Mbit ethernett med Sybase ASE 12.5 server og Windows 2000 maskin med Intel 4, 1.7Ghz prosessor og 1.5 GB ram. Disse målingene gir oss en indikasjon om hvor realistisk det er å kunne implementere *kvalitetsmoduler* i databasesystemer med bruk av triggere. Av de tester som er utført ser vi at responstiden øker med få prosent i nettverksammenheng. Brukt i transaksjonsmiljøer med relativt lite volum, som i et HR miljø[1], skulle dette ikke medføre flere ulemper enn de store fordeler det er å benytte triggere.

[1] Regner med 10% endringer (turnover).

5 Generelt om metrikker

Bruk av metrikker i denne oppgavens kontekst er omtalt blant annet i [16][31].
Avklaring av begreper måling og metrikk som benyttes kan være greit og vi holder oss til denne definisjonen på metrikk og måling [32]:

Metrikk benyttes til å analysere sammenheng mellom to eller flere målinger, inkluderer definisjon på hvordan det skal måles samt beskrivelse av relasjonen mellom gjentatte målinger.

En måling er et bilde ('snapshot') av en tilstand i et definert punkt og med bruk av diskrete numeriske verdier.

5.1 Generelt

Som vi har sett av relatert kunnskap er datakvalitet et multidimensjonalt konsept og innen dimensjonene er det igjen flere nivå. Eksempelvis kan vi betrakte datakvalitet i et konseptuelt , teknisk, operasjonelt eller driftsmessig perspektiv. I dette har vi igjen flere kvalitetsdimensjoner som: Nøyaktighet, tidsriktig, tilgjengelig, entydighet, komplett, med flere. Eksempler måling av datakvalitet:

- Tilbakemelding fra sluttbruker (eksempelvis telle klager til 'Brukerstøtte')
- Spørreundersøkelser rettet mot sluttbruker, systemdesigner, osv.
- Sammenligne alternative kilder (ansattregister mot folkeregister)
- Analyse av systemlogger (eksempelvis ETL [1] logger i datavarehusmiljø)
- Referanse integritet (på flere nivå, fysisk og logisk)
- Kompletthet og integritet (via normaliseringsregler i rdbms)
- Verdiområder, domeneområder
- Ikke eksisterende og manglende data (manglende master i master-slave forhold)
- Eksistens av metadata (både teknisk og forretningsmessig informasjon)
- Samsvarighet[2]
- Sannsynlig datadistribusjon (gjennomsnittsmålinger)

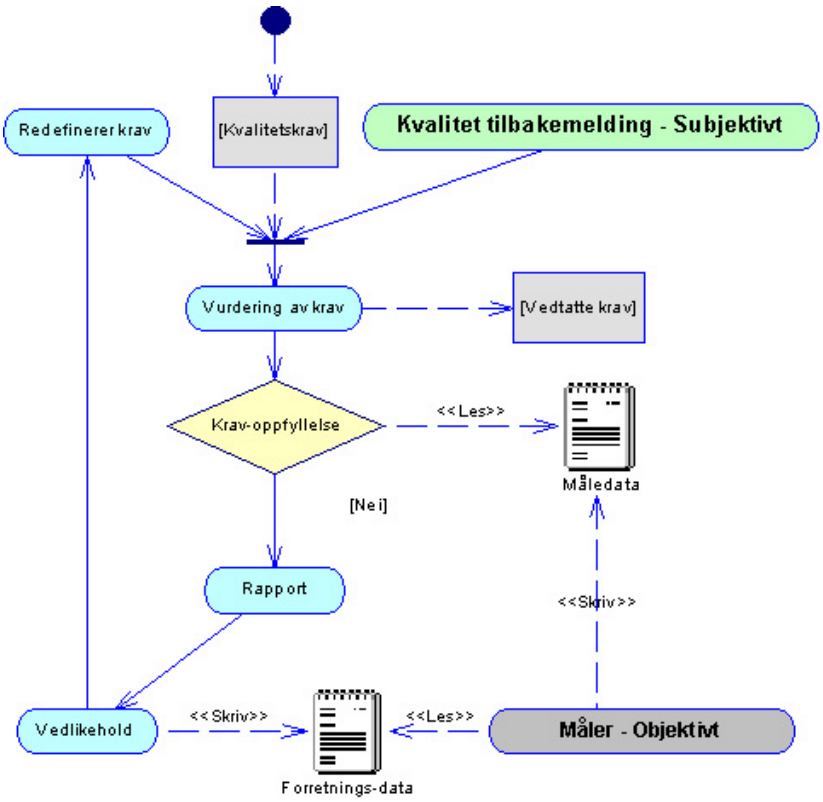
5.2 Gyldighet og pålitlighet

Pålitelighet er et resultat av hvordan vi har utført målingene. Gjentatte målinger skal gi mest mulig likt resultat. Feilene i hvert ledd i måleprosessen og senere bearbeiding, må være minst mulig. **Gyldigheten** til de målte data forteller oss om de data vi måler virkelig er de data vi ønsker å måle. Uten gyldige data spiller det mindre rolle om påliteligheten er høy og motsatt.

[1] Operasjonene som ekstraherer,transformerer og laster data i et datavarehus miljø, betegnes ETL
[2] Som er i samsvar med eksempelvis lover og regler, policy, rettingsregler og beste praksis

5.3 Metrikker i organisasjonen

Bruk av metrikker og måling har mindre nytteverdi dersom man ikke samtidig benytter den kunnskap som fremkommer videre i foredling av datakvaliteten opp mot de mål som er satt. Om man ikke har en kultur i bedriften for måling og foredling av datakvalitet, er det nødvendig å etablere en prosess for å lage metrikker som er tilfredsstillende for både brukere, forvaltere, ledere, utviklere og eventuelt eksternt kontrollerende myndigheter av system(ene). Veien frem til gode metrikker skal også oppfattes som motiverende, lærerik og bevisstgjørende for de deltagende parter. *Poenget er å være bevisst hvilken datakvalitet bedriften har og en bevisst kultur for å forvalte dette på en optimal måte*. Dette er bedriftens ansvar.
Mye av dette er generelt men nevnes likevel i denne oppgave fordi sammenhengene er så viktige i det overordnede bildet. Eksempler på kvalitetsprogram som er omtalt og utprøvd er: Total Data Quality Management (TDQM)[33] og Data Profiling Technology (DPT)[28] som er en datasentrert løsning som identifiserer unøyaktige data og tar aksjoner for å forbedre data nøyaktigheten. Prosesser som driver frem og vedlikeholder metrikker kan illustreres som vist i figur 2



Figur 2: Prosess for administrering av metrikker som kan betraktes som en del av et totalt administrasjonsmiljø. Modellen tar utgangspunkt i [34], prosesser og bevisstgjøring. Som vi har sett er det et innslag av subjektiv kvalitetsvurdering og i denne sammenheng må vi ta denne i betraktning ved vurdering og revurdering av krav.

5.4 Metrikk krav

Metrikker må være utformet slik at de er formålstjenlige i bruk. Det skal være et effektivt verktøy for å bestemme i hvilken grad man skal øke eller minske kvalitetstiltak.
Gode metrikker skal være lette å forstå, enkle og ikke måle personer men prosesser. De skal være resultatorientert og belønne de aktiviteter man selv har kontroll over og være 'SMART' [32]:

- **S**pesifikke
- **M**ålbare
- **A**nvendlig (Realistisk/Oppnåelig)
- **R**epeterbare
- **T**idsuavhengige

Metrikker kan brukes til å indikere i hvilken grad objektive kvalitetskrav blir nådd og til å forbedre virksomhetens kvalitetsprogram og planer.

5.5 Metrikkalgebra

5.5.1 Nøyaktighet

Wang m.fl.[14] beskriver dimensjonen 'nøyaktighet' som:

Nøyaktighet = 1 - ((Korrektverdi - Aktuellverdi)/Korrektverdi)

Problemet at det ikke er gitt noen beskrivelse om bruk av tekst verdier om man ønsker å måle slike data. Tekstverdier er aktuelt i HR sammenheng. Dette medfører at vi eventuelt må introdusere en funksjon for å konvertere tekstvariabel til et tall. Flere algoritmer for behandling av tekster er definert i blant annet [35]. Nøyaktighet kan være et forhold mellom tekster, relasjoner mellom entiteter, attributter i entiteter, mm. Ser vi på nøyaktighet som et forhold mellom *entiteter* er det enklere. Da kan vi beregne antall faktiske relasjoner mot aktuelle relasjoner.

5.5.2 Nyttegrad

Vekting av distinkte elementer er benyttet i mange sammenheng ([16]) og her utnytter vi teknikken i blant annet beregning av *kompletthet i informasjonsproduktet*. Av dette innfører vi begrepet 'nyttegrad' til et element i et IP. Dette kan også opptre som et viktig metadataelement da det forteller oss noe om *forventningene* til korrekthet i et IP.
Som vi erfarer fra intervju er ikke alle elementer like viktige i de forskjellige sammenheng. Eksempelvis vil ansattnummer og organisatorisk tilhørighet være MÅ elementer med krav til høyeste nyttegrad i et gitt IP. På den andre siden trenger ikke element som privatadresse ha like høy nyttegrad i samme IP. Dette kan utnyttes ved kontroll av kompletthet ved en leveranse av et IP. Det er ikke like nyttig å måle kvalitet i alle elementer i et IP. Som vi skal se i undersøkelsen er det i HR miljøet en stor andel ustrukturerte data og det er kanskje i praksis noen ganger umulig å måle datakvalitet i enkelte element.
Nyttegradselementet kan også vise seg å være en sentral komponent i et konsept for *brukerkontrollert kvalitetsmåling*. Brukeren kan da slå av og på de kontroller som for tiden er av interesse eller vekte de forskjellig i forskjellige situasjoner. Referer kap. 2.1.4 hvor Maxwell påpeker det dynamiske miljø.

Nyttegraden kan vi definere som 'nge' (nge= nyttegraden av element) og kan variere fra 0 til 1 og hvor viktige elementer i IP er lik 1 (må eksistere og må være riktige) 0 indikerer et ikke

nødvendig felt.
En nyttegradsmodell kan betraktes slik som i figur 14 (side 75) og kan være en aktiv del ved måling av IP ved at eksempelvis kun attributter med nge = 1 måles, eller tekster med nge < 1 måles med tilnærmede verdier (eksempel bruk av soundex[3] lignende funksjoner). Modellen viser hvilke roller som benytter hvilke IP og hvilke attributter som inngår i disse IP med hvilken nyttegrad. Modellen kan også inngå som et element i en metamodell. Nyttegraden kan nå beregnes ut fra

<Nyttegrad> ::= (<nge>+, . . .)1*n /<Antall attributter i IP>

Eksempel:
<KatalogIP> ::= <fornavn>, <etternavn>, <ansattnr>, <telefon>, <adresseJobb>, <adresse-Privat>
<KatalogIP> ::= ‘1’,‘1’,‘1’,‘1’,‘0,5’,‘0,5’
Gir verdien 5/6 = 0,83. Dvs nyttighetsgrad = 0,83 i et bestemt IP og i en bestemt kontekst (eksempel: telefonregister). Alle ‘1’ skal være med i IP.
I en annen kontekst kunne nyttegraden se slik ut: <KatalogIP> ::= ‘1’,‘1’,‘1’,‘0,5’,‘1’,‘1’
Gir verdien 0,92.

5.5.3 Referanseintegritet

En viktig faktor i eksempelvis systemer for tilgangskontroll, er den konseptuelle entitets integritet.
Integritetsregler ([29]):

- *Entitet integritet : Ingen komponent i en primærnøkkel kan ha NULL verdi.*
- *Referanse integritet: For hver distinkt ikke-NULL ‘fremmednøkkel’ verdi i databasen, må det eksistere en tilhørende primærnøkkel fra samme domene.*

For en forekomst a i Entiteten A skal en forekomst l i Entiteten L eksistere (referanse integritet).
Forhold som dette må kunne uttrykkes i en metrikk. Skulle ikke være så vanskelig og er en selvfølgelig del av grunnlaget til kvalitetsmåling av integritet i ansattdata.

5.6 Metrikk beskrivelse

Generell beskrivelse av sikkerhetsmetrikker som direkte omhandler konfidensialitet, integritet og tilgjengelighet er utført i [34]. I denne forbindelse modifiserer vi beskrivelsene til å omhandle egenskaper ved datakvalitet og informasjonsprodukter (Se tabell 1) og legger til 3 nye kvalitetselementer. Disse tre elementene er: trusseleksponering, målepunkt referanse og IP relasjoner. Undersøker vi andres arbeid med hensyn på beskrivelse av metrikker i HR sammenheng er det lite å finne av konkretiseringer, men det foreligger mye omkring dimensjoner og egenskaper ved disse. Eksempel er ‘Data Quality Assessment Framework’[8].

[3] Soundex funksjon beregner likhet i uttale.

Attributt	Beskrivelse (*relatert datakvalitet og informasjonsprodukt*)
Kvalitets-element	Beskrivelse av hva som skal måles
Tekstlig utdypning	Detaljert beskrivelse av kvalitets element.
Metrikk	Definisjon av kravet som stilles for å oppnå det kvantitative målet til metrikken.
Metrikk-formål	Hvorfor skal denne metrikken brukes i en kvalitetsmåling.
Krav	Konkrete forutsetninger som settes i fokus for å sikre metrikkens oppfyllelse. Hvilke kvalitetstiltak stilles det krav til å implementere?
Frekvens	Hvor ofte skal målinger foretas.
Skala	Måleskalaer forteller hvordan vi måler og tolker de målte verdier. Hvilken målestokk som benyttes for målingen. Metrikker skal kunne kvantifiseres, da kvalitative begreper er vanskeligere å måle mot. Eksempel: antall, prosent eller gjennomsnitt.
Formel	Utregninger (gjennomsnitt, tellinger, forholdstall) som ligger til grunn for å beregne metrikken.
Datakilde	Henvisning til hvor datagrunnlaget fra målingen hentes fra. Kan ha ulike kilder, eksempel: automatiserte metoder, dokumentgjennomgang, kartlegginger, intervju, spørreundersøkelse, gjennomgang av systemkonfigurasjon eller ved observasjon.
Indikatorer	Beskrivelse av hva det betyr at metrikken blir nådd eller ikke nådd samt trender for målingene.
Godhet og eierskap ved metrikker	
Pålitelighet	Målemetodens målepresisjon. Hvilken feiltoleranse aksepteres og hvor reproduserbare er målingene. Pålitelighet avhenger av den operasjonelle definisjonen av metrikken for eksempel detaljeringsgrad og kompletthet. Introduksjon av tilfeldige feil svekker påliteligheten.
Gyldighet	Måles det vi tror skal måles og det vi virkelig er interessert i å måle? Systematiske feil svekker gyldigheten av metrikken. Gyldighetsfeil kan kun reduseres gjennom forbedring av metrikkdefinisjonen vedrørende operativitet og valg av attributter som kan måles. Er det vanskelig å finne direkte målinger, som måler det en vil, kan kryssmålinger benyttes som kan gi indirekte svar på metrikken.
Gjennomfør-barhet	Hvor lett eller vanskelig det er å utføre målingene? Det kan eksempelvis være tekniske, administrative eller personalmessige problemer som gjør at målingene ikke så lett kan gjennomføres.
Konflikt-områder	Beskrivelse av konflikter ved kvalitetsmåling.
Eierskap	Hvem eier denne metrikken og har anledning til å endre den?
Kostnader og anvendelser	
Kostnader	Hvilke ekstra kostnader er forbundet med å gjennomføre målinger og metrikken. Kostnader som uansett kommer i forbindelse med kravene metrikken setter vil ikke inngå som kostnad her.
Konkrete an-vendelser	Beskrivelse av hvilke bruksområder metrikken har. I hvilken kontekst kan metrikken benyttes.
IP relasjoner	Beskriv hvilke spesielle forhold skal ivaretas i dette informasjonsproduktet.
Målepunkt referanse	Hvilket målepunkt er aktuelt i forhold til målepunkter i rammeverk (figur 8).
Trussel-eksponering	Hvilken trussel metrikken eventuelt eksponerer relatert til informasjons-sikkerhet.

Tabell 1: Beskrivelse av metrikk innhold

6 Undersøkelser

6.1 Spørreundersøkelse

Det som er viktig med denne spørreundersøkelsen er først og fremst å finne dagens status for måling av datakvalitet i HR miljø. Det er fremsatt en arbeidshypotese og forskningsspørsmål som illustrerer hvilke variabler som kan tenkes å påvirke datakvalitet i et selskap, som vist i samvirkemodellen. Når variablene skal operasjonaliseres velges ofte begrep som vi mener er enkelt å måle, eksempel antall omorganiseringer siste 3 år. Spørsmålet er da: er denne operasjonaliseringen dekkende?

6.2 Gyldighet og pålitlighet ved undersøkelsen

Vi skal ha et spørreskjema som er enkelt å besvare samtidig som det skal gi mest mulig troverdige svar. Denne sammenhengen er ikke like enkel å få til. Her har jeg valgt å vektlegge enkelhet i svarene. Påliteligheten i svarene bør bli høyere med dette utgangspunktet men den definisjonsmessige gyldighet kan bli noe lavere. Antall svar som behandles påvirker påliteligheten, jo flere svar som behandles jo bedre pålitelighet. Utvalget i denne undersøkelsen er alle innen HR området og rollene er fordelt i tre kategorier. En svakhet her er at det er få fra kategoriene ‘teknisk’ og ‘funksjonær’. Kategorien ‘HR ledelse’ er desto større. For å bøte på denne svakheten i antall spurte ble et dybdeintervju utført. Her ble 5 personer intervjuet med ønskelig rollefordeling: HR leder, teknolog og funksjonær.

6.3 Kilder til undersøkelsene

Det ble sendt ut en forespørsel til norske selskap som, etter en telefonrunde, kunne være aktuelle for å delta i undersøkelsen. Disse har et større antall ansatte og som også opererte internasjonalt. Resultatet var at 25 spørsmålsark ble sendt ut, 19 svar ble mottatt fordelt på 8 bedrifter.
Flere av kildene ønsket anonymitet. Det ble lovet full diskresjon og anonymisering i presentasjon av svar. Flere av de potensiell kandidatene frafalt nettopp deltagelse med begrunnelse i at dette er opplysninger de ikke ønsker å utlevere. Andre kunne ikke delta på grunn av prioritering av tid.

6.4 Oppfølgingsintervju

Intervjuene skulle utfylle spørreskjemaene der det var nødvendig. Alle intervju er nøytralisert med hensyn på arbeidssted, systembetegnelser og personer. Nøytralisering var et krav fra flere av de intervjuede. Intervjuene ble avtalt noen tid i forveien og foretatt via telefon. Hvert intervju varte i ca 60 - 90 minutter.

Intervju guide

Det ble lagt opp til et mer brukerstyrt intervju, men skulle også berøre flesteparten av de momenter som ble laget på forhånd og som er listet under under. Det var viktig å lage en slik intervjuguide (se vedlegg D). Dette for å kunne stille omtrent de samme spørsmål til alle de intervjuede og at vi berørte de punkter jeg syntes var viktige. Ellers fikk intervjuobjektene uttrykke sin mening med sine egne ord og uttrykk. I de tilfelle det ble uttrykt relative mål uten kjent bias som eksempel kvaliteten er *ganske god*, tok vi tak i dette og forsøkte en konkretisering. Oppsummeringen av intervjuene er samlet under overskriftene: hva karakteriserer HR data?,

viktige IP i HR, hva tror du om kvaliteten og hvilket fokus har man på kvalitet.

Siktemål

Bidra til å utfylle svar gitt i spørreskjema. Personene skal ha et høyt informasjonsnivå innen behandling av HR data og representere ytterpunkter i forståelse av arbeidsområder, for å få et så bredt perspektiv på analyseområdet som praktisk mulig.
Skal videre utfylle kapittel 1-2 i spørreundersøkelsen og det som gjelder personens forståelse av begreper som datakvalitet og måling av kvaliteten av IP.
Personen som intervjues jobber med HR data. Det er ikke noe krav om at personen har levert noe svar på spørreundersøkelsen.

6.5 Folkeregisterundersøkelse

Hypotese og forskningsspørsmål kommer inn på forholdet ansattdata og *kildenøyaktighet*. Dette medførte at det ble sendt en forespørsel til Skattedirektoratet og flere aktuelle ambassader om landets folkeregistreringsystem og mulighet for å kontrollere ansattdata opp mot eventuell registre. Fra det norske ‘Sentralkontoret for folkeregistrering’ ble ‘Folkeregistreringsforskriften 2003’ returnert. Det ble foretatt et etterfølgende kort intervju som avklarte saksbehandlingstider og mulighet for å få utlevert data.
Ellers responderte de fleste ambassader. Kun en svarte direkte på spørsmålene og det ble mange direkte og indirekte henvendelser til andre mer eller mindre vanskelig tilgjengelige institusjoner. Det viste seg at oppfølgingen av dette tema ble for omfattende og komplisert for hva jeg kunne forvente å få tilbake og jeg hadde allerede tilfredsstillende svar fra norske myndigheter.

7 Funn ved undersøkelsene

Spørreundersøkelsen benyttet for det meste grupperte svaralternativer med innbyrdes rangering samt noe klassifisering. Undersøkelsen var gruppert i fem områder:

1. Generelt om datakvalitet,
2. Hvordan benyttes datakvalitetsmåling,
3. Datakvalitetsdimensjoner,
4. Kildesystem for ansattdata og
5. Intervjuobjekt og organisasjonen.

Oppsumering av alle grupper er vedlagt, ref. vedlegg B, side 61. Ved gjennomgang av funn i undersøkelsen vil det refereres til spørsmål i dette vedlegget.

7.1 Generelt om datakvalitet

Her ble det spurt om den generelle oppfatningen av datakvalitet og bruk av offentlige registre til kontroll av ansattdata. De relaterte spørsmål er Q1 til Q11.
Spørsmålene om internasjonalisering og datautveksling ga ikke noen spesielle resultater. De fleste mener at internasjonalisering påvirker datakvaliteten. Det stemmer jo også med de svar og analyser som er gjort vedrørende kvalitet og flere datakilder, flere kilder mer problemer.
De fleste (70% av ledere og 80% av funksjonærer) påpeker forholdet mellom nasjonale lover og forskrifter og krav til kvalitet i ansattedata. At dette tallet ikke er nærmere 100% kan bero på misforståelse. Forholdet lover og forskrifter kontra kvalitet på ansattedata ble av flere også påpekt i intervjuene. Når det gjelder utveksling av data internasjonalt gjør de fleste *ikke* dette. Godt over halvparten (63%) svarer at de har en *dedikert* person eller gruppe som administrerer forbedringer ved ansattdata i organisasjonen.
I vår hypotese relaterer vi godt definerte kilder til originale data (kap.1.4). Vi kunne da tro at de offentlige registrene ble utnyttet også ved kontroll av ansattdata, men det viste seg å ikke stemme. Kun 50% bruker slike registre helt eller delvis.

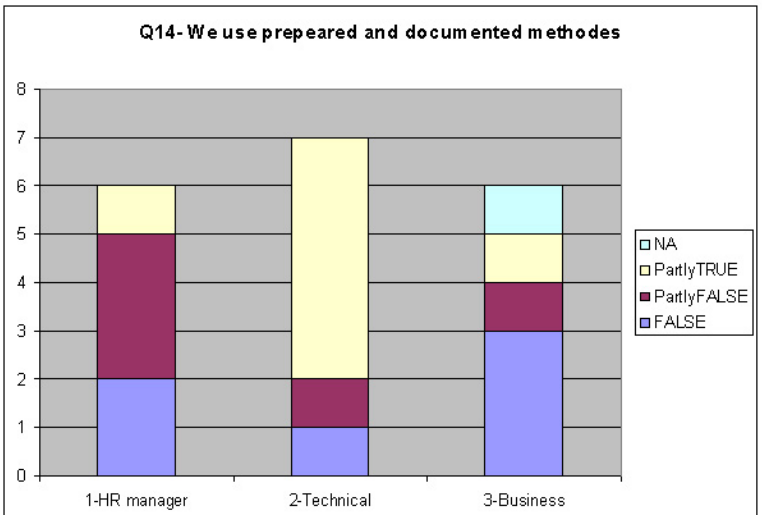
7.2 Hvordan benyttes datakvalitetsmåling

Undersøkelser om hvordan man foretar målinger av datakvalitet er relatert til spørsmålene Q12 til Q20.
Referer tabellen B.2 for detaljer i svarene og tabell C.4 for gruppering av svar i JA/NEI kategorier vs. rolle.

Måling av ansattdata (Q12 - Q15)

Utviklede og dokumenterte metoder for måling av kvalitet har *ingen*, se figur 3. Dette svaret var det klareste av alle svar i undersøkelsen. Og det ble bekreftet under intervjuene, ingen måler kvalitet på en planlagt måte. Hvorfor ikke? Av intervjuene fremkommer det at flere ikke tror på målinger, kanskje til akademiske (forskning) formål, kanskje til en viss grad for enkelte system og det må ikke koste for mye og ikke måle personers aktivitet.
Interessant er det når man spør hvor god datakvaliteten er. Svarene er da preget av subjektive

oppfatninger. I hypotesen gikk vi ut i fra at datakvalitet var preget av flere objektive enn subjektive kriterier. Dette ser ikke ut til å stemme.
Samtlige av de intervjuede ble spurt om de syntes det var ok å måle kvalitet i deres system. Ingen var i mot, noen veldig for og ønsket mer kunnskap på området, og noen litt mer skeptiske. Ikke alle måler datakvalitet i sine prosesser og dette kan komme av at man graderer systemene i viktighet og implementerer teknikker avhengig av viktighetsgrad (ref. 7.6.3). Hvordan rapporteres kvalitetsfeil? Vi kan lese at det kun er 25% av de spurte som har slike



Figur 3: Viser hvordan de forskjellige roller i organisasjonen svarer på spørsmål om forberedte og dokumenterte metoder for måling. 5 av 7 tekniske personer mener de har delvis slike metoder. Ingen av alle spurte mener de har slike metoder.

rutiner og at det er i de aller fleste tilfelle (94%) sluttbruker som helt eller delvis rapporterer feil.

Hvordan man avslører feil i ansattdata (Q16 - Q20)

De aller fleste utfører kontroller til regelmessige tidspunkt. Q17 reiser spørsmål om feil rapporteres av sluttbruker, og omtrent 100% svarer positivt på dette. Det er vel også et naturlig svar med en slik spørsmålstilling. Spørsmålet skulle vært pressisert om det *kun* er sluttbruker som rapporterer feil. Dette ble fulgt opp i intervju og der ble det bekreftet at det manglet rapporteringsrutiner.
Ser vi på Q20 oppdager vi at kun 50% sier de har en policy for feilrapportering mens de fleste (82%) hevder å ha systemer som er laget for å detektere datakvalitet. Vi kan da regne med at det detekteres feil som ikke blir rapportert.

7.3 Datakvalitetsdimensjoner

Dette er en sentral spørsmålsgruppe. Dimensjoner er viktig når vi behandler datakvalitet. Her ble det spurt om kvalitetsdimensjoner som er relevante i en HR sammenheng, se tabell 13 side 63.
Spørreundersøkelsen viser at det er særlig fem områder det er stor enighet om når det gjelder dimensjoner innen datakvalitet. Figur 4 viser antall prosent av vekting og er sortert på ‘Most Important’,5,. . . ,’Least important’. Dette er de dimensjonene som er mest vektet med hensyn på ‘Most Important’:

- Ensure that Employee data is used and stored safely (Q32).

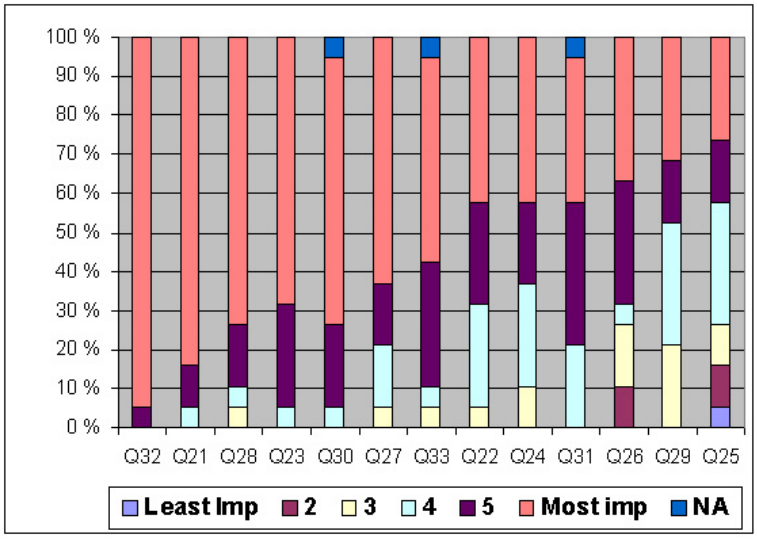
- Employee data must be accurate in relation to its source (Q21).
- Employee data is available when you need it (Q28).
- Employee data must be timely (Q23)
- Employee data has to be highly regarded related to its source and content (Q30).

Spørsmål Q23 - 'Employee data must be timely' er interessant. Sammenligner vi dette med spørsmål om å rapportere feil (Q20) ser vi at de færreste har slik policy i bedriften, noe vi skulle tro de skulle hatt utfra vekting av dimensjonen (Q23). Vi merker oss videre vektingen av dokumentasjon (Q25). Denne dimensjonen fikk lavest oppmerksomhet. Dette stemmer dårlig med de poengene blant andre Olson [28] har vedrørende viktighet av metadata i forbindelse med datakvalitet. Det ble også utført to spesielle analyser i sammenheng med dimensjoner, se vedlegg C.1 og 17. I C.1 ser vi en sammenligning av landenes gjennomsnittlige svarverdi for hver dimensjon. Nedenfor er de mest karakteriserte dimensjonene listet. Egne erfaringer viser viktighet av metadata i et miljø som utveksler informasjon på tvers av organisasjoner eller i miljø som utvikler informasjonsløsninger. Ser vi på rolle relatert til dimensjoner (ref: C.2) oppdager vi følgende interessante forhold:

Documented: De som er sluttbrukere av data ønsker best dokumentasjon. Hvorfor er ikke dette i like stor interesse blant teknikere og HR administratorer? Det samme forholde har vi ved

UniFormat: De som *bruker* data ønsker enhetlig format i størst grad.

Sourcemark:Interessant at det er teknikere som ønsker kildemerke, ikke administratorer eller funksjonærer i like stor grad. Dette kan begrunnes med at det er teknikere som først og fremst ser problemene med flere kildesystemer når it-systemer konstrueres og implementeres. Hele 82% av ledere mener det ikke er noe problem.

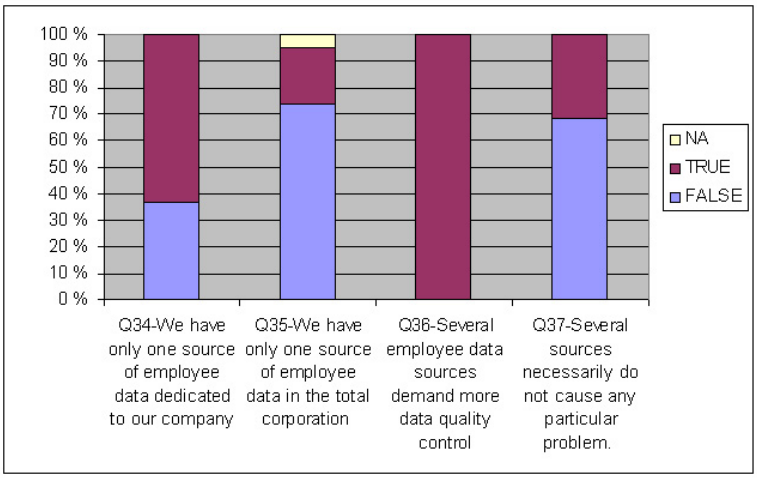


Figur 4: Viser hvilke kvalitetsdimensjoner de spurte har rangert som viktigst. Sortert på 'Most Important',5,. . . ,'Least important'. Viktige dimensjoner er Q32, 'Used and stored safely', hvor 95% mener dette er mest viktig, og Q21 'Accurate in relation to its source', hvor litt over 80% mener dette er mest viktig.

7.4 Kildesystem for ansattdata

Spørsmålsgruppe Q34 - Q37.

Referer tabell 14, side 64. Hva vi ser her er at de fleste (63%) stort sett har en kilde til ansattdata i sitt selskap men det er kun 37% som har kun *en* kilde. Ingen har kun en kilde for hele konsernet. Vi legger også merke til at mengden mener at flere kilder medfører dårligere kvalitet. Ved intervjurunden fremkom det at flere bedrifter har satt igang tiltak for å konvergere flere ansattkilder mot en felles.



Figur 5: Grafen viser et komprimert bilde (uriktig/riktig) av hvordan kilder for ansattdata blir oppfattet og benyttet. Fremstillingen baseres på data fra tabell 14. I tabellen er svaralternativene inndelt i uriktig, delvis uriktig, delvis riktig og riktig.

7.5 Intervjuobjekt og organisasjonen

Spørsmålsgruppe Q39 - Q43.

Se tabell 14, side 64. I denne seksjonen ble det spurt etter data om størrelse på organisasjonen, antall omorganiseringer siste 3 år, lederskifte, rutineendringer og om hvilken rolle svarkandidaten har i organisasjonen.

11 av de intervjuede kategoriseres som personalleder og eller HR leder, 4 personer har jobber som er teknisk orientert (IT-design, -utvikling, -arkitekt) og 4 personer er mest forretningsprosess orientert (brukere av HR systemer). Hvilken kategori som man tilhører er analysert med hensyn på forskjellige parametre (ref C, side 65 og etterfølgende sider.).

Ut fra svar som er gitt ser det ikke ut som om det er noen forskjell på antall reorganiseringer målt opp mot størrelse på organisasjonene, ref. tabell 2. Mindre avdelinger reorganiseres like mye som større. Vi skal være klar over at selv om det er en avdeling som har få ansatte kan den være styrt av større selskap og slik være en del av en mer omfattende reorganisering. Flere har erfart at

Antall ansatte	Antall reorganiseringer
250 til 1000	3
1000 til 5000	2
5000 til 12000	3

Tabell 2: Antall reorganiseringer de siste 3 år, per selskapstørrelse.

datakvaliteten har sammenheng med omorganisering, dette fremkom under intervjurunden. Som en uttrykte det: *Nye personer bruker lengre tid og blir satt til å operere systemer som krever effektivitet, settes under tidspress og registreringer feiler.*
De fleste (61%) intervjuede opplyste at arbeidsrutinene hadde endret seg etter omorganiseringen. Ser vi på hvilket fokus bedriftene har på datakvalitet ved reorganisering og sammenligner dette mellom de som har endret og de som ikke har endret arbeidsprosedyre (Q42 vs Q6), er det praktisk talt ingen forskjell.

7.6 Funn ved intervju

I det etterfølgende er en oppsummering av hva som fremkom i intervju. Oppsummeringen er gruppert i de temaområder som intervjuet omhandlet. Generelt kan vi oppsummere funnene slik:

- Ansattdata er omfattende og komplisert med mye ustrukturerte data.
- Noen data benyttes direkte mot adgangskontrollsystemer.
- Lite dokumentasjon av forretningsregler.
- Mangler rutiner for å sjekke datakvalitet.
- Generelt er det brukere av data som rapporterer dårlig kvalitet
- Temaet datakvalitet oppstår ad hoc.
- Omorganisering påvirker datakvaliteten

7.6.1 Hva karakteriserer HR data? (2)

Data som reflekterer aktuell situasjon om ansatte i et firma (oppdaterte). Personvern og personopplysningsloven (alle har forskjellige formuleringer), data må ha tillit og fortrolighet (konfidensialitet). HR data inngår som basis for aksesskontroller. Disse data må ha stor nøyaktighet. Finnes et organisasjonshierarki er det viktig at HR data er relatert til dette. Det er mye friformat felter i HR miljøet og disse kan være vanskelig å kontrollere.

7.6.2 Viktigste informasjonsprodukt i HR? (3)

Hva som er informasjonsprodukt er avhengig av hvem man spør i hierarkiet. Feil detekteres oftest av sluttbruker. Mye er underforstått med hensyn til datakvalitet. Knyttet til informasjonssikkerhet er det viktig å kunne kontrollere relasjonen mellom ansatte og organisasjonen da dette forholdet er med/kan være med å hjemle korrekt autorisering. Informasjonsprodukt som benyttes til aksesskontroll og som inneholder: Hvem som er ansatt, hvor, når begynt og når sluttet, er viktig. Lønnsdata, forfremmelsedata sykefravær, hms rapporter, evalueringsdata, ledelsekapasiteter, kostnadsstyring. Dette er produkter som spesielt nevnes av personer med ledelsesansvar. Telefonliste kan stå som et produkt men kan ha feil. Alt er altså ikke like viktig. Eksempelvis er hvor en sitter (lokalisering) viktig informasjon, ikke adresse.

7.6.3 Hvor god tror du datakvalitet er? (4)

Det er få som måler datakvalitet men man er bevisst behovet for å kunne oppnå bedre kvalitet. De fleste uttrykker et eksplisitt behov for å kunne måle kvalitet. Generelt er det brukere av data som rapporterer dårlig kvalitet. Det er uklarhet om hvem som har ansvar for data. Det hersker en subjektiv oppfatning av kvalitet. Som en uttrykte det: ‘Vet det bare’. Noen mente det var vanskelig å måle noe som ikke er der, for eksempel at en har aksess men er ikke ansatt. Noen nevner utfordringer ved sentralisert HR data behandling, det var bedre med lokal behandling av HR data, da kjente man personene.

Forretningsregler

Det er lite dokumentasjon av forretningsregler. Det er mangler på rutiner for å sjekke datakvalitet. Eksempel: Selskapskode ble endret, da gikk verdikjeden i stå. Ingen dokumentasjon over hva sluttbruker kan forvente av datakvalitet. Oppdager ikke nødvendigvis at et felt mangler, kontrollerer ikke input fra kilde, stoler på kilden. Forretningsregler er underforstått. Det finnes ingen formelle regler, ansatte kan relateres til fiktive organisasjonselementer. Mye er underforstått når vi snakker om datakvalitet. Forretningsregler er for noen ukjent innen HR området men man er sikker på at det finnes.

7.6.4 Fokus på kvaliteten (5)

Fokus varierer. Oppstår ofte ved brudd på aksepterte kvalitetsnormer. Det er behov for konkretisering og klargjøring av retningslinjer. Det snakkes mye om datakvalitet. Flere innfører nye systemer som blant annet skal bedre datakvalitet, men man har ikke klarlagt måltall for dette. Temaet datakvalitet oppstår ad hoc. Omorganisering påvirker datakvaliteten mente flere. Noen kom med eksempel fra egen organisasjon om omorganisering og feilregistreringer. Når det er organisasjonsendringer reflekteres dette ikke raskt nok i systemene noe som går utover datakvaliteten. Retningslinjer for datakvalitet er for dårlig. De er for personavhengig. Tungt å endre datakvalitet pga manglende ressurser.

7.6.5 Hva er en ansatt?

Hvordan definerer man en ansatt? Det er ikke så enkelt som skulle tro. I en HR database finner man langt flere personer enn de man kan betrakte som ‘aktive’ ansatte. I arbeidsmiljøloven [36] finner vi retningslinjer for hva som er en ansatt. Vi finner også bestemmelser om formell avtale mellom partene og hvilke opplysninger som skal finnes i en slik avtale. Med *arbeidstaker* mener loven enhver som utfører arbeid i annens tjeneste. Kapittel XI A omhandler tilsetting og lister hvilke informasjon denne skal minst omfatte. Dette er partenes identitet, arbeidsplassen, beskrivelse av arbeidet eller tittel, tidspunktet for arbeidsforholdets begynnelse, eventuelt forventet varighet, oppsigelsesfrister, lønn, arbeidstid, mm. I tillegg sier Kap. VI. noe om registrering og melding av arbeidsulykke og yrkessykdom.

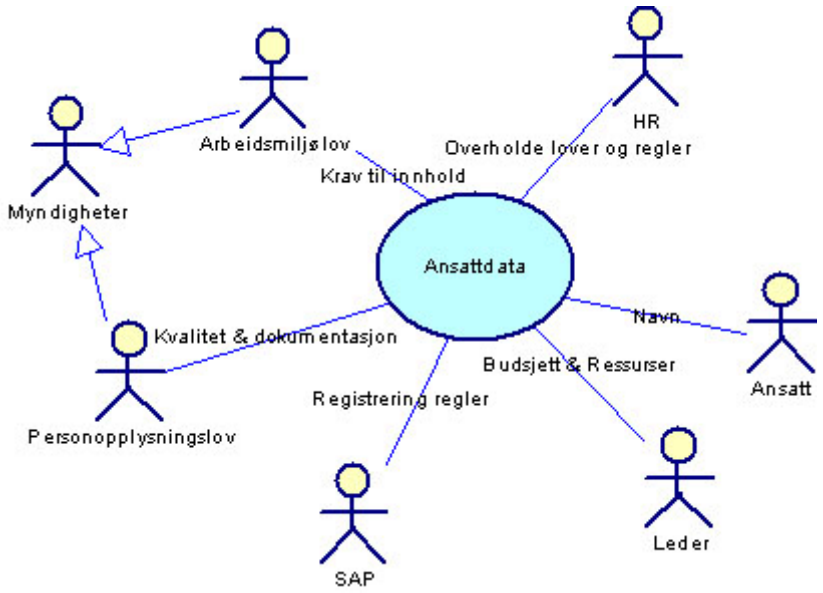
Interessenter ved ansattdata.

Som illustrert i figur 6 eksisterer forskjellige interesser rundt ansatt data. En ansatt har eksempelvis interesse av at sitt eget navn er stavet korrekt. I undersøkelser som ble gjort ble det referert til at det var den ansatte selv som var interessenten i å ha sitt navn korrekt skrevet og det var denne som skulle rapportere feil og mangler. Ledere er interessert i at ressursantall og budsjett stemmer, HR ivaretar ansvar for implementering av lover og regler osv.

7.7 Funn ved folkeregistrering

Norden har en overenskomst om folkeregistrering som ivaretar internordiske flyttinger. Alle har folkeregistre med noe forskjellige betegnelser.

Det eksisterer kun en sannhet om innbyggeres persondata og det er Folkeregisteret som opptrer som ekte kilde, kun dette og personen det gjelder selvfølgelig. Har bedrifter mulighet til å ’abonnere’ på slike data? Begrepet eksisterer ikke får vi opplyst av Skattedirektoratet. Dersom det er et juridisk behov kan man etter søknad få ut data. Se paragraf 13 i lov om folkeregistrering. Hvor lang tid tar det fra en endring skjer ved en persons data, til de er tilgjengelige? Fra 3 dager til måneder. Eksempelvis så tar skilsmissesaker lengre tid.



Figur 6: Ansattdata har mange interessenter. Den ansattes navn er det i stor grad den enkelte ansatte som har ansvar for å holde korrekt. Som vi så i undersøkelsen er det de færreste som kontrollerer ansattdata mot myndighetenes registre.

Jeg utførte selv en undersøkelse på respons ved endring av navn i det norske folkeregisteret. Fra innmeldelsen ble foretatt til bekreftelsen forelå tok det en uke.

7.8 Hva manglet ved undersøkelsen?

Det er noe vanskelig å finne 'rett' svar i en undersøkelse som omhandler et område hvor begrepet kvalitet behandles. Det er så preget av subjektive oppfatninger. En erfaring er å presisere begrepsapparatet mest mulig når man skal utarbeide slike undersøkelser. Enkelte av spørsmålene kunne vært noe mer konkretisert eller utfylt av flere spørsmål for å bedre påliteligheten. Men det er også en avveining mellom kompleksitet i spørsmål og tid som er til disposisjon, både for den som utformer og den som skal svare på dette. I denne oppgave ble kun en samvirkemodell utarbeidet. Om man har anledning og spørreundersøkelse var det sentrale, kunne man utviklet en kausalmodell med tanke på å finne effekter av skjulte variable og fortsatt med en mer inngående statistisk analyse.

Eksempler på spørsmål som kunne vært endret er 'Q20-We have a policy that tells us how to report errors in employee data'. Hvordan definerer svarkandidaten begrepet 'Policy'? Hvordan er den dokumentert? Begrepet kommer igjen i Q7 hvor vi spør etter 'Policy for data ownership'. Dette skulle vært konkretisert bedre.

Q5 -'Our corporation has well known guidelines for data quality in general' ble *for* generelt. Helt klart at begrepet 'well known' skulle vært konkretisert i større grad. Vi ville da sett større spredning i svarene for Q5.

Q14 -'We use prepared and documented methods when measure data quality', er eksempel på en god formulering.

Av intervjuene fremkommer at flere ikke helt tror på målinger. Dette skulle vært fulgt opp med spørsmål om hvilken måle prosess man egentlig har forsøkt og hvorfor man mente at dette ikke var bra. Det kan jo ha vært målemetoder som er forsøkt som denne undersøkelsen altså ikke har avdekket.

I etterkant ser vi at det skulle vært undersøkt mer omkring bruk av dokumentasjon og metadata samt rapporteringsrutiner. Eksempelvis benytter kun 50% offentlige registre på en eller annen måte for å kontrollere ansattdata men hvilke rapporteringsrutiner finnes for den ansatte når det gjelder endringer?

8 Metrikk konkretisering

Av spørreundersøkelsen har vi sett at det ikke er noen av svarkandidatene som måler datakvalitet på en dokumentert, metodisk måte. Skal vi kunne måle datakvaliteten på ansattdata må vi ha tilgjengelig metrikker som dekker de områder som vi har sett er blitt karakterisert som viktige i undersøkelsen. Vi skal her komme frem til et sett med forslag til funksjonelle metrikker. Vi konkretiserer metrikkene i henhold til den mal for metrikker som vi har laget på generell basis (ref. tabell 1).
Metrikker er hovedsakelig relatert til hva vi fant som mest vektet i undersøkelsen.

- Data er lagret og blir brukt på en sikker måte (trygg bruk).
- Data er nøyaktige i forhold til kilden (Kildenøyaktighet).
- Er tilgjengelig når man trenger dem.
- Data må være tidsriktige (oppdaterte).

At data blir benyttet og oppbevart på en sikker måte, er ikke noe spesielt tema i denne sammenheng og gjelder mer eller mindre alle data som behandles i en bedrift. Det spesielle her er at det dreier seg om personopplysninger og at *kvaliteten* blant annet er spesielt viktig for enkelte grunnleggende sikkerhetsfunksjoner som for eksempel *korrekt autorisering*. Slik sett er sikker bruk og oppbevaring, et sentralt tema i HR sammenheng. Kvalitet kan blant annet måles i forhold til hvilke lover og regler som regulerer forhold i henhold til personopplysning [6]. SOX [19] stiller krav til kontrollerte rutiner for aksesskontroll.
Nøyaktighet (Accuracy) kan karakteriseres på to måter[28]: form og innhold. Form er viktig fordi det forteller brukeren om hvordan innholdet skal benyttes/tolkes. Et eksempel er bruk av dato felter: 5/3/2004, - er det 5. mars eller 3. mai i 2004? Innholdet er ikke nøyaktig dersom brukeren ikke kan si hva det er. Metadata forteller hva det er, så dette er viktig å ha med.
Nøyaktighet er et noe diffust begrep. Dette forbindes ofte med presisjon, men presisjon er et mål på reproduserbarhet og er således forskjellig fra nøyaktighet[7].

Skepsis eksisterer til bruk av metrikker.

Potensielle svakheter og kritikk i bruk av metrikker er fremsatt av flere som at det er uetisk, misoppfatning, fordreining og unøyaktighet. Vi ser det også i intervju i forbindelse med denne oppgave at noen er skeptiske, men ikke motvillige, til metrikker og måling. Noen er også tvilende til hvilken effekt disse kan ha i det praktiske liv.

8.1 Metrikk for kompletthet i organisasjon

Se tabell 3.
En av de viktigste faktorer for kompletthet i HR data er at en ansatt virkelig tilhører et gyldig organisasjonselement og at dette forholdet er i tiden (ikke utløpt). Denne metrikken utvider de tradisjonelle deklarative referanse integritet som vi vil se i en organisasjonsstruktur ved at den tar i betraktning andre forretningsmessige forhold som eksempelvis den tilsetningsperiode som en ansatt er i. Dette betyr at vi kan ha et teknisk lovlig forhold mellom ansatt og organisasjon men et forretningsforhold som ikke er gyldig.
Ved denne metrikken kan målesettet inneholde en eller flere typer av ansatt og hvilke av disse typene som er gyldige å behandle må defineres for hvert enkelt bedriftsmiljø.
Metrikken eksponerer et potensielt trusselbilde.

Kvalitetselement	Ansatt organisasjonstilhørighet
Tekstlig utdypning	Sikre kompletthet mellom ansatt og organisasjonselement
Metrikk	IP kompletthet skal være iht bestemte kriterier.
Formål	Alle ansatte skal tilhøre et gyldig organisasjonselement.
Krav	Godkjent når ansatt tilhører et organisasjonselement og dagens dato ligger innenfor tilsetningsperioden.
Frekvens	Transaksjonsbasert (måles ved hver endring) eller database orientert 1 gang pr kvartal (frekventert måling for hele datasettet).
Skala	Antall %
Formel	Antall godkjente i målesettet / Antall ansatte totalt i målesettet * 100
Datakilde	Måling i HR database.
Indikatorer	OK dersom 100%
Godhet og eierskap ved metrikker	
Pålitelighet	Gjentatte målinger vil gi samme resultat med samme datagrunnlag. Datagrunnlaget endres ved endringer av ansattdata. Endringsfrekvensen i disse data regnes som liten bortsett fra større omorganiseringer (må regnes som unntak).
Gyldighet	Metrikken gir svar på antall ansatte som ikke har organisatorisk tilhørighet i en gyldig periode. Ansatte må defineres så samme måte og benyttes konsist.
Gjennomførbarhet	Enkel, teknisk og personellmessig.
Konfliktområder	Metrikken betraktes å ikke være kontroversiell.
Eierskap	HR
Kostnader og anvendelser	
Kostnader	Måleprosess må designes og implementeres. Betraktes som enkel.
Anvendelser	Metrikken måler kompletthet ved et sentralt område for HR data - ansattes relasjon til en lovlig organisasjon og den ansattes tilsetningsperiode.
IP relasjoner	Aksesskontroll
Målepunkt referanse	M2
Trusseleksponering	Belyser andel registrerte ansatte som ikke eksisterer i kontrollerte omgivelser (tilhører et faktisk organisasjonselement med en leder).

Tabell 3: Metrikkbeskrivelse av kompletthet i organisasjon

8.2 Metrikk for trygg behandling av ansattdata

Se tabell: 4. Overordnet metrikk som måler på et akkumulert nivå. De 'akkumulerte' krav kan igjen måles via nye metrikker men er i denne HR sammenheng ikke interessante da de ikke vil være spesielle HR metrikker. Sikker bruk av HR data inkluderer flere forhold som alle må ivaretas om konklusjonen skal være positiv. Metadata er viktig å dokumentere. Nå er det ingen fasit eller oppfattet standard på hva som er metadata i et HR miljø så her er det tatt med et antatt utvalg. Viktige egenskaper i sammenheng for sikker bruk av IP inkluderer:
Tilgang er kontrollert og dokumentert. Ingen uautorisert person eller system skal ha tilgang til HR data. Ansattdata benyttes til klart definerte formål. Forretningsregler er dokumentert. Bruk av data, kontroller, hva er gyldig og lovlig. Godkjent SLA eksisterer og inkluderer kvalitet og sikker lagring (ref. [34]).

Kvalitetselement	Trygg behandling
Tekstlig utdypning	Lover og regler sier at data behandles på en kontrollert og sikker måte.
Metrikk	Ansattdata skal benyttes og lagres på en trygg og forsvarlig måte.
Formål	Følge opp og sikre at data benyttes på en trygg måte.
Krav	Svar JA, NEI eller IA (ikke aktuelt): Har HR data en eier? Er distribusjonen (spredning) av HR data dokumentert? Er tilgang (hvilke personer og systemer) til HR data dokumentert? Benyttes HR data kun til klart definerte formål? Er formålene dokumentert? Er bedriftsinterne retningslinjer for informasjonssikkerhet tatt i bruk for å sikre HR data? Er det gjennomført kundetilfredshet-undersøkelse for opplevd kvalitet for HR data? Er forretningsreglene som regulerer bruk og kontroll av HR data dokumentert? Er det utarbeidet SLA (tjenesteavtale) mellom dataeier og driftsoperatør for HR data? Er det iverksatt dokumenterte rutiner for tilbakerapportering av brudd på datakvalitet?
Frekvens	Målingene foretas en gang pr år. (Eller etter større endringer.)
Skala	Antall svar i krav som er positive.
Formel	(Antall JA svar) / (Totalt antall krav - Antall IA svar) * 100
Datakilde	Kartlegginger ved hjelp av spørreundersøkelse.
Indikatorer	OK dersom 80%
Godhet og eierskap ved metrikker	
Pålitelighet	Målingene kan til en viss grad oppfattes subjektivt.
Gyldighet	Kravformulering kan være en medvirkende årsak til avvik ved gyldighet. Reduseres ved høyere presisjonsnivå.
Gjennomførbarhet	Teknisk gjennomførbarhet bør være enkelt.
Konfliktområder	Tolking av krav. Tolkninger unngås ved mest mulig presise kravformuleringer.
Eierskap	HR ledelse
Kostnader og anvendelser	
Kostnader	Ikke beregnet.
Anvendelser	Trendanalyser og en konkretisering av hvilket kvalitetsnivå som eksisterer for trygg behandling. Krav kan vektes. Vekting må i så tilfelle inkluderes i formel.
IP relasjoner	Metrikken kan også benyttes pr IP
Målepunkt referanse	Overordnet måling. Ingen direkte referanse.
Trusseleksponering	Ikke definert

Tabell 4: Metrikkbeskrivelse av sikker bruk

8.3 Metrikk for kompletthet i navn

Se tabell 5.
Alle felt som inngår i et IP med nyttighetsgrad = 1, skal være utfylt (må felter). I en metrikk for ansattdata tar vi med noen regler for navngivning i Norge[37] som kan implementeres i en grunnleggende metrikk. Ved måling kan vi se om de grunnleggende regler for navngivning er til stede ved å se etter fornavn, etternavn, norsk tegnsett, ikke likhet mellom for og etternavn og ikke bruk av de nevnte skilletegn.

Kvalitetselement	Kompletthet
Tekstlig utdypning	Måling av navnkompletthet iht navnelov.
Metrikk	Kompletthet ved personnavn (fornavn, mellomnavn, etternavn) skal være i henhold til krav.
Formål	Gi måltall på kompletthet ved navn i HR IP.
Krav	• Alle skal ha minst et fornavn. • Alle skal ha ett etternavn. • Navn skal være uttrykt med bokstavene i det norske alfabetet. • Talltegn kan ikke brukes i navn. • Skilletegn som komma, punktum og skråstrek kan ikke inngå i et navn. • Fornavn skal ikke være likt etternavn eller mellomnavn
Frekvens	Målingene foretas fortløpende.
Skala	Antall feil som avdekkes.
Formel	Antall logiske feil funnet i navn
Datakilde	Måling ved inngang i HR database.
Indikatorer	0 feil: OK, 1 eller flere feil: OBS
Godhet og eierskap ved metrikker	
Pålitelighet	Målemetoden kan repeteres. Samme regler benyttes ved hver måling.
Gyldighet	Gjelder kun måling av norske personnavn. Unntak kan forekomme da en kan søke om unntak.
Gjennomførbarhet	Enkel
Konfliktområder	Ikke funnet.
Eierskap	HR
Kostnader og anvendelser	
Kostnader	Marginale ytelsekostnader. Endringskostnader vurderes som små da det benyttes relativt stabile regler.
Anvendelser	En generell kontroll på logisk rett navn i databasen. Det kan kjøres en distinkt kontroll av hvert enkelt navn og da vil metrikken gi antall logiske feil i navn. Alternativt kan metrikken kjøres som et databaseorientert søk. Formel vi da være antall feil navn dividert på totalt antall navn søkt * 100. Resultatet vil da være i prosent.
IP relasjoner	Ikke relevant.
Målepunkt referanse	M1,M2
Trusseleksponering	Ikke definert

Tabell 5: Metrikkbeskrivelse av kompletthet i navn

8.4 Metrikk for kompletthet i HR IP

Se tabell 6. Benytter nyttegrad modellen som definert i figur 14 som hjelp for denne måling. Informasjonsproduktet for HR data må være komplett. Alle felt som inngår med nyttighetsgrad = 1, skal være utfylt (må felter).
På forhånd er data om de enkelte HR IP utfylt i modell for IP nyttegrad. Vi sammenligner felter i HR IP og tilsvarende verdier på nyttegradselementet (nge).

Kvalitetselement	Kompletthet
Tekstlig utdypning	Sikre kompletthet i det HR IP som behandles
Metrikk	IP kompletthet skal være iht bestemte kriterier som er dokumentert i en IP nyttegradsmodell (figur E.2).
Formål	Gi måltall på kompletthet ved HR IP.
Krav	• Kompletthetkriterier er definert og tilgjengelig (*Definer basis HR data*) • Metadata for IP er definert og dokumentert
Frekvens	Målingene foretas fortløpende.
Skala	Antall HR IP som tilfredsstiller krav.
Formel	Antall godkjente IP / Totalt antall IP * 100
Datakilde	Måling i HR database.
Indikatorer	OK dersom 80% riktig og ingen ENERE er feil
Godhet og eierskap ved metrikker	
Pålitelighet	Gjentatte målinger med samme input gir samme svar.
Gyldighet	Relatert til hvor gode oppslagsdata vi har i IP nyttegrad modell.
Gjennomførbarhet	Krever at IP nyttegradsmodell er etablert og populert. Ved grader av nyttighet for attributter er en avhengig av gode algoritmer for fuzzy målinger.
Konfliktområder	Enighet om verdiene som ligger i IP nyttegrad modellen.
Eierskap	
Kostnader og anvendelser	
Kostnader	
Anvendelser	Kompletthet i IP kan integreres i databasesystemet med NULL verdi angivelse. Her benyttes en fuzzy logikk hvor gradering kan være flytende. Alle elementer med gradering < 1 trenger ikke være helt korrekte. Elementer med gradering 1 må være tilstede og må stemme (kontrolleres via oppslag eller beregninger). Eksempel: Bruk av Soundex på navn, innenfor et tidsrom hvor 1= presis dato og 0,5 kan bety et periodisk avvik.
IP relasjoner	Alle IP
Målepunkt referanse	M3
Trusseleksponering	Ikke definert

Tabell 6: Metrikkbeskrivelse av kompletthet i HR IP

8.5 Metrikk for kildenøyaktighet

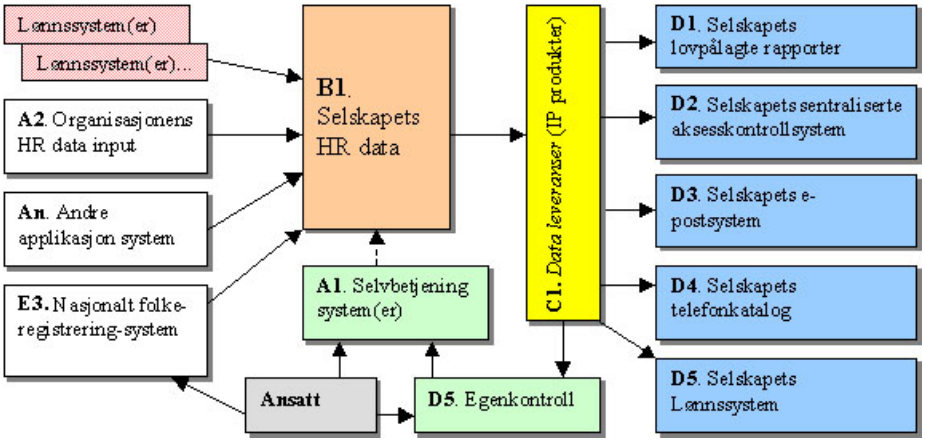
Se tabell: 7. Hva er det som er spesielt i forbindelse med HR data og kildenøyaktighet? Undersøkelsene viser bruk av flere kilder og behov for identifisering av disse og når registreringen har funnet sted. Forslag til nødvendige ekstra IP felt: Dataregistrerer (so - Source origin), når registrert (STS -source time stamp) og eventuelt en kryptografisk sjekksum (hashverdi) av IP som kan benyttes ved kontroll av kvalitet senere. Dette gir: IP + SO + STS + H(IP+SO+STS). Bruk av digitale signaturer er ikke tema i denne kontekst.

Attributt	Kildenøyaktighet
Kvalitetselement	Sikre kvalitet av en leveranse fra en datakilde.
Tekstlig utdypning	Sikre at vi behandler data som kommer fra kjente kilder.
Metrikk	Alle registreringer uansett kilde skal være identifiserbare og data korrekte ved ankomst HR base.
Formål	Gi måltall på kildeopprinnelse og tidsriktighet. Flere kilder gir større sannsynlighet for feil. Ved flere kilder kan siste registrerte velges. Kildene identifiseres og kan rettes.
Krav	• Alle som registrerer basis HR data er spesielt tiltrodd • Alle kilderegistreringer tidsstemples • Alle kilderegistreringer har en unik identifikasjon • Kilderegistrering skjer i et sikkerhetsgodkjent miljø (IT og fysisk sikkerhet)
Frekvens	Målingene foretas fortløpende eller statistisk.
Skala	Prosent av kravoppnåelse.
Formel	Antall godkjente krav / antall krav * 100
Datakilde	Måling i HR database.
Indikatorer	OK dersom 80% er riktig.
Godhet og eierskap ved metrikker	
Pålitelighet	For høy troverdighet for opprinnelsedata kan tiltrodd 3. part involveres.
Gyldighet	Krav som ikke er målbare i transaksjonsøyeblikket må finnes ved oppslag mot ansatt - organisasjonsmodell (ref 13). Modellen må inneholde nødvendige data.
Gjennomførbarhet	Må etablere regime. Metrikken er ikke ment å beskytte ondsinnede data manipuleringer.
Konfliktområder	Skal benyttes internt for opprinnelse- og tidsidentifisering. Ikke annet.
Eierskap	Eier av HR data
Kostnader og anvendelser	
Kostnader	Design og implementering ved datafangst systemer og ved måleprosess i målepunktet.
Anvendelser	Identifisere kilde og bekrefte rett innhold i IP.
IP relasjoner	Alle IP som leveres fra kildesystem for ansattdata
Målepunkt referanse	M2,M3
Trusseleksponering	Viser andel kilder som kan være upålitelige.

Tabell 7: Metrikkbeskrivelse av Kildenøyaktighet

9 Rammeverk for måling

I figur 7 vises de HR funksjonselementer som danner det aktuelle funksjonelle rammeverket. Vi legger merke til informasjonsproduktenes lokalisering mellom HR data og de forskjellige systemer i selskapet som HR leverer data til. De elementene som er skissert ivaretar et utvalg som input, kontroll og leveranse samt sentraliserte HR data. En ansatt er tatt med for å vise hvilke funksjoner denne selv kan benytte og for å illustrere at data om en ansatt kan komme fra flere kilder. C1 er det generelle grensesnitt for leveranse av forskjellige informasjonsprodukter. Det leveres også til egenkontroll. Lønnssystemene er illustrert ved flere instanser da disse kan finnes i flere av bedriftens forskjellige selskap og således kan være input til en felles HR base i bedriften. I dette bildet skal vi implementere målepunkter for måling av datakvalitet.



Figur 7: Sentralt i omgivelsen er selskapets HR database. Data kommer fra forskjellige kilder som manuelle og maskinelle registreringssystemer, selvbetjeningsystemer og lønnssystemer. Data leveres forskjellige systemer i via informasjonsprodukter.

Informasjonsproduktet sammenstilles av data fra forskjellige kilder og leveres som et produkt til det selskapssystem som behandler eksempelvis aksesskontroll. Viktige forutsetninger ved dataflyt i miljøet:

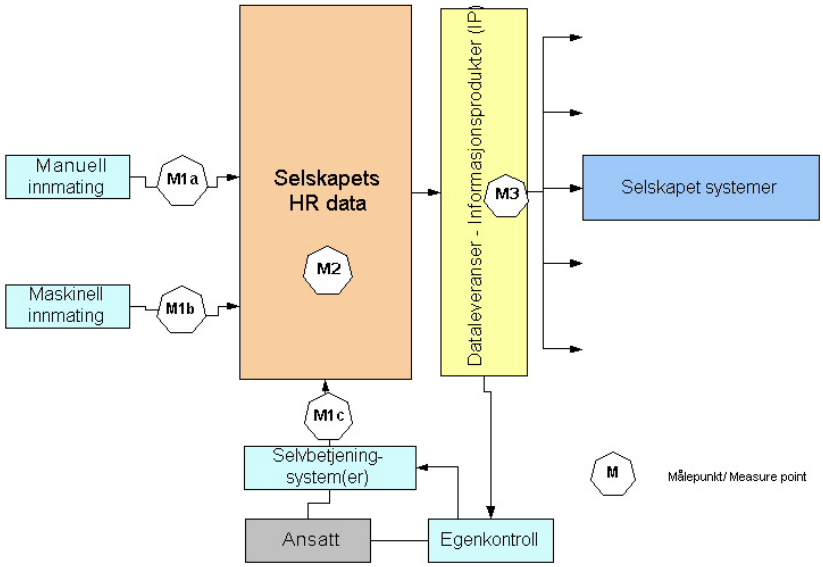
- Data flyter fra et viktigere system til et mindre viktig system
- Data etableres i systemer hvor forvaltningsansvar er tillagt aktuelle data
- Dataflyt følger en trestruktur (ikke sløyfer, gitt av punktet over)
- RC ('Race Condition') utelukkes[1]

9.1 Målepunkter

Vi har sett at granskning av data i databaser kan være databaseorientert eller transaksjons-orientert.

[1] Race conditon er en uønsket situasjon som oppstår når et system forsøker å utføre to eller flere operasjoner samtidig som egentlig er innbyrdes utelukkende og som må utføres i sekvens for å oppnå korrekt resultat.

Databaseorientert : Vurdering av hele datamengden for å måle kvalitet.
Transaksjonsorientert: Måler en enkelt transaksjon før denne medfører endring i databasen. Transaksjonsorientert måling kan etableres ett eller annet sted mellom input i databasesystemet og den endelige lagring (‘commit’ punktet). I dette forsøket vurderes transaksjonsmåling inkludert som trigger i databasesystemet. Vi trenger målepunkter, standardiserte kontroller og rapporter.



Figur 8: Vi kan måle input til databasen (transaksjonsorientert måling), data lagret i databasen (databaseorientert måling) og informasjonsprodukter som er lagret og klar for leveranse.

Ved IP leveranse skal det tas hensyn til de definerte nyttegrader og måles som bestemt (ref: kap. 5.5). Eksempel: alle elementer med nyttegrad 1 blir målt.
Målepunktene detekterer verdier og rapporterer til et rapportsenter. Uten rapportering og etterbehandling er måling mindre meningsfylt. Rapportering defineres som et eget tema utenfor denne oppgavens kontekst. Identifisering av målepunkter: Målepunktene skal bestemmes og benyttes sammen med definerte metrikker som er egnet i denne kontekst.
Det er laget et oppsett for å kunne benytte metrikk ‘Kompletthet i organisasjon’ på en transaksjonsbasert måte (8.1). Vi skal tilføre noe mer enn det som deklarativt kan introduseres i databasesystemet, det skal være enkelt å slå av og på (i tilfelle problemer med ytelse) og det skal være lett anvendbart (altså reflektere momenter Olson[28] mener om triggere).

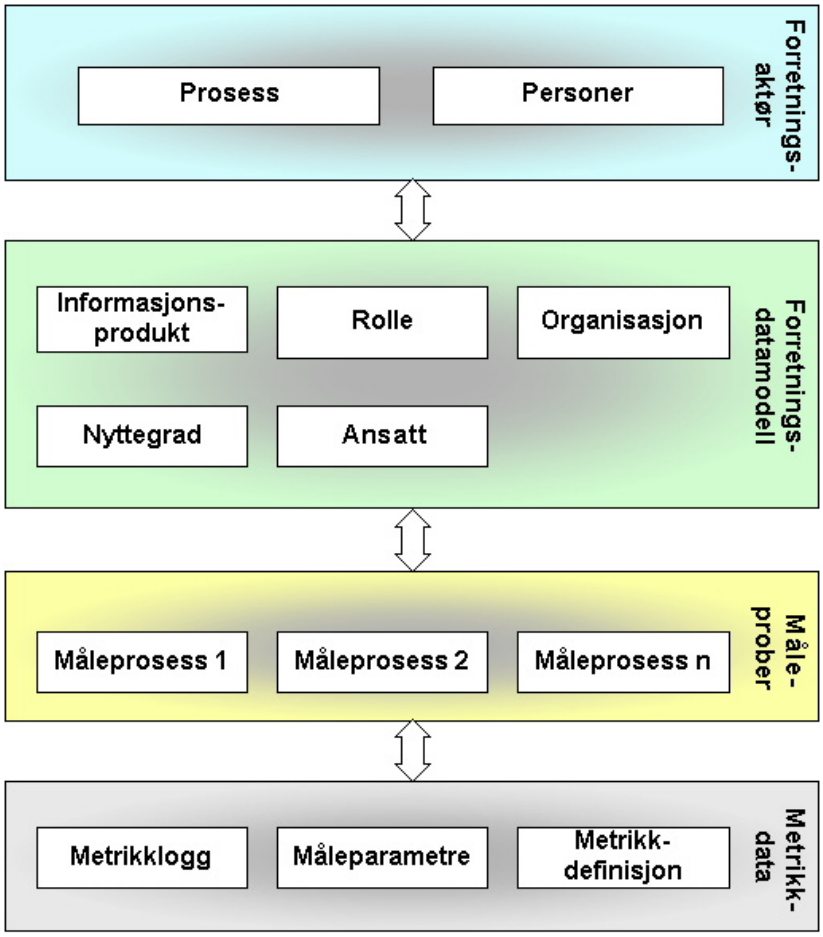
9.2 Overordnet modell for måling

Overordnet modell for måling i et HR miljø er vist i figur 9. I denne lagdeling vises brukere, forretningsdatamodell, måleprober og metrikkdata.

- Brukere er forretningsaktører som it systemer eller personer
- Datalaget viser overordnet datastruktur som er av interesse i denne kontekst og som aksesseres av forretningsaktørene. Dette ligger i forretningsdatamodell med tilhørende beskrivelse av data og regler for bruk av data (metadata).

PunktID	Navn	Hensikt
M1	Mottak	Alle nødvendige felt skal være utfylt, tidsstempel (når data ble lagret) og kildeidentifikasjon skal være lagt til (kan være benyttet hash funksjon). Kildeidentifikasjon kontrolleres.
M2a	Kontroller IP	Kontroll av ansattes organisasjonstilhørighet i henhold til metrikk 8.1 side 32.
M2b	Kontroller IP	Trigger fyres for databasesøk (statistisk kontroll) når eksempelvis 100 endringer er fullført.
M3	Kontroller IP	Kontrollerer leveranse av et gitt IP. Hva og hvordan bestemmes av aktuell metrikk beskrivelse og styres av modell

Tabell 8: Eksempler på målepunkter



Figur 9: Forretningsaktører benytter forretningsdata som er representert ved forretningsdatamodell (FDM). FDM inneholder rene forretningsdata samt informasjon om hva og hvordan informasjonsproduktene er sammensatt. Måleprobene måler kvaliteten på forretningsdata via de metrikker som er spesifisert i metrikkdata. I metrikkdata oppbevares også måleresultater fra hver enkelt måling.

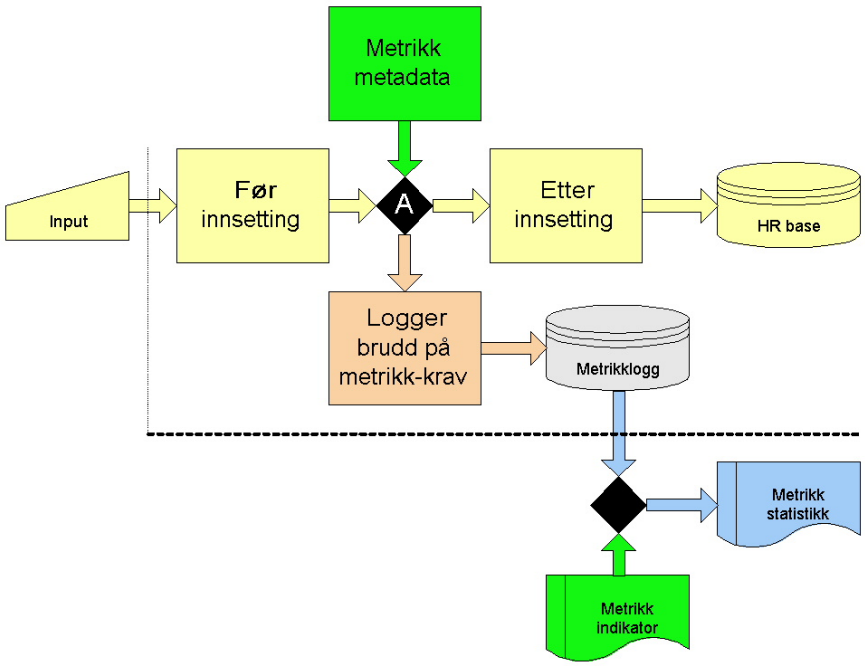
- Måleprober illustrerer bruk av målinger for å samle inn data vi kan benytte i metrikker.
- Metrikkdatalaget inneholder data fra målinger og som benyttes i metrikkene samt målepara-

metre for de enkelte måleprosesser og metrikkenes definisjoner.

Måleparametre inneholder data om måleprosessen er aktivisert, om denne er transaksjonsorientert, satsvis eller statistisk orientert. Inneholder også frekvens på måling dersom denne er statistisk. Disse data er viktige ved en differensiert bruk av en metrikk i forskjellige granskningsmiljø (ref. relatert arbeid, side 5).

10 Implementering av måling

Vi skal prøve ut en målemetode som understøtter metrikk for kompletthet i organisasjon 8.1 og setter opp en enkelt testmiljø. Vi skal samtidig benytte triggere som teknikk. Triggere blir av enkelte betegnet som problemfylt [28]. Vi se om dette stemmer i denne kontekst.
Våre testdata består av 100 insert i tabellen 'Ansatt'. Ca 6% av data har feil som aktiviserer en



Figur 10: Viser miljøet rundt implementering av trigger. En trigger tillater før- og etteranalyse ved endring av data i databasen. Dette er verdifulle egenskaper ved kontroll av forretningsregler. Ved kjøring av statistikk sammenlignes metrikkloggen med de respektive metrikkindikatorer.

eller flere av de kontroller som er relatert til kravene i metrikken. For hvert brudd på de totale krav i metrikken vil det skrives ut en linje i loggen. Loggen kan analyseres ved behov i sann tid eller i ettertid.

10.1 Implementeringsmoduler

10.1.1 Datamodeller

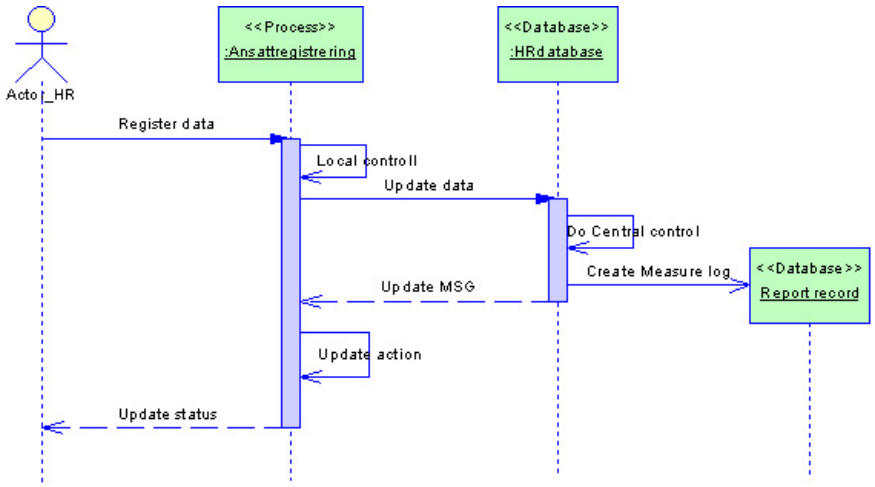
Se modeller i vedlegg E. Følgende modeller er aktuelle:

- Ansatt - organisasjonsforhold (figur 13, side 74)
- Nyttegrad ved informasjonsprodukt (figur 14, side 75)
- Måleparametre (figur 15, side 75)

Måleparametre benyttes til å kontrollere trigger utføring.

10.1.2 Sekvensdiagram for måleprosess

I modellen 11 ser vi et overordnet diagram for dataregistrering hvor forretningsregler kontrolleres i HR basen. Actor HR kan være en person eller et system, Ansattregistrering illustrerer distribuert prosess som mottar input data for registrering i HR base. Data lagres i HRdatabase og måledata lagres i 'MetrikkLogg' for senere analyser og etterbehandling. Meldingen: 'Do Central control' kan være implementert via trigger. Eksempel på utskrift fra logg og grunnlag for metrikk:



Figur 11: I sekvensdiagrammet ser vi et overordnet bilde av dataregistreringen hvor forretningsregler kontrolleres i HR basen.

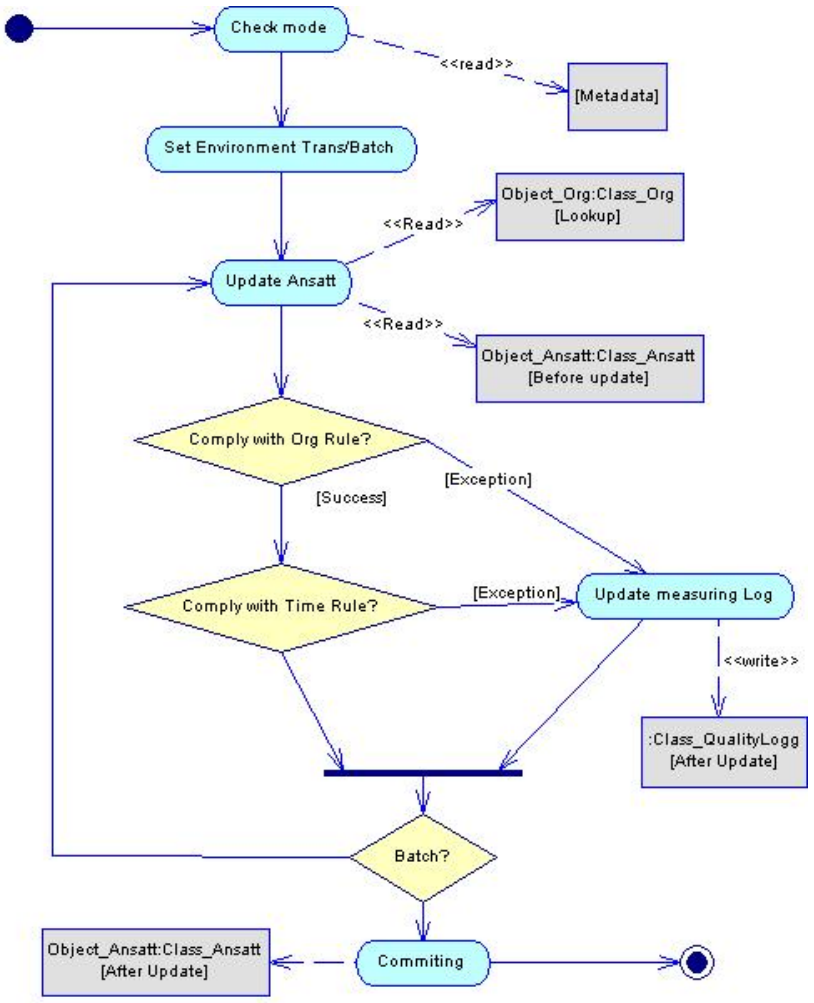
2005-05-24 17:25:46.436	Feil ved 1 , inserted.orgnr: Sluttdato mindre enn dd.
2005-05-24 17:25:46.436	Feil ved 2 , inserted.orgnr: Organisasjon relasjon, orgnr: 7 . Sluttdato mindre enn dd.
2005-05-24 17:25:46.436	Feil ved 3 , inserted.orgnr: Organisasjon relasjon, orgnr: 8 .
2005-05-24 17:25:46.450	Feil ved 9 , inserted.orgnr: Sluttdato mindre enn dd.

Tabell 9: Metrikklogg. For hver måling som krever logging vises tidspunkt, type feil og årsak (hvilket brudd på forretningsregel). Som eksempel viser siste logglinje en feil ved innsetting av ansattnummer 9 ved at sluttdato for ansettelsen er mindre enn dagens dato.

10.1.3 Programlogikk for måleprosess

Nå vil mye av programlogikken være knyttet til den enkelte metrikk. Her relaterer vi arbeidet til metrikken 'Metrikk for kompletthet i organisasjon' 8.1. Aktivitetsdiagrammet i figur 12, viser måling av ansattdata i forhold til organisasjonstilhørighet og tidsrom (ved oppdatering av ansattdata). Ved oppretting av ansattdata kontrolleres den ansatte sin tilhørighet til organisasjon og eventuelt andre relevante data i denne forbindelse. I diagrammet er også skissert hvordan en trigger kan kobles fra transaksjonsorientert til databaseorientert måling ved hjelp av en enkel bryter. Dette foregår via metadata og beslutning i aktiviteten 'Check mode'.
Legg merke til 'Batch' beslutningen nederst i diagrammet. Dette er en generell implementering som tar hensyn til større oppdateringer. I SQL kan man oppdatere mange forekomster via et utsagn og normalt vil da dette utsagnet oppfattes som en hendelse, men det kan være at utsagnet inkluderer 100-vis av endringer. Vi må da ha en teknikk som ivaretar kontrollen for alle disse

endringene.
Eksempel: Vi kan oppdatere alle ansatte med annen organisatorisk tilhørighet. Det som da skjer i dette miljøet er at man høyst sannsynlig sender en kommando til databasen for å utføre dette (et eneste utsagn), i motsetning til å sende en kommando pr ansatt som skal endres. I slike tilfelle må måleprosessen automatisk detektere masseoppdateringer.



Figur 12: Aktivitetsdiagrammet viser måling av ansattdata i forhold til organisasjonstilhørighet og tidsrom. Ved oppretting eller endring av ansattdata kontrolleres forretningsregler som beskrevet i aktuell metrikk. I diagrammet er også skissert hvordan en trigger kan kobles fra transaksjonsorientert til databaseorientert måling ved hjelp av en enkel bryter. Dette foregår via metadata og beslutning i aktiviteten 'Check mode'.

10.2 Konklusjon på responstider

I F.2 vises initielle kostnader ved bruk av trigger ved hjelp av tre målinger.
Måling 1 benyttet trigger fysisk avslått og logisk avslått. Det betyr at dbms ikke aktiviserer trigger i det hele tatt.
Måling 2 benyttet trigger fysisk på men logisk avslått. Dette betyr at dbms fyrer trigger men logikk i trigger (via metaparametre) deaktiviserer triggerens eksekvering.
Måling 3 viser fullt funksjonell trigger. All trigger logikk utføres. Respons her avhenger selv-

følgelig av implementert triggerlogikk.

Konklusjon

Trigger forbruker ca 25% initielle kostnader.

11 Diskusjoner

I denne oppgaven er det gjennomført spørreundersøkelse med etterfølgende intervju, det er utarbeidet metrikker til bruk i HR miljø og det er gjennomført en praktisk test.
I dette arbeidet fremkom en del spørsmål som vi må se litt nærmere på. Dette er samlet under et eget kapittel fordi noen diskusjoner kan omhandle flere tema i oppgaven og da kan det være praktisk å samle dette et sted.

11.1 Manglende måling av datakvalitet

Vi ser av spørreundersøkelsen og i de etterfølgende intervjuer at det er lite målinger av datakvalitet i HR miljø. Det er interessant fordi bedriftene har lange tradisjoner med måling av flere andre forretningsrelaterte områder. Når det gjelder HR området kan det være datamengden som gjør at det ikke er så interessant å måle datakvalitet. Det kan skyldes at man mener å ha oversikt over alle ansatte, ikke har kunnskap om emnet og således ikke reflekterer over dette, forretningene har hatt en stabil struktur inntil nå eller man rett og slett mener det er unødvendig. Ser vi på svarene som er kommet inn vedrørende omorganisering er det åpenbart stor aktivitet på dette området og det er ikke så mange år siden vi følte en betydelig mer stabil organisasjonsstruktur. Det kan tenkes at det er denne tidligere stabilitet som har gjort at det ikke har vært så mye fokus på kvalitetsmåling i denne kontekst men at man nå etter stadig høyere dynamikk i organisasjonsstrukturer og 'turn over' ved ansettelser, både nasjonalt og internasjonalt, vil se at ansattdata er nødt til å gjennomgå samme analyser som man her tradisjon for innen andre forretningsområder (produksjon og kunde).
Under intervjuene fremkommer at datakvaliteten nok ikke er som den bør være, så det er klart at tiltak må iverksettes.

I sammenheng med måling av HR data er det naturlig å stille seg spørsmålet om fornuften ved bruk av metrikker i motsetning til kvalitative studier som kanskje vil være et vel så godt alternativ. I den grad man har kvalitative studier er dette også et godt alternativ men slike studier vil ofte bære preg av mer subjektive oppfatninger, liten grad av formalisering og er mindre egnet til automatiserte prosesser.
Jeg mener det ikke er et enten eller men snarere et både og. Selv om kvantitative målinger kan koste mer å implementere enn kvalitative studier vil det være fornuftig å inkludere objektive målinger og som vil bidra til å øke forståelsen av måling generelt. Innen datakvalitet måling for HR data søker vi her å komme frem til et eksempel på et sett av metrikker.

11.2 Metadata

Det virker litt underlig at dokumentasjon av ansattdata (metadata) ikke blir mer fokusert av kandidatene i spørreundersøkelse og intervju enn det svarene tyder på. Faktisk *lavest* rangert er 'All use of Employee data has to be documented' (ref B.3) - krav om at all bruk av ansattdata må være dokumentert.

Hvorfor denne lave fokuseringen?

Flere forhold kan bidra til dette.
Intervjuobjektene kan ha misoppfattet spørsmålet og ment at 'rubbel og bit' er unødvendig å dokumentere eller at det er faktisk så lite fokus på dokumentasjon av bruk som egne erfaringer kan tyde på, - dokumentasjon er i beste fall et nødvendig onde som blir utført kun når man er nødt.
Det **viktigste verktøy** for å administrere kunnskap om våre data er metadatakataloger. Men det ser ut til at dette ikke helt fungerer i det virkelige liv. Av andre og egne erfaringer kan det være flere årsaker til dette:

- Bedriften benytter metadatakatalog men den er ufullstendig implementert.
- Implementasjon eller konsept blir ikke tatt alvorlig
- Har ikke ledelsens støtte, følges ikke opp
- De blir ikke oppdatert. Katalogene blir passive etter hvert og ingen bruker dem.
- Utformet av arkitekter/designere. Sluttbrukere forstår ikke eller har ikke nytte av innhold
- Programvareindustrien har ikke klart å komme frem til en implementert standard (standard finnes). Resultat: Mange proprietære metakataloger er aktive med liten grad av kooperasjon.
- Designere har ikke tid til å oppdatere metakataloger
- Personer ønsker ikke eller har ikke tid til å legge sin kunnskap i et repository
- Kunnskap er makt.
- Ledelsen blir ikke målt på metadata (*Knowledge Management* er lite fokusert tema)
- Kunnskap om Metadata kan bli mye bedre i organisasjonene.

Under intervjurunden ble det påpekt at det er 'vanskelig å måle noe som ikke er der', og det stemmer. Dette er et av de store utfordringene ved måling og meget viktig i forbindelse med dimensjonen 'nøyaktighet' (Olson [28]).
Her er vi ved et viktig poeng: *Vi kan ikke si hva som er feil når vi ikke vet hva som er rett og det er våre* metadata *som forteller hva som er rett.*
For å kunne gå videre på denne analysen skulle det vært gjort flere og grundigere undersøkelser i brukermiljøene eller det kunne vært stilt spørsmål med litt annerledes vinkling i spørreundersøkelsen.

Forslag til et utvalg metadata.

Her foreslås et utvalg metadata som bør kunne betraktes i et HR miljø (Stikkord er Hvem, Hva, Hvor og Hvorfor):

- Eier av data .
- Hvilke informasjonsprodukt inkluderes i HR data (Hva leveres?)
- Distribusjon av HR data (Hvor og hvorfor).
- Fysisk format (database skjema med godt akseptert kvalitet).
- Kildesystem (opprinnelse og hvordan kilde blir benyttet).
- Rettigheter til data.
- Bruker(grupper) som kan lese, oppdatere, initiere data.

- Kvalitets-kontroller for å fastslå hva som er rett kvalitet.
- Forretningsregler for bruk av data (beskrive data i en livssyklus).

11.3 Flaskehalsanalyse

Ved bruk av trigger til implementering av forretningsregler er det en fordel å utføre en flaskehalsanalyse[1]. En slik analyse vurderer hvor i systemet eventuelle ytelsesproblemer vil kunne oppstå i en travel periode. Når vi lager forslag til strukturkart for datamodell og sekvenskart for programlogikk kan det være vanskelig å fastslå hvor effektivt dette egentlig er fordi det ikke er utført noen god flaskehalsanalyse. Mye er skrevet om analyser i forhold til database utforming og program utforming [38]. Vi vet at det er initielle kostnader ved å ha triggere (vist i kap. 4.1), uansett hva de gjør. Det har vi sett av måleresultat i denne oppgave. Dernest er det et spørsmål om triggerens virkemåte, hvilke oppslagstabeller som benyttes, løkkestrukturer og hvilke egenskaper disse innehar med hensyn til ytelseoptimalisering. Mye kan utføres pr databasesystem (dbms) som eksempelvis optimalisering av cache, indeksering, etc. Mye avhenger av transaksjonsvolum, noe som ansees som lavt i et HR data miljø.
Vi vet ikke ikke noe om strategi for kall til trigger for det enkelte dbms. Oppslagsoptimalisering er et tall fra 1 til 0 hvor 1 tilsvarer ingen optimalisering.
Vi kan betrakte følgende parametre for beregning av en ytelsekostnad:

- Transaksjonsvolum i travel periode.
- Initielle kostnader for hver transaksjon.
- Oppslagskostnad for hvert oppslag som utføres mot andre tabeller i trigger.
- Vekting pga optimalisert oppslagsoptimalisering.
- Antallet oppslagstabeller som benyttes.

[1] Betegnes også trafikkanalyse eller lastestimering

12 Konklusjon og videre arbeid

I denne oppgaven har vi sett på følgende forskningsspørsmål:

1. Hvordan oppfattes og forvaltes datakvalitet innen HR området i dag.
2. Hvordan måle kvalitet for IP som inngår i HR området?
3. Hvilke metrikker er sentrale for å måle kvaliteten i ansattdata?
4. Hvordan kan metrikker benyttes i praksis?

12.1 Konklusjon

System- og informasjonseiere er kanskje de i organisasjonen som kan ta de første og sikreste grep for å få lagt grunnlaget for god datakvalitet samt en sikker forvaltning av dette. Som en naturlig følge av våre forskningsspørsmål, det vi har sett av undersøkelsene og de funn som er gjort, kan vi konkludere med følgende:

Ta kontroll over ansattdata.

Noen må eie ansattdata og det er naturlig at HR eier HR data. Kun halvparten av de spurte i denne undersøkelsen oppgir at de har en klar policy for dataeierskap. HR må ha kontrollen og legge føringer på produksjonsapparatet hvordan kvaliteten skal defineres og spesifiseres for ansattdata.

Mål det som er viktig.

Det viser seg at de aller fleste er for å måle datakvalitet. Forholdet mellom den ansatte og organisasjonen er satt i fokus. Vi har benyttet begrepet informasjonsprodukter (IP) og sett hvordan de viktigste deler av et IP kan måles. Muligheten for å kunne gjennomføre praktiske målinger med minimale kostnader ved implementasjon ved å benytte triggerteknologi er vist. Denne teknologien kan være nyttig i sammenheng med måling av integritet i forretningsregler.

Fokuser datakvalitet.

Datakvalitet må fokuseres i bedriften. Denne oppgaven har vist oss at for mye er underforstått når vi snakker om datakvalitet. Samtidig ser vi at det er høy omorganiseringstakt i bedriftene og det mangler altså dokumenterte forretningsregler. Vi har også sett at omorganisering påvirker datakvalitet. Kjente rutiner og regler bør gjøre datakvaliteten mindre påvirkelig av endringer i organisasjonen.
Kunnskap om hva som er god eller dårlig kvalitet er for en stor del i den intellektuelle del av bedriften og ikke nedtegnet i dokumentasjons form. Kunnskap som er nødvendig for å utføre god forvaltning må dokumenteres.

Målinger og metrikker.

Målinger utføres ikke på en systematisk og dokumentert måte i noen bedrifter. Det betyr ikke at man ikke måler, men de måleprosesser som finnes i bedriften kan være noe mangelfulle. Når noe måles er det stor sannsynlighet for at man ikke måler på samme måte neste gang. Slik sett blir det vanskeligere å sammenligne målinger for å kunne trekke slutninger av trender eller konklusjoner får å iverksette tiltak for kvalitetsforbedring.

Når man ikke måler på en systematisk og dokumentert måte, må vi kunne trekke den slutning at metrikker, eller lignende, ikke er i bruk. Metrikker er en sentral komponent i måling av datakvalitet. Utarbeidelse av metrikker kan i seg selv være motiverende, ideskapende og kunnskapshevende i en organisasjon, så dette er et viktig steg mot høyere bevissthetsnivå.
Et måleoppsett (rammeverk) for måling av datakvalitet i HR data er her foreslått. Dette inkluderer måling av mer kompliserte integritetsregler som er forretningsrelaterte og muligheter for 'hva hvis' analyser.

12.2 Videre arbeid

Mange aktiviteter kan videreføre eller supplere denne oppgaven. Det kan bidra til mer komplett og robust rammeverk for måling av datakvalitet i HR miljø og knytte dette mer direkte mot informasjonssikkerhet.

12.2.1 Metrikker i praksis

I denne oppgaven er det foreslått noen metrikker. Disse bør utprøves i et større miljø slik at man kan erfare i praksis hvordan effekten av disse er. Metrikken for 'kompletthet i organisasjon' er implementert og utprøvd i et testmiljø men bør også testes i et produksjonsmiljø. Når disse utprøves i praksis vil også metrikkens krav bli vurdert på nytt, noe som vil gi verdifull erfaring.

12.2.2 Robusthet ved rammeverket

Hvor robust er det fremlagte rammeverk mot forandringer og angrep?
For å besvare dette må vi definere hva robusthet består av ved et slikt rammeverk. Rammeverket består av et sett betingelser, definisjoner (metadata), metoder, modeller, metrikker og teknikker satt i en kontekst.
Robusthet kan være en funksjon av *konsistens* mellom definisjoner, modeller og metrikker. Robusthet kan også bestå i klarhet i fremstilling, konsistent bruk av begreper og løs kobling til teknologi.
Metrikkenes *pålitelighet og gyldighet* er sentrale områder. Likeså er forandringer i *modell-strukturer*.
Hvor godt skalerer rammeverket? Skalerbarhet skal også betraktes. Hvilke utilsiktede og ondsinnede endringer kan oppstå i et slikt miljø som beskrevet her? Dette er et stort område og kan være innspill til videre arbeid.

12.2.3 Modell for kvalitetskriteria

Datamodell for kvalitetskriteria kan utvides og gjøres mer konsistent eller kunne omfatte forskjellige deler av et informasjonsprodukt eller forskjellige informasjonsprodukt. Er det mulig å lage et rammeverk for pluggbar modell? Tenker da på en basismodell med grensesnitt for å plugge inn nye informasjonsprodukt og nye målekriteria. Motivasjonen for å reise denne problemstillingen er at det reises tvil om å kunne integrere forretningsregler tett inn til databasesystemene pga proprietær og eller til dels komplisert prosedyrespråk.

12.2.4 Rapporteringsrutiner

Kvalitetsavvik oppdages oftest av den som skal benytte produktet. Avvik i forhold til det forventede oppfattes som kvalitetsendring. Negative oppfatninger må kunne rapporteres på en enhetlig og trygg måte. Dette kan oppfattes som rapportering av sikkerhetshendelser i en organisasjon. Det er funnet lite i forbindelse med rutiner for rapportering av datakvalitet i organisasjonen.

12.2.5 Leveranseforventning av ansattdata

Er leveransen av HR data i henhold til de kvalitetskrav kunden hevder? Som vi så i 2.1.1, hevdes det ultimate forskningsprosjekt å være 'forsikringer om leveranser av et dataprodukt faktisk lever opp til de kvalitetskrav kunden hevder'. Flere andre referanser i denne rapporten peker i samme retning.

For å få til dette må forventningene til kunden være kjent og relateres til tilgjengelighet og integritet. Er ikke forventningene kjent blir dette litt vanskelig.

I vår spørreundersøkelse ble området tildels adressert via spørsmål om hvilke kvalitetskrav kunden generelt hevder er de viktigste, men for å komme videre på dette området må det gjøres mer.

I andre sammenheng, som i ved utvikling av informasjonssystemer, er dette forventningsnivået forsøkt adressert ved hjelp av mange forskjellige utviklingsmetoder, med mer eller mindre hell.

Formelle avtaler er en innfallsvinkel. Dette etablerer en kontrakt mellom leverandør og kunde. Alternativt løsere initielle avtaler hvor man benytter metoder som ivaretar hyppig interaksjon mellom leverandør og kunde i utviklingsfasen.

Kan vi ta med oss de erfaringer som ligger i utviklingen av informasjonssystemer og justere det beste til å også ivareta kvalitetsdimensjonen? Se hva som blir gjort av Wang m.fl i [14]. Skulle det være mulig å inkludere et standardisert modelleringsspråk i denne sammenhengen?

Det som er observert så langt er at det ikke er tatt i bruk noen form for de facto standard språk for modellering av datakvalitet. Vi ser at fremstilling av informasjonsprodukter baseres på spesielle løsninger. (2.1). Er UML et potensielt godt modelleringsspråk for datakvalitet? Kan dette språkets konstruksjonselementer benyttes til å beskrive kvalitetsmodeller? Det ville vært topp. Hva vi trenger minst av alt er flere modellspråk i utviklingsmiljøene.

Bibliografi

[1] PricewaterhouseCoopers, “Global data management survey.” http://www.pwc.com/, 2002. (Visited Nov.2004).

[2] ComputerWeekly and P. Brown, “Bad data eats it budgets.” http://www.computerweekly.com, 8 2004. (Visited Nov.2004).

[3] Rådet for IT-sikkerhet, “Digitale signaturer gir tillit til elektronisk kommunikasjon,” 11 1998. (Visited May 2005).

[4] T. Daler, *Håndbok i datasikkerhet : informasjonsteknologi og risikostyring*. Tapir, 82-519-1785-9 (ib.) ed., 2002.

[5] J. Bing, “Lange tak kan gi kultur for it-sikkerhet.” http://digi.no/php/art.php?id=101711, 3 2004. Microsoft seminar 22.3.2004 (web site Visited jan. 2005).

[6] Justisdepartementet, “Lov om behandling av personopplysninger (personopplysningsloven)..” Lovdata - I 2000 hefte 8, LOV-2000-04-14-31.

[7] Y. Huh, F. Keller, T. Redman, and A. Watkins, “Data quality,” *Information and Software Technology*, vol. 32, no. 8, pp. 559–565, 1990.

[8] C. Carson, “What is data quality.” the 9th Meeting of the Heads of National Statistical Offices of East Asian Countries, 5 2000.

[9] R. Y. Wang, M. Ziad, and Y. W. Lee, *Data Quality*. New York Kluwer Academic Publishers, 2002.

[10] R. Wang, V. Storey, and C. Firth, “A framework for analysis of data quality research,” *Knowledge and Data Engineering, IEEE Transactions*, vol. Volume: 7, 8 1995.

[11] D. Loshin, *Enterprise Knowledge Management: The Data Quality Approach*, vol. 1. Morgan Kaufmann, 2001.

[12] R. Y. Wang, M. P. Reddy, and H. B. Kon, “Toward quality data: An attribute-based approach,” *Decision Support Systems*, vol. 13, no. 3-4, pp. 349–372, 1995.

[13] B. Maxwell, “American management association’s hr focus,” *Personnel*, vol. 66, no. 4, pp. 48–58, 1989.

[14] G. Shankaranarayan, M. Ziad, and R. Y. W. Email, “Managing data quality in dynamic decision environments: An information product approach.” http://web.mit.edu/tdqm/www/publications.shtml, 2003. (Visited Dec.2004).

[15] P. Missier and C. Batini, “A multidimensional model for information quality.” citeseer.ist.psu.edu/705235.html. (Visited Nov.2004).

[16] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *COMMUNICATIONS OF THE ACM*, vol. 45, pp. 211–218, 4 2002.

[17] N. C. Henri Theuwissen, "The horror of bad data quality," *SUIG 28*, vol. Paper 162, pp. 1–7, 2003. Advanced Tutorials.

[18] A. Parssian, S. Sarkar, and V. S. Jacob, "Assessing data quality for information products," in *Proceeding of the 20th international conference on Information Systems*, pp. 428–433, Association for Information Systems, 1999.

[19] Sarbanes-Oxley, "Sarbanes-oxley act of 2002." http://www.sarbanes-oxley.com/index.php, 2002. (Visited Jan. 2005).

[20] Department of Commerce, "Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of disseminated information." Federal Register - http://www.osec.doc.gov/cio/oipr/iqg.html, 2 2002. (Visited Jan. 2005).

[21] The Center for Regulatory Effectiveness, "Dataqualityact.us." http://thecre.com/quality/index.html, 2004. (Visited Dec. 2004).

[22] Øyvind W. Remme, "Sjekkliste for datakvalitet i informasjonssystemer." Den Norske Dataforeningen, 04 2004.

[23] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso, and P. Angeletti, "Improving data quality in practice: A case study in the italian public administration," *Distributed and Parallel Databases*, vol. 13, no. 2, pp. 135–160, 2003.

[24] A. Umar, G. Karabatis, L. Ness, B. Horowitz, and A. Elmagardmid, "Enterprise data quality: A pragmatic approach," *Information Systems Frontiers*, vol. 1, no. 3, pp. 279–301, 1999.

[25] Hugin ASA, "Hugin online." http://www.huginonline.no/, 2004. (Visited Oct. 2004).

[26] G. Haraldsne, *Spørreskjemametodikk*. ad Notam, Gyldendal, 1999.

[27] I. M. Holme and B. K. Solvang, *Metodevalg og Metodebruk*. TANO, 2003.

[28] J. E. Olson, *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 12 2002.

[29] C. J. Date, *An Introduction to Database Systems, volume II*. Addison Wesley Publishing Company, 1982.

[30] J. Dorn and L. Rivero, *Database Integrity: Chalange and Solutions*. Idea Group Publishing, 2002.

[31] Y. L. Rita Kovac and L. Pipino, "Total data quality management: The case of iri." The MIT Total Data Quality Management Program, 10 1997.

[32] S. C. Payne, "A guide to security metrics." SANS Security Essentials GSEC Practical Assignment, 7 2001.

[33] Massachusetts Institute of Technology - MIT, "Total quality management (tqm) programs," 2004.

[34] Orderløkken, Bakås, and Hagen, “Sikkerhetsmetrikker for outsourcing av driftstjenester.” Prosjektoppgave i sikkerhetsmetrikker, Nislab/Hig, 12 2003.

[35] S. Lujan-Mora and M. Palomar, “Reducing inconsistency in integrating data from different sources.” Database Engineering & Applications, 2001 International Symposium on, 2001. Pages 209-218.

[36] Arbeids- og sosialdepartementet, “Arbeidsmiljøloven.” ISBN 82-504-1271-0, 02 1977.

[37] Justisdepartementet, “Lov om personnavn (navneloven).” Lovdata -I 2002 hefte 5, LOV-2002-06-07-19. (Visited may 10. 2005).

[38] Sølvberg and Kung, *Information Systems Engineering*. Springer-Verlag, 1993.

A Spørreundersøkelse

De følgende sidene viser det tekstlige innhold av spørreundersøkelsen, ikke form. Spørsmålene vises som de ble utsendt. Ønskes den fullstendige spørreundersøkelsen med følgeskriv kan dette hentes hos forfatteren. Spørsmålsnummereringen er vist i forbindelse med tabeller på sidene 62 til 64.

1 Your general opinion about data quality
Please express your view concerning data quality in your working environment.

1.1 Data quality and your working environment

Statement (Yes No)
1. I think that the quality of employee data is affected by internationalisation of companies.
2. National rules and regulations make demands on employee data quality.
3. Most of our employee data is exchanged with other companies in other countries (cross frontiers)
4. There is a dedicated team/person in our organisation that manages the improvement of employee data.
Statement (False Partly false Partlytrue True)
5. Our corporation has well known guidelines for data quality in general.
6. When our company undergoes reorganization, data quality is also focused upon.
7. In our company we have a well-defined policy for data ownership.
8. Data quality is a subject that has our (the organisationŠs) attention.

1.2 Employee data and national population registration system
Please tick off. (Yes No)
1. Does population registration system exists in your country?
If it exists:
False Partlyfalse Partlytrue True
2. The quality of employee data (name, address) is strictly related to our countries population registration system.
3. The quality of, or some of, the employee data is verified to our countries population registration system of regular intervals

2 How do you make use of Data Quality Measuring?
There are several methods and levels for measuring data quality and the concept of quality may also be ambiguous. How do you and your organisation interpret the concept?

2.1 Measuring employee data quality
Please tick of.
Statement False Partlyfalse Partlytrue True
1. Measuring data quality is an ongoing process in each system and follows stated procedure.
2. We make some random sampling (ad hoc) when needed.
3. We use prepared and documented methods when measure data quality
4. We do not have any special methods for measuring in this context.
Other (Please specify):

2.2 How we detect error in employee data
Statement False Partlyfalse Partlytrue True
1. At periodic interval and via controlled measures.
2. Error is reported by the person who use the data (end user) in his/hers business process, such as e.g. access control routines, pay systems etc.
3. Methods for measuring data quality are worked out and implemented in processes.
4. Our computer-based processes are designed with the intention to detect data quality.
5. We have a policy that tells us how to report errors in employee data (or data in general)

3 Data Quality dimensions
Data quality dimensions focus, more or less, on the relevant attributes of data. Some may be strictly objectiv (e.g. completeness) and some may be pure subjectiv (e.g. repudiation). These conditions have an important role concerning measuring data quality.

3.1 Data quality relevance for employee data
How do you weigh the following statements?
Please use 1 for least important and 6 for most important
Dimension
Employee data must be accurate in relation to its source.
Employee data must be complete. All relevant fields have to be supplied with values.
Employee data must be timely (changes are registered, or edited as soon as known)
Employee data field values must not interfere with each other. E.g. postal code is connected to address as first name is connected to surname.
All use of Employee data has to be documented.
It is important to document the layout of Employee data.
It is important to document the business rules that involve employee data
Employee data is available when you need it.
Employee data is presented in the same format as in different context (e.g. name has same format/layout in the payroll system as in the phone directory)
Employee data has to be highly regarded related to its source and content (data must have a good reputation).
Employee data is certificated at source (you know the exact source of origin)
Ensure that Employee data is used and stored safely.
Employee data must be non ambiguous.

4 Source of employee data
Where is your employee data created initialy the first time?
There may be more or less sources where data are created or edited (changed).

Agreement
Statement False Partlyfalse Partlytrue True
We have only one source of employee data dedicated to our company
We have only one source of employee data in the total corporation
Several employee data sources demand more data quality control
Several sources necessarily do not cause any particular problem.

5 About you and your organisation
Please give some information about your organisation and your role.

5.1 Your organisation
Variable Value
1. What country do you work in?
2. Number of employees in your company (approximately)
3. How much reorganization has your department undergone during the last 3 years?
4. How many times have you changed manager (team leader/superior) last 3 years?
5. Did your work routines change at the last reorganization? Yes No

5.2 Your key position
Please tick off the most valid alternative or specify.
Key position Tick off
ů Personnel responsibility (Personnel manager / human resources manager)
ů Non personnel responsibility Ű mostly technical oriented
ů Non personnel responsibility Ű mostly business process oriented
Other (Please specify):

5.3 Enabling more information
It would be of very great help if you would be kind enough to help me by enabling a possible interview to go into details.
The interview will take no longer than fifteen to twenty minutes.
Yes, my organisation is willing to take part in an interview with more detailed questions about measuring data quality.
Name of organisation: My name:
Phone number: E-mail:

Yes, I would like to receive a copy of the master thesis.
Name:
E-mail:

B Spørreundersøkelse - oppsummering

Spørreundersøkelsen ble distribuert på engelsk og blir gjengitt her som opprinnelig fremstilt bortsett fra tilleggsnummerering (Q..)
Som tidligere nevnt ble det utsendt 25 spørreundersøkelser. Det ble mottatt 19 svar fra 8 forskjellige firma. Noen firma returnerte svar fra flere avdelinger i selskapet, da de ble oppfordret til det. 7 av svarene er levert fra utenlandske selskap hvor hovedselskap tilhører i Norge. 4 selskap ønsket ikke å delta med begrunnelse i tidspress eller de ikke kunne pga prinsipper mot å utlevere sine oppfatninger av emnet som ble forespurt. 2 selskap har ikke svart.
Med 19 svar og 5 dybdeintervju ansees det som tilstrekkelig for å få et fundament for å kunne danne et bilde av hvordan temaet måling av datakvalitet blir oppfattet i det virkelige liv.
Spørreundersøkelsen ble etterfulgt av strukturerte dybde-intervjuer. Disse er oppsummert i 7.6.
Alle svar presenteres som summerte antall per svarkolonne. NA betyr ‘Not Answered’, ikke besvart.

B.1 Your general opinion about data quality (C1)

'Query - Answer'	'Yes'	'No'	NA
Q1-I think that the quality of employee data is affected by internationalisation of companies.	13	6	0
Q2-National rules and regulations make demands on employee data quality.	14	4	1
Q3-Most of our employee data is exchanged with other companies in other countries (cross frontiers)	3	16	0
Q4-There is a dedicated team/person in our organisation that manages the improvement of employee data.	12	7	0

Tabell 10: General opinion about data quality

'Query - Importance:'	FALSE	PF	PT	TRUE	NA
Q5-Our corporation has well known guidelines for data quality in general.	1	1	15	2	0
Q6-When our company undergoes reorganization, data quality is also focused upon.	1	4	7	7	0
Q7-In our company we have a well-defined policy for data ownership.	0	6	4	8	1
Q8-Data quality is a subject that has our (the organisationś) attention.	0	6	5	8	0
Q10-The quality of employee data (name, address) is strictly related to our countries population registration system	7	1	6	2	3
Q11-The quality of, or some of, the employee data is verified to our countries population registration system of regular intervals	6	4	2	4	3

Tabell 11: General focus about data quality

B.2 How to make use of Data Quality Measuring

'Query - Importance:'	FALSE	PF	PT	TRUE	NA
Q12-Measuring data quality is an ongoing process in each system and follows stated procedure.	1	4	8	5	1
Q13-We make some random sampling (ad hoc) when needed.	1	4	7	6	1
Q14-We use prepared and documented methods when measure data quality	6	5	7	0	1
Q15-We do not have any special methods for measuring in this context	1	6	6	5	1
Q16-At periodic interval and via controlled measures	1	3	6	8	1
Q17-Error is reported by the person who use the data (end user) in his/hers business process.	1	0	4	12	2
Q18-Methods for measuring data quality are worked out and implemented in processes.	3	5	3	7	1
Q19-Our computer-based processes are designed with the intention to detect data quality.	1	2	10	4	2
Q20-We have a policy that tells us how to report errors in employee data (or data in general)	6	3	5	4	1

Tabell 12: The use of Data Quality Measuring

B.3 Data Quality dimensions (C3)

'Query - Weigh the following statements (Importance)'	1	2	3	4	5	6	NA
Q21-Employee data must be accurate in relation to its source.	0	0	0	1	2	16	0
Q22-Employee data must be complete. All relevant fields have to be supplied with values.	0	0	1	5	5	8	0
Q23-Employee data must be timely (changes are registered, or edited as soon as known)	0	0	0	1	5	13	0
Q24-Employee data field values must not interfere with each other.	0	0	2	5	4	8	0
Q25-All use of Employee data has to be documented.	1	2	2	6	3	5	0
Q26-It is important to document the layout of Employee data.	0	2	3	1	6	7	0
Q27-It is important to document the business rules that involve employee data	0	0	1	3	3	12	0
Q28-Employee data is available when you need it.	0	0	1	1	3	14	0
Q29-Employee data is presented in the same format as in different context	0	0	4	6	3	6	0
Q30-Employee data has to be highly regarded related to its source and content.	0	0	0	1	4	13	1
Q31-Employee data is certificated at source (you know the exact source of origin)	0	0	0	4	7	7	1
Q32-Ensure that Employee data is used and stored safely.	0	0	0	0	1	18	0
Q33-Employee data must be non ambiguous.	0	0	1	1	6	10	1

Tabell 13: Data Quality dimensions

B.4 Source of employee data (C4)

'Query - Importance:'	FALSE	PF	PT	TRUE	NA
Q34-We have only one source of employee data dedicated to our company	4	3	5	7	0
Q35-We have only one source of employee data in the total corporation	11	3	4	0	1
Q36-Several employee data sources demand more data quality control	0	0	5	14	0
Q37-Several sources necessarily do not cause any particular problem.	6	7	3	3	0

Tabell 14: Source of employee data

B.5 About you and your organisation (C5)

'Query - Answer'	'max'	'min'	'avg'
Q39-Number of employees in your company (approximately)	12000	250	3745

'Query - Answer'	Nothing	Some	Much	NA
Q40-How much reorganization has your department undergone during the last 3 years?	1	5	12	1

'Query - Answer'	0	1	2	3	>3	NA
Q41-How many times have you changed manager (team leader/superior) last 3 years?	4	3	7	3	1	1

'Query - Answer'	Yes	No	NA
Q42-Did your work routines change at the last reorganization?	11	7	1

'Query - Answer'	Personnel resp	Mostly technical	Mostly business	NA
Q43-Role - Responsibility	11	4	4	0

Tabell 15: About you and your organisation

C Spesielle analyser

Det er laget et sett med spesielle tabeller som illustrerer grupperinger og sammenhenger mellom datagrupper.
To forhold er fokusert: Analyser mellom rolle i organisasjonen og svar som avgis samt gruppering mellom regioner (Norge kontra de øvrige land) i spørreundersøkelsens forskjellige spørsmålsgrupper. Motivasjonen for disse to grupperingene var å finne ut om det er spesielle forhold i besvarelsene som avhenger av roller og land.
Området land kontra kvalitetsdimensjon (ref C.1) er noe detaljert, mest med hensyn på å vise de vanligste mål for statistiske sammenligninger. Med et begrenset datavolum ser man nesten like godt forholdene direkte av matrisen.

C.1 Country vs Dimension

Country	Accurate	Complete	Timely	uninterfere	Documente	Layout	Buisn. rules	Available	Uni For-mat	Regarded	Sourcemarl	Safely	Non Am-big.
Bangladesh	6,00	5,50	6,00	5,00	4,50	5,50	5,50	6,00	5,00	6,00	4,00	6,00	6,00
Denmark	6,00	6,00	6,00	4,00	6,00	4,00	4,00	6,00	4,00	6,00	4,00	6,00	6,00
Hungary	4,00	4,00	6,00	5,00	6,00	5,00	6,00	5,00	5,00	5,00	5,00	6,00	5,00
Malaysia	6,00	6,00	6,00	5,00	6,00	6,00	6,00	6,00	6,00	6,00	5,00	6,00	6,00
Norway	5,83	4,75	5,42	4,92	3,58	4,25	5,25	5,42	4,17	5,55	5,33	5,92	5,09
Sweden	6,00	6,00	6,00	6,00	4,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00
Std avvik	0,81	0,83	0,24	0,63	1,12	0,86	0,78	0,43	0,86	0,41	0,78	0,03	0,49
Varians	0,65	0,69	0,06	0,40	1,25	0,74	0,61	0,18	0,74	0,17	0,61	0,00	0,24
Modus	6,00	6,00	6,00	5,00	6,00	6,00	6,00	6,00	5,00	6,00	4,00	6,00	6,00
Average	5,64	5,38	5,90	4,99	5,01	5,13	5,46	5,74	5,03	5,76	4,89	5,99	5,68

Tabell 16: Country vs Dimension

C.2 Role vs Dimension

'Role'	'Accurate'	'Complete'	'Timely'	'Uninterfere'	'Document	'Layout'	'BusRules'	'Avail'	'UniForma	'Regarded'	'Sourcemrk	'Safely'	'NonAmbig'
HR manag-er	5,70	5,00	5,60	4,90	4,00	4,50	5,50	5,30	4,30	5,50	5,20	6,00	5,10
mostly Tech.	5,70	5,00	5,20	5,20	3,70	5,00	5,20	5,70	4,20	5,60	5,70	5,70	5,60
mostly Busin.	6,00	5,20	6,00	4,70	5,20	4,70	5,00	6,00	5,50	6,00	4,50	6,00	5,70
Std avvik	0,17	0,12	0,40	0,25	0,79	0,25	0,25	0,35	0,72	0,26	0,60	0,17	0,32
Varians	0,03	0,01	0,16	0,06	0,63	0,06	0,06	0,12	0,52	0,07	0,36	0,03	0,10
Average	5,80	5,07	5,60	4,93	4,30	4,73	5,23	5,67	4,67	5,70	5,13	5,90	5,47

Tabell 17: Role vs Dimension

C.3 Query group2, Grouped by Staff and Position

Query - Importance - For query group-2, Grouped by staff position	GROUP MEMBER	Yes	No	NA	Yes	No
Q1-I think that the quality of employee data is affected by internationalisation of companies.	Leader	7	4	0	64%	36%
	officer	6	2	0	75%	25%
Q2-National rules and regulations make demands on employee data quality.	Leader	7	3	1	70%	30%
	officer	7	1	0	88%	13%
Q3-Most of our employee data is exchanged with other companies in other countries (cross frontiers)	Leader	0	11	0	0%	100%
	officer	3	5	0	38%	63%
Q4-There is a dedicated team/person in our organisation that manages the improvement of employee data.	Leader	7	4	0	64%	36%
	officer	5	3	0	63%	38%
Query - Importance - For query group-2, Grouped by Norway-Abroad	NATION	Yes	No	NA	Yes	No
Q1-I think that the quality of employee data is affected by internationalisation of companies.	abroad	5	2	0	71%	29%
	norway	8	4	0	67%	33%
Q2-National rules and regulations make demands on employee data quality.	abroad	6	1	0	86%	14%
	norway	8	3	1	73%	27%
Q3-Most of our employee data is exchanged with other companies in other countries (cross frontiers)	abroad	2	5	0	29%	71%
	norway	1	11	0	8%	92%
Q4-There is a dedicated team/person in our organisation that manages the improvement of employee data.	abroad	5	2	0	71%	29%
	norway	7	5	0	58%	42%

Tabell 18: Query group2, Grouped by Staff and Position

C.4 Query group4, Grouped by Staff Position

Query - Importance - For query group4, Grouped by staff position	GROUP MEMBER	FALSE	TRUE	NA	%False	%True
Q5-Our corporation has well known guidelines for data quality in general.	Leader	1	10	0	9%	91%
	officer	1	7	0	13%	88%
Q6-When our company undergoes reorganization, data quality is also focused upon.	Leader	2	9	0	18%	82%
	officer	3	5	0	38%	63%
Q7-In our company we have a well-defined policy for data ownership.	Leader	2	9	0	18%	82%
	officer	4	3	1	57%	43%
Q8-Data quality is a subject that has our (the organisationŠs) attention.	Leader	4	7	0	36%	64%
	officer	2	6	0	25%	75%
Q10-The quality of employee data is strictly related to our CPRS system	Leader	5	5	1	50%	50%
	officer	3	3	2	50%	50%
Q11-The quality of, or some of, the employee data is verified to our CPRS of regular intervals	Leader	7	3	1	70%	30%
	officer	3	3	2	50%	50%
Q12-Measuring dq is an ongoing process in each system and follows stated procedure.	Leader	3	8	0	27%	73%
	officer	2	5	1	29%	71%
Q13-We make some random sampling (ad hoc) when needed.	Leader	4	7	0	36%	64%
	officer	1	6	1	14%	86%
Q14-We use prepared and documented methods when measure data quality	Leader	6	5	0	55%	45%
	officer	5	2	1	71%	29%
Q15-We do not have any special methods for measuring in this context	Leader	6	5	0	55%	45%
	officer	1	6	1	14%	86%
Q16-At periodic interval and via controlled measures	Leader	2	9	0	18%	82%
	officer	2	5	1	29%	71%
Q17-Error is reported by the person who use the data (end user) in his/hers business process	Leader	1	9	1	10%	90%
	officer	0	7	1	0%	100%
Q18-Methods for measuring data quality are worked out and implemented in processes.	Leader	5	6	0	45%	55%
	officer	3	4	1	43%	57%
Q19-Our computer-based processes are designed with the intention to detect data quality.	Leader	3	7	1	30%	70%
	officer	0	7	1	0%	100%
Q20-We have a policy that tells us how to report errors in employee data (or data in general)	Leader	6	5	0	55%	45%
	officer	3	4	1	43%	57%
Q34-We have only one source of employee data dedicated to our company	Leader	6	5	0	55%	45%
	officer	1	7	0	13%	88%
Q35-We have only one source of employee data in the total corporation	Leader	8	2	1	80%	20%
	officer	6	2	0	75%	25%
Q36-Several employee data sources demand more data quality control	Leader	0	11	0	0%	100%
	officer	0	8	0	0%	100%
Q37-Several sources necessarily do not cause any particular problem.	Leader	9	2	0	82%	18%
	officer	4	4	0	50%	50%

Tabell 19: Query group4, Grouped by Staff Position

C.5 Query group4, Grouped by Region

Query - Importance' - For query group4, Grouped by Norway-abroad	NATION	FALSE	TRUE	NA	%False	%True
Q5-Our corporation has well known guidelines for data quality in general.	abroad	1	6	0	14%	86%
	norway	1	11	0	8%	92%
Q6-When our company undergoes reorganization, data quality is also focused upon.	abroad	0	7	0	0%	100%
	norway	5	7	0	42%	58%
Q7-In our company we have a well-defined policy for data ownership.	abroad	1	6	0	14%	86%
	norway	5	6	1	45%	55%
Q8-Data quality is a subject that has our (the organisationŠs) attention.	abroad	1	6	0	14%	86%
	norway	5	7	0	42%	58%
Q10-The quality of employee data (name, address) is strictly related to our CPRS	abroad	2	4	1	33%	67%
	norway	6	4	2	60%	40%
Q11-The quality of, or some of, the employee data is verified to our CPRS of regular intervals	abroad	3	3	1	50%	50%
	norway	7	3	2	70%	30%
Q12-Measuring data quality is an ongoing process in each system and follows stated procedure.	abroad	1	6	0	14%	86%
	norway	4	7	1	36%	64%
Q13-We make some random sampling (ad hoc) when needed.	abroad	1	6	0	14%	86%
	norway	4	7	1	36%	64%
Q14-We use prepared and documented methods when measure data quality	abroad	3	4	0	43%	57%
	norway	8	3	1	73%	27%
Q15-We do not have any special methods for measuring in this context	abroad	1	6	0	14%	86%
	norway	6	5	1	55%	45%
Q16-At periodic interval and via controlled measures	abroad	0	7	0	0%	100%
	norway	4	7	1	36%	64%
Q17-Error is reported by the person who use the data (end user) in his/hers business process	abroad	1	6	0	14%	86%
	norway	0	10	2	0%	100%
Q18-Methods for measuring data quality are worked out and implemented in processes.	abroad	2	5	0	29%	71%
	norway	6	5	1	55%	45%
Q19-Our computer-based processes are designed with the intention to detect data quality.	abroad	1	5	1	17%	83%
	norway	2	9	1	18%	82%
Q20-We have a policy that tells us how to report errors in employee data	abroad	4	3	0	57%	43%
	norway	5	6	1	45%	55%
Q34-We have only one source of employee data dedicated to our company	abroad	1	6	0	14%	86%
	norway	6	6	0	50%	50%
Q35-We have only one source of employee data in the total corporation	abroad	3	3	1	50%	50%
	norway	11	1	0	92%	8%
Q36-Several employee data sources demand more data quality control	abroad	0	7	0	0%	100%
	norway	0	12	0	0%	100%
Q37-Several sources necessarily do not cause any particular problem.	abroad	3	4	0	43%	57%
	norway	10	2	0	83%	17%

Tabell 20: Query group4, Grouped by Region

D Intervjuguide

1 Rolle og erfaring
2 Hva karakteriserer HR data fra andre data (CRM data, Operasjonelle data, ...)
    Mht. presisjon
    Viktighet
    Oppmerksomhet?
3 Hvilke Aktuelle IP i ditt miljø? (Hvilke system HR data benyttes i)
    Eksempel
        Adgangskontrollsystemer
        Intern Telefonkatalog
        Annet?
    Data i informasjonsprodukt (IP)
        Hvilke data inngår?
        Gradering av viktighet i attributter?
        Kjenner du til forretningsregler (Buisness Rules) for dette IP
            - som stadfester kvalitet i HR data?
            - som forteller hva sluttbruker kan forvente av kvalitetsnivå?
            - som sier hva som er organisatoriske og forretningsmessige krav?
4 Hvor god tror du datakvaliteten er?
    I dine HR data
    I de HR data du behandler
    Eksempler på områder som må/kan forbedres?
5 Er du fornøyd med/fokus på kvaliteten?
    Snakkes det om datakvalitet i miljøet ditt?
    Forbedringstiltak?
    Tror du datakvalitet henger sammen med organisasjonens stabilitet?
6 Har du ansvar for noen HR data?
8 Har dere kjente retningslinjer for datakvalitet generlt?
9 Hvor godt mener du dere fokuserer datakvalitet?
10 Har dere en vel definert policy for dataeierskap?(Q7)
11 Savner du metoder for å kunne måle datakvalitet (Q15)

E Datamodeller

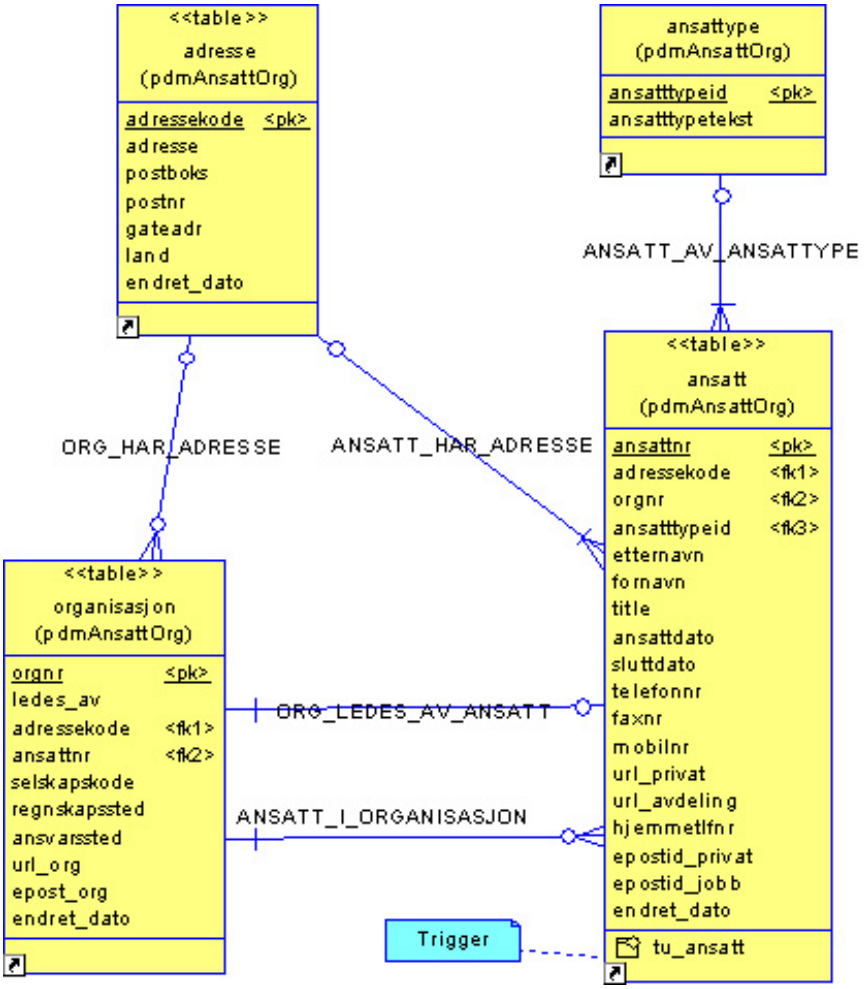
Når vi omtaler modeller i data sammenheng, er ofte disse tre nivå omtalt:

- **Konseptuelt skjema** viser datamodeller slik sluttbrukere oppfatter data. Viser logiske grupperinger og deres sammenhenger. Baseres på forretningskrav, kan inkludere arv og mange til mange forhold. Forretningsregler beskrives her i prosa. Benytter begreper som entitet og attributt. Uavhengig av databasesystem (dbms) og teknologi.
- **Logisk skjema** er mer dbms relatert (domene mapping, løser opp mange til mange forhold) men inkluderer ikke fysiske implementasjoner. Avtegner strukturer slik at både sluttbrukere forstår modellen og at den er enkel å implementere. Forretningsregler kan relateres til triggere og beskrives ved enkle modeller. Benytter termer som tabell og kolonne.
- **Fysisk nivå** er en implementering (instans) av konseptuelt eller logisk skjema i en bestemt dbms. Triggere er skrevet i et dbms spesifikt språk.

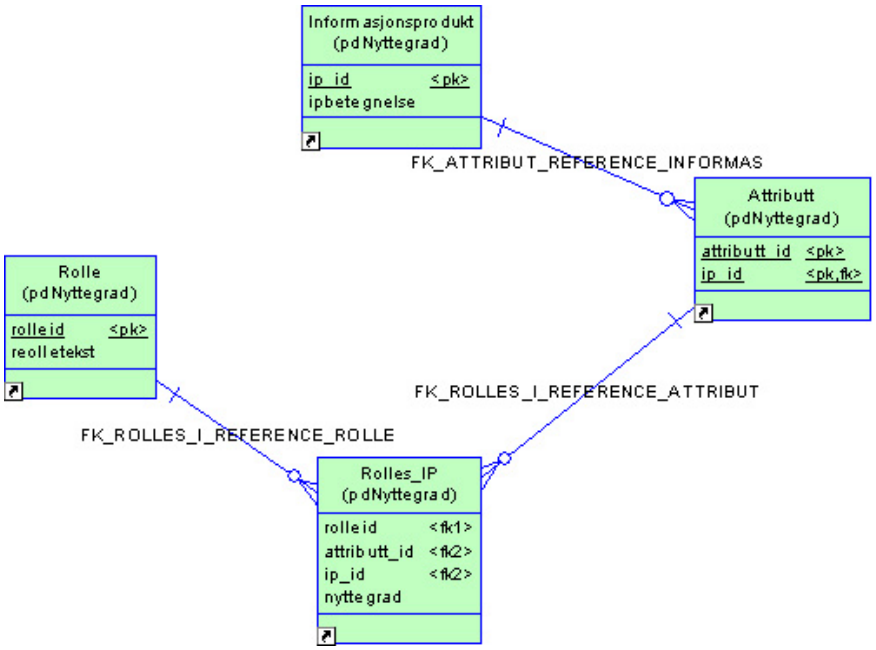
E.1 Ansatt - Organisasjon

E.2 Nyttegrad

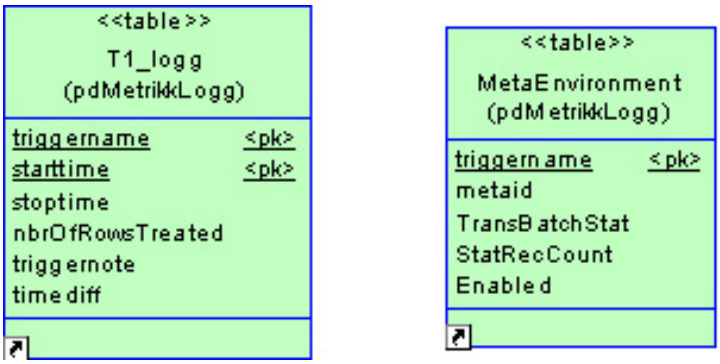
E.3 Logg og Metadata



Figur 13: Modellen implementerer integritet mellom ansatt og organisasjon deklarativt. En integritetsregel som ikke kan implementeres deklarativt er for eksempel tidsdimensjonen ved et ansatt forhold. Tidsdimensjonering ivaretas av trigger.



Figur 14: Modellen viser implementering av nyttegrad ved et informasjonsprodukt. Kan benyttes ved måling av kompletthet i et IP eller utvides til målinger som ivaretar mer fuzzy måling (graderinger av kvalitet). Ved kompletthet måles i leveransepunkt M3 (ref. figur 8 side fig:measurepoints) og alle nyttegrad kolonner i Rolles_IP som har 1-ere, skal være utfylt. Ved fuzzy måling av et element må gradering av elementet være kjent. Man må vite hva graderingselementet representerer. Dette bestemmes i den aktuelle metrikk.



Figur 15: Tabeller for logging av måleresultater (T1_logg) og parameterstyring av triggeraktivitet (MetaEnvironment).

F Triggerresponstabeller

F.1 Måling av trigger respons

Målinger utført ved å sende 100 'insert' kommandoer til databasetabell.
Måling 1 er utført på server. Vi ser da 42% økning i respons når trigger aktiv.
Måling 2 er utført fra klient over nettverk. Pga økt nettverkstrafikk ved hver insert blir trigger respons marginal, kun 4%.

Uten trigger stored procedure (uten klient trafikk)	
Run time: Apr 20 2005 3:09:12:950PM	Elapsed time 93
Run time: Apr 20 2005 3:09:39:340PM	Elapsed time 90
Run time: Apr 20 2005 3:09:52:326PM	Elapsed time 93
Run time: Apr 20 2005 3:10:02:200PM	Elapsed time 110
Run time: Apr 20 2005 3:10:15:216PM	Elapsed time 76
Run time: Apr 20 2005 3:10:27:873PM	Elapsed time 93
Snitt: 92,5(ms)	
Med trigger t1_1, stored procedure (uten klient trafikk)	
Run time: Apr 20 2005 3:11:04:340PM	Elapsed time 170
Run time: Apr 20 2005 3:11:25:936PM	Elapsed time 156
Run time: Apr 20 2005 3:11:40:670PM	Elapsed time 156
Run time: Apr 20 2005 3:11:55:920PM	Elapsed time 156
Run time: Apr 20 2005 3:11:55:920PM	Elapsed time 156
Snitt 158.8(ms)	Differanse: 42%

Tabell 21: Eksekveringstider på server

Med trigger t1_1 og med klient trafikk		Elapsed
Start: Apr 20 2005 2:46:56:076PM	Stop:2:47:02:903PM	6827
Start: Apr 20 2005 2:48:34:293PM	Stop:2:48:40:670PM	6377
Start: Apr 20 2005 2:49:12:013PM	Stop:2:49:18:483PM	6470
Start: Apr 20 2005 2:51:08:153PM	Stop:2:51:14:653PM	6500
Start: Apr 20 2005 2:51:43:640PM	Stop:2:51:50:263PM	6623
Snitt: 6559 (ms)		
Uten trigger med klient trafikk		
Start: Apr 20 2005 2:52:26:233PM	Stop:2:52:32:576PM	6343
Start: Apr 20 2005 2:52:57:750PM	Stop:2:53:03:950PM	6200
Start: Apr 20 2005 2:57:58:966PM	Stop:2:58:05:373PM	6407
Start: Apr 20 2005 2:58:31:826PM	Stop:2:58:38:030PM	6204
Start: Apr 20 2005 2:59:16:250PM	Stop:2:59:22:576PM	6326
Snitt: 6296 (ms)		Differanse: 4%

Tabell 22: Eksekveringstider via nettverk

F.2 Måling av initiell trigger kostnad

I det etterfølgende vises initielle kostnader ved bruk av trigger ved hjelp av tre målinger.

Trigger logikk

Som i aktivitetsdiagram, se fig:12, side 43.
100 poster ble satt inn i ansatt tabell. 6% av diss postene var det logiske feil på. Disse ble skrevet til en logg-tabell (T1_logg).
Testen foregikk på ASE serveren, win2000, lokal prosessering (ikke nettverk involvert). Benyttet ISQL. Database: mais

Konklusjon

Som vi har konkludert med tidligere i rapporten ser vi at en trigger forbruker ca 25% i initielle kostnader.

MÅLING 1 Målingen benyttet trigger fysisk avslått og logisk avslått. Det betyr at dbms ikke aktiviserer trigger i det hele tatt.
update MetaEnvironment set Enabled=N where triggername = org_rule alter table ansatt disable trigger
Run time: May 24 2005 5:18:06:983PM Elapsed time 126
Run time: May 24 2005 5:18:07:090PM Elapsed time 106
Run time: May 24 2005 5:18:07:200PM Elapsed time 110
Run time: May 24 2005 5:18:07:310PM Elapsed time 110
Run time: May 24 2005 5:18:07:420PM Elapsed time 110
Run time: May 24 2005 5:18:07:530PM Elapsed time 110
Run time: May 24 2005 5:18:07:640PM Elapsed time 110
Run time: May 24 2005 5:18:07:763PM Elapsed time 110
Run time: May 24 2005 5:18:07:890PM Elapsed time 126
Run time: May 24 2005 5:18:08:000PM Elapsed time 110
Snitt 112,8 ms. = 100,00% = basismåling
MÅLING 2 Målingen benyttet trigger fysisk på men logisk avslått. Dette betyr at dbms fyrer trigger men logikk i trigger (via metaparametre) deaktiviserer triggerens eksekvering.
update MetaEnvironment set Enabled=N where triggername = org_rule alter table ansatt enable trigger
Run time: May 24 2005 5:20:50:750PM Elapsed time 160
Run time: May 24 2005 5:20:50:903PM Elapsed time 140
Run time: May 24 2005 5:20:51:043PM Elapsed time 140
Run time: May 24 2005 5:20:51:186PM Elapsed time 143
Run time: May 24 2005 5:20:51:326PM Elapsed time 140
Run time: May 24 2005 5:20:51:466PM Elapsed time 140
Run time: May 24 2005 5:20:51:606PM Elapsed time 140
Run time: May 24 2005 5:20:51:750PM Elapsed time 143
Run time: May 24 2005 5:20:51:903PM Elapsed time 140
Run time: May 24 2005 5:20:52:030PM Elapsed time 126
Snitt 141,2 ms => 125,18% relatert til basismåling.
MÅLING 3 Målingen viser fullt funksjonell trigger. All trigger logikk utføres. Respons her avhenger selvfølgelig av implementert triggerlogikk.
update MetaEnvironment set Enabled=Y where triggername = org_rule alter table ansatt enable trigger
Run time: May 24 2005 5:22:50:606PM Elapsed time 156
Run time: May 24 2005 5:22:50:780PM Elapsed time 173
Run time: May 24 2005 5:22:50:966PM Elapsed time 156
Run time: May 24 2005 5:22:51:140PM Elapsed time 173
Run time: May 24 2005 5:22:51:310PM Elapsed time 170
Run time: May 24 2005 5:22:51:466PM Elapsed time 156
Run time: May 24 2005 5:22:51:653PM Elapsed time 170
Run time: May 24 2005 5:22:51:810PM Elapsed time 156
Run time: May 24 2005 5:22:52:000PM Elapsed time 173
Run time: May 24 2005 5:22:52:140PM Elapsed time 140
Snitt 162,3 ms => 143,88% relatert til basismåling

Tabell 23: Eksekveringstider trigger initiell kostnad

G BNF Notasjon benyttet

BNF er en forkortelse for ‘Backus Naur Form’. John Backus og Peter Naur introduserte en formell notasjon for å beskrive syntaks for et gitt programmeringsspråk (også SQL) og datastrukturer.
Denne oppgaven følger en BNF notasjon med disse retningslinjer:

	Forklaring	Eksempel
‘’	Verdi	‘jan’
<>	Variabel	<navn>
::=	Består av	
|	Eller	<navn> ::= ’jan’ | ’erik’
[setning]	Valgfri setning.	
setning …	Repetisjon av setning.	
tall er lik	tall ::= ’0’ | ’1’ | ’2’ | ’3’ | ’4’ | ’5’ | ’6’ | ’7’ | ’8’ | ’9’	
Bokstav er lik	bokstav ::= ‘a’...’å’	Lovlige bokstaver i norsk alfabet
bokstav+	En eller flere bokstaver	
bokstav*	0 til mange bokstaver	
Repetisjon	tegnet * indikerer repetisjon (0 til mange). <l>*<m>element indikerer minst 1 og maks m repetisjoner	<tall>4*4 alle tall består av 4 siffer

Tabell 24: BNF notasjon som er benyttet

Eksempel: Definisjon av adresse til en person:
<adresse> ::= <navn-del> (<gateadresse> | <gårdsnummer>) <postnummer> <poststed>
<navn> ::= <navn-tekst> | <initialer> ‘.’
<navn-del> ::= <fornavn> [<mellomnavn] <etternavn> [‘jr’ | ‘sr’]
<gateadresse> ::= <gatenavn> <husnummer>
<postnummer> ::= <tall>4*4
navn-tekst ::= bokstav (bokstav | ‘-’ | ‘ ‘)+