UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

# MODEL ZAGOTAVLJANJA KAKOVOSTI PODATKOV

Ljubljana, april 2010 GREGOR BOROŠA

**IZJAVA**

Študent Gregor Boroša izjavljam, da sem avtor tega magistrskega dela, ki sem ga napisal v soglasju s svetovalcem dr. Jurijem Jakličem, in da v skladu s 1. odstavkom 21. člena Zakona o avtorskih in sorodnih pravicah dovolim njegovo objavo magistrskega dela na fakultetinih spletnih straneh.

V Ljubljani, dne 7. 4. 2010 Podpis: _____

# KAZALO

## KAZALO TABEL



## KAZALO SLIK

# UVOD

## Problematika in namen magistrskega dela

Podatke kopičimo v vedno večjem obsegu, a ne le v računalniških okoljih. Zbirke podatkov so z nami že od nekdaj, saj olajšujejo in utemeljujejo odločitve, ki jih sprejemamo v življenju. Če je nek vojaški poveljnik prestregel podatek o številu nasprotnikov, je imel bistveno lažjo odločitev za vojaško operacijo, zato je kakovost podatkov v vojaških okoljih skrajno pomembna (Wang, Allen, Harris & Madnick, 2003, str. 1). Podobno, če ima nek podjetnik podatek o neizkoriščenosti nekega področja z visokim potencialom, bo usmeril energijo tja in tako imel večjo verjetnost ustvarjanja dobička.

V podjetjih je vedno večji del poslovanja zapisan v zbirkah podatkov: podrobni opisi proizvodnih procesov, zgodovina finančnih tokov, komunikacija s kupci in dobavitelji, kratkoročni in dolgoročni plani podjetja in tako dalje. Vse te množice podatkov omogočajo učinkovitejše poslovne odločitve, poleg tega pa v mnogih primerih celo pogojujejo delovni proces. Brez računalnikov je poslovanje le redko še mogoče – vendar sama infrastruktura ni ključna: strežniki, računalniki in ostala infrastruktura je enostavno zamenljiva – če se pokvari, poškoduje, ali drugače izgubi del strojne opreme, se preprosto nakupi novo. A podatki so precej kočljivejši: v primeru izgube ali vprašljive integritete podatkov je podjetje lahko eksistenčno ogroženo. Pri tem ne gre le za primere uničenja podatkovne baze, pač pa tudi za sprejemanje napačnih poslovnih odločitev na podlagi nekakovostnih podatkov in izgubo nadzora nad zasebnimi podatki, kar v nekaterih primerih lahko za vedno uniči podjetje. Takšen primer, opisan na spletni strani Wired, je podjetje CardSystem Solutions, ki se je ukvarjalo z obdelavo transakcij s kreditnimi karticami (Zetter, 2005). Leta 2005 je zaradi nepravilnega upravljanja podatkov prišlo do razlitja podatkov o 40 milijonih kreditnih karticah. Podjetje je nato odkupila družba Pay by Touch, a je nato čez 3 leta prenehala poslovati.

Podatki so eden od nosilnih temeljev poslovanja združbe, imajo strateške lastnosti (Hubley, 2001). Kadri se s časom zamenjajo, infrastruktura se nadgrajuje, menjajo se stranke, izdelki, storitve, stalno pa je prisotna le poslovna ideja v mislih zaposlenih in – podatki, v katerih je shranjeno znanje podjetja in zgodovinski zapisi poslovanja.

Nekakovostni podatki lahko povzročajo vrtoglave stroške. Po analizi Data Warehousing Institute leta 2002, objavljeni na spletni strani AD Mag (Eckerson, 2002), ki je vsebovala intervjuje z vrhunskimi strokovnjaki s področja informatike in raziskavo s 647 sodelujočimi uporabniki v ZDA, so letni stroški nekakovostnih podatkov v ZDA kar 600 milijard dolarjev.

Veliki informacijski projekti, ki so v zadnjih letih propadli zaradi nepričakovano slabe kakovosti podatkov, počasi odpirajo vodilnim ljudem pogled na stroške, ki nastajajo zaradi slabih podatkov. Tega dejstva se zavedajo tudi velika podjetja s področja informacijskih tehnologij, zato je na področju izboljšanja kakovosti podatkov zadnja leta videti precej aktivnosti. IBM je leta 2005 za dobro milijardo dolarjev prevzel podjetje Ascential Software, ki je bilo eden največjih konkurentov na področju kakovosti in integritete podatkov na svetu (IBM, 2005). SAP je leta 2007 prevzel podjetje Business Objects za malo manj kot 7 milijard dolarjev (SAP, 2007). Business Objects je tako podjetju SAP, računalniškemu velikanu na področju poslovnih informacijskih sistemov, predal skoraj 20-letne izkušnje na področju poslovne inteligence in upravljanja podatkov. Podobne aktivnosti izvajajo tudi Oracle (intenziven razvoj orodja IBM WebSphere Information Analyzer), Microsoft (razvoj orodja SQL Integration Services, prevzem podjetja Zoomix leta 2008 (Microsoft, 2008), ki se je ukvarjalo predvsem s kakovostjo podatkov) in druga velika podjetja s področja informacijskih tehnologij (v nadaljevanju IT podjetja). Na drugi strani je vedno več pobud tudi na odprto-kodnih projektih, opisanih na spletni strani podjetja Talend (Madsen, 2009). Podjetja namreč dokaj pogosto delajo integracijske podatkovne vmesnike med posameznimi informacijskimi sistemi ročno, saj jih je zaradi raznolikosti sistemov potrebno v določeni meri vsakokrat prilagoditi specifičnim potrebam. Odprto-kodna skupnost poskuša zmanjšati prepad do razmeroma dragih orodij za zagotavljanje kakovosti podatkov.

V preteklosti je bilo že razvitih nekaj modelov za zagotavljanje kakovosti podatkov, a se ob aktualnih dogajanjih v svetu s področja podatkov dogajajo vselej novi pojavi, ki zahtevajo pazljivo obravnavo. Tako npr. model CDQM z modeliranjem IP-MAP (Shankaranarayanan et al., 2000) omogoča procesni pogled na tokove podatkov, a podatke obravnava razmeroma ozko ter nepovezano med seboj in s poslovnimi procesi. Redmanov model (Redman, 1996, str. 119) in model TDQM (English, 2003, str. 6), skupaj z njunimi kasnejšimi izpeljankami, predstavljata sicer širok model za zagotavljanje kakovosti podatkov, vendar se komaj dotakneta dveh, v današnjih časih izjemno pomembnih, elementov: varnosti in zasebnosti podatkov. Izhajajoč iz teh ugotovitev, bo v magistrskem delu predstavljen nadgrajen model za zagotavljanje kakovosti podatkov, v katerem bodo vključeni tudi ti elementi.

Namen magistrskega dela je osvetliti pomembnost večplastne kakovosti podatkov, ki posredno in neposredno vplivajo na učinkovitost poslovanja. Splošni trend kopičenja podatkov v najrazličnejših podatkovnih zbirkah sam po sebi zahteva razmislek, ali se obvladovanju podatkov namenja dovolj pozornosti. Poleg tehničnega vidika kakovosti podatkov, ki je običajno najbolj izpostavljen, ogromne količine podatkov zahtevajo obravnavanje tudi z vidika varnosti, varovanja zasebnosti in nadzora nad podatki v splošnem.

V magistrskem delu bo predstavljenih nekaj načinov, kako je moč preveriti in zagotavljati dolgoročno in široko opredeljeno kakovost podatkov.

## Cilj magistrskega dela

Cilj magistrskega dela je na osnovi pregleda strokovne literature sestaviti model za vzpostavitev in zagotavljanje kakovosti podatkov in predstaviti ter empirično preveriti metode za doseganje in ohranjanje kakovosti podatkov. Pri tem je najpomembnejša analiza kakovosti obstoječih podatkov, s katerimi upravlja podjetje. Rezultat te analize je dognanje, v kolikšni meri je potrebno kakovost podatkov izboljšati in kakšni so načini zagotavljanja kakovosti podatkov.

## Metode dela

Metode dela, uporabljene pri izdelavi magistrskega dela, temeljijo na študiju literature, na podlagi katere bo razvit model in opisani načini za doseganje in ohranjevanje kakovosti podatkov. Empirični preizkus modela bo izveden s pomočjo analize primera relacijske podatkovne baze za upravljanje terenske prodaje. Pri tem bo uporabljena množica algoritmov za pregledovanje strukture, povezanosti, unikatnosti, veljavnosti in točnosti podatkov.

Magistrsko delo bo za uvodom, v prvem poglavju, obravnavalo širok pregled tematike iz več zornih kotov, skozi celoten življenjski tok podatkov, in na ta način predstavilo relevantnost podatkov v poslovnem in tudi vsakdanjem življenju. Drugo poglavje bo vsebovalo natančno teoretično opredelitev pojmov, povezanih s kakovostjo podatkov, poglobljen pregled strokovne literature in kritično analizo vplivov kakovosti podatkov na poslovanje združb. Pri tem bo predstavljen tako tehnični vidik kakovosti podatkov, v smislu tehničnih in tehnoloških okvirjev, kot tudi vplivi kakovosti podatkov, ki se odražajo znotraj združbe in v njenem odnosu do okolja, torej partnerjev in konkurentov. V tretjem poglavju bodo opisani nekateri modeli upravljanja s podatki za zagotavljanje kakovosti. Predstavljena bo še razširjena različica modela, ki bo vsebovala tudi vpetost kakovosti podatkov v samo poslovanje združbe. Naslednje, četrto poglavje bo zajemalo podroben opis tehnik za doseganje in vzdrževanje kakovostnih podatkov. V petem poglavju bo predstavljena preslikava razširjenega modela na konkretnem primeru in v šestem analiza rezultatov preizkusa modela. Zadnje poglavje bo povzelo ugotovitve in priporočilo smernice za praktično uporabo v poslovnem okolju.

# 1 ŽIVLJENJSKI KROG PODATKOV V PODJETJU

Pred analizo kakovosti podatkov je zaradi nedvoumnega razumevanja potrebno opredeliti temeljne pojme, ki se uporabljajo pri tej tematiki. Pojem podatka dojemajo strokovnjaki različnih strok z različnih vidikov, ravno tako si je pojem kakovosti moč razlagati na več načinov. Opredelitev temeljnih pojmov je zato potrebna, da se vzpostavi skupni imenovalec pri uporabi pojmov kakovosti podatkov. Za tem bo predstavljen časovni potek projektov kakovosti podatkov, ki bo vodilo nadaljevanja magistrskega dela.

## 1.1 Opredelitev temeljnih pojmov

Pod pojmom kakovost podatkov si je moč predstavljati več stvari in ravno tako kot je več opredelitev pojma podatki in informacije, različni avtorji različno opredelijo kakovost podatkov. Nekatere definicije so zelo tehnične, druge povsem splošne; in ker se podatke povsem drugače obravnava v poslovnem svetu kot v računalniškem, nemalokrat pride do nerazumevanja. Za konsistentnost tega dela je zato pomembno, da temeljne pojme opišemo takoj na začetku.

**Podatki** so običajno predstavljeni s številkami, besedami ali slikami. V računalništvu so podatki kodirani na nek vnaprej dogovorjen način, ki omogoča njihovo berljivost. Iz naslednjih opredelitev lahko razberemo bistveno razliko med podatkom in informacijo, ki se v vsakdanjem življenju včasih enačita.

- Podatek je zapis dejstva ali pojava (Kroenke, 1997, str. 133-144).
- Podatek je zapis dejstva, slike ali zvoka, ki je lahko, ali pa tudi ne, primeren za določeno uporabo (Alter, 2002, str. 35).

Pojem **informacije** ima številne pomene, ki so odvisni od konteksta uporabe. Največkrat informacije povezujemo s pojmi znanja, interpretacije, pa tudi s pridobivanjem izkušenj in uporabo intuicije. Spodnji dve opredelitvi jasno predstavita razliko med podatki in informacijo.

- Informacija je znanje, pridobljeno iz podatkov. Informacija je novo znanje (Kroenke, 1997, str. 133-144).
- Informacija je rezultat obdelave podatkov, ki je po obliki in vsebini primerna za določeno uporabo. (Alter, 2002, str. 35).

- Informacija je zmanjšanje negotovosti po sprejetju sporočila (Shannon, 1948, str. 379-423, 623-656).
  Količino informacije lahko predstavimo z enačbo (1).

$$I = log(n); \qquad (1)$$

kjer logaritemska osnova določa mersko enoto informacije (npr. z uporabo $\log_2$ merimo informacijo v bitih), $n$ pa je število možnih izidov oziroma dogodkov.

**Znanje** je množica instinktov, idej, pravil in postopkov, ki vplivajo na aktivnosti in odločitve (Alter, 2002, str. 282). Znanje lahko razvrstimo v dve skupini:

- eksplicitno znanje: je formalizirano znanje, ki ga je moč razmeroma enostavno izraziti, običajno v obliki principov, postopkov, dejstev, likov, pravil, formul itd. Sčasoma postane rutinsko in prevzame značaj podatkov;
- tacitno ali skrito znanje: takšnega znanja ni enostavno izraziti niti videti. Je subjektivno in prepleteno z vedenjem in časom. Obsega izkušnje, ideale, čustva in intuicijo. Lahko ga naprej delimo na tehnično znanje (angl. *know-how*) in na zaznavno (kognitivno).

Pojem **kakovosti** vsak strokovnjak razume na svoj način in v odvisnosti od konteksta. Standard ISO 9000, ki daje smernice za upravljanje s kakovostjo, na spletni strani International Organization for Standardization opredeli kakovost kot stopnjo, do katere lastnosti opazovanega objekta ustrezajo zahtevam (International Organization for Standardization, 2008). Pri tem so zahteve opredeljene kot potrebe ali pričakovanja). Filozofsko bi lahko kakovost opredelili kot stopnjo približka idealu.

Kakovost večinoma razumemo subjektivno (kot je npr. izpolnjevanje pričakovanj uporabnika), a tudi objektivno (npr. ustrezanje danim zahtevam). To je potrebno upoštevati pri vsakem merjenju kakovosti, saj le-te ne moremo prepustiti samodejnim meritvam, ampak morajo te meritve uporabniki subjektivno interpretirati in oceniti (Eppler, 2003, str. 17).

**Življenjski krog podatkov**

Nastanek, uporaba in arhiviranje ali uničenje podatkov so najbolj očitna življenjska obdobja podatkov. V posameznih obdobjih se podatke obdeluje na med seboj dokaj različne načine, zato je potrebno, da se upravljavci podatkov dobro zavedajo celotnega življenjskega toka podatkov. V nadaljevanju bo predstavljen model POSMAD (McGilvray, 2008, str. 24), ki življenjski tok podatkov opisuje v šestih fazah:

1. načrtovanje (angl. *plan*): priprava za podatke kot poslovni vir,
2. pridobitev (angl. *obtain*): pridobitev podatkov,

3. shranjevanje in deljenje (angl. *store and share*): shranjevanje podatkov in ureditev dostopov do njih,
4. vzdrževanje (angl. *maintain*): spreminjanje, standardiziranje, preverjanje ipd. podatkov,
5. uporaba (angl. *apply*): uporaba podatkov za doseganje poslovnih ciljev ter
6. prenehanje uporabe (angl. *dispose*): arhiviranje ali brisanje podatkov.

**Povzetek temeljnih pojmov**

Podatki tvorijo osnovo za pridobivanje informacij, informacije pa služijo kot osnova za odločanje in ukrepanje. Človek informacij ne more koristno uporabiti, če nima za to potrebnega predznanja. Npr. informacije o padcu delnic so pomembne za nekoga, ki je aktiven na borznem trgu, drugemu pa so odveč. Ljudje podatke interpretiramo, iz česar pridobimo informacije in se na tej podlagi in na podlagi obstoječega znanja odločamo in ukrepamo. Rezultati takšnih aktivnosti nato pomagajo pri akumulaciji dodatnega znanja (učenje).

## 1.2 Ohranjanje zgodovine podjetja

Arhivski oziroma zgodovinski podatki vsebujejo shranjena dejstva o poslovanju podjetja v določenih časovnih rezinah. Za te podatke je ključno, da se jih ne sme spreminjati, razen v izjemnih primerih, ko je potrebno popraviti določene napake ali neskladnosti. Zgodovinski podatki so pomembni s stališča varnosti (saj predstavljajo posnetek stanja v določenem trenutku v preteklosti, poleg tega je moč ugotavljati želene ali nenamerne spremembe podatkov s primerjavo zgodovinskih in trenutnih podatkov), pomembni pa so tudi za poročila in analize podatkov, ki vsebujejo časovno dimenzijo, npr. gibanje zalog.

S časovno pogojenimi atributi podatkov označujemo lastnosti podatkov (poslovnih entitet), ki se spreminjajo s časom. Kot se v času spreminja večina stvari v svetu, npr. otroci zrastejo v najstnike, ki odrastejo v zaposlene poslovneže in nato v modre zrele osebe, se spreminjajo tudi kadrovske strukture v podjetjih, zaposlenim se spreminja raven usposobljenosti in s tem višina plač, rojevajo se nove države, umirajo in rojevajo se zvezde, tako je neizogibno tudi časovno spreminjanje podatkov (in informacij), ki jih zbirajo podjetja. Te spremembe podatkov morajo omogočati ustrezni sistemi za upravljanje podatkovnih baz (v nadaljevanju SUPB) – ne le transakcijski sistemi, ki podpirajo najnižjo raven odločanja v poslovnih sistemih, ampak tudi sistemi za upravljanje podatkovnih skladišč na operativni in taktični ravni. Podatke v določeni časovni rezini označimo v posebnih atributih v obliki časovnih oznak (angl. *timestamp*), ki so sicer lahko shranjeni v različnih oblikah, npr. kot "dan-mesec-leto ura:minute:sekunde".

Kritičnost kakovosti zgodovinskih podatkov in njihovih časovnih oznak daje splošno priporočilo, da je časovnim oznakam potrebno dati velik pomen pri izvajanju kakršnih koli izboljšav kakovosti podatkov.

# 2 KAKOVOST PODATKOV

## 2.1 Pomen in opredelitev kakovosti podatkov in informacij

Skrb za kakovost podatkov je že desetletje prisotna na poslovnih področjih, v zdravstvu, v geografskih informacijskih sistemih in še mnogo kje. Načela kakovosti podatkov postajajo vpeta vse širše v področja, kjer se uporabljajo informacijsko-komunikacijski sistemi. Zanimivo je, da se šele v zadnjem času tej tematiki daje več poudarka na področju upravljanja taksonomij. Taksonomski podatki so podatki, s katerimi so opisane hierarhične strukture stvari (npr. v biologiji so s taksonomijami opisana razmerja med živimi organizmi). Zaradi močne razkropljenosti podatkov, nešteto načinov shranjevanja in uporabe podatkov, nenazadnje tudi zaradi razmeroma časovno zelo dolgih obdobij uporabe, bi pri taksonomskih podatki pričakovali že ustavljene načine za zagotavljanje kakovosti podatkov. Chapman (2005) takole predstavlja ta problem: "Hitro povečevanje izmenjave in dostopnosti taksonomskih podatkov in drugih podatkov opazovanja narave zahteva bistveno večji poudarek na kakovosti podatkov, saj njihovi uporabniki želijo vse več podrobnosti. Dogaja se, da so podatki, ki so jih zbrali v nekem muzeju, popolnoma nesprejemljivi ali nerazumljivi ljudem iz drugih muzejev. Vprašanje je, ali je temu tako zaradi nekakovostnih podatkov ali slabe dokumentacije teh podatkov. Ti podatki so namreč ključni in izjemnega pomena. Ker se zbirajo skozi mnogo let, predstavljajo neprecenljivo bazo podatkov o biološki raznovrstnosti v času, ko je imel na to raznovrstnost velik vpliv prav človek. So bistven vir, ki vodi vsak napor za ohranjanje okolja, saj beležijo spremembe habitatov, povzročene zaradi posekavanj in požigov gozdov za potrebe kmetijstva, ali zaradi urbanizacije, podnebnih sprememb ali česa drugega." Tematika kakovosti podatkov torej zadeva praktično vsa področja, kjer se podatki uporabljajo.

V časih, ko govorimo o podatkih, ki so dostopni preko najrazličnejših naprav, od osebnih računalnikov do mobilnih telefonov, in vedno na voljo 24 ur na dan, večinoma ne vemo, kje so ti podatki dejansko shranjeni in v kakšni obliki. Uporabnika podatkov takšna vprašanja običajno niti ne zanimajo, vendar se lahko pojavijo velike težave, če uporabnik iz takšnih ali drugačnih razlogov tem podatkom ne verjame več. Primer, ki ni tako redek, so letne analize in planiranja: skozi vse leto se uporabljajo obstoječi podatki na mikro nivoju, npr. tedenski ali mesečni pregledi cen, prodaje in stroškov. Ob celovitejšem, letnem pregledu gibanja cen pa se lahko ugotovijo večja ali manjša odstopanja od pričakovanih vrednosti. Izvor

nepričakovanih podatkov, ugotovljenih na tako visoko agregiranem nivoju, je izziv za marsikaterega poslovneža in informatika. Za letnim prikazom cen se največkrat skriva kakšno podatkovno skladišče, nekaj relacijskih baz podatkov, datotek s preglednicami … Preveč virov podatkov, da bi se jih mogli zavedati ob vsakodnevni uporabi poslovnih programov. Skupna množica podatkov v podjetju torej zajema najrazličnejše zbirke podatkov. Zato ni čudno, da je npr. prikazan skupni seštevek vseh kupcev v poslovni aplikaciji 10 tisoč kupcev, čeprav uporabniki vedo, da je dejansko kupcev samo 8 tisoč. Neusklajenost in nekonsistenca podatkov v podjetju imajo lahko daljnosežne in globoke posledice:

- nizka izkoriščenost sistemov za management odnosov z odjemalci (CRM):
  vzroki so lahko slaba opredelitev, kaj pojem kupec pomeni za podjetje, kako so kupci povezani med seboj v hierarhije in druga razmerja, kako so določeni naslovi in kontaktni podatki kupcev ...
- napačni podatki pri uporabi programov za poslovno inteligenco:
  med vzroki je pogosto nepravilno določen podatkovni model (glavni ključi na glavnih dimenzijah podatkov, kot so kupci, dobavitelji in izdelki, ki ne zagotavljajo pravilne slike poslovnega sveta), npr. mešanje različnih merskih enot med seboj ali nekonsistentne vrednosti podatkov med različnimi vrstami pregledov,
- neusklajenost, podvajanje, nepreglednost izdelkov:
  ob uporabi več podatkovnih zbirk lahko pride do podvajanj podatkov o izdelkih, če podatki niso ustrezno sinhronizirani med zbirkami, analiza stroškov kar naenkrat postane izjemno zapletena, potrebno je ročno obdelovanje podatkov.

Da je obravnavanje kakovosti podatkov vedno bolj potrebno, kaže tudi novi standard v pripravi, ISO 8000, od katerega je en del od štirih že objavljen. ISO 8000-110:2008 vsebuje strukturna in vsebinska pravila za kakovostne podatke, predvsem glavnih, matičnih podatkov ključnih poslovnih entitet v informacijskih sistemih. Z izpolnjevanjem zahtev tega standarda podjetja lahko del nadzora nad podatki prepustijo avtomatskim računalniškim obdelavam, ki skrbijo, da so podatki skladni s specifikacijami oziroma metapodatki, ki o njih obstajajo. S tem, ko vsi podatki v sistemu, ali vsaj dovolj velik del njih, ustreza določenim pogojem in omejitvam, se v podatkovnih zbirkah že odstranijo nekatere vrste anomalij, uporabnikom se poveča zaupanje v podatke in manj je težav s pripravo analiz na podlagi teh podatkov.

Kakovost podatkov je opredeljena na več načinov, ena izmed najpogosteje citiranih opredelitev je: »Kakovost je različno opredeljena kot 'primernost za uporabo', 'izpolnjevanje pričakovanj', 'stopnja odličnosti' in 'ustreznost standardom'. Uporabljajo se tudi druge

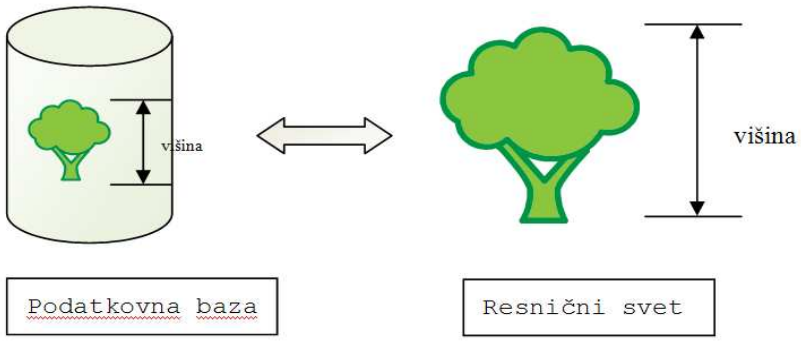
opredelitve, njihova uporaba je odvisna predvsem od uporabnika oziroma konteksta uporabe« (Hayes & Romig, 1977, str. 9).

Primernost za uporabo pomeni učinkovitost oblike in izdelave oziroma vsebine pri ustvarjanju izdelka ali storitve, ki ustreza predpisanemu oziroma želenemu namenu. To je opredelitev, na katero v literaturi zelo pogosto naletimo (Hayes & Romig, 1977, Chapman, 2005, IAIDQ na spletni strani in mnogi drugi), izkaže pa se, da njeno pokritje ni vedno polno, npr. v primerih, kjer je uporabnost določena v prihodnosti.

Primer tega je popis neke živalske vrste na območju A, ki je zabeležen z zelo visoko natančnostjo. Če ima območje B neničelni presek z območjem A, in se vprašamo, ali ta vrsta živali biva v območju A, so podatki popisa primerni za uporabo. V primeru pa, da se vprašamo, ali ta vrsta živi na območju B, podatki niso primerni za uporabo (lahko so zavajajoči). Lahko rečemo, da so podatki tega popisa potencialno uporabni, skratka imajo dobro možnost za uporabo v prihodnosti (Chapman, 2005, str. 4).

Drug primer, ki poudarja pomen opredelitve kakovosti podatkov kot »potencialno uporabnost«, so zapisi v podatkovnih bazah, ki sami kot taki nimajo nobenega pomena ali kakovosti. Imajo le potencial oziroma možnost uporabnosti, ki je udejanjena šele, ko nekdo do teh zapisov dostopa na nek uporaben način (English, 1999, str. 15-30). Z zornega kota računalniških baz podatkov lahko kakovost podatkov opredelimo kot stopnjo ujemanja podatkov z resničnim svetom, kar je prikazano na Sliki 1. Popolno ujemanje pomeni, da zapisi v podatkovni bazi povsem odražajo stanje resničnega sveta; nasprotno, če ujemanja ni, pomeni, da podatki niso usklajeni z resničnim svetom (Orr, 1997, str. 1-3).

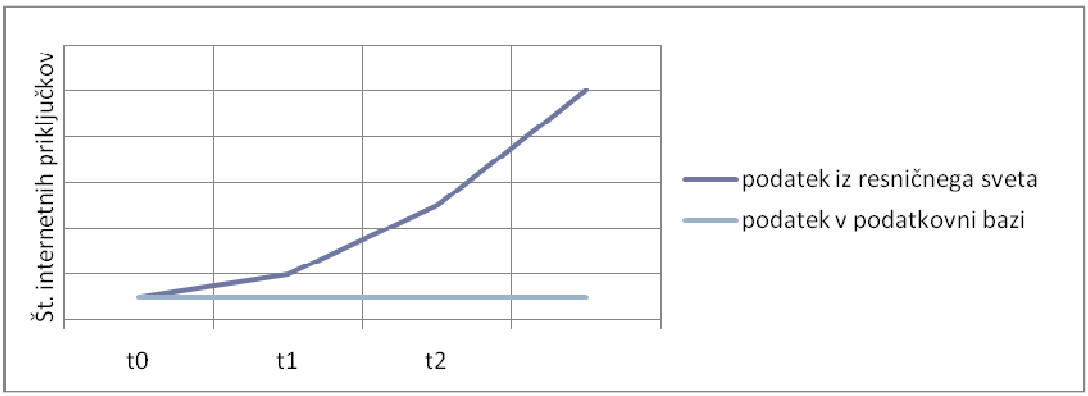
*Slika 1: Preslikava dejstva v resničnem svetu v zapis dejstva v podatkovni bazi*




Namen upravljanja s kakovostjo podatkov pa ni v tem, da podatki popolnoma odražajo stanje resničnega sveta, temveč da je to ujemanje doseženo vsaj do nekega nujnega približka. Razlog, zakaj popolnega ujemanja ni mogoče ohranjati, tudi če bi ga v neki časovni rezini

dosegli, je spreminjanje sveta – obstoječi podatki v podatkovnih bazah so (večinoma) statični in odražajo stanje obravnavanega sveta v nekem trenutku. Sledenje in obravnavanje sprememb v času pa je zapleten problem, ki v odvisnosti od posameznega primera uporabe zahteva natančno obravnavo. Primer neujemanja podatkovne baze z dejanskim svetom je razvidno s Sliko 2.

*Slika 2: Večanje razlik med resničnostjo in podatkovnimi zapisi glede na čas*




Kakovost podatkov zahteva razmeroma drugačne pristope k obravnavi kot so sicer običajni pri upravljanju podatkov v splošnem, kar ponazarjajo naslednja načela (Orr, 1997, str. 3):

- Točnost podatkov, ki se ne uporabljajo, se lahko postopoma manjša.
- Kakovost podatkov v informacijskem sistemu lahko opišemo s funkcijo uporabe podatkov in ne s funkcijo zbiranja podatkov.
- Kakovost podatkov se s časom bliža tisti stopnji uporabe podatkov v sistemu, ki je najstrožje omejena (če je slabo omejena, je tudi kakovost slaba).
- Težave zaradi kakovosti podatkov postajajo s staranjem informacijskega sistema vse večje.
- Manj, ko je verjetno, da se bo nek podatkovni atribut spremenil, bolj travmatična bo sprememba, ko se bo zgodila.
- Načela kakovosti podatkov veljajo v enaki meri za podatke in za meta podatke.

Pri časovnih oznakah, s katerimi označujemo podatke v časovnih rezinah, v praksi pogosto naletimo na naslednje težave:

- niso točne (zaradi programskih/strojnih zakasnitev v obdelavah ali drugih zakasnitev v sistemu) ali pravilne (npr. neskladnost časovnih oznak zaradi različnih datumskih oblik, ki jih uporabljajo sodelujoči programi),

- časovnih oznak ni oziroma jih ni povsod, kjer bi morale biti (npr. če se podjetje odloči shranjevati spremembe podatkov s pomočjo časovnih oznak v trenutku *t*, potem sprememb podatkov pred *t* v dnevniku še ni zapisanih; drug razlog je neusklajenost postavljanja časovnih oznak pri uporabi podatkov preko več programov),
- podatki, ki so v podatkovnih zbirkah prisotni dlje časa, imajo večjo verjetnost, da so se nad njimi izvajale nekatere obdelave, kot npr. arhiviranje ali stiskanje, pri čemer je potreben dogovor za ohranjanje ali ažuriranje časovnih oznak med potekom obdelave.

Omejitve oziroma pravila časovnih oznak zagotavljajo, da so vsi potrebni in zahtevani podatki shranjeni s pravimi časovnimi oznakami (Maydanchik, 2007, str. 12-12). Pravilo časovne veljavnosti (angl. c*urrency rule*) uveljavlja zahtevo po časovni ustreznosti podatkov, običajno v obliki omejitev na datumih zadnjih zapisov. Primer pravila časovne veljavnosti je npr. letni pregled plač delavca, kjer se zahteva, da je zadnji zapis datumsko skladen z zadnjim mesecem v zaključenem letu (in ne npr. datum pred ali po zadnjem mesecu v letu). Drugo pravilo je pravilo ohranjanja (angl. *retention rule*), ki zagotavlja ustrezno časovno globino zgodovinskih podatkov. Primer ohranjevalnega pravila so omejitve, kdaj lahko nek podatek zbrišemo oziroma koliko časa ga moramo hraniti (npr. davčna poročila hranimo 5 let). Pri podatkih, kjer se uporabnost izkazuje skozi več združevanj podatkov (npr. letno gibanje zaloge), je pomembno upoštevati še pravilo granularnosti, ki zahteva konsistentne podatke v vseh časovnih rezinah, ter pravilo polnosti, ki prepoveduje luknje in prekrivanja v podatkih.

Če sicer pri zbiranju podatkov običajno velja, da zbiramo čim več podatkov, saj nam lahko v prihodnosti »pridejo prav«, je takšno razmišljanje s stališča kakovosti podatkov načeloma odsvetovano. Razlog za to je dinamika resničnega sveta. V primeru, da z neko akcijo posredno zberemo in shranimo tudi kopico »postranskih« podatkov (npr. ogromen seznam e-poštnih naslovov), pa jih v podjetju potem leta in leta nihče ne uporablja, spremembe podatkov, ki se zgodijo v resničnem svetu, ne bodo ažurirane v podatkovni bazi in kakovost teh podatkov bo dramatično padla (čez nekaj let je prvotni seznam e-poštnih naslovov prej škodljiv kot uporaben). V bioloških sistemih je temu analogna atrofija – neuporabljeni deli živega bitja se zmanjšajo, odmrejo ali zakrnijo. V informacijskih sistemih pa ni vedno enostavno ugotoviti, ali se nekateri podatki uporabljajo ali ne. Morda se ti podatki nahajajo na nekih poročilih, ki pa jih nihče ne uporablja več.

Po drugi strani se neke podatke lahko uporablja, vendar se uporabniki ne zavedajo zgornjih načel kakovostnih podatkov, npr. tega, da je kakovost podatkov sčasoma velika le toliko, kot dopušča najstrožje izmed pravil pri upravljanju teh podatkov. Kot primer naj se vrnem k seznamu elektronske pošte: če se ne ažurira sproti, vsakokrat ob masovnem pošiljanju e-

sporočil povzroča stroške, saj obdelava na računalniku traja dlje časa, zavrnjena sporočila zahtevajo dodatno obravnavo uporabnika, zaradi pošiljanja neželene pošte lahko nastanejo celo sodni spori in tako dalje.

Potrebno je razložiti tudi pojem kakovosti informacij ter pojasniti, kako se razlikuje od kakovosti podatkov. V splošnem nam več podatkov, oziroma širša in globlja pokritost problemskega področja s podatki, olajša odločitve, ki temeljijo na njih. Vendar velike množice podatkov hitro postanejo nepregledne in posledično manj uporabne. Zato je te velike količine podatkov smiselno primerno agregirati, jih interpretirati, oziroma iz njih izvleči informacije. Iz tega stališča je bolje manj informacij – a te naj bodo kakovostnejše.Kakovost informacij je širše opredeljena kot kakovost podatkov. "Upravljanje s kakovostjo informacij je moč opredeliti kot kombiniranje med tradicionalnim upravljanjem kakovosti in upravljanjem z znanjem" (Eppler, 2003, str. 24), saj je kakovost informacij močno odvisna tudi od konteksta uporabe in vsebine. Tabela 1 prikazuje pogled na težave kakovosti z vidika podatkov in z vidika informacij.

*Tabela 1: Težave kakovosti podatkov v primerjavi s težavami kakovosti informacij*

| **Primeri težav nekakovostnih podatkov** | **Primeri težav nekakovostnih informacij** |
|---|---|
| Podvojeni zapisi, več virov podatkov | Nasprotujoča priporočila v raziskavi ali analizi |
| Manjkajoče podatkovne relacije | Nejasni vzročno-posledični učinki v diagnozi |
| Nesmiselni podatki | Dolga, nelogično strukturirana poročila |
| Črkovalne napake | Stilsko neustrezno besedilo z jezikoslovnimi napakami |
| Zastareli podatki | Neažurirane analize, ki ne upoštevajo novih odkritij |
| Nekonsistentna struktura in poimenovanje podatkov | Nekonsistentni pregledi in navigacija programa |
| Podatki shranjeni na napačnih mestih | Izgubljeni dokumenti |
| Zapletene poizvedbe za dostop do podatkov | Zahtevno iskanje želenih informacij |
| Napačno označeni podatki (neustrezni meta-podatki) | Neprimerna ali pomanjkljiva kategorizacija |
| Netočen vnos podatkov zaradi pomanjkljive kontrole | Neutemeljene odločitve na podlagi nezadostnih podatkov |
| Neželene spremembe podatkov (brisanje, spreminjanje) | Neželene spremembe v odločitvenih procesih (zavajanje, zmeda, dvoumnosti) |

*Vir: M. J. Eppler, Managing Information Quality, 2003, str. 29.*

Zagotavljanje kakovosti podatkov in nadzor kakovosti podatkov sta povezana, a različna pojma. Na prvi pogled razlika ni očitna, vendar sta drug od drugega odvisna (Taulbee, 1996, str. 47-75):

- nadzor kakovosti je presoja, temelječa na internih standardih in procesih, ki skrbi za nadzorovanje in spremljanje kakovosti,
- zagotavljanje kakovosti pa je postopek, osnovan na zunanjih standardih izven procesov in nadzora kakovosti, ki uresničuje, da končni izdelki ali storitve ustrezajo vnaprej določenim kakovostnim standardom.

Kakovost podatkov je moč dosegati na dva načina: s preprečevanjem napak v splošnem (preventiva) in s popravljanjem obstoječih napak (kurativa). Velja pravilo, da tem bolj na začetku življenjskega kroga podatkov ko napake odpravimo, manjši so stroški za odpravo napak. A samo s preprečevanjem napak ni vedno moč zagotoviti potrebne kakovosti podatkov, dasiravno temu namenjamo velik napor (Maletic & Marcus, 2000, str. 1-2). Nepravilni podatki se neredko vseeno prikradejo v sistem: razen če organizacije ne namenjajo izjemnih naporov v preprečevanje napak, je nepravilnih podatkov običajno vsaj 5 odstotkov. Prav zato odpravljanje napak v podatkih zahteva obravnavo in morajo obstajati mehanizmi za takšne postopke. Obseg podatkov nemalokrat zahteva avtomatizacijo, o čemer obstaja razmeroma veliko literature. Obširen pregled takšne literature je moč dobiti na spletni strani prof. Nuray-Turan iz kalifornijske univerze Irvine (http://www.ics.uci.edu/~rnuray/biblio_conf.html).

## 2.2 Dimenzije kakovosti podatkov

Pojem dimenzije kakovosti podatkov ima v literaturi včasih dvoumen pomen, saj dimenzija običajno pomeni množico atributov, ki predstavljajo določen aspekt oziroma konstrukt značilnosti kakovosti podatkov. Kljub temu posamezne atribute kakovosti podatkov pogosto enačijo s pojmom dimenzije (npr. Wang & Strong,, 1996, str. 1; Godnov, 2008, str. 46; McGilvray, 2008, str. 30), zato bo takšno poimenovanje uporabljeno tudi v magistrskem delu.

### 2.2.1 Predstavitev dimenzij kakovosti podatkov

Wang in Strongova (1996) sta v svoji empirični raziskavi dimenzije kakovosti podatkov raziskovala z vidika uporabnikov. Raziskava ni temeljila na tradicionalnih modelih kakovosti, kakovost podatkov sta želela predstaviti z uporabniškega vidika, uravnotežiti posamezne atribute po pomembnosti in jih razvrstiti v hierarhični okvir. Postavila sta

temeljne opredelitve in razčlenitve kakovosti podatkov, na katerih sloni večina tudi novejših raziskav (Redman leta 1996, Jarke leta 1999, Bovee leta 2001, Naumann leta 2002 itd.). Z anketiranjem uporabnikov sta preučevala 179 podatkovnih atributov, ki sta jih združila v 15 dimenzij, razporejenih v štiri kategorije, kot prikazuje Tabela 2.

*Tabela 2: Skupine dimenzij kakovosti podatkov*

| Kategorija kakovosti podatkov | Dimenzija kakovosti podatkov |
|---|---|
| Notranje dimenzije | natančnost, objektivnost, verodostojnost, ugled |
| Kontekstne dimenzije | dodana vrednost, pomembnost, pravočasnost, popolnost, ustreznost količine podatkov |
| Predstavitvene dimenzije | interpretabilnost, razumljivost, doslednost, jedrnatost |
| Dimenzije dostopnosti | dostopnost, varnost |

*Vir: R. Y. Wang & D. M. Strong, Beyond accuracy, 1996, Tabela 1.*

Kategorija notranjih dimenzij izpostavlja verodostojnost in ugled, ki ga ima nek podatkovni vir. To je bistvenega pomena, saj sta to težje merljivi, »mehkejši« značilnosti podatkov, ki ju oceni uporabnik in ne ponudnik podatkov. Za tehnično osredotočene informatike ti dve dimenziji nista vedno relevantni, čemur bi torej morali posvetiti več pozornosti.

Kontekstne dimenzije postavijo kakovost podatkov v močno odvisnost od posameznega primera uporabe. To nakazuje, da ne obstaja splošno in vseobsegajoče pravilo, kakšna je kakovost podatkov, temveč je kakovost vedno odvisna od konteksta vsake posamične uporabe.

Predstavitvene dimenzije se dotikajo oblike podatkov (doslednost in jedrnatost) ter vsebine podatkov (za podatke se zahteva, da jih znamo razložiti in razumeti). S temi vidiki kakovosti podatkov se v veliki meri in neposredno srečujemo pri oblikovanju modelov podatkovnih baz.

Dimenzije dostopnosti so se s pojavom računalniških omrežij nekoliko spremenile: če je bila prej težava dostopati do podatkov predvsem fizične narave, npr. papirni arhivi na oddaljenih lokacijah, zdaj razdalja in sam medij, kjer so podatki shranjeni, ne igrata več tako pomembne vloge. Dostopnost in varnost sta zato zdaj predvsem neločljivo povezana z nadzorovanjem dostopov do podatkovnih zbirk in imata ključno vlogo pri obravnavanju kakovosti podatkov.

Jasna opredelitev dimenzij kakovosti podatkov je za projekte kakovosti podatkov pomembna zato, ker predstavlja temeljno usmeritev projekta: kaj preučujemo, ko analiziramo kakovost,

in kakšne so standardne pričakovane vrednosti teh analiz. V nadaljevanju bo uporabljena sodobnejša opredelitev dimenzij kakovosti podatkov (McGilvray, 2008, str. 31), predstavljena v Tabeli 3, ki bo dodatno razširjena z nekaterimi dimenzijami, ki so z razmahom omrežij in socialnih mrež na internetu postale aktualne.

*Tabela 3: Dimenzije kakovosti podatkov*

| Št. | Dimenzija | Opis dimenzije |
|---|---|---|
| **1.** | Podatkovne strukture in pravila | Mera, ki vrednoti obstoj, popolnost, kakovost in dokumentacijo podatkovnih in poslovnih pravil, podatkovnih modelov, metapodatkov in referenčnih podatkov. |
| **2.** | Temeljna integriteta podatkov | Mera, ki vrednoti obstoj, veljavnost, strukturo, vsebino in druge temeljne lastnosti podatkov. |
| **3.** | Podvajanja podatkov | Mera, ki vrednoti število dvojnikov: v polju tabele, med posameznimi vrsticami tabele ali v izbrani množici podatkov. |
| **4.** | Točnost podatkov | Mera, ki vrednoti vsebinsko točnost podatkov. |
| **5.** | Skladnost podatkov (konsistenca in sinhronizacija) | Mera, ki vrednoti enakost podatkov, shranjenih ali uporabljanih v različnih podatkovnih zbirkah, programih in sistemih; poleg tega vrednoti tudi procese, ki skrbijo za skladnost podatkov. |
| **6.** | Ažurnost in dostopnost podatkov | Mera, ki vrednoti stopnjo pravočasnosti podatkov v pričakovanih časovnih okvirih in stopnjo dostopnosti podatkov glede na zahteve. |
| **7.** | Enostavnost uporabe podatkov | Mera, ki vrednoti možnost dostopa ter uporabe podatkov in stopnjo, do katere je podatke moč spreminjati, vzdrževati in upravljati. |
| **8.** | Podatkovno pokritje | Mera, ki vrednoti vsebinsko pokritje opazovanega sveta zapisanega v podatkovnih zbirkah, glede na celotno širino opazovanega sveta v resnici. |
| **9.** | Predstavitev podatkov | Mera, ki vrednoti kako so podatki prikazani in kako so zajeti s strani uporabnikov. |
| **10.** | Razumevanje, pomembnost in zaupanje podatkov | Mera, ki vrednoti razumevanje in verodostojnost podatkov v smislu pomembnosti, nujnosti in vrednosti za poslovne potrebe. |
| **11.** | Razkrojevanje podatkov | Mera neželenih sprememb podatkov, običajno opredeljena za neko časovno obdobje. |
| **12.** | Namenskost podatkov | Mera, ki vrednoti stopnjo, do katere bodo opazovani podatki uresničili pričakovane poslovne transakcije oziroma rezultate. |

*Vir: D. McGilvray, Executing data quality projects, 2008, str. 31.*

**Dimenzija 1: Podatkovne strukture in pravila**

Podatkovne strukture in podatkovna pravila predstavljajo referenčni sistem, s katerim lahko primerjamo rezultate analize kakovosti podatkov. Poleg tega služijo kot navodila: za ročne (ali izredne) posege v podatke, za opredelitev podatkovne ravni programov, za razvoj programske opreme, kot del dokumentacije ...

**Dimenzija 2: Temeljna integriteta podatkov**

Vse druge dimenzije kakovosti podatkov temeljijo na dimenziji, ki določa temeljno integriteto podatkov. Ta zajema osnovne mere kakovosti podatkov, kot so: polnost zapisov (npr. za katere kupce nimamo shranjenega e-poštnega naslova), zaloga (območja) vrednosti zapisov, ekstremne vrednosti zapisov, referenčna integriteta ... Za analizo te dimenzije kakovosti podatkov obstajajo tudi programska orodja, od najpreprostejših, ki podajo le osnovne indikatorje kakovosti, do zapletenih sistemov, ki podatke analizirajo podrobno v globino in širino.

**Dimenzija 3: Podvajanja podatkov**

Z dimenzijo podvojenih podatkov ugotavljamo, kateri podatki so shranjeni na več kot enem mestu in kakšne posledice imajo za poslovanje. Podvojeni podatki so včasih potrebni npr. zaradi porazdelitve računalniških obdelav, v splošnem pa so nezaželeni.

Normalizacija podatkovnih modelov, ki je ena od ključnih metod za zagotavljanje logično konsistentnih podatkovnih baz in eden od temeljnih načinov za preprečevanje anomalij v podatkih, že sama po sebi preprečuje podvojene zapise v podatkih. Pojem normalizacije je že leta 1970 opredelil oče relacijskih podatkovnih baz, E. F. Codd. Normalizacija v svoji osnovni ideji skrbi za to, da so podatki v podatkovni bazi vedno zapisani tako, da sistematično zagotavljajo pravilne rezultate poizvedb. To pomeni, da dodajanje, spreminjanje ali brisanje zapisov ne povzroča anomalij. Te bi npr. nastale, če bi imeli nek podatek zapisan v dveh tabelah, nato pa bi ga spremenili le v pri tabeli – poizvedbe, ki bi uporabljale drugo tabelo, bi tako dobile drugačen rezultat, kot tiste, ki bi uporabljale prvo tabelo. Normalizacijo je moč zagotavljati do različnih nivojev, pri čemer velja, da višja kot je oblika normalizacije, strožja pravila veljajo za podatkovni model.

Podvajanja podatkov imajo lahko zelo negativne posledice v poslovnem svetu. V primeru, da imamo podvojene podatke o kupcih, npr. enkrat je kupcu določen en naslov, drugič drug, potem imamo težave izterjati dolg tega kupca, saj ne vemo na kateri naslov poslati opomin. Drug primer je težavno ugotavljanje kreditnega limita kupca, če ga v sistemu vodimo pod več številkami.

**Dimenzija 4: Točnost podatkov**

Zagotavljanje točnosti podatkov zahteva primerjavo podatkov z dejstvi v resničnem življenju, ki jih ti podatki predstavljajo, ali vsaj primerjavo z nekim drugim verodostojnim virom. Preverjanje točnosti podatkov je vsebinski problem, pri katerem je pogosto potrebno sodelovanje strokovnjakov z obravnavanega področja. Če je npr. pri analizi temeljne integritete podatkov (dimenzija 2) ugotovljeno neko odstopanje od običajnih vrednosti, ni mogoče z avtomatiko te vrednosti zavreči ali normirati. Programu vrednost "500" ne pove veliko, nekemu strokovnjaku pa lahko to predstavlja izjemen dogodek in razlog za sprejem drugačnih odločitev, kot bi jih sprejel sicer. Podobno, računalniku so količine zaloge samo številke, uporabniki računalnika jim nato verjamejo ali tudi ne, a le skladiščnik v resnici ve, koliko izdelkov je dejansko na policah. Točnosti podatkov zato večinoma ni moč preverjati hitro, ampak zahteva ročno delo in posledično več časa.

**Dimenzija 5: Skladnost podatkov (konsistenca in sinhronizacija)**

Skladnost podatkov vključuje dva pogleda na medsebojno povezanost oziroma usklajenost podatkov. Podatki v eni tabeli morajo biti tako ali drugače skladni s podatki v drugih tabelah (kar med drugim zagotavlja referenčna integriteta, glavni in tuji ključi v tabelah, omejitve podatkovne sheme (angl. *constraints*) itd.). Npr. prodajni računi morajo biti nekako povezani s kupci, sicer ne bi mogli ugotoviti komu smo kaj prodali.

Po drugi strani morajo biti podatki skladni tudi med podatkovnimi bazami, med strežniki in po vsem sistemu, kjer se uporabljajo. Zaradi zagotavljanja visoke razpoložljivosti ali izboljšanja performančnih zmogljivost je včasih podatke potrebno hraniti na dislociranih sistemih. To je pogosta naloga sinhronizacijskih opravil, ki tečejo med posameznimi zbirkami podatkov. V tem primeru so potrebna jasna pravila, kaj, kdaj in kako bo sinhronizacija potekala.

**Dimenzija 6: Ažurnost in dostopnost podatkov**

Dinamika delovanja sveta nujno pomeni tudi, da se bodo podatki v zbirkah sčasoma spreminjali, brisali in dodajali. Pri tem velja, da od spremembe v resničnem življenju, do ažuriranja podatkov v podatkovnih bazah, vedno preteče nekaj časa. S tega stališča je potrebno zagotoviti dovolj sveže oziroma ažurne podatke kot jih uporabniki potrebujejo in omogočiti, da so podatki dostopni takrat in tistim uporabnikom kot je to potrebno. Z ažurnostjo podatkov izražamo hitrost, s katero se spremembe v resničnem svetu odrazijo v podatkovnih zbirkah.

**Dimenzija 7: Enostavnost uporabe podatkov**

Pojem enostavnosti uporabe podatkov ni enak pojmu dostopnosti podatkov. Drži, da dostopnost nastopi pred enostavnostjo, vendar imajo zapleteni dostopi do podatkov negativen vpliv na njihovo uporabnost in učinkovitost sistema. Ni tako redko, ko ekipa ljudi več dni zbira podatke za neko poročilo, ko bi z enostavnejšim sistemom lahko ti podatki bili z enim klikom miške pripravljeni v trenutku. Identificiranje in odprava zapletenih procesov uporabe podatkov lahko veliko pripomore k učinkovitejšemu delu zaposlenih.

**Dimenzija 8: Podatkovno pokritje**

Ta dimenzija predstavlja širino, do katere zapisi v podatkovni bazi opisujejo (zrcalijo) opazovani svet. Pri tem je potreben dogovor o tem, ali je npr. opazovani svet kupcev le množica kupcev, ki jim prodaja dotično podjetje, ali pa vsi (potencialni) kupci. Glede na ta dogovor je odvisno, ali želimo 100-odstotno pokritje, ali le dovolj veliko pokritje, ki še ustreza poslovnim zahtevam.

**Dimenzija 9: Predstavitev podatkov**

Predstavitev podatkov je pomembna predvsem pri zajemu in prikazu podatkov. Ta dimenzija vključuje kakovost uporabniških vmesnikov programov, oblike poročil ipd. Zmožnost, funkcionalnost iskanja po podatkih in prikaz zadetkov iskanja so vsaj v velikih množicah podatkov absolutno ključni za kakovost podatkov (brez iskalnikov tudi interneta ni!). Predstavitev podatkov je ocenjena subjektivno s stališča uporabnikov. Zaradi »mehkejše« narave ocenjevanja je to dimenzijo razmeroma težje določiti, a ima lahko velik vpliv na kakovost podatkov, saj enostavnejši uporabniški vmesniki že sami po sebi pritegnejo uporabnika, da izpolni več polj na obrazcu in da so podatki točni in pravilno vpisani.

**Dimenzija 10:Razumevanje, pomembnost in zaupanje podatkov**

Iz opredelitve podatkov sledi, da so le-ti zgolj zapisi takšnih ali drugačnih dejstev. Za uporabnike podatkov pa ti zapisi naj ne bi bili le neme črke in številke, temveč morajo biti razumljivi oziroma prevedljivi v okolje resničnega življenja. Iz tega sledi, da uporabnike zanimajo le podatki, ki so posredno pomembni, npr. za poslovanje.

Zaupanje v podatke je izjemnega pomena pri zagotavljanju kakovosti podatkov. To pride do izraza predvsem na višjih ravneh odločanja; na nižjih nivojih običajno poteka vnos podatkov in preproste obdelave, agregacije podatkov, kjer je zaupanje trivialno. Zaupanje je tesno povezano z razumevanjem, saj če je razumljivo, kako je nek podatek nastal, mu je lažje zaupati. Podatki, ki uživajo večje zaupanje, pravimo, da so verodostojni, in imajo zato večjo težo pri sprejemanju odločitev na njihovi podlagi.

**Dimenzija 11: Razkrojevanje podatkov**

Pri podatkih, za katere pričakujemo, da se bodo lahko spreminjali, je potrebno oceniti interval osveževanja. Vendar sprememb podatkov ni moč vedno predvideti – kadar se te spremembe nepričakovano zgodijo in imajo negativen vpliv na uporabnost podatkov, govorimo o razkrojevanju podatkov. Pogosto velja splošno pravilo, da s starostjo podatkov pada njihova točnost. Primer je sprememba številčenja področnih telefonskih številk, prehod iz 6- na 7-mestne telefonske številke, spremembe ali dopolnitve poštnih številk itd. Trivialnejši primer so naslovi kupcev v podatkovni bazi, za katere lahko in moramo pričakovati, da čez nekaj let ne bodo več točni. Razkrojevanje podatkov torej zahteva odgovor na vsaj dve vprašanji: katere podatke ter kako pogosto jih je potrebno osveževati in ali je potrebno ohranjati zgodovinske zapise ter kako.

**Dimenzija 12: Namenskost podatkov**

Vsebinska uporabnost podatkov ni samoumevna niti če pri zajemu in obdelavi podatkov sodelujejo vrhunski strokovnjaki z dotičnega področja. Npr. podatki na nekem računu so lahko veljavni, točni in v kakovostni obliki, vendar morajo biti na računu vsi podatki, ki so tudi eksplicitno namenjeni temu, da se izpišejo na računu, ki ne ovirajo procesa knjiženja računa in tako naprej.

**Dodatni dimenziji kakovosti podatkov**

Z razširjanjem svetovnega internetnega omrežja in njegovim globokim ter daljnosežnim vplivom na poslovanje podjetij pa se odpirajo nove kategorije kakovosti podatkov, ki jih v tem magistrskem delu želim predstaviti in izpostaviti. Praktično neoviran pretok podatkov ima lahko nepredstavljivo velike posledice na področju zasebnosti in varnosti uporabnikov. Gre torej za neposreden vpliv na ljudi v resničnem svetu. Zato je podan predlog, da se v zgornjo strukturo dimenzij doda vsaj še dva posebej izpostavljena pogleda na kakovost podatkov (prikazana v Tabeli 4) – to sta varnost in zasebnost.

*Tabela 4: Dodatni dimenziji kakovosti podatkov*

| Št. | Dimenzija | Opis dimenzije |
|---|---|---|
| **13.** | Varnost | Mera, ki vrednoti celostno varovanje podatkov, tako na nivoju posamičnega zapisa kot na nivoju baze podatkov. |
| **14.** | Zasebnost | Mera, ki vrednoti zagotavljanje zasebnosti ob uporabi preučevane podatkovne zbirke. |

**Dimenzija 13: Varnost**

Varnost lahko opredelimo kot "izključevanje nevarnosti" (Vidmar, 2002, str. 501). Računalniški sistemi morajo zagotavljati dostopnost podatkov in razpoložljivost virov (angl.

*resources*) za shranjevanje, obdelavo in prikazovanje podatkov. Bistvenega pomena pri tem je, da so podatki in viri varni pred nenamernimi poškodbami (npr. okvara diskov) ter nepooblaščeno ali nedovoljeno uporabo (Petković & Jonker, 2007, str. 5-6). S tem problemom se računalniški sistemi soočajo vse od svojega obstoja. Obstaja veliko mehanizmov za zagotavljanje varnosti podatkov – ena ključnih komponent pri tem je nadzorovan dostop do podatkov (angl. *access control*). Na primer, operacijski sistemi nadzorujejo dostop do datotek in map, SUPB skrbijo za nadzorovan dostop do tabel in pogledov znotraj podatkovnih baz itd. Dimenzija varnosti je v sodobni, informacijski družbi zelo pomembna, zato bo podrobneje predstavljena v poglavju 2.3.

**Dimenzija 14: Zasebnost**

Količina osebnih podatkov, ki je v zadnjih letih postala dostopna na internetu, zahteva posebno pozornost pri obravnavi kakovosti podatkov. Vpeta je med dimenzijami podatkovnih struktur in pravil, dostopnosti, predstavitve, namenskosti in varnosti. Področje zasebnosti dviguje nemalo prahu s pojavom t.i. socialnih mrež na internetu, ljudje pa se v grobem delijo na tiste, ki jim za zasebnost ni veliko mar, ter na goreče zagovornike zasebnosti – vmesnih mnenj praktično ni mogoče imeti, saj npr. svoje osebne podatke razkriješ ali pa jih ne razkriješ. Omejitve pri tem so večinoma omejene na programska orodja, ki so v splošnem dobičkonosnejša, če je zasebnosti manj, zato je npr. na velikanu socialnih mrež, Facebooku, izjemno težko omejiti prikazovanje svojih osebnih podatkov: tudi če v nekem trenutku uporabnik nastavi želen nivo razkrivanja osebnih podatkov, se le-ta lahko čez nekaj časa napovedano ali tudi nenapovedano spremeni, brez vednosti uporabnika in brez vpliva uporabnika na to spremembo. A zasebnost je temeljna človekova pravica, v Sloveniji zagotovljena v ustavi (Ustava Republike Slovenije, 1991, člen 35).

Zasebnost s stališča kakovosti podatkov sproža nekatere zanimive probleme, ki se jih v praksi pogosto ignorira:

- kako shranjevati osebne podatke, da ne bodo vidni vzdrževalcem programske in strojne opreme,
- na kakšen način omejiti podatkovno rudarjenje, orodja poslovnega obveščanja in podobna orodja, da bo osebni podatki ne bodo razkriti,
- upravljanje z digitalnimi identitetami,
- opredeliti, preučiti in omogočiti anonimno uporabo podatkov ter mnoge druge.

Zasebnost in nadzor sta pogosto nasprotujoča pojma: večja zasebnost zahteva manj vpogledov, manj nadzora nad podatki – povečevanje nadzora pa, nasprotno, največkrat pomeni izvajanje analiz podatkov in povezovanje podatkov s konkretnimi osebami. Kovačič

(2006, str. 22) pojem zasebnosti in nadzora nazorno postavi v današnji čas: "Pravzaprav gre celo pri pravici do zasebnosti za neko obliko “samonadzora” – nadzora nad informacijami o sebi, nad svojimi avtonomnimi odločitvami in nad svojo osebnostjo. Čeprav je nadzor še vedno pogosto razumljen kot nekaj negativnega, nekaj, kar prisiljuje, pa je nadzorovanje danes dobilo bolj prijazen, lahko bi rekli celo hinavski obraz. Postalo je neopazno, a povsod navzoče, postalo je prijazno in prostovoljno (npr. potrošniški nadzor po različnih karticah ugodnosti), predvsem pa je postalo tako rekoč nujno za življenje v sodobni družbi. Slogan iz Orwellovega romana "Veliki brat te opazuje! / Big Brother is watching you!" se spreminja v "Veliki brat skrbi zate / Big Brother is waching out for you" (Whitaker, 1999, str. 142).

Zaradi pomembnosti in aktualnosti bosta varnost in zasebnost podrobneje predstavljeni v ločenih poglavjih v nadaljevanju. Opisane dimenzije je moč predstaviti tudi v bolj strukturirani obliki in tako lažje razumeti njihovo povezanost (Ryu, 2006, str. 192). Pri tem je moč izhajati iz dejstva, da nekatere dimenzije predstavljajo neposredne, lažje merljive in od podatkovne zbirke neodvisne kategorije kakovosti podatkov (točnost, pokritje, pravočasnost ...), kar lahko predstavimo s "pogledom v globino". Odvisne kategorije kakovosti so širše vpete v podjetje, so skupne več bazam podatkov ali pa so povezane med več strežniki tako znotraj kot tudi zunaj podjetja (ažurnost, dostopnost, predstavitev ...), kar je analogno "pogledu v širino".

Dimenzije kakovosti podatkov lahko dalje razvrstimo v tri skupine: kakovost vrednosti podatkov, kakovost podatkovnih storitev in kakovost standardnih metapodatkov. Takšna kategorizacija omogoča visok abstraktni pregled nad kakovostjo podatkov, zato je uporabna pri uporabi kazalnikov kakovosti podatkov (angl. *dashboard*), kot prikazuje Slika 3. Zaradi učinkovite predstavitve bodo te tri skupine podrobneje predstavljene v nadaljevanju. Omeniti velja tudi to, da se analiza kakovosti podatkov lahko izvaja tudi samo za eno od teh skupin, odvisno pač od konkretnega projekta.

*Slika 3: Predstavitev kakovosti podatkov na abstraktnem nivoju*

### 2.2.2 Kakovost vrednosti podatkov

Kakovost vrednosti podatkov je kategorija, ki je večinoma omejena na točno določeno zbirko podatkov, pogosto celo na točno določeno tabelo ali stolpec v tabeli. Gre za načelno neodvisnost teh dimenzij kakovosti podatkov med posameznimi množicami preučevanih podatkov – to je torej analogno pogledu v globino. Običajno je te dimenzije lažje vrednotiti, za kar je na voljo tudi več programskih orodij, pogosto temelječih na zapletenih analitičnih in statističnih izračunih.

Kakovost vrednosti podatkov zajema naslednje dimenzije, opisane v tabeli 3 in 4:

- temeljna integriteta podatkov (dimenzija 2),
- podvajanje podatkov (dimenzija 3),
- točnost podatkov (dimenzija 4),
- podatkovno pokritje (dimenzija 8),
- razkrojevanje podatkov (dimenzija 11) in
- zasebnost (dimenzija 14).

### 2.2.3 Kakovost podatkovnih storitev

Kakovost podatkovnih storitev je analogna pogledu v širino. Gre torej za vprašanje kakovosti povezav med posameznimi podatkovnimi tabelami in zbirkami ter pregled nad podatki tudi z vidika uporabnikov, ki prihajajo iz različnih okolij. Pod kakovost podatkovnih storitev se uvrščajo naslednje dimenzije:

- skladnost podatkov (dimenzija 5),
- ažurnost in dostopnost podatkov (dimenzija 6),
- enostavnost uporabe podatkov (dimenzija 7),
- predstavitev podatkov (dimenzija 9),
- razumevanje, pomembnost in zaupanje podatkov (dimenzija 10),
- namenskost podatkov (dimenzija 12) ter
- varnost (dimenzija 13).

### 2.2.4 Kakovost standardnih metapodatkov

S kategorijama kakovosti vrednosti podatkov in podatkovnih storitev je sama kakovost podatkov že razmeroma dobro opisana, do celostne opredelitve manjka le še popis sistema, ki mora biti dovolj podroben in natančen, da lahko deluje kot "lepilo", ki povezuje posamezne

elemente kakovosti podatkov. S temeljno integriteto podatkov npr. zagotovimo "dobre" podatke, vendar morajo biti le-ti primerno medsebojno povezani. Podatkovni model je torej temelj kakovosti podatkov, saj določa kaj in kako se v podatkovno bazo shranjuje (Simsion & Witt, 2005). Za podatkovni model pa je prav tako pomembno, da je primerno dokumentiran (dovolj podrobno, razumljivo in na standarden način), kar je namen meta-podatkovnega modela. Metapodatkovni model je potemtakem bistven za kakovosten podatkovni model.

Metapodatke lahko glede na nivo obravnave razvrstimo v tri kategorije (Ryu, 2006, str. 192):

- metapodatki logičnega modela:
  gre za katalog poslovnih objektov, ki so opredeljeni na nivoju celotnega podjetja, npr. opisi poslovnih entitet in njihovih lastnosti;
- metapodatki fizičnega modela:
  na fizičnem nivoju gre za opredelitev fizičnih objektov v sistemih za upravljanje podatkovnih baz, npr. imena in načela poimenovanja baz, tabel, stolpcev;
- metapodatki preslikave iz logičnega v fizični model:
  ker gre pri uporabi elektronskih podatkovnih baz za čimboljše posnemanje resničnega sveta (modeliranega z logičnim modelom) v računalniškem okolju (ki je modeliran s fizičnim modelom), mora biti ta preslikava opredeljena, npr. z določitvijo virov, transformacij, pravil in ponorov.

Dualnost metapodatkov v smislu tehnološkega in poslovnega vidika je najbolj očitna pri metapodatkih preslikave iz logičnega v fizični model. Po eni strani izhajamo iz poslovnih pravil in zakonitosti, ki jih opišemo s semantičnimi (poslovnimi) metapodatki. Le-ti se morajo odraziti v ustreznih podatkovnih strukturah, povezavah in omejitvah, kar opišemo s tehnološkimi metapodatki. Semantični metapodatki pa niso nujno bijektivno preslikani v tehnološke, saj imajo semantični metapodatki lahko zelo različne posledice: bodisi vplivajo na podatkovni model, bodisi na aplikacijsko arhitekturo, bodisi predstavljajo le vsebinske smernice itd.

Kakovost standardnih metapodatkov zajema torej naslednje dimenzije:

- podatkovne strukture in pravila (dimenzija 1),
- varnost (dimenzija 13) in
- zasebnost (dimenzija 14).

## 2.3 Varnost podatkov

Do varnosti podatkov imamo pogosto podoben odnos kot do sklepanja zavarovanja ali do pisanja oporoke: to je nekaj, kar vemo, da moramo storiti, a s tem pogosto odlašamo. Ta opravila zahtevajo, da premislimo mnoge neprijetne scenarije, ki so lahko zelo zapleteni in imajo dolgoročne posledice. Verjetnost nevarnosti v neposredni prihodnosti večinoma ni visoka, zato se razmišljanjem o tem velikokrat izogibamo in jih prestavljamo na kasneje. Kljub vsemu vemo, da bomo mirneje spali, če bo to za nami, saj bomo v primeru katastrofe zavarovani oziroma bomo imeli že pripravljen postopek za reševanje.

Podatki v napačnih rokah so lahko zelo nevarni. Ni tako velik problem, če kdo ukrade strežnik, da so le zbirke podatkov nedotaknjene. Skrajni primeri, o katerem se trenutno veliko piše, so podatki vojske ali policije v rokah terorističnih organizacij (Wang, Allen, Harris & Madnick, 2003). Vse vojaške operacije, oprema in zmogljivosti so natanko popisane v podatkovnih bazah. Tak podatek je za vojaškega nasprotnika lahko ključ do zmage. Iz tega razloga je vse več pobud za nadzor nad internetnim prometom. Nekatere države želijo nadzorovati celoten tok podatkov po internetu. To ima poleg dobrih lastnosti, kot npr. onemogočanje delovanja terorističnih organizacij, tudi velik vpliv na kakovost življenja vseh ljudi, od katerih je velika večina povsem nedolžnih. Potrebno je omeniti, da je nedotakljivost zasebne pošte v Sloveniji zagotovljena z ustavno pravico. Pošta, ki jo pošiljamo preko interneta, se vsebinsko v ničemer ne razlikuje od papirne pošte in mora biti zato obravnavana na enak način (Informacijski pooblaščenec Republike Slovenije, 2009). Obstajajo torej dobri razlogi za nadzor nad spletnim prometom in vsebino podatkov, vendar mora biti tak nadzor vedno utemeljen, formaliziran, transparenten, podrejen zakonodaji, upoštevati mora načelo sorazmernosti (nadzorovanje vse povprek je nepotrebno, drago in tudi nevarno) in predvsem mora spoštovati temeljne človekove pravice.

Visoka raven dostopnosti podatkov vedno in povsod poleg velikega števila prednosti prinaša tudi razloge za skrb. Podatki niso več zavarovani za železnimi vrati, podatkovni strežniki niso več izolirani od zunanjega sveta. Za dostop do podatkov ni več potreben neposreden priklop na fizično varovan strežnik ali imeti vstop v lokalno računalniško omrežje podjetja. Nasprotno, podatki so razpršeni po celem svetu, po osebnih računalnikih, internetnih strežnikih, nekako visijo v zraku, obstaja mnogo kopij, pretok podatkov je hiter, poceni in skrajno neobvladljiv. V neredkih primerih pa je potrebno podatke nadzorovati, zato se uporablja raznovrstne načine nadzorovanja dostopov, preverjanja verodostojnosti in veljavnosti, ki bodo v magistrskem delu tudi podrobneje predstavljeni Med najpogosteje omenjanimi so: fizično varovanje podatkovnih strežnikov, omejitev dostopov za posamezne

uporabnike na nivoju datotečnih sistemov (angl. *file system*) in operacijskih sistemov, dodeljevanje uporabnike v hierarhične skupine z različnimi pravicami (angl. RBAC, *role-based access control*), omejevanje konteksta izvajanja operacij na računalniku (na nivoju programa, procesa ali posamezne izvajalne niti). Manj pozornosti se namenja varnosti neposredno na podatkovnih strukturah –datoteke, zapisane v razširljivem označevalnem jeziku (angl. *extensible markup language, XML*), so vse bolj prisotne pri izmenjavi podatkov znotraj in tudi zunaj posameznega podjetja in so zelo primerne za omejevanje vpogleda v dele podatkov. Poleg tega so XML datoteke nadvse priročne za vgradnjo mehanizmov za preverjanje verodostojnosti in integritete podatkov, tu velja omeniti sporočila, napisana po standardu za spletne storitve (angl. *Simple Object Access Protocol, SOAP*), in digitalne certifikate.

Vse večja uporaba informacijskih tehnologij je v podjetjih potiho razširila funkcijo nadzora iz le nekaj ključnih vodstvenih zaposlenih tudi na tiste osebe, ki skrbijo za tehnološko infrastrukturo (Petković & Jonker, 2007, str. 89). Skrbniki podatkov so lahko npr. pomočniki v oddelkih za trženje, ki urejajo rezultate vprašalnikov, anket in nagradnih iger. Pomembno je, da so te osebe vredne zaupanja, da podatkov ne bodo zlorabile. Skrbniki podatkov so sicer največkrat tehniki, skrbniki podatkovnih baz (angl. *database administrator, DBA*), ki zaradi narave njihovega dela obvladujejo podatke v njihovem celotnem življenjskem ciklu. Kdo ali kaj jim preprečuje, da ne bi zaupnih poslovnih podatkov podjetja razkrili konkurenčnemu podjetju ali pa za svoje zasebne namene? Dostope do podatkov imajo tudi sistemski administratorji (angl.*system administrator, SA*), ki vzdržujejo baze podatkov na še globljem nivoju, na nivoju strojne opreme, strežnikov. Ti imajo dostop do podatkov tudi v fizični obliki, na diskovnih poljih. Nenazadnje je presoja zaupanja potrebna še pri vzdrževalcih in razvijalcih programske opreme. Programerji razvijajo programe običajno na nekih testnih podatkih, ki so pogosto neprečiščeni in zaradi tega tudi zaupne narave. Vzdrževalci programov rešujejo programske, pa tudi vsebinske/podatkovne, napake na "živih" poslovno-informacijskih sistemih, kar pomeni, da imajo vpogled v vsaj del podatkovnih zbirk podjetja. Zaupanje je torej bistven pojem, čigar napačna presoja ima lahko velike posledice kar se tiče kakovosti podatkov. V času interneta in izredno povečane pretočnosti podatkov je zaupanje še toliko pomembnejše.

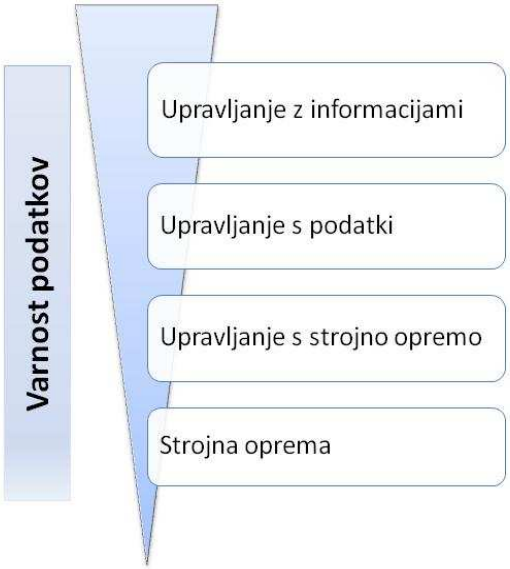
V primeru varnosti podatkov izhajamo iz že večkrat omenjenega dejstva, da obseg podatkov v elektronskih podatkovnih bazah iz leta v leto narašča, in drugega dejstva, da so podatki ključen vir znanja in delovanja podjetja. Podatkovne baze so čedalje pomembnejše za obstoj podjetja. Iz tega sledi nujno, da je tak vir potrebno ustrezno zavarovati tako pred nenamernimi napakami kot pred nepooblaščenimi dostopi. Prav to predstavlja svojevrsten izziv, saj v času informacijske družbe pogosto govorimo o skorajda neomejenemu prenosu

podatkov po spletnih mrežah in o neuničljivosti podatkov, ko se ti enkrat objavijo na internetu (povzeto po Petrocelli, 2005, str. 2).

### 2.3.1 Varnost podatkov kot temelj kakovosti

Varnost podatkov je potrebno obravnavati sistemsko in celostno, saj le tako lahko zagotovimo kakovost podatkov glede na večplastno dimenzijo varnosti. Drži, da je veriga močna največ toliko kot njen najšibkejši člen. Hierarhično lahko varnost podatkov predstavimo s Sliko 4.

*Slika 4: Hierarhičen pogled na varnost podatkov*




*Vir: T. Petrocelli, Data Protection and Information Lifecycle Management, 2005, slika 1-1.*

Najnižji nivo obravnave varnosti podatkov je **nivo strojne opreme**. To je osnova računalniških sistemov, zato na njem temeljijo vsi višji nivoji. Sestavljajo ga sistemi za shranjevanje podatkov: diskovna polja, tračne enote, NAS (angl. *Network Attached Storage*), SAN (angl. *Storage Area Network*) itd. Nevarnost na tem nivoju predstavljajo predvsem okvare strojne opreme in nepooblaščen fizičen dostop, zato je varnost podatkov zagotovljena s fizično varnostjo, kar zajema npr.: redundantnost strojne opreme, kakovostne pomnilniške medije, kakovostne diskovne kontrolerje, vso mrežno opremo, pa tudi omejen dostop do prostora, kjer so podatkovni strežniki, potresno, požarno in poplavno zavarovana soba .

Drugi nivo je **ravnanje s strojno opremo**, ki zajema obvladovanje računalniških virov, analizo omrežja in podobno. Ta nivo je z vidika varnosti podatkov manj očiten, a je pomemben člen pri celotni strategiji podatkovnih sistemov. Niti izjemno varovani diskovni strežniki in kriptirane baze podatkov ne pomagajo dosti, če so nezaščitene povezave med

perifernimi napravami (npr. tipkovnico) in računalnikom. Znan je primer prisluškovanja tipkovnicam s slabo zaščitenimi kabli na razdalji do 15 m (BBC News, 2009), ki je ne le povzročil precej razprav o varnosti podatkov na spletnih forumih, ta ranljivost je bila prikazana tudi v praksi.

Bistveno več obravnave od najnižjih dveh nivojev varnosti podatkov je namenjeno tretjemu nivoju, to je **upravljanju s podatki** (tukaj je ta pojem omejen izključno na neposredno upravljanje s podatki). Varnost podatkov se s tega stališča zagotavlja preko opravil varnostnih kopij podatkov, obnovitve podatkov iz varnostnih kopij v primerih katastrof, vzdrževanje kakovosti vrednosti podatkov, porazdeljevanje podatkov zaradi zagotavljanja visoke dostopnosti ali visoke razpoložljivosti, šifriranje podatkov in podobno. Zavarovanje pred nenamernimi napakami (npr. nepravilnim vnosom podatkov v podatkovno bazo) je na tem nivoju mogoče z ustreznim podatkovnim modelom, ki zagotavlja določeno raven podatkovne integritete in varnosti.

Fizično varnost podatkov najpogosteje predstavljajo varnostne kopije podatkov, s katerimi lahko neželene spremembe podatkov povrnemo v prvotno stanje. Najbolj neposredno varnost podatkov pred zlorabami zagotavlja omejitev dostopov do podatkov – v primerih, ko se na to ni mogoče popolnoma zanesti, se običajno uporablja šifriranje (kriptiranje) podatkov. Enkripcija podatkov se pogosto dopolnjuje z uporabo digitalnih podpisov, ki nudijo še verodostojno informacijo o lastniku podatkov in zagotavljajo, da je prejeto sporočilo res enako poslanemu (torej da ni bilo na poti spremenjeno oziroma potvorjeno).

Z vidika mrežnih plasti varnost pri upravljanju s podatki zagotavljamo z varnostnimi (požarnimi) zidovi, sestavljenih iz niza: usmerjevalnik – aplikacijski pretvornik – usmerjevalnik (Vidmar, 2002, str. 664). Filter (usmerjevalnik) nadzoruje podatkovne pakete glede na naslove izvorov in ponorov v njih. Aplikacijski pretvornik omogoča nadzor storitev omrežja (npr. podatkovni promet spletnih strani).

Najvišji nivo predstavlja **upravljanje z informacijami**, ki je za varnost podatkov pomemben zaradi osnove, ki jo informacijam predstavljajo podatki: na najvišjem nivoju morajo podatki omogočati tvorjenje informacij v uporabnih in točnih kontekstnih povezavah, primernih za poslovno uporabo. Dostop do informacij je običajno omejen, saj marsikatere informacije o podjetju niso javne, večina zaupnih je povsem nedostopna tudi večini zaposlenih v podjetju.

Iz hierarhične predstavitve varnosti podatkov je poleg več nivojev problematike razviden tudi delež (prikazan s puščico), ki je namenjen obravnavi varnosti podatkov na vsakem nivoju. Višji nivoji zahtevajo več pozornosti in prilagoditev, medtem ko najnižja nivoja običajno po

začetni nastavitvi ostaneta nespremenjena. Število sprememb oziroma prilagoditev nekega ERP sistema je nedvomno precej višje od sprememb konfiguracije strežniškega sistema.

### 2.3.2 Nadzor nad podatki

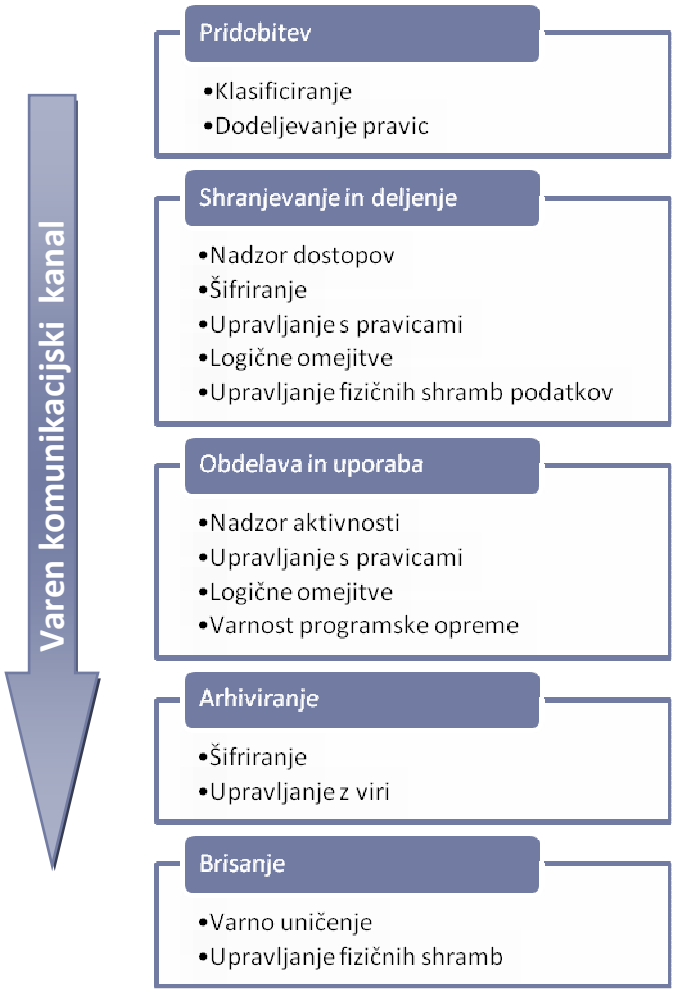
Modeli kakovosti podatkov ne morejo zagotavljati dolgoročne kakovosti podatkov, če podatke po popravkih v nekem trenutku pustijo nenadzorovane (o tem govori Dimenzija 11: razkrojevanje podatkov, opisana v poglavju 2.2.1). Iz tega sledi, da je stalen nadzor nad podatki nujen – zaradi velikih posledic, ki jih nadzor prinaša, tudi negativnih, pa ga je potrebno ustrezno obvladovati.

Nadzor je opredeljen kot usmerjen vpliv na zastavljeni cilj (Beniger, 1986, str. 7), pri čemer je pomembna povratna povezava (in primerjava) med učinki ter namenom procesa. Ločimo lahko nadzor nad ljudmi in nad stvarmi, pri čemer je moč sklepati na določene analogije. Nadzor nad ljudmi po eni strani ljudem omogoča varno življenje v družbi, saj določa meje med interakcijami ljudi, po drugi strani pa je posledica nadzora tudi večja ali, večkrat, manjša svoboda posameznika v družbi. Nadzor nad stvarmi omogoča učinkovitejše delovanje sistemov, saj s pomočjo nadzorovanja sistem ni prepuščen samemu sebi, ampak izhodi iz sistema (njegovi učinki) povratno vplivajo na delovanje (procesiranje) sistema in tako omogočajo njegovo stalno izboljševanje.

Celostni nadzor podatkov je mogoče razčleniti na več načinov, odvisno od nivojev obravnave (fizični in logični, lokalni in omrežni, konceptualni in tehnološki ...). Glede na življenjski cikel podatkov (opisanem v poglavju 1 Življenjski krog podatkov v podjetju) lahko nadzor podatkov predstavimo kot prikazuje Slika 5. Faze pridobitve podatkov, njihovega shranjevanja in deljenja, obdelave in uporabe, trajnega shranjevanja oziroma arhiviranja in na koncu morebitnega brisanja nepotrebnih podatkov, morajo potekati preko varnih komunikacijskih kanalov, ki omogočajo verodostojno in pravilno prehajanje med fazami ter potek posameznih faz.

Področje varovanja komunikacijskih kanalov je izjemno široko, saj je varnost potrebno zagotavljati na vseh mrežnih plasteh, ustrezni postopki pa so lahko zelo kompleksni. To presega okvir tega magistrskega dela, zato so predstavljeni zgolj bazični koncepti, ki problematiko le osvetlijo z namenom splošnega pregleda.

*Slika 5: Nadzor podatkov skozi njihov življenjski cikel*




*Vir: D. McGilvray, 2008, Executing data quality projects, Slika 2.4.*

**Pridobitev**: Ob nastanku oziroma pridobitvi podatkov (ki so lahko strukturirani ali nestrukturirani in v kakršni koli obliki) jih umestimo v kategorije in zavarujemo s pravicami. Običajno se ta postopek v ozadju izvede samodejno, npr. glede na tip podatkov se samodejno shranijo v ustrezne zbirke in opremijo z ustreznimi metapodatki, pravice pa se dedujejo od njihovega lastnika, kar je kasneje moč nastaviti natančneje.

**Shranjevanje in deljenje**: Shranjevanje podatkov in njihovo deljenje v uporabo že zahteva omejitev dostopov do nivoja potrebne granularnosti. Pri shranjevanju in tudi pri deljenju je možno šifriranje (kriptiranje) podatkov, ki onemogoča berljivost podatkov tudi v primerih nepooblaščenih dostopov. Deljenje podatkov sproža vprašanje zaupanja in verodostojnosti podatkov, kar lahko rešujemo z digitalno podpisanimi podatki. Dodaten nadzor podatkov so logične in fizične omejitve shranjevanja podatkov, s katerimi se zagotavlja pravilna oblika in vsebina podatkov.

V letu 2009 je na področju informatike eden najpogosteje omenjenih pojmov "računalništvo v oblakih" (angl. *cloud computing*), ki naj bi postalo temeljna tehnologija prihodnosti. Njena glavna značilnost je, da se programska oprema in podatkovne baze nahajajo "nekje na spletu", za uporabo pa uporabniki potrebujejo le spletni brskalnik. Takšno načelo zahteva določeno stopnjo zaupanja, saj so podatki podjetja zdaj shranjeni izven okvirov in nadzora tega podjetja. Varnosti podatkov "v oblakih" namreč podjetje ne more več nadzorovati samo, ampak mora to prepustiti zunanjemu partnerju, ki lahko nekoč bankrotira ali preprosto preneha ponujati storitve.

**Obdelava in uporaba**: Večina sprememb podatkov se dogaja v fazi obdelave in splošne uporabe podatkov. Te spremembe morajo biti pod nadzorom, kar je možno preko nadzora aktivnosti (angl. *activity monitoring*) ali preko preventivnih mehanizmov, kot je upravljanje s pravicami dostopov. Logične omejitve so zagotovljene s podatkovnim modelom. Nenazadnje velja posebno pozornost nameniti tudi varnosti programske opreme, ki podatke pregleduje in spreminja ter ima posreden in neposreden nadzor nad podatki. Število varnostnih popravkov, s katerimi uporabniki računalnikov redno zapolnjuje luknje v programski opremi kaže na to, da je to področje posredno izjemno pomembno tudi za kakovost podatkov.

Varnost programske opreme zajema mnoga področja informatike, nekaj pomembnejših je (Gula, 2009): nadzor nad izvajanjem programske opreme, analiziranje varnostnih lukenj in spremljanje varnostnih popravkov, pasivno spremljanje uporabe programske opreme, zaznavanje neobičajnih pojavov in obveščanje o izrednih dogodkih, sledenje spremembam, sledenje uporabnikom itn.

**Arhiviranje**: Z arhiviranjem se uporaba podatkov preneha – podatki se po potrebi preoblikujejo v najprimernejšo obliko za shranjevanje (npr. se stisnejo in šifrirajo) in prenesejo na medije za dolgoročno shranjevanje, kjer so običajno deležni standardnih postopkov za upravljanje arhiva podjetja.

**Brisanje**: Ko podatkov ne potrebujemo več, jih je potrebno izbrisati na varen način (angl. *secure shredding*), kar pomeni, da jih ni mogoče obnoviti niti s specializirano programsko opremo niti s posebnimi obdelavami fizičnih medijev. Ključnega pomena je, da so podatki izbrisani z vseh lokacij, kar dosežemo z učinkovitim upravljanjem fizičnih shramb podatkov, poleg tega je potrebno pozornost nameniti tudi vsem "vmesnim" ali začasnim shrambam podatkov, npr. indeksom podatkov, začasnim datotekam, varnostnim in arhivskim kopijam itd.

Varen komunikacijski kanal omogoča komunikacijo med izvorom in ponorom tako, da te komunikacije ne more prestreči, razumeti in potvoriti nepooblaščeni uporabnik. Izvedba varnega komunikacijskega kanala zajema ustrezne postopke na vseh arhitekturnih nivojih, v splošnem pa se uporabljajo naslednja načela (Kurose & Ross, 2009, str. 687):

- skrivanje vsebine
  - šifriranje, kjer je za razkritje vsebine potrebno imeti ustrezen dešifrirni ključ,
  - steganografija, kjer se zaupni podatki skrijejo v neko zavajajočo vsebino (npr. prikriti podatki v zvočnih datoteka),
  - zaprta omrežja, kjer so izvorne ter ponorne točke znane in zato vrednejše zaupanja,
- skrivanje izvorne in ponorne točke
  - anonimnost, pri kateri resnična imena niso znana,
  - skrivanje v množici, kjer gre za načelo, da so posamezne točke v množici bolj ali manj neopazne,
- skrivanje poteka komunikacije
  - maskiranje sporočil, ki se običajno izvaja na nižjih mrežnih plasteh, pri čemer se običajni komunikaciji dodaja lažna sporočila in tako oteži prisluškovanje.

### 2.3.3 Sledljivost dostopov in spreminjanja podatkov

Poslovni informacijski sistemi običajno ne le shranjujejo in prikazujejo podatke, ampak jih večinoma tudi spreminjajo. To na prvi pogled ni tako pomembno, saj je analiza vrste dostopov do baz podatkov potrebna večinoma le pri fizičnem načrtovanju informacijskega sistema (koliko in kakšni diski so potrebni za optimalno delovanje sistema). Vendar obstajata vsaj dva razloga, ko želimo natančno vedeti kakšne spremembe podatkov so se zgodile.

Prvi razlog je zagotavljanje ustreznosti zakonodaji in drugim predpisom ali priporočilom, ki določajo nadzorovano spreminjanje predvsem najbolj kritičnih podatkov, torej finančnih, zdravstvenih in zasebnih. V Združenih državah Amerike je najbolj odmeven zakon s tega področja Sarbanes-Oxley Act (v nadaljevanju SOX) iz leta 2002, sprejet po številnih škandalih in pretresih v velikih podjetjih, kot so Enron, Tyco International in WorldCom, ki so dodobra zamajali zaupanje v delovanje velikih podjetij in državnega nadzora. Čeprav SOX zadeva predvsem finančni nadzor nad korporacijami, njegove posledice v veliki meri nosi informacijsko-tehnološki sektor, ki mora ta nadzor zagotavljati. Druge države so kmalu sprejele podobne zakone: v Nemčiji velja Deutscher Corporate Governance Kodex, v Franciji Loi de sécurité financière itd. (Bentley & Davis, 2009, str. 13) V Sloveniji ta področja ureja predvsem Zakon o gospodarskih družbah ZGD-1 iz leta 2006 in noveli zakona ZGD-1A in ZGD-1B iz leta 2008 (Register predpisov Slovenije).

Drugi razlog, da je potrebno zagotoviti sledljivost uporabe podatkov (angl. *auditing*), je odkrivanje nenamernih sprememb podatkov, preprečevanje zlorab in zagotavljanje podatkovne integritete. Velja pravilo, da prej ko odkrijemo nepooblaščene ali neželene spremembe podatkov, manjši so stroški za vzpostavitev pravilnih podatkov. Ni malo primerov, ko nezadovoljni ali okoriščevalski delavci potvarjajo podatke v informacijskih sistemih. O tem podjetja sicer največkrat molčijo, saj se bojijo za dobro ime, na katerega bi padla temna senca, če bi se ugotovilo, da so njihovi zaposleni nezadovoljni ali udeleženi v kaznivih dejanjih. Po drugi strani se veliko piše o vdorih v informacijske sisteme od zunaj, torej s strani kriminalnih organizacij. Omeniti je potrebno, da je ta problem dvoplasten: podatke se lahko spreminja neposredno z ročnimi popravki npr. v podatkovnih tabelah, ali pa to samodejno opravi nelegalni program, kjer je človeška prisotnost lahko tudi povsem izključena.

Sledljivost uporabe podatkov nenazadnje ne pomeni le vzdrževanje dnevnika sprememb v informacijskem sistemu, saj je v nekaterih primerih smiselno analizirati tudi bralne dostope do podatkov. To je uporabno zlasti z vidika nadzorovanja dostopov do zaupnih podatkov, v manjši meri pa je koristno tudi pri optimizaciji delovanja informacijskih sistemov. Takšne funkcionalnosti omogoča več programov, omeniti velja npr. Oracle Audit Vault, ki avtomatizirano shranjuje in analizira dnevnike uporabe podatkov v varovanem repozitoriju in omogoča učinkovit pregled in obveščanje uporabnikov. Omogoča prilagodljivo sledenje spremembam glede na tip podatkovne baze (DBMS so si med seboj zelo različni in ker auditing poteka na nizkem nivoju sistema, to lahko predstavlja velik izziv za razvijalce teh orodij). Zajeti podatki se nato v kriptirani obliki pošiljajo v centralno skladišče, katero podpira učinkovit sistem poročanja in obveščanja. Orodje je namenjeno predvsem velikim podjetjem, poleg zagotavljanja visoke stopnje varnosti in podatkovne integritete pa po zaslugi avtomatiziranih poročil lahko tudi zmanjša stroške revizij. Velja pravilo, da vsakršne meritve sistemov nujno posegajo v sam sistem, kar pri sledljivost podatkov lahko predstavlja dodatno obremenitev strežnikov (vklop in analiziranje sistemskih kazalcev, sistemskih dnevnikov itd.) in podatkovnih baz (postavljanje novih sprožilcev itd.). Obstaja več načinov, kako se je takšnim težavam moč izogniti. Primer je orodje Apex SQL, ki deluje le na baznih dnevnikih (angl. *transaction log file*) in ne na podatkovnih datotekah (angl. *data file*), torej analizira dostope do podatkovne baze na mestu, ki je nižje v hierarhiji sistema DBMS in ga uporabniki ne uporabljajo neposredno. Pri sledljivosti uporabe podatkov običajno torej zbiramo podatke o tem kdo, kdaj, kakšen in iz katerega računalnika je bil dostop do podatkovne baze narejen.

Računalniški sistemi so postali kompleksna mreža najrazličnejših programov, procesov in zbirk podatkov. Pogosto je težko ugotoviti kaj delajo posamezni deli sistema, kdo jih uporablja in kdaj. Sledljivost podatkom služi mnogim koristnim namenom, kot je objavljeno na spletni strani organizacije Information Systems Audit and Control Association (v nadaljevanju ISACA) (Nair, 2007):

- optimiziranje sistemov s stališča preglednosti in zmogljivosti,
- zaznavanje nevarnih dogodkov, kršenje pravil uporabe podatkov, nelegalna dogajanja,
- izvajanje notranje in zunanje revizije sistema ali podjetja,
- izboljševanje kakovosti podatkov,
- časovna analiza uporabe podatkov ...

Iz analize sledljivosti podatkom je s pomočjo statističnih in drugih metod moč razbrati neobičajne ali neželene dostope do podatkov. V nekaterih okoljih se izvajanje takšnih analiz priporoča zelo pogosto, kar na svoji spletni strani priporoča ISACA (Nair, 2007), o čemer obstaja množica standardov in priporočil (CobiT, PCI DSS, HIPAA, CMS ARS ... ), ki zelo natančno določajo:

- pogostost pregledovanja dnevnikov dostopov,
- zapadlost časovno občutljivih podatkov,
- obravnavanje privilegiranih dostopov (npr. dostopi vzdrževalcev in skrbnikov informacijskih sistemov),
- varnost osebnih podatkov v podatkovnih bazah in
- postopke obveščanja.

### 2.3.4 Varovanje zasebnosti

Zasebnost ni nova skovanka, temveč ima dolgo zgodovino. Kovačič (2006, str. 11) pravi takole: " Različne študije kažejo, da je zasebnost nekaj, kar je medkulturno in medvrstno univerzalno in ni značilno samo za človeka. Tako nekatere navedbe v bibliji (npr. zavedanje Adama in Eve, da sta gola, zavedanje o goloti Noeta in povezovanje golote s sramoto itd., pa tudi v koranu, judovski tradiciji, antični Grčiji in starodavni Kitajski kažejo na to, da zasebno sfero poznajo različne kulture in družbe."

Vseprisotnost računalnikov je vzrok za praktično nenehno nadzorovanje ljudi. To ima globoke posledice za družbo in svobodo. Korporacije in državni represivni organi izkoriščajo vse tehnološke možnosti za nadzor ljudi, zato je za ohranjanje svobode in zasebnosti posameznikov potrebno razumeti in analizirati trende na tem področju. Nadzorovanje se je v

zadnjih letih preoblikovalo iz osredotočenega na splošnega. Če se je prej govorilo o npr. prestrezanju pošte točno določenega osumljenca, je zdaj vse več govora o "preventivnem" in "vsesplošnem" prestrezanju pošte, tudi elektronske. V preteklosti se je prisluškovalo skrbno izbranim osebam, zdaj pa lahko predvidevamo, da gre večina vseh telefonskih pogovorov skozi obdelave, ki samodejno analizirajo govor in iščejo ključne besede, glede na katere bi lahko sklepali npr. na komunikacijo dveh teroristov. Drugo vrsto grožnje zasebnosti predstavljajo elektronske sledi, ki jih uporabniki računalnikov (in še posebej uporabniki interneta) puščajo za seboj z vsakim klikom miške. Zavedati se je potrebno, da smo včasih izbor hotela za počitnikovanje izbrali na podlagi ustnih priporočil prijateljev ali glede na lastno izkušnjo, plačilo nočitev pa potem izvršili z osebno predajo gotovine – ta proces je bil praktično popolnoma anonimen. V dobi interneta bi izbiro hotela opravili z intenzivnim brskanjem po velikem številu spletnih strani, na vsaki bi pustili svojo sled, da smo jo obiskali. Plačilo nočitev bi nato izvedli elektronsko tako, da bi izbranemu hotelu posredovali svoje osebne podatke in številko kreditne kartice.

Zasebni podatki so dragoceni, ne le za posameznega človeka, pač pa tudi za policijo in tržnike. Schneier na svoji spletni strani opisuje, da je oglaševanje namreč bistveno učinkovitejše, če je ciljna publika znana in dosegljiva. Trend zmanjševanja zasebnosti na tako veliko področjih je skrb zbujajoč in zato je nujno, da se tega zavedamo in smo previdni pri tem, koliko osebnih podatkov je res potrebno deliti s celim svetom (Schneier, 2006).

#### 2.3.4.1 Zasebnost kot nova dimenzija kakovosti podatkov

S stališča kakovosti podatkov v podatkovnih bazah je zasebnost deležna manj pozornosti kot nekatere druge dimenzije. Če varnost opredelimo kot izključevanje nevarnosti, se je potrebno vprašati, ali je bolj nevarno, da s pridobljenimi podatki na široko pokrijemo svet obravnave, ali pa je morda bolj nevarno (po nepotrebnem) hraniti osebne podatke ljudi, za katere obstaja možnost zlorab? Zasebnost ima nekaj globokih in daljnosežnih vplivov na proces uporabe in upravljanja podatkov. Potrebno je odgovoriti na nekaj pomembnih vprašanj. Kako izvajati rudarjenje po podatkih tako, da zasebnost ne bo ogrožena? Kako ohranjati zasebnost pri statističnih obdelavah? Kako iskati, če sploh, po šifriranih dokumentih? Kakšen podatkovni model uporabiti, da bo del podatkov šifriran, del pa ne? In, nenazadnje, kako preprečiti odliv zaupnih in zasebnih podatkov iz podjetij?

Podatkovno rudarjenje (angl. *data mining*) s prepletanjem podatkovnih baz, umetne inteligence in statistike odpira možnosti učinkovitejšega črpanja informacij iz podatkov, zato je včasih poimenovano tudi kot "odkrivanje znanja" (angl. *knowledge discovery*). V nasprotju z običajnimi statističnimi metodami, gre pri podatkovnem rudarjenju za pregledovanje

podatkov in iskanje zanimivih pravil, vzorcev ter informacij s pomočjo strojnega učenja, ne da bi pred tem postavljali in nato preverjali določene hipoteze.

Mnogo strokovnjakov s tega področja se strinja, da bo zasebnost postajala vse večji izziv pri podatkovnem rudarjenju (Petković & Jonker, 2007). Kot že rečeno, se vse več osebnih podatkov ljudi zbira v podatkovnih bazah, nad katerimi se potem izvajajo postopki poslovnega obveščanja, podatkovnega rudarjenja, statistične analize in podobno. Rezultati teh postopkov so nadalje tudi tržno zanimivi in pravzaprav ni sledljivosti, kam se le-ti predajo ali prodajo. Nov, svojevrsten izziv glede zasebnosti predstavlja po drugi strani tudi globalizacija. Za podjetja, ki imajo poslovalnice v več državah, je značilno, da želijo imeti skupen pregled nad vsemi enotami, vendar podrobnosti o poslovanju posamezne enote lahko vidi le ta enota in matično podjetje. Drugim poslovalnicam so podrobnejši pregledi onemogočeni. Takšni in podobni primeri se običajno rešujejo z omejitvami pravic nad podatki, vendar v vseh primerih to ni mogoče iz povsem tehničnih ali administracijskih razlogov. Ta problem se rešuje s posebnimi obdelavami podatkov (angl. *privacy preserved data mining*), ki z minimalno režijo zagotavljajo, da s podatkovnim rudarjenjem vsi udeleženi pridobijo novo znanje, hkrati pa ohranjajo potreben nivo zasebnosti, torej brez razkritja osebnih podatkov. Teoretični postopki se že nekaj let izboljšujejo (Lindell & Pinkas, 2002, str. 1-2), a zaenkrat še ni bilo veliko prenosov v prakso, a glede na aktualnost področja lahko sklepamo, da bo tega vse več.

#### 2.3.4.2 Zasebnost in spletna socialna omrežja

Milijoni uporabnikov vsak mesec na internetu objavljajo dogodke iz svojega vsakdanjika v obliki člankov, fotografij in video prispevkov in jih delijo z drugimi. Vsako od spletnih socialnih omrežij omrežij je specializirano za neko področje: LinkedIn stremi k ohranjanju in grajenju poslovnih stikov, Friends Reunited omogoča ohranjanje vezi z bivšimi sošolci, Twitter je namenjen hitremu sporočanju in tako dalje.

Nedolgo nazaj so bili uporabniki spletnih socialnih omrežij večinoma povsem brezbrižni do varovanja svoje zasebnost, a zdaj, ko je število zlorab doseglo kritično mejo, je vse več govora o omejevanju dostopov do osebnih podatkov (npr. omejitev, da lahko nekatere objavljene fotografije vidijo le izbrani uporabniki). Zdi se, da se uporabniki socialnih omrežij samoorganizirajo in sami skrbijo za ohranjanje svoje zasebnosti, pravnih nastavkov zanje pa (še) ni ali pa so preohlapni in težko prenosljivi na prakso spletnih omrežij.

Na tveganja in težave varnosti in zasebnosti podatkov v socialnih mrežah med drugimi opozarja tudi Council of European Professional Informatics Societies (v nadaljevanju

CEPIS). CEPIS je neprofitno združenje 33 evropskih držav, ki si prizadeva za izboljšanje vpliva, ki ga ima informatika na zaposlovanje, poslovanje in družbo. V svoji izjavi (CEPIS, 2008) opozarjajo na velik porast objav osebnih podatkov na internetu, ki ima dve veliki posledici: množica teh podatkov je zaslepljujoče privlačna za oglaševalska podjetja, ki dobijo ciljno publiko takorekoč "na dlani"; druga posledica pa je tveganje za ta ista podjetja (in, seveda, tudi druga na splošno), saj ta omrežja postajajo integralni del poslovanja, nad katerim ima podjetje razmeroma malo nadzora. Pri tem niso mišljena le tipična spletna socialna omrežja in spletne storitve za ohranjanje stikov, temveč tudi spletne strani, ki ponujajo deljenje slikovnega ter video gradiva, igre in interaktivni virtualni svetovi. V spletnih igrah gre za neposredne interakcije med ljudmi, s tem da se osebe predstavljajo s svojimi virtualnimi alter egi, ki lahko počno praktično kar se jim zahoče – tudi napadajo druge ljudi, pa čeprav le virtualno, in s tem sprožajo vprašanja, ali je v spletni igri možno izvesti kriminalno dejanje.

# 3 MODELI ZA ZAGOTAVLJANJE KAKOVOSTI PODATKOV

Namen modelov za zagotavljanje kakovosti podatkov je vzpostaviti sistematičen način za izboljšanje in ohranjanje kakovosti podatkov, ne glede na to, v kakšni podatkovni zbirki se podatki nahajajo. Prav zato morajo biti modeli dovolj splošni, da jih je možno uporabiti na kar se da raznolikih primerih. Za vse modele velja temeljno pravilo vstopajočih in izstopajočih objektov iz sistema (v našem primeru so objekti kar podatki) GIGO: "*Garbage In – Garbage Out*", ki poudarja bistven pomen kakovostnih podatkov že na vhodu v informacijski sistem.

V majhnih podjetjih oziroma tam, kjer se uporablja razmeroma majhne in preproste zbirke podatkov, se kakovost podatkov lahko neformalno zagotavlja z neko preprosto, enkratno obdelavo. V kompleksnejših okoljih pa je potrebno pripraviti sistematičen načrt izvedbe projekta za zagotavljanje kakovosti podatkov, ki primerno razdeli delo na vse sodelujoče strokovnjake. A enkratna vzpostavitev kakovostnih podatkov sama po sebi še ne zagotavlja, da bo kakovost podatkov ostala na visokem nivoju tudi v prihodnje. Zato je smiselno vzpostaviti pravila in mehanizem, ki bo v vsakem trenutku preprečeval nekakovost podatkov, samo kakovost pa nato periodično preverjati. Naslednja ključna točka modelov je obveščanje uporabnikov o stanju podatkov, predvsem o dogodkih, ki potrebujejo ukrepanje. Nenazadnje velja tudi, da kakovosti podatkov skorajda ni mogoče vzdrževati, če le-ta ni ustrezno dokumentirana in dobro razumljena vsem skrbnikom podatkov.

## 3.1 Epplerjev model kakovosti informacij (IQF)

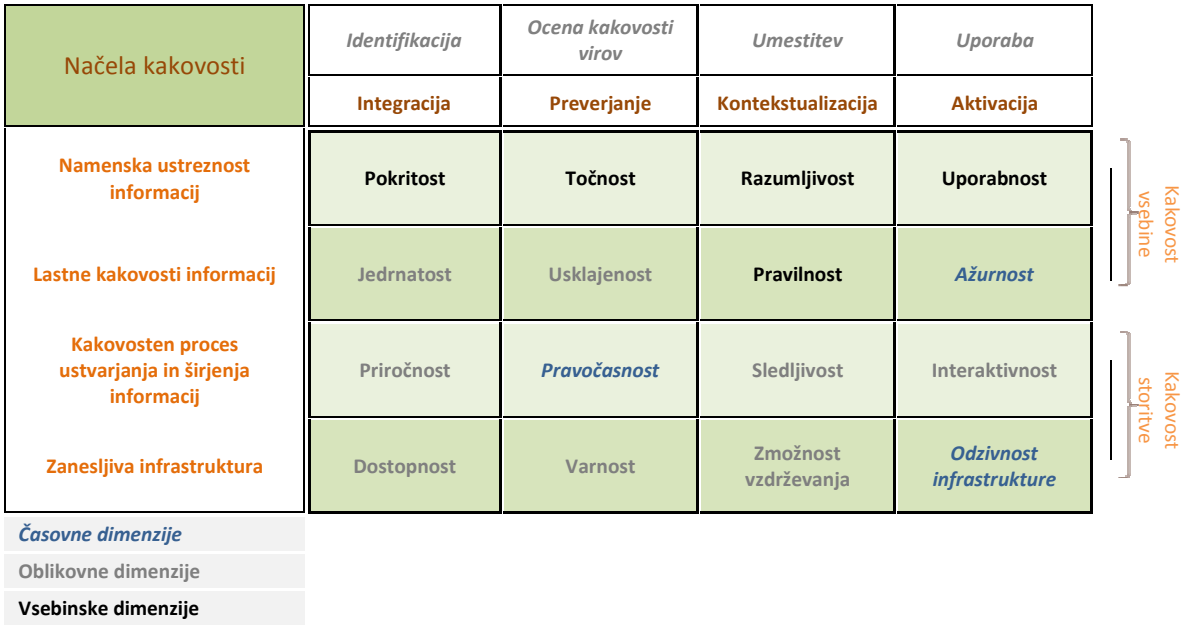
Eppler je svoj model osredotočil na informacije, saj je strogo ločil podatke od informacij (Eppler, 2003, str. 19). Pomen podatkov je skrčil na pojem "surovih, nepovezanih, kakovostnih ali količinskih dejstev". Kljub temu moramo njegov model kakovosti informacij primerjati z drugimi modeli kakovosti podatkov, saj avtorji sami večkrat omenjajo težave pri ločevanju teh izrazov: informatikom je bližji pojem podatek, poslovnežem je bližje pojem informacija, zato pojma celo zamenjujejo glede na ciljno publiko (McGilvray, 2008, English, 2003). Velik pomen, ki ga Eppler posveča informacijam, predstavlja informacije kot potencialno znanje, ki ga prejemnik lahko pridobi, če ga pravilno interpretira in poveže z obstoječim znanjem. Iz tega sledi, da kakovostni podatki olajšajo pretvorbo v informacije in nato v znanje.

Epplerjev model, ki ga je poimenoval "okvir kakovosti informacij" (v nadaljevanju IQF, angl. *Information Quality Framework*), je postavljen v kontekst procesov, intenzivnih z znanjem, saj tam potekajo aktivnosti pretvarjanja informacij in zahtevajo specializirana znanja. Procesi, ki se intenzivno ukvarjajo z znanjem, so le redko rutinski, zahtevajo veliko mero učenja in ustvarjalnosti, pri tem pa je poudarjena medosebna komunikacija in dokumentiranje informacij (Eppler, 2003, str. 22). Zato je z Epplerjevim modelom moč analizirati tri vrste znanja: upravljanje procesov oziroma znanje o procesih (angl. *know-how*), znanje, pridobljeno znotraj procesov (angl. *know-what*) in znanje, pridobljeno iz posledic procesov (angl. *know-why*).

Model IQF, ki ga predlaga Eppler, je sestavljen iz štirih glavnih elementov, predstavljenih v nadaljevanju ter prikazanih na Sliki 6:

- štirih navpičnih nivojev, ki predstavljajo sorodne skupine dimenzij (oziroma skupine kriterijev) kakovosti podatkov: namenske ustreznosti informacij, lastne (objektivne) kakovosti informacij, kakovosten proces ustvarjanja ter širjenja informacij in zanesljiva infrastruktura,
- štirih faz v življenjskem krogu informacij s stališča uporabe,
- šestnajstih dimenzij kakovosti podatkov, porazdeljenih po navpičnih nivojih ter vodoravnih korakih in
- štirih skupin načel izboljševanja kakovosti informacij v posameznem časovnem koraku.

*Slika 6: Epplerjev model kakovosti informacij*




*Vir:M. J. Eppler, Managing Information Quality, 2003, Slika 6.*

**Nivoji obravnave kakovosti informacij** so priporočilo k izvedbi projektov kakovosti podatkov od zgoraj navzdol (angl. *top-down approach*). Najprej je potrebno analizirati potrebe uporabnikov podatkov, nato sledi opredelitev značilnosti podatkov in načina oziroma procesa pridobitve teh podatkov, šele zatem se analizira in vzpostavlja kakovosten infrastrukturni nivo.

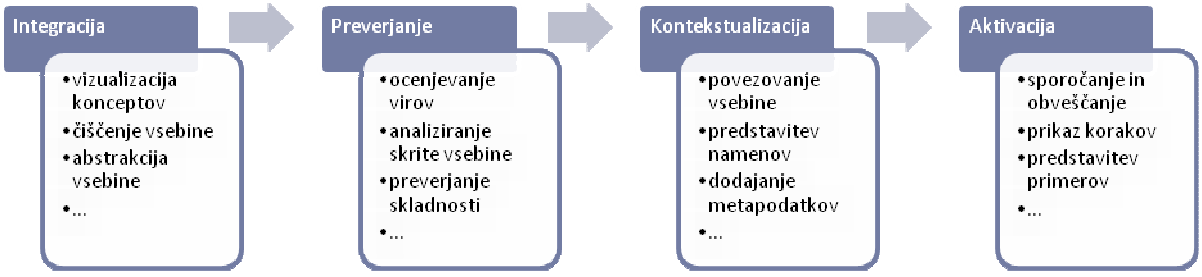
Eppler je **življenjski krog informacij** razdelil v štiri krovne faze, ki so analogne upravljanju z informacijami kot jih vidi uporabnik. Prvi korak, identifikacija, zajema umestitev v domeno iskanih informacij, iskanje ter identificiranje virov informacij in iskanje sorodnih informacij. V drugem koraku uporabnik oceni kakovost virov, v tretjem pridobljene informacije umesti v nov, aktualen kontekst. V zadnjem koraku poteka dejanska uporaba informacij za reševanje konkretnih problemov. Na tem mestu je potrebno znova omeniti bližino pojmov informacije in podatka, saj Epplerjev življenjski krog informacij lahko brez večjih prilagoditev preslikamo na življenjski krog podatkov.

**Dimenzije kakovosti**, ki jih predlaga Eppler, so preudarno skrčen nabor dimenzij mnogih avtorjev, pri čemer je Eppler dal poudarek na nepodvajanje, splošnost in uporabnost posameznih dimenzij. Rezultat je množica dimenzij kakovosti, ki je večinoma skladna z mnogimi drugimi modeli. V kontekstu magistrskega dela izpostavljam dimenziji varnosti in sledljivosti. Varnost naj zagotavlja varen, omejen dostop do informacij le pooblaščenim uporabnikom in varno shranjevanje podatkov z zaščito pred izgubo. Sledljivost, ki je manj

pogosta dimenzija, Eppler predstavlja kot zmožnost prikaza nastanka podatka oziroma informacije, kar ima posredno pozitiven vpliv na verodostojnost podatkov.

**Načela kakovosti** so zadnji element Epplerjevega modela. Nudijo praktične usmeritve za izvajanje projektov kakovosti podatkov v vsaki fazi življenjskega kroga podatkov in s tem pragmatično obogatijo model. Načela so bila odkrita z empiričnimi raziskavami in s poglobljeno študijo literature (Eppler, 2003, str. 78). Moč jih je predstaviti preko sorodnih aktivnosti, s katerimi se ta načela udejanjajo, kot prikazano na Sliki 7.

*Slika 7: Skupine aktivnosti, ki so temelj Epplerjevim načelom kakovosti*




*Vir: M. J. Eppler, Managing Information Quality, 2003, Slika 9.*

Eppler je uporabo IQF prikazal na številnih primerih (Eppler, 2003, str. 185), s čimer je utemeljil vrednost modela za namene analiziranja in izboljševanja kakovosti podatkov, poleg tega je IQF umestil v izobraževalno sfero, s čimer je zaokrožil široko uporabnost modela. Manj pozornosti pa je namenil ekonomskemu vidiku kakovosti podatkov in socialnemu vplivu, ki ga imajo hitro razširjajoči se podatki v dobi interneta.

## 3.2 Englishev model celovitega upravljanja podatkov (TDQM)

Model celovitega upravljanja informacij (v nadaljevanju TIQM, angl. *Total Information Quality Management*) se je sprva, od leta 1993, imenoval model celovitega upravljanja podatkov (v nadaljevanju TDQM, angl. *Total Data Quality Management*), ki ga je English leta 2001 preimenoval z namenom približanja modela poslušalcem iz poslovnih okolij. Zaradi skladnosti terminologije v tem magistrskem delu bodo sklici na Englishev model zapisani kot sklici na TDQM.

Eno temeljnih sporočil modela TDQM je vzpostavitev okolja in procesov, ki najprej omogočajo in po vzpostavitvi tudi ohranjajo kakovost podatkov v poslovnem sistemu. TDQM je razmeroma redek primer modela kakovosti podatkov, ki vključuje tudi stroške in tveganja nekakovostnih podatkov, kot prikazano na Sliki 8.

*Slika 8: Englishev model celovitega upravljanja podatkov (TDQM)*


*Vir: L. English, Total Information Quality Management, 2003, Slika 1.*

Kot že rečeno, nekakovostni podatki skoraj zagotovo povzročajo stroške, ki bi se jim podjetja lahko izognila. Kako veliki so ti stroški in kakšen napor zahteva dvig kakovosti podatkov, sta pomembni vprašanji, ki ju je v projektih nujno odgovoriti. Vse faze življenjskega kroga podatkov povzročajo stroške, koristi pa nastopijo šele v fazi uporabe podatkov. Ko uporabniki podatke uporabijo pri tvorjenju informacij, iz kakovostnejših podatkov izčrpajo več informacij in posledično več znanja, nekakovostni podatki pa imajo lahko ekstremno negativen učinek (McGilvray, 2008, str. 5). Aktivnosti v vseh fazah življenjskega kroga podatkov vplivajo na kakovost podatkov, a običajno (poslovne) uporabnike podatki zanimajo šele takrat, ko jih želijo uporabiti. Obravnavanje podatkov kot vir v podjetju omogoča ugotavljanje njihovih stroškov in koristi. Pri tem velja, da so podatki, za razliko od drugih virov, ponovno uporabljivi in se ne porabljajo oziroma trošijo.

Prvi korak Englishevega modela je ocena, kako dobro so podatki in informacije opredeljeni in dokumentirani ter kakšna je podatkovna arhitektura. V naslednjem koraku se izvede ocena kakovosti podatkov vseh virov (znotraj podatkovnih baz, pri zajemu podatkov ali v drugih procesih). Tretji korak zajema merjenje stroškov in tveganj, ki nastanejo kot posledica nekakovostnih podatkov (zastoji v procesih, izgubljene priložnosti, odtujitev kupcev itd.), zato je ta korak velikega pomena pri upravičevanju stroškov projektov zagotavljanja kakovosti podatkov. Naslednji, četrti korak opredeljuje čiščenje podatkov in skupaj s petim korakom, izboljšanjem procesov, daje pravo vrednost projektom in zagotavlja trajnejšo kakovost podatkov, pri čemer English priporoča uporabo cikla načrtuj-naredi-preveri-ukrepaj. Krovni proces modela pa je nenehna skrb za optimalno okolje, v katerem je

kakovost podatkov omogočena in zagotovljena že sama po sebi. Za to je potrebna primerno usmerjena kultura zaposlenih in samo delovno okolje v podjetju.

Iz zgornjega lahko rečemo, da je Epplerjev model zelo primeren za pedagoške namene, Englishev model TDQM (oziroma njegova novejša različica, TIQM) pa je za zasnovan predvsem z mislijo na uporabo v poslovnem svetu.

## 3.3 Celostni model kakovosti podatkov CDQM in vizualizacija z IP-MAP

Stremenje k izdelavi modela za zagotavljanje kakovosti podatkov je vodilo marsikaterega strokovnjaka, tako v poslovnih kot v akademskih krogih. Številčnost tako nastalih modelov sama po sebi pomeni aktualnost tematike, a se v takih primerih kmalu pojavi potreba po standardizaciji in poenoteni vizualizaciji modelov. Leta 2000 so Shankaranarayanan, Wang in Ziad predstavili tehniko IP-MAP (angl. *Information Product Map*), s katero je mogoče formalno opredeliti koncepte procesov zagotavljanja kakovosti podatkov.

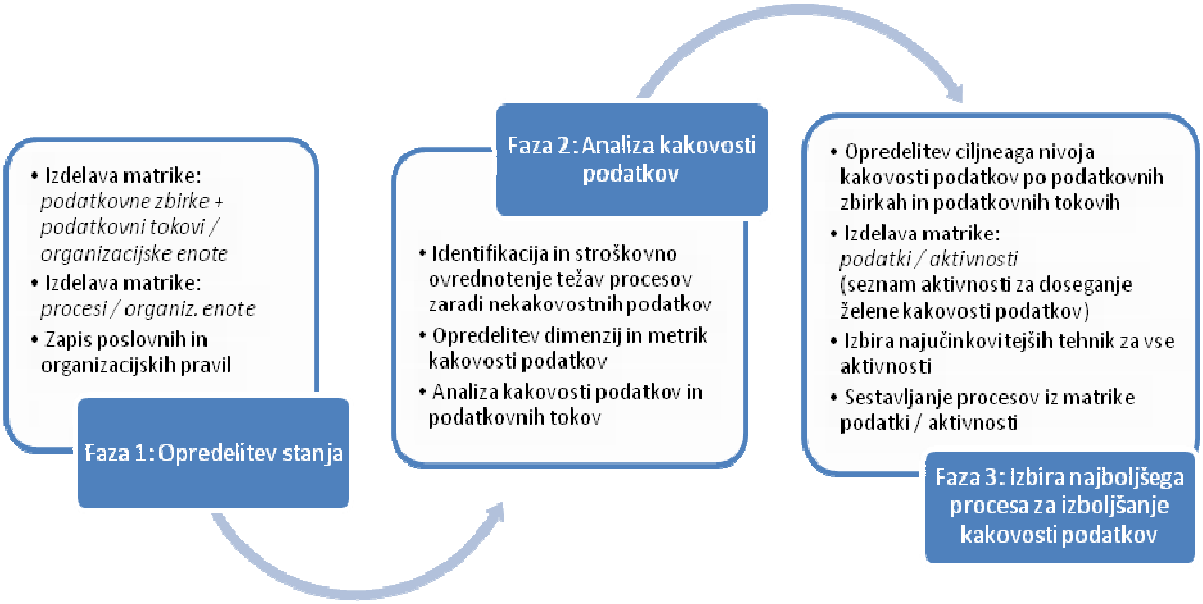
IP-MAP je razširitev sistema tvorjenja informacij (angl. *Information Manufacturing System*), ki so ga predlagali Ballou, Wang, Pazer in Tayi že leta 1998. Njegova glavna prednost pri ugotavljanju kakovosti podatkov je združevanje analize podatkov in analize procesov. Podatki so obravnavani kot poslovni učinki (angl. *Information Product*, *IP*) – procesi, ki so tvorci podatkov, so torej analizirani s stališča negativnih vplivov na kakovost pri nastajanju oziroma tvorjenju podatkov. Torej gre za analogijo modela podatkov in proizvodnje fizičnih izdelkov. Zakaj je to zanimivo, ni težko razumeti, če vemo, da je za procese v proizvodnji že zgodovinsko gledano značilno, da so deležni intenzivnih optimizacij in poglobljenih raziskav. Tako IP-MAP omogoča vpeljavo celostnega upravljanja kakovosti tudi na področje podatkov.

Na podlagi IP-MAP Batini in Scannapieca (2006) predlagata uporabo novega modela kakovosti podatkov, poimenovanega CDQM (angl. *Complete Data Quality Methodology*), pri katerem lahko IP-MAP učinkovito uporabimo. Izhodišče za CDQM je bilo osredotočanje na poslovne procese in na stroške, ki jih povzročajo nekakovostni podatki (Batini & Scannapieca, 2006, str. 89). Razdeljen je v 3 faze, kot je prikazano s Sliko 9.

V prvi fazi se analizira trenutno stanje. S pomočjo izdelave matričnih preglednic se opredeli vse pomembne povezave med organizacijskimi enotami, procesi in podatki, pri čemer velja, da se morajo neznane povezave definirati in zagotoviti. V naslednji fazi se (običajno s pomočjo intervjujev uporabnikov) identificirajo glavne težave v poslovnih procesih, ki nastajajo zaradi nekakovostnih podatkov. Nato je glede na identificirane problematične

procese potrebno izbrati ustrezne dimenzije kakovosti podatkov in ugotoviti stroške ter nove koristi izboljšanja podatkov. V zadnji fazi se poišče najučinkovitejši način za izboljšanje stanja kakovosti podatkov. Pri tem najprej za vsako podatkovno zbirko in podatkovni tok opredelimo želen nivo kakovosti oz. želeno raven znižanja stroškov. Nato je potrebno pripraviti seznam aktivnosti za doseganje tega nivoja kakovosti. Sledi pregled možnih načinov za izvedbo aktivnosti in na koncu izbira optimalnega procesa izboljšav, glede na ovrednotene stroške in koristi posameznih kandidatnih procesov.

*Slika 9: Faze modela CDQM*




*Vir: C. Batini & M. Scannapieca, Data Quality, 2003, Slika 7.15.*

## 3.4 Predlog nadgrajenega modela

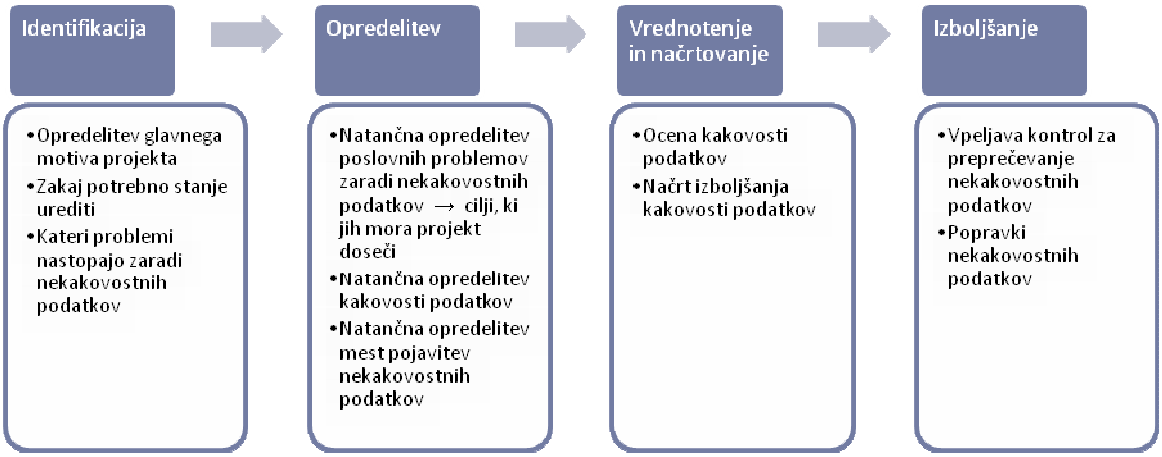
Povod za izdelavo predloga nadgrajenega modela za zagotavljanje kakovosti podatkov je predvsem v posodobitvi obstoječih modelov, kar zahtevajo spremembe poslovnih modelov zaradi globalizacije ter razmaha spletnih omrežij, in potreba po učinkoviti, praktični uporabi teh modelov v poslovnih okoljih.

V magistrskem delu so bili v predhodnih poglavjih predstavljeni trije izmed najbolj poznanih modelov za zagotavljanje kakovosti podatkov. Kot omenjeno, Epplerjev model kakovosti informacij (IQF) predstavlja dobro teoretično podlago za razumevanje tega področja, a njegova uporaba za moderne poslovne potrebe ni enostavna. Englishev model celovitega upravljanja podatkov (TDQM) je, poenostavljeno rečeno, preslikava modelov celovitega upravljanja s kakovostjo na področje kakovosti podatkov, torej je splošnejši in ne zajema nekaj specifičnih dimenzij kakovosti, ki se tičejo podatkov, npr. varnosti. Celostni model

kakovosti podatkov CDQM in vizualizacija z IP-MAP imata po drugi strani premišljeno in učinkovito uporabo s tehničnega vidika – učinkovita sta na opisnem nivoju problematike, npr. pri dokumentiranju, manj pa pri praktični uporabi.

Predlog nadgrajenega modela, predstavljenega na Sliki 10, proces vzpostavitve in zagotovitve kakovosti podatkov v določenem okolju razdeli na štiri faze. V prvi poteka identifikacija ključnih dejavnikov, ki omogoča razumevanje poslovnih težav in postavlja okvirje aktivnostim za zagotavljanje kakovosti podatkov: kaj je privedlo do potrebe po projektu kakovosti podatkov, zakaj je obstoječe nekakovostne podatke potrebno urediti in kakšni so poslovni problemi, ki se pojavljajo kot posledica nekakovostnih podatkov. Sledi druga faza, v kateri je potrebno te poslovne probleme natančno opredeliti in na tej podlagi postaviti cilje projekta. Ta faza zajema tudi natančno opredelitev kakovosti podatkov s pomočjo komunikacije z uporabniki podatkov ter izčrpno identifikacijo vseh mest, kjer se nekakovostni podatki pojavljajo in jih je potrebno urediti. V tretji fazi sta potrebna dva koraka: ovrednotenje kakovosti podatkov po izbranih dimenzijah kakovosti in izdelava načrta za izboljšanje ter zagotavljanje kakovosti podatkov v prihodnosti. Zadnja, četrta faza zajema vpeljavo tehničnih kontrol in procesnih sprememb, ki preprečujejo nastanek novih nekakovostnih podatkov, in dejanske popravke nekakovostnih podatkov z izbranimi tehnikami.

*Slika 10: Predlog nadgrajenega modela za zagotavljanje kakovosti podatkov*




Pri opredelitvi kakovosti podatkov je predlagan nabor dimenzij kakovosti, kot jih predlaga Eppler, saj je Eppler množico dimenzij izčrpno analiziral in nazorno strukturiral ter predstavil (glej poglavje 3.1). Temu naboru nadgrajeni model le dodaja dve novi, aktualni dimenziji, opisani v poglavju 2.2.1. Razdelitev po zgoraj predlaganih fazah in poudarka na ustvarjanju ustreznega okolja za zagotavljanje kakovostnih podatkov izhaja iz Englishevega

modela (glej poglavje 3.2). Predlog nadgrajenega modela poleg tega skuša faze še podrobneje razdeliti po zaporedju korakov, da se omogoči kar najpreprostejšo praktično uporabo v poslovnem svetu.

### 3.4.1 Faza 1: identifikacija ključnih dejavnikov projekta

Evidentno je, da se projekti za izboljšanje in zagotavljanje kakovosti podatkov dotaknejo podatkov v širokem obsegu in da so 100-odstotno kakovostni podatki v večini primerov praktično nedosegljivi. Zato je potrebno, da izvajalci poznajo razloge, zakaj se podjetje odloči za takšen projekt in kaj želi z njim doseči. Prvi predlagan korak takšnih projektov je torej ugotoviti, **kakšne so poslovne potrebe ali priložnosti (glavni motiv projekta)**, kar je dejansko motiv projekta, in jih lahko razdelimo na naslednje (McGilvray, 2008, str. 69):

- zmanjšan dobiček ali potopljene priložnosti (npr. potencialen kupec ni prišel v stik s podjetjem zaradi napačne telefonske številke ali napačnega e-poštnega naslova),
- izgubljen posel (npr. objava netočnih podatkov o izdelku na spletni trgovini povzroči, da je kupec odšel h konkurenci),
- nepotrebni ali previsoki stroški (npr. zaustavitev proizvodnje zaradi manjkajočih surovin, kar je bila posledica nepravilnega stanja zaloge),
- katastrofa (npr. izguba dobrega imena podjetja zaradi odtujitve podatkov kot posledice slabe varnostne politike),
- povečano tveganje (npr. nepravočasne izterjave dolga zaradi podvojenih zapisov kupcev).

V primerih, ko se s projekti želi odpraviti konkretne težave zaradi nekakovostnih podatkov, je vredno **ugotoviti, zakaj je stanje potrebno urediti**. Takšno razumevanje je koristno v nadaljevanju pri iskanju vzrokov nekakovostnih podatkov, poleg tega morajo vsi popravki podatkov v projektu podpirati doseganje poslovnih ciljev. Okvirji projekta se nato postavijo z **opisom problemov, ki nastopajo zaradi nekakovostnih podatkov.**

### 3.4.2 Faza 2: opredelitev ključnih sestavin projekta

Prvi predlagan korak v drugi fazi projekta je natančna **opredelitev poslovnih problemov zaradi nekakovostnih podatkov**, iz česar sledi **seznam ciljev**, ki jih mora projekt doseči. Projekt bo najučinkovitejši, če bo največ pozornosti namenjal reševanju najbolj škodljivim poslovnim problemom. Za to je potrebno seznam ciljev razvrstiti po prioriteti, saj bo tako dodana vrednost projekta kar največja.

V naslednjem koraku je za zagotovitev osredotočenosti na prave stvari potrebno natančno **opredeliti kakovost podatkov**, in sicer tako, da jo bosta na enak način razumela tako izvajalec kot naročnik. Ta korak je velikega pomena za uspeh projekta, saj zagotavlja, da naročnik in izvajalec "govorita isti jezik", torej da s projektom delamo prave stvari na pravi način. Potrebno je jasno razumevanje poslovnih izrazov naročnika (pojmov, ki jih uporablja naročnik za opisovanje poslovnih procesov, programskih okolij itd.), kar se lahko doseže preko intervjujev, obstoječe naročnikove dokumentacije ali npr. z opazovanjem zaslonskih mask dotičnih poslovnih sistemov.

Sledi priprava **seznama vseh mest, kjer se nekakovostni podatki pojavljajo** in nastajajo. To zajema popis vse od nivoja sistemov in programov do nivoja podatkovnih baz in posameznih tabel. Od intervjujev z naročnikom do izvedbe projekta je potrebno vzpostaviti matriko od splošnih poslovnih zahtev do podrobnega seznama nekakovostnih podatkov: kaj so in kje so. Primer takšne matrike je lahko v naslednji obliki: poslovni izraz (podatki o kupcu) → podatkovno področje/skupina (naziv podjetja, kontaktni podatki, naslov za dostavo) → zapis podatka (kratek naziv podjetja, poln naziv podjetja, ime kontakta, priimek kontakta, telefonska številka kontakta, ulica za dostavo, pošta in kraj za dostavo, država dostave).

Pri iskanju lokacij, kjer se nekakovostni podatki pojavljajo, ni vedno potrebno pregledovati celotnih podatkovnih zbirk, temveč si lahko pomagamo z vzorčenjem in tako močno zmanjšamo problemski prostor. Če že vnaprej poznamo nekaj problematičnih podatkov, jim lahko sledimo po sistemu od izvora do ponora in tako gradimo seznam potencialnih lokacij, ki jih je potrebno analizirati.

### 3.4.3 Faza 3: ovrednotenje kakovosti podatkov in načrtovanje izboljšav

**Ovrednotenje kakovosti podatkov** je prvi korak v tretji fazi projekta. Zajema dejansko analizo stanja kakovosti podatkov po vseh dimenzijah, dogovorjenih v drugi fazi projekta. Predlagan nabor dimenzij v magistrskem delu (glej poglavje 2.2 Dimenzije kakovosti podatkov) je lahko preobsežen in nepotreben za vsak projekt, zato se morata naročnik in izvajalec dogovoriti, katere dimenzije so relevantne in bodo vključene v projekt.

V naslednjem koraku sledi **načrtovanje izboljšav** kakovosti podatkov. Pomembno je, da se ovrednoteno stanje kakovosti podatkov, pridobljeno v prejšnjem koraku, koristno uporabi tako, da se pripravi akcijski načrt, kako se bo kakovost podatkov izboljšala in tako bo ustvarjen pozitiven poslovni učinek. Pri načrtovanju izboljšav lahko posamezne akcije razvrstimo po prioriteti glede na izbrane kriterije (npr. glede na razmerje korist/strošek).

Načrt naj obsega tri področja: preprečevanje nekakovostnih podatkov, popravki nekakovostnih podatkov in organiziranje komunikacije ter akcije med posameznimi udeleženci.

### 3.4.4 Faza 4: izboljšanje kakovosti podatkov

Dejanski popravki problematičnih podatkov in zagotavljanje kakovosti v prihodnje se izvede v zadnji, četrti fazi projekta. Razlog za to je v tem, da drugačna rešitev ni smiselna. Namreč, če bi se najprej lotili izboljševanja kakovosti podatkov še pred prvo fazo identifikacije, bi zelo verjetno kmalu ugotovili, da niti ne vemo, kakšne anomalije v podatkih pravzaprav iščemo, ali pa bi potratili veliko časa za popravljanje podatkov, ki nimajo bistvenega vpliva na poslovne zahteve. Ravno tako ni smiselno popravljati podatkov vse dokler nismo prepričani, da so podatkovne zbirke ustrezno zavarovane, da ne bodo že v naslednjem trenutku nekakovostni podatki znova vstopili v sistem in bi bilo potrebno podatke čistiti ponovno.

Iz teh razlogov se v zadnji fazi projekta najprej predlaga **vpeljava kontrol za preprečevanje nekakovostnih podatkov** v sistemu, ki zajema dve vrsti aktivnosti:

- vpeljava tehničnih kontrol nad podatki, ki so lahko preproste ali izjemno kompleksne; potrebno je upoštevati vse dimenzije podatkov, torej gre za vzpostavitev strožjih podatkovnih struktur in pravil v programski opremi ter v podatkovnih bazah, vzpostavitev boljšega podatkovnega modela, varnostnih ukrepov, zaščito zasebnih podatkov oseb itd.
- trajna zagotovitev kakovostnih podatkov s pomočjo vsebinskih (procesnih) sprememb v organizaciji, kar je moč doseči z zajezitvijo in nadzorom nad viri nekakovostnih podatkov. V praksi to lahko pomeni zamenjavo ročnih vnosov podatkov z avtomatiziranimi uvozi, standardiziranje obdelav podatkov, vpeljavo sistematiziranega upravljanja podatkov, vpeljavo dobrih praks s področja varnosti in zasebnosti podatkov, povečanje zavedanja in odgovornosti za kakovost podatkov ipd.

Nazadnje se izvede še zaključni korak: **popravljanje nekakovostnih podatkov**, kot je bilo načrtovano glede na ovrednoteno stanje in poslovne zahteve. Popravljanje je možno izvesti na več načinov:

- z ročnimi popravki v poslovni programski opremi (najbolj varen način, a primeren le za manjši obseg podatkov),

- z množičnimi popravki neposredno v podatkovni bazi (primerno za velik obseg podatkov, a zelo tvegano, saj s takšnimi posegi lahko zaobidemo interne omejitve in prožilce poslovne aplikacije, zato lahko pride do nekonsistentnih podatkov),
- z uporabo specializirane programske opreme za čiščenje podatkov (prednost je ponovna uporaba standardiziranih postopkov ali dobrih praks, vendar so takšna orodja lahko draga in neučinkovita v posameznih primerih oziroma programih) ali
- z uporabo namensko izdelane programske opreme za ureditev konkretnih težav (praviloma je to najbolj varen in hiter način, vendar tudi najdražji).

Predlagan model kakovosti podatkov izkorišča najboljše prakse, kot jih predlagajo že zgoraj predstavljeni modeli, predvsem poudarek na vrstnem redu posameznih faz projektov in uporabo zdaj že večinoma standardiziranih dimenzij podatkov. Bistvene prednosti nadgrajenega modela pa so v dodatnih dimenzijah kakovosti podatkov (varnost in zasebnost), kar zahteva uporaba računalniških podatkovnih baz v sodobnem času, poleg tega pa model poudarja tesno povezavo projektov kakovosti podatkov z uresničevanjem poslovnih ciljev.

# 4 TEHNIKE ZA ZAGOTAVLJANJE KAKOVOSTI PODATKOV

Kakovost podatkov lahko zagotovimo z uporabo različnih tehnik, nekaj jih bo opisanih v nadaljevanju. V splošnem lahko korake, ki so pri tem potrebni, razvrstimo na tri procese:

- opredelitev kakovosti podatkov,
- identificiranje nekakovostnih podatkov in
- popravljanje nekakovostnih podatkov.

Za vsakega od teh korakov lahko uporabimo kopico metod in tehnik. Prvi in tretji korak sta bila že predstavljena v prejšnjih poglavjih (glej poglavji 2. in 3.4.4 Faza 4: izboljšanje kakovosti podatkov). Identificiranje nekakovostnih podatkov oziroma problematičnih zapisov je pravzaprav poglavje zase, na kratko ga lahko opišemo z metodami iskanja, kot so:

- statistične (npr. iskanje ekstremnih vrednosti z uporabo povprečij, deviacij itd.),
- gručenje (npr. iskanje ekstremnih primerov glede na evklidske razdalje),
- temelječe na vzorcih (iskanje primerov, ki ne ustrezajo predvidenim oblikam – za opredelitev vzorcev uporabimo regularne izraze (angl. *regular expressions*),
- asociativna pravila (npr. iskanje osamelih zapisov, ki so šibko povezani v skupino).

## 4.1 Zagotavljanje točnosti podatkov

Točnost podatkov je pojem, pod katerim si večina uporabnikov predstavlja kakovost podatkov. Tej dimenziji kakovosti podatkov je namenjene tudi največ literature (Eppler, 2003; Batini & Scannapieca, 2006; Al-Hakim, 2007; Chapman, 2005; Guillet & Hamilton, 2007; English, 2003).

Točnost opredelimo kot stopnjo ujemanja podatkov z dejanskim stanjem (Eppler, 2003, str. 69). Preverjanje točnosti je pogosto težko meriti z avtomatiziranim postopkom, saj presoja točnosti zahteva primerjanje podatka z dejstvom iz resničnega sveta. Nasprotje točnosti je napačnost, ki jo v podatkih lahko vsaj delno identificiramo z iskanjem "čudnih", torej zelo odstopajočih vrednosti.

Iskanje potencialno netočnih podatkov lahko poteka s pregledovanjem vrednosti, ki najbolj odstopajo od povprečja. Pri tem se izkaže, da se z enačbo (2), ki je pravzaprav definicija natančnosti (DeGroot, 2004, str. 42), nazorno **izpostavi predvsem ekstremne vrednosti**. Natančnost opredelimo z razpršenostjo zapisov (enačba (2)) posameznega dejstva, torej kako velik je odklon več meritev ceteris paribus:

$$Natančnost = \frac{1}{\sigma^2} \tag{2}$$

Pri čemer je varianca izračunana kot prikazuje enačba (3):

$$\sigma^2 = \sum_{i=1}^{n} p_i \cdot (x_i - \bar{x})^2 \tag{3}$$

kjer so $x_i$ posamezne meritve in $p_i$ njihove verjetnosti.

Pri pregledovanju kakovosti podatkov na ta način najdemo tiste podatke, ki so najbolj "čudni", torej najbolj izstopajoči – prav tisti, ki imajo običajno največje negativne posledice pri njihovi uporabi. Omenimo, da popolne natančnosti v praksi največkrat ni mogoče zagotoviti, kjer pa je to vseeno potrebno, je potrebno izračun natančnosti zaščititi, da ne pride do deljenja z nič.

V programskem jeziku T-SQL bi tako natančnost lahko izračunali z uporabo vgrajenih funkcij `1/VAR(x)` (za izračun nad vzorcem) ali `1/VARP(x)` (za izračun nad celotno populacijo).

## 4.2 Zagotavljanje veljavnosti podatkov

Točnost podatkov smo opredelili v semantičnem smislu, zato za popolnejšo sliko kakovosti podatkov na veljavnost podatkov gledamo sintaktično. S preverjanjem veljavnosti zapisov (npr. pravilna oblika e-poštnih naslovov, datumov, telefonskih številk ipd.) je moč poiskati vse neveljavne zapise, kjer je veljavnost moč opredeliti. Kjer pa to ni mogoče, si lahko pomagamo z analizo oblike nizov (angl. *pattern matching*) ali podobnim iskanjem neobičajnih vrednosti. Veljavnost podatkov je moč zagotoviti z uporabo domenskih omejitev (glej 4.4 Zagotavljanje integritete podatkov).

## 4.3 Zagotavljanje unikatnosti podatkov

Ponovljene zapise podatkov, ki predstavljajo eno samo dejstvo iz resničnega sveta, imenujemo dvojniki. Pri iskanju dvojnikov se soočamo ne le s problemom iskanja in primerjanja vseh elementov v množici, temveč je zaradi nepravilnosti v podatkih pogosto potrebno uporabiti tehnike delnega (mehkega) ujemanja (angl. *fuzzy matching*).

Če bi iskali dvojnike s popolnim ujemanjem, bi spregledali vse zapise, ki se morda razlikujejo le v enem znaku ali v velikih začetnicah, vsebujejo presledke ali druge tipkarske napake. Zato se lahko poslužujemo načinov, ki primerjajo zapise na različne načine. V programskem jeziku T-SQL lahko uporabimo funkcijo `SOUNDEX(x)`, s katero lahko primerjamo dva zapisa glede na njun fonetični zapis. V praksi se algoritem te funkcije ne obnese prav dobro (npr. "Brighton" in "Bristol" imata enako vrednost funkcije), zato je njegova uporaba omejena.

Obstaja več precej kompleksnejših načinov za mehko ujemanje. Nekateri najbolj znani so:

- Levenshteinova razdalja (na voljo je tudi v orodju Microsoft SQL Server Integration Services), ki je opredeljena kot število potrebnih operacij nad posameznimi znaki, da prvi niz pretvorimo v drugega,
- Hammingova razdalja, ki je opredeljena kot število potrebnih zamenjav, da prvi niz pretvorimo v drugega,
- metrika Jaro Winkler, opredeljena je kot uteženo razmerje med ujemajočimi znaki, zamenjavami in dolžino nizov; dobro se obnese npr. pri primerjavi imen in drugih kratkih nizov.

Zgornji algoritmi so pripravljeni tudi za uporabo v jeziku T-SQL preko zunanjih knjižnic (performančno zelo dober vir izvorne kode se nahaja na spletnem naslovu http://sourceforge.net/projects/simmetrics/, ki jo je nato potrebno pripraviti za uporabo na SQL strežniku).

Ne glede na način primerjave med zapisi je potrebna opredelitev, katere lastnosti (atributi) nekega zapisa so ključ, po katerem bo primerjava dejansko potekala. Primer: za iskanje dvojnih zapisov kupcev ne moremo primerjati le kombinacij "ime + priimek", saj lahko obstaja več oseb z enakim imenom in priimkom. Potrebno je npr. razširiti ključ vsaj še na "naslov", a tudi to ni vedno prava rešitev, saj ima neka oseba lahko več naslovov v svojem življenju. Opisan primer nakazuje na kompleksnost problema dvojnikov, ki ima lahko tako drage kot tudi časovno potratne posledice.

Izkaže se, da sodobna programska orodja (npr. Microsoft Dynamics CRM) problem rešujejo s sprotnim pregledovanjem unikatnosti vsakič, ko se zapisi spreminjajo, po drugi strani pa se preverjanje unikatnosti izvaja tudi ob uvozih novih zapisov in z občasnim pregledovanjem celotne baze (npr. vsako noč se požene opravilo za iskanje dvojnikov, ki morebitne spore popravi ali sporoči skrbniku podatkov).

## 4.4 Zagotavljanje integritete podatkov

Del integritete podatkov je tudi pravilnost podatkov, za potrebe magistrskega dela pa integriteto podatkov lahko opredelimo na višjem abstraktnem nivoju: ko zadostimo kriterijem pravilnosti same strukture podatkov, moramo zagotoviti tudi kakovost podatkovnega modela. Zato se pri zagotavljanju integritete podatkov predlaga analiza podatkovnega modela, predvsem analiza entitetnih, domenskih, referenčnih in drugih omejitev.

**Entitetne omejitve** zagotavljajo integriteto podatkov na nivoju posamezne vrstice v tabeli relacijske podatkovne baze. V praksi to predstavljajo glavni in unikatni ključi (v SQL-u so to `PRIMARY CONSTRAINTS, UNIQUE CONSTRAINTS`). Entitetne omejitve se vsebine zapisov neposredno ne tičejo, skrbijo le za to, da je vsak zapis mogoče učinkovito poiskati.

**Domenske omejitve** zagotavljajo integriteto podatkov na nivoju posameznih stolpcev v tabelah. Iz tega stališča lahko torej povzamemo, da domenske omejitve najbolj neposredno skrbijo za pravilnost podatkov. V praksi to dosežemo s strogim določevanjem: podatkovnih tipov, zaloge vrednosti, privzete vrednosti ipd. (v SQL-u so to DEFAULT CONSTRAINTS, NULL CONSTRAINTS, RULES).

**Referenčne omejitve** skrbijo za integriteto podatkov med posameznimi, povezanimi tabelami. Z njimi dosežemo to, da ohranjamo razmerja med posameznimi poslovnimi entitetami, npr. zapisi računov morajo imeti ustrezne referenčne omejitve, da lahko na računu prikažemo podatke o plačniku, prejemniku in artiklih (v SQL-u to dosežemo z uporabo tujih ključev, `FOREIGN KEYS`).

Pogosto se kot dobra praksa izkaže poslovna pravila vgraditi že na nivoju podatkovnega modela, kar lahko dosežemo z "uporabniškimi omejitvami", ki jih lahko razumemo kot kompleksne, specializirane omejitve. Vendar se je potrebno zavedati, da je prenos poslovne logike na nivo podatkovne baze včasih preveč omejujoč, saj s tem poslovna pravila postanejo sicer zavarovana in neobhodna, a zelo toga. V časih, ko se podjetja osredotočajo na visoko prilagodljivost svojih poslovnih procesov glede na spremembe na trgu, takšna praksa zahteva tehten premislek.

S podatkovnim modelom moramo zadostiti naslednjim kriterijem (Simsion & Witt, 2005, str. 480-483):

- popolnost oziroma pokritost shranjenih podatkov – ali shranjujemo vsa potrebna dejstva, potrebna za poslovne potrebe,
- nereduntantnost – ali je en podatek zapisan natanko enkrat in so anomalije zaradi spreminjanj podatkov onemogočene,
- uveljavljanje in izvrševanje poslovnih pravil – kršenje poslovnih pravil na nivoju podatkovnega modela ima lahko izjemno negativen vpliv na programsko opremo, ki dostopa do podatkov,
- stabilnost in skalabilnost – kako zahtevno je podatkovni model spreminjati glede na spremenjene poslovne potrebe,
- uporabnost podatkov – kako enostavno so podatki dostopni,
- umestitev v obstoječe informacijsko okolje – kako skladni so različni podatkovni modeli v celotnem informacijskem okolju.

## 4.5 Zagotavljanje konsistentnosti podatkov

Pri uporabi podatkovnih baz je praviloma potrebno zagotoviti, da v izbranem trenutku vsi uporabniki vidijo enake podatke, in to ne glede na programsko opremo, ki jo uporabljajo. Z drugimi besedami, to pomeni, da med podatki ni (logičnih) nasprotovanj: vsi podatki skupaj so celovita in usklajena podatkovna baza. Problem zagotavljanja konsistentnosti se izkaže za še kompleksnejšega, če je podatkovna baza porazdeljena. Zaradi performančnih ali zaradi varnostnih razlogov so namreč podatki lahko zapisani na več mestih hkrati – v takih primerih

mora obstajati sistem za usklajevanje (oziroma sinhronizacijo) podatkov po vseh lokacijah. Pri tem največkrat velja, da če se podatki ne spreminjajo (ali vsaj ne zelo pogosto), se vse kopije podatkov sinhronizirajo ob določenih časovnih intervalih, kar tehnično ni težko izvedljivo. V primeru pogostih sprememb podatkov na poljubni lokaciji, pa bi se lahko zgodilo, da bi nek uporabnik prebral podatek x=1, nekdo drug pa bi trenutek prej na neki drugi lokaciji podatek spremenil v x=2. Vsak od uporabnikov bi tako imel drugačen pogled na podatek x, kar bi v skrajnem primeru lahko imelo zelo negativne posledice. Zato je potrebno vzpostaviti nek mehanizem za učinkovito sinhronizacijo podatkov v celotnem informacijskem sistemu, in sicer na tak način, da se ohranja konsistentnost podatkov vsaj do neke (vnaprej znane) mere (Simsion & Witt, 2005).

V splošnem lahko konsistentnost podatkov zagotavljamo po več arhitekturnih nivojih, od fizičnega zapisovanja podatkov na pomnilniške medije (npr. v Windowsih ukaz `chkdsk`), do organiziranja podatkov znotraj podatkovne baze (v SQL-u za to uporabljamo ukaze vrste `DBCC CHECKDB`) in do najvišjega nivoja, to je konsistentna predstavitev podatkov do uporabnika. Za potrebe magistrskega dela bo jedrnato predstavljen le najvišji nivo.
Podatke, zapisane na več lokacijah, med seboj usklajujemo na enega izmed treh logičnih načinov (uporabljeni izrazi veljajo za okolje Microsoft SQL Server, vendar je princip v splošnem enak tudi v drugih okoljih):

- replikacija posnetkov (angl. *snapshot replication*), kjer se med posameznimi lokacijami izmenjujejo natančne kopije – primerno za redko spremenljive podatke;
- transakcijska replikacija (angl. *transaction replication*), kjer se spremembe podatkov po nekih časovnih vrstah izvršijo po vseh lokacijah podatkov – primerno za hitro širjenje vseh sprememb po korakih po vseh lokacijah;
- replikacija s spajanjem (angl. *merge replication*), kjer se spremembe podatkov dogajajo po vseh lokacijah in je potreben mehanizem za njihovo usklajevanje – primerno za okolja, kjer se na vsaki lokaciji obdelujejo večinoma le posamezni deli zbirke podatkov in je z nekim zamikom spremembe potrebno spojiti s preostalimi deli podatkovne zbirke, pri čemer je potrebno zagotoviti mehanizem za reševanje konfliktnih primerov.

## 4.6 Varovanje podatkov

Varnost podatkov je potrebno zagotavljati večnivojsko, saj velja, da je veriga varnostnih mehanizmov močna natanko toliko, kot njen najšibkejši člen. Podatke lahko varujemo:

- s fizičnim varovanjem dostopa do strežnikov (npr. tako imenovane "varne sobe"),

- z elektronskim varovanjem dostopa do podatkov do poljubne granularnosti: omejitev dostopa do strežnika (`SQL LOGIN`), do posamezne podatkovne baze (`SQL USER`), do posamezne tabele (`SQL PERMISSIONS` ali `SQL ROLES`), posameznega zapisa (v okolju Oracle to lahko dosežemo z uporabo `VIRTUAL PRIVATE DATABASES`, v okolju Microsoft SQL Server pa je postopek kompleksnejši) itd.

Poleg samega omejevanja dostopov do podatkov se v praksi izkaže za uporabno tudi sledenje uporabi podatkov (angl. *data and security audit*), s čimer je moč nadzorovati učinkovitost trenutnih varnostnih nastavitev sistema. To lahko dosežemo na več načinov: z analizo dnevnika transakcij nad podatki (angl. *transaction log analysis*), z analizo dnevnika dogodkov na strežniku (angl. *event log analysis*), z uporabo posebnih procedur za sledenje dostopov (npr. preko uporabe prožilcev (angl. *triggers*) in namenskih tabel za sledenje (angl. *audit tables*)).

Pri sledenju uporabe podatkov, in tudi sicer, se je v vsakem trenutku potrebno zavedati omejitev, ki jih zahteva varovanje osebnih podatkov uporabnikov. Kovačič (2006, str. 91) pravi takole: "Kot že rečeno, sodobna informacijsko-komunikacijska tehnologija omogoča in tudi povzroča številne posege v zasebnost, saj je pogosto že zasnovana za nadzor. Vendar pa tehnologija posameznikom tudi omogoča, da se izognejo nadzorovanju. A podobno kot pri primerjavi Benthamove ideje Panoptikona in načela publicitete v delovanju politične skupščine, lahko tudi pri tehnologiji ugotovimo, da je ta večinoma uporabljena za nadzor posameznikov, ne pa toliko za njihovo zaščito pred nadzorom. Uporabo tehnologij, ki onemogočajo nadzor, skušajo države in njeni represivni organi sistematično omejevati, zato je ta tehnologija dostopna le manjšemu številu posameznikov, ki si uporabo teh tehnologij uspejo izboriti."

Zasebnost lahko ohranjamo s prikrivanjem ali šifriranjem osebnih podatkov uporabnikov (npr. podatke o imenih spremenimo v neprepoznavne), vendar moramo natančno poznati kateri podatki so opredeljeni kot osebni in nato natančno opredeliti pravilen in varen dostop do njih. Primer: če želimo v SQL-u zaščititi osebne podatke v tabeli `Kupci`, je eden od načinov uporaba pogledov (`SQL VIEW`). Postopek je tak, da vsem uporabnikom onemogočimo neposredni dostop do tabele, nato pa pripravimo ustrezen pogled nad to tabelo, v katerem vse osebne podatke prešifriramo (npr. z uporabo `HASHBYTES`), da so neprepoznavni. Uporabniki lahko tako dostopajo do tabele `Kupci` le preko tega pogleda, sicer pa ne.

# 5 UPORABA MODELA NAD PRIMEROM CRM SISTEMA

## 5.1 Opis CRM sistema

CRM sistem z izmišljenim imenom CRMVision, ki bo v magistrskem delu uporabljen za uporabo predloga nadgrajenega modela kakovosti podatkov, je tipičen primer sistema za upravljanje s strankami s področja trženja ter prodaje farmacevtskih izdelkov in spremljanje terenske prodaje za srednje velika podjetja.

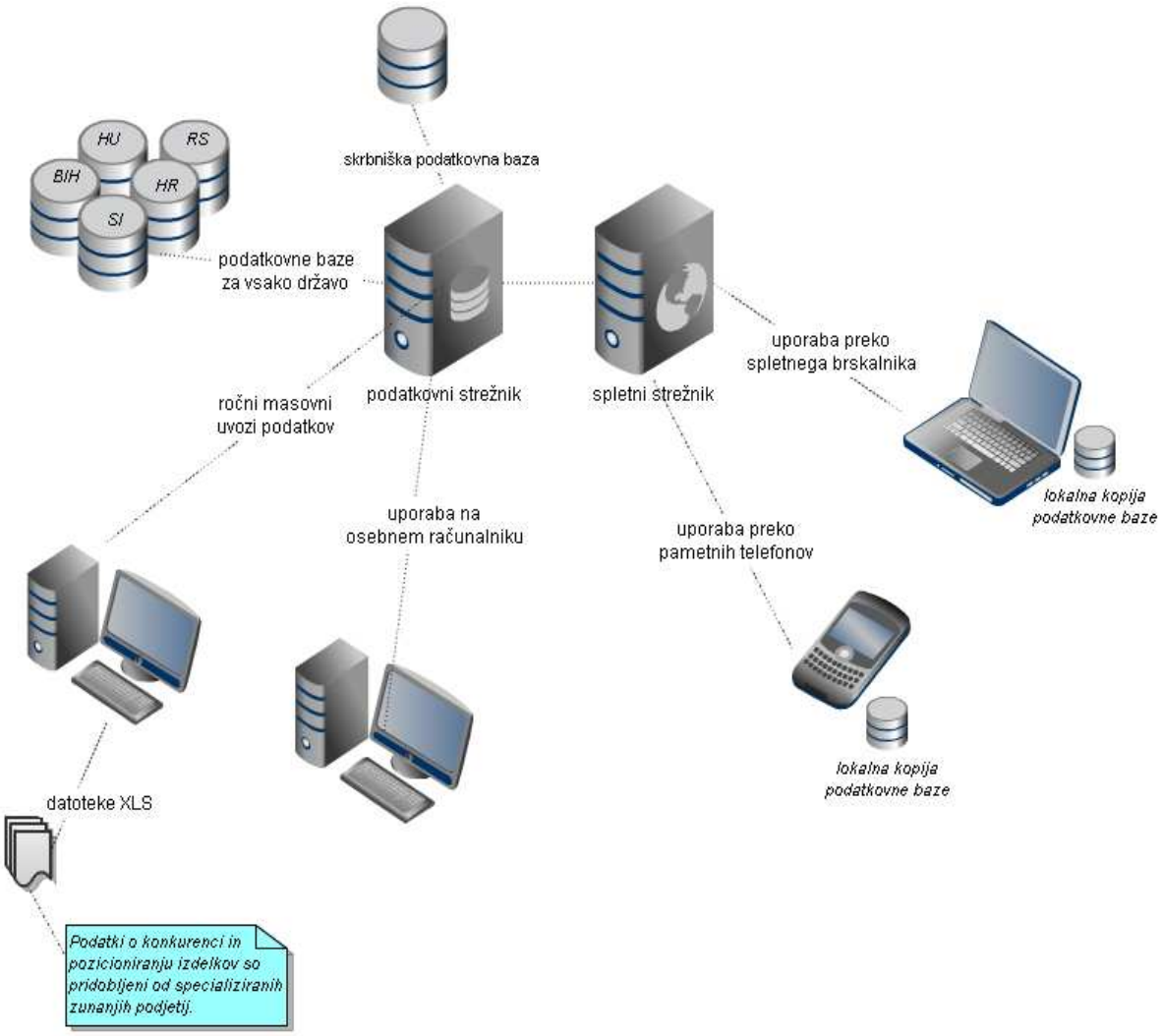
Glavni moduli sistema CRMVision omogočajo:

- samodejno in ročno zbiranje, obdelovanje in pregledovanje trženjskih podatkov o strankah in izdelkih,
- pripravo in analizo trženjskih akcij,
- spremljanje konkurence, pozicioniranje izdelkov in
- elektronsko podprto spremljanje terenske prodaje.

Tehnične lastnosti sistema CRMVision so predstavljene na Sliki 11, oziroma opisane na kratko v nadaljevanju:

- sistem temelji na dvonivojski arhitekturi odjemalec-strežnik,
- uporabljata se dve vrsti odjemalcev: uporaba preko programa na osebnem računalniku ali uporaba preko spletne aplikacije (ta je omejena na brskalnik Microsoft Internet Explorer),
- posamezno število uporabnikov po državah se giblje med 10 in 100,
- uporaba sistema je mogoča v povezanem (angl. *online*) ali nepovezanem (angl. *offline*) načinu, pri katerem se pri uporabniku vzpostavi lokalna kopija podatkovne baze, ki se ob preklopu v povezan način sinhronizira z matično podatkovno bazo, in
- podprto je lokalizirano delo v več državah.

*Slika 11: Diagram arhitekture sistema CRMVision*




Programsko orodje CRMVision je bilo izbrano za testno uporabo modela kakovosti podatkov zaradi velikega števila najraznovrstnejših zapletov, na katere uporabniki programa pogosto naletijo. Kljub temu, da je cenovni rang tega sistema od nekaj deset tisoč do sto tisoč evrov, ima nekaj tisoč uporabnikov in ga razvijajo že več kot desetletje, ga pestijo mnoge razmeroma hude težave, ki so odraz nepoznavanja ali neupoštevanja katere koli dimenzije kakovosti podatkov. En izmed takšnih problemov je izjemno nestabilen podatkovni model, zaradi katerega prihaja do anomalij podatkov v sistemu in celo do nedelovanja programa, kar ima za posledico visoke mesečne stroške vzdrževanja, težave pri analizi podatkov, praktično nemogočo integracijo sistema v preostalo programsko okolje organizacije in mnoge druge težave.

## 5.2 Uporaba predloga nadgrajenega modela kakovosti podatkov

### 5.2.1 Izvedba 1. faze: identifikacija ključnih dejavnikov projekta

#### 5.2.1.1 Opredelitev glavnega motiva projekta

Program CRMVision ima pomembno vlogo pri razporejanju virov in osredotočanju na poslovne priložnosti v organizaciji – ne le v matični državi, Sloveniji, temveč tudi po štirih sosednjih državah. Neučinkovito razporejanje človeških virov ima lahko zelo negativne posledice na poslovanje, saj se lahko s tem neko geografsko področje prenasiči s ponudbo, druga potencialno dobra področja pa tako ostanejo le slabo izkoriščena. To je bilo opredeljeno kot glavni motiv, da je potrebno zagotoviti pravilno delovanja programa, kakovostne podatke ter povečati zaupanje uporabnikov programa.

#### 5.2.1.2 Identificiranje razlogov za ureditev stanja

Podjetje se je soočalo z velikimi težavami pri pregledovanju učinkovitosti terenske prodaje in njenem upravljanju. Ni bilo jasno, ali analize, ki jih zaposlenci izvajajo v programu CRMVision, izkazujejo prave informacije, ali pa se z obdelavami in pregledovanjem podatkov v programu večinoma le trati čas. Zaupanje v program je bilo izjemno nizko, zaradi pogostih programskih napak tudi povsem upravičeno, in zato je bilo čutiti močan odpor uporabnikov do posvečanja energije v skrbno in natančno vnašanje podatkov v bazo, kar je pravzaprav pogoj za kakovostne podatke.

Velik delež zaposlenih v podjetju opravlja aktivnosti na terenu, a je zaradi same narave dela zaposlence dokaj zapleteno nadzorovati oziroma meriti njihovo učinkovitost. Na daljša obdobja se da uspešnost posameznikov sicer nazorneje razbrati iz spremljanja prodaje, na krajša obdobja pa to ni enostavno. Zaradi velike rasti podjetja je bil torej velik problem ugotavljati, ali novinci dejansko opravljajo delo na terenu ali ne. Na internih sestankih se je pogosto pojavljalo vprašanje, koliko so zaposlenci v resnici aktivni na terenu in koliko se le izgovarjajo na nedelovanje programa.

#### 5.2.1.3 Jedrnat opis problemov zaradi nekakovostnih podatkov

Z večmesečnim zbiranjem in urejanjem težav, ki so jih uporabniki javljali službi za informatiko v podjetju, ter z analizo poslovnih potreb, ki so jih glede nepravilnih in včasih povsem neuporabnih podatkov izrazili ključni uporabniki programa, je bil izdelan seznam

problemov, s katerimi so se začrtali okvirji projekta. Nekaj problemov iz seznama sledi v nadaljevanju:

- napake pri vpisovanju dnevnih aktivnosti na terenu (program ne pusti vpisovati novih aktivnosti),
- napake pri vpisovanju novih stikov (program javlja napako, da stik že obstaja),
- iskanje po obstoječih ustanovah v sistemu ne vrne pravih zadetkov,
- zelo počasno pregledovanje mesečnih podatkov,
- razmeroma pogosto nedelovanje programa (napake v podatkih onemogočajo uporabo programa za vse uporabnike neke države),
- uporabniki imajo nameščene zelo različne verzije programa (ni sistematičnega posodabljanja programske opreme), zato imajo različni uporabniki lahko različne funkcionalnosti v programu,
- varnostna politika glede uporabe programa praktično ne obstaja (ker gesla včasih delujejo, včasih pa jih program zavrne, si uporabniki med seboj izmenjujejo uporabniška imena in gesla),
- pogosto potrebno drago posredovanje dobavitelja prgorama, da program sploh deluje,
- masovni uvozi podatkov vedno povzročijo težave v programu in
- starih podatkov nihče ne arhivira, zato povzročajo dodatno zmedo v programu.

### 5.2.2 Izvedba 2. faze: opredelitev ključnih sestavin projekta

#### 5.2.2.1 Opredelitev poslovnih problemov in seznam ciljev projekta

Preko pogovorov s ključnimi uporabniki programa so bili ugotovljene glavne težave, ki jih imajo uporabniki zaradi nekakovostnih podatkov. Na podlagi seznama problemov so bili zastavljeni cilji, ki jih mora projekt za zagotavljanje kakovosti podatkov uresničiti.

Ključni problemi z negativnim učinkom na poslovanje podjetja, razvrščeni po prioriteti od najvišje do najnižje, si sledijo:

- neučinkovito razporejanje terenskih prodajalcev in neoptimalno osredotočanje na regije in ustanove zaradi izvajanja analiz nad nekakovostnimi podatki,
- nemogoča celovita analiza strank zaradi težavne integracije podatkov iz programa CRMVision z ostalimi zbirkami podatkov,
- neredno vpisovanje podatkov o obiskih na terenu zaradi nedelovanja programa (pogosta opozorila o podvojenih zapisih v programu),
- potrata časa zaradi počasnega delovanja programa.

Izhajajoč iz poslovnih problemov zaradi nekakovostnih podatkov so izpeljani cilji projekta, ki naj te probleme rešijo oziroma kar se da zmanjšajo:

- izločitev ali spojitev podvojenih zapisov,
- vgradnja mehanizma za zaznavanje podvojenih zapisov,
- očiščenje in dopolnitev podatkov,
- vzpostavitev kontrol za preprečevanje pomanjkljivih ali nepravilnih podatkov pri vnašanju v podatkovno bazo,
- s pomočjo SQL pogledov vzpostaviti nov podatkovni nivo, ki bo omogočal vsaj enosmerno (izhodno) integracijo podatkov iz CRMVisiona do drugih programov in
- performančna optimizacija podatkovne baze (do dopustnih omejitev programa).

#### 5.2.2.2 Opredelitev kakovosti podatkov

**Opredelitev točnosti podatkov**: z vidika najpomembnejših poslovnih entitet, katerim je potrebno zagotoviti razmeroma visoko točnost, je potrebno identificirati in po potrebi urediti vse izstopajoče zapise. Pri tem velja, da je zapis izstopajoč, če se bodisi njegova vrednost bodisi oblika pojavlja le redko. Redkost je opredeljena različno, odvisno od posameznega polja v tabeli, ki ga preučujemo, primeri redkosti pa so npr.: odstopanje od povprečne vrednosti za več kot tri standardne odklone, dolžina zapisa za več kot 50 odstotkov različna od najpogostejše dolžine in podobno.

**Opredelitev veljavnosti podatkov**: izbranim poljem v tabelah je potrebno določiti veljavno obliko zapisov. Primeri:

- e-poštni naslov mora ustrezati izrazu (angl. *regular expression*):
  `[A-Z0-9._-]+@[A-Z0-9.-]+\.[A-Z]{2,4}`,
- naslovi spletnih strani morajo ustrezati izrazu:
  `^((http|https|ftp):\/\/(www\.)?|www\.)[a-zA-Z0-9\_\-]+\.([a-zA-Z]{2,4}|[a-zA-Z]{2}\.[a-z-A-Z]{2})$`,
- imena in priimki oseb morajo biti dolgi vsaj 2 in največ 15 znakov, vsebujejo lahko le črke, presledke, pike in apostrofe:
  `^[a-zA-Z'.\s]{2,15}$`
- rojstni dnevi morajo biti v obdobju preteklih 200 let itd.

**Opredelitev integritete podatkov**: najpomembnejšim poljem v tabelah je potrebno določiti omejitve, ki bodo na nivoju podatkovne baze zagotavljale dobro obliko podatkov. Dodajanje novih entitetnih omejitev v obliki glavnih ključev ali unikatnih omejitev ni mogoče zaradi

posledic, ki bi jih to imelo na delovanje programa. Prav tako je zaradi nezmožnosti spreminjanja programa nemogoče spreminjati referenčne omejitve, npr. tuje ključe, saj bi s tem lahko povzročili nedelovanje programa. Tako kot edina možnost vplivanja na kakovost integritete podatkov ostajajo domenske omejitve: `DEFAULT, NULL, CHECK, RULE`. Z njimi je potrebno vpeljati privzete vrednosti najpomembnejšim poljem, jim določiti ali so obvezna, ter vpeljati poslovna pravila, kjer je to mogoče. Primeri poslovnih pravil so npr. standardizirana uporabniška imena uporabnikov programa, pravilo za minimalno dolžino opisa, s katerim se kratko zapiše potek in uspešnost posameznega obiska stranke.

**Opredelitev konsistence podatkov**: potrebno je zagotoviti, da vsi uporabniki vidijo enake podatke, upoštevajoč le razlike v pooblastilih. Izhajajoč iz izkušenj z uporabo programa CRMVision, je konsistenca podatkov tesno povezana z naslednjimi dejstvi:

- uporabniki zaradi neusklajenosti verzij programa dostopajo do različnih podatkov na različne načine, zato je potrebna sistematizacija in poenotenje posodabljanja verzij programa pri vseh uporabnikih,
- program se uporablja po več državah, vsaki je pripisana ločena podatkovna baza, a posamezne baze se razlikujejo ne le v vsebini, temveč tudi v podatkovnem modelu, zato je heterogenost podatkovnega modela potrebno upoštevati pri analizah podatkov.

**Opredelitev unikatnih zapisov**: zaradi narave programa (upravljanje odnosov s strankami), je z vidika zagotavljanja unikatnih zapisov najpomembneje, da je prav seznam stikov (oziroma oseb) kar se da kakovosten in ne vsebuje podvojenih zapisov.

S pomočjo analize različnih zapisov, objavljene na spletni strani SQL Server Central (Boroša, 2009), po posameznih poljih v tabeli stikov (`PERSON`), prikazane na Sliki 12 so bila določena polja, ki opredeljujejo unikaten zapis stika v podatkovni bazi: priimek (`LNAME`), ime (`FNAME`), naslov (`STREET1`).

*Slika 12: Analiza različnih zapisov po poljih tabele stikov*

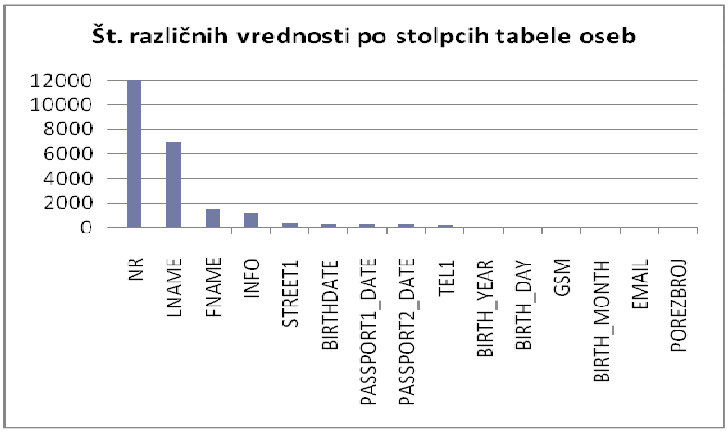


Tabela stikov (`PERSON`) obsega 71 različnih polj, vendar večina polj ni izpolnjena za vsak stik in je prazna ali irelevantna za opredelitev edinstvenosti nekega stika. Iz tega razloga so bila v opredelitev vključena polja ime, priimek in naslov. Niso bila vključena druga polja, ki bi sicer tudi lahko bila razlikovalna med stiki: polje `NR` je samodejno generirana številka stika, na katero uporabnik ne more vplivat, polje `INFO` vsebuje le opombe stika, polje `BIRTHDATE` ima premalo vrednosti, zato se očitno ne vpisuje redno itd.

**Opredelitev varovanja podatkov in zasebnosti**: na podlagi analize dostopov do podatkov s pomočjo sledenja SQL poizvedbam (angl. *SQL tracing*) je potrebno preveriti uporabo programa po posameznih uporabnikih. Zagotoviti je potrebno:

- uporabo varnih gesel uporabnikov (dolžina gesel naj bo vsaj 6 znakov, od katerih je vsaj 1 črka in vsaj 1 številka),
- med poslovnimi uporabniki določiti skrbnike, ki bodo upravljali s pravicami uporabnikov, upoštevajoč vnaprej dogovorjeno matriko pravic,
- varne komunikacijske povezave (dostop le znotraj lokalnega omrežja oziroma preko navideznega zasebnega omrežja (angl. *virtual private network* oziroma VPN tunela) in
- iz poročil umakniti osebne podatke, kjer niso nujno potrebni (npr. umakniti priimke iz poročil, kjer za analizo niso relevantni).

#### 5.2.2.3 Seznam mest pojavljanja nekakovostnih podatkov

**Opredelitev poslovnih entitet, katerim je potrebno izboljšati kakovost podpirajočih podatkov:** Iz seznama poslovnih problemov, ki je bil izdelan v predhodnem koraku, je moč izluščiti glavne poslovne entitete, ki jih je potrebno analizirati in jim zagotoviti kar največjo kakovost podatkov, ki jih opisujejo. To so:

- osebe (stranke, potencialne stranke in drugi udeleženci poslovnih procesov),
- ustanove (kjer so osebe zaposlene ali kjer se prodajajo artikli),
- artikli (izdelki, ki se jih prodaja in promovira) in
- obiski (dogodek, ko prodajalec obišče neko osebo v ustanovi in predstavi nek artikel).

**Seznam podatkovnih baz, tabel, pogledov in polj, kjer je zaznana nekakovost ključnih podatkov:** Sistem CRMVision sestavlja 5 uporabniških podatkovnih baz in ena sistemska. Zaradi podobnosti med uporabniškimi bazami bo iskanje lokacij nekakovostnih podatkov omejeno na eno podatkovno bazo, s podatki za državo Slovenijo.
Seznam tabel je pridobljen na podlagi analize strukture podatkovnega modela in analize vsebine tabel, opisane na spletni strani SQL Server Central (Boroša, 2009). Delni seznam

tabel je prikazan v Tabeli 5. Prikazanih je le približno 20 tabel od vseh 274 tabel, od katerih jih približno 200 vsebuje nič ali le eno vrstico in so kot take nepomembne s stališča kakovosti podatkov, obstajati pa morajo zaradi zagotavljanja delovanja programske opreme CRMVision. Modro obarvane vrstice so zanimive za ocenjevanje kakovosti podatkov, najpomembnejše so označene z ★.

*Tabela 5: Analiza najpomembnejših lokacij z nekakovostnimi podatki*

| Tabela | Št. vrstic | Rezervirano (kB) | Podatki (%) | Indeksi (%) | Št. odvisnih objektov | Prioriteta | *Opomba* |
|---|---|---|---|---|---|---|---|
| **SALES_INST_DATA** | 3.027.008 | 2488624 | 37 | 33 | 1 | | *podatki o prodaji ustanovam* |
| ~~**KUPDATELOG**~~ | ~~1.674.300~~ | ~~969192~~ | ~~28~~ | ~~45~~ | ~~0~~ | | *povezovalna tabela, večinoma ključi povezanih tabel* |
| **PERS_SREP_WORK** | 413.838 | 420312 | 22 | 43 | 23 | | povezovalna tabela (večinoma ključi povezanih tabel – zaradi št. odvisnih objektov zanimiva za analizo) |
| **VISITPRODUCT** | 328.530 | 208576 | 35 | 65 | 1 | | *podaki o predstavljenih izdelkih za posamezen obisk osebe v ustanovi* |
| **VISITPERS** | 201.439 | 263240 | 39 | 61 | 4 | | *podatki o obisku osebe v ustanovi* |
| **VISPERSPLAN** | 78.048 | 82552 | 29 | 45 | 1 | | *podatki o načrtovanih obiskih oseb* |
| **PRODFORM_PRICE** | 56.861 | 16488 | 38 | 61 | 0 | | *podatki o cenikih artiklov* |
| ~~**DAYREP_CODE**~~ | ~~56.852~~ | ~~13080~~ | ~~47~~ | ~~52~~ | ~~1~~ | | *povezovalna tabela dnevnih poročil o potovanjih (upoštevajoč vsebino: večinoma ključi povezanih tabel)* |
| **DAY_REPORT** | 48.446 | 25504 | 58 | 41 | 3 | | *podatki o potovanjih* |
| **REP_TEXT** | 24.747 | 4352 | 49 | 46 | 0 | | *podatki lokalizacije programa* |
| ~~**PERSPRODSTAT**~~ | ~~15.616~~ | ~~8944~~ | ~~41~~ | ~~58~~ | ~~1~~ | | *povezovalna tabela (upoštevajoč vsebino: večinoma ključi povezanih tabel)* |
| ~~**VISDAYPLAN_CODE**~~ | ~~14.885~~ | ~~4184~~ | ~~44~~ | ~~51~~ | ~~1~~ | | *povezovalna tabela (upoštevajoč vsebino: večinoma ključi povezanih tabel)* |
| **VISDAYPLAN** | 14.543 | 2832 | 69 | 30 | 2 | | *podatki o načrtovanih obiskih oseb v ustanovah* |
| **PERSON** | 12.121 | 15440 | 39 | 61 | 1 | ★ | *podatki o osebah* |
| **VISINSTPLAN** | 7.618 | 2560 | 38 | 53 | 1 | | *podatki o načrtovanih obiskih osebe v ustanovi* |
| **INST** | 4.523 | 3792 | 48 | 41 | 3 | ★ | *podatki o ustanovah* |
| **SALES_LOOKUP_PROD** | 3.760 | 1472 | 38 | 54 | 0 | | *povezovalna tabela artiklov med ustanovami (polje Status dvoumno)* |
| ~~**REP_COLUMNS**~~ | ~~2.969~~ | ~~384~~ | ~~79~~ | ~~21~~ | ~~0~~ | | *tabela z opredelitvami polj podatkovne baze (baza v bazi!)* |
| **SALES_LOOKUP_HOSP** | 2.965 | 1440 | 41 | 46 | 0 | | *podatki o ustanovah: nazivi, naslovi (nenormalizacija!)* |
| ~~**QRYDISPLAY**~~ | ~~1.831~~ | ~~448~~ | ~~39~~ | ~~46~~ | ~~0~~ | | *podatki o poizvedbah, ki jih izvaja program (baza v bazi!)* |
| **VISITINST** | 1.794 | 2320 | 29 | 53 | 7 | ★ | *podatki o obisku ustanove* |
| ~~**QRYSELCOND**~~ | ~~1.246~~ | ~~904~~ | ~~69~~ | ~~18~~ | ~~0~~ | | *podatki o poizvedbah, ki jih izvaja program (baza v bazi!)* |
| **PRODUCT** | 1.182 | 1064 | 32 | 62 | 0 | ★ | *artikli* |

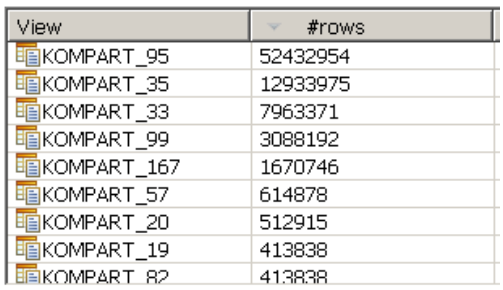
Pri pripravljanju seznama tabel so najbolj relevantne naslednje značilnosti:

- ime tabele (angl. *table name*),
- število vrstic v tabeli (angl. *rows count*),
- odstotek zasedenosti rezerviranega prostora tabele s podatki (angl. *data percent of total reserved space*),
- odstotek zasedenosti rezerviranega prostora tabele z indeksi (angl. *indexes percent of total reserved space*),
- število odvisnih objektov v podatkovni bazi (angl. *dependants count*).

Dobro poimenovanje tabel že samo po sebi pove veliko o tem, kakšni podatki so v tabeli, zato so imena tabel v vsakem podatkovnem modelu zelo pomembna. Že na podlagi imen tabel lahko nabor potencialnih kandidatov za projekt kakovosti podatkov precej zožamo. Število vrstic v tabeli nam pove, ali gre za npr. tabelo referenčnih podatkov (kot npr. šifranti) ali pa prometno tabelo, v kateri se shranjujejo transakcije ipd. S stališča kakovosti podatkov so pomembne vse tabele, vendar se je velikih tabel potrebno lotevati drugače (npr. z analizo nad vzorcem, ne nad celotno populacijo). Odstotek, ki ga predstavljajo podatki, in odstotek, ki ga predstavljajo indeksi, sta pomembna s stališča, da so morda tabele lahko majhne po številu vrstic, vendar če imajo po drugi strani veliko pripetih indeksov, to pomeni določeno obremenitev podatkovnega modela. Sicer oba odstotka skupaj ne predstavljata nujno celotnega prostora, ki ga zavzema tabela, saj je del prostora lahko še nezaseden. Zadnji pomemben podatek je število objektov, ki so od posamezne tabele odvisni: naj si bo preko tujih ključev, povezanih pogledov, prožilcev, ali kako drugače. To veliko pove o tem, kako "popularna" je tabela v podatkovnem modelu, torej kako močno je vpeta med ostale tabele.

Analiza SQL pogledov je pokazala, da obstaja 83 različnih pogledov, del njih je prikazanih na Sliki 13. Približno tretjina jih ne vrne nobenega zapisa, nekaj jih zajema več milijonov vrstic (slaba struktura pogledov), vsi pa imajo enako ime, razlikujejo se le v številki na koncu imena, iz česar je moč sklepati, da gre za poglede izključno za namene uporabe v programu. Upoštevajoč to ugotovitev, SQL pogledi ne bodo zajeti v projektu kakovosti podatkov.

*Slika 13: Del seznama SQL pogledov in št. vrstic v vsakem pogledu*



| View | #rows |
|---|---|
| KOMPART_95 | 52432954 |
| KOMPART_35 | 12933975 |
| KOMPART_33 | 7963371 |
| KOMPART_99 | 3088192 |
| KOMPART_167 | 1670746 |
| KOMPART_57 | 614878 |
| KOMPART_20 | 512915 |
| KOMPART_19 | 413838 |
| KOMPART_82 | 413838 |

Podrobnejša analiza tabel zajema seznam in opredelitev relevantnih polj v posameznih tabelah. V Tabeli 6 je prikazan primer za tabelo oseb (PERSON), v kateri so zbrani podatki o osebah in je glede na opisane poslovne zahteve v prvi fazi projekta tudi najpomembnejša.

*Tabela 6: Seznam pomembnih polj v tabeli oseb (najpomembnejša so označena z ★)*

| Ime polja | Podatkovni tip | Št. različnih vrednosti | Prioriteta | Opomba |
|---|---|---|---|---|
| ~~**NR**~~ | ~~int~~ | ~~12.100~~ | | *številka osebe, spremembe niso mogoče* |
| **LNAME** | nvarchar | 6.913 | ★ | *priimek* |
| **FNAME** | nvarchar | 1.473 | ★ | *ime* |
| **INFO** | nvarchar | 1.134 | | *opombe, lastnosti osebe* |
| **STREET1** | nvarchar | 389 | ★ | *naslov* |
| **BIRTHDATE** | datetime | 280 | ★ | *rojstni datum* |
| ~~**PASSPORT1_DATE**~~ | ~~datetime~~ | ~~257~~ | | *neuporabljana, nepomembna funkcionalnost* |
| ~~**PASSPORT2_DATE**~~ | ~~datetime~~ | ~~257~~ | | *neuporabljana, nepomembna funkcionalnost* |
| **TEL1** | nvarchar | 202 | ★ | *telefonska številka* |
| **BIRTH_YEAR** | smallint | 45 | | *leto rojstva* |
| **BIRTH_DAY** | smallint | 32 | | *dan rojstva* |
| **GSM** | nvarchar | 25 | | *mobilna telefonska številka* |
| **BIRTH_MONTH** | smallint | 13 | | *mesec rojstva* |
| **EMAIL** | nvarchar | 13 | ★ | *e-poštni naslov* |
| **POREZBROJ** | nvarchar | 6 | | *davčna številka* |

### 5.2.3 Izvedba 3. faze: ovrednotenje kakovosti podatkov in načrtovanje izboljšav

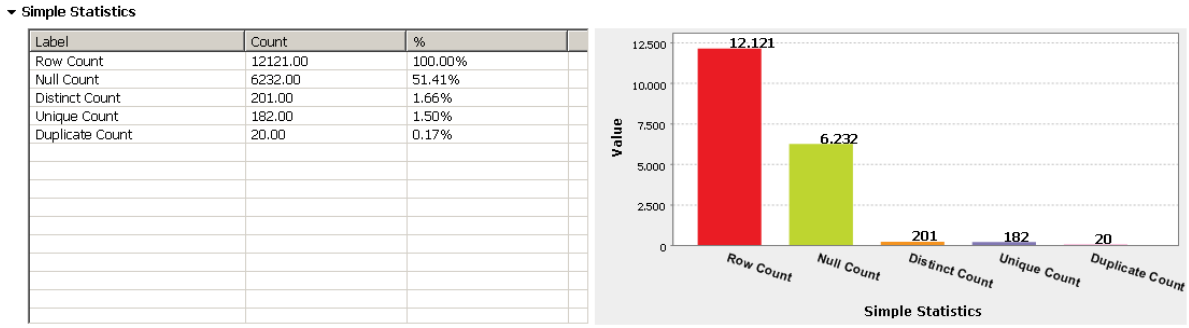
#### 5.2.3.1 Ovrednotenje kakovosti podatkov

Kakovost podatkov je bila ovrednotena upoštevajoč ugotovitve predhodnih korakov projekta. Opredeljene poslovne probleme je bilo potrebno rešiti s pomočjo izbranih dimenzij kakovosti podatkov, po vseh identificiranih lokacijah, kjer se le-ti pojavljajo. Preden je znano, kako rešiti poslovne probleme, ki nastopajo kot posledica nekakovostnih podatkov, je potrebna podrobna ocena stanja in načrt izvedbe. Ovrednotenje kakovosti podatkov je potekalo po predhodno izbranih dimenzijah kakovosti. Za namen predstavitve uporabe modela kakovosti podatkov v praksi bo v nadaljevanju poglavja opisano ovrednotenje kakovosti le na eni lokaciji za vsako dimenzijo.

**Ovrednotenje točnosti podatkov**

Točnost podatkov je bila ovrednotena s stališča izstopajočih zapisov. V nadaljevanju je opisana analiza tabele oseb (`PERSON`), konkretno polji telefonska številka (`TEL1`) in priimek (`LNAME`). Na Sliki 14 je razvidno, da je vseh zapisov v tabeli 12.121, vendar je od tega kar 6.232 nedefiniranih oziroma praznih. Zanimiva je razlika med različnimi zapisi (angl. *distinct count*) in unikatnimi zapisi (angl. *unique count*), pri čemer so unikatni zapisi opredeljeni kot različni zapisi s po le enim primerkom.

*Slika 14: Statistika telefonskih številk*



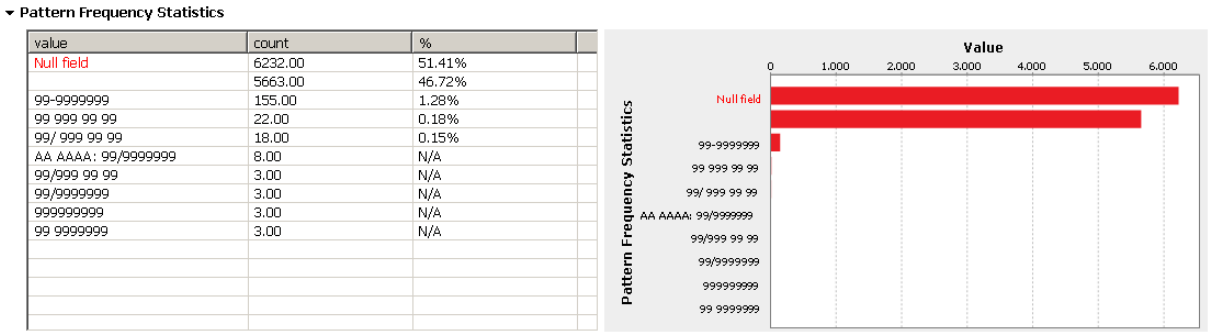
Simple Statistics

| Label | Count | % |
|---|---|---|
| Row Count | 12121.00 | 100.00% |
| Null Count | 6232.00 | 51.41% |
| Distinct Count | 201.00 | 1.66% |
| Unique Count | 182.00 | 1.50% |
| Duplicate Count | 20.00 | 0.17% |



Na Sliki 15 je moč opaziti, da sicer pri večini oseb telefonska številka ni zapisana. Pri tem je pomembna ugotovitev ta, da je nedefiniranih (angl. *null*) zapisov kar 6.232, poleg tega pa je 5.663 zapisov praznih. To nakazuje na nestandardiziran vnos telefonskih številk – včasih so manjkajoče številke vpisane kot prazen niz, drugič ostajajo nedefinirane, iz česar je mogoče sklepati, da je v nekem trenutku, morda z novo verzijo, program začel prazne številke zapisovati na drugačen način. To nakazuje na potrebno dodelavo podatkovnega modela z opredelitvijo `NULL` in `DEFAULT` omejitev. Po drugi strani lahko iz obstoječih telefonskih številk razberemo tudi prevladujočo obliko zapisov, kar je moč izrabiti za opredelitev veljavne oblike telefonskih številk v podatkovnem modelu z uporabo pravil veljavnih zapisov.

*Slika 15: Točnost podatkov o telefonskih številkah – opazna je prevladujoča oblika številk*



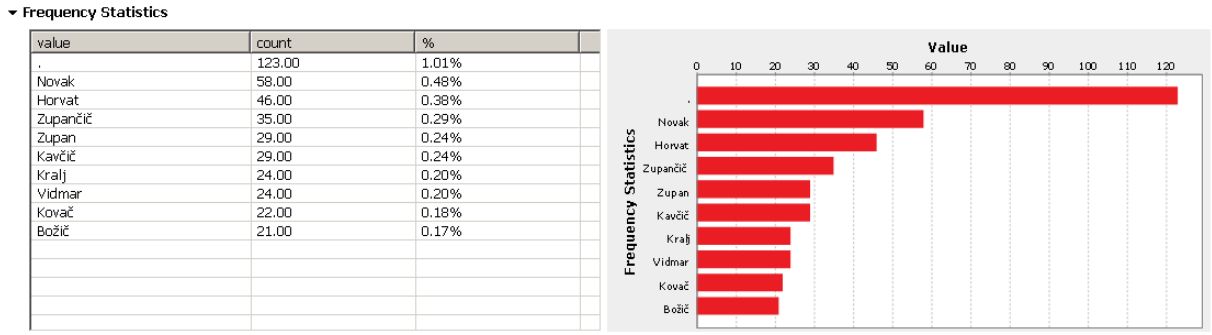
Pattern Frequency Statistics

| value | count | % |
|---|---|---|
| Null field | 6232.00 | 51.41% |
| | 5663.00 | 46.72% |
| 99-9999999 | 155.00 | 1.28% |
| 99 999 99 99 | 22.00 | 0.18% |
| 99/ 999 99 99 | 18.00 | 0.15% |
| AA AAAA: 99/9999999 | 8.00 | N/A |
| 99/999 99 99 | 3.00 | N/A |
| 99/9999999 | 3.00 | N/A |
| 999999999 | 3.00 | N/A |
| 99 9999999 | 3.00 | N/A |



Analiza priimkov je razkrila, da so uporabniki neznane priimke oseb označevali s piko ("."), kot prikazuje Slika 16. Tudi to bi bilo mogoče izkoristiti z dodelavo podatkovnega modela z uporabo `DEFAULT` omejitev. Ta ugotovitev je pomembna tudi pri vsebinskih analizah podatkov o osebah. V določenih primerih bi vse osebe s priimkom "." lahko označili kot neveljavne, saj takšne osebe ni mogoče nasloviti in so zato takšni zapisi neuporabni.

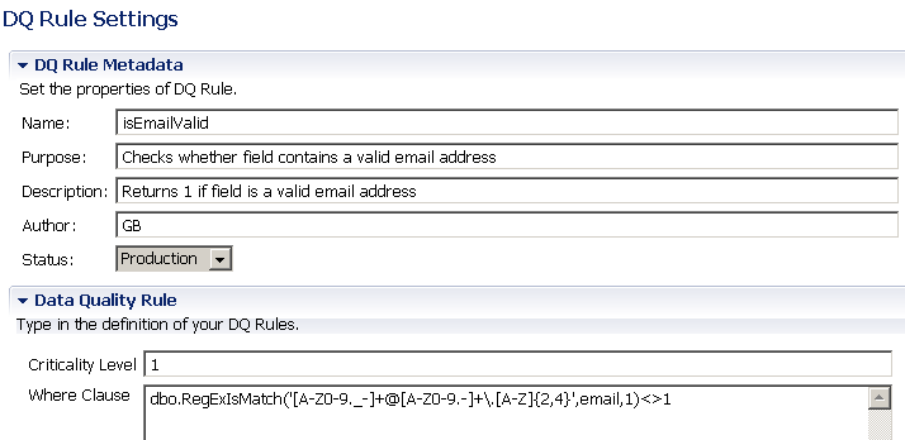
*Slika 16: Točnost priimkov – opazni so zapisi, pri katerih so neznani označein s "."*




## Ovrednotenje pravilnosti podatkov

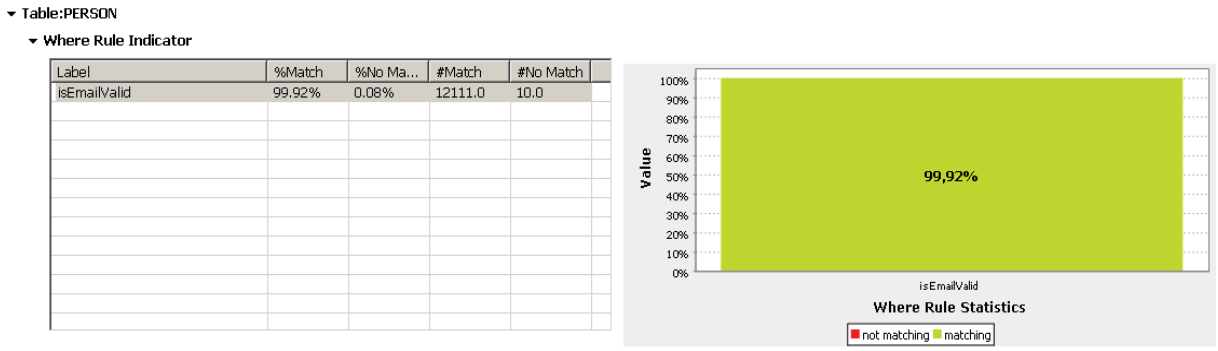
Pravilnost podatkov je bila glede na opredelitev v prejšnjem koraku izvedena s pomočjo uporabe pravil kakovosti podatkov v programu Talend Open Profiler. Primer pravilnosti e-poštnih naslovov prikazuje Slika 17.

*Slika 17: Pravilnost zapisov e-poštnih naslovov – nastavitev pravila*




Ovrednotenje pravilnosti e-poštnih zapisov je pokazalo praktično popolno pravilnost podatkov, kot prikazuje Slika 18, saj je le 0,08 % zapisov vsebovalo neveljaven naslov.

*Slika 18: Analiza veljavnosti e-poštnih naslovov*



**Ovrednotenje integritete podatkov**

V nadaljevanju je opisano vrednotenje integritete tabele oseb (`PERSON`), ki je bilo izvedeno s pomočjo analize tabele neposredno v podatkovni bazi.

Tabela oseb ima od 71 polj le eno polje opredeljeno kot obvezno polje. To je polje `PERS_SNR`, ki je tipa `uniqueidentifier`, in je označeno kot glavni ključ. Opredelitev obveznih podatkov je torej neusklajena s poslovnimi zahtevami, ki zahtevajo več obveznih podatkov. Poleg tega ima glavni ključ performančno neučinkovit podatkovni tip: vsak `uniqueidentifier` zavzema 16 bajtov pomnilnika in zaradi naključnega delovanja povzroča drobljenje (fragmentacijo) podatkov.

Tabela nima nobenih drugih omejitev, ki bi posredno lahko imeli pozitiven vpliv na podatke: ni privzetih vrednosti za nobeno polje in ni opredeljenih nobenih pravil glede vsebine podatkov. Nad tabelo je zgrajen en SQL pogled, ki povezuje osebe z regijo, vendar je slabo opredeljen in vrača več kot 400.000 vrstic, kar je veliko glede na dobrih 12.000 oseb. Na tabeli je opredeljenih tudi 19 preprostih indeksov s po enim poljem. Glede na velikost tabele je vzdrževanje tako velikega števila indeksov lahko performančno potratno.

**Ovrednotenje konsistence podatkov**

Najpomembnejši del analize konsistence podatkov je bil pregled vseh potencialnih ponornih točk, kjer se uporabljajo podatki iz programa CRMVision. Ugotovljeno je bilo, da v uporabi ni drugih orodij in da se vse analize podatkov izvajajo izključno v programu CRMVision. S tem je odpadla potreba po analizi podatkovnih tokov izven tega programa.

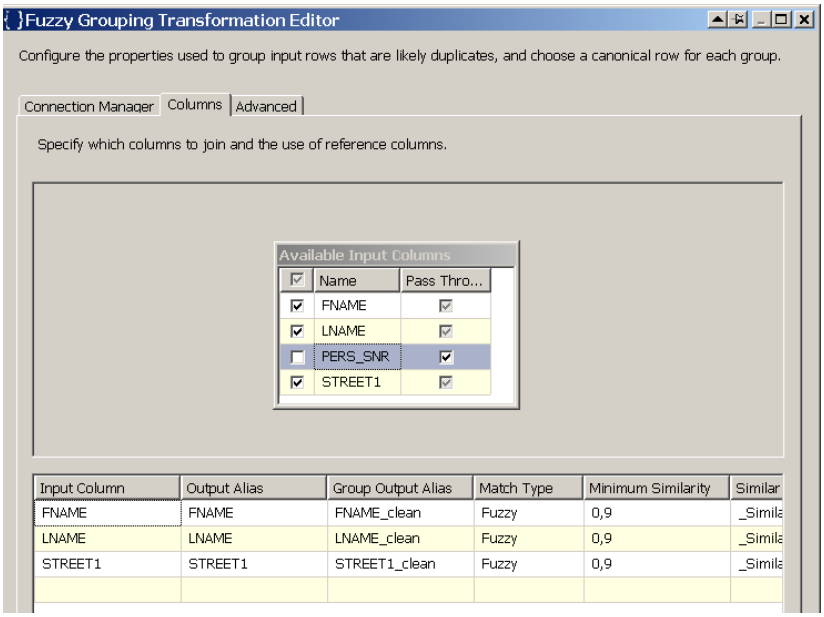
Drugi del vrednotenja konsistence podatkov je zajemal pregled verzij programa pri uporabnikih. Znano je bilo, da različne verzije programa delujejo bistveno drugače, kar je v določenih primerih povzročajo nekonsistentnost podatkov. Izhajajoč iz te ugotovitve se je naredil popis verzij programa v celotni organizaciji po vseh državah. Ugotovljeno je bilo, da je v vsakodnevni uporabi več verzij programa, nekateri uporabniki so uporabljali celo več let stare verzije. Dodatno je bilo ugotovljeno tudi to, da namestitve programa novim uporabnikom niso usklajene in standardizirane. Tako se je dogajalo, da so uporabniki v tujih državah dobili povsem druge verzije programa kot uporabniki v Sloveniji. Ta proces je bil označen kot kritičen in nujno potreben ureditve.

**Ovrednotenje unikatnih zapisov**

V nadaljevanju je predstavljena analiza unikatnih zapisov oseb, s poslovnega vidika opredeljenih kot množica {ime, priimek, naslov} oziroma v podatkovnem modelu označenih kot {`FNAME`, `LNAME`, `STREET1`}. Upoštevaje neizogibne tipkarske napake, ki jih

povzročajo ročni vnosi podatkov, je bilo pri iskanju dvojnikov uporabljeno mehko ujemanje podatkov z uporabo orodja Microsoft SQL Server Integration Services. Na Sliki 19 je na levi strani prikazana nastavitev ujemanja in na desni število dvojnikov, kot rezultat analize. Identificiranih je bilo 719 dvojnikov, ki jih bo v zaključni fazi potrebno spojiti.

*Slika 19: Analiza unikatnih zapisov – levo nastavitev unikatnosti, desno št. dvojnikov*




| | Identifikator | StDvojnikov |
|---|---|---|
| 1 | 3776 | 4 |
| 2 | 8982 | 4 |
| 3 | 5837 | 3 |
| 4 | 3864 | 3 |
| 5 | 7084 | 3 |
| 6 | 10828 | 3 |
| 7 | 8675 | 3 |
| 8 | 7049 | 3 |
| 9 | 9090 | 3 |
| 10 | 6285 | 3 |
| 11 | 10384 | 3 |
| 12 | 9946 | 3 |
| 13 | 6846 | 3 |
| 14 | 10696 | 3 |
| 15 | 10613 | 3 |
| 16 | 7977 | 3 |
| 17 | 1950 | 3 |
| 18 | 10566 | 3 |

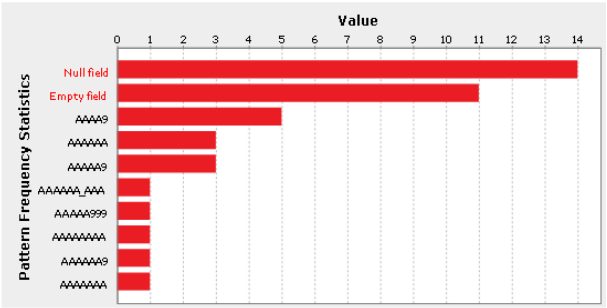
**Ovrednotenje varovanja podatkov in zasebnosti**

Prvi del analize varnosti podatkov je zajemal pregled zapisov uporabniških imen in gesel v programu. S pomočjo ročnega pregledovanja podatkovne baze in z uporabo programa SQL Profiler je bilo ugotovljeno, da so uporabniki zapisani v 4 različnih in nepovezanih tabelah. Gesla so kot prosto besedilo (angl. *plain text*) vpisana v treh od teh štirih tabelah. Ugotovljeno je bilo tudi nelogično spreminjanje gesel v programu, saj se gesla iz neznanih razlogov včasih zapišejo v eno, včasih v drugo, včasih pa v več tabel hkrati. Sama struktura gesel, kjer so že vpisana, je razmeroma zadovoljiva, saj kot prikazuje Slika 20, gesla vsebujejo v večini po 4 črke in 1 številko.

*Slika 20: Analiza oblike gesel uporabnikov*

Pattern Frequency Statistics

| value | count | % |
|---|---|---|
| Null field | 14.00 | 31.11% |
| Empty field | 11.00 | 24.44% |
| AAAA9 | 5.00 | 11.11% |
| AAAAAA | 3.00 | 6.67% |
| AAAAA9 | 3.00 | 6.67% |
| AAAAAA_AAA | 1.00 | 2.22% |
| AAAAA999 | 1.00 | 2.22% |
| AAAAAAAA | 1.00 | 2.22% |
| AAAAAA9 | 1.00 | 2.22% |
| AAAAAAA | 1.00 | 2.22% |



V nadaljevanju analize je bil pregledan dostop do programa izven lokalnega omrežja podjetja. Prenos podatkov je sicer potekal preko šifriranega SSL protokola, vendar se je vodstvo odločilo, da bo dostop potrebno omejiti z obvezno uporabo VPN tunela.

Nadalje se je s stališča varovanja osebnih podatkov pregledalo poročila, ki jih omogoča program. Analiza je pokazala, da so osebni podatki vidni na vseh poročilih, ki se vsaj posredno tičejo oseb in da kdorkoli ima dostop do programa, ima dostop do vseh podatkov ter poročil. Glede na različna opravila, ki jih izvajajo uporabniki programa, je bilo ugotovljeno, da je potrebno vzpostaviti model varnostnih vlog (angl. *security role*), kjer ima vsaka vloga omejen dostop do podatkov, posamezni uporabniki pa se razporedijo po skupinah v te vloge.

#### 5.2.3.2 Načrt izboljšanja kakovosti podatkov

Na podlagi zaznanih poslovnih problemov in opredelitve kakovosti podatkov iz druge faze projekta ter ovrednoteno kakovostjo iz prejšnjega koraka je bil izdelan načrt za izboljšanje kakovosti podatkov v sistemu. Ker je bilo pri samih izboljšavah potrebno sodelovanje večje množice uporabnikov, ne le članov projektne skupine, je bil poseben poudarek dan na sistematičnem komuniciranju: načrtovanju tedenskih sestankov o napredovanju aktivnosti, sprejete so bile predloge dokumentov za dokumentiranje posegov v sistem idr.

Načrt izboljšanja kakovosti podatkov je bil jedrnato predstavljen s pomočjo tabelaričnega prikaza, kot prikazuje Tabela 7 in so ga morali podpisati vsi sodelujoči na projektu, s čimer se je potrdilo razumevanje projekta in so uporabniki sprejeli odgovornost za svoje naloge.

*Tabela 7: Načrt izboljšanja kakovosti podatkov*

| Št. | Predlagan korak | Prioriteta | Opombe |
|---|---|---|---|
| 1 | Ureditev seznama stikov | ① | izločitev dvojnikov, očiščenje in oplemenitenje podatkov, arhiviranje starih zapisov |
| 2 | Omogočiti boljši vpogled v podatke | ② | vpeljava dodatnega nivoja podatkov preko SQL pogledov |
| 3 | Vpeljava kontrol podatkov | ③ | omejitve dopustnih podatkov na nivoju podatkovne baze |
| 4 | Izboljšanje internih procesov | ④ | izobraževanje uporabnikov o pravilni uporabi programa, sistematično posodabljanje verzij programa pri vseh uporabnikih |
| 5 | Zmanjšati odvisnost od dobavitelja programa | ⑤ | pritisk na dobavitelja, da v večji meri zagotovi delovanje programa in paketno, ne več posamično nameščanje posodobitev |

Prioriteti 1 (urediti stike) in 2 (izboljšanje analitičnih možnosti) sta bili najpomembnejša koraka projekta, imela sta tudi najbolj daljnosežen vpliv na podjetje, zato jima je bilo posvečeno največ pozornosti.

### 5.2.4 Izvedba 4. faze: izboljšanje kakovosti podatkov

#### 5.2.4.1 Vpeljava kontrol za preprečevanje nekakovostnih podatkov

**Tehnične kontrole nad podatki**

Z množico tehničnih mehanizmov so se vzpostavile kontrole podatkov na nivoju podatkovne baze. S tem so se postavila rigorozna pravila glede vnosov in sprememb podatkov. Kljub učinkovitosti takšnih kontrol v splošnem, pa so pri naknadnem postavljanju takšnih kontrol določene težave, ki jih ne bi bilo, če bi se kontrole postavljale ob samem kodiranju programa. Tako pa je obseg možnih sprememb podatkovnega modela močno zmanjšan, saj je potrebno ohraniti transparentnost sprememb do aplikacijskega nivoja programa.

Prva skupina tehničnih kontrol je zajemala dimenzijo pravilnosti podatkov. Uporabljena je bila knjižnica `RegExFunction`, objavljena na spletni strani Simple Talk (Factor, 2009), ki razširja funkcionalnost Microsoft SQL strežnika. Primer uporabe te knjižnice za zagotavljanje pravilnosti e-poštnih naslovov:

```
ALTER TABLE [PERSON]
ADD CONSTRAINT chkEmail CHECK
(dbo.RegExIsMatch('[A-Z0-9._-]+@[A-Z0-9.-]+\.[A-Z]{2,4}',[EMAIL],1)=1)
```

Druga skupina tehničnih kontrol je temeljila na zagotavljanju zadostnega podatkovnega pokritja in skladnosti podatkov. Zajemala je opredelitev obveznih podatkov, privzetih vrednosti in najmanjše dolžine nizov v izbranih poljih.

Primer takšnih kontrol na tabeli oseb:

```
--zagotovitev obveznega polja priimek (LNAME)
ALTER TABLE PERSON WITH NOCHECK
ALTER COLUMN [LNAME] nvarchar(100) NOT NULL
--zagotovitev privzete vrednosti polja, ali osebo lahko kontaktiramo
ALTER TABLE PERSON
ADD CONSTRAINT df_Person_Dontsendmail DEFAULT (0) FOR [DONTSENDMAIL]
```

Izmed ostalih tehničnih kontrol, ki so bile vpeljane tekom projekta, omenimo še nekaj varnostnih mehanizmov:

- opredelitev dopustne oblike gesel uporabnikov (z uporabo `RegExFunction),`
- oddaljeni dostop do programa le preko navideznega zasebnega omrežja – VPN,
- zanimiv je tudi preprost ukrep, ki se je izkazal za izjemno učinkovitega: onemogočenje dostopa dobavitelja programske opreme do strežnika, s čimer se je omejilo ad hoc programske spremembe in discipliniralo dobavitelja k bolj nadzorovanemu delu.

**Vsebinske, procesne spremembe, ki onemogočajo pojav nekakovostnih podatkov**

Tehnične kontrole nad podatki same po sebi še ne zagotavljajo kakovosti podatkov, vsaj ne v popolni meri. Kot je že razvidno iz Englishevega modela, opisanega v poglavju 3.2, je v praksi nujno vzpostaviti tudi vsebinske mehanizme, torej ustrezno okolje, ki bo implicitno zagotavljalo kakovostnejše podatke. To je tudi način za dolgoročnejšo uspešnost projektov. Izhajajoč iz te ugotovitve so bila v organizaciji sprejeta naslednja navodila oziroma:

- V vsaki državi je vzpostavljena matrika oseb, s katero je opredeljeno, katere osebe podatke vnašajo v program, katere osebe podatke obdeluje in spreminja ter katere osebe podatke pregleduje oziroma izvaja analize.
- Za glavne poslovne entitete, opisane v 2. fazi projekta, so sprejeta pravila za vnašanje podatkov v program (npr. katera polja so obvezna, kakšna je oblika polj, kako je potrebno beležiti posamezne dogodke ipd.).
- Vzpostavljen je nov, dodatni podatkovni nivo, ki omogoča preverjeno zanesljive analize. Uporabljene so vrtilne tabele v orodju Microsoft Excel in razmeroma kompleksni SQL pogledi, s katerimi se je delno obvladalo slab podatkovni model.
- Vpeljan je sistematičen postopek posodabljanja programske opreme pri uporabnikih. Zagotovljeno je, da je najnovejša verzija programa dostopna na datotečnem strežniku, starejše verzije se arhivirajo, posodabljanje programa pri uporabnikih pa poteka hkrati za vse uporabnike z uporabo skupinskih pravilnikov (angl. *group policy*).

Zgoraj opisana pravila so bila uporabnikom izčrpno skomunicirana, saj morajo vsako spremembo pri procesu vnašanja ali obdelave podatkov uporabniki sprejeti in se nanjo navaditi. Ko so uporabniki razumeli, kako pomembni so podatki, ki jih obdelujejo, za poslovanje, so dobili pozitivno motivacijo za kakovostnejše delo in s tem neposredno pripomogli k izboljšanju informacijske kulture v podjetju.

#### 5.2.4.2 Popravki slabe kakovosti podatkov

Dejanski korektivni ukrepi so zajemali naslednje aktivnosti v ustreznih tabelah in poljih:

- čiščenje podatkov:
  - uskladitev velikih in malih začetnic (pri imenih, priimkih, naslovih),
  - popravki e-poštnih naslovov, da ustrezajo pravilni obliki,
  - popravki nepravilnih sklicev na šifrant poštnih številk in krajev,
  - brisanje nerazpoznavnih nazivov oseb in ustanov ter njihovih naslovov,
  - arhiviranje oziroma deaktiviranje oseb, ki ne izpolnjujejo naslednjih pogojev:
    - v zadnjih dveh letih vsaj dve zapisani aktivnosti,
    - imajo vpisane vse obvezne podatke za enolično identifikacijo osebe,
    - obstaja podatek, kdaj je bila oseba vpisana (preprečevanje fantomskih oseb);
- dopolnitve podatkov:
  - ročna dopolnitev manjkajočih opredelitev spola pri osebah,
  - dopolnitev in popravki naslovov oseb z ročnim pregledovanje zapisov in konsolidacijo z vsemi ad hoc zapisanimi podatki uporabnikov (podatki v beležkah, v e-poštnih odjemalcih, ...);
- spojitve dvojnikov:
  - podvojene zapise, identificirane s pomočjo mehkega ujemanja po ustreznih poljih, delno spojili s pomočjo računalniške obdelave, delno pa ročno;
- varnostni ukrepi:
  - menjava gesel uporabnikov, skladno z varnostnimi pravili za dodeljevanje gesel,
  - izdelava poročila za pregledovanje aktivnosti uporabnikov,
  - umik priimkov oseb z vseh poročil in analiz, kjer je mogoče in le-ti niso nujno potrebni.

Posledica zgornjih ukrepov je bilo precej manjše število razpoložljivih (aktivnih) stikov oziroma oseb v programu. V nekaterih podatkovnih bazah (predvsem tistih za tuje države) se je število stikov zmanjšalo tudi za 90 odstotkov (iz več kot 20 tisoč oseb na manj kot tisoč), saj večinoma ni bilo jasno, od kje in zakaj so bili ti zapisi vpisani v bazo in jih torej ni bilo mogoče učinkovito uporabljati. Celo več, nepravilna uporaba kakršnih koli seznamov oseb je lahko škodljiva in tudi nezakonita. To sta bila dovolj tehtna razloga, da so ključni uporabniki programa sprejeli zgoraj opisane ukrepe. Nenazadnje je podjetje s tem pridobilo sicer manjšo bazo (potencialnih) strank, vendar mnogo bolj kakovostno.

# 6 UGOTOVITVE

Predlog nadgrajenega modela se je v uporabi izkazal za praktičnega in učinkovitega. V prvem delu projekta so se sodelujoči sporazumeli o težavah, s katerimi se srečujejo pri uporabi programa za upravljanje odnosov s strankami. Pri tem so spoznali pomen kakovosti podatkov za podjetje ter predvsem kako lahko kakovostnejši podatki pripomorejo k učinkovitejšemu poslovanju: manjšanju stroškov in sprejemanju pravilnejših odločitev.

Novi model sestavlja jasno zaporedje logičnih korakov, začenši z razumevanjem negativnih poslovnih učinkov, ki jih povzročajo nekakovostni podatki, in na podlagi tega postavljanje ciljev, ki jih mora udejaniti projekt, temelječ na opisanem modelu. Zavedati se je potrebno, da brez jasno zastavljenih ciljev zvodeni vsak projekt. Ker je zagotavljanje kakovosti podatkov stalen proces in ne le množica enkratnih aktivnosti, je bistveno zavedanje njegovega daljnosežnega pomena in pripravljenost na določene procesne spremembe v organizaciji.

V prvih dveh fazah model predlaga identifikacijo ključnih dejavnikov in sestavin projekta. Kakovost podatkov si je moč predstavljati na različne načine, zato je potrebno, da se naročnik in izvajalec sporazumeta o temeljnih pojmih. Poseben izziv je nato identificiranje vseh lokacij, kjer se nekakovostni podatki pojavljajo. Možna sta dva primera: če je poslovni proces dobro dokumentiran in podatkovni model znan, se je dejanski analizi strukture podatkov moč izogniti in takoj pričeti z izboljševanjem kakovosti podatkov. V nasprotnem primeru je potrebno skrbno pregledati podatkovne zbirke, v kakšnih poslovnih procesih se uporabljajo, kako so sestavljene ter kako povezane med seboj. To je lahko zelo zamudno, zato je moč uporabiti nekatere tehnike, opisane v poglavju 5.

V tretji fazi modela se je kakovost podatkov ovrednotila po vseh izbranih dimenzijah kakovosti, določenih v predhodni fazi. Za samo vrednotenje je nabor uporabljenih pristopov povsem odvisen od konkretne situacije in programskih orodij, ki so na voljo. S stališča učinkovitosti modela zato to ne predstavlja odločilnega kriterija – pomembneje je, da se, upoštevajoč identificirane poslovne probleme, opredelijo in izberejo ustrezne dimenzije kakovosti podatkov, ki bodo omogočile relevanten vidik problematike podatkov . Vrednotenju sledi izdelava načrta izboljšav, kar se je izkazalo za smiselno in ustrezno, saj je v trenutku po zaključenem ovrednotenju kakovosti zavedanje o stanju največje in je torej moč podati najučinkovitejša priporočila oziroma smernice za izboljšanje stanja.
Omeniti je potrebno, da nekaterih kontrol ni mogoče vpeljati pred popravki nekakovostnih podatkov, saj te kontrole vsiljujejo tudi obliko vseh že obstoječih podatkov. V takšnih

primerih se v tem koraku tehnične kontrole zgolj opredeli in pripravi za izvedbo. Nato se izvedejo korektivni ukrepi nad podatki in šele nato se vzpostavijo tehnične kontrole podatkov. V praksi se pogosto izkaže, da se oba koraka četrte faze projekta izvajata sočasno oziroma izmenjujoče. Tudi vpeljava tehničnih kontrol nad podatki ni vedno mogoča pred korektivnimi ukrepi. Razlog za to je v tem, da popravki lahko trajajo dalj časa. Dokler niso izvedeni, ni mogoče vzpostaviti določenih vrst tehničnih kontrol. Po drugi strani pa popravljanje nekakovostnih podatkov ni vedno smiselno izvajati v celoti, vse dokler niso zagotovljene tehnične kontrole, ki bi preprečevale, da se v vmesnem času v podatkovno bazo ne bi zopet prikradli nekakovostni podatki. Delno rešitev je mogoče doseči tako, da se tehnične kontrole nad podatki sicer vzpostavijo na podatkovni bazi, vendar na način, da ignorirajo obstoječe podatke in torej veljajo le za nove (obstoječe podatke je v tem primeru potrebno popraviti v naslednjem koraku). To je moč doseči z uporabo zastavice `NOCHECK`, npr.

```
ALTER TABLE <table_name> WITH NOCHECK ADD CONSTRAINT ...
```

Z uporabo predlaganega modela se je v podjetju ponovno vzpostavilo zaupanje v podatke v programu za upravljanje odnosov s strankami. Uporabnikom se je omogočilo pravilnejše vnašanje podatkov, z dodelavami podatkovnega modela pa je program postal stabilnejši. Najpomembnejša pozitivna posledica izvedbe projekta pa je bila zagotovitev verodostojnih poročil o delu terenskih prodajalcev in analizi prodaje. Novi uvidi v podatke, ki jih je model omogočil, so povzročili tudi nekaj posrednih, pozitivnih procesnih sprememb v podjetju. Izgovori uporabnikov o nedelovanju programa so se precej zmanjšali, zato bo zelo verjetno sčasoma prišlo do pozitivnih poslovnih učinkov.

Iz zgornjih dejstev lahko sklepamo, da je nadgrajeni model moč učinkovito uporabiti pri zagotavljanju kakovosti podatkov. Iz zaporedja posameznih faz se projekt zagotavljanja kakovosti podatkov izvede tako tehnično nedvoumno, kot uspešno s poslovnega vidika.

## SKLEP

Kakovost podatkov je v organizacijah bistvenega pomena, kar velja še posebej za podjetja v sodobnem času. Pogosto se govori, da zdaj živimo v informacijski družbi, kjer je pretok količine podatkov mnogokrat že kar nepredstavljiv. Že samo to dejstvo, podkrepljeno z znastvenimi raziskavami, nakazuje na problematiko, koliko so vsi ti podatki sploh resnični, uporabni in koristni To je vprašanje, ki si ga morajo zastavljati vsi lastniki in uporabniki podatkov. Glede na praktično popolno odvisnost podjetij od podatkov, je zdaj prav gotovo čas, da se raznovrstna zagotavljanja kakovosti v podjetjih razširijo tudi na podatke.

V preteklosti je bilo na temo zagotavljanja kakovosti podatkov razvitih več modelov, pri katerih je jasno čutiti prehajanje iz akademske sfere na poslovne aplikacije. Iz tega je mogoče pričakovati, da imajo sodobni modeli dovolj teoretične podlage in so zreli za uporabo v poslovnem svetu. V magistrskem delu sem podal predlog nadgrajenega modela za zagotavljanje kakovosti podatkov, katerega prednost je predvsem enostavna praktična uporaba. Obenem nadgrajeni model daje poudarek dvema dodatnima dimenzijama kakovosti, ki postajata z vse večjo vključenostjo interneta v življenja ljudi vse pomembnejši. To sta varnost in zasebnost podatkov, ki ju lahko razumemo tudi kot vzvod za ohranjanje dostojanstva in varnosti ljudi. Menim namreč, da informatika (in še posebej internet) samo po sebi teži k obravnavi ljudi kot številk, to pa nujno zahteva previdnost in varovanje pred zlorabami.

V praksi je opisanih že veliko primerov, ko so podjetja z izboljšanjem kakovosti podatkov dosegla velike pozitivne učinke na poslovanje. Pri tem je iz študije primerov v praksi opaziti, da so projekti za zagotavljanje kakovosti podatkov v veliki meri poslovno uspešni, kar v splošnem na področju informatike ni tako pogosto. Bolj kot tveganja pred stroškovno neupravičenostjo takšnih projektov je zaznati določen upor pri uporabnikih podatkov, saj kakovostni podatki nujno zahtevajo ustrezno informacijsko kulturo. To lahko pripišemo dejstvu, da si zavedanje o pomenu podatkov šele zdaj počasi vtira pot v razmišljanja uporabnikov računalnikov. V svetu je na voljo tudi vse več izobraževanj in drugih aktivnosti s tega področja. Uveljavljajo se celo posebej sistematizirana delovna mesta, namenjena izključno za zagotavljanje in kontrolo kakovosti podatkov. Sklepati je mogoče, da bo v prihodnosti to področje še pridobivalo na pomenu. Ker podatki v raznovrstnih podatkovnih zbirkah vse bolj neposredno vplivajo na življenja ljudi, je pomembno, da na podatke ne gledamo več le s tehničnega vidika, ki je sicer neizogibno potreben, temveč se zavedamo celotne vpetosti podatkov v poslovne procese podjetij.

Cilj magistrskega dela je sestaviti model za vzpostavitev in zagotavljanje kakovosti podatkov. Ta cilj je dosežen s predstavljenim modelom, ki je dovolj širok za splošno uporabo in vsebuje praktičen pristop k izboljševanju kakovosti podatkov. Predstavljeni model nadgrajuje že obstoječe modele in obenem izpostavlja dve aktualni dimenziji kakovosti, varnost in zasebnost, katerih obravnavo zahteva vsesplošna razširjenost spletnih omrežij.

# LITERATURA IN VIRI


1. Al-Hakim, L. (2007). *Information Quality Management*. Hershey: Idea Group Publishing.
2. Alter, S. (2002). *Information Systems: Foundation of E-Business*. New Jersey: Prentice Hall.
3. Batini, C., & Scannapieca, M. (2006). *Data Quality, Concepts, Methodologies and Techniques.* Berlin: Springer Verlag.
4. Beniger, R. J. (1986). *The Control Revolution*. London: Harvard University Press.
5. Bentley, W., & Davis, P. T. (2009). *Lean Six Sigma secrets for the CIO*. Kentucky: CRC Press.
6. Boroša, G. (2009). *Getting a clue about your databases.* Najdeno 1. decembra 2009 na spletnem naslovu http://www.sqlservercentral.com/articles/Data+Quality/65326/
7. Chapman, A.D. (2005). *Principles of Data Quality*. Copenhagen: Global Biodiversity Information Facility.
8. DeGroot, M. H. (2004). *Optimal statistical decisions*. New Jersey: Wiley Classics Library.
9. Eckerson. W. (2002). *Data quality and the bottom line*. Najdeno 20. decembra 2009 na spletnem naslovu http://adtmag.com/Articles/2002/05/01/Data-Warehousing-Special-Report-Data-quality-and-the-bottom-line.aspx?Page=1
10. English, L. (2003). Total Information Quality Management – A Complete Methodology for IQ Management. *International Association for Information and Data Quality*. Najdeno 1. februarja 2009 na spletnem naslovu http://www.iaidq.com/bookclub/doc/
11. English, L. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profit*. New York: John Wiley & Sons, Inc.
12. English, L. (1998). *DAMA awards recipients.* Najdeno 23. februarja 2009 na spletnem naslovu http://www.dama.org/i4a/pages/index.cfm?pageid=3379
13. Eppler, M. J. (2003). *Managing Information Quality.* Berlin: Springer Verlag.
14. Factor, Phil (2009): *CLR Assembly RegEx Functions for SQL Server*. Najdeno 10. februarja 2010 na spletnem naslovu http://www.simple-talk.com/sql/t-sql-programming/clr-assembly-regex-functions-for-sql-server-by-example/
15. Godnov, U. (2009). *Vpliv implementacije relacijskega podatkovnega modela na kakovost spodatkov v poslovnih informacijskih sistemih* (doktorska disertacija). Ljubljana: Ekonomska fakulteta.
16. Godnov, U., & Knaflič, A. (2008). Kakovost podatkov – iskanje celovite opredelitve. *27. Mednarodna konferenca o razvoju organizacijskih znanosti, (str. 1-7)*. Portorož: Fakulteta za organizacijske vede.

17. Guillet, F., & Hamilton, H. (2007). *Quality Measures in Data Mining*. Berlin: Springer Verlag.
18. Gula, R. (2009). Securing your data assets. *Source Conference*. Boston: Source Boston 2009.
19. Hayes, G. E., & Romig, H. G. (1977). *Modern Quality Control*. Wisconsin: Bruce.
20. Hubley, J. (2001). Data quality: The foundation for business intelligence. Najdeno 22. februarja 2009 na spletnem naslovu http://searchcrm.techtarget.com/news/interview/0,289202,sid11_gci754429,00.html#
21. *IBM, objava za medije.* Najdeno 20. marca 2010 na spletnem naslovu http://www-03.ibm.com/press/us/en/pressrelease/7561.wss
22. Informacijski pooblaščenec Republike Slovenije. Pogosta vprašanja. Najdeno 10. decembra 2009 na spletnem naslovu http://www.ip-rs.si/pogosta-vprasanja/varstvo-osebnih-podatkov
23. *International Organization for Standardization*. Najdeno 20. marca 2010 na spletnem naslovu http://www.iso.org/iso/iso_catalogue/management_standards/iso_9000_iso_14000/iso_9000_essentials.htm
24. Kovačič, M. (2006). *Nadzor in zasebnost v informacijski družbi*. Ljubljana: Univerza v Ljubljani, Fakulteta za družbene vede.
25. Kroenke, D. M. (1997). *Database Processing: Fundamentals, Design, and Implementation*. New Jersey: Prentice-Hall, Inc.
26. Kurose, J. F., & Ross, K. W. (2009). *Computer networking: A top-down approach*. New Jersey: Pearson Education, Inc.
27. Lindell, Y., & Pinkas, B. (2000). *Privacy preserving data mining*. Berlin: Springer Verlag.
28. Madsen, M. (2009). *The Role of Open Source in Data Integration*. Third Nature: Technology report. Najdeno 21. februarja 2009 na spletnem naslovu http://www.talend.com/library/reflibrary.php
29. Maletic, J. I., & Marcus, A. (2000). *Data Cleansing: Beyond Integrity Analysis*. Memphis: The University of Memphis.
30. Maydanchik, A. (2007). *Data Quality Assessment*. New Jersey: Technics Publications, LLC.
31. McGilvray, D. (2008). *Executing Data Quality Projects*. San Francisco: Morgan Kaufmann.
32. *Microsoft, objava za medije.* Najdeno 20. marca 2010 na spletnem naslovu http://www.microsoft.com/Presspass/press/2008/oct08/10-06BI08PR.mspx
33. Nair, S. (2007). *Mining log files for gold*. Najdeno 3. julija 2009 na spletnem naslovu http://www.isaca-washdc.org/pages/articles/article-aug2007-print.htm

34. Nuray-Turan, R. (2010). *Data cleaning publications grouped by conferences*. Najdeno 20. marca 2010 na spletnem naslovu http://www.ics.uci.edu/~rnuray/biblio_conf.html
35. Orr, K. (1997). *Data Quality and Systems Theory*. Topeka: The Ken Orr Institute.
36. Petković, M., & Jonker, W. (2007). Security, Privacy and Trust in Modern Data Management. Berlin: Springer Verlag.
37. Petrocelli, T. (2005). *Data protection and information lifecycle management*. New Jersey: Pearson Education, Inc.
38. Redman, T. C. (1996). *Data Quality for the Information Age*. Norwood, MA: Artech House.
39. Redman, T. C. (2004). Confronting Data Demons. *Six Sigma Forum Magazine*, (maj) 9-10.
40. Ryu, K., Park, J.-S., & Park, J.-H. (2006). A Data Quality Management Maturity Model. *ETRI Journal*, 28 (2), 191-204.
41. Zakon o gospodarskih družbah. *Uradni list RS* št. 42/2006, 10/2008-ZGD1A, 68/2008-ZGD-1B.
42. *Objava za medije [podjetja SAP]*. Najdeno 20. marca 2010 na spletnem naslovu http://www.sap.com/about/newsroom/news-releases/press.epx?pressid=8360:
43. Scannapieco, M., Pernici, B., & Pierce, E. (2003). *IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework.* Članek predstavljen na konferenci International Conference on Information Quality (ICIQ-02). Cambridge: MIT.
44. Schneier, B. (2009, 4. junij). Be careful when you come to put your trust in the clouds. *The Guardian*. Najdeno 21. junija 2009 na spletnem naslovu http://www.guardian.co.uk/technology/2009/jun/04/bruce-schneier-cloud-computing/print
45. Schneier, B. (2006, 6. marec). The future of privacy. *Schneier on security*. Najdeno 15. junija 2009 na spletnem naslovu http://www.schneier.com/blog/archives/2006/03/the_future_of_p.html
46. Shankaranarayanan, G., Wang, R. Y., & Ziad, M. (2000). *IP-MAP: Representing the Manufacture of an Information Product*. Članek predstavljen na konferenci Information Quality na Massachusetts Institute of Technology.
47. Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Labs Technical Journal*, 27, 379-423.
48. Simsion, G. C., & Witt, G. C. (2005). *Data modeling essentials*. San Francisco: Morgan Kaufmann.
49. *Snooping through the power socket*. BBC News. Najdeno 20. junija 2009 na spletnem naslovu http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/8147534.stm
50. Taulbee, S. M. (1996). *Implementing data quality systems in biomedical records*. Boca Raton: CRC Press.
51. Ustava RS. *Uradni list RS* št. 33/1991.

52. *Varstvo osebnih podatkov* (Informacijski pooblaščenec Republike Slovenije). Najdeno 30. aprila 2009 na spletnem naslovu http://www.ip-rs.si/pogosta-vprasanja/varstvo-osebnih-podatkov
53. Vidmar, T. (2002). *Informacijsko-komunikacijski sistem*. Ljubljana: Pasadena.
54. Wang, R., Allen, T., Harris, W., & Madnick, S. (2003). *An Information Product Approach for Total Information Awareness*. Članek predstavljen na IEEE Aerospace Conference. Texas: IEEE Aerospace.
55. Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, XII (4).
56. Whitaker, R. (1999). *The End of Privacy*. New York: The New Press.
57. Zetter, K. (2005). CardSystems' Data Left Unsecured. *Wired.* Najdeno 21. januarja 2009 na spletnem naslovu http://www.wired.com/science/discoveries/news/2005/06/67980